X-Ray Crystallography

“If a picture is worth a thousand words, then a macromolecular structure is priceless to a physical biochemist.” – *van Holde*

Questions:

1. How is an image formed? What is the difference between images by Kodak / Light Microscope / EM / X-ray / NMR ?
2. What is a Crystal? How are they obtained? Materials / Methods
3. What is a Crystal Lattice? - Lattice Constants / Space Groups / Asymmetric Unit
4. What are X-rays? How are they produced?
5. What is the Bragg Equation? What can we learn from it?
6. What do we measure experimentally? How?
7. Phase Problem: What is the “phase” part and what is the “magnitude” part?
8. How do we “solve” the phase problem? What does “solving” it get us?
9. How is a protein “model” obtained?
10. How do I read a “crystallographic” paper?
11. What tools are available to help me understand protein structures?

Diffraction: Scattering from “atoms”

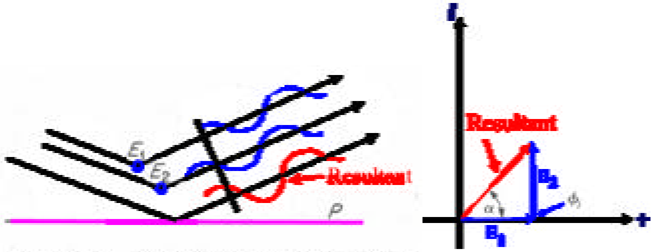


Figure 2.10. Diffraction from E_1 and E_2 as if reflected from plane P.

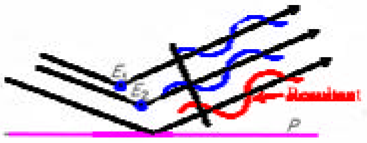
Scattering from “many atoms”

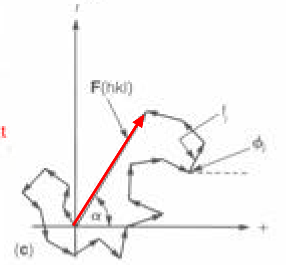
F(*hkl*) = SQRT [cI(*hkl*)]

Experimental

Calculated

$$\mathbf{F}(hkl) = F(hkl)e^{i\alpha(hkl)} = \sum_{j=1}^{N'} \mathbf{f}_j(hkl) = \sum_{j=1}^{N'} f_j(hkl)e^{i\phi_j(hkl)}$$

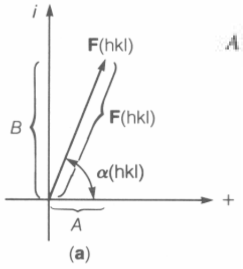




The structure factor for a reflection may be thought of as the vector sum of the x-ray scattering contributions from many atoms.

Each of the j contributions may be represented as a vector in the complex plane, with amplitude f_j and phase phi_j.

$$\mathbf{F} = A + iB$$



$$A = F\cos\alpha \quad \text{and} \quad B = F\sin\alpha$$

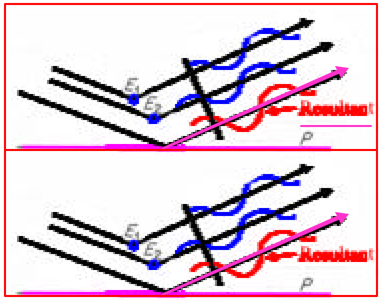
$$F = |\mathbf{F}| = \sqrt{A^2 + B^2} = \sqrt{\mathbf{FF^*}}$$

$$\alpha = \tan^{-1}\frac{B}{A}$$

The structure factor magnitude F(hk*l*) is represented by the length of a vector in the complex plane.

The phase angle α(hk*l*) is given by the angle. measured counterclockwise, between the positive real axis and the vector F.

Scattering from “atoms in two unit cells”

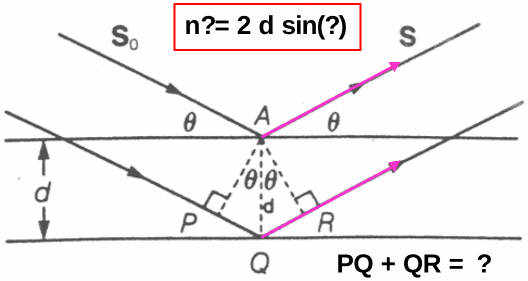


Crystals: Scattering from “planes”

Resultant scattering of resultant scattering!

Bragg Equation

n?= 2 d sin(?)



PQ + QR = ?

→ **Scattering** will only be **“observed”** at discrete **Bragg angles** (q)

The spacings of the Bragg reflections → **Lattice Constants**

Bragg Planes

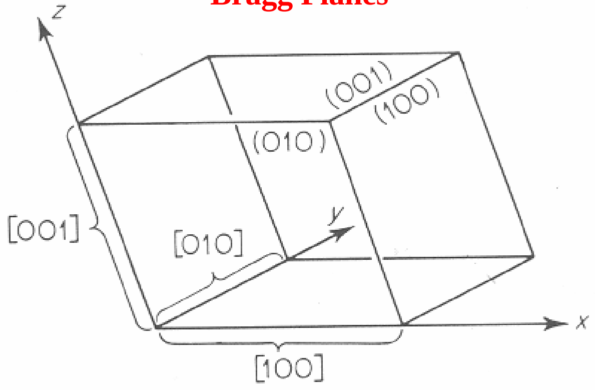


Figure 2.7. Unit cell showing bounding planes and edges.

1 1 0 **1 3 0** **-2 1 0**

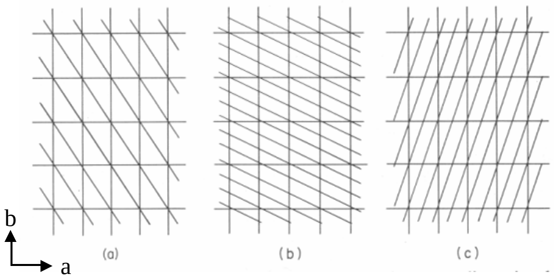


Figure 2.5. Three families of lattice "planes" in a two-dimensional lattice.

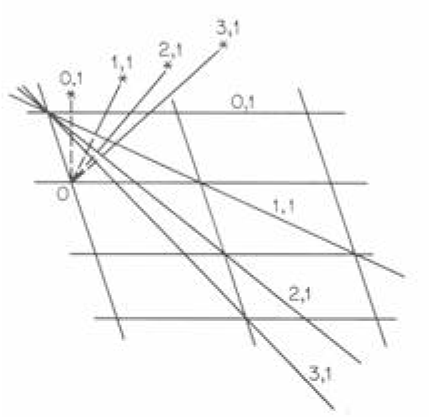
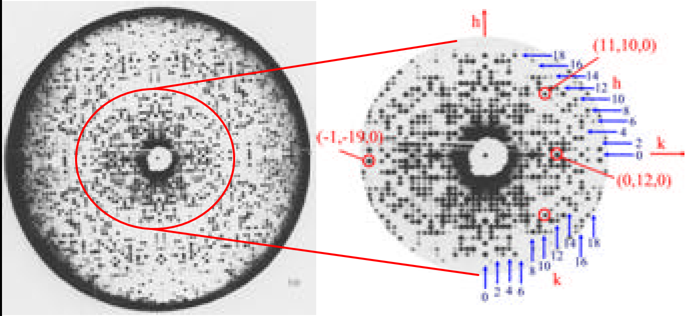


Figure 2.11. Planes in direct space represented by points in reciprocal space.

Electron Density Function

$$\rho(X,Y,Z) = \frac{1}{V}\sum_h\sum_k\sum_l \underline{F(hkl)}\ \underline{\exp[i\alpha(hkl)]}\exp[-2\pi i(hX + kY + lZ)]$$

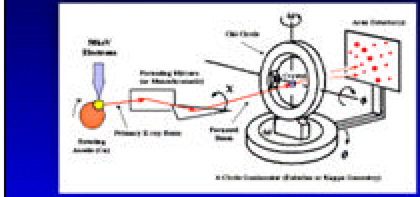


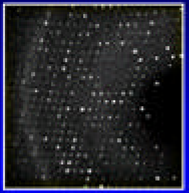
Measure thousands of **Amplitudes** - [F_{hkl}]'s - ?? How do we obtain **Phases** α_{hkl} ??

→ **Phase Problem**

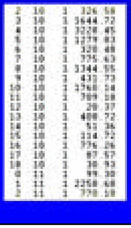
Advanced Methods in Modern Biomolecular Crystallography

The information we get from a single diffraction experiment……

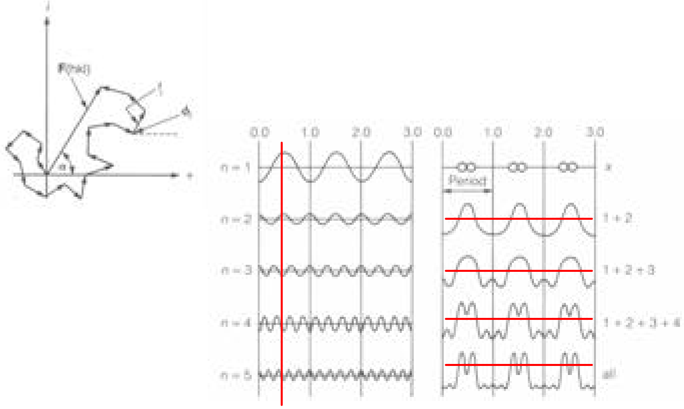




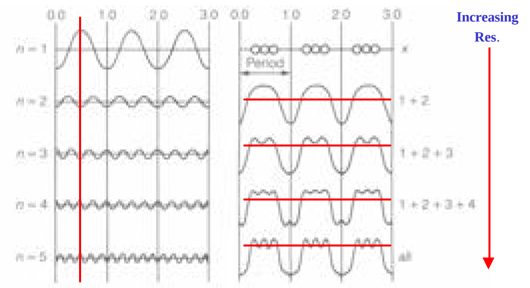
The reflections are indexed (consistent assignment of reciprocal cell indices h,k,l) and all we get for the money is a long list of intensities from several ten thousand reflections



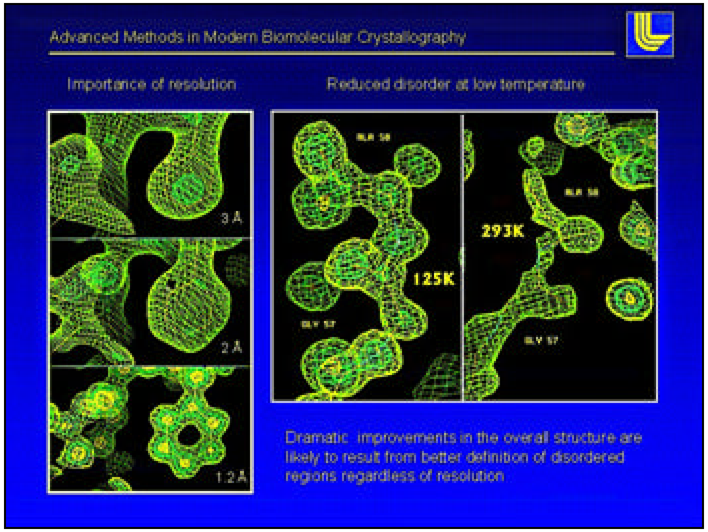
2	10	1	326	58
3	10	1	3644	72
4	10	1	3228	45
5	10	1	3279	83
6	10	1	328	48
7	10	1	775	43
8	10	1	3344	55
9	10	1	431	73
10	10	1	3760	34
11	10	1	789	38
12	10	1	30	37
13	10	1	408	72
14	10	1	53	36
15	10	1	334	72
16	10	1	776	26
17	10	1	87	57
18	10	1	30	93
0	11	1	99	30
1	11	1	3258	68
2	11	1	778	18



Representation of the electron density of a one-dimensional "crystal" by a superposition of waves. The crystal is formed by a periodic repetition of a diatomic molecule, as shown at the top of the right-hand column. The component waves, each with proper phase and amplitude, are on the left. The curves on the right show the successive superposition of the five waves on the left. (From Waser, 1968.)



Representation of another one-dimensional crystal, this one containing a triatomic molecule. Note that this crystal is built up from the same waves as the crystal of (a) ; only the amplitudes and phases have been changed. (From Waser, 1968.)



Advanced Methods in Modern Biomolecular Crystallography

Importance of resolution

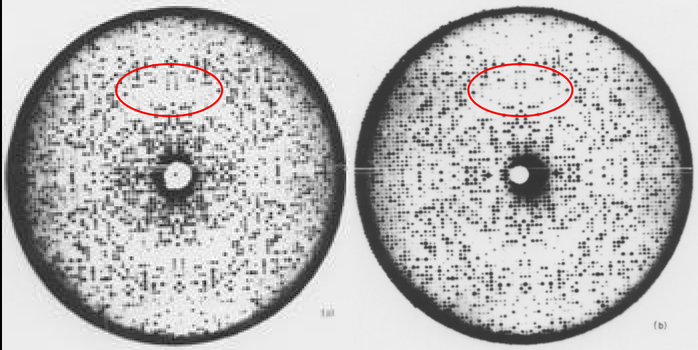
Reduced disorder at low temperature

Dramatic improvements in the overall structure are likely to result from better definition of disordered regions regardless of resolution

Solving the Phase Problem

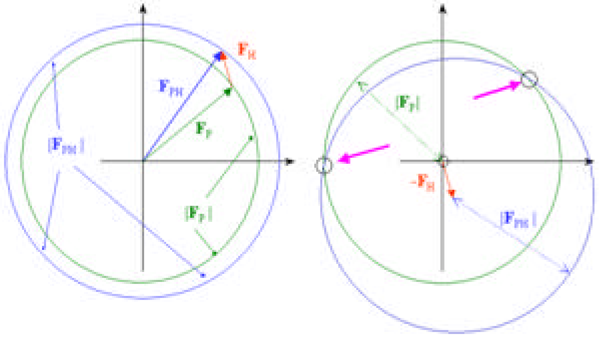
1. **MIR**: **Multiple Isomorphous Replacement (Heavy Atom)**
2. **MR**: **Molecular Replacement**
3. **MAD**: **multiwavelength anomolous dispersion**

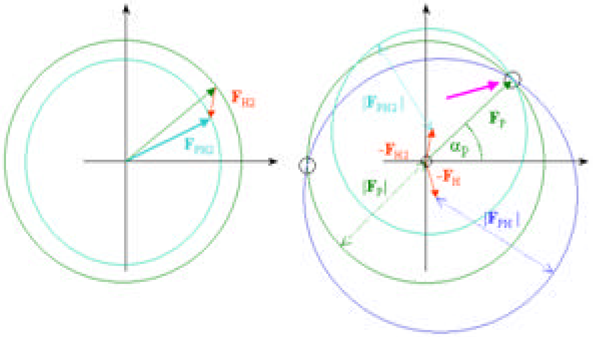
Use of Heavy Metal Ions for Phasing by MIR Methods



Native Phosphorylase

Phosphorylase + Ethyl Hg thiosalicylate





Solving the phase problem by "**Molecular Replacement**".

If an approximate model of the protein structure is known in advance, approximate phases can be guessed, and the unknown parts of the structure can be calculated in an iterative procedure.

No heavy atom derivative required.

BUT – need starting model and orientation (rotation and translation)

For example, molecular replacement can be used to determine the structure of an complex with inhibitor bound to an enzyme active site, if the structure of the enzyme itself is already known. Also, MR is often used to solve the structures of closely related proteins in a superfamily.

Rotation function

$$R(\kappa,\phi,\psi)=\int_{r_{\min}}^{r_{\max}} \mathbf{P}_{nat}(\mathbf{u})\mathbf{P}_{\text{mod}}(\kappa,\phi,\psi,\mathbf{u})d\mathbf{u}$$

Translation function

$$\mathbf{T}(\mathbf{t}) = \int_{cell} \mathbf{P}_{2\to1}(\mathbf{u}-\mathbf{t})\mathbf{P}_{nat}(\mathbf{u})d\mathbf{u}$$

$$= \tfrac{1}{V}\sum_{\mathbf{h}}\left(\mathbf{F}_1(\mathbf{h})\mathbf{F}_2^*(\mathbf{h})\right)^*\left|\mathbf{F}_O(\mathbf{h})\right|^2 \exp(-2\pi i\mathbf{h}\cdot\mathbf{t})$$

$$= \tfrac{1}{V}\sum_{\mathbf{h}}\mathbf{F}_1^*(\mathbf{h})\mathbf{F}_2(\mathbf{h})\left|\mathbf{F}_O(\mathbf{h})\right|^2 \exp(-2\pi i\mathbf{h}\cdot\mathbf{t})$$

"Multiwavelength Anomolous Dispersion" (MAD) methods

Additional information used in calculating phases can be obtained if x-ray diffraction intensities can be measured at wavelengths near the absorption edge of the heavy atom derivative.

A tunable x-ray source is required (provided by a synchrotron). In a synchrotron, accelerated electrons traveling near the speed of light emit intense x-rays.

a) often only a single heavy atom derivative is required to solve a structure (selenomethionine).

b) it is possible to solve structure of higher molecular weight molecules (such as the ribosome, at MW = 2,500,000).

Difference Fourier

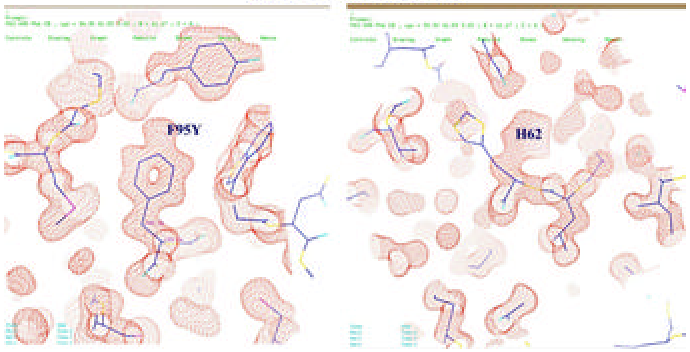
Obs. $\rho_o(x, y, z) = \frac{1}{V}\sum_h\sum_k\sum_l F_{o,hkl}e^{-2\pi i(hx+ky+lz)} + R$

Calc. $\rho_c(x, y, z) = \frac{1}{V}\sum_h\sum_k\sum_l F_{c,hkl}e^{-2\pi i(hx+ky+lz)} + R'$

$$\rho_o(x, y, z) - \rho_c(x, y, z) = \frac{1}{V}\sum_h\sum_k\sum_l (F_o - F_c)_{hkl}e^{-2\pi i(hx+ky+lz)} + R - R'$$

$$\rho_o - \rho_c = \frac{1}{V}\sum_h\sum_k\sum_l \Delta F_{hkl}e^{-2\pi i(hx+ky+lz)}$$

MIF - 1.5A 2Fo-Fc



Least Squares Refinement

$$\sum_{r=1}^{m} w_r \left(\frac{\partial|kF_{c,r}|}{\partial p_1}\right)^2 \Delta p_1 + \sum_{r=1}^{m} w_r \frac{\partial|kF_{c,r}|}{\partial p_1}\frac{\partial|kF_{c,r}|}{\partial p_2}\Delta p_2 + \cdots$$

$$+ \sum_{r=1}^{m} w_r \frac{\partial|kF_{c,r}|}{\partial p_1}\frac{\partial|kF_{c,r}|}{\partial p_n}\Delta p_n = \sum_{r=1}^{m} w_r \Delta F_r \frac{\partial|kF_{c,r}|}{\partial p_1}$$

$$\sum_{r=1}^{m} w_r \frac{\partial|kF_{c,r}|}{\partial p_2}\frac{\partial|kF_{c,r}|}{\partial p_1}\Delta p_1 + \sum_{r=1}^{m} \left(\frac{\partial|kF_{c,r}|}{\partial p_2}\right)^2 \Delta p_2 + \cdots$$

$$+ \sum_{r=1}^{m} w_r \frac{\partial|kF_{c,r}|}{\partial p_2}\frac{\partial|kF_{c,r}|}{\partial p_n}\Delta p_n = \sum_{r=1}^{m} w_r \Delta F_r \frac{\partial|kF_{c,r}|}{\partial p_2}$$

$$\vdots$$

$$\sum_{r=1}^{m} w_r \frac{\partial|kF_{c,r}|}{\partial p_n}\frac{\partial|kF_{c,r}|}{\partial p_1}\Delta p_1 + \sum_{r=1}^{m} w_r \frac{\partial|kF_{c,r}|}{\partial p_n}\frac{\partial|kF_{c,r}|}{\partial p_2}\Delta p_2 + \cdots$$

$$+ \sum_{r=1}^{m} w_r \left(\frac{\partial|kF_{c,r}|}{\partial p_n}\right)^2 \Delta p_n = \sum_{r=1}^{m} w_r \Delta F_r \frac{\partial|kF_{c,r}|}{\partial p_n}$$

Energy Refinement

$E_{TOTAL} = E_{EMPIRICAL} + E_{EFFECTIVE}$

$E_{EFFECTIVE} = E_{XREF} + E_{NOE} + E_{HARM} +$
$E_{CDIH} + E_{NCS} + E_{DG} + E_{RELA} + E_{PLAN}$

$E_{EMPIRICAL} = S^{N}_{p=I}\ [w^{p}_{BOND}E_{BOND} + w^{p}_{ANGL}E_{ANGL} +$
$w^{p}_{DIHE}E_{DIHE} + w^{p}_{IMPR}E_{IMPR} +$
$w^{p}_{VDW}E_{VDW} + w^{p}_{ELEC}E_{ELEC} +$
$w^{p}_{PVDW}E_{PVDW} + w^{p}_{PELE}E_{PELE} +$
$w^{p}_{HBON}E_{HBON}]$.

Bonded Energy Terms

$E_{BOND} = \underset{bonds}{S}\, k_b(t - t_o)^2$

$E_{ANGL} = \underset{angles}{S}\,(k_{?}\,(? - ?_0)^2 + k_{ub}(r_13 - r_{ub})^2)$

$E_{DIHE} = \underset{dihedrals}{S}\ \underset{i=1,m}{S}\, k_{f\,i}\,(1 + COS(nf_{\,i} + d_i))$ if $n_i > 0$

$\underset{dihedrals}{S}\ \underset{i=1,m}{S}\, k_{f\,i}\,(f_{\,i} - d_i)^2$ if $n_i = 0$

$E_{IMPR} = \underset{impropers}{S}\ \underset{i=1,m}{S}\, k_{f\,i}\,(1 + COS(nf_{\,i} + d_i))$ if $n_i > 0$

$\underset{impropers}{S}\ \underset{i=1,m}{S}\, k_{f\,i}\,(f_{\,i} - d_i)^2$ if $n_i = 0$

Nonbonded Energy Terms

$$E_{ELEC} = \sum_{i<j} f_{ELEC}(R_{ij}) + e_{14} \sum_{(i,j)\in\{1-4\}} f_{ELEC}(R_{ij})$$

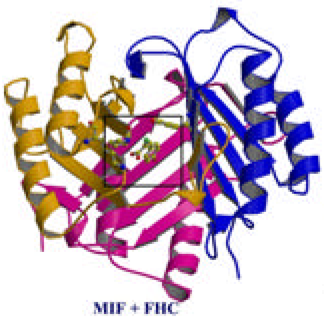
$$E_{VDW} = \sum_{i<j} f_{VDW}(R_{ij}) + \sum_{(i,j)\in\{1-4\}} f_{VDW}(R_{ij})$$

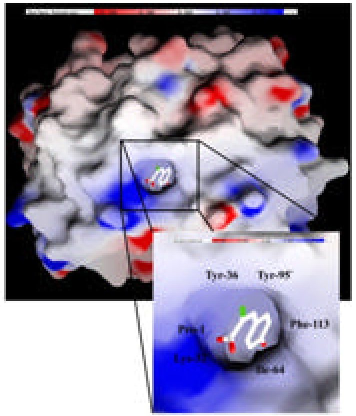
$$E_{PVDW} = \sum_{S=1}^{N_S} \sum_{i<j} f_{VDW}(R_{Sij})$$

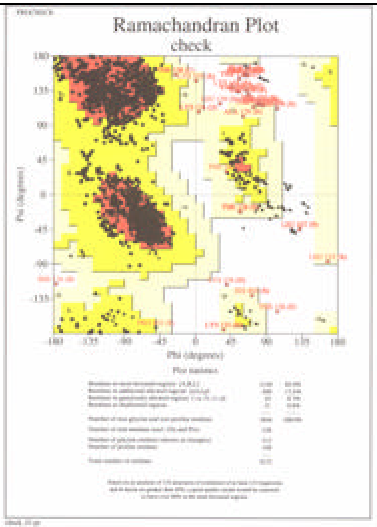
$$E_{PELE} = \sum_{S=1}^{N_S} \sum_{i<j} f_{ELEC}(R_{Sij}) \quad R_{Sij} = |\mathcal{F}^{-1} \cdot MinG(\mathcal{F} \cdot \vec{r}_i - O_s \cdot \mathcal{F} \cdot \vec{r}_j + \vec{t}_s)|$$

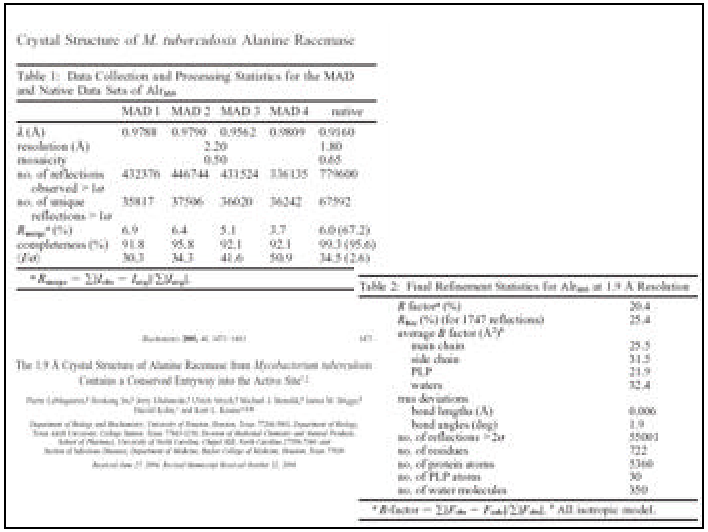
$$E_{PVDW} = \sum_{S\in NCS} \sum_{i<j} f_{VDW}(r_i - Sr_j)$$

$$E_{PELE} = \sum_{S\in NCS} \sum_{i<j} f_{ELEC}(r_i - Sr_j)$$









Crystal Structure of *M. tuberculosis* Alanine Racemase

Table 1: Data Collection and Processing Statistics for the MAD and Native Data Sets of Alr_{Mtb}

	MAD 1	MAD 2	MAD 3	MAD 4	native
λ (Å)	0.9788	0.9790	0.9562	0.9809	0.9160
resolution (Å)			2.20		1.80
mosaicity			0.50		0.65
no. of reflections observed > 1σ	432376	446744	431524	336135	779600
no. of unique reflections > 1σ	35817	37506	36020	36242	67592
R_{merge}[a] (%)	6.9	6.4	5.1	3.7	6.0 (67.2)
completeness (%)	91.8	95.8	92.1	92.1	99.3 (95.6)
⟨I/σ⟩	30.3	34.3	41.6	50.9	34.5 (2.6)

[a] $R_{merge} = \Sigma|I_{obs} - I_{avg}|/\Sigma I_{avg}$.

Biochemistry 2005, 44, 1471–1481 1471

The 1.9 Å Crystal Structure of Alanine Racemase from *Mycobacterium tuberculosis* Contains a Conserved Entryway into the Active Site[†,‡]

Pierre LeMagueres,[§] Hookang Im,[§] Jerry Ebalunode,[§] Ulrich Strych,[§] Michael J. Benedik,[§] James M. Briggs,[§] Harold Kohn,[∥] and Kurt L. Krause*,[§,⊥]

Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204-5001, Department of Biology, Texas A&M University, College Station, Texas 77843-3258, Division of Medicinal Chemistry and Natural Products, School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599-7360, and Section of Infectious Diseases, Department of Medicine, Baylor College of Medicine, Houston, Texas 77030

Received June 27, 2004; Revised Manuscript Received October 12, 2004

Table 2: Final Refinement Statistics for Alr_{Mtb} at 1.9 Å Resolution

R factor[a] (%)	20.4
R_{free} (%) (for 1747 reflections)	25.4
average *B* factor (Å^2)[b]	
main chain	25.5
side chain	31.5
PLP	21.9
waters	32.4
rms deviations	
bond lengths (Å)	0.006
bond angles (deg)	1.9
no. of reflections > 2σ	55001
no. of residues	722
no. of protein atoms	5360
no. of PLP atoms	30
no. of water molecules	350

[a] *R*-factor = $\Sigma|F_{obs} - F_{calc}|/\Sigma|F_{obs}|$. [b] All isotropic model.

Analyze – structure (Ramachandran Plot) and biochemistry

Publish in leading biochemical or structural biology journal

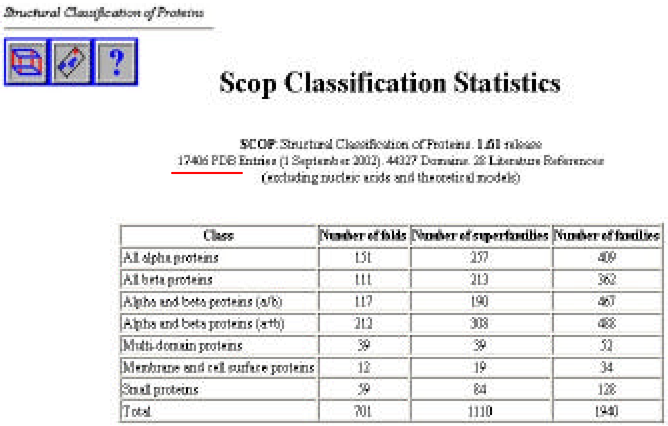
Contribute results (coordinates, etc.) to PDB

****

Data Mining

- **Visualization programs (Cn3D / RasMol / SwissPDBV / etc)**
- **SCOP – Structural Classification of Proteins**
- **CATH – Classification / Arch / Topology**

SCOP Structural Classification of Proteins



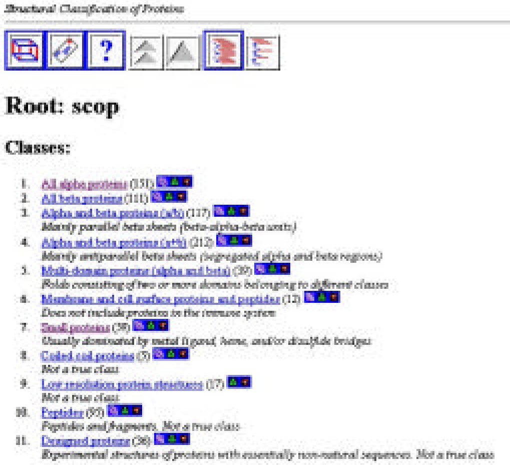
Structural Classification of Proteins

Scop Classification Statistics

SCOP: Structural Classification of Proteins. **1.61** release
17406 PDB Entries (1 September 2002). 44327 Domains. 28 Literature References
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	151	237	409
All beta proteins	111	213	362
Alpha and beta proteins (a/b)	117	190	467
Alpha and beta proteins (a+b)	212	308	488
Multi-domain proteins	39	39	52
Membrane and cell surface proteins	12	19	34
Small proteins	59	84	128
Total	701	1110	1940

SCOP Structural Classification of Proteins



Structural Classification of Proteins

Root: scop

Classes:

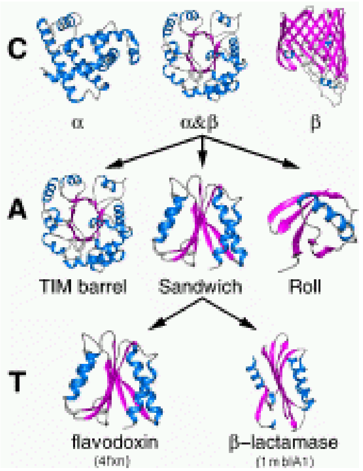
1. All alpha proteins (151)
2. All beta proteins (111)
3. Alpha and beta proteins (a/b) (117)
   Mainly parallel beta sheets (beta-alpha-beta units)
4. Alpha and beta proteins (a+b) (212)
   Mainly antiparallel beta sheets (segregated alpha and beta regions)
5. Multi-domain proteins (alpha and beta) (39)
   Folds consisting of two or more domains belonging to different classes
6. Membrane and cell surface proteins and peptides (12)
   Does not include proteins in the immune system
7. Small proteins (59)
   Usually dominated by metal ligand, heme, and/or disulfide bridges
8. Coiled coil proteins (5)
   Not a true class
9. Low resolution protein structures (17)
   Not a true class
10. Peptides (95)
    Peptides and fragments. Not a true class
11. Designed proteins (36)
    Experimental structures of proteins with essentially non-natural sequences. Not a true class

CATH - Protein Structure Classification

CATH is a novel hierarchical classification of protein domain structures, which clusters proteins at four major levels: Class (C), Architecture (A), Topology (T), and Homologous (H) Superfamily

Class, derived from secondary structure content, is assigned for more than 90% of protein structures automatically. **Architecture**, which describes the gross orientation of secondary structures, independent of connectivities, is currently assigned manually. The **topology** level clusters structures according to their topological connections and numbers of secondary structures. The **homologous superfamilies** cluster proteins with highly similar structures and functions. The assignments of structures to toplogy families and homologous superfamilies are made by sequence and structure comparisons.

CATH


C
α
α&β
β
A
TIM barrel
Sandwich
Roll
T
flavodoxin
(4fxn)
β–lactamase
(1mbIA1)