Macromolecular structure determination by X-ray crystallography

Arjan Snijder
Bioinformatics III
Department of Chemistry and Bioscience
Molecular Biotechnology
arjan.snijder@chembio.chalmers.se

Macromolecular structure determination by X-ray crystallography

Overview of X-ray structure determination
Protein data-bank coordinate files
Structure quality

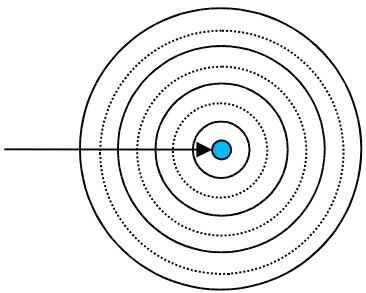
Structural Biology

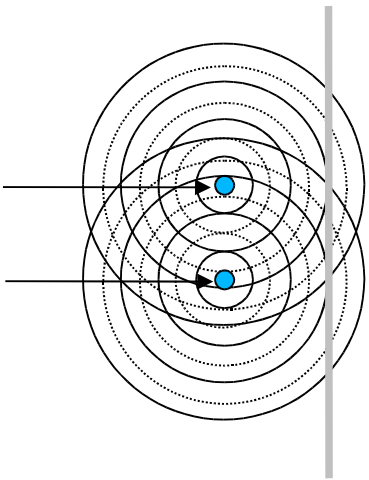
- Insight in biological processes
- Three-dimensional structure of biomacromolecules

  - NMR spectroscopy
  - AFM
  - Electron microscopy
  - X-ray crystallography

Basis of X-ray crystallography

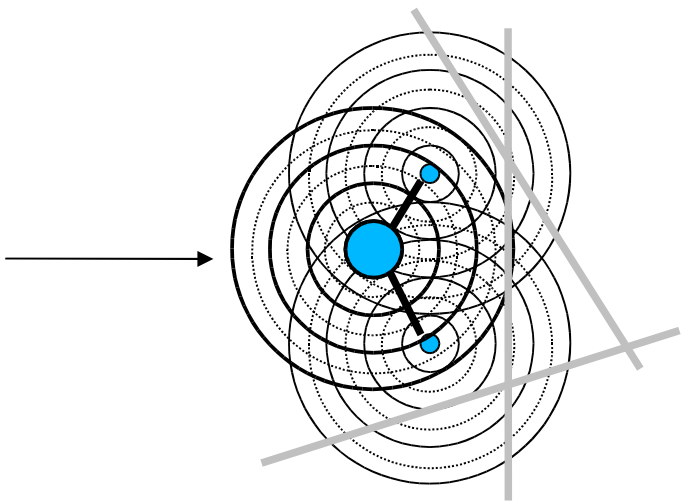
- Irradiation of electrons with X-ray wave
- Electrons emits X-rays
- Interference

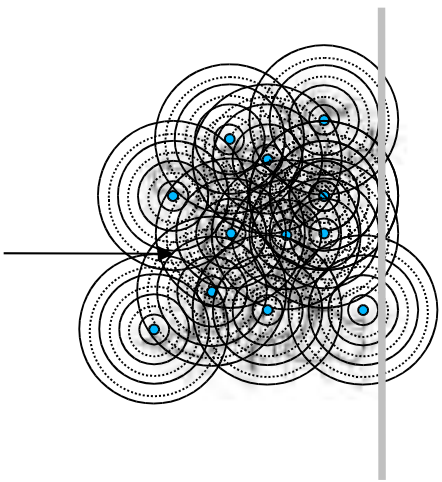




Basis of X-ray crystallography

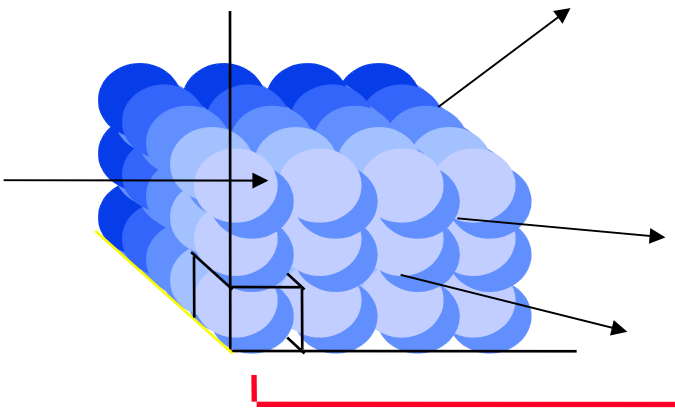
- Interference
- Structure dependent
- Scattering related to number of electrons



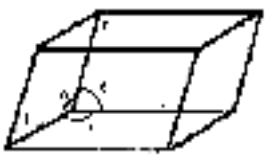


Crystals

- Weak diffraction from single molecules
- Crystals are regular arrangements of molecules
- Interference by molecules
- Amplification of signal in certain directions
- Signal strong enough to measure

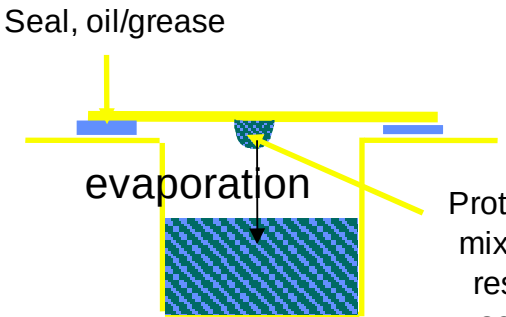


Unit cell:
smallest repeating unit

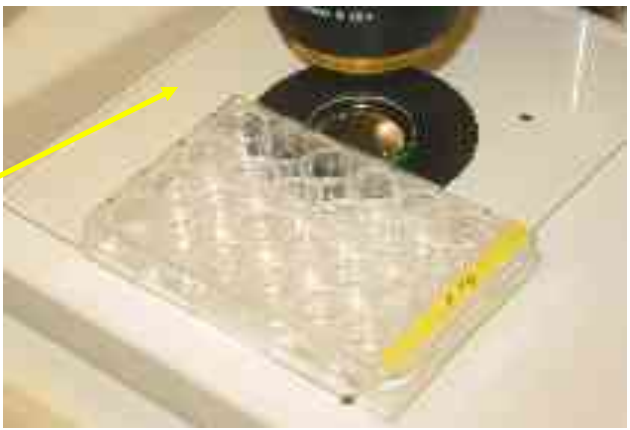


Crystals, Strength and Weakness

Crystallization: controlled decrease of solubility

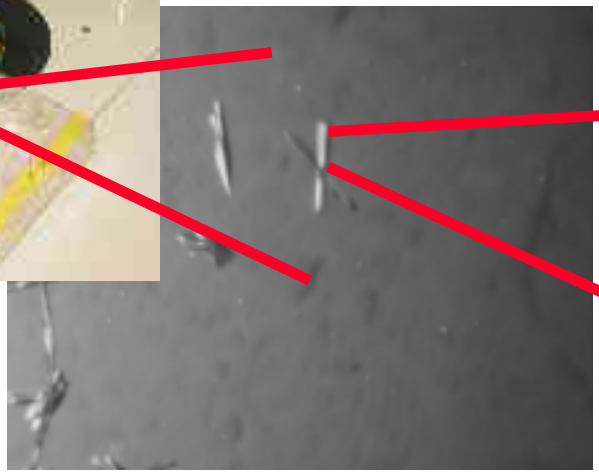


Hanging drop set-up

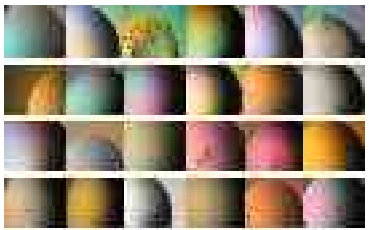


Parameters: Salt concentration, type of salts, kinetics, pH, protein concentration, additives, protein purity, temperature

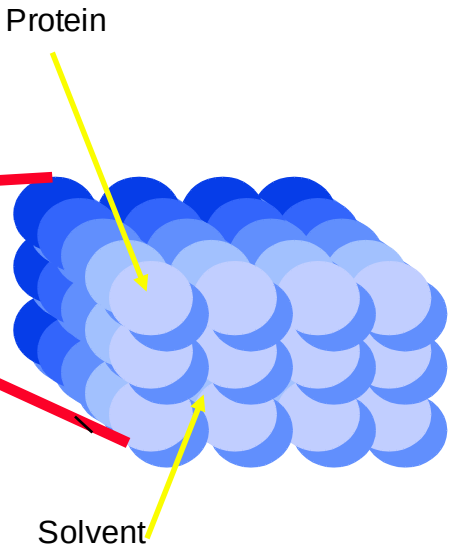
Protein crystals



Number of experiments 24-100.000

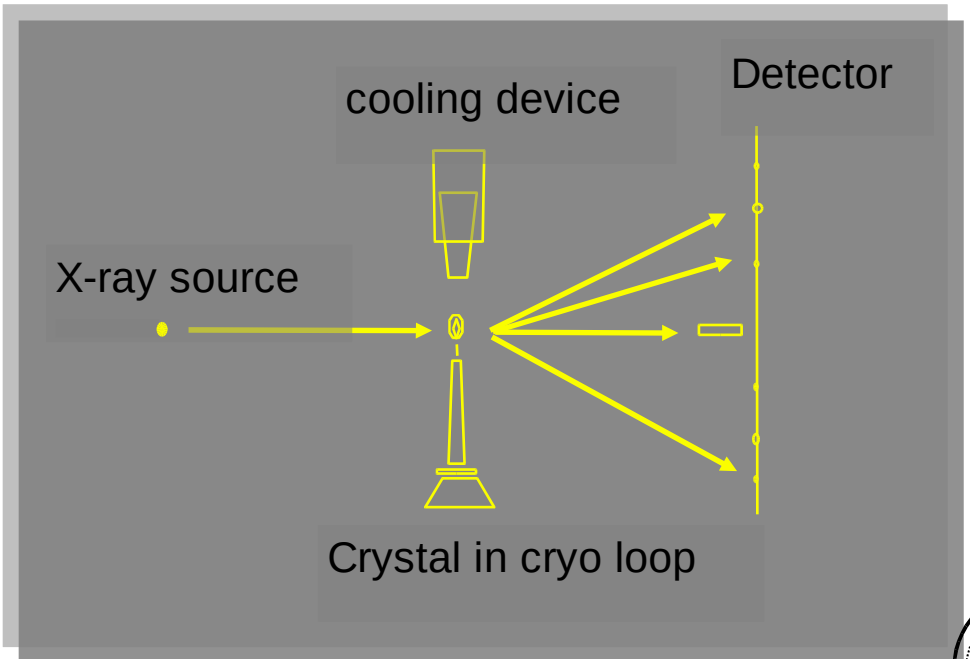


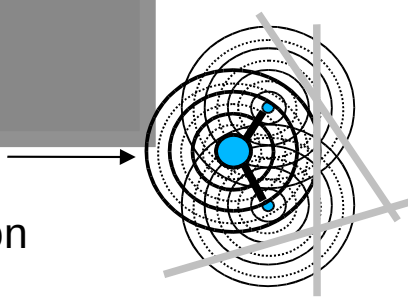
Dimensions 0.1 x 0.05 x 0.05 mm^3



Solvent content: 20% to 80%

Diffraction experiment

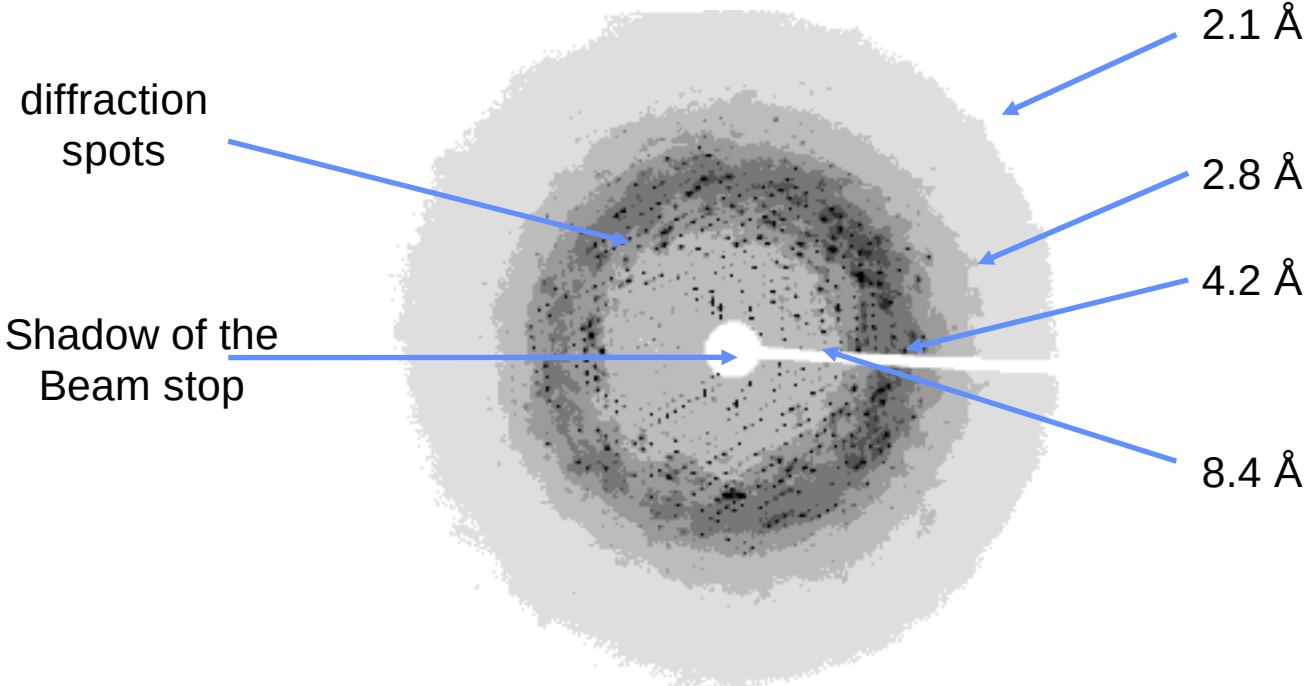




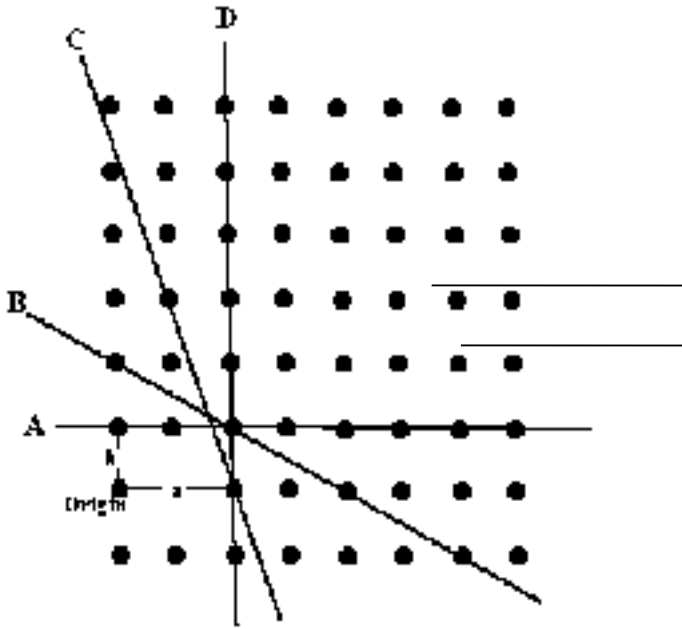
Crystal rotates during data collection

The X-ray diffractometer

X-ray diffraction image

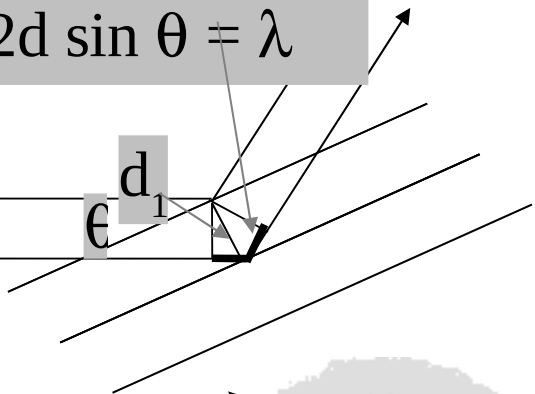


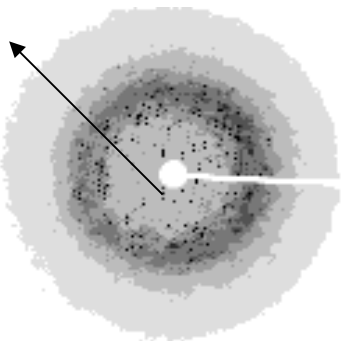
Reflection against planes



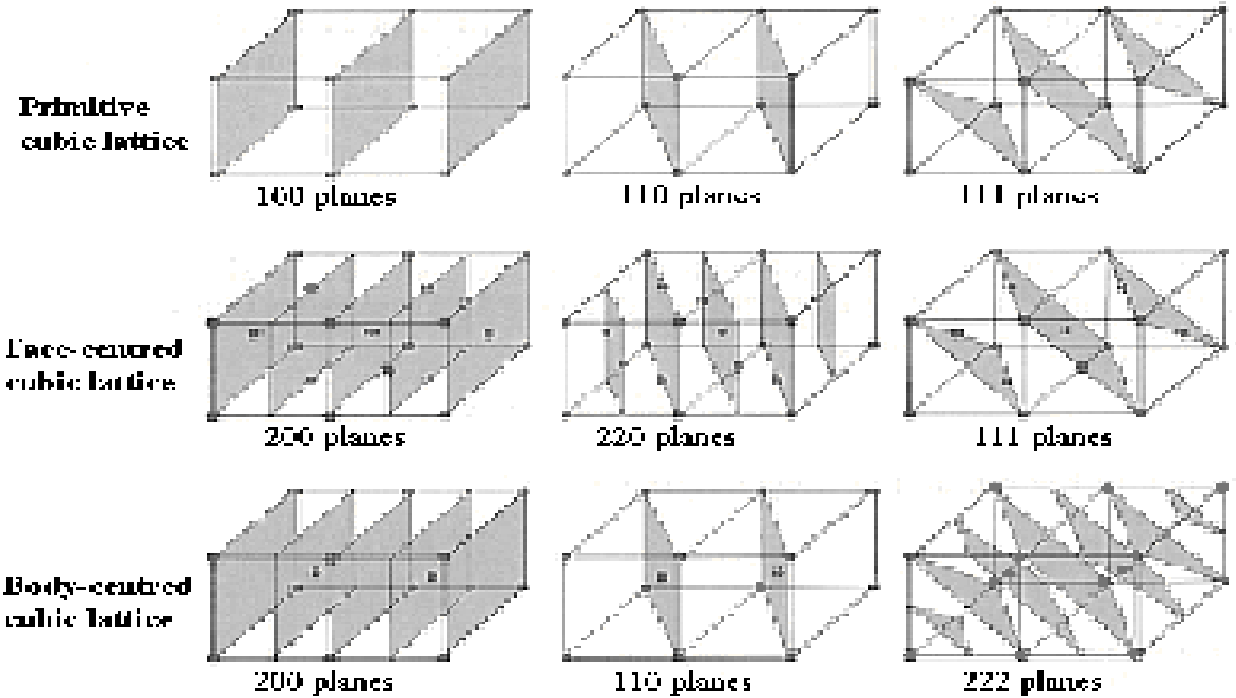
Characterization of a two-dimensional lattice in terms of a set of planes.

path difference
$2d \sin \theta = \lambda$





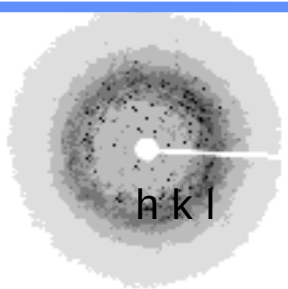
Reflection against planes

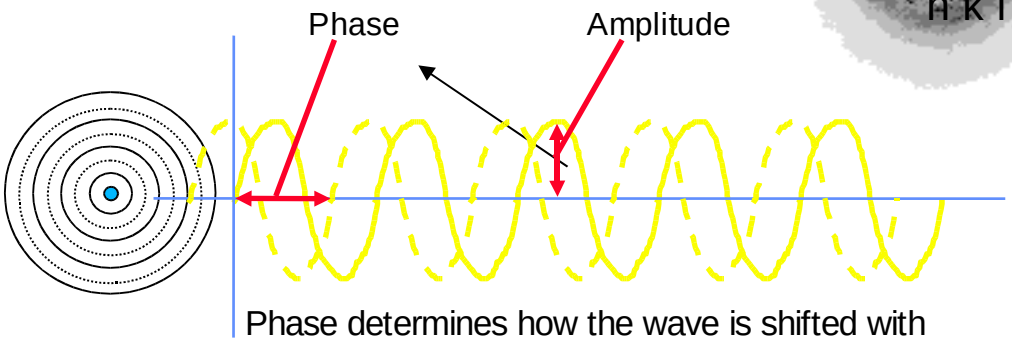


Miller indices for three types of cubic lattices.

Fourier synthesis

- X-ray wave, amplitude and phase
- Intensities measured, amplitudes
- Phases cannot be measure directly



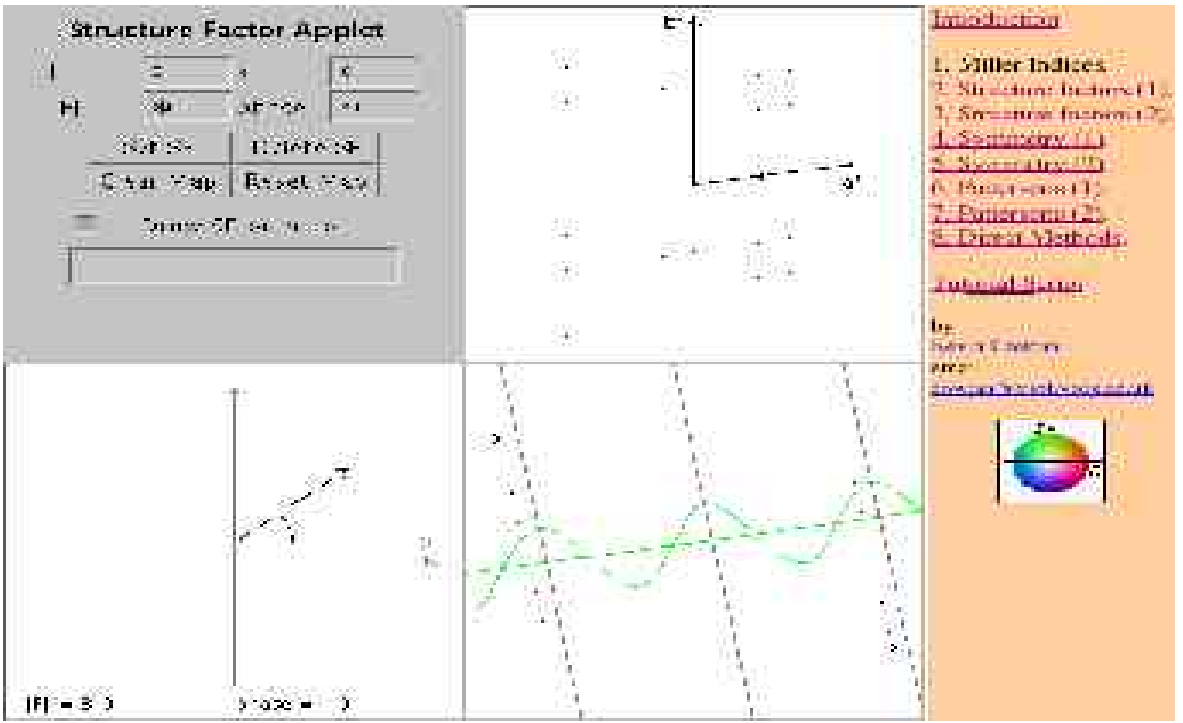


Phase determines how the wave is shifted with respect to the origin

$$\rho(x\ y\ z) = 1/V \sum\sum\sum |F(h\ k\ l)| \exp[-2\pi i(hx+ky+lz)+i\alpha(h\ k\ l)]$$

electron density — structure factor amplitudes — phase angle

Structure Factors



http://www.embl-hamburg.de/~tilo/sfapplet/bigframe_direct.html

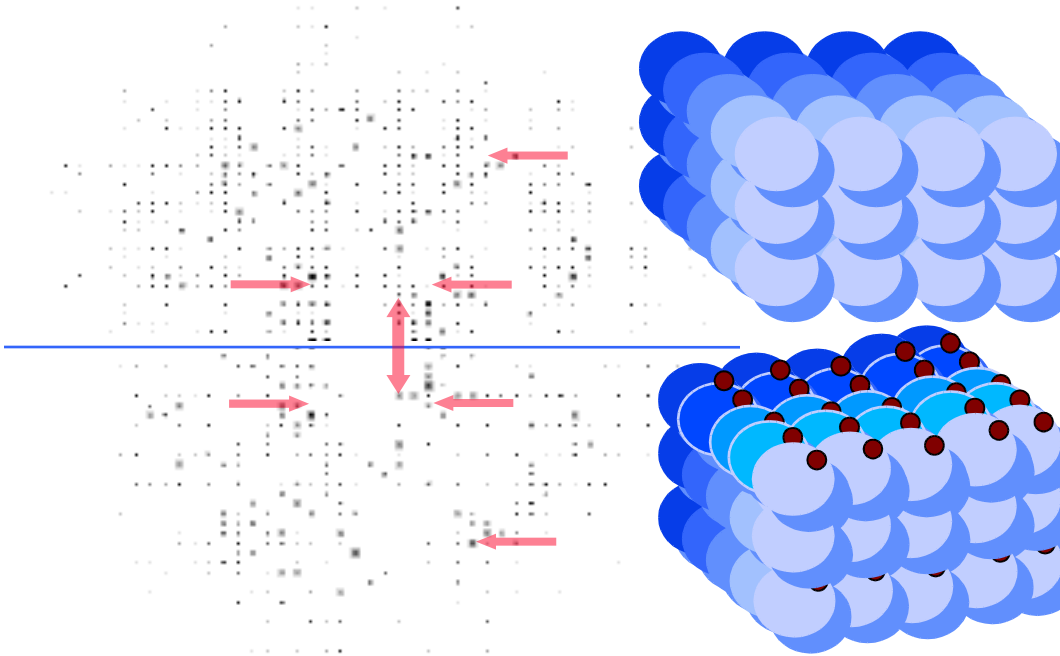
Phases

Method	Principal	Requirement
Isomorphous Replacement	Scattering difference of extra added electrons is exploited to obtain phases	Specific binding of heavy atoms
Multiple-wavelength Anomalous Dispersion	Wavelength dependent scattering of heavy atoms results in phases	Anomalous scatterer, synchrotron radiation
Molecular Replacement	Phases calculated from highly homologous protein with known structure.	Homologous structure
Direct Methods	Phases are calculated without further information than amplitudes	High obs/parameter ratio

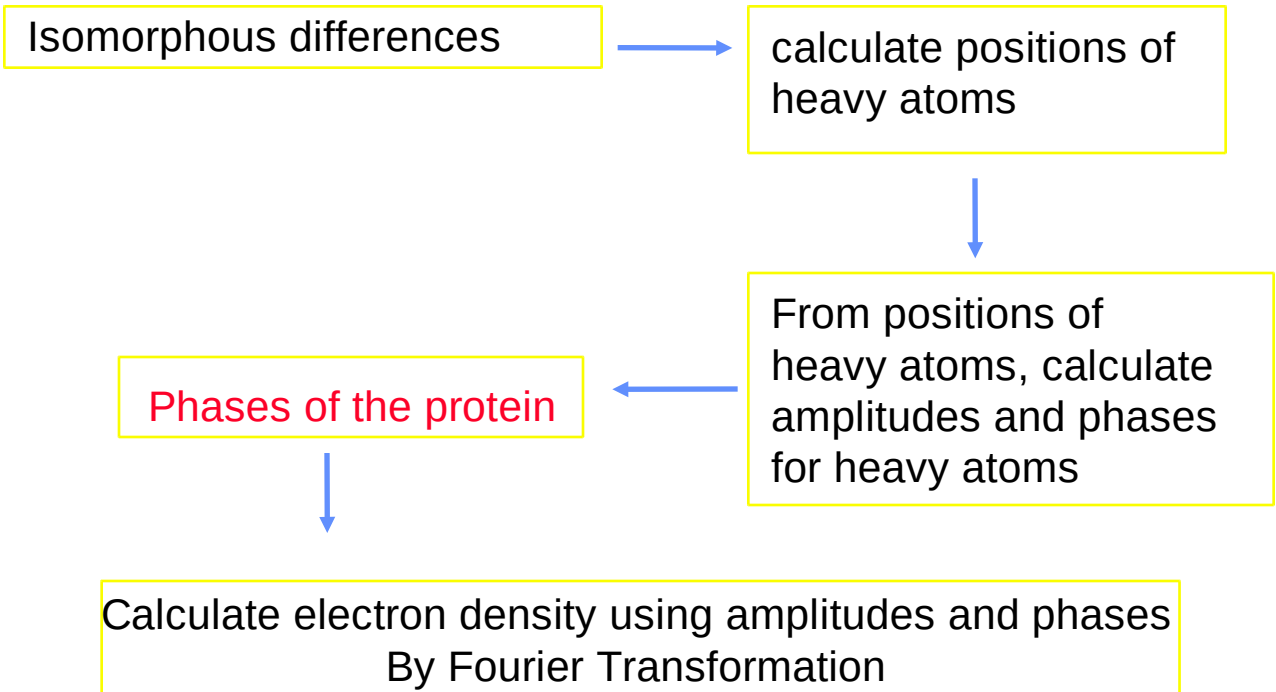
Isomorphous differences

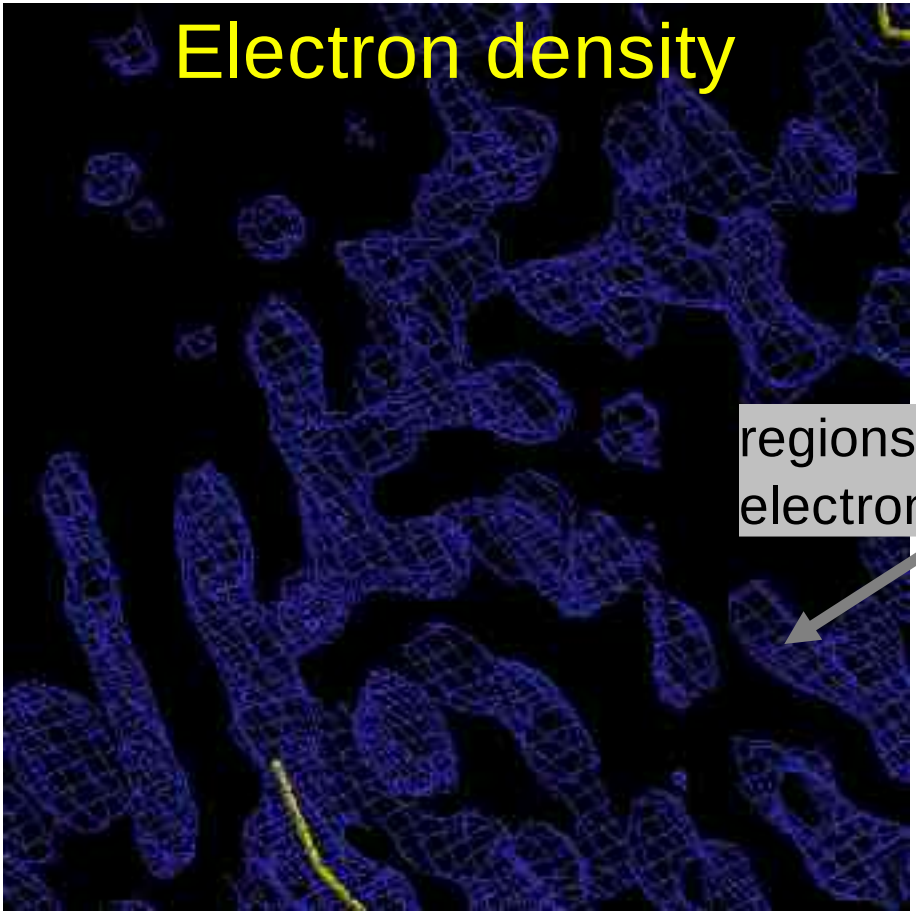
Native

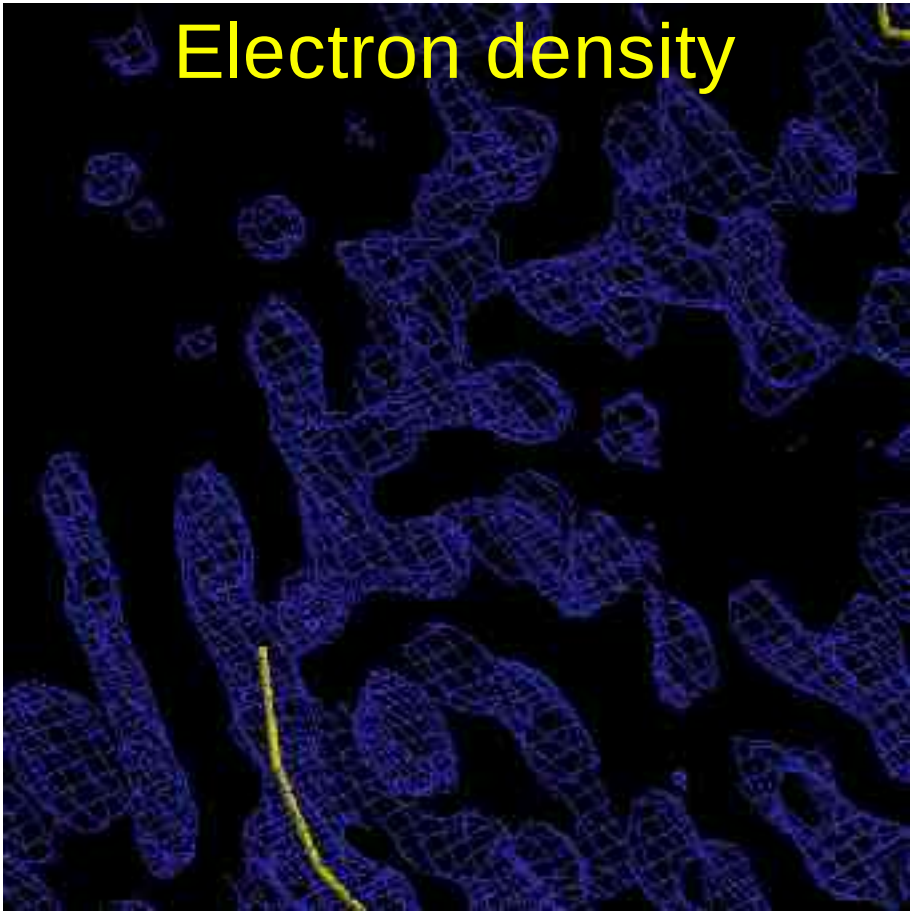
Ir derivative

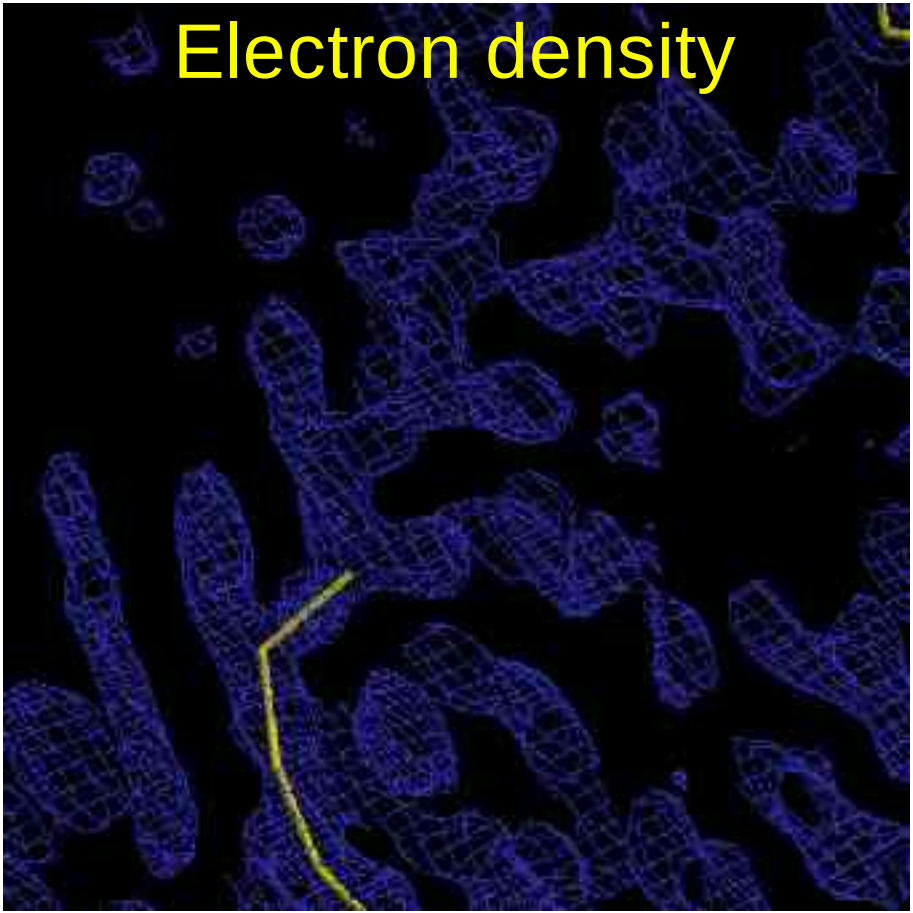


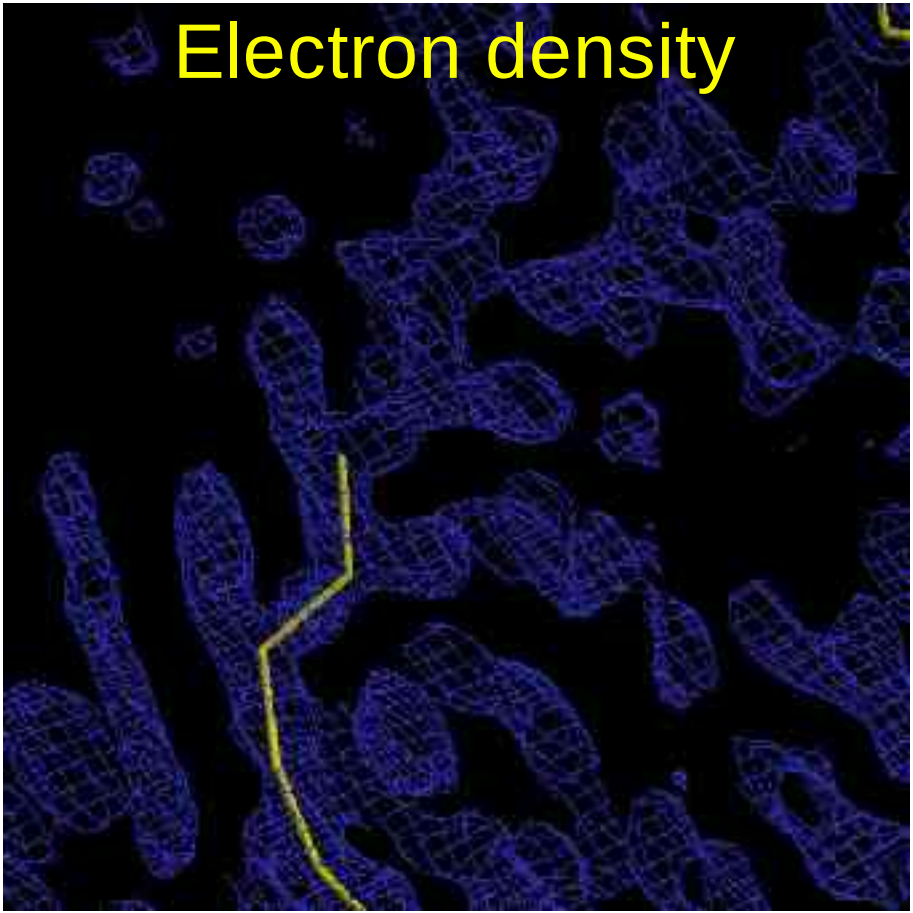
Isomorphous differences

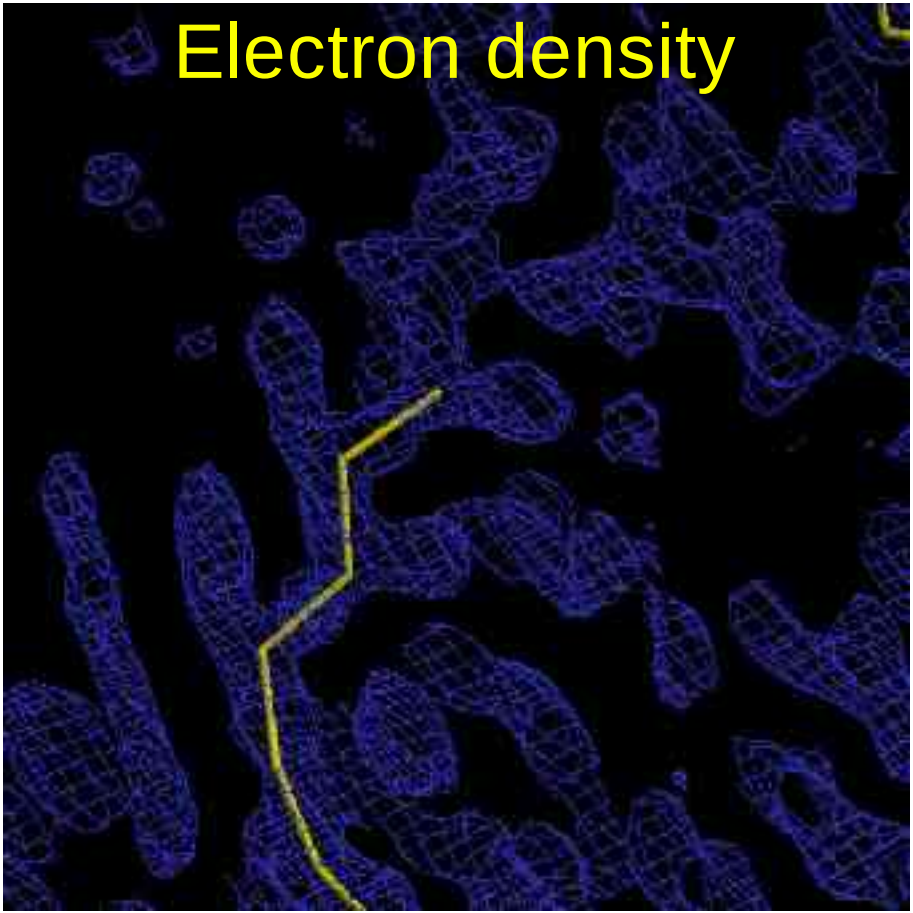


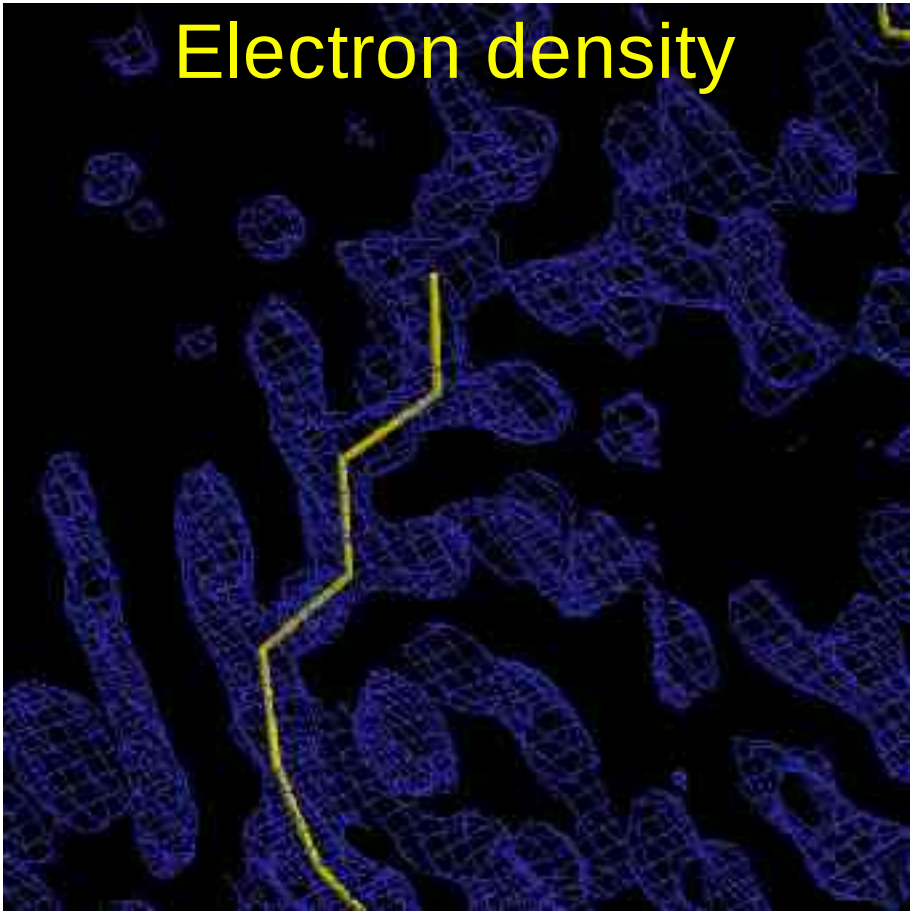

Electron density
regions with high electron density


Electron density


Electron density

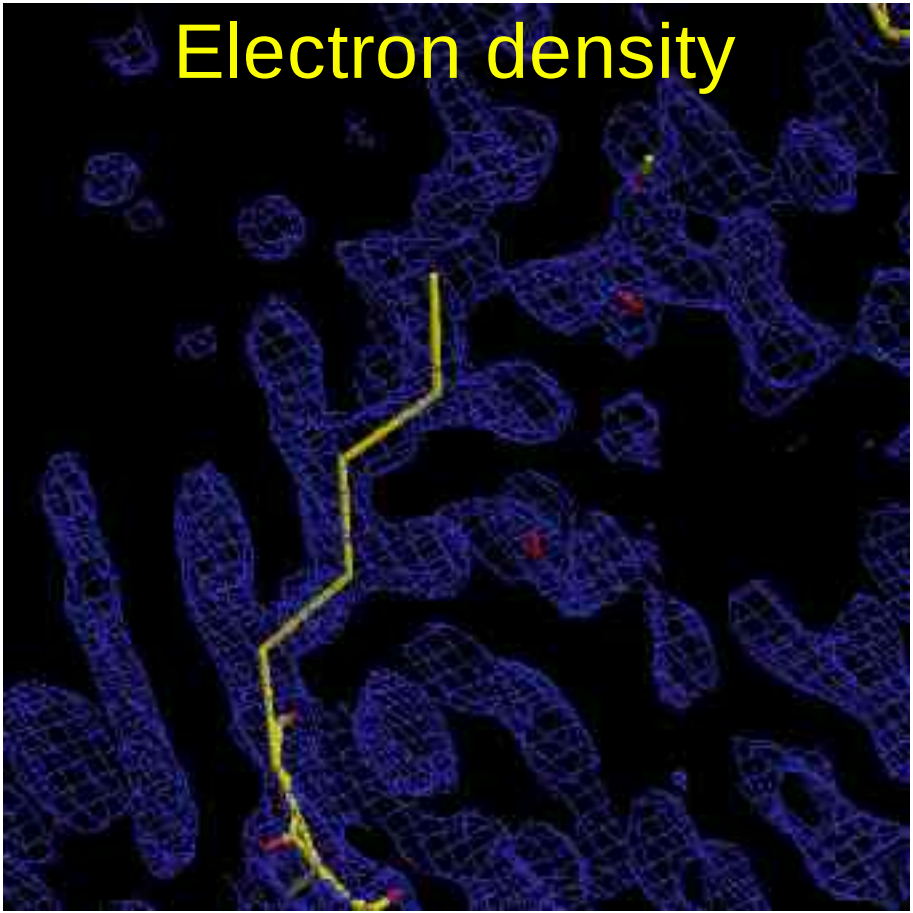

Electron density

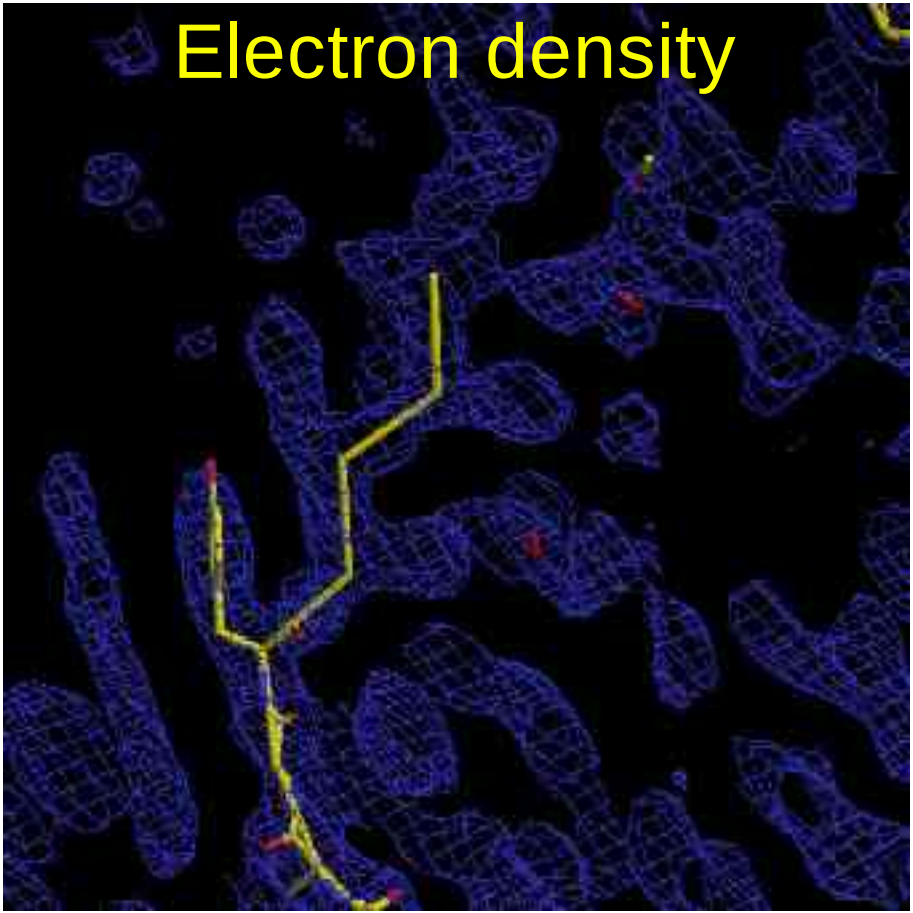

Electron density

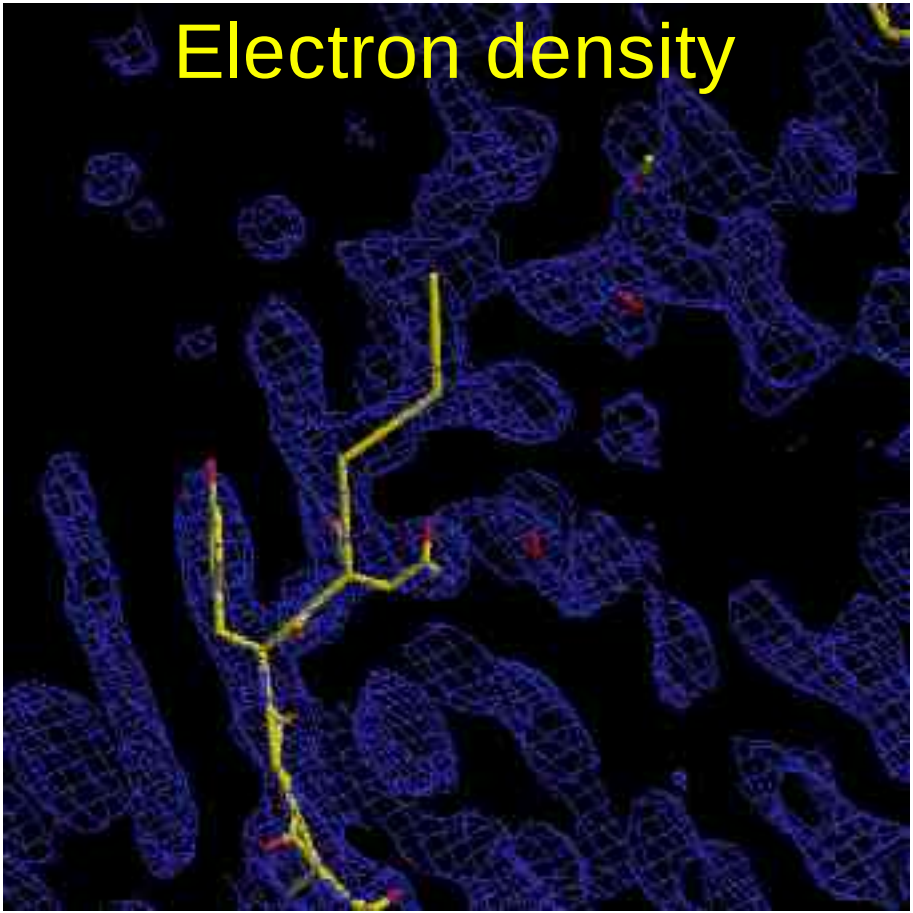

Electron density

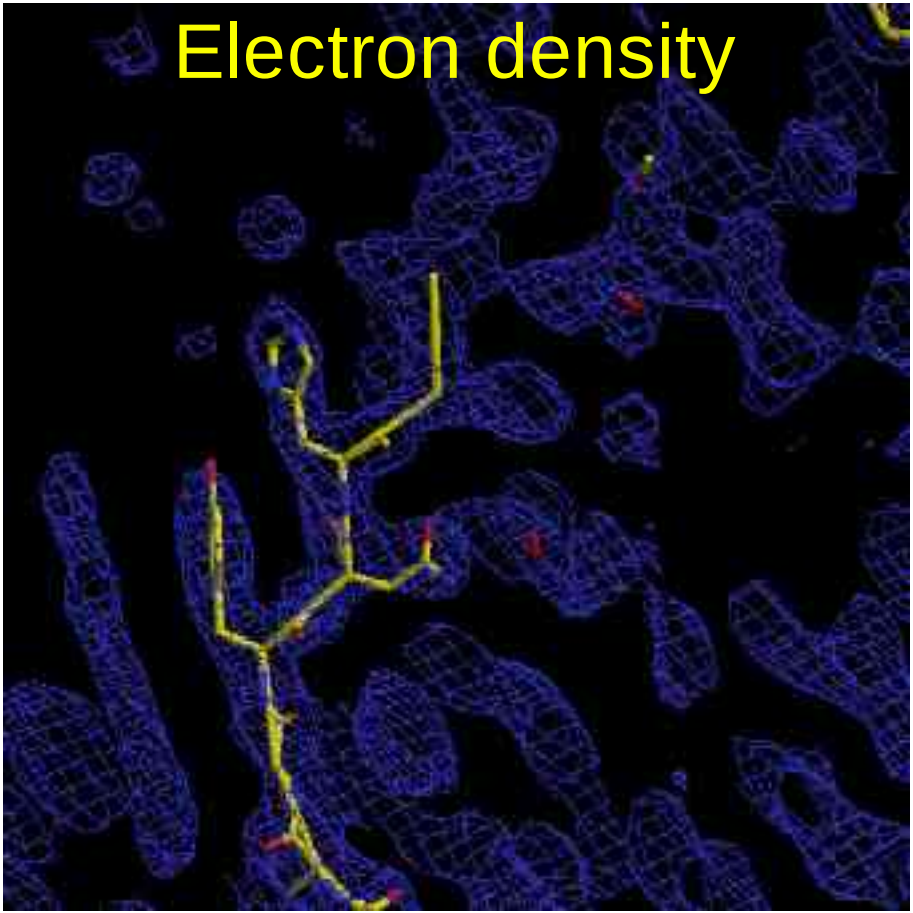
Sequence

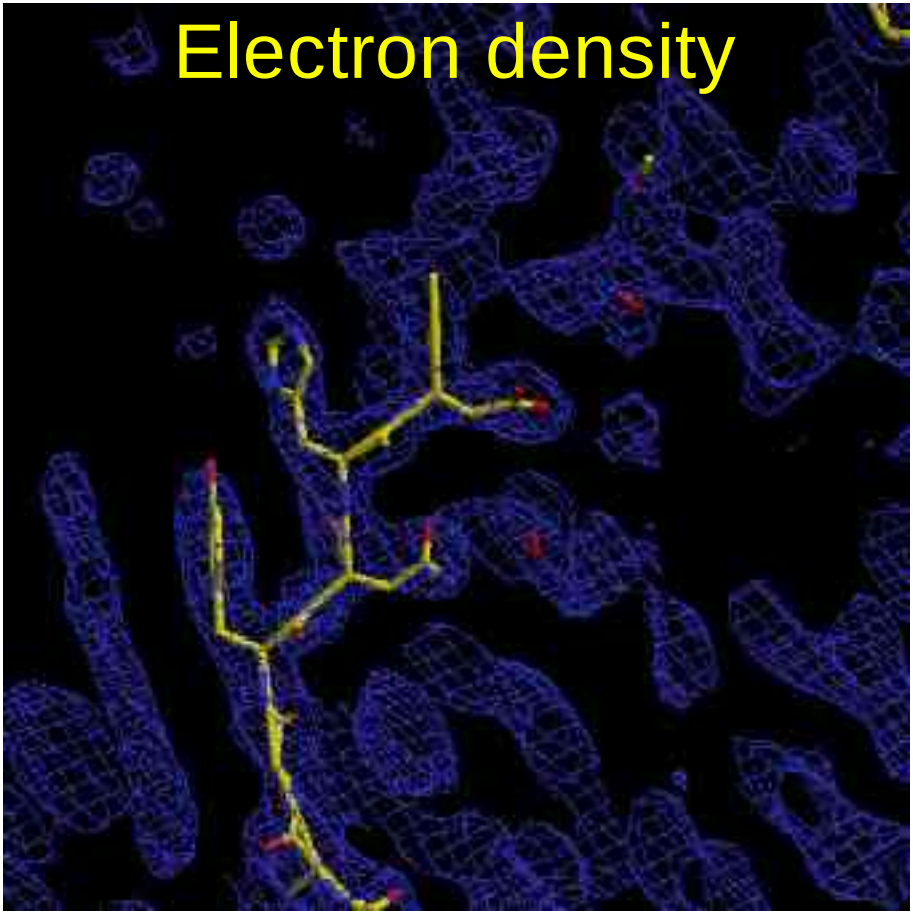
```
QEATVKEVHD APAVRGSIIA NMLQEHDNPF TLYPYDTNYL
1                                          40
IYTQTSDLNK EAIASYDWAE NARKDEVKFQ LSLAFPLWRG
41                                         80
ILGPNSVLGA SYTQKSWWQL SNSEESSPFR ETNYEPQLFL
81                                        120
GFATDYRFAG WTLRDVEMGY NHDSNGRSDP TSRSWNRLYT
121                                       160
RLMAENGNWL VEVKPWYVVG NTDDNPDITK YMGYYQLKIG
161                                       200
YHLGDAVLSA KGQYNWNTGY GGAELGLSYP ITKHVRLYTQ
201                                       240
VYSGYGESLI DYNFNQTRVG VGVMLNDLF
241                          269
```

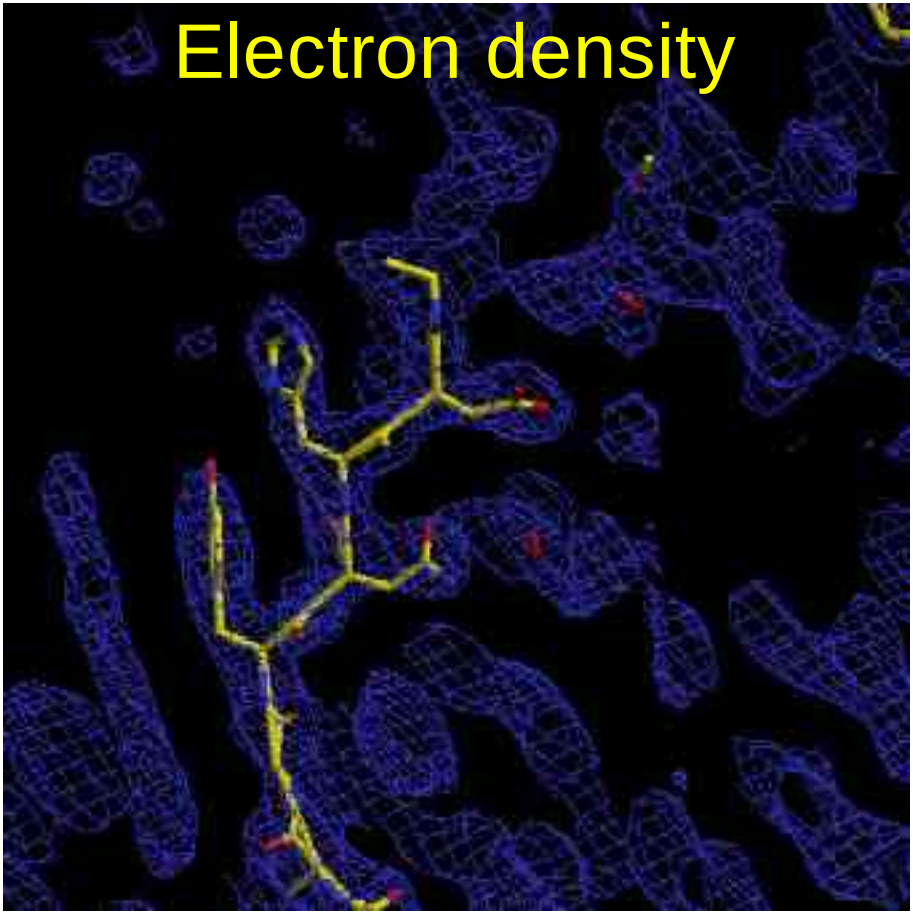

Electron density


Electron density


Electron density

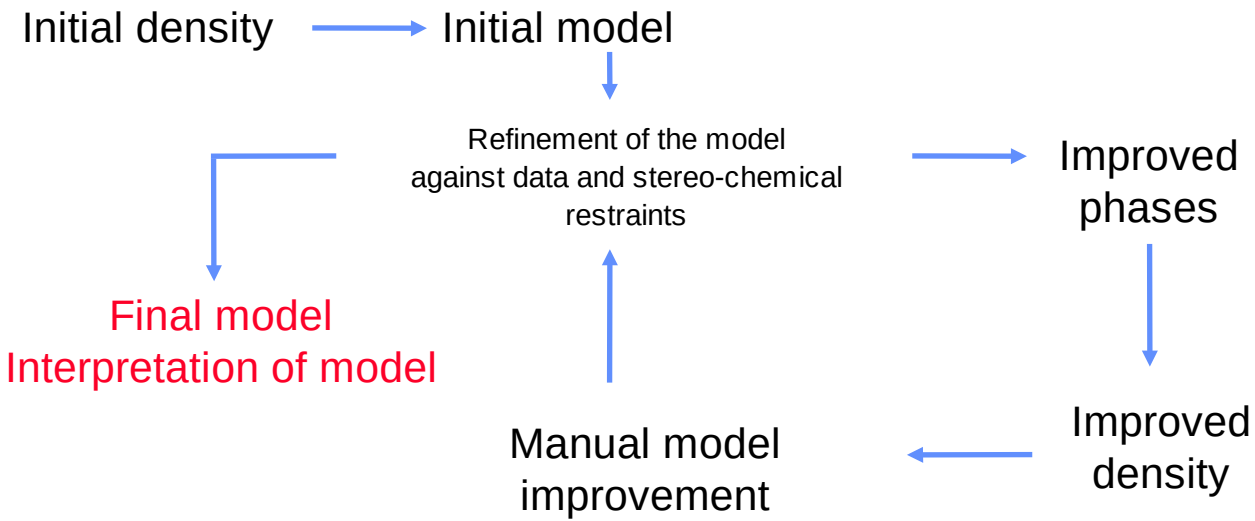

Electron density


Electron density


Electron density

Refinement

Refinement: Adjustment of the model for better agreement with experimental data and stereo-chemical restraints



Refinement

Minimization: $E_{total} = (1-w_x)E_{geometry} + w_x E_{Xray}$

w_x weighting term depending on data resolution

$E_{Xray} = \Sigma_{hkl} (|F_{obs}(hkl)-k|F_{calc}(hkl)|)^2$

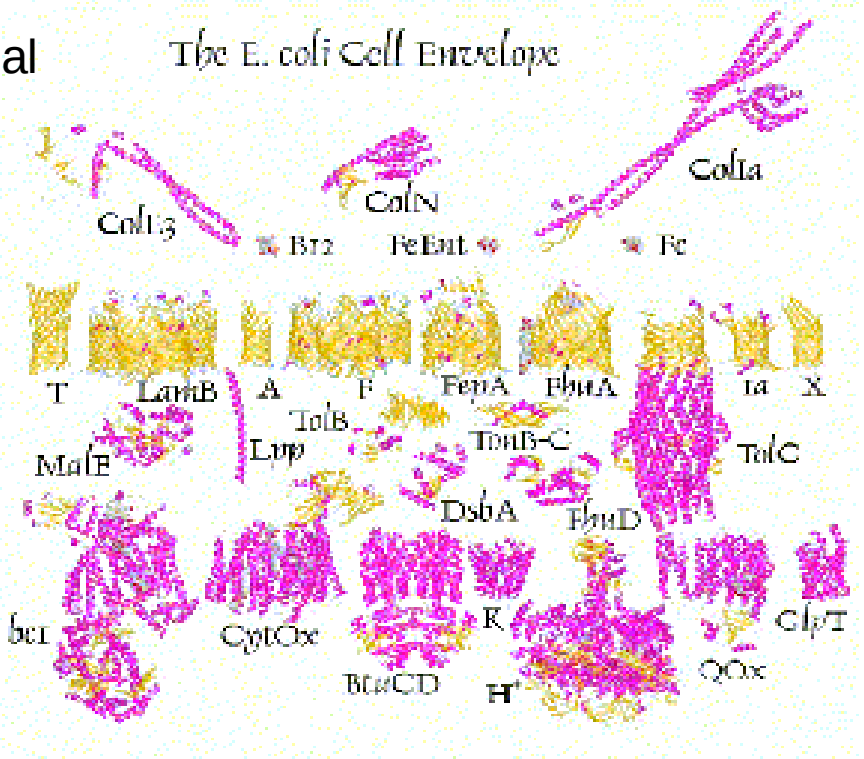
$E_{geometry} = E_{bond} + E_{bond\ angle} + E_{torsion\ angle} + E_{dihedral} + E_{chirality} + E_{van\ der\ Waals}$

Engh&Huber parameters

Small molecule crystallography

Structures

Structural and functional analysis, biochemical and biomedical implications



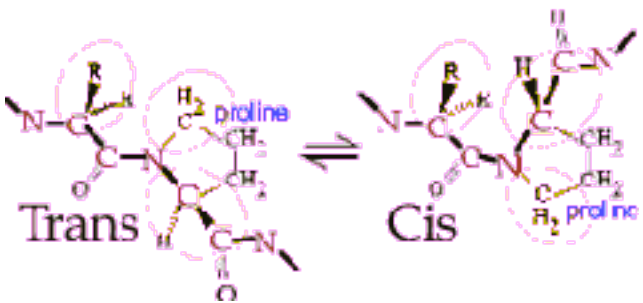
Search
Back
Forward
Reload
Home
Bookmarks
DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
BETA XML files
Current Holdings
24358 Structures
Last Update: 17-Feb-2004
PDB Statistics
PDB
PROTEIN DATA BANK
Did you find what you wanted?
Search the Archive
Query Tutorial
Search
News
Complete News
Newsletter
pdb-l Archive
Subscribe
17-Feb-2004
PDB Mirrors
San Diego Supercomputer Center, UCSD
Rutgers University
Center for Advanced Research in Biotechnology, NIST
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Max Delbrück Center for Molecular Medicine, Germany
RCSB and External Software Links

Protein coordinate file

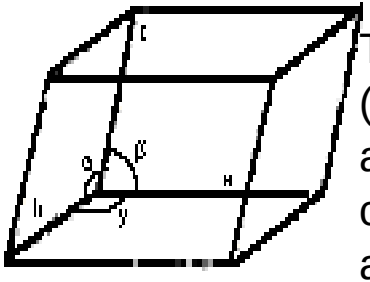

```
HEADER    MEMBRANE PROTEIN                        09-JUL-99   1QD6
TITLE     OUTER MEMBRANE PHOSPHOLIPASE A FROM ESCHERICHIA COLI
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: OUTER MEMBRANE PHOSPHOLIPASE (OMPLA);
COMPND   3 CHAIN: A, B;
COMPND   4 FRAGMENT: RESDIUES 33-45;
COMPND   5 ENGINEERED: YES;
COMPND   6 MOL_ID: 2;
CRYST1   80.067   84.038   95.245  90.00  90.00  90.00 P 21 21 21
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012490  0.000000  0.000000        0.00000
SCALE2      0.000000  0.011899  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010499        0.00000
ATOM      1  N   ALA A  13     -20.453   8.418  34.649  1.00 64.99
ATOM      2  CA  ALA A  13     -20.653   9.129  35.951  1.00 66.50
ATOM      3  C   ALA A  13     -19.548   8.788  36.956  1.00 65.58
ATOM      4  O   ALA A  13     -18.956   7.710  36.890  1.00 65.60
ATOM      5  CB  ALA A  13     -22.027   8.797  36.535  1.00 66.84
ATOM      6  N   VAL A  14     -19.319   9.713  37.892  1.00 64.40
ATOM      7  CA  VAL A  14     -18.259   9.678  38.920  1.00 62.17
ATOM      8  C   VAL A  14     -17.035  10.508  38.525  1.00 59.47
CONECT 4157 3104 4154
MASTER      301    0    4   10   30    0    0    6 4220    4   42    4
END
```

Header

DBREF 1QD6 D 30 269 SWS P00631 PA1_ECOLI 50 289
SEQRES 1 A 13 ALA VAL ARG GLY SER ILE ILE ALA ASN MET LEU GLN GLU
SHETNAM HDS 1-HEXADECANOSULFONIC ACID
FORMUL 5 CA 2(CA1 2+)
HELIX 10 10 SER D 248 TYR D 252 5 5
SHEET 15 B15 TYR D 33 PRO D 34 -1 O TYR D 33 N ALA D 74
LINK CB SER D 144 O3S HDS D 144
CISPEP 1 ASP C 149 PRO C 150 0 1.48



Protein coordinates



The unit cell is composed of asymmetric units (asu). Application of symmetry operators to the asymmetric unit generate the unit cell. The pdb-coordinates describe the contents of the asymmetric unit

```
                a       b         c      alpha  beta   gamma spacegroup    asu
CRYST1   80.067   84.038   95.245  90.00  90.00  90.00 P 21 21 21    8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012490  0.000000  0.000000        0.00000
SCALE2      0.000000  0.011899  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010499        0.00000
ATOM      1  N   ALA A  13     -20.453   8.418  34.649  1.00 64.99           N
ATOM      2  CA  ALA A  13     -20.653   9.129  35.951  1.00 66.50           C
ATOM      3  C   ALA A  13     -19.548   8.788  36.956  1.00 65.58           C
ATOM      4  O   ALA A  13     -18.956   7.710  36.890  1.00 65.60           O
ATOM      5  CB  ALA A  13     -22.027   8.797  36.535  1.00 66.84           C
ATOM      6  N   VAL A  14     -19.319   9.713  37.892  1.00 64.40           N
ATOM      7  CA  VAL A  14     -18.259   9.678  38.920  1.00 62.17           C
ATOM      8  C   VAL A  14     -17.035  10.508  38.525  1.00 59.47           C
```

Protein coordinates

The coordinates are given in Ångström, scale cards give a matrix to convert coordinates to fractional coordinates

```
            a        b        c      alpha  beta   gamma spacegroup    asu
CRYST1   80.067   84.038   95.245  90.00  90.00  90.00 P 21 21 21    8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012490  0.000000  0.000000        0.00000
SCALE2      0.000000  0.011899  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010499        0.00000
ATOM      1  N   ALA A  13     -20.453   8.418  34.649  1.00 64.99           N
ATOM      2  CA  ALA A  13     -20.653   9.129  35.951  1.00 66.50           C
ATOM      3  C   ALA A  13     -19.548   8.788  36.956  1.00 65.58           C
ATOM      4  O   ALA A  13     -18.956   7.710  36.890  1.00 65.60           O
ATOM      5  CB  ALA A  13     -22.027   8.797  36.535  1.00 66.84           C
ATOM      6  N   VAL A  14     -19.319   9.713  37.892  1.00 64.40           N
ATOM      7  CA  VAL A  14     -18.259   9.678  38.920  1.00 62.17           C
ATOM      8  C   VAL A  14     -17.035  10.508  38.525  1.00 59.47           C
```

0.012490 = 1/ 80.067

atom number, atom-type, residue type, chain, residue number, x, y,z, occupancy and B-factor, atom

Judging quality

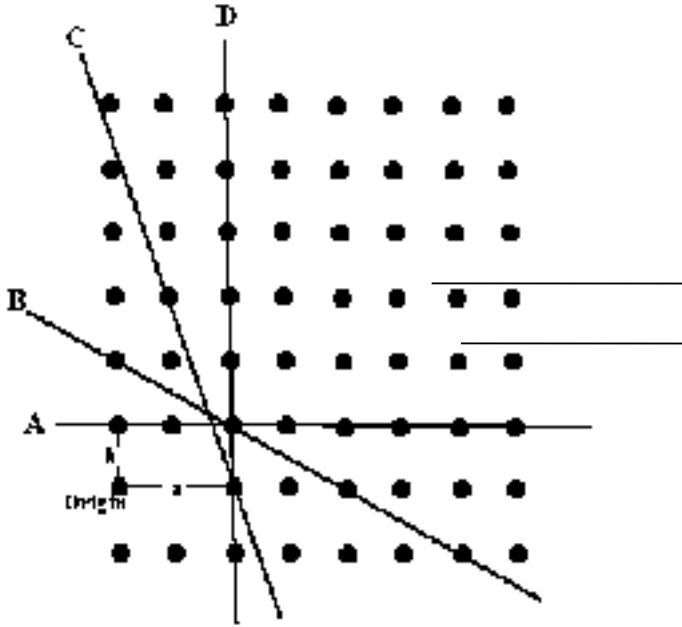
Not one parameter can describe the quality

- Resolution 3.5Å ~0.6 Å
- Overall B-factor
- Data quality
- Model quality

pdb-headers contain information to judge the quality of the structure

- Some statistics on data collection
- Model quality report

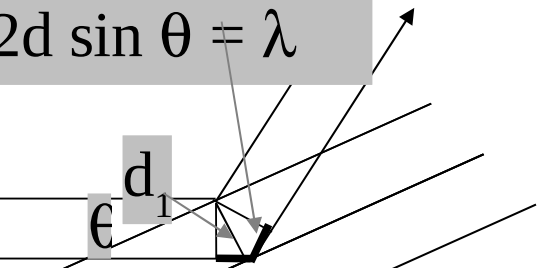
Resolution: Reflection against planes

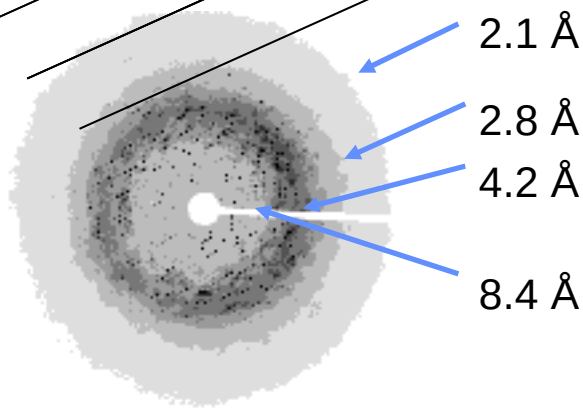


Characterization of a two-dimensional lattice in terms of a set of planes.

path difference

$2d \sin \theta = \lambda$





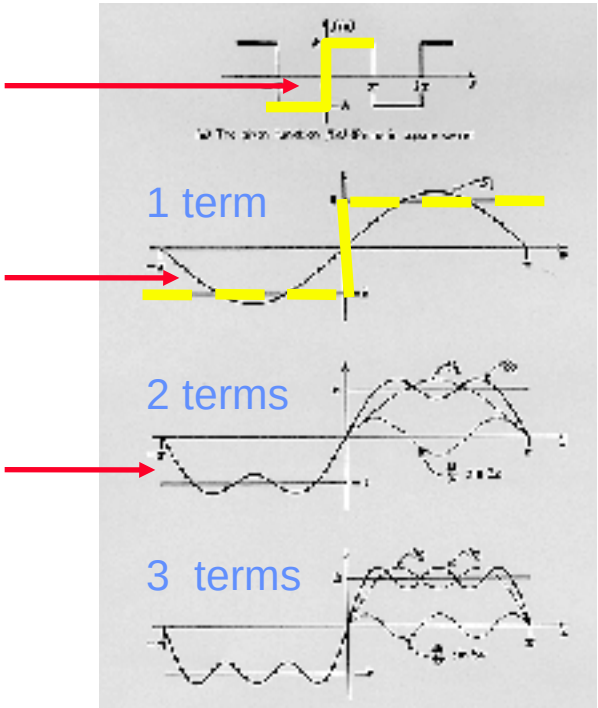
1-D Fourier synthesis of square

Square function which is approximated

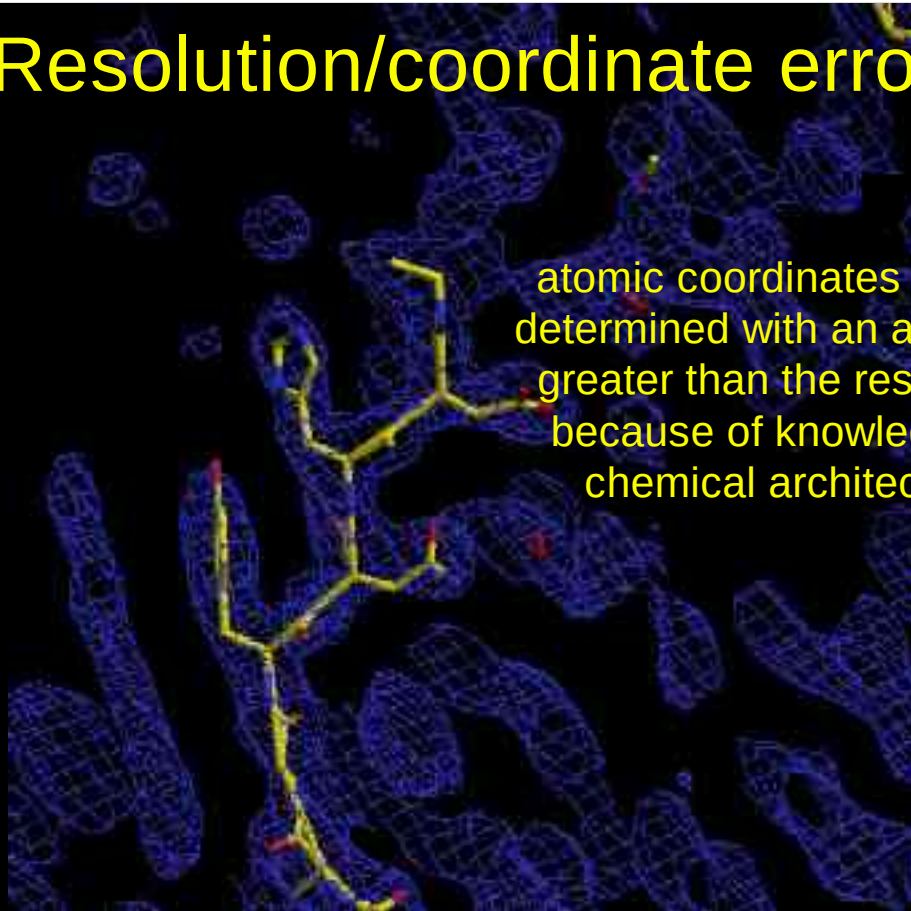
Waves: amplitudes |F| and phases α

Better approximation

More terms



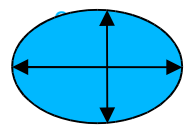
Resolution/coordinate error

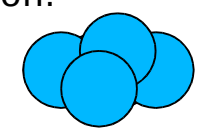


atomic coordinates can be determined with an accuracy greater than the resolution because of knowledge of chemical architecture

B-factors or temperature factors

```
            a       b       c     alpha  beta   gamma spacegroup    asu
CRYST1   80.067   84.038   95.245  90.00  90.00  90.00 P 21 21 21    8
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012490  0.000000  0.000000        0.00000
SCALE2      0.000000  0.011899  0.000000        0.00000
SCALE3      0.000000  0.000000  0.010499        0.00000
ATOM      1  N   ALA A  13     -20.453   8.418  34.649  1.00 64.99           N
ATOM      2  CA  ALA A  13     -20.653   9.129  35.951  1.00 66.50           C
ATOM      3  C   ALA A  13     -19.548   8.788  36.956  1.00 65.58           C
ATOM      4  O   ALA A  13     -18.956   7.710  36.890  1.00 65.60           O
ATOM      5  CB  ALA A  13     -22.027   8.797  36.535  1.00 66.84           C
ATOM      6  N   VAL A  14     -19.319   9.713  37.892  1.00 64.40           N
ATOM      7  CA  VAL A  14     -18.259   9.678  38.920  1.00 62.17           C
ATOM      8  C   VAL A  14     -17.035  10.508  38.525  1.00 59.47
```





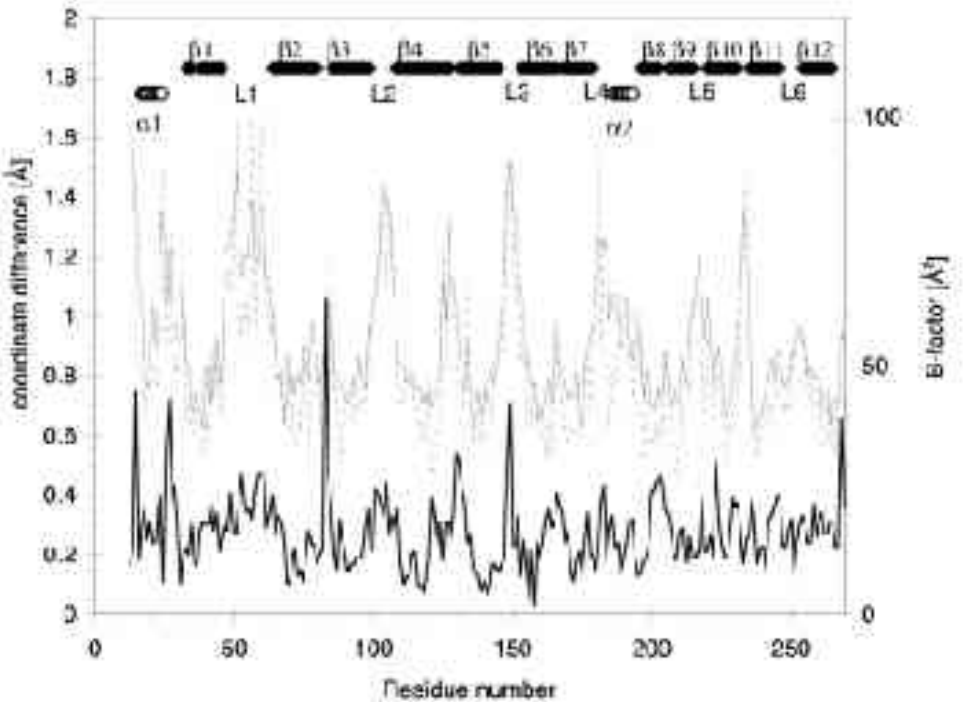
- B-factors: indicates how “ smeared” the electron density is.
- X-ray scattering by a crystal temperature dependent vibrational motion.
- Atoms in unit cells not at exactly the same position.
- Reduce scattering particular at high resolution.
- B related to atomic vibration $B = 8 \pi^2 \times \langle u^2 \rangle$
- B expressed in $Å^2$ ranges from ~2-100$Å^2$

(root mean square displacement ~0.3-1.99Å

B-factors or temperature factors

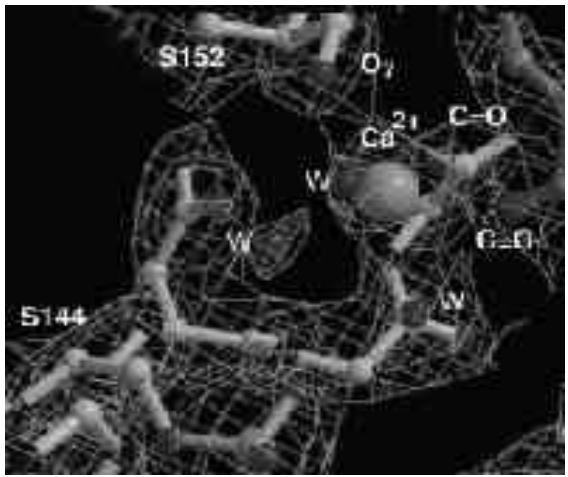
High B-factors correlate with flexible loops.

core regions, lower B-factors



Water, ions, small molecules


```
....
ATOM   4114  OXT PHE D 269     -12.967   4.388  10.176  1.00 42.79
TER    4115      PHE D 269
HETATM 4116 CA    CA     1      17.541  -0.885  29.854  1.00  2.32
HETATM 4117 CA    CA     2      18.918  17.071  18.829  1.00 10.71
HETATM 4118  C1  HDS C 144      11.063   0.452  28.241  1.00 38.03
HETATM 4119  C2  HDS C 144       9.650  -0.100  28.331  1.00 35.97
HETATM 4120  C3  HDS C 144       8.775   0.483  27.240  1.00 28.66
```

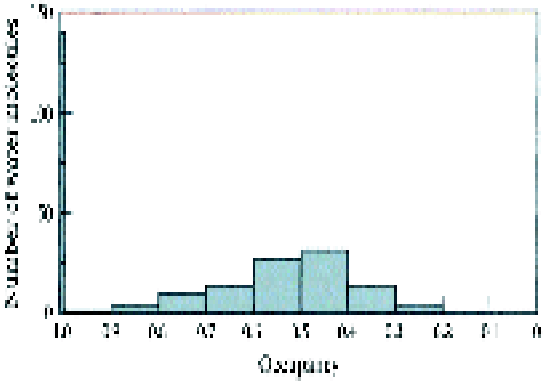


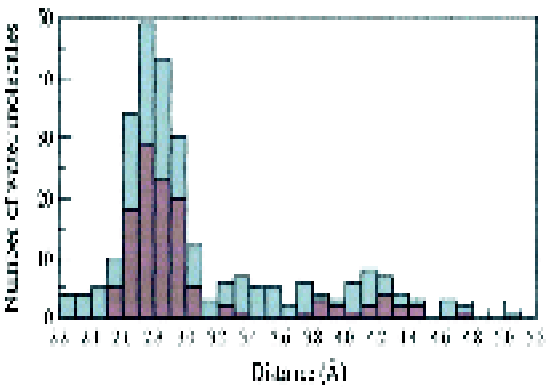
Water molecules spherical density with options for hydrogen bonding

Resolution better than 2.5Å

Ions: large spherical density with coordination

Water, ions, small molecules





Water molecules crucial role in activity or have structural roles

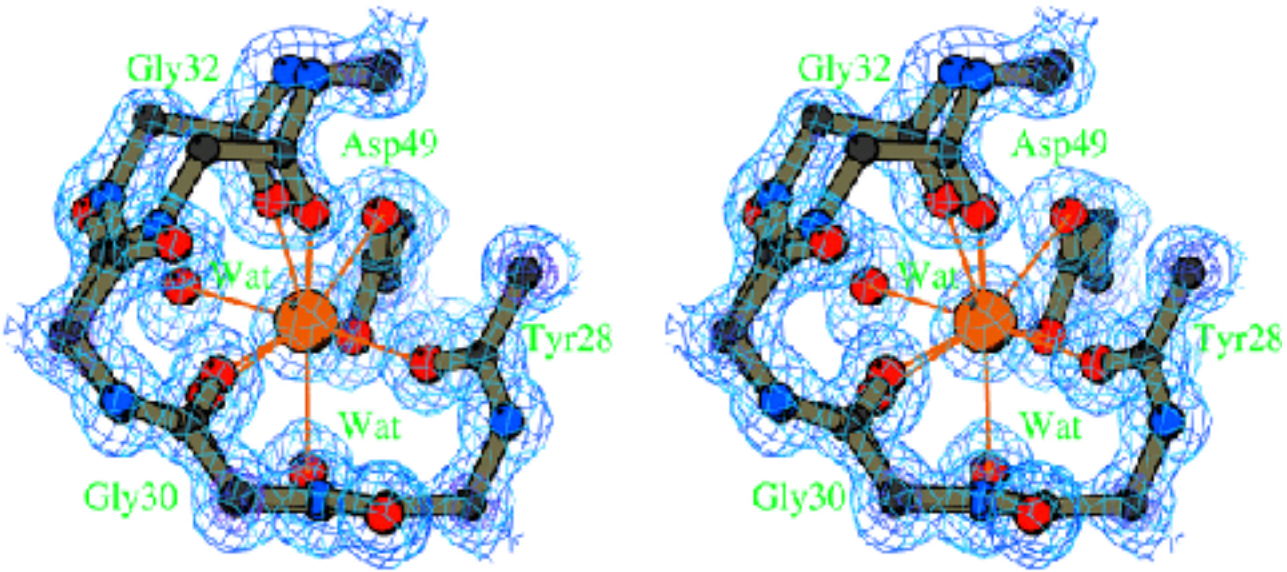
Waters on protein surface

Not completely occupied waters

(Steiner et al. 2001 Acta D)

Alternative conformations

Protein residues or regions may have two or more discrete conformations which, depending on the resolution, can be modelled (Side chains on surfaces)



(Steiner et al. 2001 Acta D)

Table 1. Data collection and refinement statistics

	Monomeric OMPLA		Dimeric OMPLA
	Ca^{2+}	Sm^{3+}	
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Resolution (Å)	30–2.6 (2.61–2.60 Å)	34–3.3 (3.4–3.3 Å)	37.9–2.8 (2.90–2.79 Å)
Unit cell (a, b, c) (Å)	47.83, 77.83, 110.83	75.66, 75.68, 101.98	83.54, 84.97, 91.40
Reflections	47 201	20 753	53 984
Unique reflections	13 975	5791	14 611
Completeness (%)	89.5 (67.3)	97.9 (98.2)	81.0 (57.7)
Completeness at Friedel pairs (%)		91.0 (74.3)	
I/σ(I)	6.3 (1.3)	14.1 (3.5)	8.3 (2.5)
R_{sym} (%)	3.1 (8.1)	5.3 (10.5)	12.1 (25.6)

number of reflections: actual number of intensitys of diffracted X-rays measured
unique reflections, reflections in the asymmetric unit of the reciprokal lattice
redundancy: number of reflections/number of unique reflections
completeness: number of unique reflections/ theoretical number of unique reflections
I/sigma: Intensity of the reflections/noise level
R sym: indicates how accurate identical reflections are measured
$R_{sym} = \Sigma_{hkl} \Sigma_i ||F_i(hkl)|-<|F_i(hkl)|>| / \Sigma_{hkl} \Sigma_i |F_i(hkl)|$

(values in brackets indicate the highest resolution shell)

Data collection

```
REMARK 200  NUMBER OF UNIQUE REFLECTIONS   : 28336
REMARK 200  RESOLUTION RANGE HIGH      (A) : 2.100
REMARK 200  RESOLUTION RANGE LOW       (A) : 20.300
...
REMARK 200 OVERALL.
REMARK 200  COMPLETENESS FOR RANGE     (%) : 74.1
REMARK 200  DATA REDUNDANCY                : 2.690
REMARK 200  R MERGE                    (I) : 0.07700
REMARK 200  R SYM                      (I) : NULL
REMARK 200  <I/SIGMA(I)> FOR THE DATA SET  : 10.9000
REMARK 200
REMARK 200 IN THE HIGHEST RESOLUTION SHELL.
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE HIGH (A) : 2.10
REMARK 200  HIGHEST RESOLUTION SHELL, RANGE LOW  (A) : 2.17
REMARK 200  COMPLETENESS FOR SHELL     (%) : 66.7
REMARK 200  DATA REDUNDANCY IN SHELL       : NULL
REMARK 200  R MERGE FOR SHELL          (I) : 0.26500
REMARK 200  R SYM FOR SHELL            (I) : NULL
REMARK 200  <I/SIGMA(I)> FOR SHELL         : NULL
...
```

Table 1. Data collection and refinement statistics

	Monomeric OMPLA Ca^{2+}
Space group	$P2_12_12_1$
Resolution (Å)	30–2.6 (2.61–2.65 Å)
Unit cell (a, b, c) (Å)	47.53, 97.53, 110.63
Reflections	47 301
Unique reflections	12 675
Completeness (%)	99.5 (97.3)
Completeness at Friedel pairs (%)	
$I/\sigma(I)$	8.3 (1.3)
R_{sym} (%)	3.1 (8.1)

Refinement	Monomeric OMPLA Ca^{2+}
Protein atoms	2094
R-factor (%)	22.2 (33.6)
R_{free} (%)	26.6 (38.5)
B-factor (Å^2)	45.9
Non-protein atoms	51
Error estimate (Å)	0.4
r.m.s.d. bond lengths (Å)	0.008
r.m.s.d. bond angle (deg.)	1.2
r.m.s.d. dihedrals (deg.)	26.4
r.m.s.d. improper dihedrals (deg.)	1.23

Top numbers in parentheses are for the outermost shell of data.

Protein atoms: parameter observation ratio
xyzB 8376/9975 = 0.84

Geometric restrains

Model calculated structure factors:

$R = \Sigma_{hkl} \; ||Fobs(hkl)|-k|F_{calc}(hkl)|| / \Sigma_{hkl} |F(hkl)|$

Rfree, test set not used in refinement

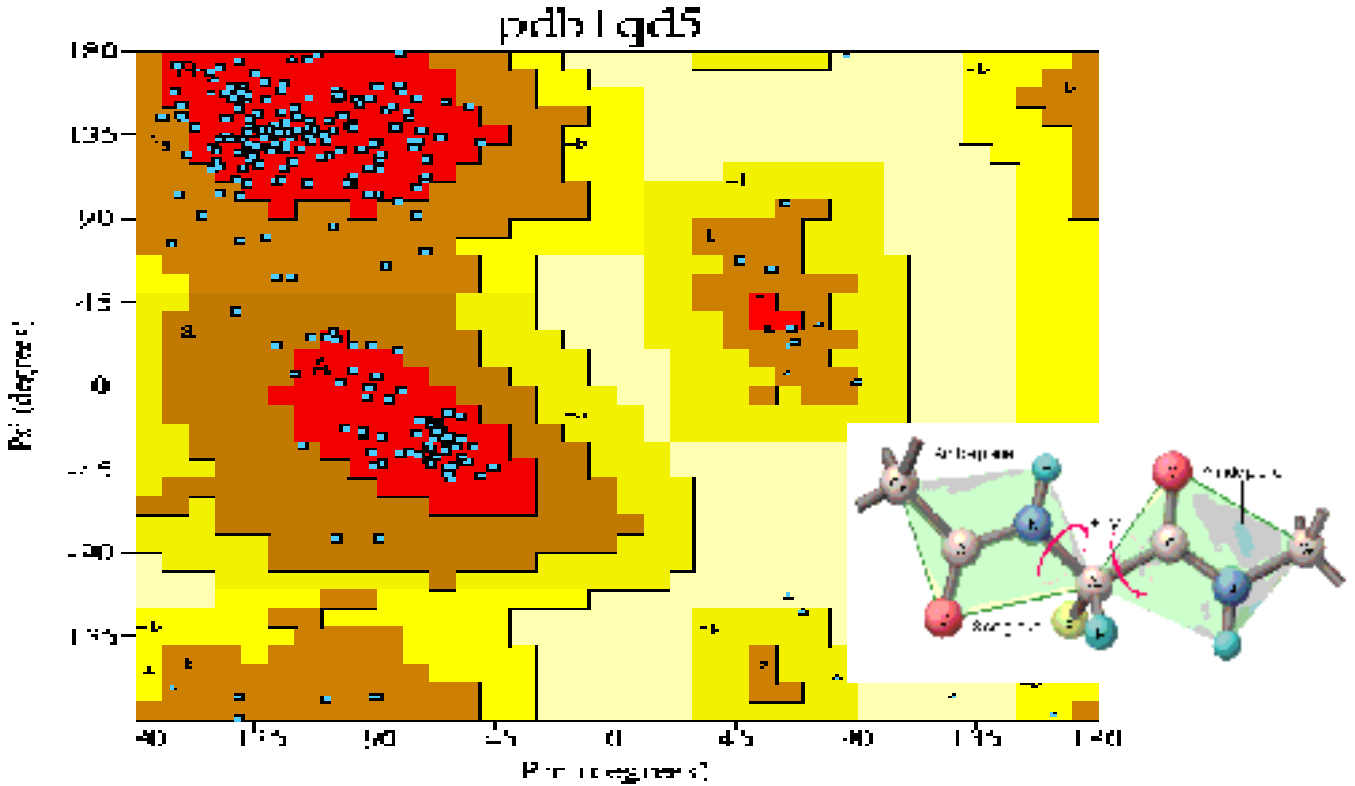
Refinement

```
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.10
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 20.30
REMARK   3   DATA CUTOFF            (SIGMA(F)) : 0.000
REMARK   3   DATA CUTOFF HIGH         (ABS(F)) : 100000.000
REMARK   3   DATA CUTOFF LOW          (ABS(F)) : 0.0000
REMARK   3   COMPLETENESS (WORKING+TEST)   (%) : 74.1
REMARK   3   NUMBER OF REFLECTIONS             : 28947
REMARK   3
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD           : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION   : RANDOM
REMARK   3   R VALUE            (WORKING SET) : 0.226
REMARK   3   FREE R VALUE                      : 0.283
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 8.300
...
REMARK   3  ESTIMATED COORDINATE ERROR.
REMARK   3   ESD FROM LUZZATI PLOT        (A) : 0.27
REMARK   3   ESD FROM SIGMAA              (A) : 0.300
...
```

Model quality

- Does the model reflect our chemical knowledge
  - location of water molecules
  - biochemical evidence
  - comparison with homologous proteins
- Refinement with geometrical constraints
  - bond lengths, angles, dihedrals etc.
  - high resolution, unrestrained refinement
- Indicators not used in refinement
  - Ramachandran plot
- Violations
- Electron density quality

Ramachandran plot

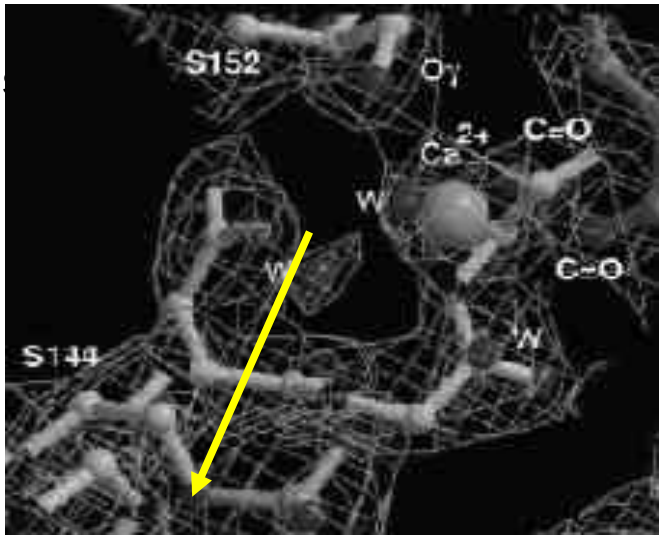


Violations

```
REMARK 500 THE STEREOCHEMICAL PARAMETERS OF THE FOLLOWING RESIDUES
REMARK 500 HAVE VALUES WHICH DEVIATE FROM EXPECTED VALUES BY MORE
REMARK 500 THAN 6*RMSD (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN
REMARK 500 IDENTIFIER; SSEQ=SEQUENCE NUMBER; I=INSERTION CODE).
REMARK 500
REMARK 500 STANDARD TABLE:
REMARK 500 FORMAT: (10X,I3,1X,2(A3,1X,A1,I4,A1,1X,A4,3X),F6.3)
REMARK 500
REMARK 500 EXPECTED VALUES: ENGH AND HUBER, 1991
REMARK 500
REMARK 500  M RES CSSEQI ATM1   RES CSSEQI ATM2   DEVIATION
REMARK 500    MET A  22   CE    MET A  22   SD    -0.117
REMARK 500    MET B  22   CE    MET B  22   SD    -0.065
...
REMARK 500  M RES CSSEQI ATM1   ATM2   ATM3
REMARK 500    LEU C  71   N    -  CA   -  C   ANGL. DEV. = -7.4 DEGREES
REMARK 500    LEU C  88   N    -  CA   -  C   ANGL. DEV. =-10.2 DEGREES
REMARK 500    LEU C 158   N    -  CA   -  C   ANGL. DEV. = -7.6 DEGREES
...
REMARK 500    LEU D  71   N    -  CA   -  C   ANGL. DEV. =-10.0 DEGREES
REMARK 500    LEU D  88   N    -  CA   -  C   ANGL. DEV. = -9.8 DEGREES
REMARK 500    LEU D 120   N    -  CA   -  C   ANGL. DEV. = -7.7 DEGREES
```

Missing atoms

```
REMARK 470 MISSING ATOM
REMARK 470 THE FOLLOWING RESIDUES HAVE MISSING ATOMS(M=MODEL
NUMBER;
REMARK 470 RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE
NUMBER;
REMARK 470 I=INSERTION CODE):
REMARK 470   M RES CSSEQI  ATOM
REMARK 470     SER C 144    OG
REMARK 470     SER D 144    OG
```



Model quality programs

- **Procheck**
- What if
- MD programs
- SF-Check
- O (Xray model building program)
- Oops

Crystals, structure and interpretation