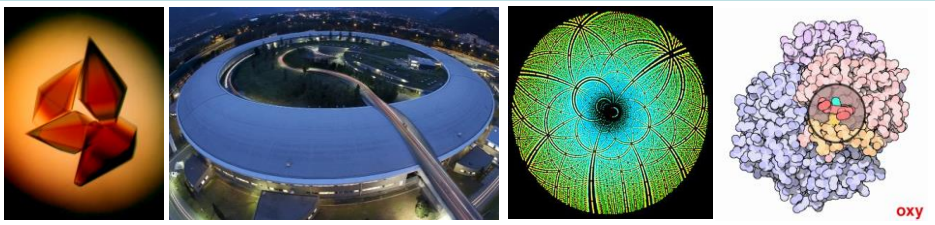
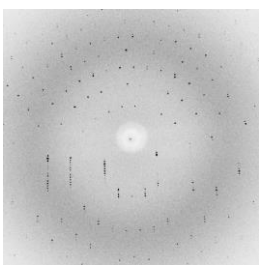


Data phasing and crystallographic refinement

Patrice Gouet, Université de Lyon

A large part of the slides from C. Mayer and J.L. Ferrer, Renafobis 2018

Observed amplitudes from merged intensities



Data collection
$I_{obs}(hkl) = F^2(hkl)$

Data scaling and merging →

Unique reflections in the a.u.

h	k	l	F	SIGF
0	0	1	0.00	0.00
0	0	2	-1.00	0.00
0	0	3	-1.00	0.00
0	0	4	101.12	6.29
0	0	5	5087.18	868.91
0	0	6	-1.00	868.91
0	0	7	-1.00	868.91
0	0	8	712.77	26.26
0	0	9	251303.12	24365.59
0	0	10	-1.00	24365.59
0	0	11	-1.00	24365.59
0	0	12	374.42	11.63
.....				
.....				
36	20	1	239.06	4.01

PHI_{obs} missing

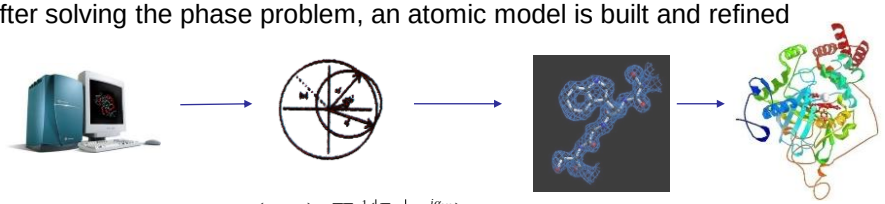
resolution limit

Methods used to solve the phase problem

Three methods can be used to **obtain phase information** and determine macromolecular structures by X-ray crystallography

1. Molecular Replacement (**MR**) with the help of a known structure
2. Multiple Isomorphous Replacement (**MIR**) with the help of heavy atoms
3. Multiple or Single Anomalous Dispersion (**MAD, SAD**) by using a significant dependence of some diffusion factors to wavelength

After solving the phase problem, an atomic model is built and refined



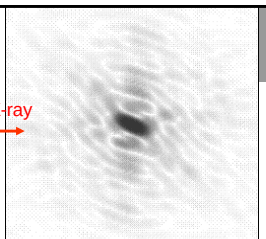
Data treatment and phasing $\rho(x,y,z) = TF^{-1}(|F_{hkl}| \times e^{i\alpha_{hkl}})$ electron density map 3D structure

Remark: one can also combine phasing methods (SIRAS, MIRAS)

1. Phasing by molecular replacement

Crystal of an unknown structure

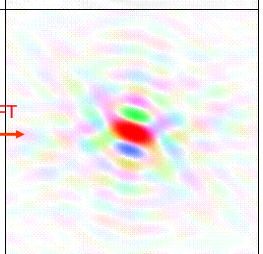
X-ray



Experimental **Fobs**

Homologous of coordinates x_i, y_i, z_i

FT

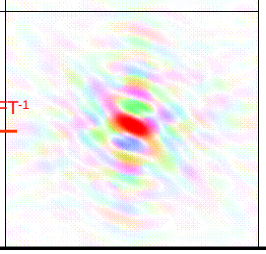


e.g. a protein of known structure homologous in sequence to the query (Id>30%) and detected from the PDBaa with BLAST

Calculated Fcalc, **ϕcalc**

Resulting electron density map

FT^{-1}



Calculation of FT^{-1}(**Fobs, ϕcalc**)

Remark : calculation of FT^{-1}(**Fobs-Fcalc, ϕcalc**) gives the cat's tail

1. The Patterson function in Molecular Replacement

- The Patterson function is the FT^{-1} of intensities

$$P(\mathbf{u}) = \frac{1}{V_c}\sum_{\mathbf{h}} |F(\mathbf{h})|^2 \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}) \quad \text{with} \quad \mathbf{u} = u\mathbf{a} + v\mathbf{b} + w\mathbf{c}$$

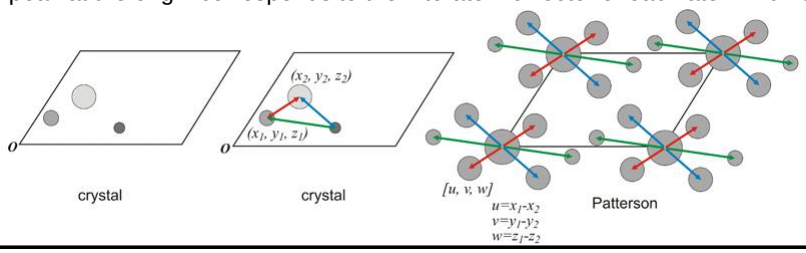
- It is also equivalent to the electron density convolved with its inverse:

$$P(\mathbf{u}) = \rho(\mathbf{r}) \otimes \rho(-\mathbf{r})$$

- Considering that a macromolecular structure is made of point atoms of Z electrons

$$\rho(\mathbf{r}) = \sum_j Z_j \delta(\mathbf{r} - \mathbf{r}_j) \qquad P(\mathbf{u}) = \sum_j \sum_{j'} Z_j Z_{j'} \delta(\mathbf{u} + \mathbf{r}_j - \mathbf{r}_{j'})$$

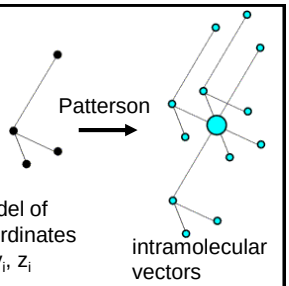
The unit cell of the Patterson function contains N^2-N+1 peaks of height $Z_jZ_{j'}$ positioned on interatomic vectors **r**j'-**r**j. This function has **centrosymmetry** and the high peak at the origin corresponds to the interatomic vector of each atom with itself



1. The rotation function

Let's suppose that the identified PDB **model** is made of four atoms and give it an arbitrary orientation

→ a Patterson function of the model (**$P_B$model)** is calculated in a crystal with arbitrary large unit cell parameters

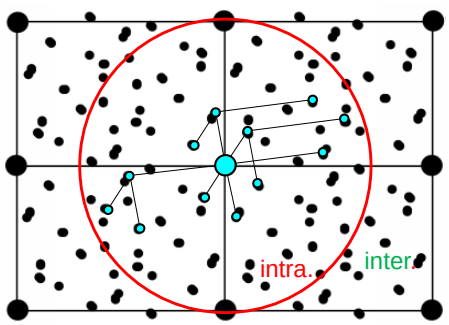


$P_B$model is superposed onto the experimental **$P_A$native** and a maximum overlay ***R*([R])** is calculated within a region of radius R centered at uvw=0

$$R([\mathrm{R}]) = \int_U \underbrace{P_A(u)}_{\text{native}} \underbrace{P_B([C]u)}_{\text{model}} du$$

FT^{-1}(I$_{obs}$(hkl))

Intramolecular vectors are kept within R
intermolecular vectors are excluded

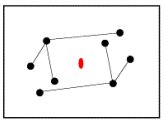


The orientation of the model in the unit cell is found at the **maximum** of the rotation function ***R*([R])**

1. The translation function

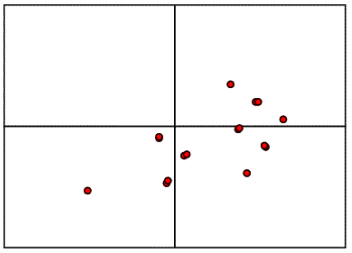
The oriented model is arbitrary positioned in the real unit cell (**a**) and intermolecular vectors ($P_{1,2}$) are calculated (**b**).
$P_{1,2}$ is superimposed onto the experimental Patterson and a translation function *T* is calculated (**c**)

a

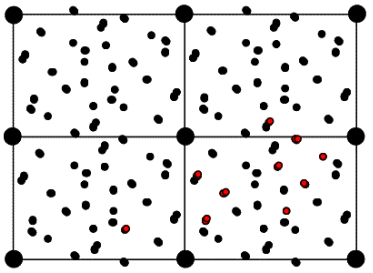


Space group P2

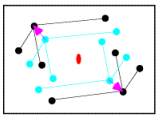
b



c



d



$$T(\mathbf{t}) = \int_V P_{1,2}(\mathbf{u}, \mathbf{t}) \times P(\mathbf{u})\, d\mathbf{u}$$

calculated

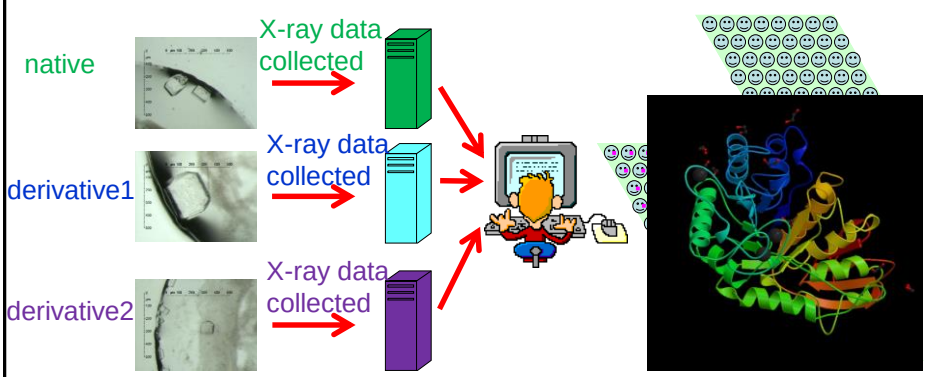
The oriented model is then moved along u,v,w (**d**) giving rise to a new set of intermolecular vector.

The solution in translation will correspond to the maximum of function *T*
→ The homologuous structure is **oriented** and **positioned** in the unit cell

2. The multiple isomorphous replacement (MIR) phasing method

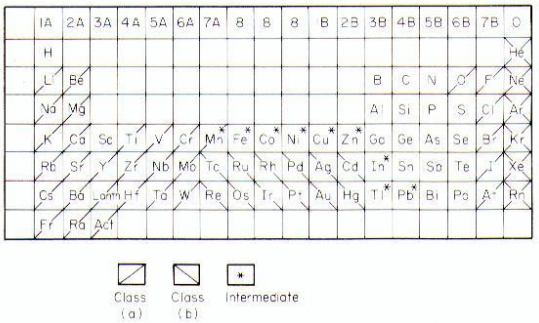
- **Principle**: Incorporation to the macromolecular structure of heavy atoms (Sm Z_e=62 or Au Z_e=79 or…) whose positions will be determined from Patterson maps
- **Method**: soaking crystals with various heavy-atom salt solutions ($AuCl_3$, $KAu(CN)_2$, $Sm(NO_3)_3 \cdot 6H_2O$,…)

  Then data collection, processing and phasing



2. Crystal soaking with heavy-atom salt solutions

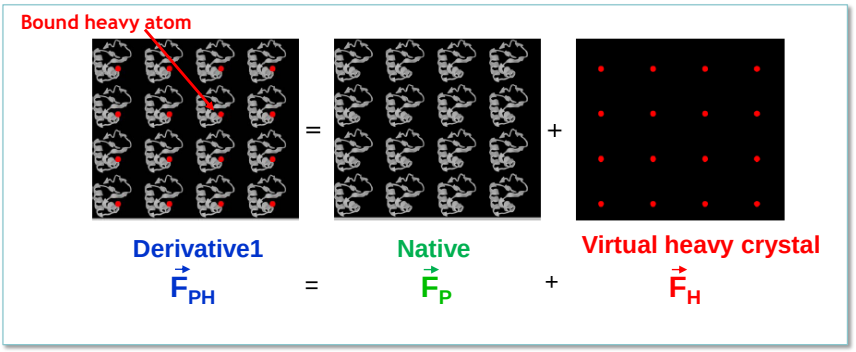
- Protein crystals are composed of 27-65% solvent and can be soaked in heavy-atom salts that will diffuse into the crystal assembly
- The reaction of heavy-atom salts with protein residues is rather aspecific.
- Yet, one may distinguish **class A** heavy atoms (cationic alkali or lanthanides Sm_{62}, U_{92}…) which bind to "hard" electronegative ligands (**Glu**, **Asp, Ser…**)
- from **class B** heavy atoms (Pt_{78}, Au_{79}, Hg_{80}…) which bind to "soft" polarizable ligands (**Met**, **Cys, His**…)



- Most derivatives are hydrophilic and bind to protein surfaces. Some such as EtHgCl are hydrophobic and bind in protein cavities.

→ **Lot** of soaking experiments needed to find the "good" derivative

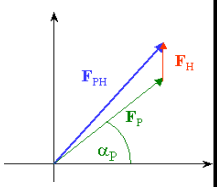
2. First heavy-atom derivative crystal



- **Isomorphous replacement**: Heavy atoms bind to the protein crystal without affecting its assembly (unit cell parameters remain similar)

$$\vec{F}_P(hkl) = \underbrace{\Sigma_P f_P e^{i2\pi(hx_P + ky_P + lz_P)}}_{|F_P| e^{i\varphi_P}}$$

$$\vec{F}_{PH}(hkl) = \underbrace{\Sigma_P f_P e^{i2\pi(hx_P + ky_P + lz_P)}}_{|F_P| e^{i\varphi_P}} + \underbrace{\Sigma_H f_H e^{i2\pi(hx_H + ky_H + lz_H)}}_{|F_H| e^{i\varphi_H}}$$



2. Determination of the coordinates of the bound heavy atoms

- We know $|\mathbf{F_P}|$ and $|\mathbf{F_{PH}}|$ of each reflection hkl collected (resolution likely to be lower for the soaked derivative crystal)

Calculation of an isomorphous difference Patterson map

$\Delta|F|_{iso} = |F_{PH}| - |F_P|$ and $(\Delta|F|_{iso})^2 \sim |F_H|^2$

$P(uvw) = FT^{-1}[(\Delta|F|_{iso})^2] = \sum_{uvw} (\Delta|F|_{iso})^2 e^{-i2\pi(hu+kv+lw)}$

→ This Patterson function defines a map whose peaks are **interatomic vectors between** "only" **heavy atoms**

- Programs such as **SHELXD** can derive the atomic positions (coordinates) in the crystal from the coordinates of the maxima of a Patterson map (**direct method**). The Patterson map has N^2-N+1 peaks and this is very difficult for $N_{atom} > 50$.

→ **Direct methods** are not suitable to solve macromolecular structures but are very likely to determine heavy atom positions **(x_H, y_H, z_H)**

→ $|F_H|e^{i\varphi_H}$ can then be calculated: $\vec{F}_H(hkl) = \Sigma_H f_H e^{i2\pi(hx_H + ky_H + lz_H)}$

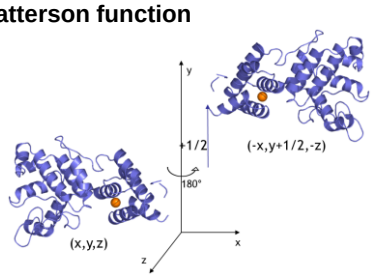
2. Determination of the coordinates of the bound heavy atoms

Harker section

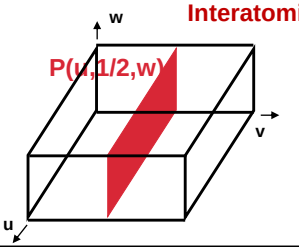
Certain locations (lines or planes) in the Patterson Function contain information about the interatomic vectors between equivalent atoms (atoms related by symmetry operations) and can **help to solve the Patterson function**

Exemple : space group $P2_1$
(2 equivalent positions x,y,z and –x, y+1/2, -z)

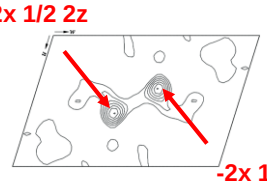
	x y z	-x y+1/2 -z
x y z	0 0 0	-2x +1/2 -2z
-x y+1/2 -z	2x -1/2 2z	0 0 0



Interatomic vectors located in the Harker section v = 1/2



2x 1/2 2z



u = ±2x v = ±1/2 w ± 2z

2 equivalent peaks

1 heavy atom

-2x 1/2 -2z

2. Centric reflections hkl

Now we know: F_{PH} in magnitude, F_P in magnitude, F_H in magnitude and phase

- A **reflection** is termed **centric** if there is a reciprocal space symmetry operator which maps it onto its Friedel mate (-h,-k,-l).

Exemple for the space groupe **$P2_1$**(screw axis along **b**)
x,y,z equivalent to –x,y+1/2,-z
h,k,l equivalent to –h,k,-l
and **h0l** reflections are **centric**

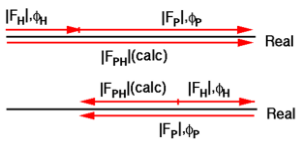
For the space group **$P2_12_12_1$** centric reflections are on planes 0kl, h0l, 0kl and some space groups such as **R3** have no centric reflections.

- Centric reflections have only two phase possibilities: **0** and **π**

Assuming that φ_H is known and that φ_P is shifted by 0 or π in phase

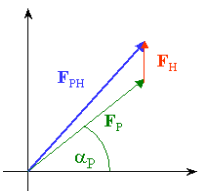
If $\varphi_P = \varphi_H$ $\quad |\boldsymbol{F}_{PH}|$ (calc) $= |\boldsymbol{F}_P| + |\boldsymbol{F}_H|$

If $\varphi_P = \varphi_H + \pi$ $\quad |\boldsymbol{F}_{PH}|$ (calc) $= |-|\boldsymbol{F}_P| + |\boldsymbol{F}_H||$

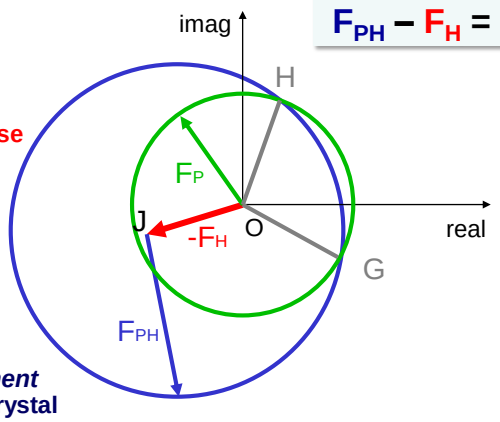


Comparison between the known $|\boldsymbol{F}_{PH}|$(obs) and $|\boldsymbol{F}_{PH}|$(calc) give the phase φ_P of each centric reflections → **great but** generally not sufficient to solve the phase problem.

2. Acentric reflections hkl



We know:
F_P in magnitude
F_H in magnitude and phase
F_{PH} in magnitude



$F_{PH} - F_H = F_P$

SIR

Single Isomorphous Replacement
1 native crystal + 1 derivative crystal

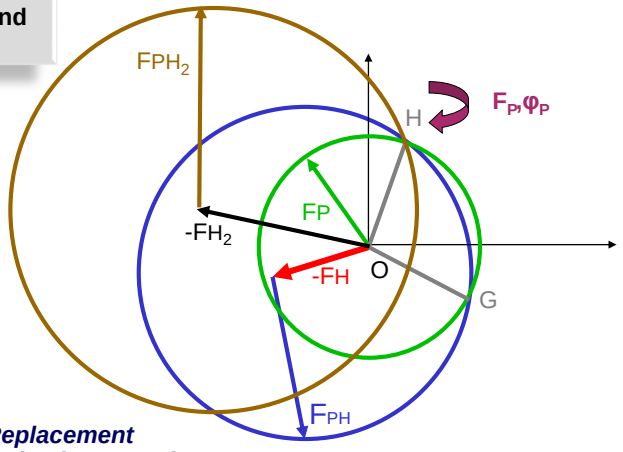
→ Phase ambiguity for acentric reflections: φ_P can be the H or the G solution

2. Acentric reflections hkl

Removing the phase ambiguity with a second derivative crystal

$F_{PH} - F_H = F_P$

$F_{PH2} - F_{H2} = F_P$



MIR

Multiple Isomorphous Replacement
1 native + at least two derivative crystals

→ Heavy atoms **H** and **H2** must bind at **different sites** to have different structure factors so as to remove the phase ambiguity
→ one usually uses **class A** and **class B** heavy atoms

3. Experimental phasing by anomalous dispersion

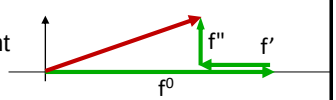
Correction to the **diffusion factor f** for a $\lambda_{incident}$ **dependent** anomalous scatterer

$$f = {}^{0}f + {}^{\lambda}f' + i\,{}^{\lambda}f''$$

Absorption component (imaginary)

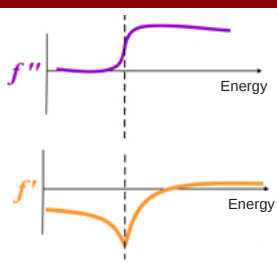
Number of electrons at θ=0

Dispersion component (real and negative)



At $\lambda_{incident}$ around 1 Å, anomalous signal is negligible for H, C, N, O and can be significant for heavy atoms

Anomalous signal occurs near electronic absorption edge

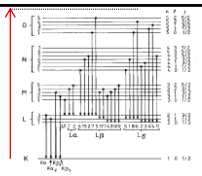


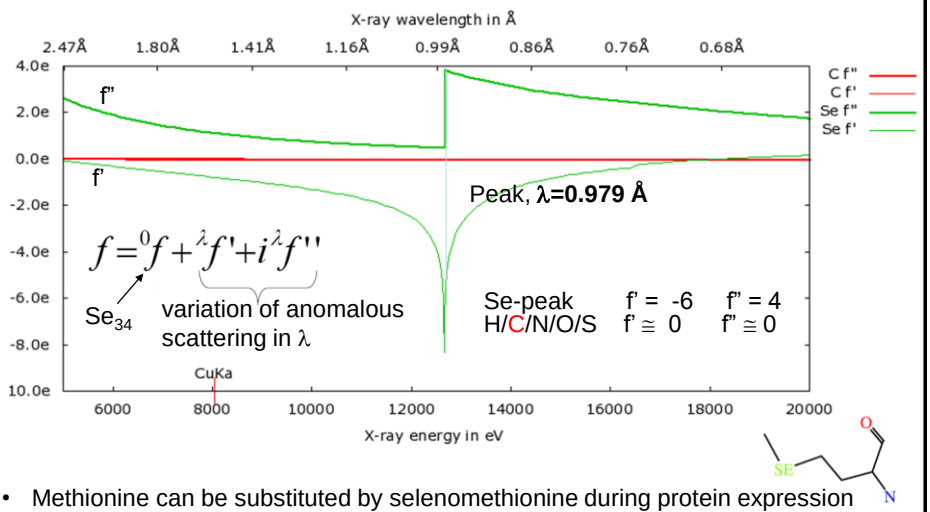
At **f"** maximum, the wavelength corresponds to the energy required to eject inner electrons from their K, LI, LII or LIII shells

$$f'(E) = \frac{2}{\pi}\int_0^\infty \frac{E' f''(E')}{E^2 - E'^2}\,dE'$$

→ The imaginary component **f"** can be **measured** by **X-ray fluorescence** and the unmeasurable real component **f'** is **calculated** by the **Kramer-Kronig** relation.
→ f" inflexion point and minimal f' occur at the same $\lambda_{incident}$

Anomalous signal of Selenium at K-edge



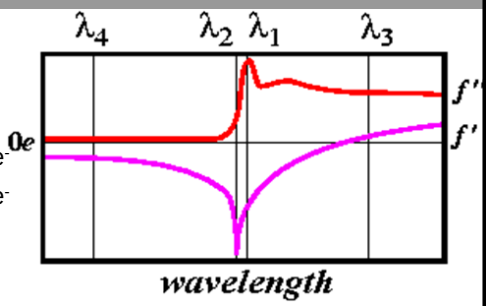


- Methionine can be substituted by selenomethionine during protein expression

3. MAD data collection at the Se-K edge

- Data collection from one SelenoMet protein crystal at >3 different wavelengths

1. λ_1=0.979 Å, Se-peak f'= -6e$^-$, f''= 4e$^-$
2. λ_2=0.983 Å , Se-inflection f'= -8e$^-$, f''= 2e$^-$
3. and high remote λ_3 or/and low remote λ_4



Advantages of MAD

- no problem of isomorphism (only 1 crystal needed)
- no problem of occupation (correct S → Se substitution can be checked by Mass Spectrometry)
- Selenomethionine is an anchor for structure building

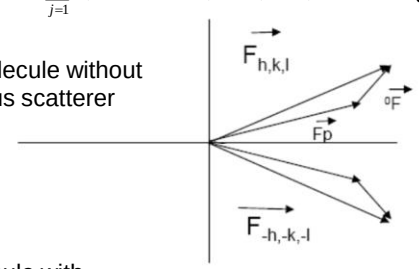
Disadvantages

- Anomalous signal is weak
- Presence of ~1 SelenoMet every ~80 protein residues for phasing
- Access to a synchrotron source is mandatory to select the incident wavelength

3. The breakdown of the Friedel's law for the mates hkl and -h-k-l

$$F_{hkl}\exp i\phi_{hkl} = \sum_{j=1}^{N} f_j[\exp i2\pi(hx_j + ky_j + lz_j)] = \overrightarrow{Fp} + \overrightarrow{F}_{anomalous\ scatterer}$$

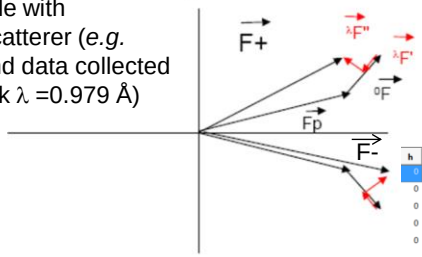
Macromolecule without anomalous scatterer



$|F_{h,k,l}| = |F_{-h,-k,-l}|$

and $I_{hkl} = I_{h\text{-}k\text{-}l}$ (*Friedel True*) reflections have centrosymmetry

Macromolecule with anomalous scatterer (*e.g.* SelenoMet and data collected at the Se-peak λ =0.979 Å)

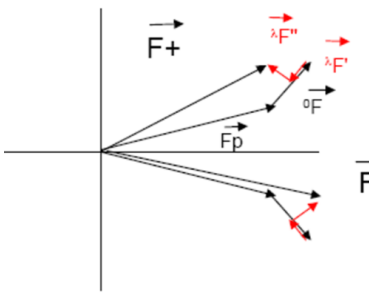


$|F_{h,k,l}| \neq |F_{-h,-k,-l}|$

and $I_{hkl} \neq I_{h\text{-}k\text{-}l}$ (*Friedel False*)

h	k	l	F	SIGF	DANO	SIGDANO	F(+)	SIGF(+)	F(-)	SIGF(-)
0	0	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0	0	2	-1.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00
0	0	3	-1.00	0.00	-1.00	0.00	-1.00	0.00	0.00	0.00
0	0	4	101.12	6.29	0.00	0.00	100.92	9.00	100.05	9.11
0	0	5	5087.18	868.91	5087.18	868.91	5087.18	868.91	5004.75	871.44

- The difference between centrosymmetric F(h,k,l) and F(-h,-k,-l) is ΔF_{ANO}



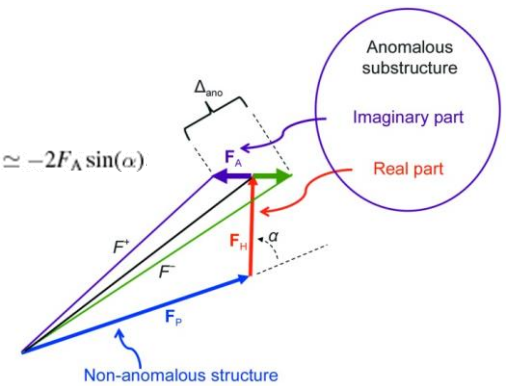
$|F_{h,k,l}| \neq |F_{-h,-k,-l}|$

and $I_{hkl} \neq I_{h\text{-}k\text{-}l}$ (*Friedel False*)

Macromolecule with anomalous scatterer

$$\Delta_{ano} \equiv F^{+} - F^{-} \simeq -2F_A \sin(\alpha)$$

Better to start collecting at λ_{peak}

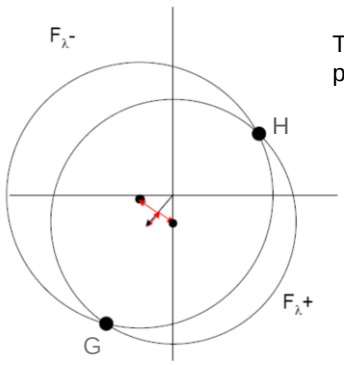


- A **Patterson anomalous difference** $FT^{-1}[(\Delta|F|_{ANO})^2]$) is calculated to determine the positions of the anomalous scatterer (*e.g.* with SHELXD)

$\vec{F_\lambda} = \vec{Fp} + \overrightarrow{F_0 - F_\lambda' + iF_\lambda''}$ $\quad \vec{Fp} = \vec{F_\lambda} - \overrightarrow{(F_0 - F_\lambda' + iF_\lambda'')}$ ←

protein | anomalous scatterer | **only amplitude known**; represented with a circle of radius F_λ

positions of the **anomalous scatterer** in the unit cell estimated with an anomalous Patterson; amplitude and phase known.



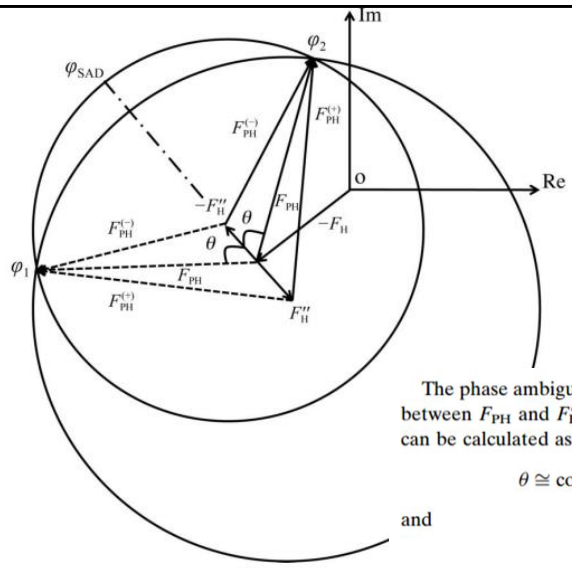
Two possible phases ϕ_H and ϕ_G for protein **Fp.**

→ be **MAD** (two more data sets collected at $\lambda_{inflexion}$ and λ_{remote})

→ or remain **SAD (**phase ambiguity broken through a density-modification procedure**)**

SAD: one takes the mean of the two possible phases and use it with a weight *m*:

$$\boldsymbol{Fp} = m\,|F_P|\,\exp(i\phi_{mean})$$



The phase ambiguity arises from the existence of an angle θ between F_{PH} and F_H'', related to $|F_{PH}^{(+)}|$, $|F_{PH}^{(-)}|$ and $|F_H''|$, which can be calculated as (Blundell & Johnson, 1976)

$$\theta \cong \cos^{-1}\{[|F_{PH}^{(+)}| - |F_{PH}^{(-)}|]/2|F_H''|\},$$

and

$$\varphi = \varphi_{SAD} \pm \theta,$$

Figure of merit (fom)

$$\boldsymbol{Fp} = \cos\theta\,|F_P|\,\exp(i\phi_{SAD})$$

Chen *et al.* • Direct phase selection of initial phases from SAD *Acta Cryst.* (2014). D**70**, 2331–2343

3. Phases improvement by solvent flattening

Calculating a **mask** where the protein zone is set to 1 and the solvent to 0

```
11212321112212112
21245475122256232
12367867784567631
32867559877685432
12325677987754322
12234221221132141
```

Initial electron density map

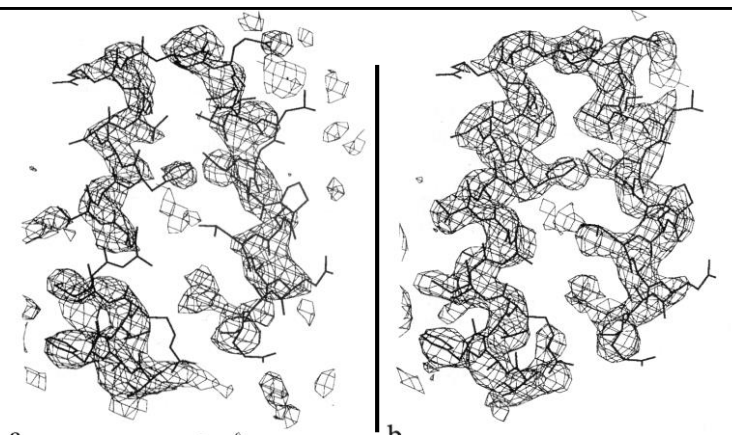
```
00000000000000000
00011111000011000
00111111111111100
00111111111111000
00001111111110000
00000000000000000
```

mask

Determination of the average density in the solvent mask region set to 0 (ρ_s = 2).

```
22222222222222222
22245475222256222
22367867784567622
22867559877685222
22225677987752222
22222222222222222
```

Better phases will be obtained in turn by Fourier Transform and cycling (program *Parrot*) . **The figure of merit** of each hkl must tend towards 1.



a SAD map at 3.3 Å

b SAD map after solvent flattening

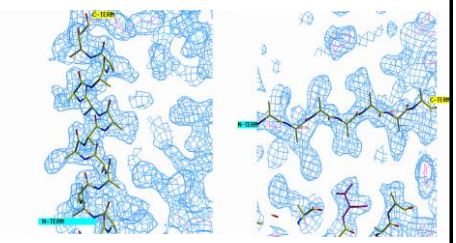
The procedure of solvent flattening can be improved in case of **non-crystallographic** symmetry.

The density of the protein is considered to be identical between regions linked together by non-crystallographic symmetry and is **averaged** within masks (powerful for **icosahedral viruses**)

$$\rho_{avg}(x) = \sum_{i=1}^{N} \frac{1}{N} M_i(x) \sum_{j=1}^{N} \rho(x_{ij})$$

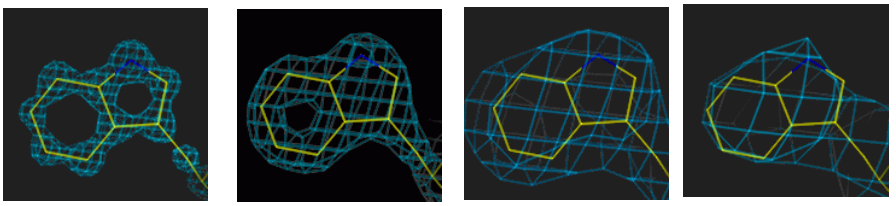
Model Building: Steps in making the first trace in electron density

- Generating Ca chain trace
- Recognize secondary structures
- Identifying chain direction
- Sequence assignment
- Add water molecules, etc...



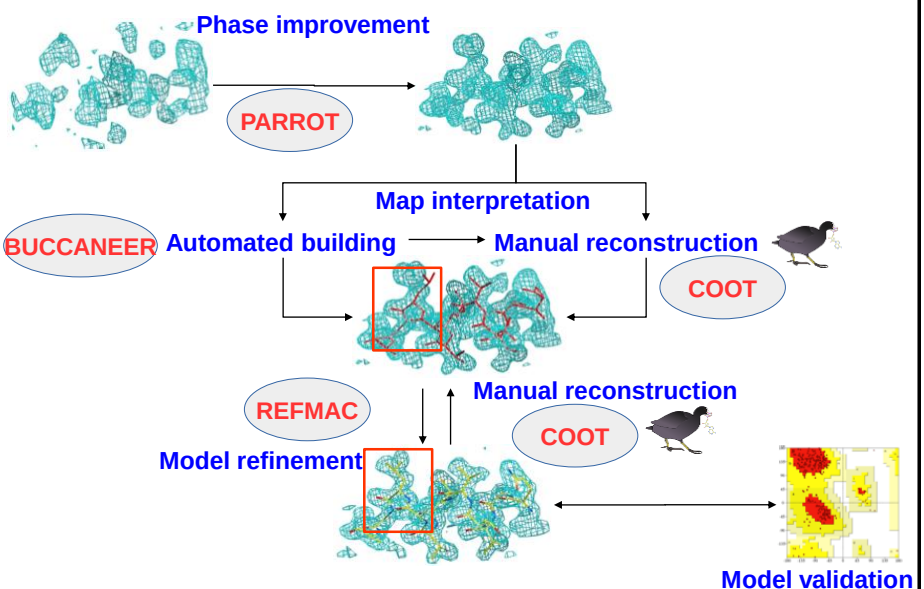
The sulfur or Se-methionines are the perfect starting point for the sequence fitting if the map is from sulfur-SAD or Se-MAD phases.

The large tryptophan may well be recognized in electron density maps.



1.0 Å | 1.8 Å | 3.0 Å | 4.0 Å

***Structure refinement*: iterative process of changing model parameters to optimally describe experimental data**



Crystallographic refinement: optimizing a scoring function in reciprocal space (*Model = Calc*)

$$T_{\mathrm{DATA}}(F_{\mathrm{OBS}}, F_{\mathrm{MODEL}})$$

- **Least-Squares**

$$T_{\mathrm{DATA}} = \sum_s w\left(F_{\mathrm{OBS}} - F_{\mathrm{MODEL}}\right)^2$$

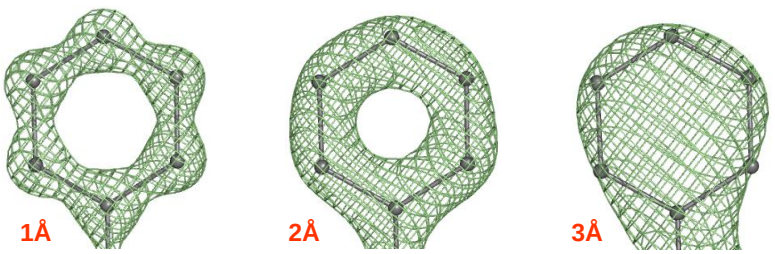
  - **Used in small molecule crystallography**
  - **Used in macromolecular crystallography in the past**

- **Maximum-Likelihood**

$$T_{\mathrm{DATA}} = \sum_s (1 - K_s^{cs})\left(\frac{a_s^2\left(F_s^{MODEL}\right)^2}{e_s b_s} + \ln\left(I_0\left(\frac{2 a_s F_s^{MODEL} F_s^{OBS}}{e_s b_s}\right)\right)\right) + K_s^{cs}\left(\frac{a_s^2\left(F_s^{MODEL}\right)^2}{2 e_s b_s} + \ln\left(\cosh\left(\frac{a_s F_s^{MODEL} F_s^{OBS}}{e_s b_s}\right)\right)\right)$$

  - **Option of choice for macromolecules**

Restraints for coordinate refinement



- **Lower the resolution, less detailed the map**
- **Need extra information to keep correct geometry during refinement**

The resolution the highest, the **w**eight the lowest

$$T = T_{\mathrm{DATA}}(F_{\mathrm{OBS}}, F_{\mathrm{MODEL}}) + w T_{\mathrm{RESTRAINTS}}$$

$$T_{\mathrm{RESTRAINTS}} = T_{\mathrm{BOND}} + T_{\mathrm{ANGLE}} + T_{\mathrm{DIHEDRAL}} + T_{\mathrm{PLANARITY}} + T_{\mathrm{NONBONDED}} + T_{\mathrm{CHIRALITY}}$$

$$T_{\mathrm{BOND}} = \Sigma_{\text{all bonded pairs}} w(d_{\mathrm{ideal}} - d_{\mathrm{model}})^2$$

e.g. d_{ideal} = 1.5 Å for C-C covalent

Crystallographic assessment and refinement

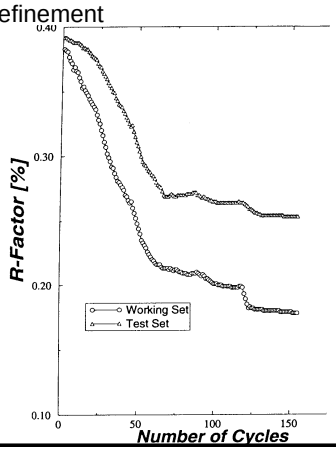
An essential validation of the 3D crystal structure is the confidence factor $\mathbf{R_{factor}}$

$$R = \frac{\sum_{hkl} \left| |F_{obs}| - k|F_{calc}| \right|}{\sum_{hkl} |F_{obs}|}$$

To avoid bias, a second confidence factor, $\mathbf{R_{free}}$, is calculated on 5% of the data, which are not included in the restrained positional refinement

$$R_T^{free} = \frac{\sum_{hkl \subset T} \left| |F_{obs}| - k|F_{calc}| \right|}{\sum_{hkl \subset T} |F_{obs}|}$$

If the restrained refinement is correct (*performed with Refmac5*), the $\mathbf{R_{free}}$ must decrease with the $\mathbf{R_{factor}}$ and tends towards <28%



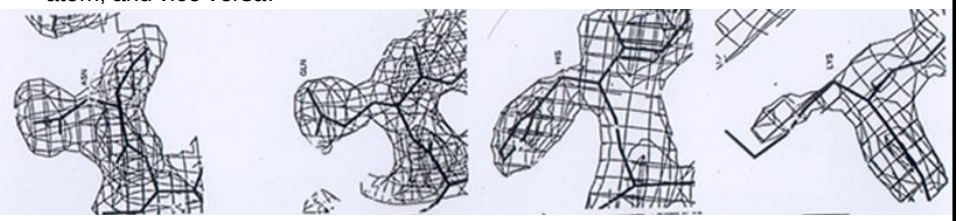

```
HEADER    HYDROLASE(O-GLYCOSYL)                   20-JAN-92   1HEW
COMPND    2 MOLECULE: HEN EGG WHITE LYSOZYME;
JRNL        AUTH   J.C.CHEETHAM,P.J.ARTYMIUK,D.C.PHILLIPS
REMARK    2 RESOLUTION.    1.75 ANGSTROMS.
...........................................................................
DBREF  1HEW A    1   129  UNP    P00698   LYC_CHICK       19    147
SEQRES   1 A  129  LYS VAL PHE GLY ARG CYS GLU LEU ALA ALA ALA MET LYS
SEQRES   2 A  129  ARG HIS GLY LEU ASP ASN TYR ARG GLY TYR SER LEU GLY
...........................................................................
HET    NAG  A 201      15
HET    NAG  A 202      14
HET    NAG  A 203      14
HETNAM     NAG N-ACETYL-D-GLUCOSAMINE
CRYST1   78.860   78.860   38.250  90.00  90.00  90.00 P 43 21 2      8
ATOM      1  N   LYS A   1       3.398   9.981  10.408  1.00 30.48           N
ATOM      2  CA  LYS A   1       2.459  10.365   9.364  1.00 28.03           C
ATOM      3  C   LYS A   1       2.458  11.880   9.149  1.00 21.93           C
ATOM      4  O   LYS A   1       2.481  12.672  10.100  1.00 14.10           O
ATOM      5  CB  LYS A   1       1.026   9.935   9.695  1.00 30.54           C
ATOM      6  CG  LYS A   1       0.028  10.169   8.558  1.00 37.93           C
ATOM      7  CD  LYS A   1      -1.415  10.089   9.048  1.00 33.23           C
ATOM      8  CE  LYS A   1      -2.357  10.822   8.082  1.00 32.17           C
ATOM      9  NZ  LYS A   1      -3.661  10.090   8.025  1.00 31.92           N
ATOM     10  N   VAL A   2       2.429  12.232   7.880  1.00 17.30           N
ATOM     11  CA  VAL A   2       2.395  13.653   7.465  1.00 14.47           C
...........................................................................
ATOM   1000  CD2 LEU A 129     -13.441  19.891   8.982  1.00 29.73           C
ATOM   1001  OXT LEU A 129     -17.993  19.662   8.407  1.00 31.81           O
TER    1002      LEU A 129
HETATM 1003  C1  NAG A 201       5.991  25.237  25.980  1.00 32.10           C
HETATM 1004  C2  NAG A 201       4.850  24.302  26.455  1.00 29.05           C
HETATM 1005  C3  NAG A 201       4.046  24.991  27.538  1.00 14.31           C
HETATM 1006  C4  NAG A 201       5.038  25.548  28.618  1.00 41.63           C
...........................................................................
HETATM 1046  O   HOH A 204     -16.295  29.471   0.511  1.00 18.64           O
HETATM 1047  O   HOH A 205      -1.660  14.995   1.659  1.00 45.86           O
END
```

positions X, Y, Z in Å

occupancy

temperature factor in Å² = 8 π² <U²>

- In crystallography, **uncertainty** in the positions of atoms increases with **disorder** in the macromolecular crystal.
- **Resolution** represents the average uncertainty for all atoms
- In contrast, the ***temperature factor*** or ***B factor*** quantitates the uncertainty for each atom. A high temperature factor reflects a low **empirical electron density** for the atom, and *vice versa*.



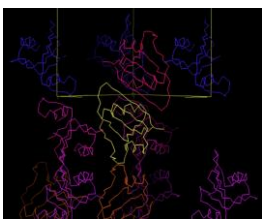
- The B-factor is given by $\mathbf{B_j=8\,\pi^2\,\langle U_j^2\rangle}$ where $\langle Uj^2\rangle$ is the mean square displacement of atom j. As U increases, the B-factor increases and the contribution of the atom to the scattering is decreased.

$$\exp\left(-B_j\frac{\sin^2\Theta}{\lambda^2}\right)$$

- For a B-factor of 15 Å^2, the mean square displacement of an atom from its equilibrium position is 0.4 Å, and 0.9 Å for a B-factor of 60 Å^2.

Structure Validation and Deposition

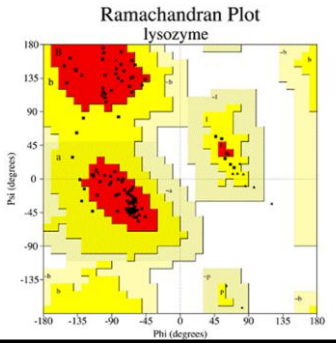
Generate symmetry related molecules and check contacts



The model should make chemical sense

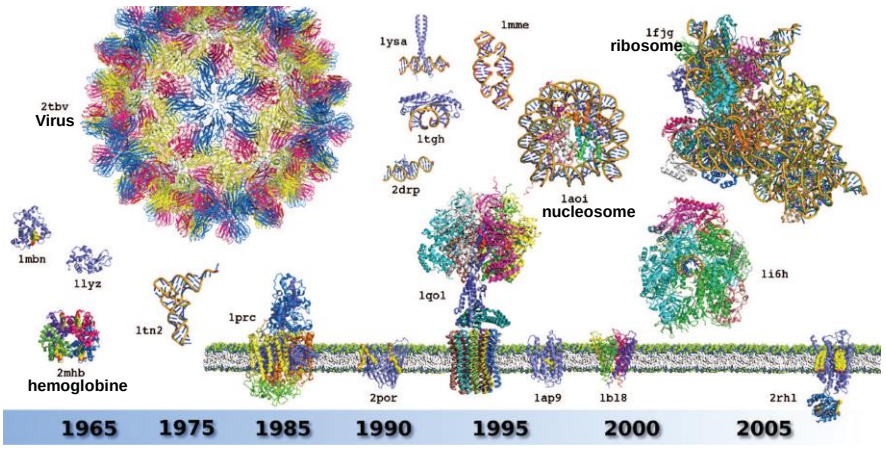
The stereochemical parameters such as bond length, bond angle etc, should be within the standard deviation from their ideal values

The Ramachandran Plot should be normal



==> WHATCHECK, MOLPROBITY, ...

Protein Data Bank, > 150 000 atomic structures



Why water boils at 100ºC and methane at -161ºC, why blood is red and grass is green, why diamond is hard and wax is soft… The answers to all these problems have come from structural analysis.

***Max Perutz**, July 1996, Churchill College, Cambridge*