Direct Methods

Carmelo Giacovazzo
Istituto di Cristallografia, CNR,
Bari University, Italy

carmelo.giacovazzo@ic.cnr.it
\[ \rho(r) = T^{-1}[F] \]

\[ F = T[\rho(r)] \]
Let us answer the following questions:

\[
\text{crystal structure } \Rightarrow \{ |F|^2 \} ?
\]

\[
\{ |F|^2 \} \Rightarrow \text{crystal structure } ?
\]

\[
|F_h|^2 = \sum_{j=1}^{N} f_j \exp(2\pi i h r_j) \sum_{j=1}^{N} f_j \exp(-2\pi i h r_j)
\]

\[
= \sum_{i,j=1}^{N} f_i f_j \exp[2\pi i h (r_i - r_j)]
\]

\[
= \sum_{j=1}^{N} f_j^2 + \sum_{i \neq j=1}^{N} f_i f_j \exp[2\pi i h (r_i - r_j)]
\]

As a consequence:

\[
\{ |F|^2 \} \leftrightarrow \rho(r)
\]
A third question: structure $\iff \{\varphi\}$?

$$F_h = \sum_{j=1}^{N} f_j \exp\left(2\pi i h r_j\right) = \sum_{j=1}^{N} f_j \exp\left(2\pi i h (X_0 + r_j')\right)$$

$$= \exp\left(2\pi i h X_0\right)\sum_{j=1}^{N} f_j \exp\left(2\pi i h r_j'\right)$$

$$= \exp\left(2\pi i h X_0\right)F_h'$$

$$\rightarrow F_h' = F_h \exp\left(-2\pi i h X_0\right) = |F_h|\exp i(\varphi_h - 2\pi h X_0)$$
A fourth basic question

How can we derive the phases from the diffraction moduli? This seems contradictory: indeed

*The phase values depend on the origin chosen by the user, the moduli are independent of the user.*

The moduli are *structure invariants*, the phases are not structure invariants.

Evidently, from the moduli we can derive information only on those combinations of phases (if they exist) which are structure invariants.
The simplest invariant: the triplet invariant

Use the relation

\[ F'_h = F_h \exp (-2\pi i\hbar X_0) \]

to check that the invariant \( F_h F_k F_{-h-k} \) does not depend on the origin.

\[
F'_h F'_k F'_{-h-k} = |F'_h| \exp i(\phi_h - 2\pi \hbar X_0) |F'_k| \exp i(\phi_h - 2\pi k X_0) \\
| F_{-h-k} | \exp i[\varphi_{-h-k} + 2\pi (h + k) X_0] \\
= |F_h| |F_k| |F_{-h-k}| \exp i(\varphi_h + \varphi_k + \varphi_{-h-k})
\]

The sum \( (\varphi_h + \varphi_k + \varphi_{-h-k}) \) is called triplet phase invariant.
Structure invariants

Any invariant satisfies the condition that the sum of the indices is zero:

- **doublet invariant**: $F_h F_{-h} = |F_h|^2$

- **triplet invariant**: $F_h F_k F_{-h-k}$

- **quartet invariant**: $F_h F_k F_l F_{-h-k-l}$

- **quintet invariant**: $F_h F_k F_l F_m F_{-h-k-l-m}$

...............
The prior information we can use for deriving the phase estimates may be so summarised:

1) **atomicity**: the electron density is concentrated in atoms:

\[
\rho(r) = \sum_{j=1}^{N} \rho_{aj} (r - r_j)
\]

2) **positivity** of the electron density:

\[
\rho(r) > 0 \Rightarrow f > 0
\]

3) **uniform distribution of the atoms in the unit cell.**
The Wilson statistics

- Under the above conditions Wilson (1942,1949) derived the structure factor statistics. The main results were:
  
  \[ <|F_h|^2| = \sum_{j=1}^{N} f_j^2 \]  

- Eq.(1) is:
  
  a) resolution dependent \((f_j \text{ varies with } \theta)\),
  
  b) temperature dependent: \(f_j = f_j^0 \exp(-B_j \sin^2 \theta / \lambda^2)\)
  
- From eq.(1) the concept of normalized structure factor arises:

\[ E_h = F_h / (\sum_{j=1}^{N} f_j^2)^{1/2} \]
The Wilson Statistics

- $|E|$-distributions:

$$P_1(|E|) = 2|E| \exp(-|E|^2)$$

$$P_1(|E|) = \sqrt{\frac{2}{\pi}} \exp(-|E|^2 / 2)$$

and

$$<|E|^2> = 1$$

in both the cases. The statistics may be used to evaluate the average theme factor and the absolute scale factor.
The Wilson plot

\[ F_h = \sum_{j=1}^{N} f_j \exp(2\pi i \hbar r_j) = \sum f_j^0 \exp \left( -B_j \frac{\sin^2 \theta}{\lambda^2} \right) \exp(2\pi i \hbar r_j) \]

\[ |F_h|_{obs}^2 = K |F_h|^2 = K |F_h^0|^2 \exp(-2Bs^2) \]

\[ < |F_h|_{obs}^2 > = K < |F_h^0|^2 > \exp(-2Bs^2) = K\Sigma_s^0 \exp(-2Bs^2) \]

\[ \ln \left( \frac{< |F_h|_{obs}^2 >}{\Sigma_s^0} \right) = \ln K - 2Bs^2 \]
The Cochran formula

\[ \Phi_{h,k} = \phi_h + \phi_k + \phi_{-h-k} = \phi_h + \phi_k - \phi_h \]

\[ P(\Phi_{hk}) \approx [2\pi I_0]^{-1} \exp(G \cos \Phi_{hk}) \]

where \( G = 2 | E_h E_k E_{h+k} | / N^{1/2} \)

Accordingly:

\[ \phi_h + \phi_k - \phi_{h+k} \approx 0 \quad \propto \quad G = 2 | E_h E_k E_{h+k} | / N^{1/2} \]

\[ \phi_h - \phi_k - \phi_{h-k} \approx 0 \quad \propto \quad G = 2 | E_h E_k E_{h-k} | / N^{1/2} \]

\[ \phi_h \approx \phi_k - \phi_{h-k} \quad \propto \quad G = 2 | E_h E_k E_{h-k} | / N^{1/2} \]
The tangent formula

A reflection can enter into several triplets. Accordingly

\[ \varphi_h \approx \varphi_{k1} + \varphi_{h-k1} = \theta_1 \text{ with } P_1(\varphi_h) \propto G_1 = 2 \mid E_h E_{k1} E_{h-k1} \mid / N^{1/2} \]

\[ \varphi_h \approx \varphi_{k2} + \varphi_{h-k2} = \theta_2 \text{ with } P_2(\varphi_h) \propto G_2 = 2 \mid E_h E_{k2} E_{h-k2} \mid / N^{1/2} \]

\[ \varphi_h \approx \varphi_{kn} + \varphi_{h-kn} = \theta_n \text{ with } P_n(\varphi_h) \propto G_n = 2 \mid E_h E_{kn} E_{h-kn} \mid / N^{1/2} \]

Then

\[ P(\varphi_h) \approx \prod_j P_j(\varphi_h) \approx L^{-1} \prod_j \exp \left[ G_j \cos (\varphi_h - \theta_j) \right] \]

\[ = L^{-1} \exp \left[ \alpha \cos (\varphi_h - \theta_h) \right] \]

where

\[ \tan \theta_h = \frac{\sum G_j \sin \theta_j}{\sum G_j \cos \theta_j} = \frac{T}{B}, \quad \alpha_h = \left( T^2 + B^2 \right)^{1/2} \]
A geometric interpretation of $\alpha$
The random starting approach

To apply the tangent formula we need to know one or more pairs \((\phi_k + \phi_{h-k})\). Where to find such an information?

The most simple approach is the **random starting approach**. Random phases are associated to a chosen set of reflections. The tangent formula should drive these phases to the correct values. The procedure is cyclic (up to **convergence**).

How to recognize the correct solution? **Figures of merit** can or cannot be applied.
Tangent cycles

- $\phi_1 \quad \phi'_1 \quad \phi''_1 \ldots \quad \phi^c_1$
- $\phi_2 \quad \phi'_2 \quad \phi''_2 \ldots \quad \phi^c_2$
- $\phi_3 \quad \phi'_3 \quad \phi''_3 \ldots \quad \phi^c_3$
- $\ldots$
- $\phi_n \quad \phi'_n \quad \phi''_n \ldots \quad \phi^c_n$
- **Ab initio** phasing
- **SIR2009** is able to solve
  - *small size structures* (up to 80 atoms in the a.u.);
  - *medium-size structures* (up to 200);
  - *large size* (no upper limit)
- It uses
  - **Patterson deconvolution techniques**
  - *(multiple implication transformations)*
- as well as
  - **Direct methods**
- to obtain a starting set of phases. They are extended and refined *via*
  - **electron density modification techniques**
• Direct methods limits for proteins:

• 1) the large size (proteins range from 300 atoms in the asymmetric unit to several thousands). The G factor in the Cochran formula are very small.

• 2) data resolution
• To overcome the limits one is obliged to:
• - increase the **number of direct methods trials**. The cost to pay concerns the computing time.
• - **improve and extend the the poor phases** available by DM by exploiting some specific features of the proteins (e.g., the solvent, etc.).
About the data resolution limit

Atomic resolution at length was considered a necessary (and not sufficient) condition for *ab initio phasing* (Sheldrick rule), condition relaxed later on (up to 1.2 Å). If it is not satisfied:

- the atomicity condition is violated;
- the number of reliable triplet invariants exploitable by the tangent procedure is small.
- Patterson and EDM procedures are less effective;
- the small ratio

*number of observations/ number of parameters*

- make least squares unreliable.