About the basic postulate of the structural crystallography

Carmelo Giacovazzo
Istituto di Cristallografia, CNR,
Bari, Italy

carmelo.giacovazzo@ic.cnr.it
\[ \rho(r) = T^{-1}[F] \]
\[ F = T[\rho(r)] \]
Let us answer the following questions:

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

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

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

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

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

As a consequence:

\[ \left\{ |F|^2 \right\} \Leftrightarrow \left\{ r_i - r_j \right\} \]
If we accept that only one structure exists with a given set of interatomic vectors, then

\[ \rho(r) \Leftrightarrow \{ \mathbf{r}_i - \mathbf{r}_j \} \Leftrightarrow \left\{ |F|^2 \right\} \]

Then for only one structure the set of calculated structure factor moduli fits the set \(|F_{\text{obs}}|\). That is, with sufficiently small

\[ R_{\text{cryst}} = \frac{\sum_h \| F_{\text{obs}} - |F_{\text{calc}}| \|}{\sum_h |F_{\text{obs}}|} \]
What about homometric structures?
If they exist, the preceding assumptions are false.

What about electron density maps?
Can they exist different electron density maps with the same $|F|$?
Infinite electron density maps exist compatible with the same set of $|F|^2$

How to distinguish the correct map from the wrong ones?
The basic postulate of the structural crystallography

Only one chemically sound crystal structure exists compatible with the experimental diffraction data.

This is a postulate and chemically sound is a necessary attribute.
It is on the basis of this postulate that we can accept that the about one million of structures deposited in the DataBanks are really and correctly solved.
Structures deposited in the CSDB up to 01/01/2012
For each range of $R_{\text{cryst}}$, $N_{\text{str}}$ and % are the number of structures and percentage, respectively

<table>
<thead>
<tr>
<th>$\Delta R_{\text{cryst}}$</th>
<th>$N_{\text{str}}$</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01-0.03</td>
<td>62774</td>
<td>10.5</td>
</tr>
<tr>
<td>0.03-0.04</td>
<td>122706</td>
<td>20.6</td>
</tr>
<tr>
<td>0.04-0.05</td>
<td>135525</td>
<td>22.7</td>
</tr>
<tr>
<td>0.05-0.07</td>
<td>163269</td>
<td>27.4</td>
</tr>
<tr>
<td>0.07-0.09</td>
<td>60651</td>
<td>10.2</td>
</tr>
<tr>
<td>0.09-0.10</td>
<td>13370</td>
<td>2.2</td>
</tr>
<tr>
<td>0.10-0.15</td>
<td>18353</td>
<td>3.1</td>
</tr>
<tr>
<td>0.15- ....</td>
<td>3835</td>
<td>0.6</td>
</tr>
</tbody>
</table>
For the practical daily work the basic postulate of the structural crystallography may be transformed into a working rule:

*in a diffraction experiment the phase information is not lost, it is only hidden in the diffraction amplitudes.*

Accordingly, *any phasing approach is nothing else but a method for recovering the hidden phases from the set of diffraction amplitudes.*
Is the amount of information stored in the diffraction amplitudes always sufficient to define the structure?

Since the data resolution defines the quantity of information provided by the diffraction experiment, we try to understand how this quantity varies with the resolution.
Crystal with *P1* symmetry.

\[ N = \text{number of non-H atoms in the unit cell.} \]

The diffraction experiment provides data up to \( d_{\text{min}} \). \n
\[ N_{\text{ref}} = \text{number of measurable reflections} \]

\[
N_{\text{ref}} = \frac{\Phi_{\text{meas}}^*}{V^*} = \Phi_{\text{meas}}^* V = \frac{4\pi}{3d_{\text{min}}^3} kN
\]

Here \( V = kN \), where \( k \) is usually between 15.5 and 18.5 for a small or medium size molecule; for a protein, owing to the presence of the solvent, \( k \) may be significantly larger, up to about 40 and more.
A criterion for estimating the experimental information:

\[ R_{inf} = \frac{\text{number of measured symmetry independent reflections}}{\text{number of structural parameters}} \]

\[ R_{inf} = \frac{N_{ref}}{(4N)} = \frac{4\pi}{kN} \cdot \frac{1}{4N} \]

\[ 4N = \text{number of structural parameters necessary for defining the structure.} \]

By including the Friedel Law

\[ R_{inf} \approx \frac{\pi}{6d_{min}^3} k \]
$R_{inf}$ in $P1$ some values of $d_{min}$ and of $k$.

<table>
<thead>
<tr>
<th>$d_{min}$</th>
<th>$k=17$</th>
<th>$k=25$</th>
<th>$k=35$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>139</td>
<td>204</td>
<td>286</td>
</tr>
<tr>
<td>0.6</td>
<td>41</td>
<td>60.6</td>
<td>84.7</td>
</tr>
<tr>
<td>0.8</td>
<td>17.4</td>
<td>25.6</td>
<td>35.7</td>
</tr>
<tr>
<td>1.0</td>
<td>8.9</td>
<td>13.1</td>
<td>18.3</td>
</tr>
<tr>
<td>1.4</td>
<td>3.2</td>
<td>4.8</td>
<td>13.7</td>
</tr>
<tr>
<td>1.8</td>
<td>1.5</td>
<td>2.2</td>
<td>3.1</td>
</tr>
<tr>
<td>2.2</td>
<td>0.8</td>
<td>1.2</td>
<td>1.7</td>
</tr>
<tr>
<td>2.5</td>
<td>0.6</td>
<td>0.8</td>
<td>1.2</td>
</tr>
<tr>
<td>3.0</td>
<td>0.3</td>
<td>0.5</td>
<td>0.7</td>
</tr>
<tr>
<td>3.5</td>
<td>0.2</td>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>4.0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Modern *ab initio* phasing methods

- Patterson techniques
- Direct Methods
- Charge Flipping
- VLD (Vive la Difference)

Non *ab initio* phasing Methods

- Molecular replacement
- Isomorphous Replacement
- Anomalous Dispersion Techniques
About Direct Methods: they directly go from moduli to phases.

They are the main responsible for the solution of the phase problem for small molecules.
We positively answered the two basic questions:

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

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

Two other questions are however necessary for starting the phasing approach.
The third question:  \( \text{structure} \iff \{\varphi\}? \)

\[ F_h = \sum_{j=1}^{N} f_j \exp(2\pi i h r_j) = \sum_{j=1}^{N} f_j \exp(2\pi i h (X_0 + r'_j)) \]

\[ = \exp(2\pi i h X_0) \sum_{j=1}^{N} f_j \exp(2\pi i h r'_j) \]

\[ = \exp(2\pi i h X_0) F'_h \]

\[ \rightarrow F'_h = F_h \exp(-2\pi i h X_0) = |F_h| \exp(i(\varphi_h - 2\pi h X_0)) \]
The 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

\[ F'_h = F_h \exp (-2\pi i hX_0) \]

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

\[
F'_h F'_k F'_{-h-k} = \left| F_h F_k F_{-h-k} \right| \exp i(\phi'_h + \phi'_k + \phi'_{-h-k})
\]

\[
F_h \left| \exp i(\phi_h - 2\pi hX_0) \right| F_k \left| \exp i(\phi_k - 2\pi kX_0) \right|
\]

\[
\left| F_{-h-k} \right| \exp i[\phi_{-h-k} + 2\pi(h + k)X_0]
\]

\[
= \left| F_h \right| \left| F_k \right| \left| F_{-h-k} \right| \exp i(\phi_h + \phi_k + \phi_{-h-k})
\]

The sum \((\phi_h + \phi_k + \phi_{-h-k})\) is called *triplet phase invariant*.
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** : $F_h F_k F_l F_m F_{-h-k-l-m}$

Structure invariants however may also be

$$F_h F_{-ph} = |F_h F_{-ph}| \exp[i(\phi_h - \phi_{ph})]$$

so widely used in *EDM* techniques
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:

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

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 where:

\[ \langle |F_h|^2 \rangle = \sum_{j=1}^{N} f_j^2 \]  

(1)

- 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:

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

- \(|E|\)-distributions:

\[
P_{\bar{1}}(|E|) = 2 |E| \exp(-|E|^2)
\]

\[
P_{\bar{1}}(|E|) = \sqrt{\frac{2}{\pi}} \exp(-|E|^2 / 2)
\]

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

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

\[ A \]

\[ \exp \left( -B \frac{\sin^2 \theta}{\lambda^2} \right) \sum f_j^0 \exp 2\pi i h r_j \]

\[ F_h^0 \]

\[ |F_h|^2_{obs} = K |F_h|^2 = K |F_h^0|^2 \exp \left(-2B s^2\right) \]

\[ A \]

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

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

\[ y \]

\[ x \]
The Cochran formula

$$\Phi_{h,k} = \varphi_h + \varphi_k + \varphi_{-h-k} = \varphi_h + \varphi_k - \varphi_{h+k}$$

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

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

Accordingly:

$$\varphi_h + \varphi_k - \varphi_{h+k} \approx 0 \quad \varphi_h - \varphi_k - \varphi_{h-k} \approx 0 \quad \varphi_h \approx \varphi_k + \varphi_{h-k}$$

$$G = 2 | E_h E_k E_{h+k} | / N^{1/2}$$

$$G = 2 | E_h E_k E_{h-k} | / N^{1/2}$$

$$G = 2 | E_h E_k E_{h-k} | / N^{1/2}$$
The tangent formula

A reflection can enter into several triplets. Accordingly

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

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

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

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 \((\varphi_k + \varphi_{h-k})\). 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

• $\phi_1$ $\phi'_1$ $\phi''_1$ ............ $\phi^c_1$

• $\phi_2$ $\phi'_2$ $\phi''_2$ ............ $\phi^c_2$

• $\phi_3$ $\phi'_3$ $\phi''_3$ ............ $\phi^c_3$

• ...

• $\phi_n$ $\phi'_n$ $\phi''_n$ ............ $\phi^c_n$
• A suitable **Figure of merit** is able to rank the various trials in order of reliability.
• The most probable ones are those explored first.

• **Phase extension** and **least squares** usually follow direct solutions.
Ab initio phasing

- SIR2011 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)
- Direct methods
- the algorithm VLD (vive la difference)