Introduction to Sir2011

Giovanni Luca Cascarano
CNR– Istituto di Cristallografia – Bari – Italy
The crystal structure solution

To solve a crystal structure, i.e. to obtain the atomic coordinates, different methods are available.

We’ll focus our attention on Direct Methods, very popular and effective, able to give in a short time and in automatic way the solution to our problem: to compute the phases lost during the diffraction experiment (the so called Phase Problem).

Once the phases are available it is possible to apply the Fourier Transform and get the electron density map whose maxima correspond to the atomic positions.
The crystal structure solution

Crystallization

Experiment

Beam (X rays, electrons) → Crystal → Diffraction pattern → I = |F|^2

\( \varphi = ? \)

Atomic model

Crystallographic Methodologies → Software → Electron density map → Molecular model
The crystal structure solution

Crystallization

Experiment

Beam (X rays, electrons) → Crystal → Diffraction pattern → I = |F|^2

Atom model → The phase problem

Crystallization

Crystallographic Methodologies → Software → Electron density map
The crystal structure solution

Crystallization

Esperiment

Beam (X rays, electrons) → Crystal → Diffraction pattern

I = |F|^2
φ = ?

The phase problem

Crystallographic Methodologies

Software

Electron density map

Atomic model
The crystal structure solution

Crystallization

Experiment

I = |F|^2
\varphi = ?

Beam (X rays, electrons) → Crystal → Diffraction pattern

Crystallographic Methodologies → Software → Electron density map → Atomic model
The phase problem

If we define the structure factor $F_h$ with indices $h=(hkl)$

$$F_h = \sum_{j=1}^{N} f_j \exp(2\pi hr_j) = |F_h| \exp(i\phi_h)$$

the electron density $\rho(r)$ for every point $r$ in the unit cell is defined as:

$$\rho(r) = V^{-1} \sum_{h} |F_h| \exp(i\phi_h) \exp(-2\pi ihr)$$

In this formula the $|F_h|$ value is known from experiment, the phase of the structure factor $\phi_h$ has to be recovered.
Statistical analysis

The statistical analysis of the structure factor amplitudes gives useful information about the presence or absence of the inversion center; furthermore it allows the estimate of the overall isotropic thermal factor (B) and of the scale factor (K), necessary to transform the structure factor $F_h$ in the normalized structure factor $E_h$ used in Direct Methods.

To derive the theoretical probability distributions it is assumed that the atomic positions are random variables, uniformly distributed in the unit cell.

Wilson (1942) proposed a simple method to derive K and B, based on the statistical analysis of the intensities.
Statistical analysis

The experimental structure factors are on a relative scale. If we assume that the thermal motion is isotropic and equal for all atoms:

\[ F_h = \exp(-Bs^2) \sum_{j=1}^{N} f_j^o \exp(2\pi i hr_j) = \exp(-Bs^2) F_h^o \]

where \( s = \text{sint}(\theta)/\lambda \).

The experimental structure factors \( |F_{obs}| \) are related to the calculated structure factors \( |F_h| \) through a scale constant \( K \):

\[ |F_{obs}|^2 = K |F_h|^2 = K |F_h^o|^2 \exp(-2Bs^2) \]
Statistical analysis

Sir2011 divides the reflections in \( s^2 \) shells and computes the average values of the intensities in every shell; assuming that the atoms are uniformly distributed in the unit cell:

\[
\langle |F_h^0|^2 \rangle = \sum_{j=1}^{N} \left( f_j^0 \right)^2 = \sum_{s}^0
\]

from which:

\[
\ln\left(\frac{\langle |F_h^0|^2 \rangle}{\sum_{s}^0}\right) = \ln K - 2Bs^2
\]
Statistical analysis

If the values of the logarithm are plotted against $s^2$ and the least square line is computed, its intercept on the vertical axis will give the value of $\ln K$, its slope the value of $2B$.

From the scaled intensities, corrected using the thermal factor, we obtain the normalized structure factors:

$$|E_h|^2 = \frac{|F_h|^2}{\varepsilon \Sigma}$$

These magnitudes do not depend on resolution and on the atomic species, so they correspond to point atom structure. They are needed in the Direct Methods procedures.
In this example using the least square line

\[
\ln\left(\frac{\langle |F_h|^2 \rangle}{\Sigma_s^o}\right) = \ln K - 2Bs^2
\]

we obtain the following values:

Intercept = \ln K = -0.08737

Slope = -2B = -5.27746

K = 1.09130 \quad B = 2.63873
Structure Invariants

The moduli of the observed structure factors are independent from the origin but the phases depend on it.

Starting from the observed structure factor it is possible to obtain information only on linear combinations of phases independent from the origin choice called *Structure Invariants*.

The most used structure invariant is formed by three phases and is called *triplet*:

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

where the sum of the indices is equal to zero.
Structure Invariants

Let’s consider the structure factor referred to a new origin shifted by \( r_0 \):

\[
F'_H = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{H} \cdot (r_j - r_0))
\]

\[
= F_H \exp(-2\pi i \mathbf{H} \cdot r_0)
\]

\[
= |F_H| \exp(i(\Phi_H - 2\pi i \mathbf{H} \cdot r_0))
\]

The modulus does not change, the phase differs:

\[
\Delta \Phi_H = 2\pi \mathbf{H} \cdot r_0
\]

If we compute the structure factor respect to the new origin

\[
F'_{H_1} F'_{H_2} ... F'_{H_n} = F_{H_1} F_{H_2} ... F_{H_n} \exp\left[i \left( \Delta \Phi_{H_1} \Delta \Phi_{H_2} ... \Delta \Phi_{H_n} \right) \right]
\]

\[
= F_{H_1} F_{H_2} ... F_{H_n} \exp\left[ i2\pi \sum_{j=1}^{n} \mathbf{H}_j \cdot r_0 \right] = F_{H_1} F_{H_2} ... F_{H_n}
\]
Triplet estimates: Cochran formula

Triplets can be estimated using the classical Cochran formula. In this case the value of the triplet phase $\Phi_{h,k}$ depends only on three phases $\varphi_h, \varphi_k, \varphi_{h+k}$ and its estimate on three normalized structure factors

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

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

where $G = 2 | E_h E_k E_{h+k}|/N^{1/2}$ is the concentration parameter.
Triplet estimates: Cochran formula

Looking at the figure, we can deduce that larger is the value of $G_{HK}$, higher is the probability that the phase $\Phi_{HK}$ is close to zero. If we compute the triplet invariants using a subset of reflections with the higher values of $|E|$, we can be more confident on the value of $\Phi_{HK}$.
Triplet estimates: P10 formula

Triplet invariants may be replaced by a family of quintets:

\[ \Phi_{h,k} = \varphi_h + \varphi_k + \varphi_{-h-k} + \varphi_l - \varphi_l \]

the values of which is perfectly equal to the basis triplet.

The vector \( l \) may freely vary over the set of measured reflections.

Now \( \Phi_{h,k} \) depends, for each \( l \) vector, on 10 moduli

\[ |E_h|, |E_k|, |E_{h+k}|, |E_l|, |E_{h+l}|, |E_{h-l}|, |E_{k+l}|, |E_{k-l}|, |E_{h+k+l}|, |E_{h+k-l}|. \]
The triplet phase $\Phi_{h,k}$ may be estimated via

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

but this time

$$G_{h,k} = \frac{2 \left| E_h E_k E_{h+k} \right|}{\sqrt{N}} \left(1 + \frac{1}{N} C\right)$$

C is a summation of contributions deriving from the ten E’s. If $C / N < -1$ the triplet is expected to be negative.
The tangent formula

We can write the triplet phase relationship as $\varphi_H \approx \varphi_K + \varphi_{H-K}$; the distribution of $\varphi_H$ with respect to the expected value

$$\theta_H = \varphi_K + \varphi_{H-K}$$

is

$$P(\varphi_H) = K \exp(G_{HK} \cos(\varphi_H - \theta_H))$$

Let’s consider all the triplets related to a reflection $H$

$$\varphi_H \approx \varphi_{K1} + \varphi_{H-K1}$$
$$\varphi_H \approx \varphi_{K2} + \varphi_{H-K2}$$
$$\cdots$$
$$\varphi_H \approx \varphi_{Kn} + \varphi_{H-Kn}$$
The tangent formula

If we assume that the phase indications are statistically independent, the joint probability of the phase \( \varphi_H \) will be

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

\[
\approx L^{-1} \exp \left[ \sum_j G_j \cos(\varphi_h - \varphi_j) \right]
\]

\[
\approx L^{-1} \exp \left( \cos \varphi_h \sum_j G_j \cos \varphi_j + \sin \varphi_h \sum_j G_j \sin \varphi_j \right)
\]
The tangent formula

If we set
\[ \Sigma G_j \sin \theta_j = \alpha_H \sin \theta_j \quad \text{and} \quad \Sigma G_j \cos \theta_j = \alpha_H \cos \theta_j \]
we obtain
\[ P(\varphi_H) \approx K \exp(\alpha_H \cos(\varphi_H - \theta_H)) \quad \text{where} \quad \alpha_H = (T^2 + B^2)^{1/2} \]
and
\[ \tan \theta_h = \frac{\Sigma G_j \sin \theta_j}{\Sigma G_j \cos \theta_j} = \frac{T}{B} \]

This equation gives the most probable value of \( \theta_j \) and is known as the tangent formula.
Using this formula is possible to derive the phase values of a given reflection when an estimation of \( n \) couples of phases is known.
The tangent formula

Vector representation, in the complex plane, of the combination of five triplets with concentration parameter $G_1, \ldots, G_5$
The multisolution approach

Usually random phase values are assigned to a subset of *strong reflections* (those with the higher of $|E|$); by means of the triplets and of the tangent formula almost all the strong reflections can be phased.

This procedure is repeated $n$ times changing the values of the random phases and, for every obtained set of phases a suitable *Figure of Merit* (eFOM) is computed. Higher is the eFOM value, more likely the corresponding phase set is correct.

These trials, sorted with respect to the eFOM value, will be submitted to the automatic Direct Space Refinement (DSR).
The multisolution approach

Usually random phase values are assigned to a subset of strong reflections (those with the higher of $|E|$); by means of the triplets and of the tangent formula almost all the strong reflections can be phased. This procedure is repeated $n$ times changing the values of the random phases a suitable Figure of Merit ($eFOM$) is computed. Higher is the $eFOM$ value, more likely the corresponding phase set is correct. These trials, sorted with respect to the $eFOM$ value, will be submitted to the automatic Direct Space Refinement (DSR).
The multisolution approach

Usually random phases are assigned to a subset of strong reflections (those with the highest |E|); by means of the triplets and of the tangent formula almost all the strong reflections can be phased. This procedure is repeated n times changing the values of the random phases. For every obtained set of phases a suitable Figure of Merit (eFOM) is computed. Higher is the eFOM value, more likely the corresponding phase set is correct. These trials, sorted with respect to the eFOM value, will be submitted to the automatic Direct Space Refinement (DSR).
The structural complexity

In Sir2011 the parameter related to the structural complexity governs most of the algorithms used for the automatic structure solution.

In the program there is the following classification:

- **Very small** size structures: up to 6 non-H atoms in asymmetric unit
- **Small** size structures: 6-80 non-H atoms in asymmetric unit
- **Medium** size structures: 81-300 non-H atoms in asymmetric unit
- **Large** size structures: more than 300 non-H atoms in asymmetric unit
The phase extension procedure

The number of reflections phased by the tangent formula is, in general, a small percentage of all reflections available.

To extend and refine for all suitable reflections (i.e. those for which $F_{\text{obs}} > 3\sigma(F_{\text{obs}})$) an iterative procedure is used, the DSR (Direct Space Refinement).

The procedure is constituted by $n$ cycles

$$\rho \rightarrow \{\varphi\} \rightarrow \rho$$

where $\rho$ is the electron density map and $\{\varphi\}$ is the set of calculated phases. The number of cycles depends on the structural complexity.
The electron density maps are calculated by using a number of reflections cyclically increasing; a small fraction of $\rho$ is used during the inversion of the electron density map, the remaining part is set to zero. To process a trial, EDM (Electron Density Modification) is applied a number of times, depending on structural complexity.
The automatic model refinement

At this stage both modulus and phase for an adequate number of reflections are available, therefore it is possible to apply the Fourier transform to the best set of phases (in terms of eFOM), extended using the DSR procedure, to obtain an electron density map.

Its peaks are labelled in terms of atomic species according to their intensity and to the chemical content of the unit cell.
The automatic model refinement

At this stage both modulus and phase for an adequate number of reflections are available, therefore it is possible to apply the Fourier transform to the best set of phases (in terms of eFOM), extended using the DSR procedure, to obtain an electron density map.

Its peaks are labelled in terms of atomic species according to their intensity and to the chemical content of the unit cell.
The automatic model refinement

In the automatic model refinement there are two steps:

- six cycles of Fourier Refinement (FR)
- n cycles of Diagonal Least Squares (DLSQ)

The FR procedure modifies the isotropic thermal parameter of the atoms according to the Fourier peak height: to lower values of intensity correspond an higher value of $B_{\text{iso}}$.

During the DLSQ procedure the atomic coordinates and the isotropic thermal factor are refined. This procedure stops when the convergence has been attained.
The automatic model refinement

During these cycles, new values of phases and weights are computed to produce a new electron density map using \((2F_{\text{obs}} - F_{\text{calc}})\) coefficients. A new map is calculated and this iterative procedure stops when the crystallographic residual \(R\) increases, where:

\[
R = \frac{\sum w(|F_o| - K|F_c|)}{\sum w(|F_o|)}
\]

If the R value is below a given threshold the program stops, otherwise a new set of phases is considered.
During these cycles, new values of phases and weights are computed to produce a new electron density map using \((2F_{\text{obs}} - F_{\text{calc}})\) coefficients. A new map is calculated and this iterative procedure stops when the crystallographic residual 

\[ R = \text{Residual} \]

If the R value is below a given threshold the program stops, otherwise a new set of phases is considered.
The program

This program will be available, free of charge for academic institutions, for Linux, Mac and MS Windows systems.
The program

Sir2011: a new program for crystal structure solution
M.C. Burla, R. Caliandro, M. Camalli, B. Carrozzini,
G.L. Cascarano, C. Giacovazzo, A. Mazzone, G. Polidori, R. Spagna

This program will be available, free of charge for academic institutions, for Linux, Mac and MS Windows systems.
The authors

R.Caliandro, B.Carrozzini, G.L.Cascarano, C.Giacovazzo, A. Mazzone
*Istituto di Cristallografia del CNR – Bari*

M.Camalli, R.Spagna
*Istituto di Cristallografia del CNR – Rome*

M.C.Burla, G.Polidori
*Dipartimento di Scienze della Terra, University of Perugia*