Curso 1995/96
CIENCIAS Y TECNOLOGÍAS

JAVIER GONZÁLEZ PLATAS

Advances in direct methods for small molecules and proteins

Director
CARMELO GIACOVAZZO

SERVICIO DE PUBLICACIONES
UNIVERSIDAD DE LA LAGUNA

SOPORTES AUDIOVISUALES E INFORMÁTICOS
Serie Tesis Doctorales

Acknowledgements

I would like to thank all the people that in one or other form cooperated for the accomplishment this Doctoral Thesis.

For the italian part, my director of Thesis, Prof. Carmelo Giacovazzo: it is a great pleasure to work with him. For the rest of the SIR Team, I thank Gianluca Cascarano, Angela Altomare, Antonella Guagliardi, Anna Grazia Moliterni, Caterina Chiarella and in special form Dritan Siliqi. I felt at home when I was in Bari.

For the spanish part, I would to thank all members of my Department of Physics of the University of La Laguna, in special form for my X-rays group, Dra. Catalina Ruiz, Cristina González, María Hernández and Víctor Rodríguez, that in any form stimulated me during the thesis period.

I want to express my gratitude to the Consejería de Educación y Ciencia del Gobierno Autónomo de Canarias for the partial financial support in this year for my stay in Bari.

To my Family: I can only say to them: Thanks. You are very special for me.

Finally, I thank my wife to which this Thesis is dedicated. She faced a lot of difficulties when I wasn't at home. Cande, T.Q. M^m.

September, 1995

Contents

Chapter II: The Formula P_{13}

Chapter III: The normalization procedure for Proteins

Chapter IV: Phasing up to derivative resolution

Chapter V: The use of the partial structure information in ab-initio solution of Proteins by Direct Methods

The scope of this Thesis

In the last years the probability methods have seen new important developments. At the basis of the new approach the following principle may be stated: *It's possible to obtain good estimates of structure invariants (s.i.) or structure semi-invariants (s.s.) provided appropiate sets of normalized structure factor moduli are exploited, which are statistically the most effective in determining the value of the given s.i. or s.s*. In this sense we have studied a new formula that we have called P_{13} and it has been applied to practical cases for small molecules.

Traditional direct methods based on the tangent formula and/or on Sayre's equation cannot solve *ab-initio* the large majority of protein crystal structures, but they seem able to solve proteins if isomorphous derivative data are available. We have settled new direct techniques which minimize errors in the normalization procedure, produce high quality phases up to derivative resolution and seem able to extend phase information up to the native protein resolution.

Symbols and abbreviations

$\mathbf{C} \equiv (\mathbf{R}, \mathbf{T})$	Symmetry operator; $\mathbf{R}$ is the rotation component, $\mathbf{T}$ the translation component.
$D_i(\mathrm{x}) = I_i(\mathrm{x}) / I_o(\mathrm{x})$	I_i is the modified Bessel function of the order *i*
$\Delta = S' - R'$	
$\Delta\,' = S'T - R'$	
$E_d = F_d / \Sigma_d^{1/2} = S\exp(\mathrm{i}\psi)$	Normalized structure factor of the isomorphous derivative
$E'_d = F_d / \Sigma_H^{1/2} = S'\exp(\mathrm{i}\psi)$	Derivative pseudo-normalized structure factor (with respect to heavy atom substructure)
$E_{\mathbf{h}} = R_{\mathbf{h}} \exp(\mathrm{i}\phi_{\mathbf{h}})$	Normalized structure factor
$E''_{\mathbf{h}} = F_{\mathbf{h}} / \Sigma_q^{1/2}$	Pseudo-normalized structure factor of the native protein
$E_p = F_p / \Sigma_p^{1/2} = R\exp(\mathrm{i}\phi)$	Normalized structure factor of the native protein
$E'_p = F_p / \Sigma_H^{1/2} = R'\exp(\mathrm{i}\phi)$	Native protein pseudo-normalized structure factor (with respect to heavy atom substructure)
$E''_{\pi,\mathbf{h}} = F_{\pi,\mathbf{h}} / \Sigma_q^{1/2}$	Pseudo-normalized structure of the partial structure
$\varepsilon_{\mathbf{h}} = R_{\mathbf{h}}^2 - 1$	
f_j	Atomic scattering factor of the *j*th atom
$^{\circ}f_j$	Atomic scattering factor of the *j*th atom at rest.
$F_d = \|F_d\| \exp(\mathrm{i}\psi)$	Structure Factor of the isomorphous derivative
$F_H = \begin{cases} \|F_H\| \exp(\mathrm{i}\phi_H) \\ F_d - F_p \end{cases}$	Structure factor of the heavy-atom (added to the native protein)
$F_p = \|F_p\| \exp(\mathrm{i}\phi)$	Structure factor of the native
$F_\pi = \|F_\pi\| \exp(\mathrm{i}\phi_\pi)$	Structure factor of a partial structure
$\Phi_3 \begin{cases} = \phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h-k}} \\ = \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \end{cases}$	with $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$.
$G = 2\|R_{\mathbf{h}} R_{\mathbf{k}} R_{\mathbf{h-k}}\| \left[\sigma_3 / \sigma_2^{3/2}\right]_p$	

$I_i(\mathrm{x})$	I_i is the modified Bessel function of the order i
m	Number of symmetry operators.
N	Number of atoms in the primitive unit cell.
N_d	Number of atoms in the primitive unit cell for the derivative structure.
$N_{eq} = \sigma_2^3 / \sigma_3^2$	Statistically equivalent number of atoms in the primitive unit cell
N_H	Number of heavy atoms in the primitive unit cell.
N_p	Number of atoms in the primitive unit cell for native structure.
p	Number of atoms (symmetry equivalents included) whose positions are a priori known
q	Number of atoms (symmetry equivalents included) whose positions are unknown: $q=N-p$
$\sigma_i = \sum_{j=1}^{N} Z_j^i$	Z_j is the atomic number of the *j*th atom
$\left[\sigma_2^3 / \sigma_3^2\right]_H$	Value of N_{eq} relative to the heavy-atom structure
$\left[\sigma_2^3 / \sigma_3^2\right]_p$	Value of N_{eq} for the native protein
$\left[\sigma_2^3 / \sigma_3^2\right]_\pi$	Value of N_{eq} for the partial structure for the primitive unit cell
$\left[\sigma_2^3 / \sigma_3^2\right]_q$	Value of N_{eq} for the difference structure obtained by subtracting the partial from the protein structure
$\Sigma_d = \sum_{j=1}^{N_d} f_j^2$	
$\Sigma_H = \sum_{j=1}^{N_H} f_j^2$	
$\Sigma_p = \sum_{j=1}^{N_p} f_j^2$	

$$\Sigma_q = \sum_{j=1}^{q} f_j^2$$

$$T = D_1(2\,R'\,S')$$

Chapter I

General aspects

The phase problem

Crystal structure analysis is usually based on diffraction phenomena caused by the interaction of matter with some type of radiation , in general X-rays, electrons, or neutrons. The diffracted intensity (under the kinematic approximation) is simply related to the squared structure factor $\left|F_{\mathbf{h}}\right|^2$, where

$$F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp(2\pi \, \mathrm{i} \, \mathbf{h} \cdot \mathbf{r}_j) \qquad \text{(I.1)}$$

where $\mathbf{h} \equiv h\,\mathbf{a}^* + k\,\mathbf{b}^* + l\,\mathbf{c}^*$ is a vector in the reciprocal space and $\mathbf{r}_j \equiv x_j\,\mathbf{a} + y_j\,\mathbf{b} + z_j\,\mathbf{c}$ the positional vector of the *j*th atom in the unit cell. $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ are the elementary translations of the direct lattice and $\mathbf{a}^*$, $\mathbf{b}^*$, $\mathbf{c}^*$ are the translations of the reciprocal lattice. As it is well known, the two lattices are related by the relation

$$\mathbf{a}^* \cdot \mathbf{b} = \mathbf{a}^* \cdot \mathbf{c} = \mathbf{b}^* \cdot \mathbf{a} = \mathbf{b}^* \cdot \mathbf{c} = \mathbf{c}^* \cdot \mathbf{a} = \mathbf{c}^* \cdot \mathbf{b} = 0$$

$$\mathbf{a}^* \cdot \mathbf{a} = \mathbf{b}^* \cdot \mathbf{b} = \mathbf{c}^* \cdot \mathbf{c} = 1 \qquad \text{(I.2)}$$

The crystal structure is defined if the electron density distribution ρ is known. ρ may be written in the form:

$$\rho(\mathbf{r}) = \mathrm{T}\left[F_{\mathbf{h}}\right] = \frac{1}{V}\sum_{\mathbf{h}} F_{\mathbf{h}} \exp(-2\pi \,\mathrm{i}\, \mathbf{h}\cdot\mathbf{r}) \qquad \text{(I.3)}$$

where T is the Fourier transform function. Viceversa, $F_{\mathbf{h}}$ can be written in terms of ρ by calculating the inverse Fourier transform:

$$F_{\mathbf{h}} = \mathrm{T}^{-1}\left[\rho(\mathbf{r})\right] = \frac{1}{V}\int_V \rho(\mathbf{r})\exp(2\pi \,\mathrm{i}\, \mathbf{h}\cdot\mathbf{r}) \qquad \text{(I.4)}$$

Since the diffraction data provide the magnitudes of the structure factors but not their phases, (I.3) cannot be used to directly obtained the electron density distribution from the experimental data. This is the so called *phase problem*.

The problem must in principle have a solution (even if not necessarily unique), but in this case the unknown quantities (the atomic positions $\mathbf{r}_j$) appear as argument of trigonometric functions:

$$\left|F_{\mathbf{h}}\right|^2 = \sum_{j=1}^{N} f_j^2 + 2\sum_{j>k=1}^{N} f_j f_k \cos 2\pi\; \mathbf{h}\cdot(\mathbf{r}_j - \mathbf{r}_k) \qquad \text{(I.5)}$$

(I.5) is a system of non-linear equations, and its solutions cannot be obtained in any analytical way, even though the number of relationships greatly exceeds the number of unknowns.

If an approximate solution, is available, that is if an *initial structural model* has been obtained, then the system of non-linear equations may be solved. This can then be refined applying least squares methods on minimizing the quantity

$$W = \sum\left(|F_{\mathrm{obs}}| - |F_{\mathrm{cal}}|\right)^2 \tag{I.6}$$

until the best agreement with the experimental data is achieved.

As a measure of agreement between the observed and calculated values of the structure factors, crystallographers commonly use a quantity called reliability index or residual R defined by

$$R = \frac{\sum_{\mathbf{h}} \left| |F_{\mathbf{h}}|_{\mathrm{obs}} - |F_{\mathbf{h}}|_{\mathrm{cal}} \right|}{\sum_{\mathbf{h}} |F_{\mathbf{h}}|_{\mathrm{obs}}} \tag{I.7}$$

Direct methods

Methods which try to derive the structure factor phases directly from the observed amplitudes through mathematical relationships are called *direct methods*. They are responsible of the great deal of crystal structures solved in these last years. Historically, the first mathematical relationships capable of giving phase information were obtained by Harker and Kasper (1948) in the form of inequalities. In 1953 Hauptman and Karle established the basic concepts and the probabilistic foundations of direct methods. Also in 1952 Sayre was able to derive a very important relationships

$$F_{\mathbf{h}} = \vartheta_{\mathbf{h}} \sum_{\mathbf{k}} F_{\mathbf{k}}\, F_{\mathbf{h-k}} \tag{I.8}$$

where $\vartheta_{\mathbf{h}}$ is a known value.

A crucial role in direct methods is played by the *structure invariants*. They are linear combinations of phases which are independent of the choice of origin. Therefore their value depends only on the crystal structure, and therefore may be estimated (in principle) from the observed magnitudes diffraction.

The most general structure invariants is represented by the product

$$F_{\mathbf{h}_1} F_{\mathbf{h}_2} \cdots F_{\mathbf{h}_n} = \left| F_{\mathbf{h}_1} F_{\mathbf{h}_2} \cdots F_{\mathbf{h}_n} \right| \exp\left[\mathrm{i}(\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \ldots + \phi_{\mathbf{h}_n}) \right] \tag{I.9}$$

when

$$\mathbf{h}_1 + \mathbf{h}_2 + \ldots + \mathbf{h}_n = 0 \qquad \text{(I.10)}$$

The simplest structure invariant is F_{000}, its phase is always zero. The structure invariant $F_{\mathbf{h}}\, F_{-\mathbf{h}} = |F_{\mathbf{h}}|^2$ does not contain any phase information, while the triplet invariant $F_{\mathbf{h}} F_{-\mathbf{k}} F_{-\mathbf{h}+\mathbf{k}}$ with phase $\phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h}-\mathbf{k}}$, plays a primary role in the probabilistic procedures for phase determination. With a simple extension it is possible to define the quartet invariant $F_{\mathbf{h}}\, F_{-\mathbf{k}}\, F_{-\mathbf{l}}\, F_{-\mathbf{h}+\mathbf{k}+\mathbf{l}}$ with phase $\phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{l}} - \phi_{\mathbf{h}-\mathbf{k}-\mathbf{l}}$. In the same form we can define quintets, sextets, etc.

Structure semi-invariants are single phases or linear combinations of phases which are invariant with respect to a shift of permissible origin. A basic property of a structure semi-invariants (Giacovazzo, C., 1980) is its capability of being transformed into a structure invariant by adding one or more pairs of symmetry-equivalent phases. For instance, in a given space group possessing the symmetry operator $\mathbf{C} \equiv (\mathbf{R}, \mathbf{T})$, the phase $\phi_{\mathbf{H}}$ is a semi-invariant if it possible to find a reflection $\mathbf{h}$ such that

$$\Psi = \phi_{\mathbf{H}} - \phi_{\mathbf{h}} + \phi_{\mathbf{hR}} \qquad \text{(I.11)}$$

is an invariant, this is $\mathbf{H} - \mathbf{h} + \mathbf{hR} = 0$.

The normalized structure factors

A central role in direct methods is played by the normalized structure factors, defined as

$$|E_{\mathbf{h}}| = |F_{\mathbf{h}}| \big/ \sqrt{\varepsilon_{\mathbf{h}}\, \Sigma} \qquad \text{(I.12)}$$

where $\varepsilon_{\mathbf{h}}$ is the Wilson coefficient depending on the specific indices $(h\ k\ l)$ (for the tabulation of $\varepsilon_{\mathbf{h}}$ see Giacovazzo, C., 1980). From (I.12) one can immediately obtain (under the hypothesis that the atomic positions are random variables with uniform distribution throughout the unit cell)

$$< |E_{\mathbf{h}}|^2 >= 1 \qquad \text{(I.13)}$$

So far we have implicitly assumed that the observed structure factor moduli are on the absolute scale, but in general the values of $\left|F_{\mathbf{h}}\right|^2_{\text{obs}}$ obtained from the intensities are on a relative scale. In the case that we assume an overall isotropic thermal motion equal for all the atoms, we may write

$$\left|F_{\mathbf{h}}\right|^2_{\text{obs}} = K\left|{}^{o}F_{\mathbf{h}}\right|^2 \exp(-2\,Bs^2) \qquad \text{(I.14)}$$

where K is a scale factor, $\left|{}^{o}F_{\mathbf{h}}\right|$ is the structure amplitude in absolute scale for atoms at rest, B is the overall isotropic tempreature factor, and $s = \sin\theta\ /\ \lambda$.

Wilson (1942) proposed a method to derive the values K and B. From (I.14) one obtains

$$\text{Ln}\left(\frac{<\left|F_{\text{obs}}\right|^2>}{\Sigma_s}\right) = \text{Ln}\,K - 2\,B < s^2 > \qquad \text{(I.15)}$$

where

$$\Sigma_s = \sum_{j=1}^{N} {}^{o}f_j^2 \qquad \text{(I.16)}$$

where ${}^{\circ}f_j$ denote the atomic scattering factor for *j*th atom at rest.

If $\text{Ln}\left(\frac{<\left|F_{\text{obs}}\right|^2>}{\Sigma_s}\right)$ is plotted againts $< s^2 >$ and then the best straight line is derived, the intercept of the line on the vertical axis will give us $\text{Ln}\ K$ and its slope the value of $2B$. This is called the *Wilson plot*.

Probabilistic methods

The basic probability formula for triplet invariants was derived by Cochran (1955):

$$\text{P}(\boldsymbol{\Phi}) = \frac{1}{2\pi\, I_o(G)} \exp(G\cos\boldsymbol{\Phi}) \qquad \text{(I.17)}$$

where

$$G = 2\left[\sigma_3 / \sigma_2^{-3/2}\right] \left| E_{\mathbf{h}}\ E_{\mathbf{k}}\ E_{\mathbf{h-k}} \right| \tag{I.18}$$

and

$$\sigma_i = \sum_{j=1}^{N} Z_j^i \tag{I.19}$$

where Z_j is the atomic number of the *j*th atom. $\mathrm{P}(\Phi)$ is a so-called von Mises distribution and G is its concentration parameter.

If more than one pair of phases $\phi_{\mathbf{k}_j}$, $\phi_{\mathbf{h-k}_j}$, with *j*=1, 2,..., *r*, are known, all defining the same $\phi_{\mathbf{h}}$ through triplet relations such us $\phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h-k}}$, then (Karle & Hauptman , 1956)

$$\tan\theta_{\mathbf{h}} = \frac{\sum_{j=1}^{r} G_j \sin(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h-k}_j})}{\sum_{j=1}^{r} G_j \cos(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h-k}_j})} \tag{I.20}$$

where

$$G_j = 2\left[\sigma_3 / \sigma_2^{-3/2}\right] \left| E_{\mathbf{h}} E_{\mathbf{k}_j} E_{\mathbf{h-k}_j} \right| \tag{I.21}$$

gives the most probable value of $\phi_{\mathbf{h}}$. The relation (I.20) is known as the *tangent formula*. This formula plays a central role in the phase determination process.

In the last years probability methods have seen new important developments. Not only it has been possible to improve the estimate of the triplets, but also to derive reliable estimates of other phase relationships. The first step of the new methods is to identify the moduli $\{|E|\}$ that provide information on the structure invariant or semi-invariant Φ, and the second step consist in deriving the probability distribution $\mathrm{P}(\boldsymbol{\Phi}\,|\{|E|\})$, where the vertical bar after $\boldsymbol{\Phi}$ stands for: "given all magnitudes in the set $\{|E|\}$".

Solving crystal structures

The following general scheme for a crystal structure solution may be used:

1. *Normalization*

   The values of the normalized structure factors are calculated as we have described before.
2. *Setting up phase relationships*

   Triplet and quartet relationships are calculated among the reflexions with large $|E|$ values in order to estimate those with the highest reliability. The triplet and quarted search must take into account the space group symmetry.
3. *Definition of an optimun starting set of phases*

   The choice of the reflections in the starting set is very important because the success of the process will depend on them. In general the starting set is formed by the reflections which fix the origin (up to three reflections), one reflection if it is neccesary to fix the enantiomorph, and a limited number of reflections (usually five or more) which are assigned by different techniques (symbolic addition, magic integers permutation, etc). Given the phases in the starting set, all the other phases can be determined one after the other in a chain process.
4. *Figures of merit*

   They are functions which allow an a priori estimate of the goodness of each phase set as representative of the correct solution.
5. *Electron density maps*

   The phase set with the highest CFOM (combined figure of merit) is used for obtain the electron density map. The most modern programs find the peaks and supply a list of maxima sorted in decreasing order of height. This list may then be analysed in terms of distances and angles.
6. *Completing and refining the structure*

   If the model obtained according to the above procedures is not complete, there are differents methods which recovery the complete structure. The most widely used method of structure refinement are the least-squares Fourier-methods.

Protein crystallography

Protein crystallography is a specialized branch of crystallography that investigates, by using diffraction techniques on single crystals, the three-dimensional structure of biological macromolecules. The solvent content makes the difference between a classical molecular crystal and a protein crystal: the protein crystals are much less ordered than classical crystals, not only for the large amount of unordered material present in the crystal itself, but also because surface groups of the macromolecule in contact with the solvent can show a great mobility. As a consequence, diffraction data cannot be measured to the resolution normally attainable with small molecules, and this reduces the quality of information available for direct methods. A futher problem is the large number of atoms in the unit cell: the consequence is that Cochran distribution is rather flat for all the triplet invariants, and therefore it is of very limited usefulness.

The isomorphous derivative

The term isomorphous derivative ideally indicates a crystal where some solvent molecules have been replaced by a group of atoms with more electrons, without any alteration of the structure of the protein or of the crystal lattice itself. In practice, this never happens: the introduction of a bulky compound which interacts with some of the atoms on the surface of the protein will give rise to local movements and displacements of atomic groups, at least in the close vicinity of the binding site. Lack of isomorphism can be confirmed by differents parameters: a change in unit cell parameters or the comparison of structure data between the native a derivative structures.

The solution of the phase problem in proteins

There are two main methods for solving the phase problem in protein crystallography: isomorphous replacement and anomalous scattering techniques. Other methods like molecular replacement, can in some cases help to solve the phase problem, coupled with suitable, translation and rotation functions. Recent developments of direct methods seem able to offer the possibility of *ab-initio* solution of protein crystal structures if at least diffraction data of one isomorphous derivative is available (Giacovazzo, Guagliardi, Ravelli & Siliqi, 1994; Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995; Giacovazzo, Siliqi & Gonzalez-Platas, 1995). As usual for direct methods, we speak of *ab-initio* crystal structure solution when phases are directly derived from diffraction data without any supplementary prior information.

In this case, a mathematical technique can be used (Hauptman, 1982) that integrates direct-methods and isomorphous-replacement techniques. A perfect isomorphism is assumed: accordingly $F_d = F_p + F_H$.

The triplet phase invariants of the protein may then be estimated *via* the following probabilistic formula:

$$\mathrm{P}(\Phi \,|R_{\mathbf{h}}, R_{\mathbf{k}}, R_{\mathbf{h-k}}, S_{\mathbf{h}}, S_{\mathbf{k}}, S_{\mathbf{h-k}}) \cong \left[2\pi\, I_o(A)\right]^{-1} \exp(A\cos\Phi) \qquad \text{(I.22)}$$

where A is a positive or negative term, the value of which depends on an intricate interrelationship among the six moduli, $R_{\mathbf{h}}$, $R_{\mathbf{k}}$, $R_{\mathbf{h-k}}$, $S_{\mathbf{h}}$, $S_{\mathbf{k}}$, $S_{\mathbf{h-k}}$ are the moduli of the native and of the derivative structure factors respectively. Hauptman's approach has been reconsidered and generalized by Giacovazzo, Cascarano & Zheng (1988), who derived a more simply and effective expression for A:

$$A = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_p R_{\mathbf{h}} R_{\mathbf{k}} R_{\mathbf{h-k}} + 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \Delta_{\mathbf{h}} \Delta_{\mathbf{k}} \Delta_{\mathbf{h-k}} \qquad \text{(I.23)}$$

where

$$\Delta = \left(\left|F_d\right| - \left|F_p\right|\right) \Big/ \Sigma_H^{1/2} \qquad \text{(I.24)}$$

In the phasing process a modified tangent formula can be applied, according to which the most probable value of $\phi_{\mathbf{h}}$ is given by

$$\tan\theta_{\mathbf{h}} = \sum_{j=1}^{r} A_j \sin(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h-k}_j}) \Big/ \sum_{j=1}^{r} A_j \cos(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h-k}_j}) = T_{\mathbf{h}} / B_{\mathbf{h}}$$

(I.25)

where now the reliability parameter is

$$\alpha_{\mathbf{h}} = \left(T_{\mathbf{h}}^2 + B_{\mathbf{h}}^2\right)^{1/2} \qquad \text{(I.26)}$$

References

Cochran, W. (1955). *Acta Cryst*. **8**, 473-478.

Giacovazzo, C. (1980). *Direct Methods in Crystallography*. Academic London.

Giacovazzo, C., Cascarano, G. & Zheng, C.-D. (1988). *Acta Cryst*. A**44**, 45-51.

Giacovazzo, C., Guagliardi, A., Ravelli, R. & Siliqi, D. (1994). *Z. Kristallogr.* **209**, 136-142.

Giacovazzo, C., Siliqi, D. & Ralph. A. (1994). *Acta Cryst*. A**50**, 503-510.

Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst*. A**50**, 609-621.

Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *Acta Cryst*. A**51**, 177-188.

Giacovazzo, C. & Gonzalez-Platas, J. (1995). *Acta Cryst*. A**51**, 398-404.

Giacovazzo, C., Siliqi, D. & Gonzalez-Platas, J. (1995). *Acta Cryst*. A**51**, 000-000.

Harker, D. & Kasper, J. S. (1948). *Acta Cryst*. **1**, 70.

Hauptman, H. (1982). *Acta Cryst*. A**38**, 289-294, 632-641.

Hauptman, H. & Karle, J. (1953). *The solution of the phase problem. I. The centrosymmetric crystal*, ACA Monograph, No. 3 Polycrystal Book Service, New York.

Karle, J. & Hauptman, H. (1956). *Acta Cryst*. **9**, 635-651.

Sayre, D. (1952). *Acta Cryst*. **5**, 60-65.

Wilson, A. J. C. (1942). *Nature*, **150**, 151.

Chapter II

The Formula P_{13}

Introduction

The triplet relationship

$$\Phi_3 = \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \cong 0 \quad (\mathrm{mod}\ 2\pi) \tag{II.1}$$

with $\mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_3 = 0$ is the most widely used phase relationship for solving crystal structures. It's very important to use good triplets in the phase determination process because the occurrence of a few bad triplets (i.e. triplets for which (II.1) is violated) in the early stages of a phasing procedure can lead to wrong results even if the set of starting phases is relatively accurate. The probability of finding the correct solution is enhanced if the bad triplet relationships are recognized: then they may be excluded from the structure-solving process or suitably used.

Φ_3 is traditionally estimated by the Cochran (1955) formula (denoted by P_3 from now on):

$$\mathrm{P}_3(\Phi_3) \cong \frac{1}{2\pi\ I_{\mathrm{o}}(C)} \exp\left(C \cos\Phi_3\right) \tag{II.2}$$

where I_o is the modified Bessel function of order zero, $C = 2\sigma_3\, \sigma_2^{-3/2} |E_{\mathbf{h}_1}\, E_{\mathbf{h}_2}\, E_{\mathbf{h}_3}|$, $\sigma_n = \sum_{j=1}^{N} Z_j^n$, Z_j is the atomic number of *j*th atom of the structure, N is the number of atoms in the unit cell. Cochran's formula estimates Φ_3 by exploiting only the information contained in the three moduli $|E_{\mathbf{h}_1}|$, $|E_{\mathbf{h}_2}|$, $|E_{\mathbf{h}_3}|$.

In the last years probability methods have seen new important developments. Not only it has been possible to improve the estimate of the triplets, but also to derive reliable estimates of other phase relationships. The theory of representations proposed by Giacovazzo (1977,1980) is a useful framework for designing the procedures devoted to estimating structure inviariants.

In accordance with this theory we can use the second representation of Φ_3 to obtain a better estimate of Φ_3. The second representation in this case is the collection of special quintets

$$\Psi_2 = \Phi_3 + \phi_{\mathbf{kR}_i} - \phi_{\mathbf{kR}_i}, \qquad i = 1,\ldots,m \tag{II.3}$$

where $\mathbf{k}$ is a free vector in the reciprocal space and m is the number of symmetry operators $\mathbf{C}_j = (\mathbf{R}_j, \mathbf{T}_j)$ not related by a centre of symmetry ($\mathbf{R}_j$ is the rotational part, $\mathbf{T}_j$ the translational part of the symmetry operator). Since $\Psi_2 \equiv \Phi_3$, estimating any Ψ_2 is perfectly equivalent to estimating Φ_3. The basis magnitudes of Ψ_2 are the magnitudes $|E_{\mathbf{h}_1}|$, $|E_{\mathbf{h}_2}|$, $|E_{\mathbf{h}_3}|$, $|E_{\mathbf{k}}|$ and the cross magnitudes are the six terms $|E_{\mathbf{h}_1 \pm \mathbf{kR}_i}|$, $|E_{\mathbf{h}_2 \pm \mathbf{kR}_i}|$, $|E_{\mathbf{h}_3 \pm \mathbf{kR}_i}|$. The collection of the basis and of the cross magnitudes of various quintents Ψ_2 is called the second phasing shell of Φ_3:

$$\{B\}_2 = \left\{ R_{\mathbf{h}_1}, R_{\mathbf{h}_2}, R_{\mathbf{h}_3}, R_{\mathbf{k}}, R_{\mathbf{h}_1 \pm \mathbf{kR}_i}, R_{\mathbf{h}_2 \pm \mathbf{kR}_i}, R_{\mathbf{h}_3 \pm \mathbf{kR}_i} \right\} \quad i = 1,\ldots m$$

where R is the modulus of E. A formula was derived (Cascarano, Giacovazzo, Camalli, Spagna, Burla, Nunzi & Polidori, 1984) which is able to estimate Φ_3 given the moduli in $\{B\}_2$:

$$\mathrm{P}(\Phi_3 | \{B\}_2) \cong \frac{1}{2\pi\, I_o(G)} \exp\left(G \cos\Phi_3\right) \tag{II.4}$$

where

$$G = C(1+Q)$$

$$C = 2\,R_{\mathbf{h}_1}\;R_{\mathbf{h}_2}\;R_{\mathbf{h}_3}\Big/\sqrt{N}$$

$$Q = \sum_{\mathbf{k}}\left[\frac{\displaystyle\sum_{i=1}^{m}{}'\,A_{\mathbf{k},i}\,/\,N}{1+\left(\varepsilon_{\mathbf{h}_1}\varepsilon_{\mathbf{h}_2}\varepsilon_{\mathbf{h}_3}+\displaystyle\sum_{i=1}^{m}{}'\,B_{\mathbf{k},i}\right)/\,2\,N}\right]$$

$$\begin{aligned}A_{\mathbf{k},i} = \varepsilon_{\mathbf{k}}\Big[&\varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}\left(\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}\right)+\\ &\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}\left(\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}\right)+\\ &\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}\left(\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}\right)\Big]\end{aligned}$$

$$\begin{aligned}B_{\mathbf{k},i} = \varepsilon_{\mathbf{h}_1}&\Big[\varepsilon_{\mathbf{k}}\left(\varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}\right)\\ &+\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}\;\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}\;\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}\Big]\\ &+\varepsilon_{\mathbf{h}_2}\Big[\varepsilon_{\mathbf{k}}\left(\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}\right)\\ &+\varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}\;\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}\;\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}\Big]\\ &+\varepsilon_{\mathbf{h}_3}\Big[\varepsilon_{\mathbf{k}}\left(\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i}\right)\\ &+\varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}\;\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}\;\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}\Big]\end{aligned}$$

The prime to summation warns the reader that precautions have to be taken in order to avoid duplications of contributions. In our notation $\varepsilon_{\mathbf{h}}$ stays for $R_{\mathbf{h}}^2 - 1$. The conditional distribution $\mathrm{P}(\Phi_3 | \{B\}_2)$ was denoted P_{10} in order to emphasize the fact that the formula explores reciprocal space by means of ten-nodes figure. Unlike in Cochran formula, G may be positive or negative: in particular if $G < 0$ the triplet is estimated negative. The accuracy with which the value is estimated by (II.4) strongly depends on $\varepsilon_{\mathbf{k}}$: large $\varepsilon_{\mathbf{k}}$ will provide higher contributions to the formula. In practice, only a subset of magnitudes (the reflections $\mathbf{k}$ with large values of ε) may be used for estimating Φ_3.

The long experience with *SIR88* (Burla, Camalli, Cascarano, Giacovazzo, Polidori, Spagna & Viterbo, 1989) and *SIR92* (Altomare, Cascarano, Giacovazzo, Guagliardi, Burla, Polidori & Camalli, 1994), two packages for direct phasing of crystal structures, proved that P_{10} is much more efficient than P_3 formula. Its use often makes the difference between success and failure.

The basic reason for the succes of P_{10} may be described in the following way. For any $\mathbf{k}$ vector the 6*m* cross magnitudes of Φ_3 are listed below:

$$\begin{array}{cccccc}
R_{\mathbf{h}_1+\mathbf{kR}_1} & R_{\mathbf{h}_1-\mathbf{kR}_1} & R_{\mathbf{h}_2+\mathbf{kR}_1} & R_{\mathbf{h}_2-\mathbf{kR}_1} & R_{\mathbf{h}_3+\mathbf{kR}_1} & R_{\mathbf{h}_3-\mathbf{kR}_1} \\
R_{\mathbf{h}_1+\mathbf{kR}_2} & R_{\mathbf{h}_1-\mathbf{kR}_2} & R_{\mathbf{h}_2+\mathbf{kR}_2} & R_{\mathbf{h}_2-\mathbf{kR}_2} & R_{\mathbf{h}_3+\mathbf{kR}_2} & R_{\mathbf{h}_3-\mathbf{kR}_2} \\
. & & & & & . \\
. & & & & & . \\
. & & & & & . \\
R_{\mathbf{h}_1+\mathbf{kR}_m} & R_{\mathbf{h}_1-\mathbf{kR}_m} & R_{\mathbf{h}_2+\mathbf{kR}_m} & R_{\mathbf{h}_2-\mathbf{kR}_m} & R_{\mathbf{h}_3+\mathbf{kR}_m} & R_{\mathbf{h}_3-\mathbf{kR}_m}
\end{array} \qquad \text{(II.5)}$$

From the magnitudes in the *i*th line of the matrix (II.5) six quadrupoles arise,

$$\text{a)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_1+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2} + \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_2-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{h}_1+\mathbf{kR}_i} - \phi_{\mathbf{h}_2-\mathbf{kR}_i} \end{cases} \qquad \text{b)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_1+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{h}_1+\mathbf{kR}_i} - \phi_{\mathbf{h}_3-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_3-\mathbf{kR}_i} \end{cases}$$

$$\text{c)}\begin{cases} \phi_{\mathbf{h}_1}+\phi_{\mathbf{h}_2}+\phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1}+\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_1-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2}-\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_2+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3}-\phi_{\mathbf{h}_1-\mathbf{kR}_i}-\phi_{\mathbf{h}_2+\mathbf{kR}_i} \end{cases} \qquad \text{d)}\begin{cases} \phi_{\mathbf{h}_1}+\phi_{\mathbf{h}_2}+\phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1}-\phi_{\mathbf{h}_2+\mathbf{kR}_i}-\phi_{\mathbf{h}_3-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2}-\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_2+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3}+\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_3-\mathbf{kR}_i} \end{cases} \qquad \text{(II.6)}$$

$$\text{e)}\begin{cases} \phi_{\mathbf{h}_1}+\phi_{\mathbf{h}_2}+\phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1}+\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_1-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2}-\phi_{\mathbf{h}_1-\mathbf{kR}_i}-\phi_{\mathbf{h}_3+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3}-\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_3+\mathbf{kR}_i} \end{cases} \qquad \text{f)}\begin{cases} \phi_{\mathbf{h}_1}+\phi_{\mathbf{h}_2}+\phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1}-\phi_{\mathbf{h}_2-\mathbf{kR}_i}-\phi_{\mathbf{h}_3+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2}+\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_2-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3}-\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_3+\mathbf{kR}_i} \end{cases}$$

each of which giving a well recognizable contribution to P_{10}. In conclusion the P_{10} formula is an efficient way for simultaneously exploiting the information contained in a quite large number of quadrupoles.

Beyond the formula P_{10}

The question is: does (II.6) represent the only quadrupoles exploitable *via* the second representation of Φ_3 ? In case of affirmative answer, P_{10} is a limit formula to which no other information may be added *via* the second representation of the triplet. If other quadrupoles might be identified some supplementary information could be should to improve the triplet estimates.

Other types of quadrupoles do exist. For example,

$$\begin{cases} \phi_{\mathbf{h}_1}+\phi_{\mathbf{h}_2}+\phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_2}-\phi_{\mathbf{kR}_i}+\phi_{\mathbf{h}_2+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3}+\phi_{\mathbf{kR}_j}+\phi_{\mathbf{h}_3-\mathbf{kR}_j} \\ -\phi_{\mathbf{h}_1}-\phi_{(\mathbf{h}_2+\mathbf{kR}_i)\mathbf{R}_p}-\phi_{(\mathbf{h}_3-\mathbf{kR}_j)\mathbf{R}_s} \end{cases} \qquad \text{(II.7)}$$

is a quadrupole too, provided $\mathbf{h}_1+\mathbf{h}_2\mathbf{R}_p+\mathbf{h}_3\mathbf{R}_s+\mathbf{kR}_i\mathbf{R}_p-\mathbf{kR}_j\mathbf{R}_s=0$. (II.7) is structurally different from quadrupoles (II.6) because it involves magnitudes contained in two lines of the matrix

(II.5), and also because the sum of the four triplets in (II.7) is no longer strictly equal to zero. Indeed if the sum D of the four triplet phases in (II.7) is calculated, one obtains:

$$D = 2\pi \left[\mathbf{h}_2\mathbf{T}_p + \mathbf{h}_3\mathbf{T}_s + \mathbf{k}\left(\mathbf{T}_i - \mathbf{T}_j + \mathbf{R}_i\mathbf{T}_p - \mathbf{R}_j\mathbf{T}_s\right)\right]$$

When D=0, the quadrupole is called consistent (Viterbo & Woolfson, 1973), inconsistent in the other cases. Since quadrupoles (II.6) are all consistent, P_{10} cannot exploit any inconsistent quadrupole. It's therefore of some interest to understand if some quintets exist which are referred to quadrupoles (II.7), and to introduce a formalism able to involve such quintets. In this cases, both quadrupole types could be simoultaneously exploited.

Algebraic considerations

Let

$$\Psi_2 = \phi_{\mathbf{h}_1\mathbf{R}_p} + \phi_{\mathbf{h}_2\mathbf{R}_q} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i} - \phi_{\mathbf{hR}_j} \qquad \text{(II.8)}$$

where $\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3$ and $\mathbf{k}$ are chosen so as to satisfy the condition

$$\mathbf{h}_1\mathbf{R}_p + \mathbf{h}_2\mathbf{R}_q + \mathbf{h}_3 + \mathbf{k}(\mathbf{R}_i - \mathbf{R}_j) = 0 \qquad \text{(II.9)}$$

The quintet (II.8) differs from Φ_3 by a know symmetry phase-shift:

$$\Psi_2 = \Phi_3 - 2\pi \left[\mathbf{h}_1\mathbf{T}_p + \mathbf{h}_2\mathbf{T}_q + \mathbf{k}(\mathbf{T}_i - \mathbf{T}_j)\right]$$

Therefore any method estimating Ψ_2 also provides an estimate of Φ_3. Finding for each set of four vectors $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$, $\mathbf{k}$, all the combinations of 4 matrices $\mathbf{R}_p$, $\mathbf{R}_q$, $\mathbf{R}_i$, $\mathbf{R}_j$ for which (II.9) is satisfied is a too long job, even for fast computers. Thus we prefer to limit our study to three subsets of quintets (II.8), more precisely to the following cases:

Case I:

$$\left\{\phi_{\mathbf{h}_1\mathbf{R}_p}+\phi_{\mathbf{h}_2}+\phi_{\mathbf{h}_3}+\phi_{\mathbf{kR}_i}-\phi_{\mathbf{kR}_i\mathbf{R}_p}\right\} \tag{II.10}$$

under the condition

$$(\mathbf{h}_1-\mathbf{kR}_i)(\mathbf{R}_p-\mathbf{I})=0 \tag{II.11}$$

Case II:

$$\left\{\phi_{\mathbf{h}_1}+\phi_{\mathbf{h}_2\mathbf{R}_p}+\phi_{\mathbf{h}_3}+\phi_{\mathbf{kR}_i}-\phi_{\mathbf{kR}_i\mathbf{R}_p}\right\} \tag{II.12}$$

under the condition

$$(\mathbf{h}_2-\mathbf{kR}_i)(\mathbf{R}_p-\mathbf{I})=0 \tag{II.13}$$

Case III:

$$\left\{\phi_{\mathbf{h}_1}+\phi_{\mathbf{h}_2}+\phi_{\mathbf{h}_3\mathbf{R}_p}+\phi_{\mathbf{kR}_i}-\phi_{\mathbf{kR}_i\mathbf{R}_p}\right\} \tag{II.14}$$

under the condition

$$(\mathbf{h}_3-\mathbf{kR}_i)(\mathbf{R}_p-\mathbf{I})=0 \tag{II.15}$$

For each set of four vector $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$, $\mathbf{k}$, we need now to identify only the two matrices $\mathbf{R}_p$ and $\mathbf{R}_i$ which satisfy (II.11), (II.13) or (II.15). Since the three cases have similar properties we focus our attention to the case I: results will then be easily extended to cases II and III.

Let us consider for the case I the generic quintet (II.10). Its phasing shell is constituted by the following 13 magnitudes:

$$\left\{ R_{\mathbf{h}_1}, R_{\mathbf{h}_2}, R_{\mathbf{h}_3}, R_{\mathbf{k}}, R_{\mathbf{h}_1\mathbf{R}_p+\mathbf{kR}_i}, R_{\mathbf{h}_1-\mathbf{kR}_i}, R_{\mathbf{h}_2+\mathbf{kR}_i}, R_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p}, \right. \qquad \text{(II.16)}$$
$$\left. R_{\mathbf{h}_3+\mathbf{kR}_i}, R_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p}, R_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2}, R_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})-\mathbf{h}_2}, R_{\mathbf{kR}_i(\mathbf{R}_p-\mathbf{I})} \right\}$$

In accordance with (II.11) the vector $\mathbf{kR}_i(\mathbf{R}_p - \mathbf{I})$ can be rewritten as $\mathbf{h}_1(\mathbf{R}_p - \mathbf{I})$. This notation emphasizes the fact that when $\mathbf{k}$ varies over reciprocal space and p is kept constant then $\mathbf{kR}_i(\mathbf{R}_p - \mathbf{I})$ remains constant. Therefore the last three terms in (II.16) depend only on $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{h}_3$, $\mathbf{R}_p$.

When $\mathbf{R}_p = \mathbf{I}$ the 13 magnitudes reduce to 10, which constitute the second phasing shell of Φ_3. Accordingly the formulation used for deriving the P_{10} expression is a particular case of the theory described here.

It should be noted that because of (II.11) $\mathbf{h}_1 - \mathbf{kR}_i$ is a special reflexion marked by a Wilson coefficient η≠1 (i.e., η rotation matrices exist for which $(\mathbf{h}_1 - \mathbf{kR}_i)\mathbf{R}_p = \mathbf{h}_1 - \mathbf{kR}_i$. In literature the Wilson coefficient is usually called ε or p. Here it's called η to avoid a conflict with other symbols). According to (II.10) the phase shift between Ψ_2 and Φ_3 reduces to $\Delta = 2\pi\ (\mathbf{h}_1 - \mathbf{kR}_i)\mathbf{T}_p$; therefore Δ≠0 only if $(\mathbf{h}_1 - \mathbf{kR}_i)$ is a systematic absence, otherwise Δ=0.

We also note that equation (II.11) may be written as

$$\mathbf{h}_1\mathbf{R}_p + \mathbf{kR}_i = \mathbf{h}_1 + \mathbf{kR}_i\mathbf{R}_p \qquad \text{(II.17)}$$

If $\mathbf{C}_p$ represents a symmetry operator of order 2 then also $(\mathbf{h}_1\mathbf{R}_p + \mathbf{kR}_i)$ is a special reflexion with η≠1. Indeed

$$\mathbf{h}_1\mathbf{R}_p + \mathbf{kR}_i = (\mathbf{h}_1 + \mathbf{kR}_i\mathbf{R}_p^{-1})\mathbf{R}_p = (\mathbf{h}_1 + \mathbf{kR}_i\mathbf{R}_p)\mathbf{R}_p$$

and, according to (II.17),

$$(\mathbf{h}_1 + \mathbf{kR}_i\mathbf{R}_p)\mathbf{R}_p = \mathbf{h}_1 + \mathbf{kR}_i\mathbf{R}_p \qquad \text{for} \quad \mathbf{R}_p \neq \mathbf{I}$$

We conclude that al least one cross reflexion of the quintets (II.10) is always special: when it coincides with a systematic absence, the phase shift between Ψ_2 and Φ_3 is different from zero.

Quintet (II.10) exploits the following 14 quadrupoles (for a simpler writing we will often denote $\mathbf{h}_1(\mathbf{R}_p - \mathbf{I}) - \mathbf{h}_2$ as $\mathbf{h}_1\mathbf{R}_p + \mathbf{h}_3$):

$$\text{a)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1} + \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_1-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_2+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{h}_1-\mathbf{kR}_i} - \phi_{\mathbf{h}_2+\mathbf{kR}_i} \end{cases} \qquad \text{b)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1} + \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_1-\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{h}_1-\mathbf{kR}_i} - \phi_{\mathbf{h}_3+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_3+\mathbf{kR}_i} \end{cases}$$

$$\text{c)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{h}_1\mathbf{R}_p} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} + \phi_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \end{cases} \qquad \text{d)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} + \phi_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{h}_1\mathbf{R}_p} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \end{cases}$$

$$\text{e)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_3} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \\ -\phi_{\mathbf{h}_2} + \phi_{\mathbf{kR}_i\mathbf{R}_p} + \phi_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p} \\ -\phi_{\mathbf{kR}_i} - \phi_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \end{cases} \qquad \text{f)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_3} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_2+\mathbf{kR}_i} \\ +\phi_{\mathbf{kR}_i\mathbf{R}_p} - \phi_{\mathbf{h}_2+\mathbf{kR}_i} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \end{cases}$$

(II.18)

$$\text{g)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2} + \phi_{\mathbf{kR}_i\mathbf{R}_p} + \phi_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{kR}_i} - \phi_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p} \end{cases} \qquad \text{h)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{kR}_i} - \phi_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p} \\ -\phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i\mathbf{R}_p} + \phi_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p} \end{cases}$$

$$\text{i)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \\ \phi_{\mathbf{kR}_i\mathbf{R}_p} - \phi_{\mathbf{h}_3+\mathbf{kR}_i} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_3+\mathbf{kR}_i} \end{cases} \qquad \text{j)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \\ -\phi_{\mathbf{kR}_i} - \phi_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \\ -\phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i\mathbf{R}_p} + \phi_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p} \end{cases}$$

$$\text{k)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_2+\mathbf{kR}_i} - \phi_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_2+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i\mathbf{R}_p} + \phi_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p} \end{cases} \qquad \text{l)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p} - \phi_{\mathbf{h}_3+\mathbf{kR}_i} \\ -\phi_{\mathbf{h}_2} + \phi_{\mathbf{kR}_i\mathbf{R}_p} + \phi_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{kR}_i} + \phi_{\mathbf{h}_3+\mathbf{kR}_i} \end{cases}$$

$$\text{m)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{kR}_i} + \phi_{\mathbf{kR}_i\mathbf{R}_p} - \phi_{\mathbf{kR}_i(\mathbf{R}_p-\mathbf{I})} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \\ -\phi_{\mathbf{h}_3} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} + \phi_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \end{cases} \qquad \text{n)}\begin{cases} \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} \\ -\phi_{\mathbf{kR}_i} + \phi_{\mathbf{kR}_i\mathbf{R}_p} - \phi_{\mathbf{kR}_i(\mathbf{R}_p-\mathbf{I})} \\ -\phi_{\mathbf{h}_1\mathbf{R}_p} - \phi_{\mathbf{h}_3} + \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \\ -\phi_{\mathbf{h}_2} - \phi_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} + \phi_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \end{cases}$$

The reader will easily verify that some of the quadrupoles (II.18) are in common with the quadrupoles (II.6). Such an overlapping will be reflected in the probabilistic formula estimating triplets *via* 13 moduli, which will therefore have terms in common with P_{10}.

The conditional probabilistic formula P_{13}

Let us consider the quintet (case I of the preceding section)

$$\phi_{\mathbf{h}_1\mathbf{R}_p} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i} - \phi_{\mathbf{kR}_i\mathbf{R}_p} \qquad \text{(II.19)}$$

The method of joint probability distribution function of structure factors (Hauptman & Karle, 1953; Klug, 1958; Giacovazzo, 1980) will be used to derive the conditional probability

$$P\left(\Phi_3 | R_{\mathbf{h}_1}, R_{\mathbf{h}_2}, R_{\mathbf{h}_3}, R_{\mathbf{k}}, R_{\mathbf{h}_1\mathbf{R}_p+\mathbf{kR}_i}, R_{\mathbf{h}_1-\mathbf{kR}_i}, R_{\mathbf{h}_2+\mathbf{kR}_i}, R_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p}, \right.$$
$$\left. R_{\mathbf{h}_3+\mathbf{kR}_i}, R_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p}, R_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2}, R_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3}, R_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})}\right)$$

under the condition (II.11).

For the sake of simplicity we will not give any detail about the mathematical derivation. We only underline that $\mathbf{k}$ is a free vector which can vary over reciprocal space and that $\mathbf{R}_i$ is a rotation matrix which can freely vary over the set of rotation matrices included in the space group. Once the vector $\mathbf{kR}_i$ and the matrix $\mathbf{R}_p$ satisfies (II.11) then the conditional probability of Φ_3 given 13 magnitudes is calculated. Contributions arising from different $\mathbf{k}$ and different $\mathbf{R}_i$ can be combined with each other to give the general formula:

$$P(\Phi_3 | \ldots) \cong \frac{1}{2\pi\, I_o(G')} \exp(G' \cos\Phi_3) \qquad \text{(II.20)}$$

where

$$G' = C(1 + Q')$$

$$Q' = \sum_{\mathbf{k}} \left[\frac{\sum_{i=1}^{m}{}' A'_{\mathbf{k},i} / N}{1 + \left(\varepsilon_{\mathbf{h}_1} \varepsilon_{\mathbf{h}_2} \varepsilon_{\mathbf{h}_3} + \sum_{i=1}^{m}{}' B'_{\mathbf{k},i} \right) / 2N} \right]$$

$$
\begin{aligned}
A'_{\mathbf{k},i} = \sum_{p=1}^{m}{}' \Bigg\{ & \varepsilon_{\mathbf{k}} \Bigg[\varepsilon_{\mathbf{h}_1-\mathbf{k}\mathbf{R}_i} \Big(\varepsilon_{\mathbf{h}_2+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_3+\mathbf{k}\mathbf{R}_i} \Big) \\
& + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{k}\mathbf{R}_i} \Big(\varepsilon_{\mathbf{h}_2-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} + \varepsilon_{\mathbf{h}_3-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Big) \\
& + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \Big(\varepsilon_{\mathbf{h}_3+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_3-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Big) \\
& + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \Big(\varepsilon_{\mathbf{h}_2+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_2-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Big) \\
& + \varepsilon_{\mathbf{h}_2+\mathbf{k}\mathbf{R}_i}\ \varepsilon_{\mathbf{h}_3-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} + \varepsilon_{\mathbf{h}_3+\mathbf{k}\mathbf{R}_i}\ \varepsilon_{\mathbf{h}_2-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Bigg] \\
& + \frac{1}{4} \Big(\varepsilon_{\mathbf{k}} - 2 \Big) \varepsilon_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \Big(\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \Big) \Bigg\} \cos\mathit{\Delta} \\
& + \frac{1}{4} \Big(\varepsilon_{\mathbf{h}_1} - 2 \Big) \varepsilon_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \Big(\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \Big)
\end{aligned}
$$

$$
\mathit{\Delta} = 2\pi \Big(\mathbf{h}_1 - \mathbf{k}\mathbf{R}_i \Big) \mathbf{T}_p
$$

$$
\begin{aligned}
B'_{\mathbf{k},i} = \sum_{p=1}^{m}{}' \Bigg\{ & \varepsilon_{\mathbf{h}_1} \Bigg[\varepsilon_{\mathbf{k}} \Big(\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_1-\mathbf{k}\mathbf{R}_i} \Big) \\
& + \varepsilon_{\mathbf{h}_2+\mathbf{k}\mathbf{R}_i}\ \varepsilon_{\mathbf{h}_3-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} + \varepsilon_{\mathbf{h}_2-\mathbf{k}\mathbf{R}_i\mathbf{R}_p}\ \varepsilon_{\mathbf{h}_3+\mathbf{k}\mathbf{R}_i} \\
& + \varepsilon_{\mathbf{h}_2}\ \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} + \varepsilon_{\mathbf{h}_3}\ \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \Bigg] \\
& + \varepsilon_{\mathbf{h}_2} \Bigg[\varepsilon_{\mathbf{k}} \Big(\varepsilon_{\mathbf{h}_2+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_2-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Big) + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{k}\mathbf{R}_i}\ \varepsilon_{\mathbf{h}_3-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \\
& + \varepsilon_{\mathbf{h}_1-\mathbf{k}\mathbf{R}_i}\ \varepsilon_{\mathbf{h}_3+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3}\ \varepsilon_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \Bigg] \\
& + \varepsilon_{\mathbf{h}_3} \Bigg[\varepsilon_{\mathbf{k}} \Big(\varepsilon_{\mathbf{h}_3+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_3-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Big) + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{k}\mathbf{R}_i}\ \varepsilon_{\mathbf{h}_2-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \\
& + \varepsilon_{\mathbf{h}_1-\mathbf{k}\mathbf{R}_i}\ \varepsilon_{\mathbf{h}_2+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2}\ \varepsilon_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})} \Bigg] \\
& + \varepsilon_{\mathbf{k}} \Bigg[\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3} \Big(\varepsilon_{\mathbf{h}_2+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_2-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Big) \\
& + \varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2} \Big(\varepsilon_{\mathbf{h}_3+\mathbf{k}\mathbf{R}_i} + \varepsilon_{\mathbf{h}_3-\mathbf{k}\mathbf{R}_i\mathbf{R}_p} \Big) \Bigg] \Bigg] \Bigg\}
\end{aligned}
$$

Let us consider now the case II. The quintet (II.12) may be written down as

$$\Psi_2 = \phi_{\mathbf{h}_2\mathbf{R}_p} + \phi_{\mathbf{h}_1} + \phi_{\mathbf{h}_3} + \phi_{\mathbf{kR}_i} - \phi_{\mathbf{kR}_i\mathbf{R}_p}$$

and depends on the following ordered set of magnitudes:

$$\Big\{R_{\mathbf{h}_2}, R_{\mathbf{h}_1}, R_{\mathbf{h}_3}, R_{\mathbf{k}}, R_{\mathbf{h}_2\mathbf{R}_p+\mathbf{kR}_i}, R_{\mathbf{h}_2-\mathbf{kR}_i}, R_{\mathbf{h}_1+\mathbf{kR}_i}, R_{\mathbf{h}_1-\mathbf{kR}_i\mathbf{R}_p}, \\ R_{\mathbf{h}_3+\mathbf{kR}_i}, R_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p}, R_{\mathbf{h}_2\mathbf{R}_p+\mathbf{h}_1}, R_{\mathbf{h}_2\mathbf{R}_p+\mathbf{h}_3}, R_{\mathbf{h}_2(\mathbf{R}_p-\mathbf{I})}\Big\} \qquad \text{(II.21)}$$

The quintet (II.14) may be written down as

$$\Psi_2 = \phi_{\mathbf{h}_3\mathbf{R}_p} + \phi_{\mathbf{h}_2} + \phi_{\mathbf{h}_1} + \phi_{\mathbf{kR}_i} - \phi_{\mathbf{kR}_i\mathbf{R}_p}$$

and depends on the following ordered set of magnitudes:

$$\Big\{R_{\mathbf{h}_3}, R_{\mathbf{h}_2}, R_{\mathbf{h}_1}, R_{\mathbf{k}}, R_{\mathbf{h}_3\mathbf{R}_p+\mathbf{kR}_i}, R_{\mathbf{h}_3-\mathbf{kR}_i}, R_{\mathbf{h}_2+\mathbf{kR}_i}, R_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p}, \\ R_{\mathbf{h}_1+\mathbf{kR}_i}, R_{\mathbf{h}_1-\mathbf{kR}_i\mathbf{R}_p}, R_{\mathbf{h}_3\mathbf{R}_p+\mathbf{h}_2}, R_{\mathbf{h}_3\mathbf{R}_p+\mathbf{h}_1}, R_{\mathbf{h}_3(\mathbf{R}_p-\mathbf{I})}\Big\} \qquad \text{(II.22)}$$

The ordered set (II.16) has its image in the ordered set (II.21) if $\mathbf{h}_1$ is replaced by $\mathbf{h}_2$ and viceversa. Accordingly the ordered set (II.22) is the image of the ordered set (II.16) if $\mathbf{h}_1$ and $\mathbf{h}_3$ change their roles. In conclusion the same formula holds for all the cases if the position of magnitudes in (II.16), (II.21) and (II.22) are considered rather than their indices. Then the final formula P_{13}, collecting contributions from different cases and from different $\mathbf{k}$'s, may be so written:

$$\mathrm{P}_{13} = \mathrm{P}(\Phi_3|\ldots) \cong \frac{1}{2\pi\, I_o(G'')} \exp(G'' \cos\Phi_3) \qquad \text{(II.23)}$$

where

$$G'' = C(1+Q'')$$

$$Q'' = \sum_{\textbf{cases}}{}' \sum_{\mathbf{k}}{}' \left[\frac{\sum_{i=1}^{m}{}' A'_{\mathbf{k},i} / \mathbf{N}}{1 + \left(\varepsilon_{\mathbf{h}_1} \varepsilon_{\mathbf{h}_2} \varepsilon_{\mathbf{h}_3} + \sum_{i=1}^{m}{}' B'_{\mathbf{k},i} \right) / 2N} \right]$$

$$\begin{aligned} A'_{\mathbf{k},i} = \sum_{p=1}^{m} \Big\{ & \varepsilon_4 \big[\varepsilon_6\varepsilon_7 + \varepsilon_6\varepsilon_9 + \varepsilon_5\varepsilon_8 + \varepsilon_5\varepsilon_{10} + \varepsilon_9\varepsilon_{11} \\ & + \varepsilon_{10}\varepsilon_{11} + \varepsilon_7\varepsilon_{12} + \varepsilon_8\varepsilon_{12} + \varepsilon_7\varepsilon_{10} + \varepsilon_8\varepsilon_9 \big] \\ & + \frac{1}{4}(\varepsilon_4 - 2)\varepsilon_{11}\varepsilon_{13} + \frac{1}{4}(\varepsilon_4 - 2)\varepsilon_{12}\varepsilon_{13} \Big\} \cos\mathit{\Delta} \\ & + \frac{1}{4}(\varepsilon_1 - 2)\varepsilon_{11}\varepsilon_{13} + \frac{1}{4}(\varepsilon_1 - 2)\varepsilon_{12}\varepsilon_{13} \end{aligned}$$

$$\begin{aligned} B'_{\mathbf{k},i} = \sum_{p=1}^{m}{}' \Big\{ & \varepsilon_1 \big[\varepsilon_4\varepsilon_5 + \varepsilon_4\varepsilon_6 + \varepsilon_7\varepsilon_{10} + \varepsilon_8\varepsilon_9 \\ & + \varepsilon_2\varepsilon_{11} + \varepsilon_3\varepsilon_{12} \big] \\ & + \varepsilon_2 \big[\varepsilon_4\varepsilon_7 + \varepsilon_4\varepsilon_8 + \varepsilon_5\varepsilon_{10} + \varepsilon_6\varepsilon_9 + \varepsilon_{12}\varepsilon_{13} \big] \\ & + \varepsilon_3 \big[\varepsilon_4\varepsilon_9 + \varepsilon_4\varepsilon_{10} + \varepsilon_5\varepsilon_8 + \varepsilon_6\varepsilon_7 + \varepsilon_{11}\varepsilon_{13} \big] \\ & + \varepsilon_4 \big[\varepsilon_7\varepsilon_{12} + \varepsilon_8\varepsilon_{12} + \varepsilon_9\varepsilon_{11} + \varepsilon_{10}\varepsilon_{11} \big] \Big\} \end{aligned}$$

Pedices of ε indicate the position of the related magnitude in (II.16), (II.21), (II.22).

First applications of the formula P$_{13}$

In order to check the practical effectiveness of P$_{13}$ for $\mathbf{R}_p \neq \mathbf{I}$ we have suitably modified the *SIR92* program. Triplets are sought among the NLAR reflexions with largest *R* values (NLAR is fixed by the program) and estimated according to Cochran (1955) formula (denoted here as P_3), and P_{10} respectively.

Nine test structures were used: for them we give in Table I references, space groups and main crystal data.

Structure Code	Space Group	Molecular Formula	Z
AX118[(1)]	Pccn	$C_{19}H_{21}N_2O_3Cl$	8
AZET[(*)]	$Pca2_1$	$C_{21}H_{16}Cl\,N\,O$	8
BED[(*)]	I4	$C_{26}H_{26}N_4O_4$	8
BOBBY[(*)]	$P2_13$	$Na^+Ca^{++}N(CH_2CO_2)_3^{3-}$	4
CUIMID[(*)]	$P3_221$	$C_6H_8N_4Cl\,Cu$	6
DIAM[(*)]	$P4_2\,/\,n$	$C_{14}H_{20}O$	8
DIOLE[(*)]	$I\bar{4}2d$	$C_{10}H_{18}O_2$	16
FEGAS[(2)]	$P6_3\,/\,mmc$	$Fe_2Ga_2S_5$	2
INOS[(*)]	$P2_1\,/\,n$	$C_6H_{12}O_6 \cdot H_2O$	8

Table I: Code name, space group crystallochemical data for the test structures.
(*) Complete references for such structures are not given for the sake of brevity. The reader is referred to magnetic tape distribuited by crystallographic group in Göttingen.
(1) Distributed by the Crystallographic Group of York.
(2) Cascarano, Douggy-Smiri & Nguyen-Huy Dung, (1987).

P_{10} and P_{13} formulas require that the vector $\mathbf{k}$ is allowed to vary over a number of reflexions. *SIR92* fixed for P_{13} the same number of $\mathbf{k}$ vectors used for P_{10}: let $N_{\mathbf{k}}$ be this value (see Table II) and let $< N_{\mathbf{kR}_i} >$ be the average number of the $\mathbf{kR}_i$ vectors involved in quintets (II.3) exploited for each triplet by P_{10}.

CODE	NLAR	NTRIP	$\mathrm{N}_{\mathbf{k}}$	$<\mathrm{N}_{\mathbf{kR}_i}>$	$<\mathrm{N}'_{\mathbf{kR}_i}>$	ρ_3	ρ_{10}	ρ_{13}	ρ_{10+13}
AX118	300	7025	46	156	60	0.200	0.635	0.503	0.644
AZET	342	8000	60	186	50	0.156	0.406	0.284	0.413
BED	286	4585	64	186	1	0.176	0.282	0.183	0.281
BOBBY	68	2217	32	196	21	0.329	0.732	0.427	0.736
CUIMID	198	4204	33	120	5	0.224	0.657	0.284	0.657
DIAM	260	6455	46	162	41	0.193	0.553	0.430	0.571
DIOLE	182	6508	42	230	34	0.179	0.275	0.212	0.280
FEGAS	71	1334	30	180	320	0.311	0.773	0.543	0.785
INOS	304	3572	56	99	17	0.188	0.660	0.412	0.656

Table II: NTRIP is the number of triplets calculated by SIR92. For the other symbols see the main text.

Because of algebraic reasons only a subset of the vectors $\mathbf{kR}_i$ will satisfy (II.11) or (II.13) or (II.15): we denoted by $<\mathrm{N}'_{\mathbf{kR}_i}>$ the average number of quintets (II.10) or (II.12) or (II.14) exploited for each triplet by P_{13}.

The relative efficiency of P_3 and P_{13} can be deduced by Tables III , IV where triplets estimates are ranked as a function of ARG (ARG equal to C or G' or G'' according to circumstances). In Table III n is the number of triplets with $|C|$ or $|G''|$ larger than ARG, nw is the number of wrong estimates. In Table IV $<|\Phi|^{\mathrm{o}}>$ is the average absolute deviation of the triplet phase from 2π. We have calculated P_{13} only for the cases in which $\mathbf{R}_p \neq \mathbf{I}$, in order to check the usefulness of the terms not included in the P_{10} formula.

Tables III and IV show that P_{13} is an efficient tool both for ranking positive triplets and for picking up negatives ones. It proves to be a formula more accurate than P_3, and therefore may constitute a useful alternative to it. In the same tables we show the corresponding statistics obtained for P_{10}. It is inmediately seen that P_{13}, calculated for $\mathbf{R}_p \neq \mathbf{I}$ is not better than P_{10} and is highly correlated with it. In order to have a simple figure for measuring the relative efficiency of the three formulas we calculated for each structure the correlation coefficient

$$\rho = \frac{<\left(\cos\phi_{\mathrm{T}} - <\cos\phi_{\mathrm{T}}>\right)\left(D_1(\mathrm{ARG}) - <D_1(\mathrm{ARG})>\right)>}{\left[<\left(\cos\phi_{\mathrm{T}} - <\cos\phi_{\mathrm{T}}>\right)^2>\right]^{1/2}\left[<\left(D_1(\mathrm{ARG}) - <D_1(\mathrm{ARG})>\right)^2>\right]^{1/2}}$$

P_3			P_{13}				P_{10}			
			POSITIVE EST. TRIP.		**NEGATIVE EST. TRIP.**		**POSITIVE EST. TRIP.**		**NEGATIVE EST. TRIP.**	
ARG	**n**	**(nw)**	**n**	**(nw)**	**n**	**(nw)**	**n**	**(nw)**	**n**	**(nw)**
0.4	6823	(843)	3435	(53)	93	(14)	3794	(17)	131	(19)
0.8	2672	(143)	2672	(24)	14	(1)	2919	(8)	11	(0)
1.2	740	(13)	1514	(11)	3	(0)	1547	(1)	1	(0)
1.6	197	(0)	831	(1)	1	(0)	675	(0)		
2.0	56	(0)	425	(0)	1	(0)	276	(0)		

Table III: AX118 - Triplet statistics. The P_{13} formula has been calculated for $\mathbf{R}_p \neq \mathbf{I}$.

P_3			P_{13}				P_{10}			
			POSITIVE EST. TRIP.		**NEGATIVE EST. TRIP.**		**POSITIVE EST. TRIP.**		**NEGATIVE EST. TRIP.**	
ARG	**n**	$<\|\Phi\|>°$	**n**	$<\|\Phi\|>°$	**n**	$<\|\Phi\|>°$	**n**	$<\|\Phi\|>°$	**n**	$<\|\Phi\|>°$
0.4	8000	50.5	5708	44.4	87	87.6	5373	40.3	55	109.9
1.2	1409	38.8	1972	35.5	9	121.2	2341	32.0	1	122.0
2.0	95	27.2	521	29.4	2	89.5	597	25.8		
3.2			111	25.4			67	22.3		
4.4			25	26.9			4	23.0		

Table IV: AZET - Triplet statistics. The P_{13} formula has been calculated for $\mathbf{R}_p \neq \mathbf{I}$.

where $\cos\phi_T$ is the true cosine of the triplet and $D_1(\mathrm{ARG})$ is the expected value of the triplet cosine according to P_3, P_{10} and P_{13}. Accordingly, for each structure three correlation factors ρ_3, ρ_{10} and ρ_{13} are calculated which correspond to Cochran (P_3), P_{10} and P_{13} formulas respectively (see Table II). It's easy seen that:

1. ρ_{13} is always much higher than ρ_3. The indications of Tables III and IV are therefore confimed.

2. ρ_{10} is always larger than ρ_3. This corroborates the well documented higher efficiency of P_{10} with respect to Cochran formula.
3. For low symmetry space groups $< \mathrm{N}'_{\mathbf{kR}_i} >$ is very small: consequently P_{13} does not provide a relevant improvement of P_3 perfomances. We then decided to recalculate P_{13} after having increased the value of $\mathrm{N}_{\mathbf{k}}$ and, as a consequence, the value of $< \mathrm{N}'_{\mathbf{kR}_i} >$.The results are in Table V and show that ρ_{13} generally increases even if, for low symmetry space groups (i.e., for BED), conditions (II.11) or (II.13) or (II.15) are hardly satisfied.
4. In most cases ρ_{13} is significantly close to ρ_{10}. This correlation seems not casual and suggests a supplementary algebraic and statistical analysis of the P_{13} formula.

CODE	$\mathrm{N}_{\mathbf{k}}$	$< \mathrm{N}'_{\mathbf{kR}_i} >$	ρ_{13}
AX118	150	185	0.556
AZET	150	126	0.314
BED	150	3	0.175
BOBBY	68	45	0.447
CUIMID	150	24	0.344
DIAM	150	116	0.464
DIOLE	150	129	0.233
FEGAS	71	770	0.635
INOS	150	44	0.440

Table V: For the symbols see the main text. The correlation coefficient has been calculated using P_{13} formula with $\mathbf{R}_p \neq \mathbf{I}$.

Algebraic and statistical analysis of the P_{13} formula

If $\mathbf{R}_p = \mathbf{I}$ condition (II.11) is verified for any $\mathbf{kR}_i$: then $A'_{\mathbf{k},i}$ and $B'_{\mathbf{k},i}$ terms coincide with terms $A_{\mathbf{k},i}$ and $B_{\mathbf{k},i}$ in P_{10}. Thus the present formulation encompasses the P_{10} formalism. However some theoretical and practical drawbacks limit the usefulness of the present theory. For example, as for

P_{10}, the prime to the summation warns the reader that precautions have to be taken in order to avoid duplications of contributions(i.e., if $\mathbf{R}_p$ is a symmetry operator of order two ($\mathbf{R}_p = \mathbf{R}_p^{-1}$), then $A'_{\mathbf{k},i}$ and $B'_{\mathbf{k},i}$ do not change when $\mathbf{R}_i$ is replaced by $-\mathbf{R}_i\mathbf{R}_p$). While duplications of contributions can be easily avoided for P_{10}, a computer program able to eliminate all of them from (II.23) is too time consuming even for fast computer. Thus (II.23) would result less efficient in practice than theoretical expected.

Let us now compare (II.20) with P_{10}, with special attention to the comparison between $A_{\mathbf{k},i}$ and $A'_{\mathbf{k},i}$ ($A_{\mathbf{k},i}$ and $A'_{\mathbf{k},i}$ influence the sign of $\cos\Phi_3$, while $B_{\mathbf{k},i}$ and $B'_{\mathbf{k},i}$ are only scaling factors). We note:

1. The term $\varepsilon_{\mathbf{k}}\varepsilon_{\mathbf{h}_1-\mathbf{kR}_i}(\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i})$ is both in $A_{\mathbf{k},i}$ and in $A'_{\mathbf{k},i}$ (the quadrupoles (II.18a) and (II.18b) are also in (II.6)), but is multiplied in $A'_{\mathbf{k},i}$ by $\cos\Delta$. This is not a contradiction. Indeed $\Delta \neq 2\pi$ n only if the reflexion with vectorial index $\mathbf{h}_1 - \mathbf{kR}_i$ is systematically absent: but in this case the term itself vanishes.
2. If $\mathbf{R}_i$ is replaced by $\mathbf{R}_i\mathbf{R}_p^{-1}$ and $\mathbf{R}_p$ is a symmetry operator of order two then the term $\varepsilon_{\mathbf{k}}\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{kR}_i}(\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p}+\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p})$ (from quadrupoles (II.18g) and (II.18h)) is replaced by $\varepsilon_{\mathbf{k}}\varepsilon_{\mathbf{h}_1+\mathbf{kR}_i}(\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i})$ which is also included in $A_{\mathbf{k},i}$ (that is not true if $\mathbf{R}_p$ is not a symmetry operator of order two). The fact that in $A'_{\mathbf{k},i}$ the term is multiplied by $\cos\Delta$ is not a contradiction. In the section dedicated to algebraic considerations we showed that if $\mathbf{R}_p^{-1} = \mathbf{R}_p$ then $\mathbf{h}_1\mathbf{R}_p + \mathbf{kR}_i$ is a special reflexion with η≠1. Therefore $\Delta \neq 2\pi$ n only if the reflexion with vectorial index $\mathbf{h}_1\mathbf{R}_p + \mathbf{kR}_i$ is a systematically absent reflexion: but in this case the term itself vanishes.
3. For a fixed $\mathbf{R}_p$ the term $\frac{1}{4}\left(\varepsilon_{\mathbf{h}_1}-2\right)\varepsilon_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})}\left(\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2}+\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3}\right)$ (from quadrupoles (II.18c) and (II.18d)) does not change with $\mathbf{k}$. Consequently its role in (II.20) is statistically not relevant. In addition it is based on special quadrupoles (see relationships (II.18)) involving Σ_1 relations which are unreliable for complex structures. An analogous conclusion holds for the term $\frac{1}{4}\left(\varepsilon_{\mathbf{k}}-2\right)\varepsilon_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})}\left(\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2}+\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3}\right)$ arising from quadrupoles (II.18m) and (II.18n).

4. The term

$$\varepsilon_{\mathbf{k}}\Big[\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_2}\Big(\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p}\Big)$$
$$+\varepsilon_{\mathbf{h}_1\mathbf{R}_p+\mathbf{h}_3}\Big(\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}+\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p}\Big)\Big]$$

arising from quadrupoles (II.18i), (II.18j), (II.18f) and (II.18e), is present both in $A'_{\mathbf{k},i}$ and in $B'_{\mathbf{k},i}$. When it is large it contributes to $A'_{\mathbf{k},i}/N$ meaningfully, but at the same time it makes $B'_{\mathbf{k},i}/2N$ large so auto-reducing its own influence on the formula (II.20). Generally speaking large value of this term are asociated to large variance values. This strange behaviour may be so explained: while typical quadrupoles strengthening Φ_3 involve triplets each of which containing $\mathbf{h}_1$ or $\mathbf{h}_2$ or $\mathbf{h}_3$, quadrupoles (II.18e), (II.18f), (II.18i) and (II.18j) do not satisfy this condition. They are based on two-phase seminvariants (for example, the sum of the last two triplets in the quadrupole (II.18e) is the two-phase seminvariant $\phi_{\mathbf{h}_1(\mathbf{R}_p-\mathbf{I})-\mathbf{h}_2}-\phi_{\mathbf{h}_2}-2\pi\,\mathbf{kR}_i\mathbf{T}_p$) which are unreliable for complex structures.

5. The term $\varepsilon_{\mathbf{k}}\Big[\varepsilon_{\mathbf{h}_2+\mathbf{kR}_i}\varepsilon_{\mathbf{h}_3-\mathbf{kR}_i\mathbf{R}_p}+\varepsilon_{\mathbf{h}_3+\mathbf{kR}_i}\varepsilon_{\mathbf{h}_2-\mathbf{kR}_i\mathbf{R}_p}\Big]$ arising from quadrupoles (II.18k) and (II.18l), is not present in $A_{\mathbf{k},i}$: it contains useful information supplementary to that provided by P_{10}. In particular it is able to exploit quadrupoles of type (II.7) (not accessible to P_{10}) because it involves magnitudes contained in two different lines of the matrix (II.5). It should be worthwhile calculating the role of the quadrupoles (II.18k) and (II.18l) in the P_{13} formula. We neglect in $A'_{\mathbf{k},i}$ all the terms but $\varepsilon_4\left(\varepsilon_7\varepsilon_{10}+\varepsilon_8\varepsilon_9\right)$ and in $B'_{\mathbf{k},i}$ all the terms but $\left\{\varepsilon_1\left(\varepsilon_7\varepsilon_{10}+\varepsilon_8\varepsilon_9\right)+\varepsilon_2\left(\varepsilon_4\varepsilon_7+\varepsilon_4\varepsilon_8\right)+\varepsilon_3\left(\varepsilon_4\varepsilon_9+\varepsilon_4\varepsilon_{10}\right)\right\}$. The results for AX118 and AZET are shown in Table VI and VII.

AX118 P_{13}				
	POSITIVE ESTIMATED TRIPLETS		NEGATIVE ESTIMATED TRIPLETS	
ARG	**n**	**(nw)**	**n**	**(nw)**
0.4	3512	(144)	62	(21)
0.8	2527	(70)	13	(11)
1.2	1338	(23)	2	(0)
1.6	738	(12)		
2.0	400	(6)		

Table VI: Triplet statistics calculated using only the contrubition of quadrupoles (II.18k) and (II.18l).

AZET P_{13}				
	POSITIVE ESTIMATED TRIPLETS		NEGATIVE ESTIMATED TRIPLETS	
ARG	**n**	$<\|\Phi\|>°$	**n**	$<\|\Phi\|>°$
0.4	6435	47.7	17	96.0
1.2	1805	39.2	4	139.2
2.0	343	34.3		
3.2	56	29.5		
4.4	5	25.2		

Table VII: Triplet statistics calculated using only the contribution of quadrupoles (II.18k) and (II.18l).

We see that the simple use of the information contained in the quadrupole (II.7) is able to identify negative triplets and to efficiently rank the positive ones. The phase indications provided by quadrupoles (II.18k) and (II.18l) well agree with those obtained through the complete P_{13} formula (see Tables III and

IV). We have combined the contributions from P_{10} just by adding the corresponding numerators and respective denominators (terms can be considered as statistically independent). The correlation coefficient ρ was then calculated for the test structures and is shown in the last column of Table II. It is seen that ρ_{10+13} is slightly better than ρ_{10} but improvement is not really significant.

Conclusions

The main purpose was to investigate about the limits of accuracy to which one can go by embedding triplet invariants in quintet invariants. P_{13} provides a slightly better information then P_{10} but the additional contribution does not seem of sufficient quality for justifying the quite larger amount of computing time. However it’s too early to conclude that the limits have seen reached.

References

Altomare, A., Cascarano, G., Giacovazzo, C., Guagliardi, A., Burla, M. C., Polidori, G. & Camalli, M. (1994). *J. Appl. Cryst.* **27**, 435.

Burla, M. C., Camalli, M., Cascarano, G., Giacovazzo, C., Polidori, G., Spagna, R. & Viterbo, D. (1989). *J. Appl. Cryst.* **22**, 389-393.

Burla, M. C., Giacovazzo, C., Moliterni, A. G. G. & González-Platas, J. (1994). *Acta Cryst.* A**50**, 771-778.

Cascarano, G., Douggy-Smiri, L. & Nguyen-Huy Dung (1987). *Acta Cryst.* C**43**, 2050-2053.

Cascarano, G., Giacovazzo, C., Camalli, M., Spagna, R., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* A**40**, 278-283.

Cochran, W. (1955). *Acta Cryst.* **8**, 473-478.

Giacovazzo, C. (1977). *Acta Cryst.* A**33**, 933-944.

Giacovazzo, C. (1980). *Direct Methods in Crystallography, Academic Press London*.

Hauptman, H. & Karle, J. (1953). *The solution of the phase problem. I. The Centrosymmetric Crystal,* ACA Monograph Nº3, Polycrystal Book Service, New York.

Klug, A. (1958). *Acta Cryst.* **11**, 515-543.

Viterbo, D. & Woolfson, M. M. (1973). *Acta Cryst.* A**29**, 205-208.

Chapter III

The normalization procedure for Proteins

Introduction

In recents works (Giacovazzo, Siliqi & Ralph, 1994; Giacovazzo, Siliqi & Spagna, 1994; Giacovazzo, Siliqi & Zanotti, 1995) a probabilistic approach has been described for the *ab-initio* crystal structure solution of proteins. The method integrates direct methods and isomorphous techniques and requires diffraction data from the native protein and from one isomorphous derivative. It is based on the formula obtained by Giacovazzo, Cascarano & Zheng Chao de (1988) estimating three-phase invariants given six magnitudes. The results may be summarized as follows:

1. the multisolution technique is applied to random starting phases. A small number of trials is sufficient for obtaining the correct solution.
2. Proper figures of merit rank the trials: the correct solution is often found among the trials characterized by the largest values of the combined figure of merit CFOM.
3. About 40% of the reflections up to derivative resolution can be phased with good reliability. The process needs a relative short computing time.
4. The accuracy of the phasing process relies on the quality of the heavy-atom derivative.

Quite small phase errors can be obtained in case of good isomorphism. Severe lack of isomorphism degrades the accuracy of the triplet invariant estimates and therefore the quality of the assigned phases. The process proved sufficiently robust against experimental errors but it may still be improved in several ways. Our attention will be focused on the following ones:

a) the procedure is based on Δ (or Δ') values which are obtained by a statistical treatment of the experimental data. Each Δ may be considered as sum of a signal (i.e., the heavy atom scattering) and of a noise (arising from the disordered water distribution, lack of isomorphism, error in measurements, etc.). Since noise is unavoidable, the following question arises: is the procedure yielding Δ values optimally designed to face a large noise? If not, can new criteria be fixed to design a robust procedure accurately working in severe conditions?
b) The phasing procedure was applied to different structures for which derivative data are available, whose resolutions are 2Å for CARP and APP, and 3Å for E2 and M-FABP (see Table I and Table II for more details) and it was proved that success for direct methods can be obtained even at non-atomic resolution. However it is not unfrequent that only isomorphous data up to 4Å resolution are available. Can the phasing process successfully work at such a low resolution where the scaling Wilson procedure is rather inaccurate?
c) The various trials solutions are ranked by proper FOM's, which are extremely efficient when perfect isomorphism occurs. For real cases one can expect that the correct solution is among the trials with the highest values of CFOM, however various experimental errors and lack of isomorphism can heavily reduce the discriminating power of the various FOM's. The search of FOM's less sensitive to the various "errors" is a topic of enormous importance for the success of direct methods applied to macromolecules. An alternative way for contributing to the solution of the problem may consist in answering the following question: do criteria exist which are able to discard, among the various trials with the highest values of CFOM, the trial solutions devoid of structural meaning?

We will describe some techniques which will provide efficient solution to the problems described in a), b) and c).

The test structures

In order to check the phasing process, we use the test structures defined by the code names **APP**, **CARP**, **E2**, **M-FABP**, **BPO**, **FIX**, **NOX**, **TAQ**.

APP data were collected using a four-cicle diffractometer for the native and a $HgCl_2$ derivative (Blundell, Pitts, Tickle, Wood & Wu, 1981). The structure was solved by applying SIRAS (single isomorphous replacement including anomalous scattering) techniques to 2Å resolution data. Phases were extended to 1.4Å resolution by using a modified tangent formula. New data for the native protein up to

0.98Å resolution were collected by a four-circle diffractometer (Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell, 1983).

For **CARP** (carp muscle calcium-binding protein), isomorphous and anomalous scattering data were measured (Kretsinger & Nockolds, 1973) up to 2.0Å resolution using precession photography; three heavy-atom derivatives were used. In our calculations, we only make use of the (3-chloromercurio-2-methoxypropyl)urea (CMMPU) derivative.

Diffraction data of **E2** (catalytic domain of *Azotobacter vinelandii* dihydrolipoyl transacetylase) were collected on a fast television area derector (Mattevi, Obmolova, Schulze, Kalk, Westphal, De Kok & Hol, 1992). One mercury and two platinum derivatives were used for phasing: data include anomalous-dispersion effects (multiple isomorphous replacement including anomalous scattering, MIRAS). We only make use of the mercury derivative, which, as stated by Mattevi *et al.*, is of excellent quality.

The structure of **M-FABP** (recombinant human-muscle fatty-accid-binding protein) was originally solved using both multiple isomorphous replacement and molecular replacement procedures (Zanotti, Scapin, Spadon, Veerkamp & Scchettini, 1992). Data for native and two isomorphous derivatives were collected with Siemens X1000 area detector system and on a SDMS area detector system coupled with a rotating-anode generator. For our calculations, we used the HgAc_2 derivative.

Data for **BPO** (bromoperoxidase A2 from Streptomyces aureofaciens ATCC 10762 (Hecht, Sobek, Haag, Pfeifer & van Pee, 1994) have been collected at room temperature. Native data were measured to 2.05Å resolution, two derivative data sets were measured to 2.6Å resolution. The resulting MIR map was easily interpretable and allowed a complete chain tracing in the asymmetric unit.

The factor for inversion stimulation, **FIS**, was determined by multiple isomorphous replacement (Kostrewa, Granzin, Stock, Choe, Labahn & Saenger, 1992): we will only use native data (up to 2Å resolution) and the $\left[\mathrm{PtCl}_2\left(\mathrm{C}_2\mathrm{H}_4\right)\right]_2$ derivative data (resolution up to 3.3Å).

NOX is the code name for NADH oxidase from Thermus thermophilus (Hecht, Erdmann, Park, Sprinzl, Schmid & Schonburg, 1993; Hecht, Erdmann, Park, Sprinzl, Schmid, 1994). Native data were collected up at 12°C to 2.3Å resolution. Potencial derivative data sets were collected up to 2.9Å resolution, but we will use only $\left[\mathrm{PtCl}_2\left(\mathrm{NH}_2\right)\right]$ derivative.

TAQ is the code name for Adenine-N6-DNA-methytransferase (Labahn, Granzin, Schluckebier, Robinson, Jack, Schildkraut & Saenger, 1994). Native data were collected up to 2.4Å resolution. Three derivatives were used for crystal structure determination, $\mathrm{K}_2\mathrm{PtCl}_4$, $\left[\mathrm{PtCl}_2\left(\mathrm{C}_2\mathrm{H}_4\right)\right]_2$,

$(C_7H_5O_3)HgCl$. We will only use the $[PtCl_2(C_2H_4)]_2$ and $(C_7H_5O_3)HgCl$ derivative in the calculations.

In Table I we collect the main crystallochemical data for all the test structures, and in Table II we show some relevant parameters of the diffraction data we used.

Structure Code	Space Group	Molecular Formula	Z
APP (1)	C 2	$C_{190}N_{53}O_{58}Zn$	4
CARP (2)	C 2	$C_{513}N_{131}O_{121}Ca_2S$	4
E2 (3)	F 4 3 2	$C_{1170}N_{310}O_{366}S_7$	96
M-FABP (4)	$P\ 2_1\ 2_1\ 2_1$	$C_{667}N_{170}O_{261}S_3$	4
BPO (5)	$P\ 2_1\ 3$	$C_{2744}N_{712}O_{1073}$	12
FIS (6)	$P\ 2_1\ 2_1\ 2_1$	$C_{783}N_{224}O_{1312}S_{10}$	4
NOX (7)	$P\ 4_1\ 2_1\ 2$	$C_{1034}O_{704}N_{299}S_2P_{1/8}$	8
TAQ (8)	$P\ 2_1\ 2_1\ 2$	$C_{4390}N_{1174}O_{1240}S_8$	4

Table I: Code name, space group and crystallochemical data for test structures.
(1) Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983)
(2) Kretsinger & Nockolds (1973)
(3) Mattevi, Oblomova, Schulze,Kalk, Westphal, De Kok & Hol (1992)
(4) Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992)
(5) Hetch *et al*. (1994)
(6) Kostrewa *et al*. (1991)
(7) Hetch *et al* (1993); Hetch *et al* (1993)
(8) Labahn *et al*. (1994)

The normalization process at 4Å resolution

We applied the normalization procedure described by Giacovazzo, Siliqi & Spagna (1994) to TAQ by using Hg-derivative data (experimental data up to 4Å resolution). The procedure is a two-step method: first, the standard Wilson method is applied to native protein data truncated at derivative resolution to obtain $(K_{\mathrm{Dw}})_p$ and $(B_{\mathrm{Dw}})_p$, where K and B indicate scale and temperature factor respectively. Then, in the second step, estimates of the ratio (K_d / K_p) and $(B_d - B_p)$ are obtained by a differential Wilson plot (Blundell & Johnson, 1976) through the equation

$$\ln\left[\left(\Sigma_p + \Sigma_H\right) < F_p^2 > / \left(\Sigma_p < F_d^2 >\right)\right] = \\ \ln\left(K_p / K_d\right) + 2\left(B_d - B_p\right) \sin^2\theta / \lambda^2$$

Structure Code	Native		Derivative			
	RES(Å)	**NREFL**	**Heavy atom**	$[\sigma_2]_H/[\sigma_2]_p$	**RES(Å)**	**NREFL**
APP	0.99	17058	Hg	0.23	2.00	2086
CARP	1.71	5056	Hg	0.09	1.71	4687
E2	3.00	10388	Hg	0.08	3.00	9179
M-FABP	2.14	7595	Hg	0.06	2.15	7125
BPO	2.76	16348	Au	0.06	2.76	16348
			Pt	0.06	2.76	15291
FIS	2.00	12846	Pt	0.35	3.30	2983
NOX	3.00	4295	Pt	0.08	3.00	4295
TAQ	2.40	37268	Pt	0.09	3.20	15620
			Hg	0.04	4.00	8484

Table II: Relevant parameters for diffraction data of test structures.

The scaling $(K_{\mathrm{Dw}})_d$ and thermal $(B_{\mathrm{Dw}})_d$ parameters for the derivative are used to calculate the first estimates of the Δ' values, which are scaled by the factor

$$S = \left(< |E_d'|^2 + |E_p'|^2 - 2\,|E_p' E_d'|\, T > \right)^{-1/2}$$

to make the experimental distribution of $|\Delta'|$ closer to expected one. This procedure was used with a relative success in APP, CARP, E2 and M-FABP (Giacovazzo, Siliqi & Spagna, 1994) but the results for TAQ were in some way surprising: strongly negative values of the termal factors were obtained by the Wilson procedure both for the native $(B_{\mathrm{Dw}})_p$= -20.81 and for de derivative $(B_{\mathrm{Dw}})_d$= -8.55 data. The Wilson plot is highly non-linear and is shown in Fig.1.

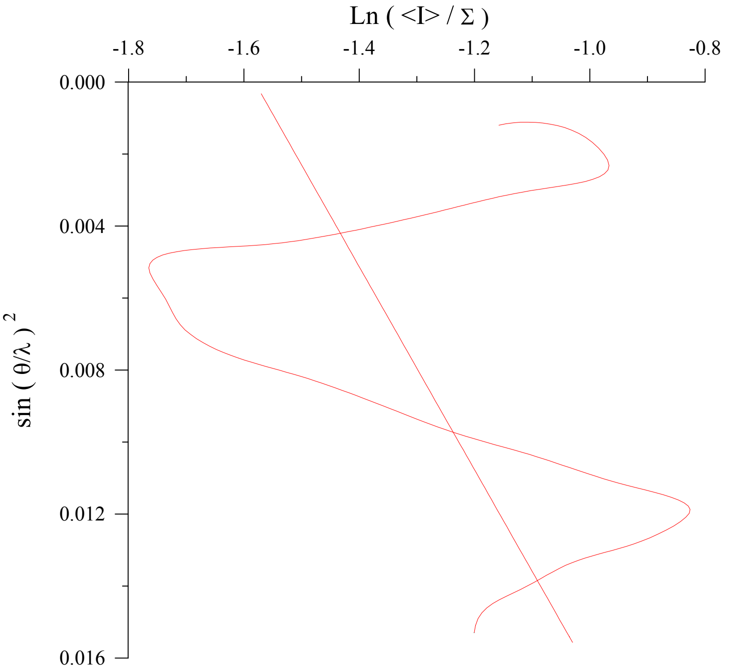


Fig. 1: TAQ - Wilson plot for native diffraction data up to derivative resolution (4Å).

Structure Code	**4Å Resolution Data**		**Derivative Resolution Data**	
	$(K_{\mathrm{Dw}})_p$ $(K_{\mathrm{Dw}})_d$	$(B_{\mathrm{Dw}})_p$ $(B_{\mathrm{Dw}})_d$	$(K_{\mathrm{Dw}})_p$ $(K_{\mathrm{Dw}})_d$	$(B_{\mathrm{Dw}})_p$ $(B_{\mathrm{Dw}})_d$
APP	0.27	-18.86	0.14	9.78
	0.34	-14.50	0.19	12.57
CARP	3.87	-22.88	2.12	5.69
	3.92	-18.60	2.22	7.69
E2	27783.82	-11.89	19178.18	9.14
	28088.40	-6.33	20429.74	11.41
M-FABP	433.78	-31.41	211.37	6.11
	65.48	-27.99	33.70	6.54
BPO (Au)	0.091	-37.56	0.037	6.71
	0.093	-34.30	0.039	8.11
BPO (Pt)	0.091	-37.56	0.037	7.30
	0.101	-37.95	0.039	7.70
FIS	0.57	-10.61	0.39	10.12
	4.30	29.71	3.57	39.64
NOX	0.036	-39.52	0.0128	10.25
	0.035	-35.92	0.0130	11.86
TAQ (Pt)	4.88	-18.09	2.94	10.92
	0.031	-11.26	0.019	16.32
TAQ (Hg)	5.23	-20.81	5.23	-20.81
	0.036	-8.55	0.036	-8.55

Table III: Scale and thermal factors for the test structures obtained by the procedure described by Giacovazo, Siliqi & Spagna (1994).

In order to check if such a result was casual or representative of a systematic behaviour of protein data at 4Å resolution, we cut at 4Å the data of all the other test structures. The results are shown in Table III: the thermal factors B are all negative for the protein data $(B_{\mathrm{Dw}})_p$<< 0, differences $(B_{\mathrm{Dw}})_d$ - $(B_{\mathrm{Dw}})_p$ are all positive. If the same procedure is applied to the other test data up to derivative resolution the results fit better with expectations (see Table III again). Indeed positive B_p values are now obtained and the differences $(B_{\mathrm{Dw}})_d$ - $(B_{\mathrm{Dw}})_p$ are again all positive (and highly correlated with the corresponding differences obtained at 4Å resolution).

The reason of the "anomalous" behaviour at 4Å resolution can be in mediately understood from Figs.2 and 3 where Wilson plots for native APP and FIS data are shown.

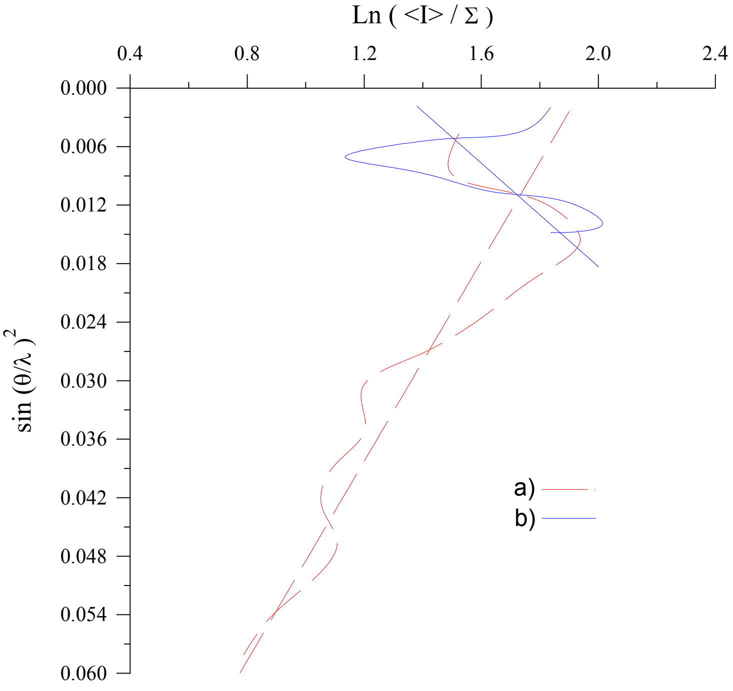


Fig. 2: APP - Wilson plot. a) Is the native driffraction data up to derivative resolution (2Å); b) Is the native diffraction data up to 4Å resolution

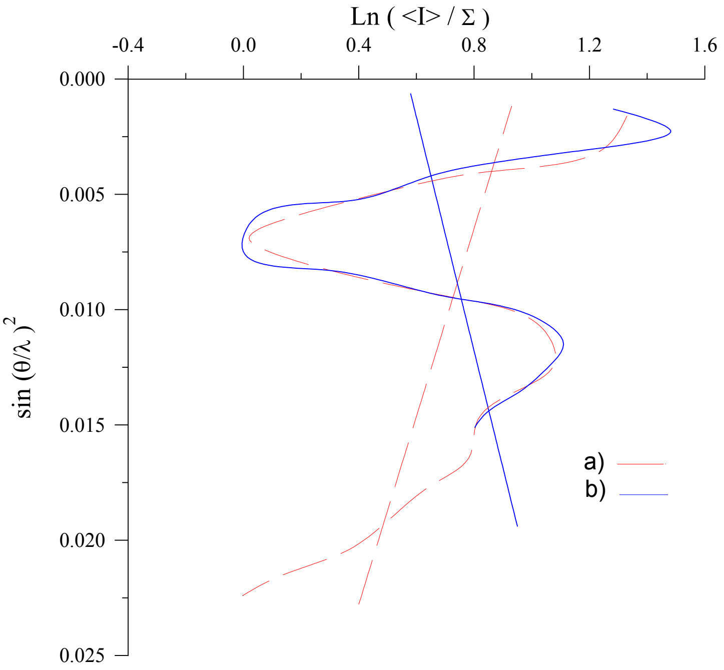


Fig. 3: FIS - Wilson plot. a) Is the native driffraction data up to derivative resolution (3.2Å); b) Is the native diffraction data up to 4Å resolution

In each figure Wilson plots for the data up to 4Å resolution and up to derivative resolution are shown together with the corresponding least squares straight lines. It is easy seen that the impressive errors in the estimated K and B values at 4Å resolution are consequence of Debye effects. Indeed the radial distribution of diffracted intensities of proteins has always a through at about 6Å and a peak at about 4.5Å (Richardson & Richardson, 1985). The problem is now to understand if errors in the normalizing step can hinder the success of the phasing process or damage its efficiency. That really occurs for small molecules (Subramanian & Hall, 1982; Hall & Subramanian, 1982a,b; Cascarano, Giacovazzo & Guagliardi, 1992). Does the same occur for macromolecules, where Δ (and not $|E|$) quantities are used? This is not so obvious since Δ parameters are more sensitive to the ratio K_d / K_p and to the difference $B_d - B_p$ rather than to their absolute values.

In order to answer the above question we used Δ' values of TAQ (4Å resolution data) to estimate, *via*

$$A = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_p R_{\mathbf{h}} R_{\mathbf{k}} R_{\mathbf{h-k}} + 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \varDelta'_{\mathbf{h}} \varDelta'_{\mathbf{k}} \varDelta'_{\mathbf{h-k}}, \qquad \text{(III.1)}$$

the triplet invariants among the 986 reflections with the largest values of $|\varDelta'|$. A is the concentration parameter of the distribution

$$\mathrm{P}(\varPhi \mid R_{\mathbf{h}}, R_{\mathbf{k}}, R_{\mathbf{h-k}}, S_{\mathbf{h}}, S_{\mathbf{k}}, S_{\mathbf{h-k}}) \cong \left[2\pi\ I_o(A)\right]^{-1} \exp(A \cos\varPhi)$$

derived by Giacovazzo, Cascarano & Zheng (1988). Short statistics are shown in Table IV-(a). Triplets are divided into two subsets, positive and negative estimated triplets: Nr is the number of triplets having $|A|$>ARG, % is the percentage of triplets whose cosine sign is correctly estimated, <$|\varPhi|$°> is the average of the absolute values of the triplet phase $\varPhi$.

It is inmediately seen that the number of triplets estimated negative is abnormally higher than the number of triplets estimated positive. This has no physical meaning and is mainly due to errors in the normalizing procedure. As a consequence the percentage of correctly estimated negative triplets is smaller than 50%, and this seriously endangers the success of the phasing process. A similar result is obtained for FIS at 4Å resolution (see Table IV-(b)) by using the 643 reflections with the largest value of $|\varDelta'|$.

Different statistics are obtained for NOX and BPO (see Tables IV-(c),(d),(e)).

For NOX a too large percentage of the triplets found among the 564 reflections with the largest values of $|\varDelta'|$ have $|A|$ value between 0.0 and 0.2, but there is no systematic error in the estimation of the sign of the triplets. For BPO (Pt derivative), only 35 triplets have $|A|\geq 0.2$, and for BPO (Au derivative) only 3455 have $|A|\geq 0.2$. Again no systematic error is found in the estimation of the sign of the triplets.

The above results indicate that the normalizing procedure described before is unable to carefully control at 4Å resolution the various parameters playing a role in the scaling process. That constitutes a bad premise for the success of the subsequent phase determination procedure. This was confirmed when we applied the phasing process described by Giacovazzo, Siliqi & Spagna, (1994) and Giacovazzo, Siliqi & Zanotti, (1995) to data up to 4Å resolution for all the test structures (see Apendix I).

TAQ (Hg derivative)						
	Positive estimated triplets			Negative estimated triplets		
|ARG|	**Nr**	**%**	<|Φ|°>	**Nr**	**%**	<|Φ|°>
1.6	6255	52	87	32302	48	88
2.0	5389	53	86	28693	48	88
2.6	3470	53	86	20132	48	88
3.2	1998	54	85	12620	48	88
3.8	842	57	82	5997	48	88
4.4	282	60	78	2680	47	88
5.5	70	60	82	561	47	88

Table IV-(a): Statistical calculations for triplets invariants estimated via (III.1) at 4Å resolution.

FIS						
	Positive estimated triplets			Negative estimated triplets		
|ARG|	**Nr**	**%**	<|Φ|°>	**Nr**	**%**	<|Φ|°>
0.4	7660	55	84	42340	49	89
0.8	5947	56	83	29277	49	89
1.2	3232	57	81	14135	48	88
1.6	1523	58	80	6144	48	88
2.0	724	56	82	2537	49	90
2.6	223	53	85	669	44	85
3.2	62	48	88	148	42	81
3.8	7	43	96	13	31	70

Table IV-(b): Statistical calculations for triplets invariants estimated via equation (III.1) at 4Å resolution

NOX						
	Positive estimated triplets			**Negative estimated triplets**		
|ARG|	**Nr**	**%**	<|Φ|°>	**Nr**	**%**	<|Φ|°>
0.0	31477	54	85	18523	54	95
0.2	100	76	52	23	61	101
0.4	1	0	178			

Table IV-(c): Statistical calculations for triplets invariants estimated via (III.1) at 4Å resolution

BPO (Au derivative)						
	Positive estimated triplets			**Negative estimated triplets**		
|ARG|	**Nr**	**%**	<|Φ|°>	**Nr**	**%**	<|Φ|°>
0.0	24151	68	69	25849	67	110
0.2	2348	78	58	1107	80	124
0.4	120	86	44	75	84	139

Table IV (d): Statistical calculations for triplets invariants estimated via equation (III.1) at 4Å resolution

BPO (Pt derivative)						
	Positive estimated triplets			**Negative estimated triplets**		
|ARG|	**Nr**	**%**	<|Φ|°>	**Nr**	**%**	<|Φ|°>
0.0	33101	64	73	16899	70	113
0.2	28	93	42	7	86	137

Table IV (e): Statistical calculations for triplets invariants estimated via equation (III.1) at 4Å resolution

Results are shown in Table V. It is not a surprise that the procedure does not succeed in the majority of the cases (i.e., for APP, M-FABP, BPO, FIS, NOX, and TAQ(Hg)). Only E2 and CARP are satisfactorily phased while for TAQ(Pt) a solution is found but with an appreciable mean phase error.

Structure Code	Order of Solution	NPHAS	Error (w-Error)
APP	-	-	-
CARP	2	348	42 (37)
E2	2	1696	30 (30)
M-FABP	-	-	-
BPO (Au)	-	-	-
BPO (Pt)	-	-	-
FIS	-	-	-
NOX	-	-	-
TAQ (Pt)	1	3194	66 (57)
TAQ (Hg)	-	-	-

Table V: Application of the phasing procedure described by Giacovazzo, Siliqi & Spagna (1994); Giacovazzo, Siliqi & Zanotti (1995) at 4Å resolution data. Order of solution is the order of the trial solution as ranked by CFOM. NPHAS is the number of the phased reflexions, Error is the average phase error calculated with respect to published phases values.

The question is now if a more accurate normalizing procedure can be found which is able to overcome the difficulties met with 4Å resolution data and possibly to improve the accuracy of the phasing process also with data at higher resolution. Such a technique is described in the next sections.

The normalization procedure by histogram matching

For R and S larger than or close to unity the factor T is so close to unity that Δ'may be replaced by Δ. The advantage of the quantity Δ is that its distribution may be readily calculated (see Apendix II). This has been done by Giacovazzo, Siliqi & Zanotti (1995), where, in order to guess about the number of phases to involve in the phasing process, the probability distribution function $P(\Delta)$ has been obtained as a function of the parameter $\sigma = [\sigma_2]_H / [\sigma_2]_p$. We will see now that $P(\Delta)$ can play a basic role also in the normalizing process.

Let Δ_T be a positive threshold for Δ, $n^+_{\Delta_T}$ be the number of positive Δ's for which $\Delta > \Delta_T$, $n^-_{\Delta_T}$ be the number of negative Δ for which $|\Delta| > \Delta_T$. Since $P(\Delta)$ is not an even function, the ratio

$$\mathrm{RPM} = n^+_{\Delta_T} / n^-_{\Delta_T}$$

is expected to be larger than unity for any value of σ and for any Δ_T. In Fig.4 we show RPM curves for different values of σ.

RPM increases with σ, and, for a given σ, increases with Δ_T. Its value is strictly correlated with the ratio K_d / K_p: errors in the estimate of this ratio will produce anomalous values of RPM. For example, if F_d values are scaled so as they are larger than their true values the number of positive Δ's will exceed the expected value. In the converse case the number of negative Δ's will be larger than the expected value. In general the experimental $P(\Delta)$ curve is modelled by different sources of errors: besides the scaling error, also wrong estimates of the difference $B_d - B_p$ (as a consequence of the scaling error), errors in measurements, lack of isomorphism, etc. will generate anomalies in $P(\Delta)$. It is therefore instructive to compare for all the test structures (see Figs.5 - 8) the theoretical Δ curves with those obtained from measurements at 4Å resolution in accordance with the normalization procedure described by Giacovazzo, Siliqi & Spagna (1994). APP and FIS curves are shown in Fig.5,together with the teoretical curve at $\sigma \cong 0.35$. CARP, E2 and NOX curves are shown in Fig.6 together with the theoretical curve at $\sigma \cong 0.08$. TAQ curves are shown in Fig.7, together with expected curve at $\sigma \cong 0.06$. M-FABP and BPO curves are shown in Fig.8 together with expected curve at $\sigma \cong 0.06$.

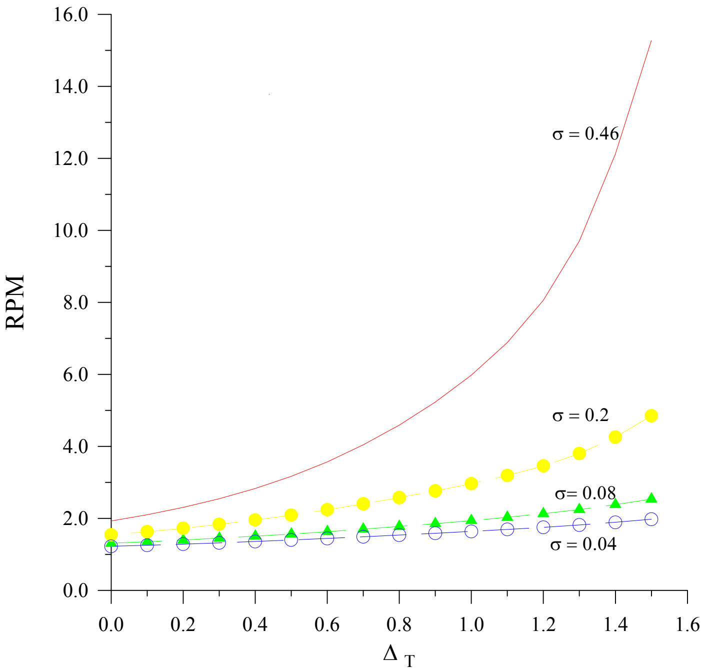


Fig.4: RPM curves for some representative values of σ against the threshold Δ_T

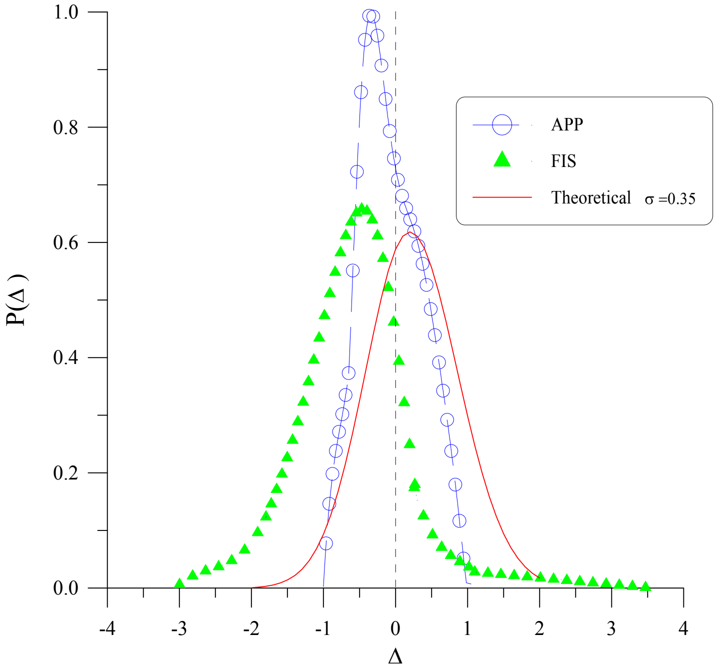


Fig. 5: P(Δ) distribution curve theoretically expected at σ=0.35 and corresponding experimental curves of APP and FIS, obtained by the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at 4Å resolution.

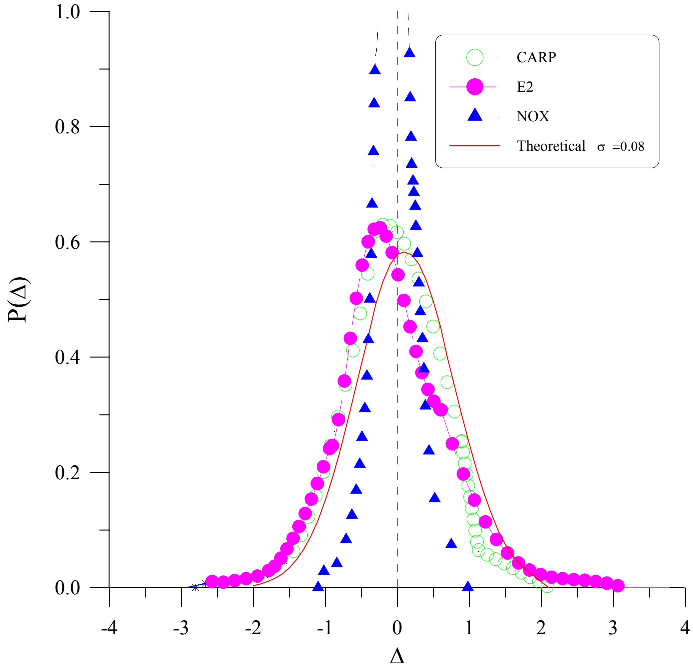


Fig. 6: P(Δ) distribution curve theoretically expected at σ=0.08 and corresponding experimental curves of CARP, E2 and NOX, obtained by the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at 4Å resolution.

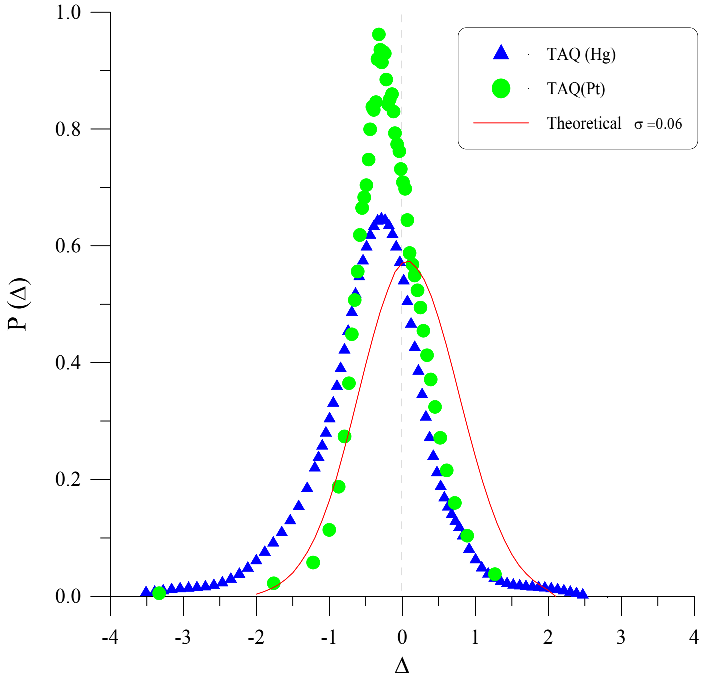


Fig. 7: P(Δ) distribution curve theoretically expected at σ=0.06 and corresponding experimental curves of TAQ (Pt derivative) and TAQ (Hg derivative), obtained by the

normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at 4Å resolution.

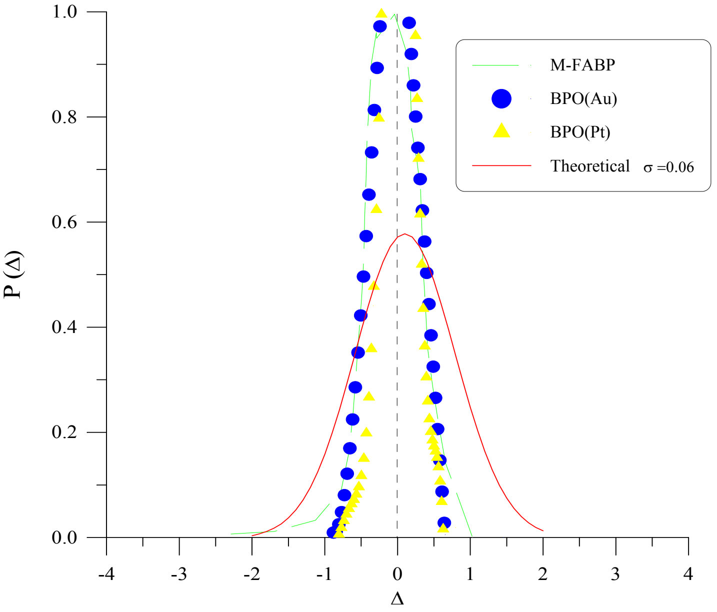


Fig. 8: P(Δ) distribution curve theoretically expected at σ=0.06 and corresponding experimental curves of BPO (Au and Pt derivative) and M-FABP, obtained by the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at 4Å resolution.

We note that:

a) APP, NOX, M-FABP, TAQ (Pt) and BPO curves are too sharp. As a consequence the $|\varDelta|$'s are underestimated and reliable triplets are weakly discriminated from unreliable ones. This explains the anomalous triplet statistics of NOX and BPO shown in Table IV-(c),(d),(e).

b) FIS and TAQ (Hg) curves are markedly shifted towards left. As a consequence, the ratio RPM will generally be smaller than its expected value and the percentage of negative triplets will be abnormally high. This explains the bad triplet statistics shown in Tables IV-(a) and (b) for TAQ (Hg) and FIS.

c) E2 and CARP curves are sufficiently close to the theoretical ones. It is therefore not surprising that among the test structures, only E2 and CARP (see Table V) were satisfactorily phased by the procedure described by Giacovazzo, Siliqi & Spagna (1994) and Giacovazzo, Siliqi & Zanotti (1995).

The above observations suggest that $P(\varDelta)$ may be conveniently used as a target distribution with which the experimental curves should comply. We do this according to the following procedure:

1. the B_p value is found by standard Wilson method using all the reflections up to native resolution.
2. $\varDelta\ B = B_d - B_p$ and $R_K = K_d / K_p$ are found by differential Wilson plot. Then B_d and K_d are set to $B_d\ = B_p + \varDelta\ B$ and $K_d\ = K_p \cdot R_K$.
3. The scale factor K_d is suitably modified in order to satisfy the expected RPM at the chosen σ value for $\varDelta_T$=0.
4. Histogram matching techniques are (Zhang & Main, 1990) applied to transform the experimental curve into the P($\varDelta$) distribution expected at the chosen σ value. The next equation

$$A = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_p R_\mathbf{h} R_\mathbf{k} R_\mathbf{h-k} + 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \varDelta_\mathbf{h} \varDelta_\mathbf{k} \varDelta_\mathbf{h-k} \qquad \text{(III.2)}$$

is then applied to the $\varDelta$ values so obtained for estimating triplet invariants.

It is instructive to compare triplet statistics obtained by the new procedure (see Table VI) with statistics shown in Table IV. It is immediately seen that systematic errors in the triplet sign estimation for FIS and TAQ (Hg) are avoided by the new normalizing procedure. Furthermore triplet statistics for NOX and BPO is largely improved: the range of G values in Tables VI-(c),(d),(e) is quite reasonable, and good triplets are more efficiently discriminated from unreliable ones.

The application of the phasing procedure to the new $\varDelta$'s at 4Å resolution data gives the results shown in Table VII.

We note:

a) CARP and E2, for which a solution was found in Table V, are again solved.

b) A satisfactory solution is now found for APP, M-FABP, BPO and NOX. It is worthwhile stressing the spectacular result obtained for BPO. A noisy solution is also found for FIS and TAQ (Pt) even if with the penalty of a large mean phase error.

TAQ (Hg derivative)						
	Positive estimated triplets			**Negative estimated triplets**		
|ARG|	**Nr**	**%**	**<|Φ|°>**	**Nr**	**%**	**<|Φ|°>**
0.4	22413	53	86	14520	51	91
0.8	21982	53	86	14103	51	91
1.2	12875	53	85	7088	52	92
1.6	57608	54	84	2731	52	92
2.0	2260	55	84	919	52	92
2.6	484	54	85	183	50	90
3.2	82	54	87	26	54	99
3.8	5	40	105	3	0	43

Table VI-(a): Statistical calculations for triplets invariants estimated via (III.2) after histogram normalizing procedure at 4Å resolution.

FIS						
	Positive estimated triplets			**Negative estimated triplets**		
|ARG|	**Nr**	**%**	**<|Φ|°>**	**Nr**	**%**	**<|Φ|°>**
0.2	31019	55	84	18981	54	95
0.4	16948	56	84	7094	56	96
0.8	3232	59	80	923	57	97
1.2	657	59	80	141	55	96
1.6	148	56	82	22	64	106
2.0	27	59	81	1	100	138

Table VI-(b): Statistical calculations for triplets invariants estimated via (III.2) after histogram normalizing procedure at 4Å resolution.

NOX						
	Positive estimated triplets			Negative estimated triplets		
|ARG|	Nr	%	<|Φ|°>	Nr	%	<|Φ|°>
0.2	25465	56	83	24535	55	96
0.4	24365	58	82	20432	56	97
0.8	6519	62	76	5211	61	103
1.2	1901	66	70	1457	67	110
1.6	593	70	67	473	70	114
2.0	190	76	56	146	80	124
2.6	33	79	54	21	95	143
3.2	4	100	33	3	100	152

Table VI-(c): Statistical calculations for triplets invariants estimated via (III.2) after histogram normalizing procedure at 4Å resolution.

BPO (Au derivative)						
	Positive estimated triplets			Negative estimated triplets		
|ARG|	Nr	%	<|Φ|°>	Nr	%	<|Φ|°>
0.4	25072	69	68	24928	68	111
0.8	4936	77	56	3891	77	121
1.2	788	84	60	84	89	114
1.6	139	90	39	96	93	137
2.0	18	94	33	14	100	144

Table VI-(d): Statistical calculations for triplets invariants estimated via (III.2) after histogram normalizing procedure at 4Å resolution.

BPO (Pt derivative)						
	Positive estimated triplets			Negative estimated triplets		
|ARG|	Nr	%	<|Φ|°>	Nr	%	<|Φ|°>
0.4	24937	71	65	24028	71	114
0.8	3809	81	53	3161	82	127
1.2	542	87	44	437	91	139
1.6	92	92	34	56	96	149
2.0	12	100	28	3	100	162

Table VI-(e): Statistical calculations for triplets invariants estimated via (III.2) after histogram normalizing procedure at 4Å resolution.

Structure Code	Order of Solution	NPHAS	Error (w-Error)
APP	2	143	38 (36)
CARP	3	391	44 (40)
E2	1	1750	35 (32)
M-FABP	1	580	58 (54)
BPO (Au)	1	2583	30 (23)
BPO (Pt)	1	2442	24 (19)
FIS	43	835	68 (60)
NOX	1	740	58 (42)
TAQ (Pt)	5	3867	71 (70)
TAQ (Hg)	-	-	-

Table VII: The phasing procedure is applied by using the histogram matching normalizing procedure described in the text, at 4Å resolution data. Order of solution is the order of solution as ranked by CFOM. NPHAS is the number of the phased reflexions, Error is the average phase error calculated with respect to published phase values, w-Error is the weighted error.

It is useful to check if the normalizing procedure described by Giacovazzo, Siliqi & Spagna, (1994) might be successfully applied at derivative resolution to all our test proteins, and also to cases like FIS, NOX and TAQ (Pt derivative) for which only low quality derivatives are available. Whe show in Figs. 9 - 12 the experimental curves $P(\Delta)$ at derivative resolution together with the theoretical curve calculated for the representative σ value.

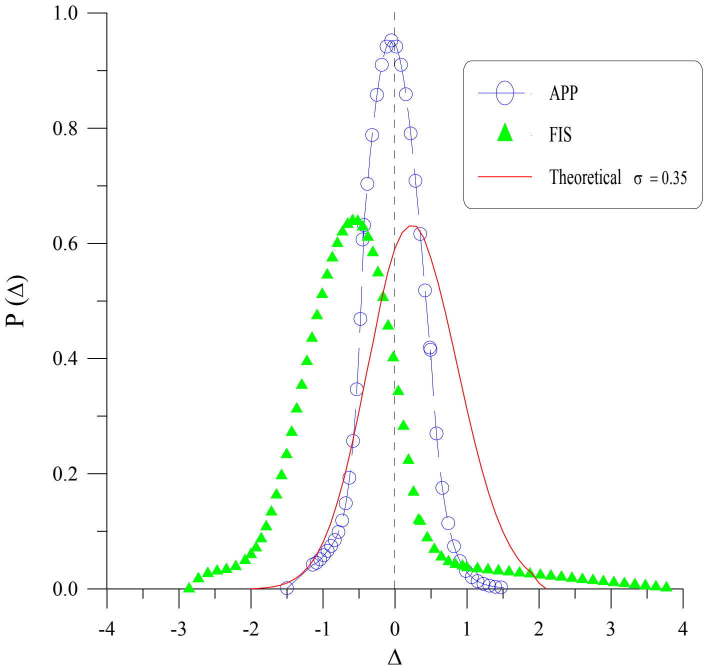


Fig. 9: P(Δ) distribution curve theoretically expected at σ=0.35 and corresponding experimental curves of APP and FIS, obtained by the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at derivative resolution data.

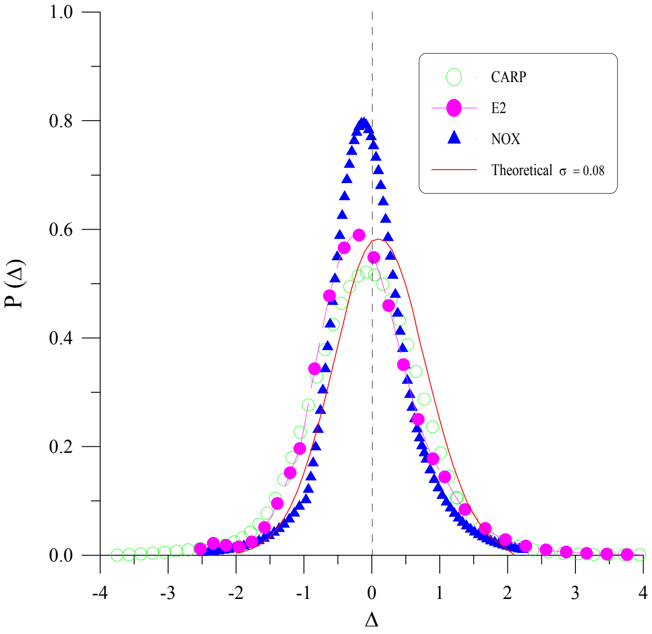


Fig. 10: P(Δ) distribution curve theoretically expected at σ=0.08 and corresponding experimental curves of CARP, E2 and NOX, obtained by the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at derivative resolution data.

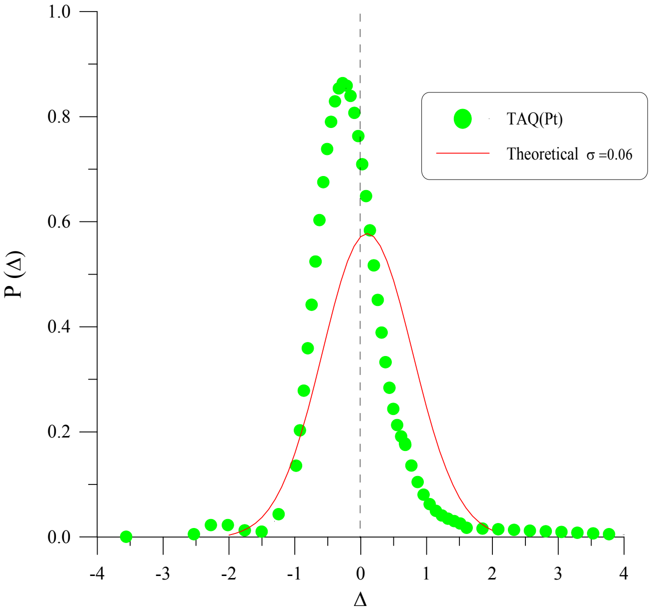


Fig. 11: P(Δ) distribution curve theoretically expected at σ=0.06 and corresponding experimental curves of TAQ (Pt derivative), obtained by the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at derivative resolution data.

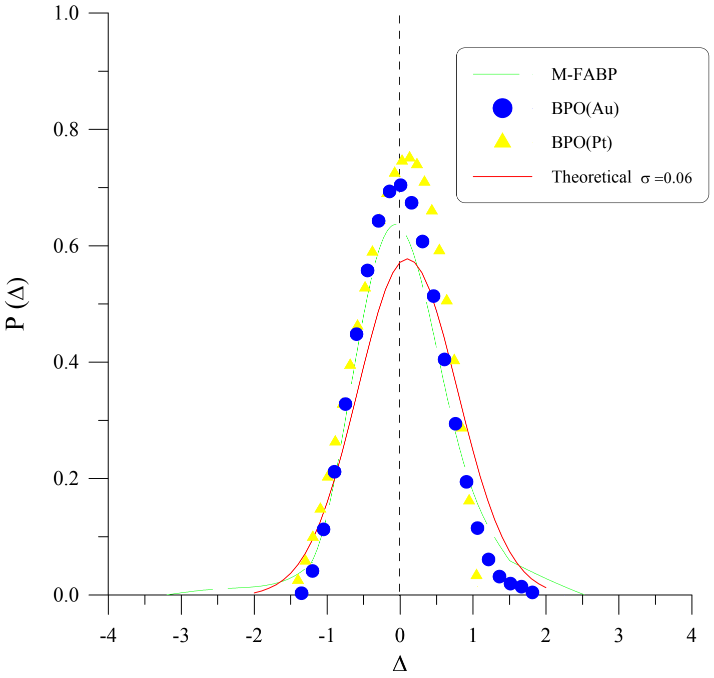


Fig. 12: P(Δ) distribution curve theoretically expected at σ=0.06 and corresponding experimental curves of BPO (Au and Pt derivative) and M-FABP, obtained by the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) at derivative resolution data.

Comparison of Figs. 5 - 8 with Figs. 9 - 12 shows that:

a) the APP curve in Fig. 9 is shifted towards right, and the fit with the theoretical curve improves. On the contrary the FIS curve does not remarkably change with resolution.

b) While E2 and CARP do not change with resolution (they remain sufficiently good) the fit of NOX and BPO is remarkably better at derivative resolution.

c) The M-FABP experimental curve strongly improves at high resolution.

One can conclude that the normalizing procedure proposed by Giacovazzo, Siliqi & Spagna, (1994) improves as resolution increases. It is not a surprise then that the phasing process described in papers Giacovazzo, Siliqi & Spagna, (1994) and Giacovazzo, Siliqi & Zanotti, (1995) works well at 3Å or higher resolution, even if for NOX with the penalty of a large mean phase error (see Table VIII).

If the histogram matching normalizing procedure is used the phasing process produces the result shown in Table VIII. For NOX it should be noticed that the order of solution is 7. However the three trials with order 1,2,3, are also solutions, even if with a larger phase error (73°, weighted error 67°). By comparing the effects of the old normalizing procedure with those produced by the new one we conclude that the new normalizing procedure is preferable.

Structure Code	Procedure	Order of Solution	NPHAS	Error (weighted)
APP	old	3	810	46 (43)
	new	2	988	45 (41)
CARP	old	2	2111	50 (46)
	new	2	2443	51 (48)
E2	old	3	3218	40 (37)
	new	2	3662	41 (36)
M-FABP	old	1	1231	50 (47)
	new	1	3330	54 (51)
BPO (Au)	old	1	8307	28 (19)
	new	1	8343	28 (20)
BPO (Pt)	old	-	-	-
	new	1	7908	34 (25)
FIS	old	-	-	-
	new	92	1242	67 (69)
NOX	old	4	1827	67 (60)
	new	7	1842	68 (61)
TAQ (Pt)	old	-	-	-
	new	-	-	-

Table VIII: The phasing procedure is applied at derivative resolution data by: a) using the normalizing procedure described by Giacovazzo, Siliqi & Spagna (1994) and Giacovazzo, Siliqi & Zanotti (1995) (old in the Table); b) the histogram matching normalizing procedure described in the text (new in the Table). Order of solution is the order of solution as ranked by CFOM. NPHAS is the number of the phased reflexions, ERR is the average phase error calculated with respect to published phase values. 100 trials have been calculated.

Discarding false solutions

When we applied the normalizing procedure give by Giacovazzo, Siliqi & Spagna, (1994) we calculated various FOM's for differents trials and then we combined them in order to pick up the correct solution. Recognizing the correct solution among different trials is not a simple task for protein structures (Woolfsoon & Yao, 1990; Giacovazzo, Guagliardi, Ravelli & Siliqi, 1994). Figures of merit used in our procedure for picking the correct solution from the trial solutions are based on the theory described in two papers (Cascarano, Giacovazzo & Viterbo, 1987; Cascarano, Giacovazzo & Guagliardi, 1992 b) with some modifications in order to take advantage of the information contained in the derivative data. We will briefly introduce them:

The first FOM is MABS. It is defined as:

$$\mathrm{MABS} = \sum_{\mathbf{h}} \alpha_{\mathbf{h}} \Big/ < \sum_{\mathbf{h}} \alpha_{\mathbf{h}} >$$

where

$$\alpha_{\mathbf{h}} = \left\{ \left[\sum_j A_j \sin(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h}-\mathbf{k}_j}) \right]^2 + \left[\sum_j A_j \cos(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h}-\mathbf{k}_j}) \right]^2 \right\}^{1/2}$$

and

$$A_j = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_p R_{\mathbf{h}} R_{\mathbf{k}_j} R_{\mathbf{h}-\mathbf{k}_j} + 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \mathit{\Delta}_{\mathbf{h}} \mathit{\Delta}_{\mathbf{k}_j} \mathit{\Delta}_{\mathbf{h}-\mathbf{k}_j}$$

MABS gives a measure of the consistency of the triplet estimates but it is not used as an active FOM for picking (in combination with others) the correct solution.

The second FOM is ALFCOMB where

$$\mathrm{ALFCOMB} = \sum_{\mathbf{h}} \left(\alpha_{\mathbf{h}} - < \alpha_{\mathbf{h}} > \right) \Big/ \sum_{\mathbf{h}} \sigma_{\alpha_{\mathbf{h}}}$$

and

$$< \alpha_{\mathbf{h}} >= \sum_j A_j D_1(A_j)$$

$$\sigma^2_{\alpha_{\mathbf{h}}} = \frac{1}{2} \sum_j A_j^2 \left[1 + D_2(A_j) - 2 D_1^2(A_j) \right]$$

This expression for the variance holds in the absence of errors in measurements and in their mathematical treatment as well as in the presence of perfect isomorphism between native and derivative structures. If this is not the case, as for real data, the variance cannot be perfectly calculated and is probably underestimated by $\sigma_{\alpha_{\mathbf{h}}}$. Accordingly, we used $2\sigma_{\alpha_{\mathbf{h}}}$ instead of $\sigma_{\alpha_{\mathbf{h}}}$ in ALFCOMB.

The third FOM is PSICOMB. It depends on the ratios $\alpha'_{\mathbf{h}} / \sigma_{\alpha'_{\mathbf{h}}}$, where

$$\alpha'_{\mathbf{h}} = \left\{ \left[\sum_j A'_j \sin(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h}-\mathbf{k}_j}) \right]^2 + \left[\sum_j A'_j \cos(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h}-\mathbf{k}_j}) \right]^2 \right\}^{1/2}$$

$$A'_j = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \Delta_{\mathbf{k}_j} \Delta_{\mathbf{h}-\mathbf{k}_j}$$

$$\sigma_{\alpha'_{\mathbf{h}}} = \left(\sum_j A'^2_j \right)^{1/2}$$

This expression relies on the expectation that the distribution of the psi-zero triplets should be as random as possible. The weak reflections that constitute psi-zero triplets with the NLAR reflections are characterized by small values of both R and $|\Delta'|$.

The fourth FOM is CPHASE. It is based on negative and positive estimated triplets phases Φ_j. Therefore we calculate the ratio

$$\sum_j A_j \cos\Phi_j / \sum_j A_j < \cos\Phi_j >$$

Finally, a combined figure of merit, CFOM, integrates the indications arising from ALFCOMB, PSICOMB and CPHASE.

Each FOM must lie between zero and one and is expected to be one for the correct solution. In the practice cases, the FOM's often are not maximal for the correct structure: This suggests to the reader some inefficiency, but they are sufficiently good for most practical purposes. When applied to our test structures the FOM's produce the following results.

Only in the case for M-FABP the correct solution correspondes with highest value for CFOM, but that seemed to be the exception, not the role. Indeed for CARP and E2 the solutions with highest value of CFOM were devoid of structural meaning, while for APP the correct solution had CFOM quite similar to that of a false solution (Giacovazzo, Siliqi & Spagna,1994). More efficient FOM's seem necessary for discriminating the correct solution from the false. In their absence it would be useful to be able to fix some criteria which could help to discard the false solutions with high values of CFOM. The scenario may be the following. Suppose that the phasing procedure has produced various trial solutions at the end of the phase extension process described by Giacovazzo, Siliqi & Zanotti, (1995). They are ranked in order of CFOM. Then:

1. difference Fourier syntheses with coefficients

$$(F_d - F_p)\exp(\mathrm{i}\phi_p)$$

   are calculated for the solutions with the highest values of CFOM. The maxima in the map should provide heavy-atom positions.
2. Such parameters are defined according to the phase refinement process (Dickerson, Kendrew & Strandberg, 1961).

Structure Code	**Set**	**CFOM**	**Heavy atom positions**			**Height of the peaks**
			x	**y**	**z**	
APP	1	0.68	0.50	0.29	0.50	369
	2	**0.57**	**0.75**	**0.45**	**0.23**	**250**
	3	0.36	0.50	0.31	0.33	243
CARP	1	0.77	0.00	0.00	0.00	112
	2	**0.57**	**0.76**	**0.17**	**0.09**	**192**
	3	0.39	0.47	0.31	0.38	133
E2	1	1.00	0.00	0.00	0.50	185
	2	**0.96**	**0.21**	**0.07**	**0.20**	**397**
	3	0.55	0.09	0.00	0.09	119
M-FABP	**1**	**0.40**	**0.89**	**0.06**	**0.74**	**648**
	2	0.35	0.09	0.15	0.59	670
BPO (Au)	**1**	**0.89**	**0.41**	**0.03**	**0.78**	**446**
			0.78	**0.11**	**0.81**	**320**
	2	**0.88**	**0.59**	**0.03**	**0.28**	**636**
			0.21	**0.11**	**0.31**	**541**
	3	0.63	0.15	0.15	0.15	410
			0.94	0.06	0.55	293
BPO (Pt)	**1**	**0.85**	**0.41**	**0.03**	**0.78**	**446**
			0.78	**0.11**	**0.81**	**320**
	2	0.74	0.04	0.04	0.04	1304
			0.11	0.11	0.11	246
NOX	**1**	**0.77**	**0.24**	**0.11**	**0.47**	**526**
	2	**0.61**	**0.76**	**0.11**	**0.27**	**537**
	7	**0.56**	**0.74**	**0.11**	**0.77**	**775**

Table IX: Heavy atom positions from Fourier synthesis with coefficients $(F_d - F_p)\exp(i\phi_p)$ for the highly ranked trial solutions (derivative resolution data).

3. If the refined positional parameters concide with an allowed origin of the protein space group then the trial solution is discarded from the set of reliable ones.

Steps 1, 2 and 3 are executed in sequence without user intervention.

Why such a process should work? Readers customed with direct phasing of small molecules know that in symmorphic space groups the so called “uranium solution” occurs quite frequently. It is marked by a high consistency of triplet phases, which are all close to zero. An observed Fourier synthesis would produce a huge maximum at an *allowed origin*. This type of false solution may be recognized and therefore discarded by special FOM’s like the psi-zero and negative quartet criteria. Since the psi-zero FOM described in Figures of merit is not highly discriminating for macromolecules and the negative quartet criterion is not among the used FOM’s, the calculation of the difference Fourier synthesis is an efficient substitute of the specific FOM’s. It is worthwhile emphasizing that a difference Fourier synthesis for proteins should not provide huge maxima at the allowed origins as for small molecules: since our phasing procedure uses a nearly equivalent number of positive and negative triplets, peak intensities in the maps corresponding to the “uranium solutions” are similar to peak intensities corresponding to true heavy-atom positions.

In Table IX we show, for each test structure and for the trial solutions highly ranked by CFOM, the heavy atom positions as obtained after some cycles of Fourier-least squares calculations. Data corresponding to the correct solution are in bold character. If use is made of the information in Table IX the correct solution is unambiguoisly recognized.

Conclusions

A more robust normalizing procedure has been designed which makes explicit use of the distribution $P(\Delta)$. Histogram matching procedures are used to obtain an optimal fit of the observed Δ distribution with the expected one. The new Δ‘s are statistically more meaningful and are able in most cases to overcome the disturbing effects provoked on the Wilson method for data up to 4Å resolution by the presence of strong Debye effects.

A method is also suggested for discarding some false solutions provided by our multisolution technique. Since FOM’s cannot safely work for isomorphous data where the signal is often comparable with the noise, an a-posteriori check on the heavy atom positions allows to discard those trials which correspond to what are called “uranium solution” in the small molecule direct methods applications.

Appendix I

The phasing procedure

We can described the procedure in differents steps in order to better understanding the process.

Step 1. Selection of the reflections to phase

The reflections to phased should be characterized by:

(a) high values of $|\mathit{\Delta}|$, in order to guarantee a reliable phase assignment.

(b) Non-vanishing values of R, in order to provide, once phased, useful information for electron density maps. Accordingly, the $NREFL$ reflections (those for which both $|F_p|$ and $|F_d|$ are available from measurements and symmetry independent) are partitioned into two subsets:

1.- The subset including the reflections with the smallest R values. Their number is chosen to be the minimum between 1000 and 25% of $NREFL$. Some of these reflections, *i.e.* those with $|\mathit{\Delta}|$≤0.2, will be used for constructing $PSI0$ triplets (it is made with two larger value and one small of R). Let $NPSI$ be the number of reflections with small values of $|R|$ and $|\mathit{\Delta}|$ that are actually involved in $PSI0$ triplets.

2.- The subset $\{\gamma_1\}$ including $NREFL$ - $NPSI$ reflections. According to the preceding section, we should try to phased about 52% of the $NREFL$ reflections, *i.e.* those characterized by the largest $|\mathit{\Delta}|$ values (accessible phases).

Step 2. The first batch

By default, 60% of the reflections in $\{\gamma_1\}$ (those with the largest R values) are selected. The cumulative distributions of the $|\mathit{\Delta}|$'s relative to such reflections is calculated, giving the number n of reflections with $|\mathit{\Delta}|$ larger than a fiven value. The threshold $TR\,\mathit{\Delta}_1$ is chosen as the value of $|\mathit{\Delta}|$ corresponding to n ≈800. The *statistical solvavility criterion* is applied: if it is satisfied then $NLAR_1 = n$ is the number of reflections which will be phased first, otherwise $NLAR_1$ is increased until the criterion is satisfied. The $NLAR_1$ reflections are said to constitute the subset $BATCH_1$.

Step 3. The next batch

Let $NLAR_2$ be the number of reflections (among the $NREFL$ - $NPSI$ - $NLAR_1$ reflections) with $|\Delta|>TR\,\Delta_2 \equiv TR\,\Delta_1$. They will constitute the subset $BATCH_2$.

The remaining $NREFL$ - $NPSI$ - $NLAR_1$ - $NLAR_2$ reflections are divided into subsets (*i.e.* each $BATCH_i$ for *i* > 2 contains about 400 reflections), the *i*th subset being associated with a given threshold $TR\,\Delta_i$ for $|\Delta|$. Since $TR\,\Delta_{i+1} \leq TR\,\Delta_i$, the reflections in $BATCH_i$ will have $|\Delta|$ larger than the reflections in $BATCH_{i+1}$. The last $TR\,\Delta$ value will coincide with a minimal value of $|\Delta|$ that we can be used in the phasing procedure $TR\,\Delta_F$.

Step 4. A supplementary batch

In order to impove the continuity in the Fourier map, an additional number of reflections in the low $\sin\theta\,/\,\lambda$ range is phased. The corresponding subset (*i.e.* $BATCH_{Last}$) will involve reflections with $\sin\theta\,/\,\lambda \leq (\sin\theta\,/\,\lambda)_{\max}\,/\,2$, provided

$|\Delta| \geq TR\,\Delta_F$ x 0.95 x 0.85 for reflections with restricted phase value

$|\Delta| \geq TR\,\Delta_F$ x 0.95 for reflections of general type.

Step 5. Triplet calculation

Let $\{T_{ii}\}$ be the set of triplet invariants among the reflections in $BATCH_i$ and let $\{T_{ij}\}$ be the set of triplets constituted by one reflection in $BATCH_i$ and two reflections in $BATCH_j$. In the procedure, we only calculated the sets $\{T_{i1}\}$ for i=1,2,... and we store for each *i*th set up to 50000 triplets (the most reliable ones).

Step 6. The phasing procedure

The phasing procedure is a multisolution one, where starting sets of phases are generated by a random process (Baggio, Woolfson, Declercq & Germain, 1978). Random phases are given to $NLAR/\,2$ reflections (Burla, Cascarano, & Giacovazzo, 1992) with unit weights for the origin and enantiomorph-fixing reflections and with weight equal 0.8 for the rest. Cycles of weighted tangent refinement are first applied to the $NLAR/\,2$ reflections and, after convergence the phasing process is extended to $NLAR$ reflections. As in *SIR88* (Burla, Camalli, Cascarano, Giacovazzo, Polidori, Spagna

& Viterbo, 1989) and *SIR92* (Altomare, Cascarano, Giacovazzo, Guagliardi, Burla, Polidori & Camalli, 1994), a weighted tangent formula is used for phase extension and refinement:

$$\begin{aligned}\tan\phi_{\mathbf{h}} &= \sum_j \beta_j \sin(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h-k}_j}) / \sum_j \beta_j \cos(\phi_{\mathbf{k}_j} + \phi_{\mathbf{h-k}_j}) \\ &= T_{\mathbf{h}} / B_{\mathbf{h}}\end{aligned} \qquad \text{(A.I.1)}$$

where β_j is defined by the equation

$$D_1(\beta_j) = D_1(A)\, D_1(\alpha_{\mathbf{k}_j})\, D_1(\alpha_{\mathbf{h-k}_j}) \qquad \text{(A.I.2)}$$

and

$$\alpha_{\mathbf{h}} = (T_{\mathbf{h}}^2 + B_{\mathbf{h}}^2)^{1/2} \qquad \text{(A.I.3)}$$

The reliability parameter $\alpha_{\mathbf{h}}$ of any determined phase $\phi_{\mathbf{h}}$ is modified according to the agreement between the calculated and the expected value of $\alpha_{\mathbf{h}}$. In particular, if $\alpha_{\mathbf{h}}$ is larger than the expected value

$$<\alpha_{\mathbf{h}}>= \sum_j A_j\, D_1(A_j) \qquad \text{(A.I.4)}$$

then the calculated $\alpha_{\mathbf{h}}$ is replaced by

$$<\alpha_{\mathbf{h}}> \exp\left[-\left(\alpha_{\mathbf{h}} - <\alpha_{\mathbf{h}}>^2\right) / 2\sigma_{\alpha_{\mathbf{h}}}^2\right]^{1/3} \qquad \text{(A.I.5)}$$

where

$$\sigma_{\alpha_{\mathbf{h}}}^2 = \frac{1}{2}\sum_j A_j^2 \left[1 + D_2(A_j) - 2\, D_1^2(A_j)\right]$$

The weighting scheme is designed to drive phases towards values that minimize the difference between α and $< \alpha >$ by reducing in the tangent refinement the importance of the phases with too large values of α.

The $NLAR_1$ reflections in $BATCH_1$ are phased according to above descriptions. Among the various trials provided by the multisolution approach, the most probable one is chosen as a seed for the subsequent phase extension.

The set $BATCH_2$ is phased from $BATCH_1$ by using the $\{T_{21}\}$ triplets: phases are then refined by making use of the triplets $\{T_{11}\} \cup \{T_{21}\}$. Since $TR\,\Delta_2 \equiv TR\,\Delta_1$, the average accuracy of the phases in $BATCH_2$ is expected to be very close to that of the reflections in $BATCH_1$. Therefore, for $i > 2$, the set $BATCH_i$ is phased from $BATCH_1$ by using the $\{T_{i1}\}$ triplets: phases are then refined by using the set of triplets $\{T_{11}\} \cup \{T_{21}\} \cup \{T_{i1}\}$. It is worthwhile noting that every set of phases so obtained is referred to the same origin, that fixed for set $BATCH_1$.

Appendix II

The probability distribution function P(|Δ|)

According to Hauptman (1982),

$$\begin{aligned} \mathrm{P}(R,S) = &\left[(4\,RS) / (1-\alpha^2)\right] \exp - \left[(R^2 + S^2) / (1-\alpha^2)\right] \\ &\mathrm{x}\; I_o\left[(2\alpha\, RS) / (1-\alpha^2)\right] \end{aligned} \qquad \text{(A.II.1)}$$

where

$$\alpha \cong \left\{ \left[\sigma_2\right]_p / \left[\sigma_2\right]_d \right\}^{1/2}$$

We first express (A.II.1) in terms of the pseudo-normalized structure factors S' and R',

$$\begin{aligned} \mathrm{P}(R',S') = &\, 4\,R'\,S'\left(\Sigma_H / \Sigma_p\right) \exp\left\{-\left[R'^2\left(\Sigma_d / \Sigma_p\right) + S'^2\right]\right\} \\ &\mathrm{x}\; I_o(2\,R'\,S') \end{aligned} \qquad \text{(A.II.2)}$$

Then we introduce the change of variable $\mathit{\Delta} = S' - R'$ and (A.II.2) becomes

$$\begin{aligned} \mathrm{P}(R',\mathit{\Delta}) = &\, 4\left(\Sigma_H / \Sigma_p\right) R'(R' + \mathit{\Delta}) \\ &\mathrm{x}\quad \exp\left\{-\left[2\,R'^2 + R'^2\left(\Sigma_H / \Sigma_p\right) + 2\,R'\,\mathit{\Delta} + \mathit{\Delta}^2\right]\right\} \\ &\mathrm{x}\quad I_o\left[2\,R'(R' + \mathit{\Delta})\right] \end{aligned} \qquad \text{(A.II.3)}$$

For $-3.75 \leq x \leq 3.75$, $I_o(x)$ may be approximated by a polynomial in even powers of t (see Abramowitz & Stegun, 1972), where $t = x/3.75$. For large values of R' it is not easy to compute (A.II.3) directly. For

$3.75 < x < \infty$, we approximate $I_o(x)$ by $Q(t)\exp(x)\,x^{-1/2}$, where Q is a suitable polynomial of order 8 in terms of t^{-1} .

We obtain

$$\begin{aligned} \mathrm{P}(R',\Delta) &= 2(2)^{1/2}\left(\Sigma_H/\Sigma_p\right)\exp\,(-\Delta^2)\left[R'(R'+\Delta)\right]^{1/2} \\ &\quad \mathrm{x}\ Q(t)\exp\left[-R'^2\left(\Sigma_H/\Sigma_p\right)\right] \end{aligned} \qquad \text{(A.II.4)}$$

Then,

$$\mathrm{P}(\Delta) = \int_0^\infty \mathrm{P}(R',\Delta)\,\mathrm{d}\,R' \qquad \text{(A.II.5a)}$$

for positive values of Δ, and

$$\mathrm{P}(\Delta) = \int_{-\Delta}^\infty \mathrm{P}(R',\Delta)\,\mathrm{d}\,R' \qquad \text{(A.II.5b)}$$

for negative values of Δ (the limits of integration are because $R' = S' - \Delta$ has be positive). Finally,

$$\mathrm{P}(|\Delta|) = \mathrm{P}(+|\Delta|) + \mathrm{P}(-|\Delta|) \qquad \text{(A.II.6)}$$

The distribution $\mathrm{P}(\Delta)$ normaly is calculated by numerical methods and for reason of simplicity the $\left(\Sigma_H/\Sigma_p\right)$ is replaced by $\sigma = \left[\sigma_2\right]_H / \left[\sigma_2\right]_p$. Curves corresponding to various values of σ are shown in Fig. A.II.1.

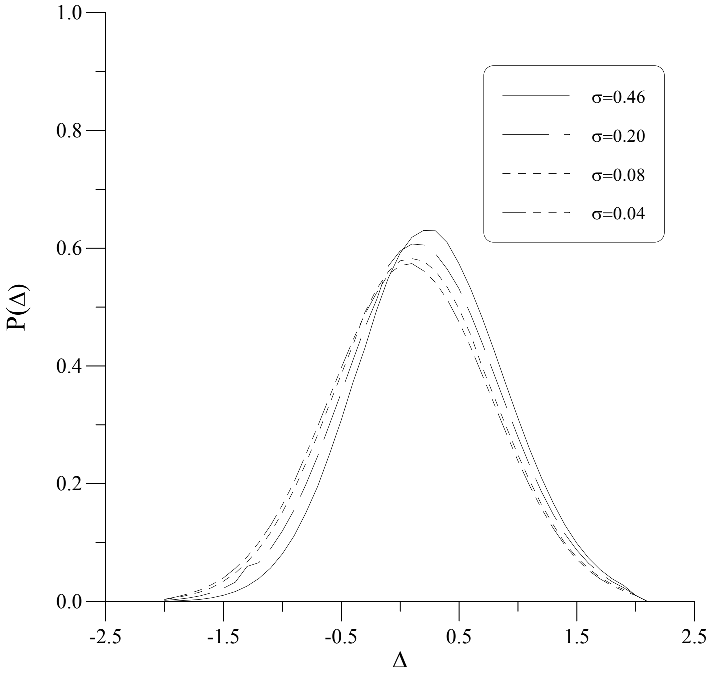


Fig. A.II.1: $P(\Delta)$ distribution for selected values of σ

References

Abramowitz, M. & Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover Publications, Inc.

Altomare, A., Cascarano, G., Giacovazzo, C., Guagliardi, A., Burla, M. C., Polidori, G. & Camalli, M. (1994). *J. Appl. Cryst.* **27**, 435.

Baggio, R., Woolfson, M. M., Declercq, J.-P. & Germain, G. (1978). *Acta Cryst.* A**34**, 883-892.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography,* **p**.336. *Academic Press, London.*

Blundell, T. L., Pitts, J. E., Tickle, I. J., Wood, S. P. & Wu, C. W. (1981). *Proc. Natl. Acad. Sci. USA*, **7**, 4175-4179.

Burla, M. C., Camalli, M., Cascarano, G., Giacovazzo, C., Polidori, G., Spagna, R. & Viterbo, D. (1989). *J. Appl. Cryst.* **22**, 389-393.

Burla, M. C., Cascarano, G., Giacovazzo, C. (1992). *Acta Cryst.* A**48**, 906-912.

Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1992 a). *Z. f. Kristallogr.* **200**, 63-71.

Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1992 b). *Acta Cryst.* A**48**, 859-865.

Cascarano, G., Giacovazzo, C. & Viterbo, D. (1987). *Acta Cryst.* A**43**, 22-29.

Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Acta Cryst.* **14**, 1188-1195.

Giacovazzo, C., Cascarano, G. & Zheng, C. (1988). *Acta Cryst*, A**44**, 45-51.

Giacovazzo, C., Guagliardi, A., Ravelli, R. & Siliqi, D. (1994). *Z. f. Kristallogr.* **209**, 136-142.

Giacovazzo, C., Siliqi, D. & Ralph, A. (1994). *Acta Cryst.* A**50**, 503-510.

Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst.* A**50**, 609-621.

Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *Acta Cryst.* A**51**, 177-188.

Glover, I., Haneef, I., Pitts, J., Woods, S., Moss, D., Tickle, I. & Blundell, T. L. (1983). *Biopolymers*, **22**, 293-304.

Hall, S. R. & Subramanian, Y. (1982a). *Acta Cryst.* A**38**, 590-598.

Hall, S. R. & Subramanian, Y. (1982b). *Acta Cryst.* A**38**, 598-608.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 289-294.

Hecht, H., Erdmann, H., Park, H., Sprinzl, M. & Schmid, R. D. (1995). *In preparation*

Hecht, H., Erdmann, H., Park, H., Sprinzl, M., Schmid, R. D. & Schomburg, D. (1993). *Acta Cryst.* A**49**, *Suppl.* 86.

Hecht, H., Sobek, H., Haag, T., Pfeifer, O. & Van Pee, K. H. (1994). *Nature (London) Struct. Biol.* **1**, 532-537.

Kostrewa, D., Grazin, J., Stock, D., Choe, H., Labahn, J. & Saenger, W. (1992). *J. Mol. Biol.* **226**, 209-226.

Kretsinger, R. H. & Nockolds, C. E. (1973). *J. Biol. Chem.* **248**, 3313-3326.

Labahn, J., Granzin, J., Schluckebier, G., Robinson, D. P., Jack, W. E., Schildkraut, I. & Saenger, W. (1994). *Proc. Natl. Acad. Sci. In press*.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., De Kok, A. & Hol, W. G. J. (1992). *Science*, **255**, 1544-1550.

Richarson, J. S. & Richardson, D. C. (1985). *Method in Enzymology*, **115**B, 189-206. De. by H. W. Wyckoff, C. H. W. HIRS & S. N. Timasheff, *Academic Press, Inc., Orlando*.

Subramanian, V. & Hall, S. R. (1982). *Acta Cryst*. A**38**, 577-590.

Woolfson, M. M. & Yao, J.-X. (1990). *Acta Cryst*. A**46**, 409-413.

Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541-18550.

Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst*. A**46**, 41-46.

Chapter IV

Phasing up to derivative resolution

Introduction

Two pioneristic papers by Hauptman (1982 a,b) showed how direct methods may be integrated with isomorphous replacement techniques. The triplet phase invariant of the native protein $\varPhi = \phi_{\mathbf{h}} - \phi_{\mathbf{k}} - \phi_{\mathbf{h-k}}$ was estimated *via* a von Mises distribution whose reliability coefficient A depends on an intricate interrelationship among the six moduli $R_{\mathbf{h}}$, $R_{\mathbf{k}}$, $R_{\mathbf{h-k}}$, $S_{\mathbf{h}}$, $S_{\mathbf{k}}$, $S_{\mathbf{h-k}}$. Hauptman approach has been reconsidered by Giacovazzo, Cascarano & Zheng (1988): a simpler distribution

$$\mathrm{P}(\varPhi \,|\, R_{\mathbf{h}}, R_{\mathbf{k}}, R_{\mathbf{h-k}}, S_{\mathbf{h}}, S_{\mathbf{k}}, S_{\mathbf{h-k}}) \cong \left[2\pi\ I_o(A)\right]^{-1} \exp(A\cos\varPhi) \qquad \text{(IV.1)}$$

was obtained, where

$$A = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_p R_{\mathbf{h}} R_{\mathbf{k}} R_{\mathbf{h-k}} + 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \varDelta_{\mathbf{h}} \varDelta_{\mathbf{k}} \varDelta_{\mathbf{h-k}} \qquad \text{(IV.2)}$$

and $\varDelta = \left(\left|F_d\right| - \left|F_p\right|\right) / \Sigma_H^{1/2}$ is the pseudo-normalized difference (with respect to the heavy-atom structure). Since $\left[\sigma_3 / \sigma_2^{3/2}\right]_H >> \left[\sigma_3 / \sigma_2^{3/2}\right]_p$, the Cochran parameter is often negligible with respect to the term including pseudonormalized differences: this last may attain large values even for large proteins. Since $\varDelta_{\mathbf{h}} \varDelta_{\mathbf{k}} \varDelta_{\mathbf{h-k}}$ may be positive or negative, positive as well as negative triplets can be identified *via* (IV.2).

Hauptman's as well as Giacovazzo, Cascarano & Zheng formulas succeded when applied to calculated data, but failed when applied to real experimental data. The common believe was that the experimental data were too inaccurate to be used in direct methods applications: in particular the general feeling was that lack of isomorphism between native and derivative structure combined with errors in the experimental data and/or in their mathematical treatment hinder any success when direct phasing procedures are applied to experimental data even if the structure solution could be straightforwardly solved *via* ideal error-free data.

The entire situation has been reconsidered by Giacovazzo, Siliqi & Ralph (1994), which focused the attention onto the case in which diffraction data of one isomorphous derivative are available. It was shown that in such a case the direct *ab-initio* solution of protein structures is feasible in principle. Giacovazzo, Siliqi & Spagna (1994) described a direct procedure successfully applying the formula by Giacovazzo, Cascarano & Zheng to real data. The approach was remarkably different from the typical procedures used for small-molecule crystal structure solution: the crucial innovations concerned the normalization process (of the derivative with respect to the native protein), the phasing procedure, and the figures of merit for finding the correct among the various solutions provided by the random starting phase approach. It was shown that a certain number of reflexions (roughly speaking, less than 0.15 of the total number of reflexions up to the derivative resolution) could be phased with a limited phase error.

The main limitation of the procedure described by Giacovazzo, Siliqi & Spagna (1994) was the small number of phased reflexions (rather than the quality of the assigned phases). As a consequence, the corresponding electron density maps suffered by severe series truncation effects. However the assigned phases were of such a high quality that they could be used as a starting point for a reliable phase extension process. This was just the aim of the paper by Giacovazzo, Siliqi & Zanotti (1995): the probability distribution function $\mathrm{P}(\varDelta)$ was derived, which suggested, as a rule of thumb, to extend the phasing process to reflexions with $|\varDelta| \geq 0.5$. The result was that about 40% of the measured reflections (up to derivative resolution) could be phased without paying too much in terms of the quality of the new phases. The phase extension process was fast and could be run in a completely automatic way.

The following drawbacks were still limiting the usefulness of the procedure described by Giacovazzo, Siliqi & Zanotti (1995): (a) a non negligible number of reflections with $|\Delta| < 0.5$ but large *R* value remained unphased: phasing them could valuably contribute to make interpretable the electron density maps. (b) The solution with the highest figure of merit was not always the correct solution; (c) The phasing procedure could not be applied when the derivative resolution was about 4Å or lower. (d) Pseudo-centrosymmetric phases were provided in specific space groups.

The points (b) - (c) were discussed in other work of Giacovazzo, Siliqi & Gonzalez-Platas (1995): a more robust normalizing procedure was designed which made explicit use of the distribution $\mathrm{P}(\Delta)$. Histogram matching procedures were used to obtain an optimal fit of the observed Δ distribution with the expected one. The new Δ's proved statistically more meaningful and were able in most of the cases to overcome, for data up to 4Å resolution , the disturbing consequences provoked on the Wilson plot by strong Debye effects. A method was also suggested for discarding the false solutions: an *a posteriori* check of the heavy atom positions allows to discard the trials corresponding to the so called (in small molecule direct methods applications) "*uranium solutions*".

It may be shown now (Giacovazzo, Siliqi, Gonzalez-Platas, Hecht, Zanotti, Krauss & York, 1995) show that all the reflexions up to derivative resolution may be phased in principle, so improving the quality of the final electron density map. The various sources of errors are shortly analysed, in order to provide higher insight into the limits and the advantages of the phasing method. Owing to such errors it is also suggested that about 0.20 of the total number of reflexions up to derivative resolution can be omitted from the phasing procedure without causing valuable impoverishment of the quality of the electron density map.

In Table I we show the code name, the space group and the crystallochemical data of our test structures, in Table II the relevant parameters concerning the diffraction data are given.

The scaling procedure and the sign inversion of Δ

According to (IV.1) and (IV.2) the expected value of the triplet phase Φ mostly depends on the value of $\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{h-k}}$. If $|\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{h-k}}|$ is sufficiently large and $\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{h-k}} > 0$ then Φ is expected close to zero; if $|\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{h-k}}|$ is large and $\Delta_{\mathbf{h}}\Delta_{\mathbf{k}}\Delta_{\mathbf{h-k}} < 0$ then Φ is expected close to π. Relationships (IV.1) and (IV.2) were obtained on assuming perfect isomorphism between native and derivative structures.

Structure Code	Space Group	Molecular Formula	Z
APP [1]	C 2	$C_{190}N_{53}O_{58}Zn$	4
BPO [2]	P 2_1 3	$C_{2744}N_{712}O_{1073}$	12
E2 [3]	F 4 3 2	$C_{1170}N_{310}O_{366}S_7$	96
M-FABP [4]	P 2_1 2_1 2_1	$C_{667}N_{170}O_{261}S_3$	4
NOX [5]	P 4_1 2_1 2	$C_{1034}O_{704}N_{299}S_2P_{1/8}$	8

Table I: Code name, space group and crystallochemical data for test structures.
(1) Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983)
(2) Hetch *et al*. (1994)
(3) Mattevi, Oblomova, Schulze,Kalk, Westphal, De Kok & Hol (1992)
(4) Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992)
(5) Hetch *et al* (1993); Hetch *et al* (1993)

Structure Code	Native		Derivative			
	RES(Å)	**NREFL**	**Heavy atom**	$[\sigma_2]_H/[\sigma_2]_p$	**RES(Å)**	**NREFL**
APP	0.99	17058	Hg	0.23	2.00	2086
BPO	2.35	23956	Au	0.06	2.78	15741
			Pt	0.06	2.78	14786
E2	3.00	10388	Hg	0.08	3.00	9179
M-FABP	2.14	7595	Hg	0.06	2.15	7125
NOX	3.00	4619	Pt	0.08	3.00	4619

Table II: Relevant parameters for diffraction data of test structures.

Owing to lack of isomorphism and/or errors in the experimental data and/or in their mathematical treatement one Δ could invert its sign with respect to the sign corresponding to ideal error-free data: then the expected triplet phase should change by π.

High frequency of the sign inversion for Δ deteriorates the efficiency of our direct phasing procedure. In Fig.1 (this is Fig.8 in Giacovazzo, Siliqi & Spagna, 1994) the percentage of the reflexions that undergo sign inversion for Δ' was shown for APP, CARP, E2 and M-FABP.

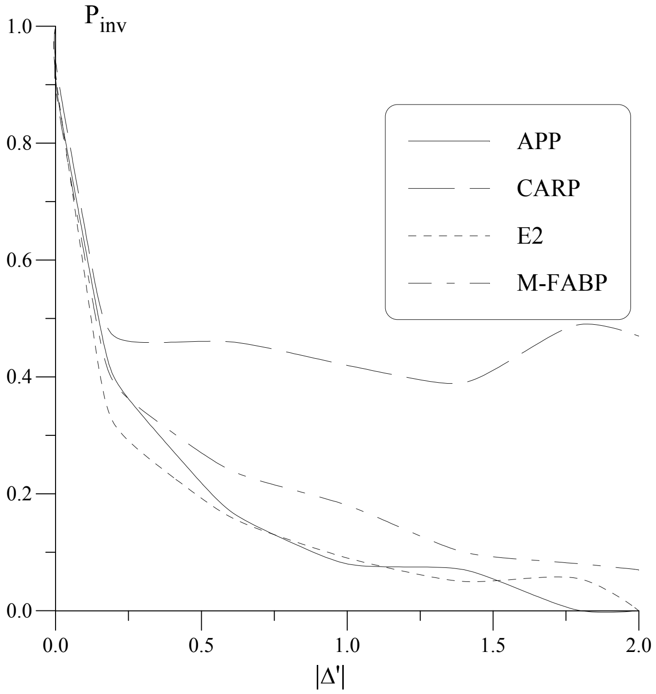


Fig.1: Percentage of reflections that undergo sign inversion for Δ' as a result of the normalization process in Giacovazzo, Siliqi & Spagna (1994) and of physical sources of error (mostly lack isomorphism and error in measurements). Experimental data are used.

Such a figure discouraged any attempt at phasing reflexions with small $|\Delta'|$ values because the frequency of the sign inversion for them is too large; in particular it was larger than 0.5 for very small $|\Delta|$'s, that is, worse than for randomly distributed signs. In order to understand the reason of such a systematic error let us compare Fig.1 with Fig.2 (this is the Fig.7 in Giacovazzo, Siliqi & Spagna, 1994), where, for calculated error free data, the percentage of Δ' which undergo sign inversion as a result of the mere normalization process is given.

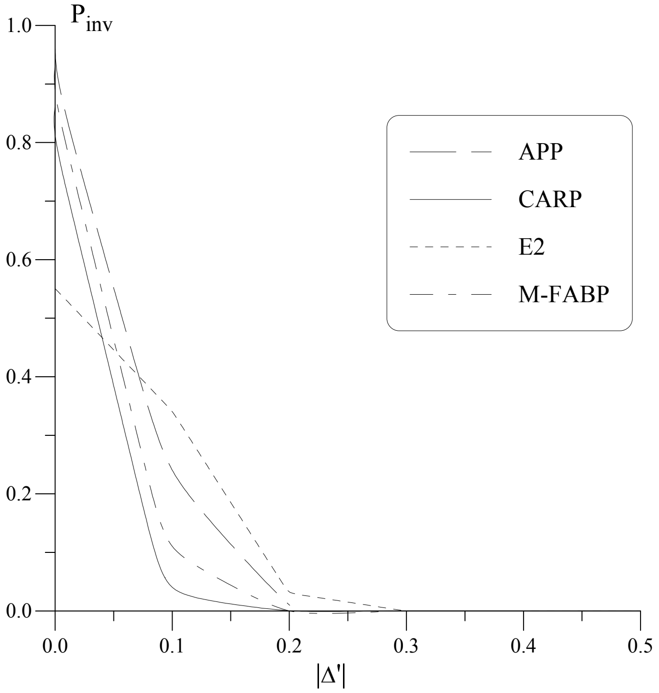


Fig.2: Percentage of reflections that undergo sign inversion for Δ' as a result of the normalization process derived by Giacovazzo, Siliqi & Spagna (1994). (Calculated error-free data).

We observe that the frequency of the sign inversion at small $|\Delta|$ values in Fig.2 is extremely large, with a trend very close to the inversion frequency depicted in Fig.1. As a consequence, phase extension to reflexions with small $|\Delta'|$ value is fruitful in the practice only if a normalization procedure is available which avoids the systematic errors at small $|\Delta'|$ values shown in Fig.2.

Let us now apply the normalization process described by Giacovazzo, Siliqi & Gonzalez-Platas (1995) to ideal error-free data in order to evaluate the percentage of the reflections which undergo sign inversion for Δ as a consequence of the mere mathematical data treatment. The results is shown in Fig.3: the inversion frequency [called here $(P_{inv})_n$] is practically negligible up to $|\Delta|$ = 0.08, and is never larger than 0.4, so confirming the higher efficiency of the new normalization process.

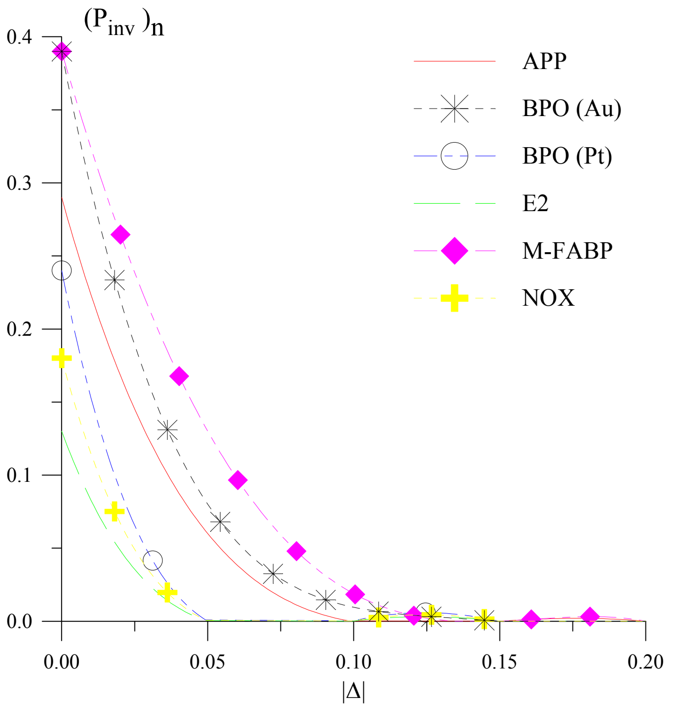


Fig.3: Percentage of reflexions which undergo sign inversion for Δ as a result of the normalization process described by Giacovazzo, Siliqi & Gonzalez-Platas (1995) (calculated error-free data).

In order to have a countercheck that such a normalization process allows a fruitful phase extension to small $|\Delta|$ values we show in Fig.4 the inversion frequency for the test structures in Table I as obtained by applying to the experimental data the normalization procedure described by Giacovazzo, Siliqi & Gonzalez-Platas (1995).

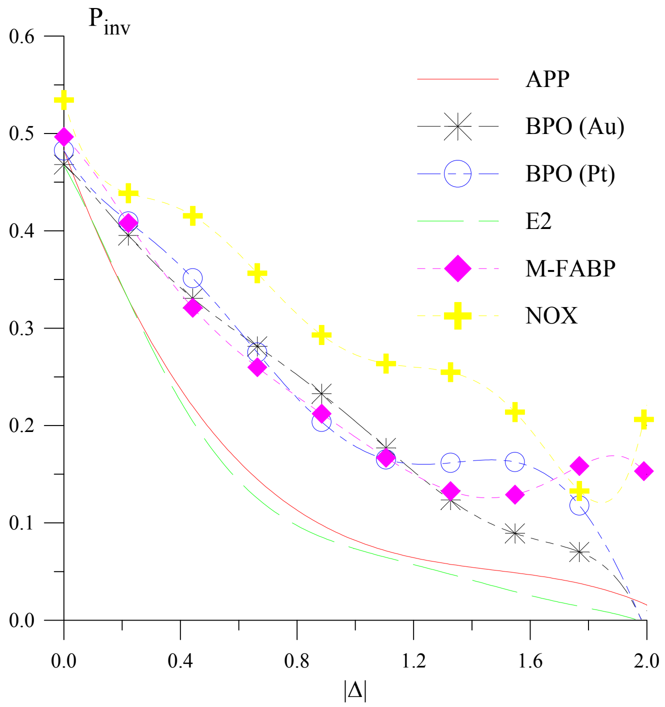


Fig.4: Percentage of reflexions that undergo sign inversion for Δ as a result of the combination of mathematical data treatment (i.e., the normalization process), block of isomorphism and errors in measurements.

The inversion frequency in Fig. 4 is dramatically smaller at low $|\Delta|$ than in Fig.1; this confirms the higher quality of the normalization procedure proposed by Giacovazzo, Siliqi & Gonzalez-Platas (1995), and suggests that even reflexions with $|\Delta|$ smaller than 0.5 could be conveniently phased.

The problems are now: Is it worthwhile phasing all the reflexions up to derivative resolution, or can some reflexions be excluded from the phasing process without detriment for the quality of the electron density map? How to select the reflexions to which it is worthwhile extending the phasing process? Can some general criteria be fixed? We will show that a minimum threshold for $|\Delta|$ and the standard deviation associated to reflection intensity measurements are sensible criteria to be applied.

The average phase error as a function of $|\Delta|$

Figure 4 shows that for experimental data $\mathrm{P}_{inv} \leq 0.50$. The maximum value is attained close to $|\Delta| \cong 0$, where the combined effect of the mathematical treatement of the data (see Fig. 3) and of the lack of isomorphism make the experimental sign of Δ completely unreliable. A reasonable criterion has to be found balancing the advantage of employing more data and the disadvantage that the extra data may have a higher phase error.

The phasing process described by Giacovazzo, Siliqi & Zanotti (1994) and Giacovazzo, Siliqi & Gonzalez-Platas (1995) extends step by step the phase determination to batches of reflexions with progressively smaller values of $|\Delta|$. The trend of the phase extension process as a function of a threshold $TR\,\Delta$ may be deduced from Table III, where, for each batch, we give the number of reflections (NREFL), the cumulative average phase error (ERR) and the weighted cumulative phase error (W-ERR). A better insight can be obtained by calculating the differential average phase error (ERRD). While ERR refers to all the reflexions with $|\Delta| > TR\,\Delta$, ERRD refers only to the NREFLD reflexions with $|\Delta|$ between the current $TR\,\Delta$ value and the preceding one.

Both ERR and ERRD increase by decreasing $TR\,\Delta$ values, but ERRD in a dramatical way. However extending phases to a larger number of reflexions improves the quality of our electron density map ρ (Lunin & Woolfson, 1993). This may be monitored by calculating the correlaction factor (CORR) between ρ and the "correct" map ρ_{mod} (obtained *via* model phases, all reflexions up to derivative resolution included):

$$\mathrm{CORR} = \frac{\langle \rho\ \rho_{mod} \rangle - \langle \rho \rangle \langle \rho_{mod} \rangle}{\left(\langle \rho^2 \rangle - \langle \rho \rangle^2 \right)^{1/2} \left(\langle \rho_{mod}^2 \rangle - \langle \rho_{mod} \rangle^2 \right)^{1/2}}$$

In Table III CORR is given for the various maps calculated by using reflexions with $|\Delta| > TR\,\Delta$. CORR increases with decreasing values of $TR\,\Delta$ except for very small $TR\,\Delta$ values. It may be argued from Table III that reflexions with $|\Delta| \leq 0.1$ do not provide valuable additional information to the electron density map, so that they could be skipped from the phasing process.

APP					
TRΔ	**NREFL**	**ERR (W-ERR)**	**NREFLD**	**ERRD (W-ERRD)**	**CORR**
0.0	2107	62.4 (57.9)	-	-	0.494
0.1	1864	60.1 (56.6)	243	75.5 (79.6)	0.494
0.2	1623	56.9 (54.5)	241	81.7 (81.3)	0.496
0.3	1396	54.1 (52.8)	227	73.9 (72.7)	0.496
0.4	1173	50.4 (49.9)	223	73.7 (73.5)	0.494
0.5	980	48.2 (48.1)	193	61.6 (61.1)	0.487

BPO (Au)					
TRΔ	**NREFL**	**ERR (W-ERR)**	**NREFLD**	**ERRD (W-ERRD)**	**CORR**
0.0	15731	62.3 (56.3)	-	-	0.461
0.1	13929	59.2 (54.5)	1802	86.6 (85.6)	0.464
0.2	12159	56.2 (52.6)	1770	75.1 (74.4)	0.461
0.3	10452	53.1 (50.4)	1707	68.6 (67.3)	0.459
0.4	8865	50.3 (48.3)	1587	68.2 (67.3)	0.448
0.5	7389	47.7 (46.2)	1476	62.4 (61.6)	0.429

BPO (Pt)					
TRΔ	**NREFL**	**ERR (W-ERR)**	**NREFLD**	**ERRD (W-ERRD)**	**CORR**
0.0	14777	62.9 (57.0)	-	-	0.468
0.1	13089	59.8 (55.2)	1688	86.5 (86.5)	0.469
0.2	11427	56.6 (53.1)	1662	81.8 (81.7)	0.468
0.3	9832	53.8 (51.1)	1595	74.1 (73.4)	0.462
0.4	8329	50.9 (48.9)	1503	69.8 (68.6)	0.450
0.5	6971	48.4 (46.8)	1358	63.6 (62.3)	0.437

E2					
TRΔ	**NREFL**	**ERR(W-ERR)**	**NREFLD**	**ERRD (W-ERRD)**	**CORR**
0.0	7756	59.8 (53.4)	-	-	0.539
0.1	6868	56.2 (51.6)	888	87.6 (87.6)	0.539
0.2	5991	52.3 (49.2)	877	82.6 (82.3)	0.538
0.3	5143	48.8 (46.7)	848	73.5 (74.5)	0.532
0.4	4369	45.5 (44.3)	774	67.3 (67.3)	0.527
0.5	3644	42.6 (42.0)	725	60.3 (60.4)	0.507

M-FABP					
TRΔ	**NREFL**	**ERR (W-ERR)**	**NREFLD**	**ERRD (W-ERRD)**	**CORR**
0.0	7122	69.4 (64.0)	-	-	0.401
0.1	6301	66.4 (62.3)	821	92.6 (92.4)	0.401
0.2	5499	64.2 (60.8)	802	81.5 (81.0)	0.398
0.3	4727	61.3 (58.9)	772	81.5 (79.7)	0.393
0.4	3998	58.9 (56.9)	729	74.8 (73.2)	0.379
0.5	3337	56.1 (54.7)	661	73.0 (70.9)	0.367

NOX					
TRΔ	**NREFL**	**ERR (W-ERR)**	**NREFLD**	**ERRD (W-ERRD)**	**CORR**
0.0	4613	77.7 (72.6)	-	-	0.281
0.1	4084	75.7 (71.4)	529	93.1 (92.8)	0.281
0.2	3559	74.7 (70.5)	525	82.3 (81.6)	0.280
0.3	3065	72.9 (69.1)	494	86.2 (84.2)	0.280
0.4	2595	71.6 (67.8)	470	80.3 (79.4)	0.277
0.5	2163	69.5 (65.7)	432	82.1 (80.0)	0.269

Table III: Cumulative average phase error (ERR) and differential average phase error (ERRD) when the phasing process is extended to reflexions with $|\Delta| > TR\Delta$ (experimental data). For each threshold $TR\Delta$ the value of the correlation factor (CORR) between our electron density map and the "correct" map obtained via model phases is given.

The role of the standard deviation of the intensity measurements

While the effects of the lack of isomorphism are *a-priori* unpredictable, the effects of the errors in measurements may be partially controlled by exploiting the standard deviation $\sigma(|F|)$ usually associated to the structure factor modulus $|F|$. The value of $\sigma(|F|)$ is probably an underestimate of the total error since it takes into account only errors coming out from the statistical fluctuations in the X-ray intensity and neglects systematic variations like inaccuracy in absorption corrections, misalignement of the crystal, etc. However $\sigma(|F|)$ provides a reasonable estimate of the relative reliability of different measurements and therefore can be used for estimating the percentage of sign inversion for Δ caused by counting errors in intensity memasurements.

Let us suppose that errors in measurements are distributed according to the normal distribution: then

$$\sigma(\Delta) \cong \left[\sigma\left(\left|F_p\right|\right) + \sigma\left(\left|F_d\right|\right)\right] \Sigma_H^{-1/2}.$$

On the first approximation the variable $y = \Delta / \sigma(\Delta)$ may be considered to be distributed, in absence of systematic errors, according to $\mathrm{N}[y; y_m, 1]$ where $y_m = [\Delta / \sigma(\Delta)]_m$ is the value of y obtained from measurements. For positive Δ values (but the final results will hold also for negative values of Δ) the probability of the sign inversion is equal to the integral

$$\int_{-\infty}^{0} \mathrm{N}[y; y_m, 1] \;\mathrm{d}\,y \tag{IV.3}$$

If the normalized variable $z = y - y_m$ is substituted to y , the probability of the sign inversion due to counting errors is equal to

$$\left(\mathrm{P}_{inv}\right)_c = \int_{-\infty}^{-y_m} \mathrm{N}[z;0,1] \;\mathrm{d}\,z = 1 - \left\{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y_m} \exp\left(-z^2/2\right) \;\mathrm{d}\,z\right\} = 1 - \Phi(y_m) \tag{IV.4}$$

which is tabulated in standard books.

$\Phi(y_m)$ is the well known error function: for reader usefulness we show $\left(\mathrm{P}_{inv}\right)_c$ in Fig. 5. The Figure suggests that: (a) $\left(\mathrm{P}_{inv}\right)_c$is about 0.15 for reflections with $|\Delta|/\sigma(|\Delta|)=1$. Since the robustness of the phasing method can bear larger sign inversion frequency, even reflections with $|\Delta|/\sigma(|\Delta|)\leq 1$ can be fruitfully involved into the phasing process. A more reasonable criterion may be to include reflexions with $|\Delta|/\sigma(|\Delta|)\geq 0.80$: that corresponds to $\left(\mathrm{P}_{inv}\right)_c \cong 0.20$. Such a criterion does not characterize a fixed percentage of the reflexions, since this number depends on the quality of the data.

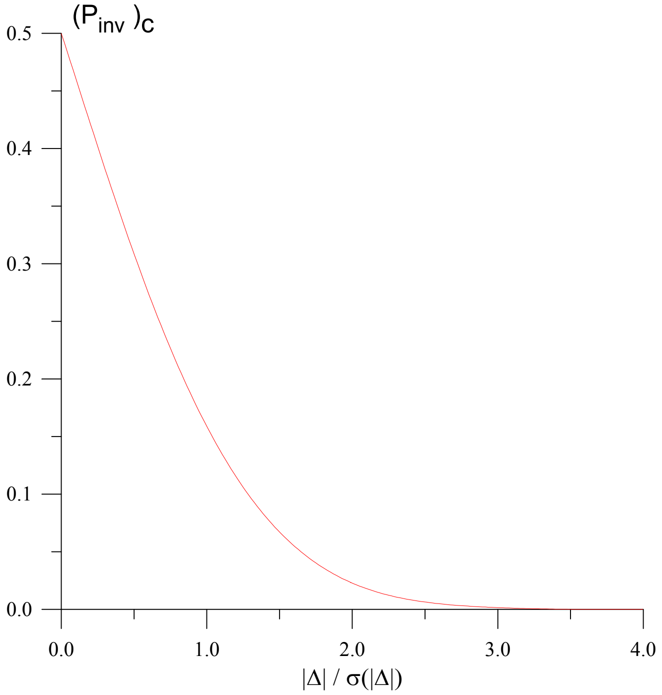


Fig.5: The sign inversion frequency for Δ as a result of the counting errors in intensity measurements (experimental data).

The efficiency of the criterion may be judged from Table IV, where, for the various test structures, the average phase error (as calculated by our phasing procedure) is given for the reflexions with $|\Delta|/\sigma(|\Delta|)\leq 0.80$. It is immediately seen that such reflexions, once phased, should not add relevant information to the electron density map, and could therefore be excluded from the phasing process.

Structure Code	NREF	ERR (W-ERR)
APP(*)	-	-
BPO (Au)	3196	82.6 (81.4)
BPO (Pt)	2455	85.2 (84.4)
E2	1352	90.8 (90.8)
M-FABP	1376	89.7 (89.5)
NOX	440	95.4 (95.4)

Table IV: NREF is the number of reflexions with $|\Delta|/\sigma(|\Delta|) < 0.8$; ERR is the average phase error for such reflexions (W-ERR) is the weighted average phase error.
(*) $\sigma(|F|)$ are not available for this structure.

When reflexions with $|\Delta|/\sigma(|\Delta|) \leq 0.80$ or $|\Delta| \leq 0.1$ are excluded from the procedure the CORR values shown in Table V are obtained. The Table shows that the average phase error of the phased reflexions is markedly smaller than in Table III at $TR\,\Delta = 0$, and that no structural information is lost by applying the above omit criteria.

Structure Code	NREFL	ERR (W-ERR)	CORR
APP(*)	1864	60.1 (56.6)	0.494
BPO (Au)	12449	56.9 (52.9)	0.463
BPO (Pt)	11736	57.2 (53.4)	0.470
E2	6240	53.2 (49.5)	0.540
M-FABP	5664	64.1 (60.5)	0.402
NOX	4039	75.4 (71.2)	0.285

Table V: For each test structure the number of reflexions with $|\Delta|/\sigma(|\Delta|) > 0.8$ and $|\Delta| > 0.1$ are given, together with the average phase error and the correlation coefficient CORR of our map with the model electron density map.
(*) Only the criterion $|\Delta| > 0.1$ is used.

The relation between traditional isomorphous derivative techniques and direct methods

Our procedure (generically defined as DMT, direct methods techniques, from now on) is able to estimate Φ in absence of any information on the heavy-atom structure. Traditional isomorphous derivative techniques (from now on referred as AT, algebraic techniques) estimate single phases provided the heavy-atom structure is known. A first question may be: are DMT and AT estimates consistent? In order to answer the above question we will analyze a few practical cases.

Let us suppose that Δ_1, Δ_2, Δ_3 are all positive and large. Then DMT suggest $\Phi \approx 0$, where "≈" stays for "probably equal to". Once the heavy-atoms have been located the AT suggest:

$$\text{if } \Delta_1 > 0\,,\ \phi_{p_1} \approx \phi_{H_1} \tag{IV.5a}$$

$$\text{if } \Delta_2 > 0\,,\ \phi_{p_2} \approx \phi_{H_2} \tag{IV.5b}$$

$$\text{if } \Delta_3 > 0\,,\ \phi_{p_3} \approx \phi_{H_3}\,. \tag{IV.5c}$$

Summing (IV.5a)-(IV.5c) gives

$$\Phi \approx \Phi_H \tag{IV.6}$$

where $H_1 + H_2 + H_3 = 0$ and $\Phi_H = \phi_{H_1} + \phi_{H_2} + \phi_{H_3}$. If $|\Delta_1|$, $|\Delta_2|$, $|\Delta_3|$ are sufficiently large also $|E_{H_1}|$, $|E_{H_2}|$, $|E_{H_3}|$ will be large (by definition $|\Delta_i| \le |E_{H_i}|$). If the number of heavy atoms in the unit cell is small (as usual) then Φ_H is expected close to zero, so that DMT and AT triplet estimates agree. When one of the Δ_i's is relatively small DMT and AT may diverge since Φ_H is expected far from zero.

Suppose now that $\Delta_1 > 0$, $\Delta_2 > 0$ and $\Delta_3 < 0$. Then DMT suggest $\Phi \approx \pi$ and AT provide, once the heavy atoms have been located, the following indications:

$$\text{if } \Delta_1 > 0\,,\ \phi_{p_1} \approx \phi_{H_1} \tag{IV.7a}$$

$$\text{if } \Delta_2 > 0\,,\ \phi_{p_2} \approx \phi_{H_2} \tag{IV.7b}$$

$$\text{if } \Delta_3 < 0\,,\ \phi_{p_3} \approx \phi_{H_3} + \pi\,. \tag{IV.7c}$$

Summing (IV.7a)-(IV.7c) gives

$$\Phi \approx \Phi_H + \pi\,, \tag{IV.8}$$

which again agrees with DMT estimates only if $\Phi_H \approx 0$. Two questions now arise: (1) Is the condition $\Phi_H \approx 0$ fulfilled in the practice for most of the cases? (2) If DMT and AT diverge which ones should be considered more reliable?

As for the first question we show in Table VI for *E2* the average value of $< |\Phi_H| >$ calculated for all the reflexions with $|\Delta| > 0.84$. It is easy seen that $< |\Phi_H| >$ is far from beeing close to zero in most of the cases. Since the number of heavy atoms in the unit cell is small the high value of $< |\Phi_H| >$ is rather surprising. However all becomes clear when one considers that, because of the "*errors*" above discussed, large experimental values of $|\Delta|$ are often associated to smaller values of R_H. In the practice a non-negligible percentage of reflexions with $|\Delta| > 0.84$ show R_H values markedly smaller than 0.84.

G_H	**NTRIP**	$< \|\Phi_H\| >$
0.0 - 0.2	15292	85.9
0.2 - 0.4	25129	79.2
0.4 - 0.8	66819	68.0
0.8 - 1.2	51528	53.7
1.2 - 1.6	30043	41.8
1.6 - 2.0	17011	33.6
2.0 - 15.0	18163	25.9

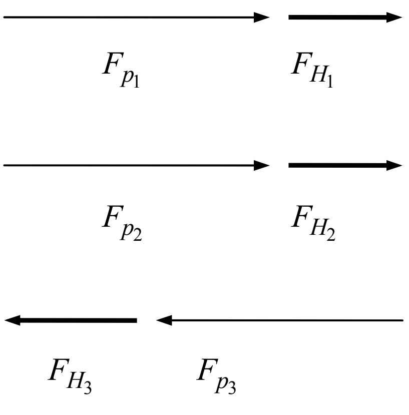
Table VI: E2: The values $<|\Phi_H|>$ for selected ranges of $G_H = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \left|E_{H_1} E_{H_2} E_{H_3}\right|$. NTRIP is the number of triplet invariants for which $<|\Phi_H|>$ is calculated (TRΔ=0.84).

The statistics does not qualitatively change (see Table VII) if calculations are made for all the triplets used in the phasing process (i.e., for $TR\,\Delta = 0$).

G_H	**NTRIP**	$<\lvert\Phi_H\rvert>$
0.0 - 0.2	176086	85.6
0.2 - 0.4	224541	78.8
0.4 - 0.8	446161	67.4
0.8 - 1.2	295015	53.1
1.2 - 1.6	165372	41.5
1.6 - 2.0	90630	33.7
2.0 - 15.0	105796	25.3

Table VII: E2: The values $<\lvert\Phi_H\rvert>$ for selected ranges of $G_H = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \left|E_{H1} E_{H2} E_{H3}\right|$. NTRIP is the number of triplet invariants on which $<\lvert\Phi_H\rvert>$ is calculated (TRΔ=0.0).

We have so far proved that DMT and AT are not equivalent methods even if they are correlated with each other. If their estimates diverge which one should be considered more efficient? The results shown in the preceding section suggest the larger efficiency of DMT methods: which is the source of such an effectiveness?



$$\Delta_1 > 0,\ \phi_{p_1} = 0,\ \phi_{H_1} = 0$$

$$\Delta_2 > 0,\ \phi_{p_2} = 0,\ \phi_{H_2} = 0$$

$$\Delta_3 > 0,\ \phi_{p_3} = \pi\ ,\ \phi_{H_3} = \pi$$

Fig.6 a: Triplets with phase value symmetry restricted to (0,π), having $\Phi_H = \phi_{H1} + \phi_{H2} + \phi_{H3} = \pi$. $\Delta_1 \Delta_2 \Delta_3 > 0$: according to (IV.1) the triplet phase is expected to be 2π but the true value is π.

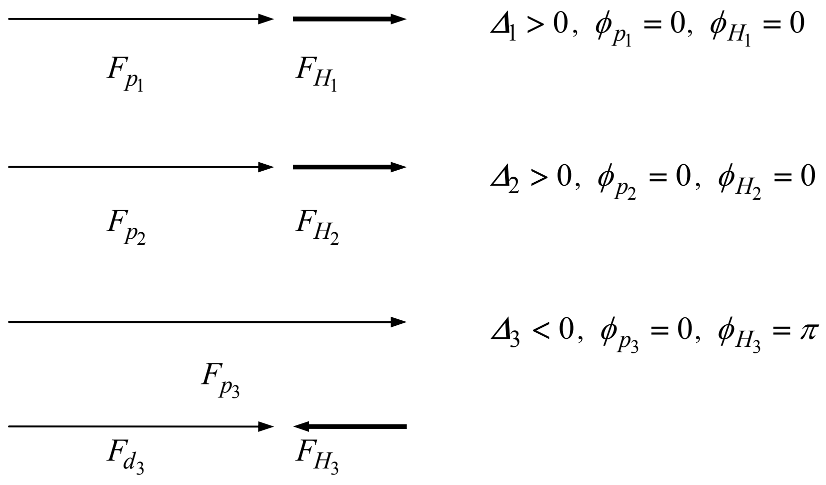

$\Delta_1 > 0,\ \phi_{p_1} = 0,\ \phi_{H_1} = 0$
F_{p_1}
F_{H_1}
$\Delta_2 > 0,\ \phi_{p_2} = 0,\ \phi_{H_2} = 0$
F_{p_2}
F_{H_2}
$\Delta_3 < 0,\ \phi_{p_3} = 0,\ \phi_{H_3} = \pi$
F_{p_3}
F_{d_3}
F_{H_3}

Fig.6 b: Triplets with phase value symmetry restricted to (0,π), having $\Phi_H = \phi_{H1} + \phi_{H2} + \phi_{H3} = \pi$. $\Delta_1\Delta_2\Delta_3 < 0$: according to (IV.1) the triplet phase is expected to be π but the true value is 2π.

Let us first examine the case of triplets with phase value symmetry restricted to (0,π), and having $\Phi_H = \pi$. In Fig. 6a we show the case in which $\Delta_1\Delta_2\Delta_3 > 0$: according to DMT Φ is expected to be zero while its true value is π. In Fig. 6b the case $\Delta_1\Delta_2\Delta_3 < 0$ is depicted: according to (IV.1) Φ is expected to be π while the true value is 0. In both the cases the AT successfully phase the three reflexions while DMT fail.

Let us now consider the so called "*cross-over*" case: despite the condition $\Delta > 0$ the sign of F_p is reversal with respect to F_H, and $\cos\left(\psi_d - \phi_p\right) < 0$. Triplets with one or more reflexions showing cross-over cannot be correctly estimated by (IV.1) even if $\Phi_H = 0$.

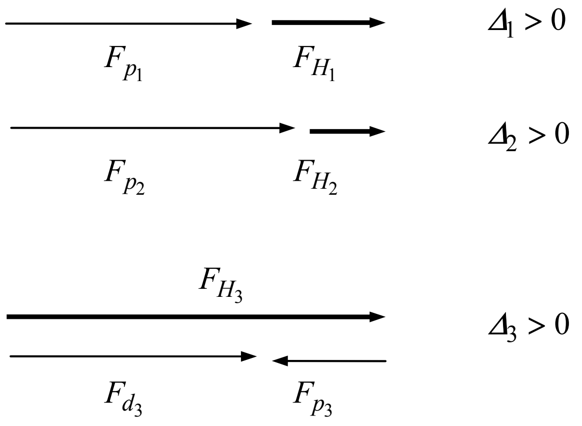


Fig.7: Triplet with phase values symmetry restricted to (0,π) having $\Phi_H = 0$. The third reflexion shows a "*cross-over*".

As an example, let us consider (see Fig. 7) the case in which: (a) the three reflexions forming the triplet have symmetry restricted phase values to (0,π); (b) the Δ_i, *i*=1, 2, 3 are all positive; (c) the third reflexion shows a cross-over. Then (IV.1) will always estimate $\Phi \approx 0$ while, correctly, AT would provide

$$\phi_{p_1} \approx \phi_{H_1} \tag{IV.9a)}$$

$$\phi_{p_2} \approx \phi_{H_2} \tag{IV.9b}$$

$$\phi_{p_3} \approx \phi_{H_3} + \pi \tag{IV.9c)}$$

from which, by summation,

$$\Phi \approx \Phi_H + \pi = \pi \, . \tag{IV.10)}$$

What can we conclude about the relative reliability of DMT and AT from the above examples? Unlike DMT, AT require the prior information on the heavy atom structure. Extracting it from the experimental data requires a preliminary work, but, once available, the knowledge of F_H constitutes a valuable supplementary information. Thus, if we focuse our attention on a single triplet there is no doubt that the DMT estimate of Φ, derived in absence of any information on the heavy-atom structure, is equivalent or inferior to the AT estimate. However DMT have a basic advantage: reflexions are phased by combining the indications provided by hundred, and often by thounsand triplets. This cooperative action can drive phases towards reliable values, so compensating the lack of information on the heavy-atom substructure and generating the supplemental advantage of DMT. Furthermore, direct methods algorithms, so efficient for the crystal structure solution of small molecules, can be used to automatize the entire phasing procedure. In conclusion, AT and DMT are expected to provide highly correlated electron density maps of similar quality. Once the information on the heavy atom structure becomes available it may be hoped that such a supplemental source of information will improve the phase prediction by DMT.

About the quality of the electron density maps

For some of the test structures the mean-phase error is sufficiently low to suggest that the electron density maps could be directly interpreted. We discuss here the quality of the various maps calculated at the end of our procedure.

APP - Fig. 4 suggests high quality experimental data and good isomorphism between native and derivative. The map resulted easily interpretable.

BPO - Fig. 4 indicates that experimental data are of high quality, with good isomorphism between native and derivative. The structure was solved *via* two derivatives (Au and Pt): the heavy-atom positions were identified from difference Patterson maps. The native Fourier maps, calculated for the resolution range 10.0 to 2.8Å from MIR phases was interpretable for both derivatives and showed clear separation of solvent and protein regions.

E2 - Fig. 4 suggests very high quality of the experimental data and very good isomorphism between native and derivative. The structure was originally solved *via* one Hg and two Pt derivatives: in particular the Hg derivative was of excellent quality and has been used in the tests here described. The SIR map is inmediately interpretable.

M-FABP - Fig. 4 indicates a relatively good isomorphism between native and derivative. The structure was solved using both multiple isomorphous replacement (Hg, Pt) and molecular replacement procedures. Our map is close to be interpretable.

NOX - Fig. 4 indicates a quite bad isomorphism between native and derivative. The structure was solved by using multiple isomorphous replacement techniques (5 derivatives). The P6 cis-Pt derivative is used in our calculations. The map if for from beeing interpretable.

Conclusions

This paper is the conclusion of a series which, starting from the theoretical results obtained by Hauptman (1982a,b) and by Giacovazzo, Cascarano & Zheng (1988) has been devoted to make those achievements practicable. The phasing procedure designed throughtout this series can be commented as follows: without any information on the heavy-atom positions the phasing process is able to provide in favourable cases electron density maps which nay be directly interpreted. The process is able in principle to phase all the reflexions up to the derivative resolution and may be accomplished in a fully automatic way, so adding appeal to the method. Too bad isomorphism between the native and the derivative hinder the success. However the practical tests here described suggest that our method has a reserve of power with with respect to standard SIR methods.

References

Giacovazzo, C., Cascarano, G. & Zheng, C.-D. (1988). *Acta Cryst.* A**44**, 45-51.

Giacovazzo, C. & Gonzalez-Platas, J. (1995). *Acta Cryst.* A**51**, 398-404.

Giacovazzo, C., Siliqi, D. & Gonzalez-Platas, J. (1995). *Acta Cryst.* A**51**, 000-000.

Giacovazzo, C., Siliqi, D., Gonzalez-Platas, J., Hecht, H., Zanotti, G., Krauss, N & York, B. (1995). *In preparation.*

Giacovazzo, C., Siliqi, D. & Ralph, A. (1994). *Acta Cryst.* A**50**, 503-510.

Giacovazzo, C., Siliqi, D. & Spagna, R. (1994). *Acta Cryst.* A**50**, 609-621.

Giacovazzo, C., Siliqi, D. & Zanotti, G. (1995). *Acta Cryst.* A**51**, 177-188.

Glover, I., Haneef, I., Pitts, J., Woods, S., Moss, D., Tickle, I. & Blundell, T. L. (1983). *Biopolymers*, **22**, 293-304.

Hauptman, H. (1982a). *Acta Cryst.* A**38**, 289-294.

Hauptman, H. (1982b). *Acta Cryst.* A**38**, 632-641.

Hecht, H., Erdmann, H., Park, H., Sprinzl, M., Schmid, R. D. & Schomburg, D. (1993). *Acta Cryst.* A**49**, *Suppl*. 86.

Hecht, H., Sobek, H., Haag, T., Pfeifer, O. & Van Pee, K. H. (1994). *Nature Struct. Biol.* **1**, 532-537.

Lunin, V., Y. & Woolfson, M., M. (1993). *Acta Cryst.* D**49**, 530-533.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., De Kok, A. & Hol, W. G. J. (1992). *Science*, **255**, 1544-1550.

Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem.* **267**, 18541-18550.

Chapter V

The use of the partial structure information in ab-initio solution of Proteins by Direct Methods

Introduction

According to the tangent formula (Karle & Hauptman, 1956),

$$\tan\theta_{\mathbf{h}} = \frac{\sum_{j=1}^{r} G_j \sin(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}_j})}{\sum_{j=1}^{r} G_j \cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h}-\mathbf{k}_j})} = \frac{T_{\mathbf{h}}}{B_{\mathbf{h}}} \tag{V.1}$$

$\theta_{\mathbf{h}}$ is the most probable value of $\phi_{\mathbf{h}}$. Its reliability depends on the concentration parameter

$$\alpha_{\mathbf{h}} = \left(T_{\mathbf{h}}^2 + B_{\mathbf{h}}^2\right)^{1/2} \tag{V.2}$$

Relationship (V.1) has practically solved the phase problem for small molecules. Its application to two small proteins (i.e., APP, a 36-residue hormone, and rubredoxin from Desulfovibrio vulgaris), both with data up to atomic resolution, attained a notable succes. Relation (V.1) is strictly connected with Sayre (1952) equation

$$E_{\mathbf{h}} = \frac{1}{L} \sum_{\mathbf{k}} E_{\mathbf{k}} \, E_{\mathbf{h-k}} \qquad \text{(V.3)}$$

which, with respect to (V.1), imposes additional restraints on the moduli of the structure factors.

Specific reasons make difficult the application of (V.1) to proteins of usual size: a) the flattness of the probability distribution $\mathrm{P}(\Phi)$; b) the limited data resolution; c) the difficulty in finding the correct phase set, if obtained, among the various trial solutions.

The problem has been reconsidered by Giacovazzo, Guagliardi, Ravelli & Siliqi (1994). Their results may be so summarized:

a) in absence of phase information

$$z_{\mathbf{h}} = < \alpha_{\mathbf{h}} > / \sigma_{\alpha_{\mathbf{h}}} \qquad \text{(V.4)}$$

may be considered as the ratio “signal-to noise”. $< \alpha_{\mathbf{h}} >$ is the expected value of $\alpha_{\mathbf{h}}$, given by

$$< \alpha_{\mathbf{h}} > = \sum_{j=1}^{r} G_j D_1(G_j) \qquad \text{(V.5)}$$

and $\sigma^2_{\alpha_{\mathbf{h}}}$ is the variance of $\alpha_{\mathbf{h}}$, given by (Cascarano, Giacovazzo, Burla, Nunzi & Polidori, 1984)

$$\sigma^2_{\alpha_{\mathbf{h}}} = \frac{1}{2} \sum_{j=1}^{r} G_j^2 \left[1 + D_2(G_j) - 2\, D_1^2(G_j) \right] \qquad \text{(V.6)}$$

b) the statistical solvability criterion was formulated according to which (V.1) can be successfully applied to a given set of diffraction data if the relation

$$z \geq \mathrm{T_r} \qquad \text{(V.7)}$$

is satisfied by a sufficiently high percentage of large normalized structure factors. $\mathrm{T_r}$ represents an acceptable inferior limit for the "signal-to-noise" ratio (say $\mathrm{T_r} \cong 3$).

c) For proteins of usual size $z \leq \mathrm{T_r}$ for a large percentage of reflexions. In these conditions Sayre relation is not satisfied, the application of the tangent formula does not succeed and the correct solution cannot be recognized among the others.

When the diffraction data of one isomorphous derivative are available one can use a mathematical technique (Hauptman, 1982) that integrates direct methods and isomorphous replacement techniques. Then then triplet reliability parameter G is replaced by (Giacovazzo, Cascarano & Zheng Chao-de, 1988)

$$A = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_p R_{\mathbf{h}} R_{\mathbf{k}} R_{\mathbf{h-k}} + 2\left[\sigma_3 / \sigma_2^{3/2}\right]_H \Delta'_{\mathbf{h}} \Delta'_{\mathbf{k}} \Delta'_{\mathbf{h-k}} \;.$$

The α parameter is consequently modified into

$$< \alpha_{\mathbf{h}} >= \sum_{j=1}^{r} A_j D_1(A_j)$$

When A is used, the condition (V.7) is satisfied by a sufficiently high percentage of large normalized structure factors. This suggested that *ab-initio* direct solution of proteins is feasible when diffraction data from one isomorphous derivative are additionally available.

Some drawbacks still limit the usefulness of the procedure for the *ab-initio* phasing of the protein in order to provide an interpretable electron density maps. These are:

1. the quality of the final phased depends on the quality of the atom derivative.
2. Even if the number of phases reflexions is sufficiently large for several practical purposes a non negligible number of reflections with $|\Delta| \cong 0$ but large R values remain unphased. Their contribution to the electron density map is therefore lost.
3. The overall phase error is moderately large: its reduction should provide a better definition of the protein envelop.
4. No method is suggested for extending phases beyond the derivative resolution.
5. Pseudo-centrosymmetrical phases are obtained in specific space groups.

We wants to check the feasibility of a phasing method which exploits as prior information the electron density map eventually available after the application of the techniques involving isomorphous derivative data. Therefore, in order to better understand the present theory we will briefly introduce some general aspects of the theories that recover the entire structure from a partial one.

Recovery of the full structure from a partial one: current methods

We quote the various methods which are currently used for completing the crystal structure from a fragment.

Weighted Fourier syntheses

Woolfson (1956) and Sim (1960) (see also Main, 1979) suggested that the use of Fourier syntheses with weighted terms $W\,|F|\exp(\mathrm{i}\phi_\pi)$ would reveal the unknown atomic positions better than the usual syntheses with $|F|\exp(\mathrm{i}\phi_\pi)$. These and related Fourier methods use essentially the information contained in the distributions

$$\mathrm{P}(\phi_{\mathbf{h}}\,|R_{\mathbf{h}},R_{\pi,\mathbf{h}},\phi_{\pi,\mathbf{h}}) \cong \frac{1}{2\pi\ I_o(G)}\exp\left[G\cos(\phi_{\mathbf{h}}-\phi_{\pi,\mathbf{h}})\right] \qquad \text{(V.8)}$$

where $G = 2\,R_{\mathbf{h}}R_{\pi,\mathbf{h}}/b_{\mathbf{h}}$ and $b_{\mathbf{h}} = \Sigma_q \Big/ \left[|F_{\pi,\mathbf{h}}|^2 + \Sigma_q\right]$. This equation is a von Mises distribution: $\phi_{\pi,\mathbf{h}}$ is the expected value of $\phi_{\mathbf{h}}$ and the reliability of the distribution $\phi_{\mathbf{h}} \cong \phi_{\pi,\mathbf{h}}$ increases with G.

Tangent recycling methods

According to Karle (1970) (see also Hull & Irwin, 1978), a phase ϕ_π is accepted if $|F_\pi| > \eta\,|F|$, where η is the fraction of the total scattering power contained in the fragment and where $|F|$ is associated with an $|E| \geq 1.5$. This approach emplirically exploits the same distribution as in the above case, because it aims to select high products $|E_\pi\ E|$. Tangent recycling uses a large starting set of ϕ_π's in order to compensate for wrong estimations. In each tangent cycle the *a priori* structural information is only used for defining a good starting set.

Tangent recycling methods applied to difference structure factors

In the procedure proposed by Beurskens, Prick, Doesburg & Gould (DIRDIF, 1979) difference structure factors $\Delta\ F = \left(|F| - |F_\pi|\right)\ \exp\,(\mathrm{i}\phi_\pi)$ are calculated and, in favourable cases, accepted for a first estimation of F_q. Weighted tangent formula is applied to the ΔF values in order to convert them to more probable F_q values. It may be observed: (a) The difference structure factors are used in the tangent formula instead of the structure factors: the information about the correlation between E and E_π is obtained by suitable statistical criteria based on (V.8); (b) unfortunately the difference structure factors are unable to use the true F_q's and strict theoretical distributions involving, E's, E_π's, ϕ's, ϕ_π's simultaneously.

Joint probability distribution methods

The prior information is exploited in order to obtain more accurate probabilistic formulas estimating Φ. Main (1976) generalized Cochran's (1955) formula for the phase probability of a triplet in order to exploit some *a priori* knowledge about the structure. He considered different kinds of information: (a) randomly positioned atoms; (b) randomly positioned and randomly oriented atomic groups; (c). randomly positioned but correctly oriented atomic groups; (d) correctly positioned atoms.

A mathematical derivation of Main's formula was given by Heinerman (1977). In his formulation the normalized structure factor $E_\mathbf{h}$ is defined

$$E_\mathbf{h} = \frac{F_\mathbf{h}}{<|F_\mathbf{h}|^2>^{1/2}_{\text{p.r.v}}} \tag{V.9}$$

where $<|F_\mathbf{h}|^2>_{\text{p.r.v}}$ denotes the average of $|F_\mathbf{h}|^2$, the variables being the primitive random variables.

If a group of p atoms is assumed to be correctly positioned then $q = N - p$ atomic positions are the primitive random variables. Then (V.9) may be written as

$$E_\mathbf{h} = \frac{F_\mathbf{h}}{\left[|F_{\pi,\mathbf{h}}|^2 + \Sigma_q\right]^{1/2}} \tag{V.10}$$

When we know the position of p atoms, Main's formula reduces to

$$\mathrm{P}(\Phi\,|\,R_\mathbf{h}, R_\mathbf{k}, R_{\mathbf{h-k}}) \cong \frac{1}{2\pi\, I_o(Q)} \exp\left[Q\cos(\Phi - q)\right] \tag{V.11}$$

where

$$Q\exp(\mathrm{i}\,q) = 2\,R_\mathbf{h} R_\mathbf{k} R_{\mathbf{h-k}} (R_{\pi,\mathbf{h}} R_{\pi,\mathbf{k}} R_{\pi,\mathbf{h-k}} \exp \mathrm{i}\phi_\pi + c)$$

$$c = \sum_{j=p+1}^{N} t_{j,\mathbf{h}} t_{j,\mathbf{k}} t_{j,\mathbf{h\text{-}k}}$$

$$t_{j,\mathbf{h}} = \frac{f_j}{\left[|F_{\pi,\mathbf{h}}|^2 + \Sigma_q\right]^{1/2}}$$

Heinerman observed that (V.11) was not quite satisfactory in the practice and concluded that only high-order terms of the distributions could improve the accuracy of the formula.

New insight into this method was afforded by Giacovazzo (1983). He considered that: (a) the atomic positions are assumed to be the random variables; (b) any normalized structure factor $E_{\mathbf{h}}$ is considered as the sum of a fixed term $E_{\pi,\mathbf{h}}$ arising from the atoms with known positions and of a random term $E_{q,\mathbf{h}}$ arising from the atoms with unknown positions. Then

$$E_{\pi,\mathbf{h}} = \frac{F_{\pi,\mathbf{h}}}{\left[|F_{\pi,\mathbf{h}}|^2 + \Sigma_q\right]^{1/2}} \qquad E_{q,\mathbf{h}} = \frac{F_{q,\mathbf{h}}}{\left[|F_{\pi,\mathbf{h}}|^2 + \Sigma_q\right]^{1/2}}$$

For the sake of simplicity we will denote $E_{\pi,\mathbf{h}}$ and $E_{q,\mathbf{h}}$ as pseudo-normalized structure factors.

The basic result of this work was the conditional joint probability (we have replaced the indices **h**, **k**, **h-k** by 1, 2, 3 respectively)

$$\begin{aligned}
&\mathrm{P}(R_1, R_2, R_3, \phi_1, \phi_2, \phi_3 \mid R_{\pi,1}, R_{\pi,2}, R_{\pi,3}, \phi_{\pi,1}, \phi_{\pi,2}, \phi_{\pi,3}) \\
&\cong (\pi)^{-3} (b_1 b_2 b_3)^{-1}\, R_1 R_2 R_3 \exp\Bigg\{ -\sum_{i=1}^{3} \frac{1}{b_i} \Big[R_i^2 + R_{\pi,i}^2 \\
&-2\, R_i R_{\pi,i} \cos(\phi_i - \phi_{\pi,i}) \Big] \\
&+ \frac{2c}{b_1 b_2 b_3} \Big[R_1 R_2 R_3 \cos(\phi_1 + \phi_2 + \phi_3) - \\
&- R_{\pi,1} R_{\pi,2} R_{\pi,3} \cos(\phi_{\pi,1} + \phi_{\pi,2} + \phi_{\pi,3}) \\
&- R_{\pi,1} R_2 R_3 \cos(\phi_{\pi,1} + \phi_2 + \phi_3) \\
&- R_1 R_{\pi,2} R_3 \cos(\phi_1 + \phi_{\pi,2} + \phi_3) \\
&- R_1 R_2 R_{\pi,3} \cos(\phi_1 + \phi_2 + \phi_{\pi,3}) \\
&+ R_{\pi,1} R_{\pi,2} R_3 \cos(\phi_{\pi,1} + \phi_{\pi,2} + \phi_3) \\
&+ R_{\pi,1} R_2 R_{\pi,3} \cos(\phi_{\pi,1} + \phi_2 + \phi_{\pi,3}) \\
&+ R_1 R_{\pi,2} R_{\pi,3} \cos(\phi_1 + \phi_{\pi,2} + \phi_{\pi,3}) \Big] \Bigg\}
\end{aligned} \qquad \text{(V.12)}$$

It coincides with the classical trivariate distribution of Karle & Hauptman when $p = 0$.

From (V.12) one can obtain:

$$\mathrm{P}(\phi_{\mathbf{h}} \mid \ldots) \cong \left[2\pi\; I_o(\alpha) \right]^{-1} \exp\left[\alpha\, \cos(\phi_{\mathbf{h}} - \xi) \right] \qquad \text{(V.13)}$$

where $\alpha^2 = \alpha_1^2 + \alpha_2^2$,

$$\begin{aligned}
\alpha_1 = 2\, R_{\mathbf{h}} \Bigg\{ & \left(R_{\pi,\mathbf{h}} / b_{\mathbf{h}} \right) \cos\phi_{\pi,\mathbf{h}} + \sum_{\text{triplets}} \frac{c}{b_{\mathbf{h}} b_{\mathbf{k}} b_{\mathbf{h-k}}} \\
\text{x} \quad & \Big[R_{\mathbf{k}} R_{\mathbf{h-k}} \cos(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}}) - R_{\pi,\mathbf{k}} R_{\mathbf{h-k}} \cos(\phi_{\pi,\mathbf{k}} + \phi_{\mathbf{h-k}}) \\
- \quad & R_{\mathbf{k}} R_{\pi,\mathbf{h-k}} \cos(\phi_{\mathbf{k}} + \phi_{\pi,\mathbf{h-k}}) + R_{\pi,\mathbf{k}} R_{\pi,\mathbf{h-k}} \cos(\phi_{\pi,\mathbf{k}} + \phi_{\pi,\mathbf{h-k}}) \Big] \Bigg\}
\end{aligned}$$

$$\alpha_2 = 2\,R_{\mathbf{h}}\Big\{\left(R_{\pi,\mathbf{h}}/b_{\mathbf{h}}\right)\sin\phi_{\pi,\mathbf{h}} + \sum_{\text{triplets}} \frac{c}{b_{\mathbf{h}}b_{\mathbf{k}}b_{\mathbf{h-k}}}$$

$$\text{x} \quad \left[-R_{\mathbf{k}}R_{\mathbf{h-k}}\sin(\phi_{\mathbf{k}}+\phi_{\mathbf{h-k}}) + R_{\pi,\mathbf{k}}R_{\mathbf{h-k}}\sin(\phi_{\pi,\mathbf{k}}+\phi_{\mathbf{h-k}})\right.$$

$$+ \quad \left. R_{\mathbf{k}}R_{\pi,\mathbf{h-k}}\sin(\phi_{\mathbf{k}}+\phi_{\pi,\mathbf{h-k}}) - R_{\pi,\mathbf{k}}R_{\pi,\mathbf{h-k}}\sin(\phi_{\pi,\mathbf{k}}+\phi_{\pi,\mathbf{h-k}})\right]\Big\}$$

$$\cos\xi = \alpha_1/\alpha \qquad \sin\xi = \alpha_2/\alpha\ .$$

From (V.13) the special tangent formula (V.14) arises:

$$\tan\phi_{\mathbf{h}} \cong \alpha_2/\alpha_1 \tag{V.14}$$

The distribution (V.13) can be simplified if we replace $E_{\mathbf{h}}$ and $E_{\pi,\mathbf{h}}$ by the psuedo-normalized structure factors

$$E''_{\mathbf{h}} = F_{\mathbf{h}}\Big/\Sigma_q^{1/2} \qquad , \qquad E''_{\pi,\mathbf{h}} = F_{\pi,\mathbf{h}}\Big/\Sigma_q^{1/2}\,.$$

Then we can rewrite (V.13) as

$$\mathrm{P}(\phi_{\mathbf{h}}\,|\ldots) \cong \left[2\pi\ I_o(\alpha)\right]^{-1}\exp\left[\alpha\ \cos(\phi_{\mathbf{h}}-\theta_{\mathbf{h}})\right] \tag{V.15}$$

where

$$\alpha_{\mathbf{h}}^2 = \alpha_1''^2 + \alpha_2''^2$$

$$\begin{aligned}\alpha_1'' = {} & 2\,R_{\mathbf{h}}''\Big\{R_{\pi,\mathbf{h}}''\cos\phi_{\pi,\mathbf{h}} \\ & + q^{-1/2}\sum_{\mathbf{k}}\Big[R_{\mathbf{k}}''R_{\mathbf{h-k}}''\cos(\phi_{\mathbf{k}}+\phi_{\mathbf{h-k}}) \\ & - R_{\pi,\mathbf{k}}''R_{\mathbf{h-k}}''\cos(\phi_{\pi,\mathbf{k}}+\phi_{\mathbf{h-k}}) - R_{\mathbf{k}}''R_{\pi,\mathbf{h-k}}''\cos(\phi_{\mathbf{k}}+\phi_{\pi,\mathbf{h-k}}) \\ & + R_{\pi,\mathbf{k}}''R_{\pi,\mathbf{h-k}}''\cos(\phi_{\pi,\mathbf{k}}+\phi_{\pi,\mathbf{h-k}})\Big]\Big\} \\ = {} & 2\,R_{\mathbf{h}}''\left\{\mathcal{R}\left[E_{\pi,\mathbf{h}}''+\mathrm{q}^{-1/2}\sum_{\mathbf{k}}\left(E_{\mathbf{k}}''-E_{\pi,\mathbf{k}}''\right)\left(E_{\mathbf{h-k}}''-E_{\pi,\mathbf{h-k}}''\right)\right]\right\}\end{aligned} \tag{V.16}$$

$$\begin{aligned}\alpha_2' = {} & 2\,R_{\mathbf{h}}''\Big\{R_{\pi,\mathbf{h}}''\sin\phi_{\pi,\mathbf{h}} \\ & + q^{-1/2}\sum_{\mathbf{k}}\Big[R_{\mathbf{k}}''R_{\mathbf{h-k}}''\sin(\phi_{\mathbf{k}}+\phi_{\mathbf{h-k}}) \\ & - R_{\pi,\mathbf{k}}''R_{\mathbf{h-k}}''\sin(\phi_{\pi,\mathbf{k}}+\phi_{\mathbf{h-k}}) - R_{\mathbf{k}}''R_{\pi,\mathbf{h-k}}''\sin(\phi_{\mathbf{k}}+\phi_{\pi,\mathbf{h-k}}) \\ & + R_{\pi,\mathbf{k}}''R_{\pi,\mathbf{h-k}}''\sin(\phi_{\pi,\mathbf{k}}+\phi_{\pi,\mathbf{h-k}})\Big]\Big\} \\ = {} & 2\,R_{\mathbf{h}}''\left\{I\left[E_{\pi,\mathbf{h}}''+\mathrm{q}^{-1/2}\sum_{\mathbf{k}}\left(E_{\mathbf{k}}''-E_{\pi,\mathbf{k}}''\right)\left(E_{\mathbf{h-k}}''-E_{\pi,\mathbf{h-k}}''\right)\right]\right\}\end{aligned} \tag{V.17}$$

and

$$\tan\theta_{\mathbf{h}} = \alpha_2''/\alpha_1'' \tag{V.18}$$

In (V.16) and (V.17) $\mathscr{R}$ and $\mathscr{I}$ stand for *'real part of'* and *'imaginary part of'* respectively. This form of the distribution allows understanding the next properties:

a) the new formulas are now easily applicable;

b) The triplet contribution is of order $q^{-1/2}$ (it's replaced by $\left[\sigma_3/\sigma_2^{3/2}\right]_q$ if atoms are not equal). It's worthwhile observing that q is just the number of independent random variables in our problem in analogy with N in the Cochran's formula.

c) $E_{\mathbf{h}}''$ and $E_{\pi,\mathbf{h}}''$ are not normalized structure factors: indeed

$$<|E''|^2>=1+\Sigma_p/\Sigma_q \qquad <|E''_\pi|^2>=\Sigma_p/\Sigma_q$$

so that $<|E''|^2>$ is always larger than unit, while $<|E''_\pi|^2>$ is larger or smaller than unity according to whether Σ_p is larger or smaller than Σ_q. The larger the *a priori* information, the smaller $<|E'-E'_\pi|^2>$ is.

d) Equation (V.15) reduces to Sim's (1959) formula if triplet contribution is not taken into account.

e) The best estimate for $\phi_\mathbf{h}$, i.e. $\theta_\mathbf{h}$, is the phase of the complex vector

$$E''_{\pi,\mathbf{h}}+q^{-1/2}\sum_\mathbf{k}\left(E''_\mathbf{k}-E''_{\pi,\mathbf{k}}\right)\left(E''_\mathbf{h-k}-E''_{\pi,\mathbf{h-k}}\right). \qquad \text{(V.19)}$$

The larger its modulus α is, the larger the expected accuracy of the estimation. According to (V.19), the vectorial differences $(E''_\mathbf{k}-E''_{\pi,\mathbf{k}})$ and $(E''_\mathbf{h-k}-E''_{\pi,\mathbf{h-k}})$ do influence the value $\theta_\mathbf{h}$.

f) If $p\rightarrow 0$, then $q\rightarrow N$, $E''_\pi\rightarrow 0$ and (V.15) reduces to the classical Karle & Karke's (1966) relationships.

g) The larger is p, the less important in the average is the triplet contribution compared with Sim's contribution. In particular, because of point (c), the triplet contribution vanishes when $q\rightarrow 0$. Then $R''_\mathbf{h}$ and $R''_{\pi,\mathbf{h}}$ are both infinite and, (V.15) approximates the Dirac $\delta(\phi_\mathbf{h}-\phi_{\pi,\mathbf{h}})$.

h) From the above considerations, the following probabilistic relation is suggested:

$$E''_\mathbf{h}\cong E''_{\pi,\mathbf{h}}+q^{-1/2}\sum_\mathbf{k}(E''_\mathbf{k}-\mathrm{E}''_{\pi,\mathbf{k}})(E''_\mathbf{h-k}-\mathrm{E}''_{\pi,\mathbf{h-k}}) \qquad \text{(V.20)}$$

which may be considered as a generalized Sayre's (1952) equation emphasizing the fact that part of the structure is known. When $q=0$ the equation (V.20) reduces to the trivial identity $E''_\mathbf{h}=E''_{\pi,\mathbf{h}}=\infty$. When $q=N$ then $E''_\pi=0$ and (V.20) reduces to the classical Sayre's equation. When $q\neq 0$, the prior information introduces new algebraic and probabilistic constraints so as to recentre $E''_\mathbf{h}$ around $E''_{\pi,\mathbf{h}}$.

Direct methods for high-resolution phase refinement for proteins

While the *ab-initio* solution of protein structures is not within the capacity of traditional direct methods, their efficiency for phase refinement and extension is still under discussion. Since the first trials by Reeke and Lipscomb (1969), Weinzierl, Einsenberg & Dickerson (1969) and Coulter (1971), it was clear that a characteristic feature of the tangent formula is the following: a possible moderate improvement of the phases is frequently followed, after few cycles of refinement, by their deterioration. Phases diverge to a self-consistent incorrect set. The application of Sayre equation proved more stable even if much more time consuming: therefore some programs (for example, SAYTAN, see Woolfson, 1993) introduce Sayre's formula restraints in the tangent formula framework. A more general approach has been followed by Main (1990): the electron density map is improved by combining information from real and reciprocal spaces. The solution of large non-linear systems, as required by Sayre equation, is circumvented by the use of the conjugate-gradient method to calculate shifts of the electron density map. The information so obtained is combined (Cowtan & Main, 1993) with solvent flattening techniques (Wang, 1985), histogram matching (see Lunin, 1993), non-crystallographic symmetric averaging (Bricogne, 1974), and the use of a partial structure according to the method of Read (1986). The applications of such a method to practical cases show that the improvement of the electron density map is a product of the simultaneous use of the different techniques.

A different point of view may be introduced. Let us suppose that the phase estimates (for example, by isomorphous derivative techniques) are available for a subset of reflexions and that the calculated electron density map is able to reveal the main features of the structure. The map may be supposed not interpretable in terms of chain tracing but showing the general envelope of the molecule. This envelope may be considered as the prior information for the subsequent steps: in particular, its inverse Fourier transform may be calculated and the values F_π (structure factor of a partial structure) are derived for the various structure factors. Then triplet invariants can be estimated via distribution like (V.12) derived by Giacovazzo (1983), instead than *via*

$$\mathrm{P}\left(\phi_{p,\mathbf{h}}, \phi_{p,\mathbf{k}}, \phi_{p,\mathbf{h-k}} \middle| \, |F_{p,\mathbf{h}}|, |F_{p,\mathbf{k}}|, |F_{p,\mathbf{h-k}}|\right) \tag{V.21}$$

used by the tangent formula. The advantage of (V.12) with respect to the other methods may be so summarized:

a) the electron density map is divided into two regions; the first coincides with the assumed partial structure, the second is "flattened" to zero and gives vanishing contribution to the values of F_π.

b) The distribution can take full advantage of the known partial structure, which on the contrary is neglected in (V.21).
c) The prior information proved to lead , by tangent refinement, to highly accurate estimates of the phases (Camalli, Giacovazzo & Spagna, 1985; Burla, Cascarano, Fares, Giacovazzo, Polidori & Spagna, 1989) at least for small molecule structures.

The problem in now: is the supplementary information provided by (V.12) sufficient for reliably extending and refining the phases of macromolecules? The answer is not easy: the effectiveness of the process depends on the accuracy of the starting phases ϕ_π , on the general correctness of the envelop, on the complexity of the entire structure. We want to explore first the feasibility of the method, by working in an ideal and therefore well controlled situation (Giacovazzo & Gonzalez-Platas, 1995).

The probabilistic formula to apply is the equation (V.20) derived from (V.12): in terms of phases that is equivalent to

$$\tan\theta_{\mathbf{h}} = T_\pi / B_\pi \tag{V.22}$$

where

$$\begin{aligned} T_\pi = 2R''_{\mathbf{h}}\Big\{ & R''_{\pi,\mathbf{h}}\sin\phi_{\pi,\mathbf{h}} + \left[\sigma_3/\sigma_2^{3/2}\right]_q \sum_{\mathbf{k}}\Big[R''_{\mathbf{k}}R''_{\mathbf{h-k}}\sin\left(\phi_{\mathbf{k}}+\phi_{\mathbf{h-k}}\right) \\ & - R''_{\pi,\mathbf{k}}R''_{\mathbf{h-k}}\sin\left(\phi_{\pi,\mathbf{k}}+\phi_{\mathbf{h-k}}\right) - R''_{\mathbf{k}}R''_{\pi,\mathbf{h-k}}\sin\left(\phi_{\mathbf{k}}+\phi_{\pi,\mathbf{h-k}}\right) \\ & + R''_{\pi,\mathbf{k}}R''_{\pi,\mathbf{h-k}}\sin\left(\phi_{\pi,\mathbf{k}}+\phi_{\pi,\mathbf{h-k}}\right)\Big]\Big\} \end{aligned}$$

$$\begin{aligned} B_\pi = 2R''_{\mathbf{h}}\Big\{ & R''_{\pi,\mathbf{h}}\cos\phi_{\pi,\mathbf{h}} + \left[\sigma_3/\sigma_2^{3/2}\right]_q \sum_{\mathbf{k}}\Big[R''_{\mathbf{k}}R''_{\mathbf{h-k}}\cos\left(\phi_{\mathbf{k}}+\phi_{\mathbf{h-k}}\right) \\ & - R''_{\pi,\mathbf{k}}R''_{\mathbf{h-k}}\cos\left(\phi_{\pi,\mathbf{k}}+\phi_{\mathbf{h-k}}\right) - R''_{\mathbf{k}}R''_{\pi,\mathbf{h-k}}\cos\left(\phi_{\mathbf{k}}+\phi_{\pi,\mathbf{h-k}}\right) \\ & + R''_{\pi,\mathbf{k}}R''_{\pi,\mathbf{h-k}}\cos\left(\phi_{\pi,\mathbf{k}}+\phi_{\pi,\mathbf{h-k}}\right)\Big]\Big\} \end{aligned}$$

$\theta_{\mathbf{h}}$ is the most probable value of $\phi_{\mathbf{h}}$, and

$$\alpha_{\pi,\mathbf{h}} = \left(T_\pi^2 + B_\pi^2\right)^{1/2} \tag{V.23}$$

is its reliability parameter.

The statistical z test when a partial structure is available

In order to estimate the efficiency of (V.20) we should calculate, in accordance with (V.4),

$$z_{\pi,\mathbf{h}} =< \alpha_{\pi,\mathbf{h}} > / \sigma_{\alpha_{\pi,\mathbf{h}}} \qquad \text{(V.24)}$$

$< \alpha_{\pi,\mathbf{h}} >$ and $\sigma_{\alpha_{\pi,\mathbf{h}}}$ may be derived according to the following procedure:

1. first we derive, from equation .(V.12), the marginal distribution

$$\mathrm{P}(\phi_{\mathbf{k}}, \phi_{\mathbf{h-k}} \big| R''_{\mathbf{h}}, \mathrm{R}''_{\mathbf{k}}, \mathrm{R}''_{\mathbf{h-k}}, \mathrm{R}''_{\pi,\mathbf{h}}, \mathrm{R}''_{\pi,\mathbf{k}}, \mathrm{R}''_{\pi,\mathbf{h-k}}, \phi_{\pi,\mathbf{h}}, \phi_{\pi,\mathbf{k}}, \phi_{\pi,\mathbf{h-k}})$$

The phase $\phi_{\mathbf{h}}$ is supposed unknown, as it occurs when we are interesed to $< \alpha_{\pi,\mathbf{h}} >$. Neglecting terms of order $\left[\sigma_3 / \sigma_2^{3/2}\right]_q$ we obtain that $\left(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}}\right)$ is distributed according to the von Mises distribution

$$\mathrm{M}(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}}; \phi_{\pi,\mathbf{k}} + \phi_{\pi,\mathbf{h-k}}, q_{1,\mathbf{k}})$$

where $q_{1,\mathbf{k}}$ satisfies the relation

$$D_1(q_{1,\mathbf{k}}) = D_1(2\, R''_{\mathbf{k}}\, R''_{\pi,\mathbf{k}})\, D_1(2\, R''_{\mathbf{h-k}}\, R''_{\pi,\mathbf{h-k}})$$

$(\phi_{\pi,\mathbf{k}} + \phi_{\pi,\mathbf{h-k}})$ is the expected value of $(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}})$ and $q_{1,\mathbf{k}}$ is the concentration parameter of the distribution.

2. In an analogous way we obtain that $(\phi_{\pi,\mathbf{k}} + \phi_{\mathbf{h-k}})$ is distributed according to the Von Mises distribution

$$\mathrm{M}(\phi_{\pi,\mathbf{k}} + \phi_{\mathbf{h-k}}; \phi_{\pi,\mathbf{k}} + \phi_{\pi,\mathbf{h-k}}, q_{2,\mathbf{k}})$$

where

$$q_{2,\mathbf{k}} = 2\, R''_{\mathbf{h-k}}\, R''_{\pi,\mathbf{h-k}}.$$

3. Also $(\phi_{\mathbf{k}} + \phi_{\pi,\mathbf{h-k}})$ is distributed according to

$$\mathrm{M}(\phi_{\mathbf{k}} + \phi_{\pi,\mathbf{h-k}};\phi_{\pi,\mathbf{k}} + \phi_{\pi,\mathbf{h-k}},q_{3,\mathbf{k}})$$

   where

$$q_{3,\mathbf{k}} = 2\, R''_{\mathbf{k}} R''_{\pi,\mathbf{k}}.$$

4. We recall that the distribution of the modulus α of the resultant of r complex vectors $Q_j \exp(\mathrm{i}\,\nu_j)$ under the hypothesis that Q_j are distrubuted according to the von Mises distribution $\mathrm{M}(\nu_j;\theta,q_j)$ is the normal distribution (Cascarano, Giacovazzo & Guagliardi, 1992) $\mathrm{N}(\alpha;<\alpha>,\sigma^2)$ where

$$<\alpha>=\sum_{j=1}^{r} Q_j D_1(q_j)$$

$$\sigma^2 = \frac{1}{2}\sum_{j=1}^{r} Q_j^2\left(1 + D_2(q_j) - 2\,D_1^2(q_j)\right) \qquad \text{(V.25)}$$

5. We apply the above results to the sets of vectors

$$\begin{aligned}&\sum_{\mathbf{k}} Q_{1,\mathbf{k}} \exp\left[\mathrm{i}(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}})\right],\\ &\sum_{\mathbf{k}} Q_{2,\mathbf{k}} \exp\left[\mathrm{i}(\phi_{\pi,\mathbf{k}} + \phi_{\mathbf{h-k}})\right],\\ &\sum_{\mathbf{k}} Q_{3,\mathbf{k}} \exp\left[\mathrm{i}(\phi_{\mathbf{k}} + \phi_{\pi,\mathbf{h-k}})\right],\end{aligned} \qquad \text{(V.26)}$$

where

$$Q_{1,\mathbf{k}} = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_q R''_{\mathbf{k}} R''_{\mathbf{h-k}}$$

$$Q_{2,\mathbf{k}} = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_q R''_{\pi,\mathbf{k}} R''_{\mathbf{h-k}}$$

$$Q_{3,\mathbf{k}} = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_q R''_{\mathbf{k}} R''_{\pi,\mathbf{h-k}} \,.$$

Then

$$< \alpha_{\pi,\mathbf{h}} >= 2\, R''_{\pi,\mathbf{h}} + \sum_{\mathbf{k}} \left[Q_{1,\mathbf{k}}\, D_1(q_{1,\mathbf{k}}) - Q_{2,\mathbf{k}} D_1(q_{2,\mathbf{k}}) - Q_{3,\mathbf{k}} D_1(q_{3,\mathbf{k}}) + Q_{4,\mathbf{k}} \right] \qquad \text{(V.27)}$$

where

$$Q_{4,\mathbf{k}} = 2\left[\sigma_3 / \sigma_2^{3/2}\right]_q R''_{\pi,\mathbf{k}} R''_{\pi,\mathbf{h-k}}$$

$< \alpha_{\pi,\mathbf{h}} >$ reduces to $<\alpha_{\mathbf{h}} >$ when no partial structure is available.

6. The value of $\sigma^2_{\alpha_{\pi,\mathbf{h}}}$ may be derived by applying in turn (V.25) to the terms (V.26) and then summing the contributions.

The statistical solvability criterion (V.4) have been applied to the experimental data of the structures quoted in Table I (Giacovazzo & Gonzalez-Platas, 1995).

Structure Code	Space Group	Molecular formula	Z
APP[(1)]	C2	$C_{190}\ N_{53}\ O_{58}$ Zn	4
BPTI[(2)]	$P2_1\ 2_1\ 2_1$	$S_9\ O_{149}\ N_{84}\ C_{289}$	4
LYSO[(3)]	$P4_3\ 2_1\ 2$	$S_{10}\ O_{286}\ N_{193}\ C_{613}$	8
MYO[(4)]	$P2_1$	Fe $S_4\ O_{389}\ N_{220}\ C_{817}$	2
M-FABP[(5)]	$P2_1\ 2_1\ 2_1$	$C_{667}\ N_{170}\ O_{261}\ S_3$	4
E2[(6)]	F4 3 2	$C_{1170}\ N_{310}\ O_{366}\ S_7$	96

Table I: Code name, space group and crystallochemical data for test structures.
(1) Glover, Haneef, Pitts, Wood, Moss, Tickle & Blundell (1983).
(2) Data by courtesy of R. Huber, MPI Martinsried, FRG.
(3) Data from courtesy of C. Betzel, ENBL, Hamburg, FRG.
(4) Hartmann, Steigemann, Reuscher & Parak (1987).
(5) Zanotti, Scapin, Spadon, Veerkamp & Sacchettini (1992).
(6) Mattevi, Obmolova, Schulze, Kalk, Westphal, De Kok & Hol (1992)

CODE	RES(Å)	N	NREFL	NLAR	NTRIP
APP	0.99	413	17058	700	8907
BPTI	1.00	1860	17300	700	21759
LYSO	1.69	7720	13622	700	24899
MYO	1.50	2648	15903	1100	26796
M-FABP	2.14	4076	7769	700	23012
E2	3.00	40783	8136	400	23820

Table II: Parameters defining protocol for calculations (see the main text for the symbols).

For each test structure we give in Table II the resolution of the diffraction data (RES), the number of atoms (statistically calculated) in the primitive unit cell (N), the number of measured reflexions (NREFL), the number of large normalized structure factors (NLAR) among which triplet invariants are calculated, the total number of triplets which contribute to the various α values (NTRIP). For each structure we calculated the $z_{\mathbf{h}}$ values corresponding to the NLAR reflexions according to the definition (V.4), and the $z_{\pi,\mathbf{h}}$ values according to the definition (V.24).

Two different amounts of prior information were used for $z_{\pi,\mathbf{h}}$ corresponding to different values of the diffraction ratio $DR_{\pi} = \left[\sigma_2\right]_{\pi} / \left[\sigma_2\right]_{p}$ = 0.20, 0.40. In Figs. 1 - 6 we show the $\mathrm{P}(z)$ and $\mathrm{P}(z_{\pi})$ curves. In general $\mathrm{P}(z)$ curves do not satisfy the statistical solvability criterion: on the contrary the $\mathrm{P}(z_{\pi})$ curves are remarkarbly shifted towards the right and satisfy the criterion.

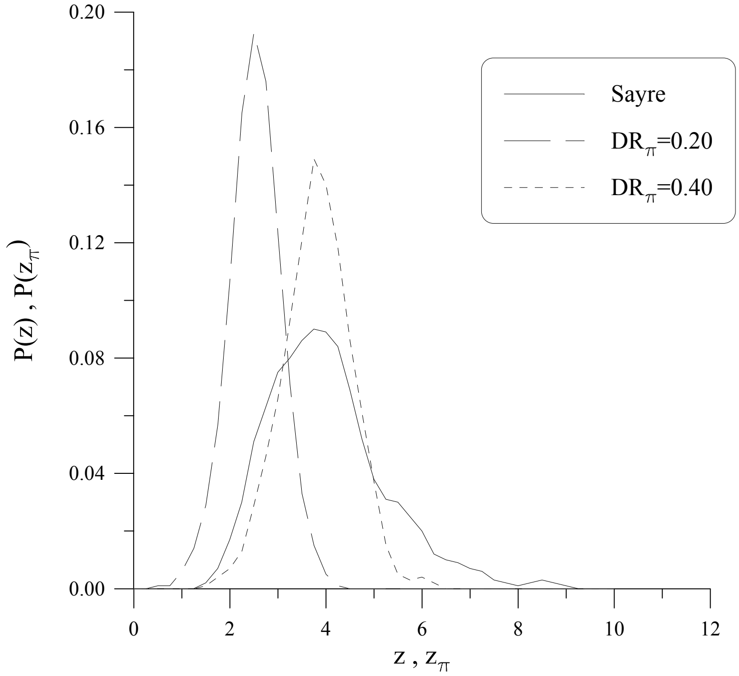


Fig.1: APP: $\mathrm{P}(z)$ and $\mathrm{P}(z_{\pi})$ curves

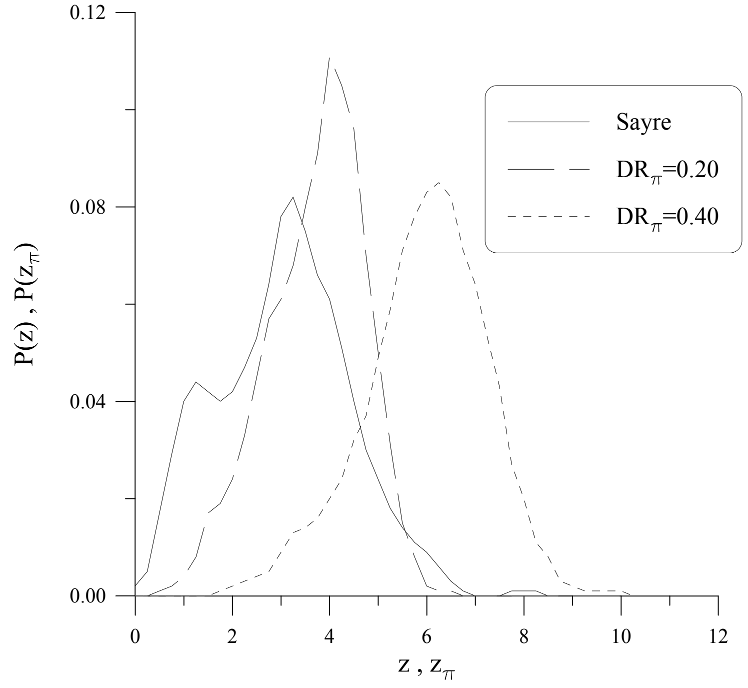


Fig.2: BPTI: $P(z)$ and $P(z_\pi)$ curves

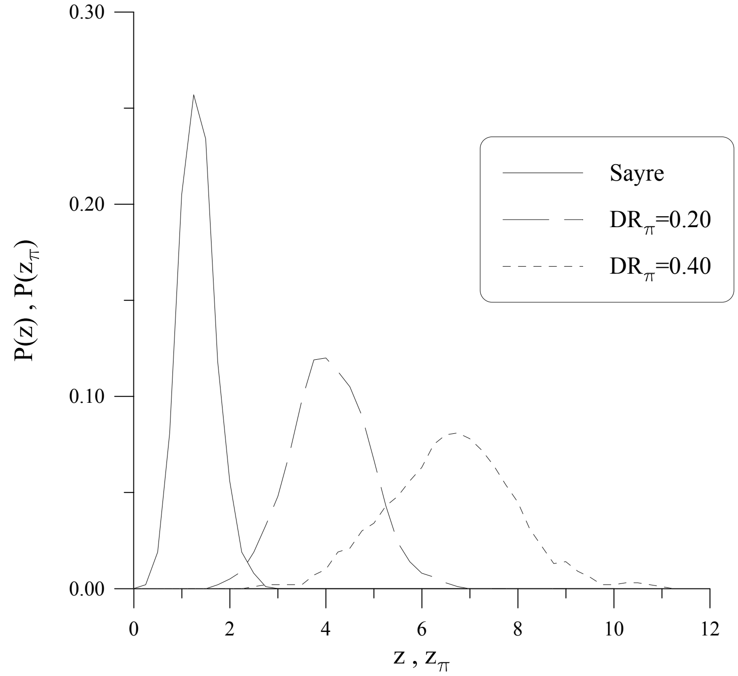


Fig.3: LYSO: $P(z)$ and $P(z_\pi)$ curves

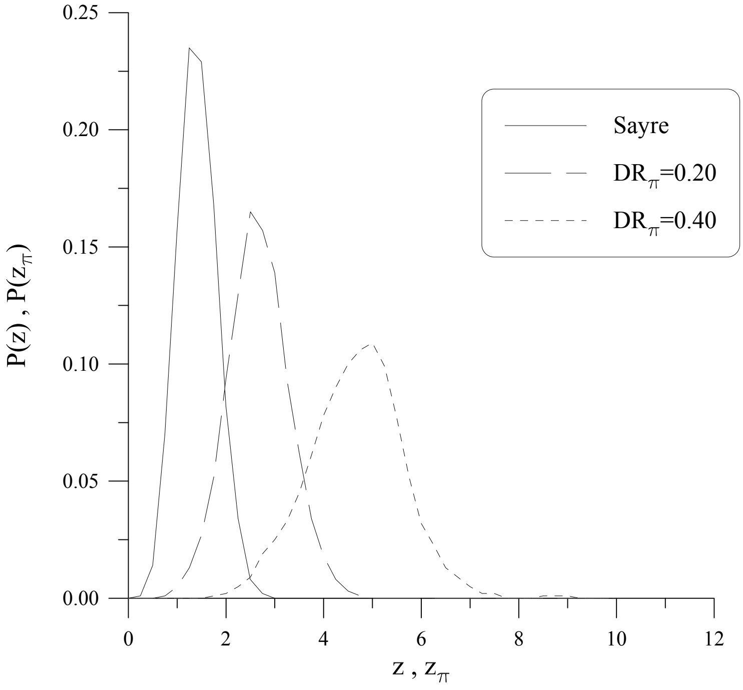


Fig.4: MYO: $P(z)$ and $P(z_\pi)$ curves

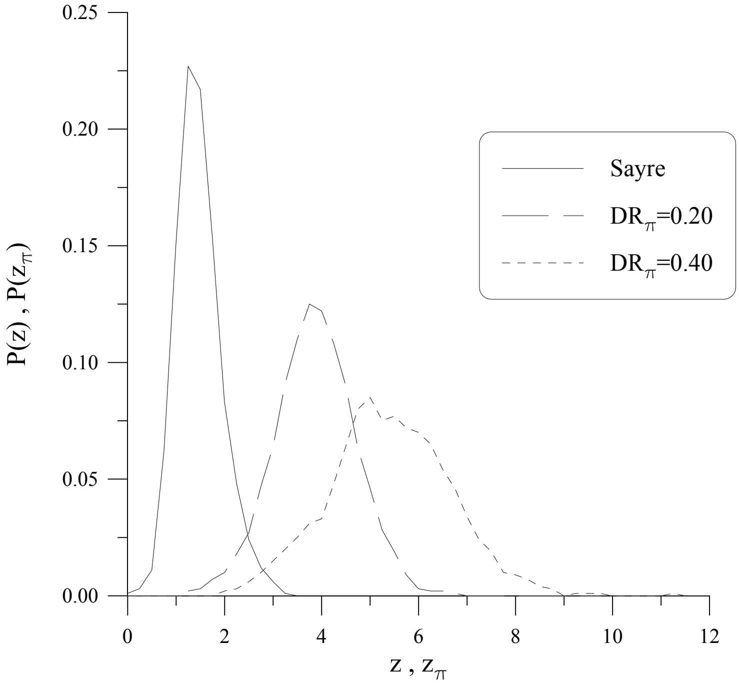


Fig.5: M-FABP: $P(z)$ and $P(z_\pi)$ curves

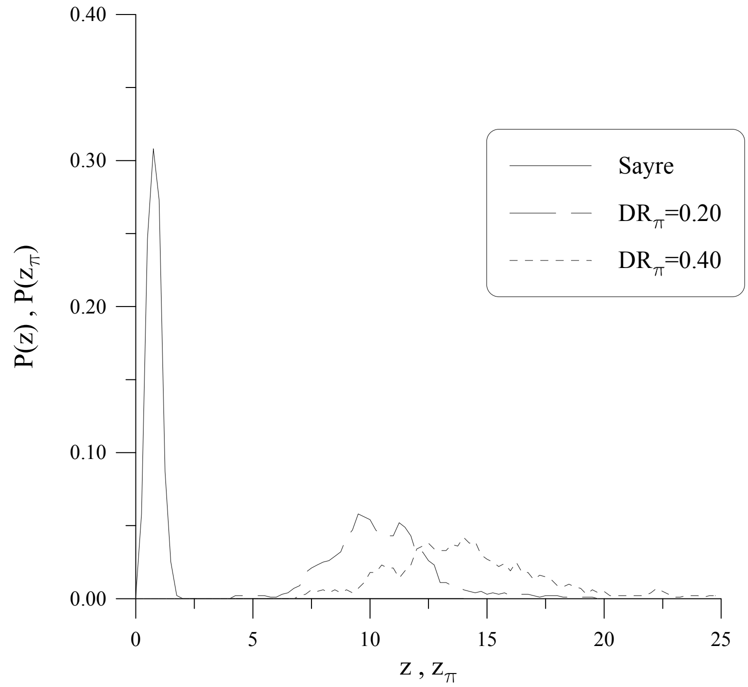


Fig.6: E2: P(z) and P(z$_\pi$) curves

The only exception occurs for APP for which the prior information does not improve the distribution of the ratio "signal-to-noise". At the moment we are unable to explain this discordant result. The shifts relative to the $\mathrm{P}(z_\pi)$ curves increase with the amount of prior information: therefore larger information should make (V.20) more efficient.

In order to collect the above observations in a simple sentence we can so conclude: Sayre relation (V.3) is expected to be violated for all test structures but for APP, while relation (V.20) is expected to be satisfied for all the test structures. In order to check this conclusion we calculated the following figures of merit.

$$\mathrm{FOM} = \frac{\sum_{\mathbf{h}} \left| E_{\mathbf{h}} - E_{\mathbf{h}\,\mathrm{cal}} \right|}{\sum_{\mathbf{h}} \left| E_{\mathbf{h}} \right|} \qquad \mathrm{FOM}_{\pi} = \frac{\sum_{\mathbf{h}} \left| E''_{\mathbf{h}} - E''_{\mathbf{h}\,\mathrm{cal}} \right|}{\sum_{\mathbf{h}} \left| E''_{\mathbf{h}} \right|} \qquad \text{(V.28)}$$

In (V.28) $E_{\mathbf{h}}$ and $E''_{\mathbf{h}}$ stay for $R_{\mathbf{h}}\exp(\mathrm{i}\phi_{\mathbf{h}})$ and $R''_{\mathbf{h}}\exp(\mathrm{i}\phi_{\mathbf{h}})$ respectively; $\phi_{\mathbf{h}}$ is the true phase (derived from the refined published crystal structure). R and R'' are respectively normalized and pseudo-normalized magnitudes derived from measurements.

When FOM is calculated, the Sayre relationship is applied, and $E_{\mathbf{h}\,\mathrm{cal}}$ is obtained from the right-hand side of (V.3) by using true phases $(\phi_{\mathbf{k}} + \phi_{\mathbf{h-k}})$. When FOM_{π} is calculated, the formula (V.20) is applied, and $E''_{\mathbf{h}\mathrm{cal}}$ is obtained from the right-hand side of (V.20) by using the true values $\phi_{\mathbf{k}}, \phi_{\mathbf{h-k}}, \phi_{\pi,\mathbf{h}}, \phi_{\pi,\mathbf{k}}, \phi_{\pi,\mathbf{h-k}}$. Large values of FOM and FOM_{π} involve remarkable deviations (in terms of moduli and phases) of the calculated E's from the observed ones and therefore indicate violation of (V.3) and (V.20). The results are shown in Table III.

CODE	**FOM**	**FOM$_{\pi}$** (DR$_{\pi}$ = 0.20)	**FOM$_{\pi}$** (DR$_{\pi}$ = 0.40)
APP	0.512	0.515	0.490
BPTI	0.783	0.579	0.496
LYSO	0.847	0.696	0.598
MYO	0.800	0.632	0.497
M-FABP	0.863	0.695	0.570
E2	0.910	0.567	0.477

Table III: FOM and FOM$_{\pi}$ values for the test structures.

The FOM values are quite large, thus confirming that the Sayre relation is not satisfied. The FOM_{π}'s are remarkably smaller than the corresponding FOM's: the situation is still improved when DR_{π} increases. The comparison between the numerical values of FOM and FOM_{π} confirms the significance of Figs. 1 - 6: a good correlation can be found between the shifts to right of the distribution $\mathrm{P}(z_{\pi})$ and the differences FOM - FOM_{π}. For example $\mathrm{FOM}_{\pi} \approx \mathrm{FOM}$ for APP.

The values of FOM and FOM_{π} in Table IV suggest that the accuracy of the phases determined via (V.22) is expected to be higher than for phases fixed by (V.3); the average accuracy should increase with DR_{π}. This expectation is confirmed by a supplementary test (see Table IV): we extracted the average phase error $<|\Delta\phi^{\mathrm{o}}|> = <|\phi^{\mathrm{o}}_{\mathrm{cal}} - \phi^{\mathrm{o}}_{\mathrm{true}}|>$ from FOM and FOM_{π}. The larger the amount of prior information is, the better the phase estimates are.

CODE	$<|\Delta\phi^{\circ}|>$	$<|\Delta\phi^{\circ}|>_{\pi}$ ($DR_{\pi} = 0.20$)	$<|\Delta\phi^{\circ}|>_{\pi}$ ($DR_{\pi} = 0.40$)
APP	23.7	20.0	18.4
BPTI	27.1	21.5	16.4
LYSO	45.1	32.1	25.4
MYO	42.6	28.9	20.6
M-FABP	46.3	32.7	23.3
E2	52.8	23.4	17.7

Table IV: $<|\Delta\phi^{\circ}|>=<|\phi^{\circ}_{\mathrm{cal}} - \phi^{\circ}_{\mathrm{true}}|>$: mean phase error for Sayre relationship. $<|\Delta\phi^{\circ}|>_{\pi}$: mean phase error for (V.22).

A question could arise: is the value $\phi_{\mathbf{h}\,\mathrm{cal}}$, as calculated from (V.22), closer in the average to $\phi_{\mathbf{h}\,\mathrm{true}}$ than the value $\phi_{\pi,\mathbf{h}}$? In Table V we show the values $<|\phi^{\mathrm{o}}_{\mathrm{true}} - \phi^{\mathrm{o}}_{\pi}|>$ for the different values of DR_{π}. If these are compared with the average phase errors in Table IV, the important role of the triplet contribution in the phasing process is realized. Thus, phases are estimated through (V.22) much better than through the Sim (1959) relationship.

CODE	$\lvert \phi°_{\text{true}} - \phi°_{\pi} \rvert$ (DR$_{\pi}$ = 0.20)	$\lvert \phi°_{\text{true}} - \phi°_{\pi} \rvert$ (DR$_{\pi}$ = 0.40)
APP	39.5	29.8
BPTI	47.1	24.4
LYSO	46.5	30.1
MYO	41.6	25.5
M-FABP	43.2	29.4
E2	30.7	20.8

Table V: Average phase errors according to Sim (1959) relationship.

Tangent refinement

It is wortwhile noting than in Table IV $<|\Delta\phi°|>$ is rather small even for the Sayre relationship. However it should not be concluded that Sayre relationship is satisfied. Indeed the correct criterion for deciding about the violation of (V.3) or (V.20) is the inspection of FOM and FOM$_{\pi}$ because they simultaneously involve phases and moduli. If this is true, the values of FOM and FOM$_{\pi}$ should be useful indicators for foreseeing the behaviour of the tangent procedures. In particular they should measure the tendency of the tangent formulas (V.1) and (V.22) to diverge to self-consistent incorrect sets of phases.

In order to confirm this property we started phase refinement from correct phase values according to (V.1) and to (V.22), and we checked the average phase error after convergence is attained. The threshold value TR$_{\alpha}$ (i.e., a reflexion is considered "phased" if $\alpha \geq \text{TR}_{\alpha}$) is multiplied by 0.65 each new cycle. The process stops for (V.3) if $\left(\sum_{\mathbf{h}}\alpha - \sum_{\mathbf{h}}\alpha_{\text{pc}}\right) \Big/ \sum_{\mathbf{h}}\alpha \leq 0.02$ where $\alpha_{\text{pc}}$is the α value in the preceding refinement cycle. When (V.22) is used α_{π} replaces α. The results are shown in Table VI.

CODE	<|Δϕ°|> (%) Sayre relation	<|Δϕ°|> (%) Rel. (V.22) (DR_π = 0.20)	<|Δϕ°|> (%) Rel. (V.22) (DR_π = 0.40)
APP	43.2 (99)	41.7 (99)	43.5 (99)
BPTI	79.4 (93)	66.4 (98)	65.9 (99)
LYSO	40.2 (88)	31.2 (91)	24.5 (95)
MYO	68.0 (98)	62.1 (99)	57.6 (100)
M-FABP	41.8 (87)	30.0 (89)	23.1 (93)
E2	35.2 (53)	26.1 (91)	19.2 (98)

Table VI: <|Δϕ°|> after the application of the tangent formula to true phases. (%) is the percentage of the NLAR reflexions phased by the process.

It may be noted:

i) Not all the NLAR reflexions are phased at the end of the process. The percentage of phased reflexions is small for E2 when Sayre relation is used.

ii) Phases remarkably diverge, except for LYSO, M-FABP and E2 when $DR_\pi \neq 0$.

iii) As a general trend (V.22) is more efficient than (V.1) but is still not satisfactory.

Much better results are obtained by slightly modifying the refinement process. The progman stops when $\mathrm{TR}_\alpha \leq 1.5$ (this condiction prevents unreliable phase assignments) or when the number of phased reflexions is larger than 0.85 * NLAR. This last condition avoids repeated cycles of refinement on the same set of phases: in these conditions, phases usually move towards autoconsistency and diverge from the true values. The results of the new procedure are shown in Table VII.

CODE	$<\|\Delta\phi°\|>$ (%) Sayre relation	$<\|\Delta\phi°\|>$ (%) Rel. (V.22) ($DR_\pi = 0.20$)	$<\|\Delta\phi°\|>$ (%) Rel. (V.22) ($DR_\pi = 0.40$)
APP	27.3	25.9	25.6
	(85)	(88)	(90)
BPTI	36.4	24.5	14.6
	(88)	(92)	(88)
LYSO	32.9	27.5	22.9
	(72)	(86)	(93)
MYO	40.7	25.5	17.4
	(85)	(88)	(92)
M-FABP	33.4	26.1	19.5
	(68)	(84)	(85)
E2	30.8	23.3	16.9
	(32)	(87)	(93)

Table VII: $<|\Delta\phi°|>$ after the application of the tangent formula to true phases. (%) is the porcentage of the NLAR reflexions phased by the process. The minimum threshold for TR_α is 1.5.

It may be noted:

i) the number of phased reflexions is larger when relation (V.22) is used. In particular, the Sayre equation is still unable to fix for E2 the phases of about 0.68*NLAR reflexions.

ii) The phase error is remarkably smaller for the relation (V.22): the error decreases when DR_π increases.

The above conclusions confirm that the suggestions we derived from Figs. 1 - 6 are sound: the prior information on part of a crystal structure allows the successful application of (V.20) to macromolecules, that is the complete crystal structure may be in principle recovered when a partial structure is available.

Conclusions

It has been shown that relationship (V.20) is potentially able to estimate accurately the phases of a relevant number of reflexions provided some prior information is available on part of the structure. As a rule of thumb, a prior information about 30-40% of the structure should make (V.20) highly efficient.

Relationship (V.20) can be used in two different ways:

a) combined with the probabilistic techniques to improve the phasing process for reflexions up to the isomorphus resolution. In this case, the partial structure constitutes a supplementary derivative, the quality of which depends on the accuracy with which the partial structure is defined. Tangent refinement of this second derivative will produce phases that may be usefully combined with MIR phases.

b) As a stand-alone technique that is particulary useful at resolution higher than the derivative resolution.

In both cases, the use of (V.20) should be cyclic: the initial prior information is used for phase extension and refinement, which, in turn, should provide a better electron density map and therefore a better partial structure to use as new prior information.

References

Beurskens, P. T., Prick, A. J., Doesburg, H. M. & Gould, R. O. (1979). *Acta Cryst.* A**35**, 765-772.

Bricogne, G. (1974). *Acta Cryst.* A**30**, 395-405.

Burla, M. C., Cascarano, G., Fares, V., Giacovazzo, C., Polidori, G. & Spagna, R. (1989). *Acta Cryst.* A**45**, 781-786.

Camalli, M., Giacovazzo, C., & Spagna, R. (1985). *Acta Cryst.* A**41**, 605-613.

Cascarano, G., Giacovazzo, C., Burla, M. C., Nunzi, A. & Polidori, G. (1984). *Acta Cryst.* A**40**, 389-394.

Cascarano, G., Giacovazzo, C. & Guagliardi, A. (1992). *Acta Cryst.* A**48**, 859-865.

Cochran, W. (1955). *Acta Cryst.* **8**, 473-478.

Coulter, C. L. (1971). *Acta Cryst.* B**27**, 1730-1740.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148-157.

Giacovazzo, C. (1983). *Acta Cryst.* A**39**, 685-692.

Giacovazzo, C., Cascarano, G. & Zheng, C. (1988). *Acta Cryst*, A**44**, 45-51.

Giacovazzo, C., Guagliardi, A., Ravelli, R. & Siliqi, D. (1994). *Z. f. Kristallogr.* **209**, 136-142.

Giacovazzo, C. & González-Platas, J. (1995). *Acta Cryst.* A**51**, 398-404.

Glover, I., Haneef, I., Pitts, J., Wood, S. Moss, D., Tickle, I. & Blundell, T. (1983). *Biopolymers*, **22**, 293-304.

Hartmann, H., Steigemann, W., Reuscher, H. & Parak, F. (1987). *Eur. Biophys. J.* **14**, 337-348.

Hauptman, H. (1982). *Acta Cryst.* A**38**, 289-294.

Heinerman, J. J. L. (1977). *Acta Cryst.* A**33**, 100-106.

Hughes, E. W. (1953). *Acta Cryst.* **6**, 871.

Hull, S. E. & Irwin, M. J. (1978). *Acta Cryst.* A**34**, 863-870.

Karle, J. (1970). *In Crystallographic Computing*. Copenhagen: Munksgaard.

Karle, J. & Karle, I. (1966). *Acta Cryst.* **21**, 849-859.

Karle, J. & Hauptman, H. (1956). *Acta Cryst* **9**, 635-651.

Lunin, V. Yu. (1993). *Acta Cryst.* D**49**, 90-99.

Main, P. (1976). *Crystallographic Computing Techniques,* edited by F. R. Ahmed, pp. 97-105. Copenhagen: Munksgaard.

Main, P. (1979). *Acta Cryst.* A**35**, 779-785.

Main, P. (1990). *Acta Cryst.* A**46**, 372-377.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., De Kok, A. & Hol, W. G. J. (1992). *Science*, **255**, 1544-1550.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140-149.

Reeke, G. N. & Lipscomb, W. N. (1969). *Acta Cryst.* B**25**, 2614-2623.

Sayre, D. (1952). *Acta Cryst*. **5**, 60-65.

Sim, G. A. (1959). *Acta Cryst*. **12**, 813-815.

Sim, G. A. (1960). *Acta Cryst*. **13**, 511-512.

Wang, B. C. (1985). *Methods Enzymol*. **115**, 90-112.

Weinzierl, J. E., Eisenberg, D. & Dickerson, R. E. (1969). *Acta Cryst*. B**25**, 380-387.

Woolfson, M. M. (1956). *Acta Cryst*. **9**, 804-810.

Woolfson, M. M. (1993). *Acta Cryst*. D**49**, 13-17.

Zanotti, G., Scapin, G., Spadon, P., Veerkamp, J. H. & Sacchettini, J. C. (1992). *J. Biol. Chem*. **267**, 18541-18550.