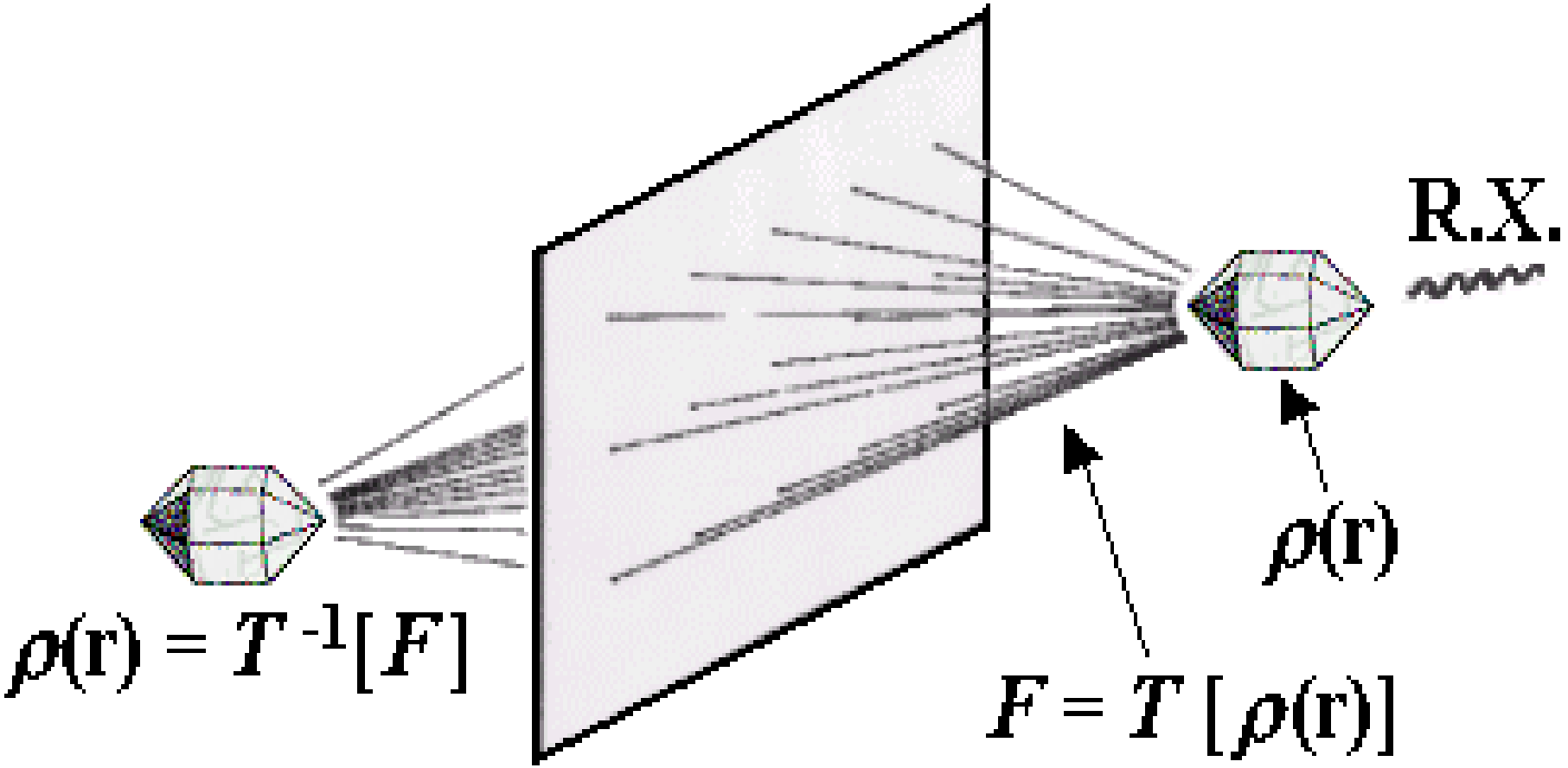
workshop *Facets of Electron Crystallography*
Berlin 7-9 july 2010

Direct Methods

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

carmelo.giacovazzo@ic.cnr.it


R.X.
$\rho(\mathbf{r}) = T^{-1}[F]$
$F = T[\rho(\mathbf{r})]$
$\rho(\mathbf{r})$

Let us answer the following questions:

crystal structure $\Rightarrow \left\{ |F|^2 \right\}$?

$\left\{ |F|^2 \right\} \Rightarrow$ crystal structure ?

$$|F_{\boldsymbol{h}}|^2 = \sum_{j=1}^{N} f_j \exp(2\pi i \boldsymbol{h}\boldsymbol{r}_j) \sum_{j=1}^{N} f_j \exp(-2\pi i \boldsymbol{h}\boldsymbol{r}_j)$$

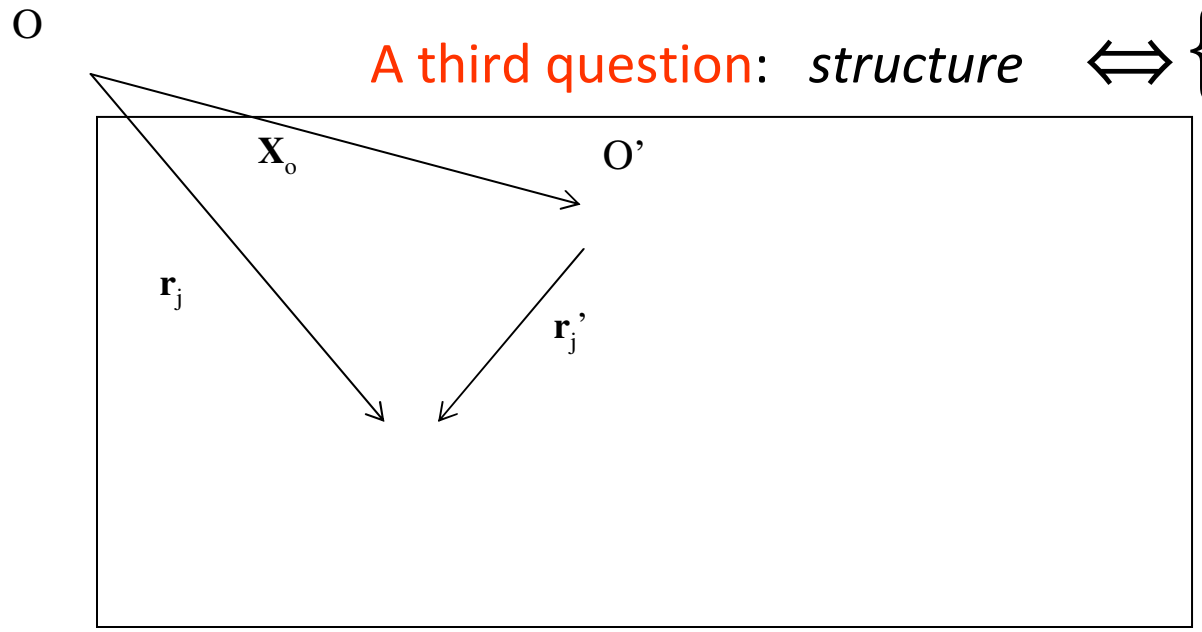
$$= \sum_{i,j=1}^{N} f_i f_j \exp[2\pi i \boldsymbol{h}(\boldsymbol{r}_i - \boldsymbol{r}_j)]$$

$$= \sum_{j=1}^{N} f_j^2 + \sum_{i \neq j=1}^{N} f_i f_j \exp[2\pi i \boldsymbol{h}(\boldsymbol{r}_i - \boldsymbol{r}_j)]$$

As a consequence :

$$\left\{ |F|^2 \right\} \Leftrightarrow \rho(\boldsymbol{r})$$

A third question: *structure* $\Leftrightarrow \{\varphi\}$?



$$F_{\boldsymbol{h}} = \sum_{j=1}^{N} f_j \exp\left(2\pi i \boldsymbol{h}\boldsymbol{r}_j\right) = \sum_{j=1}^{N} f_j \exp\left(2\pi i \boldsymbol{h}(\boldsymbol{X}_0 + \boldsymbol{r}_j')\right)$$

$$= \exp\left(2\pi i \boldsymbol{h}\boldsymbol{X}_0\right) \sum_{j=1}^{N} f_j \exp\left(2\pi i \boldsymbol{h}\boldsymbol{r}_j'\right)$$

$$= \exp\left(2\pi i \boldsymbol{h}\boldsymbol{X}_0\right) F_{\boldsymbol{h}}'$$

$$\rightarrow F_{\boldsymbol{h}}' = F_{\boldsymbol{h}} \exp\left(-2\pi i \boldsymbol{h}\boldsymbol{X}_0\right) = \left|F_{\boldsymbol{h}}\right| \exp i\left(\varphi_h - 2\pi \boldsymbol{h}\boldsymbol{X}_0\right)$$

A fourth basic question

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

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

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

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

The simplest invariant : *the triplet invariant*

Use the relation

$$F'_{\mathbf{h}} = F_{\mathbf{h}} \exp\left(-2\pi i \mathbf{h} \mathbf{X}_0\right)$$

to check that the invariant $F_{\mathbf{h}} F_{\mathbf{k}} F_{-\mathbf{h}-\mathbf{k}}$ does not depend on the origin.

$$F'_{\boldsymbol{h}} F'_{\boldsymbol{k}} F'_{-\boldsymbol{h}-\boldsymbol{k}} = |F_{\boldsymbol{h}}| \exp i\left(\phi_{\boldsymbol{h}} - 2\pi \boldsymbol{h} \boldsymbol{X}_0\right) |F_{\boldsymbol{k}}| \exp i\left(\phi_{\boldsymbol{h}} - 2\pi \boldsymbol{k} \boldsymbol{X}_0\right)$$

$$|F_{-\boldsymbol{h}-\boldsymbol{k}}| \exp i\left[\varphi_{-\boldsymbol{h}-\boldsymbol{k}} + 2\pi(\boldsymbol{h}+\boldsymbol{k})\boldsymbol{X}_0\right]$$

$$= |F_{\boldsymbol{h}}| \, |F_{\boldsymbol{k}}| \, |F_{-\boldsymbol{h}-\boldsymbol{k}}| \exp i\left(\varphi_{\boldsymbol{h}} + \varphi_{\boldsymbol{k}} + \varphi_{-\boldsymbol{h}-\boldsymbol{k}}\right)$$

The sum ($\varphi_{\mathbf{h}} + \varphi_{\mathbf{k}} + \varphi_{-\mathbf{h}-\mathbf{k}}$) is called ***triplet phase invariant***

Structure invariants

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

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

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

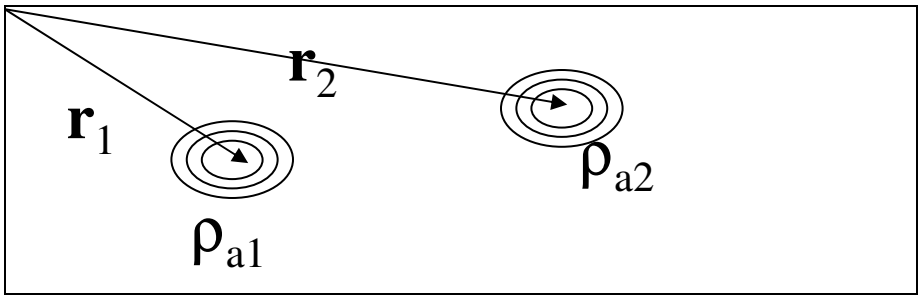
quartet invariant : $F_h F_k F_l F_{-h-k-l}$

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

.....

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

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



$$\rho(\boldsymbol{r}) = \sum_{j=1}^{N} \rho_{aj}(\boldsymbol{r} - \boldsymbol{r}_j)$$

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

$$\rho(\mathbf{r}) > 0 \quad \Rightarrow \quad f > 0$$

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

The Wilson statistics

- Under the above conditions Wilson (1942,1949) derived the structure factor statistics. The main results where:
- $$<|F_{\mathbf{h}}|^2>=\sum_{j=1}^{N} f_j^2 \qquad (1)$$
- Eq.(1) is :
- a) resolution dependent (f_j varies with θ),
- b) temperature dependent: $f_j = f_j^0 \exp(-B_j \sin^2\theta/\lambda^2)$
- From eq.(1) the concept of normalized structure factor arises:

$$E_{\mathbf{h}} = F_{\mathbf{h}} / (\sum_{j=1}^{N} f_j^2)^{1/2}$$

The Wilson Statistics

- |E|-distributions:

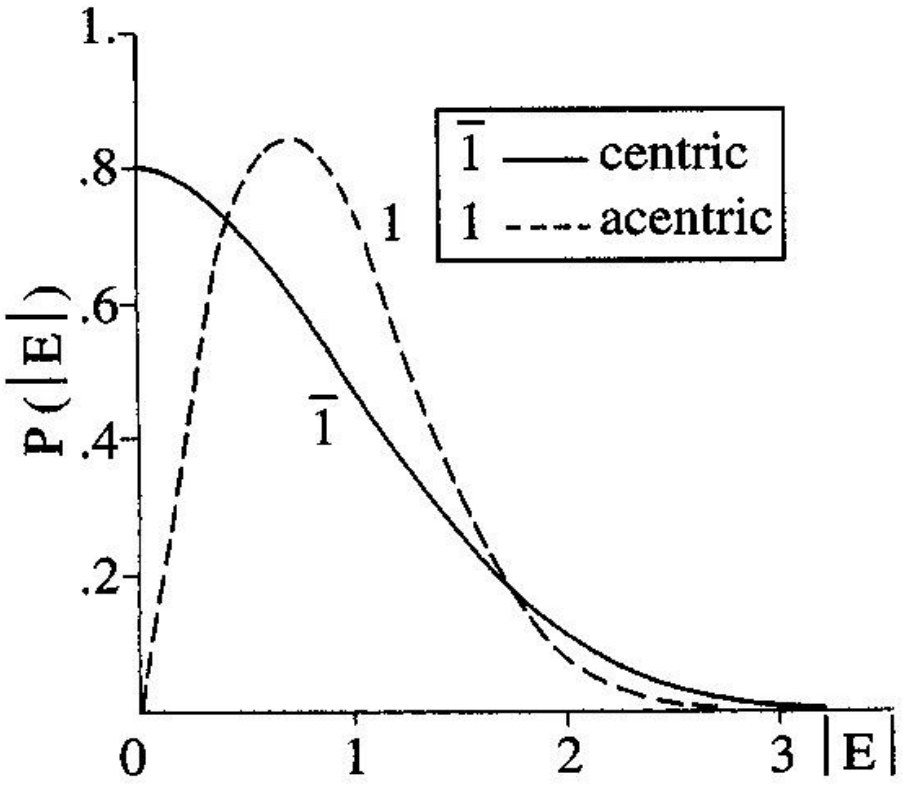
$$P_1(|E|) = 2|E|\exp(-|E|^2)$$

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

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

in both the cases.

The statistics may be used to evaluate the average themel factor and the absolute scale factor.



The Wilson plot

$$F_{\mathbf{h}} = \sum_{j=1}^{N} f_j \exp 2\pi i \mathbf{hr}_j = \sum f_j^0 \exp\left(-B_j \frac{\sin^2\theta}{\lambda^2}\right) \exp 2\pi i \mathbf{hr}_j$$

A

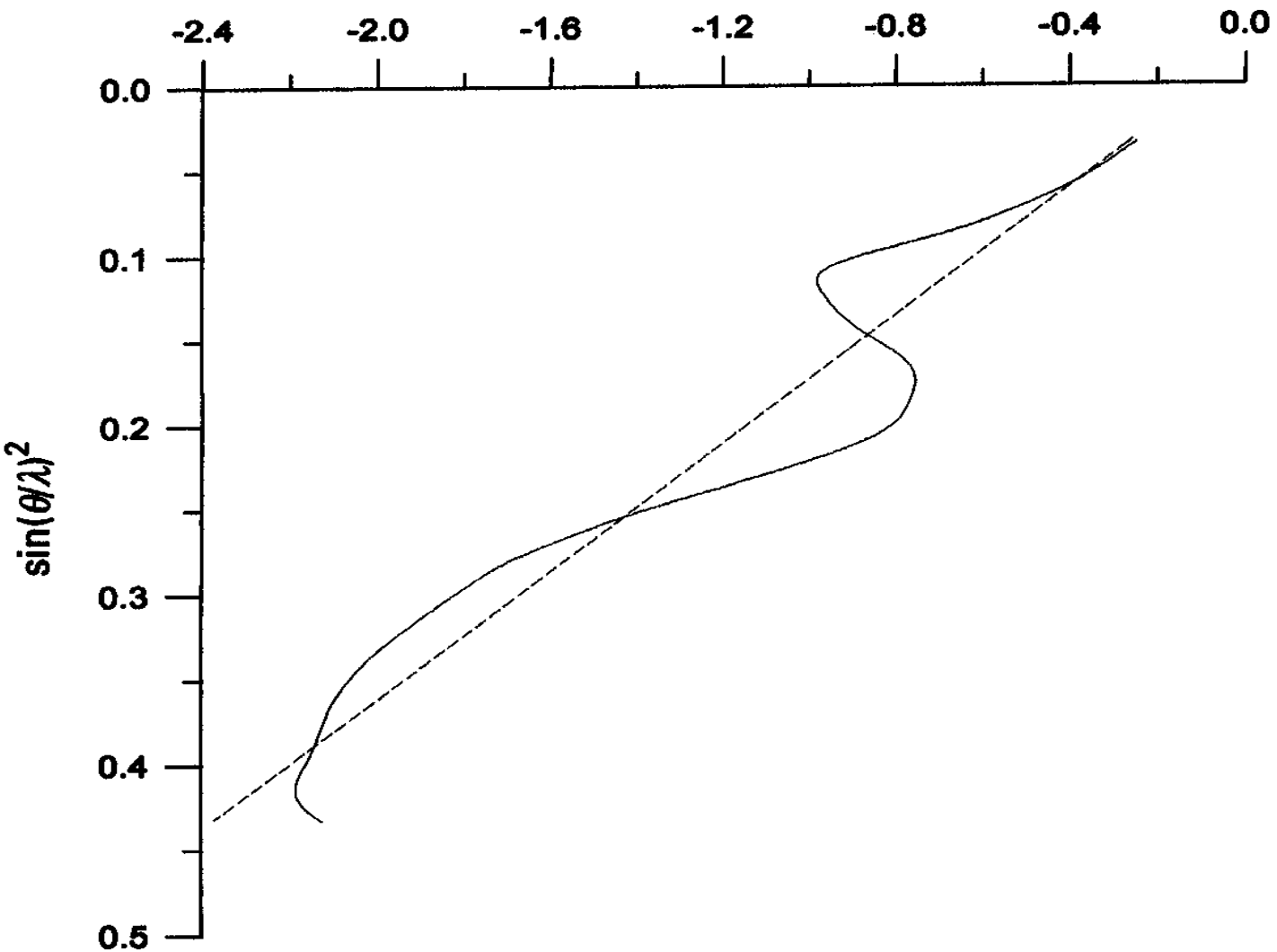
$$\exp\left(-B \underbrace{\frac{\sin^2\theta}{\lambda^2}}_{s^2}\right) \underbrace{\sum f_j^0 \exp 2\pi i \mathbf{hr}_j}_{F_h^0}$$

$$|F_h|^2_{obs} = K|F_h|^2 = K\left|F_h^0\right|^2 \exp\left(-2Bs^2\right)$$

A

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

$$\underbrace{\ln\left(\frac{<|F_h|^2_{obs}}{\Sigma_s^0}\right)}_{y} = \ln K - 2B\underbrace{s^2}_{x}$$

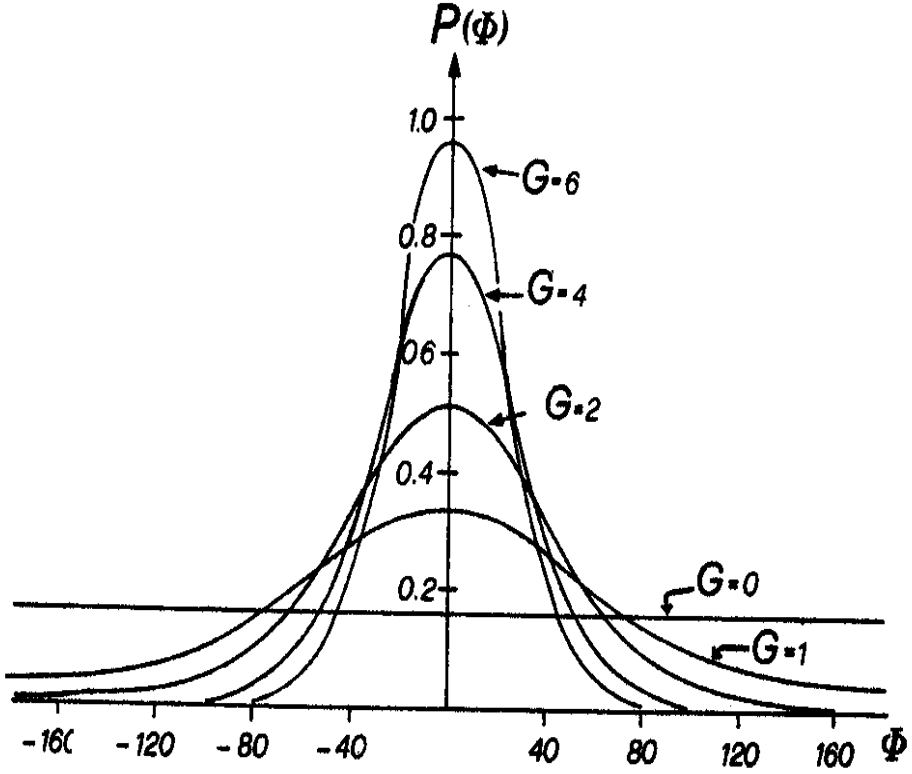

-2.4
-2.0
-1.6
-1.2
-0.8
-0.4
0.0
0.0
0.1
0.2
0.3
0.4
0.5
sin(θ/λ)²

The Cochran formula

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

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

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



Accordingly:

$\varphi_h + \varphi_k - \varphi_{h+k} \approx 0 \quad \propto \quad G = 2 \mid E_h E_k E_{h+k} \mid / N^{1/2}$

$\varphi_h - \varphi_k - \varphi_{h-k} \approx 0 \quad \propto \quad G = 2 \mid E_h E_k E_{h-k} \mid / N^{1/2}$

$\varphi_h \approx \varphi_k - \varphi_{h-k} \quad \propto \quad G = 2 \mid E_h E_k E_{h-k} \mid / N^{1/2}$

The tangent formula

A reflection can enter into several triplets.Accordingly

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

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

……

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

Then

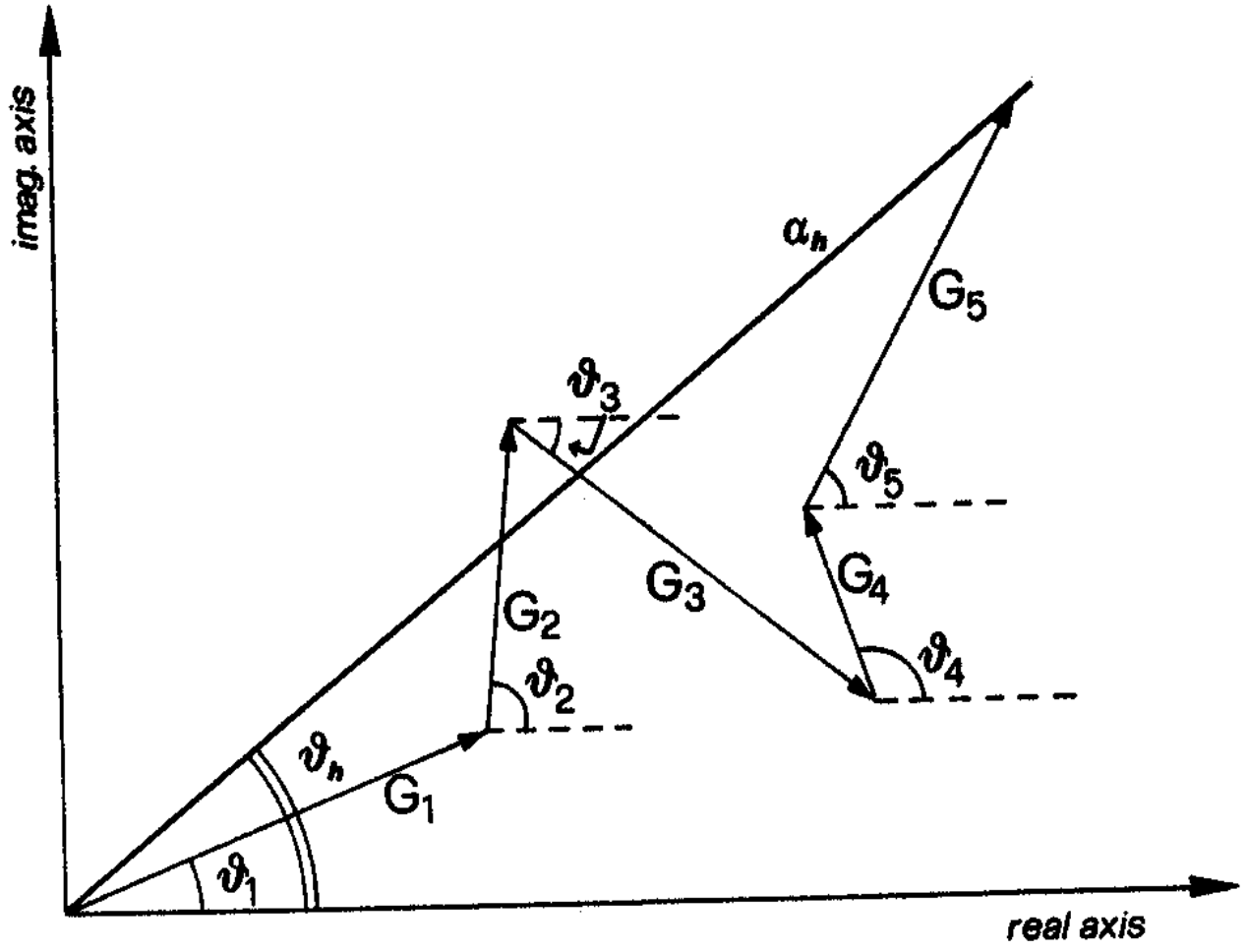
$P(\varphi_h) \approx \prod_j P_j(\varphi_h) \approx L^{-1} \prod_j \exp\,[G_j \cos\,(\varphi_h - \theta_j)]$

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

where

$$\tan\,\theta_h = \frac{\sum G_j \sin\,\theta_j}{\sum G_j \cos\,\theta_j} = \frac{T}{B}, \qquad \alpha_h = \left(T^2 + B^2\right)^{1/2}$$

A geometric interpretation of α



The random starting approach

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

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

How to recognize the correct solution?
Figures of merit can or cannot be applied

Tangent cycles

- ϕ_1 ϕ'_1 ϕ''_1 …… ϕ^c_1
- ϕ_2 ϕ'_2 ϕ''_2 …… ϕ^c_2
- ϕ_3 ϕ'_3 ϕ''_3 …… ϕ^c_3
- ……
- ϕ_n ϕ'_n ϕ''_n…… ϕ^c_n
-

- ***Ab initio*** **phasing**
- **SIR2009** is able to solve

-small size structures (up to 80 atoms in the a.u.);

-medium-size structures (up to 200);

-large size (no upper limit)

- It uses
- **Patterson deconvolution techniques**
- (***multiple implication transformations***)
- as well as
- **Direct methods**
- to obtain a starting set of phases. They are extended and refined *via*
- ***electron density modification techniques***
-

- **Direct methods limits for proteins**:

- 1) the large size (proteins range from300 atoms in the asymmetric unit to several thousands). The G factor in the Cochran formula are very small.

- 2) data resolution

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

About the data resolution limit

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

- - the atomicity condition is violated;
- - the number of reliable triplet invariants exploitable by the tangent procedure is small.
- - Patterson and EDM procedures are less effective;
- - the small ratio
- ***number of observations/ number of parameters***
- make least squares unreliable.