

# Maximum likelihood refinement

Garib N. Murshudov

More to come here

## 1 Experiment and model

Treatment of experimental data could be considered as following way:

$$\text{new knowledge} = \text{known} \times \text{data} \quad (1)$$

Where

new knowledge is what experimenter wants to know i.e the aim of the experiment

known - what is already known. It helps to design and to treat experiment. In crystallography it might be information about protein geometry or the known structure of the protein under investigation.

data - the result of the experiment. In crystallography the intensities (amplitudes of structure factors) collected from the crystals.

The treatment of the experimental data will change with time, but the data themselves do not change. This fact gives extra responsibility to the experimenter and to deposition centres. Experimenter should take care to collect the best

experimental data which the equipment and his skills allow. For example in macromolecular crystallography it is important to collect all data (high as well as low resolution data and anomalous pairs). Since the direct results of experiments are images they should be kept. Deposition centres and journals publishing direct or indirect result of experiments should require deposition of the data. Ideally images should be deposited.

What is experiment? In macromolecular crystallography schematically primary treatment of experiment could be written as:

$$\text{Images} \Rightarrow \text{integrated data} \Rightarrow \text{intensities} \Rightarrow \text{amplitudes} \quad (2)$$

Every step here uses some assumption and thus after each step, depending on the assumption, there might be some loss of information. For example after integration the diffuse scattering which gives information about correlated motion of atoms will be lost. In the derivation of intensities and amplitudes there are problems with estimating uncertainty of the data ( $\sigma_{exp}$ ). To reduce the loss of information images themselves should be considered as data. But modern techniques do not yet allow their direct use, for example for the refinement. But in future this might change. There is already some work in this direction.

(1) could be written in a different way (it is famous Bayes's theorem)

$$P(\text{model};\text{experiment}) = P_{\text{prior}}(\text{model})P(\text{experiment};\text{model}) \quad (3)$$

$P(\text{model};\text{experiment})$  reflects knowledge about model after experiment,

$P_{prior}(\text{model})$  what is known before experiment. For example bond lengths, bond angles, chiral volumes, torsion angles etc

$P(\text{experiment};\text{model})$  behaviour of experiment if model would be known. In other words conditional probability distribution of experimental data (e.g. amplitudes of structure factors) if coordinates with errors are known.

Examples of the model and experiment:

1) model -  $\{X\}$  coordinates of protein atoms

experiment -  $\{|F_{nati}|, (\phi_{nati})\}$  amplitudes and possibly some information about the phases of the structure factors

2) model -  $\{\text{heavy atoms}\}$  heavy atom positions

experiment -  $\{|F|_{nati}, |F|_{derivatives}\}$  amplitudes of structure factors for native and heavy atom derivative crystals.

3) model -  $\{X_{nati}, X_{complex}\}$  coordinates of the native protein atoms and the protein complexed with some ligands.

experiment -  $\{|F|_{nati}, |F|_{complex}\}$  amplitudes of structure factors of native and complexed crystals

This case could be rewritten in a different way. In principle the experimenter might be interested not in coordinates themselves but differences between the two coordinate sets. In this case:

model -  $\{X_{differences}, X_{common}\}$  differences and similarity between native and complexed structures.

In principle 2) is a particular case of 3), where the ligand is the heavy atom

compound.

Treatment of 1) and 2) are already ready in the first approximation. Programs like REFMAC (Murshudov et al 1997), CNS, Xplor (Pannu and Read 1996, Adams et al 1998), BUSTER/TNT (Bricogne and Irwin 1996) can refine macromolecules using one or another implementation of this treatment. SHARP (de la Fortelle and Bricogne 1997) can refine heavy atom parameters. In near future we can expect that 3) will be available.

So for the treatment of experimental data it is important to have expressions for  $P(\text{experiment};\text{model})$  and  $P_{\text{prior}}(\text{model})$ . I.e. we have to rationalise available knowledge on model we are interested in, and the behaviour of the experimental data if the model were known.  $P_{\text{prior}}(\text{model})$  in current implementations uses restraints (REFMAC, TNT, SHELXL, RESTRAIN) or energy (CNS, Xplor), although there are still unsolved problems. For example many restraints assume that their deviation from mean value is distributed by following a Gaussian law, which is not always true. Another problem is that the treatment of the restraints (and energy) do not take into account the B-values of the atoms. It is known from small molecular crystallography that B-values do affect observed bond lengths (for example see Giacovazzo et al, Fundamentals of Crystallography ). Observed bond lengths may become shorter than ideal value. B-values might affect bond angles also. So far there are no completely satisfactory approaches to the problems of restraints.

## 2 Likelihood function in crystallography

$P(\text{experiment}; \text{model})$  is a likelihood function. Luzzati (1952) gave the first derivation of this. In fact he derived the probability distribution of structure factors when model with an error is known using the central limit theorem.

### *Outline of central limit theorem*

If there are a large number of random variables with definite mean and dispersion, then their average value could be approximated by a Gaussian distribution with mean value of the average of the means and with dispersion of the average of the dispersions.

In macromolecules there are a large number of atoms and if their positions are known approximately, then structure factors derived from these atoms could be considered as random variables with definite means and dispersion. Then distribution of the sum of these structure factors could be approximated by a Gaussian:

$$P(F; (F_c)) = \begin{cases} \frac{1}{\pi(\Sigma)} e^{-\frac{|F-D F_c|^2}{\Sigma}} & \text{acentric} \\ \frac{1}{\sqrt{2\pi}\Sigma} e^{-\frac{|F-D F_c|^2}{2\Sigma}} & \text{centric} \end{cases} \quad (4)$$

where:

$F$  is true structure factor

$F_c$  is structure factor from incomplete model with an error

$$\Sigma = \varepsilon (\Sigma_q + (1 - D^2)\Sigma_p)$$

$D = \cos(2\pi\Delta x)$ . It reflects degree of accuracy of known coordinates.  $D = 1.0$  correspond to the model without error.

$\varepsilon$  is multiplicity of reflecting plane

$\Delta x$  is error in atomic positions.

$\Sigma_q = \sum_{\text{unknown atoms}} f_j^2$ . It reflects the amount of atoms we have not modelled yet.

$\Sigma_p = \sum_{\text{known atoms}} f_j^2$ . It reflects the amount of atoms we already know.

$f_j$  is atomic form factor of  $j$ s atom. It reflects the scattering power or possible contribution of atom  $j$ .

To take into account the experimental uncertainties we increment  $\Sigma$  given above by the experimental ( $\sigma_{exp}^2$ ) (See for details Murshudov et al 1997).

The distribution of the structure factors themselves is Gaussian with mean at the calculated structure factor reduced by the factor  $D$  which reflects the degree of the accuracy of the model. Its dispersion is a function of the unknown (unmodelled), known part and the error in the known part of the structure.

In this equation we should take into account the fact that structure factor  $F$  is a complex number and can be expressed as  $|F| \exp i\phi$  where  $\phi$  is phase of structure factor and  $|F|$  is amplitude of structure factor. The result of experiment is usually  $|F|$ , the amplitude of the structure factor. But sometimes when we use heavy atom derivative methods for structure solutions we have some information about phases also. To derive distribution of the amplitudes we have to integrate phase out. At this stage we can use any available information about phases (MIR, MIRAS, SAD, NCS or other sources). See for example Murshudov et al (1997).

This approach already gives dramatic improvement of refinement behaviour (see examples below). The better approach is to treat all experimental data

(derivatives, native, liganded, mutants) simultaneously. Derivation of necessary equations already has been done, but the implementation will take some time. This will increase the stability of the refinement and the reliability of the derived model.

### 3 Comparison of maximum likelihood and least-squares methods

To compare maximum likelihood and least-squares residuals we can analyse the behaviour of their gradients. In principle least-squares is a special case of the maximum likelihood methods. If the **amplitudes** of the structure factors were distributed as Gaussians with known dispersion then maximum likelihood would become least-squares. But it is structure factors themselves that are distributed according to the Gaussian law, not the amplitudes. But at the end stages of refinement when the model is complete and has a small error then the maximum likelihood could be approximated by the least-squares (see Murshudov et al 1997).

#### *Gradients of least-squares and maximum likelihood functions*

For least square gradients map with coefficients:

$$w(|F| - |F|_c) \exp(i\phi_c) \quad (5)$$

is used, where  $w$  is the weight to be applied. In most cases  $w = 1$ .

In the maximum likelihood case for the gradient calculations a map with

coefficients is calculated:

$$\frac{(m_{comb}|F| \exp(\iota\phi_{comb}) - D|F|_c \exp(\iota\phi_c))}{\Sigma} \quad (6)$$

where  $m_{comb}$  is figure of merit of the combined phases

$\phi_{com}$  are combined phases.

$D$  reflects degree on accuracy of model

$\Sigma$  reflects degree of knowledge about model (see above)

Thus. unlike least-squares, maximum likelihood maps take into account the degree of accuracy of the model ( $D$ ), the degree of accuracy of the phases ( $m_{comb}$ ) and the completeness of the model ( $\Sigma$ ).

A similar type of coefficient (without the factor  $\Sigma$ ) is used for the difference map calculations for inspection of the model under study.

These coefficients are:

$$FWT = \begin{cases} 2m_{comb}|F| \exp(\iota\phi_{comb}) - D|F_c| \exp(\iota\phi_c) & \text{if reflection was included in refinement} \\ D|F_c| \exp(\iota\phi_c) & \text{otherwise} \end{cases} \quad (7)$$

$$DELFWT = \begin{cases} m|F| \exp(\iota\phi_{comb}) - D|F_c| \exp(\iota\phi_c) & \text{if reflection was included in refinement} \\ 0 & \text{otherwise} \end{cases}$$

where  $FWT$  is coefficients for the map similar to the  $2F - F_c$  and  $DELFWT$  is coefficient for the map similar to the  $F - F_c$ .

In this treatment absent reflections or those chosen for the free R set are restored to avoid unnecessary noises in the map.

## 4 Examples of application of maximum likelihood refinement

Two examples are considered.

- 1) Bias removal during refinement. cytochrome c'.
- 2) Application of phased refinement for phase improvement
- 1) *Cytochrome c'. Bias removal after molecular replacement*

The structure was solved by Baker et al. (1995). This starting model was based on a molecular replacement (MR) solution where the model used had only 25% homology to Cytochrome c'. Although a solution was found ten residues had out of register errors and another 10 were completely misplaced. In such cases where an extensive rebuilding is necessary, the problem of map bias is very serious. Automatic cycles of FLSQ (amplitude based least-squares) refinement using all reflections had reduced the R-value to 34.8, and this partially refined model was used as a starting point for MLKF (maximum likelihood based on amplitudes of structure factors).

At this point 5% of reflections were assigned as "free" and were used for estimation of the overall likelihood parameters. At first the  $\sigma_A$  weighting was overestimated and the initial  $m$  was much higher than the  $\langle \cos \Delta\phi \rangle$ , indicating the importance of assigning "free" reflections at the beginning of refinement. REFMAC was able to refine the FLSQ model further and the phase error was reduced by 6 degrees. The behaviour of  $\langle \cos \Delta\phi \rangle$  vs resolution (Figure 1) shows that during refinement the phases for the high resolution data were improved

most, and  $m$  and  $\langle \cos \Delta\phi \rangle$  converged.

It is interesting to inspect the maps (figure 2) which were available for correcting the model. These illustrate a section where the initial model was completely wrong. The  $(2F^o - F^c)$  map (figure 2a) and  $(3F^o - 2F^c)$  map (not shown) based on the FLSQ model are seriously biased and noisy and it would be easy to trace the chain perpendicular to its true direction. The SIGMAA maps for the initial FLSQ model (figure 2b) correlated with the final  $F^c$  model map better but still there is a break in the main chain and the electron density could be interpreted wrongly. The map after REFMAC had a map correlation coefficient 5% higher than that for the map calculated by SIGMAA coefficients, showed less ambiguous connectivity, and density for side chains and water molecules had appeared (figure 2c).

## 2) Mannanose. Phased refinement and phasing.

This structure was solved by Sabini, Schubert, Murshudov, Wilson, Siika-Aho & Penttila.

Space group of crystals was  $P2_1$ . There were three datasets. One from native crystal at 2Å with cell dimensions 50.0, 54.3 60.2 and  $\beta$  angle 111.3. Second from the crystal soaked in Pt compound at 1.65Å resolution with cell dimensions 51.1, 54.3, 61.0 with  $\beta=110.2$ . Third dataset was also from the crystal soaked in Pt compound at the resolution 1.5Å with cell dimensions 44.7, 54.6, 60.8,  $\beta=111.2$ . The R-value between the second data and the native (first data) was 40.7% showing serious problem of nonisomorphism. The first and the third data are virtually identical. Attempts to solve structure using SIRAS (single isomorphous

and anomalous scattering method) failed. Then it was decided to try to solve the second crystal structure alone using SAD (single anomalous dispersion) and to use it as an initial structure for the solution of the others. For this the following procedure was used:

- 1) Find positions of platinum from anomalous Patterson map
- 2) Refine Pt sites with anomalous information. The program MLPHARE was used.
- 3) Produce SAD (single anomalous dispersion) phases. Check solvent flattened map. At this stage handedness was clear.
- 4) ARPP. Add free atoms to the model
- 5) Phased refinement of the model.
- 6) Goto 4) if convergence has not been achieved
- 7) Final map
- 8) Build model.
- 9) Solvation and phased anisotropic refinement
- 10) Use resultant model to solve native structures

Map was so good that automatic model building procedure in QUANTA failed. Building of model was very fast. Almost all residues could be identified without looking amino acid sequence.

## References

- Adams, P.D., Pannu, N.S., Read, R.J. & Brünger, A.T. (1997). *Proc. Nat'l. Acad. Sci. (USA)*, **94**, 5018–5023.
- Baker, E.N., Anderson, B.F., Dobbs, A.J. & Dodson, E.J. (1995) *Acta Cryst.*, **D51** 282-289
- Luzzati, V. (1952) *Acta Cryst.* **5**, 802-810
- Box, G.E.P. & Tiao, G.C. (1973) *Bayesian Inference in Statistical Analysis*, Addison-Wesley publishing Company.
- Bricogne, G. & Irwin, J. *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend January 1996*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pages 85–92, Daresbury, UK. Central Laboratory of the Research Councils.
- CCP4. Collaborative Crystallographic Project, Number 4. (1994) *Acta Cryst.* **D50**, 760-763
- de La Fortelle, E. & Bricogne, G. (1997) *Methods in Enzymology*, **276**, 472–494.
- Murshudov, G.N., Dodson, E.J. & Vagin, A.A. *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend January 1996*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pages 85–92, Daresbury, UK. Central Laboratory of the Research Councils.
- Murshudov, G.N., Vagin, A.A. & Dodson, E.J. (1997). *Acta Cryst.*, **D53**, 240–255.

- Sabini, E., Schubert H., Murshudov, G.N, Wilson, K.S, Siika-Aho, M & Penttila, M. (1998) unpublished
- Pannu, N.S. & Read, R.J. (1996) *Acta Cryst.*, **A52**,659-668.
- Pannu, N., Murshudov, G.N., Dodson, E.J. & Read, R. (1998) *ActaCryst.*
- Srinivasan, R. & Ramachandran, G.N. (1965) *Acta Cryst.* **19**, 1008-1014