Crystallography Reports, Vol. 45, No. 2, 2000, pp. 195–198. Translated from Kristallografiya, Vol. 45, No. 2, 2000, pp. 227–230.

THEORY OF CRYSTAL STRUCTURES

Reliability of Maximum Likelihood-Based Figures of Merit

T. P. Skovoroda and V. Yu. Lunin

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow oblast, 142292 Russia

Received February 19, 1997; in final form, February 3, 1998

Abstract—Various schemes for determining the maximum likelihood-based figures of merit for phases of structure factors have been considered. It is shown that the use of the likelihood function of all the available structure factors provides the adequate estimates of the accuracy of phases calculated for the atomic models with independent errors in the coordinates, but, at the same time, systematically overestimates the figures of merit for models preliminarily refined in the reciprocal space. It is shown that the use of the marginal likelihood function calculated from the control set of reflections allows the elimination of the systematic bias estimates. A method for reducing the statistical dispersion of the estimates based on a small number of control reflections is suggested. © *2000 MAIK "Nauka/Interperiodica".*

INTRODUCTION

In crystallography of macromolecules, the Fourier maps of electron density are often calculated with the use of the coefficients

$$m_s F_s^{obs} \exp(i\varphi_s^{mod}), \quad (1)$$

where **s** is the reciprocal-lattice point, F_s^{obs} is the experimentally determined structure-factor modulus, and φ_s^{mod} is the structure-factor phase calculated for a certain preliminarily chosen atomic model of the structure. The weighting factor m_s (the so-called figure of merit of the phase determination) is introduced to compensate possible errors in the Fourier coefficient caused by the discrepancy between the experimentally determined phase φ_s^{mod} and the true phase φ_s^{true}. The probabilistic assumptions about the accidental nature of the errors in the preliminary atomic model of the structure allow one to determine the weights m_s as the mathematical expectations of the quantities $\cos(\varphi_s^{true} - \varphi_s^{mod})$ [1] and lead to the widely used formula

$$m_s = \Lambda(t_s F_s^{obs} F_s^{mod}/\varepsilon_s), \quad (2)$$

where the function $\Lambda(x)$ is either the hyperbolic tangent (for centrosymmetric reflections) or the ratio of the modified Bessel functions $I_1(2x)/I_0(2x)$ (for the noncentrosymmetric ones), F_s^{mod} is the structure-factor modulus calculated for the preliminarily model, and ε_s is the coefficient compensating the differences between the average intensities of reflections of various types. The parameter t_s in (2) reflects the errors made in the construction of the preliminary model. The correct estimate of this error (i.e., the estimate of the adequacy of the preliminary model) is the key factor in the determination of the weights m_s.

On the average, the "appropriately calculated" weights m_s should correspond (be close) to the real values of $\cos(\varphi_s^{true} - \varphi_s^{mod})$. The degree of this correspondence can be checked in some test situations; i.e., in the situations where, in addition to the preliminary model, one also uses sufficiently reliable "sought" atomic structure such that the phases determined for this structure can be assumed to be the true ones. Such a test performed by Read [2] showed that many of the procedures suggested earlier for calculating the parameters t_s yielded unrealistic figures of merits, and that the best results are obtained within the maximum likelihood approach [2–5]. However, as was indicated in [2–5], the use of this approach gives rise to some difficulties in the work with the preliminary models subjected earlier to crystallographic refinement. Below, we analyze some modifications of this method that allow one to obtain the realistic figure of merits for the phases calculated for the preliminarily refined atomic models.

The tests were performed on the experimental data for Protein *G* (sp. gr. $P2_12_12_1$, the unit cell dimensions $34.9 \times 40.3 \times 42.2$ Å) with the known atomic structure determined at a high resolution.

The refinement of the test preliminary models was performed with the use of the FROG complex of programs [6], the maximum likelihood-based estimates (the ML estimates) and the estimates based on the maximization of the marginal likelihood function (the MML estimates) and the LBEST program [7] specially

1063-7745/00/4502-0195$20.00 © 2000 MAIK "Nauka/Interperiodica"

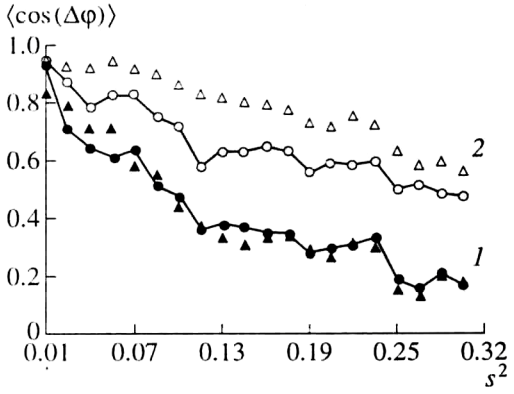

Fig. 1. Cosines of phase errors predicted from ML estimates (triangles) and the real values (circles) for (*1*) the model with independent errors in the atomic coordinates and (*2*) for the same model upon 12 cycles of its refinement in the reciprocal space.

designed in this study in accordance with the method considered elsewhere [3, 5].

ML-ESTIMATES OF THE PARAMETERS OF PROBABILITY DISTRIBUTIONS

The essence of the probabilistic approach to the estimation of the errors in the phases [8–10] is the allowance for the additional information about the object under study in the form of a statistical hypothesis on the character of error distribution in the preliminary model. The simplest and the most rigorous variant of this hypothesis reduces to the assumption that all the errors are reduced to the errors in the atomic coordinates and that they are independent and distributed according to the normal law with the nonzero mean and the same dispersions (more complicated examples were considered in [5, 7, 11]). Similar hypotheses allow one to consider the values of the structure factors as random quantities and obtain the expressions for their probability distributions. The resulting distributions for a large class of initial hypotheses are described by the same formula,

$$P(F_s, \varphi_s) \propto F_s \exp\{-|F_s \exp(i\varphi_s) - \alpha_s F_s^{mod} \exp(i\varphi_s^{mod})|^2/\varepsilon_s\beta_s\} \quad (3)$$

and differ only by the form of the functional dependences of α_s and β_s on the parameters describing the error distribution in the atomic model. Therefore, there is no need to fix any concrete hypothesis about the atomic model. We proceed from the assumption that the probability distribution for a structure factor is described by formula (3) with certain parameters α_s and β_s. The mathematical expectation of the cosine of the phase error in this case is given by formula (2) with $t_s = \alpha_s/\beta_s$. Thus, we arrive at one of the standard problems of the mathematical statistics—the problem of selection of the distribution from class (3) or the problem of estimating the parameters α_s and β_s.

The maximum likelihood-based methods of estimating the α_s and β_s parameters were described elsewhere [2–5]. The method reduces to the following. The estimates of the α_s and β_s quantities for a thin spherical layer of the reciprocal space are taken to be the constants maximizing the probability of coincidence of the structure-factor moduli distributed according law (3) and the $\{F_s^{obs}\}$ values obtained in the real experiment.

ESTIMATES OF FIGURES OF MERITS

Within the framework of the approach used, the figures of merits calculated by formula (2) are also random quantities possessing all the characteristics typical of random quantities. One of the most important statistical characteristics of such estimates is the bias—the deviation of the expected value of the estimate from its true value. Another important characteristic is the expected root-mean-square deviation of the estimate from its true value (the estimate dispersion).

The first run of the tests was devoted to the study of the bias of the figure of merit calculated from ML estimates for the α_s and β_s parameters. In order to exclude the effect of the experimental errors and inaccuracy of the initial model, we used the data calculated for the structure of *Protein G* without water molecules as the exact values of the phases and the experimental values of the structure factor moduli. The same model but with deliberately introduced independent random errors in the atomic coordinates (the mean absolute value 0.8 Å) was used as the starting model for the further refinement. Figure 1 shows the results of the use of the ML estimates for the starting model and for the mode upon twelve cycles of its refinement in the reciprocal space. It is clearly seen that the figures of merits based on ML estimates show no bias in independent errors in the atomic coordinates of the model, but are systematically overestimated for the preliminarily refined models.

ELIMINATION OF BIAS. MML ESTIMATES

Earlier, we suggested the modified method for determining the α_s and β_s parameters [5] based on the Brünger *R*-free likelihood-based method [12] also used in [13–15]. The likelihood function in this method is calculated using only the reflections, which were preliminarily excluded from the refinement process. Hereafter the parameter values obtained by maximization of this marginal likelihood function are called the MML estimates. Our further experiments were directed to the determination of the bias of the figures of merit based on the MML estimates for the α_s and β_s parameters. In each test, the refinement procedure was preceded by the selection of a control set of reflections that should be excluded from the refinement and should be used for

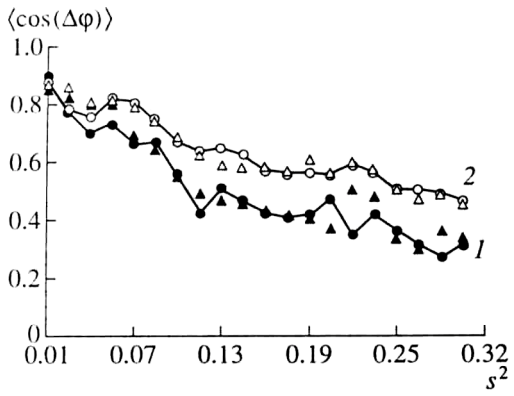

Fig. 2. Cosines of phase errors predicted from MML estimates (triangles) and the real values (circles) for the refinement (*1*) without and (2) with the stereochemical restraints. Model data.

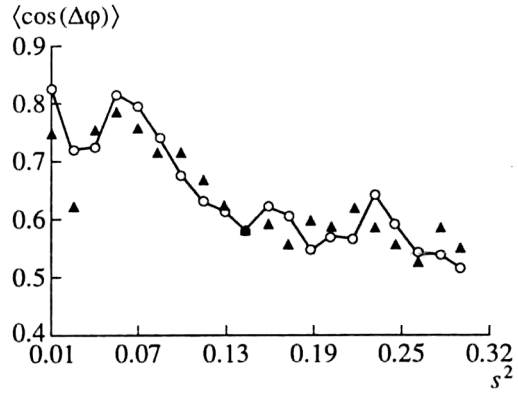

Fig. 3. Cosines of phase errors predicted from MML estimates (triangles) and the real values (circles) for the model refined over 50% of the experimental data.

the subsequent estimation of the α_s and β_s parameters. Then the figure of merits were calculated for all the reflections using formula (2) with $t_s = \alpha_s/\beta_s$.

Figure 2 shows the results of such calculation for two different refinement strategies. Using the first strategy, we performed 24 cycles of refinement not imposing any the stereochemical restraints; the second strategy included 12 cycles of refinement under the stereochemical restraints. In order to reduce the dispersion in the estimates arising in the determination of the distribution parameters from a small number of experimental data, we first used quite a large number of control reflections (up to 50% of all the reflections from the zone of a 1.8 Å resolution). It is seen from Fig. 2, that both strategies resulted in no bias; the predicted mean values of cos(Δφ) are rather close to the true ones and reflect different quality of the models obtained upon the refinement.

All the subsequent tests were performed using the same starting model, but the refinement, the estimation of the α_s and β_s parameters, and the calculation of the figures of merit were made with the use of the real experimental data. In these cases, the exact phase values were taken to be the phases calculated using all the atoms from the model of *Protein G* (including water molecules). It should be emphasized, that in this case, the sources of the errors in the phases calculated according to the model were both positional errors for the atoms of the model and some missing atoms (those of water molecules). The results obtained in this case with the use of MML estimates are shown by triangles in Figs. 3 and 4. The control set of reflections in these tests attained about 50 (Fig. 3) or 10% (Fig. 4) of the total number of reflections. It is seen that there the estimates have no bias with respect to their true values.

REDUCTION OF STATISTICAL DISPERSION OF MML ESTIMATES

As showed our calculations, a reduction of the percentage of control reflection from 50% to a more realistic 10% of the total number of reflections without introducing any bias into the estimates results in an almost double increase of the dispersion in the predicted figures of merit. The statistical dispersion in such estimates can be reduced by different methods. The first method consists in the use of some additional hypotheses on the character of the errors in the atomic model and, thus, in a decrease of the number of the parameters to be determined and, at the same time, an increase of the ratio of the number of measurements to the number of the parameters to be determined. For example, one can assume that the model includes all the structure atoms and that the errors in their coordinates are independent and distributed over the radial-symmetric normal law with the same (although unknown)

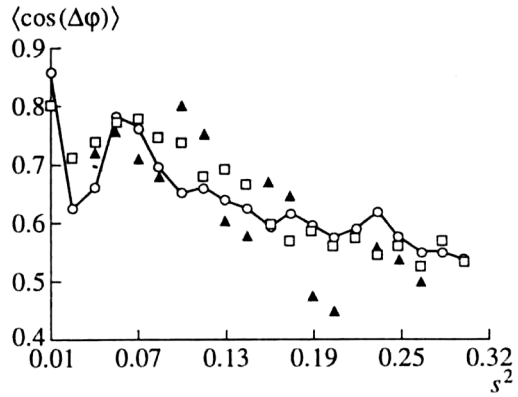

Fig. 4. Cosines of phase errors predicted from MML (triangles) and smoothed MML (squares) estimates and their true values (circles) for the model refined over 90% of the experimental data.

standard deviation ν. Assuming that the structure factors are normalized, we can show that the following relationships are valid

$$\alpha_s = \exp(-2\pi^2\nu^2 s^2), \quad \beta_s = 1 - \alpha_s^2. \qquad (4)$$

Thus, we reduced the problem to the determination of one parameter ν instead of two (α and β) for each zone of the reciprocal space. A somewhat more general variant of this approach was considered in [15], where instead of the additional assumption on the character of the coordinate errors, it was assumed that the α_s and β_s parameters obey a relationship of type (4). The advantage of the latter approach consists in the considerable improvement of the ratio of the number of measurements to the number of the parameters to be determined, which, in turn, reduces the statistical dispersion of the estimates. On the other hand, this approach requires the use of some new, not always obvious, assumptions. The problem of search for the optimum parameters becomes much more complicated and hinders the search for the global minimum. Thus, in such a situation, Brigogne and Irwin [15] had to use only the local likelihood maximization.

Another approach to the reduction of the statistical dispersion of the estimates consists in the change of the requirement of a fixed functional dependence of the α and β parameters on s to the requirement that these dependences should be smooth. Such a requirement can be taken into account, e.g., by introducing a simple correction in the functional to be maximized, in which a penalty is applied where the value lies far from the line connecting two neighbors [13]. However, in this case as well, one has to use only the local optimization.

However, the additional requirement of smoothness can be implemented in a simpler way. As earlier, the initial values of the α and β parameters for each zone are determined from the condition of the global maximum. Then, the thus obtained values are smoothed out using one of the standard schemes. The simplest smoothing procedure is illustrated by Fig. 4. The ratio $t = \alpha/\beta$, initially determined for each zone of the reciprocal space, was then substituted by the values averaged over the previous, the current, and the following zones. The thus obtained smoothed t values were used to calculate figures of merit by formula (2). It is also seen from Fig. 4 that this procedure considerably reduces the errors in the calculated estimates and allows one to use a relatively small number of control reflections without any negative effect on the refinement process.

ACKNOWLEDGMENTS

This study was supported by the Russian Foundation for Basic Research, project nos.94-04-12844 and 97-04-48319.

REFERENCES

1. D. M. Blow and F. H. C. Crick, Acta Crystallogr. **12**, 794 (1959).
2. R. J. Read, Acta Crystallogr., Sect. A: Found. Crystallogr. **42**, 140 (1986).
3. V. Yu. Lunin, Preprint No. 22, NTsBI AN SSSR (Center of Biological Studies, Soviet Academy of Sciences, Pushchino, Moscow Area, 1982).
4. V. Yu. Lunin and A. G. Urzhumtsev, Acta Crystallogr., Sect. A: Found. Crystallogr. **40**, 269 (1984).
5. V. Yu. Lunin and T. P. Skovoroda, Acta Crystallogr., Sect. A: Found. Crystallogr. **51**, 880 (1995).
6. A. G. Urzhumtsev, V. Yu. Lunin, and E. A. Vernoslova, J. Appl. Crystallogr. **22**, 500 (1989).
7. A. G. Urzhumtsev, T. P. Skovoroda, and V. Yu. Lunin, J. Appl. Crystallogr. **29**, 741 (1996).
8. V. Luzzati, Acta Crystallogr. **5**, 802 (1952).
9. G. A. Sim, Acta Crystallogr. **12**, 813 (1959).
10. R. Srinivasan and S. Parthasarathy, *Some Statistical Applications in X-ray Crystallography* (Pergamon, Oxford, 1976; Mir, Moscow, 1979).
11. R. J. Read, Acta Crystallogr., Sect. A: Found. Crystallogr. **46**, 900 (1990).
12. A. T. Brünger, Nature **355**, 472 (1992).
13. N. S. Pannu and R. J. Read, Acta Crystallogr., Sect. A: Found. Crystallogr. **52**, 659 (1996).
14. G. N. Murshudov, E. J. Dodson, and A. A. Vagin, in *Macromolecular Refinement. Proceedings of the CCP4 Study Weekend*, January, 1996, p. 93.
15. G. Bricogne and J. Irwin, in *Macromolecular Refinement. Proceedings of the CCP4 Study Weekend*, January, 1996, p. 85.

Translated by L. Man