Biophysics, Vol. 44, No. 1, 1999, pp. 18–22. Translated from Biofizika, Vol. 44, No. 1, 1999, pp. 22–26.

MOLECULAR BIOPHYSICS

Maximum Likelihood Approach to Choosing a Prior Distribution of Atomic Coordinates in Macromolecular Structures

T. E. Petrova, V. Yu. Lunin, N. L. Lunina, and T. P. Skovoroda

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Puschino, Moscow Region, 142292 Russia

Received October 10, 1997

Abstract—We discuss the potentiality of the maximal likelihood approach to selection of a prior distribution for the coordinates of atoms of macromolecular structures in solving the phase problem of X-ray analysis. Different approaches in selection of the likelihood function are considered. It is shown that the maximum likelihood in some cases fails to provide a satisfactory prior distribution, and a comprehensive study of the limits of applicability is necessary in every case.

Key words: structures of biopolymers, phase problem, probabilistic approach

INTRODUCTION

One of the key problems of X-ray crystallography is the phasing problem, or the problem of assessing the phases of complex coefficients (structural factors) the of Fourier expansion of electron distribution in the studied crystal. Absolute values of these coefficients are measured in the experiment. In the statistical approach, practically, the phasing problem of X-ray crystallography is assessed in three stages. At the first stage, an ensemble of potential structures is introduced, and probabilities are assigned to these structures. The following model is usually used therewith: the structure under study is considered as a series of independent tests; each test consists in putting any atom of the structure, at random and independently, in an elementary cell with some prior probability distribution. At the second stage, the joint distribution of the structural factors is obtained, basing on which conditional distributions for phase invariants are calculated under the assumption that the absolute values of structural factors take their experimental values. At this stage, *a priori* information concerning the structures is transformed into the information on phase invariant distributions. The third stage consists in evaluation of the phase invariants of the structure with the help of the distribution obtained at the second stage.

The choice of a prior for atomic coordinates is essential in this technique. The uniform prior, which is successfully employed for the low-molecular structures, is in poor agreement with the experimental data on macromolecular structures for low and medium resolution. In recent years, the maximum likelihood principle was extensively used in choosing the prior that agrees best with the experimental data [1, 2, 5]. To every prior, a value called likelihood can be assigned, defined as the probability, or, more rigorously, the probability density, of the event that the absolute values of the structural factors are equal to their observed experimental values:

$$L(q(\mathbf{r})) = P\left\{|F_i| = |F_i^{\text{obs}}|,\ i = 1, M\right\}. \qquad (1)$$

According to this principle, as a prior for atomic coordinates of the studied structure one should choose the distribution $q(\mathbf{r})$ for which the likelihood attains its maximum. However, a more detailed investigation indicates that this principle should be used with caution and only after a careful study of its applicability limits. Below, we consider three examples, in which the results provided by this principle are correct only within certain limits.

CALCULATION OF THE LIKELIHOOD FUNCTION

Mathematically, the problem of calculation of the prior that corresponds to the maximal likelihood is quite complicated. If there were an analytical expression for a joint distribution of absolute values and phases of structural factors, then it would be possible to calculate the likelihood by integrating the joint distribution over phases with the absolute values given. However, it is very difficult to obtain an exact expression for the distribution of absolute values and phases. An asymptotic Edgeworth series provides sufficient accuracy only for small deviations of unitary structural factors from their expected values: $\Delta U \sim N^{-0.5}$ [3]. For larger deviations $\Delta U \sim 1$, there is an approximate expression for the joint distribution of structural factors, which was obtained by Bricogne using the saddle-point method [4]. However, his estimation is difficult to employ in practice, since the final expression contains an implicit function that is the solution of a large set of nonlinear equations.

Mathematical problems also arise in integrating the distribution obtained over phases, for which substantial approximations are necessary. For instance, many authors used "diagonal approximation" [1], in which the nondiagonal elements of a covariance matrix are put equal to zero. This implies that structural factors are independent variables, an assumption that results in the loss of a substantial part of phase information.

An alternative approach to likelihood calculation is to proceed from the ordinary likelihood function (1) to the so-called generalized likelihood, which is calculated in computer simulations [5]. By definition, the generalized likelihood is the probability of the absolute values of structural factors being sufficiently close to the experimental absolute values:

$$L(q(\mathbf{r}),\omega) = P\left\{C\left(\left\{F_{\mathbf{h}}^{\text{calc}}\right\},\left\{F_{\mathbf{h}}^{\text{obs}}\right\}\right) \geq \omega\right\}, \quad (2)$$

where ω is the approximation parameter, which is chosen in the particular study; C is a certain measure of distance between two sets of structural factors. In computer simulations, a large number of model structures are generated, with atomic coordinates distributed over the elementary cell according to the given prior $q(\mathbf{r})$, whereupon the probability (2) is approximately calculated as the ratio of the number of models with C greater than the given level to the total number of models generated. It should be noted that this technique demands substantial computing power and also requires an additional study of the dependence of the outcome on the model parameters.

Below we illustrate these methods of likelihood calculation with examples.

EVALUATION OF EXPECTED PHASE ERRORS USING AN APPROXIMATE ATOM MODEL OF THE STRUCTURE

The phases of structural factors at different stages of determining the structure are often calculated for an approximate atomic model. In so doing, it is necessary to estimate the expected phase errors. To this end, several authors of the present paper have employed a probabilistic model [6], in which the coordinates of every atom were distributed with the same average $\mathbf{r}_j^{\text{mod}}$, and the mean error was unknown but equal for each atom [6]. For example, the error distribution can be Gaussian, and in this case by choosing the mean error one simultaneously defines the prior for the atomic coordinates. It has been shown mathematically that all unknown errors in the model as well as the scale coefficients for reflexes belonging to the same resolution zone comply with a two-parameter distribution. The unknown parameters were estimated using the maximum likelihood principle. The joint distribution of the absolute values of structural factors was obtained in the diagonal approximation [1, 2]. The algorithm was tested on known structures; in these tests the expected phase errors were calculated using the estimated parameter values and compared with the observed phase errors. The calculated average phase errors were found to be close to their correct values. However, after the model was refined, the calculated estimates were significantly below the correct values (Fig. 1). It should be noted that the disagreement between calculated and observed errors can be reduced by selecting a set of reflexes and saving them from refinement [7]; the likelihood in this case is calculated taking into account only the selected reflexes.

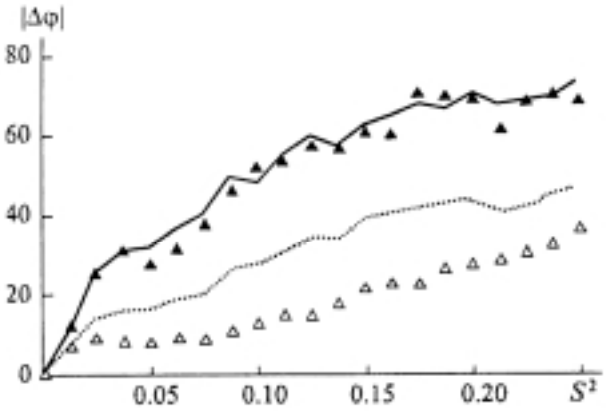


Fig. 1. Zone average of the predicted and the observed phase errors for the actinidine model. The predicted (dark triangles) and observed (solid line) phase errors for the model with the average absolute error of atomic coordinates equal to 0.79 Å. The predicted (light triangles) and the observed (dotted line) phase errors for the same model after refinement.

GENERALIZED LIKELIHOOD FOR REGION MASK SELECTION FROM SEVERAL ALTERNATIVES

A phasing problem, when it is solved using FAM technique [9] or other *ab initio* methods of phase estimation, requires selection of a single region mask from several alternative region masks calculated on different phase sets [5]. A prior distribution of atomic coordinates corresponds to each region mask, this distribution is constant within this mask and equal to zero beyond it. Therefore, the problem of selection of a mask from several alternatives is equal to the problem of prior selection from several alternative distributions.

For each potential prior, the generalized likelihood (defined above) was calculated in computer simulations. We expected that the best region mask would correspond to the maximal number of FAM variants with a high correlation of magnitudes, i.e., to the maximal value of the generalized likelihood. This was confirmed in tests with the experimental data obtained for RNase SA with 16 Å resolution using the masks isolating 60% of the elementary cell volume (Fig. 2a), which is approximately equal to the molecule volume. However, when the masks built for the same phase sets but isolating 30% of the elementary cell volume were compared, the maximal value of the general likelihood did not correspond to the best mask (Fig. 2b).

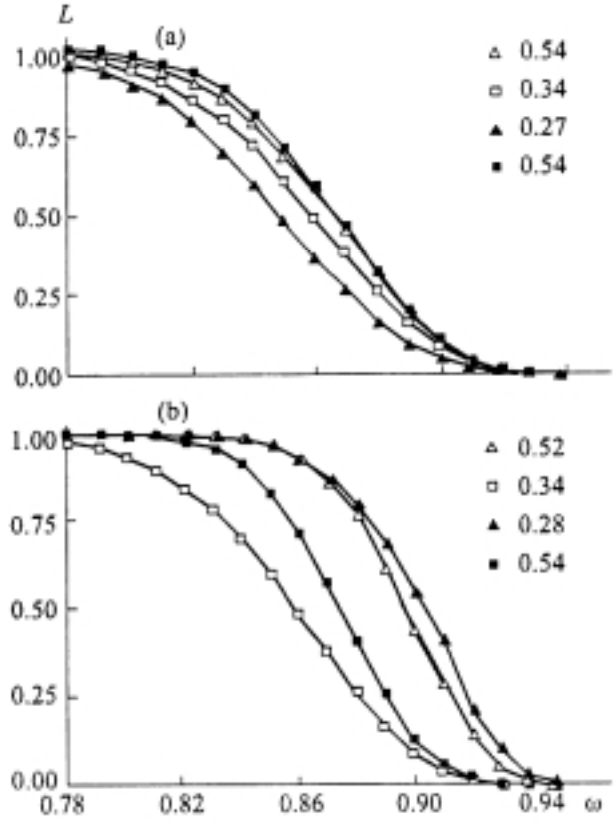
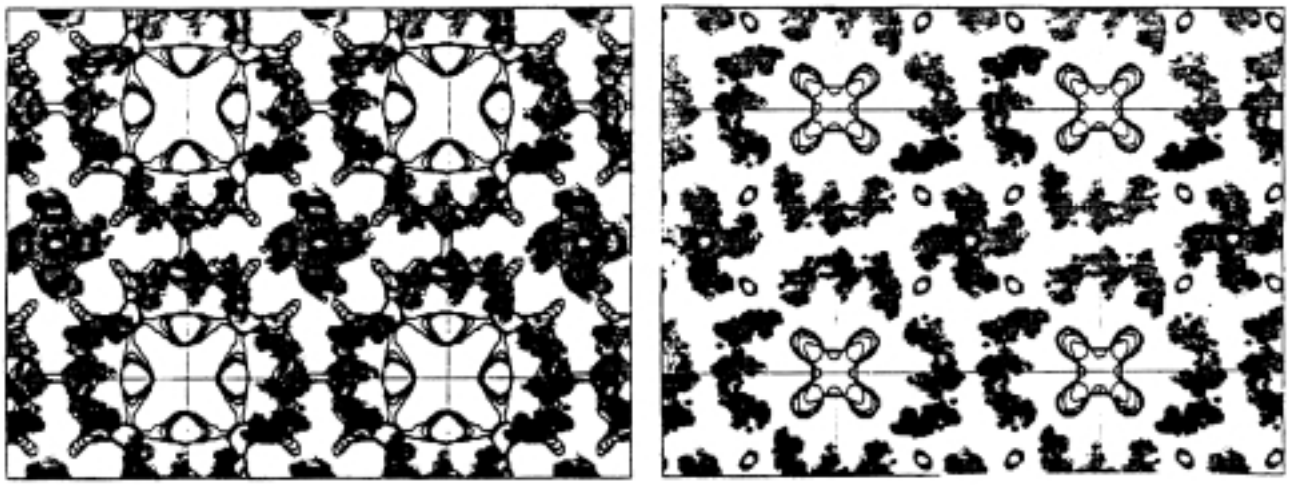


Fig. 2. The ω-dependence of generalized likelihood values calculated for 100 atoms generated according to different region masks; (a) the masks isolate 60% of the elementary cell volume; (b) the masks isolate 30% of the elementary cell volume. Different curves correspond to different values of the mask quality, the average coefficient of phase correlation (specifed at the symbols) during generation of atoms controlled by the mask.

APPROXIMATION OF THE PRIOR WITH THE LIKELIHOOD GRADIENT

For low resolutions, the prior map can be directly used to determine the supposed regions with high and low atom concentration. Since calculation of a prior distribution corresponding to the global minimum of the maximum likelihood function is a complicated problem, we have tried to start from an easier problem: to find a distribution $q(\mathbf{r})$ for which the likelihood value obtained would be greater than the value obtained for the uniform distribution. Such a distribution can be obtained if one moves in the prior distribution space following the likelihood gradient, starting from the point corresponding to the uniform distribution. Thus, the desired distribution is sought in the form:



(a) (b)

Fig. 3. Maps of the likelihood gradient. The black contour corresponds to the lines of the level isolating (a) 25% of the points of the elementary cell with maximal values of the likelihood gradient function, and (b) 5% of the points of the elementary cell with minimal values of the likelihood gradient function. The atoms of the model are shown in grey.

$$q_\lambda(\mathbf{r}) = q_0(\mathbf{r}) + \lambda \frac{\delta L}{\delta q(\mathbf{r})} \tag{3}$$

where λ should be sufficiently small. The function $q(\mathbf{r})$ can be calculated via its Fourier coefficients. The likelihood gradient can be expressed through the derivatives of the likelihood with respect to the Fourier coefficients of the prior:

$$\frac{\delta L}{\delta q(\mathbf{r})} = \sum_{\mathbf{h}} \frac{\partial L}{\partial G_{\mathbf{h}}} \exp(2\pi i(\mathbf{h},\mathbf{r})). \tag{4}$$

It is necessary to remark that the gradient of the likelihood function is not zero only when the origin of coordinates and the enantiomorph are fixed. One can easily see from (3) that functions $q_\lambda(\mathbf{r})$ and $\frac{\delta L}{\delta q(\mathbf{r})}$ have coinciding maxima and minima. Therefore, in order to find the regions of the elementary cell with supposedly high and low atom concentration, it is sufficient to build the map of the gradient of the likelihood function, which can be calculated with equation (4).

In this study we have paid special attention to obtaining the analytical expressions for the function $\frac{\partial L}{\partial G_{\mathbf{h}}}$. We have demonstrated that, for the case when the deflections of unitary structural factors from their expected values are on the order of $\Delta U - N^{-4}$, an asymptotic estimate can be obtained that does not include implicit functions in the expressions for joint distributions of structural factors. It should be stressed that when the order of deflection of the unitary structural factors for their expected values comply with the limits above, the large enough values of structural factors are not excluded from consideration.

In order to give insight into the equations obtained, we present the expression for $\frac{\partial L}{\partial G_{\mathbf{h}}}$ in the case when the phase of the reflex $\mathbf{h}$ is an absolute semi-invariant:

$$\frac{\partial L}{\partial G_{\mathbf{h}}} = \begin{cases} M(\mathbf{h}), & \mathbf{h} \notin S \\ E_{\mathbf{h}}^3 M(\mathbf{h}) + \omega(\mathbf{h},\mathbf{h})(E_{\mathbf{h}}^4 - 3), & \mathbf{h} \in S, \end{cases} \tag{5}$$

where

$$M(\mathbf{h}) = \sum_{\mathbf{k} \neq \mathbf{h}} \omega(\mathbf{k},\mathbf{h})(E_{\mathbf{k}}^2 - 1),$$

$$\omega(\mathbf{k},\mathbf{h}) = \frac{\{\text{number of such pair}(\mu,\nu), \text{that } \mathbf{R}_\mu^T\mathbf{k} + \mathbf{R}_\nu^T\mathbf{k} + \mathbf{h} = 0\}}{2\{\text{number of such } \mu, \text{that } \mathbf{R}_\mu^T\mathbf{k} = \mathbf{k}\}\{\text{number of such } \nu, \text{that } \mathbf{R}_\nu^T\mathbf{h} = \mathbf{h}\}}, \tag{6}$$

where E_h is the normalized structural factor; $\mathbf{R}_\mu$ is the symmetry matrix.

In Fig. 3 we show the maps of the likelihood gradient calculated by equations (5)–(6) taking into account 24 central symmetrical reflexes, which are the semi-invariants of space group I432, and using the positions of C_α atoms in the known complex AspRS–tRNA [10, 11]. Figure 3 shows that the function thus calculated has its maxima in the regions where atoms are found, and has its minima in the regions where atoms are almost absent. However, additional maxima of the likelihood function gradient are found also in the regions of the elementary cell where the atom concentration is insignificant.

CONCLUSIONS

The three examples considered demonstrate that the acceptance of the maximum likelihood principle for selecting a prior distribution of atomic coordinates is justified. In many cases, with this principle one can obtain the correct result. However, it is necessary to take into account that, as with any statistical principle, the maximum likelihood principle does not guarantee a correct result in every specific case, as revealed in the tests above. Addressing any practical problem, it is necessary to study in detail the applicability limits of this principle.

ACKNOWLEDGMENTS

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

REFERENCES

1. Bricogne, G. and Gilmore, C.J., *Acta Cryst.*, 1990, A46, p. 284.
2. Lunin, V.Yu. and Urzhumtsev, A.G., *Acta Cryst.*, 1984, A40, p. 269.
3. Klug, A., *Acta Cryst.*, 1958, vol. 11, p. 515.
4. Bricogne, G., *Acta Cryst.*, 1984, A40, p. 410.
5. Lunin, V.Yu., Lunina, N.L., Petrova, T.E., Urzhumtsev, A.G., and Podjarny, A., *Acta Cryst.*, 1998, D54, p. 726.
6. Lunin, V.Yu. and Skovoroda, T.P., *Acta Cryst.*, 1995, A51, p. 880.
7. Brunger, A.T., *Nature* (London), 1992, vol. 355, p. 472.
8. Read, R.J., *Acta Cryst.*, 1986, A42, p. 140.
9. Lunin, V.Yu., Lunina, N.L., Petrova, T.E., Vernoslova, E.A., Urzhumtsev, A.G., and Podjarny, A., *Acta Cryst.*, 1995, D51, p. 896.
10. Moras, D., Lorber, B., Romby, P., Ebel, J.-P., Giegé, R., Lewitt-Bentley, A., and Roth, M., *J. Biomol. Struct. Dynam.*, 1983, vol. 1, p. 209.
11. Urzhumtsev, A.G., Podjarny, A.D., and Navaza, J., *J. CCP4 ESF-EACBM Newslett. Protein Crystallog.*, 1994, vol. 30, p. 29.