АКАДЕМИЯ НАУК СССР
НАУЧНЫЙ ЦЕНТР БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ
НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

ПРЕПРИНТ

В.Ю.ЛУНИН

ИСПОЛЬЗОВАНИЕ МЕТОДА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ ДЛЯ ОЦЕНКИ ОШИБОК ПРИ ОПРЕДЕЛЕНИИ ФАЗ В КРИСТАЛЛОГРАФИИ БЕЛКА

ПУЩИНО · 1982

УДК 548.73

Рассматривается вопрос об оценке надежности фазовой информации, полученной по атомной модели структуры. Предложен алгоритм оценки параметров, характеризующих величину ошибок в модели структуры. Получаемые значения параметров позволяют оценить ошибку в фазах структурных факторов, рассчитанных по модели.

ВВЕДЕНИЕ

1. Одним из этапов работы по расшифровке структуры белка методом рентгеноструктурного анализа является определение фаз структурных факторов (фаз коэффициентов Фурье функции распределения электронной плотности в кристалле белка). В настоящее время для определения этих фаз существует ряд методов, опирающихся на различного рода дополнительную информацию о структуре белка (изоморфное замещение, аномальное рассеяние, некристаллографическая симметрия, уравнения Сейра и т.д.). В силу различных причин, связанных как с ошибками рентгеновского эксперимента, так и с приближенным характером доступной нам информации о структуре, эти методы часто не дают возможности определить точное значение фазы, а дают лишь область наиболее вероятных ее значений. На практике информацию о фазе, полученную на основе какого-то метода, обычно характеризуют плотностью распределения вероятностей $P(\varphi)$ значений фазового угла /6,7,12/. Минимальную среднеквадратичную ошибку при расчете карт электронной плотности

$$\rho(\vec{r}) = \frac{1}{V} \sum_{\vec{s}} F(\vec{s})\, e^{i\varphi(\vec{s})}\, e^{-2\pi i(\vec{s},\vec{r})}$$

($F(\vec{s})$ – экспериментально полученные модули структурных факторов, $\vec{s}$ – точки "обратной" решетки) дает в этом случае синтез Фурье с коэффициентами

$$F(\vec{s})\, m(\vec{s})\, e^{i\varphi_{best}(\vec{s})}$$

где

$$m e^{i\varphi_{best}} = \langle e^{i\varphi} \rangle = \int_0^{2\pi} e^{i\varphi} P(\varphi)\, d\varphi .$$

Введенная таким образом величина m – "показатель достоверности определения фазы" – часто используется как характеристика надежности определения фазы. Другой характеристикой надежности может служить средняя величина ошибки, получаемой при использовании фазы φ_{best} :

$$\langle |\varphi - \varphi_{best}| \rangle = \int_0^{2\pi} |\varphi - \varphi_{best}| \, P(\varphi) \, d\varphi .$$

2. В данной работе мы обсудим, как установить, насколько близки к "истинным" фазы структурных факторов, рассчитанных по атомной модели структуры, если эта модель содержит не все атомы и в координатах включенных в нее атомов есть ошибки. Для этого мы попытаемся найти распределение $P(\varphi)$ фазы структурного фактора, соответствующее такой ситуации.

На протяжении ряда лет в серии работ /8, 9, 1/ были изучены распределения вероятностей для модуля F и фазы φ структурного фактора при некоторых вероятностных предположениях о структуре и причинах, вызывающих ошибки в фазах. При этом оказалось, что в значительном числе случаев плотность этих распределений имеет одинаковый вид:

$$P(F,\varphi) = \frac{F}{\pi\beta} \exp\left\{-\frac{F^2+\alpha^2 F_c^2}{\beta}\right\} \exp\left\{\frac{2\alpha}{\beta} F F_c \cos(\varphi-\varphi_c)\right\}, \quad (1)$$

где F_c и φ_c – известные величины (например, модуль и фаза структурного фактора, рассчитанного по модели), а α и β – некоторые параметры, характеризующие источник ошибок в определении фазы (например, параметры, характеризующие распределение ошибки в координатах атомов модели, или параметры, характеризующие коэффициенты приведения данных к абсолютной шкале, и т.п.). Поскольку эксперимент дает нам значения модулей структурных факторов, то фазовую информацию можно задать условным распределением вероятностей фазы φ при условии, что модуль F принял измеренное в эксперименте значение:

$$P(\varphi|F) = P(F,\varphi) \Big/ \int_0^{2\pi} P(F,\varphi)\,d\varphi = \frac{1}{2\pi I_0\left(\frac{2\alpha}{\beta} F F_c\right)} \exp\left\{\frac{2\alpha}{\beta} F F_c \cos(\varphi-\varphi_c)\right\} . \quad (2)$$

Отметим, что такой же вид имеют распределения $P(\varphi)$ предложенные Симом и Бриконем /7, 12/.

Основным препятствием при попытке использовать это условное распределение (2) служит то, что мы, как правило, не знаем значения параметров α и β. Поэтому возникает задача определения параметров распределения (1), одному из способов решения которой и посвящена данная работа.

В качестве основы для определения параметров α и β мы будем использовать тот факт, что из совместного распределения (1) величин F и φ следует распределение для модуля структурного фактора:

$$P(F) = \int_0^{2\pi} P(F,\varphi)\,d\varphi = \frac{2F}{\beta} \exp\left\{-\frac{F^2+\alpha^2 F_c^2}{\beta}\right\} I_0\left(\frac{2\alpha}{\beta} F F_c\right) \quad (3)$$

с теми же параметрами α и β. С другой стороны, эксперимент дает нам значения модулей структурных факторов $F_e(\vec{s})$ для различных точек $\vec{s}$ обратного пространства. Рассматривая эти экспериментальные значения $F_e(\vec{s})$ как реализации независимых случайных величин $F(\vec{s})$, распределенных по закону (3), мы применим в § 2 для оценки параметров распределения (3) принцип максимума функции правдоподобия /3/.

3. Приведенные в § 5 результаты тестов показывают, что распределение (2) с параметрами α и β, определенными из максимума функции правдоподобия, хорошо описывают реальную ситуацию в случае, когда ошибки в координатах атомов модели случайны и независимы. В этом случае имеется хорошее соответствие между теоретической оценкой средней ошибки $|\varphi - \varphi_c|$ (исходя из распределения $P(\varphi)$) и реально возникающей ошибкой в фазах. В то же время в случае, когда φ_c является фазой структурного фактора, рассчитанного по модели, прошедшей уточнение в обратном пространстве /10/, теоретическая оценка ошибки имеет тенденцию оказываться примерно в два раза ниже реально возникающей ошибки. По-видимому, это связано с тем, что после уточнения модели ошибки в координатах атомов не являются независимыми и предположение о независимости структурных факторов $F(\vec{s})$ при различных $\vec{s}$ становится слишком грубым.

Отметим еще, что в случае, когда модель содержит слишком большие ошибки, определение параметров α и β из максимума функции правдоподобия выливается в шкалирование данных по Вильсону /11/.

Автор благодарен Г.Н.Борисюк за ценные консультации по теории вероятностей и статистике и Е.А.Вернословой и А.Г.Уржумцеву за помощь в проведении тестов и обсуждении результатов.

§ 1. Вывод основного распределения

1. В этом параграфе мы рассмотрим некоторые ситуации, когда совместное распределение модуля и фазы отдельного структурного фактора имеет вид:

$$P(F,\varphi) = \frac{F}{\pi\beta} \exp\left\{-\frac{F^2+\alpha^2 F_c^2}{\beta}\right\} \exp\left\{\frac{2\alpha}{\beta} F F_c \cos(\varphi-\varphi_c)\right\}.$$

В 1949 г. Вильсон /11/ показал, что если рассмотреть все структуры как равновероятные (более точно: если рассматривать координаты всех атомов структуры как независимые равномерно распределенные случайные величины), то для фиксированной точки обратного пространства $\vec{s}$ плотность распределения вероятностей величины $I = F_a^2(\vec{s})$ есть

$$P(I) = \frac{1}{\sigma_N^2} e^{-I/\sigma_N^2}, \qquad (4)$$

где

$$\sigma_N^2 = \sigma_N^2(\vec{s}) = \sum_{j=1}^{N} f_j^2(s)\, e^{-B_j^t \frac{s^2}{2}}, \quad s = |\vec{s}|, \tag{4}$$

$f_j(s)$ – фактор атомного рассеяния j –го атома, B_j^t – температурный фактор j –го атома. (Говоря более строго, правая часть равенства (4) представляет собой главный член асимптотики плотности распределения вероятностей величины I, когда число атомов в структуре $N \to \infty$).

Предполагая, что в эксперименте измерение модулей структурных факторов происходит в относительной шкале (то есть мы получаем в результате эксперимента величину $F(\vec{s}) = k F_a(\vec{s})$, где $F_a(\vec{s})$ – значение модуля структурного фактора в абсолютной шкале, а k – некоторый шкальный коэффициент, который, вообще говоря, неизвестен), мы из распределения (4) получаем плотность распределения вероятностей величины $F(\vec{s}) = k\sqrt{I}$ в виде

$$P(F) = \frac{2F}{k^2 \sigma_N^2}\, e^{-\frac{F^2}{k^2 \sigma_N^2}}, \tag{5}$$

то есть плотность вида (1) с $\alpha = 0$ и $\beta = k^2 \sum_{j=1}^{N} f_j^2(s)\, e^{-B_j^t \frac{s^2}{2}}$.

2. Предположим теперь, что у нас имеется некоторая частичная атомная модель структуры, то есть мы предположим, что нам известны (возможно, с ошибкой) координаты части атомов структуры. Пусть $F_c(\vec{s})\, e^{i\varphi_c(\vec{s})}$ – структурные факторы, рассчитанные по такой модели. Частные случаи такой ситуации рассматривались Луззати в /9/ (в модель включены все атомы, но координаты содержат ошибку) и Симом /7/ (модель неполная, но координаты без ошибок). Общий случай рассмотрен в /1/. Аналогично этой работе (см. также /5/) можно показать, что если ошибки в координатах атомов модели считать независимыми одинаково распределенными случайными величинами, то совместное распределение вероятностей модуля F и фазы φ отдельного структурного фактора (мы обозначим $F_c\, e^{i\varphi_c}$ значение соответствующего структурного фактора, рассчитанного по модели) имеет плотность вида (1) с $\alpha = \mu$, $\beta = \sigma_N^2 - \mu^2 \sigma_M^2$. Здесь параметр $\mu = \langle \cos 2\pi(\vec{s}, \vec{\Delta}) \rangle$ характеризует грубость ошибки в координатах ($\mu = 1$, если ошибок нет),

$$\sigma_N^2 = \sum_{j=1}^{N} f_j^2(s)\, e^{-B_j^t \frac{s^2}{2}}, \quad \sigma_M^2 = \sum_{j=1}^{M} f_j^2(s)\, e^{-B_j^t \frac{s^2}{2}},$$

N – число атомов в структуре, M – число атомов в модели.

Переход к относительной шкале измерения модулей структурных

факторов приводит к плотности вида (1) с $\alpha = k\mu$, $\beta = k^2(\sigma_N^2 - \mu^2\sigma_M^2)$.

В работе Луззати /8/ рассмотрен еще ряд ситуаций, приводящих к распределению вида (1).

3. Приведенные выше плотности $P(F, \varphi)$ представляют собой плотности распределения вероятностей для модуля и фазы отдельного структурного фактора $F(\vec{s})\, e^{i\varphi(\vec{s})}$. При этом параметры α, β, F_c, φ_c, фигурирующие в (1), зависят от точки обратного пространства $\vec{s}$, то есть для разных $\vec{s}_1$ и $\vec{s}_2$ мы имеем различные распределения вероятностей. Однако при сферически симметричных факторах атомного рассеяния $f_j(s)$, изотропных температурных факторах B_j^t и сферически симметричном распределении ошибки $\vec{\Delta}$ в координатах атомов модели функции α и β зависят только от $s = |\vec{s}|$ и мы (с определенной погрешностью) можем считать, что в "тонком" сферическом слое $s_1 \leq |\vec{s}| \leq s_2$ функции $\alpha(s)$ и $\beta(s)$ постоянны.

Мы будем далее рассматривать такой слой. Пусть в нем содержится K точек обратной решетки: $\{\vec{s}_j\}_{j=1}^K$. Будем обозначать

$$F_j = F(\vec{s}_j),\ \varphi_j = \varphi(\vec{s}_j),\ F_{c,j} = F_c(\vec{s}_j),\ \varphi_{c,j} = \varphi_c(\vec{s}_j).$$

§ 2. Функция правдоподобия

1. Итак, в рамках рассмотренной в предыдущем параграфе ситуации мы имеем K случайных величин $\{F_j\}_{j=1}^K$ каждая из которых имеет плотность распределения вероятностей:

$$P_j(F) = \frac{2F}{\beta} \exp\left\{-\frac{F^2 + \alpha^2 F_{c,j}^2}{\beta}\right\} I_0\left(2\frac{\alpha}{\beta} F F_{c,j}\right), \qquad (6)$$

где $\{F_{c,j}\}_{j=1}^K$ – известные нам величины, α и β – неизвестные параметры, подлежащие определению, I_0 – модифицированная функция Бесселя первого рода.

Мы будем предполагать далее, что выполнено

У с л о в и е 1. Случайные величины $\{F_j\}_{j=1}^K$ независимы и распределены по закону (6).

Отметим, что, строго говоря, величины F_j, возникающие в § 1, не являются независимыми, так как они выражаются через одни и те же исходные случайные величины $\{\vec{\Delta}_j\}_{j=1}^M$. Однако поскольку коэффициенты корреляции между ними имеют порядок $N^{-\frac{1}{2}}$, то (ограничиваясь далее везде при рассмотрении плотностей распределений вероятностей главными членами асимптотики при $N \to \infty$) мы будем применять результаты, вытекающие из условия (1), к ситуациям, рассмотренным в предыдущем параграфе.

Отметим также, что излагаемые далее результаты применимы

не только в ситуациях, рассмотренных в § 1, но и в любом случае, когда выполнено Условие 1.

2. Мы будем теперь рассматривать экспериментально полученные модули структурных факторов $F_{e,j} = F_e(\vec{\gamma}_j)$ как реализации независимых случайных величин F_j с распределениями (6). В этом случае функция правдоподобия имеет вид:

$$\Psi(\alpha,\beta) = \prod_{j=1}^{K} P_j(F_{e,j}) \quad .$$

Принцип максимума функции правдоподобия предлагает в качестве оценки параметров распределений (6) значения α и β, при которых функция $\Psi(\alpha,\beta)$ достигает максимума. Мы будем искать этот максимум в области $\alpha \geqslant 0$, $\beta > 0$.

Вместо максимума функции $\Psi(\alpha,\beta)$ нам будет удобнее искать максимум функции

$$\tilde{\Psi}(\alpha,\beta) = \frac{1}{K}\left[\ln\Psi(\alpha,\beta) - K\ln 2 - \sum_{j=1}^{K}\ln F_{e,j}\right] =$$
$$= \ln\frac{1}{\beta} - \frac{1}{\beta}\cdot\frac{1}{K}\sum_{j=1}^{K}F_{e,j}^2 - \frac{\alpha^2}{\beta}\cdot\frac{1}{K}\sum_{j=1}^{K}F_{c,j}^2 + \frac{1}{K}\sum_{j=1}^{K}\ln I_0\left(\frac{2\alpha}{\beta}F_{e,j}F_{c,j}\right),$$

который достигается при тех же значениях α и β, что и максимум функции $\Psi(\alpha,\beta)$.

Введем обозначения

$$A = \frac{1}{K}\sum_{j=1}^{K}F_{c,j}^2 \quad , \quad B = \frac{1}{K}\sum_{j=1}^{K}F_{e,j}^2 \, ,$$
$$C = \frac{1}{K}\sum_{j=1}^{K}F_{c,j}F_{e,j} \, , \quad D = \frac{1}{K}\sum_{j=1}^{K}F_{c,j}^2F_{e,j}^2 \qquad (7)$$

и перейдем к новым переменным

$$u^2 = \frac{1}{\beta} \quad , \quad v^2 = \frac{\alpha^2}{\beta} \quad (u \geqslant 0 \, , \, v \in (-\infty, \infty)) \, .$$

В таком случае мы приходим к задаче: найти максимум функции

$$Q(u,v) = 2\ln u - Bu^2 - Av^2 + \frac{1}{K}\sum_{j=1}^{K}\ln\left[I_0(2uvF_{e,j}F_{c,j})\right]$$

в области $u \in [0,\infty)$, $v \in (-\infty,\infty)$.

3. Функция $Q(u,v)$ определена и непрерывно дифференцируема при $u > 0$ и четна по v: $Q(u,-v) = Q(u,v)$. Кроме того, можно видеть (см. Лемму 1 в Приложении 1), что $Q(u,v)$ стремится к $-\infty$ при приближении аргументов к границе рассматриваемой области по u и v. Поэтому максимум функции $Q(u,v)$ достигается во внутренней точке области и в точке максимума выполнены необходимые условия максимума – частные производные $\partial Q/\partial u$ и $\partial Q/\partial v$ обращаются в нуль.

З а м е ч а н и е. Здесь и далее мы будем вести рассмотрение для случая, когда нельзя подобрать такой единый для всех рефлексов шкальный коэффициент k, что $F_{e,j} = k F_{c,j}$ при всех $j = 1, \dots, K$, то есть модель не идентична структуре. Мало интересный с практической точки зрения случай, когда такое значение k существует, разобран в Приложении 1.

Дифференцируя функцию $Q(u,v)$ (используя при этом равенство $[I_0(t)]'_t = I_1(t)$) и вводя функцию

$$\Lambda(t) = \frac{2}{K} \sum_{j=1}^{K} F_{c,j} F_{e,j} \frac{I_1(2t F_{c,j} F_{e,j})}{I_0(2t F_{c,j} F_{e,j})}, \tag{8}$$

запишем необходимые условия максимума функции $Q(u,v)$ в виде:

$$\begin{cases} \frac{1}{u} - Bu + \frac{1}{2} v \Lambda(uv) = 0 \\ -Av + \frac{1}{2} u \Lambda(uv) = 0 \end{cases} . \tag{9}$$

4. Нетрудно видеть, что одним из решений системы (9) является: $u = 1/\sqrt{B}$, $v = 0$, что соответствует $\alpha = 0$, $\beta = B$. Чтобы установить, является ли эта точка точкой максимума для функции $Q(u,v)$, мы разложим $Q(u,v)$ в ряд Тейлора в окрестности этой точки. Учитывая, что /4/

$$I_0(t) = 1 + \frac{1}{4} t^2 + O(t^4) \quad \text{при} \quad t \to 0,$$

мы получаем, проведя вычисления, что

$$Q(u,v) = (-\ln B - 1) - 2B\left(u - \frac{1}{\sqrt{B}}\right)^2 - \frac{AB - \mathcal{D}}{B} v^2 + \tag{10}$$

$$+ \text{ члены более высокого порядка малости при } v \to 0, u \to 1/\sqrt{B},$$

где величины A, B, $\mathcal{D}$ определены равенствами (7).

Из разложения (10) видно, что характер точки $(1/\sqrt{B}, 0)$ определяется знаком величины

$$\Omega = AB - \mathcal{D} . \tag{11}$$

При $\Omega > 0$ точка $(1/\sqrt{B}, 0)$ является точкой (локального) максимума функции $Q(u,v)$. Если же $\Omega < 0$, то точка $(1/\sqrt{B}, 0)$ является седловой точкой и функция $Q(u,v)$ достигает максимума в другой точке.

§ 3. Шкалирование по Вильсону

1. Итак, мы установили, что одной из стационарных точек функции правдоподобия $\Psi(\alpha, \beta)$ всегда является точка $(0, B)$.

Этой точке соответствует распределение величины F вида

$$P(F) = \frac{2F}{\beta} e^{-\frac{F^2}{\beta}},$$

то есть распределение Вильсона (5) с

$$k^2 \sum_{j=1}^{N} f_j^2(s) e^{-B_j^t \frac{s^2}{2}} = \frac{1}{K} \sum_{j=1}^{K} F_{e,j}^2 . \qquad (12)$$

Считая, что температурные факторы всех атомов одинаковы, мы получаем, логарифмируя равенство (12), для каждого сферического слоя в обратном пространстве уравнение

$$2 \ln k - B^t \frac{s^2}{2} = \ln \left[\frac{1}{K} \sum_{j=1}^{K} F_{e,j}^2 \Big/ \sum_{j=1}^{N} f_j^2(\hat{s}) \right] , \qquad (13)$$

где $\hat{s}$ – среднее значение величины $|\vec{s}|$ для данного слоя.

Решение в смысле "метода наименьших квадратов" системы (13) есть не что иное, как общепринятое шкалирование экспериментальных данных "из графика Вильсона" /2/.

2. Как было установлено, точка $(0, B)$ является точкой локального максимума функции правдоподобия при

$$\Omega = \left(\frac{1}{K}\sum_{j=1}^{K} F_{c,j}^2\right)\left(\frac{1}{K}\sum_{j=1}^{K} F_{e,j}^2\right) - \frac{1}{K}\sum_{j=1}^{K} F_{c,j}^2 F_{e,j}^2 > 0$$

и является седловой точкой при $\Omega < 0$. В Приложении 2 показано, что если мы находимся в рамках ситуации, рассмотренной в /1/ (координаты атомов $\vec{r}_j$ – независимые равномерно распределенные случайные величины, ошибки $\vec{\Delta}_j$ – независимые одинаково распределенные величины), то при некоторых упрощающих предположениях

$$\langle \Omega \rangle \simeq -\left(1 - \frac{1}{K}\right) \mu^2 k^2 \sigma_M^2 .$$

Это означает, что, как правило, величина Ω отрицательна, то есть точка $(0, B)$ не является точкой максимума функции правдоподобия. При увеличении ошибок в модели структуры величина μ стремится к нулю, то есть и $\langle \Omega \rangle$ стремится к нулю. Поэтому положительные значения величины Ω можно ожидать для достаточно грубых моделей, когда модель перестает нести информацию об истинной структуре. Таким образом, при слишком грубой модели (или при отсутствии ее) оценка параметров распределения (6) из максимума функции правдоподобия приводит к шкалированию данных из графика Вильсона.

§ 4. Алгоритм поиска максимума функции правдоподобия

1. Вернемся теперь к системе уравнений для определения стационарных точек функции правдоподобия:

$$\begin{cases} \frac{1}{u} - Bu + \frac{1}{2} v \Lambda(uv) = 0, \\ -Av + \frac{1}{2} u \Lambda(uv) = 0. \end{cases}$$

Считая, что $v \neq 0$, мы преобразуем систему к виду:

$$\begin{cases} Bu^2 - \frac{1}{2} uv \Lambda(uv) = 1, \\ Av^2 - \frac{1}{2} uv \Lambda(uv) = 0. \end{cases} \tag{14}$$

Введем новую переменную

$$t = uv,$$

тогда, так как из (14) следует, что $Bu^2 - Av^2 = 1$ и $u^2v^2 = t^2$, то

$$v^2 = \frac{1}{2A}\left(\sqrt{1+4ABt^2} - 1\right)$$

$$u^2 = \frac{1}{2B}\left(\sqrt{1+4ABt^2} + 1\right)$$

и для определения t имеем уравнение

$$\sqrt{1+4ABt^2} - 1 - t\Lambda(t) = 0.$$

Введя функцию

$$f(t) = \frac{\sqrt{1+4ABt^2} - 1}{t}, \tag{15}$$

получаем для определения параметра t уравнение

$$G(t) \equiv f(t) - \Lambda(t) = 0. \tag{16}$$

Таким образом, задача поиска максимума функции правдоподобия свелась к поиску ненулевых решений уравнения (16), где функции f и Λ определены равенствами (15) и (8). Функция $G(t)$ нечетна, поэтому мы будем далее при рассмотрении уравнения (16) считать, что $t > 0$.

2. При больших t мы имеем

$$f(t) = 2\sqrt{AB} - \frac{1}{t} + O\left(\frac{1}{t^2}\right)$$

и

$$\Lambda(t) = \frac{2}{K}\sum_{j=1}^{K} F_{e,j} F_{c,j} - \frac{1}{2t} + O\left(\frac{1}{t^2}\right).$$

Поскольку (неравенство Коши):

$$C = \frac{1}{K}\sum_{j=1}^{K} F_{e,j} F_{c,j} \leq \sqrt{\frac{1}{K}\sum_{j=1}^{K} F_{e,j}^2 \cdot \frac{1}{K}\sum_{j=1}^{K} F_{c,j}^2} = \sqrt{AB},$$

то при росте t функция $G(t) = f(t) - \Lambda(t)$ имеет горизонтальную асимптоту на высоте $2(\sqrt{AB} - C) > 0$.

При $t \to 0$ мы имеем

$$f(t) = 2ABt + O(t^3), \quad \Lambda(t) = 2Dt + O(t^3),$$

где величины A, B, D определены равенством (7), то есть при малых t:

$$G(t) = f(t) - \Lambda(t) = 2(AB - D)t + O(t^3).$$

Мы ввели равенством (11) величину $\Omega = AB - D$ и установили, что она определяет характер стационарной точки $\alpha = 0$, $\beta = B$ функции правдоподобия. Из проведенного здесь исследования вытекает, что при $\Omega < 0$ (то есть когда точка $\alpha = 0$, $\beta = B$ – седловая) уравнение (16) имеет по крайней мере одно ненулевое решение. В случае же $\Omega > 0$ уравнение (16) может не иметь ненулевых решений. Эскизы графиков функции $y = G(t)$ для случаев $\Omega > 0$, и $\Omega < 0$ даны на рис. 1.

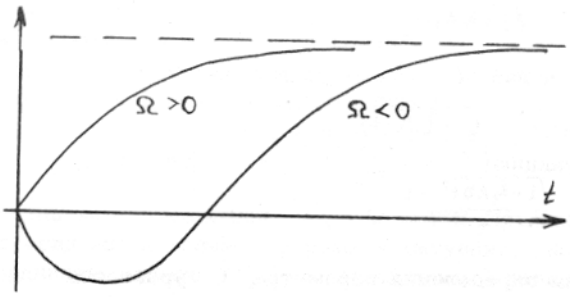

Рис. 1. Поведение функции $y = G(t)$

3. Вопрос о единственности решения в случае $\Omega > 0$ и о единственности ненулевого решения в случае $\Omega < 0$ остается открытым. Здесь автору не удалось установить строгого результата. Однако попытки решения уравнения (16) в практической ситуации не привели к нахождению более одного ненулевого решения.

4. После того, как ненулевой корень t уравнения (16) найден, искомые значения параметров α и β, максимизирующие функцию правдоподобия, выражаются через t по формулам:

$$\alpha = \sqrt{\frac{B}{A}\,\frac{\sqrt{1+4ABt^2}-1}{\sqrt{1+4ABt^2}+1}}, \quad \beta = \frac{2B}{\sqrt{1+4ABt^2}+1}.$$

§ 5. Тестирование метода

1. Предложенный способ оценки параметров распределения

$$P(F) = \frac{2F}{\beta} \exp\left\{-\frac{F^2 + \alpha^2 F_c^2}{\beta}\right\} I_0\left(\frac{2\alpha}{\beta} F F_c\right) \qquad (17)$$

был проверен на ряде тестов.

Для проведения теста были рассчитаны точные значения структурных факторов $F e^{i\varphi}$ по модели актинидина (координаты атомов были взяты из банка белковых молекул, при этом всем атомам был приписан одинаковый температурный фактор $B^t = 20$). Модули F этих структурных факторов имитировали в тесте экспериментально полученные величины F_e. Далее были рассчитаны структурные факторы $F_c e^{i\varphi_c}$ по этой же модели, но со внесенными в координаты атомов случайными ошибками. Ошибки $\vec{\Delta}_j$ в координатах атомов были взяты как независимые случайные величины с плотностью распределения вероятностей:

$$P(\vec{\Delta}) = \frac{1}{(2\pi\nu)^{3/2}} \exp\left\{-\frac{|\vec{\Delta}|^2}{2\nu^2}\right\},$$

где параметр ν менялся в разных тестах. Как было указано в § 1 п. 2 в этом случае величина F распределена по закону (17) с

$$\alpha = \alpha_{теор} = \langle \cos 2\pi(\vec{s}, \vec{\Delta}) \rangle,$$

$$\beta = \beta_{теор} = \left[1 - \langle \cos 2\pi(\vec{s}, \vec{\Delta}) \rangle^2\right] \sum_{j=1}^{N} f_j^2(s) e^{-B^t \frac{s^2}{2}}.$$

При этом непосредственное вычисление показывает, что

$$\langle \cos 2\pi(\vec{s}, \vec{\Delta}) \rangle = \exp\left\{-\frac{\pi^3}{4}(\omega |\vec{s}|)^2\right\},$$

где ω характеризует среднюю абсолютную ошибку в координатах атомов:

$$\omega = \langle |\vec{\Delta}| \rangle = \int_{\mathbb{R}^3} |\vec{\Delta}| P(\vec{\Delta})\, dV = \frac{4}{\sqrt{2\pi}} \nu.$$

Оба набора структурных факторов $\{F(\vec{s}) e^{i\varphi(\vec{s})}\}$ и $\{F_c(\vec{s}) e^{i\varphi_c(\vec{s})}\}$ были рассчитаны для точек $\vec{s}$ обратной решетки с $0 \le s^2 \le 0.25$ (что соответствует разрешению 2,0 Å). Далее интервал (0., 0.25) был разбит на 20 равных подинтервалов и для каждой группы структурных факторов (имеющих значение s^2 из одного и того же подинтервала) было определено по описанной в § 2 п. 4 методике значение параметров α, β и параметра $t = \alpha/\beta$. Результаты сравнения этих значений с теоретическими значениями приведены на рис. 2.

2. После того, как из максимума функции правдоподобия были определены параметры α и β для каждой зоны по s^2, мы проверили, насколько точно полученные распределения

$$P(\varphi | F) = \frac{1}{2\pi I_0\left(\frac{2\alpha}{\beta} F F_c\right)} \exp\left\{\frac{2\alpha}{\beta} F F_c \cos(\varphi - \varphi_c)\right\}$$

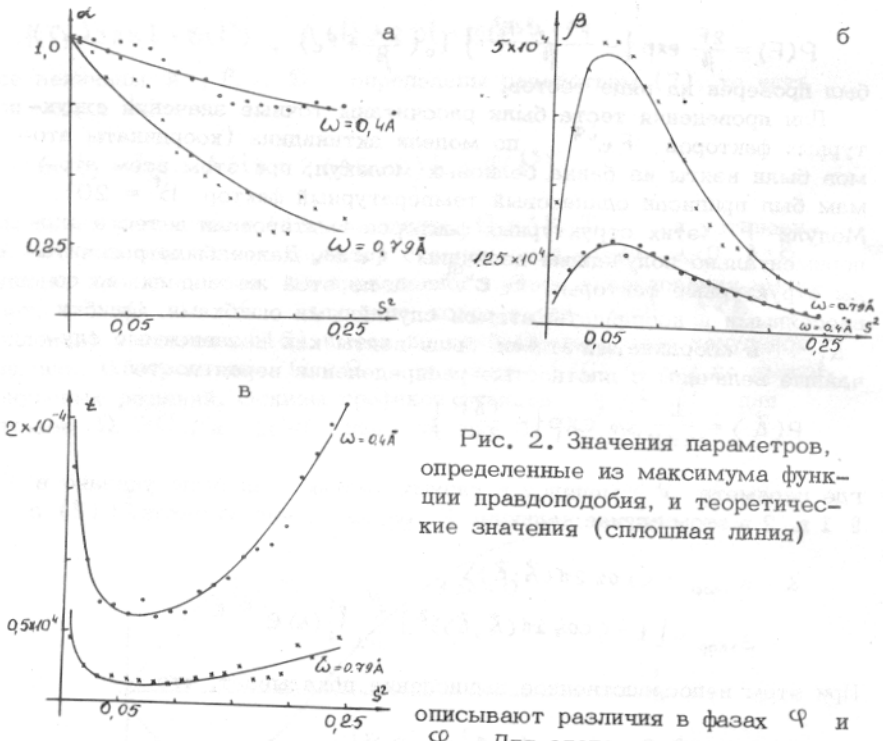

Рис. 2. Значения параметров, определенные из максимума функции правдоподобия, и теоретические значения (сплошная линия)

описывают различия в фазах φ и φ_c. Для этого для каждого из структурных факторов были вычислены величина

$$\delta\varphi = \langle |\varphi - \varphi_c| \rangle = \int_0^{2\pi} |\tau - \varphi_c| \, P(\tau | F) \, d\tau$$

(то есть "предсказываемая" средняя ошибка) и величина $|\varphi - \varphi_c|$ (то есть ошибка при данной модели). Сравнение средних арифметических (по всем нецентросимметричным рефлексам) значений "предсказанных" и реальных ошибок дано в табл. 1.

3. Следующий эксперимент относился к проверке возможности аналогичной оценки ошибок в фазах, вычисленных по модели, подвергшейся уточнению в обратном пространстве. Для этого модель со случайными ошибками, внесенными описанным в п. 1 способом, подвергалась уточнению (с использованием программы Айзекса, написанной согласно алгоритму /10/ и исправленной и поставленной на ЭВМ серии ЕС А.Г.Уржумцевым). По уточненной модели рассчитывались величины $F_c e^{i\varphi_c}$ и далее повторялась процедура, описанная в п. 1 и п. 2. Результаты тестов приведены в табл. 2.

Здесь следует заметить, что для модели, подвергшейся уточнению, предсказываемая ошибка близка к действительной лишь для

Таблица 1

Сравнение предсказанных средних ошибок ($\delta\varphi$) с реальными ($|\varphi-\varphi_c|$)

s^2_{min}	s^2_{max}	Модель 1*		Модель 2**		Число рефлексов
		$\delta\varphi$***	$\lvert\varphi-\varphi_c\rvert$***	$\delta\varphi$***	$\lvert\varphi-\varphi_c\rvert$***	
0.000	0.013	8	5	12	12	105
0.013	0.025	14	13	25	26	215
0.025	0.038	16	16	31	31	322
0.038	0.050	15	16	28	32	359
0.050	0.063	18	19	32	37	454
0.063	0.075	19	22	38	40	503
0.075	0.088	26	26	47	50	515
0.088	0.100	26	25	52	48	563
0.100	0.113	31	30	54	55	636
0.113	0.125	31	34	58	60	642
0.125	0.138	33	31	57	57	720
0.138	0.150	36	36	61	63	755
0.150	0.163	37	37	60	65	738
0.163	0.175	39	39	71	68	764
0.175	0.188	39	39	70	67	871
0.188	0.200	40	42	70	71	820
0.200	0.213	37	38	62	68	920
0.213	0.225	38	40	69	69	863
0.225	0.238	39	40	71	70	975
0.238	0.250	43	43	69	74	973
0.000	0.250	33	34	58	60	12713

* – модель со случайными ошибками с ω = 0.4 Å; ** – модель со случайными ошибками с ω = 0.79 Å; *** – дано среднее арифметическое значение для данной зоны по s^2 (в градусах).

Таблица 2

Сравнение предсказанных ошибок ($\delta\varphi$) с реальными ($|\varphi-\varphi_c|$) для модели, подвергшейся уточнению

$\sin^2\theta_{min}$	$\sin^2\theta_{max}$	Модель 3*		Модель 4**		Число рефлексов
		$\delta\varphi$***	$\lvert\varphi-\varphi_c\rvert$***	$\delta\varphi$***	$\lvert\varphi-\varphi_c\rvert$***	
0.000	0.013	8	9	7	7	105
0.013	0.025	11	21	9	14	215
0.025	0.038	11	23	8	16	322
0.038	0.050	10	24	8	16	359
0.050	0.063	12	26	9	19	454
0.063	0.075	12	30	9	20	503
0.075	0.088	16	38	11	26	515
0.088	0.100	17	38	13	28	563
0.100	0.113	23	45	15	31	636
0.113	0.125	48	48	15	34	642
0.125	0.138	50	47	18	34	720
0.138	0.150	52	55	22	39	755
0.150	0.163	54	56	23	41	738

0.163	0.175	61	58	23	42	764
0.175	0.188	56	60	27	43	871
0.188	0.200	59	62	28	44	820
0.200	0.213	55	58	29	41	920
0.213	0.225	61	59	31	42	863
0.225	0.238	58	60	33	45	975
0.238	0.250	67	64	37	47	973
0.000	0.250	45	50	22	36	12713

* – модель 2 (со случайными ошибками с ω = 0.79 Å) после 3 циклов уточнения координат атомов по зоне $0.0 \leqslant \lambda^2 \leqslant 0.111$; ** – модель 2 после 3 циклов уточнения координат атомов по зоне $0. \leqslant \lambda^2 \leqslant 0.111$, 3 циклов уточнения координат по зоне $0.111 \leqslant \lambda^2 \leqslant 0.16$, 3 циклов уточнения координат по зоне $0.16 \leqslant \lambda^2 \leqslant 0.207$ и 3 циклов уточнения по зоне $0.207 \leqslant \lambda^2 \leqslant 0.25$; *** – дано среднее арифметическое значение для данной зоны по λ^2 (в градусах).

тех зон по s^2, модули структурных факторов из которых не включались в уточнение. Для зон же, по которым уточнение проводилось, предсказываемая погрешность фаз имеет тенденцию быть ниже реальной. По всей вероятности, это связано с тем, что рефлексы из зон, включенных в уточнение, не могут более рассматриваться как независимые случайные величины с распределением (6), то есть не выполнено Условие 1 из § 2.

Приложение 1. Поведение функции $Q(u,v)$ на границе.

Рассмотрим в области $u>0$, $v\in(-\infty,\infty)$ функцию

$$Q(u,v)=2\ln u-Bu^2-Av^2+\frac{1}{K}\sum_{j=1}^{K}\ln I_0(2uvF_{e,j}F_{c,j}),$$

где $A>0$, $B>0$, $F_{e,j}>0$, $F_{c,j}>0$ при всех $j=1,\dots,K$.

У с л о в и е 2. Существует такое число k, что

$$F_{e,j}=kF_{c,j} \quad \text{при всех} \quad j=1,\dots,K.$$

Л е м м а 1. Если Условие 2 не выполняется, то функция $Q(u,v)$ стремится к $-\infty$ при приближении аргументов к границе области определения.

Д о к а з а т е л ь с т в о. Ясно, что если приближение к границе осуществляется так, что величина $|uv|$ остается ограниченной, то при этом $Q(u,v)\to-\infty$.

Если приближение к границе происходит так, что $|uv|\to\infty$, то используя асимптотику (см., например, /4/)

$$I_0(t)=\frac{e^t}{\sqrt{2\pi t}}\left[1+O\left(\frac{1}{t}\right)\right] \quad \text{при} \quad t\to\infty,$$

имеем

$$\frac{1}{K}\sum_{j=1}^{K}\ln I_0(2uvF_{e,j}F_{c,j})=2|uv|\frac{1}{K}\sum_{j=1}^{K}F_{e,j}F_{c,j}-\frac{1}{2}\ln|uv|+O(1).$$

Поскольку Условие 2 не выполняется, то

$$\frac{1}{K}\sum_{j=1}^{K}F_{e,j}F_{c,j}<\sqrt{\frac{1}{K}\sum_{j=1}^{K}F_{e,j}^2}\sqrt{\frac{1}{K}\sum_{j=1}^{K}F_{c,j}^2},$$

(неравенство Коши становится строгим), то есть

$$\frac{1}{K}\sum_{j=1}^{K}F_{e,j}F_{c,j}=(1-\gamma^2)\sqrt{AB} \quad \text{где} \quad \gamma>0.$$

Поэтому

$$Q(u,v) = -\gamma^2 B u^2 + 2\ln u - \gamma^2 A v^2 - \frac{1}{2}\ln|uv| - (1-\gamma^2)(\sqrt{B}u - \sqrt{A}|v|)^2 + O(1) \to -\infty,$$

то есть лемма доказана.

Л е м м а 2. Пусть Условие 2 выполнено, тогда при приближении к границе области определения вдоль любой прямой $v = \lambda u$ с $\lambda \neq \pm k$ имеем $Q(u, \lambda u) \to -\infty$ при $u \to \infty$. Если же приближение к границе происходит вдоль одной из прямых $v = \pm k u$, то имеем $Q(u, \pm k u) \to \infty$ при $u \to \infty$.

Д о к а з а т е л ь с т в о. Действуя аналогично доказательству Леммы 1, получаем, что в этом случае

$$Q(u,v) = -A(ku - |v|)^2 + \ln\frac{u\sqrt{u}}{\sqrt{|v|}} + O(1) \qquad \text{при} \quad |uv| \to \infty.$$

Таким образом,

$$Q(u, \lambda u) = -A(k - |\lambda|)^2 u^2 + \ln u + O(1) \to -\infty$$

и

$$Q(u, \pm k u) = \ln u + O(1) \to +\infty \quad \text{при} \quad u \to \infty.$$

Лемма доказан

З а м е ч а н и е. Из Леммы 2 следует, что в случае выполнения Условия 2 для функции правдоподобия $\Psi(\alpha, \beta)$ максимум не достигается внутри рассматриваемой области $\alpha \geq 0$, $\beta > 0$, а при приближении к границе области в точке $\alpha = k$, $\beta = 0$ функция $\Psi(\alpha, \beta)$ неограниченно возрастает. Это соответствует тому, что исходное распределение для величины F_j не имеет более плотности (6), а становится сингулярным и сосредоточено в одной точке $k F_{c,j}$.

Приложение 2. Среднее значение величины Ω.

Пусть имеет место ситуация, рассмотренная в /1/, то есть

$$F_j e^{i\varphi_j} = \sum_{q=1}^{N} f_q(s_j) e^{-B_q^t \frac{s_j^2}{4}} e^{2\pi i(\vec{s}_j, \vec{\tau}_q)},$$

$$F_{c,j} e^{i\varphi_{c,j}} = \sum_{q=1}^{M} f_q(s_j) e^{-B_q^t \frac{s_j^2}{4}} e^{2\pi i(\vec{s}_j, \vec{\tau}_q + \vec{\Delta}_q)},$$

где $\vec{\tau}_q$ – независимые случайные вектора, равномерно распределенные в единичном кубе; $\vec{\Delta}_q$ – независимые одинаково распределенные случайные вектора с распределением вероятностей, симметричным относительно нуля. Отметим, что здесь мы в отличие от § 1 п. 2 считаем случайными величинами не только коорди-

наты атомов модели, но и координаты атомов структуры, то есть рассматриваем "всевозможные" структуры со случайными ошибками. Пусть $F_{e,j} = k F_j$.

В /1/ показано, что в этом случае

$$\langle F_{c,j}^2 \rangle = \sigma_M^2, \quad \langle F_j^2 \rangle = \sigma_N^2, \quad \text{то есть}$$

$$\langle F_{e,j}^2 \rangle = k^2 \sigma_N^2.$$

В /1/ показано также, что случайная величина $z = \frac{F_j^2}{\sigma_N^2} \cdot \frac{F_{c,j}^2}{\sigma_M^2}$ распределена с плотностью

$$P(z) = \frac{2}{\sigma_B^2} I_0\left(\frac{2\sigma_A}{\sigma_B^2}\sqrt{z}\right) K_0\left(\frac{2}{\sigma_B^2}\sqrt{z}\right),$$

где

$$\sigma_A = \frac{\sigma_M}{\sigma_N}\mu, \quad \sigma_B^2 = 1 - \frac{\sigma_M^2}{\sigma_N^2}\mu^2, \quad \mu = \langle \cos 2\pi(\vec{x}, \vec{\Delta}) \rangle$$

и I_0, K_0 – модифицированные функции Бесселя первого и второго рода.

Отсюда, используя /4/ (формулы 6.576.5 и 9.131.1), имеем при $\sigma_A < 1$:

$$\langle z \rangle = \int_0^\infty z P(z)\, dz = \frac{\sigma_B^6}{4} \int_0^\infty \tau^3 I_0(\sigma_A \tau) K_0(\tau)\, d\tau = 1 + \sigma_A^2,$$

то есть

$$\langle F_{e,j}^2 F_{c,j}^2 \rangle = k^2 \sigma_N^2 \sigma_M^2 \left(1 + \frac{\sigma_M^2}{\sigma_N^2}\mu^2\right).$$

Используя технику работы /1/, можно также показать, что корреляция между величинами $F_{e,j}$ и $F_{c,j}$ имеет порядок $1/\sqrt{N}$, поэтому, ограничиваясь главными членами, имеем

$$\begin{aligned}
\langle \Omega \rangle &= \frac{1}{K^2} \sum_{j,q=1}^{K} \langle F_{e,j}^2 F_{c,q}^2 \rangle - \frac{1}{K} \sum_{j=1}^{K} \langle F_{e,j}^2 F_{c,j}^2 \rangle \simeq \\
&\simeq \frac{1}{K} \sum_{j=1}^{K} \langle F_{e,j}^2 \rangle \cdot \frac{1}{K} \sum_{j=1}^{K} \langle F_{c,j}^2 \rangle + \\
&+ \frac{1}{K^2} \sum_{j=1}^{K} \left[\langle F_{e,j}^2 F_{c,j}^2 \rangle - \langle F_{e,j}^2 \rangle \langle F_{c,j}^2 \rangle \right] - \\
&- \frac{1}{K} \sum_{j=1}^{K} \langle F_{e,j}^2 F_{c,j}^2 \rangle = \\
&= k^2 \sigma_N^2 \sigma_M^2 + \frac{1}{K}\left[k^2 \sigma_N^2 \sigma_M^2 \left(1 + \frac{\sigma_M^2}{\sigma_N^2}\mu^2\right) - k^2 \sigma_N^2 \sigma_M^2 \right] - \\
&- k^2 \sigma_N^2 \sigma_M^2 \left(1 + \frac{\sigma_M^2}{\sigma_N^2}\mu^2\right) = \\
&= -\left(1 - \frac{1}{K}\right) \mu^2 k^2 \sigma_M^4.
\end{aligned}$$

ЛИТЕРАТУРА

1. Сринивасан Р., Партасарати С. Применение статистических методов в рентгеновской кристаллографии. М., Мир, 1979.

2. Бландел Т., Джонсон Л. Кристаллография белка. М., Мир, 1979.

3. Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. М., Мир, 1980.

4. Градштейн И.С., Рыжик И.М. Таблицы интегралов, сумм, рядов и произведений. М., 1962.

5. Лунин В.Ю., Уржумцев А.Г. Повышение разрешения карт электронной плотности белков путем уточнения модельной структуры. I. Описание метода. Препринт. Пущино, ОНТИ НЦБИ АН СССР, 1981.

6. Blow D.M., Crick F.H.C. - Acta Cryst., 1959, v. 12, p. 794-802.

7. Sim G.A. - Acta Cryst., 1959, v. 12, p. 813-815.

8. Luzzati V. - Acta Cryst., 1955, v. 8, p. 795-806.

9. Luzzati V. - Acta Cryst., 1952, v. 5, p. 802-810.

10. Agarwal R.C. - Acta Cryst., 1978, v. A34, p. 791-809.

11. Wilson A.J.C. - Acta Cryst., 1949, v. 2, p. 318-321.

12. Bricogne G. - Acta Cryst., 1976, v. A32, p. 832-847.

СОДЕРЖАНИЕ

Владимир Юрьевич Лунин

ИСПОЛЬЗОВАНИЕ МЕТОДА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ ДЛЯ ОЦЕНКИ ОШИБОК ПРИ ОПРЕДЕЛЕНИИ ФАЗ В КРИСТАЛЛОГРАФИИ БЕЛКА

Препринт

Отредактировано и подготовлено к печати в ОНТИ НЦБИ АН СССР

Редактор Т.В.Букина
Технический редактор С.М.Ткачук
Корректоры Т.К.Крейцер, Л.М.Орлова

Подписано в печать 31.08.82 г. Т16425. Уч.-изд.л. 1,4
Формат 60x90/16. Тираж 100 экз. Бумага офсетная.
Заказ 2565Р. Бесплатно. Изд. № 288.

Отпечатано на ротапринте в Отделе научно-технической информации Научного центра биологических исследований АН СССР в Пущине