КРИСТАЛЛОГРАФИЯ, 2000, том 45, № 2, с. 227–230

ТЕОРИЯ КРИСТАЛЛИЧЕСКИХ СТРУКТУР

УДК 548.737

НАДЕЖНОСТЬ ПОКАЗАТЕЛЕЙ ДОСТОВЕРНОСТИ, ВЫЧИСЛЯЕМЫХ НА БАЗЕ ПРИНЦИПА МАКСИМУМА ПРАВДОПОДОБИЯ

© 2000 г. Т. П. Сковорода, В. Ю. Лунин

Институт математических проблем биологии РАН, Пущино

Поступила в редакцию 19.02.97 г.
После доработки 03.02.98 г.

Рассмотрены различные схемы определения показателей достоверности фаз структурных факторов, базирующиеся на принципе максимума правдоподобия. Продемонстрировано, что включение в расчет функции правдоподобия всех имеющихся в наличии модулей структурных факторов позволяет получать адекватные оценки точности для фаз, рассчитанных по атомным моделям с независимыми ошибками в координатах, но систематически завышает показатели достоверности, если модель подвергалась уточнению в обратном пространстве. Показано, что использование рассчитанной по контрольному набору рефлексов маргинальной функции правдоподобия позволяет устранить систематическое смещение оценок. Предложен способ уменьшения статистического разброса оценок при использовании небольшого числа контрольных рефлексов.

ВВЕДЕНИЕ

В кристаллографии макромолекул для расчета синтезов Фурье электронной плотности общепринятым является использование коэффициентов

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

где **s** – узел решетки обратного пространства, F_s^{obs} – экспериментально определенное значение модуля структурного фактора, φ_s^{mod} – рассчитанная по некоторой предварительной атомной модели фаза структурного фактора. Весовой множитель m_s, называемый показателем достоверности определения фазы, введен для компенсации неточности коэффициента Фурье, вызванной расхождением между φ_s^{mod} и точным значением фазы φ_s^{true}. Вероятностные предположения о случайном характере ошибок в предварительной атомной модели позволяют определять веса m_s как математические ожидания величин $\cos(\varphi_s^{true} - \varphi_s^{mod})$ [1] и приводят к широко применяемой формуле

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

где функция $\Lambda(x)$ – гиперболический тангенс для центросимметричных рефлексов и отношение модифицированных функций Бесселя $I_1(2x)/I_0(2x)$ для нецентросимметричных рефлексов; F_s^{mod} – модуль структурного фактора, рассчитанного по предварительной модели; ε_s – коэффициент, компенсирующий различия в средней интенсивности для разных типов рефлексов. Параметр t_s, входящий в (2), отражает величину ошибок, допущенных при построении предварительной модели, и его правильная оценка (что по существу есть оценка качества предварительной модели) играет ключевую роль при определении величин весов m_s.

Вычисленные "правильно" значения весов m_s в среднем должны быть близки к реальным величинам $\cos(\varphi_s^{true} - \varphi_s^{mod})$. Степень этого соответствия можно проверить в тестовых ситуациях, когда помимо предварительной модели известна достаточно надежная "искомая" атомная структура, и вычисленные по ней фазы могут рассматриваться как точные. Такая проверка, проведенная Ридом [2], показала, что многие из предложенных ранее способов вычисления параметров t_s приводят к значениям показателей достоверности, далеким от реальности, и что наилучшие результаты достигаются при использовании подхода [2–5], основанного на принципе максимального правдоподобия. Тем не менее, как отмечено в [2–5], применение этого метода вызывает сложности при работе с предварительными моделями, подвергавшимися кристаллографическому уточнению. В данной работе анализируются модификации метода, позволяющие получать правильные значения показателей достоверности для фаз, рассчитанных по уточнявшимся атомным моделям.

В процессе тестирования использовались экспериментальные данные для белка *Protein G* (пр. гр. $P2_12_12_1$, размеры элементарной ячейки 34.9 × 40.3 × 42.2 Å) и его известная атомная структура, определенная ранее при высоком разрешении.

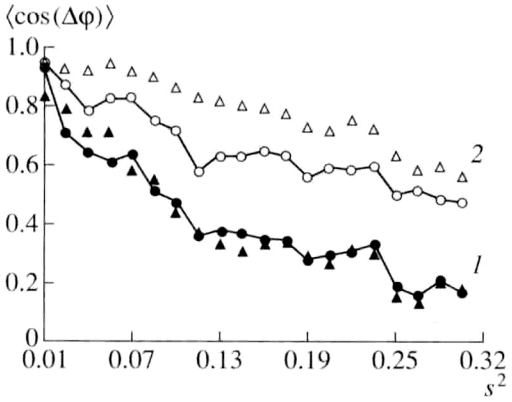

Рис. 1. Предсказанные на базе МП-оценок (треугольники) и действительные значения (кружки) косинусов фазовых ошибок; *1* – модель с независимыми ошибками в координатах, *2* – та же модель после 12 циклов уточнения в обратном пространстве.

Уточнение тестовых предварительных моделей проводилось с помощью программного комплекса FROG [6], а оценки на основе принципа максимизации правдоподобия (**МП**-оценки) и максимизации маргинальной функции правдоподобия (**ММП**-оценки) вычислялись программой LBEST [7], созданной авторами в соответствии с методикой [3, 5].

МП-ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ ВЕРОЯТНОСТЕЙ

Сутью вероятностного подхода к проблеме оценки фазовых ошибок [8–10] является включение дополнительной информации об изучаемом объекте в виде статистической гипотезы о характере распределений ошибок в предварительной модели. Простейший и наиболее жесткий вариант такой гипотезы состоит в том, что ошибки присутствуют только в координатах атомов, независимы и распределены по нормальному закону с нулевыми средними и одинаковыми дисперсиями (более сложные примеры рассмотрены в [5, 7, 11]). Такого рода гипотезы позволяют рассматривать значения структурных факторов как случайные величины и получать выражения для их распределений вероятностей. Для широкого класса исходных гипотез результирующие распределения описываются одной и той же формулой

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

и различаются лишь видом функциональных зависимостей α_s и β_s от параметров, описывающих распределения ошибок в атомной модели. Поэтому, не фиксируя конкретную гипотезу относительно атомной модели, будем исходить из того, что распределение вероятностей для структурного фактора имеет вид (3) с некоторыми параметрами α_s и β_s. Математическое ожидание косинуса фазовой ошибки в этом случае дается формулой (2) с $t_s = \alpha_s/\beta_s$, и возникает одна из стандартных задач математической статистики – задача выбора распределения из класса (3) или задача оценки параметров α_s и β_s.

Методы МП-оценки параметров α_s и β_s описаны в [2–5]. Суть метода состоит в том, что в каждом тонком сферическом слое обратного пространства в качестве оценок для величин α_s и β_s берутся константы, максимизирующие вероятность совпадения модулей структурных факторов, распределенных по закону (3) и реально полученных в эксперименте величин $\{F_s^{obs}\}$.

ОЦЕНКИ ДЛЯ КОЭФФИЦИЕНТОВ ДОСТОВЕРНОСТИ

В рамках рассматриваемого подхода показатели достоверности, вычисляемые по формуле (2), сами являются случайными величинами и обладают всеми характеристиками, присущими случайным величинам. Одной из наиболее важных статистических характеристик оценок является смещение – отклонение ожидаемого значения оценки от ее истинной величины. Другой существенной характеристикой является ожидаемое среднеквадратичное отклонение оценки от истинного значения (разброс оценки).

Первая серия тестов была посвящена исследованию смещенности показателей достоверности, вычисляемых на базе МП-оценок для параметров α_s и β_s. Чтобы исключить влияние экспериментальных ошибок и неточностей исходной модели, в этих тестах в качестве точных значений фаз и вместо экспериментальных значений модулей использовались величины, рассчитанные по структуре *Protein G*, из которой удалены молекулы воды. Эта же модель, но с внесенными в координаты атомов случайными независимыми ошибками (средней абсолютной величины 0.8 Å) играла роль стартовой модели для дальнейшего уточнения. На рис. 1 представлены результаты использования МП-оценок для стартовой модели и после 12 циклов ее уточнения в обратном пространстве. Отчетливо видно, что показатели достоверности, вычисленные на базе МП-оценок, не имеют смещения в случае независимых ошибок в модели, но становятся систематически завышенными для модели, прошедшей уточнение.

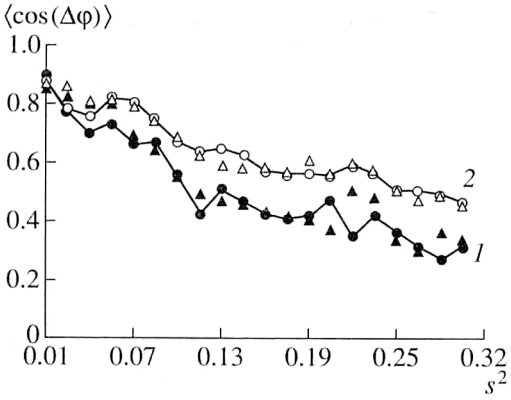

Рис. 2. Предсказанные на базе ММП-оценок (треугольники) и действительные значения (кружки) косинусов фазовых ошибок; *1* – уточнение без учета стереохимических ограничений, *2* – уточнение с учетом стереохимических ограничений. Модельные данные.

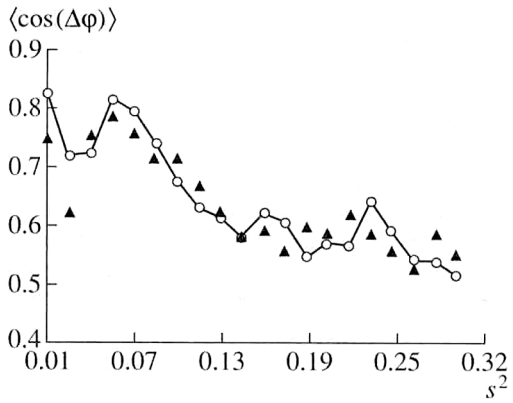

Рис. 3. Предсказанные на базе ММП-оценок (треугольники) и действительные значения (кружки) косинусов фазовых ошибок, для модели, уточненной по 50% экспериментальных данных.

УСТРАНЕНИЕ СМЕЩЕНИЯ. ММП-ОЦЕНКИ

В [5] предложена модификация метода нахождения параметров α_s и β_s, базирующаяся на *R-free* методологии Брюнгера [12]. При таком подходе, примененном нами и в [13–15], функция правдоподобия вычисляется только по рефлексам, предварительно исключенным из процесса уточнения. Значения параметров, полученные в результате максимизации такой маргинальной функции правдоподобия, и называют ММП-оценками. Целью следующих тестов было исследование смещенности оценок для коэффициентов достоверности, рассчитанных на основе ММП-оценок для параметров α_s и β_s. В каждом месте до начала процедуры уточнения случайным образом выбиралось контрольное множество рефлексов, которые исключались из уточнения и использовались для последующего определения оценок α_s и β_s, после чего коэффициенты достоверности вычислялись для всех рефлексов по формуле (2) с $t_s = \alpha_s/\beta_s$.

Рисунок 2 показывает результаты таких расчетов для двух стратегий уточнения. В первом случае было проведено 24 цикла уточнения без учета стереохимических ограничений, а во втором 12 циклов с их учетом. Чтобы уменьшить разброс оценок, возникающий при определении параметров распределения по небольшому числу экспериментальных данных, первоначально взяли достаточно большое контрольное множество, составляющее 50% рефлексов из зоны разрешения 1.8 Å. Как следует из рис. 2, в обоих случаях смещенность оценок устраняется, предсказанные средние величины cos(Δφ) достаточно близки к реальным и отражают разное качество моделей, полученных в результате уточнения.

Все следующие тесты проводились с той же стартовой моделью, но для уточнения, для оценки параметров α_s и β_s и для вычисления коэффициентов достоверности использовались настоящие экспериментальные данные. В этом случае за точные значения фаз принимались фазы, рассчитанные по всем атомам модели белка *Protein G*, включая молекулы воды. Следует отметить, что в этом случае источниками ошибок в рассчитанных по модели фазах являлись как позиционные ошибки атомов модели, так и отсутствие в модели ряда атомов (молекул воды). На рис. 3 и 4 показаны (треугольниками) результаты использования ММП-оценок в этом случае. Контрольное множество в этих тестах составляло 50 (рис. 3) либо 10% (рис. 4) рефлексов. Оба рисунка демонстрируют несмещенность полученных оценок относительно их истинных значений.

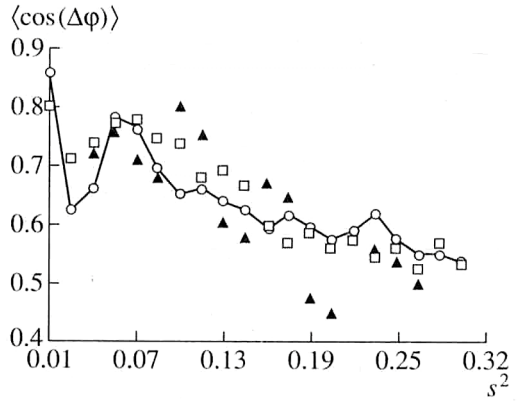

Рис. 4. Предсказанные на базе ММП-оценок (треугольники) и на базе сглаженных ММП-оценок (квадраты) значения косинусов фазовых ошибок и их действительные значения (кружки) для модели, уточненной по 90% экспериментальных данных.

УМЕНЬШЕНИЕ СТАТИСТИЧЕСКОГО РАЗБРОСА ММП-ОЦЕНОК

Как показывают проведенные расчеты, уменьшение доли контрольных рефлексов от 50% набора до более реалистичного на практике числа 10%, сохраняя несмещенность оценок, приводит к фактическому удвоению амплитуды разброса предсказываемых значений показателей достоверности. Возможны несколько подходов к проблеме уменьшения статистического разброса таких оценок. Первый подход состоит во введении дополнительных гипотез о характере ошибок в атомной модели, что позволяет уменьшить число определяемых параметров и тем самым увеличить отношение числа измерений к числу определяемых параметров. Предположим, что в модели присутствуют все атомы и ошибки в координатах этих атомов независимы и распределены по радиально-симметричному нормальному закону с одинаковым (но неизвестным нам) стандартным отклонением ν. Предполагая структурные факторы нормализованными, можно показать, что в этом случае выполняются соотношения

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

и поставить задачу определения одного значения параметра ν вместо определения значений α и β для каждой из зон обратного пространства. Несколько более общий вариант этого подхода предложен в [15], где дополнительное предположение о характере ошибок в координатах сразу заменено на предположение о наличии для параметров α_s и β_s зависимости типа (4). Достоинством такого подхода является то, что существенно улучшается соотношение числа измерений и числа определяемых параметров – тем самым снижается статистический разброс оценок. С другой стороны, он требует введения новых предположений, достоверность которых не всегда очевидна. При этом задача поиска оптимальных значений параметров существенно усложняется и делает труднореализуемым поиск глобального минимума. Так, в [15] авторы в такой ситуации ограничиваются лишь локальной максимизацией правдоподобия.

Другой подход к уменьшению статистического разброса оценок заключается в замене требования наличия фиксированной функциональной зависимости параметров α и β от s на более мягкое требование гладкости этих зависимостей. Такое требование может быть учтено, например, введением в максимизируемый функционал дополнительного штрафа за сильные локальные отклонения в получаемых зависимостях [13]. Однако в этом случае опять-таки приходится обходиться локальной оптимизацией.

Тем не менее дополнительное условие гладкости может быть реализовано и более простым путем, при котором начальные значения параметров α и β по-прежнему определяются для каждой зоны из условия глобального максимума, а затем осуществляется сглаживание полученных значений по одной из стандартных схем. На рис. 4 показаны (квадратами) результаты применения простейшей процедуры сглаживания, при которой отношение $t = \alpha/\beta$, определенное первоначально независимо для каждой из зон обратного пространства, заменялось затем на его среднее значение в текущей, предыдущей и последующей зонах. Полученные таким образом сглаженные значения t использовались для вычисления показателей достоверности по формуле (2). Из рисунка видно, что такой прием заметно снижает ошибки вычисляемых оценок и делает возможным использование сравнительно небольшого контрольного множества рефлексов, что не будет оказывать существенного влияния на ход процесса уточнения.

Работа поддержана грантами № 94-04-12844 и 97-04-48319 Российского фонда фундаментальных исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Blow D.M., Crick F.H.C.* // Acta Cryst. 1959. V. 12. P. 794.
2. *Read R.J.* // Acta Cryst. A. 1986. V. 42. P. 140.
3. *Лунин В.Ю.* Использование метода максимального правдоподобия для оценки ошибок при определении фаз в кристаллографии белка. Препринт НЦБИ АН СССР. Пущино, 1982. С. 22.
4. *Lunin V.Yu., Urzhumtsev A.G.* // Acta Cryst. A. 1984. V. 40. P. 269.
5. *Lunin V.Yu., Skovoroda T.P.* // Acta Cryst. A. 1995. V. 51. P. 880.
6. *Urzhumtsev A.G., Lunin V.Yu., Vernoslova E.A.* // J. Appl. Cryst. 1989. V. 22. P. 500.
7. *Urzhumtsev A.G., Skovoroda T.P., Lunin V.Yu.* // J. Appl. Cryst. 1996. V. 29. P. 741.
8. *Luzzati V.* // Acta Cryst. 1952. V. 5. P. 802.
9. *Sim G.A.* // Acta Cryst. 1959. V. 12. P. 813.
10. *Сринивасан Р., Партасарати С.* // Применение статистических методов в рентгеновской кристаллографии. М.: Мир, 1979. С. 312.
11. *Read R.J.* // Acta Cryst. A. 1990. V. 46. P. 900.
12. *Brünger A.T.* // Nature (London). 1992. V. 355. P. 472.
13. *Pannu N.S., Read R.J.* // Acta Cryst. A. 1996. V. 52. P. 659.
14. *Murshudov G.N., Dodson E.J., Vagin A.A.* // Macromolecular Refinement. Proceedings of the CCP4 Study Weekend. January, 1996. P. 93.
15. *Bricogne G., Irwin J.* // Macromolecular Refinement. Proceedings of the CCP4 Study Weekend. January, 1996. P. 85.