МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА, 2006, том 1, №1, с.17-26,
http://www.matbio.org/downloads/Lunin_rus2006(1_17).pdf

==========МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ==========

УДК 548.737

Статистическое моделирование и принцип максимального правдоподобия в кристаллографии макромолекул

Лунин В.Ю. [*], Афонин П.В.[1], Уржумцев А.Г.[***2]**

Институт Математических Проблем Биологии, Российская Академия Наук, 142290 Пущино, Россия

[1]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, BLDG 64R0121, Berkeley, CA 94720 USA

[2]Département de Physique, Université H.Poincaré Nancy 1, 54506 Vandoeuvre-lès-Nancy, France

Аннотация. *При статистическом моделировании изучаемая структура рассматривается как элемент некоторого ансамбля структур. Это позволяет распространять общие свойства ансамбля на конкретную исследуемую структуру. Такие общие свойства иногда устанавливаются более легко для всего ансамбля, нежели для отдельной структуры. Полезность получаемой информации зависит от того, насколько адекватно выбирается ансамбль структур. Статистическое правдоподобие может быть использовано как основа для выбора наиболее адекватной статистической модели. Обсуждаются несколько примеров использования такого похода к изучению биологических макромолекул методами рентгеноструктурного анализа.*

Ключевые слова*: структура биологических макромолекул, рентгеноструктурный анализ, статистическое моделирование*

1 Введение

1.1 Основные понятия рентгеноструктурного анализа

Картина рассеивания рентгеновских лучей определяется распределением электронов в исследуемом образце и описывается ***функцией распределения электронной плотности*** $\rho(\mathbf{r})$, так что $\rho(\mathbf{r})dV_\mathbf{r}$ есть усредненный за время эксперимента заряд в элементарном объеме $dV_\mathbf{r}$**.** Нахождение этого распределения и интерпретация его в структурных терминах является конечной целью рентгеновского исследования. При изучении кристаллического образца распределение электронной плотности описывается трехмерной периодической функцией с периодами **a**,**b**,**c**:

$$\rho(\mathbf{r}) = \rho(\mathbf{r}+\mathbf{a}) = \rho(\mathbf{r}+\mathbf{b}) = \rho(\mathbf{r}+\mathbf{c}). \qquad (1)$$

Параллелепипед V, построенный на этих трех векторах, называется ***элементарной ячейкой*** кристалла. Поскольку функция $\rho(\mathbf{r})$ является периодической, она может быть представлена в виде трехмерного ряда Фурье:

$$\rho(\mathbf{r}) = \frac{1}{|V|}\sum_{\mathbf{s}\in S'} F(\mathbf{s})\exp[i\varphi(\mathbf{s})]\exp[-2\pi i(\mathbf{s},\mathbf{r})] \qquad (2)$$

$$F(\mathbf{s})\exp[i\varphi(\mathbf{s})] = \int_V \rho(\mathbf{r})\exp[2\pi i(\mathbf{s},\mathbf{r})]dV_\mathbf{r}\,, \qquad (3)$$

[*] lunin@impb.psn.ru, http://www.impb.ru/lmc
[**] PAfonine@lbl.gov
[***] Alexander.Ourjoumtsev@stmp.uhp-nancy.fr

где суммирование в (2) ведется по всем векторам $\mathbf{s}=h\mathbf{a}^*+k\mathbf{b}^*+l\mathbf{c}^*$ с целыми h,k,l, а $\{\mathbf{a}^*,\mathbf{b}^*,\mathbf{c}^*\}$ - базис, сопряженный с $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. Комплексные коэффициенты Фурье $F(\mathbf{s})\exp[i\varphi(\mathbf{s})]$ называются ***структурными факторами***, а их модули $F(\mathbf{s})$ и аргументы $\varphi(\mathbf{s})$ называются ***модулями структурных факторов*** и ***фазами структурных факторов*** соответственно. Для краткости ниже мы их будем называть просто модулями и фазами. Мы называем вектор ***s рефлексом***, а целые числа h,k,l ***индексами рефлекса***. Традиционный рентгеновский эксперимент позволяет получить для некоторого множества рефлексов S модули структурных факторов, в то время как значения фаз в эксперименте теряются. Восстановление значений фаз составляет центральную проблему рентгеноструктурного анализа, называемую ***фазовой проблемой***. Помимо того, что в эксперименте не регистрируются значения фаз, часть модулей структурных факторов также теряется (в основном, для высокочастотных гармоник Фурье). Поэтому, даже восстановив значения $\{\varphi(\mathbf{s})\},\mathbf{s}\in S$, невозможно точно рассчитать распределение электронной плотности (2). Частичная сумма $\rho_S(\mathbf{r})$ ряда (2), рассчитанная с имеющимся набором модулей и восстановленными значениями фаз, называется ***синтезом Фурье электронной плотности***. Такое ***изображение*** электронной плотности $\rho_S(\mathbf{r})$ содержит искажения, вызванные как эффектом обрыва ряда, так и ошибками в значениях восстановленных фаз.

Гипотеза "атомности" предполагает, что искомое распределение электронной плотности является не произвольной функцией, а суммой локальных атомных вкладов:

$$\rho(\mathbf{r})=\sum_{j=1}^{N}\rho^{atom}(\mathbf{r}-\mathbf{r}_j)\ , \qquad (4)$$

где N - число атомов, из которых состоит изучаемый объект (десятки тысяч для больших белковых молекул), а $\{\mathbf{r}_j\}$ - их декартовы координаты. В этой заметке для упрощения формул мы предполагаем, что все атомы одинаковы. Это не является чересчур грубым приближением для белковых молекул. Распределение электронной плотности в отдельном атоме $\rho^{atom}(\mathbf{r})$ предполагается известным и сферически симметричным, поэтому $\rho^{atom}(\mathbf{r})\equiv\rho^0(r)$. Трехмерная гауссова функция дает приблизительный вид распределения электронов в отдельном атоме. Если координаты атомов известны, суммарное распределение электронной плотности может быть вычислено посредством (4). Обратная задача, т.е. разложение распределения электронной плотности в сумму вкладов отдельных атомов, может оказаться значительно более сложной, особенно когда в качестве входной информации используется не истинное распределение, а некоторое его изображение $\rho_S(\mathbf{r})$. Такое разложение называется ***интерпретацией*** ряда Фурье или ***построением атомной модели***.

"Атомность" позволяет представить структурные факторы в виде

$$F^{mod}(\mathbf{s})\exp\left[i\varphi^{mod}(\mathbf{s})\right]=f^{0}(s)\sum_{j=1}^{N}\exp\left[2\pi i(\mathbf{s},\mathbf{r}_j)\right]\ , \qquad (5)$$

где $f^{0}(s)$ - синус-Фурье преобразование распределения плотности одного атома $\rho^0(r)$, т.е. известная функция. Задача определения атомных координат $\{\mathbf{r}_j\}$ может теперь быть сформулирована, формально, как задача минимизации в пространстве размерности $3N$:

$$\sum_{\mathbf{s}\in S}\left(F^{mod}(\mathbf{s};\{\mathbf{r}_j\})-F^{obs}(\mathbf{s})\right)^2 \quad\Rightarrow\quad \min\ , \qquad (6)$$

где $F^{obs}(\mathbf{s})$ означает экспериментально определенное значение модуля для рефлекса $\mathbf{s}$. Высокая размерность (N порядка 10^4-10^5 для макромолекулярных структур) и высокочастотный колебательный характер правой части выражения (5) не позволяют практически реализовать столь прямолинейный подход. В то же время с учетом современных алгоритмов и компьютерных мощностей могут быть проведены

процедуры локальной минимизации, которые позволяют уточнить предварительные значения координат при наличии приближенной атомной модели.

Традиционное рентгеновское определение макромолекулярной структуры состоит из трех основных этапов:

- восстановление (приближенных) значений фаз структурных факторов и вычисление соответствующего синтеза Фурье $\rho_S(\mathbf{r})$;
- построение приближенной атомной модели путем разложения синтеза Фурье в сумму вкладов атомов (4);
- уточнение координат атомов (и других параметров, в общем случае) путем минимизации (6).

1.2 Статистическое моделирование

Исходная цель рентгеновского исследования вполне "детерминистская" - определение координат атомов конкретного объекта. Однако при решении этой проблемы оказался полезным ряд подходов, в которых координаты атомов выступают как случайные величины. В этой заметке мы не будем обсуждать "традиционное" использование статистических методов - анализ и учет экспериментальных ошибок. Цель данной работы - обсудить некоторые вероятностные подходы в ситуациях, которые, на первый взгляд, "вероятностными" не являются.

Основная идея подхода может быть сформулирована так:

- изучаемая структура рассматривается не изолировано, а как элемент некоторого ансамбля возможных структур; при этом слово "ансамбль" подразумевает не только набор разных структур, но и заданные вероятности элементов ансамбля;
- изучение общих свойств ансамбля является иногда более простой задачей, чем изучение отдельной структуры, и это позволяет вывести некоторые "типичные" соотношения для параметров включенных в ансамбль структур;
- предполагается, что отдельная изучаемая структура подчиняется этим общим свойствам, что позволяет сделать некоторые выводы о значениях ее параметров.

Следующий пример иллюстрирует эти идею [1,4-6,21,23,24].

Пусть N - количество атомов в изучаемой структуре, и пусть все структуры из N атомов рассматриваются как равновероятные. Более строго, рассмотрим координаты атомов как независимые случайные величины, равномерно распределенные в элементарной ячейке. В этом случае структурные факторы, рассчитанные по формуле (5), становятся случайными величинами. Для каждого рефлекса может быть поставлена (и, в некотором приближении, решена) задача получения совместной функции распределения модуля и фазы $P(F,\varphi)$. Найденное совместное распределение вероятности может быть использовано для получения двух распределений $P(F)$ и $P(\varphi)$. Эти два распределения отражают, как часто могут появляться те или иные значения модуля и фазы данного рефлекса при случайном переборе всевозможных структур. Получаемое таким путем распределение модулей $P(F)$ используется на практике для шкалирования экспериментальных данных. Например, если экспериментальные значения получены в некоторой относительной шкале (числа квантов, оптические плотности пятен на рентгенограммах и т.п.), то необходимый шкальный коэффициент λ может быть выбран так, чтобы обеспечивать максимальную близость отшкалированных величин $\lambda F^{obs}(\mathbf{s})$ к математическим ожиданиям значений модулей $\langle F(\mathbf{s})\rangle$ относительно распределений $P_{\mathbf{s}}(F)$. Что же касается распределения $P(\varphi)$, то для отдельной фазы оно является равномерным и не несет никакой информации. Ситуация меняется, если поставить несколько иные вопросы к ансамблю структур.

Пусть три рефлекса $\mathbf{s}_1,\mathbf{s}_2,\mathbf{s}_3$ связаны соотношением $\mathbf{s}_1+\mathbf{s}_2+\mathbf{s}_3=\mathbf{0}$ и модули соответствующих структурных факторов известны из рентгеновского эксперимента.

Введем в рассмотрение ***фазовый инвариант*** $T=\varphi_1+\varphi_2+\varphi_3$. Рассмотрим теперь задачу нахождения условного распределения вероятностей для этого инварианта при условии, что три соответствующих модуля фиксированы результатами эксперимента. Эта условная вероятность (в некотором приближении) имеет вид распределения фон Мизеса (нормального кругового распределения)

$$P\left(T\middle|F_1,F_2,F_3\right)\propto \exp\left[\kappa\cos T\right] \quad , \quad \kappa=\frac{2F_1F_2F_3}{f^0(\mathbf{s}_1)f^0(\mathbf{s}_2)f^0(\mathbf{s}_3)N^2} \ . \tag{7}$$

Если величина параметра κ мала, то такое распределение близко к равномерному распределению. Однако при больших значениях κ, т.е. когда экспериментальные значения модулей F_1, F_2, F_3 велики, оно имеет существенный максимум при T=0. Таким образом, установлено общее свойство ансамбля структур: если у триплета рефлексов большие значения модулей, то значение соответствующего фазового инварианта, "как правило", близко к нулю.

Приписывание этого общего свойства конкретной изучаемой структуре позволяет получить практическую процедуру решения фазовой проблемы:

- отбираются тройки рефлексов, удовлетворяющие условию $\mathbf{s}_1+\mathbf{s}_2+\mathbf{s}_3=0$ и обладающих большими значениями экспериментальных модулей;
- для соответствующих фаз выписываются приблизительные равенства;

$$\varphi(\mathbf{s}_1)+\varphi(\mathbf{s}_2)+\varphi(\mathbf{s}_3)\approx 0 \ ; \tag{8}$$

- полученная система уравнений используется для определения приближенных значений фаз.

Следует заметить, что такая процедура содержит очевидный "волюнтаризм". Условие (8) выполняется статистически для случайно сгенерированных структур, и нет гарантии, что оно будет выполняться для конкретной реализации – изучаемой структуры. Тем не менее, мы постулируем его для этой структуры. Несмотря на такие методологические ограничения, указанные подходы (во множестве вариаций) широко используются в кристаллографической практике и дают осмысленные результаты, когда ансамбль структур выбирается адекватно. На выбор ансамбля можно смотреть как на средство преобразования дополнительной информации об изучаемой структуре в математическую форму, и, конечно же, результат будет зависеть от привлекаемой информации.

Ниже мы остановимся на простейшем способе введения ансамбля структур: предполагается, что координаты различных атомов независимы, и для каждого атома задано распределение вероятности $p_j(\mathbf{r})$ его координат (в общем случае эти вероятности различны для разных атомов).

1.3 Принцип максимального правдоподобия

В этой работе мы обсуждаем подход к выбору ансамбля структур, основанный на принципе максимального правдоподобия [10]. Статистическое правдоподобие - широко распространенный инструмент математической статистики и теории вероятностей, и дискуссия о математических аспектах этого инструмента лежит далеко за пределами данной статьи. Мы напомним только основную идею использования концепции правдоподобия для выбора статистической гипотезы из некоторого набора гипотез. Применительно к нашим исследованиям стандартную ситуацию вкратце можно описать так:

- существует множество "экспериментальных" измерений $x_1^{obs},x_2^{obs},...,x_M^{obs}$;
- существует гипотеза H, заключающаяся в том, что эти величины получены как результат независимой генерации случайных чисел $x_1,x_2,...,x_M$ с распределением вероятностей $p_1(x),p_2(x),...,p_M(x)$, соответственно.

Возможный подход к количественной оценке разумности этой гипотезы - оценить, как велика вероятность воспроизведения результата $x_1^{obs}, x_2^{obs}, ..., x_M^{obs}$ в рамках гипотезы H (т.е. при генерации независимых случайных величин с вероятностями $p_1(x), p_2(x), ..., p_M(x)$), например, рассчитать величину

$$L = p_1\left(x_1^{obs}\right) p_2\left(x_2^{obs}\right) \cdots p_M\left(x_M^{obs}\right) . \quad (9)$$

Величина $L=L(H)$ называется правдоподобием гипотезы H. Если рассматриваются несколько альтернативных гипотез $H_1, H_2, ..., H_K$ для объяснения результата $x_1^{obs}, x_2^{obs}, ..., x_M^{obs}$, то правдоподобие позволяет ранжировать гипотезы, и та из них, которая приводит к максимальному правдоподобию, может быть рассмотрена в качестве наиболее разумного объяснения экспериментального результата. Конечно же, метод максимального правдоподобия является лишь одним из многих методов, используемых в математической статистике для оценки статистических гипотез. Если множество альтернативных гипотез H_t является бесконечным, например, параметризовано непрерывным параметром t, то правдоподобие становится функцией $L(t)$ непрерывного переменного (или переменных).

Предположим теперь, что существует некоторая предварительная информация, дающая предпочтения некоторым гипотезам, и эта информация представлена в форме "распределения вероятностей для гипотез" (более строго, в форме "априорного" распределения вероятностей $P_{prior}(t)$ для параметра t). В таком случае функция правдоподобия может быть использована в рамках "Байесовского подхода" для вывода апостериорного распределения вероятностей гипотез (параметра t в данном случае), учитывающего результаты эксперимента

$$P_{post}(t) \propto L(t) P_{prior}(t) . \quad (10)$$

Трактовка правдоподобия как вероятности воспроизведения экспериментальных результатов позволяет конструировать простые компьютерные процедуры Монте-Карловского типа для оценки величины правдоподобия [9,17,18]. С другой стороны, такие процедуры требуют больших временных затрат, и наличие аналитического выражения для функции правдоподобия (когда его удается получить) может существенно облегчить выбор гипотезы.

2 Выбор молекулярной оболочки и определение фаз на основе правдоподобия

На начальной стадии рентгеновского исследования макромолекулярной структуры существенную роль может играть информация о молекулярной оболочке. Молекулярная оболочка - это область Ω в элементарной ячейке кристалла, в которой содержится основная часть атомов молекулы исследуемого белка (около половины объема элементарной ячейки в кристаллах белка занимает неупорядоченный растворитель). Математически оболочка может быть описана как бинарная (характеристическая) функция $\chi(\mathbf{r})$. Молекулярная оболочка дает информацию о положении и ориентации молекулы в элементарной ячейке кристалла и о внешних очертаниях молекулы. Иногда в качестве кандидатов на решение выступают несколько альтернативных оболочек. Ранжирование оболочек с помощью правдоподобия дает возможность сделать выбор одной из них [9,11,17].

Предположим, что экспериментальные значения модулей структурных факторов $\{F^{obs}(\mathbf{s})\}, \mathbf{s} \in S$ известны, и рассмотрим две альтернативные оболочки Ω_1 и Ω_2. Мы можем связать с оболочкой Ω_i статистическую гипотезу H_i:

- значения $\{F^{obs}(\mathbf{s})\}, \mathbf{s} \in S$ были получены в результате расчета по формуле (5), где координаты атомов $\{\mathbf{r}_j\}$ были выбраны случайно (независимо и равномерно) в области Ω_i.

Основанный на величине правдоподобии выбор оболочки в этом случае означает выбор оболочки, которая даст максимальную вероятность воспроизвести экспериментальные модули при помещении атомов случайным образом внутрь нее.

Ранжирование оболочек в соответствии с их правдоподобием может быть использовано для решения фазовой проблемы [2,18]. Предположим, что полное множество рефлексов с известными модулями разделено на два подмножества: S_1 (рабочие рефлексы) и S_2 (контрольные рефлексы). Рассмотрим задачу определения фаз для рефлексов из рабочего множества. Для каждого пробного набора фаз мы можем определить соответствующую ему оболочку как область наибольших значений в синтезе Фурье электронной плотности (2), рассчитанном с экспериментальными значениями модулей $\{F^{obs}(\mathbf{s})\}, \mathbf{s} \in S_1$ и пробными фазами

$$\Omega = \Omega\left(\{\varphi(\mathbf{s})\}\right) = \{\mathbf{r} : \rho_S(\mathbf{r}) \geq \rho_{crit}\} . \qquad (11)$$

Вероятность воспроизвести модули из множества S_2 по формуле (5), помещая атомы случайным образом в эту оболочку, становится в таком случае оценкой осмысленности пробных фаз и может быть использована в качестве критерия для выбора наилучшего набора фаз.

3 Оценка фазовых ошибок

3.1 Статистическое моделирование источника фазовых ошибок

Координаты атомов в моделях, построенных на промежуточных стадиях рентгеновского исследования структуры, обычно содержат некоторые ошибки. Более того, такая модель часто является неполной, то есть в ней не представлена часть атомов изучаемого объекта. Фазы, рассчитанные по такой модели по формуле (5), содержат ошибки. Статистическое моделирование может быть использовано для оценки точности рассчитанных по предварительной модели фаз [7,19].

Введем ансамбль структур следующим образом. Пусть $\{\mathbf{r}_j^{mod}\}, j = 1,..., M$ обозначают координаты атомов предварительной модели, и N - полное число атомов в изучаемой молекуле. Рассмотрим все структуры, состоящие из N атомов, при этом для j=1,...,M вероятность нахождения j-ого атома в позиции $\mathbf{r}$ зададим как $p^0\left(\left|\mathbf{r} - \mathbf{r}_j^{mod}\right|\right)$, где распределение (ошибок в координатах) $p^0(r)$ предполагается известным. Для j=M+1,...,N; координаты $\mathbf{r}_j$ будем считать распределенными равномерно в элементарной ячейке. Распределение ошибок в координатах модели $p^0(r)$ (предполагается, что оно является изотропным для всех атомов) характеризует качество модели. Эта информация об ошибках координат может быть трансформирована в оценки точности фаз, рассчитанных по предварительной атомной модели. В частности, введенный в рассмотрение ансамбль позволяет получить для каждого рефлекса распределение вероятностей для соответствующей фазы структурного фактора:

$$P_{\mathbf{s}}(\varphi) \propto \exp\left[2\frac{\alpha_s}{\beta_s} F^{mod}(\mathbf{s}) F^{obs}(\mathbf{s}) \cos\left(\varphi - \varphi^{mod}(\mathbf{s})\right)\right] . \qquad (12)$$

Эта формула показывает, что значение фазы $\varphi^{mod}(\mathbf{s})$ является наиболее вероятным (что неудивительно), и позволяет оценить ожидаемое отклонение истинного значения фазы структурного фактора от среднего значения, получаемого расчетом по предварительной модели.

3.2 Выбор статистического ансамбля (модели ошибок)

Ожидаемые отклонения истинных значений фаз от модельных определяются величинами параметров $\alpha_{\mathbf{s}}$ и $\beta_{\mathbf{s}}$. Эти параметры могут быть вычислены как

$$\alpha_{\mathbf{s}} = \int \cos 2\pi(\mathbf{s}, \mathbf{r}) p^0(r) dV_{\mathbf{r}} \ , \ \beta_{\mathbf{s}} = \left(1 - \alpha_{\mathbf{s}}^2\right) \sum_{k=1}^{M} f_k^2(s) + \sum_{k=M+1}^{N} f_k^2(s) \qquad (13)$$

в случае, когда число отсутствующих атомов и распределение ошибок в модели $p^0(r)$ известно. Однако в реальной ситуации это распределение заранее неизвестно, и адекватный выбор этого распределения является ключевым шагом на пути использования распределения (12) для оценки надежности фаз. Заметим, что, строго говоря, для использования формулы (12) нам не требуется знание самого распределения $p^0(r)$, а требуются лишь значения двух параметров (для каждого рефлекса), связанных с этим распределением. Более того, являясь формально различными для разных рефлексов, эти параметры могут рассматриваться как постоянные внутри сферического слоя $s \approx const$, поэтому необходимо определить два параметра α и β для каждого такого слоя. Максимизация правдоподобия является одним из возможных подходов к решению этой проблемы.

Введенная статистическая модель ошибок позволяет получить распределение вероятностей не только для фаз, но и для модулей структурных факторов $F(\mathrm{s})$:

$$P_{\mathbf{s}}(F) \propto \exp\left[-\frac{F^2 + \alpha_{\mathbf{s}}^2\left(F^{mod}(\mathbf{s})\right)^2}{\beta_{\mathbf{s}}}\right], \tag{14}$$

где $\alpha_{\mathbf{s}}$ и $\beta_{\mathbf{s}}$ – те же самые параметры, что и в (12). Вместе с экспериментально определенными модулями эти распределения позволяют вычислить правдоподобие

$$L = \prod_{\mathbf{s}\in S} P_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right), \tag{15}$$

которое отражает вероятность воспроизвести экспериментальные значения модулей после случайных исправлений, введенных в координаты модели, и случайного добавления необходимого числа утерянных атомов. Максимизация правдоподобия (15) позволяет получить значения параметров $\alpha_{\mathbf{s}}$ и $\beta_{\mathbf{s}}$ и использовать их посредством (12) для получения оценок надежности фаз $\varphi^{mod}(\mathbf{s})$.

4 Уточнение атомной модели, основанное на максимизации правдоподобия

В традиционном уточнении по "методу наименьших квадратов" каждому множеству параметров модели (например, текущим значениям координат атомов) ставится в соответствие множество рассчитанных по модели структурных факторов. Целью уточнения является выбор параметров атомов таких, что соответствующие рассчитанные модули наилучшим образом соответствуют экспериментальным данным. Традиционно эта согласованность выражается разницей

$$\begin{aligned} Q_{LSQ} &= \sum_{\mathbf{s}\in S} w_{\mathbf{s}}\left(\kappa F^{mod}(\mathbf{s}) - F^{obs}(\mathbf{s})\right)^2 \\ &= const + \sum_{\mathbf{s}\in S}\left\{w_{\mathbf{s}}\kappa^2\left(F^{mod}(\mathbf{s})^2\right) - 2w_{\mathbf{s}}\kappa F^{mod}(\mathbf{s})F^{obs}(\mathbf{s})\right\} \end{aligned}, \tag{16}$$

где $w_{\mathbf{s}}$ – некоторые веса, а κ – шкальный коэффициент. Предполагается, что критерий (16) достигает своего минимума (в идеальном случае равного нулю) для точных значений параметров. Необходимость в статистическом моделировании появляется, когда атомная модель содержит неустранимые в процессе минимизации ошибки и, следовательно, структурные факторы, рассчитанные по соответствующей модели, отличаются от экспериментальных значений даже тогда, когда значения параметров модели являются точными [13]. Простейший пример - уточнение неполной атомной модели. В этом случае модули структурных факторов, рассчитанные по точным координатам частичной модели, все еще отличаются от "правильных" значений. Источник такого расхождения неустраним, пока не изменен характер атомной модели, т.е. пока в нее не включены дополнительные атомы.

При "статистическом" уточнении для описания исследуемой структуры используется одновременно два объекта: традиционная атомная модель и

статистическая модель для компенсации неустранимых ошибок. Например, можно ожидать, что рассчитанные по частичной модели модули будут равны экспериментальным, если частичную модель с точными атомными координатами дополнить необходимым количеством отсутствующих атомов, расположив их в правильных местах. Это равенство не может быть достигнуто без точного определения координат этих отсутствующих атомов. С другой стороны, можно оценить вероятность получить это равенство, точно или хотя бы приближенно, после того, как отсутствующие атомы будут добавлены случайным образом к частичной модели. Можно ожидать, что эта вероятность будет наибольшей, когда мы попытаемся дополнять точную частичную модель, и будет ниже, если частичная модель содержит ошибки в положениях атомов. Следовательно, эта вероятность (правдоподобие) может быть использована как целевая функция для оценки качества частичной модели.

Более формально, при таком рассмотрении каждому множеству атомных параметров ставится в соответствие совместное распределение вероятностей модулей структурных факторов (а не отдельный набор рассчитанных модулей, как при стандартном уточнении). Структурные факторы являются теперь случайными величинами, поскольку они отвечают текущей частичной атомной модели с добавленными случайными исправлениями. Распределения вероятностей этих модулей являются различными для различных наборов атомных параметров, и, конечно же, зависят от заданной схемы (вероятностной модели) необходимых исправлений. Цель статистического уточнения атомных параметров может быть сформулирована как выбор множества параметров, для которых соответствующее совместное распределение вероятности для модулей наилучшим образом соответствует экспериментальным данным. Значение правдоподобия (т.е. вероятности воспроизвести множество экспериментальных данных в рамках этого распределения вероятностей) является примером количественной оценки степени этого соответствия.

В процедуре, называемой ML-уточнение [3,12,13,15,16], где *ML* обозначает *Maximum Likelihood*, минимизируемый критерий - это взятый с обратным знаком логарифм правдоподобия [7,9] (точнее, некоторая аналитическая аппроксимация этой функции). Часть критерия, зависящая от параметров атомной модели, может быть при этом представлена как

$$Q_{ML} = \sum_{\mathbf{s}\in S} \left\{ \frac{\alpha_{\mathbf{s}}^2 \left(F^{mod}(\mathbf{s})\right)^2}{\beta_{\mathbf{s}}} - \ln\left(I_0\left(\frac{2\alpha_{\mathbf{s}} F^{mod}(\mathbf{s}) F^{obs}(\mathbf{s})}{\beta_{\mathbf{s}}} \right) \right) \right\} . \qquad (17)$$

Здесь, как и в (16), $F^{obs}(\mathbf{s})$ - экспериментальное значение модуля структурного фактора для рефлекса **s,** а $F^{mod}(\mathbf{s})$ - соответствующее значение, рассчитанное по неполной атомной модели. Такой тип функции правдоподобия был получен сначала в предположении, что отсутствующие атомы добавляются независимо и равномерно в элементарную ячейку, и что ошибки в координатах атомов имеют одно и то же радиальное распределение для всех атомов модели. Можно показать, что некоторые более сложные статистические модели, рассмотренные в [8,14,20,22], приводят к такому же типу функции правдоподобия. Параметры $\alpha_{\mathbf{s}}$ и $\beta_{\mathbf{s}}$ здесь те же, что и в (12, 14). Они отражают ожидаемый характер неустранимых ошибок в ***уточненной*** модели. Необходимо подчеркнуть, что различные предположения о характере неустранимых ошибок в уточненной атомной модели ведут к различным функциям правдоподобия, хотя и представляемым в одной и той же форме (17). Качество атомной модели, получаемой минимизацией выражения (18), может сильно зависеть от величин используемых параметров $\alpha_{\mathbf{s}}$ и $\beta_{\mathbf{s}}$, и неадекватный выбор гипотезы о характере ошибок модели может существенно осложнить работу. Для оценки этих параметров может быть использована изложенная выше (п.3) процедура.

Работы по данной тематике были поддержаны грантами РФФИ. Частично исследования были поддержаны программой межакадемического сотрудничеством РАН – CNRS (Франция). А.Уржумцев благодарен Pole “Intelligence Logicielle” и CRVHP, LORIA, Nancy за финансовую поддержку. Авторы благодарны Н.Л.Луниной, Т.Е.Петровой, Т.П.Сковорода, Е.А.Вернословой и A.D.Podjarny за их вклад в различные этапы исследований, представленных в этой статье.

ЛИТЕРАТУРА

1. Bricogne G. 1984. Maximum Entropy and the Foundations of Direct Methods. *Acta Cryst.* **A40.** 410-455.
2. Bricogne G., Gilmore C. J. 1990. A multisolution method of phase determination by combined maximization of entropy and likelihood. I. *Theory, algorithms and strategy*. *Acta Cryst*. **A46**. 284-29.
3. Bricogne G., Irwin J. 1996. Maximum-Likelihood Refinement of incomplete models with BUSTER + TNT. Proceedings of the CCP4 Study Weekend. Daresbury Laboratory, Warrington. England. 85-92.
4. Cochran W. 1955. Relations between the phases of structure factors. *Acta Cryst*. **5**. 473–478.
5. Giacovazzo C. 1999. *Direct Phasing in Crystallography* .Oxford University Press. Oxford.
6. Hauptman, H., Karle, J.: Solution of the phase problem. I. The centrosymmetric crystal. American Crystallographic Association Monograph, 3. Wilmington: The Letter Shop (1953)
7. Lunin V.Yu., Urzhumtsev A.G. 1984. Improvement of Protein Phases by Coarse Model Modification. *Acta Cryst.* **A40**. 269-277.
8. Lunin V.Yu., Skovoroda T.P. 1995. R-free Likelihood-Based Estimates of Errors for Phases Calculated from Atomic Models. *Acta Cryst*. **A51.** 880-887.
9. Lunin V.Yu., Lunina N.L., Petrova T.E., Urzhumtsev A.G., Podjarny A.D. 1998. On the Ab initio solution of the Phase Problem for Macromolecules at Very Low Resolution. II. Generalized Likelihood Based Approach to Cluster Discrimination. *Acta Cryst*. **D54**. 726-734.
10. Lunin V.Y. 1997. The likelihood based choice of priors in statistical approaches to the phase problem. In: S.Fortier (ed.): Direct Methods for Solving Macromolecular Structures, NATO ASI Series C, Vol.507. Kluwer Academic Publishers, the Netherlands 451-454
11. Lunin V.Y., Lunina N.L., Petrova T.E., Skovoroda T.P., Urzhumtsev A.G., Podjarny A.D. 2000. Low-resolution ab initio phasing: problems and advances. *Acta Cryst.* **D56**. 1223-1232.
12. Lunin V.Y., Urzhumtsev A.G. 1999. Maximal Likelihood Refinement. It works, but why? CCP4 Newsletter on Protein Crystallography. **37**. 14-28.
13. Lunin V.Y., Afonine P.V., Urzhumtsev A.G. 2002. Likelihood-based refinement. I. Irremovable model errors. *Acta Cryst.* **A58**. 270-282.
14. Luzzati V. 1952. Traitement Statistique des Erreurs dans la Determination des Structures Cristallines. *Acta Cryst.* **5.** 802-810.
15. Murshudov G. N., Vagin A. A., Dodson E. J. 1997. Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Cryst.* **D53**. 240-255.
16. Pannu N. S. & Read R. J. 1996. Improved Structure Refinement Through Maximum Likelihood. *Acta Cryst.* **A52**. 659-668.
17. Petrova T.E., Lunin V.Y., Podjarny A.D. 1999. A likelihood-based search for the macromolecular position in the crystalline unit cell. *Acta Cryst.* **A55**. 739-745

18. Petrova T.E., Lunin V.Y., Podjarny A.D. 2000. Ab initio low-resolution phasing in crystallography of macromolecules by maximization of likelihood. *Acta Cryst*. **D56.** 1245-1252.
19. Read R. J. 1986. Improved Fourier Coefficients for Maps Using Phases from Partial Structures with Errors. *Acta Cryst.* **A42**. 140-149.
20. Read R.J. 1990. Structure-Factor Probabilities for Related Structures. *Acta Cryst.* **A46.** 900-912.
21. Sheldrick, G.M., Hauptman, H.A., Weeks, C.M., Miller, R., Usón, I.: 2001. Direct methods. In: Rossmann, M., Arnold, E. (eds.): *International Tables for Crystallography*. Vol.F. Kluwer Academic Publishers, Dordrecht Boston London. 333-345
22. Srinivasan R., Parthasarathy S. 1976. *Some Statistical Applications in X-ray Crystallography*. Pergamon Press. Oxford.
23. Wilson A.J.C. 1949. The Probability Distribution of X-ray Intensities. *Acta Cryst.* **2.** 318-321.
24. Woolfson M.M. 1954. The statistical theory of sign relationship. *Acta Cryst.* **7.** 61-64.

Материал поступил в редакцию 21 апреля 2006 г., опубликован 5 мая 2006 г.