РОССИЙСКАЯ АКАДЕМИЯ НАУК

ИНСТИТУТ ТЕОРЕТИЧЕСКОЙ И ЭКСПЕРИМЕНТАЛЬНОЙ БИОФИЗИКИ

На правах рукописи
УДК 577.32

ПЕТРОВА Татьяна Евгеньевна

ИСПОЛЬЗОВАНИЕ ПРИНЦИПА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ ПРИ РЕШЕНИИ ФАЗОВОЙ ПРОБЛЕМЫ В КРИСТАЛЛОГРАФИИ МАКРОМОЛЕКУЛ

Специальность 03.00.02 – Биофизика

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Пущино 2000

Работа выполнена в Институте математических проблем биологии РАН (Пущино).

Научный руководитель: доктор физико-математических наук
Лунин В.Ю.

Официальные оппоненты: доктор биологических наук
проф. Чиргадзе Ю.Н.,
доктор физико-математических наук
Мосунов А.С.

Ведущая организация: Институт кристаллографии РАН, г.Москва.

Защита диссертации состоится "____" _____ 2000 г. в _____ часов на заседании диссертационного совета Д 200.22.01 в Институте теоретической и экспериментальной биофизики РАН по адресу:
142290 Пущино, Московская обл., ИТЭБ РАН.

С диссертацией можно ознакомиться в библиотеке ИТЭБ РАН.

Автореферат разослан "____" _____ 2000 г.

Ученый секретарь
диссертационного совета Д 200.22.01
кандидат биологических наук П.А.Нелипович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Одной из центральных проблем при определении неизвестной структуры методом рентгеновского анализа является так называемая фазовая проблема. Создание подходов к решению этой проблемы, не требующих проведения дополнительных дорогостоящих и трудоемких экспериментов или знания модели структуры, гомологичной исследуемой структуре, является одним из наиболее важных направлений современной белковой кристаллографии. Это связано прежде всего с тем, что традиционные методы решения фазовой проблемы становятся трудно применимыми при работе с большими макромолекулярными комплексами. В то же время, попытки распространения в макромолекулярную область методов, с помощью которых решается фазовая проблема для низкомолекулярных структур, приводят к успеху лишь для сравнительно небольших белков, и при этом для их использования необходимо наличие данных очень высокого разрешения, что не всегда возможно.

Цель работы. Задачей исследования являлась разработка нового подхода к решению фазовой проблемы в кристаллографии макромолекул на низком разрешении, не требующего для своего применения наличия тяжелоатомных производных или моделей известных гомологичных структур.

Новизна. Впервые в кристаллографической практике для получения значений фаз структурных факторов низкого разрешения для макромолекулярных комплексов предложен подход, основанный на фильтрации случайно сгенерированных вариантов с помощью критерия, являющегося аналогом статистического правдоподобия. Для оценки величины этого критерия предложена и реализована компьютерная процедура Монте-Карловского типа.

Научно-практическое значение работы. Предложен новый подход к определению области, занимаемой молекулой в элементарной ячейке. Подход основан на статистическом моделировании исследуемой структуры и выборе одного из нескольких альтернативных вариантов решения на основе принципа, аналогичного статистическому принципу максимального правдоподобия. Показано, что аналогичная процедура может быть применена и к задаче выбора наиболее адекватного набора значений фаз структурных факторов из ансамбля альтернативных наборов. Предложен *ab-initio* подход к решению фазовой проблемы при низком разрешении, основанный на фильтрации случайно сгенерированных фазовых наборов при помощи критерия обобщенного правдоподобия и усреднения отобранных вариантов. Предложенный подход позволяет получить приближенные значения фаз структурных факторов низкого разрешения, которые могут быть использованы в качестве стартовых для процедур расширения и уточнения наборов фаз. Синтезы Фурье, построенные с использованием полученных значений фаз, могут быть использованы в рамках метода молекулярного замещения для определения положения и ориентации модели молекулы в элементарной ячейке. Эти синтезы представляют самостоятельный интерес в случае исследования больших макромолекулярных

комплексов, когда модель низкого разрешения может нести важную структурную информацию. Показано, что предложенный подход может быть также применен для решения задачи расширения фазового набора.

Апробация работы. Результаты исследований, изложенные в работе, докладывались и обсуждались на следующих конференциях:

15-ая Европейская кристаллографическая конференция, Дрезден, Германия, 1994;

Ежегодный конгресс Американской кристаллографической ассоциации (Атланта, США, 1994);

Ежегодный конгресс Французской кристаллографической ассоциации (Гренобль, Франция, 1995);

4-ый Европейский семинар по структуре биологических макромолекул (Комо, Италия, 1995);

Международная школа по прямым методам решения макромолекулярных структур (Эриче, Италия, 1997);

Восемнадцатая Европейская кристаллографическая конференция (Прага, Чешская республика, 1998);

Восемнадцатый всемирный кристаллографический конгресс (Глазго, Шотландия, 1999);

Совещание по проблемам решения структур на низком разрешении (Йорк, Великобритания, 2000)

Публикации. По материалам диссертации опубликовано 15 работ, список которых приведен в конце автореферата.

Структура и объём диссертации.

Диссертация состоит из введения, пяти глав, двух приложений, основных выводов и результатов и списка цитируемой литературы (85 названий). Она изложена на 132 страницах, содержит 19 рисунков и 8 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Определение неизвестной структуры в кристаллографии макромолекул состоит из трех этапов. На первом этапе выращиваются кристаллы исследуемого объекта и проводится эксперимент по рассеянию на этих кристаллах рентгеновского, нейтронного или электронного излучения. На втором этапе по полученной в эксперименте дифракционной картине рассчитывается функция распределения электронной плотности в кристалле. Конечным этапом является интерпретация функции распределения электронной плотности, то есть построение и уточнение атомной модели исследуемого вещества. Настоящая работа посвящена разработке методов, используемых на втором этапе исследования. Так как искомая функция распределения электронной плотности в кристалле является периодической, она может быть представлена в виде комплексного ряда Фурье:

$$\rho(\mathbf{r}) = \frac{1}{V}\sum_{\mathbf{h}} \mathbf{F}_{\mathbf{h}} \exp(-2\pi i \mathbf{h};\mathbf{r}), \qquad (1)$$

где V- объем элементарной кристаллической ячейки, а трехмерный вектор $\mathbf{h}=(h_1,h_2,h_3)$ принимает все целочисленные значения. Комплексный коэффициент $\mathbf{F}_{\mathbf{h}} = |F_{\mathbf{h}}|e^{i\varphi}$ называется структурным фактором. Значения модулей структурных факторов определяются в эксперименте. Они пропорциональны корню квадратному из значений интенсивностей пятен (рефлексов) на дифракционной картине рассеяния излучения кристаллом исследуемого вещества. Однако эксперимент не позволяет получить необходимые для расчета функции (1) значения фаз структурных факторов $\{\varphi_{\mathbf{h}}\}$. Их определение является центральной проблемой этого этапа - фазовой проблемой. Основными подходами к решению фазовой проблемы в кристаллографии макромолекул в настоящее время являются:

- Метод множественного изоморфного замещения
- Метод молекулярного замещения
- Многоволновая аномальная дифракция

Однако, с помощью перечисленных методов фазовую проблему далеко не всегда удается решить. Для использования метода молекулярного замещения требуется знание модели структуры, гомологичной исследуемой структуре. Применение методов изоморфного замещения и аномальной дифракции наталкивается на большие трудности при получении тяжелоатомных производных с требуемыми свойствами. Поэтому в настоящее время существует значительный интерес к разработке новых подходов к решению фазовой проблемы для макромолекул, свободных от этих трудностей.

Одним из возможных подходов является использование вероятностных моделей исследуемой структуры. На основе такого подхода были созданы процедуры решения низкомолекулярных соединений. Однако их непосредственное использование применительно к макромолекулам к успеху не приводит из-за принципиальных ограничений на число атомов в исследуемом объекте, лежащих в основе этих процедур. Поэтому возникает потребность в разработке новых методик, применимых для работы с биологическими макромолекулами и их комплексами, в которых число атомов может превышать десятки и сотни тысяч. Представляемая работа посвящена разработке подхода к решению фазовой проблемы для макромолекул на низком разрешении. Под рефлексами низкого разрешения в кристаллографии понимаются рефлексы, находящиеся в малоугловой зоне рассеяния. Соответствующие структурные факторы отвечают гармоникам Фурье с большими "длинами волн", передающим грубые детали объекта. Задача определения фаз низкого разрешения, в первую очередь, представляет интерес как начальный этап решения фазовой проблемы. Согласно одному из возможных подходов к решению фазовой проблемы для макромолекул, сначала определяются значения фаз структурных факторов только низкого разрешения, а затем этот набор расширяется и уточняется. Кроме того, знание фаз низкого разрешения

представляет самостоятельный интерес в ряде случаев, поскольку позволяет определять местоположение исследуемого объекта в элементарной ячейке кристалла и его внешние очертания.

ГЛАВА I. ЛИТЕРАТУРНЫЙ ОБЗОР

В этой главе изложены идеи, лежащие в основе вероятностного подхода к решению фазовой проблемы. Кратко рассмотрены методы определения структуры низкомолекулярных соединений, а также анализируются новые идеи по развитию вероятностного подхода к решению фазовой проблемы для макромолекул.

ГЛАВА II. ВЫБОР ОБЛАСТИ МОЛЕКУЛЫ. ВЫБОР АПРИОРНОГО РАСПРЕДЕЛЕНИЯ КООРДИНАТ АТОМОВ НА ОСНОВЕ ПРИНЦИПА МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ. ПОНЯТИЕ ОБОБЩЕННОГО ПРАВДОПОДОБИЯ

При низком разрешении элементарная ячейка может быть приблизительно разделена на две области: область молекулы и область растворителя. В данной главе рассматривается задача выбора области молекулы из набора альтернативных областей. Далее, в главе 4, показывается, как к этой задаче может быть сведена основная проблема, которой посвящена работа, а именно, - проблема определения значений фаз низкого разрешения.

Выбор области молекулы из набора альтернативных областей

Подход основан на предположении, что в ситуации, когда область молекулы известна, существуют шансы воспроизвести с разумной точностью набор экспериментальных модулей структурных факторов низкого разрешения в результате следующего компьютерного эксперимента. Необходимое число атомов размещается в области молекулы случайным образом, и по такой конфигурации рассчитываются структурные факторы.

При сравнении нескольких альтернативных областей каждая из них поочередно выбирается в качестве гипотетической области молекулы и рассчитывается вероятность того, что при случайном размещении атомов именно в этой области рассчитанные модули структурных факторов будут близки экспериментальным. Эту вероятность можно приближенно рассчитать, повторяя многократно процесс генерации атомов в тестируемую область. Та область, для которой эта вероятность оказывается наиболее высокой, выбирается в качестве ответа.

Предлагаемый подход к выбору области молекулы из набора альтернативных областей имеет аналогию со статистическим принципом максимального правдоподобия, который в нашем случае может быть сформулирован следующим образом. Пусть $\{F^{obs}(\mathbf{h})\}$ обозначает набор модулей

структурных факторов, полученных в рентгеновском эксперименте. Отвлечемся от реального происхождения этих величин и рассмотрим следующую математическую процедуру:

- координаты каждого из N атомов исследуемого объекта выбираются случайно и независимо от других с бинарной плотностью априорного распределения вероятностей:

$$q(\mathbf{r}) = \begin{cases} 1\ V & \text{в области молекулы} \\ 0 & \text{в области растворителя} \end{cases} \qquad (2)$$

- по полученной конфигурации атомов рассчитываются модули структурных факторов $\{F^{obs}(\mathbf{h})\}$.

Получающиеся в результате выполнения такой процедуры модули структурных факторов $\{F^{calc}(\mathbf{h})\}$ являются случайными величинами, и можно говорить об их распределении вероятностей. Предположим теперь, что набор модулей $\{F^{obs}(\mathbf{h})\}$ был получен в результате применения процедуры такого типа, и ставится вопрос, какое именно априорное распределение $q(\mathbf{r})$ из распределений вида (2) было использовано в этом процессе. Этот вопрос, по существу, является вопросом выбора статистической гипотезы (определяемой в данном случае функцией $q(\mathbf{r})$ на основе реализации $\{F^{obs}(\mathbf{h})\}$. Одним из распространенных в математической статистике подходов к решению подобных задач является подход, основанный на выборе гипотезы, обладающей наибольшим значением величины правдоподобия. Правдоподобие определяется в данном случае, как плотность вероятности получения величин $\{F^{calc}(\mathbf{h})\}$, совпадающих с величинами $\{F^{obs}(\mathbf{h})\}$:

$$L(q(\mathbf{r})) = P\left\{F_{\mathbf{h}} = F_{\mathbf{h}}^{o},\ \textit{для всех}\ \mathbf{h}\right\}. \qquad (3)$$

Задача выбора области молекулы среди альтернативных областей сводится к задаче выбора априорного распределения координат атомов исследуемой структуры среди распределений вида (2). В качестве оптимального выбирается то распределение, которому соответствует максимальное значение правдоподобия (3). Сформулированный подход к выбору области молекулы мог бы рассматриваться как реализация принципа максимального правдоподобия, если бы набор $\{F^{obs}(\mathbf{h})\}$ и в самом деле был получен как результат выполнения указанной математической процедуры. Однако экспериментальные значения $\{F^{obs}(\mathbf{h})\}$ имеют иное, более сложное, происхождение. Поэтому одной из главных задач диссертации являлось исследование применимости этого подхода в работе с реальными объектами.

Обобщенное правдоподобие

Применение принципа максимального правдоподобия требует умения вычислять значение величины правдоподобия, в предположении, что априорное распределение $q(\mathbf{r})$ задано. Эта задача является весьма сложной, если пытаться

решать ее аналитически. В данной работе вместо статистического правдоподобия (3) предлагается использовать другой критерий, являющийся его аналогом, но допускающий существенно более легкое практическое вычисление. Мы называем этот критерий обобщенным правдоподобием и определяем его как вероятность того, что рассчитанные модули структурных факторов достаточно близки экспериментальным модулям:

$$\mathrm{GL}_{\omega} = \mathrm{P}\left\{C\left(\left\{\mathrm{F}^{\mathrm{calc}}(\mathbf{h})\right\},\left\{\mathrm{F}^{\mathrm{obs}}(\mathbf{h})\right\}\right) \geq \omega\right\}, \qquad (4)$$

где C - некоторая мера близости двух наборов модулей структурных факторов, а ω - параметр, называемый уровнем срезки, задающий приемлемую степень близости двух наборов модулей структурных факторов. Очевидно, что величина критерия (4) зависит от выбора меры близости C и параметра ω. При ужесточении требования близости наборов величина обобщенного правдоподобия стремится в пределе к величине обычного правдоподобия. В проведенных тестах в качестве меры близости C использовался коэффициент корреляции модулей:

$$C_F\left(\left\{F(\mathbf{h})\right\},\left\{F^{obs}(\mathbf{h})\right\}\right) = \frac{\sum_{\mathbf{h}}\left(F(\mathbf{h})-\langle F\rangle\right)\left(F^{obs}(\mathbf{h})-\langle F^{obs}\rangle\right)}{\sqrt{\sum_{\mathbf{h}}\left(F(\mathbf{h})-\langle F\rangle\right)^2\sum_{\mathbf{h}}\left(F^{obs}(\mathbf{h})-\langle F^{obs}\rangle\right)^2}}, \qquad (5)$$

где $\langle F\rangle$ - среднее значение модуля для всех рассматриваемых рефлексов.

Преимуществом использования критерия (4) является то, что его приближенные значения могут быть легко вычислены в ходе следующей компьютерной процедуры. Случайным образом генерируется большое количество моделей, состоящих из N_{glob} псевдоатомов, расположенных в тестируемой области. Для каждой генерируемой модели рассчитывается набор модулей структурных факторов $\{F(\mathbf{h})\}$ и коэффициент их корреляции с наблюдаемыми модулями (5). Приближенное значение обобщенного правдоподобия вычисляется как отношение числа моделей, для которых значение C_F превышает выбранное значение ω, к общему числу сгенерированных моделей:

$$\mathrm{GL}_{\omega} \approx \frac{\text{Количество моделей, для которых } C_F \geq \omega}{\text{Общее число моделей}} \qquad (6)$$

Общая схема расчета критерия обобщенного правдоподобия для тестируемой области молекулы представлена на рис. 1.

Предлагаемый подход к определению области молекулы на основе максимизации обобщенного правдоподобия тестировался на известных структурах, когда правильный ответ известен, т.е. известны истинные значения фаз структурных факторов и область, занятая молекулой. Для сравнения результатов применения тестируемого подхода с точными значениями в работе применялись различные критерии, называемые далее контрольными критериями. Мы будем ссылаться ниже на два таких критерия. Первый из них -

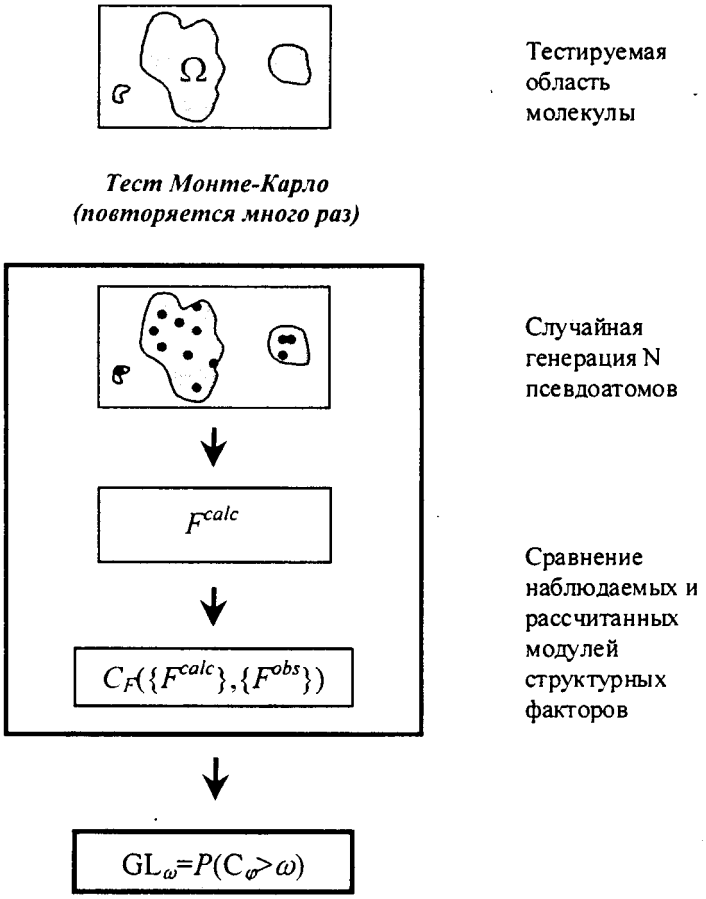


Рис. 1. Схема расчета величины обобщенного правдоподобия для гипотетической области молекулы

это коэффициент корреляции значений определенных нами фаз с правильными фазами [8]:

$$C_\varphi = \sum_{\mathbf{h}} (F_{\mathbf{h}}^{o})^2 \cos(\varphi_{\mathbf{h}}^{true} - \varphi_{\mathbf{h}}) \Big/ \sum_{\mathbf{h}} (F_{\mathbf{h}}^{o})^2 \qquad (7)$$

Второй критерий – "функция захвата" T, определяемая как отношение числа атомов модели, попадающих внутрь тестируемой области, к общему числу атомов модели. Заметим, что каждый раз перед вычислением этих критериев осуществлялась процедура выравнивания соответствующих синтезов Фурье или соответствующих областей [7]. Необходимость такого выравнивания связана с тем, что две на первый взгляд сильно различающиеся карты (или области) могут

оказаться почти идентичными при соответствующем изменении начала координат или энантиоморфа

ГЛАВА III. ОПРЕДЕЛЕНИЕ ПОЗИЦИИ МАКРОМОЛЕКУЛЫ В ЭЛЕМЕНТАРНОЙ ЯЧЕЙКЕ С ПОМОЩЬЮ КРИТЕРИЯ ОБОБЩЕННОГО ПРАВДОПОДОБИЯ.

Первой задачей, к которой был применен критерий обобщенного правдоподобия была задача определения положения центра масс макромолекулы в элементарной ячейке. Для решения этой задачи неоднократно разными исследователями использовался поиск с помощью одного Гауссова атома или с помощью одной сферы [6], [3].Однако, наряду с успешными примерами, такой поиск часто приводит к неправильным результатам или же результаты трудно поддаются интерпретации.

Подход основан на предположении, что при низком разрешении форму молекулы можно считать почти сферической. В таком случае вопрос о положении центра молекулы может быть переформулирован как вопрос выбора одной из всевозможных сферических областей. В качестве критерия, на основании которого выбирается лучшая сферическая область, в данной работе использовалось значение обобщенного правдоподобия. Исследовалось, может ли такой поиск обеспечить более стабильные результаты по сравнению с простым поиском с помощью одного Гауссова атома.

Поиск с помощью одного Гауссова атома представляет собой следующую процедуру. В элементарной ячейке вводится некоторая сетка, в узлы этой сетки поочередно помещается один Гауссов атом, и по такой модели рассчитываются значения структурных факторов. В качестве правильной позиции центра выбирается точка, соответствующая максимальному соответствию между рассчитанными и наблюдаемыми модулями структурных факторов, например, максимуму коэффициента корреляции рассчитанных и наблюдаемых модулей (5).

В тестах с комплексом аминоацил-тРНК-синтетазы с тРНК [7] поиск с помощью одного Гауссова атома давал ложные решения, расположенные на осях симметрии (положения p2-p5 Таб.1),

Пики	Правдоподобие GL ω=0.60	Функция захвата T	Фазовая корреляция $\langle C_\varphi \rangle$	C_F для одного Гауссова атома
p1	0.48	0.46	0.76	0.63
p2	1.00	0.0	-0.15	0.70
p3	1.00	0.05	-0.03	0.71
p4	1.00	0.03	-0.11	0.67
p5	1.00	0.02	0.18	0.66

Таблица 1. Результаты поиска позиции центра комплекса аминоацил-тРНК-синтетазы с тРНК (d=40 Å) с помощью критерия обобщенного правдоподобия

для областей с фиксированным радиусом сферы и с помощью одного Гауссова атома.

а правильное решение (p1) было только пятым по величине коэффициента корреляции (5). Аналогичные результаты дал и поиск с помощью критерия обобщенного правдоподобия, в случае, когда области строились как объединение сфер фиксированного радиуса. Однако, когда фиксировался объем области, являющейся объединением всех симметрично связанных сфер, подход, основанный на максимизации обобщенного правдоподобия, позволил избежать этих ложных решений и получить правильноерешение (Таб.2).

Пики	$V/V_{cell}=0.30$		
	GL $\omega=0.58$	T	$\langle C_\varphi \rangle$
p1	0.30	0.54	0.72
p2	0.00	0.00	-0.17
p3	0.10	0.18	-0.15
p4	0.00	0.15	-0.01
p5	0.02	0.15	0.30

Таблица 2. Результаты поиска позиции центра комплекса аминоацил-тРНК-синтетазы с тРНК (d=40 Å) с помощью критерия обобщенного правдоподобия для областей с фиксированным объемом.

При исследовании структуры рибосомной частицы T50S для *Thermus Thermophilus* [13] результаты, полученные с помощью критерия обобщенного правдоподобия, совпали с результатами, полученными независимо другими методами (FAM, метод молекулярного замещения [12]). Результаты совпали также и с результатами, полученными с помощью одного Гауссова атома. Однако в случае использования критерия обобщенного правдоподобия был получен гораздо более высокий контраст сигнала для правильного решения, чем при поиске с помощью одного атома. Аналогично с данными для белка Protein G [5] критерий обобщенного правдоподобия позволил получить решение с более высоким контрастом, чем поиск с помощью одного Гауссова атома. Как и ожидалось, такой подход позволяет найти центр масс макромолекулы в элементарной ячейке в тех случаях, когда область молекулы имеет приблизительно сферическую форму. С помощью предложенного метода не удалось получить правильного решения при работе с данными для РНКазы SA [10], γ-кристаллина IIIb [1], рибосомного фактора элонгации G [2]. Возможными причинами являются наличие двух молекул в независимой части элементарной ячейки кристаллов РНКазы и γ-кристаллина и вытянутая форма области молекулы на низком разрешении для фактора элонгации G.

ГЛАВА IV. *AB-INITIO* ОПРЕДЕЛЕНИЕ ЗНАЧЕНИЙ ФАЗ СТРУКТУРНЫХ ФАКТОРОВ ДЛЯ РЕФЛЕКСОВ НИЗКОГО РАЗРЕШЕНИЯ С ПОМОЩЬЮ КРИТЕРИЯ ОБОБЩЕННОГО ПРАВДОПОДОБИЯ

Данная глава посвящена определению значений фаз $\{\varphi(\mathbf{h})\}_{\mathbf{h}\in S}$ структурных факторов для некоторого фиксированного набора рефлексов S. Предполагается, что модули структурных факторов этого набора $\{F^{obs}(\mathbf{h})\}_{\mathbf{h}\in S}$ известны из эксперимента.

Для поиска решения применялся подход, при котором вначале рассматривается большое количество пробных значений фаз $\{\varphi(\mathbf{h})\}_{\mathbf{h}\in S}$ и на основании некоторого критерия из них выбирается решение. При этом в качестве критерия отбора рассматривался критерий обобщенного правдоподобия, для чего задача определения набора фаз была переформулирована как рассмотренная выше задача выбора области в элементарной ячейке. Любому пробному набору фаз низкого разрешения можно поставить в соответствие область молекулы. Это можно сделать, например, следующим образом. Рассматриваемый набор фаз используется совместно с экспериментальными модулями для расчета синтеза Фурье (1). В качестве области молекулы в элементарной ячейке выбирается область заданного объема, содержащая точки с максимальными значениями синтеза (Рис.2). Таким образом, альтернативные фазовые наборы приводят к альтернативным гипотетическим областям молекулы. Задача выбора наилучшего фазового набора формулируется теперь как задача выбора области молекулы из набора альтернативных областей, для решения которой может быть использован критерий обобщенного правдоподобия (4).

Выбор наилучшего фазового набора среди нескольких альтернативных фазовых наборов

Первая задача, к которой была применена изложенная выше схема, - это выбор из нескольких альтернативных фазовых наборов. Такая задача представляет интерес в связи с тем, что на начальной стадии решения фазовой проблемы при низком разрешении различные методы не дают, как правило, однозначного ответа, а приводят к небольшому числу альтернативных решений. Так, в тестах с экспериментальными данными для РНКазы SA разрешения 16 Å FAM (Few Atoms Model) метод привел к четырем альтернативным фазовым наборам. На рис.3 представлены значения критерия обобщенного правдоподобия при различных значениях параметра уровня срезки для этих четырех решений. (При расчете величины правдоподобия рассматривались области, выделяющие 60% объема элементарной ячейки, что приблизительно соответствует объему молекулы). Легенда содержит значения корреляции каждого из фазовых наборов с правильными фазами. Как следует из этого рисунка, в данном случае критерий обобщенного правдоподобия позволяет

идентифицировать лучшее решение. При использовании областей, построенных по тем же фазовым наборам, но выделяющим 30% элементарной ячейки, максимальное значение обобщенного правдоподобия получалось не для

Рис. 2. Схема перехода от задачи выбора наилучшего фазового набора к выбору одной из альтернативных масок области молекулы.

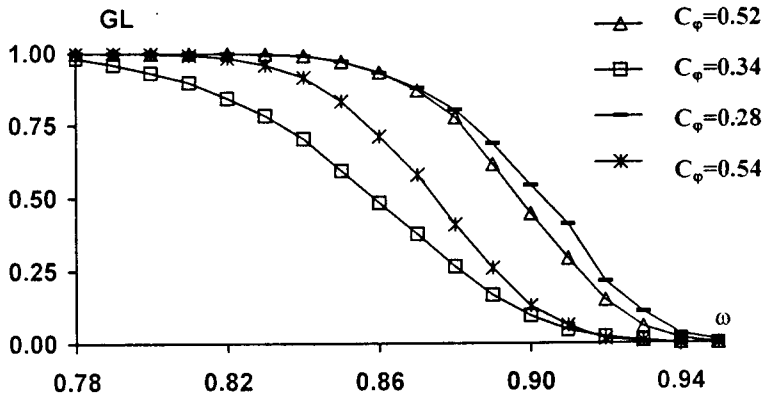


Рис. 3. Зависимость величины критерия обобщенного правдоподобия от уровня срезки ω для четырех альтернативных масок, полученных с помощью FAM-метода. Дифракционные данные для RNAse Sa разрешения d=16Å-inf.

лучшей области. Следует заметить, что объем, занимаемый молекулой в элементарной ячейке кристалла, определяется, как правило, достаточно точно на основании значения молекулярного веса, то есть эта информация при проведении реального исследования структуры известна.

***Ab-initio* определение фаз низкого разрешения**

Подход, примененный выше к выбору одного из нескольких альтернативных фазовых наборов, может быть расширен до выбора из большого числа фазовых вариантов, а в благоприятной ситуации и из всех возможных наборов фаз. В проведенных тестах использовались два способа генерации фазовых наборов: "полный" перебор и случайная генерация. В общем случае значение фазы структурного фактора может иметь любое значение в интервале от 0 до 2π. Однако из-за наличия кристаллографической симметрии фазы некоторых рефлексов, называемых центросимметричными, могут принимать только одно из двух возможных значений. Заметим, что во многих пространственных группах среди рефлексов низкого разрешения значительное количество составляют именно центросимметричные рефлексы. Свободу выбора значений фаз для нецентросиммтеричных рефлексов можно также ограничить, разрешив им, например, принимать только значения: $\pm\pi/4$, $\pm 3\pi/4$. (Это ограничивает возможную ошибку в значении фазы величиной 45 градусов, что приемлемо при низком разрешении.) В таком случае число потенциально возможных фазовых наборов становится конечным, и можно пытаться организовать их полный перебор. Такой способ перебора был использован, когда рассматриваемый набор рефлексов представлял собой небольшое количество только сильных рефлексов. В тестах, описанных ниже, при работе со всеми рефлексами из некоторой зоны разрешения использовалась также случайная генерация фазовых наборов.

В первом тесте с данными для комплекса аминоацил-тРНК-синтетазы с тРНК было сформировано 4096 фазовых наборов путем полного перебора для 12 сильнейших рефлексов в диапазоне разрешения 68Å-inf. На рис.4 представлено распределение значений обобщенного правдоподобия и фазовой корреляции для различных фазовых наборов. Можно видеть, что вариант, наиболее близкий к правильному варианту, имеет одно из самых высоких значений правдоподобия. Однако существуют плохие варианты с высоким значением правдоподобия и, наоборот, хорошие варианты с низким значением правдоподобия.

Аналогичные результаты были получены при полном переборе значений фаз сильных рефлексов для γ-кристаллина IIIb (d=29 Å) и фактора элонгации G (d=34 Å). Таким образом, не наблюдается однозначной зависимости между качеством фазового набора и значением исследуемого критерия, и задача определения фаз не может быть сформулирована, в общем случае, как задача поиска глобального максимума функции правдоподобия.

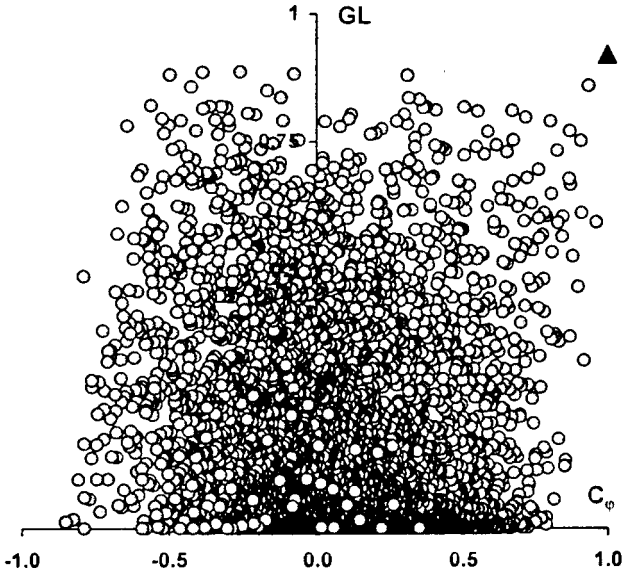


Рис. 4. *Ab-initio* определение фаз. Комплекс аминоацил-тРНК-синтетазы с тРНК. Каждый фазовый набор представлен точкой, координаты которой равны коэффициенту фазовой корреляции с правильными фазами (7) и критерию обобщенного правдоподобия (6), рассчитанных для этого набора. Треугольник соответствует варианту, ближайшему к правильному решению.

Использование дополнительной информации

Наличие дополнительной информации об объекте может в значительной мере повысить эффективность применения предложенного критерия отбора. Так, в тесте с комплексом аминоацил-тРНК-синтетазы с тРНК было показано, что при наличии некоторой дополнительной информации об исследуемой структуре критерий обобщенного правдоподобия может однозначно указывать на правильное решение. В этом случае было известно, что молекула является димером. Из этого следовало, что не должно быть областей высокой плотности на осях вращения 3-го и 4-го порядков. Учитывая этот факт, из всех 4096 фазовых наборов были отобраны только 250 вариантов, для которых построенные маски имеют наименьшее количество точек, лежащих на осях вращения 3-го и 4-го порядков. На рис.5. представлены только эти варианты. Вариант, ближайший к правильному, имеет значение правдоподобия, значительно превышающее правдоподобие для остальных рассмотренных вариантов.

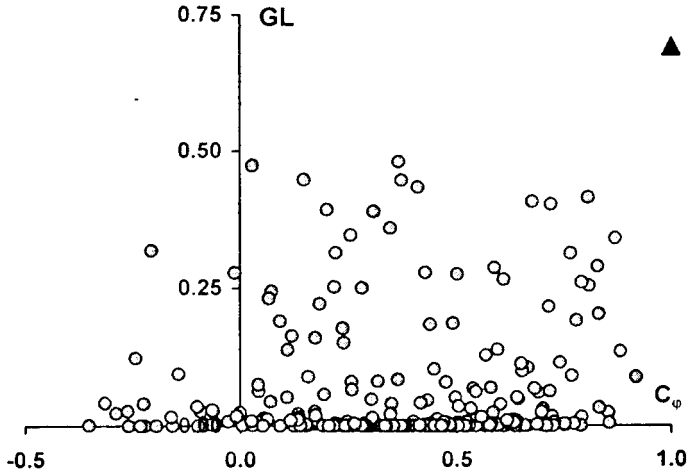


Рис. 5. *Ab-initio* определение фаз. Использование дополнительной информации о структуре. Комплекс аминоацил-тРНК-синтетазы с тРНК. Каждый фазовый набор представлен точкой, координаты которой равны коэффициенту фазовой корреляции (7) и обобщенному правдоподобию (6), рассчитанных для этого набора. Треугольник соответствует варианту, ближайшему к правильному решению.

Обогащение ансамбля фазовых наборов.

Несмотря на то, что в общем случае не обнаруживается отчетливой связи между качеством набора фаз и соответствующей величиной правдоподобия, тем не менее, обобщенное правдоподобие может быть успешно использовано для решения фазовой проблемы при использовании подхода, основанного на фильтрации исходного ансамбля фазовых наборов. При этом больше не ставится задача нахождения одного варианта, обладающего максимальным значением критерия отбора. Вместо этого формулируется задача получить ансамбль вариантов, обладающий более высоким процентным содержанием "хороших" вариантов по сравнению со стартовым ансамблем. Эта идея иллюстрируется следующим тестом. Для белка γ–кристаллин IIIb было случайно сгенерировано большое количество фазовых наборов, рассчитано значение обобщенного правдоподобия для всех вариантов, а затем были отобраны только варианты с высоким значением обобщенного правдоподобия. Сравнение распределений значений фазовой корреляции среди всех сгенерированных и среди отобранных фазовых наборов показало, что относительное число вариантов с высоким значением фазовой корреляции намного выше среди отобранных вариантов, чем среди всех (Рис.6). Аналогичный результат был получены в тестах с фактором элонгации G и комплексом аминоацил-тРНК-синтетазы с тРНК. Таким образом, критерий

обобщенного правдоподобия может служить как фильтр для отбора ансамбля вариантов с более высокой концентрацией "хороших" вариантов.

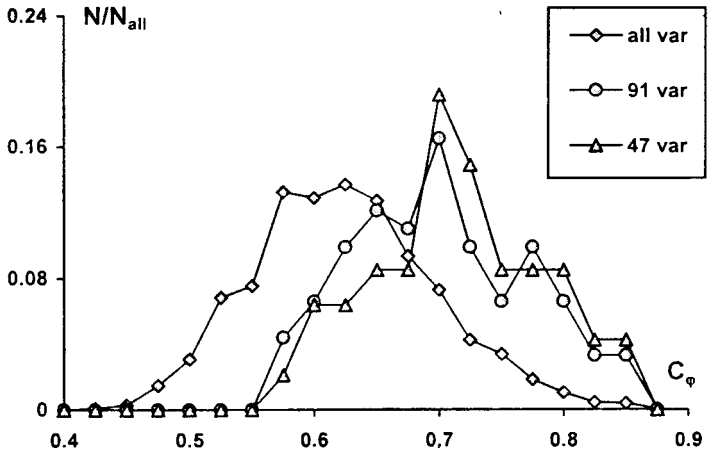


Рис. 6 Сравнение распределения значений фазовой корреляции C_φ по вариантам для разных ансамблей фазовых наборов: 2000 случайно сгенерированных вариантов, 91 вариант с GL >=0.47 и 47 вариантов с GL>=0.5. γ-кристаллин IIIb.; d=23 Å-inf,

Усреднение отобранных вариантов позволяет получить приближенные значения фаз факторов и соответствующие им показатели достоверности
определения этих фаз:

$$m(\mathbf{h})\exp\left[\varphi^{best}(\mathbf{h})\right] = \frac{1}{M}\sum_{j=1}^{M}\exp\left[i\varphi_j(\mathbf{h})\right]. \qquad (8)$$

Фазы, полученные усреднением отобранных вариантов, имеют более высокое значение фазовой корреляции с правильными фазами, чем решение, полученное усреднением по всем случайно сгенерированным вариантам (Таб. 3).

	γ-кристаллин IIIb C_φ; d=23 Å		*Фактор элонгации G* C_φ; d=29 Å		*комплекс AspRS* C_φ; d=50 Å
	по всем рефл.	без 4 рефл. с фикс. фазами	по всем рефл.	без 4 рефл. с фикс. фазами	по всем рефл.
v1	0.71	-0.01	0.60	0.21	0.22
v2	0.83	0.64	0.61	0.47	0.49

Таблица.3. Сравнение решений, полученных усреднением 2000 случайно сгенерированных вариантов (v1) и усреднением 100 вариантов с

максимальными значениями обобщенного правдоподобия (v2) для трех структур.

Использование полученных фаз для расчета синтеза Фурье (1) позволяет определить внешние очертания молекул и их расположение в элементарной ячейке кристалла. На рис.7 показан результат применения предложенный процедуры к исследованию структуры γ-кристаллина IIIb.

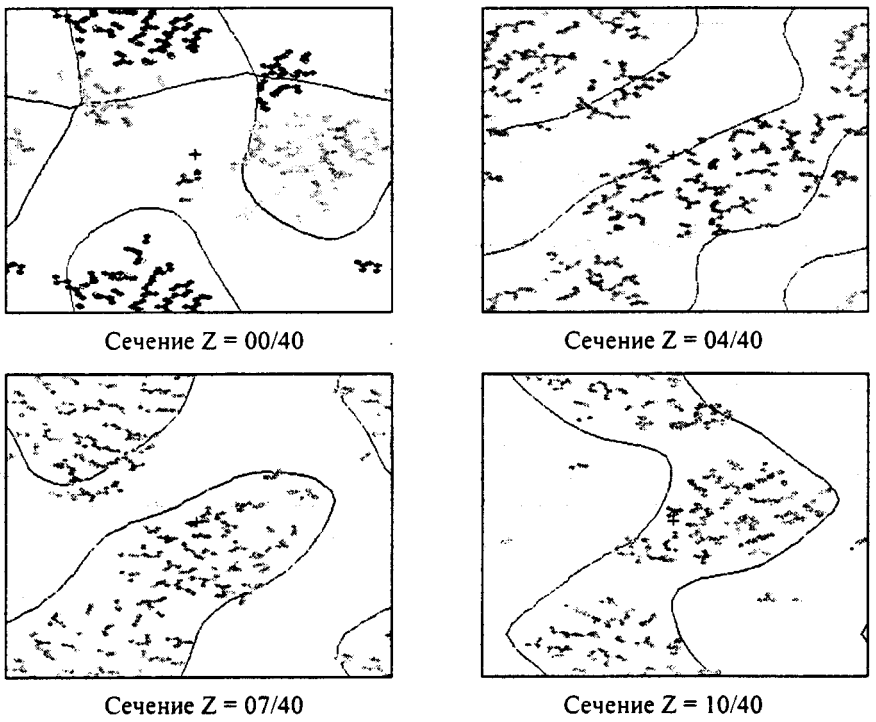


Рис. 7. Границы области молекулы, определенные с помощью процедуры отбора и усреднения вариантов с максимальными значениями правдоподобия (черный контур), и действительные положения атомов для γ-кристаллина IIIb.

Одновременное использование нескольких критериев отбора.

В практической работе более эффективно использовать не один, а несколько независимых критериев отбора вариантов одновременно. Тесты показали, что использование критерия обобщенного правдоподобия в комбинации с критерием связности [11], основанном на анализе топологических свойств синтезов низкого разрешения, позволяет получать

решения лучшего качества. Из таблицы видно, что усреднение решений, полученных независимо с помощью критерия правдоподобия и критерия связности, дает решение лучшего
качества, чем каждое из этих решений в отдельности.

	γ-кристаллин IIIb d=24Å		*Фактор элонгации G d=29Å*	
	C_φ по всем рефл.	*C_φ искл. 4 рефл. с фикс. фазами*	*C_φ по всем рефл.*	*C_φ искл. 4 рефл. с фикс. фазами*
Решение, полученное с помощью критерия правдоподобия	0.82	0.64	0.39	0.02
Решение, полученное с помощью критерия связности	0.82	0.64	0.51	0.32
Усредненное решение	0.92	0.84	0.52	0.41

Таблица 4. Усреднение решений, полученных с помощью критерия правдоподобия и критерия связности.

Расширение набора фаз

В диссертации исследуется также возможность использования критерия обобщенного правдоподобия при расширении фазового набора. В этом случае определенные ранее фазы структурных факторов самого низкого разрешения фиксируются, а перебираются значения фаз структурных факторов более высокого разрешения. В тестах наблюдалась такая же тенденция, как в и случае случайной генерации фазовых наборов: отбор вариантов с высокими значениями правдоподобия давал в результате множество фазовых наборов с более высокой концентрацией хороших вариантов. Однако этот эффект наблюдался в узком диапазоне разрешения, только для зон разрешения, содержащих не более 30-ти рефлексов.

Для комплекса аминоацил-тРНК-синтетазы с тРНК удалось в ходе двухступенчатой процедуры расширения фазового набора выделить лучший вариант на разрешении 68 Å-inf и расширить его до разрешения 48 Å-inf. В результате был получен фазовый набор для 23-х структурных факторов разрешения 68 Å-inf., имеющий фазовую корреляцию C_φ = 0.98 с точными фазами. Синтез рассчитанный с этими фазами и экспериментальными модулями хорошо очерчивает общие контуры комплекса в элементарной ячейке (рис.8).

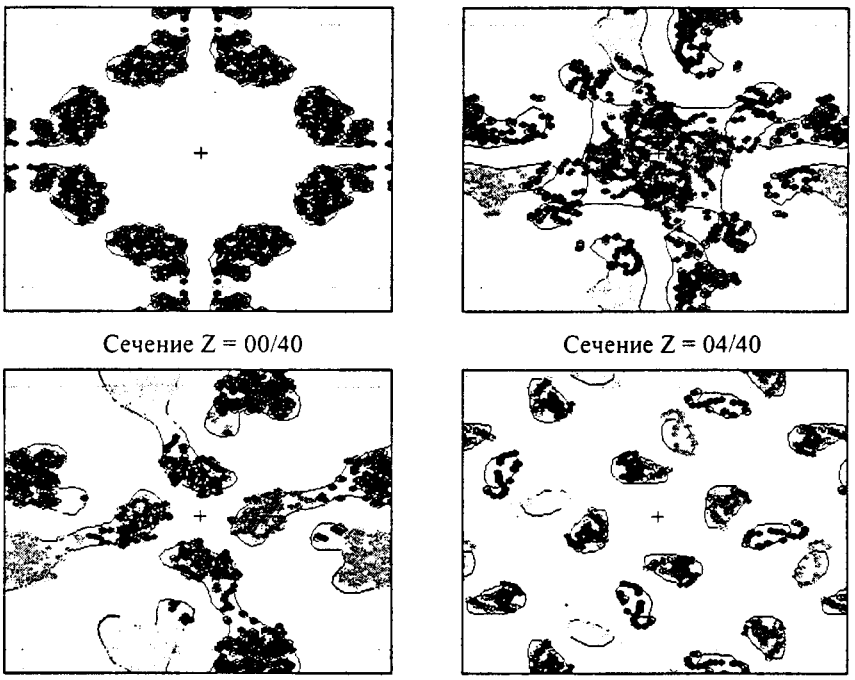


Рис. 8. Границы области молекулы, определенные в ходе двухступенчатой процедуры расширения фазового набора (черный контур), и действительные положения атомов для комплекса аминоацил-тРНК-синтетазы с тРНК.

ГЛАВА V. ПОЛУЧЕНИЕ АПРИОРНОГО РАСПРЕДЕЛЕНИЯ КООРДИНАТ АТОМОВ АНАЛИТИЧЕСКИМ ПУТЕМ

В этой главе рассмотрена возможность получения аналитически априорного распределения, удовлетворяющего принципу максимального правдоподобия. Полученное априорное распределение может быть непосредственно использовано для определения предполагаемых областей высокой и низкой концентрации атомов, то есть областей молекулы и растворителя.

Учитывая сложность нахождения априорного распределения, соответствующего глобальному максимуму функции правдоподобия, рассматривалась более частная задача: найти такое распределение $q(r)$, которому бы соответствовало значение правдоподобия, превышающее значение правдоподобия для равномерного распределения. Получить такое распределение можно, двигаясь из точки, соответствующей равномерному распределению в пространстве априорных распределений, вдоль градиента

функции правдоподобия. Таким образом, искомое распределение может быть представлено в виде:

$$q_\lambda(\mathbf{r}) = q_0(\mathbf{r}) + \lambda \frac{\delta L}{\delta q(\mathbf{r})} , \qquad (9)$$

где значение λ достаточно мало. Задача определения функции $q(r)$ эквивалентна задаче определения его коэффициентов Фурье. Градиент функции правдоподобия может быть выражен через производные функции правдоподобия по коэффициентам Фурье априорного распределения координат атомов:

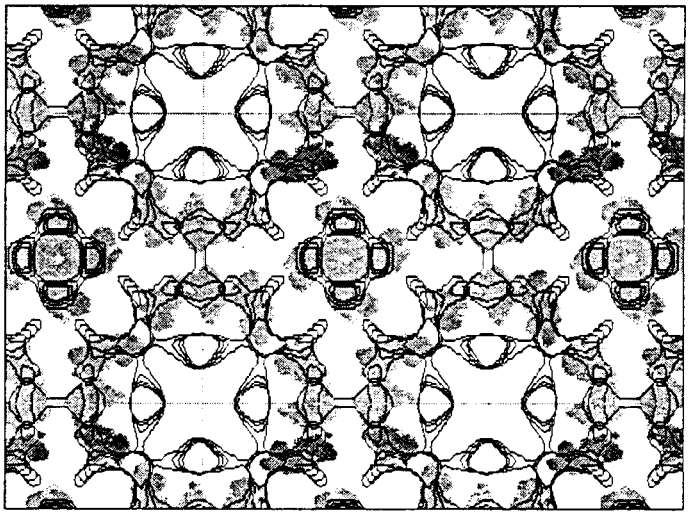
$$\frac{\delta L}{\delta q(\mathbf{r})} = \sum_{\mathbf{h}} \frac{\partial L}{\partial G_{\mathbf{h}}} \exp(2\pi i(\mathbf{h},\mathbf{r})) . \qquad (10)$$

Максимумы и минимумы функций $q_\lambda(\mathbf{r})$ и $\delta L/\delta q(\mathbf{r})$ достигаются одновременно. Следовательно, для того, чтобы выявить в элементарной ячейке предполагаемые области максимальной и минимальной концентрации атомов, достаточно построить карту градиента функции правдоподобия, рассчитав ее по формуле (10).

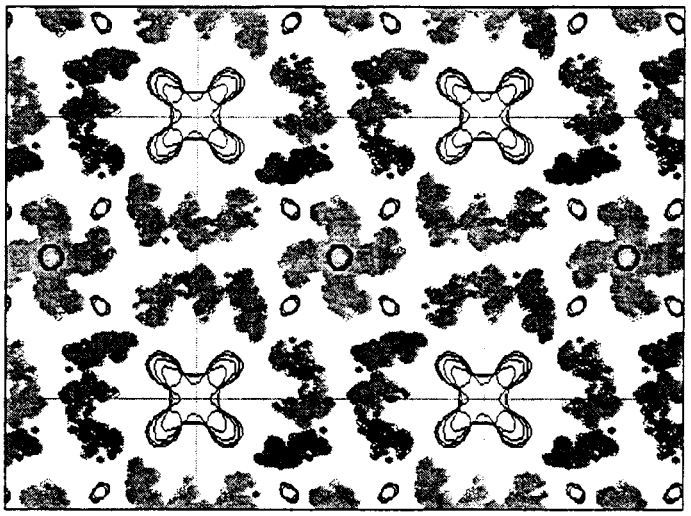
Первая серьезная трудность при получении аналитического выражения для функции правдоподобия - это получение приближенных формул для совместного распределения структурных факторов. Для больших отклонений структурных факторов от ожидаемых значений $\Delta F \sim N$ приближенное выражение было получено Бриконем [4] с помощью метода перевала. Однако полученные им формулы выражаются через неявные функции, являющиеся решением системы нелинейных уравнений. Последнее обстоятельство существенно затрудняет их применение. Для случая, когда отклонения структурных факторов от ожидаемых значений $\Delta F \sim N^{1-\varepsilon}$, в диссертации были получены приближенные формулы для совместного распределения структурных факторов, не содержащие неявных компонент. Используя эти формулы, были получены выражения для производных функции правдоподобия по коэффициентам Фурье априорного распределения $G(h)$ в случае, когда фаза рефлекса h является абсолютным семиинвариантом. Для построения карт градиента функции правдоподобия были использованы данные нейтронного рассеяния для комплекса комплекса аминоацил-тРНК-синтетазы с тРНК. Для работы были отобраны 24 центросимметричных рефлекса, являющиеся семиинвариантами. На рис.9 приведены рассчитанные карты градиента функции правдоподобия и действительные положения C_α- атомов модели. Из рис.9 видно, что рассчитанная функция имеет максимумы в областях, где находятся атомы, а в областях, где атомов почти нет – минимумы.

Использование аналитического подхода к максимизации функции правдоподобия является альтернативой компьютерному Монте-Карловскому моделированию и позволяет достичь некоторого прогресса. Однако сложность математического аппарата делает цену этого продвижения достаточно высокой. В связи с этим основные усилия были сконцентрированы на развитии

компьютерных процедур, реализующих расчет обобщенного правдоподобия и его максимизацию, поскольку на сегодняшний день именно этот путь представляется наиболее эффективным.



а)



б)

Рис. 9. Карты градиента функции правдоподобия. Черный контур соответствует линиям уровня, выделяющим: 25% точек элементарной ячейки с максимальными значениями градиента правдоподобия (а); 5% точек

элементарной ячейки с минимальными значениями градиента правдоподобия (б). Серым цветом обозначены атомы модели.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

1. Предложен подход к выбору области, занимаемой макромолекулой в элементарной ячейке кристалла из класса альтернативных областей, основанный на статистическом моделировании, принципе максимального правдоподобия и компьютерной процедуре Монте-Карловского типа, позволяющей оценивать величину правдоподобия. Применимость подхода продемонстрирована в тестовых расчетах на белках с известной атомной структурой.

2. Предложен метод трансформирования фазовой проблемы рентгеноструктурного анализа в проблему выбора области, занимаемой макромолекулой, из класса альтернативных областей.

3. Предложен подход к *ab-initio* к решению фазовой проблемы при низком разрешении, основанный на фильтрации случайно сгенерированных фазовых наборов при помощи критерия отбора, являющегося аналогом статистического правдоподобия, и усреднении отобранных вариантов. С помощью предложенного подхода для тестируемых структур получены значения фаз, которые можно использовать в качестве первого приближения к решению фазовой проблемы.

4. Получены явные асимптотические формулы для совместного распределения вероятностей набора структурных факторов, пригодные для использования при работе с сильными рефлексами. Найденные формулы применены для расчета градиента функции правдоподобия и приближения априорного распределения градиентом функции правдоподобия.

ЦИТИРУЕМАЯ ЛИТЕРАТУРА

1. Чиргадзе Ю.Н., Невская Н.А., Фоменкова Н.П., Никонов С.В., Сергеев Ю.В., Бражников Е.В., Гарбер М.Б., Лунин В.Ю., Уржумцев Ф.Г., Вернослова Е.А. (1986) Доклады АН СССР, т.290, в.2, 492-495.
2. AEvarsson, A., Braznihnikov, E., Garber, M., Zhelnotsova, J.,Chirgadze, Yu., al-Karadaghi, S., Svensson, L.A. and Liljas, A. (1994), *Embo Journal*, 13, 3669-3677.
3. Andersson, K.M. (1999). *J. Appl. Cryst.* , **32**, 530-535.
4. Bricogne G. (1984). Acta Cryst. **A40**, 410-445.
5. Derrick, J.P., Wingley, D.B. (1994). *J.Mol.Biol.*, **243**, 906.
6. Kraut, J. (1958). Biochim. Bioph. Acta, **30**, 265-270.
7. Lunin, V.Y., Lunina, N.L. (1996). Acta Cryst., A**52**, 365-368.
8. Lunin, V.Y. & Woolfson, M.M. (1993). Acta Cryst., D**49**, 530-533.
9. Moras D., Lorber B., Romby P., Ebel J.-P., Giegé R., Lewitt-Bentley A., Roth M. (1983). J.Biomol. Structure & Dynamics **1**, 209-223.
10. Sevcik, J., Dodson, E. & Dodson, G.G. (1991). Acta Cryst., B**47**, 240-253.
11. Urzhumtsev, A.G., Lunin , V.Y. & Lunina, N.L. (1998). *ECM-18 Abstracts, XVIIIth European Cryst.Meeting, 16-20 August 1998, Prague, Republic Czeck, E5-P7, Bulletin of the Czeck and Slovak Crystallographic Association,* 5B, 482-483.
12. Urzhumtsev, A.G., Vernoslova, E.A. & Podjarny, A.D. (1996). Acta Cryst. **D52**, 1092-1097.
13. Volkmann, N., Hottenträger, S., Hansen, H.A.S., Zaytsev-Bashan, A., Sharon, R., Yonath, A. & Wittmann, H.G. (1990). *J. Mol. Biol.* **216**, 239-241.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Lunin, V.Yu., Lunina, N.L., **Petrova, T.E.**, Vernoslova, E.A., Urzhumtsev, A.G., Podjarny, A.D. (1994) "On the ab-initio solution of the phase problem for macromolecules at very low resolution". *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography,* **30**, 37-44
2. Urzhumtsev, A.G., Lunin, V.Yu., Vernoslova, E.A., Lunina, N.L., **Petrova, T.E.**, Navaza, J., Podjarny, A.D. (1994) "Very low resolution phasing with simple models". *15th Eur. Cryst. Meeting, Dresden, Germany, 28.8-2.9/1994, Coll.Abstr., R.Oldenbourg Verlag, Munchen,* 217
3. Lunin, V.Yu., Lunina, N.L., **Petrova, T.E.**, Vernoslova, E.A., Urzhumtsev, A.G., Podjarny, A.D. (1994) "A Monte Carlo Approach to low resolution phasing in protein crystallography". *ACA Annual Meeting, June 25 - July 1, 1994, INFORUM, Atlanta, Georgia, Atlanta Convention Center,* 178
4. Lunin, V.Yu., Lunina, N.L., **Petrova, T.E.**, Vernoslova, E.A., Urzhumtsev, A.G., Podjarny, A.D. (1995) "On the ab-initio Solution of the Phase Problem for Macromolecules at Very Low Resolution: the Few Atoms Model Method". *Acta Cryst.,* **D51**, 896-903.

5. Lunin, V.Yu., Urzhumtsev, A.G., Vernoslova, E.A., Lunina, N.L., **Petrova,** T.E., Podjarny, A.D. (1995) "La methode FAM pour le phasage ab-initio : Theorie.". *AFC95, 24-27 Jan. 1995*, Grenoble, P07a.
6. Lunin, V.Yu., Lunina, N.L., **Petrova, T.E.,** Vernoslova, E.A., Urzhumtsev, A.G., Podjarny, A.D. (1995) "On the ab-initio solution of the phase problem for macromolecules at very low resolution: the Few Atoms Models method". 4th European Workshop on Crystallography of Biological Macromolecules, Como (Italy), May 21-25, O19.
7. Lunin V.Yu., Lunina N.L., **Petrova T.E.,** Urzhumtsev A.G. & Podjarny A.D. (1997) "A generalized likelihood and the phase cluster choice problem". In "Direct methods for solving macromolecular structures", Lecture Notes, Erice, Italy, May 22 - June 02 1997, 485.
8. 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 the Cluster Discrimination.". *Acta Cryst.,* **D53,** 726-734
9. Lunin, V.Y., Lunina, N.L., **Petrova, T.E.,** Urzhumtsev, A.G., Podjarny, A.D. (1998) "Very low resolution ab-initio phasing. Problems and advances". *ECM-18 Abstracts, XVIIIth European Cryst.Meeting, 16-20 August 1998, Prague, Republic Czeck, E5-P7, Bulletin of the Czeck and Slovak Crystallographic Association,* 5A, 131-132
10. **Petrova, T.E.,** Lunin, V.Y., Podjarny, A.D. (1998) "Likelihood based search of the centre of a macromolecular object". *ECM-18 Abstracts, XVIIIth European Cryst.Meeting, 16-20 August 1998, Prague, Republic Czeck, E5-P4, Bulletin of the Czeck and Slovak Crystallographic Association,* 5B, 481-482.
11. Podjarny, A.D., Urzhumtsev, A.G., Vernoslova, E.A., **Petrova, T.,** Lunina, N., Lunin, V. (1998) "Recent advances on the application of the Few Atom Model method to the T50S ribosomal particle". *AFC98, 24-27 Feb. 1998*, Orlean, BSG-A02.
12. Podjarny, A.D., Lunina, N.L., Urzhumtsev, A.G., Vernoslova, E.A., **Petrova, T.E.,** Lunin, V.Y. (1998) "Ab-initio phasing at 30 A of the t50S ribosomal particle with the Few Atoms Model method". *ACA Annual Meeting, in press*
13. **Петрова, Т.Е.,** Лунин, В.Ю., Лунина, Н.Л., Сковорода, Т.П. (1999). "Выбор априорного распределения координат атомов для макромолекулярных структур на базе принципа максимального правдоподобия". Биофизика, том 44, вып.1, с.22-26.
14. **Petrova, T.E.,** Lunin, V.Yu., Podjarny, A.D. (1999) " A likelihood-based search for the macromolecular position in the crystalline unit cell". *Acta Cryst.,* **D53,** 726-734.
15. **Petrova, T.E.,** Lunin, V.Yu., Podjarny, A.D. (1999) "Ab-initio calculation of envelopes for macromolecules by maximisation of likelihood". Collected *Abstracts, XVIIIth IUCr Congress & General Assembly, 4-13 August 1999, Glasgow, Scotland,* M12.BB.003, 183.