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

ПРЕПРИНТ

В.Ю.ЛУНИН

РЕНТГЕНОСТРУКТУРНЫЙ АНАЛИЗ БЕЛКОВ. ПРОБЛЕМЫ РАСШИФРОВКИ СТРУКТУРЫ

ДОКЛАД НА III ВСЕСОЮЗНОМ СОВЕЩАНИИ «МАТЕМАТИЧЕСКИЕ МЕТОДЫ ДЛЯ ИССЛЕДОВАНИЯ ПОЛИМЕРОВ»

(Пущино, 21—23 июня 1983 г.)

ПУЩИНО·1983

УДК 548.73

Рассмотрены основные математические и вычислительные проблемы, возникающие в процессе определения пространственной структуры белка методом рентгеноструктурного анализа.

Работа будет доложена на III Всесоюзном совещании "Математические методы для исследования полимеров".

ВВЕДЕНИЕ

В этой работе делается попытка взглянуть на современный рентгеноструктурный анализ белков "извне"- с точки зрения математика.В ней почти не затрагиваются биохимические и физические аспекты проблемы,и упор сделан на математическую интерпретацию методов,применяемых при установлении пространственной структуры белка.Такой подход,разумеется,несколько формализует изложение,лишая его деталей конкретной экспериментальной работы. Однако,с другой стороны,он позволяет освободиться от некоторой традиционности и взглянуть на проблему под другим углом зрения. Более того,автор полагает,что базирующиеся в основном на "здравой мысли",а не на математике методы,сделавшие рентгеноструктурный анализ рабочим инструментом для определения пространственной структуры белка,близки сегодня к своему пределу,и дальнейшее развитие методики будет во многом идти за счет средств вычислительной математики и компьютерной техники.Это заставляет более пристально смотреть на математическую основу метода.

На сегодняшний день рентгеноструктурный анализ является единственным методом,позволяющим определить пространственную структуру сложных биологических макромолекул (белков,РНК,вирусов).В первых двух параграфах сообщено в каких терминах описывается пространственная структура исследуемого объекта в рамках метода рентгеноструктурного анализа,что определяет круг вопросов,на которые может дать ответ этот метод.Там же описана основная схема определения структуры,применяющаяся сегодня для исследования макромолекул.В третьем параграфе дано краткое описание одного из современных направлений развития рентгеноструктурного анализа - методов,позволяющих повысить разрешение,с которым мы "видим" структуру,дать возможность определить большее число её деталей.

Отметим еще,что говоря далее для краткости о рентгенострук-

турном анализе белков,мы подразумеваем везде работу с макромолекулами не обязательно белковой природы (РНК,вирусами и т.п.).

§ I. Модели,используемые в рентгеноструктурном анализе.

I.I Модель исследуемого объекта.

Поскольку экспериментальной основой обсуждаемой методики является рассеяние рентгеновских лучей электронами атомов исследуемого объекта,то моделью объекта в рамках задач рентгеноструктурного анализа является модель распределения электронной плотности в нем.При этом,так как продолжительность рентгеновского эксперимента велика (может достигать десятков часов),то ясно, что речь может идти лишь об распределении,усредненном по времени эксперимента.В рамках метода рентгеноструктурного анализа считается,что электронная плотность объекта есть простое объединение электронных плотностей отдельных атомов

$$\rho(\vec{r}) = \sum_j \rho_j(\vec{r}) \qquad (I)$$

(здесь $\rho(\vec{r})$ - распределение электронной плотности в исследуемом объекте, $\rho_j(\vec{r})$ - распределение электронной плотности в отдельном атоме, $\vec{r}$ - радиус-вектор в точку наблюдения),то есть,что деформация "облаков электронной плотности" при объединении атомов в молекулу незначительна *.Далее считается,что распределение электронной плотности покоящегося атома,находящегося в начале координат,описывается сферически симметричной функцией $\rho_j^0(r)$ ($r=|\vec{r}|$), которая обычно предполагается известной.Тепловое движение атомов моделируется "размазыванием" распределения $\rho_j^0(r)$ путем свертки его с гауссовой функцией $(4\pi/B_j)^{3/2}\exp\{-4\pi r^2/B_j\}$ **.Здесь B_j - инди-

---

* Существуют работы,в которых делались попытки изучения методами рентгеноструктурного анализа деформации распределения электронной плотности отдельного атома при образовании химической связи.Однако,эти работы относились только к низкомолекулярным соединениям и требовали точности проведения исследования, недостижимой на сегодняшний день для белков.

** Такой путь учета тепловых колебаний содержит в себе предположение об их изотропности.Для учета более тонких эффектов иногда вместо одного параметра B_j вводится тензор,отражающий анизотропию тепловых колебаний.Однако это приобретает смысл при точности проведения рентгеноструктурного эксперимента,редко достижимой для белков.

видуальный для каждого атома параметр, описывающий амплитуду тепловых колебаний центра атома *. Следует отметить, что к "размазыванию" плотности отдельного атома приводит и влияние некоторых экспериментальных погрешностей, так что получаемые в процессе расшифровки структуры значения B_j могут не иметь строгого физического смысла, а лишь отражают относительную подвижность отдельных атомов молекулы.

Таким образом, в рамках сделанных допущений распределение электронной плотности в отдельном атоме имеет вид :

$$\rho_j(\vec{r}) = \int_{R^3} \rho_j^n(|\vec{u}-\vec{r}_j|)\left(\frac{4\pi}{B_j}\right)^{3/2} \exp\left\{-\frac{4\pi|\vec{r}-\vec{u}|^2}{B_j}\right\} dV_{\vec{u}} \quad , \qquad (2)$$

то есть характеризуется четырьмя параметрами : координатами "центра атома" $\vec{r}_j = (r_{xj}, r_{yj}, r_{zj})$ и параметром тепловых колебаний B_j. Поскольку предполагается, что объединение атомов в молекулу не меняет их распределение плотности, то тем самым исследуемый объект - распределение электронной плотности $\rho(\vec{r})$ в молекуле - задается набором координат центров атомов $\vec{r}_j$ и параметрами B_j ($j = 1, \ldots, N$) тепловых колебаний атомов. Задача, которую решает рентгеноструктурный анализ, заключается в определении всех этих величин. Принятая модель исходного объекта определяет и круг вопросов, на которые может дать ответ его рентгеноструктурное исследование. Это в основном стереохимические вопросы - вопросы, которые могут быть сформулированы в терминах координат атомов, длин связей, валентных углов и т.п.

Здесь сразу следует сделать несколько замечаний. Во-первых, определение параметров $\vec{r}_j, B_j$ действительно <u>всех</u> атомов молекулы осуществляется, как правило, лишь для низкомолекулярных соединений. Для белков под определением параметров всех атомов обычно понимают определение их для всех неводородных атомов. Атомы водорода ввиду малости их заряда при существующей точности метода рентгеноструктурного анализа для белков не различимы на фоне "экспериментального шума". Существуют лишь единичные работы, где для белков была локализована часть водородных атомов.

Далее считается, что нам известны распределения плотности для атомов всех типов, присутствующих в молекуле. Более строго говоря, считается, что нам известны факторы атомного рассеяния $f_j^n(s)$ различных атомов, связанные взаимно-однозначно с распределениями $\rho_j^n(r)$ синус-преобразованием Фурье :

---

* Атомы в молекуле обладают различной подвижностью, что приводит к различиям в значениях параметров B_j для разных атомов.

$$f_j^n(s) = \frac{2}{s}\int_0^\infty r\,\rho_j^0(r)\sin 2\pi sr\,dr \,. \qquad (3)$$

Учет тепловых колебаний приводит к домножению функции $f_j^n(s)$ на $\exp\{-Bs^2/4\}$, то есть фактор атомного рассеяния $f_j(s)$ для атома, совершающего тепловые колебания, есть

$$f_j(s) = f_j^n(s)\,e^{-B\frac{s^2}{4}} \,. \qquad (4)$$

Значения $f_j^n(s)$ для различных типов атомов рассчитаны методами квантовой механики и табулированы.

Наконец, следует подчеркнуть, что определение пространственной структуры (в терминах параметров $\vec{r}_j$ и B_j) возможно на сегодняшний день лишь для кристаллических структур. Это с одной стороны ставит перед исследователем проблему кристаллизации изучаемого объекта, а с другой стороны открывает простор для дискуссий на тему: "Насколько структура кристаллизованного белка идентична структуре белка, находящегося в биологически активном состоянии?" Принятая сейчас точка зрения - небольшие конформационные различия между белком в кристалле и белком в растворе нельзя исключить, но резкое изменение конформации молекулы при кристаллизации маловероятно. Это объясняется тем, что силы, связывающие молекулы в кристалле, значительно слабее сил, определяющих структуру молекулы, и подтверждается тем, что в некоторых случаях одни и те же белки, кристаллизованные в различных условиях, обнаружили очень близкую кристаллическую структуру.

Также необходимо отметить, что метод рентгеноструктурного анализа статичен. Рентгеновский эксперимент требует десятков часов, и получаемый результат вынужденно усреднен по этому времени. Поэтому анализ "быстрых" перестроек пространственной структуры в рамках современного рентгеноструктурного анализа невозможен. Здесь требуются иные методы.

I.2 Модель рентгеновского эксперимента.

В процессе рентгеновского эксперимента образец исследуемого вещества помещается в пучок рентгеновских лучей, и регистрируется интенсивность излучения, рассеянного в различных направлениях. Для обработки результатов обычно используется следующая модель процессов, протекающих при эксперименте. Считается, что первичный пучок рентгеновских лучей есть плоская монохроматическая волна (с длиной волны λ), распространяющаяся в направлении, задаваемом единичным вектором $\vec{e}_0$. Под воздействием этой волны каждый электрон приходит в движение и становится источником вторичной сфери-

ческой волны. Сложение этих вторичных волн и определяет интенсивность рассеяния в направлении, определяемом единичным вектором $\vec{\sigma}$ *:

$$I(\lambda) = \frac{A}{R^2} \left| \int \rho(\vec{r}) e^{2\pi i (\vec{\lambda}, \vec{r})} dV_{\vec{r}} \right|^2 . \qquad (5)$$

Здесь $\lambda = \frac{\vec{\sigma} - \vec{\sigma}_0}{\lambda}$, R – расстояние от образца до точки наблюдения, A – коэффициент пропорциональности. Варьируя направление первичного пучка $\vec{\sigma}_0$ и направление $\vec{\sigma}$, в котором измеряется интенсивность рассеяния, мы в состоянии определить, таким образом, квадрат модуля преобразования Фурье функции $\rho(\vec{r})$ для "частот", удовлетворяющих ограничению

$$\lambda = |\vec{\lambda}| \leq \frac{2}{\lambda} . \qquad (6)$$

Для того, чтобы получить регистрируемые экспериментально значения интенсивностей $I(\lambda)$, требуется участие в рассеянии первичного пучка большого числа молекул исследуемого вещества. Понятно, что если эти молекулы имеют произвольную ориентацию (газы, растворы), то регистрируемая интенсивность может нести информацию лишь о каких-то средних по всевозможным ориентациям характеристиках объекта. Такие характеристики изучаются методами малоуглового рентгеновского рассеяния и о них далее речь идти не будет. Другая возможность – определения положения всех атомов структуры – открывается, когда объект кристаллизован, то есть все молекулы находятся в одинаковой ориентации.

Для кристалла функция распределения электронной плотности $\rho(\vec{r})$ является периодической функцией с некоторыми периодами $\vec{a}, \vec{b}, \vec{c}$, то есть она может быть представлена в виде ряда Фурье

$$\rho(\vec{r}) = \frac{1}{|V|} \sum_{\vec{\lambda} \in Q'} F(\vec{\lambda}) e^{i\varphi(\vec{\lambda})} e^{-2\pi i (\vec{\lambda}, \vec{r})} , \qquad (7)$$

где коэффициенты Фурье (называемые в кристаллографии структурными факторами) есть

---

* Существуют более тонкие модели, учитывающие взаимодействие рассеянной и падающей волн (динамическая теория рассеяния). Однако для "реальных" кристаллов лучшее соответствие эксперименту дает более грубая модель (кинематическая теория рассеяния), описанная выше. Причина в том, что "реальный" кристалл состоит из мелких блоков, слегка дезориентированных друг относительно друга, что и уничтожает тонкие эффекты. Преимущества же динамической теории рассеяния сказываются лишь при изучении "почти идеальных" кристаллов типа алмаза, либо искуственно полученных крупных совершенно моноблочных кристаллов.

$$F(\vec{\lambda})e^{i\varphi(\vec{\lambda})} = \int_V \rho(\vec{r})\, e^{2\pi i(\vec{\lambda},\vec{r})}\, dV_r \,, \qquad (8)$$

V – параллелепипед, построенный на векторах $\vec{a},\vec{b},\vec{c}$, $|V|$ – его объем, суммирование в (7) идет по множеству $\mathcal{R}'$ векторов $\vec{\lambda}$ таких, что проекции вектора $h=(\vec{\lambda},\vec{a})$, $k=(\vec{\lambda},\vec{b})$, $l=(\vec{\lambda},\vec{c})$ есть произвольные целые числа *. (В кристаллографии множество $\mathcal{R}'$ носит название решетки обратного пространства).

Таким образом, рентгеновский эксперимент дает теоретическую возможность измерения модулей некоторого ограниченного набора коэффициентов Фурье функции распределения электронной плотности в исследуемом объекте. Следует отметить, что принципиальная граница набора величин $F(\vec{\lambda})$ (6), связанная с длиной волны используемого излучения, редко достигается на практике. Как правило, удается экспериментально определить лишь набор величин $F(\vec{\lambda})$ с $|\vec{\lambda}| \le \lambda_{max} < \frac{2}{\lambda}$, где величина λ_{max} определяется степенью совершенства кристалла и для белков может быть в несколько раз меньше теоретической границы $\frac{2}{\lambda}$.

I.3 Система уравнений для параметров структуры.

Итак, в рамках описанных моделей, рентгеновский эксперимент дает нам возможность получить значения модулей коэффициентов Фурье $F_э(\vec{\lambda})$ функции распределения электронной плотности $\rho(\vec{r})$ в кристалле исследуемого вещества. С другой стороны, мы можем выразить эти величины, используя формулы (I)-(4), через параметры $\vec{r}_j$, B_j модели электронной плотности $\rho(\vec{r})$. Это приводит к системе нелинейных уравнений для определения величин $\vec{r}_j, B_j$, $j=1,\dots,N$:

$$\left| \sum_{j=1}^{N} f_j^n(\lambda)\, e^{-B\frac{\lambda^2}{4}}\, e^{2\pi i(\vec{\lambda},\vec{r}_j)} \right| = F_э(\vec{\lambda}) \,, \quad |\vec{\lambda}| \le \lambda_{max} \,. \qquad (9)$$

В этой системе уравнений величины $f_j^n(\lambda)$ и $F_э(\vec{\lambda})$ предполагаются нам известными, а $\vec{r}_j, B_j$ – неизвестные, подлежащие определению.

---

* Строго говоря, в силу конечных размеров реального кристалла, функция $\rho(\vec{r})$ не является периодической, а может быть представлена в виде $\tilde{\rho}(\vec{r})\chi(\vec{r})$, где $\tilde{\rho}(\vec{r})$ – периодическая во всем пространстве функция, а $\chi(\vec{r})$ – функция, равная единице внутри пространства, занятого образцом, и равная нулю вне его. Это приводит к тому, что коэффициент пропорциональности A в (5) теряет δ-образный характер – дифракционные максимумы (локализованные в точках $\vec{\lambda}\in\mathcal{R}'$) "расплываются". Этот эффект учитывается при первичной обработке данных, и далее работа идет с модулями коэффициентов Фурье идеально-периодической функции $\tilde{\rho}(\vec{r})$.

Число уравнений в системе (9), как правило, превышает число неизвестных (для белка среднего размера при среднем разрешении число неизвестных около 6000, число уравнений около 15000), поэтому решение системы (9) понимается в смысле "метода наименьших квадратов", как такой набор величин $\vec{r}_j, B_j$, при котором минимальна невязка

$$R_2 = \sum_{|\vec{h}| \le h_{max}} \left(|\sum_j f_j^n(h) e^{-B\frac{h^2}{4}} e^{2\pi i(\vec{h},\vec{r}_j)}| - F_э(\vec{h}) \right)^2 . \qquad (10)$$

Понятно, что громадное число уравнений и неизвестных и сложный осциллирующий характер уравнений (9) не оставляют надежд на "лобовое" решение системы, путем поиска глобального минимума функции (10)*. Единственное на что мы можем надеяться – это нахождение локального минимума функции (10). Поэтому в кристаллографии белка система уравнений (9) используется только на заключительных этапах работы. Когда у нас имеется хорошее приближение $\vec{r}_j^{\,0}, B_j^0$ для искомых величин, мы можем пытаться определить величины $\vec{r}_j, B_j$ более точно путем локальной минимизации функционала (10) в окрестности приближенных значений параметров $\vec{r}_j^{\,0}, B_j^0$. Такая процедура носит название уточнения структуры белка**.

Следует заметить, что точность, с которой на заключительном этапе мы можем определить значения координат $\vec{r}_j$ атомов, зависит только от точности экспериментального определения величин $F_э(\vec{h})$ и точности моделей, опираясь на которые были получены уравнения (9). Увеличение числа уравнений несколько повышает потенциальную точность их решения, так как ошибка эксперимента усредняется в статистическом смысле. Однако этот эффект не принципиален и в известной мере может ослабляться за счет роста относительной ошибки определения величин $F_э(\vec{h})$ с ростом $|\vec{h}|$. При реальной точности экспериментальных данных (порядка 5 – 10%) удается достичь точности определения координат порядка 0.01 Å.

При общем рассмотрении системы (9) естественно возникает вопрос о том, единственно ли её решение. (Вопрос о существовании решения при реальных $F_э(\vec{h})$ мы аннулировали сведением задачи к минимизации функции (10)). Очевидно, что сдвиг всей функции на произвольный вектор $\vec{u}$ (то есть переход к функции $\rho(\vec{r}-\vec{u})$)

---

* Для расшифровки структур низкомолекулярных соединений использовался поиск глобального минимума функции (10) методом оврагов. Однако область применения такого подхода ограничивается структурами с очень небольшим числом независимых параметров.

** На этом этапе в уточнение часто вводят дополнительные стереохимические ограничения на взаимное расположение атомов.

не меняет значения модулей структурных факторов $F(\vec{\chi})$,то есть не меняет и значение критерия (I0).Аналогично,не меняет их и переход к функции $\rho(-\vec{r})$* .Это тривиальные случаи неединственности,и проблема такой неединственности снимается путем фиксации фаз четырех структурных факторов (эти структурные факторы и значения фаз выбираются почти произвольно).Некоторым потрясением является то,что могут существовать существенно различные структуры,удовлетворяющие одним и тем же уравнениям (9).

Уравнения (9) строго эквивалентны уравнениям

$$\sum_{j,k} f_j(\chi) f_k(\chi) e^{2\pi i(\vec{\chi},\vec{r}_j-\vec{r}_k)} = F_\chi^2(\vec{\chi}) , \qquad (II)$$

то есть рентгеновский эксперимент накладывает ограничения только на набор межатомных векторов $\{\vec{r}_j-\vec{r}_k\}_{j,k}$,и если структуры имеют одинаковый набор межатомных векторов (такие структуры носят название гомометричных),то эти структуры имеют идентичную картину рассеяния.

С проблемой гомометрии столкнулись в I930 г. при расшифровке структуры низкомолекулярного соединения биксбиита.В I939 г. Патерсон построил ряд примеров гомометричных структур на окружности.Один из них дан на рисунке I.В I974 г. Франклин предложил

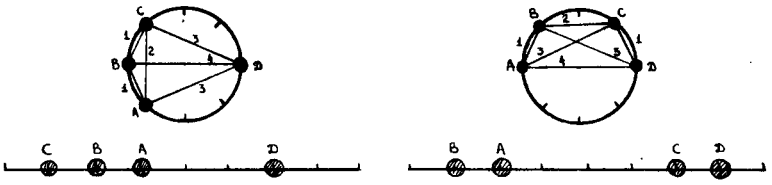

Рис.I Одномерные структуры,дающие одинаковую дифракционную картину (цифрами обозначены относительные межатомные расстояния)

очень простой способ построения любого числа примеров гомометричных структур в пространстве любой размерности : если X и Y два произвольных множества точек,не имеющие центра симметрии,то множества $X+Y$ и $X-Y$ гомометричны (под суммой $X+Y$ понимается

---

* Выбор абсолютной конфигурации из двух возможных "энантиоморфных" вариантов $\rho(\vec{r})$ и $\rho(-\vec{r})$ может быть осуществлен привлечением данных по аномальному рассеянию рентгеновских лучей.Для белков выбор правильной конфигурации может быть также сделан из тех соображений,что аминокислоты должны иметь L-конфигурацию, а α-спирали - "правую закрутку".

множество всевозможных элементов вида $x+y$,где x лежит в X , а y в Y).Простейший пример такого построения дан на рис.2.

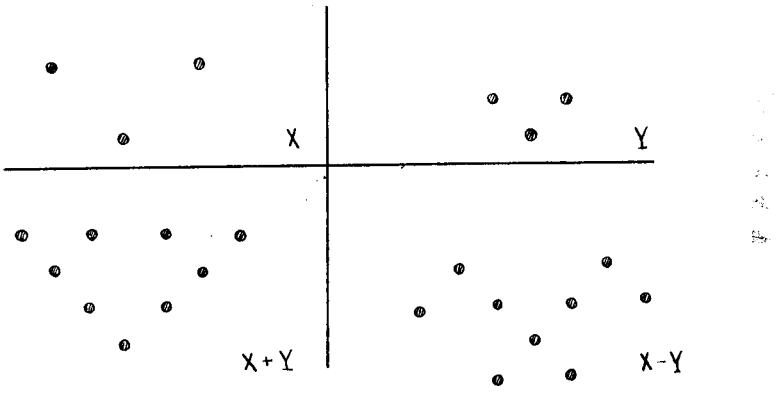

Рис.2 Структуры $X+Y$ и $X-Y$ дают одинаковую картину дифракции .

Результаты Патерсона и Франклина позволяют строить примеры,но решение более интересного вопроса - для данной структуры найти гомометричные ей - автору не известно.

Отметим,что для белков основной метод,применяемый при расшифровке структуры - метод изоморфного замещения - может снимать эту неоднозначность.

§ 2 Построение "изображения" изучаемого объекта.

2.I Поиск изображения распределения электронной плотности исследуемого объекта.

Отказавшись от попытки определить параметры $\vec{r}_j$, B_j исследуемого объекта непосредственно из системы уравнений (9),связывающих искомые величины с результатами эксперимента,мы временно изменим постановку задачи."Забыв" о координатах атомов,мы поставим себе пока целью попытаться найти саму функцию распределения электронной плотности $\rho(\vec{r})$ (точнее какое-нибудь её приближение)*. Смысл такой смены цели простой - не умея определить сразу пара-

---

* Заметим,что при известных $\vec{r}_j$, B_j распределение $\rho(\vec{r})$ легко определяется по формулам (I) и (2).

метры искомого объекта $\vec{r}_j$, B_j, мы пытаемся для начала "увидеть" его. Такое "визуальное" изучение распределения $\rho(\vec{r})$ может принести по крайней мере двоякую пользу. Во-первых, мы можем описать какие-то более грубые характеристики исследуемой молекулы - внешнюю форму, ход полипептидной цепи в белке и т.п.. Во-вторых, при хорошем качестве восстановления функции $\rho(\vec{r})$ мы можем просто увидеть сгустки электронной плотности, соответствующие отдельным атомам (либо группам атомов) и тем самым приближенно определить их координаты, которые далее могут быть уточнены минимизацией функции (10). Таким образом, мы пытаемся разделить решение проблемы, поручив человеческому глазу и опыту исследователя увидеть детали структуры, а ЭВМ - уточнить эти детали, "перемалывая" систему уравнений (9).

Первым препятствием на пути осуществления такой программы является то, что эксперимент дает нам только модули коэффициентов Фурье функции $\rho(\vec{r})$. Поэтому для расчета функции по формуле (7) нам необходимо решить фазовую проблему рентгеноструктурного анализа - задачу определения фаз $\varphi(\vec{s})$ коэффициентов Фурье (8) функции $\rho(\vec{r})$. В следующем пункте будут обозначены основные способы решения этой задачи. Здесь же мы только отметим, что несмотря на все трудности решения фазовой проблемы этот путь является на сегодняшний день единственным, позволяющим подойти к определению пространственной структуры белка.

Второе принципиальное ограничение заключается в том, что мы в состоянии определить лишь ограниченный набор коэффициентов Фурье с $|\vec{s}| \leq s_{max}$. Как известно обрыв ряда (7) приводит к расплыванию пиков и появлению "волн обрыва". Таким образом, определив как-то фазы $\varphi(\vec{s})$, мы получаем суммированием ряда (7) в пределах $|\vec{s}| \leq s_{max}$ не саму функцию $\rho(\vec{r})$, а её "размытое" и зашумленное изображение $\tilde{\rho}(\vec{r})$. Дополнительные искажения вносят в это изображение неточности в измерении величин $F(\vec{s})$ и в расчете фаз $\varphi(\vec{s})$.

В кристаллографии для каждой гармоники $e^{2\pi i(\vec{s},\vec{r})}$ ряда Фурье (7) вводится понятие разрешения $d = \frac{1}{|\vec{s}|}$. Геометрически разрешение d -это расстояние между двумя соседними плоскостями $e^{2\pi i(\vec{s},\vec{r})} = C$ (то есть "длина волны"). Далее говорят, что изображение $\tilde{\rho}(\vec{r})$ построено с разрешением d_0, если в ряд (7) включены слагаемые с $d = \frac{1}{|\vec{s}|} \geq d_0$. Геометрически это означает, что два "точечных" атома (атома, вклады которых в функцию $\rho(\vec{r})$ представляются δ-функциями), находящиеся на расстоянии меньшем d_0, на изображении $\tilde{\rho}(\vec{r})$ представятся в виде одного широкого пика. В кристаллографии белка принято считать "низким" разрешение порядка 6 Å (при таком разрешении изображение $\tilde{\rho}(\vec{r})$ дает представление лишь о форме моле-

кулы),"средним" - разрешение порядка 3 Å (такое разрешение дает возможность проследить ход полипептидной цепи белка) и "высоким" порядка 2 Å и выше.

Следует подчеркнуть,что в публикуемых результатах по рентгеноструктурному анализу белков фраза например о том,что структура получена при разрешении 2.0 Å,означает только,что использовался набор экспериментальных данных с $|\vec{s}| \leq 0.5$.Это с одной стороны еще не означает,что на полученном изображении $\tilde{\rho}(\vec{r})$ распределения электронной плотности были видны детали размером 2.0 Å - качество изображения сильно зависит от качества определения фаз $\varphi(\vec{s})$. С другой стороны,набор данных с $d = \frac{1}{|\vec{s}|} \geq 2.0$ не препятствует получить значения координат атомов с точностью 0.01 Å,путем минимизации функции (10).

2.2 Подходы к решению фазовой проблемы

2.2.1 Основным методом решения фазовой проблемы при расшифровке структуры белка является привлечение дополнительной информации путем проведения эксперимента по рассеянию рентгеновских лучей не только с исходным (нативным) белком,но и с рядом его модификаций (с изоморфными производными).Если мы знаем,в чем заключалась модификация белка,то есть распределение электронной плотности $\rho_{PH}(\vec{r})$ такого модифицированного белка отличается от распределения $\rho(\vec{r})$ известной добавкой $\rho_H(\vec{r})$:

$$\rho_{PH}(\vec{r}) = \rho(\vec{r}) + \rho_H(\vec{r}) , \qquad (12)$$

то для искомой фазы φ имеет место соотношение :

$$|Fe^{i\varphi} + F_H e^{i\varphi_H}| = F_{PH} . \qquad (13)$$

В этом соотношении $F_H e^{i\varphi_H}$ - коэффициент Фурье известной нам функции $\rho_H(\vec{r})$,а модули коэффициентов Фурье F и F_{PH} функций $\rho(\vec{r})$ и $\rho_{PH}(\vec{r})$ могут быть получены из эксперимента.Уравнение (13) имеет, вообще говоря,два различных решения φ_1 и φ_2 ,то есть неопределенность фазы сводится к выбору одного из двух возможных вариантов (для каждого коэффициента Фурье).Проведение рентгеновского эксперимента с несколькими различными модификациями нативного белка позволяет получить систему уравнений типа (13),однозначно определяющую фазу φ .

Как правило модифицированный белок (изоморфное производное) получается путем присоединения химическим путем к молекуле белка тяжелого атома,либо группы атомов.Если это присоединение произошло так,что положение атомов молекулы белка не изменилось и кристаллическая упаковка не нарушилась,то функция $\rho_H(\vec{r})$ есть

просто сумма распределений электронной плотности в присоединенных атомах и может быть рассчитана,если нам удалось каким-то образом определить координаты этих атомов (и параметры их температурных колебаний)*.Определение координат присоединившихся атомов является ключевым местом при использовании этой методики расчета фаз и определяет успех дальнейшей работы.
Заметим,что формально ситуация здесь еще более тяжелая,чем в проблеме определения координат атомов нативного белка.Если для функции $\rho(\vec{r})$ нам были известны модули её коэффициентов Фурье,то для коэффициентов Фурье $F_H e^{i\varphi_H}$ функции $\rho_H(\vec{r})$ нам известно только,что они есть разность двух комплексных чисел $F_{PH} e^{i\varphi_{PH}}$ и $Fe^{i\varphi}$ с известными нам модулями F_{PH} и F (то есть мы даже не знаем модуля F_H).Однако число присоединившихся атомов,как правило,невелико,что позволяет привлекать для определения их координат арсенал методов,накопленных для расшифровки низкомолекулярных соединений (о некоторых из этих методов будет упомянуто ниже).

2.2.2 Основной подход к проблеме определения фаз при расшифровке структур низкомолекулярных соединений заключается в исключении из системы уравнений

$$\begin{cases} \left| \sum_{j=1}^{N} f_j(\lambda)\, e^{2\pi i(\vec{\lambda},\vec{r}_j)} \right| = F_3(\vec{\lambda}) \\ \arg\left\{ \sum_{j=1}^{N} f_j(\lambda)\, e^{2\pi i(\vec{\lambda},\vec{r}_j)} \right\} = \varphi(\vec{\lambda}) \end{cases}, \quad |\vec{\lambda}| \le \lambda_{max} \qquad (14)$$

координат атомов $\vec{r}_j$ и получение тем самым системы уравнений для фаз $\varphi(\vec{\lambda})$**.Спецификой такого исключения является то,что в получаемые уравнения для фаз входят все коэффициенты Фурье (для бесконечного разрешения).Поэтому такие уравнения имеют приемлемую точность,когда набор экспериментальных данных собран для доста-

---

* Такая ситуация является,конечно,идеализацией реальной обстановки.Определенная деформация молекулы белка при присоединении к ней тяжелых атомов всегда имеет место,и возможность использовать соотношение (13) для расчета фаз определяется тем,насколько сильна эта деформация.Отметим,что чувствительность уравнений (13) к деформации молекулы растет с ростом "частоты" $|\vec{\lambda}|$,что приводит в реальной ситуации к тому,что уравнения типа (13) оказываются применимыми лишь при определенном разрешении $d \ge d_{min}$.

** Обычно в кристаллографии такое исключение координат атомов производится в виде вывода вероятностных соотношений для значений определенных комбинаций фаз при переборе всех коэффициентов Фурье функции $\rho(\vec{r})$.

точно высокого разрешения,и их применимость сильно ограничена для белков.Тем не менее в последние годы делаются попытки распространения этих методов и на белковые структуры,правда здесь, как правило,речь идет лишь об уточнении уже имеющегося грубого набора фаз.

2.2.3 Еще один путь расшифровки структуры,обходящий проблему расчета фаз - построение функции Патерсона :

$$P(\vec{r}) = \frac{1}{V} \sum_{\vec{s}} F^2(\vec{s})\, e^{-2\pi i(\vec{s},\vec{r})}. \qquad (15)$$

Как видно из равенства (II),функция Патерсона $P(\vec{r})$ представляет собой распределение электронной плотности для некой гипотетической структуры из N^2 атомов,характеризующихся распределениями электронной плотности $\rho_j * \rho_k$ и координатами центров $\vec{r}_j - \vec{r}_k$ ($j,k = 1,\ldots,N$).Если функция $P(\vec{r})$ построена при достаточно высоком разрешении,и число атомов не слишком велико,то её анализ может позволить определить весь набор векторов $\{\vec{r}_j - \vec{r}_k\}_{j,k}$. Далее по существующим алгоритмам по набору межатомных векторов $\{\vec{r}_j - \vec{r}_k\}_{j,k}$ могут быть восстановлены искомые координаты $\vec{r}_j$. Такой подход требует локализации N^2 пиков функции $P(\vec{r})$,поэтому его возможности сильно ограничены числом N атомов,и для белковых структур его прямое применение нереалистично.Тем не менее,этот метод является в кристаллографии белка основным способом определения координат атомов,присоединившихся к молекуле белка при получении изоморфного производного (поскольку число присоединившихся атомов невелико).

§ 3. Повышение качества изображения $\tilde{\rho}(\vec{r})$ функции распределения электронной плотности.

Практическое определение пространственной структуры белка методом рентгеноструктурного анализа предоставляет исследователю широкий спектр всевозможных трудностей.Достаточно заметить,что на сегодняшний день определение структуры белка требует,как правило,нескольких лет работы.Мы не в состоянии здесь коснуться всех проблем современного рентгеноструктурного анализа белков и остановимся лишь на одном вопросе.

3.I.I Как уже говорилось,основной метод расчета фаз коэффициентов Фурье функции распределения электронной плотности $\rho(\vec{r})$ базируется на предположении,что при получении используемого тяжелоатомного производного присоединение тяжелой группы

атомов не вызвало смещение атомов молекулы белка. При этом ошибка в определении фаз $\varphi(\vec{s})$, вызванная подвижкой атомов белка (и не вполне точным определением координат тяжелых атомов), нарастает с ростом $|\vec{s}|$. Это приводит к тому, что практически определить фазы удается лишь для набора коэффициентов Фурье с $|\vec{s}| \le s_{out}$, где s_{out} меньше s_{max} - предельного значения $|\vec{s}|$ для собранного набора величин $F_э(\vec{s})$, то есть для части коэффициентов Фурье функции $\rho(\vec{r})$ с известными модулями $F_э(\vec{s})$ фазы $\varphi(\vec{s})$ определить так и не удается *. Таким образом, мы имеем возможность построить изображение $\tilde{\rho}(\vec{r})$ функции $\rho(\vec{r})$ лишь с разрешением, отвечающим степени идеальности тяжелоатомных производных. Далее, для коэффициентов Фурье, включенных в расчет $\tilde{\rho}(\vec{r})$, ошибка в фазах нарастает с ростом $|\vec{s}|$, что приводит к зашумлению изображения $\tilde{\rho}(\vec{r})$. Переход к исследованию "тех белков, которые интересно исследовать, а не тех которые удается исследовать" привел к тому, что работа все чаще стала идти с худшими по качеству производными нежели раньше. Это в свою очередь привело к проявлению интереса к проблеме улучшения набора фаз. Под улучшением набора фаз мы понимаем, во-первых, определение фаз для тех коэффициентов Фурье, для которых их ранее не удавалось определить (расширение набора фаз), а, во-вторых, уточнение тех значений фаз, которые ранее были определены с ошибкой (уточнение фаз).

3.I.2 Первые методы, применявшиеся для улучшения набора фаз - это методы, характерные для рентгеноструктурного анализа низкомолекулярных соединений, упомянутые в п.2.2.2. Однако, как показал опыт работы, попытка решения возникающей в этих методах системы уравнений (Сейра) для фаз

$$F_э(\vec{s})\, e^{i\varphi(\vec{s})} = \alpha(\vec{s}) \sum_{\vec{s}_1 \in R'} F_э(\vec{s}-\vec{s}_1) F_э(\vec{s}_1)\, e^{i[\varphi(\vec{s}-\vec{s}_1)+\varphi(\vec{s}_1)]} \qquad (16)$$

(здесь $\alpha(\vec{s})$ - известная нам функция) приводила к успеху лишь при наличии набора величин $F_э(\vec{s})$ достаточно высокого разрешения (порядка 2.0 Å). Попытки же решения этой системы при среднем разрешении приводили к "самосогласованному" набору фаз, не дающему правильной функции $\tilde{\rho}(\vec{r})$.

3.I.3 Более успешными оказались попытки использования при

---

* Более того, из-за частичного нарушения кристаллической упаковки при внедрении тяжелых атомов часто величины $F_{рн}(\vec{s})$ удается измерить для меньшего набора узлов $\vec{s} \in R'$, нежели для нативного белка.

среднем разрешении (порядка 3.0 Å) дополнительной геометрической информации о расположении молекулы исследуемого белка в кристалле. Довольно распространенной для белков (а особенно для вирусов) ситуацией является наличие в элементарной ячейке нескольких идентичных молекул, связанных "некристаллографической" симметрией (не обусловленной свойствами кристаллической решетки). Также распространена ситуация, когда на определенном этапе исследования мы можем приближенно определить область, занятую молекулой, и тем самым определить "пустые" области, где нет атомов исследуемого белка (для белков эти пустоты могут занимать до половины объема элементарной ячейки). Требование, чтобы искомые фазы были таковы, что функция $\varrho(\vec{r})$, построенная с ними согласно (7), имеет предписанную некристаллографическую симметрию и равна нулю вне границ молекулы, приводит к системе уравнений для фаз, которая имеет гораздо лучшие вычислительные свойства, нежели система уравнений Сейра. Итерационное решение такой системы уравнений некристаллографической симметрии является сегодня распространенным инструментом для уточнения набора фаз.

Другой тип дополнительной геометрической информации возникает на более поздних этапах работы, когда про некоторые области элементарной ячейки мы знаем, какие аминокислотные остатки в них расположены. Тогда на искомый набор фаз может быть наложено дополнительное требование, чтобы функция $\varrho(\vec{r})$ в заданных областях имела плотность, отвечающую плотности идентифицированных групп атомов.

3.I.4 Еще одним источником дополнительной информации служит предположение об "атомности" структуры или, иными словами, о том, что фазы $\varphi(\vec{s})$ выражаются согласно (I4) через некоторые неизвестные нам параметры $\vec{r}_j, B_j$. Поэтому можно пытаться улучшить набор фаз, введя некоторые "фиктивные" (не совпадающие с $\vec{r}_j, B_j$) параметры $\vec{r}_j', B_j'$, воспроизводящие согласно (I4) стартовый набор фаз (например, полученный методом изоморфного замещения). Параметры $\vec{r}_j', B_j'$ модифицируются далее так, чтобы попасть в локальный минимум функции (I0), и рассчитать далее улучшенный набор фаз по модифицированным значениям параметров. Такой подход был реализован практически и дал полезные результаты.

3.I.5 Сказанное выше относилось к проблеме определения и улучшения набора фаз. Когда резервы улучшения набора фаз исчерпаны, то есть получен некоторый набор структурных факторов, не улучшаемый доступными нам средствами, возникает проблема построения

"наилучшего изображения $\tilde{\rho}(\vec{r})$ по этим данным.Идейно наиболее простой путь - это построение $\tilde{\rho}(\vec{r})$ в виде частичной суммы ряда Фурье (7).Требование наименьшей среднеквадратичной ошибки построения этой функции (при некоторых вероятностных предположениях о характере экспериментальных ошибок) приводит к введению в сумму весов,отражающих надежность определения конкретных фаз. Такой путь является на сегодняшний день доминирующим.Другой путь - добавить в ряд (7) высокочастотные гармоники $e^{-2\pi i(\vec{h},\vec{r})}$ с $|\vec{h}|>h_{max}$ (про их коэффициенты Фурье нам ничего не известно,то есть любые значения этих коэффициентов не противоречат эксперименту) так,чтобы добиться от $\tilde{\rho}(\vec{r})$ выполнения некоторых предписанных свойств,например,условия $\tilde{\rho}\geq 0$ или наличия некристаллографической симметрии.Разрешив коэффициентам Фурье,модули которых эксперимент нам не дал,принимать в сумме (7) всевозможные значения,мы даже при дополнительных требованиях на $\tilde{\rho}(\vec{r})$ можем получить много функций,удовлетворяющих всем этим ограничениям. Поэтому необходим дополнительный критерий отбора решения.В качестве такого критерия может быть выбран,например,критерий наибольшей гладкости функции $\tilde{\rho}(\vec{r})$ либо наибольшей равномерности ("максимум информационной энтропии").Работы в этом направлении для белковых структур только начинаются.

СОДЕРЖАНИЕ

Т02493. 22.02.83 г. Тираж 250 экз. Заказ 3128Р.
Уч.-изд. л. 1,0. Бесплатно. Изд. № 73.

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