РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ МАТЕМАТИЧЕСКИХ ПРОБЛЕМ БИОЛОГИИ

На правах рукописи
УДК 548.737

Вернослова Елена Анатольевна

РАЗРАБОТКА И ПРИМЕНЕНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ЗАДАЧ БЕЛКОВОЙ КРИСТАЛЛОГРАФИИ

01.04.18 - Кристаллография и кристаллофизика

Диссертация
на соискание ученой степени
кандидата физико-математических наук

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

Пущино - 1996

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ.

Рентгеноструктурный анализ в настоящее время является одним из важнейших методов исследования пространственной структуры макромолекул, позволяющий получить подробную информацию о строении биологически важных молекул на атомном уровне. Знание трехмерной структуры дает возможность проводить дальнейшее исследование объекта с точки зрения его функционирования, механизма активности и протекания биологических процессов. Развитие вычислительной техники, автоматизация рентгеновского эксперимента, совершенствование существующих методов и разработка новых подходов к расшифровке структуры привело к тому, что количество расшифрованных структур быстро растет.

Рентгеноструктурный анализ включает в себя ряд последовательных стадий, начиная от выращивания кристалла и проведения с ним рентгеновского эксперимента и кончая получением трехмерной атомной структуры. Этими этапами являются:

- кристаллизация нативного белка;
- получение тяжелоатомных производных;
- сбор и обработка дифракционных данных;
- определение положений тяжелых атомов;
- расчет фаз структурных факторов;
- построение карт распределения электронной плотности;
- интерпретация электронной плотности и построение атомной модели;
- кристаллографическое уточнение структуры.

Последний этап - кристаллографическое уточнение атомной модели - отнимает большую часть компьютерного времени, необходимого для расшифровки структуры. Поэтому программа уточнения, которую применяют к большим объектам, должна быть максимально эффективной, что приводит к довольно сложным алгоритмам и архитектуре таких программ. Возможности автоматической корректировки сильно зависят от типа информации, которую использует программа

уточнения. Поскольку информация обычно комплексная и происходит из различных источников, важным условием также становится удобство использования программы: простой способ накладывания различных требований к модели, удобные средства подготовки исходных файлов и данных, ясные управляющие параметры.

С появлением машинной графики возможности применения компьютеров в области исследования структуры и функционирования белковых макромолекул получили новое развитие. Начиная с 70-х годов компьютерная графика все активнее используется в рентгеноструктурном анализе макромолекул для построения начальной атомной модели, визуального анализа результатов работы, поиска и исправления ошибок в структуре и решения множества других задач в этой области. Развитие аппаратного и программного обеспечения, предоставляющих пользователю возможность интерактивно управлять изображением, его формой, размером, цветом позволило использовать машинную графику не только в целях иллюстрации, но и для активной работы по созданию и анализу моделей молекул с использованием графических терминалов. Появился целый ряд сложных графических программ и комплексов, без которых ныне немыслим весь процесс расшифровки исследуемой структуры. Современные зарубежные лаборатории разрабатывают и используют мощные графические системы, требующие специализированных дорогостоящих аппаратных средств, высокопроизводительных дисплеев с совершенной архитектурой и соответствующих сложных программных комплексов. В то же время быстрое развитие персональных компьютеров делает актуальным для них разработку собственного программного обеспечения, что требует иного подхода и специальных алгоритмов. Постоянно растущие мощность и объем доступной памяти персональных компьютеров позволяют решать на них все больше задач из круга проблем рентгеноструктурного анализа.

Данная работа посвящена созданию и использованию программного обеспечения для задач белковой кристаллографии для персональных компьютеров класса IBM PC : комплекса программ уточнения FROG и серии программ, использующих компьютерную графику для расчета и визуализации синтезов электронной плотности и атомных моделей.

Диссертация состоит из введения, четырех глав и заключения. В первой главе дается постановка задачи кристаллографического уточнения, описываются основные проблемы и сложности уточнения белковых структур и способы их преодоления. Обсуждается необходимость повышения надежности и требований, предъявляемых к полученным в процессе уточнения атомным моделям, и приводится целый ряд современных критериев проверки качества моделей. В конце главы дается краткий обзор и анализ существующих методов и программ уточнения макромолекул, а также некоторых наиболее распространенных и типичных графических программ, применяемых в рентгеновской кристаллографии.

Во второй главе описывается разработанный с участием автора в лаборатории кристаллографии макромолекул ИМПБ РАН комплекс программ уточнения FROG, предназначенный для решения задач рентгеновской кристаллографии, конформационного анализа, расчетных задач белковой инженерии. Комплекс FROG представляет собой пример разработки программного обеспечения, которое можно легко адаптировать к новым постановкам задач, сохраняя при этом высокую эффективность работы. Его можно применять на любых типах ЭВМ (в том числе и на персональных компьютерах) в любой операционной системе для решения задач уточнения структуры макромолекулы по данным рентгеновского эксперимента с возможным использованием стереохимических ограничений и жестких групп; энергетического уточнения структуры; уточнения ориентации и положения молекулы как целого; уточнение структуры с использованием данных о фазах структурных факторов.

Основными особенностями комплекса FROG являются:

- быстродействие, основанное на последовательном применении алгоритма быстрого дифференцирования и быстрого преобразования Фурье;
- атомно-блочная модель, состоящая из свободных атомов и/или из жестких групп, которыми могут быть вся молекула, отдельные домены, спиральные участки, пептидные звенья, боковые цепи и т.д;
- задание критериев уточнения внешними файлами, что позволяет менять требования, предъявляемые к модели, без корректировки текста программы;
- применение R-free фактора как одного из критериев оценки качества модели;
- адаптируемость к размеру доступной памяти.

В комплексе FROG имеются программы сбора статистики, позволяющие получить информацию о нарушении стереохимических соотношений и невалентных контактов в имеющейся модели молекулы белка.

Глава 3 посвящена описанию графических программ для персональных компьютеров типа IBM PC для расчета, визуализации и исследования синтезов распределения электронной плотности и атомных моделей молекул белка. Все программы этой серии имеют единый графический интерактивный интерфейс, простой и удобный для пользователя, и взаимодействуют друг с другом посредством файлов фиксированных форматов. Набор возможных операций в каждой программе достаточно разнообразен для решения поставленных задач, но это не делает их громоздкими и запутанными при их эксплуатации, что свойственно многим большим программным комплексам.

В описываемую серию входят программы FFT, FAN, CAN и FANS, позволяющие рассчитать синтезы Фурье и Паттерсона любого типа в любой пространственной группе; вывести синтезы на экран монитора в режиме моно или стерео и провести их визуальное исследование;

сравнить несколько синтезов, накладывая их один на другой; одновременно с синтезом вывести на экран атомную модель исследуемого объекта и провести с ней некоторые манипуляции (вращение, трансляцию) и т.д.

В главе 4 приводятся примеры применения разработанного программного обеспечения. По запросам пользователей графические программы были разосланы более чем по 20 адресам во многие зарубежные университеты, институты и научно-исследовательские центры. Графические программы применяются для визуализации и анализа нескольких вариантов рассчитанных карт электронной плотности с различными весовыми схемами на разном разрешении, т.е. в тех случаях, когда приходится иметь дело с многочисленными расчетами и просмотрами карт электронной плотности; для изображения моделей молекул белка вместе с плотностью и их пространственной упаковки в кристалле.

Программа уточнения FROG применялась при расшифровке структуры γ-кристаллина IIIb из хрусталика глаза теленка на разрешении 2.5Å совместно с сотрудниками Института белка РАН; для компьютерного моделирования белковых гибридов (лизоцима фага T4 с CA^{2+}- связывающим модулем); для улучшения качества синтеза и расширения набора фаз с 3Å до 2.4Å комплекса углеводспецифичного белка лектина из семян гороха с глюкозидом совместно с сотрудниками Института молекулярной генетики; при уточнении положения модели и расширения набора фаз с 18Å до 6Å структуры *CMtV* капсида вируса крапчатости гвоздики совместно с сотрудниками Института кристаллографии РАН.

ГЛАВА 1.

Уточнение макромолекул в белковой кристаллографии.

Термин "уточнение структуры макромолекул" в широком смысле означает серию этапов исследования, включающих целый ряд различных действий: расчеты синтезов Фурье, интерактивную корректировку модели, автоматическую модификацию атомной модели. В более узком смысле уточнением атомной модели является только автоматическая корректировка, выполняемая специальными компьютерными программами. Ставя задачу уточнения в широком смысле, мы должны ответить на вопросы:

- каким **требованиям** должна удовлетворять модель,
- какими **средствами** можно достичь поставленной задачи,
- как **контролировать** ход уточнения и качество результата.

Кристаллографическое уточнение является завершающей стадией исследования структуры макромолекул. При этом есть несколько возможностей уточнения моделей белка: можно уточнить фазы; расширить набор фаз на область более высокого разрешения; добиться более лучшего соответствия модели и рассчитанной электронной плотности. В процессе автоматической модификации на атомную модель могут накладываться самые различные требования: соответствие рентгеновским или нейтронным дифракционным данным, удовлетворение энергетическим и стереохимическим ограничениям, близость к гомологам и т.д. Эти требования отражают доступную информацию об объекте в количественной форме.

Для улучшения стартовой атомной модели разработан и реализован целый ряд алгоритмов и программ уточнения. Они включают использование при уточнении "мягких" (restraints) ограничений на стереохимические параметры модели /1-3/, жесткого (constraints) соблюдения стереохимических условий /4-5/, точных молекулярно-механических силовых полей /6/, методов быстрого преобразования Фурье для ускорения вычислений /7/, методов молекулярной

динамики, основанных на энергетических потенциалах /8-11/, моделирование посредством компьютерной графики на основе карт электронной плотности /12/ и т.д.

После того, как выбраны формальные критерии и варьируемые параметры модели, проблема уточнения представляет собой задачу локальной минимизации функции большого числа переменных. Локальный характер минимизации определяется тем, что функция имеет большое число почти равноценных минимумов, и в силу экспериментальных ошибок и неточности теории рассеяния правильному решению может соответствовать не самый глубокий минимум.

И, наконец, в процессе расшифровки структуры молекулы могут возникать как ошибки, связанные с экспериментом, так и ошибки при интерпретации промежуточных результатов работы. Поэтому важным элементом исследования является оценка качества модели и возможность идентификации участков структуры, требующих более тщательного изучения.

1.1. Метод наименьших квадратов (МНК) и его применение в кристаллографическом уточнении макромолекул.

Процедура уточнения заключается в систематическом варьировании атомных параметров таким образом, чтобы получить наилучшее согласование между амплитудами структурных факторов, вычисленных для текущей модели, и наблюдаемыми амплитудами. При этом используется серия последовательных циклов уточнения до тех пор, пока на некоторой стадии не будет достигнута сходимость, а изменение атомных параметров станет пренебрежимо мало по сравнению с ожидаемыми ошибками.

Количественной мерой близости расчета и эксперимента может служить фактор расходимости (R-фактор)

$$R = \sum_H \left\| F_H^o \right| - \left| F_H^c \right\| \Big/ \sum_H \left| F_H^o \right| \qquad (1)$$

или корреляционная функция общего вида

$$Q = \sum_H w_H \left(k \left|F_H^o\right|^\alpha - \left|F_H^c\right|^\alpha\right)^\beta , \qquad (2)$$

где W_H - весовой множитель, характеризующий, например, точность измерения $|F_H^o|$ или достоверность вычисления $|F_H^c|$;

k - коэффициент приведения $|F_H^o|$ к абсолютной шкале;

α, β - некоторые константы.

Правильной модели соответствует абсолютный минимум R-фактора, а процесс уточнения заключается в нахождении наиболее низкой точки корреляционной функции Q при, например, $\alpha = 1$ и $\beta = 2$, методом наименьших квадратов и градиентного спуска. Степень сложности такой задачи зависит от числа минимизируемых параметров.

Сформулируем задачу в общем виде. Пусть n неизвестных параметров $\{x_1, \ldots, x_n\}$ связаны с m наблюдаемыми величинами $\{g_1, \ldots, g_m\}$ системой линейных уравнений

$$\begin{cases} g_1 = a_{11} x_1 + \ldots + a_{1n} x_n \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ g_m = a_{m1} x_1 + \ldots + a_{mn} x_n \end{cases} , \quad \text{где } a_{ij} \text{ - константы.} \qquad (3)$$

Эти уравнения называются *условными* уравнениями. Итак, имеется n неизвестных параметров и m наблюдений. При n>m задачу нельзя решить; при n=m имеется единственное решение; при n<m задача переопределена, и возникает проблема нахождения наилучших значений параметров. Предположим, что известны приближенные значения параметров $x_1, \ldots, x_n$. Тогда ошибки будут

$$\begin{cases} E_1 = a_{11} x_1 + \ldots + a_{1n} x_n - g_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ E_m = a_{m1} x_1 + \ldots + a_{mn} x_n - g_m \end{cases} \qquad (4)$$

Наилучшими значениями параметров будут те, для которых

$\sum_j^m E_j^2 \Rightarrow \min$, т.е.

$$\partial(\sum_j^m E_j^2)/\partial x_1 = \partial(\sum_j^m E_j^2)/\partial x_2 = \ldots = \partial(\sum_j^m E_j^2)/\partial x_n = 0 \qquad (5)$$

Теперь имеется n уравнений для n неизвестных (*нормальные* уравнения):

$$\partial(\sum_j^m E_j^2)/\partial x_k = \sum_j^m a_{j1}(a_{j1}x_1 + \ldots + a_{jn}x_n - g_j) = 0\ ,\quad k=1, \ldots, n \qquad (6)$$

В матричной форме условные уравнения записываются как $\mathbf{Ax} = \mathbf{g}$, а нормальные уравнения как $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{g}$, где $\mathbf{A}^T$ - транспонированная матрица. Тогда система из n уравнений решается с помощью обращения матрицы $\mathbf{U}=\mathbf{A}^T\mathbf{A}$ и решение выглядит как $\mathbf{x}=(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{g}$. Часто некоторые наблюдения оказываются более надежными, чем другие, и тогда это учитывается введением в уравнения весовых множителей. Более точным измерениям приписываются большие веса.

В рентгеновской кристаллографии для приведения в соответствие наблюдаемых и вычисленных интенсивностей наилучшими параметрами будут такие, которым отвечает минимум функционала

$$Q = \sum_H w_H (|F_H^o| - |F_H^c|)^2 \qquad (7)$$

Обозначим координаты $\mathbf{r} = (x, y, z)$ всех n независимых атомов ячейки и другие возможные параметры (константы тепловых колебаний атомов, коэффициент приведения к абсолютной шкале), от которых зависят $|F_H^c|$, через $\{\xi_i\}_{i=1,..,3n,...,q}$. Задача заключается в нахождении значений $\{\xi_i\}$, отвечающих минимуму функционала Q, т.е. в решении системы уравнений $\partial Q/\partial\xi_i = 0$. Поскольку искомые параметры входят в $|F_H^c|$, то система приводится к виду

$$\sum_H w_H \left(|F_H^o| - |F_H^c|\right) \frac{\partial |F_H^c|}{\partial \xi_i} = 0 \qquad (i=1,2,...,3n,...,q) \tag{8}$$

Здесь $|F_H^c| = |\sum_H f_k \tau_k e^{2\pi i(Hr_k)}|$, (9)

f_k - фактор атомного рассеяния k-го атома;

$\tau_k = e^{-B_k(\sin\theta/\lambda)^2}$ для изотропных тепловых колебаний или

$\tau_k = e^{-(b_{1k}h^2+b_{2k}k^2+b_{3k}l^2+b_{4k}hk+b_{5k}hl+b_{6k}kl)}$ для анизотропных колебаний;

r_k, B_k и b_{ik} - уточняемые параметры.

Т.к. искомые параметры ξ_i входят в аргументы тригонометрических функций, из которых построены $|F_H^c|$, то в общем виде система не решается. Уравнения (8) не являются линейными относительно параметров, а для непосредственного применения МНК требуется система линейных уравнений.

Если выбрать пробную структуру достаточно близкой к искомой, то можно осуществить переход к линейной системе уравнений, но уже относительно сдвигов параметров, а не самих параметров. Итак, пусть найдено приближенное решение структуры, т.е. известны значения параметров ξ_i^o, достаточно близкие к искомым ξ_i. Разложив $|F_H^c|$ в ряд Тейлора, можно оборвать ряд на втором члене:

$$|F_H^c| = |F_H^c|_o + \sum_j^q \left(\frac{\partial |F_H^c|}{\partial \xi_j}\right) \Delta\xi_j , \qquad \text{где } \Delta\xi_j = \xi_j - \xi_j^o \tag{10}$$

Справедливость этого приближения зависит от того, насколько близка пробная структура к истинной. При неблагоприятных условиях процедура уточнения сходится не к правильному решению, а к ложному локальному минимуму, т.е. успех этого метода определяется удачным выбором пробной структуры. Такое приближение означает,

что в пределах изменения параметров функция $|F_H^c|$ меняется линейно, и, следовательно, можно считать, что

$$\frac{\partial |F_H^c|}{\partial \xi_j} = \left(\frac{\partial |F_H^c|}{\partial \xi_j}\right)_0$$ - значение производной в точке $\{\xi_j^o\}$. (11)

Подставив (10) в (8) и с учетом (11), получим: (12)

$$\sum_H w_H \left(|F_H^o| - |F_H^c|_o\right)\left(\frac{\partial |F_H^c|}{\partial \xi_i}\right)_0 - \sum_j^q \sum_H w_H \left(\frac{\partial |F_H^c|}{\partial \xi_i}\right)_0 \left(\frac{\partial |F_H^c|}{\partial \xi_j}\right)_0 \Delta\xi_j = 0$$

Перейдя к другим обозначениям, получим систему из q линейных уравнений:

$$\sum_j^q a_{ij}\,\Delta\xi_j - c_i = 0, \qquad i=1, \ldots, q \qquad (13)$$

что позволяет найти все $\xi_j = \xi_j^o + \Delta\xi_j$. В матричном виде система уравнений выглядит как $\mathbf{A}\Delta\xi = \mathbf{C}$, где матрица нормальных уравнений

$$\mathbf{A}=\begin{pmatrix} a_{11}\, a_{12} \cdots a_{1q} \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ a_{q1}\, a_{q2} \cdots a_{qq} \end{pmatrix}, \quad \Delta\xi=\begin{pmatrix} \Delta\xi_1 \\ \ldots \\ \Delta\xi_q \end{pmatrix}, \quad \mathbf{C}=\begin{pmatrix} C_1 \\ \ldots \\ C_q \end{pmatrix}.$$

При такой линеализации матрица **A** будет содержать недиагональные элементы, включающие множители вида $\frac{\partial |F_H^c|}{\partial \xi_i}\,\frac{\partial |F_H^c|}{\partial \xi_j}$. Поскольку для макромолекул число наблюдений и число уточняемых параметров может достигать многих тысяч, то полная норамальная матрица может иметь миллионы элементов, каждый из которых содержит несколько тысяч слагаемых, и при этом становятся неизмеримо громоздкими как вычисления, так и хранение и обработка матриц, поэтому необходимо использовать подходящие аппроксимации.

Для сокращения вычислений в первом приближении можно считать, что корреляция между атомными параметрами отсутствует, и тогда при $i \neq j$ недиагональными членами нормальной матрицы можно пренебречь. Такая аппроксимация называется диагональным приближением МНК, уравнения (12) при этом принимают вид:

$$\sum_H w_H \left(|F_H^o| - |F_H^c|\right) \left(\frac{\partial |F_H^c|}{\partial \xi_i}\right)_o = \Delta\xi_i \sum_H w_H \left(\frac{\partial |F_H^c|}{\partial \xi_i}\right)_o^2 \qquad (14)$$

Другим способом аппроксимации является сохранение в нормальной матрице кроме диагональных еще и тех элементов, которые вносят вклад в стереохимические условия, связывающие атомы (если такие условия применяются при уточнении). Это ведет к сильно разреженной матрице, и для решения таких уравнений подходит метод сопряженных градиентов.

Вследствие нелинейности системы уравнений (8) и последующей аппроксимации вектор сдвигов $\Delta\xi$ не всегда является оптимальным (с точки зрения снижения значения минимизируемой функции (7)). Поэтому при использовании МНК необходимо определять оптимальное значение шкального коэффициента K для вектора сдвигов: $\Delta\xi_{opt}=K\Delta\xi$. Выполнить это можно путем многократного вычисления выражения (7) или кристаллографического R-фактора (1) при различных значениях K, и наиболее низкому значению R-фактора будет соответствовать оптимальное значение K.

Для определения q параметров необходимо иметь по крайней мере q экспериментальных наблюдений, но на самом деле, т.к. наблюдения содержат экспериментальные погрешности, число наблюдений m должно существенно превышать число варьируемых параметров. В большинстве случаев при работе с трехмерными рентгеновскими данными для малых молекул, кристаллы которых позволяют собрать дифракционный набор до разрешения 1Å и выше, отношение m/q

составляет 5 ÷10, так что уравнения существенно переопределены. Столь высокое соотношение обеспечивает точность определения координат атомов с ошибкой, не превышающей тысячных долей Å.

Однако при использовании МНК для уточнения структур белков возникают трудности. Во-первых, число экспериментальных модулей структурных факторов может оказаться близко к числу параметров, что происходит из-за ограниченности дифракционного поля кристаллов макромолекул, и не получается необходимой переопределенности уравнений МНК. Кроме того, из-за размера матриц нельзя воспользоваться преимуществами полноматричного МНК. В то же время положения ковалентно связанных атомов сильно коррелированы, и применение диагонального варианта необосновано. Компромиссом в таком случае служит вариант расчленения полной матрицы на блоки, и параметры одного блока уточняются полноматричным МНК. Параметры валентно связанных атомов включаются в один блок, а разбиение на блоки производится с перекрыванием, чтобы ввести некоторую корреляцию между валентно связанными атомами, отнесенными в разные блоки.

Т.о. весь процесс уточнения по методу наименьших квадратов можно представить в виде следующих шагов:

1) Расчет структурных факторов по координатам модели.

2) Расчет частных производных по атомным параметрам.

3) Составление матрицы **A** нормальных уравнений. Обычно для белков оставляют только диагональные элементы матрицы, но в отдельных случаях включают и некоторые недиагональные члены, соответствующие наиболее сильной корреляции между уточняемыми параметрами.

4) Решение системы нормальных уравнений, для чего применяется целый ряд методов в зависимости от количества ненулевых элементов в матрице нормальных уравнений. В случае расчета лишь главной

диагонали обращение матрицы представляет собой тривиальную процедуру. В результате решения системы нормальных уравнений определяется вектор сдвигов исходных параметров $\Delta\xi$.

5) Определение оптимального шкального коэффициента K.

6) Применение оптимального сдвига к исходным параметрам $\xi=\xi_0+\Delta\xi$.

Перечисленная последовательность шагов имеет итерационный характер, и процесс продолжается до достижения сходимости.

Приведенная процедура уточнения является лишь некоторой общей схемой работы алгоритма той или иной программы уточнения, основанной на методе наименьших квадратов. В зависимости от свойств программы некоторые из перечисленных шагов осуществляются одновременно, а в ряде случаев могут появляться еще промежуточные шаги, необходимые для работы.

1.2. Проблема переопределенности задачи кристаллографического уточнения.

При уточнении макромолекул возникает несколько проблем, характерных именно для этого класса структур.

a). Кристаллы макромолекул имеют большую элементарную ячейку, и требуется собрать огромное количество экспериментальных данных, которым свойственно низкое значение сигнал/шум отношения. Поэтому обычно трудно собрать набор данных атомного разрешения, как это делается для низкомолекулярных структур. Кроме того, доступный набор данных часто содержит как систематические, так и случайные ошибки из-за размеров кристалла и технических проблем съемки.

b). Кристаллы белков создают дополнительные трудности за счет того, что обычно в объеме этих кристаллов содержится большой процент растворителя. Благодаря высокому содержанию растворителя в макромолекулярных кристаллах им свойственна большая тепловая подвижность и динамическая разупорядоченность возможных локальных конформаций. Эти физические особенности кристаллов

макромолекул оказывают влияние на быстрое падение интенсивностей дифрагирующих пучков с ростом угла рассеивания, что приводит к ограничению предельного разрешения, на котором можно измерить экспериментальные данные.

c). Перечисленные выше причины ведут к тому, что удается собрать набор экспериментальных данных ограниченного разрешения: фактически для больших белков, вирусов, нуклеиновых кислот набор данных выше 2.0Å является недоступным.

d). Из-за ограниченности разрешения доступного набора данных отношение числа экспериментальных параметров к числу уточняемых параметров становится слишком малым для того, чтобы обеспечить сходимость и устойчивость обычного метода минимизации, используемого при уточнении - метода наименьших квадратов.

В то же время точность и надежность атомной модели, полученной в процессе кристаллографического уточнения, сильно зависит от степени переопределенности задачи минимизации МНК. Ошибки в минимизиуемых переменных тем меньше, чем выше отношение числа экспериментальных данных к числу уточняемых параметров модели. В случае низкомолекулярных структур это отношение достигает 10:1, даже если в набор переменных входят 6 анизотропных тепловых параметра для каждого атома, и уточнение, проводимое с такой избыточностью данных, дает решение с большой точностью. Однако для кристаллов макромолекул такой переопределенности системы уравнений удается достичь очень редко.

Ограниченность разрешения макромолекулярных структур также оказывает влияние на качество стартовой модели для уточнения. В то время как стартовые атомные позиции для малых органических молекул обычно отклоняются в пределах 0.1Å от своих конечных значений, то для белков стартовая точность атомных координат в среднем 0.5Å. Это означает, что линейная аппроксимация (12) системы

нелинейных уравнений вряд ли допустима, и резко повышается вероятность попадания решения в локальный минимум.

1.3. Уточнение с ограничениями ("constraints" и "restraints")

Можно рассмотреть два способа улучшения недостаточной переопределенности задачи: сократить число уточняемых параметров либо увеличить количество независимых наблюдений. Общеупотребимым средством для достижения этого является привлечение дополнительной информации, в первую очередь данных о химических и физических закономерностях (атомность, связность, геометрия, стереохимия, упаковка, некристаллографическая симметрия), свойственных макромолекулярной структуре. Дифракционные данные и дополнительная информация поэтому часто комбинируются в процессе расшифровки структуры объекта.

Имеется два способа, с помощью которых стереохимическая информация вводится в процесс уточнения: ее можно добавить как дополнительные наблюдения (мягкие ограничения типа "restraints"), которые входят в минимизируемый функционал с соответствующим взвешиванием, либо модель может быть параметризована таким образом, чтобы геометрия отдельных ее частей оставалась всегда идеальной, а соответствующие переменные исключались из уточнения (жесткие ограничения типа "constraints").

"Constraints" - это точные математические условия, которые приводят к исключению из рассмотрения (из процедуры минимизации) соответствующие переменные, поскольку их можно выразить через остальные параметры. "Restraints" - это дополнительные условия, которые не являются точными, а применяются в форме добавочных уравнений и дают вклад в минимизируемый критерий, например, в виде $w\Delta r^2$, где w - весовой множитель, Δr - отклонение параметра от его точного значения. "Restraints" удерживают параметры модели в некоторой физически разумной области возможных значений, в то

время как “constraints” приписывают им определенные конкретные значения. В принципе, “restraints” становятся “constraints”, если задать для них очень высокий вес, но они оставляют количество формальных параметров без изменения, в то время как “constraints” сокращают число уточняемых переменных.

Хотя использование “constraints” более эффективным образом улучшает соотношение числа наблюдаемых и уточняемых параметров, у “restraints” есть свои преимущества. Модели с “restraints” ограничениями белее реалистичны, и различные типы “restraints” соотношений можно по-разному взвешивать, чем достигается более гибкая процедура уточнения. Как правило, эти ограничения входят в программы уточнения в виде критериев, формально имеющих такой же вид, как и рентгеновский критерий, что делает процедуру минимизации единообразной и упрощает вычисления.

Что касается конкретного вида используемой стереохимической информации, то в настоящее время накоплено много данных о геометрии и конформации отдельных компонент, из которых построены макромолекулы, и об определенных стереохимических особенностях самих полимеров. Эта информация происходит из различных источников, включающих химический анализ, теоретические исследования, кристаллографическое определение структур фундаментальных химических единиц и олигомеров, и содержит точные значения длин валентных связей и углов, хиральность асимметричных центров, планарность отдельных групп, конформационную предпочтительность торсионных углов, Ван-дер-Вальсовые взаимодействия, возможные водородные связи, геометрию элементов вторичной структуры, некристаллографическую симметрию. Геометрические соображения также имеют место при задании границ изменения температурных параметров и коэффициентов заполнения.

При фиксации геометрических характеристик улучшение переопределенности может быть значительным. Например, если

валентные связи, углы и плоские группы в модели фиксированы, то в качестве переменных остаются только конформационные углы вращения. Другим способом сокращения числа формальных параметров является переход от индивидуальных атомных характеристик к параметрам, описывающим жесткие группы атомов (вектор трансляции, ориентации и температурные факторы для атомной группы, либо конформационные углы вращения групп вокруг одинарных связей). В качестве жестких групп можно задавать и крупные элементы вторичной структуры, но такой подход становится неэффективным при уточнении на высоком разрешении.

1.4.Надежность результатов кристаллографического уточнения

Чтобы иметь возможность сделать предположения и выводы о механизме функционирования исследуемой молекулы, необходима надежная структурная информация о ее пространственном строении. Рентгеновский структурный анализ является на сегодняшний день единственным экспериментальным методом, позволяющим определять атомную структуру как низкомолекулярных соединений, так и больших макромолекулярных комплексов. Но рентгеновский эксперимент часто бывает настолько сложным и требующим больших затрат времени и сил, что его трудно повторить независимым образом другим исследователям. Это накладывает большую ответственность на кристаллографов, работающих в области расшифровки структур макромолекул. В работе /13/ Е.Додсон дает подробный обзор и анализ проблем, связанных в первую очередь с необходимостью адекватной оценки правильности и надежности расшифрованных структур, а также выборе эффективных критериев для этой оценки.

1.4.1. Что такое “хорошая” и “плохая” модель?

Итак, пусть на некоторой стадии расшифровки структуры (вписывание модели в электронную плотность, кристаллографическое уточнение, определение структуры из двумерного ЯМР, белковая

инженерия, предсказание структуры мутантов и т.д.) получена полная атомная модель. После этого встает вопрос, как можно проверить модель на ее корректность? Правильно ли установлен ход полипептидной цепи, правильно ли построены все боковые цепи аминокислотных остатков, т.е. насколько хороша полученная модель.

Если коротко дать определение хорошей модели, то это модель, имеющая разумные во всех смыслах характеристики:

- **химические**: валентные связи и углы имеют приемлемые значения; хиральность имеет правильный знак; ароматические кольца и пептидные звенья плоские; атомы, связанные водородными связями, действительно могут их образовывать.
- **физические**: в модели нет слишком близких контактов; связанные каким-либо образом атомы (н/к симметрией, ковалентными или водородными связями) имеют близкие температурные факторы; заряженные группы спрятаны в гидрофобном окружении.
- **закономерности строения белковых молекул**: белок имеет определенную вторичную структуру; распределение торсионных углов φ и ψ на карте Рамачандрана /14/ не содержит запрещенных конформаций; большинство боковых цепей имеют разрешенные поворотные конформации; молекулы воды и ионы правильно размещены; большинство (или все) пептидные группы находятся в транс-конфигурации.
- **статистические**: модель наилучшим образом соответствует экспериментальным данным.

На каждом этапе построения, корректировки и уточнения в модель могут быть внесены ошибки. В работах /15, 16/ проведена классификация возможных ошибок кристаллографических моделей:

- полностью неправильная модель или субъединица, в которой вся или существенная часть главной цепи проведена неверно;

- частично неверный ход главной цепи, обычно из-за ошибок при соединении элементов вторичной структуры;
- локальные ошибки из-за неаккуратного построения модели либо из-за недостаточности данных;
- неверная конформация боковой цепи;
- неправильная ориентация пептида.

В статье Г.Клейвета & Т.Джонса /16/ приводится целый список примеров структур, взятых из PDB, с перечисленными ошибками. Одной из наиболее частых ошибок является переизбыток параметров модели. Если при уточнении используется гораздо больше параметров, чем это оправдано имеющимся набором экспериментальных данных и другой дополнительной информацией, то кристаллографический R-фактор можно снизить почти до произвольной величины, не улучшив при этом качества модели. Например, этого можно достичь, проводя уточнение молекул, связанных н/к симметрией, без ограничений; уточняя индивидуальные температурные факторы на низком разрешении; сажая без всякого на то основания множество молекул воды; уточняя коэффициенты заполнения и альтернативные конформации на среднем и низком разрешении.

1.4.2. Необходимость применения адекватных критериев при оценке качества моделей.

Некоторые журналы ("Nature", "Science" и др.) удовлетворяются минимальным списком требований к качеству предъявленной структуры: разрешение, стандартный R-фактор, среднее значение температурного фактора, среднеквадратичное отклонение от идеальных значений длин валентных связей и углов. Этих характеристик абсолютно недостаточно, чтобы не только оценить степень надежности модели, но и отличить правильную модель от неправильной. В работе /16/ приведен пример двух моделей, одна из которых является

реально расшифрованной структурой, а другая сконструирована полностью неверно.

Молекула	"X"	"Y"
Разрешение (Å)	3.0	2.9
R-фактор	0.214	0.251
rmsd длин связей (Å)	0.009	0.009
rmsd валентных углов (Å)	2.1	1.6
<B> (Å^2)	13.4	49.2

Табл.1. Список стандартных показателей качества модели для двух белковых структур.

В табл.1 приведены их характеристики, из которых можно сделать вывод, что модель "X" вполне удовлетворительная, а модель "Y" хуже, т.к. у нее выше как R-фактор, так и средний температурный фактор B. Тем не менее неверной структурой является модель "X", а большие значения R и B для второй модели объясняются плохим качеством данных (эффективное разрешение ~ 3.2Å, структура уточнялась с жестко зафиксированной н/к симметрией).

В другой работе /17/ А.Уржумцевым был проведен анализ взятых из PDB /18/ файлов для проверки геометрии расшифрованных структур. Из 100 файлов, подряд идущих в PDB, были исключены файлы со структурами ДНК, с неполными, неуточненными или частично уточненными моделями. Для оставшихся 50 структур был проведен анализ распределения ошибки в валентных связях, который подтвердил наличие отдельных больших отклонений в моделях. Так, для 30 из 50 моделей $\Delta d_{max} > 0.1$Å; для 9 из 50 моделей $\Delta d_{max} > 0.2$Å. Проведенный анализ показывает, что в процессе уточнения следует контролировать не только средние отклонения в геометрии модели, но

и следить за максимальными величинами. Комплекс программ уточнения FROG /19-23/, например, дает возможность получить гистограммы распределения всех стереохимических характеристик и распечатать отдельные нарушения, превышающие заданный пользователем порог. Ниже приводится целый ряд критериев качества моделей, которые дают более объективную оценку, особенно в сочетании друг с другом.

1.4.3. Критерии оценки качества атомной модели.

В работах /13,16,24,25/ проанализированы наиболее широко используемые подходы и дается сравнение эффективности разнообразных критериев, применяемых в настоящее время для обнаружения ошибок и оценки качества моделей молекул белка.

1). Стандартный кристаллографический R-фактор.

При оценке общего качества структуры общеупотребимыми показателями являются кристаллографический R-фактор и разрешение. Чем выше разрешение набора экспериментальных данных, тем выше считается точность определения структуры. Но для R нет однозначной зависимости между его величиной и надежностью структуры: низкое значение R само по себе является необходимым, но не достаточным условием точности модели. Его следует рассматривать как функцию, зависящую от разрешения и полноты набора экспериментальных данных. Как уже отмечалось выше, им можно легко манипулировать, исключая некоторые данные или добавляя несвойственные модели параметры, т.е. его можно снизить различными способами, не улучшив при этом качество структуры. Можно получить совершенно неправильную структуру, но имеющую низкий R. Несколько исследований, обобщенных в работе С.Брандена & Т.Джонса /15/, показали, что среднее значение R=0.25 не является гарантией правильности модели. Тем не менее структура считается достаточно надежно определенной при разрешении выше 2Å и с R не более 20%.

Существует серия других критериев, которые можно подразделить на две большие категории. Первая категория включает в себя такие характеристики, для вычисления которых требуется привлечение экспериментальных данных (оценка средней ошибки в координатах атомов; величина Rfree фактора; различные критерии, описывающие соответствие атомной модели и синтеза электронной плотности). Вторая категория критериев вычисляется непосредственно по атомным координатам структуры и не требует экспериментальных данных. Это всевозможные стереохимические и энергетические функционалы и методы, основанные на статистической информации, полученной по известным структурам из PDB, которые оценивают пространственную укладку модели и окружение отдельных аминокислотных остатков.

2). Оценка ошибок в координатах атомов модели.

В.Лузатти /26/ предложил способ оценки ошибки в координатах атомов модели $\Delta\mathbf{r}$ на основе статистического метода А.Вильсона /27/. В предположении независимого нормального распределения ошибок в координатах атомов метод Лузатти дает возможность приближенной оценки $|\Delta\mathbf{r}|$ по величинам R-фактора. При этом R-фактор рассматривается как функция разрешения, и считается, что все ошибки, которые берутся в рассмотрение, сосредоточены только в ошибках координат. Позднее было обнаружено, что график Вильсона дает достаточно грубые приближения, и было предложено несколько более точных методов как координатных, так и фазовах оценок/28 -31/.

3). Соответствие модели и синтеза электронной плотности.

Карта распределения электронной плотности позволяет визуально оценить, насколько хорошо атомная модель согласуется с синтезом. На основе этого можно получить и количественные характеристики отдельных частей структуры (R-фактор в прямом пространстве, коэффициент корреляции и т.д.).

Критерий, предложенный Т.Джонсом и др. /12/, измеряет соответствие между картой распределения электронной плотности и атомной моделью. Для каждого остатка все атомы или их некоторое подмножество преобразуются в карту путем задания каждого атома в виде Гауссовой функции распределения электронной плотности с заданным температурным фактором. Затем модельная карта сравнивается с экспериментальной по совокупности точек сетки в окрестности исследуемых атомов. Величина рассчитанного таким образом R-фактора изменяется от 0.0 (при идеальном совпадении) до 1.0 и обычно бывает около 0.25 для правильной структуры, но может достигать и 0.5 для подвижных боковых цепей, контактирующих с растворителем. Структура с ошибочной укладкой характеризуется высоким значением R-фактора для большинства остатков.

Комплементарным подходом является вычисление R-фактора в обратном пространстве с поочередным удалением аминокислотных остатков из модели - метод "скользящего окна". Каждый из проверяемых остатков удаляется из модели и по ней рассчитываются структурные факторы. Если удаленный остаток содержал значительные ошибки в координатах атомов, то при такой процедуре R снизится. Описанный алгоритм можно осуществить, например, с помощью программы X-PLOR А.Брюнгера /32/. В ней рассчитывается общий R для полной модели, затем методом "скользящего окна" определяются R-факторы для моделей с исключенными по-очереди остатками и вычисляется разность с общим R (нормированная на массу удаленных атомов). Чем хуже был вписан остаток, тем выше величина этого критерия.

4. Rfree фактор.

Независимым показателем соответствия рассчитанных по модели и экспериментальных амплитуд является Rfree фактор, введенный в 1992г. А.Брюнгером /33,34/. Он базируется на общем статистическом принципе "cross validation", когда модель должна воспроизводить не только те экспериментальные данные, на основе которых она

строилась, но и независимый набор данных, исключенных из процесса расшифровки модели. Критерий использует случайно выбранный набор экспериментальных данных, не участвующих в уточнении модели, только для оценки надежности определения атомных параметров, шкальных коэффициентов и т.д.

Rfree является полезным объективным средством оценки хода и результатов уточнения, эффективность которого можно продемонстрировать на следующем примере. Если введение новых параметров модели (посадка молекул воды или переход от изотропных температурных факторов к анизотропным) приводит к одинаковому снижению как R, так и Rfree, то такая процедура улучшает модель. Если же R падает, а Rfree остается тем же самым или возрастает, то введение новых параметров является неправомерным и не улучшает модель. Более подробно способы применения Rfree описаны ниже в п.1.5.

5. Оценка геометрических параметров модели. Словари стандартной стереохимии. Программа PROCHECK.

С задачей оценки и учета конкретных стереохимических условий тесно связана проблема идентичности словарей стандартной стереохимии, используемых в разных программах. Для правильного учета каждого конкретного условия для него нужно определить две величины - его идеальное значение и индивидуальный весовой коэффициент, с которым данное ограничение входит в общий критерий.

В работе /35/ в 1993 г. Д.Пристли был проведен анализ длин валентных связей, относящихся только к остатку триптофану, по словарям, используемым в программах PROLSQ, TNT, X-PLOR и FRODO. Анализ показал, что разброс величин идеальных параметров, заложенных в эти программы, в отдельных случаях превышал 0.1Å. Следовательно, при регуляризации модели то, что является идеальным для одной программы, будет испорченной моделью с точки зрения другой программы. Поэтому возникла необходимость в стандартизации

стереохимических параметров для всех программ уточнения и построения моделей.

В 1991 г. Р.Энг и Р.Хубер /36/ систематически исследовали около 80000 низкомолекулярных кристаллических структур, имеющихся в базе данных, и получили набор значений длин валентных связей, валентных углов и весовые оценки для них. Д.Пристли провел состыковку /37/ геометрических стандартов для программ X-PLOR, PROTIN, TNT, O и FRODO с параметрами Энга и Хубера, так что после этого среднеквадратичное отклонение идеальных длин связей, заданных во всех словарях, снизилось с 0.031Å до 0.003Å, и для валентных углов с 1.9° до 0.3°.

Продолжением работы по стандартизации стереохимических параметров для белков явилось исследование, проведенное в 1994 г. В.Ламзиным и др. /38/, в котором геометрия нескольких белковых структур, расшифрованных по данным атомного разрешения (1.2Å и выше), сравнивалась с общепринятыми стандартами Энга и Хубера, полученными для низкомолекулярных соединений. Поскольку отбирались модели белков, для которых уточнение проводилось без использования стереохимических ограничений или с применением слабых ограничений, то по результирующим моделям можно было получить стереохимическую информацию, свойственную именно белкам. В геометрии белков были отмечены некоторые закономерности, отличающие их параметры от соответствующих параметров малых структур. Например, валентные связи в белках имеют тенденцию быть короче; есть некоторые отличия для аминокислотных остатков Pro и Gly, которые не учитываются в словаре Энга и Хубера. Эти исследования дают основание предположить, что работа по окончательному созданию словаря идеальной стереохимии для белков еще не завершена, и с накоплением статистической информации о геометрии белков он будет обновляться.

Многие перечисленные выше критерии дают оценку общего качества модели целиком, и поэтому информация об отдельных частях структуры остается недоступной. Однако строение некоторых областей молекулы устанавливается более надежно, чем других. Например, внутренние области ядра белковой глобулы имеют тенденцию показывать более четкую картину распределения электронной плотности, которую легче интерпретировать в процессе расшифровки структуры. В 1993 г. Р.Ласковским и др. /24/ был разработан комплекс программ PROCHECK, который производит детальную проверку стереохимии как для всей белковой структуры в целом, так и для каждого аминокислотного остатка цепи. В основном контролируются параметры, которые обычно не включены в широкоиспользуемые процедуры уточнения и поэтому слабо поддаюся улучшению с их помощью. Стереохимическими параметрами в PROCHECK являются все двугранные углы, Cα-хиральность, планарность пептидной группы, энергия водородных связей в главной цепи, а также длины валентных связей и валентные углы главной цепи. Результатом работы программ является целый ряд карт, гистограмм и детальных листингов с информацией по каждому остатку. Карты выдаются в формате PostScript, их можно напечатать на лазерном принтере или вывести на экран графической станции типа Sun workstation или Silicon Graphics IRIS-4D. Аналогичным образом работает программа GEOM /38/, разработанная Г.Кохеном.

6. Распределение температурного фактора и сдвига координат при уточнении.

Известно, что более подвижные области молекулы характеризуются более высоким температурным фактором, но чаще всего высокий B фактор указывает на ошибку в структуре. Поэтому одним из критериев состояния модели является распределение температурного фактора. Для остатков, расположенных на поверхности молекулы, эта величина

может быть высокой, но для внутренних частей глобулы наличие большого B может свидетельствовать об ошибке в модели.

Другим признаком, основанным на распределении B, является тот факт, что для связанных каким-либо образом атомов (ковалентной или водородной связью, н/к симметрией) соответствующие температурные факторы не должны сильно различаться.

Еще одним критерием правильности модели может служить величина сдвига координат атомов при уточнении. Опыт показывает, что правильно определенные координаты атомов модели не дают значительных сдвигов при последующем уточнении, что позволяет их использовать как один из показателей качества модели.

7. Предпочтительные конформации.

Одним из эффективных приемов является проверка, насколько хорошо параметры, задающие конформацию модели, согласуются со стандартными значениями. Для торсионных углов φ и ψ главной цепи, определяющих вторичную структуру молекулы белка, имеются предпочтительные и разрешенные области значений. Они объясняются энергетическими соображениями /14/ и представляются в виде карты Рамачандрана распределения углов φ/ψ. Одним из способов независимой проверки правильности структуры является процедура, при которой торсионные углы φ и ψ, чувствительные к ошибкам в структуре, не включаются в уточнение и используются только для оценки качества модели. Если комбинация этих углов приводит структуру к запрещенной конформации, то уже одно это позволяет сделать вывод об ошибках в модели.

Двугранные углы χ боковых цепей молекул белка также имеют предпочтительные области значений, при которых молекула принимает энергетически выгодные конформации. Исследования хорошо решенных структур, взятых из PDB, для которых в процессе уточнения не применялись ограничения на торсионные углы,

позволили установить "библиотеку ротамеров" конформаций боковых цепей молекул белка /39/, которую можно использовать для проверки правильности торсионных углов χ модели.

8. Потенциальная энергия; правильность молекулярной укладки.

В работе Д.Новотни и др. /40/ для отбора эффективных энергетических критериев проверки правильности структуры испытывались различные подходы: величина свободной энергии; электростатические потенциалы; упаковка атомов; поверхность, доступная растворителю; распределение зарядов на поверхности молекулы и т.д. Была сгенерирована модель с неправильным ходом цепи (путем встраивания α-спирального участка в β-слой). Оказалось, что если при этом провести минимизацию потенциальной энергии, то по одному этому критерию нельзя различить исходную правильную и испорченную структуры, т.е. общеупотребимый критерий потенциальной энергии является недостаточным при оценке правильности структуры. Из проведенных исследований было установлено, что удачными критериями являются неполярная поверхность боковых групп, доступная растворителю, а также количество укрытых ионизированных групп.

Другим подходом к проблеме идентификации неправильной пространственной укладки структуры является опубликованный в 1992г. в журнале "Nature" метод /41/, разработанный Д.Эйзенбергом с сотрудниками. Критерий, основанный на вычислении "3D-профилей", проверяет, совместима ли пространственная укладка данной структуры с ее первичной последовательностью. Для каждого остатка молекулы анализируется его окружение: локальная вторичная структура; область этого остатка, доступная растворителю; наличие вблизи него полярных атомов других остатков. Совокупность этих признаков позволяет приписать остатку один из классов окружения, и затем рассчитывается вероятность нахождения остатка данного типа в данном классе

окружения. При расчете вероятности используется статистическая информация, полученная по хорошо уточненным моделям из PDB. Т.о. для каждого остатка можно определить показатель, характеризующий вероятность нахождения остатка в данном окружении, и просуммировать его по всем остаткам. Общий показатель высок для структур с правильной пространственной укладкой (около 0.4) и падает ниже нуля для ошибочных структур.

Похожий подход предложен в работе Г.Враенда и др./42/, где разумность укладки белковой молекулы оценивается путем вычисления критерия, характеризующего атомные контакты. Исходя из гипотезы, что невалентные взаимодействия определяют пространственную укладку молекулы в кристалле, вычисляется величина, измеряющая степень соответствия между распределением атомов вокруг фрагментов аминокислотных остатков в модели и эквивалентным эталонным распределением, полученным статистически из PDB для известных структур, расшифрованных на высоком разрешении. Чем лучше выполняется соответствие, тем выше рассчитанный показатель качества контактов, который не зависит от экспериментальных данных. Контакты в белковых структурах анализировались во многих работах (взаимодействия между отдельными остатками; попарные межатомные взаимодействия; ориентация остатков относительно друг друга и т.д.). В данной работе одним партнером в контакте является определенный фрагмент аминокислотного остатка, другим партнером - атом определенного химического типа. Т.о. каждый случай контакта характеризуется типом фрагмента, типом атома и взаимным расположением атома и фрагмента.

Перечисленные критерии, применяемые в подходящем сочетании в зависимости от каждой конкретной ситуации, позволяют сделать достаточно объективную и надежную оценку правильности исследуемой структуры.

1.5. Использование R-free фактора в качестве объективной статистической оценки в кристаллографии.

В 1990 г. С.Бранден и Т.Джонс опубликовали в журнале "Nature" статью, озаглавленную "Between objectivity and subjectivity" /15/, в которой была продемонстрирована необходимость в повышении требований, предъявляемых к расшифрованным структурам, и в применении более надежных методов оценки результатов кристаллографических исследований макромолекул.

Обычно для оценки степени соответствия модели данным дифракционного рассеяния используется величина R-фактора (1), но как критерий качества модели он обладает следующими недостатками:

- несмотря на введение стереохимических ограничений в процессе уточнения можно получить модель с низким значением R и хорошей геометрией, но неправильную;
- можно добиться низкого значения R путем увеличения числа свободных параметров модели, но не улучшив при этом ее качества.

Типичная проблема такого сорта возникает при посадке большого количества молекул воды, чем достигается достаточно хорошая подгонка модели к экспериментальным данным за счет компенсации реальных ошибок в модели. Основная причина указанных недостатков R-фактора кроется в его тесной связи с минимизируемым при уточнении функционалом (7), т.е. R не является независимым критерием. Величину Q и, следовательно, R можно сделать как угодно малыми, необоснованно увеличивая количество параметров модели.

1.5.1. R-free как критерий оценки качества модели в процессе уточнения.

В 1992 г. Брюнгер /33/ предложил ввести новый более надежный критерий оценки точности модели - R-free фактор, рассчитанный по контрольному набору рефлексов T, которые исключаются из процесса уточнения и построения модели и используются только для контроля.

Все множество рефлексов, таким образом, разбивается на два непересекающихся подмножества: A (рабочий набор) и T (контрольный набор). Минимизируемым критерием по-прежнему остается

$$Q_A = \sum_{s \in A} (|Fobs(s)| - k|Fcalc(s)|)^2 , \qquad (15)$$

но он рассчитывается не по всему набору рефлексов, а только по рабочему набору A. При этом может возникнуть вопрос о том, достаточно ли останется данных в наборе A для проведения полноценного уточнения. Кристаллографические дифракционные данные, как правило, содержат избыточную информацию даже для макромолекул. На практике определенную долю рефлексов исключают из расчетов из-за ошибок измерения, поэтому область T можно рассматривать как такой набор, для которого дифракционные данные исключены из-за ошибок измерения. Важным является лишь алгоритм выбора рефлексов в это подмножество.

В качестве контрольного Брюнгер предлагает использовать случайный набор рефлексов в объеме 10% от полного набора. Размер контрольного набора является компромиссом между желанием свести к минимуму статистические флуктуации при расчете R-free и необходимостью избежать эффекта неполноты набора при исключении из уточнения слишком большого количества данных. Обычно способ выбора контрольного набора T не оказывает большого влияния на поведения R-free, если выборка осуществляется случайным образом и контрольный набор содержит достаточное количество элементов. Но при работе на низком разрешении поведение R-free может существенно меняться в зависимости от конкретного выбора T, поскольку в нем содержится сравнительно мало рефлексов. Ниже будет описан подход, предложенный Брюнгером, для использования R-free на низком разрешении.

На примере структур, взятых из PDB, А.Брюнгер продемонстрировал чувствительность R-free к ошибкам в модели (табл.2):

	Правильная модель	Неправильная модель
R	0.16	0.20
R-free	0.34	0.47
r.m.s. bonds	0.02Å	0.03Å
r.m.s. angles	4°	5°

Табл.2. Чувствительность R и R-free к ошибкам в модели.

Из таблицы видно, что в то время как R для правильной и неправильной модели различается на 4%, то R-free - на 13% при сравнимой геометрии.

Кроме того, была проведена серия интересных экспериментов для исследования зависимости R-free от качества уточняемой модели и оценки его корреляции с фазовой ошибкой.

1). 2365 атомов, случайно размещенных в независимой части элементарной ячейки, после уточнения рентгеновского критерия $Q_A(3)$ на разрешении 1.8Å приводят к очень низкому значению R=0.16, но R-free=0.54 и $<|\Delta\varphi|>=90°$ (близкие к параметрам случайного распределения).

2). 2365 атомов, случайно размещенных вблизи положений атомов известной структуры, после уточнения Q_A дают R=0.14, а R-free=0.43 и $<|\Delta\varphi|>=48°$. R-free "различает" распределение атомов, близкое к кристаллической структуре белка, и случайное распределение, причем R в обоих случаях можно довести до очень низких значений.

3). Если стартовать с предыдущей модели и добавить в процесс уточнения стереохимические ограничения, то R увеличивается до 0.21, но зато значительно снижаются R-free и $<|\Delta\varphi|>$ (0.27 и 37°), что соответствует общей ситуации улучшения качества модели.

4). Добавление 314 присутствующих в модели молекул воды снижает все три параметра: R=0.16, R-free=0.20 и $<|\Delta\varphi|>=33°$.

5). Если к последней модели добавить 1850 атомов кислорода, случайно размещенных в области растворителя, то после уточнения резко возрастут R-free и $<|\Delta\varphi|>$ (0.35 и 44°), зато R уменьшится до 0.13, хотя качество модели при этом ухудшается.

На протяжении всех тестов коэффициент корреляции между R-free и $<|\Delta\varphi|>$ оставался около 0.98.

Т.о. R-free представляет собой надежный и объективный критерий оценки качества модели, полученной методами РСА. Область его применения не ограничивается только данными высокого разрешения: тесты, проводившиеся как на разрешении 1.8Å, так и 2.8Å, дают большую корреляцию между R-free и $<|\Delta\varphi|>$. Тот факт, что R-free различает случайный набор структурных факторов и набор данных, характерных для рассеивания молекулой белка, дает возможность использовать R-free фактор для определения фаз *ab initio*. Рост R-free при моделировании области растворителя атомами с фиксированными координатами подтверждает неупорядоченный характер растворителя.

Предложенная Брюнгером идея применения R-free фактора при уточнении модели исследуемой структуры в последнее время получила широкое распространение в среде кристаллографов, заинтересованных в применении наиболее объективных критериев проверки качества моделей структур, объявленных расшифрованными. Только в одном номере журнала "Acta Crystallographica" (D51) за 1995 г. было опубликовано 4 работы по расшифровке белковых структур с использованием R-free фактора в процессе уточнения /43-45/. При этом критерий оказался полезным как при работе на среднем разрешении 2.5Å /44/, так и на высоком - 1.9Å /47/ и 1.2Å /48/.

1.5.2. Улучшение фаз методом модификации плотности с использованием "complete cross-validation".

Идею деления всех данных на контрольный и рабочий набор (cross validation), на которой основано использование R-free, можно приме-

нять не только к вычислению R-фактора при уточнении модели исследуемой структуры, но и к любым другим статистическим критериям, описывающим соответствие модели экспериментальным данным. Например, при вычислении figures of merit или коэффициента корреляции в прямом или обратном пространстве, или в ситуации, когда атомную модель еще невозможно построить и работа происходит с электронной плотностью. В этом случае при использовании метода модификации электронной плотности для улучшения и расширения набора фаз также можно применить "cross validation", чтобы оценить, улучшает ли используемая процедура фазовую ошибку. Из-за неизбежных ошибок в стартовых фазах необходимым этапом решения структуры является уточнение и расширение набора фаз до более высокого разрешения. Одним из основных приемов улучшения фаз является использование различных схем модификации синтеза электронной плотности, которые основаны на введении физических или химических ограничений типа малой контрастности электронной плотности в области растворителя (solvent flatness), связности электронной плотности молекулы, атомности структуры, предсказание распределения значений функции электронной плотности на основании гистограмм, использование н/к симметрии в прямом пространстве и т.д.

Общепринятой процедурой при модификации электронной плотности является метод выравнивания плотности в области растворителя (solvent flattening), который часто используется в комбинации с молекулярным усреднением /49/, применением эталонных гистограмм /50-54/, уравнений Сейра /55,56/, методами максимума энтропии /57/, срезки отрицательной плотности /58/ и т.д. Solvent flattening является итерационной процедурой модификации электронной плотности в реальном пространстве и комбинирования фаз в обратном пространстве. Электронная плотность в области растворителя заменяется, например, на ее усредненную величину, что приводит к ее выравниванию в этой области. Фазы, рассчитанные

путем обратного преобразования Фурье по модифицированному синтезу, комбинируются со стартовой фазовой информацией для получения более точных фазовых оценок. Общая схема процедуры модификации электронной плотности изображена на рис.1.

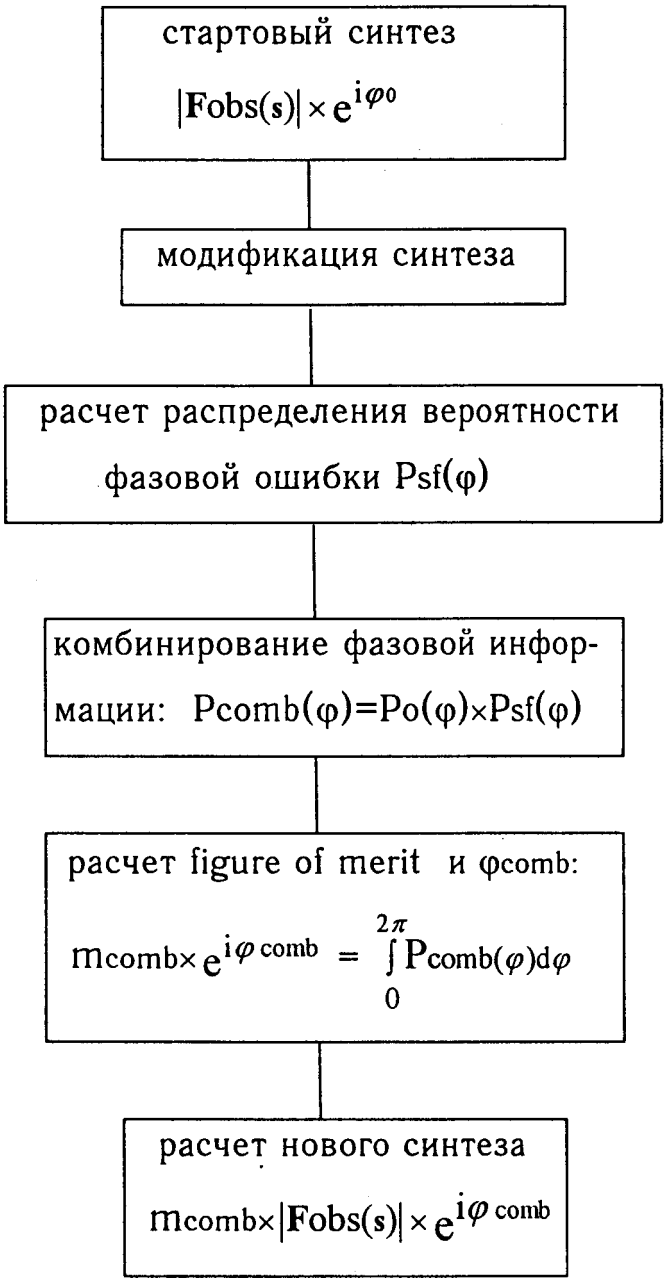

Рис.1. Схема процедуры модификации электронной плотности.

Однако в этой процедуре сложно подобрать объективный критерий оценки фазовой ошибки, т.е. оценки того, улучшаются ли фазы в процессе работы. Часто результативность метода оценивается визуаль-но

путем просмотра карт электронной плотности, что может быть сделано довольно субъективно. Либо используется стандартный R-фактор

$$R = \frac{\sum_{s} m_{comb} \left| |Fobs(s)| - k|Fcalc(s)| \right|}{\sum_{s} m_{comb} |Fobs(s)|} . \qquad (16)$$

А.Брюнгер в своих работах 1993-1995 г. /34,59-61/ предлагает в качестве критерия эффективности процедуры модификации электронной плотности использовать R-free фактор, основываясь на высокой корреляции R-free с фазовой ошибкой. Применение R-free в этом случае состоит в выделении некоторого подмножества рефлексов (контрольного набора) T из всего набора данных, модификации плотности, рассчитанной по оставшемуся набору A, вычислению нового набора структурных факторов и расчету R-free по контрольной группе рефлексов T.

В кристаллографическом уточнении низкоугловые рефлексы обычно отсутствуют, но их наличие является жизненно важным при модификации электронной плотности, поскольку они определяют форму оболочки молекулы и связность областей электронной плотности. Но при работе на низком разрешении поведение R-free существенно зависит от конкретного выбора T, поскольку в нем содержится сравнительно мало рефлексов. Чтобы избавиться от флуктуаций R-free на низком разрешении, Брюнгер предложил использовать метод "complete cross-validation" /59-61/. При этом полный набор данных разбивается на n непересекающихся подмножеств (T_1, ..., T_n), где T_i - контрольный набор. Каждому T_i соответствует рабочий набор рефлексов A_i такой, что в сумме с T_i они дают полный набор рефлексов. Процедура модификации плотности выполняется n раз для каждого набора A_i, т.е. при вычислении синтеза используются только рефлексы с индексами, входящими в подмножество A_i. Но структурные факторы и показатели достоверности вычисляются по модифицированному синтезу и сохраняются как

выходные данные только по контрольным наборам Ti. После завершения процедуры в сумме получаются значения всех структурных факторов, и по ним рассчитывается "complete cross-validation" фактор

$$R_{comp} = \frac{\sum_s \left| |Fobs(s)| - k|\overline{F}calc(s)| \right|}{\sum_s |Fobs(s)|}, \qquad (17)$$

который более адекватно оценивает точность полученных фаз.

В работе К.Ковтана /62/ приводится описание комплекса программ, встроенного в библиотеку CCP4 /63/, который представляет собой автоматическую процедуру для использования всех перечисленных выше методов модификации электронной плотности в любой их комбинации. Как пишет автор, наиболее полезным и надежным индикатором результативности каждой процедуры является R-free фактор.

И еще одно применение cross-validation методики следует здесь отметить. При построении модифицированного синтеза используется показатель достоверности m_{comb}, для определения которого нужно уметь вычислять функцию распределения вероятности значений фаз структурных факторов $P(\varphi)$. Это распределение содержит параметры, отражающие уровень фазовой ошибки. В работах В.Лунина и др. /64,65/ для получения адекватных оценок фазовых ошибок был предложен метод, основанный на максимизации функции правдоподобия для нахождения параметров распределения вероятности и применении R-free методологии при ее вычислении.

Таким образом, R-free ("cross validation") является важным понятием в кристаллографии макромолекул, которое позволяет ввести более объективный критерий корректности и точности модельного представления кристаллической структуры.

1.6. Разнообразие схем и программ уточнения.

Уточнение атомных моделей давно уже стало стандартной частью кристаллографического исследования макромолекул, хотя при этом часто возникает целый ряд сложностей: низкое соотношение количества наблюдаемых и уточняемых параметров, большой объем вычислений, высокие требования к стартовой модели, от которой зависит сходимость процесса уточнения, и т.п. Большие объемы данных и комплексный характер используемых ограничений требуют значительных компьютерных ресурсов при уточнении макромолекул, которое отнимает большую часть общего компьютерного времени, необходимого для расшифровки структуры. Поэтому программы уточнения, которые применяют к большим объектам, должны быть эффективными и удобными в использовании, что приводит к их довольно сложным алгоритмам и архитектуре.

При уточнении можно применить целый ряд различных подходов и опций, и разнообразные их комбинации использовались в тех или иных пакетах программ. Наиболее важные из них следующие: прямое или обратное пространство; Фурье-методы, МНК или другие процедуры оптимизации; свободные атомы либо модель с "constraints" или "restraints" ограничениями; одновременное использование рентгеновского и стереохимического критериев либо периодическое восстановление "идеальной" геометрии; общий температурный фактор, групповые, индивидуальные изотропные или анизотропные тепловые параметры и т.д. Например, процедура Р.Даймонда /66/ уточнения в реальном пространстве использует "constrained" геометрию, основанную на конформационных углах. А.Джек & М.Левитт /6/ применяют энергетическое уточнение с "restrained" геометрией, чтобы избежать близких невалентных контактов. Ж.Суссман и др. /4/ ввели жесткие группы, чтобы использовать ограничения типа "constraints" в МНК в обратном пространстве. Р.Агарвал & Н.Айзекс /7,67/ разработали метод вычисления производных и структурных факторов с помощью

преобразования Фурье и после серии циклов свободного уточнения периодически осуществляли восстановление стереохимии. В.Хендриксон & Д.Коннерт /1-3/ использовали стереохимические ограничения в качестве "restraints" при уточнении МНК в обратном пространстве. А.Брюнгер /8-10/ ввел в процесс уточнения применение метода молекулярной динамики, и его программа X-PLOR /32/ сейчас одна из наиболее популярных среди кристаллографов.

Анализ некоторых методов и программ уточнения дается в диссертации Б.В.Строкопытова /68/. С тех пор появилось несколько новых подходов и программных разработок для уточнения структур макромолекул, и ниже приводится их краткий обзор и сравнение.

1. Программа Р.Даймонда уточнения в прямом пространстве.

В 1966 г. Р.Даймондом /69/ была создана первая программа уточнения, предназначенная для построения моделей белков, и на ее основе в 1971 г. была создана вторая версия, позволяющая проводить кристаллографическое уточнение атомных моделей /66/. Целью метода является нахождение наилучшего соответствия модели белка и электронной плотности. Уточнение модели происходит в прямом пространстве и осуществляется путем минимизации критерия

$$R = \int_V (\rho_0 - \rho_m)^2 dV , \qquad (18)$$

где ρ_o - наблюдаемая электронная плотность, ρ_m - плотность, вычисляемая по модели.

Первоначально стартовую модель строят, приблизительно вписывая атомы в пики электронной плотности и собирая из аминокислотных остатков цепь с жесткими длинами связей и валентными углами (кроме угла при Cα-атоме и иногда при Cβ-атоме). Допускается свободное вращение вокруг всех одинарных связей. Полипептидная цепь максимально приближается к реперным точкам (приблизительным положениям отдельных атомов) путем варьирования двугранных углов

в главной цепи и боковых группах. Программа построения модели последовательно добавляет одну боковую группу и одно звено главной цепи к уже построенной структуре, с учетом двух предыдущих аминокислотных остатков, подгоняя до тех пор, пока не достигается хорошее соответствие для всех трех остатков.

При уточнении минимизируется объемный интеграл (18) при варьировании координат модели таким образом, чтобы сохранялась геометрия полипептида, а менялись двугранные и некоторые валентные углы. Основная сложность такой процедуры в том, что изменение одного двугранного угла означает поворот всей полипептидной цепи. Эту трудность можно преодолеть, вводя "зону расплава", содержащую 5-10 аминокислотных остатков. Меняются параметры атомов в пределах этой зоны, а положения атомов на концах - фиксированы, чтобы сохранить непрерывность цепи. Затем "зону" перемещают по цепи на один остаток и подгоняют следующие участки.

Первоначально метод Даймонда предполагал, что и модули, и фазы структурных факторов являются наблюдаемыми величинами (взятыми, например, из изоморфного замещения). Ж.Дайзенхофер & В.Штейгеман /70/ первыми предложили использовать фазы, рассчитанные по модели. Модельные фазы пересчитывались после каждого цикла уточнения, а уточнение проводилось по картам распределения электронной плотности, рассчитанным с коэффициентами Фурье вида

$$[nFo - (n-1)Fc] \exp(i\varphi c). \qquad (19)$$

Основное достоинство программы Даймонда - возможность улучшения модели путем использования карт электронной плотности, построенных по изоморфным фазам, когда можно получить оптимальную интерпретацию карты до замены изоморфных фаз на расчетные. Но этот метод можно применять только для карт высокого разрешения, иначе из-за перекрывания электронных облаков различных боковых

цепей боковые группы могут попасть в одну и ту же электронную плотность. А для моделей больших белков, как правило, удается получить набор изоморфных фаз на среднем или низком разрешении, что существенно ограничивает область применимости этого метода.

Кроме того, чрезмерная жесткость модели не всегда позволяет атомам занять оптимальные положения, а также приводит к накоплению стереохимических нарушений в варьируемых валентных углах. К указанным недостаткам следует добавить также сложность модифицирования или исключения остатка из последовательности. Все это привело к тому, что в настоящее время программа Даймонда RLSP не используется для уточнения макромолекул, хотя с вычислительной точки зрения уточнение в прямом пространстве выгоднее.

2.Использование стереохимической информации при уточнении. Программа PROLSQ В.Хендриксона & Д.Коннерта.

Программа В.Хендриксона & Д.Коннерта PROLSQ основана на использовании при уточнении кроме стантартного рентгеновского критерия еще и стереохимических ограничений типа “restraints”. Все слагаемые минимизируемой функции имеют квадратичный вид, что упрощает расчет и позволяет применить стандартный МНК.

Минимизируется функция вида

$$Q = \sum_H W_H \left(|F_H^o| - |F_H^c|\right)^2 + \sum_l W_l \left(d_l^o - d_l^c\right)^2 , \qquad (20)$$

где d_l^o - идеальные значения стереохимических параметров модели;

d_l^c - соответствующие параметры, вычисленные по модели;

$W_l = \frac{1}{\sigma_l^2}$ (σ - стандартное отклонение).

<u>Стереохимическими параметрами</u> являются длины валентных связей; валентные углы (попарные расстояния между тройками атомов, образующих валентный угол); торсионные углы (расстояние между 1-м и 4-м атомом в четверке атомов, образующих двугранный угол);

плоские группы атомов; хиральные объемы (тройное скалярное произведение векторов, направленных от центрального атома к трем связанным с ним атомам); индивидуальные изотропные температурные факторы; коэффициенты заполнения; водородные связи; положения атомов, связанных локальной симметрией; отталкивающий потенциал для невалентно связанных атомов.

В программе применяется разработанный авторами эффективный алгоритм к составлению элементов матрицы нормальных уравнений: в расчетах используются только диагональные элементы и элементы, входящие в перечисленные выше геометрические ограничения. При таком подходе вместо полноматричного уточнения происходит работа с сильно разреженной матрицей (по оценкам Хендриксона и Коннерта, остается менее 1% всего набора элементов норамальной матрицы), и для решения такой системы наиболее эффективным является метод сопряженных градиентов, важнейшей особенностью которого является сохранение без изменения матрицы А нормальных уравнений в процессе вычислений. Это позволяет хранить в одномерном массиве только ее ненулевые элементы, а при вычислении использовать систему ссылок на эти элементы.

Т.о., метод уточнения Хендриксона и Коннерта характеризуется следующими свойствами: это уточнение в обратном пространстве на основе МНК с одновременным сохранением идеальной геометрии структуры моделей белков (а не периодическим ее восстановлением после серии циклов уточнения), где геометрическая информация вводится в виде мягких ограничений типа “restraints”. Программу уточнения PROLSQ активно используют в настоящее время, получая модели со стереохимическими параметрами высокого качества.

Е.Вестхоф и др. /71/ на основе программы PROLSQ /1-3/ разработали свою версию программы NUCLSQ специально для уточнения нуклеиновых кислот. В ней аналогичным образом устанавливаются ограничения на длины валентных связей, углы,

плоские группы, хиральность атомов сахара, отталкивающие ограничения для невалентных контактов, длины водородных связей, конформационные торсионные углы, псевдоротационные параметры.

3. Применение "constraints" и "restraints" ограничений в программе CORELS Ж.Суссмана.

Программа CORELS Ж.Суссмана /4,5/ уточнения белков и нуклеиновых кислот использует оба типа ограничений (как "constraints", так и "restraints") для снижения числа уточняемых параметров и повышения сходимости процедуры уточнения. Минимизируемую функцию можно разбить на три квадратичных члена:

$$Q = \sum_H W_H \left(|F_H^o| - |F_H^c|\right)^2 + \sum_l W_l \left(d_l^o - d_l^c\right)^2 + \sum_i W_i \sum_j \left(X_{T,i,j} - X_{i,j}\right)^2 \qquad (21)$$

Первый член представляет собой рентгеновскую часть минимизируемого функционала; второй соответствует геометрическим ограничениям, накладываемым на модель (валентные связи, валентные и торсионные углы, отталкивающие потенциалы); третий член отвечает за рассогласование между координатами текущей модели $X_{i,j}$ и реперными координатами $X_{T,i,j}$.

Атомную структуру в программе можно представить в виде набора определенных жестких групп, которые связаны между собой специальными геометрическими соотношениями. Поэтому в общем виде критерий Q является функцией групповых позиционных и тепловых параметров Q(**T**i, **R**i, ψi1, ψi2, ..., Bi1, Bi2,...,k), где **T**i и **R**i - вектор трансляции и вращения i-ой жесткой группы; ψij - двугранные углы внутри i-ой группы, если они есть; Bij - соответствующие групповые тепловые параметры; k - шкальный коэффициент, связывающий Fo и Fc. Программа использует сильно разреженную матрицу нормальных уравнений и итерационный метод сопряженных градиентов, что позволяет сократить размер хранимой и используемой при вычислениях информации.

Программа CORELS эффективно регулирует соотношение между числом экспериментальных данных и параметров уточнения: возможность зафиксировать в качестве жестких групп атомы боковых и главных цепей, отдельные домены, всю молекулу в целом позволяет значительно повысить это соотношение по сравнению с программами, использующими только "restraints" ограничения. Она сохраняет большую часть геометрических параметров модели идеальными (т.к. внутри жесткой группы все валентные связи и углы фиксированы). Основным недостатком является накопление нарушений в тех стереохимических параметрах, которые варьируются в процессе уточнения (двугранные углы, валентные связи и углы между группами). Кроме того, использование прямых методов расчета структурных факторов и их производных требует значительного времени, что становится особенно критичным при уточнении моделей макромолекул.

4. Программа RESTRAIN Д.Мосса.

В разработанной Д.Моссом с коллегами /72-74/ программе RESRTAIN уточнения белков и нуклеиновых кислот используется похожий на CORELS подход к применению "constraints"/"restraints" ограничений с некоторыми дополнительными возможностями. Уточняется рентгеновский критерий для модулей и фаз структурных факторов совместно с псевдоэнергетическим критерием ("restraints") и использованием жестких групп ("constraints"). Жесткими могут быть объявлены определенные группы атомов (боковые цепи Phe, Pro, Tyr, Trp, Val и т.п. для белков; основания, фосфатные группы, рибозы для нуклеиновых кислот), которым приписывается фиксированная локальная геометрия. В общем виде минимизируемый критерий выглядит как

$$Q = \sum_H W_H \left(|F_H^o| - |F_H^c|\right)^2 + \sum_P W_P (\varphi_o - \varphi_c)^2 + \sum_l W_l (d_l^o - d_l^c)^2 , \qquad (22)$$

где первый член суммы - обычный рентгеновский критерий для модулей структурных факторов; второй член отвечает за соответствие фаз, вычисленных по модели, набору предписанных значений фаз

(например, из изоморфного замещения и/или аномального рассеяния); третий член связывает идеальные геометрические параметры и параметры, вычисленные по модели (валентные связи и углы, торсионные углы, хиральность, планарность, отталкивающие потенциалы).

Переменными минимизации являются атомные координаты; общий шкальный коэффициент, температурный фактор B и коэффициент заполнения C; параметры жестких групп; уравнения локальной симметрии; индивидуальные изотропные и анизотропные B и C; групповые анизотропные тепловые параметры (TLS тензор) и коэффициенты заполнения. Параметрами жестких групп являются вектор сдвига и, в отличие от общеупотребимых в качестве параметров вращения углов Эйлера, более наглядные переменные, связывающие угол поворота группы с направляющими косинусами оси вращения.

Т.о. программа RESRTAIN позволяет использовать как "constraints", так и "restraints" ограничения, что дает ей соответствующие преимущества перед программами, имеющими только один из типов ограничений. Кроме того, применение индивидуальных анизотропных тепловых параметров и TLS тензоров для жестких групп дает возможность проводить уточнение с данными высокого разрешения.

5. Программа SFRF Агарвала & Айзекса.

Большой вклад в разработку и развитие математической и вычислительной стороны процедуры уточнения внесли Р.Агарвал и Н.Айзекс /7,67/, впервые применив алгоритм быстрого преобразования Фурье /75-77/ для расчета структурных факторов и их производных, который существенно увеличивает скорость расчетов. Минимизируемой функцией в их программе SFRF является стандартный рентгеновский критерий

$$Q = \sum_{H} W_H \left(\left|F_H^o\right| - \left|F_H^c\right|\right)^2 \qquad (23)$$

Варьируемые параметры - координаты атомов, индивидуальные изотропные температурные факторы и коэффициенты заполнения.

В настоящее время программа SFRF не используется при уточнении макромолекул, т.к., хотя высокая скорость расчетов является ее несомненным достоинством, но отсутствие применения какой-либо дополнительной информации не позволяет провести эффективного уточнения моделей. Исключение из процесса уточнения геометрических ограничений приводит к необходимости чередовать минимизацию кристаллографического критерия с восстановлением стереохимии, что сильно замедляет в целом процесс расшифровки структуры. Тем не менее идеи и алгоритмы Агарвала и Айзекса послужили хорошей основой для разработки новых программ, использующих быстрое преобразование Фурье.

6. Применение энергетического критерия. Программа EREF Джека & Левитта.

А.Джек и М.Левитт /6/ в программе EREF использовали энергетическое уточнение с "restraints" геометрией. Характерной особенностью этой программы является сочетание достоинств применения быстрого преобразования Фурье для расчета структурных факторов и их производных и подключение к уточнению функции полной потенциальной энергии, включающей Ван-дер-Вальсовые взаимодействия. Минимизируемый функционал имеет вид

$$Q = E + kR \quad , \qquad (24)$$

где R - обычный кристаллографический критерий, E - потенциальная энергия молекулы:

$$E = \sum k_b (b_i - b_o)^2 + \qquad \text{(валентные связи)}$$

$$\sum k_\tau (\tau_i - \tau_o) + \qquad \text{(валентные углы)}$$

$$\sum k_\theta (1+\cos(m\theta+\sigma)) + \qquad \text{(торсионные углы)}$$

$$\Sigma\left(\frac{A}{r_{ij}^{12}}+\frac{B}{r_{ij}^{6}}\right)$$ (невалентные взаимодействия),

где k_b, k_τ, k_θ - силовые константы,

m, σ - параметры,характеризующие энергию связей торсионных углов,

A, B - константы Леннарда-Джонса,

b_o, τ_o - стандартные значения валентных связей и углов,

b_i, τ_i, θ - вычисленные по модели величины валентных связей, валентных и торсионных углов.

Силовые константы для стереохимических параметров получены из колебательных спектров малых молекул /78/ и скорректированы с учетом отсутствия в модели атомов водорода. Коэффициенты A и B в потенциале Леннарда-Джонса выбраны эмпирическим путем из условия минимума энергии невалентного взаимодействия. Функция (24) минимизируется методом сопряженных градиентов с диагональным приближением матрицы нормальных уравнений. Уточняемыми параметрами являются координаты атомов, индивидуальные тепловые факторы и коэффициенты заполнения.

Главным достоинством программы EREF является использование физически разумных энергетических параметров, в том числе учет Ван-дер-Вальсовых взаимодействий, заложенных в программу и зависящих от типа взаимодействующих атомов. В рамках программы легко восстанавливаются нарушения геометрии модели и положения невалентно связанных атомов, и она активно используется в настоящее время при уточнении макромолекул.

7. Пакет программа TNT. Универсальность его архитектурной структуры.

Пакет программ TNT для уточнения МНК с использованием стереохимических ограничений и возможностью фиксации жестких групп, разработанный Д.Тронрудом с сотр. /79,80/, использует алгоритм

быстрого преобразования Фурье во всех кристаллографических преобразованиях. Специальные усилия разработчиков пакета TNT были направлены на то, чтобы создать как можно более универсальный комплекс программ, гибкий и эффективный при его использовании. Например, стереохимические ограничения задаются таким образом, чтобы их можно было применять к любым структурам - белкам, нуклеиновым кислотам, молекулам растворителя, любым комплексам и образованиям. Программа позволяет легко изменять идеальные геометрические параметры или вводить их для новых или нестандартных химических групп. Геометрическая информация внутри программы задается в общем виде, а во внешнем файле описываются конкретные компоненты (аминокислотные остатки, нуклеотиды, кофакторы), и для каждого из них определяются геометрические соотношения.

Одним из существенных ограничений многих программ уточнения является их негибкость: нельзя ни изменить, ни добавить новые функции в минимизируемый критерий без глобальной модификации всей программы. Это ограничивает возможности пользователя экспериментировать с различными стратегиями уточнения, весовыми схемами, критериями и т.п. Пакет TNT создан таким образом, чтобы снять по возможности эти ограничения. Процесс уточнения разбит на нескольлко шагов, каждую задачу выполняют отдельные программы, а функциональные единицы связаны между собой посредством файлов. Чтобы заменить или модифицировать любую программу, не требуется больших усилий. Например, чтобы оптимальным образом реализовать зависящий от пространственной группы алгоритм (типа быстрого преобразования Фурье), нужно внести минимальные изменения в определенную часть соответствующей программы.

Архитектура пакета программ TNT позволяет оптимально реализовать каждый блок операций, используя по мере их возникновения новые более эффективные алгоритмы, внося усовершенствования в отдельные части комплекса и не меняя всего пакета в целом.

8. Программа X-PLOR А.Брюнгера.

Одним из наиболее часто используемых в настоящее время комплексов программ уточнения макромолекул является пакет программ X-PLOR, разработанный А.Брюнгером /32/. В этом пакете используется новый способ кристаллографического уточнения - метод моделированного отжига ("simulated annealing") /8-11/, основанного на применении молекулярной динамики.

Обычное уточнение на основе МНК имеет ограниченный радиус сходимости. Как правило, этот метод не может восстановить положение остатка, которое отличается более чем на 1Å от правильного, при том что точность стартовой атомной модели бывает невысокой из-за ошибок при ее построении и недостатка фазовой информации. Кроме того, минимизация МНК легко может привести в локальный минимум, после чего требуется ручная правка модели. Все это в совокупности приводит к большим временным затратам. "Simulated annealing" (SA) метод, основанный на использовании молекулярной динамики, помогает преодолеть проблему локальных минимумов, значительно увеличить радиус сходимости, ускорить процесс уточнения и уменьшить необходимость в ручной перестройке модели.

Моделирование методом молекулярной динамики предполагает решение классических уравнений движения Ньютона в силовом поле, которое описывается эмпирической потенциальной энергией стереохимических, электростатических и невалентных взаимодействий в молекуле. Брюнгер, кроме того, вводит в расчет полной энергии стандартный рентгеновский критерий, описывающий расхождение экспериментальных и рассчитанных по модели модулей структурных факторов:

$$E_{pot} = E_{chem} + E_{X\text{-}ray} \qquad (25)$$

где E_{chem} включает в себя энергию валентных (валентные связи, углы, торсионные углы, хиральные центры, планарность) и невалентных взаимодействий (Ван-дер-Вальсовая функция, электростатический по-

тенциал, кристаллографические и н/к взаимодействия). Кинетическая энергия обеспечивается молекулярно-динамическим моделированием движения системы атомов.

Для замкнутой системы, состоящей из N атомов массой mi и координатами $\{ \mathbf{r}_i \}_{i=1}^{N}$, ее движение в силовом поле описывается системой уравнений Ньютона:

$$m_i \frac{\partial^2 r_i}{\partial t^2} = - \nabla_i E_{pot} . \qquad (26)$$

Каждому атому, исходя из распределения Максвелла, задается начальная скорость, соответствующая стартовой температуре. Система уравнений решается численными методами с шагом по t, и в результате получается траектория движения системы. Контроль за температурой в процессе расчетов молекулярной динамики осуществляется путем периодического изменения атомных скоростей:

$$V_i^{new} = Sc \times V_i^{old} , \qquad (27)$$

где шкальный коэффициент Sc задается как

$$Sc = \left\langle \sum_i m V_i^{old}(t)^2 \right\rangle / NkT. \qquad (28)$$

Здесь суммирование производится по всем атомам системы, k - постоянная Больцмана, T - температура, скобки $\langle \ \rangle$ означают усреднение по интервалу времени между моментами перешкалирования скоростей.

Первоначально устанавливается очень высокий температурный режим, что обеспечивает большую стартовую кинетическую энергию и подвижность системы, а затем постепенно температура понижается. При этом в процессе охлаждения система попадает в локальный минимум, и процедуру можно повторить снова. Этот алгоритм позволяет преодолевать высокий энергетический барьер при попадании системы в локальный минимум и значительно увеличивает радиус сходимости. Успех процедуры зависит от температурного режима,

относительного взвешивания всех компонент потенциальной энергии, скорости охлаждения системы и т.д.

Общая стратегия уточнения с использованием SA-процедуры состоит в следующем: на первом этапе производится минимизация функционала E_{pot} методом сопряженных градиентов, чтобы нормализовать структуру с точки зрения стереометрии и избавиться от недопустимо близких невалентных контактов. Затем каждому атому приписывается начальная скорость, соответствующая заданной температуре. Система уравнений, описывающая движение N атомов, решается численными методами, используя шаг по t при высокой температуре ($\Delta t \sim 0.25$fc при 3000-9000 K) и получая траекторию движения. Каждые 25-50 шагов проверяется температура системы, и, если необходимо, скорости движения атомов шкалируются, чтобы поддерживать температуру на заданном уровне. Для получения существенной перестройки структуры достаточно небольшой траектории (2-3 ps). Затем система постепенно охлаждается до ~ 300 K и остальные 0.5 - 1 ps происходит движение при этой температуре. В конечном счете скорости движения атомов становятся равными нулю, и на последнем этапе уточнения опять осуществляется процедура минимизации методом сопряженных градиентов функционала E_{pot} для оптимизации структуры и восстановления нарушенной стереохимии, которая портится при высокотемпературном движении системы.

Следует подчеркнуть, что термин "температура" при использовании процедуры SA-уточнения и движения системы атомов в методе молекулярной динамики не имеет физического смысла, т.к. общая энергетическая функция представляет собой гибрид, состоящий из эмпирической потенциальной энергии и гипотетического члена, заданного кристаллографическим критерием. Температура является только параметром, характеризующим высоту локального энергетического барьера, который может быть преодолен в процессе SA-уточнения.

Кроме нового алгоритма уточнения программа Брюнгера имеет целый ряд таких особенностей, как использование фазового критерия; уточнение молекулы как твердого тела; возможность фиксировать отдельные атомы, валентные связи и углы; применение быстрого преобразования Фурье. Имеется возможность провести тщательный контроль и проверку полученной модели: r.m.s. геометрия, R-free фактор, карты Рамачандрана, невалентные контакты, кристаллическая упаковка, водородные связи, доступная растворителю область поверхности и т.д.

Хотя метод SA-уточнения и не является полностью автоматической процедурой, но он представляет собой мощное средство из арсенала макромолекулярной кристаллографии, демонстрирует гораздо больший радиус сходимости, чем процедуры, основанные на МНК, повышает интерпретируемость карт электронной плотности и существенно облегчает и уменьшает необходимый объем ручной правки модели. Метод SA-уточнения в последние года успешно применялся для расшифровки целого ряда структур: aldose reductase /81/, hemichrome hemoglobin /82/, hexagonal turkey egg-white lysozime /83/, Cu-substituted alcohol dehydrogenase /84/, troponin C at 1.78Å /85/, myosin subfragment-1 /86/ и т.д.

Программа XPLOR рассчитана на ее использование на мощных компьютерах типа VAX, CRAY, CONVEX в VMS и UNIX операционных системах. XPLOR, как и программы PROLSQ, TNT, CORELS, EREF, входит в библиотеку программ макромолекулярной кристаллографии CCP4 (Collaborative Computing Project number 4), организованную в 1979 г. и функционирующую под эгидой Европейской ассоциации кристаллографии биологических макромолекул /63/.

9. Программа SHELX Г.Шелдрика.

Первоначально программа SHELX /87-90/ была создана для расшифровки и уточнения низкомолекулярных структур, но в связи с

появлением все более мощных компьютеров, усовершенствованием техники съемки и получением данных высокого разрешения для макромолекул ее можно использовать и для уточнения небольших белков и олигонуклеотидов. Программа применяет прямой расчет структурных факторов, не обращаясь к суммированию на основе алгоритма быстрого преобразования Фурье, и поэтому работает гораздо медленнее, чем специально разработанные для уточнения макромолекул программы. Это является платой за более высокую точность и универсальность, но в некоторой степени компенсируется лучшей сходимостью, что в конечном счете уменьшает объем необходимого ручного вмешательства в процесс построения модели. Большим достоинством программы является простота ее использования, понятный интерфейс и возможность применения практически на любых типах компьютеров, в том числе на IBM PC.

Уточнение всегда проводится с величинами F^2, при этом в процедуру можно ввести Rfree фактор (п.1.5), причем используется полный набор данных, что существенно улучшает условия минимизации при работе на высоком разрешении.

SHELX дает возможность установить ограничения типа “restraints” для длин валентных связей и углов, планарности и хиральных объемов. Торсионные углы не входят в этот список, но их можно использовать для проверки качества уточняемой структуры, поскольку они никак не ограничены в процессе уточнения. Можно также ввести расталкивающие ограничения для невалентно связанных атомов (задается пороговая величина, до которой разрешается сближение атомов); жестко фиксировать позиционные и температурные параметры отдельных атомов и групп атомов; осуществлять процедуру посадки молекул воды. Тепловое движение атомов контролируется путем введения как изотропных, так и анизотропных температурных факторов, причем имеется возможность ввода таких ограничений, при

которых соответствующие компоненты близки для атомов, валентно связанных или пространственно близких.

Алгоритм минимизации методом сопряженных градиентов, используемый в SHELX, основан на процедуре, описанной Хендриксоном & Коннертом /3/. Кроме того, можно использовать полноматричное уточнение.

В работе /90/ описывается успешное применение программы SHELX для расшифровки прямыми методами и уточнения структуры небольшого белка цитохрома c6 (89 остатков) на разрешении 1.1Å, а в работе /91/ - первое успешное полноматричное уточнение МНК белковой структуры (крамбин, 46 остатков, разрешение 0.83Å, R=9.0%) с помощью программы SHELX.

10. ARP - программа автоматическое уточнение белковых структур.

Разработанная в 1993 г. В.Ламзиным и К.Вильсоном программа ARP /92/ является попыткой создания средства, позволяющего осуществлять автоматическое уточнение модели молекулы белка, не прибегая к правке модели с помощью интерактивной машинной графики, которая является стандартным этапом процедуры уточнения. В результате процесс уточнения исследуемой структуры в целом значительно сокращается. Естественно, что при этом область применения программы сильно ограничена: требуется наличие данных хорошего качества и высокого разрешения (> 2Å), а стартовая модель должна содержать не менее 75% атомов в приблизительно правильных позициях.

Как известно, эффективность любой процедуры уточнения зависит от близости стартовой модели к идеальной и от радиуса сходимости данного метода. Стандартные алгоритмы минимизации МНК имеют небольшой радиус сходимости (d/3 - d/4) и не могут автоматически осуществить большие сдвиги в координатах неправильно посаженных атомов. Метод “simulated annealing” А.Брюнгера /8-11/ путем приме-

нения молекулярной динамики значительно увеличивает радиус сходимости при минимизации. Тем не менее и у этого метода ограниченные возможности для автоматической корректировки неправильно построенных фрагментов или пространственной укладки структуры, и, кроме того, он требует большого объема вычислительных ресурсов.

В программе ARP делается попытка осуществления процедуры одновременного уточнения и перестройки модели по разностным картам, не прибегая к ручному этапу интерактивной правки, т.е. попытка реализовать полностью автоматическую процедуру уточнения модели. Для этого требуется, чтобы стартовая модель и экспериментальные данные воспроизводили синтез, позволяющий различить на нем положения отдельных атомов.

Схематично алгоритм уточнения можно представить следующим образом. Сначала все атомы исходной модели преобразуются в атомы одного типа (например, им приписывается одинаковое количество электронов и один температурный фактор). Затем производится стандартная минимизация МНК рентгеновского критерия без каких-либо ограничений по полному набору экспериментальных данных, где параметрами уточнения являются атомные координаты и B-факторы. После каждого цикла уточнения либо через несколько циклов происходит автоматическая перестройка модели следующим образом. На основе карт разностного синтеза электронной плотности небольшая часть атомов (0.1 - 1%, если процедура осуществляется на каждом цикле уточнения, или 1 - 10% после нескольких циклов), оказавшаяся в плотности с наиболее низкими значениями, удаляется из модели, а в наиболее высокие положительные пики добавляются новые атомы с учетом близости к уже имеющимся атомам модели. Эти шаги повторяются до достижения сходимости алгоритма (критерием сходимости может служить кристаллографический R-фактор и величина абсолютного значения плотности на разностной карте).

Уточнение модели, расчет карт электронной плотности и быстрого преобразования Фурье производится при помощи программ из библиотеки CCP4 /63/. Все остальные действия - анализ карт электронной плотности, удаление атомов, расположенных в плотности с низким уровнем, посадка новых атомов в положительные пики с анализом окружающих пик атомов и переименование атомов осуществляет программа ARP. В работе /92/ приводится несколько успешных примеров применения описанной процедуры к реальным структурам.

1.7. Краткий обзор графических программ, применяемых в рентгеновской кристаллографии.

Интерпретация карт электронной плотности для получения начальной модели по-прежнему остается узким местом в процедуре кристаллографической расшифровки структуры. На этом этапе в структуру могут быть внесены ошибки, которые либо сохранятся в дальнейшем в процессе уточнения, либо их исправление потребует проведения чересчур большого количества циклов уточнения и ручной перестройки модели. Поэтому разработка и появление новых методов и программ построения и визуального анализа атомной модели, карт распределения электронной плотности, манипуляции с синтезом и моделью вызывают большой интерес в среде кристаллографов.

Для построения моделей по MIR картам электронной плотности в настоящее время активно используется трехмерная компьютерная графика. Особенно широко употребимой является программа Т.А.Джонса **FRODO**/93,94/ и ее более поздняя модификация - программа **O**, которые применяются на многих типах графических станций (Evans & Sutherland Picture Systems; Silicon Graphics Challenge machine /Crimson workstation; Indigo 2 graphics terminal; Biographics и т.д.). Практически все опубликованные в последние годы работы по расшифровке структур макромолекул использовали на соответствующих этапах решения структуры программу FRODO или ее

модификации. Программа предоставляет пользователю в удобном графическом интерактивном режиме широкий набор возможностей построения и корректировки атомной модели на основе карт распределения электронной плотности (синтезов Фурье или разностных синтезов Фурье). На экран выводится часть электронной плотности с наложенным на нее молекулярным фрагментом, и оба объекта можно двигать, вращать, менять интенсивность окрашивания и т.д. При этом используются как программные, так и аппаратные средства контроля и управления изображением (шкалирование, перемещение, вращение трехмерного объекта, изменение яркости и интенсивности цвета и т.п. при помощи управляющих рукояток). Аппаратные средства позволяют осуществлять все указанные операции очень быстро, в режиме реального времени. Для объекта, состоящего из трехмерных элементов, манипулируя интенсивностью и яркостью цвета, можно подчеркнуть эффект трехмерности (более "близкие" к зрителю элементы изображаются на экране более яркими). Серия команд позволяет вывести на экран очередную часть модели и/или синтеза, задавая либо координаты центра и радиус изображаемого объекта, либо номера остатков молекулы. Последний способ более удобен при систематическом вписывании атомной модели в плотность: можно двигаться постепенно по цепи, добавляя по одному остатку с одного конца цепи и удаляя с другого.

С помощью светового карандаша можно указать конкретные атомы на экране, для которых затем можно измерить расстояния и углы между ними, определить их координаты, тип атома и остатка, установить список ближайших соседей к указанному атому. Другие команды позволяют установить или ликвидировать валентную связь между двумя атомами, передвигать атом в новую позицию, двигать и вращать группу связанных атомов. С помощью потенциометров (рукояток) можно манипулировать торсионными углами. После построения вручную фрагмента модели можно провести его

автоматическую регуляризацию для восстановления идеальной стереохимии. При этом стратегия уточнения базируется на том, чтобы не допускать больших сдвигов атомов от своих стартовых положений, но весь фрагмент в целом должен удовлетворять локальным ограничениям на длины валентных связей, углов и торсионных углов. Одни атомы могут не ограничиваться при движении, другие быть зафиксированы, а координаты третьих могут определяться на основе словаря идеальных для белков геометрических величин. После уточнения структура фрагмента на экране приобретает налучшую для данной плотности стереометрическую конформацию. Подробное описание возможностей программы дается в работе /94/.

TURBO-FRODO /95/ - аналогичная программа трехмерной молекулярной графики для компьютеров класса Silicon Graphics. С помощью этой программы, имея доступ к PDB, можно осуществить визуальный анализ и сравнение молекул и фрагметов из разных белков, исследовать структурные различия в семействах мутантов и проверить различные химические гипотезы. Можно интерактивно строить модель из фрагментов, используя вращения и сдвиги фрагментов. Работа с картами электронной плотности происходит аналогично программе FRODO. Имеется широкий диапазон как стилей представления моделей, так и окрашивания различных объектов изображения.

MACINPLOT - изображение электронной плотности и атомных моделей на персональных компьютерах типа Macintosh /96/. Программа считывает подготовленные в FRODO файлы с атомными моделями, электронной плотностью и молекулярными объектами и выводит на экран компьютера Macintosh соответствующие рисунки, позволяя размещать на них диаграммы, подписи и другую информацию, необходимую для подготовки публикации. Изображения можно выводить в моно- и стереорежимах. Для каждого атома на экране можно задать его тип, номер остатка и тип остатка различными

способами (варьируется размер, цвет и стиль шрифта). Можно изменять размер изображения и его расположение на экране, а также вращать рисунок вокруг трех осей координат.

PLOTQ - изображение электронной плотности в двух- и трехмерном пространстве /97/. Задача визуального изучения синтезов распределения электронной плотности возникает на всех этапах расшифровки структуры рентгеновскими методами. Программа PLOTQ позволяет вывести на экран монитора одно или несколько сечений вдоль любой оси либо трехмерную оболочку синтеза. Программа работает в UNIX и VMS операционных системах на терминалах типа Tektronix, Cifer, Autograph.

ORTEX - интерактивная версия программы **ORTEP** /98/ с использование графического терминала /99/. Программа С.Джонсона ORTEP, созданная в 1976 г., позволяет вывести на плоттер шаро-стержневую модель с анизотропными температурнами параметрами с атомами в виде эллипсоидов, соединенных валентными связями. Интерактивная графическая версия ORTEX дает возможность вращать такое изображение, расставлять метки атомов, выводить содержимое экрана в файлы, пригодные для печати на лазерном принтере. Программа работает на VAX - станциях.

MOLSCRIPT - программа, разработанная Р.Краулисом /100/, позволяет сгенерировать и вывести на экран дисплея графическое изображение белковой структуры в виде комбинации различных типов представления объекта: схематическая укладка элементов вторичной структуры; шаро-стержневая или состоящая из шаров модель; проволочная модель. Выбор подходящего типа представления объекта зависит от того, какие именно аспекты изображения структуры вызывают интерес. Например, активный или связывающий центр молекулы белка можно изобразить в виде шаро-стержневой модели, в то время как общую укладку молекулы лучше представить схематичным рисунком. Возможность комбинирования разных типов

представления, использование световых бликов, теней и различной интенсивности окрашивания для создания эффекта трехмерности объекта, проставление текстовых меток для идентификации атомов и остатков позволяет эффективно конструировать высококачественное изображение модели для исследования и публикаций.

Схематический рисунок белковой структуры использует изображение спиральной ленты для спиральных участков вторичной структуры, плоские стрелки для β-слоев и гибкие цилиндрические шнуры для петель, которые строятся автоматически по координатам Cα-атомов. Программа имеет удобный для пользователя интерактивный интерфейс, позволяющий легко манипулировать с многочисленными параметрами и формировать необходимое изображение. MOLSCRIPT позволяет использовать либо красно-сине-зеленую палитру цветов, либо варьировать параметры оттенка-насыщенности-яркости.

Программа работает на графических станциях типа Silicon Graphics Iris-4D System. На выходе получается файл в формате PostScript, который можно распечатать на лазерном принтере.

MOLDRAW - программа, позволяющая вывести на экран и произвести различные манипуляции с изображениями моделей молекул на персональных компьютерах IBM PC /101/, используя систему меню и "мышь". Программа позволяет:

- вывести на экран шаро-стержневую или построенную из сфер модель молекулы как из PDB, так и из базы данных неорганических структур;
- выбирать любую часть структуры и менять масштаб изображения;
- выводить изображение вдоль любого кристаллографического направления или вдоль нормали к любой кристаллографической плоскости;
- вращать изображение модели;
- добавлять, удалять, переименовывать, соединять и разъединять атомы;

- интерактивно управлять химическими (ковалентные и Ван-дер-Вальсовые радиусы) и графическими парметрами изображения;
- генерировать файлы для вывода на плоттер или лазерный принтер.

CRYSTRUCT - графическая система изображения пространственной структуры неорганических соединений для компьютеров серии SUN в системе UNIX /102/. Программа позволяет реализовать различные типы представления моделей (скелетную, шаро-стержневую, шаро-скелетную, полиэдрическую, состоящую из сфер и т.д.). Входной файл должен содержать полную информацию о каждом атоме модели: атомные координаты, радиус, коэффициент заполнения, метку, номер цвета, пары валентно связанных атомов, толщину и цвет для изображения связи.

RIBBONS - представление структур макромолекул (белков, нуклеиновых кислот) в виде ленточной модели на графической станции типа Silicon Graphics Iris-4D System /103/. Программа является удобным средством изображения пространственной укладки полипептидной цепи и вторичной структуры белков, для чего имеется большой набор способов варьирования стилей представления ленточных моделей и широкая цветовая палитра для характеристики геометрических и биологических свойств молекулы. Ленту можно разбить на отдельные остатки, задавая различные схемы окрашивания в зависимости от типа остатка, распределения температурного фактора, энергии, значений торсионных углов φ/ψ на карте Рамачандрана и т.п. Для оптимального изображения требуется указать вторичную структуру молекулы: каждый остаток классифицируется по его принадлежности к спирали, стренду или петле. Затем каждый класс можно изобразить различным стилем или текстурой (например, спираль можно представить в виде цилиндра).

CHANNEL - пакет программ /104/ для построения и анализа каналов и полостей в белковых кристаллах для персональных компьютеров типа IBM PC.

Программа позволяет:

- построить пространственную сеть помеченных точек в элементарной ячейке таким образом, что если поместить в такую точку центр сферы заданного радиуса R, то она не будет пересекаться ни с одним атомом молекулы (каждый атом молекулы можно представит в виде сферы с соответствующим Ван-дер-Вальсовым радиусом);
- выбрать из всего множества помеченных точек связанные подмножества и разбить их на две группы: полости и каналы;
- количественно оценить объем и площадь поверхности каналов;
- построить набор аппроксимирующих каналы сфер для их визуализации.

Программа работает на любом персокомпьютере класса IBM PC, если имеется достаточно памяти.

Перечисленный ряд графических комплексов и программ представляет собой далеко не полную картину всего набора современных средств визуализации и манипулирования с кристаллографическими объектами. Здесь представлены лишь наиболее характерные тенденции в применении компьютерной графики к задачам кристаллографии макромолекул, и количество такого рода программ все возрастает в связи с развитием и усовершенствованием вычислительной и графической техники.