АКАДЕМИЯ НАУК СССР
НАУЧНО - ИССЛЕДОВАТЕЛЬСКИЙ ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР

на правах рукописи

Л У Н И Н
ВЛАДИМИР ЮРЬЕВИЧ

УДК 548.737

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК СИНТЕЗОВ ФУРЬЕ ЭЛЕКТРОННОЙ ПЛОТНОСТИ ДЛЯ РЕШЕНИЯ ФАЗОВОЙ ПРОБЛЕМЫ В КРИСТАЛЛОГРАФИИ БЕЛКА

01.04.18 - Кристаллография, физика кристаллов

Диссертация на соискание ученой степени
доктора физико-математических наук

Пущино, 1991

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Метод рентгеноструктурного анализа (РСА) постоянно привлекает внимание широкого класса исследователей, работающих в областях науки, использующих информацию о структуре вещества на атомно-молекулярном уровне. Это связано с тем, что метод РСА является основным методом, позволяющим определять структуру исследуемого объекта на атомном уровне, то есть определять координаты в трехмерном пространстве всех атомов, входящих в состав исследуемого вещества. Таким образом, РСА позволяет давать ответы на вопросы о структуре и функционировании молекул, формулируемые в терминах геометрических характеристик молекулы (расстояний между атомами, длин связей, валентных или двугранных углов и т.п.) или характеристик, связанных с геометрией (распределение поверхностного заряда, "доступные" области молекулы и т.п.).

Принципиальной особенностью метода РСА является неполнота данных, получаемых непосредственно в рентгеновском эксперименте. Эксперимент позволяет измерить лишь величины модулей F_s комплексных коэффициентов (структурных факторов) в разложении функции распределения электронной плотности в ряд Фурье :

$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{s} F_s \, e^{i\varphi_s} \, e^{-2\pi i(\mathbf{s},\mathbf{r})} .$$

Определение фаз φ_s структурных факторов представляет собой центральную проблему ("фазовую проблему ") рентгеноструктурного анализа, и успех всей работы по

определению структуры определяется во многом тем, насколько точно удалось решить фазовую проблему.

В силу неполноты данных рентгеновского эксперимента для решения фазовой проблемы необходима какая-то дополнительная информация об исследуемом объекте. В настоящее время на практике применяются, в основном, два класса методов решения фазовой проблемы. Для определения структуры низкомолекулярных соединений (100-150 атомов в независимой части элементарной ячейки) применяются "прямые методы определения фаз" ([1],[3]), дополнительной информацией для которых является "атомность" исследуемого объекта. Основой для решения фазовой проблемы при исследовании биологических макромолекул является метод изоморфного замещения [2], в котором в качестве добавочной информации об исследуемом объекте (нативном белке) выступают данные по дополнительным рентгеновским экспериментам с близкими соединениями (изоморфными производными), отличающимися от нативного белка локальными добавками ("тяжелыми" атомами), не искажающими нативную структуру. В этом методе тяжесть решения фазовой проблемы смещается в биохимическую область. Получение изоморфных тяжелоатомных производных является сложной задачей, которую не всегда удается удовлетворительно решить. Даже в тех случаях, когда производные соединения удается получить, изоморфизм может иметь место лишь приближенно - присоединяемые тяжелые атомы могут вносить некоторые искажения в нативную структуру. Это приводит (в зависимости от степени нарушения изоморфизма) к более или менее грубым ошибкам при расчете фаз и, как следствие, к усложнению задачи интерпретации в структурных терминах получаемой

функции распределения электронной плотности (вплоть до полной неинтерпретируемости). Аналогичные сложности возникают в случае, когда удается получить только одно изоморфное производное (для однозначного решения фазовой проблемы необходимы по крайней мере два производных соединения или наличие в исследуемом объекте аномально рассеивающих центров).

Следует отметить, что если в методе изоморфного замещения дополнительная информация, необходимая для решения фазовой проблемы, поставляется дополнительным рентгеновским экспериментом, то в "прямых методах" источником дополнительной информации являются математические свойства функции распределения электронной плотности.

Указанные сложности применения метода изоморфного замещения стимулировали в последние годы значительные усилия по поиску дополнительных источников информации о структуре макромолекул и методов использования такой дополнительной информации для решения двух задач:

I) повышения интерпретируемости синтезов (уточнение значений фаз, определенных с ошибкой и, возможно, доопределение значений некоторых фаз, не определенных ранее);

2) решения фазовой проблемы для макромолекул при отсутствии тяжелоатомных производных.

Гистограмма синтеза Фурье электронной плотности - новый источник информации о кристаллах белков.

При поиске подходов к уточнению значений фаз структурных факторов различными исследователями неоднократно

делались попытки использовать в качестве дополнительной информации об исследуемом объекте некоторые ограничения на область возможных значений искомой функции распределения электронной плотности. Наиболее очевидным и популярным ограничением такого рода является требование неотрицательности функции распределения электронной плотности : $\rho(\mathbf{r})\geq 0$. Из попыток использовать это ограничение развились такие методы, как неравенства, связанные с положительной определенностью матрицы Карля-Хауптмана [4],[7-9], метод максимума детерминанта [10-22], а также ряд других подходов [23-27]. Другими примерами "явных" ограничений на область возможных значений, использованных на практике, являются двусторонняя ограниченность $0\leq\rho(\mathbf{r})\leq\rho_{max}$, а также требование, чтобы функция $\rho(\mathbf{r})$ принимала только два значения - 0 и I [6I-63]. Менее явными попытками наложить ограничения на возможные величины ρ являются разнообразные методы модификации электронной плотности [28-58], в которых по ходу работы рассчитанные значения электронной плотности заменяются на "более правильные" значения. Поскольку широко используемое на практике уточнение значений фаз с использованием "tg-формулы" может рассматриваться как одна из реализаций метода модификации электронной плотности [59,60], то сюда примыкают и многочисленные работы, выполненные с использованием разнообразных модификаций "tg-формулы".

Более тщательным анализом области возможных значений является попытка установить не только то, какие значения может принимать искомая функция $\rho(\mathbf{r})$, но и то, как часто принимает функция $\rho(\mathbf{r})$ каждое из возможных значений. (Этот

подход был предложен независимо автором [64-66], а также рядом зарубежных исследователей [78],[82-84]). Анализ "гистограмм" - распределений частот , с которыми встречаются те или иные значения электронной плотности в кристаллах белков, показывает, что эти гистограммы имеют характерную асимметричную форму . Более того, оказалось, что эта форма чувствительна к наличию ошибок в фазах структурных факторов и к отсутствию части структурных факторов при расчете функции распределения электронной плотности. Это свойство гистограмм дает возможность использовать их как дополнительную информацию об объекте при рентгеноструктурном исследовании белков. (Варианты набора необходимых значений фаз, не приводящие к правильной гистограмме, должны отвергаться как ошибочные.)

Предсказание гистограмм для белков с неизвестной пространственной структурой.

При попытке использовать гистограммы синтезов электронной плотности в качестве дополнительного источника информации об исследуемом объекте немедленно возникает вопрос : откуда взять эталонную гистограмму для объекта, структура которого еще не определена ? Нами предложена простая эмпирическая модель для описания формы гистограмм синтезов электронной плотности для кристаллических белков. Предлагаемая модель не претендует на детальное описание тонких особенностей гистограмм, но дает достаточную точность предсказания, чтобы быть успешно используемой в практической работе.

Указанная модель была получена следующим образом.

Вначале были рассмотрены гистограммы, соответствующие белкам с уже известной пространственной структурой. Более точно, из банка белковых молекул был выбран "базисный набор" белков и рассчитан соответствующий "базисный" набор гистограмм. Анализ этих гистограмм позволил выдвинуть гипотезу о том, что любая из них может быть с разумной точностью смоделирована как линейная комбинация двух стандартных распределений. При этом коэффициенты комбинации выражаются через такие параметры кристалла, как объем элементарной ячейки $|V|$ и количество электронов в элементарной ячейке F_{ooo}, а сами стандартные распределения могут быть проинтерпретированы как нормализованные гистограммы, рассчитываемые отдельно по области молекулы и межмолекулярной области. Для определения точного вида стандартных распределений далее решается задача на минимизацию - эти распределения подбираются таким образом, чтобы составленные из них комбинации наилучшим образом описывали гистограммы из базисного набора.

После того как рассчитаны стандартные распределения, мы можем смоделировать гистограмму для любого белка, если только известны параметры $|V|$ и F_{ooo}, необходимые для расчета коэффициентов, с которыми включаются в модель стандартные распределения.

Тесты на не вошедших в базисный набор белках с известной структурой показывают, что предложенная модель гистограммы удовлетворительно "предсказывает" гистограммы синтезов среднего и высокого разрешения. Что же касается низкого разрешения, то здесь модель становится излишне груба. Частично проблема предсказания гистограммы синтезов

низкого разрешения снимается в ситуации, когда имеется гомолог исследуемого белка или известны (например, из электронной микроскопии) внешние очертания и размеры молекулы. В этом случае приемлемую гистограмму можно получить, "размещая" модель гомолога в элементарной ячейке исследуемого белка без самоналезаний и рассчитывая гистограмму, отвечающую такому вспомогательному объекту.

Использование гистограмм для восстановления недостающих структурных факторов.

В практической работе набор структурных факторов, использующихся при расчете синтеза Фурье, не всегда является полным. Для части рефлексов не удается сколь-нибудь надежно определить значения фаз, а для части могут быть даже не определены значения модулей структурных факторов. Обычная практика - синтез Фурье рассчитывается без этих рефлексов, то есть значения соотвествующих структурных факторов заменяются нулями. Однако исключение из расчета даже небольшого числа рефлексов может оказать существенное влияние на качество синтеза, если это исключение носило "систематический" характер. Особенно важны в этом отношении низкоугловые рефлексы, определяющие в значительной мере внешние очертания молекулы [85-88],[97]. В ситуации, когда нам известна гистограмма, которой должен обладать "правильный" синтез, мы можем попытаться подобрать для неизвестных структурных факторов такие значения (взамен нулевых), которые приведут к синтезу, имеющему более правильную гистограмму. Ниже будет более строго изложен такой подход, описаны пути его компьютерной реализации и

приведены результаты опробования метода на тестах и в практической работе.

Говоря об определении недостающих структурных факторов из условия наилучшего согласования теоретической и рассчитанной гистограмм, следует, разумеется, понимать, что речь идет об определении сравнительно небольшого числа неизвестных величин при условии, что большая часть необходимых для расчета синтеза структурных факторов известна. Особенно это относится к ситуации, когда делается попытка доопределить модули структурных факторов, не определенные по каким-то причинам экспериментально.

Использование гистограмм в задаче уточнения значений фаз структурных факторов.

Отдельный интерес представляет вопрос, какое место занимают процедуры уточнения значений фаз структурных факторов за счет дополнительной информации, поставляемой гистограммой, среди других методов фазового уточнения.

Оказывается, что широкий класс различных типов дополнительной информации может быть представлен математически как свойство распределения электронной плотности $\rho(\mathbf{r})$ не изменяться под воздействием некого преобразования $\tau[\rho]$ (своего для каждого из типов дополнительной информации). Свойство $\rho = \tau[\rho]$ в свою очередь эквивалентно системе комплексных уравнений для соотвествующих структурных факторов. Как показал наш анализ, большая часть используемых методов уточнения значений фаз (уточнение по тангенс-формуле, усреднение по некристаллографической симметрии, "метод Wang'a", метод

модификации электронной плотности и др.) могут рассматриваться как решение методом последовательных приближений фазовой части уравнений для структурных факторов, отвечающих свойству $\rho=\tau[\rho]$ с надлежащим преобразованием τ.

Далее, было установлено, что свойство функции иметь предписанную гистограмму также может быть сформулировано в виде требования $\rho=\tau[\rho]$, где преобразование τ имеет вид $\rho(\mathbf{r}) \to \lambda_\rho(\rho(\mathbf{r}))$, с некоторой специальным образом построенной функцией $\lambda_\rho(t)$ (своей для каждой из преобразуемых функций $\rho(\mathbf{r})$). Итерационное решение фазовой части соответствующих уравнений для структурных факторов совпадает с предложенными недавно "Histogram specification" [84] и "Histogram matching" [78] методами. Использованный нами минимизационный метод - подбор значений фаз, приводящих к наиболее правильной гистограмме, является попыткой использовать полную систему уравнений для структурных факторов.

Проведенный анализ позволяет по-новому взглянуть на широко распространенный метод уточнения значения фаз - метод модификации электронной плотности [28-58]. Этот метод основан на свойстве искомой функции $\rho(\mathbf{r})$ не изменяться при преобразовании $\rho \to \lambda(\rho(\mathbf{r}))$ (наиболее популярна при этом функция $\lambda(\rho)=3\rho^2-2\rho^3$). Однако оставалось неясным, какую информацию о синтезах среднего или низкого разрешения несет в себе это свойство. Изучение в модельных ситуациях функций $\lambda_\rho(t)$, восстанавливающих предписанную гистограмму, показало, что функция $3t^2-2t^3$ весьма похожа на них. Поэтому метод модификации электронной плотности можно трактовать как упрощенную реализацию "Histogram specification" метода (в

которой модифицирующая функция берется постоянной, отражающей основные черты функций, восстанвливающих гистограмму), и информация, "за счет которой" работает метод модификации электронной плотности, есть, по-существу, специфичность формы гистограмм синтезов электронной плотности в кристаллах белков.

Прямое решение фазовой проблемы для низкоугловых рефлексов.

В главе 5 ниже предлагается новый подход к прямому решению фазовой проблемы для низкоугловых рефлексов при условии, что известна гистограмма, которой должен обладать искомый синтез электронной плотности.

Внешне идея подхода проста - сгенерируем множество различных фазовых наборов (например, датчиком случайных чисел) и выберем тот, который приводит к гистограмме, наиболее близкой к эталону. Можно было бы ожидать, что, когда число сгенерированных вариантов велико, среди них с большой вероятностью встретится набор фаз, близких к настоящим значениям, и он может быть опознан по "хорошей" гистограмме соответствующего синтеза.

Однако, в действительности, ситуация более сложная. Во-первых, решение задачи - найти фазы, приводящие при заданных модулях структурных факторов к предписанной гистограмме синтеза - вообще говоря, с практической точки зрения, неединственно. Оказывается, что могут существовать сильно различающиеся наборы фаз, приводящие к достаточно хорошим гистограммам. Поскольку, предсказывая "эталонную" гистограмму для белка с еще неизвестной структурой, мы

всегда делаем это с некоторой ошибкой, то любой набор фаз, приводящий к гистограмме, отличающейся от эталона в пределах этой ошибки, должен рассматриваться как возможный. Таким образом, отбор наборов фаз по соответствующей им гистограмме приводит к множеству вариантов, среди которых встречаются и резко различающиеся между собой.

Методы кластерного анализа позволяют более тщательно исследовать множество отобранных вариантов (с хорошей гистограммой) и разбить его на кластеры, группирующиеся вокруг различных вариантов решения фазовой проблемы. Усреднение вариантов внутри каждого из выделенных кластеров позволяет выделить небольшое число (два-три) возможных вариантов решения фазовой проблемы.

Проведенные нами тесты показали, что точное решение попадает в число отобранных вариантов и, более того, может быть, как правило, выделено из остальных по характеристикам разброса точек внутри кластера относительно их среднего.

Вычислительные проблемы.

В последней главе работы рассматриваются некоторые вычислительные проблемы, возникающие при практической реализации подходов, изложенных в предыдущих разделах. Наиболее существенная из них связана с задачами минимизации функций, зависящих от большого числа переменных. Многие из разобранных выше подходов сводились в конечном счете к минимизации некоторого сложным образом определенного критерия качества пробного набора фаз. При этом вычисление каждого значения такого критерия требует обычно ощутимых затрат процессорного времени компьютера. До последнего

времени считалось, что наиболее трудноразрешимой проблемой при минимизации такого рода функций является расчет градиента минимизируемой функции (необходимого при минимизации, поскольку именно он определяет направление, в котором надо сдвигать варьируемые параметры для того, чтобы уменьшить значение целевой функции). Так, при вычислении частных производных по разностным формулам для расчета градиента требуется в n раз больше времени, нежели для расчета одного значения функции (n - число варьируемых параметров). При времени расчета одного значения критерия, исчисляемого минутами, и значительном числе переменных (в задаче уточнения атомной структуры это число может достигать десятков тысяч) задача расчета градиента может выглядеть неразрешимой. Однако, оказывается, что для любой функции $f(\mathbf{x})$, зависящей от сколь угодно большого числа переменных n можно построить алгоритм, требующий для расчета значений всех компонент градиента практически такое же время, которое нужно для расчета одного значения функции $f(\mathbf{x})$. Этот факт имеет чрезвычайно большое методологическое значений для задач рентгеноструктурного анализа. Из него следует, что любая характеристика объекта, вычисляемая на доступном исследователю компьютере, может быть локально улучшена на компьютере той же мощности. В главе 6 изложена реализация этой общей идеи для задач, возникающих при определении пространственной структуры вещества методами рентгеноструктурного анализа.

Глава I. ГИСТОГРАММА СИНТЕЗА ЭЛЕКТРОННОЙ ПЛОТНОСТИ КОНЕЧНОГО РАЗРЕШЕНИЯ - НОВЫЙ ИСТОЧНИК ИНФОРМАЦИИ О КРИСТАЛЛАХ БЕЛКОВ .

I.I. Введение

При поиске подходов к уточнению значений фаз структурных факторов различными исследователями неоднократно делались попытки использовать в качестве дополнительной информации об исследуемом объекте некоторые ограничения на область возможных значений искомой функции распределения электронной плотности. При этом применялись ограничения следующих типов :

I) $\rho(\mathbf{r}) \geq 0$ (ограниченность области возможных значений снизу;

2) $\rho_{min} \leq \rho(\mathbf{r}) \leq \rho_{max}$ (ограниченность области возможных значений с двух сторон);

3) $\rho(\mathbf{r}) = \{ 0$ или $I \}$ (конечное множество допустимых значений).

Более тщательным анализом области возможных значений является попытка установить не только то, какие значения может принимать искомая функция $\rho(\mathbf{r})$, но и то, как часто принимает функция $\rho(\mathbf{r})$ каждое из возможных значений . Анализ "гистограмм"- распределений частот , с которыми встречаются те или иные значения электронной плотности в кристаллах белков - показывает, что эти гистограммы имеют характерную асимметричную форму (Рис.I.Iа). Более того, оказалось, что

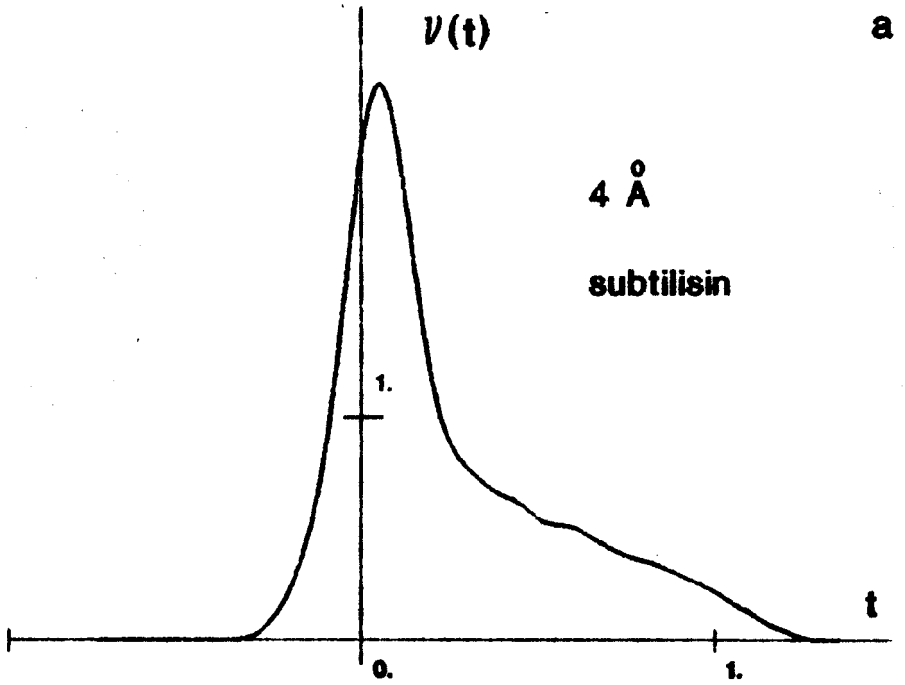
эта форма чувствительна к наличию ошибок в фазах структурных факторов и к отсутствию части структурных факторов при расчете функции распределения электронной плотности (Рис.I.Ib). Это свойство гистограмм дает возможность использовать их как дополнительную информацию об объекте при рентгеноструктурном исследовании белков. (Варианты набора необходимых значений фаз, не приводящие к правильной гистограмме, должны отвергаться как ошибочные.)

I.2. Гистограмма - упрощенный подход к введению меры на области значений исследуемой функции.

Рассмотрим ситуацию, когда в ограниченной области V трехмерного пространства определена некоторая функция $\rho(\mathbf{r})$, и мы интересуемся тем, какие значения принимает эта функция и насколько часто она принимает каждое из своих значений. Наиболее прямой практический подход к ответу на этот вопрос может быть следующим. Введем в области V равномерную сетку, и пусть $\{\rho_j\}_{j=1}^{N}$ - совокупность значений функции $\rho(\mathbf{r})$, вычисленных в узлах этой сетки. Разобьем интервал (ρ_{min},ρ_{max}), в котором лежат значения $\{\rho_j\}$, на заданное число K равных частей (бинов) и для каждого бина определим частоту попадания значений ρ_j в этот бин

$$\hat{\nu}_k = n_k / N \quad , \quad k = I,\ldots,K \ . \qquad (I.I)$$

Здесь n_k - число точек сетки с соответствующими значениями



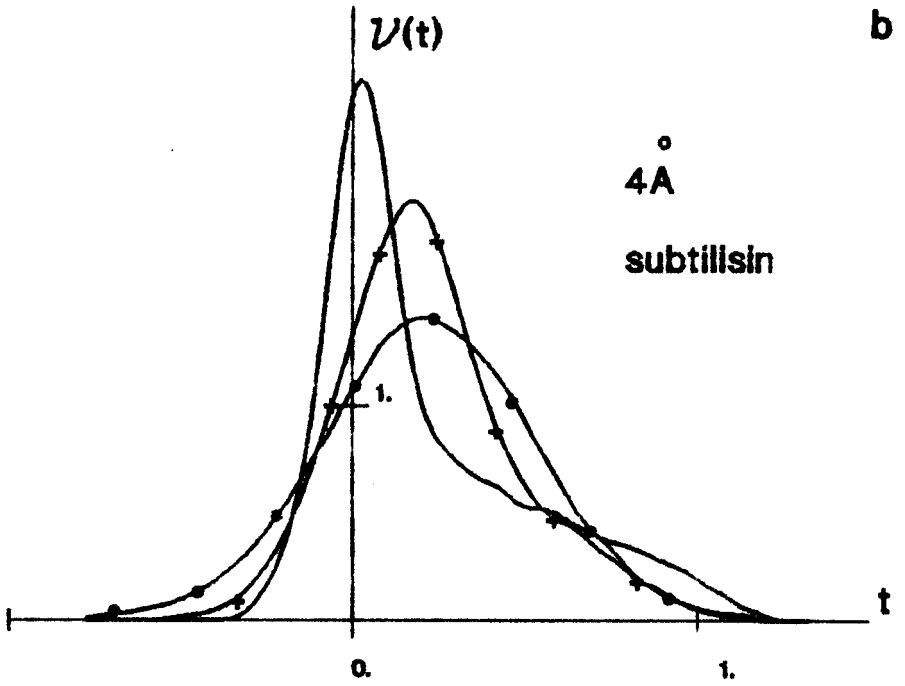


Рис.I.I. a) Гистограмма, отвечающая синтезу Фурье разрешения 4А для белка субтилизина, рассчитанному с точными значениями модулей и фаз структурных факторов;

b) гистограммы, отвечающие точному синтезу (————); синтезу, рассчитанному с точными значениями модулей и случайными значениями фаз (– · – · –); синтезу, при расчете которого исключено I7% структурных факторов (–+–+–+–).

ρ_j , лежащими в k-ом бине, то есть таких точек, что

$$\rho_{min} + (j-1)\frac{\rho_{max}-\rho_{min}}{K} \le \rho_j \le \rho_{min} + j\frac{\rho_{max}-\rho_{min}}{K}, \qquad (I.2)$$

N – общее число точек сетки. Совокупность (распределение) частот $\{\hat{\nu}_k\}_{k=1}^{K}$ мы будем называть гистограммой, отвечающей функции $\rho(\mathbf{r})$.

Иногда более удобна для работы нормированная гистограмма

$$\nu_k = n_k / (\Delta_k N), \quad k = 1,\dots,K, \qquad (I.3)$$

где Δ_k обозначает длину k-того бина. В этом случае вероятность обнаружения точки (при случайном выборе узла сетки в области V) со значением ρ_j, лежащим в k-ом бине, равна $\hat{\nu}_k\Delta_k$. При расчете нормированной гистограммы мы можем использовать разбиения интервала (ρ_{min},ρ_{max}) на бины разной длины.

I.3. Введение меры на области значений исследуемой функции. Строгий подход.

Полученные расчетом по формуле (I.I) значения частот зависят, строго говоря, не только от исследуемой функции $\rho(\mathbf{r})$, но и от того, как введены бины. Чтобы избавиться от этой зависимости, мы введем более строго меру на области значений исследуемой функции . Определим для функции $\rho(\mathbf{r})$ кумулятивную функцию

$$N(t) = \frac{1}{|V|}\,\mathrm{mes}\,\{\,\mathbf{r} : \rho(\mathbf{r}) \le t\,\} \qquad (I.4)$$

и плотность кумулятивной функции

$$\nu(t) = \frac{d}{dt}N(t) = \frac{1}{|V|}\frac{d}{dt}\,\mathrm{mes}\{\mathbf{r} : \rho(\mathbf{r}) \le t\,\}. \qquad (1.5)$$

Здесь и далее $\{\mathbf{r}: \mathcal{A}\}$ обозначает множество точек из области

V, для которых выполнено условие $\mathscr{A}$, mes S обозначает объем области S, $|V|$ = mes V – объем области V. Функции N(t) и $\nu(t)$ зависят только от исследуемой функции $\rho(\mathbf{r})$ и не связаны с выбором сетки в исследуемой области и разбиения вещественной оси на бины. Величина $\nu(t)\Delta t$ при малых Δt представляет собой вероятность встретить (при случайном выборе точки из области V) значение $\rho(\mathbf{r})$, лежащее в интервале $(\rho,\rho+\Delta t)$.

Нетрудно видеть, что нормированные частоты ν_k, вычисляемые по формуле (I.3), являются приближенными значениями функции $\nu(t)$ в точках t_k, отвечающих серединам бинов :

$$\nu(t_k) = \lim_{\Delta_k \to 0,\ N\to\infty} \nu_k \ . \qquad (I.6)$$

Мы будем далее использовать термин "гистограмма" как для обозначения наборов частот $\{\nu_k\}_{k=1}^{K}$ и $\{\hat{\nu}_k\}_{k=1}^{K}$, так и для обозначения плотности $\nu(t)$ кумулятивной функции.

Приближенные значения функции $\nu(t)$ могут вычисляться, естественно, не только по формулам (I.3),(I.6). Мы приведем сейчас класс формул для приближенного вычисления $\nu(t)$, которые будут полезны далее при решении задач уточнения набора структурных факторов (см.главы 3 и 4 ниже).

Для произвольной функции $\lambda(\rho)$ при каждом значении t справедливо соотношение

$$\frac{I}{|V|}\int_V \lambda(\ t - \rho(\mathbf{r})\ )\ dV_{\mathbf{r}} = \int_{-\infty}^{\infty} \lambda(\ t - \tau\ )\ \nu(\tau)\ d\tau\ , \qquad (I.7)$$

выражающее равенство средних "по пространству" и "по мере". Если теперь мы выберем последовательность функций $\lambda_n(\rho)$, $n = I,\ldots,\infty$, сходящуюся к δ-функции, то переходя в правой части равенства (I.7) к пределу, мы получим функцию $\nu(t)$, то

есть

$$\nu(t) = \frac{1}{|V|} \lim_{n\to\infty} \int_V \lambda_n(t-\rho(\mathbf{r}))\, dV_{\mathbf{r}} , \qquad (1.8)$$

Вводя в области V равномерную сетку и применяя для вычисления интеграла простейшую квадратурную формулу, получаем

$$\nu(t_k) \approx \frac{1}{N} \sum_{j=1}^{N} \lambda_n(t_k - \rho_j) , \qquad (1.9)$$

если N и n достаточно велики (здесь $\{\rho_j\}$ - значения функции $\rho(\mathbf{r})$ в узлах сетки).

Выбирая функции

$$\lambda_n(t) = \begin{cases} 1 / \Delta_n & \text{при } |t| \le \Delta_n/2 , \\ 0 & \text{при } |t| > \Delta_n/2 , \text{ где } \Delta_n \to 0 , \end{cases} \qquad (1.10)$$

мы получаем формулу (1.3) для нормированных частот.

1.4. Зависимость гистограммы от разрешения синтеза Фурье.

Рассмотрим теперь ситуацию, когда исследуемая функция описывает распределение электронной плотности в кристалле, то есть когда функция $\rho(\mathbf{r})$ является периодической. В этом случае $\rho(\mathbf{r})$ может быть представлена в виде трехмерного ряда Фурье

$$\rho(\mathbf{r}) = \frac{1}{|V|} \sum_{\mathbf{s}\in\mathcal{R}'} F_{\mathbf{s}} e^{i\varphi_{\mathbf{s}}} e^{-2\pi i(\mathbf{s},\mathbf{r})} , \qquad (1.11)$$

где

$$F_{\mathbf{s}} e^{i\varphi_{\mathbf{s}}} = \int_V \rho(\mathbf{r})\, e^{2\pi i(\mathbf{s},\mathbf{r})}\, dV_{\mathbf{r}} , \qquad (1.12)$$

$\mathcal{R}'$ - решетка обратного пространства, $|V|$ - объем элементарной ячейки . На практике мы всегда имеем в распоряжении конечный набор структурных факторов $F_{\mathbf{s}}\exp(i\varphi_{\mathbf{s}})$ и поэтому в состоянии рассчитать лишь некоторое приближение к искомой функции $\rho(\mathbf{r})$

- синтез Фурье конечного разрешения

$$\rho_d(\mathbf{r}) = \frac{1}{|V|} \sum_{|\mathbf{s}|\le 1/d} F_{\mathbf{s}} e^{i\varphi_{\mathbf{s}}} e^{-2\pi i(\mathbf{s},\mathbf{r})} . \qquad (1.13)$$

Разрешение получаемого синтеза характеризуется при этом величиной d.

Оказалось , что вид гистограммы $\nu(t)$ существенным образом зависит от того, синтезу какого разрешения она соответствует (Рис.1.2). Из этого следует два важных вывода:

а) говоря о конкретной гистограмме синтеза электронной плотности, мы должны сознавать, о синтезе какого разрешения идет речь;

б) дополнительная информация об исследуемом объекте, вносимая в процесс исследования гистограммами - это не одна гистограмма, а СЕРИЯ ГИСТОГРАММ, отвечающих синтезам электронной плотности исследуемого объекта разного разрешения.

1.5. Зависимость гистограммы от тепловой подвижности атомов.

Рис.1.2 демонстрирует эффект изменения формы гистограммы при уменьшении числа слагаемых, включаемых в расчет синтеза электронной плотности (понижения разрешения синтеза). Аналогичное воздействие оказывает на гистограмму повышение тепловой подвижности атомов, входящих в состав молекулы . При рентгеноструктурном исследовании тепловая подвижность атомов характеризуется обычно индивидуальными температурными параметрами B_j (при уточнении структуры при высоком разрешении для атомов вводятся индивидуальные матрицы, описывающие анизотропию тепловых колебаний). На

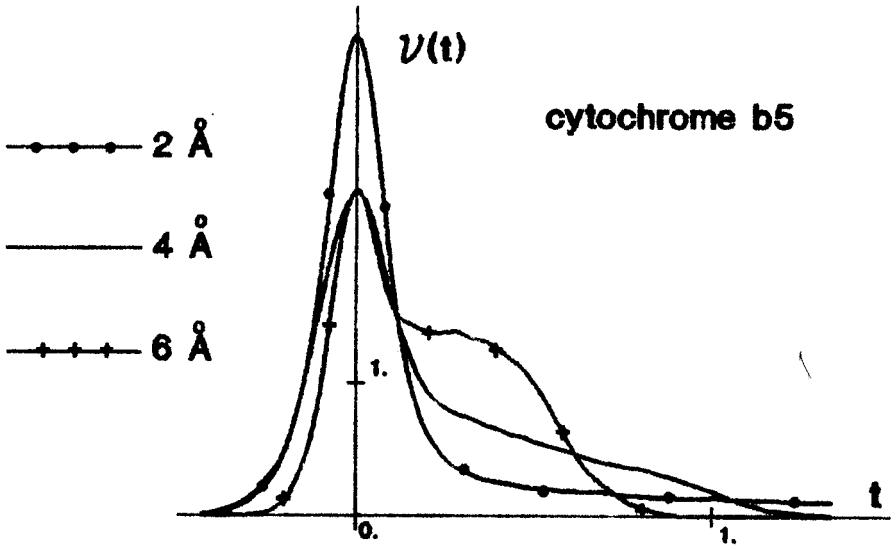


Рис.I.2. Зависимость гистограммы от разрешения синтеза Фурье.

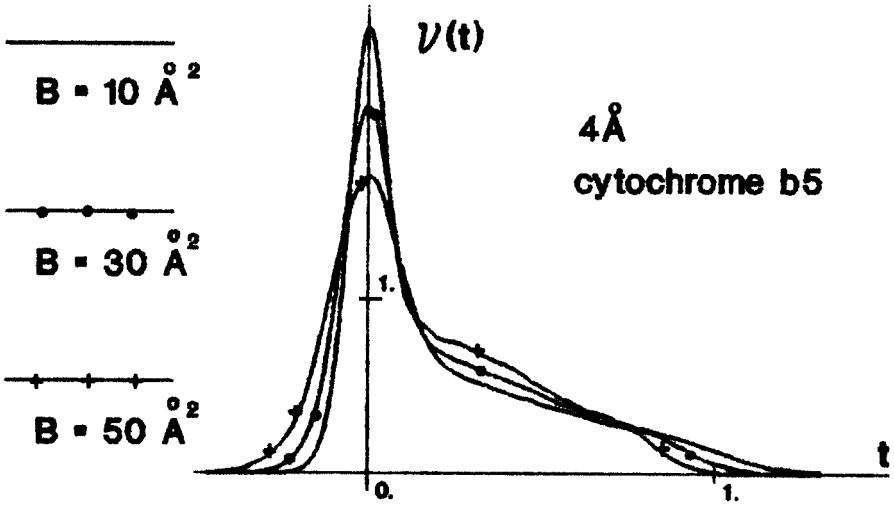


Рис.I.3. Зависимость гистограммы от значения температурного фактора. (В предположении, что все атомы имеют одинаковые значения температурных факторов).

Рис.I.3 показаны гистограммы, отвечающие синтезам одного и того же разрешения для структур с одними и теми же координатами, но разными температурными факторами атомов. Сходность характеров изменений гистограмм имеет простое объяснение. Синтез конечного разрешения $\rho_d(\mathbf{r})$ может быть представлен в виде свертки

$$\rho_d(\mathbf{r}) = \rho_o(\mathbf{r}) * \gamma_d(\mathbf{r}) \qquad (I.I4)$$

где $\rho_o(\mathbf{r})$ - точное распределение электронной плотности в кристалле (отвечающее бесконечному числу членов ряда в (I.II)), а ядро свертки имеет вид

$$\gamma_d(\mathbf{r}) = \frac{\sin(2\pi|\mathbf{r}|/d)-(2\pi|\mathbf{r}|/d)\cos(2\pi|\mathbf{r}|/d)}{2\pi^2|\mathbf{r}|^3} \qquad (I.I5)$$

Увеличение же температурных факторов всех атомов структуры на одну и ту же величину B_{add} приводит к модифицированному распределению электронной плотности

$$\rho_{mod}(\mathbf{r}) = \rho(\mathbf{r}) * \gamma^t_{B_{add}}(\mathbf{r}) \qquad (I.I6)$$

с ядром свертки

$$\gamma^t_B(\mathbf{r}) = (4\pi/B)^{3/2} \exp(-4\pi|\mathbf{r}|^2/B) \quad . \qquad (I.I7)$$

Несмотря на схожесть характера изменения гистограмм на Рис.I.2 и I.3, они имеют разную природу. Разрешение синтеза электронной плотности - это "вычислительная" характеристика, известная исследователю, работающему с этим синтезом. Параметры тепловой подвижности атомов задают физические характеристики реальной структуры, и эти параметры, вообще говоря, не известны точно до окончания работы по расшифровке пространственной структуры. Поэтому при работе с гистограммами, соответствующими белку с еще не известной структурой, мы должны либо оценить заранее средние значения температурных факторов (для белков они обычно бывают порядка

10-20 $Å^2$), либо рассматривать их как неизвестные параметры, требующие дальнейшего определения в процессе работы.

I.6. Чувствительность гистограммы к ошибкам в значениях структурных факторов.

Главное свойство гистограмм, отвечающих синтезам электронной плотности для белков, делающее гистограммы перспективным источником дополнительной информации об исследуемом объекте, является то, что они оказываются чувствительными к ошибкам в значениях структурных факторов $F_s \exp(i\varphi_s)$, используемых при расчете синтеза (I.3). Следующие два теста иллюстрируют это утверждение.

При проведении этих тестов по координатам атомов молекулы субтилизина (взятым из банка белковых молекул [91]) были рассчитаны точные значения структурных факторов (значение температурного параметра для всех атомов было взято равным 10 $Å^2$). Далее были рассчитаны три синтеза Фурье разрешения 4 Å :

а) синтез $\rho^{ex}(\mathbf{r})$, при расчете которого был использован полный набор точных структурных факторов $F_s \exp(i\varphi_s)$ (все слагаемые в (I.13), отвечающие разрешению 4 Å);

б) синтез $\rho^{sr}(\mathbf{r})$, при расчете которого были использованы все слагаемые разрешения 4Å, но точными были взяты лишь значения модулей структурных факторов F_s; фазы же φ_s были взяты случайными (сгенерированы при помощи датчика случайных чисел);

в) синтез $\rho^{r}(\mathbf{r})$, при расчете которого использовались точные значения модулей и фаз структурных факторов, но не

все слагаемые были включены в расчет; около 18 % от общего числа слагаемых (352 слагаемых) в расчет включены не были, то есть заменены нулями.(Исключенные из расчета слагаемые соответствовали узлам обратной решетки, сгруппированным вдоль оси 1 обратного пространства, и были выбраны исходя из реального рентгеновского эксперимента, в котором модули именно этих структурных факторов по техническим причинам измерить не удалось.)

Для указанных выше этих трех синтезов были рассчитаны гистограммы, приведенные на Рис I.Ib. Из этого рисунка видно, что форма гистограммы существенно меняется как при замене правильных значений фаз структурных факторов на случайные, так и при систематическом удалении части случайных факторов из расчета синтеза. Это означает, что гистограмма, рассчитанная для синтеза, построенного с какими-то пробными значениями фаз, может, вообще говоря, служить индикатором правильности определения фаз. Во всяком случае, некоторые варианты решения фазовой проблемы, приводящие к "неправильной" гистограмме, могут быть отброшены как ошибочные. Аналогичная ситуация имеет место и при неточном задании некоторых модулей структурных факторов (например, равными нулю в случае, когда эти модули не измерены экспериментально).

Как будет показано ниже (см.главы 4,5), правильность гистограммы не может служить однозначным критерием правильности решения фазовой проблемы - могут существовать сильно различающиеся наборы фаз, приводящие к близким гистограммам. Однако используемая совместно с другими видами информации об исследуемом объекте гистограмма дает полезную

дополнительную информацию, позволяющую в ряде случаев существенно продвинуться в решении фазовой проблемы.

Предложение использовать информацию, заключенную в гистограмме синтеза Фурье, для определения и уточнения значений фаз было высказано независимо и получило развитие в работах [64-84].

Глава 2. ПРЕДСКАЗАНИЕ ГИСТОГРАММ ДЛЯ БЕЛКОВ С НЕИЗВЕСТНОЙ ПРОСТРАНСТВЕННОЙ СТРУКТУРОЙ.

2.I. Введение.

При попытке использовать гистограммы синтезов электронной плотности в качестве дополнительного источника информации об исследуемом объекте немедленно возникает вопрос : откуда взять эталонную гистограмму для объекта, структура которого еще не определена ? В этой главе мы предложим простую эмпирическую модель для описания формы гистограмм синтезов электронной плотности для кристаллических белков и выясним, насколько хорошо такая модель описывает реальные гистограммы. Предлагаемая модель не претендует на детальное описание тонких особенностей гистограмм, но дает достаточную точность предсказания, чтобы быть успешно используемой в практической работе. Некоторые примеры использования предсказанных гистограмм будут даны в следующих главах.

Общая последовательность действий будет следующей. Сначала мы рассмотрим гистограммы, соответствующие белкам с уже известной пространственной структурой. Более точно, мы выберем из банка белковых молекул "базисный набор" белков и рассчитаем соответствующий "базисный" набор гистограмм. Анализ этих гистограмм позволяет выдвинуть гипотезу о том, что любая из них может быть с разумной точностью смоделирована как линейная комбинация двух стандартных

распределений. При этом коэффициенты комбинации выражаются через такие параметры кристалла, как объем элементарной ячейки $|V|$ и заряд элементарной ячейки F_{000}, а сами стандартные распределения могут быть проинтерпретированы как нормализованные гистограммы, рассчитываемые отдельно по области молекулы и межмолекулярной области. Для определения точного вида стандартных распределений решается далее задача на минимизацию - эти распределения подбираются таким образом, чтобы составленные из них комбинации наилучшим образом описывали гистограммы из базисного набора.

После того как рассчитаны стандартные распределения, мы можем смоделировать гистограмму для любого белка, если только известны параметры $|V|$ и F_{000}, необходимые для расчета коэффициентов, с которыми включаются в модель стандартные распределения.

Тесты на не вошедших в базисный набор белках с известной структурой показывают, что предложенная модель гистограммы удовлетворительно "предсказывает" гистограммы синтезов среднего и высокого разрешения. Что же касается низкого разрешения, то здесь модель становится излишне груба. Частично проблема предсказания гистограммы синтезов низкого разрешения снимается в ситуации, когда имеется гомолог исследуемого белка или известны (например, из электронной микроскопии) внешние очертания и размеры молекулы. В этом случае приемлемую гистограмму можно получить, "размещая" модель гомолога в элементарной ячейке исследуемого белка без самоналезаний и рассчитывая гистограмму, отвечающую такому вспомогательному объекту.

2.2. Двухкомпонентная модель гистограммы. Эмпирический подход.

Как было показано в предыдущей главе, форма гистограммы существенно зависит от разрешения синтеза и значений температурных параметров атомов, входящих в исследуемую молекулу. Поэтому задача, ответ на которую мы попытаемся найти, будет сформулирована так : предсказать гистограмму, отвечающую синтезу заданного разрешения d_{min}, если известно дополнительно, что значения температурных параметров атомов приблизительно одинаковы и равны данной величине B_{temp}.

Как и ранее, введем в рассмотрение нормированные частоты

$$\nu_k = n_k / (\Delta_k N), \qquad k=1,\ldots K, \tag{2.I}$$

(здесь по-прежнему n_k- число узлов сетки в элементарной ячейке со значениями функции $\rho_d(\mathbf{r})$, попадающими в k-ый бин, Δ_k- длина k-ого бина, N - общее число узлов сетки) и плотность кумулятивной функции

$$\nu(t) = \lim_{\Delta\to 0} \frac{\text{mes } \{ \mathbf{r} : t-\Delta/2 \le \rho_d(\mathbf{r}) \le t+\Delta/2 \}}{\Delta \text{ mes } \{V\}}, \tag{2.2}$$

которые, как и ранее в главе I, будем называть гистограммой, отвечающей синтезу $\rho_d(\mathbf{r})$. На Рис.2.I показаны гистограммы, отвечающие синтезам одного и того же разрешения 4A для разных белков (значения температурных параметров для всех атомов всех белков были взяты одинаковыми и равными $10A^2$). Из этого рисунка видно, что даже при одинаковом разрешении синтезов и значениях температурных параметров гистограммы для различных белков различны, и мы не можем непосредственно использовать гистограмму, отвечающую какому-то белку с

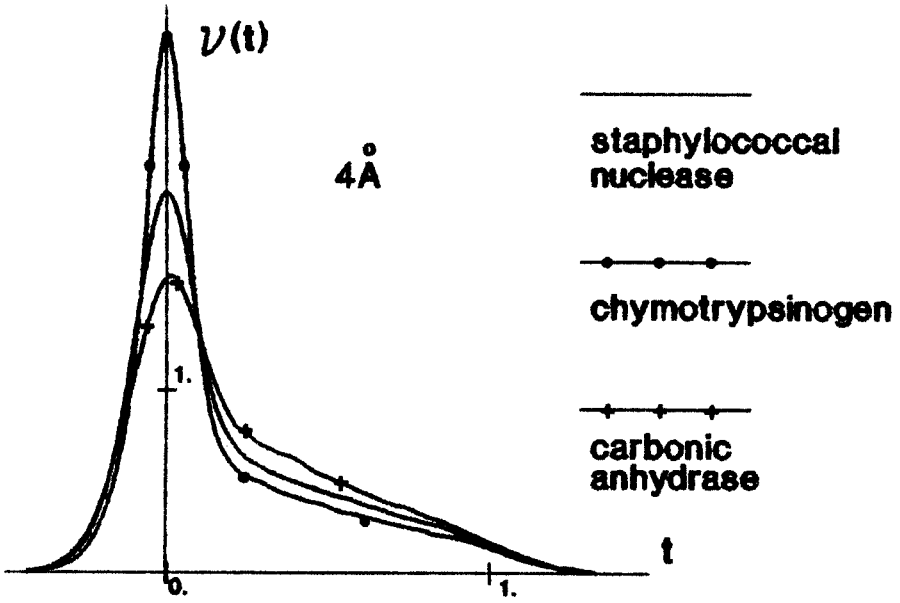


Рис.2.I. Гистограммы синтезов Фурье одинакового разрешения для разных белков.

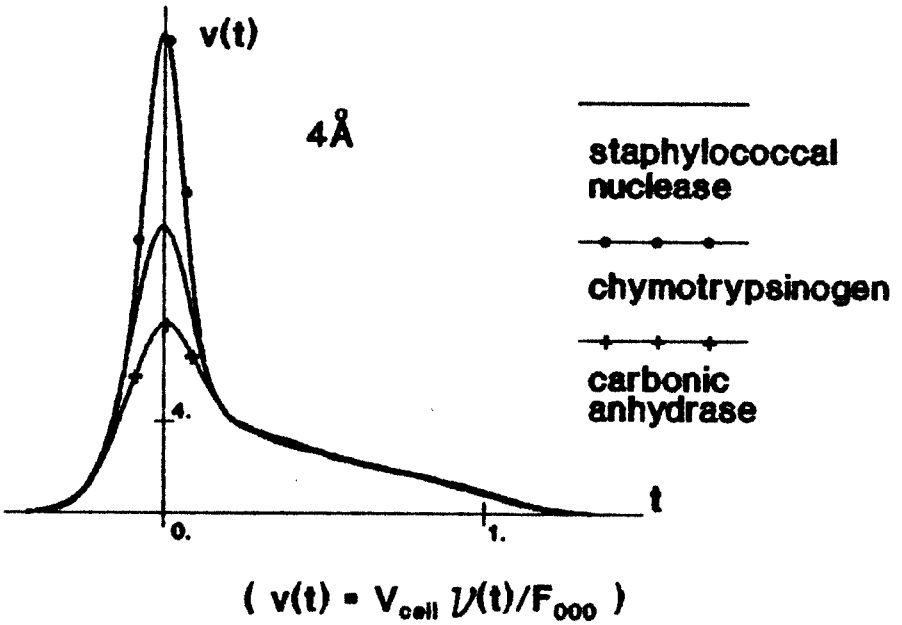


Рис.2.2. Нормализованные гистограммы, отвечающие тем же синтезам Фурье, что и на Рис.2.I.

известной структурой, в качестве эталонной гистограммы для синтеза электронной плотности другого белка.

Гистограммы для разных белков становятся гораздо более похожими друг на друга , если ввести дополнительную нормализацию гистограмм - перейти к "нормированным объемам"

$$v_k = \nu_k |V| \ / \ F_{ooo} \ , \qquad k=1,\ldots,K, \tag{2.3}$$

$$v(t) = \nu(t) \ |V| \ / \ F_{ooo} \ . \tag{2.4}$$

Здесь $|V|$ - объем и F_{ooo} - заряд элементарной ячейки. (При малых Δt, $v(t)\Delta t$ - это объем "приходящейся на один электрон заряда ячейки" области в элементарной ячейке, в которой значения $\rho_d(\mathbf{r})$ лежат в интервале $(t,t+\Delta t)$). Как видно из Рис.2.2, нормализованные гистограммы имеют похожие участки, соответствующие средним и большим значения ρ_d, также наименьшим значениям ρ_d. (Подчеркнем, что поскольку мы исследуем синтезы конечного разрешения, на этих синтезах обязательно будут области с отрицательными значениями $\rho_d(\mathbf{r})$ и, более того, эти области будут сосредоточены в основном в области молекулы [92]).

Рис. 2.2 позволяет высказать гипотезу , что "нормированный объем" $v(\rho)$ для значений ρ, встречающихся только в области молекулы, одинаков для всех белков и может быть описан стандартным (одинаковым для всех белков) распределением $v^o(t)$. При этом, конечно, распределение $v^o(t)$ меняется при изменении разрешения синтеза или значений температурных параметров атомов. Мы не можем определить непосредственно из графиков, приведенных на Рис.2.2, величины $v^o(t)$ для близких к нулю значений t. Такие значения встречаются не только в области молекулы, но и в "волнах обрыва ряда Фурье", и в кривой $v(t)$ содержится информация о

"смеси" таких значений. Однако более аккуратный анализ кривых $v(t)$ для разных белков позволит нам ниже определить значения стандартного распределения $v^o(t)$ для всех значений t.

Анализ функций $v(t)$ для разных белков позволяет заметить, что высота центрального пика оказывается прямо пропорциональной величине $|V|/F_{ooo}$, то есть чем "свободней" размещаются молекулы белка в кристаллической ячейке, тем болше высота этого пика. Это позволило выдвинуть гипотезу , что объем области в межмолекулярном пространстве элементарной ячейки, в которой значения функции $\rho_d(\mathbf{r})$ попадают в некоторый интервал $(t,t+\Delta t)$, прямо пропорционален объему межмолекулярного пространства.

Сформулированные выше гипотезы приводят к следующей эмпирической модели для распределения нормированных объемов

$$v(t) = v^o(t) + \left(|V|/F_{ooo} - \int_{-\infty}^{\infty} v^o(x)dx \right) q^o(t), \qquad (2.5)$$

или, что то же, к модели для гистограммы

$$\nu(t) = \frac{F_{ooo}}{|V|} v^o(t) + \left(1 - \frac{F_{ooo}}{|V|} \int_{-\infty}^{\infty} v^o(x)dx \right) q^o(t). \qquad (2.6)$$

Здесь $v^o(t)$ - одинаковое для всех белков распределение, описывающее распределение значений функции $\rho_d(\mathbf{r})$ внутри области молекулы, а $q^o(t)$ - одинаковое для всех белков распределение, описывающее распределение значений функции $\rho_d(\mathbf{r})$ в межмолекулярной области. (Фактически, $q^o(t)$ описывает распределение значений для "волн обрыва ряда Фурье".)

Дискретным аналогом формул (2.5)-(2.6) являются

$$v_k = v_k^o + \left(|V|/F_{ooo} - \sum_{j=1}^{K} v_j^o \Delta_j \right) q_k^o \qquad (2.7)$$

$$\nu_k = \frac{F_{ooo}}{|V|} v_k^o + \left(1 - \frac{F_{ooo}}{|V|} \sum_{j=1}^{K} v_j^o \Delta_j \right) q_k^o . \qquad (2.8)$$

2.3. Двухкомпонентная модель гистограммы. Идентификация параметров.

Формулы (2.7),(2.8) содержат не определенные пока стандартные распределения. Мы покажем сейчас, как эти распределения могут быть найдены, основываясь на белках, пространственная структура которых уже определена. После этого формулы (2.7)-(2.8) дадут возможность предсказывать гистограмму для новых белков, если только для этих белков известны значения $|V|$ и F_{ooo}.

Заметим, прежде всего, что из определения (2.2) функции $\nu(t)$ следует, что

$$\int_{-\infty}^{\infty} \nu(t)dt = 1 \quad , \quad \int_{-\infty}^{\infty} t\nu(t)dt = F_{ooo}/|V| \ , \qquad (2.9)$$

или дискретный аналог

$$\sum_{k=1}^{K} \nu_k \Delta_k = 1 \quad , \quad \sum_{k=1}^{K} t_k \nu_k \Delta_k = F_{ooo}/|V| \ . \qquad (2.10)$$

(Здесь t_k- середины бинов). Поэтому для того, чтобы модель (2.7) приближала гистограмму любого белка, необходимо, чтобы при всех значениях $|V|$ и F_{ooo} выполнялись равенства

$$\begin{aligned} &\sum_{k=1}^{K} \{ v_k^o + (|V|/F_{ooo} - \sum_{j=1}^{K} v_j^o \Delta_j) q_k^o \} \Delta_k = |V|/F_{ooo} \ , \\ &\sum_{k=1}^{K} t_k \{ v_k^o + (|V|/F_{ooo} - \sum_{j=1}^{K} v_j^o \Delta_j) q_k^o \} \Delta_k = 1. \end{aligned} \qquad (2.11)$$

Введя обозначения

$$\begin{aligned} S_v &= \sum_{k=1}^{K} v_k^o \Delta_k \ , \quad R_v = \sum_{k=1}^{K} t_k v_k^o \Delta_k , \\ S_q &= \sum_{k=1}^{K} q_k^o \Delta_k \ , \quad R_q = \sum_{k=1}^{K} t_k q_k^o \Delta_k \ , \end{aligned} \qquad (2.12)$$

мы можем переписать эти требования в виде

$$F_{ooo}S_v(1-S_q) + |V|(S_q-1) = 0 \quad ,$$
$$F_{ooo}(R_v-S_vR_q-1) + |V|R_q = 0 \quad . \qquad (2.13)$$

Выполнение этих равенств при всех F_{ooo} и $|V|$ возможно лишь в том случае, когда все коэффициенты при F_{ooo} и $|V|$ в этих уравнениях равны нулю. Это означает,что распределения $\{v_k^o\}_{k=1}^K$ и $\{q_k^o\}_{k=1}^K$ должны обладать свойствами

$$S_q = \sum_{k=1}^{K} q_k^o \Delta_k = I \ , \qquad (2.14)$$
$$R_v = \sum_{k=1}^{K} t_k v_k^o \Delta_k = I \quad , \quad R_q = \sum_{k=1}^{K} t_k q_k^o \Delta_k = 0 \quad .$$

Отметим, что система уравнений (2.14) никак не фиксирует значение S_v. Более того, переписав равенство (2.7) в виде

$$v_k = (v_k^o - S_v q_k^o) + (|V|/F_{ooo})\, q_k^o \quad , \qquad (2.15)$$

нетрудно видеть, что если равенства (2.14)-(2.15) выполняются при каких-то $\{v_k^o\}_{k=1}^K$ и $\{q_k^o\}_{k=1}^K$, то они будут по-прежнему выполняться при замене v_k^o на $v_k^o+\lambda q_k^o$ $(k=1,\ldots K)$ с любым λ. Это означает, что параметры $\{v_k^o\}_{k=1}^K$ и $\{q_k^o\}_{k=1}^K$ во введенной нами модели (2.7) могут быть взяты неоднозначно (при замене в (2.7) v_k^o на $v_k^o+\lambda q_k^o$ значения величин v_k, вычисляемых по этой формуле, будут теми же самыми). Для того, чтобы снять эту неоднозначность, мы можем фиксировать произвольным образом значение S_v, например, положив

$$S_v = \sum_{k=1}^{K} v_k^o \Delta_k = 0 \quad . \qquad (2.16)$$

Подчеркнем, что в отличие от условий (2.14), которые вытекают из общих свойств (2.10), условие (2.16) введено нами произвольно, чтобы фиксировать один из возможных наборов параметров $\{v_k^o\}_{k=1}^K$, $\{q_k^o\}_{k=1}^K$. С равным успехом мы могли бы потребовать равенства S_v любому другому числу.

Для нахождения параметров $\{v_k^o\}_{k=1}^K$ и $\{q_k^o\}_{k=1}^K$ в модели (2.7) мы отобрали из банка белковых молекул [91] набор из J (=15) базисных белков (их список дан в Табл.2.1.). Для каждого из базисных белков по координатам атомов были рассчитаны структурные факторы, построен синтез электронной плотности $\rho_d(\mathbf{r})$ разрешения 4Å и определены величины $\{v_k^{(j)}\}_{k=1}^K$ согласно (2.1)-(2.3). (Здесь верхний индекс (j) - порядковый номер белка в базисном наборе, k - номер бина). Величины $\{v_k^o\}_{k=1}^K$ и $\{q_k^o\}_{k=1}^K$ были определены из требования, чтобы модель

$$v_{k,t}^{(j)} = v_k^o + (|V|^{(j)}/F_{ooo}^{(j)} - \sum_{j=1}^{K} v_j^o \Delta_j)\, q_k^o \qquad (2.17)$$

наилучшим образом описывала гистограммы базисных белков при дополнительных условиях

$$\sum_{k=1}^{K} v_k^o \Delta_k = 0 , \qquad \sum_{k=1}^{K} q_k^o \Delta_k = 1 ,$$
$$\sum_{k=1}^{K} t_k v_k^o \Delta_k = 1 , \qquad \sum_{k=1}^{K} t_k q_k^o \Delta_k = 0 . \qquad (2.18)$$

Здесь $F_{ooo}^{(j)}$, $|V|^{(j)}$- заряд и объем элементарной ячейки для j-ого белка.

В качестве критерия соответствия теоретически рассчитываемых гистограмм $\{v_{k,t}^{(j)}\}_k$ гистограммам базисного набора $\{v_k^{(j)}\}_k$ использовалась величина

$$Q = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{N^{(j)} F_{ooo}^{(j)} \Delta_k}{|V|^{(j)}} \frac{(v_{k,t}^{(j)} - v_k^{(j)})^2}{v_k^{(j)}} . \qquad (2.19)$$

Здесь $N^{(j)}$- число точек сетки при расчете гистограммы для j-ого белка, весовые множители в критерии отвечают переходу от величин v_k к числам n_k узлов сетки, значения в которых попадают в k-ый бин. То есть, иными словами, использовался критерий

Таблица 2.I. Базовый набор белков.

Белок	File [a]	NRes [b]	Критерий соответствия реальной и модельной гистограмм	
			Q [c]	Q_g [d]
Carbonic anhydrase B	2CAB	261	1649	1.15
Chymotrypsinogen A	1CHG	245	1768	0.94
Cytochrome B5	2B5C	93	3500	0.93
HIPIP	1HIP	85	692	0.52
Bence- Jones protein	2PHE	114	1417	1.14
Insulin	1INS	21+ 30	4260	1.47
Lysozyme	1LZ1	130	1303	0.54
Ovomucoid third domain	1OVO	4* 56	1788	0.71
Phospholipase	1BP2	123	865	0.65
Plastocyanin	1PSY	99	1221	0.89
Prealbumin	2PAB	2* 127	1240	0.55
Proteinase A	2SGA	181	2603	1.29
Ribonuclease A	1RN3	124	2140	0.68
Staphylococ - cal nuclease	2SNS	149	7180	1.55
Myoglobin	1MBD	153	3641	1.99

a) Имя файла в Brookhaven Protein Data Bank.
b) Число остатков в асимметричной части ячейки.
c) Согласно (2.2I).
d) Согласно (2.23).

$$Q = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{k,t}^{(j)} - n_{k}^{(j)})^2}{n_{k}^{(j)}} . \qquad (2.20)$$

Для минимизации критерия (2.20) при дополнительных условиях (2.18) был применен метод множителей Лагранжа. В Таблице 2.2. приведены полученные в результате такой минимизации распределения $\{v_k^o\}_{k=1}^K$ и $\{q_k^o\}_{k=1}^K$. Графики функций $v^o(t)$ и $q^o(t)$ даны на Рис.2.3. На рис 2.4 показаны теоретические и "реальные" гистограммы, отвечающие некоторым из базисных белков.

2.4. Двухкомпонентная модель гистограммы. Точность предсказания гистограммы.

При решении задачи на условную минимизацию (2.18)-(2.20) было достигнуто значение Q_{min}= 0.36 * 10^5. Значения частных критериев соответствия "теоретических" и "реальных" гистограмм

$$Q^{(j)} = \sum_{k=1}^{K} \frac{(n_{k,t}^{(j)} - n_{k}^{(j)})^2}{n_{k,t}^{(j)}} . \qquad (2.21)$$

приведены в Таблице 2.1.

Полученные значения критериев Q и $Q^{(j)}$ показывают, что величины $n_k^{(j)}$ (или $n_{k,t}^{(j)}$) не являются правильными оценками средних квадратов отклонения реальных величин $n_k^{(j)}$ от теоретических $n_{k,t}^{(j)}$. Это отклонение вызвано в первую очередь грубостью модели (2.7), а не статистическим разбросом величин $n_k^{(j)}$ вокруг средних.

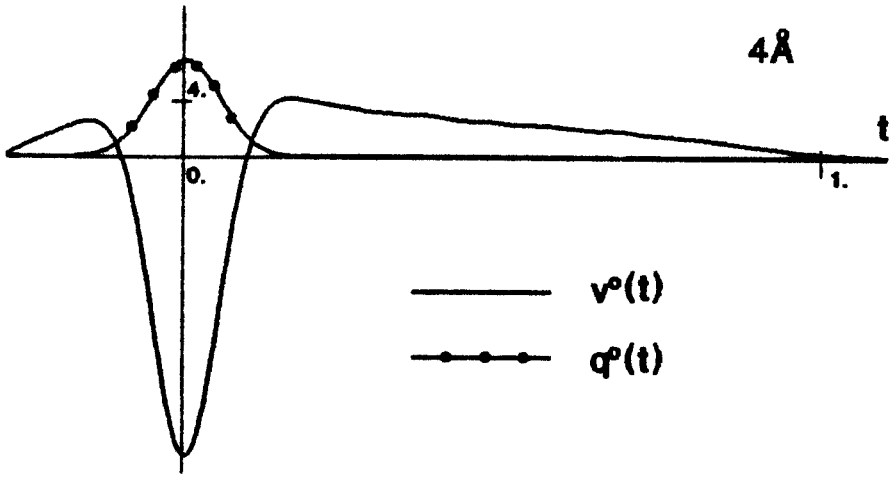
Некоторое представление о точности предсказания гистограмм моделью (2.7) можно получить, вычисляя среднеквадратичные погрешности предсказания значений частот

Таблица 2.2. Стандартные распределения, входящие в двухкомпонентную модель.

t^*	v^o	q^o	α^{**}	t^*	v^o	q^o	α^{**}
-0.450	0.12	-0.01	201.2	0.465	2.64	-0.01	10.6
-0.285	0.83	-0.05	48.9	0.495	2.55	-0.01	9.6
-0.255	1.31	-0.08	34.0	0.525	2.59	-0.05	7.7
-0.225	1.84	-0.08	35.5	0.555	2.31	-0.01	12.9
-0.195	2.37	-0.05	58.6	0.585	2.26	-0.02	8.9
-0.165	2.64	0.09	114.7	0.615	2.28	-0.04	6.4
-0.135	2.03	0.49	175.5	0.645	2.13	-0.03	6.5
-0.105	-1.04	1.48	166.5	0.675	2.02	-0.03	11.4
-0.075	-7.58	3.28	32.3	0.705	1.84	-0.01	6.3
-0.045	-16.41	5.56	68.0	0.735	1.92	-0.04	6.0
-0.015	-22.23	7.05	210.9	0.765	1.82	-0.04	7.9
0.015	-21.04	6.78	238.	0.795	1.58	-0.01	9.3
0.045	-13.84	4.99	66.4	0.825	1.52	-0.01	9.3
0.075	-4.93	2.72	35.6	0.855	1.38	-0.00	6.4
0.105	0.94	1.15	44.4	0.885	1.30	-0.00	8.7
0.135	3.55	0.34	54.9	0.915	1.15	0.00	7.4
0.165	4.26	0.04	31.0	0.945	1.02	0.00	8.3
0.195	4.10	-0.02	17.1	0.975	0.81	0.02	10.7
0.225	3.87	-0.03	12.3	1.005	0.69	0.02	19.4
0.255	3.67	-0.03	12.2	1.035	0.55	0.02	18.5
0.285	3.46	-0.02	11.5	1.065	0.34	0.04	19.8
0.315	3.20	-0.00	18.5	1.095	0.29	0.03	38.1
0.345	3.15	-0.02	16.0	1.125	0.19	0.02	36.4
0.375	2.98	-0.01	22.2	1.155	0.15	0.01	64.9
0.405	3.06	-0.04	11.2	1.185	0.06	0.02	57.6
0.435	2.86	-0.03	11.9	1.5	0.00	0.00	191.2

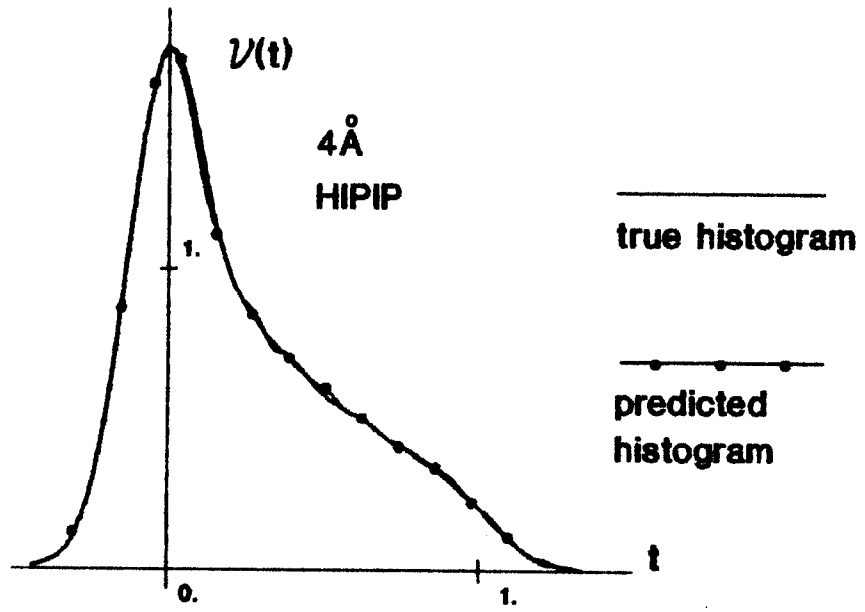
*) Величины t, отвечающие серединам бинов.

**) Характеристика точности модели, согласно (2.22).



$$\mathcal{V}(t) = (F_{000}/V_{cell})\, v^{o}(t) + q^{o}(t)$$

Рис.2.3. Стандартные распределения $v^{o}(t)$ и $q^{o}(t)$, отвечающие разрешению 4А, определенные по I5 базисным белкам.



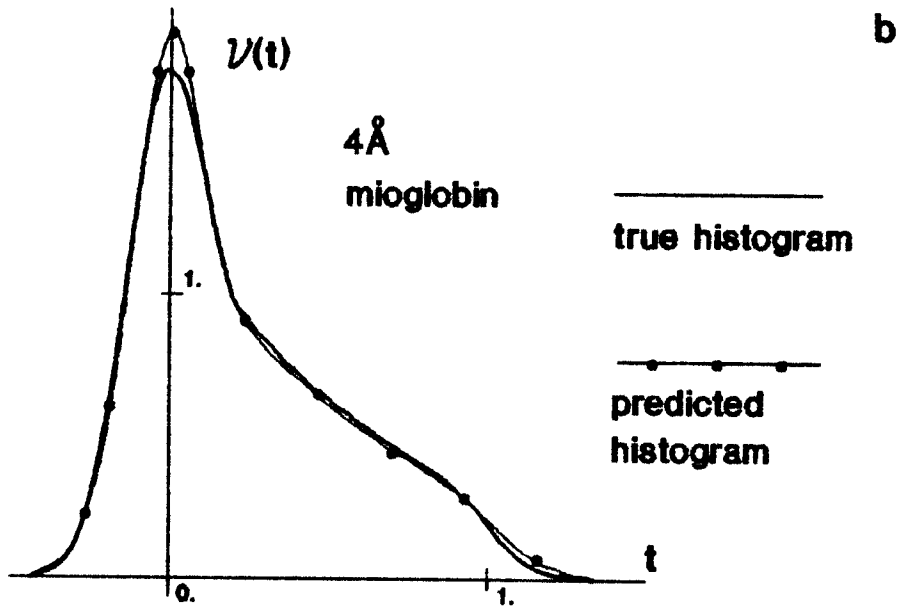


Рис.2.4. Точность предсказания гистограмм синтезов Фурье при помощи двухкомпонентной модели.

по базовому набору белков. Определим поправочные коэффициенты

$$\alpha_k = \frac{1}{J}\sum_{j=1}^{J} \frac{(n_{k,t}^{(j)} - n_k^{(j)})^2}{n_{k,t}^{(j)}}, \qquad k=1,\dots,K. \qquad (2.22)$$

и введем критерий близости теоретической и рассчитанной гистограмм в виде

$$Q_g = \frac{1}{K}\sum_{k=1}^{K} \frac{(n_{k,t} - n_{k,o})^2}{\alpha_k \; n_{k,t}} = \qquad (2.23)$$

$$= \frac{1}{K}\sum_{k=1}^{K} \frac{N\,\Delta_k}{\alpha_k} \frac{(\nu_{k,t} - \nu_{k,o})^2}{\nu_{k,t}}$$

В таком случае среднее значение этого критерия для базисных белков будет равно 1 и все пробные наборы структурных факторов, приводящие к значениям критерия Q_g, не превосходящим 1, следует считать не противоречащими гистограмме $\{\nu_{k,t}\}$. Напротив, если какой-то набор структурных факторов привел к гистограмме $\{\nu_{k,o}\}$, для которой значение критерия (2.23) заметно больше 1, то такой набор структурных факторов следует считать противоречащим гистограмме $\{\nu_{k,t}\}$. Значения поправочных коэффициентов даны в таблице 2.2. В Таблице 2.1 приведены значения критериев (2.21) и (2.23) для белков базисного набора.

В некоторых ситуациях более удобным оказывается использование другого критерия близости гистограмм

$$Q_h = \sum_{k=1}^{K} |\, \nu_{k,t} - \nu_{k,o} \,|, \qquad (2.24)$$

или его непрерывного аналога

$$Q_h = \int_{-\infty}^{\infty} |\, \nu_t(t) - \nu_o(t) \,|\, dt. \qquad (2.25)$$

В таблице 2.1 приведены значения этого критерия для базисных белков. Критерий (2.24) имеет те преимущества, что он

имеет более простую структуру (не содержит весовых множителей) и не так чувствителен к изменению разбиения на бины, как критерий (2.23). Его недостаток - он содержит операцию взятия модуля, что делает его неудобным в задачах, требующих вычисления производных.

Заметим еще, что, формально говоря, функция $\nu(t)$, рассчитываемая по формуле (2.6), может принимать и отрицательные значения (в силу неизбежных погрешностей аппроксимации). Однако на практике эти значения близки к нулю, и для получения физически осмысленных значений функции $\nu(t)$ они должны быть заменены нулем.

2.5. Анализ размерности множества гистограмм в k-мерном пространстве.

Трактовка гистограмм базисных белков как точек многомерного пространства позволяет с другой стороны взглянуть на проблему построения модели, описывающей гистограммы.

2.5.I. Формулировка задачи.

Будем рассматривать гистограмму $\nu(t)$ как "точку" некоторого пространства функций $\mathcal{L}$, например, пространства L_2 функций со скалярным произведением

$$(f_1, f_2) = \int_{-\infty}^{\infty} f_1(t)\, f_2(t)\, dt$$

[5]. (Дискретный аналог - будем рассматривать набор частот $\{\nu_k\}_{k=1}^{K}$ как точку K-мерного координатного пространства.) Множество базисных гистограмм $\{h^{(j)}\}_{j=1}^{J}$ представляет собой в

такой трактовке некое множество точек $H = \{ \mathbf{h}^{(j)} \}_{j=1}^{J}$ в пространстве $\mathscr{L}$, и наличие модели (2.6) означает, что все точки множества H лежат вблизи одномерного линейного многообразия (прямой) в пространстве $\mathscr{L}$. Это многообразие представляет из себя множество точек вида

$$\mathfrak{M}^{(1)} = \{ \mathbf{a} + \alpha\, \mathbf{e} \}_{\alpha \in R} , \tag{2.26}$$

где $\mathbf{a}$ - "точка" пространства $\mathscr{L}$, отвечающая распределению

$$(F_{ooo}/|V|) \; (v^{o}(t) - \int_{-\infty}^{\infty} v^{o}(\tau)\, d\tau \; q^{o}(t)),$$

а $\mathbf{e}$ отвечает распределению $q^{o}(t)$.

Справедливо и обратное: если все точки множества H лежат вблизи некоторой прямой пространства $\mathscr{L}$, то существует линейная аппроксимация элементов множества H

$$h^{(j)} \cong \mathbf{a} + \alpha_j\, \mathbf{e} . \tag{2.27}$$

Мы можем рассмотреть и более общую задачу - попытаться найти приближенную формулу для гистограмм в виде линейной комбинации не двух, а большего числа стандартных распределений, то есть попытаться найти линейное многообразие

$$\mathfrak{M}^{(m)} = \{ \mathbf{a} + \sum_{i=1}^{m} \alpha_i \mathbf{e}^{i} \}_{\alpha_1, \alpha_2, \ldots, \alpha_m \in R} \tag{2.28}$$

большей размерности m, такое, что все точки множества H лежат вблизи этого многообразия. Качество приближения можно оценить суммой квадратов отклонений точек $\mathbf{h}^{(j)}$ от многообразия $\mathfrak{M}^{(m)}$:

$$Q = \sum_{j=1}^{J} \| \mathbf{h}^{(j)} - \mathbf{a} - \sum_{i=1}^{m} (\mathbf{h}^{(j)} - \mathbf{a}, \mathbf{e}^{i})\, \mathbf{e}^{i} \|^{2} . \tag{2.29}$$

Для многообразия, наилучшим образом приближающего множество H, эта величина есть

$$Q_{opt}^{(m)} = \min_{\substack{\mathbf{a},\mathbf{e}^1,\dots\mathbf{e}^m \\ (\mathbf{e}^i,\mathbf{e}^j)=\delta_{ij}}} \sum_{j=1}^{J} \| \mathbf{h}^{(j)} - \mathbf{a} - \sum_{i=1}^{m} (\mathbf{h}^{(j)} - \mathbf{a}, \mathbf{e}^i)\, \mathbf{e}^i \|^2, \tag{2.30}$$

где минимум берется по всем $\mathbf{a}$ и всем ортонормированным наборам векторов $(\mathbf{e}^1,\dots,\mathbf{e}^m)$. Исчерпывающий ответ на вопрос о возможностях линейной аппроксимации множества H дает вычисление цепочки значений $Q_{opt}^{(1)}$, $Q_{opt}^{(2)}$, ... $Q_{opt}^{(J)}$, характеризующей изменение точности аппроксимации по мере увеличения размерности приближающего многообразия $\mathfrak{M}_{opt}^{(m)}$. (Ясно, что $Q_{opt}^{(J)} = 0$.)

Как будет показано в разделе 6.2, величины $Q_{opt}^{(m)}$ могут быть вычислены по формуле

$$Q_{opt}^{(m)} = \sum_{j=1}^{J} \| \mathbf{h}^{(j)} - \mathbf{h}^{av} \|^2 - \mu_1 - \mu_2 - \dots - \mu_m, \tag{2.31}$$

где $\mu_1 \geq \mu_2 \geq \dots \geq \mu_J$ – расположенные в порядке убывания собственные значения матрицы G, составленной из элементов

$$g_{ij} = (\mathbf{h}^{(i)} - \mathbf{h}^{av}, \mathbf{h}^{(j)} - \mathbf{h}^{av}) \quad , \tag{2.32}$$

$$\mathbf{h}^{av} = (\sum_{j=1}^{J} \mathbf{h}^{(j)}) / J \quad . \tag{2.33}$$

При этом в качестве базиса $(\mathbf{e}^1,\dots,\mathbf{e}^m)$ оптимального многообразия должны быть взяты вектора вида

$$\mathbf{e}^i = \sum_{k=1}^{J} \varepsilon_k^i\, \mathbf{e}^k, \qquad i=1,\dots,m \quad , \tag{2.34}$$

где $\{\varepsilon_k^i\}_{k=1}^{J}$ удовлетворяют системам уравнений

$$\sum_{k=1}^{J} g_{nk}\varepsilon_k^i = \mu_i \varepsilon_n^i \ , \quad n=1,\dots J \ , \tag{2.35}$$

отвечающим собственным значениям μ_1, μ_2, ... , μ_m. Собственные значения μ_i характеризуют степень "вытянутости" множества H по направлению $\mathbf{e}^i$, и $\mathbf{e}^1$ есть направление наибольшей "вытянутости" множества H.

Альтернативный подход (в случае, когда пространство $\mathcal{X}$

имеет конечную размерность K) – вычисление по формуле

$$Q_{opt}^{(m)} = \sum_{j=1}^{J} \| \mathbf{h}^{(j)} - \mathbf{h}^{av} \|^2 - \lambda_1 - \lambda_2 - \ldots - \lambda_m, \qquad (2.36)$$

где $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_J$ – расположенные в порядке убывания собственные значения матрицы B, составленной из элементов

$$b_{mn} = \sum_{j=1}^{J} x_m^j \, x_n^j \quad , \quad m,n=1,\ldots K. \qquad (2.37)$$

Здесь $\{x_k^j\}_{k=1}^{K}$ – координаты вектора $\mathbf{x}^j$ в некотором заданном ортонормированном базисе пространства $\mathscr{X}$. Собственные вектора матрицы B дают при этом координаты в этом же базисе векторов $\mathbf{e}^1,\ldots,\mathbf{e}^m$, определяющих оптимальное многообразие $\mathfrak{M}_{opt}$.

2.5.2. Примеры. Поиск аппроксимирующих многообразий.

Изложенный выше подход был применен к анализу множества гистограмм, отвечающих разрешениям 4A и 10A. В Таблице 2.3 дан перечень базисных белков, использованных для этой цели. В таблицах 2.4 и 2.5 приведены величины собственных значений $\lambda_1, \lambda_2, \ldots, \lambda_J$, характеристик качества приближения $Q_{opt}^{(1)}, Q_{opt}^{(2)}, \ldots Q_{opt}^{(J)}$ и "коэффициентов качества" μ_m, описывающих "относительную" точность приближения множества H многообразием $\mathfrak{M}_{opt}^{(m)}$:

$$\mu_m = \frac{\lambda_1+\lambda_2+\ldots+\lambda_m}{\lambda_1+\lambda_2+\ldots\ldots\ldots+\lambda_J} =$$

$$= (\lambda_1+\lambda_2+\ldots+\lambda_m) \Big/ \sum_{j=1}^{J} \|\mathbf{h}^{(j)} - \mathbf{h}^{av}\|^2. \qquad (2.38)$$

В таблицах 2.6 и 2.7 приведены характеристики точности приближения отдельных базисных гистограмм элементами оптимальных приближающих многообразий. Приведены значения погрешностей аппроксимации в нормах

Таблица 2.3. Базисные белки для определения размерности множества гистограмм.

1. Carbonic anhydrase.
2. Chymotrypsinogen.
3. Cytochrome .
4. HIPIP.
5. B-J Protein.
6. Insulin.
7. Lysozyme.
8. Mioglobin.
9. Neurotoxin.
10. Ovomucoid.
11. Phospholipase.
12. Plastocyanin.
13. Prealbumin.
14. Proteinase A.
15. Ribonuclease.
16. Staphilococcal nuclease.
17. Ubiquitin.
18. Crambin.
19. Avian pancreatic polypeptide.
20. Rubredoxine.
21. Concanavalin.

Таблица 2.4. Аппроксимация множества гистограмм синтезов разрешения 4А линейными многообразиями.

Размерность многообразия $\mathfrak{M}^{(m)}_{opt}$	Собственное значение λ_m	Коэффициент качества (2.38)	Относительная точность приближения Q_{rel}
I	28.15	0.970	0.029
2	0.5156	0.988	0.012
3	0.1400	0.993	0.007
4	0.05961	0.995	0.005
5	0.05497	0.997	0.003
6	0.02378	0.998	0.002
7	0.01529	0.998	0.002
8	0.01243	0.999	0.001
9	0.00910	0.999	0.001
10	0.00731	0.999	0.001

$$Q_{rel} = Q^{(m)}_{opt} \Big/ \sum_{j=1}^{J} \|\mathbf{h}^{(j)} - \mathbf{h}^{av}\|^2$$

Таблица 2.5. Аппроксимация множества гистограмм синтезов разрешения IOА линейными многообразиями.

Размерность многообразия $\mathfrak{M}^{(m)}_{opt}$	Собственное значение λ_m	Коэффициент качества (2.38)	Относительная точность приближения
I	53.50	0.722	0.278
2	12.67	0.893	0.107
3	3.263	0.937	0.062
4	1.621	0.959	0.041
5	1.017	0.973	0.027
6	0.8008	0.982	0.016
7	0.3561	0.989	0.011
8	0.2602	0.992	0.007
9	0.1989	0.995	0.005
10	0.1559	0.997	0.003

$$Q_{rel} = Q^{(m)}_{opt} \Big/ \sum_{j=1}^{J} \|\mathbf{h}^{(j)} - \mathbf{h}^{av}\|^2$$

Таблица 2.6. Точность приближения гистограмм элементами линейных многообразий (разрешение 4А).

N п/п [a]	размерность m приближающего многообразия $\mathfrak{M}^{(m)}$							
	m = 1		m = 2		m = 3		m = 4	
	Δ_1	Δ_2 [б]	Δ_1	Δ_2	Δ_1	Δ_2	Δ_1	Δ_2
1	.021	.022	.021	.022	.014	.014	.013	.013
2	.019	.023	.013	.015	.011	.011	.007	.007
3	.046	.055	.013	.013	.013	.013	.012	.012
4	.024	.025	.023	.024	.021	.021	.014	.013
5	.026	.030	.024	.027	.015	.014	.015	.014
6	.023	.023	.023	.022	.022	.021	.022	.020
7	.017	.017	.016	.016	.015	.015	.012	.013
8	.027	.028	.024	.022	.024	.022	.023	.020
9	.046	.042	.045	.041	.032	.027	.032	.026
10	.022	.024	.012	.013	.011	.011	.010	.010
11	.021	.023	.017	.017	.015	.014	.015	.014
12	.028	.029	.028	.027	.013	.012	.013	.012
13	.014	.015	.010	.010	.010	.010	.009	.010
14	.022	.023	.022	.022	.020	.019	.010	.009
15	.044	.050	.013	.012	.009	.009	.010	.009
16	.046	.060	.018	.019	.008	.007	.007	.007
17	.035	.037	.032	.031	.024	.023	.010	.010
18	.022	.022	.020	.020	.020	.019	.019	.019
19	.038	.035	.028	.026	.026	.023	.025	.023
20	.044	.043	.026	.024	.023	.023	.014	.014
21	.046	.057	.017	.021	.012	.014	.010	.011

а) согласно таблице 2.3;
б) согласно (2.39) .

Таблица 2.7. Точность приближения гистограмм элементами линейных многообразий (разрешение IOA).

N п/п [a]	размерность m приближающего многообразия $\mathfrak{M}^{(m)}$							
	m = 1		m = 2		m = 3		m = 4	
	Δ_1	Δ_2 [б]	Δ_1	Δ_2	Δ_1	Δ_2	Δ_1	Δ_2
1	.093	.130	.054	.081	.029	.037	.028	.037
2	.045	.058	.045	.058	.034	.048	.024	.031
3	.078	.111	.080	.101	.079	.101	.057	.075
4	.096	.140	.096	.140	.054	.084	.054	.084
5	.077	.107	.064	.089	.059	.082	.053	.072
6	.042	.521	.074	.096	.031	.035	.028	.031
7	.054	.072	.048	.068	.036	.052	.036	.050
8	.034	.045	.028	.037	.027	.036	.026	.034
9	.192	.257	.105	.139	.094	.123	.095	.123
10	.102	.145	.101	.134	.094	.130	.059	.072
11	.086	.111	.077	.105	.044	.056	.043	.055
12	.068	.104	.062	.092	.063	.087	.049	.069
13	.063	.084	.063	.082	.046	.066	.042	.057
14	.129	.171	.079	.110	.053	.072	.037	.050
15	.111	.130	.079	.092	.073	.091	.051	.060
16	.070	.099	.063	.096	.059	.087	.044	.052
17	.142	.186	.133	.178	.029	.039	.029	.039
18	.069	.089	.069	.089	.064	.085	.064	.085
19	.100	.138	.090	.115	.058	.077	.059	.076
20	.135	.177	.117	.146	.098	.137	.060	.078
21	.072	.096	.057	.080	.047	.067	.045	.065

а) согласно таблице 2.3;
б) согласно (2.39) .

$$\| \Delta \|_2 = (\int [\Delta(t)]^2 dt)^{1/2}$$
$$\| \Delta \|_1 = \int | \Delta(t) | dt \quad , \qquad (2.39)$$

где $\Delta(t)$ обозначает разность между рассчитанной и точной гистограммами.

2.5.3. Связь коэффициентов аппроксимации с параметрами исследуемого объекта.

Нахождение линейного многообразия, наилучшим образом приближающего базисный набор гистограмм, не решает еще проблемы предсказания гистограмм. Для решения этой проблемы надо еще научиться определять коэффициенты разложения гистограммы по базисным векторам многообразия. Эта задача решена в настоящее время удовлетворительно лишь для случая аппроксимации линейным многообразием. В этом случае удалось установить, что коэффициент α в разложении гистограммы

$$\mathbf{h} = \mathbf{a} + \alpha\, \mathbf{e} \qquad (2.40)$$

является (с разумной точностью) монотонной функцией от величины $F_{ooo}/|V|$ и, значит, может быть приближенно определен для нового белка как $\phi(F_{ooo}/|V|)$, где $\phi(x)$ – некоторая эмпирическая функция, определенная по коэффициентам $\alpha^{(j)}$, отвечающим базисным белкам. В таблицах 2.8 и 2.9 приведены сведения о точности приближения гистограмм в случаях, когда в качестве функции $\phi(x)$ берется линейная или квадратичная функция. При этом при проведении расчетов в качестве базового был взят набор из 16 гистограмм (они идут первыми в таблицах), а оставшиеся пять гистограмм "предсказывались" только на основе величины $F_{ooo}/|V|$ и гистограмм 16 базовых белков.

Таблица 2.8. Точность приближения гистограмм двухкомпонентной моделью. (Разрешение 4А).

N п/п [а]	$\frac{F_{000}}{V}$	приближение гистограммы одномерным линейным многообразием		линейная аппроксимация величин α		квадратичная аппроксимация величин α	
		α^* [б]	Δ_1^* [в]	$\lvert\alpha-\alpha^*\rvert$ [г]	Δ_1 [д]	$\lvert\alpha-\alpha^*\rvert$ [е]	Δ_1 [ж]
белки базисного набора							
1	.237	−0.86	.022	.017	.022	.007	.022
2	.170	1.97	.022	.247	.033	.018	.022
3	.205	0.003	.039	.398	.053	.201	.041
4	.241	−1.09	.022	.102	.028	.134	.031
5	.181	1.25	.023	.051	.023	.075	.024
6	.242	−0.99	.025	.034	.024	.012	.025
7	.239	−0.81	.023	.094	.025	.095	.026
8	.235	−1.07	.025	.320	.056	.262	.049
9	.257	−1.07	.050	.544	.088	.186	.056
10	.167	2.07	.022	.228	.030	.071	.025
11	.223	−0.26	.019	.026	.019	.198	.031
12	.244	−1.12	.032	.008	.032	.092	.035
13	.186	1.22	.016	.126	.020	.177	.025
14	.201	0.63	.021	.074	.023	.254	.039
15	.214	−0.15	.036	.208	.038	.002	.036
16	.201	0.27	.039	.269	.044	.088	.039
белки, не входившие в базисный набор							
17	.252	−1.53	.040	.130	.045	.360	.067
18	.256	−1.29	.031	.288	.044	.045	.032
19	.256	−1.42	.049	.158	.051	.176	.061
20	.269	−1.52	.055	.536	.076	.139	.060
21	.192	1.35	.050	.477	.064	.594	.074

а) согласно таблице 2.3;
б) коэффициент в разложении (2.40);
в) погрешность (2.39) приближения гистограммы элементом оптимального одномерного линейного многообразия;
г) погрешность при вычислении α как линейной функции от F_{000}/V;
д) точность модельной гистограммы при вычислении α как линейной функции от F_{000}/V;
е) погрешность при вычислении α как квадратичной функции от F_{000}/V;
ж) точность модельной гистограммы при вычислении α как квадратичной функции от F_{000}/V;

Таблица 2.9. Точность приближения гистограмм двухкомпонентной моделью. (Разрешение IOA).

N п/п [а]	$\frac{F_{000}}{V}$	приближение гистограммы одномерным линейным многобразием		линейная аппроксимация величин α		квадратичная аппроксимация величин α	
		α^* [б]	Δ_1^* [в]	$\lvert\alpha-\alpha^*\rvert$ [г]	Δ_1 [д]	$\lvert\alpha-\alpha^*\rvert$ [е]	Δ_1 [ж]
белки базисного набора							
1	.237	-0.97	.095	.144	.098	.181	.100
2	.170	2.06	.044	.208	.051	.553	.079
3	.205	-0.13	.072	.656	.106	.359	.088
4	.241	-1.70	.121	.397	.135	.444	.138
5	.181	1.61	.076	.100	.074	.137	.074
6	.242	-1.29	.409	.062	.408	.008	.409
7	.239	-0.97	.054	.223	.064	.226	.064
8	.235	-1.03	.039	.045	.039	.043	.039
9	.257	-1.62	.200	.500	.212	.039	.200
10	.167	3.35	.092	.931	.145	.481	.116
11	.223	-0.06	.083	.315	.091	.575	.107
12	.244	-1.25	.086	.209	.090	.083	.086
13	.187	1.67	.061	.225	.063	.302	.064
14	.201	0.80	.130	.065	.130	.336	.141
15	.214	-1.03	.100	.110	.162	.800	.142
16	.201	0.55	.065	.161	.066	.113	.068
белки, не входившие в базисный набор							
17	.252	-2.33	.168	.488	.176	.836	.202
18	.256	-2.23	.076	.166	.076	.669	.116
19	.256	-2.09	.129	.021	.128	.524	.135
20	.269	-2.06	.153	.646	.170	.372	.158
21	.192	1.44	.075	.296	.073	.472	.081

а) согласно таблице 2.3;
б) коэффициент в разложении (2.40);
в) погрешность (2.39) приближения гистограммы элементом оптимального одномерного линейного многообразия;
г) погрешность при вычислении α как линейной функции от F_{000}/V ;
д) точность модельной гистограммы при вычислении α как линейной функции от F_{000}/V;
е) погрешность при вычислении α как квадратичной функции от F_{000}/V ;
ж) точность модельной гистограммы при вычислении α как квадратичной функции от F_{000}/V;

2.6.Предсказание гистограмм для синтезов низкого разрешения.

Из таблиц 2.8 и 2.9 видно, что качество предсказания гистограмм при помощи линейной модели при низком разрешении становится "неравномерным" - в некоторых случаях предсказание дает разумную точность, а в некоторых (гистограммы 6, 9 в Таблице 2.9) предсказанные гистограммы плохо совпадают с реальными. Качество предсказания становится еще хуже при дальнейшем понижении разрешения. Указанные обстоятельства заставляют искать альтернативные пути предсказания гистограмм для исследуемых белков при низком разрешении.

Дополнительным подспорьем в предсказании гистограмм может служить наличие гомолога с известной пространственной структурой. В этом случае задача предсказания гистограммы может быть решена в два этапа :

а) "размещение" атомной модели гомолога без самоналезаний в элементарной ячейке исследуемого белка;

б) расчет гистограммы, отвечающей такой гипотетической модели.

Тесты показали, что рассчитываемые таким образом гистограммы слабо зависят от возможных различных способов упаковки модели и могут быть использованы в практической работе.

Близкий подход может быть предложен в случае, когда отсутствуют известные гомологи, но внешние очертания и размеры молекулы известны, например, из электронно-микроскопических исследований. В этом случае мы можем пытаться "собрать" молекулу нужной формы из подходящих по геометрическим размерам кусков известных структур и использовать такое искуственное образование в

качестве замены гомолога. Такой подход был применен нами при исследовании структуры фактора элонгации G (см.раздел 5.5.).