В.Ю.Лунин
Введение в кристаллографию макромолекул

Эти три лекции были прочитаны студентам второго курса факультета биоинженерии и биоинформатики МГУ осенью 2006. Цель лекций - неформальный рассказ о том, каким образом возникает информация о структуре белков, представленная в Protein Data Bank. Лекции не предполагали знаний математики, физики и химии, выходящих за пределы школьной программы.

Данный текст является записью звуковых комментариев к содержанию диапозитивов файла Lunin_MSU06_Pictures_3.pdf. Запись сделана в процессе подготовки лекций. Она не подвергалась литературной обработке и не является записью собственно лекции. Цель записи - обозначить круг вопросов, планируемых к обсуждению при показе того или иного слайда.

Цифры в тексте дают ссылку на номер слайда, соответствующего данному куску текста.

Лекция 3.

1

Основная идея рентгеновского эксперимента

2

На прошлых лекциях мы говорили о том, что рентгеновский эксперимент заключается в том, что образец исследуемого вещества помещается в пучок рентгеновских лучей, и мы пытаемся измерить интенсивность электромагнитных волн, распространяющихся от образца во всевозможных направлениях. Мы говорили, что причина появления новых электромагнитных волн в направлениях, отличающихся от направления падающего пучка, может быть описана достаточно просто. Падающий рентгеновский пучок представляет из себя электромагнитную волну. Под влиянием этой электромагнитной волны электроны, входящие в состав исследуемого вещества, приходят в движение, начинают осциллировать. Каждый осциллирующий электрон становится источником новой электромагнитной волны, распространяющейся уже во все стороны. Эти новые волны, излучаемые каждым электроном, складываются. И поэтому, если мы берем какое-то направление, мы можем пытаться измерить интенсивность волны, распространяющейся в этом направлении. Проблема заключается в том, что интенсивности таких рассеянных волн чрезвычайно малы, даже если мы используем очень мощные рентгеновские пучки. Существует средство радикально усилить интенсивность этих рассеянных волн. Оно заключается в том, что исследуемый объект приготавливается в виде монокристалла. В монокристалле мы имеем большое количество одинаковых молекул, расположенных в правильном порядке в узлах трехмерной кристаллической решетки в одинаковой ориентации. И такое большое количество регулярно уложенных одинаковых молекул позволяет существенно усилить интенсивность рассеянных волн. Но, правда, не всех рассеянных волн, а только волн, распространяемых в направлениях, которые подчиняются

определенным условиям, которые называются условиями Лауэ. Здесь эти условия выписаны.
Здесь сигма нулевое - это единичный вектор, показывающий направление падающего пучка; сигма - это вектор, задающий направление на детектор, которым мы меряем интенсивность рассеянной волны. Ну, и разность вот этих векторов, деленная на длину волны, называется вектором рассеяния, и это важная величина, она постоянно присутствует в различных формулах, которые постоянно возникают в теории рентгеноструктурного анализа.
Условия, при которых возникает такая усиленная электромагнитная волна, интенсивность которой мы можем измерить, заключаются в том, что проекция этого вектора рассеяния на базисные вектора элементарной ячейки должны являться целыми числами.

3

Если мы выпишем уравнения падающей и рассеянной электромагнитных волн, мы увидим, что уравнение рассеянной электромагнитной волны отличается от первичной волны только множителем F в амплитуде волны и сдвигом их фазы. Это мы получили формулу, каким образом это этот множитель в амплитуде и фаза связаны с распределением электронной плотности в исследуемом образце, и выяснили, что эта связь дается такими вот формулами. При этом величины F и фи называются модулем и фазой структурного фактора, и то, что мы в состоянии померить в эксперименте - это интенсивность, которая пропорциональна квадрату модуля структурного фактора. То есть, вообще говоря, из стандартного рентгеновского эксперимента мы можем извлечь значения модулей этих структурных факторов, но при этом теряются значения фаз. Тем не менее, если мы знаем значения модулей структурных факторов, то мы можем использовать эти формулы для проверки различных гипотез о том, как устроен наш объект, и для выбора из них наиболее подходящей гипотезы. Если у нас есть какая-то гипотеза, как устроен объект, мы можем, тем самым, сказать, как в нем распределена электронная плотность, и мы можем посчитать соответствующие этой гипотезе значения модулей структурных факторов и сравнить их с экспериментом. Если совпадение хорошее, то, значит, это вполне хорошая гипотеза. А если совпадение плохое - значит, надо выдвигать какие-то другие гипотезы о том, как наш объект устроен.

4

На предыдущей лекции я говорил о том, что любая функция трех переменных может быть единственным образом представлена в виде ряда Фурье. То есть представлена в виде таких стандартных функций косинусов с коэффициентами - множителями F и со сдвигами фазы φ. При этом, если мы знаем функцию ρ, то эти коэффициенты легко вычисляются, для них есть известные формулы, они здесь написаны. И, как мы можем заметить, эти формулы в точности совпадают с теми формулами, которые я на предыдущем слайде выписал для вычисления модуля и фазы структурного фактора рассеянной волны.

5

Осознание того факта, что это одни и те же формулы, позволило коренным образом изменить подход к определению структуры методами рентгеноструктурного анализа. Наличие такой связи приводит к тому, что, если мы знаем модули и фазы структурных факторов, то мы можем рассчитать распределение электронной плотности в изучаемом объекте. Просто

просуммировав этот ряд, мы можем для любой точки пространства посчитать, какое было значение электронной плотности в этой точке.
Рассчитанная таким образом сумма называется синтезом Фурье электронной плотности. Название "синтез" используется для того, чтобы подчеркнуть, что это не совсем настоящее распределение электронной плотности. Потому что в настоящем распределении электронной плотности, если мы его представляем в виде ряда Фурье, присутствует бесконечное количество таких слагаемых. А в реальной жизни, естественно, когда мы рассчитываем синтез Фурье, мы используем только имеющиеся в нашем распоряжении значения модулей и фаз структурных факторов, это всегда конечное число таких коэффициентов (хотя и, может быть, очень большое). И поэтому то, что мы получаем таким расчетом - это не совсем настоящее распределение электронной плотности, а некоторое приближение к этому распределению.

Разрешение

6

Чтобы поговорить более подробно о том, насколько хорошо это приближение совпадает с настоящим распределением, я должен остановиться на очень важном понятии, о котором мы уже говорили прошлый раз - понятии разрешения синтеза Фурье. Но сначала я начну с понятия разрешения, отвечающего гармонике Фурье. Одна такая косинусоида в этом разложении называется гармоникой Фурье. И если мы посмотрим, как эта функция зависит от вектора рассеяния s и зависит от вектора пространства R, если мы рассмотрим какое-то конкретное значение s, выберем направление, в котором мы меряем рассеяние, то эта функция является функцией трех переменных в пространстве. Но на самом деле эта функция устроена таким достаточно специальным образом.
Если мы возьмем направление s, то вдоль направления s эта функция просто меняется, как график функции косинус. Если мы возьмем теперь любую плоскость, перпендикулярную направлению s, то в этой плоскости значение функции не меняется, функция сохраняет одно и то же значение во всех точках плоскости, перпендикулярной направлению s. Поэтому, хоть формально такая гармоника Фурье - трехмерная функция, но на самом деле реально это одномерная функция, поскольку все изменения происходят только вдоль вот этого направления s.
Теперь, если мы рассмотрим, как меняется эта функция вдоль направления s - это вот такая косинусоида, и разрешением называется расстояние между двумя соседними максимумами этой косинусоиды. Можно вычислить, чему равна эта величина d, и для такой функции эта величина d - это есть просто величина, обратная величине модуля вектора рассеяния **s**. Вот эта величина d, равная единице, деленной на длину вектора **s**, называется разрешением, отвечающим гармонике Фурье. Когда мы рассчитывает синтез Фурье, мы говорим, что синтез Фурье рассчитан с каким-то разрешением, например, 16Å, если в его расчет включены все гармоники Фурье, у которых вот это расстояние d или больше, или равно 16Å. Соответственно, чем меньше разрешение синтеза, тем больше членов включается в расчет ряда Фурье, и тем более точно мы воспроизводим наше распределение электронной плотности.

Примеры синтезов разного разрешения

7

Вот здесь, на этих рисунках для белка Protein G, показано, как выглядят синтезы Фурье, посчитанные с разным разрешением. Ну, как всегда, мы на таких картинках показываем область, в которой находятся достаточно высокие значения синтеза Фурье.

Если мы берем разрешение совсем низкое, то есть включаем мало слагаемых в расчет синтеза, например, 16Å, то синтез Фурье нам показывает только общую форму молекулы и ее месторасположение в элементарной ячейке, но не более того. Если мы повышаем разрешение, берем большее количество членов ряда Фурье, например, разрешение 8Å ангстрем, то при таком разрешении мы, вообще говоря, уже можем видеть альфа-спирали, которые представляются вытянутыми цилиндрическими областями. Дальше, повышая разрешение, где-то при разрешении порядка 3.5Å мы уже можем видеть ход полипептидной цепи и такие вот общие облака для боковых групп полипептидной цепи. Ну и, наконец, если мы возьмем совсем высокое разрешение, 1 Å, то при таком разрешении мы уже видим, что каждый атом изображается здесь отдельным шариком, и мы здесь можем различать и определять местоположение отдельных атомов.

Теоретические ограничения метода

8

Теперь, как я уже сказал, формально для данной гармоники Фурье разрешение - это есть величина, обратная величине длине вектора рассеяния **s**. А сам вектор рассеяния получается как разность векторов, показывающих направление на счетчик и направление первичного пучка, деленная на длину волны λ. Поскольку эти вектора в нашем рассмотрении - σ и σ_0- имеют единичную длину, то понятно, что разница векторов - σ и σ_0 (длина этого вектора) никак не может быть больше двух (ровно 2 мы получаем в крайнем случае, когда они смотрят строго в противоположные стороны). Соответственно, длина вектора **s** не может превосходить $2/\lambda$. Тем самым, самое высокое разрешение, которое мы можем получить теоретически, равно половине длины волны падающего излучения. Как я уже говорил, для рентгеновских волн, используемых для таких экспериментов, длина волны лежит в промежутке где-то между 0.5Å и 2Å. Характерное значение - 1Å. Примерно такая длина волны часто используется на практике в рентгеновских экспериментах. То есть теоретически мы можем получить разрешение синтеза Фурье электронной плотности не лучше, чем пол-ангстрема. И, соответственно, изображение, которое мы видим на синтезе Фурье, не может нам передать детали существенно меньшие, чем вот эти пол-ангстрема.

9

Теперь, когда мы говорим о реальном исследовании в кристалле, мы далеко не всегда можем получить синтезы Фурье такого предельного разрешения. Дело в том, что в зависимости от качества кристалла дифракционное поле кристалла, то есть набор этих рассеянных лучей, интенсивности которых мы можем измерить, может быть существенно ограничен. И, соответственно, чем выше качество кристалла, чем лучше упорядочен кристалл, тем большее количество модулей структурных факторов мы можем измерить. Реально - из того, что имеет место на сегодняшний день - вот здесь выписаны три белка, для которых

достигнуто наивысшее разрешение наборов экспериментальных данных. Это маленький белок крамбин. Для него набор экспериментальных данных был собран до разрешения 0.54Å. Антифриз-протеин - 0.62Å. И довольно большой (по сравнению с ними; примерно 350 остатков) белок альдоз-редуктаза - для него был собран набор разрешением 0.66Å. Это рекордные случаи.
В остальных случаях разрешение меньше. И в современной белковой кристаллографии считается, что кристалл очень хороший и набор собран с очень хорошим, высоким разрешением, если разрешение набора структурных факторов лежит где-то в районе одного ангстрема.
С другой стороны, если вы посмотрите в PDB-банке, там есть много структур, определенных с гораздо худшим разрешением, потому что это зависит от того, какого качества кристаллы удалось получить для проведения рентгеновского эксперимента.
Информация о полноте такого набора - это важная информация. Она обычно отражается в файле, который хранится в PDB-банке. Но для этой информации не предусмотрено специального фиксированного места. И обычно она помещается в разделе этого файла, который содержит мету REMARK в поле "код записи". И здесь в достаточно свободном формате эта информация может располагаться.
Здесь вот пример такой информации. Я хочу обратить внимание, что здесь указываются две границы. Во-первых, здесь указана граница 1.4Å - это наивысшее разрешение, до которого удалось собрать набор экспериментальных данных. Вторая граница показывает, что, на самом деле, не все структурные факторы, отвечающие этому разрешению, реально используются в работе. А часть из них с самым низким разрешением, до 30Å, в работе также не участвует. Это связано, с одной стороны, с техническими причинами. Такие рефлексы, наблюдаемые под малыми углами к первичному пучку, их чисто технически трудно измерить. Ну, и в силу ряда исторических причин такие рефлексы не всегда включаются в работу.

10

Я хочу подчеркнуть, что вот это разрешение набора структурных факторов - оно показывает нам размер деталей, которые мы потенциально можем увидеть, посчитав синтез Фурье соответствующего разрешения. Оно напрямую не связано с тем, насколько точно мы можем определить координаты атомов.
Ну, вот здесь одна из картинок, которые я вам показывал на прошлой лекции, это иллюстрирует. Здесь одномерная картинка. Синтез Фурье разрешения 5Å, посчитанный для такой модельной структуры, где 2 атома находятся на расстоянии 1.5Å и третий атом находится на расстоянии 3Å от этого атома. Во-первых, мы видим, что на таком синтезе разрешения 5Å мы еще можем увидеть, что здесь у нас есть отдельные группы атомов, поскольку эта кривая содержит понижение, и мы видим эти два пика. С другой стороны, эти атомы, которые находятся на расстоянии 1.5Å, мы видим как отдельный пик, и мы не видим, что здесь на самом деле 2 атома. Тем не менее, если даже мы для каждого из этих атомов определим координаты как положение в центре этого пика, то ошибка в координатах будет существенно меньше, чем эти номинальные 5Å. Поэтому я хочу еще раз подчеркнуть, что разрешение синтеза - это характеристика визуальная. Она определяет, что именно мы визуально можем на этом синтезе увидеть. А координаты атомов мы можем определить гораздо более точно.
Итак, суммируя это повторение сказанного на прошлых лекциях -

рентгеновский дифракционный эксперимент с монокристаллом позволяет измерить модули коэффициентов в разложении функции распределения электронной плотности в ряд Фурье;
наличие модели структуры дают возможность рассчитывать гипотетические значения модулей и сравнивать модули с экспериментальными значениями;
а если мы знаем и модули, и фазы структурных факторов, то мы можем просто восстановить распределение электронной плотности, посчитав синтез Фурье с коэффициентами, используя эти известные значения модулей и фаз структурных факторов.

Метод Патерсона

11

Но для того, чтобы сделать такую вещь, нам надо каким-то образом восстановить значения фаз, которые теряются в эксперименте, и об этом я поговорю чуть позже. А пока я немножко отвлекусь в сторону и обсужу такой вопрос: а что еще мы могли бы, какие синтезы мы могли бы посчитать с экспериментальными данными и что мы могли бы на них увидеть. И я обсужу такую на первый взгляд достаточно странную вещь: что будет, если мы посчитаем некоторый синтез Фурье, но в качестве коэффициентов возьмем не модули и фазы, как для настоящего распределения электронной плотности, а возьмем в качестве модулей интенсивности, квадраты модулей структурных факторов, а фазы возьмем просто равные нулю. Поскольку интенсивности - это то, что мы меряем в эксперименте, то такой вот синтез Фурье просто эквивалентным образом передает ту информацию, которую мы получаем в рентгеновском эксперименте.

12-15

Такой синтез Фурье называется синтезом Патерсона. Соответствующая функция называется функцией Патерсона. И эта функция может быть непосредственно рассчитана по экспериментальным данным.
Функция Патерсона обладает очень важным для нас свойством: если у нас распределение электронной плотности - это не какая-нибудь произвольная функция, а атомная функция, то есть функция, состоящая из какого-то числа пиков, отвечающих атомам с координатами $\mathbf{r}_1$,..., $\mathbf{r}_n$, то соответствующая этому распределению функция Патерсона тоже будет состоять из некоторого числа пиков. Но теперь эти пики будут более сильно (примерно в 2 раза) размазаны. И расположены эти пики будут в точках, отвечающим разностям координат пар атомов $\mathbf{r}_j$ - $\mathbf{r}_k$.
Допустим, что функция электронной плотности имеет один атом в начале координат, а остальные в точках $\mathbf{r}_2$, $\mathbf{r}_3$, $\mathbf{r}_n$. Здесь составлена табличка, в каких точках будут располагаться пики на функции Патерсона. Ну, во-первых, мы видим, что здесь вот на диагонали стоит ряд нулей, конкретно - n штук нулей. Они отвечают тому, что если мы берем какой-то атом и в качестве второго атома берем этот же самый атом, то мы получаем нулевой вектор. Поэтому для этой функции Патерсона n штук пиков попадет в точку 0, и мы будем иметь в начале координат такой мощный усиленный пик, который слагается из n одинаковых пиков.

16

Далее. Если мы посмотрим, как устроены координаты остальных пиков, то надо учитывать, что у нас один из атомов находится в начале координат. Поэтому,

если мы возьмем поочередно координаты каждого атома и вычтем нулевые координаты, то мы получим ровно эти самые координаты, что здесь и написано. Теперь, если мы возьмем вектор $\mathbf{r}_2$ и начнем вычитать его из всех этих координат, то мы получим точки $-\mathbf{r}_2$, 0, $\mathbf{r}_3-\mathbf{r}_2$ и так далее. Ну, и, таким образом, мы получаем эту табличку. Теперь, если мы посмотрим на эту табличку, мы видим, что на самом деле среди вот этих n^2 пиков (на самом деле не n^2, а n^2-1, учитывая, что n пиков у нас совпадают в нуле) есть n пиков, просто отвечающих положениям атомов а нашей исходной структуре. То есть на самом деле в этой картине из n^2 пиков функции Патерсона спрятано изображение нашей исходной структуры. Но это изображение структуры среди пиков Патерсона встречается не один раз, а:
встречается изображение структуры;
встречается изображение структуры, сдвинутой на вектор $\mathbf{r}_2$ (то есть мы видим нашу структуру, сдвинутую как жесткое тело на вектор $\mathbf{r}_2$);
изображение структуры, сдвинутой на вектор $\mathbf{r}_3$,
и так далее. То есть на самом деле среди системы пиков функции Патерсона присутствует n раз изображение искомой структуры, но сдвинутое на разные вектора.
Теперь - можем ли мы из этой информации о пиках функции Патерсона извлечь информацию о том, как устроена наша структура? В благоприятных случаях - да.

17-18

Рассмотрим опять этот простой пример. Вот у нас есть функция Патерсона. И нас интересует, а как же устроена наша структура.
Мы знаем, что в нашей структуре есть 3 атома, и каждый из атомов лежит в каком-то из пиков функции Патерсона. Ну, допустим, давайте предположим, что вот эти 3 атома образуют искомую структуру. Тогда мы можем посмотреть, а какая в таком случае должна была бы быть функция Патерсона. Здесь показано положение пиков функции Патерсона, соответствующей такому расположению атомов. Мы видим, что это не совпадает с той картинкой, которую мы исходно имеем. В этой точке функции Патерсона пика нет, а он должен быть. А здесь наоборот - есть пик, а он отсутствует в нашей модели. Значит, это была плохая гипотеза.

19-20

Давайте возьмем другую гипотезу. Давайте предположим, что на самом деле атомы нашей структуры расположены вот в этих точках. Тогда, опять-таки, вычисляя соответствующую функцию Патерсона и табличку координат атомов, мы увидим, что у нас есть такая картина, и мы видим, что на этот раз это хорошо совпадает с тем, что мы видим в функции Паттерсона. Значит, это может рассматриваться как решение, как та структура, которую мы имеем.
То есть теоретически так, перебирая все комбинации из n пиков, выбранных из функции Патерсона, мы можем их все проверить и найти тот, который является настоящим решением.

21

Но на самом деле такой подход существует, и он иногда приводит к успеху, но приводит к успеху только для структур, состоящих из очень небольшого числа атомов. Причина здесь вот в чем. Для того, чтобы проделывать такую процедуру, нам надо идентифицировать на функции Патерсона n^2 пиков. Если, допустим, у нас структура состоит из десяти тысяч атомов, и эти атомы

достаточно тесно напиханы в элементарную ячейку кристалла, n^2 уже будет сто миллионов. При этом пики, отвечающие атомам, более широкие, чем настоящие. Понятно, что при этом все эти пики будут перекрываться между собой. И мы, если посчитаем для белка такую функцию Патерсона, мы увидим совершенно смазанную, слипшуюся картину. Мы просто не в состоянии определить координаты этих пиков. Но, если число атомов невелико, то такой подход, в принципе, может работать. И такие Патерсоновские методы используются для решения структур, состоящих из небольшого числа атомов. Для белковой кристаллографии это имеет большое значение, поскольку для того, чтобы применять ряд подходов к решению фазовой проблемы, надо предварительно определить координаты некоторых специальных атомов, содержащихся в молекуле белка. Ну, например координаты так называемых "тяжелых атомов" - тяжелых меток, присоединенных к молекуле белка. Этих атомов уже в элементарной ячейке мало, и поэтому для определения их координат, для определения подструктуры могут применяться такие Патерсоновские методы.

Методы решения фазовой проблемы

22

Теперь я перехожу к обсуждению подходов, которые применяются для решения фазовой проблемы. Таких подходов существует несколько. Это связано с тем, что на самом деле ни один из них не дает гарантированного определения структуры. И поэтому в реальной жизни, если не проходят одни подходы, один путь решения, пытаются попробовать другой до тех пор, пока не найдут какой-то метод, который даст это решение.

Метод изоморфного замещения

23

Сейчас я очень вкратце остановлюсь на этих методах. Первый метод, о котором я хочу поговорить - это метод изоморфного замещения. Это на самом деле первый метод, которым научились определять пространственные структуры белковых молекул. Идея метода заключается вот в чем. Если у нас есть нативный белок, то для него есть модули и фазы структурных факторов, при этом в эксперименте мы часть информации теряем. Мы, грубо говоря, меряем только половину информации - модули структурных факторов. Теперь предположим, что нам удалось получить кристаллы модифицированного белка. Такие кристаллы, в которых сама молекула белка не изменилась, находится на том же самом месте. Но в каком-то определенном месте к каждой молекуле белка присоединилась тяжелая метка. В таком случае говорят, что получено изоморфное производное (полное название - изоморфное производное соединение, поэтому такие здесь окончания). И с этим изоморфным производным тоже можно провести эксперимент и получить модули структурных факторов уже для такого модифицированного объекта, модули структурных факторов, которые отвечают системе "белок плюс тяжелый атом". Теперь, если мы возьмем разности этих измеренных модулей структурного фактора для тяжелоатомного производного и для нативного белка, то эти разности примерно равны значениям модулей структурных факторов, отвечающих распределению электронной плотности в тяжелых атомах. И, используя эти модули, поскольку тяжелых атомов у нас мало, мы можем

пытаться определить координаты этих тяжелых атомов, используя те самые Патерсоновские методы, о которых я только что рассказывал. То есть в данном случае, используя эти подходы, мы можем определить не всю структуру, но мы можем определить, имея пару "нативный белок - изоморфное производное", координаты тяжелых атомов, тяжелых меток, присоединенных к молекулам белка. А если мы можем определить координаты этих меток, то мы можем определить и модули, и фазы структурных факторов, отвечающих только подсистеме из тяжелых атомов.

24

Теперь, чтобы двинуться дальше, я должен сделать такое замечание. Вообще говоря, для всяких вычислений очень удобно рассматривать модуль и фазу структурного фактора просто как модуль и аргумент комплексного числа и работать с комплексными коэффициентами ряда Фурье. Я в этом курсе не буду использовать комплексных чисел, поэтому про это я говорить не буду. Но, тем не менее, я хочу сделать вот какое замечание.

25

Вот эту пару - модуль и фаза структурного фактора - удобно изображать на плоскости в виде вектора, у которого длина вектора совпадает со значением модуля структурного фактора, а фаза определяет угол этого вектора с осью x. Теперь, если мы имеем структурные факторы, изображенные таким вектором, отвечающие нативному белку, и имеем структурный фактор, отвечающий тяжелым атомам, то суммарная электронная плотность (электронная плотность тяжелоатомного производного) будет векторной суммой структурных факторов, отвечающих нативному белку и тяжелому атому. То есть в случае наличия тяжелоатомного производного мы имеем такую вот векторную диаграмму. Эта диаграмма своя для каждого структурного фактора. Если к структурному фактору нативного белка прибавить структурный фактор тяжелого атома, то мы получим структурный фактор тяжелоатомного производного.
Допустим, эксперимент нам дает значения модулей структурных факторов для нативного белка и тяжелоатомного производного. А для структурного фактора, отвечающего тяжелому атому, мы можем рассчитать и модуль, и фазу, ну, вот определив координаты тяжелых атомов, как я говорил раньше.
Теперь смотрите, что у нас происходит дальше.

26

Если мы знаем модуль структурного фактора вектора, отвечающего нативному белку, единственное, что мы можем сказать - это что конец этого вектора лежит где-то на этой окружности. А где именно - мы не знаем, поскольку мы не знаем фазу. Теперь дальше. Мы знаем, что если к тяжелому атому прибавим вектор, отвечающий нативному белку, то мы получим вектор, отвечающий тяжелоатомному производному, про который мы также знаем модуль этого вектора. Поэтому, если точка, отвечающая концу вектора структурного фактора FP, должна лежать на окружности с центром в конце вектора FH и радиуса FPH. Поэтому, с одной стороны, наше решение должно лежать на этой окружности, с другой стороны - на этой, поэтому единственный вариант, который у нас остается - это одно из этих двух решений.

27

Здесь вот эти два решения нарисованы. Ну, соответственно, мы можем показать, какое значение фазы структурного фактора возникает для каждого из этих решений. То есть наличие тяжелоатомного производного позволяет для каждого

структурного фактора (по крайней мере, в теории) свести неопределенность в фазе до одного из двух значений. То есть сначала мы не знали о фазу вообще ничего, а теперь мы знаем, что она принимает одно из этих двух значений. Ну, вот эта неопределенность остается. Но она уже существенно меньше. И ее можно совсем снять, если мы получили второе тяжелоатомное производное.

28-29

В этом случае, повторяя эту схему для второго производного, мы снова получим пару допустимых значений фазы, из которых одно настоящее значение будет совпадать с одним из этих значений, а второе будет отличаться. Таким образом, мы уже сможем однозначно определить фазу структурного фактора.

Суммируя это, можно сказать, что метод изоморфного замещения позволяет решать фазовую проблему. Эти изоморфные производные удается практически получать, потому что, как я уже говорил, в кристаллах белка примерно половина объема ячейки занята растворителем, водой. В результате в кристаллах есть достаточно большие полости и каналы, по которым ионы тяжелых металлов могут подойти к молекуле белка и с ней связаться. Поэтому наличие каналов позволяет получать производные. Ну, например, распространенный метод получать производные - это метод вымачивания, когда кристаллы нативного белка помещаются просто в раствор, содержащий ионы солей тяжелых металлов. И через некоторое время эти ионы проникают в кристалл и присоединяются к молекулам белка. Обычно используются такие металлы как золото, платина, уран, вольфрам.

Недостатки метода изоморфного замещения

30

С какими сложностями мы сталкиваемся, применяя этот метод? Ну, во-первых, изоморфизм имеет место только с некоторой степенью точности. Дело в том, что, конечно, когда в кристалл начинают внедряться эти тяжелые атомы, присоединяются к молекуле белка, происходят какие-то, по крайней мере, небольшие подвижки атомов молекулы белка. И поэтому сама идея изоморфизма - это такая вот теоретическая идея. А на самом деле, конечно, у нас всегда при внедрении тяжелых атомов нативный белок тоже несколько как-то меняет свою конформацию подвигается в элементарной ячейке. Отсутствие изоморфизма приводит к тому, что фазы, рассчитанные методом изоморфного замещения, имеют достаточно большие ошибки. И, соответственно, синтезы Фурье электронной плотности становятся трудно интерпретируемыми. Вот эти ошибки в фазах приводят к плохому качеству карт, соответствующих синтезам Фурье.

Вторая проблема заключается в том, что, чтобы применять этот метод, мы должны предварительно определить места присоединения тяжелых атомов. Ну, я уже сказал, что, например, атом можно сделать, используя Патерсоновские методы. Но на самом деле в жизни все не так просто. И определение координат тяжелых атомов - оно тоже представляет собой некоторую проблему, которая иногда решается достаточно просто, а иногда требует больших усилий для ее решения.

Следующая проблема возникает при попытке использовать этот метод для больших макромолекулярных комплексов. Дело в том, что, чтобы иметь регистрируемую разницу между модулем структурного фактора нативного белка и модулем для тяжелоатомного производного, для того, чтобы эта

разница была достаточно большая, чтобы мы могли ее зарегистрировать с учетом точности эксперимента, эта метка должна быть достаточно тяжелой. Если для белка небольшого размера атом золота или атом вольфрама является достаточно тяжелым атомом, чтобы какую-то эту разность создать, то когда мы используем такой большой комплекс, как рибосома, понятно, что единичный атом слишком мелок, чтобы заметить эффект от его присоединения. Поэтому в этом случае пытаются найти какие-то специальные подходы. Например, присоединять не отдельные атомы, а кластеры из тяжелых атомов. Так, при работе с рибосомой использовались кластеры, состоящие из нескольких десятков атомов золота. Но это - отдельная проблема.
Ну и последнее. Как я уже говорил, при наличии одного производного фазы определяются неоднозначно. Чтобы снять эту неоднозначность, необходимо использовать несколько производных, что, соответственно, умножает все сложности, о которых я говорил только что.
Но, тем не менее, этот метод применялся и применяется. Это один из рабочих методов.

Метод аномального рассеяния

31-32

Следующий метод, о котором я не буду говорить сколь-нибудь подробно - это так называемый метод аномального рассеяния. Он связан с тем, что та картина рассеяния, о которой я говорил - она имеет место для большинства атомов. Но в некоторых случаях, когда речь идет о рассеянии тяжелыми атомами и длинах волн, по частотам падающей волны близким к собственным частотам атомов, возникают некоторые дополнительные эффекты, сравнительно небольшие, но которые, тем не менее, возникают, которые можно зарегистрировать в эксперименте. В эксперименте этот эффект проявляется в нарушении так называемого закона Фриделя, который заключается вот в чем:
из формул, которые я вам выписывал, следует, в частности, например, такая вещь:
если мы рассмотрим модуль структурного фактора для рефлекса с индексами h, k, l и рассмотрим модуль структурного фактора для рефлекса с индексами -h, -k, -l, то мы получим одно и то же значение. Понятно, почему. Изменение всех знаков под знаком косинуса косинус не меняет, под знаком синуса приводит к появлению минуса. Но, когда мы возводим в квадрат, этот минус пропадает, и мы получаем то же самое значение. Поэтому в реальном эксперименте, чтобы собрать полный набор данных, нужно, например, померить рефлексы только с индексами, у которых только l неотрицательно, потому что вторая половина рефлексов восстанавливается по этой формуле.
А в случае нарушений закона Фриделя, в случае наличия аномально рассеивающих атомов эти величины становятся различными, и недостающую новую экспериментальную информацию дает независимое измерение рефлексов, связанных вот таким соотношением, как **s** и **-s** с их партнерами. Теперь, сравнивая разницу между модулями, отвечающими вектору рассеяния **s** и **-s**, можно, используя методы типа Патерсоновских, определить положение этих рассеивающих атомов. И дальше, работая на этих разницах, строить диаграммы примерно такие, как я строил для случая изоморфного замещения, и можно тоже определить фазу (ну, опять-таки определить с точностью до одного из двух значений). Впрочем, эта неоднозначность снимается, если удалось получить кристаллы с другими аномальными рассеивателями. Или если

скомбинировать данные по аномальному рассеиванию с данными изоморфного замещения.
Плюсы - это то, что этот метод позволяет решать фазовую проблему. Он был использован в ряде случаев. В белках иногда в составе молекулы белка встречаются аномально рассеивающие атомы. Ну, например, атом железа в гемах. Бывают медесодержащие белки. Часто в кристаллах белков находят связанные ионы магния или марганца. Ну, а кроме того, можно пытаться получить тяжелоатомные производные с аномально рассеивающими атомами.
Проблемы - те же самые, что с методом изоморфного замещения. Во-первых, надо найти места аномально рассеивающих атомов. Есть неоднозначность в определении фаз. Ну, и еще одна сложность заключается в том, что эффект аномального рассеяния очень слабый. И он требует гораздо более высокой точности проведения эксперимента, чем в случае изоморфного замещения.

Метод многоволнового аномального рассеяния

33

Следующий метод, о котором я хочу упомянуть - это очень важный метод. Сейчас это один из наиболее популярных методов. Это метод многоволнового аномального рассеяния. И в основе этого подхода лежит изменение интенсивности аномального рассеяния при изменении длины волны.
Для того, чтобы его применять, опять-таки надо иметь аномально рассеивающие атомы. И серия экспериментов заключается в том, что меряется интенсивность рассеянных волн. Но эксперимент повторяется несколько раз с изменением длины волны падающего рентгеновского излучения.
Этот метод сейчас является одним из наиболее перспективных. И расцвет этого метода связан с двумя обстоятельствами.
Первое обстоятельство - это то, что рентгеновский эксперимент стали проводить с использованием синхротронных ускорителей. Дело в том, что лабораторные рентгеновские трубки дают интенсивное рентгеновское излучение только с одной фиксированной длиной волны. Поэтому на лабораторно рентгеновской установке невозможно провести эксперимент при нескольких разных длинах волн. Синхротронный ускоритель позволяет получать интенсивные пучки с разными длинами волн, и поэтому позволяет варьировать эту длину волны и проводить эксперимент по многоволновому аномальному рассеянию. Вторая причина, сделавшая этот метод широко применимым, заключается в том, что научились делать следующую вещь. В молекулах белка, как правило, есть аминокислотные остатки метионина, содержащие в себе серу. Сама по себе сера практически не дает эффекта аномального рассеяния, и ее использование в этом методе крайне проблематично. Но была наработана методика, которая позволяет атомы серы в этих остатках заменить на атомы селена. А атом селена дает уже достаточно сильный эффект аномального рассеяния и может применяться для решения фазовой проблемы. Поэтому методика здесь заключается в следующем: в исследуемом белке все атомы серы в метионинах меняются на селен, и получается так называемое селен-метиониновое производное, и вот для этого селен-метионинового производного проводится эксперимент при нескольких длинах волн. Проблема заключается в чем? Опять-таки в нахождении мест аномально рассеивающих атомов. И здесь эта проблема осложняется тем, что остатков метионина в молекуле может быть достаточно много, и, соответственно, этих атомов селена, аномально рассеивающих, тоже может

быть очень много. А как мы говорили, Патерсоновские методы хороши, когда атомов мало. А когда атомов много - пики начинают на синтезах Патерсона перекрываться, и с ними гораздо сложнее работать. Поэтому здесь по-прежнему есть проблема нахождения мест аномально рассеивающих атомов, и она сложнее, чем в методе изоморфного замещения, и поэтому этот метод - тоже не всегда удается его применить, хотя, повторяю, это сейчас широко распространенный метод.

34

Ну, здесь вот та карта, что я вам уже показывал. Это синтез Фурье электронной плотности для белка альдоз-редуктаза. Разрешение синтеза - 0.9Å. Это как раз пример такого синтеза, где фазы были определены методом многоволнового аномального рассеяния с использованием селен-метиониновых производных. Но я еще раз повторяю, что это в некотором плане рекордный синтез по качеству. Так хорошо получается далеко не всегда. Даже наоборот, как правило, получается хуже.

Метод молекулярного замещения

35

Теперь я перехожу еще к одному очень важному методу, широко применяемому на практике - метод молекулярного замещения. Идея применения метода заключается в том, что если мы исследуем какой-то белок, и мы знаем, что у этого белка есть гомологичный белок, который уже был исследован и координаты атомов которого уже известны, то наша задача становится проще, потому что мы уже примерно знаем, как должен быть устроен наш белок. Он должен быть устроен как гомолог. И задача заключается только в том, чтобы модель гомологичного белка несколько поправить так, чтобы она стала такой, как она есть в исследуемом объекте.

Отступление. Комбинированные синтезы.

36-37

Такую правку можно осуществить, например, при помощи так называемых комбинированных синтезов Фурье. Сейчас я хочу немного об этом поговорить. Итак, давайте предположим, что у нас есть атомная модель, мы знаем координаты атомов некоторого белка. По этим координатам мы рассчитали модули и фазы и построили синтез Фурье. Ну, если этот синтез Фурье достаточно высокого разрешения, то полученная плотность достаточно хорошо воспроизводит модель, и мы видим, что модель соответствует этому синтезу.

38

Теперь - что будет, если мы из нашей модели исключим часть атомов (здесь, например, исключен остаток триптофан) и посчитаем модули и фазы по такой частичной модели, и потом посчитаем синтез Фурье. Ну, понятно, что этот синтез Фурье воспроизведет ту часть модели, которая была включена в расчет, и ничего не покажет про этот остаток триптофан, потому что ни модули, ни фазы "про него ничего не знают", координаты атомов этого остатка никак не использовались в их расчете.

39-40

Теперь возникает вопрос - а что будет, если мы возьмем модули, которые были рассчитаны по полной модели, а фазы возьмем те, которые были рассчитаны по

частичной модели? Ну, естественно, фазы "ничего не знают" о вот этих недостающих атомах. Но в модулях, в принципе, какая-то информация об этих атомах присутствует, поскольку при расчете этих модулей использовались все координаты атомов, в том числе, и вот этого выброшенного триптофана.
Ну, если мы прямо так вот возьмем и посчитаем синтез Фурье и посмотрим его в тех же условиях, мы на самом деле серьезных изменений не увидим. Ну, вот здесь у нас появилось небольшое пятнышко, но оно сравнительно небольшое, и, в целом, для остатка триптофана по-прежнему видно плохо.

41

Но ситуация изменится, если мы опустим уровень срезки, которая используется, чтобы нарисовать эту картину. То есть мы покажем области, которые отвечают более низким значениям электронной плотности, нежели на предыдущем рисунке. Ну, если мы опять построим синтез, где и модули, и фазы построены по частичной модели, мы здесь, естественно, опять ничего не увидим, поскольку этого триптофана здесь и нет. Но если мы понизим этот уровень критический на комбинированном синтезе, мы увидим, что уже на таком синтезе у нас появились очертания облака, которое достаточно хорошо воспроизводит этот остаток триптофана и позволяет его сюда вписать. То есть если мы имеем модули, отвечающие полной модели, а в эксперименте мы получаем модули, отвечающие всем атомам настоящего объекта, и имеем фазы, которые отвечают только части модели, то, рассчитывая такой комбинированный синтез, мы можем получить изображение и не включенных в модель частей структуры тоже. И, таким образом, провести корректировку модели, добавив туда то, чего там не хватало.

42-43

Ну, вот, более формально здесь то же самое и написано:
Предположим, что у нас в модели есть N атомов, а в частичной модели есть только первые n из них. Если мы строим синтез Фурье, где и модули, и фазы рассчитаны по полной модели, то мы, естественно увидим пики в местах, отвечающим всем атомам. Если мы модули и фазы берем по частичной модели, то мы, естественно, увидим те атомы, которые входят в состав частичной модели, и не увидим ничего в местах положений тех атомов, которые в расчете не участвуют. Теперь - что будет, если мы возьмем модули, посчитанные по всем атомам, а фазы, посчитанные по части атомов? Существует некоторая математическая теория, которая говорит, что в этом случае картина будет примерно следующая: те атомы, которые использовались и для расчета фаз, и для расчета модулей - они будут видны, так сказать, "в полный рост". А вот те атомы, которые участвовали в определении модулей структурных факторов, но не участвовали в определении фаз - они тоже будут видны, но с половинной высотой. Что, собственно говоря, на предыдущих картинках я вам и показал. Если мы брали изначально картинку, мы не видели, а когда мы понизили критический уровень электронной плотности, выше которого мы рисуем нашу область, - мы увидели появление вот этих атомов, но, так сказать, "более низкого роста".

44

Теперь, когда мы используем в качестве стартовой точки модель гомологичного белка, ситуация может быть более сложной. Поскольку в гомологичном белке некоторые атомы могут отсутствовать, которые нам реально нужны. Но, с другой стороны, там могут присутствовать и атомы, которые нам не нужны,

которых вообще нет в нашем исследуемом объекте, или они расположены не так, как в исследуемом объекте. Поэтому давайте теперь рассмотрим вот такую ситуацию.
Вот у нас есть полная модель, в которой есть N атомов. А частичная модель устроена так: в нее включены, во-первых, не все атомы из полной модели, а во-вторых - включено еще какое-то количество "ложных" атомов, которых в исходной модели и не было. Что у нас будет в таком случае?
Если мы построим синтез Фурье по модулям и фазам, рассчитанным по настоящей модели, мы увидим пики в местах, отвечающих атомам этой модели, и, соответственно, не будет никаких пиков в местах, отвечающих "ложным атомам". Если мы возьмем модули и фазы, посчитанные по нашей пробной модели, то мы увидим пики в тех местах, которые отвечают тем атомам настоящей модели, которые сюда включены; не увидим пиков в местах отсутствующих атомов и увидим пики в местах соответствующих "ложных атомов". А если мы возьмем комбинированный синтез, а именно возьмем модули от настоящей модели, а фазы от частично искореженной модели, то мы увидим атомы "в полный рост" там, где они есть одновременно и в пробной модели, и в настоящей структуре. Кроме того, мы увидим атомы "в половинный рост" для тех положений, которые не были включены в модель, и для тех, которые были включены в модель неправильно, то есть там, где их на самом деле в исходной структуре не существует.

45-46

Дальше для удобства был придуман такой трюк. Давайте возьмем коэффициенты этого синтеза с коэффициентом двойка, а этого возьмем с коэффициентом минус единица. Тогда, устроив вот такую комбинацию этих синтезов, мы получим, что высота пиков, отвечающих тем атомам, которые есть в настоящей структуре и включены в модель, будет 2*1-1=1; высоты пиков атомов, которые присутствовали в настоящей структуре, но отсутствовали в модели, будут 2*1/2-0=1; а высоты пиков, которые были ложно включены в модель, будут 2*1/2-1=0. То есть, построив синтез Фурье вот с таким коэффициентами, мы получим синтез, на котором будут видны в полную силу все атомы настоящей структуры - и те, которые присутствовали в модели, и те, которые в модели не присутствовали.
Такие комбинированные синтезы часто применяются в кристаллографии. Читая статьи про определение структур белков, часто можно встретить ссылки на такие синтезы. Они обычно так и обозначаются "$(2F_{obs} - F_{calc})$, φ_{calc} "; здесь F_{obs} - то, что мы меняем в эксперименте, что отвечает настоящей структуре, а F_{calc} и φ_{calc} - это то, что рассчитано по частичной модели, в которой могут присутствовать и какие-то и совсем посторонние атомы, которых на самом деле в исследуемой структуре нет.

Метод молекулярного замещения. Продолжение

47

Теперь я снова возвращаюсь к методу молекулярного замещения. План действий при исследовании этим методом такой:
Мы хотим исследовать некоторый новый белок. Прежде всего, мы пытаемся найти в PDB-банке белок, у которого координаты уже определены, про который мы думаем, что его структура похожа на структуру исследуемого белка. Ну, например, этот белок имеет высокую степень гомологии первичной

последовательности с первичной последовательностью исследуемого нами сейчас белка. Дальше эту пробную модель мы пытаемся разместить в элементарной ячейке так, чтобы она лежала в том же месте и в той же ориентации, как и молекула нашего исследуемого белка. Как это сделать - я скажу чуть позже. Ну, и после того, как мы эту молекулу гомологичного белка разместили, мы можем рассчитать по ней значения модулей и фаз структурных факторов, и дальше использовать комбинированные синтезы Фурье для того, чтобы откорректировать эту пробную модель. То есть мы строим разностные комбинированные синтезы Фурье и, исходя из них, убираем те атомы, которые надо убрать, добавляем те, которые надо добавить. То есть проводим ручную правку этой модели.
Но, как я сказал, для того, чтобы это сделать, мы должны найденную в банке пробную модель, разместить в элементарной ячейке в том же месте и в той же ориентации, как и лежит модель исследуемого белка, про которую мы пока, естественно, ничего не знаем - ни где она лежит, ни как она повернута. Как можно пытаться решить такую задачу? Идея заключается в том, что мы можем перебрать все возможные случаи размещения пробной модели в элементарной ячейке и ее ориентации; для каждого такого варианта посчитать соответствующие модули структурных факторов и сравнить их с экспериментом. Если различие очень большое - значит, вариант такого размещения плохой. А если совпадение хорошее - значит, такой вариант размещения этой модели приемлем, и с ним можно начинать работать.

48

Формально положение вот этого твердого тела определяется шестью параметрами - это углы вращения и вектор трансляции. И, на самом деле, то, что реально делается на практике - наилучшее значение этих углов находится, по существу, тупым перебором. То есть сканируются все возможные значения углов с шагом, скажем, в один градус; перебираются всевозможные положения центра тяжести модели, скажем, с шагом там один ангстрем или пол-ангстрема. И, посчитав очень большое количество таких вариантов, выбирается тот, для которого рассчитанные модули наиболее хорошо соответствуют эксперименту. Ну, здесь вот это написано более формально. Для каждого допустимого набора параметров можно рассчитать соответствующие значения модулей Fcalc.

49

Дальше можно посчитать некоторый критерий - сумму квадратов отклонений - насколько хорошо эти рассчитанные значения модулей соответствуют эксперименту. И теперь мы пытаемся, перебирая все возможные углы альфа, бета, гамма и компоненты вектора трансляции, найти такой вариант, который обеспечивает наилучшее совпадение. Но, на самом деле, существуют некоторые математические трюки, которые позволяют этот шестимерный поиск разделить на два независимых трехмерных поиска: сначала найти углы, а потом найти трансляцию. Но это уже такая, более техническая деталь. Я о ней не буду говорить.
По существу, идея заключается в том, чтобы просто перебрать все возможные положения и найти то, при котором наилучшее совпадение.
Ну, и еще раз я подчеркиваю, что после того, как такое положение ориентации этой пробной модели найдено, работа не закончена, а во много еще только начинается, потому что вот эту модель надо откорректировать.

50

Суммируя про метод молекулярного замещения, я хочу еще раз сказать, что это один из наиболее используемых методов решения фазовой проблемы. Изучая PDB-банк, вы найдете там очень много структур, решенных методом молекулярного замещения. Возможности этого метода с течением времени расширяются, поскольку, чем больше становится известных структур белков, тем больше шансов, что ваш новый исследуемый белок будет гомологичен одному из известных. Проблема заключается в том, что для применения этого метода нужно наличие достаточно гомологичной модели. И результат зависит от степени обоснованности этой гипотезы о гомологии.

51

Ну, еще серьезная проблема, связанная с этим методом - он имеет такую тенденцию, что вы оказываетесь захваченными структурой гомолога и получаете ответ очень близкий к структуре гомолога, хотя на самом деле отличие структуры гораздо больше. Но просто вы работаете с этими разностными картами, занимаясь корректировкой модели, не сумели увидеть, что эти отличия есть. Поэтому, в общем, этот метод содержит определенную опасность. Вы можете получить структуру неожиданно близкую к гомологу, хотя на самом деле там разница может быть существенно больше.

52

Теперь - откуда взять эту гомологичную модель? Ну, первый ответ - это среди структур, определенных методом рентгеноструктурного анализа. Как правило, так и делают. В принципе, можно искать структуру среди структур, определенных методами ядерно-магнитного резонанса. Этот вопрос уже не столь простой. Он активно дискутируется в печати. Существуют примеры, когда такие ЯМР-овские модели успешно использовались для решения фазовой проблемы. Но все-таки это достаточно редкая ситуация. С использованием ЯМР-овских моделей здесь все не так просто. То есть иногда можно, но сложности возникают.

Другие методы решения фазовой проблемы

53

Следующий случай - это трехмерная реконструкция модели по электронно-микроскопическим изображениям. Такие вещи применялись, например, при исследовании структуры рибосомы. Но здесь тоже есть свои сложности, прежде всего, связанные с тем, что трехмерная реконструкция электронно-микроскопических изображений дает некоторую информацию только для очень крупных комплексов, и информацию достаточно низкого разрешения, поэтому здесь тоже большие сложности.

Ну, и, наконец, как некоторая голубая мечта, теоретически существует такая возможность: попытаться предсказать пространственную структуру теоретическими методами. Вот это теоретически предсказанную структуру попытаться использовать для методов молекулярного замещения. Ну, и, соответственно, ее уточнить, получить правильное решение.

В таком направлении работают. Но это пока только как некоторая идея на будущее. Ни одна реальная структура таким методом еще решена не была.

Ннаконец, последние в моем списке так называемые "прямые методы". Это методы в которых не делается попытка получить какие-нибудь дополнительные экспериментальные данные, а в которых пытаются для решения фазовой проблемы привлечь какие-то дополнительные сведения об исследуемом объекте

в виде каких-то его общих свойств. Здесь приведены примеры таких общих свойств. Например, мы знаем, что функция распределения электронной плотности должна быть не какой угодно, а состоять из конечного числа пиков, отвечающих атомам нашей структуры; что эта функция должна быть неотрицательна (ну, или почти неотрицательна, поскольку функция распределения электронной плотности - неотрицательная функция). Ну, и ряд таких перечисленных здесь свойств.
Такие прямые методы рутинно используются для решения структуры низкомолекулярных соединений. Структуры низкомолекулярных соединений по данным рентгеновского рассеяния определяются без использования всех вот этих трюков, о которых я вам говорил, а на основе именно общих свойств. Для небольших белков при высоком разрешении был рад успешных попыток применения таких методов. Ну и, наконец, при работе с большими комплексами при среднем и низком разрешении такие методы разрабатываются. Есть отдельные успешные попытки. Но это совершенно единичные случаи. То есть пока что успешность применения таких прямых методов для белковой кристаллографии не слишком велика.

Структура информации в PDB-файле

54

Теперь я собираюсь вернуться к тому, какая информация приведена в PDB-файле, и поговорить о некоторых типах информации, о которых мы не говорили. Пока что мы говорили о том, что мы хотим определить координаты атомов. Координаты атомов могут определяться двумя способами: могут определяться относительные координаты в системе координат, связанной с осями элементарной ячейки. Но это, вообще говоря, неудобные координаты. И более удобно эти координаты атомов приводить в так называемой абсолютной системе координат. Это ортогональная система координат, в которой координаты атомов меряются в ангстремах.
В файле PDB, как мы уже говорили, содержится карта CRYST1, которая дает описание параметров элементарной ячейки - длины ребер элементарной ячейки и соответствующие углы. И содержится матрица, которая позволяет перейти от абсолютных координат, когда нужно, к относительным координатам атомов.

55

Теперь, говоря раньше о том, как происходит рассеяние на электронах, входящих в атом, я в неявном виде предполагал, что атом покоится. Но на самом деле в веществе атомы не находятся в состоянии покоя, а они совершают определенные тепловые колебания, то есть они находятся в таком осциллирующем состоянии. Это приводит к тому, что вот это распределение электронной плотности в атоме - оно, так сказать, не находится все время на одном месте, а сдвигается в разные точки пространства, что эффективно приводит к тому, что, на самом деле, это распределение несколько "размазывается" по пространству. То есть мы имеем, если у нас был сферический атом, и если он совершает колебания с равной амплитудой во всех направлениях, то у него будет такого сферического характера облако, но оно будет иметь гораздо большие размеры в пространстве, оно будет размазано. Как мы уже говорили, распределения электронной плотности в покоящемся атоме, принято описывать при помощи гауссовых кривых. Вот такое размазывание можно также смоделировать в терминах гауссовых распределений электронной

плотности путем введения в этот параметр, который отвечает за ширину гауссовой кривой, дополнительной компоненты, которая и определяет вот эту степень "размазанности" атома. Здесь вот изображено распределение электронной плотности для атома азота. И теперь я показываю, как меняется это распределение, если мы начинаем добавлять в этой модели компоненту, отвечающую за размазывание атома по пространству. Понятно, что чем больше эта компонента, тем шире эти гауссовы кривые и тем более размазан этот атом.

56-62

Значение этого параметра B, который называется параметром тепловых колебаний или сейчас его более принято называть параметром атомных смещений, его можно связать со средним смещением атома в процессе этого движения. Существует такая формула, связывающая их. Мы видим, что если атом смещается в среднем на четверть ангстрема, то этому соответствует значение параметра B, равное 5, а если значение параметра B равно 100, то это значит, что атом в среднем смещается больше, чем на ангстрем.

63

Если имеет место такая схема "размазывания" электронной плотности, то говорят, что атом имеет изотропные температурные колебания, и величина B называется изотропным температурным фактором. Эта величина является одной из характеристик атома в молекуле и она записывается в соответствующую позицию в карте ATOM в сответствующем PDB-файле.

64

Но такая картина колебаний, при которой атом смещается равномерно во все стороны - она достаточно идеалистичная. В реальности ситуация более сложная, поскольку не все направления смещения атома эквивалентны, потому что, например, атом связан химическими связями с соседними атомами. И понятно, что, скажем, вдоль этой химической связи у него степень подвижности сильно ограничена. Поэтому существует более точная модель - так называемая модель анизотропных тепловых колебаний, при котором считается, что размазанное облако плотности имеет форму эллипсоида, и этот эллипсоид может быть вытянут, сплющен, то есть это не обязательно сфера. В таком случае этот эллипсоид описывается шестью параметрами. Эти шесть параметров, описывающие эллипсоид тепловых колебаний, приводятся в специальной карте ANISOU, которая тоже присутствует в PDB - файле для соответствующих атомов..

Я должен заметить, что если вам придется работать для каких-нибудь целей с анизотропными температурными факторами, надо работать очень аккуратно. Ситуация с использованием анизотропных параметров смещения еще не совсем устоялась, и в разных статьях вы можете встретить разные схемы введения анизотропных параметров тепловых колебаний, которые не всегда в точности совпадают друг с другом. И при записи в файл PDB тоже существует некоторые соглашения, о которых полезно знать. Например, что эти значения приводятся умноженными на тысячу и в целом формате. То есть со всем этим надо аккуратно разбираться, если вы вознамерились работать с анизотропными температурными факторами. Чтобы у вас не возникло рассогласования между тем, что вы думаете и что на самом деле записано в этом PDB-файле.

Теперь я хочу сказать, что хотя я здесь говорил "температурный фактор", этого названия сейчас пытаются в меру сил избегать. То название, которое используется сейчас - это "atomic displacement parameter". Дело в том, что вот

такое размазывание электронной плотности - оно связано не только с тепловыми колебаниями, но и вообще с общей неупорядоченностью атомов. Поэтому этот параметр, который присутствует в PDB-файле, он говорит не столько именно о тепловой подвижности, сколько вообще о неопределенности в координатах атома. То есть если вы встречаете координаты какого-нибудь атома с большим значением температурного фактора, то это значит - этот атом плохо локализован в пространстве, и к этим координатам надо относиться с соответствующей степенью доверия.

65

Теперь следующий параметр, который характеризует каждый атом - это так называемый коэффициент заполнения. Его иногда называют также заселенность, по-английски occupancy. И он появляется вот в связи с чем. Вообще-то, когда мы говорим про идеальный кристалл, мы подразумеваем, что содержание всех элементарных ячеек кристалла идентично. Но при реальной работе встречаются ситуации, когда эта идентичность несколько нарушается.

66

Первая распространенная ситуация такая: допустим, у вас есть молекула белка, которая лежит каждая в своей элементарной ячейке, и, кроме того, в кристаллах белка существует растворитель, существуют молекулы воды, и некоторые из этих молекул воды вполне хорошо и стабильно связаны с молекулами белка. Я имею в виду под этим то, что существуют определенные места на поверхности белка, с которыми связывается эта вода, и она связывается практически с каждой молекулой белка. Если она связана с каждой молекулой белка, то все хорошо, мы по-прежнему имеем ситуацию идеального кристалла. Но часто возникают ситуации, что эта связь не очень прочная, и поэтому для части молекул белка в этом месте вода присутствует, а для части молекул она отсутствует, поскольку здесь она не связалась, а находится где-нибудь в другом месте. Вот для того, чтобы описать такую ситуацию, применяется коэффициент заполнения, или заселенность. Этот коэффициент численно показывает, в каком проценте элементарных ячеек кристалла в указанной позиции этот атом имеется. То есть, например, если там для семидесяти процентов молекул белка в этом месте есть молекула воды, то у атомов этой молекулы воды коэффициент заполнения будет 0.7. А если она есть для всех, то он будет равен единице.

67

Коэффициент заполнения - он тоже является одним из параметров, характеризующих атом, и для него тоже здесь отводится специальная позиция в записи.

Как правило, просматривая глазом PDB-файл, по крайней мере, в начале этого файла, вы будете видеть, что этот коэффициент заполнения равен единице, поскольку в норме молекула белка присутствует во всех элементарных ячейках, и коэффициент заполнения у них посему единица. Но вот если вы будете смотреть дальше, в конце файла, координаты атомов молекул воды, то для молекул воды очень часто эта величина выставляется не равной единице и показывает, с какой точность эта вода связывается в этом месте.

68

Второй важный случай использования коэффициента заполнения - это так называемые альтернативные конформации полипептидной цепи или боковых цепей. Это связано с тем, что, вообще говоря, боковая цепь, да и основная цепь, они имеют довольно значительную конформационную подвижность, то есть

могут существовать в разных конформациях. И в то время как в центре молекулы белка все боковые цепи, как правило, зажаты со всех сторон и присутствуют в довольно стабильном состоянии, то боковые группы, смотрящие в растворитель, часто имеют большую подвижность. Для них нередко возникает ситуация, когда в разных копиях молекулы белка эта боковая цепь присутствует в разных конформациях. Для того, чтобы описать такую ситуацию, также используется аппарат коэффициента заполнения. То есть если какая-то боковая группа в части молекул белка находится в одной конформации, а в другой части - в другой конформации, то в PDB-файл с координатами атома заносятся атомы и той и другой конформации, но при этом этим атомам выставляются неединичные коэффициенты заполнения. То есть эта единица делится между этими двумя конформациями. Здесь она разделена строго пополам. Но, вообще говоря, соотношение может быть и другое: скажем, в семидесяти процентах случаев остаток находится в такой конформации, а в тридцати - в другой. В этом случае координаты для обеих конформаций будут выписаны, но коэффициенты будут 0.7 и 0.3.
Способ описания разных конформаций тоже еще не совсем устоялся. Иногда здесь буквы "A" и "B" различают (эти буквы соответствуют идентификатору цепи), иногда это делается по-другому - это зависит уже от конкретного автора.

Уточнение
69
Теперь я перехожу к следующему разделу исследования - уточнению параметров модели.
Когда получится синтез электронной плотности, его можно попытаться проинтерпретировать, то есть построить первоначальную атомную модель структуры. Раньше она строилась целиком руками, теперь существует масса автоматических программ, которые по электронной плотности эту модель пытаются построить. Но это еще достаточно приближенная модель, и она обладает еще не очень высоким качеством. Следующая стадия работы - это уточнение модели. Идея уточнения заключается в следующем: если у нас есть какие-то параметры модели (координаты атомов, температурный фактор, коэффициенты заполнения), то мы по этим данным можем рассчитать распределение электронной плотности и посчитать соответствующие значения модулей структурных факторов. После этого мы эти рассчитанные значения модулей структурных факторов можем сравнить с экспериментом. Ну, например, посчитать некоторый интегральный критерий - сумму квадратов отклонений рассчитанных значений от экспериментальных. После этого мы можем поставить перед компьютером задачу: пошевелить координаты наших атомов (ну, и, может, другие параметры) таким образом, чтобы рассчитанные значения наилучшим образом соответствовали эксперименту. Современные компьютеры такие задачи решать умеют. Эта процедура автоматического уточнения параметров модели является обязательной в каждом процессе определения структуры. То есть все структуры, помещенные в PDB-банк на финальной стадии работы прошли такую автоматическую процедуру уточнения структуры.

Минимизируемый критерий
70

Для того, чтобы характеризовать то, насколько хорошо рассчитанные модули совпадают с экспериментальными, исторически в кристаллографии принято применять так называемый R-фактор, который вычисляется вот по такой формуле. Это - сумма модулей отклонений рассчитанных модулей структурных факторов от экспериментальных, деленная на сумму экспериментальных значений модулей. С математической точки зрения это очень неудобный показатель, и поэтому реально при уточнении используется не он, а тот, который я вам показывал. Минимизируется такая функция. Но просто так исторически возникло, что точность воспроизведения эксперимента стали вычислять по такой формуле, и поэтому во всех статьях обязательно в качестве характеристики точности определения структуры присутствует эта характеристика.
Здесь приводятся параметры типичной задачи по уточнению структуры. Сравнительно небольшой белок - эндонуклеаза SM. Число атомов - 3694. Число независимых параметров - 36 940. Число померенных рефлексов - 108 000. То есть 108000 слагаемых в таких суммах. И вот эта задача решается.
Какие здесь возникают, прежде всего, проблемы?
Минимизируемая функция - она имеет очень много локальных минимумов, и поэтому возможно только немножко пошевелить эти параметры, потому что если пытаться их сильно изменить, то просто мы сваливаемся в какие-то ложные минимумы и там застреваем.
Ну, и второе, что на самом деле, если практически взять и применить при умеренном разрешении (скажем, при разрешении 2 ангстрема или полтора ангстрема) такую процедуру уточнения прямо в таком виде, то тоже ничего хорошего не получится. У нас R-фактор резво упадет вниз, но при этом модель "рассыплется", станет совершенно бессмысленной с химической точки зрения. Чтобы этого избежать, делаются следующая вещь:

71

Приступая к изучению структуры белка, мы, кроме данных, полученных в рентгеновском эксперименте, довольно много знаем о том, как должна была бы быть устроена эта структура. Ну, мы знаем, что это полипептидная цепь. Знаем обычно последовательность аминокислотных остатков в этой цепи. Кроме того, мы знаем, как локально устроены кусочки этой цепи. Например, вот стандартное пептидное звено, элемент полипептидной цепи. Мы знаем из исследования отдельных пептидов расстояние между всеми атомами этой пептидной группы, длины ковалентных связей. Поэтому, приступая к уточнению модели, мы можем сказать, что мы хотим не только иметь соответствие рентгеновскому эксперименту. Мы хотим еще, чтобы между атомами в нашей модели расстояние между атомами соответствовало тому, что мы уже знаем. То есть мы хотим иметь расстояние между этими атомами как можно ближе к 1.46Å, здесь - к 1.37Å и так далее. Для каждой пробной модели мы можем эти расстояния посчитать. Мы хотим, чтобы они совпадали с эталоном, который мы уже знаем из исследования малых структур. То есть формально мы можем, скажем, потребовать, чтобы вот такая функция - сумма по всем парам атомов отличий рассчитанных от точных значений была как можно меньше.

72-74

Но, поскольку мы, кроме того, хотим, чтобы у нас еще было максимально хорошее соответствие рентгеновскому эксперименту, то на практике берется

такой составной критерий: эти две суммы суммируются с некоторыми весами, и минимизируется эта сумма.

75-76

Кроме длин связи, знаем также, какие должны быть величины валентных углов. Поэтому мы можем потребовать опять-таки, чтобы в нашей модели и валентные углы тоже как можно меньше отличались от эталона, который мы знаем из исследования низкомолекулярных соединений. То есть вот в этот минимизируемый критерий мы добавляем еще один член, который требует, чтобы отклонение валентных углов в модели как можно меньше отличались от предписанных точных значений.

77

Ну, и так мы можем так еще добавлять разные типы информации. Мы можем потребовать, чтобы двугранные углы, задающие конформацию цепи, лежали в разрешенных пределах. Мы можем потребовать, чтобы атомы пептидной группы находились в одной плоскости.

78

Есть еще такое свойство - хиральность, связанное с тем, что аминокислоты химические могут существовать в форме двух изомеров: L-аминокислоты и D-аминокислоты. В то же время в живой природе в белках встречаются только L-аминокислоты. Поэтому мы можем потребовать, чтобы все аминокислотные остатки находились в L-конформации. Ну, и все это вместе собирается в такой составной критерий, и вот этот критерий мы и пытаемся минимизировать в процессе уточнения структуры.

Контрольные рефлексы

79

Теперь следующая вещь, которую я хочу сказать: поскольку математически задача уточнения очень сложная, поскольку у нас очень большое число параметров, которые мы можем менять, и у нас очень сложный характер функции, которую мы минимизируем, то, вообще говоря, в результате долгих усилий мы можем иногда попадать в ложные минимумы. То есть мы можем добиться, скажем, хороших значений R-фактора и других критериев, но при этом иметь достаточно неправильную структуру. Чтобы по возможности пытаться избежать этого, не так давно, примерно 10 лет назад, в уточнение структуры была введена идея так называемого контрольного набора рефлексов. Идея заключается в следующем. У нас есть набор экспериментальных величин, которые мы померили в эксперименте. Теперь этот набор делится на две группы, большая часть которых называется рабочими рефлексами, которая дальше используется для уточнения модели, и меньшая часть (ну, например, 5 или 10 процентов набора) - это контрольные рефлексы. Идея далее заключается в следующем. Когда мы уточняем модель, мы разрешаем двигать атомы и требуем при этом обеспечить соответствие рассчитанных и экспериментальных величин только для рабочей группы рефлексов. А контрольные мы как бы забыли и вообще отодвинули в сторону. А вот после того, как уточнение уже закончено, мы смотрим для контрольных рефлексов - улучшилось ли дли них соответствие эксперименту или нет? Если для них соответствие улучшилось, мы говорим, что да, хорошо, значит, уточнение шло в правильном направлении. Если соответствие не улучшилось - мы говорим, что наше уточнение нас завело куда-то в сторону, и это плохой результат.

80
Сейчас такая методология повсеместно принята. И, опять-таки, если вы посмотрите описание PDB-файла, то тут, как правило, приводится, по крайней мере, 2 значения R-фактора. Вот значение R-фактора для рабочего набора и R-free-value - это значение R-фактора, посчитанного только по контрольным рефлексам. Мы видим, что он реально больше, потому что для этих рефлексов мы не пытались принудительно их подогнать, они "сами" пришли к экспериментальным значениям. Но не так хорошо, как те рефлексы, которые мы принудительно подгоняли к этим значениям.

Частичная модель

81
Теперь я хочу сказать несколько слов еще об одной новой идее, которая появилась в последние годы в уточнении структур, которая связана вот с чем. В стандартном уточнении, повторю, мы делаем что:
у нас есть какая-то модель;
мы по ней рассчитываем модули структурных факторов и пытаемся подвигать координаты атомов модели так, чтобы эти модули совпали с экспериментальными максимально точно.
Это все хорошо и логично, если в модель включены все атомы структуры.
А теперь давайте себе представим такую ситуацию, что по каким-то причинам (ну, скажем, синтезы электронной плотности были недостаточно хорошие) мы не смогли увидеть всю структуру, и в модель включена только часть структуры. Настоящий структурный фактор в этом случае складывается из двух частей: структурный фактор той части структуры, которая включена в модель, и структурный фактор, отвечающий потерянной части модели. И в среднем, обычно, величина модуля настоящего структурного фактора будет больше, чем величина модуля структурного фактора, отвечающего частичной модели.
Теперь, если, тем не менее, мы начнем вот эту нашу процедуру уточнения, даже если мы предположим, что у нас частичная модель совершенно правильная, атомы находятся на правильных позициях, то в процессе уточнения мы начинаем требовать, чтобы величина этого модуля, тем не менее, совпадала с экспериментальным значением модуля, которое больше, и этого можно достигнуть, только как-то сдвинув атомы с правильных позиций. То есть в процессе уточнения атомы, даже если они были вначале на правильных позициях, начнут с этих правильных позиций уходить.

82
Чтобы бороться с такой ситуацией, была предложена концепция, которая называется Likelihood-based
Refinement. Это уточнение, основанное на максимизации правдоподобия. Здесь главная идея заключается вот в чем: мы меняем нашу цель уточнения. Мы больше не пытаемся рассчитанные по модели модули структурных факторов подогнать к экспериментальным, двигая координаты атомов. Наша цель ставится такая: найти такие координаты атомов частичной модели, которые позволят нам наиболее легко получить экспериментальные значения модулей после того, как мы дополним эту частичную модель некоторым количеством потерянных атомов.
Для этого обычно применяется смешанная детерминистско-вероятностная модель. И само правдоподобие - это есть вероятность того, что мы

воспроизведем правильные значения модулей структурных факторов после того, как в нашей частичной модели мы случайно добавим недостающее количество атомов. Если у нас частичная модель хорошая, атомы находятся в правильных позициях, то у нас есть некоторые шансы вот таким образом, случайно добавляя атомы, получить более или менее приемлемую модель. Если частичная модель плохая, то у нас шансы очень малы, что нас удастся ее дополнить, потому что она уже сама плохая. Ну, и такая идея, связанная с максимизацией правдоподобия, то есть максимизации того, что нам удастся получить правильные значения модулей после того, как мы добавим недостающие атомы - она сейчас получила широкое распространение, и наиболее популярные программы используют именно такой метод уточнения.

83

В описании, в разделе замечаний, это тоже часто отражается. Здесь вот эта самая запись - REFINEMENT TARGET : MAXIMUM LIKELIHOOD.

Процент успеха при расшифровке

84

Теперь, в заключение, я хочу просто сказать немного о том, насколько легко определить структуру интересующего нас белка, скажем, методом рентгеноструктурного анализа.

После того, как был определен геном человека, воодушевленные этим успехом люди выдвинули новую программу, которая называется "Структурный геном". В изначальной, наиболее максималистской форме это формулировалось так: давайте теперь определим структуру всех белков, присутствующих в организме человека. Ну, или в каком-то более простом организме.

Более острожные коллеги сформулировали более ограниченные задачи: давайте попробуем определить структуру всех белков, участвующих в каком-нибудь конкретном жизненно важном процессе.

Здесь приведены результаты исследований в одном из таких пробных проектов, который заключался в попытке определить структуру всех белков, участвующих в процессе развития пневмонии человека.

Всего были получены плазмиды для 274 генов. При этом удалось добиться экспрессии белка для 168 из этих 274. Удалось очистить до нужной степени белок для 43. Кристаллизовать - 24. В результате при этом для десяти удалось определить методами рентгеновского анализа структуру белков. К этим цифрам можно подходить, в зависимости от темперамента, по-разному. Можно говорить - целых 10 структур удалось определить. А можно сказать - из 274 только 10. Но это показывает современный уровень и возможности.

Хотя, должен сказать, что это был тестовый случай. И здесь авторы старались работать честно в том плане, что в этом процессе использовались те методы, автоматические, которые пригодны для такого массового определения структуры белков. То есть я думаю, что если еще взять некоторые отдельные белки, которые не удалось определить структуры, и с ними персонально повозиться достаточно долгое время, то еще какое-то количество структур определить можно.

Но, тем не менее, на таком автоматическом уровне пока что возможности этих структурно-геномных исследований примерно такие.

Но, тем не менее, эта программа сейчас широко обсуждается, и есть ряд консорциумов, которые ведут эти исследования, они определяют структуры,

помещают их в банк. Поэтому сейчас уже нередко можно встретить в банке помещенную какую-нибудь структуру белка, у которой в разделе "функции белка" значится "неизвестна". То есть определяется структура белков, про которые даже неизвестно, для чего они есть. А вот просто он есть - его определяют. Ну, это такое сейчас состояние дел.

Общая схема метода

85

Ну, и в заключение я хочу еще раз показать эту картинку с общей схемой метода. Мы получаем кристаллы исследуемого белка. Помещаем их в пучок рентгеновских лучей, меряем интенсивности рассеянных отражений. Извлекая квадратный корень из интенсивности, мы получаем модули коэффициентов Фурье в разложении в ряд Фурье функции распределения электронной плотности.

86

Далее, решая каким-то образом фазовую проблему, мы получаем фазы структурных факторов. После чего можем рассчитать синтез Фурье и проинтерпретировать его, получить предварительную модель. Дальше эта модель уточняется автоматическими программами уточнения. В результате чего мы и получаем, в конце концов, этот файл, который помещается в Protein Data Bank.

И здесь для каждого атома мы имеем вот эти 5 параметров - три координаты, коэффициент заполнения и температурный фактор.

87

Ну, и, наконец, здесь приведен список книг, которые можно использовать, если вам понадобится узнать более подробно про метод рентгеноструктурного анализа. Здесь приведена фотография, чтобы вы знали, что при этом вас ждет, если вы закажете себе ту или иную книгу. Я хотел бы отметить здесь персонально две книги.

Первая - это книга Михаила Александровича Порай-Кошица. Михаил Александрович Порай-Кошиц - это один из основателей рентегноструктурного анализа в Советском Союзе. Это небольшая, очень хорошо написанная книжка. Михаил Александрович долгие годы работал профессором химфака МГУ. И эта книжка - это учебник для студентов химфака. Она хорошо написана, но она посвящена рентгеноструктурному анализу вообще и не рассматривает особенности рентгеноструктурного анализа для белков.

Ну и вторая книга, о которой я хочу упомянуть - это вот этот фолиант, который носит несколько обманчивое название "Международные кристаллографические таблицы". На самом деле это никакие не таблицы. Это - энциклопедия белковой кристаллографии. Эта книга содержит много статей, написанных наиболее крупными современными специалистами в области белковой кристаллографии. И она содержит уже достаточно профессиональное описание всех проблем и стадий рентегноструктурного анализа белков.

На этом наш курс лекций закончен. Всего хорошего.