ФАКУЛЬТЕТ
БИОИНЖЕНЕРИИ И
БИОИНФОРМАТИКИ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ

МОДЕЛИРОВАНИЕ СТРУКТУР БИОПОЛИМЕРОВ

ГОЛОВИН
АНДРЕЙ ВИКТОРОВИЧ

ФББ МГУ

Головин Андрей Викторович

Конспект лекций

МОДЕЛИРОВАНИЕ СТРУКТУР БИОПОЛИМЕРОВ

Оглавление

Лекция 1. Вводная лекция

Познание строения вещества

Людей давно интересует, как устроено вещество. Достаточно давно было предположено, что вещество состоит из частиц, и только около ста лет назад были получены экспериментальные данные о том, как эти частицы расположены друг относительно друга в пространстве. Для этого использовали рентгеновское облучение кристаллов, которое давало некоторые рефлексы, и на основе этих рефлексов определяли структуру.

Тем не менее, рентгеноструктурный анализ не даёт информации о функционировании исследуемой структуры, и можно лишь предполагать, за что могла бы отвечать часть структуры, на основе накопленного опыта и информации о работе изучаемых структур.

Для чего нужны модели?

Методы моделирования позволяют:

- упростить сложный объект до анализа только той части, которая предположительно является объектом интереса, и аппроксимировать условия;
- проводить дедуктивный анализ очень сложных или многочисленных явлений
- достичь понимания рассматриваемой системы, несмотря на то, что отражают реальность не полностью

Часто очень точные параметры системы, например, электронные эффекты, сглаживаются и становятся не очень значимыми, если брать среднее значение. Поэтому следует начинать с самых грубых моделей. Более того, грубые модели требуют меньше затрат ресурсов, так что это оптимизация процесса.

Самый важный последний этап моделирования – дизайн, то есть проектирование и создание новых белков. Это уже возможно, однако до сих пор остаётся довольно сложной проблемой.

Масштабы в моделировании

Самый маленький значимый элемент с точки зрения химии – это электрон, поскольку он образует ковалентные связи. Распределение электронной плотности приводит к электростатическим взаимодействиям или Ван-дер-Ваальсовым взаимодействиям и так далее. Таким образом, электронная плотность является самым маленьким интересующим нас объектом для исследования структур белков.

Белки существуют при температурах больше нуля и меньше ста градусов Цельсия и не обладают такими энергиями, при которых свойства элементарных частиц имеют хоть какое-то влияние на макросвойства системы. В связи с этим нас не интересует физика квантовых явлений элементарных частиц для исследования структур белков.

Электроны в данном случае рассматриваются как участники образования ковалентных связей, которые интересны нам, поскольку изрядное количество белков

являются ферментами – катализаторами химических реакций, которые протекают с разрывами и образованиями химических связей.

К сожалению, компьютеры могут обсчитывать системы ограниченного размера, будь то электроны, атомы или липиды, так как продолжительность обсчёта прямо зависит от числа частиц. В связи с этим электроны считаются лишь в небольших системах: максимум десятки ангстрем в размере и наблюдать за ними можно всего несколько фемтосекунд (Рис. 1.1).

В моделировании очень важно исследовать движения атомов, движение их окружения, так как энтропийные эффекты очень важны для биополимеров (к примеру, наибольшее влияние на структуру белков оказывает гидрофобный эффект, который имеет энтропийную природу).

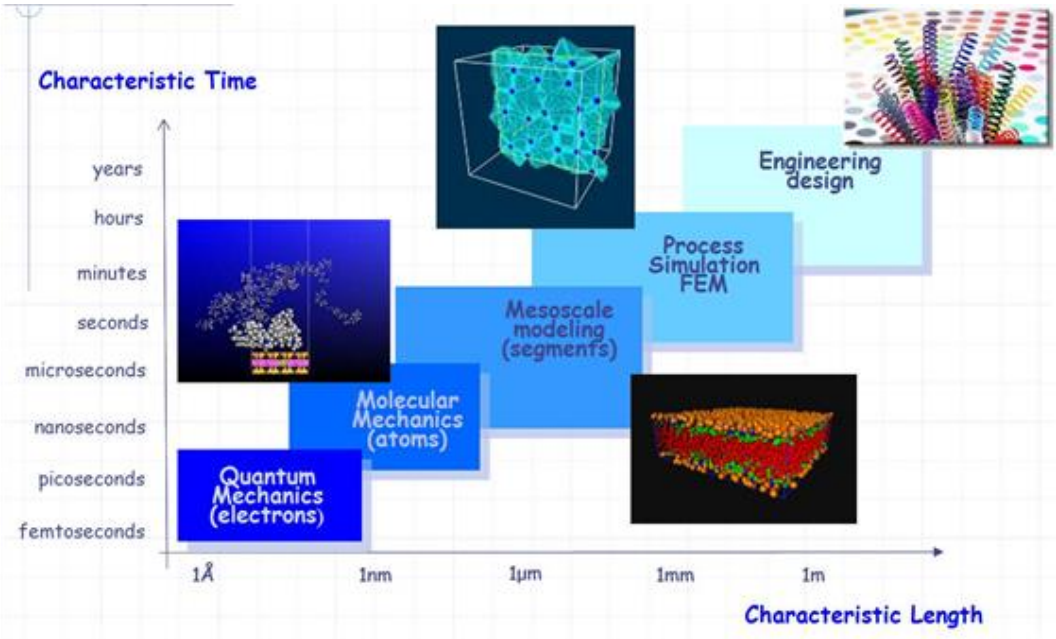


Рисунок 1.1 - Масштабы в моделировании.

Представление атома в качестве одной частицы в моделировании – некая попытка суммировать свойства ядра и электронов, которые его окружают. Естественно, суммирование сильно зависит от окружения, так как доноры электронной плотности увеличивают количество электронов на атоме, а акцепторы уменьшают. В результате описание атома зависит от химического контекста.

Поскольку количество объектов, которые нужно считать, уменьшаются, системы увеличиваются до размеров сотен нанометров и время наблюдения увеличивается до микросекунд (Рис. 1.1). Характерное время биологических процессов происходит в диапазоне до микросекунд (внутри белка).

Представление группы атомов в виде одной частицы называется крупнозернистым моделированием. В настоящий момент данный способ моделирования позволяет строить довольно крупные системы, такие как липосомы.

Под моделированием сплошных сред подразумеваются программы по типу CAD, которые позволяют рисовать некие детали, которые между собой взаимодействуют.

Формально для решения задач компьютер не является обязательным элементом. В частности, в терминах современного мира существует тенденция вместо глобального счёта использовать нейронные сети и машинное обучение.

Быстрый компьютер значительно увеличивает точность и широту исследования, и, следовательно, достоверность моделирования.

Количество вычислений отражает степень исследования конформационного пространства.

Однако важно иметь в виду, что любая программа выдаёт ответ на входные данные, и ответственность за релевантность полученных результатов лежит на исследователе, поэтому всегда стоит относиться к наблюдениям критически и проверять данные.

PyMOL

PyMOL – программа, осуществляющая следующие функции:

- визуализация моделей молекул: pdb и прочих файлов с координатами атомов
- изготовление высококачественных изображений
- начальное редактирование структур

Что касается системных требований, то желательна хорошая видеокарта (особенно для работы с четырёхмерными данными в трёхмерном пространстве, например, отображение электронной плотности) и большой объём оперативной памяти (для отображения моделирований с большим количеством состояний, так как PyMOL объективирует фреймы). В Linux использование памяти лучше, 3D монитор не обязателен, но поддерживается.

Установить PyMOL можно следующими способами:

- компиляция из исходников
- установка бинарных пакетов из репозиториев дистрибьютора
- установка с Conda

PyMOL – это GPL программа.

Меню объекта/выборки (Рис. 1.2):

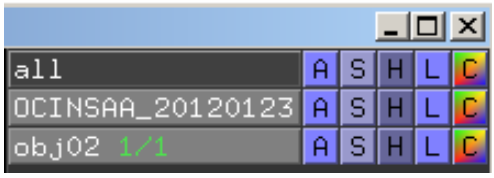


Рисунок 1.2 - Меню объекта/выборки, где A – Action; S – Show; L – Label; C – Color.

Выборки можно задать с помощью кликов мыши, удерживая SHIFT или при помощи выражений в командной строке.

Например: s*elect backbone, name ca+c+n*

Для множеств (выборок) можно применять стандартные логические операторы AND, OR (можно записывать как «,»), NOT и оператор WITHIN(…)

Пример: *select s1, (byres n. ca) within 3.5 of resn LIG*

byres расширяет выборку до остатка.

Актуальную информацию по операциям с выборками можно узнать в https://pymolwiki.org/index.php/Selection_Algebra

Диапазоны (выделить все цепи с A по C) обозначаются знаком «минус (A-C).

Оператор OR можно обозначать знаком «+».

У всех атомов есть полный объектно-ориентированный идентификатор, который можно увидеть, кликнув правой кнопкой мыши по атому. Этими иерархическими идентификаторами можно пользоваться как выборками.

Пример: *sel s1, a/102/cz* – атом cz в остатке 102.

Иерархия задаётся следующим образом: идентификатор объекта/идентификатор сегмента/идентификатор цепи/идентификатор (номер) остатка/имя атома.

По иерархии при создании выборки можно идти как слева, так и справа.

Команда *ray* создаёт лучи (световые), при помощи неё можно задавать несколько источников света, длину тени, длину пробега луча и прочее. Важный параметр – тип трассировки лучей (*ray_trace_mode*) (Рис. 1.3):

0 – простой тип (улучшает изображение)

1 – чёрная обводка

2 – есть обводка, но цвета не учитываются (белый)

3 – «индексированные» цвета

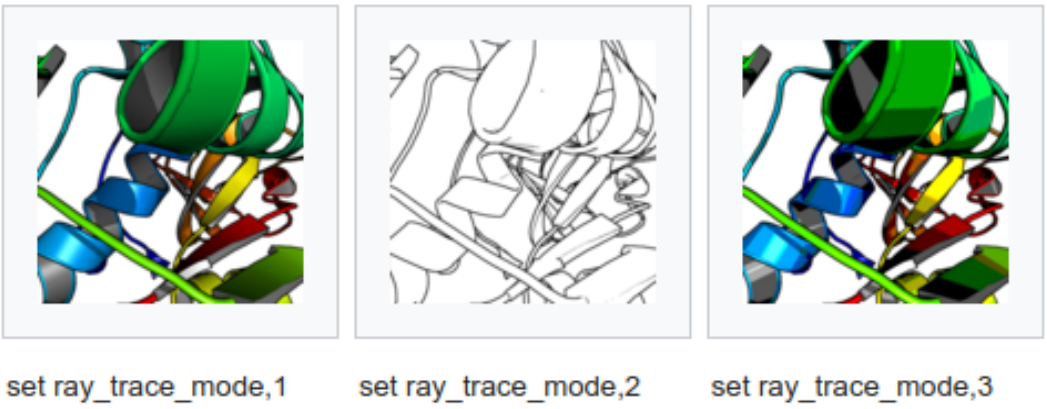


Рисунок 1.3 - Типы трассировки лучей в PyMOL.

PyMOL содержит порядка 600 настроек изображения (сейчас, наверно, уже больше). Не все из них документированы, большинство интуитивно понятны. Они доступны через меню или в командной строке:

set первые буквы опции и клавиша Tab для достроения

В файле pdb, содержащем результаты ядерно-магнитного резонанса, имеется несколько моделей. Их можно проиграть одну за другой – это основной вид анимации, хорошо реализованной в PyMOL.

Терминология анимации:

- объект и выборка – такие же, как и в отображении
- states – конформация или набор координат
- scene – положение камеры и отображение объекта
- frames – кадры в анимации, содержит state и scene

Команды анимации (примеры):

mset 1 -50 : задать анимацию от 1 до 50 state на 50 кадров (frames)

mset 1 x90 : задать анимацию первого state от 1 до 90 кадров

mset 1 x30 1 -15 15 x30 15 -1 : первые 30 кадров state 1, следующие 15 кадров – состояния 1-15, следующие 30 кадров состояние 15, следующие 15 кадров – состояния с 15 до 1.

mview : команда для создания ключевых точек

Актуальная полезная информация по этой теме:

https://pymolwiki.org/index/php/MovieSchool

Возможности моделирования и редактирования в PyMOL:

- перемещение объектов и сохранение их новых координат
- рассчёт вторичной структуры
- изменение координат отдельных атомов
- внесение мутаций в белок
- конвертирование L→D аминокислот
- добавление протонов
- выравнивание молекул в пространстве
- добавление фрагментов из библиотеки и собственных, из которых можно строить любые полимеры

При перемещении объектов рекомендуется начинать с сохранения порядка атомов (*set retain_order*).

Для манипуляций лучше всегда создавать новый объект (*create newobj, sele*), дальше его можно перемещать (*translate [0,10,0], newobj*), вращать (*rotate x,90,newobj*), и сохранить новый файл (*save newfile.pdb, newobj*). При манипуляциях всегда стоит указывать объект, с которым производится работа, иначе переместятся/повернутся все объекты системы.

Команда *alter_state* позволяет изменять координаты отдельных атомов и объектов.

В режиме редактирования структур можно двигать структуры мышкой и удалять связи, но не атомы, выбрав первый атом (Ctrl+middle click), второй атом (Ctrl+middle click) и unbond или Ctrl+D.

Мутация аминокислот осуществляется при помощи wizard → mutagenesis.

Для добавления протонов используется команда *h_add*, также это можно сделать через меню action объекта. При добавлении протонов нужно быть аккуратным, так как PyMOL не всегда правильно определяет валентности.

Суперпозиция хорошо работает для белков (можно использовать *align, super, fit*), поскольку в них есть аминокислоты, где атомы идентичны, и по ним можно найти похожие участки последовательностей. Для суперпозиции органических молекул стоит использовать команду *pair_fit*. Желательно указывать родственные атомы в молекулах.

Также в PyMOL есть широкие возможности для добавления органических фрагментов или аминокислот. Для этого используется меню Build.

Sculpting – алгоритм PyMOL, который старается сохранить значения длины связей, углов, торсионных углов при изменении координат.

Скриптование в PyMOL предполагает возможность создавать как скрипты из команд (*@ myfile.pml*), так и скрипты на Python (*run myfile.py*). Более того, возможны гибриды: внутри скрипта с командами PyMOL можно вызывать Python, в скрипт на Python можно импортировать PyMOL.

Объекты из PyMOL можно экспортировать в форматы для программ трёхмерной графики, благодаря чему можно применять более сложные структуры при визуализации.

Лекция 2. Хемоинформатика

Активные молекулы

В данной лекции речь пойдёт о такой области науки, как хемоинформатика, которая занимается моделированием небольших органических молекул: ингибиторов белков и лекарств.

Живая клетка активно работает с органическими молекулами, стимулирует их химические превращения, использует сложные каскады химического синтеза для построения белков, нуклеиновых кислот и других веществ. Естественно, небольшие органические молекулы играют важную роль в этих процессах: как субстраты для энзиматического синтеза или как регуляторы.

Чаще всего небольшие молекулы взаимодействуют с биополимерами только нековалентным способом. К одному из случаев ковалентного взаимодействия относятся интермедиаты энзиматических превращений: небольшая молекула ковалентно присоединяется к ферменту, ждёт следующую (целевую) молекулу, и происходит перенос группы на неё. В результате фермент остаётся неизменным, хотя и образовывалась ковалентная связь. Однако зачастую регуляторные и иные взаимодействия биополимеров с небольшими органическими молекулами происходят за счёт нековалентных механизмов.

Номенклатура активных молекул:

- Агонисты связываются как нативные лиганды и вызывают такой же эффект
- Антагонисты конкурируют или препятствуют связыванию нативного лиганда, тем самым блокируя эффект, который вызывают нативные лиганды
- Обратные агонисты связываются и оказывают эффект, обратный эффекту нативного лиганда

Поскольку номенклатура берётся из области взаимодействий молекул с рецепторами (основные системы передачи сигнала в эукариотических клетках), она ориентирована на лекарства, и «обратный эффект» в данном случае – передача такого типа сигнала в клетку, который даёт изменение сигнального пути (обратный эффект).

В связи с тем, что существует большое количество белков, и они обладают большими поверхностями, для специфичного связывания лигандов необходима высокая комплементарность поверхности лиганда и белка. Под «специфичным» связыванием обычно понимается, что афинность к данному месту для специфических молекул является максимальной по отношению к другим местам, существующим на всех поверхностях всех белков. Но это может быть и не так: иногда молекулы, например, занимающиеся транспортом, могут эффективно взаимодействовать, однако под «специфичными» взаимодействиями будут пониматься те, которые приводят к нужному результату. Желательно, чтобы они были самые лучшие по афинности, но это не всегда реализуемо.

Таким образом, хорошие лекарства – не обязательно те, которые лучше всех связываются. Важен эффект, который оно оказывает на белок, что можно точно проверить только функциональными тестами.

Лекарство должно иметь приемлемую растворимость в воде, но иногда бывает, что ему нужно проходить через мембрану, если мишень находится внутри клетки. В таком случае это вещество должно быть не настолько хорошо растворимо в воде, чтобы иметь возможность десольватироваться для проникновения сквозь мембрану.

Самое идеальное лекарство – то, которое будет метаболизироваться, превращаться на безопасные компоненты, желательно, на метаболиты (вещества, не накапливающиеся в организме, а участвующие в синтезе, расходующиеся, легко выводящиеся и так далее). Плохо метаболизирующиеся лекарства оказывают нагрузку на почки и печень, что особенно важно при употреблении лекарств при хронических заболеваниях.

Как искать активные молекулы?

Раньше все активные молекулы искали в биоматериалах (преимущественно в растениях).

На данный момент можно с уверенностью утверждать, что вся биомасса на Земле уже просканирована на молекулы, которые могли бы быть активными. Возможно, остались какие-то минорные соединения, но присутствующие в явном виде активные вещества в растениях и других организмах уже давно извлечены и хранятся в базах данных, либо модифицированы для использования.

Из экспериментальных методов развивается робототехника, которая позволяет определённую библиотеку соединений просканировать на активность в разных тестах. Пропускная способность подобных экспериментов может быть очень большой – десятки тысяч соединений в день. Однако не всегда можно адаптировать тесты под робота. Самый простой пример – антидепрессанты, поскольку они вызывают сложный каскад сигналов. В таком случае проще работать с целыми животными.

Другая проблема роботизированных сканирований: неспецифические взаимодействия сканируемых веществ с белком (агрегирование соединения, различное распределение по фазам и так далее). Их довольно сложно выявить на начальном этапе, поскольку используется большая библиотека. В связи с этим может возникать довольно большой шум, особенно в небинарных системах (где компонентов больше, чем два – белок и небольшая молекула).

Здесь возникают вопросы, которые могли бы быть решены на основе компьютерного анализа: фильтрация по подобию соединений.

Особенности деятельности фарм-производителей

Дженерик – лекарство, срок патентной защиты которого вышел. Основную прибыль приносят новые лекарства, поскольку рынок высоко конкурентен, их разработка дорого стоит, и они защищаются патентом. Большую часть ресурсов создания новых лекарств занимает исследование (в основном, токсичности и альтернативных способов действия вещества), в результате чего сам синтез вещества стоит только 5%, остальное – покрытие расходов на исследование и прибыль.

Четыре основные фазы разработки нового препарата: открытие, разработка, испытания, продажа.

Обычно сначала проводят исследование, выявляющее причину болезни. Далее для белка, который отвечает за данную патологию, применяют методы для установления его структуры и ищут ингибитор. Если структуру удалось установить, для поиска ингибитора применяют компьютерные методы. В противном случае можно попытаться найти ингибитор с помощью скрининга соединений. На этом заканчиваются академические исследования и начинаются медицинские, основная задача которых не понять, как лекарство работает, а определить, насколько оно ядовито и можно ли его применять на человеке.

Первым этапом медицинских исследований является испытание на животных (относительно недорогая и быстрая стадия). Далее идёт самая сложная стадия – попытка сделать конкретное лекарство, масштабировать, организовать синтез, причём в рамках российского законодательства для того, чтобы начать клинические исследования на людях, производство должно находиться в России. Этот этап требует довольно больших и рискованных инвестиций, поскольку неизвестно, будет ли лекарство работать на людях. Дальше идут клинические испытания, после успешного прохождении которых лекарство попадает на рынок.

Современные технологии позволяют добиться довольно большой скорости на академическом этапе (в основном, на стадии поиска ингибитора):

- Чипы: экспрессия генов
- Структуры: роботизированный поиск комплексов с кристаллом белка
- Высокопроизводительный поиск ингибиторов
- Виртуальный поиск
- Комбинаторная химия

Практически во всех этих методах может помочь хемоинформатика:

- Разработка методов и управление информацией о лигандах (построение баз данных)
- Оценка данных *in silico* для минимизации рисков:
  - Разработка библиотеки
  - Виртуальный поиск
  - Оценка стоимости выгоды
- Организация доступа к информации
- Интеграция процессов

High-throughput screening (HTS, высокопроизводительный поиск ингибиторов) – типичный пример такой технологии, возможна обработка до 100000 соединений в день.

Хемоинформатика использует схожие с биоинформатикой подходы:

- Исполнение HTS
- Решить, какие соединения активны, а какие нет
- Кластеризация активных соединений в классы
- Визуализация

- Идентификация «основы» для каждого класса
- Поиск причин, элементов структуры, которые приводят к «не активности»
- Использование структурной информации для объяснения активности

Комбинаторная химия – способ создания библиотеки соединений, для построения которых используются «строительные блоки», комбинирующиеся между собой на некой «основе».

Хемоинформатика позволяет адаптировать комбинаторный синтез при помощи создания виртуальных библиотек.

Компьютерное представление молекул

Хранение в компьютере молекулы как изображения имеет малую ценность для компьютерного анализа. Большинство современных баз данных представляет молекулу как граф, в котором рёбрами являются связи. Типичный способ представления молекулы в виде графа приведён на Рис. 2.1: описываются атомы, их положение и связи (на рисунке внизу).

Популярный способ описания органических молекул в линейном представлении – SMILES. В данном случае молекула представляется в виде графа, в котором каждый атом обходится только один раз. Таким образом, можно описать молекулу как одну строчку (string). SMILES позволяет описывать большое разнообразие особенностей строения химических веществ, химические реакции.

Очевидно, что одну молекулу можно описать разными способами, и в 1965 году Морган предложил рассматривать каждый атом по свойству его окружения, благодаря чему можно унифицировать запись SMILES. Такие стандартные SMILES называют Unique.

Описание атомов в SMILES осуществляется следующим образом: однобуквенные атомы записываются как есть, одним символом, остальные атомы записываются в квадратных скобках. Атомы Cl и Br можно записывать без скобок.

```
Marvin  04200617372D
  4  3  0  0  0  0            999 V2000
    0.0000    0.0000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.4125    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.8250    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
  1  4  2  0  0  0  0
  2  1  1  0  0  0  0
  3  1  1  0  0  0  0
M  END
```

Рисунок 2.1 - Компьютерное представление молекулы уксусной кислоты в виде графа.

Так как атомы водорода обычно не указываются, «валентность» атомов определяется как наименьшая из ближайших. «Валентности», отличные от «нормальных», указывают в скобках: [H+], [Fe+2], [Fe++].

Принципиальное отличие ситуации с водородами графов от 3D-структуры заключается в том, что при анализе 3D-структуры кратность связи устанавливается на основе расстояния между атомами, из-за чего возможны ошибки (короткая двойная связь, длинная одинарная) при добавлении водородов. В случае графов такой ошибки быть не может, поскольку связи прописаны в явном виде.

Описание связей в SMILES:

- одинарные – не пишутся (CC, этан, формула: $H_3C\text{-}CH_3$)
- двойные – знак «равно» (C=C, этилен, формула: $H_2C{=}CH_2$)
- тройные – знак «решётка» (C#N, HCN)
- ветвление указывается с помощью круглых скобок

Для описания циклов в SMILES используются численные индексы. Между атомами с одинаковым численным индексом образуется одинарная связь, при этом атом может иметь более одного индекса.

Примеры:

C1CCCCC1 – циклогексан, более сложные примеры на Рис. 2.2, 2.3.

Атомы, написанные маленькими буквами в SMILES считаются ароматичными. Таким образом можно упростить запись ароматических циклов. При этом для определения ароматичности используется расширенный алгоритм Хюккеля, и если молекула ему не удовлетворяет, запись маленькими буквами эквивалентна записи алифатического соединения.

Ароматичными атомами могут быть: C, N, O, P, S, As, Se и * (произвольный атом).

Пример: c1cnc[nH]c(=O)1 (Рис. 2.4).

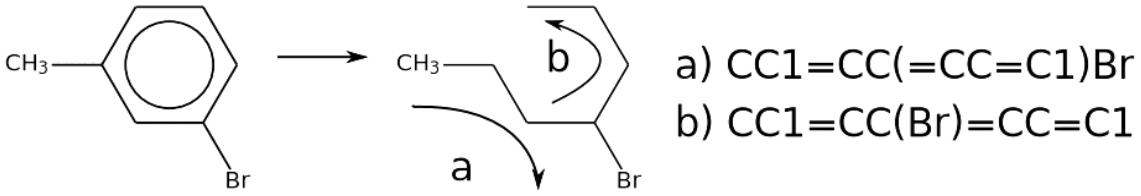


Рисунок 2.2 - Два способа замкнуть цикл в SMILES на примере 3-бромтолуола.

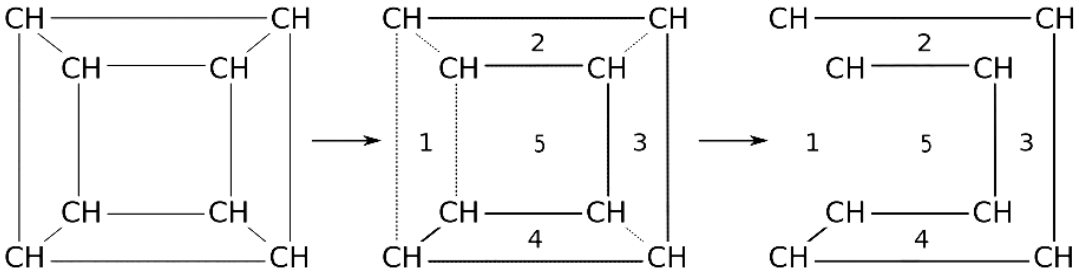


Рисунок 2.3 - Кубан в SMILES: C12C3C4C1C5C4C3C25.

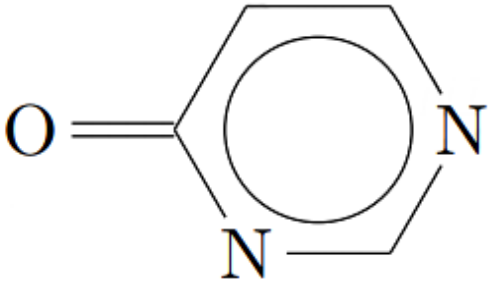


Рисунок 2.4 - c1cnc[nH]c(=O)1 в SMILES. Один из азотов имеет протон.

Нековалентные взаимодействия в SMILES никак не описываются, но можно описать смесь двух веществ через «точку»: [Na+].[O-]c1ccccc1 или c1cc([O-].[Na+])ccc1 (Рис. 2.5).

Изотопы указываются при помощи номера перед обозначением элемента: [12C], [13C],

цис-транс изомеры различаются при помощи слэшей, направленных в разные стороны: F/C=F/C и F\C=F\C – транс-изомеры, F/C=F\C и F\C=F/C – цис-изомеры.

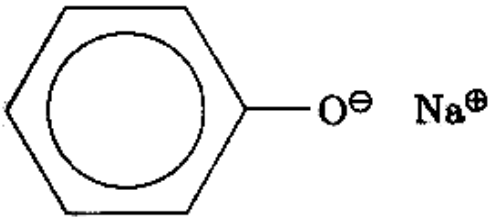


Рисунок 2.5 - Описание смеси веществ в SMILES: [Na+].[O-]c1ccccc1.

Для описания хиральности используется знаки «@»:

- @ – расположение лигандов по часовой стрелке
- @@ – расположение лигандов против часовой стрелки

SMARTS – паттерны для SMILES, то есть операторы логики и варианты в позициях.

Пример для атомов:

- C – алифатический углерод
- c – ароматический углерод
- a – любой ароматический атом
- [#6] – любой атом углерода
- [++] – атом с зарядом 2+
- [R] – атом в кольце
- [D3] – атом с тремя связями (не с атомами водорода)
- [X3] – атом с тремя связями, включая атомы водорода
- [v3] – атом с валентностью водорода

Логические операторы:

- !e1 – не e1
- e1&e2 – e1 и e2
- e1,e2 – e1 или e2
- e1;e2 – e1 и e2

Примеры:

- [!C;R] – не алифатический C в кольце
- [n;H1],[n&H1],[nH1] – H в пирроле
- [c,n&H1] – C или H в пирроле
- [X3&H0] – атом с тремя связями не с H
- [c,n;H1] – N или C в связи с одним H1

Таким образом, SMARTS представляет собой способы поиска атомов по паттернам в молекулах.

Другой способ линейного представления молекул – InChI (IUPAC International Chemical Identifier). Структура молекул описывается слоями, которые разделены слэшами:

- Основной слой – брутто формула, дальше идёт описание связанности (c), связей с водородами (h). Пример: C2H6O/c1-2-3/h3H,2H2,1H3
- Слой с описанием заряда (p) и кратности связей
- Слой с описанием стереохимии и связей

Дискрипторы, правило Липински

Дискрипторы – аналог features в машинном обучении. Некоторые примеры дискрипторов:

- Водородные связи (доноры, акцепторы водородных связей в молекуле)
- Гибкость молекулы
- Гидрофобность

Хорошее лекарство должно удовлетворять неким правилам. Так называемое «правило Липински»:

- в молекуле не должно быть больше пяти доноров водородной связи

- не больше десяти акцепторов водородной связи
- молекулярная масса молекулы должна быть меньше, чем 500 дальтон
- коэффициент гидрофобности должен быть больше, чем 5

Коэффициент гидрофобности (P) – отношение содержания вещества в воде и октаноле после помещения исследуемого вещества в данную смесь и перемешивания. Для него используется логарифмическая шкала.

Данные дискрипторы можно легко получить из формулы вещества, построенной и обработанной, например, на основании SMILES, так как доноры и акцепторы водородной связи прописываются паттернами, масса молекулы легко высчитывается, для высчитывания гидрофобности у каждого радикала существует некая оценка (score) в данном коэффициенте (log P), и коэффициент гидрофобности считается как сумма этих оценок радикалов.

Поиск по 3D-базам данных

Поиск в 2D-пространстве хорош для поиска подобных молекул, но биологически активные молекулы действуют благодаря специфической 3D-структуре.

Взаимодействие с биополимером может происходить благодаря нужному расположению в пространстве некоторых групп. При этом различие в 2D-структуре может быть весьма существенным.

Фармакофор – набор свойств, которые являются общими для некоторой группы активных молекул. Фармакофоры обычно строятся на основе анализа библиотек соединений.

Проблемы с фармакофорами:

- Если молекулы более или менее подвижны, это накладывает дополнительные требования на учёт конформационных превращений.
- Для определения фармакофора надо определить, какой набор групп располагается идентично.
- Надо быть уверенным, что выбранный набор молекул связывается с белком в одном и том же месте. Однозначное указание на это можно получить только экспериментально.

Базы данных:

- PubChem: можно искать молекулы, содержащие искомый SMILES, доступен с помощью PUG (Power User Interface) через pubchempy
- Cambridge database
- Inorganic structural database

В последнее время эта область получила большое ускорение развития за счёт использования машинного обучения для поиска ингибиторов, однако на сегодняшний момент остаются важные вещи, которые требуют изменений:

- Методы, основанные на подобии, рассматривают подобные вещества и белки, равномерное распределение отсутствует. Это приводит к переобучению.
- Описание features должно быть не бинарным, а количественным.

- Методы основаны на datasets. Нужна адаптация под успешные предсказания.
- Объединение баз данных. Комбинирование максимально доступного количества данных белок-ингибитор.
- Правильное включение структурно-функциональных данных для лигандов и белков.

Лекция 3. Введение в квантовую химию

Введение в квантовую химию

Поскольку ферменты осуществляют химические реакции, которые происходят с помощью электронов – квантовых частиц, важно изучать квантовую химию. Результаты квантово-химических расчётов формируют наши представления о молекулах, в частности, они используются при изучении белков, ферментов.

Квантовая химия – направление химии, рассматривающее строение и свойства химических соединений, реакционную способность, кинетику и механизм химических реакций на основе квантовой механики.

Эта вычислительная наука зародилась на стыке с методами спектроскопии (IR, NMR), поскольку необходимо соотносить получаемые спектры со строением вещества. Для этого необходимо понимать, как устроена электронная плотность, и квантовая химия помогает ответить на этот вопрос.

Нередко основой вычислений квантовой химии являются методы ab *initio* (из первых принципов), то есть это методы, основанные на представлении о строении окружающего мира. В данном случае имеются в виду базовые физические константы: скорость света, постоянная Планка, масса и заряд электрона. Многие расчёты квантовой химии даже не нуждаются в экспериментальных подтверждениях, так как они не являются эмпирическими, а рассчитаны из первых принципов.

Энергия и вещество

Наши представления о мире подразумевают корпускулярно-волновой дуализм. В физике это означает, что некие частицы могут обладать как корпускулярными свойствами, то есть свойствами частицы, так и свойствами волны. Естественно, чем меньше частицы, тем более явно проявляются свойства волны.

В частности, электроны находятся где-то на середине этой шкалы, они обладают явно выраженными волновыми и корпускулярными свойствами. Существует эксперимент, в котором электроны ведут себя частицы при свечении ими на узкую щель, однако при размещении рядом второй щели возникает дифракционная картина, что свидетельствует о волновых свойствах электрона.

Связь между волновым и корпускулярным описаниями вещества выражается соотношением де Бройля:

$$p = \frac{h}{\lambda} \tag{3.1}$$

где p – импульс, h – постоянная Планка, λ– длина волны.

Волна де Бройля:

$$f = Ae^{ikx} \tag{3.2}$$

где f – плотность вероятности, x – координата пространства, k – константа.

Отсюда возникает понятие волновой функции.

Волновая функция

Каждое состояние системы из n частиц в каждый момент времени может быть описано некой волновой функцией. Это комплексная функция координат частиц x_i и времени t, $\psi(x_1, x_2, \ldots, x_n, t)$, она должна быть непрерывна и дифференцируема на всём протяжении базиса, в котором рассматривается поведение системы.

Выражение $\psi * (\{x\}, t)\psi(\{x\}, t)dx_i$, где $\{x\}$ – совокупность координат частиц, имеет смысл вероятности того, что в момент t частица i находится в интервале $x_i, x_i + dx_i$. Здесь речь идёт о неком интервале возможных положений частицы, поскольку существует принцип неопределённости Гейзенберга (невозможно одновременно с точностью определить координаты и скорость квантовой частицы).

Физический смысл волновой функции заключается в том, что согласно копенгагенской интерпретации квантовой механики плотность вероятности нахождения частицы в данной точке пространства в данный момент времени считается равной квадрату абсолютного значения волновой функции этого состояния в координатном представлении.

Уравнение Шредингера

Суть уравнения Шредингера сводится к тому, что электрон (вариант его движения – вектор *r*) движется во внешнем поле *V*. Не релятивистский вариант для одной частицы:

$$i\hbar\frac{\partial}{\partial t}\psi(r,t) = \left[\frac{-\hbar^2}{2m}\nabla^2 + V(r,t)\right]\psi(r,t);\ \nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{3.3}$$

$\hbar$– постоянная Планка; ψ– волновая функция.

Выражение в квадратных скобках называется гамильтонианом и является оператором полной энергии (суммы кинетической и потенциальной (*V(r,t)*). Его собственное значение является полной энергией.

Оператор Гамильтона *H* (гамильтониан):

$$H = \frac{-\hbar^2}{2m}\nabla^2 + V \tag{3.4}$$

тогда уравнение Шредингера примет вид:

$$H\psi = E\psi \tag{3.5}$$

Для решения этого уравнения надо найти значение *E* и волновую функцию так, чтобы уравнение выполнялось.

Это уравнение относится к типу дифференциальных уравнений с собственными значениями, где оператор, действующий на функцию, возвращает произведение скалярной величины на функцию. То есть при решении мы получим и волновую функцию, и значение энергии для какого-то состояния, которое описывается этой функцией. Эта волновая функция может использоваться для многих состояний или экстраполироваться.

Типичный пример того, как работает оператор (производная по *x*):

$$\frac{d}{dx}(y) = ry;\ \ if\ y = e^{ax}\ \ then:\ \ r = a$$

Ожидаемое значение (можно рассматривать как среднее значение) какого-либо свойства: энергии, положения, линейного момента, можно определить с помощью оператора.

Пример: ожидаемое значение, используя гамильтониан (оператор энергии):

$$E = \frac{\int \psi * H\psi \partial r}{\int \psi * \psi \partial r} \tag{3.6}$$

Для подсчёта вероятностей используют квадрат модуля волновой функции.

Атомные единицы

В системе Си размер и заряд электрона очень малы, поэтому в квантовой химии используются атомные единицы:

- 1 Au, mass $9{,}1093826(16) * 10^{-31} г$ атомная единица массы – масса электрона
- 1 Au, charge $1{,}60217653(14) * 10^{-19} Кл$ атомная единица заряда – заряд электрона
- 1 Hartree, energy $4{,}35974417(75) * 10^{-18} Дж$ атомная единица энергии – электрическая потенциальная энергия атома водорода в основном состоянии
- 1 Bohr, length $5{,}291772108(18) * 10^{-11} м$ атомная единица длины – расстояние от центра ядра до орбитали 1s

Одно-электронный атом

$H = \frac{-\hbar^2}{2m}\nabla^2 - \frac{Ze^2}{4\pi\epsilon_0 r}$ или в упрощённых единицах $H = \frac{1}{2}\nabla^2 - \frac{Z}{r}$

Так как система имеет сферическую симметрию, можно представить волновую функцию в сферических координатах.

$$\left(-\frac{\hbar}{2}\nabla^2 - \frac{Ze^2}{4\pi\epsilon_0 r}\right)\psi(r,\theta,\phi) = E\psi(r,\theta,\phi) \tag{3.7}$$

раскроем оператор Лапласса:

$$\frac{\hbar^2}{2}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}\right] - \frac{Ze^2}{4\pi\epsilon_0 r}\psi = E\psi \tag{3.8}$$

разделив переменные:

$$\psi(r,\theta,\phi) = R(r)Y(\theta,\phi) \tag{3.9}$$

$$\left[\frac{\hbar^2}{2}\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right) - \frac{Ze^2}{4\pi\epsilon_0 r}\right]R(r) = \lambda R(r) \tag{3.10}$$

$$\frac{\hbar^2}{2}\left[\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}\right]Y(\theta,\phi) = -\lambda Y(\theta,\phi) \tag{3.11}$$

Накладывая стандартные условия: периодичность и нормировку (модуль квадрата интеграла волновой функции должен равняться единице), можно получить решения.

Радиальная функция:

$$R_{n,l}(r) = R_\infty(r)b_0\exp\left(\frac{\mu Ze^2 r}{2\pi\epsilon_0}\hbar^2 n\right) \tag{3.12}$$

Зенитная часть:

$$P_l^m = (1-x^2)^{\frac{m}{2}}\left(a_0\sum_{n=0}^{\infty}\frac{a_{2n}}{a_0}x^{2n} + a_1\sum_{n=1}^{\infty}\frac{a_{2n+1}}{a_1}x^{2n+1}\right) \tag{3.13}$$

где

$$a_{n+2} = \frac{(n+m)(n+m+1)-A}{(n+1)(n+2)}a_n$$

В зенитной части применяются полиномы Лагерра, содержащие атомные числа (n – основное квантовое число, l – орбитальное квантовое число, m – магнитное квантовое число, s – спиновое квантовое число).

Азимутальная часть:

$$\Phi_m(\phi) = c_l e^{im\phi} \tag{3.14}$$

Показывает косинус угла, как ориентирована та или иная магнитная орбиталь.

Получаем общее уравнение:

$$\psi_{n,l,m}(r,\vartheta,\varphi) = \sqrt{\left(\frac{2}{na_0}\right)^3\frac{(n-l-1)!}{2n(n+l)!}}e^{-\rho/2}\rho^l L_{n-l-1}^{2l+1}(\rho)Y_l^m(\vartheta,\varphi); \tag{3.15}$$

$L_{n-l-1}^{2\epsilon+1}(\rho)$– Обобщённый полином Лагерра степени n-l-1; $\rho = \frac{2r}{na_0}$

$Y_l^m(\vartheta,\varphi)$– Сферическая гармоника (не зависит от положения электрона и его расстояния от ядра, отражает факт того, как эта орбиталь или её вид будут колебаться);

где n, l, m – основные квантовые числа.

Пример того, как уравнение Шредингера приводит к конкретным описаниям волновых функций (Рис. 3.1, 3.2, 3.3), их можно интегрировать и считать энергии.

n	l	m	$\Psi, \rho = \frac{Z}{a_0} r$
1	0	0	$\Psi_{1s} = \frac{1}{\sqrt{\pi}} \left(\frac{Z}{a_0}\right)^{3/2} \exp(-\rho)$
2	0	0	$\Psi_{2s} = \frac{1}{4\sqrt{2\pi}} \left(\frac{Z}{a_0}\right)^{3/2} (2-\rho)\exp(-\rho/2)$
2	1	0	$\Psi_{2p_0} = \frac{1}{4\sqrt{2\pi}} \left(\frac{Z}{a_0}\right)^{3/2} \rho \exp(-\rho/2)\cos\theta$
2	1	1	$\Psi_{2p_1} = \frac{1}{4\sqrt{2\pi}} \left(\frac{Z}{a_0}\right)^{3/2} \rho \exp(-\rho/2)\sin\theta \exp(i\varphi)$
2	1	−1	$\Psi_{2p_{-1}} = \frac{1}{4\sqrt{2\pi}} \left(\frac{Z}{a_0}\right)^{3/2} \rho \exp(-\rho/2)\sin\theta \exp(-i\varphi)$

Рисунок 3.1 - Волновые функции одно-электронного атома.

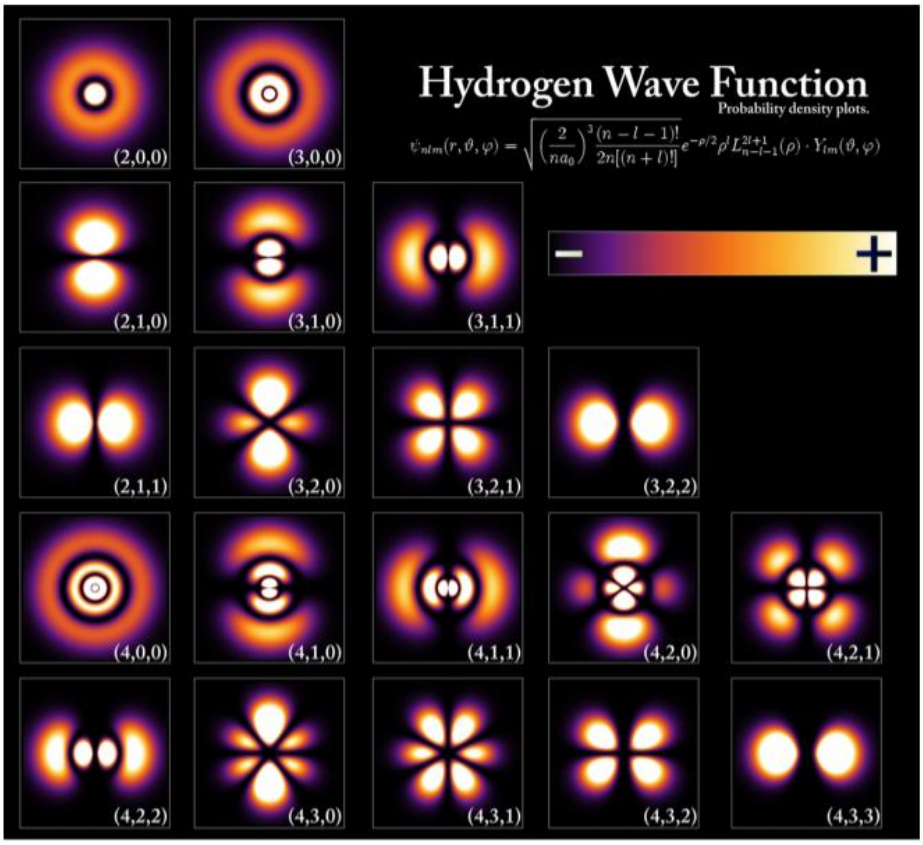


Рисунок 3.2 - Визуализация волновых функций для одно-электронного атома.

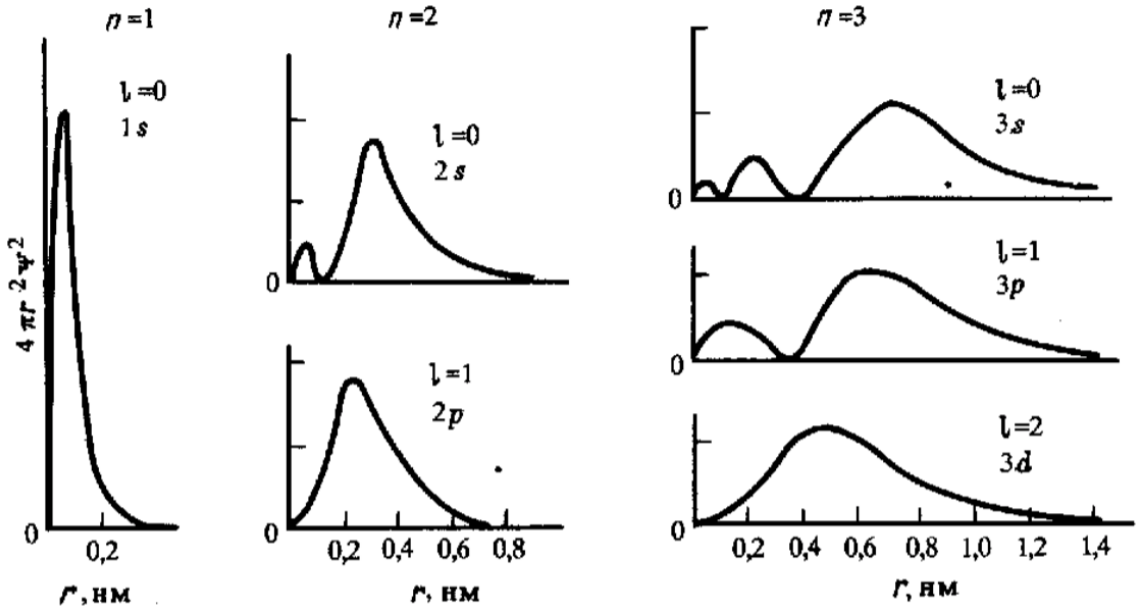


Рисунок 3.3 — Зависимость плотностей вероятности обнаружения электрона от координаты для одно-электронного атома.

Многоэлектронный атом

Полное решение уравнение Шредингера для многоэлектронного атома затруднено по ряду причин:

- N-body problem, задача – предсказать движение трёх и более тел на всём течении времени, если известны положение и скорости на текущий момент (нет в данный момент математического аппарата для решения такой задачи).
- Добавление четвёртого экспериментального квантового числа, спина, создаёт необходимость различать электроны.
- Квадрат волновой функции равен плотности. Трактование волновой функции как плотности электрона в данном месте означает, что плотность может быть образована любым электроном. Этот факт сильно затрудняет расчёты.

Приближение Борна-Оппергеймера

Поскольку ядра двигаются сильно медленнее, чем электроны (разница в массе – 3 порядка), можно считать:

$$\psi_{total} = \psi_{electronic}\psi_{nucleic} \tag{3.16}$$

$$E_{total} = E_{electronic} + E_{nucleic} \tag{3.17}$$

Атом гелия

Система, в которой одно ядро (считаем его практически неподвижным) и два электрона. Тогда общеволновая функция электронов будет линейной комбинацией их спиновых орбиталей.

Уравнение Шредингера для такой системы:

$$(H_1 + H_2)\psi(r_1, r_2) = E\psi(r_1, r_2) \tag{3.18}$$

Спин-орбиталь это:

$$\chi_i(x_i) = \chi_i(r_i)\sigma(s_i) \tag{3.19}$$

Волновая функция электрона на самой низкой орбитали:

$$\chi_1(x_1)\chi_2(x_2); \chi_1(x_2)\chi_2(x1)$$

тогда:

$$\psi = \frac{1}{\sqrt{2}}[\chi_1(x_1)\chi_2(x_2) - \chi_1(x_2)\chi_2(x_1)] \tag{3.20}$$

но

$$\psi = \frac{1}{\sqrt{2}}[\chi_1(x_1)\chi_2(x_2) - \chi_1(x_2)\chi_2(x_1)] \tag{3.21}$$

Это *принцип Паули* (два электрона не могут иметь абсолютно все одинаковые квантовые числа). Можем записать орбитали и спиновые орбитали отдельно, тогда для 1s получится матрица:

$$\psi = \frac{1}{\sqrt{2}}\begin{vmatrix}\chi_1(r_1)\sigma_1(\alpha) & \chi_1(r_2)\sigma_2(\alpha)\\ \chi_1(r_1)\sigma_1(\beta) & \chi_1(r_2)\sigma_2(\beta)\end{vmatrix} = \frac{1}{\sqrt{2}}\chi_{1s}(r_1)\chi_{1s}(r_2)\begin{vmatrix}\sigma_1(\alpha) & \sigma_2(\alpha)\\ \sigma_1(\beta) & \sigma_2(\beta)\end{vmatrix}, \tag{3.22}$$

где $\chi(r)$– орбитали, $\sigma(\alpha/\beta)$– спиновые орбитали.

При переходе к N электронам получаем *определитель Слейтера* – орбитальное приближение волновой функции:

$$\psi = \frac{1}{\sqrt{N}}\begin{pmatrix}\chi_1(1) & \chi_1(2) & \cdots & \chi_1(N)\\ \chi_2(1) & \chi_2(2) & \cdots & \chi_2(N)\\ \vdots & \vdots & \ddots & \vdots\\ \chi_N(1) & \chi_N(2) & \cdots & \chi_N(N)\end{pmatrix} \tag{3.23}$$

Расчёт энергии для молекулы водорода

В данной системе существуют два ядра и два электрона в основном состоянии.

Воспользуемся теорией молекулярных орбиталей, то есть будем считать, что каждая молекулярная орбиталь есть линейная комбинация атомных орбиталей:

$$\psi_i = \sum_{\mu=1} K\, c_{\%mui}\phi_\mu \tag{3.24}$$

Линейная комбинация двух 1s орбиталей: $1\sigma_g = A(1s_A + 1s_B)$, где A – коэффициент нормализации.

$$\psi = \begin{vmatrix} \chi_1(1) & \chi_2(1) \\ \chi_2(1) & \chi_2(2) \end{vmatrix} \tag{3.25}$$

где $\chi_1(1) = 1\sigma_g(1)\alpha(1)$и т. д.

Подставляем полученную матрицу в гамильтониан:

$$H = \frac{-1}{2}\nabla_1^2 - \frac{1}{2}\nabla_2^2 - \frac{Z_A}{r_{1A}} - \frac{Z_B}{r_{1B}} - \frac{Z_A}{r_{2B}} - \frac{Z_B}{r_{2A}} + \frac{1}{r_{12}} \tag{3.26}$$

$$E = \int \psi \int \int \psi \int \tag{3.27}$$

$$E = \frac{1}{2}\int\int \partial\,\tau_1\tau_2[\chi_1(1)\chi_2(2) - \chi_2(1)\chi_1(2)] \tag{3.28}$$

$$\left[-1/2\nabla_1^2 - 1/2\nabla_2^2 - \frac{1}{r_{1A}} - \frac{1}{r_{1B}} - \frac{1}{r_{2A}} - \frac{1}{r_{2B}} + \frac{1}{r_{12}}\right][\chi_1(1)\chi_2(2) - \chi_2(1)\chi_1(2)] \tag{3.29}$$

Разделим на компоненты.

Энергия взаимодействия с ядрами:

$$E_{ij}^{core} = \int \partial\,\tau_1\chi_i(1)\left(\frac{-1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}}\right)\chi_j(1) \tag{3.30}$$

Кинетическая и потенциальная энергия для мультиэлектронной системы:

$$E_{total}^{core} = \sum_{i=1} N\int \partial\,\tau_1\chi_i(1)\left(\frac{-1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}}\right)\chi_j(1) = \sum_{i=1}^{N} H_{ij}^{core} \tag{3.31}$$

Электростатическое отталкивание электронов:

$$E_{ij}^{Coulomb} = \int\int \partial\,\tau_1\partial\tau_2\chi_i(1)\chi_j(1)\frac{1}{r_{12}}\chi_i(2)\chi_j(2) \tag{3.32}$$

$$E_{total}^{Coulomb} = \sum_{i=1}^{N}\sum_{j=i+1}^{N}\int\int \partial\,\tau_1\partial\tau_2\chi_i(1)\chi_j(1)\frac{1}{r_{12}}\chi_i(2)\chi_j(2) = \sum_{i=1}^{N}\sum_{j=i+1}^{N} J_{ij} \tag{3.33}$$

Особенность квантовой природы электрона

Два электрона с одинаковыми спинами в рамках принятой модели имеют поправку к электростатическому отталкиванию, делая его менее значимым:

$$K_{ij} = \int\int \partial\,\tau_1 \partial\tau_2 \chi_i(1)\chi_j(2)\frac{1}{r_{12}}\chi_i(2)\chi_j(1) \tag{3.34}$$

$$E_i^{exchange} = \sum_{j\neq i}^{N} \int\int \partial\,\tau_1 \partial\tau_2 \chi_i(1)\chi_j(2)\frac{1}{r_{12}}\chi_i(2)\chi_j(1) \tag{3.35}$$

$$E_{total}^{exchange} = \sum_{i}^{N}\sum_{j'=i+1}^{N} \int\int \partial\,\tau_1 \partial\tau_2 \chi_i(1)\chi_j(2)\frac{1}{r_{12}}\chi_i(2)\chi_j(1) = \sum_{i}^{N}\sum_{j'=i+1}^{N} K_{ij} \tag{3.36}$$

Упрощённые записи.

Общая энергия:

$$E = \int_{-\infty}^{\infty} \psi * (x) H \psi(x) \partial x \equiv \langle \psi | H | \psi \rangle \tag{3.37}$$

Кулоновское отталкивание:

$$J_{ij} = \left\langle \chi_i \chi_j \middle| \frac{1}{r_{12}} \right\rangle \chi_i \chi_j \tag{3.38}$$

Обменные интегралы:

$$K_{ij} = \left\langle \chi_i \chi_j \middle| \frac{1}{r_{12}} \right\rangle \chi_j \chi_i \tag{3.39}$$

Таким образом, общая энергия для невозбуждённых состояний:

$$\sum_{i=1}^{N/2} 2\, H_{ii}^{core}$$

Поскольку существует четыре способа взаимодействия электронов с одной орбитали с электронами с другой орбитали и всего два способа получить спаренные электроны:

$$E = 2\sum_{i=1}^{N/2} 2\, H_{ii}^{core} + \sum_{i=1}^{N/2}\sum_{j=i+1}^{N/2} \left(4J_{ij} - 2K_{ij}\right) + \sum_{i=1}^{N/2} J_{ij} \tag{3.40}$$

или при $J_{ij} = K_{ij}$:

$$E = 2\sum_{i=1}^{N/2} 2\, H_{ii}^{core} + \sum_{i=1}^{N/2}\sum_{j=i+1}^{N/2} \left(J_{ij} - K_{ij}\right) \tag{3.41}$$

Лекция 4. Квантовая химия (продолжение)

Квантовая химия, продолжение

Фермионы (протоны, электроны) – вещества, спин которых равняется ½.

Бозоны – вещества, спин которых равняется 1. Это значит, что при вращении на 360º рассматриваемый симметричный объект обратится сам в себя. Таким образом, объекты, обладающие половинным спином, обращаются сами в себя при повороте на 720º.

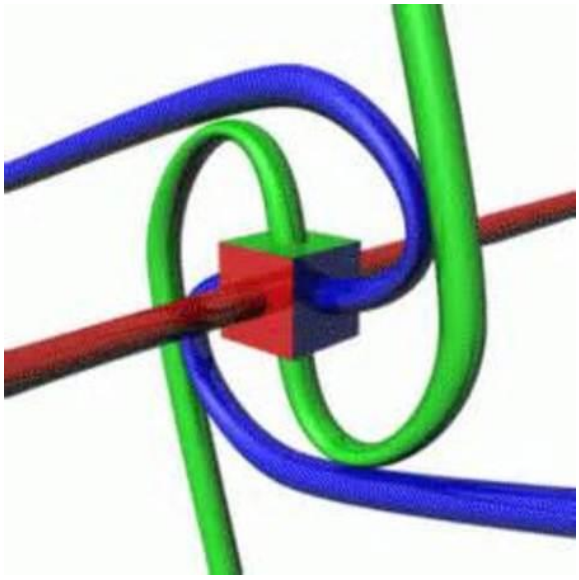


Рисунок 4.1 - Пример объекта, спин которого равен ½.

Из этого свойства спина вытекает обменная составляющая (K) многоэлектронного уравнения, учитывающая, что два электрона с одинаковыми спинами на разных орбиталях отталкиваются хуже, чем два электрона с разными спинами.

Метод самосогласованного поля, SCF

Метод самосогласованного поля – упрощение для работы с многоэлектронными системами, заключающееся в том, что межэлектронное отталкивание вычисляется как влияние общего (среднего) поля на данный электрон, и это зависит только от положения данного электрона.

Такое приближение позволяет повторить разделение переменных в сферических координатах.

Составляющие гамильтониана рассчитываются следующим образом:

$$H_i = \frac{-\hbar^2}{2m}\nabla^2 - \frac{Ze^2}{4\pi\epsilon} + \sum_{j\neq i}^{N}\left\langle\left(\frac{e^2}{4\pi\epsilon_0 r_{ij}}\right)\right\rangle_j \quad (4.1)$$

Эти уравнения называются одноэлектронными.

Суть решения состоит в итеративном изменении параметров в функциях до тех пор, пока изменение энергии не станет незначительным.

Метод Хартри-Фока

Метод Хартри-Фока предполагает, что, аппроксимируя волновые функции в молекуле с помощью орбитальных функций, полученных из водорода, можно получить волновую функцию для молекулы, и из неё же получить энергию и другие свойства молекулы.

Уравнение Фока (оператор Фока заменяет точный гамильтониан):

$$F_i = -\frac{1}{2}\nabla_i^2 - \frac{Z}{r_i} + \sum_j^N \left[\int \chi_j(x_j)\frac{1}{r_{ij}}\chi_j(x_j)\partial x_j - \int \chi_j(x_j)\frac{1}{r_{ij}}\chi_i(x_j)\partial x_j\right] = \epsilon_i\chi_i(x_i) \quad (4.2)$$

энергия электрона на орбитали χ_i:

$$\epsilon_i = H_i + \sum_{j\neq i}^{N}[J_{ij} - K_{ij}], \quad (4.3)$$

где H_i – составляющая гамильтониана, отвечающая за притяжение к ядру, J_{ij} – кулоновское отталкивание, K_{ij} – обменная составляющая.

Что касается молекул, решать напрямую уравнения Хартри-Фока по отношению к ним тяжело. Одна из успешных стратегий – введение базисных функций, то есть волновая функция представляется как комбинация одноэлектронных базисных функций и некоторых коэффициентов:

$$\psi_i = \sum_{v=1}^{K} c_{vi}\,\psi_v; \partial\frac{E}{\partial c_{vi}} = 0$$

Подход Рутхана-Хола

Подход Рутхана-Хола позволяет находить коэффициенты для базисных функций при составлении волновой функции молекулы методом Хартри-Фока и заключается в перемножении матриц:

$$FC = SCE, \quad (4.4)$$

где F – оператор Фока (фокиан), C – матрица коэффициентов, S – интеграл перекрывания (степень перекрывания орбиталей), E – собственное значение энергий.

$$\mathrm{C} = \begin{vmatrix} c_{1,1} & c_{1,2} & \cdots & c_1,k \\ c_{2,1} & c_{2,2} & \cdots & c_2,k \\ \vdots & \vdots & \ddots & \vdots \\ c_{k,1} & c_{k,2} & \cdots & c_k,k \end{vmatrix} \quad S_{ij} = \langle b_i | b_j \rangle = \int \chi_i\chi_j\partial r \quad E = \begin{vmatrix} e_1 & 0 & \cdots & 0 \\ 0 & e_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_i \end{vmatrix}$$

Общий алгоритм расчёта

- Посчитать интегралы для заполнения матрицы F
- Посчитать матрицу перекрывания S
- Диагонализируем матрицу S
- Строим $S^{-1/2}$
- Угадываем или рассчитываем матрицу плотности P
- Строим матрицу F, заполняя значениями интегралов и P
- Строим $F` = S^{-1/2}FS^{-1/2}$
- Решаем $|F`-EI|=0$ для поиска собственных значений E и C` диагонализации F`
- Рассчитываем орбитальные коэффициенты $C=S^{-1/2}C`$
- Считаем новую матрицу плотности из матрицы C
- Если плотность изменилась незначительно, заканчиваем или продолжаем заполнять матрицу F

Таким образом, электронную структуру молекулы можно посчитать, зная только основные физические константы. Такие подходы называются *ab initio*.

Базисные наборы

Для упрощения вычислений орбитальные функции Слейтора аппроксимируют гауссиановскими функциями. В общем виде это:

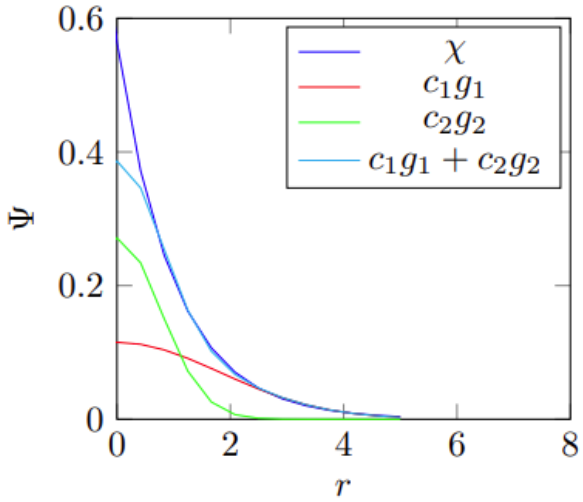
$$x^a y^b z^c e^{-\alpha r^2}$$

Важное свойство: сумму двух функций можно представить одним гауссианом, то есть

$$\phi_\mu = \sum_{i=1}^{L} d_{i\mu}\, \phi_i(\alpha_{i\mu}) \tag{4.5}$$

При выборе базисных функций надо выполнить следующие условия: они должны иметь физический смысл, и их интегралы должны быть сходящимися.

Пример для 1s:



$$\psi = \sum_{1}^{2} ce^{-\alpha r^2}$$

$$\psi = c_1 g_1 + c_2 g_2$$

	α	c
1	0.151623	0.678914
2	0.851819	0.430129

Рисунок 4.2 - Сравнение слейтеровской функции и комбинации из двух гауссианов для 1s орбитали.

Чем больше гауссианов описывают основные орбитали атомов, тем ближе это описание к орбитали Слейтера.

Базисные функции с коэффициентами называются *базисными наборами*. Есть два типа таких наборов:

- contracted – при оптимизации варьируются только множители перед гауссианами
- uncontracted – при оптимизации варьируются и множители перед гауссианами, и коэффициенты в экспоненте

Гауссианы выглядят следующим образом:

$$1s = Ne^{\alpha r^2}; \quad 2p_x = Ne^{\alpha r^2}x; \quad 2p_y = Ne^{\alpha r^2}y; \quad 2p_z = Ne^{\alpha r^2}z;$$

$$3d_{xx} = Ne^{\alpha r^2}x^2; \quad 3d_{xy} = Ne^{\alpha r^2}xy; \quad 3d_{xz} = Ne^{\alpha r^2}xz;$$

$$3d_{yy} = Ne^{\alpha r^2}y^2; \quad 3d_{yz} = Ne^{\alpha r^2}yz; \quad 3d_{zz} = Ne^{\alpha r^2}z^2;$$

$$4f_{xxx} = Ne^{\alpha r^2}x^3; \quad 4f_{xxy} = Ne^{\alpha r^2}x^2y; \quad 4f_{xxz} = Ne^{\alpha r^2}x^2z;$$

и так далее.

Сокращения базисных наборов (basis set)

Общепринятая система сокращений, называемая «сокращения базисных наборов», используется для предоставления информации об использованных базисах, например, в серии программ Gaussian.

- *Минимальные базисные наборы: Sto-nG*, где n – количество гауссианов в базисном наборе для каждой орбитали. Рассматриваются только орбитали атомов. Для

элементов, не содержащих d орбиталей, STO-3G является абсолютным минимумом. Плохо работает для несферических орбиталей и элементов в конце периода.

- *Double zetta basis:* zetta – коэффициент в экспоненте гауссиана. Таким образом, double zetta bsis – это линейная комбинация contracted и uncontracted гауссианов. Даёт более точный результат, чем STO, коэффициенты считаются в ходе итераций, поэтому можно работать с анизотропией px, py, pz.
- *Split valence double zetta:* валентные оболочки описываются с использованием uncontracted гауссианов и большим количеством функций, чем остальные электроны.

Описания базисных наборов для программы GAUSSIAN имеют следующий вид обозначений: *M-ijk...G*, где

- M – количество ограниченных гауссианов на один невалентный электрон
- наличие двух и более букв после “-” означает, что валентные электроны описываются 2 и более функциями, каждая из которых состоит из линейной комбинации i,j,k гауссианов
- * означает, что для тяжёлых атомов используются не только гауссианы, характерные для конкретной орбитали, но и гауссианы следующей орбитали
- ** – то же самое, что и * ,но добавляются 3 гауссиана для p-орбиталей к гауссианам H и He
- + означает добавление дополнительных гауссианов тех же орбиталей, но с маленьким значением α. Этот шаг нужен для точного счёта систем, где
- значительная электронная плотность удалена от ядра: электронные пары, анионы

Например, для углерода в 3-21*G: у валентных электронов с 3 гауссианами прибавляется 6 гауссианов для d-орбиталей.

Таблица 4.1 - Базисные наборы.

Запись Базиса	Количество гауссианов на орбиталь, для **C**	Применение
STO-3G	3	Большие системы
6-31G	6-не валентные 3+1-валентные	Системы без поляризации
6-31*G	То же самое + 6 функций типа l+1	Системы с анизотропией заряда
6-31**G	см. выше + 3 p-функции для H и He	Водородная связь
6-31+G(2df)	Поляризационные: 2*6 d-функций + 7 f- функций (см *) Диффузных : 4	Нужно там, где важно точно рассчитать высокую плотность электронов
6-311++ G(3df,3pd)	Диффузные на все атомы и поляризационные на C: 3*6 d типа + 7 f типа H: 3*3 p типа + 1 d типа	Если всё, что было до этого было не достаточно точным.

Семи-эмпирические методы

Семи-эмпирические методы подразумевают, что помимо расчётов *ab initio,* используются упрощения, основанные на эмпирических исследованиях.

Основную сложность представляют расчёты взаимоэлектронного отталкивания и их обменная составляющая, так как это двойные суммы, число двухэлектронных интегралов пропорционально M*4, где M – количество или размерность атомного базиса. В связи с этим сложно считать большие электронные системы.

Основная идея заключается в том, чтобы за счёт коэффициентов и аппроксимаций избежать расчёта взаимного отталкивания электронов для большинства из них.

Таким образом, основные приближения класса семи-эмпирических методов:

- рассматриваются только валентные электроны
- в молекулярных орбиталях учитываются AO с n, соответствующим высшим заселённым орбиталям
- для двухэлектронных интегралов вводят приближение нулевого дифференциального перекрывания

$$\chi_\mu(r)\chi_v(r)dr = 0, \mu \neq v$$

- двухэлектронные интегралы зависят только от природы атомов, на которых центрированы орбитали χ_μ и χ_v, и не зависят от конкретного вида орбиталей. Для обозначения среднего значения интегралов используют γ_{AB}

Основная идея нулевого дифференциального перекрывания (ZDO) заключается в уменьшении количества интегралов за счёт обнуления перекрывания некоторых орбиталей.

Например, если два ядра находятся далеко друг от друга, вероятно, центрированные к ядрам функции не перекрываются.

В таком случае, матрица обмена S=1, и уравнение Рутхана-Хола будет иметь следующий вид:

$$FC = CE \tag{4.6}$$

Однако, это приближение оказалось слишком сильным для атомов, объединённых в молекулу, и появилось несколько модификаций:

- *CNDO:* идея применяется ко всем парам функций, ненулевыми остаются только кулоновские интегралы
- *INDO, MINDO/x:* учитывается перекрывание всех функций, центрированных на одном ядре
- *NDDO, MNDO:* пренебрегается двухатомное перекрывание
- *AM1, PM3:* улучшенные и современные варианты параметризации MNDO, могут использоваться для расчёта белков целиком

Программы, в которых реализованы такие методы: MOPAC, OPENMOPAC, AMPAC.

У этих методов различные источники параметризации:

- CNDO/2 – электронная плотность, электроны спарены
- CNDO/S – спектры, электроны спарены
- INDO – электронная плотность, электроны не спарены
- INDO/S – спектры, электроны не спарены
- ZNDO и ZNDO/S – то же самое, только для переходных элементов
- MINDO/3 – теплоты образования

Таблица 4.2 - Сравнение семи-эмпирических методов .

Метод	Параметризуемое свойство	Хорошо воспроизводимые свойства	Плохо воспроизводимые свойства
CNDO/2	Разности энергий между занятыми МО	Дипольные моменты, длины связей, валентные углы, силовые константы	Теплоты образования, потенциал ионизации, сродство к электрону, спектры, реакции
CNDO/S, INDO/S, ZINDO	Электронный спектр	Спектр	Теплоты образования, геометрия молекул, реакции
INDO	Спиновые плотности	Спиновые плотности, константы сверхтонкого	Теплоты образования, потенциалы ионизации, сродство

		взаимодействия, геометрия молекул	к электрону, электронные спектры
MINDO/3	Потенциал атом-атомного взаимодействия	Теплоты образования, потенциалы ионизации, длины связей	Электронные спектры, водородная связь
MINDO	Теплоты образования	Теплоты образования, геометрия молекул	Электронные спектры, водородная связь
AM1	Теплоты образования	Теплоты образования, геометрия молекул	Электронные спектры
PM3	Теплоты образования, параметры межмолекулярного взаимодействия	Теплоты образования, геометрия молекул, водородная связь, межмолекулярные взаимодействия	Электронные спектры

Недостаток подхода Фока

Важным недостатком метода Фока является то, что он работает для систем, близких к минимуму энергии и не работает для возбуждённых систем, систем, где наблюдается переход из одного стабильного состояния в другое стабильное и так далее.

Для этих возбуждённых состояний нужно добавлять новые орбитали и новые функции в подход Фока, и далее их оптимизировать.

$$\psi(1,2 \dots N) = \sum_{k_1} \sum_{k_2} \cdots \sum_{k_n} c\,(k1 \dots K_N)\psi_{k1}(1)\psi_{k2}(2) \dots \psi_{kN}(N) \tag{4.7}$$

$$\psi(1,2 \dots N) = \sum_{K}^{C_K} \psi_K\,(1 \dots N) \tag{4.8}$$

Можно повысить точность счёта, учтя некоторые компоненты ряда. Для этого нужно минимизировать:

$$E = E_{real} - E_{HF} \tag{4.9}$$

Существуют три основных способа минимизировать энергию возбуждённых состояний в методе Харти-Фока:

- метод конфигурационного взаимодействия

- метод самосогласованного поля
- теория возмущения Меллера-Плессе

Метод конфигурационного взаимодействия

Метод конфигурационного взаимодействия (Configuration Interaction или CI) заключается в том, что электроны привязываются к более внешним орбиталям, и далее происходит оптимизация энергии с учётом этой привязки.

Например, мы нашли M спиновых орбиталей, и N из них заняты электронами. Именно их мы и используем для расчётов и подменим некоторые орбитали в найденной матрице новыми. То есть создадим возбуждённую молекулу.

$$\psi_{CI} = \sum_{K=0}^{\infty} C_K \psi_K \tag{4.10}$$

Далее применяем подход Рутана-Хола и находим коэффициенты C. Необходимо заметить, что для каждого возбуждённого состояния нужен свой детерминант Слейтора и нужно заново считать орбитали.

Метод самосогласованного поля

Иногда корреляционный эффект вносит большой вклад во взаимодействия электронов, например, в ароматических системах. В методе Хартри-Фока он не учитывается, поскольку используется приближение, в котором электрон движется в поле других электронов.

Для того, чтобы преодолеть ошибку (~1% общей энергии, но это химически значимо), возникающую из-за того, что метод Харти-Фока игнорирует мгновенное кулоновское отталкивание, используют разные подходы, в том числе и метод самосогласованного поля.

В данном методе предлагается варьировать не только коэффициенты C, но и форму орбиталей:

$$\psi = \sum_{K=0}^{\infty} C_K \psi_K(\delta\psi_1 \delta\psi_2 \dots \delta\psi_K) \tag{4.11}$$

Это приводит к тому, что матрицы вычислений становятся более громоздкими, нужно вычислять комбинации, поэтому рассматривают только однократные и двухкратные состояния возбуждения. Можно явно указать, на какой уровень будет переходить возбуждаемый электрон.

Однако в связи с сложностями вычислений этот метод редко применяют к биомолекулам.

Теория функционала плотности, DFT

Это наиболее популярный подход, идея которого состоит в том, что:

- система описывается не волновой функцией, а функцией электронной плотности, включающей вклад всех электронов
- предполагается, что для любой реальной системы с потенциалом и плотностью существует такая «невзаимодействующая» система (то есть система, в которой отсутствует межэлектронное взаимодействие) с некоторым одноэлектронным потенциалом, электронная плотность которой совпадает с точной электронной плотностью реальной системы
- точное решение многоэлектронного уравнения Шрёдингера представляется слэйтеровским детерминантом, состоящим из одноэлектронных орбиталей

$$\rho(r) = \sum_{i}^{N} |(\chi_i)|^2 \tag{4.12}$$

Эта идея доказывается с помощью теоремы *Хоэнберга-Кона*:

Если в системе постоянное количество электронов, их взаимное взаимодействие не зависит от внешнего потенциала.

Тогда можно разбить энергию электронов:

$$E = E^T + E^V + E^J E^{XC}, \tag{4.13}$$

где T – кинетическая составляющая, V – потенциальная энергия (ядра), J – отталкивание электронов, XC – обменно-корреляционная составляющая.

Главная проблема и основная задача этого подхода – оптимизировать обменно-корреляционную составляющую.

Хоненберг и Кохн показали, что E^{XC}зависит только от электронной плотности.

$$E^{XC} = E^X + E^C$$

Обе части зависят от плотности, и их часто представляют как:

- локальные (зависят только от плотности)

$$E^X_{LDA} = 3/2\left(\frac{3}{4\pi}\right)^{1/3} \int \rho^{4/3}\, \partial^3 r \tag{4.14}$$

- градиент-корректированные (зависят от плотности и её градиента)

Уравнения, связывающие электронную плотность и обменно-корреляционную составляющую называют функционалами плотности, их придумано довольно много. Как пример, широко используемый функционал Bekke 1988 года:

$$E^X_{Bekke88} = E^X_{LDA} - \gamma \frac{\int \rho^4/3x^2}{1 + 6\gamma \sin h^{-1} x} \partial^3 r;\ x = \frac{|\nabla^2 \rho|}{\rho^{4/3}} \tag{4.15}$$

Существуют корреляционные функционалы: Pedrew и Wang (1991), Vosko, Wilk и Nusair (1980). Часто применяются гибридные функционалы:

$$E_x^{HF} = \frac{1}{2}\sum_{i,j}\int\int \psi_i(r_1)\psi_j(r_1)\frac{1}{r_{12}}\psi_i(r_2)\psi_j(r_2)dr_1dr_2 \quad (4.16)$$

$$E_{xc}^{B3LYP} = E_{xc}^{LDA} + a_0(E_x^{HF} - E_x^{LDA}) + a_x(E_x^{GGA} - E_x^{LDA}) + a_c(E_c^{GGA} - E_c^{LDA}) \quad (4.17)$$

где:

$$E_{XC}^{GGA}[n_\uparrow, n_\downarrow] = \int \epsilon_{XC}(n_\uparrow, n_\downarrow, n_\uparrow, n_\downarrow \nabla() \nabla n(\vec{r})d^3r) \quad (4.18)$$

Современные функционалы дают хорошую производительность при хорошей точности результатов. Одними из самых распространённых видов обменных функционалов в расчётах квантовой химии являются BLYP и B3LYP.

Однако основной проблемой в DFT является описания Ван-дер-Ваальсовых взаимодействий или дисперсионного взаимодействия, то есть стэкинг и $\pi-\pi$ взаимодействия. В принципе, их можно компенсировать аналитическими потенциалами (простые функции, зависящие от расстояния между ядрами).

Текущее состояние метода теории функционала плотности таково, что невозможно оценить погрешность расчёта, не сравнивая его результаты с другими подходами или с результатами экспериментов.

Лекция 5. Молекулярная механика биополимеров

Молекулярная механика

Используя методы квантовой химии, можно установить электронную структуру молекул, однако эти расчёты весьма трудоёмки. Так, современные базисы предполагают примерно 60 функций на атом, то есть примерно 900 функций на аминокислоту. Понятно, что такой метод не подходит для описания биополимеров.

Электронную плотность можно аппроксимировать уравнениями классической физики, и на этом основывается молекулярная механика: электронная структура атома заменяется на достаточно простые уравнения с параметрами. Наборы параметров называются *силовыми полями*. Важный принцип молекулярной механики заключается в том, что она основана на приближении Борна-Оппенгеймера (электроны быстро адаптируются к движению ядер, то есть в процессе их движения электронная плотность не изменяется). Энергии рассчитываются на основе положения ядер.

Такие приближения позволяют работать с большими системами (систем из миллионов атомов рутинны для расчётов), при том, что в некоторых случаях подходы молекулярной механики могут давать результаты, сравнимые по точности с методами квантовой механики.

Силовое поле

Силовое поле – набор уравнений и параметров к ним, описывающий взаимодействие частиц в системе. Пример простейшего уравнения силового поля:

$$U = \sum_{bonds} \frac{k_i}{2}(l_i - l_0)^2 + \sum_{angles} \frac{k_i}{2}(\phi - \phi_0)^2 + \sum_{torsions} \frac{V_n}{2}\big(1 + \cos(n\omega - \gamma)\big) + \\ + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right) \quad (5.1)$$

Здесь можно выделить ковалентные взаимодействия (связи, углы, торсионные углы) и нековалентные взаимодействия (кулоновские и Ван-дер-Ваальсовы взаимодействия). На Рис. 5.1 схематически показаны описываемые взаимодействия.

Основные особенности силовых полей:

- Большинство параметров неотделимо от поля
- Параметризация сильно зависит от целей исследования
- Большинство силовых полей параметризованы для воспроизведения структуры
- Силовые поля – результат оптимизации параметров
- Силовые поля – подход, основанный на эмпирических данных

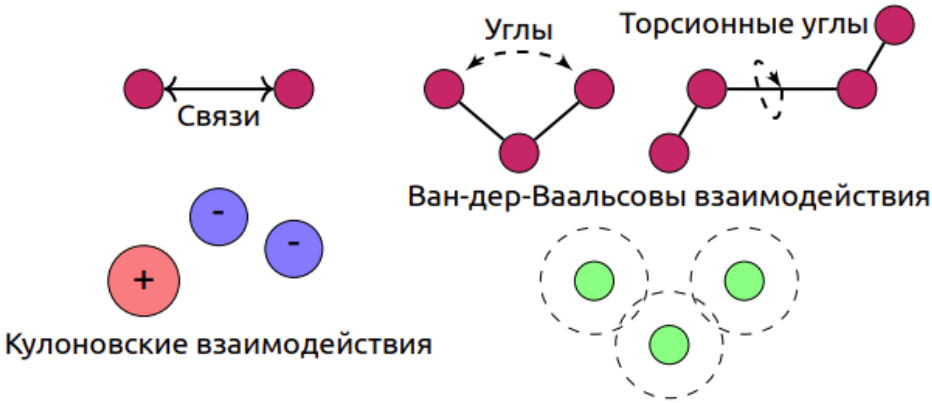


Рисунок 5.1 - Схематически показаны взаимодействия, описываемые силовыми полями.

Типы атомов в силовых полях

В отличие от квантовой химии, где тип атома определяется только массой и количеством электронов, в молекулярной механике играет роль химическое окружение рассматриваемого атома (например, азот, имеющий двойную связь, химически отличается от азота, не имеющего двойной связи). На Рис. 5.2 изображена модель гистидина, в котором все три атома азота являются разными с точки зрения молекулярной механики.

Однако типов атомов углерода, водорода, кислорода в белках и близких к ним соединениях оказывается ограниченное количество: примерно пятьдесят.

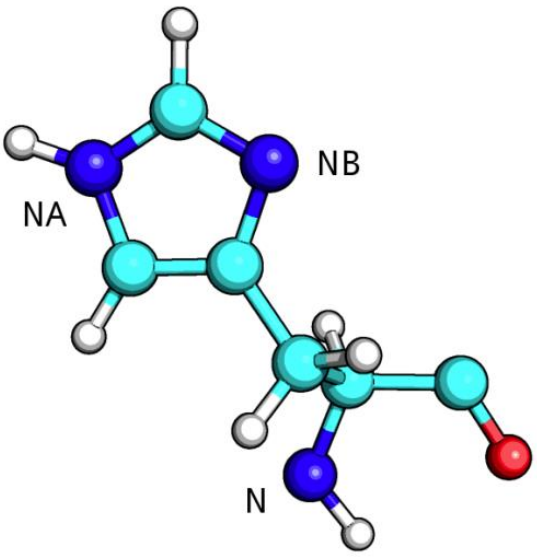


Рисунок 5.2 - Молекула гистидина, подписаны различные типы атомов азота.

Таким образом, составляются таблицы стартовых условий для моделирования, где различным атомам в молекулах сопоставляются соответствующие им типы.

Описание связей

Наиболее хорошо связи описываются потенциалом Морзе (Рис. 15):

$$U(l) = D_e\left\{1 - e^{[-a(l-l_0)]}\right\}^2 \tag{5.2}$$

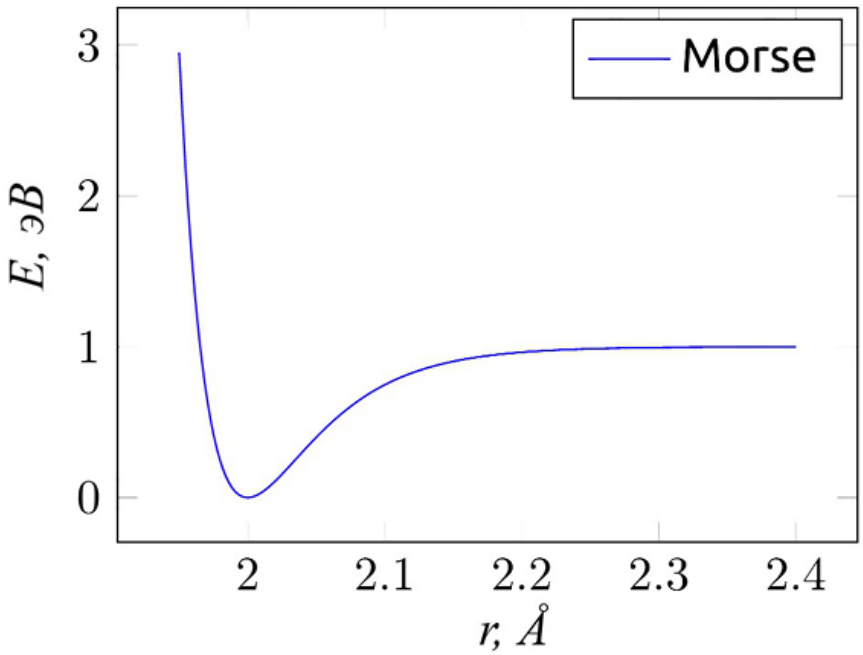


Рисунок 5.3 - Потенциал Морзе (зависимость потенциальной энергии от расстояния).

На расстоянии больше двух ангстрем (на Рис. 5.3) описывается процесс разрыва связи. В случае белков это совершенно не востребовано, поскольку моделирование белков с разрывом С-С связей моделируется очень редко. Иногда потенциал Морзе задаётся для отдельных частей, например, при исследовании энзиматических реакций, чтобы определить какое усилие необходимо для разрыва той или иной связи. Для подавляющего большинства систем для описания связей используется просто закон Гука:

$$U(l) = \frac{k_i}{2}(l_i - l_0)^2 \tag{5.3}$$

Параметры (константа жёсткости) при описании связей позволяют получить энергию при умножении на квадрат расстояния.

Таблица 5.1 — Примеры параметров связей при описании их законом Гука.

связь	r_0,A	k, kcal mol^{-1}A^{-2}
C_{sp3}-C_{sp3}	1.523	317
C_{sp2}-C_{sp2}	1.337	690
C_{sp2}-O_{sp2}	1.208	777
C_{sp3}-N_{sp3}	1.438	367

Можно воспользоваться рядами Тейлора для более точного описания связи (Рис. 5.4):

$$U = \frac{k_i}{2}(l_i - l_0)^2(1 - k'(l_i - l_0) - k''(l_i - l_0)^2 - k'''(l_i - l_0)^3 - k''''(l_i - l_0)^4 \dots) \quad (5.4)$$

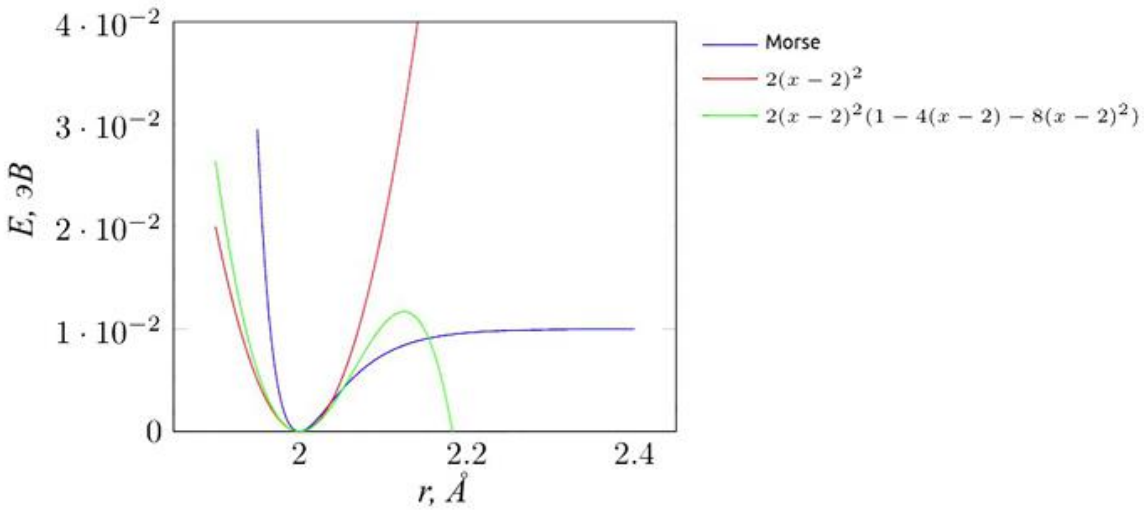


Рисунок 5.4 - Кубический и другие варианты описания связей.

Но в большинстве случаев такой метод не даёт явного увеличения точности, поэтому останавливаются на первом приближении (красная линия на Рис. 16). Оно хорошо аппроксимирует растягивание связи, но плохо аппроксимирует сжатие, поэтому в моделированиях, где важно сжатие связи либо адаптируют параметры, либо используют потенциал Морзе.

Описание углов

Потенциал валентного угла очень хорошо описывается также гармоническим потенциалом, поскольку угол не может порваться:

$$U(\phi) = \frac{k_i}{2}(\phi_i - \phi_0)^2 \tag{5.5}$$

или

$$U(\phi) = \frac{k_i}{2}(\phi_i - \phi_0)^2(1 - k'(\phi_i - \phi_0) - k''(\phi_i - \phi_0)^2 - k'''(\phi_i - \phi_0)^3 - k''''(\phi_i - \phi_0)^4 \dots) \tag{5.6}$$

С торсионными углами всё не так просто, они могут быть очень разными, поскольку вокруг связи может быть много разных заместителей.

Потенциал торсионного угла:

$$U(\omega) = \sum_{torsions} \frac{V_n}{2}\big(1 + \cos(n\omega - \gamma)\big) \tag{5.7}$$

В этой формуле

- 1+cos, чтобы потенциальная энергия угла не была отрицательной
- ω – торсионный угол
- n – число минимумов энергии при вращении вокруг рассматриваемой связи
- γ – смещение по фазе

Например, в случае этана n=3, γ=0.

Аппроксимировать одним косинусом молекулы с большим количеством заместителей, например, сахара, достаточно сложно. Поэтому используются комбинации из таких потенциалов с различными n и γ, что позволяет добиться довольно сложных функций (Рис. 5.5). При помощи квантовой химии узнают значения потенциальной энергии при вращении вокруг торсионного угла, и дальше аппроксимируют полученную функцию суммой косинусов.

В различных силовых полях используются различные потенциалы торсионного угла.

Например, в поле MM2 используется три члена:

$$U(\omega) = \frac{V_1}{2}(1 + \cos\omega) + \frac{V_2}{2}(1 + \cos 2\omega) + \frac{V_3}{2}(1 + \cos 3\omega) \dots$$

Поле OPLS использует ряды с четырьмя слагаемыми:

$$U(\omega) = \frac{1}{2}[F_1(1 + \cos\omega) + F_2(1 - \cos 2\omega) + F_3(1 + \cos 3\omega) + F_4(1 - \cos 4\omega)]$$

Для удержания нескольких атомов в одной плоскости используют «неправильные» торсионные углы. Например, для циклобутанона (Рис. 5.6) атомы 1, 2, 3 и 4 должны находиться в одной плоскости. Для этого потенциал торсионного угла задаётся не по связям, а по комбинациям автомов – 1-2-3-4. То есть здесь создаётся виртуальная связь между атомами 3 и 4, которые на самом деле напрямую не связаны. Запрет вращения

вокруг связи 3-4 приводит к тому, что все эти атомы (1, 2, 3, 4) остаются в одной плоскости.

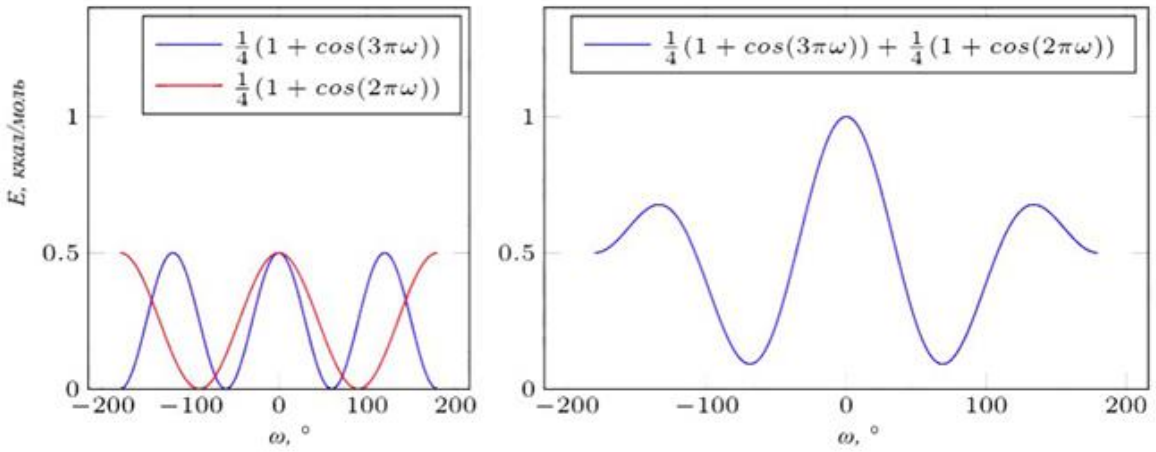


Рисунок 5.5 - Описание потенциала торсионного угла O-C-C-O (сахара в ДНК) при помощи суммы двух косинусов.

Симметричные колебания по связи приводят к тому, что жёсткость угла должна меняться – для этого существуют кросс составляющие в силовых полях (отражают зависимость состояния одной связи или угла от состояния соседней связи). Однако такие формулы почти не используются при описании белков, поскольку они нужны для того, чтобы моделировать спектры в инфракрасном диапазоне, и только там.

$$U(l_1, l_2) = \frac{K_{l1l2}}{2}(l1 - l_{1,0})^2(l2 - k_{2,0})^2$$

$$U(l_1, l_2, \phi) = \frac{K_{l1l2\varphi}}{2}\left[(l1 - l_{1,0})^2 + (l2 - l_{2,0})^2\right](\phi - \phi_0)$$

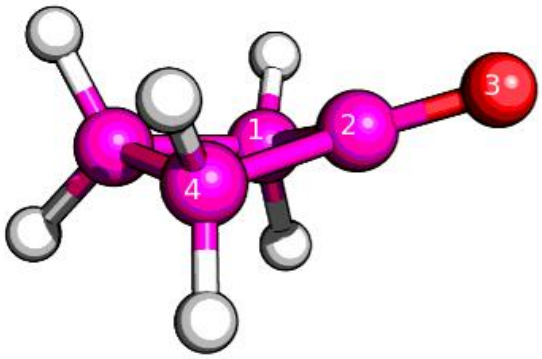


Рисунок 5.6 - Молекула циклобутанона, цифрами отмечены атомы, лежащие в одной плоскости.

Нековалентные взаимодействия

Нековалентные взаимодействия являются определяющими в структуре биополимеров. Так как эти взаимодействия реализуются через пространство, они часто описываются как функции обратно пропорциональные расстоянию между двумя атомами.

Электростатические взаимодействия

Электростатические взаимодействия являются наиболее простыми из нековалентных взаимодействий. Обычно используется допущение, что поверхность единичного потенциала можно представить зарядами в центрах атомов, и электростатические взаимодействия описываются по закону Кулона:

$$U(q_{1,}q_2) = \frac{q_1 q_2}{4\pi\epsilon_0\epsilon_r r_{ij}}; \tag{5.8}$$

$$U = \sum_{i=1}^{N_A}\sum_{j=1}^{N_B}\frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{5.9}$$

где q_1 и q_2 – заряды атомов, r – расстояние между ними, ε_0 – диэлектрическая проницаемость.

Очевидно, что количество вычислений растёт значительно быстрее количества частичных зарядов.

Существуют следующие упрощения:

- Разрастание центрального мультиполя (ММ малых молекул)
- Двойное обрезание
- Потенциал реакционного поля
- Суммирование Эвальда

Упрощение двойного обрезания заключается в следующем. Вокруг каждого атома есть две сферы: сфера ближайших взаимодействий (меньшего радиуса), для которой чётко рассчитываются кулоновские взаимодействия атом-атом, и вторая сфера (большего радиуса), для которой рассчитывается взаимодействия атома с группой атомов (для которых заранее рассчитан некий общий заряд, например, «+1» у аминогруппы). На расстоянии больше, чем радиус большей сферы, считается, что нет взаимодействий, в результате чего возникают неточности при моделировании («краевой эффект»).

$$U_1 = \sum_{i=1}^{N_A}\frac{q_1 q_i}{4\pi\epsilon_0\epsilon_r r_1} + \sum_{j=1}^{N_{group}}\frac{q_1 q_j}{4\pi\epsilon_0\epsilon_r r_{1j}} \tag{5.10}$$

Чтобы справиться с ними, было придумано другое упрощение – потенциал реакционного поля. Его основная идея заключается в том, что за некоторым расстоянием

плотность заряда одинаковая, и, следовательно, известна некая диэлектрическая проницаемость среды.

$$U_{ij} = \frac{q_! q_i}{4\pi\epsilon_0\epsilon_r r_{1i}}\left[1 + \frac{\epsilon_{rf} - \epsilon_r}{2\epsilon_{rf} + \epsilon_r}\frac{r_{ij}^3}{r_c^3}\right] - \frac{q_1 q_i}{4\pi\epsilon_0\epsilon_r r_c}\frac{3\epsilon_{rf}}{2\epsilon_{rf} + \epsilon_r} \tag{5.11}$$

Самый точный способ расчёта электростатических взаимодействий – это суммирование Эвальда. В нём учитываются не только заряды в ближайшем окружении, но и, как в кристалле, заряды, находящиеся в соседних ячейках.

$$U_{ij} = \sum_{x=1}^{N_x}\sum_{y=1}^{N_y}\sum_{z=1}^{N_z}\sum_{i=1}^{N}\sum_{j=1}^{} N\frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}} \tag{5.12}$$

Получилась сумма всех против всех да ещё и по всем направлениям, то есть пять сумм. Считать это очень тяжело, поэтому Эвальд предложил перевести этот ряд в сумму двух быстро сходящихся рядов и константы.

$$U = U_{dir} + U_{rec} + U_0 \tag{5.13}$$

$$U_{dir} = \frac{f}{2}\sum_{i,j}^{N}\sum_{x=1}^{N_x}\sum_{y=1}^{N_y}\sum_{z=1}^{N_z} q\frac{erfc(\beta_{rij,n})}{r_{ij,n}} \tag{5.14}$$

$$U_{rec} = \frac{f}{2}\pi V\sum_{i,j}^{N} q_i\, q_j \sum_{m_x}\sum_{m_y}\sum_{n_z}\frac{\exp(-\pi m/\beta)^2 + 2\pi m\left(r_i r_j\right)^2}{m} \tag{5.15}$$

$$U_0 = \frac{f\beta}{\sqrt{\pi}}\sum_{i}^{N} q_i^2, \tag{5.16}$$

где β – параметр, определяющий соотношение прямого и обратного взаимодействий.

Из-за краевого эффекта самосборка липидного бислоя приводит к формированию липидной сферы в периодических граничных условиях в случае использования двойного обрезания при моделировании электростатических взаимодействий в то время как при использовании суммирования Эвальда самосборка липидного бислоя проходит нормально (Рис. 5.7).

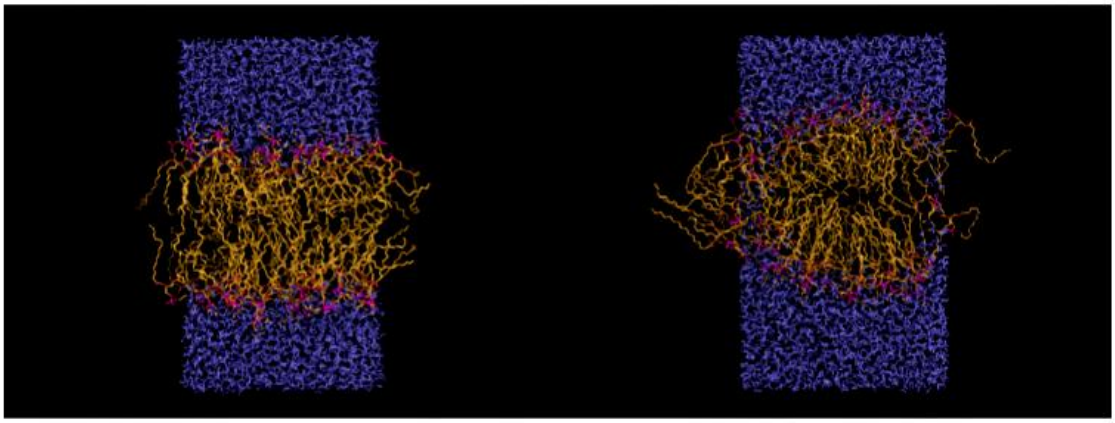


Рисунок 5.7 - Самосборка бислоя фосфолипидов с использованием суммирования Эвальда (слева) и двойного обрезания (справа) для моделировании электростатических взаимодействий.

Ван-дер-Ваальсовы взаимодействия

В основе природы Ван-дер-Ваальсовых взаимодействий лежат дисперсионные и обменные электронные эффекты, из-за чего их очень трудно считать. Такие эффекты возможно рассчитать методами квантовой механики, но это требует очень больших времён расчёта. В молекулярной механике для быстрого приближённого рассчёта этих взаимодействий используется потенциал Леонарда-Джонса:

$$U_{VdW} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} 4\,\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{5.17}$$

Для учёта взаимодействий между разными типами атомов при моделировании биологических систем наиболее часто используют следующие правила смешивания:

$$\sigma_{AB} = \frac{1}{2}(\sigma_{AA} + \sigma_{BB}) \tag{5.18}$$

$$\epsilon_{AB} = \sqrt{\epsilon_{AA}\epsilon_{BB}} \tag{5.19}$$

1-4 взаимодействия могут быть уже учтены в описании торсионного угла, поэтому в некоторых силовых полях такие нековалентные взаимодействия не учитываются.

В полях семейства AMBER, 1-4 VdW взаимодействия всё-таки учитываются, но их потенциал делится на 2.

Водородные связи

Водородные связи – одно из ключевых нековалентных взаимодействий для биополимеров. В силовых полях водородная связь часто описывается как комбинация Ван-дер-Ваальсовых и Кулоновских взаимодействий.

Однако в некоторых силовых полях используется описание водородной связи на основе потенциала Леонарда-Джонса:

$$U_{HB} = \frac{A^{10}}{r} - \frac{C^{12}}{r} \tag{5.20}$$

Для более точного описания водородной связи также вносят поправки, учитывающие её геометрию, но при моделировании биополимеров такая большая точность обычно не востребована.

Эффективный парный потенциал и модели воды

Парные потенциалы обычно не отражают свойства молекул как среды, в связи с чем возникает проблема обсчёта тройных взаимодействий, а это тройные суммы, и считать это сложно. Поэтому для воспроизведения свойств воды как среды используют парные потенциалы с «правильной» параметризацией, например, использование больших частичных зарядов для описания фазы, чем для отдельной молекулы. Так делают с водой (1,85 D для отдельной молекулы и 2,6 D для воды как фазы).

Таким образом, возникает задача моделирования воды. Модель должна воспроизводить как свойства одной молекулы, так и свойства фазы.

Существуют три основных класса моделей:

- простые модели
- поляризуемые модели
- *ab initio* модели

Простые модели не учитывают колебания по связи O-H: SPC (single-point-charge), TIP3P (трёхчастичное отображение воды).

Иногда вводят дополнительную частицу (или частицы), имеющую нулевую массу, но обладающую зарядом, чтобы аннулировать электронную плотность при распределении заряда: TIP4P, TIP5P.

В поляризуемых моделях положение дополнительной частицы (имеющей нулевую массу, но обладающей зарядом) может изменяться. Существует два подхода: смещать центр заряда кислорода относительно центра атома или добавить точки вокруг кислорода, в которых может находиться заряд.

Ab initio модели базируются на квантово-химических вычислениях как одной, так и нескольких молекул воды (например, NCC модель), однако они плохо работают, поскольку не отражают свойства воды как фазы.

Силовые поля с объединёнными атомами

Основная идея обобществления атомов в силовых полях заключается в том, что атомы водорода, не принимающие участия в образовании водородной связи, не учитываются, а к массе атома без водорода просто прибавляется 1.

Однако подобные типы моделирования могут приводить к допущению L-D переходов.

Крупнозернистое иловое поле Martini характеризуется тем, что в нём четыре тяжёлых атома и связанные ими атомы водорода объединяются в одну частицу. В результате падает число частиц в системе, что позволяет легче обсчитывать потенциальные энергии и моделировать очень крупные системы, но в результате такого упрощения теряется большое количество информации, что может приводить к неправильным ответам.

Лекция 6. Оптимизация геометрии молекулярной динамики

Минимизация энергии и другие методы исследования поверхности потенциальной энергии

С помощью силовых полей производится упрощение описания больших систем, которые свойственны для моделирования биополимеров. Силовое поле позволяет описать все взаимодействия в системе и посчитать её энергию. Естественно, можно произвести оптимизацию энергии, которая происходит за счёт изменения положения атомов, то есть оптимизации геометрии.

У биополимеров очень сложная поверхность потенциальной энергии, в связи с чем возникают сложности оптимизации энергии системы. Мы можем минимизировать функцию, если знаем её вид, однако функция энергии молекулярных систем как правило сложна, и мы её не знаем, и переменных как минимум 3 (координаты).

Суть любой минимизации сводится к тому, что мы должны сделать так, чтобы производные по координатам равнялись нулю.

Для минимизации энергии существует два типа алгоритмов:

- алгоритмы с использованием производных
- алгоритмы без использования производных

В данном случае производные являются не аналитическими, а численными, то есть мы смещаем часть системы, смотрим, как меняется от этого энергия, и делим одно на другое. Определение численных производных – ключевой метод исследования.

Но также существуют алгоритмы без использования производных, они используются намного реже.

Использование производных может предоставить информацию о форме поверхности потенциальной энергии, можно определить точки перегиба и узнать, являются они минимумами или максимумами, с помощью оценки производных вокруг точки перегиба. При этом большинство методов минимизации энергии способны двигаться только вниз по поверхности.

Необходимо понимать, что оптимизация геометрии в силовых полях характеризуется локальным эффектом, так как сворачивание белков, в основном, происходит под воздействием гидрофобного эффекта, который возможен только при наличии температуры, которой нет при таком обнаружении минимумов потенциальной энергии.

Не все методы оптимизации геометрии одинаково эффективны для квантовых и молекулярно-механических систем в связи с различным количеством частиц в этих системах.

Алгоритмы без использования производных

Самый простой метод – это Симплекс – метод изменения одной переменной. Суть метода заключается в том, что мы смотрим, как изменение одной из координат атомов влияет на общую энергию системы, и с небольшим шагом постепенно перетасовываем систему так, чтобы суммарная энергия уменьшилась.

Алгоритмы с использование производных

Алгоритмы с использованием производных опираются на ряды Тейлора:

$$U(x) = U(x_k) + (x - x_k)U'(x_k) + (x - x_k)^2 U''(x_k)/2 \ldots \quad (6.1)$$

где x_k – это матрица векторов текущего состояния системы.

Каждый элемент матрицы U`(x_k) – это первая производная по одной из переменных (x,y,z), и размерность матрицы 3Nx1.

Каждый элемент i,j матрицы U``(x_k) – это вторая производная по $\partial x_i \partial x_j$. Таким образом, размерность матрицы 3Nx3N. Эта матрица называется Гессиан или «матрица сил».

Алгоритмы с использованием производных первого порядка

Наиболее часто, особенно для небольших систем, применяется *алгоритм наискорейшего спуска*. Суть его заключается в том, что движение происходит вдоль общей силы систе мы с шагом s_k, индивидуальным для каждого атома (элемента системы):

$$s_k = \frac{F_k}{max|F_k|} \quad (6.2)$$

$$x_{k+1} = x_k + l_k * s_k \quad (6.3)$$

Таким образом, при уменьшении энергии шаг будет увеличиваться, а при увеличении – уменьшаться или будет происходить движение в обратном направлении.

Другим распространённым алгоритмом с использованием производных первого порядка является *алгоритм сопряжённых градиентов*. Этот алгоритм позволяет двигаться не как в методе наискорейшего спуска, а по градиенту уменьшения силы. Этот метод чуть менее быстрый, чем алгоритм наискорейшего спуска.

$$v_k = -g_k - \gamma v_{k-1} \quad (6.4)$$

$$\gamma_k = \frac{g_k \circ g_k}{g_{k-1} \circ g_{k-1}} \quad (6.5)$$

Алгоритмы с использованием производных второго порядка

Алгоритмы с использованием производных второго порядка основаны на построении матрицы вторых производных – матрицы Гессиан, и в неком минимуме энергии вторая производная должна равняться нулю, а значения вторых производных, геометрически находящихся рядом с этим минимумом, должны быть положительными. Считать вторые производные «в лоб» медленно и долго, особенно для белков и больших систем, поэтому существуют оптимизированные методы. Самые распространённые – *квази-Ньютоновские методы*, которые подразумевают использование аппроксимаций, и расчёт обратного Гессиана производится только для успешных итераций.

Квази-Ньютоновские методы:

- Девидсно-Флетчер-Пауер (DFP)
- Бройден-Флетчер-Голдфарб-Шано (BFGS) – наиболее часто используемый
- Муртуаг-Саргент (MS)

Минимумы, максимумы, стационарные точки, переходные состояния

В биологических системах важно понимать не только положение глобального минимума, причём в большинстве случаев методы оптимизации геометрии не могут привести к нему. Часто важны точки перегиба, поскольку в большинстве случаев это локальные максимумы, через которые должна перейти система, чтобы оказаться в другом состоянии.

В терминах матриц производных, если $f`(x)=0$:

- в максимуме все собственные значения Гессиана отрицательные
- в минимуме собственные значения Гессиана либо 0, либо положительные
- в стационарной точке не менее одного собственного значения должно быть отрицательным

Что касается переходных состояний, направление реакции определяется разницей в энергии между состояниями, и высота барьера определяет скорость реакции.

Для того, чтобы алгоритмы поиска переходного состояния сработали хорошо, нужно построить систему так, чтобы она была в *квадратичной области переходного состояния*. Эта область расположена вблизи переходного состояния и характеризуется тем, что одно из собственных значений Гессиана в ней становится отрицательным.

Существует несколько методов поиска переходных состояний:

- Метод сканирования поверхности энергии: генерируются координаты в округе стартового состояния и производится расчёт энергии. Работает для малых систем.
- Методы с движением только по одной координате (не обязательно пространственной).

Важный момент: методы минимизации энергии с использованием первых производных могут принять переходное состояние за минимум.

Введение времени и температуры в молекулярной динамике

Понятно, что одна структура с минимизированной энергией не является отображением состояния всех молекул вещества в данных условиях, так как активность биологических объектов связана с изменениями конформаций, и рассчитанные на основе этой структуры свойства, скорее всего не будут соответствовать эксперименту. В связи с этим для адекватного описания системы необходим ансамбль конформаций молекул при данной температуре и давлении.

Если мы говорим о температуре, то мы говорим и о времени, так как температура связана с кинетической энергией, которая в свою очередь связана со скоростью, которая является изменением координат во времени. Таким образом, в молекулярной динамике биологических структур необходимо следить за эволюцией системы во времени. Это

делается довольно просто, так как сила порождает ускорение, а ускорение обозначает наличие скорости и смещения, значит, время учитывается:

$$F = m\frac{\partial^2 r}{\partial t^2} \tag{6.6}$$

Два основных метода оценки энергии для текущего набора координат с применением силового поля: молекулярная динамика и метод Монте-Карло.

Метод Монте-Карло – случайное изменение координат и исследование пространства таким образом.

Метод молекулярной динамики – итеративное наблюдение за эволюцией системы во времени.

Молекулярная динамика

Алгоритм молекулярной динамики заключается в том, что рассчитываются силы в стартовом состоянии, далее через небольшое время Δt эти силы приводят к смещению атомов. Время смещения должно быть порядка 1 фс – период колебания атома водорода. После этого происходит пересчёт сил, происходят новые смещения, и этот цикл повторяется очень много раз, в результате чего можно наблюдать за эволюцией системы во времени (Рис. 6.1).

Предсказыванием новых координат на основе действующих на атом сил занимается *интегратор*. Предсказывание необходимо, поскольку Δt – не бесконечно малая величина. Рассмотрим наиболее распространённые алгоритмы интеграторов.

Ряд Тейлора:

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + \frac{1}{6}\delta t^3 b(t) \ldots \tag{6.7}$$

Алгоритм Верле:

$$r(t+\delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) \tag{6.8}$$

$$r(t-\delta t) = r(t) - \delta t v(t) + \frac{1}{2}\delta t^2 a(t) \tag{6.9}$$

сложив эти два выражения, получим:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + \delta t^2 a(t) \tag{6.10}$$

То есть алгоритм Верле основан на предсказании координат на основе действующих сил, используя предыдущие, текущие и будущие значения координат (так определяются направления сил.

Наиболее часто используется алгоритм leap-frog. Он является быстрым вариантом алгоритма Верле. Основная суть в том, что обсчитываются скорости не в конечных точках Δt, а в середине, что позволяет хорошо аппроксимировать сложную функцию положения частицы от времени:

$$r(t+\delta t) = r(t) + \delta t v\left(t + \frac{1}{2}\delta t\right) \tag{6.11}$$

$$v\left(t + \frac{1}{2}\delta t\right) = v\left(t - \frac{1}{2}\delta t\right) + \delta t a(t) \tag{6.12}$$

тогда скорость в момент *t*:

$$v(t) = \frac{1}{2}\left[v\left(t + \frac{1}{2}\delta t\right) + v\left(t - \frac{1}{2}\delta t\right)\right] \tag{6.13}$$

Этот алгоритм является наиболее эффективным на данный момент.

Рисунок 6.1 - Алгоритм молекулярной динамики.

Периодические граничные условия

Периодические граничные условия – подход в молекулярной динамике, суть которого сводится к тому, что атом, достигающий границы ячейки, переносится через эту границу и появляется с другой стороны ячейки. Таким образом, сохраняется число атомов в системе и избегаются краевые эффекты, то есть ячейка виртуально становится бесконечно большой по всем направлениям (Рис. 6.2).

Однако важен размер ячейки, поскольку возникает вероятность контакта белка самого с собой через границу ячейки. На примере поли-аланина было показано, что при маленьких размерах ячейки происходит искусственная стабилизация α-спирали поли-аланина (Рис. 6.3), то есть ячейка должна быть достаточно большой: в общем случае примерно 20 ангстрем отступа от белка до края ячейки. Это позволяет избегать артефактных явлений при моделировании.

В программе gromacs существует довольно большое количество возможных форм моделируемой ячейки (Рис. 6.4), главное, чтобы для них могли работать периодически граничные условия (PBC).

Также существуют системы, для которых неудобно использование PBC (капли жидкости, Ван-дер-Ваальсовы кластеры, гетерогенные системы при неравновесии, моделирование в вакууме), и тогда используют *сферические граничные условия*.

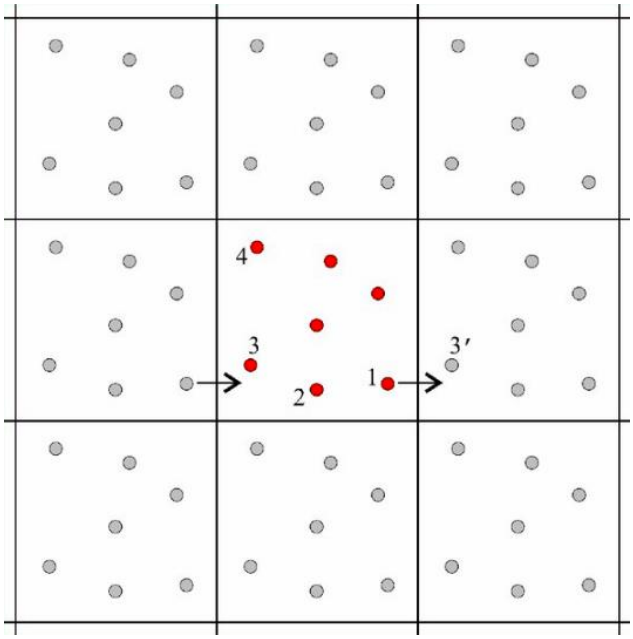


Рисунок 6.2 - Периодические граничные условия. Цифрами показано, что число частиц в ячейке неизменно.

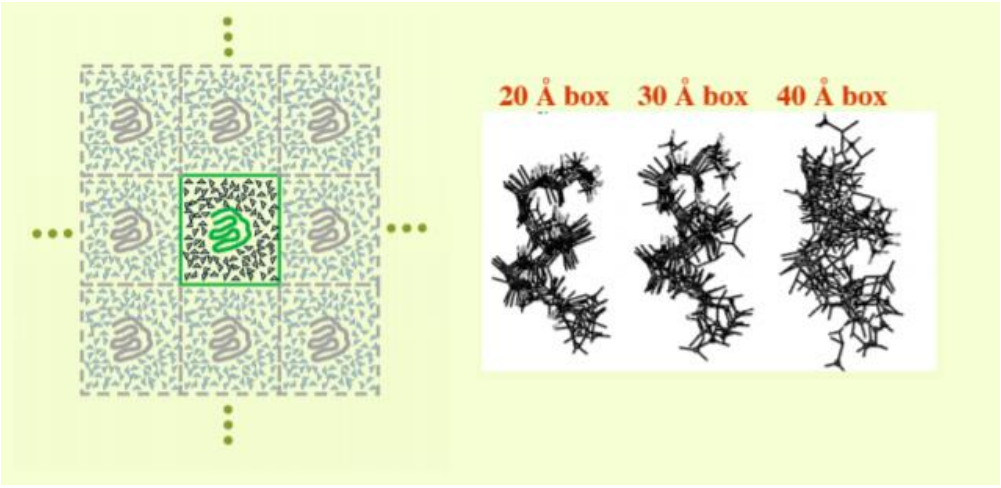


Рисунок 6.3 - Искусственная стабилизация α-спирали поли-аланина за счёт взаимодействия белка самого с собой через край ячейки в периодически граничных условиях при маленьких размерах ячейки.

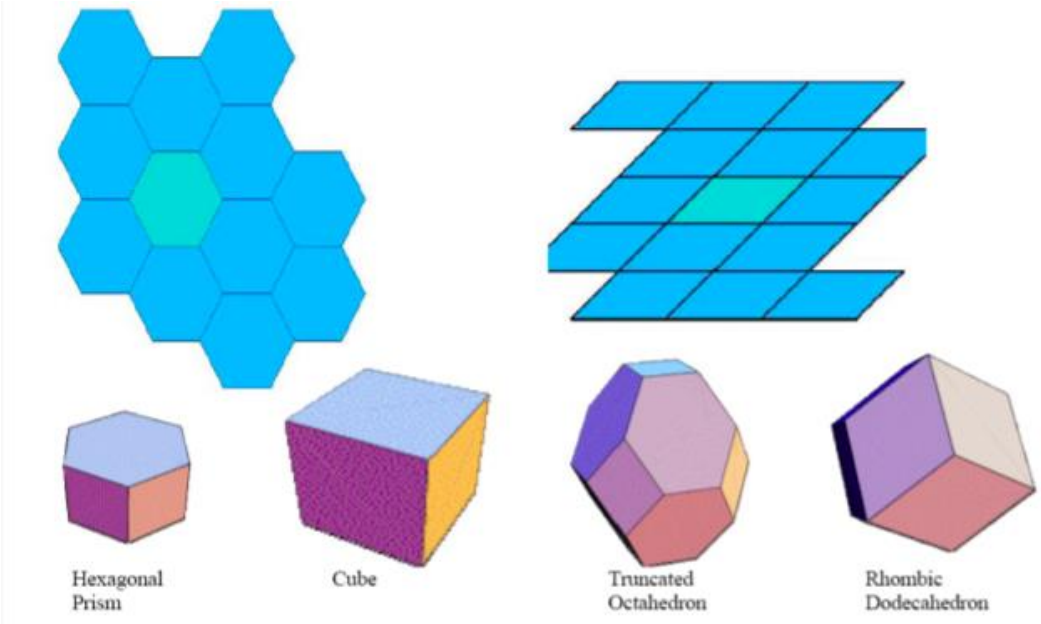


Рисунок 6.4 - Возможные формы ячеек.

Метод ближайших соседей

Главное в моделировании таких крупных систем как биополимеры – это производительность, поскольку от неё зависит насколько большие по времени движения белков можно рассчитать. Обычно нас интересуют микросекунды и миллисекунды, и требуется огромное количество шагов (порядка миллиарда), поскольку мы интегрируем с шагом в 1 фс. Для повышения производительности очень важна оптимизация. Метод ближайших соседей позволяет оптимизировать расчёт нековалентных взаимодействий, в которых состоит основная тяжесть расчёта, так как это расчёт через пространство, и там используются двойные суммы. Естественно, применение обрезания непринципиально меняет скорость счёта, так как посчитать расстояние – это почти посчитать энергию. Белки моделируются в вязкой среде, и основная идея списка соседей заключается в том, чтобы обновлять его не на каждом шагу, а раз в 10-20 шагов, так как за этот период окружение атома в жидкостях меняется незначительно.

Увеличение шага интегратора МД

Повышение шага интегрирования значительно увеличивает эффективность моделирования (так как можно добиться больших времён МД с меньшим количеством расчётов), однако оно ограничено периодом колебаний наиболее быстро колеблющихся атомов, то есть самых лёгких – атомов водорода (C-H, N-H, O-H связи).

Существуют алгоритмы *ограничения быстрых колебаний*: SHAKE (Рис. 6.5) и Links, – суть которых заключается в том, что ускорение быстро движущегося атома водорода «размазывается» по тяжёлому атому и другим атомам водорода, которые присоединены к этому тяжёлому атому, что позволяет увеличить шаг с 1 фс до 2 фс.

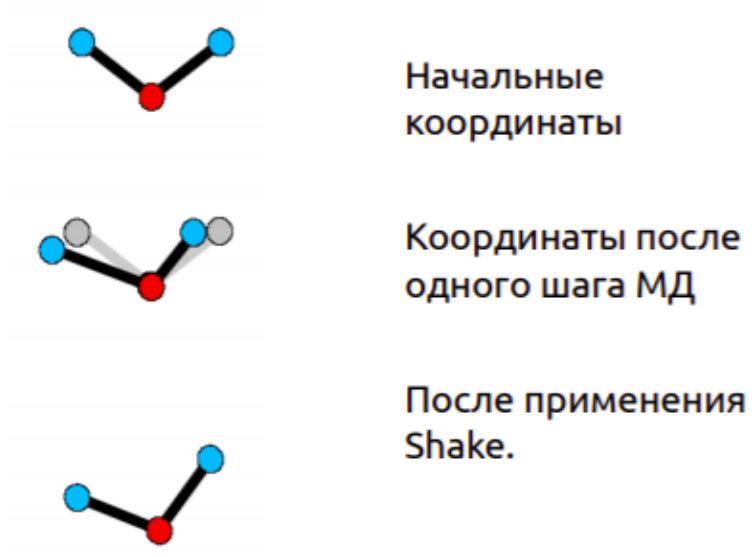


Рисунок 6.5 - SHAKE-алгоритм.

Такой подход более генерализованно осуществлён с помощью такого понятия как *пустышки* (*Dummies*). Пустышки – некие атомы, которые определяют смещение группы атомов в зависимости от такого, какие силы на них действуют. Таким образом, все атомы в группе смещаются скоординировано, поскольку есть атом (пустышка), объединяющий их в одно и определяющий смещение всей группы в зависимости от того, какие силы действуют на эту группу атомов.

Существуют различные типы атомов-пустышек (Рис. 6.6).

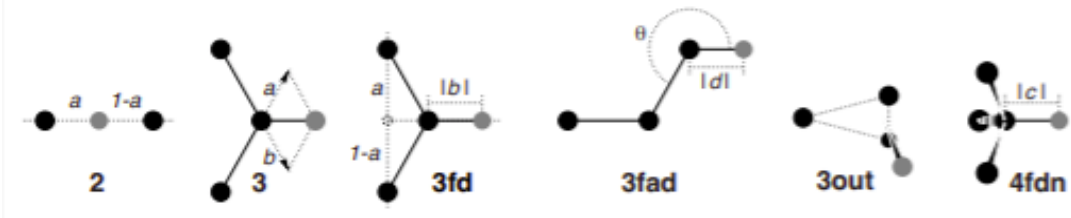


Рисунок 6.6 - Конструкции атомов-пустышек в GROMACS. Серым показаны атомы-пустышки, чёрным – реальные атомы, входящие в конструкцию.

Используя атомы-пустышки, можно увеличить шаг интегрирования до 5-7 фс. Однако сам обсчёт движения атомов-пустышек довольно сложен и затруднён на GPU, поэтому они не дают большого прироста к скорости расчёта, и чаще проще посчитать на GPU системы без их использования.

Температура и давление

Температура – очень важный момент в молекулярной динамике, поскольку она моделируется явно.

При релаксации системы появляется излишек кинетической энергии, что должно приводить к нагреву системы. Самый простой способ убирать этот излишек: при помощи масштабирования скоростей – *термостат Берендсена*.

$$\frac{\partial T}{\partial t} = \frac{T_0 - T}{\tau} \tag{6.14}$$

$$\lambda = \left[1 + \frac{n_{TC} \triangle t}{\tau_T}\left\{\frac{T_0}{T\left(t - \frac{1}{2}\triangle t\right)} - 1\right\}\right]^{1/2}, \tag{6.15}$$

где n_{TC} – частота; λ – коэффициент масштабирования.

Термостат – алгоритм, позволяющий сохранить температуру на приемлемом уровне при релаксации системы. То есть его задача – избегать перегрева при переходе потенциальной энергии системы в кинетическую.

Кроме масштабирующих термостатов, существуют столкновительные термостаты и термостаты с дополнительной степенью свободы.

Также осуществляется контроль давления в системе. Определяется количество атомов, пересекающих границу ячейки за определённое время и, в зависимости от референсного давления, размер ячейки может адаптироваться.

Баростат Берендсена:

$$\frac{\partial P}{\partial t} = \frac{P_0 - P}{\tau_p} \tag{6.16}$$

Баростат Паринелло-Рахмана:

$$\frac{\partial b^2}{\partial t^2} = VW^{-1}b^{t-1}\left(P - P_{ref}\right), \tag{6.17}$$

где b – матрица векторов ячейки, V – объём ячейки, W – матрица параметров, определяющих силу сопряжения.

Методология подготовки системы для МД

1. Построение топологии молекулы на основе координат.
2. Выбор формы и размера ячейки.
3. Минимизация энергии структуре в вакууме.
4. Добавление растворителя и ионов в ячейку.
5. "Утряска" воды и ионов вокруг неподвижной молекулы.

Воду обычно добавляют во всю ячейку и далее удаляют молекулы воды, пересекающиеся с белком. При этом часть воды попадает в полости белка, где её не должно быть, поэтому на небольшой промежуток времени в систему добавляют давление и температуру и разрешают двигаться воде и запрещают двигаться белку,

чтобы она вытекла из этих мест и расположилась менее периодично, чем когда её добавляют изначально. Также воду можно добавлять не по одной молекуле, а по заранее уравновешенным кубикам (Рис. 6.7).

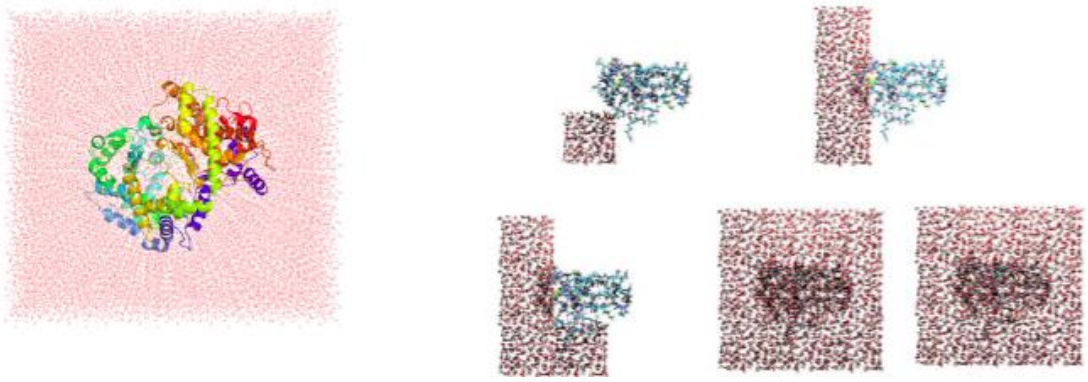


Рисунок 6.7 - Добавление воды в ячейку: слева показано, как вода добавляется во всю ячейку, в том числе и в полости белка; справа показано добавление воды в ячейку по кубикам.

Неявный растворитель

Довольно часто в молекулярной динамике используется модель неявного растворителя. В таком случае молекулы воды заменяют потенциалами. Ключевой способ обсчёта эффектов такой модели воды – расчёты поверхности, доступной растворителю, и штрафование за появление гидрофобных участков на этой поверхности. Также необходимо скалирование электростатики, то есть применение различных штрафов в зависимости от полярности групп.

В качестве примера можно привести метод Пуассона-Больцмана. Он является довольно точным, но технически сложным для расчёта.

Основными недостатками неявного растворителя являются:

- в основном, учитывается электростатическая составляющая
- гидрофобный эффект не учитывается
- вязкость как результат столкновений и скоростей не рассчитывается и не учитывается
- водородные связи воды с исследуемым объектом не могут быть учтены
- исчезает возможность учёта водных мостиков

Длина траектории МД

Длина траектории, которая необходима для наблюдения интересующего нас явления зависит от потенциального барьера события, которое мы хотим наблюдать. Чем больше барьер, тем длиннее нужна траектория (Рис. 6.8).

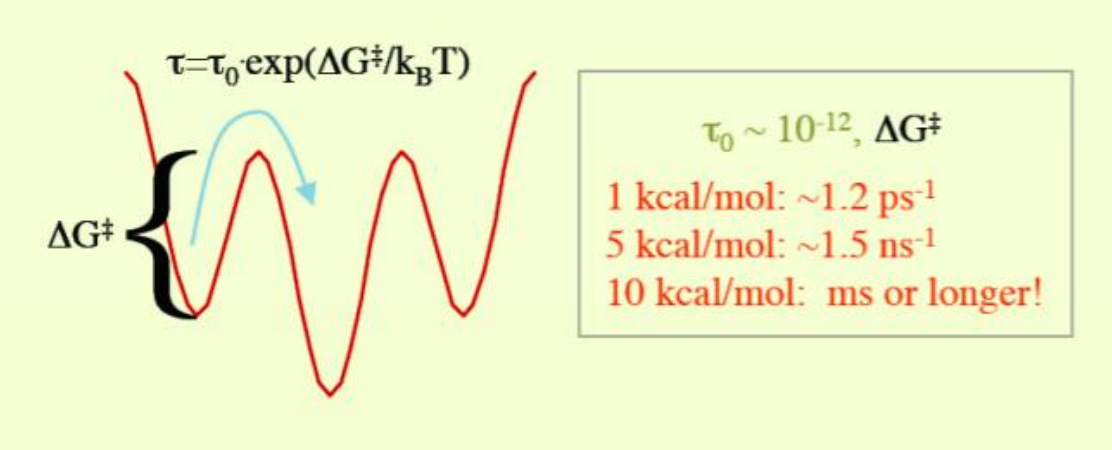


Рисунок 6.8 - Связь потенциального барьера с необходимой длиной траектории МД.

Лекция 7. Модификации молекулярной динамики

Уравнение Шредингера

Уравнение Шредингера представлено в виде:

$$i\hbar\frac{\partial}{\partial t}\psi(r,t)=\left[\frac{-\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+V(r,t)\right]\psi(r,t) \quad (7.1)$$

Или:

$$H\Psi=E\Psi; H=-\frac{\hbar^2}{m}\nabla^2-\frac{Ze^2}{4\pi\varepsilon_0 r} \quad (7.2)$$

В молекулярной механике, где аппроксимируем электронную плотность уравнениями классической физики:

$$F=m\frac{\partial^2 r}{\partial t^2} \quad (7.3)$$

Замечание: необходимо учитывать гидрофобный эффект, который важен в формировании структуры биополимеров (белки, ДНК). Осталось придумать как следить за эволюцией системы во времени. Представим простое уравнение силового поля:

$$U=\sum_{bonds}\frac{k_i}{2}(l_i-l_0)^2+\sum_{angles}\frac{k_i}{2}(\phi_i-\phi_0)^2+\sum_{torsions}\frac{V_n}{2}(1-\cos(n\omega-\gamma))\quad+$$

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N}\left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]+\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}\right) \quad (7.4)$$

Для того, чтобы следить за эволюцией системы у нас есть только один способ – молекулярная динамика. Методом молекулярной динамики можно преобразовать выражение 7.4 в:

$$v\left(t+\frac{\Delta t}{2}\right)=v\left(t-\frac{\Delta t}{2}\right)+\frac{F(t)}{m}\Delta t \quad (7.5)$$

Кроме того, можно набрать ансамбль конформаций, который говорит о том, какие состояния может занимать система в определенных условиях (T, p, pH), тогда можно воспользоваться методом Монте-Карло:

$$acc(0\to\eta)=\min(1,\exp\{-\beta[u(r_0^N-U(r^N))]\}) \quad (7.6)$$

В отличие от метода молекулярной динамики метод Монте Карло отличается тем, что это не эволюция системы, а сканирование состояний. При этом сканирование может быть осуществлено любым способом в том числе и с помощью “рандомайзера”.

Молекулярная динамика

В методе молекулярной динамики мы запускаем расчет силы, силы порождают ускорение, а ускорение порождает смещение, а смещение порождает новые силы

(рис.7.1.). Это и есть эволюция системы во времени. В этой эволюции есть и скорости и силы, и все, что необходимо для наблюдения за этой системой (T,p).

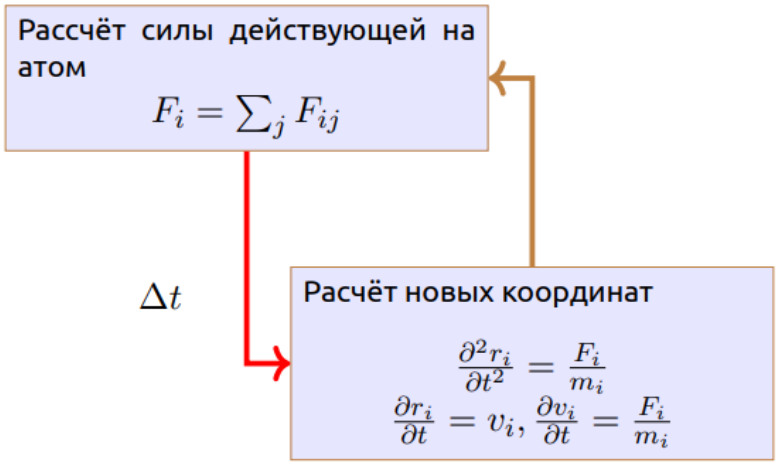


Рисунок 7.1. Основа метода молекулярной динамики.

В прошлом разделе мы пытались понять какие сколько нужно наблюдать за системой, чтобы преодолеть барьер какой-либо величины. Например, для того чтобы преодолеть барьер в 10 ккал, нужно наблюдать за системой больше, чем 1 мксек. При этом шаг больше, чем 1 фсек сделать невозможно, так как необходимо наблюдать и учитывать колебания легких атомов (водорода). Поэтому количество итераций необходимо сделать 10^9. Это выглядит весьма затратно, но современные технологии помогают решить данный вопрос.

Гибридное QM/MM моделирование

Иногда интересно узнать не только эволюция системы во времени, но и как реализуется функция ферментов. При ферментативной реакции происходит химическая реакция. Химическая реакция – это перекомпоновка систем, образование, разрыв связей и т.д. А в уравнении силового поля не предусмотрен разрыв связей. Как решают эту проблему:

- Основная идея: разделить большую систему на квантовую и молекулярную части. Грубо говоря, будем считать с помощью QM только активный центр.
- Электростатическое окружение из MM части чувствуется QM частью.
- MM часть принимает силы из QM части и соответственно адаптируется.

Простейший Гамильтониан для QM/MM системы:

$$H = -\frac{1}{2}\sum_i^{elect}\nabla^2 + \sum_i^{nucl}\sum_j^{elect}\frac{1}{r_{ij}} + \sum_i^{nucl}\sum_i^{nucl}\frac{Z_i Z_j}{R_{ij}} - \sum_i^{elect}\sum_j^{MM_q}\frac{Q_i}{R_{ij}} + \sum_i^{nucl}\sum_j^{MM_q}\frac{Z_i Q_i}{R_{ij}} \qquad (7.7)$$

К QM/MM части можно добавить VdW составляющую:

$$H_{\frac{QM}{MM}} = -\sum_i^{elect}\sum_j^{MM_q}\frac{Q_i}{R_{ij}} + \sum_i^{nucl}\sum_j^{MM_q}\frac{Z_i Q_i}{R_{ij}} + \sum_i^{nucl}\sum_j^{MM_{atoms}} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \qquad (7.7.1)$$

Активный центр находится в некой упаковке из зарядов и Ван-дер-Вальсовых взаимодействий. Данные заряды влияют на ядра и электроны квантовой системы.

Реализации описания QM/MM

В данной системе надо учитывать встраивание электронов. Активный центр белка состоит из аминокислотных остатков белка и лиганда. Для включения аминокислотных остатков белка в квантовую систему необходимо разорвать где-то ковалентную связь. Зачастую ее разрывают в $C\alpha - C\beta$. С точки зрения электронной системы у нас возникает радикал. На место разрыва необходимо вставить H, получается на месте разрыва связи – CH_3.

Встраивание электронов:

$$H_{\frac{QM}{MM}} = H_e^{QM} - \sum_i^n \sum_J^M \frac{e^2 Q_J}{4\pi\epsilon_0 r_{i_J}} + \sum_A^N \sum_J^M \frac{e^2 Z_A Q_J}{e\pi\epsilon_0 R_{AJ}} \quad (7.8)$$

- Ковалентные взаимодействия между QM и MM системами описываются соответствующими параметрами из MM.
- Для QM системы в месте разрыва связи добавляют протон для восстановления системы до полного состояния.
- Сила, действующая на этот протон “заглушку” в QM системе, распределяется между атомами, между которыми происходит раздел QM и MM системами (Оценка градиента).

ONIOM

Более подробно это реализовано в методе ONIOM, который подразумевает то, что мы можем иметь дело не только с QM/MM. Основная идея:

- Рассчитываем энергию и градиент для QM системы с желаемым уровнем теории.
- Рассчитываем энергию и градиент для MM системы с учетом ранее рассчитанных данных для QM системы.
- Рассчитываем MM энергию и градиенты для QM системы и вычитаем.

$$E_{tot} = E_I^{QM} + E_{I+II}^{MM} - E_I^{MM} \quad (7.9)$$

Этот подход можно использовать не только для двух уровней теории, но и больше. Идеология данного метода очень простая, так как квантово-механический “движок” считает энергии, а двигает атомы молекулярно-механический “движок”.

Реализация QM/MM в Gromacs

Алгоритм действий в программе Gromacs:

- Добавляем атомы ”заглушки”:

[virtual_sites2]

LA QMatom MMatom 1 0.65

0.65 – доля длины связи $C - H$ от $C - C$.

- а связь описывается:

[constraints]
QMatom MMatom 2 0.153
0.153 – постоянное расстояние между атомами

- Для атомов в QM системе надо поправить описание ковалентных связей:

[bonds]
QMatom1 QMatom2 5
QMatom2 QMatom3 5

- В mdp файле описываем параметры для QM системы.

Гибридное QM/MM моделирование

Представим, как выглядит данная система издалека. На рисунке 7.2. представлена данная система. Слева изображен фермент в воде. Посередине активный центр в представлении молекулярно-механического “движка”. Справа изображено, как видит квантовый “вычислитель” то, что ему пришло на “вход”. Грубо говоря, квантовый “калькулятор” видит, что у тирозина $C_\beta - CH_3$.

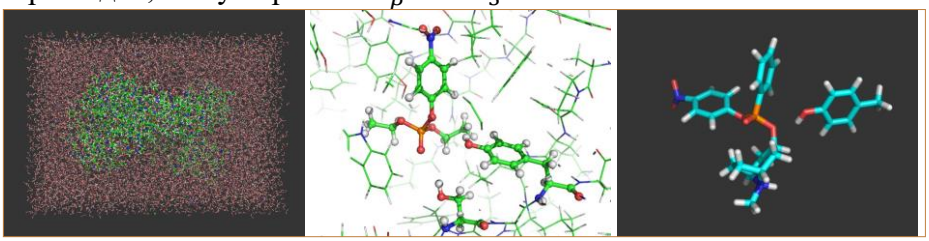


Рисунок 7.2. Пример гибридного QM/MM моделирования.

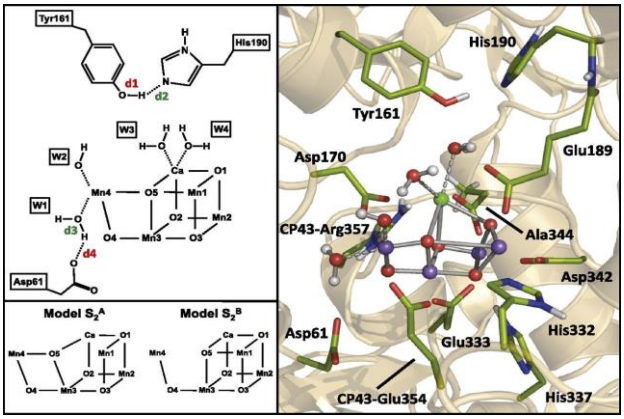


Рисунок 7.3. Пример расчета QM/MM для фотосистемы.

QM/MM можно использовать не только для того, чтобы моделировать реакции, но и исследовать передачу заряда по сложным путям (фотосистемы). Пример такого расчета представлен на рисунке 7.3. Такой расчет произвести крайне сложно, но именно для этих целей QM/MM единственный инструмент для того, чтобы понять как “путешествует” электрон в такой системе.

Adaptive resolution scheme

Часто встает вопрос не о том, как посчитать реакцию, а о том, как просто посчитать белок. Такая же идея может быть реализована в обратную сторону:

- Описание взаимодействия между полноатомным и крупнозернистым описаниями системы.
- В отличие от QM/MM уровень описания молекул системы может меняться “на лету”.

Реализация AdResS:

Система состоит в том, чтобы смешать два описания. В какой-то момент существует половина полноатомной воды, и половина крупнозернистой воды:

$$\overrightarrow{F_{\alpha\beta}} = \omega_\alpha\omega_\beta\overrightarrow{F_{\alpha\beta}^{ex,mol}} + \left[1 - \omega_\alpha\omega_\beta\right]\overrightarrow{F_{\alpha\beta}^{cg,mol}} \tag{7.10}$$

$$\omega(x) = \begin{cases} 0 : x > d_{ex} + d_{hy} \\ \cos^2\left(\frac{\pi}{2d_{hy}}(x - d_{ex})\right) : d_{ex} + d_{hy} > x > d_{ex} \\ 1 : d_{ex} > x \end{cases}$$

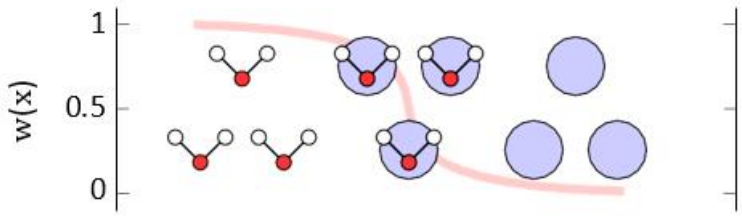


Рисунок 7.4. Зависимость количества полноатомной воды от расстояния между частицами.

Метод обмена репликами (REMD)

- Основная идея: запустить параллельно несколько счётов с разными температурами. Чем выше температура, тем быстрее происходит обмен конформациями.
- Мы можем выбрать правило, когда производить обмен конформациями.
- Если мы проводим обмен, когда потенциальная энергия одной из реплик ниже, чем других, то это похоже на моделирование отжига.
- Такой подход часто используется для моделирования самосборки.

Цель метода – это ускорить сканирование (sampling) конформационного пространства. Применимо к переходам через значимые энергетические барьеры. Кроме того, в Gromacs обмен между репликами происходит случайно по условию:

$$P(1 \leftrightarrow 2) = min\left(1, exp\left[\left(\frac{1}{k_b T_1} - \frac{1}{k_b T_2}\right)(U_1 - U_2]\right)\right) \tag{7.11}$$

При этом скорости масштабируются: $\left(\frac{T_1}{T_2}\right)^{\pm 0,5}$. Высокая температура нужна для того, чтобы преодолеть энергетический барьер.

Пример:

При самообразовании (“фолдинге”) структуры происходит множественные перестройки и как следствие энергетические барьеры необходимо преодолеть несколько раз (в отличии от химической реакции):

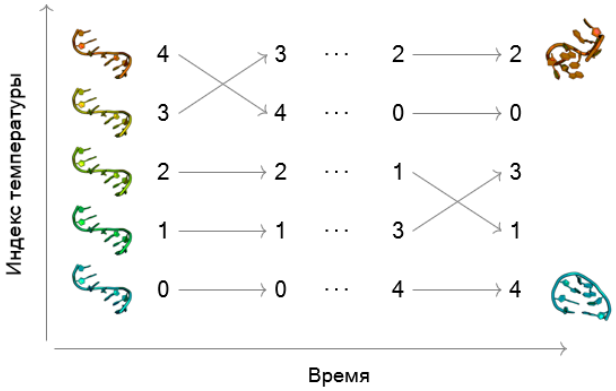


Рисунок 7.5. Зависимость индекса температуры от времени в самосборке ДНК.

Возникает вопрос: сколько необходимо сделать запусков программы, чтобы самосборка работала? В теории для того, чтобы программа работала, ансамбль состоянии при одной температуре должен пересекаться с ансамблем в другой температуре. Разница между температурами (репликами):

$$U_1 - U_2 = \frac{N_{df}c}{2} k_B (T_1 - T_2) \tag{7.12}$$

где N_{df} – это количестве степеней свободы, c –это величина от 1 до 2 для системы белок вода.

Если $T_2 = (1 + \epsilon)T_1$ тогда вероятность обмена:

$$P(1 \leftrightarrow 2) = exp\left(-\frac{\epsilon^2 c N_{df}}{2(1+\epsilon)}\right) \simeq \exp\left(-\frac{\epsilon c}{2N_{df}}\right) \tag{7.13}$$

Таким образом для вероятности обмена $e^{-2} \approx 0.135$ получаем $\epsilon \approx \frac{2}{\sqrt{cN_{df}}}$. И если мы контролируем длину связей, то: $N_{df} \approx 2\,N_{atoms}$ и при $c = 2$ надо использовать: $\epsilon = \frac{1}{\sqrt{N_{atoms}}}$.

Пример результата REMD представлен на рисунке 7.6. При данном моделировании берется диапазон температур. На картинке представлены два варианта результата подсчета. Они отличаются тем, что нижний диапазон температур отличается. У правой картинке более низкая начальная температура, так как гидрофобные взаимодействия более эффективны при низких температурах ($T\Delta S$).

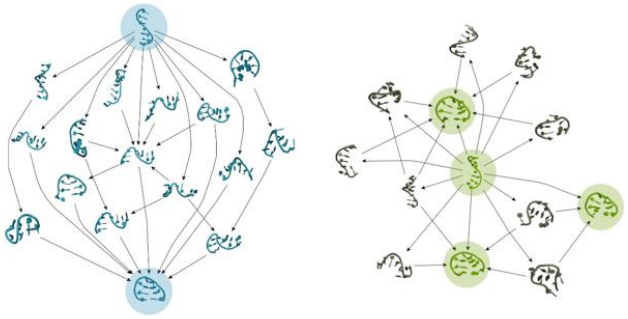


Рисунок 7.6. Результат REMD.

Коллективные переменные (CV)

Всегда ли необходимо исследовать все систему при анализе, например, фолдинга системы? Бывает так, что не нужно считать все конформации и изменения. Зачастую функционирование биополимеров сводится к тому, что мы рассчитываем небольшие, локальные переменные. Основные идеи:

- Положение всех атомов в пространстве явно избыточная информация для описания некоторых простых процессов.
- Коллективные переменные – это некоторые значения, которые наиболее значительно изменяются в ходе процесса, который нас интересует.
- Трудно предположить заранее все важные коллективные переменные.
- Недостаток описания приводит к гистерезису.
- Примеры коллективных переменных представлены в таблице 7.1.

ALPHABETA	ALPHARMSD	ANGLE
ANTIBETARMSD	CELL	CH3SHIFTS
CONSTANT	CONTACTMAP	COORDINATION
DHENERGY	DIHCOR	DIPOLE
DISTANCE	ENERGY	GYRATION
NOE	PARABETARMSD	PATHMSD
PATH	POSITION	PROPERTYMAP
RDC	TORSION	VOLUME

Таблица 7.1. Примеры коллективных переменных

Пример анализа на основе CV:

На рисунке 7.7. представлен пример анализа на основе CV. Возьмем две переменные: расстояния между двумя атомами и рассчитаем количество энергии: $E \approx \frac{1}{\beta}\log(P)$. В примере рассматривается отрыв протона от тирозина. Фосфат приближается к протону, далее увеличивается расстояние протона от тирозина и протон удаляется все

дальше и дальше. С помощью этих двух расстояний мы можем рассчитать барьер реакции.

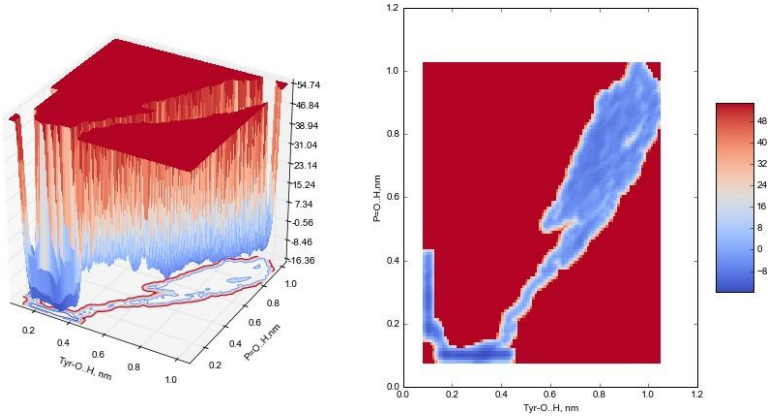


Рисунок 7.7. Пример использования коллективных переменных. Отрыв протона от тирозина.

Использование CV для влияния на МД, ABMD

adiabatic biased MD

$$V(\rho(t) = \begin{cases} \frac{K}{2}\left(\rho(t) - \rho_m(t)\right)^2, \rho(t) > \rho_m(t) \\ 0, \quad \rho(t) > \rho_m(t) \end{cases}, \text{где } \rho(t) = (CV(t) - TO)^2 \qquad (7.14)$$

Суть метода состоит в движении системы к заданным занчениям CV используя гармонический потенциал, который изменяет термические флуктуации если система не движется к заданным значениям CV.

Использование CV для влияния на МД, steered MD

Позволяет добавлять временно зависимый гармонический потенциал на одну или несколько переменных:

$$V(\vec{s}, t) = \frac{1}{2}k(t)\left(\vec{s} - (\vec{s}_o(t)\right)^2 \qquad (7.15)$$

Для примера представим некоторую молекулу в белке. К одному из атомом можно “прицепить” пружинку, а за другую часть пружинки нужно постоянно тянуть. Это приведет к тому, что пружинка растянется до какого-то состояния. Потом молекула выскочит и снова зацепится. И тем самым, зная с какой силой мы действуем на пружинку, можем оценить, как хорошо связывается молекула в том или ином месте. Это применимо к любой коллективной переменной.

Использование CV для влияния на МД, WALLS

Зачастую необходимо ограничить передвижение воды от активного сайта. Можно поставить WALLS. Суть очень проста – это выставление “стен” в виде штрафа за пределами значения CV:

$$\sum_i k_i \left(\frac{x_i - a_i + o_i}{s_i}\right)^e_i \qquad (7.16)$$

Метадинамика

Представим, что вы ночью идете по болоту и перед вами яма. Обойти ее нельзя, а глубину оценить невозможно. Это пример того, как изучаются молекулярные системы. Любая структура и ΔG состояния, а также переход из одного состояния в другое – неизвестны. Данный пример приведен на рисунке 7.8. Мы изменяем энергию системы и смотрим после какого значения изменяется какая-либо коллективная переменная.

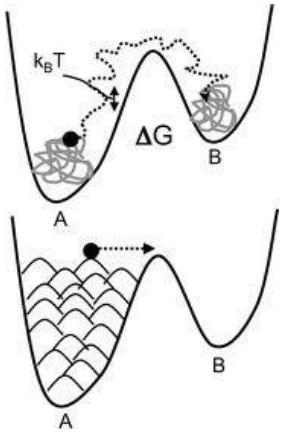


Рисунок 7.8. Пример изучения полимерных молекулярных структур.

Well-tempered metadynamics

Этот вариант метадинамики позволяет добавлять энергию более аккуратно, что приводит к тому, что добавленное количество энергии начинает стремиться к энергии перехода между состояниями.

Обсуждение

Преимущества:

- Ускорение событий путём выталкивания системы из известной области
- Знание результата не нужно, возможно все, что может быть в этой системе
- Возможность восстановить профиль поверхности потенциальной энергии

Недостатки:

- Динамика и температура изменены человеком.
- Данные о кинетике процессов не доступны (а может и нет: arXiv:1309.5323, Tiwari & Parrinello)

Выбор CV:

Правильный выбор CV это критический момент, вот некоторые требования:

- CV должны описывать процесс интереса.
- Включать все медленно изменяющиеся степени свободы
- Количество CV должно быть не большим
- Используйте химическую/физическую интуицию
- Подход проб и ошибок

Примеры с использованием метадинамики может служить образование структуры белка на основе знания химических сдвигов и реконструкция FES белка:

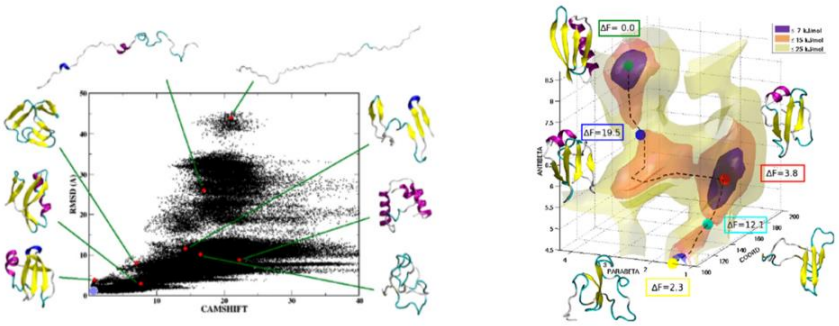


Рисунок 7.9. Образование структуры белка на основе химических сдвигов.

Как это работает: берем на старте некую коллективную переменную, которая показывает, насколько отклоняется наблюдаемые химические сдвиги от теоретических химических сдвигов. Далее делаем метадинамику по данному значению. Цветом обозначается уровень энергии.

Еще одним примером является изучение процесса фолдинга белка. Данный процесс достаточно сложно сделать с помощью метадинамики, так как для одного состояния свернутого белка можно подобрать множество состояний развернутого белка.

Использование метадинамики в определении пути лиганда к сайту связывания. По сути, данный метод — это метадинамика по координатам. У нас есть некий потенциал, который не позволяет лиганду выскочить за пределы некой воронки. Лиганд передвигается только в ограниченном пространстве. Пример приведен на рисунке 7.10. Этот способ может найти альтернативные места связывания. Кроме того, можно определить механизм связывания лиганда.

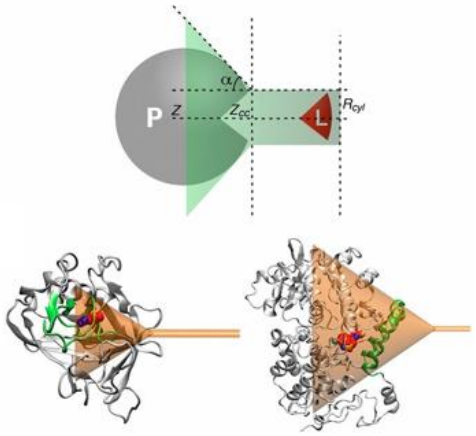


Рисунок 7.10. Определение пути лиганда к сайту связывания. .

Лекция 8. Расчет свободной энергии

Свободная энергия

Свободная энергия – это тот численный параметр, который позволяет соединить вычисление и эксперимент. Характеристика свободной энергии:

- Изменение свободной энергии определяет направление реакции.
- Сравнивая изменение свободной энергии, можно изучать эффективность ингибиторов.
- Определение значений свободной энергии позволяет соотносить результаты моделирования с экспериментальными результатами.

Большинство данных получено при постоянном давлении, то наиболее распространено выражение свободной энергии как энергии Гиббса:

$$\Delta G = \Delta H - T\Delta S \tag{8.1}$$

$$\Delta G = -RTlnK \tag{8.2}$$

Где, $K = \frac{[p_1]}{[p_2]} = \frac{N_1}{N_2} \approx 10^5$, тогда $\Delta G \approx 14$ кДж

МД и МК трудно использовать для расчёта энергии Гиббса, так как оба метода “не любят” те места фазового пространства, где энергия не минимальна:

$$A = k_B T ln\left(\int\int \partial p^N \partial r^N e^{+\frac{H(p^n, r^N)}{k_B T}} \rho(p^N, r^N)\right) \tag{8.3}$$

Придумаем способы для расчета свободной энергии:

Сравним свободные энергии этанола и этантиола в воде. Для решения этой задачи МД и МК могут подойти. Рассмотрим три метода:

- Термодинамическая пертурбация.
- Термодинамическое интегрирование.
- Метод медленного роста.

Термодинамическая пертурбация

Упростим задачу и представим две молекулы: X — это этанол в кубике воды; Y — это этантиол в кубике воды:

$$\Delta A = A_x - A_y = k_B T ln\left(\frac{Q_Y}{Q_X}\right) \tag{8.4}$$

$$\Delta A = k_B T \int\int \frac{\partial p^N \partial r^N e^{+\frac{H_Y(p^n, r^N)}{k_B T}}}{\partial p^N \partial r^N e^{+\frac{H_X(p^n, r^N)}{k_B T}}} \tag{8.5}$$

В терминах средних значений по ансамблю:

$$\Delta A = k_B T \left\langle \exp\left(\frac{H_Y(p^n, r^N)}{k_B T} - \frac{H_X(p^n, r^N)}{k_B T}\right)\right\rangle_0 \tag{8.6}$$

Тогда можно посчитать работу, которую необходимо совершить:

$$-\Delta A = k_B T \left\langle \exp\left(\frac{H_X(p^n, r^N)}{k_B T} - \frac{H_Y(p^n, r^N)}{k_B T}\right)\right\rangle_1 \quad (8.7)$$

Реализация уравнений проста:

Нам надо посчитать поведение этанола в воде и для найденных конформаций посчитать энергию этантиола. Для контроля, превращение, наоборот.

Однако, предыдущий подход хорош, если фазовые пространства молекул похожи или значимо пресекаются. Если же пространства не пересекаются, давайте введём состояние 1 между состояниями X и Y, то есть:

$$k_B T \ll |H_X - H_Y|$$

$$-\Delta A = k_B T \left\langle \exp\left(\frac{H_Y - H_1}{k_B T} - \frac{H_1 - H_X}{k_B T}\right)\right\rangle \quad (8.8)$$

Задаём значение λ от 0 до 1 и делаем МД для каждого λ:

$$k_b(\lambda) = \lambda k_b(Y) + (1 - \lambda) k_b(X) k_a(\lambda) = \lambda k_a(Y) + (1 - \lambda) k_a(X)$$

.....

$$q(\lambda) = \lambda_q(Y) + (1 - \lambda) q(X) \qquad \epsilon(\lambda) = \lambda_\epsilon(Y) + (1 - \lambda)_\epsilon(X)$$

$$\sigma(\lambda) = \lambda_\sigma(Y) + (1 - \lambda)\sigma(X) \quad (8.9)$$

На самом деле для каждого значения параметра сопряжения (λ) надо сначала уравновесить систему и только после этого снимать значения энергий.

- Расчёт с λ от 0 до 1 это прямая выборка
- Расчёт с λ от 1 до 0 это обратная выборка
- Бывает двойная выборка

Для расчета ΔG необходимо взять n количество точек с разными состояниями и проинтегрировать (рис. 8.1).

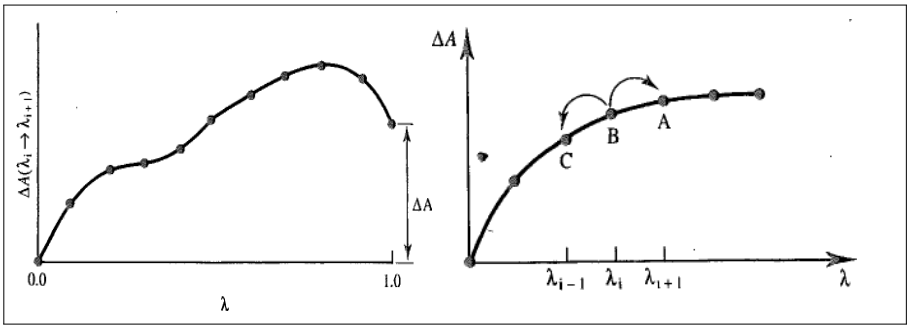


Рисунок 8.1. Пример расчета ΔG с помощью термодинамической пертурбации.

Все точки независимы. Каждая из них обозначает, что λ чему-то равна. Когда мы идет по выборке, то начинаем считать какое количество энергии затратится при переходе из состояния, например, 3 в 4. Дальше делаем переход со смещением в 0.5 и вернемся

назад (3.5 в 2.5). И так можно проверить правильно ли попадает изменение энергии Гиббса в общий тренд.

Можно интегрировать изменение энергии по параметру сопряжения:

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \frac{\langle \partial H(p^n, r^N}{\partial \lambda \rangle}_{\lambda} \partial\lambda; \quad \frac{\partial H}{\partial \lambda} \approx \frac{\Delta H}{\Delta \lambda} \tag{8.10}$$

При интегрировании, мы получим некую площадь (рис.8.2.) и можем объявить это работой по превращению из одного вещества в другое:

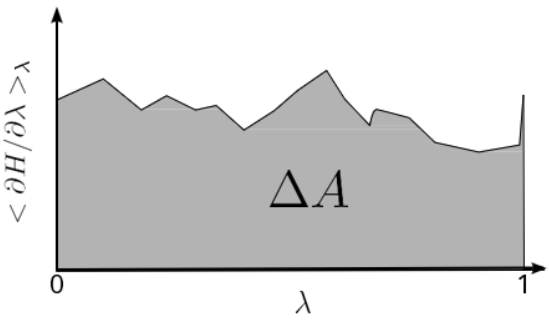


Рисунок 8.2. Расчет площади изменения энергии по параметру сопряжения

Здесь необходимо следить за тем, чтобы $\Delta\lambda$ не было настолько большим, чтобы фазовые состояния не пересекались. Самое неприятное явление может состоять в том, что при подсчете можно увидеть, что гистограммы выбранного параметра не пересекаются. Для решения этой проблемы необходимо сгенерировать новую точку между двумя предыдущими состояниями и пересчитывать.

Метод медленного роста

Суть метода состоит в том, что необходимо осуществлять маленькие шаги так, чтобы Гамольтиан следующего шага был близок к текущему:

$$\Delta A = \sum_{i=0;\lambda=0}^{i=N;\lambda=1} (H_{i+1} + H_i) \tag{8.11}$$

Однако λ в данном методе выбирается “на лету” и данный подход не нашел применения, так как невозможно заранее угадать сколько нужно считать, чтобы получилось сходимость гистограмм.

Термодинамические циклы

Таким образом, можно рассчитывать переходы между разными состояниями. И теперь переходим к расчету термодинамических циклов (рис. 8.3.). Необходимо представить некие состояния, которые нужно взять для того, чтобы правильно рассчитать энергию Гиббса.

- Исследователей часто интересует энергия нековалентного связывания лиганда с рецептором.

- Допустим есть два лиганда. Можно посчитать их $\Delta\Delta G$ *просто* промоделировав процесс связывания, но это трудно исполнимо.

Свободная энергия — это функция состояния, то надо всего-то посчитать переход из одного лиганда в другой, как в растворе, так и в белке.

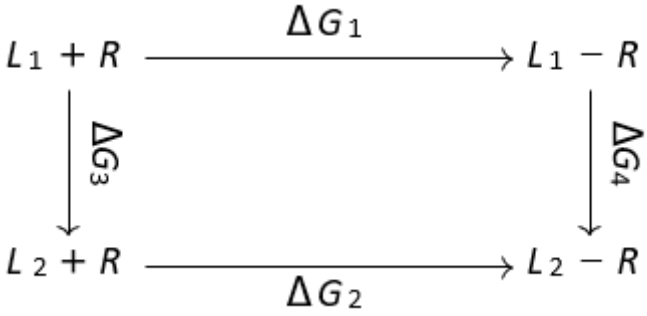


Рисунок 8.3. Пример термодинамического цикла.

Расчёт абсолютного изменения свободной энергии

Основная идея — это делать термодинамические циклы через состояния комплекса в растворе и газовой фазе:

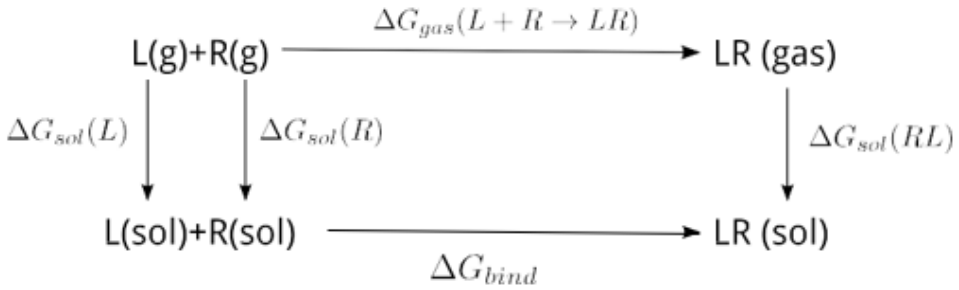


Рисунок 8.4. Схема расчета термодинамического цикла через состояния комплекса в растворе и в газовой фазе.

Давайте рассчитаем ΔG связывания белка с лигандом:

$$\Delta G_{bind} = \Delta G_{gas}(R + L \rightarrow RL) + \Delta G_{sol}(RL) - \Delta G_{sol}(R) - \Delta G_{sol}(L)$$

подставляем:

$$\Delta G_{sol}(X) = \Delta G_{gas}(X \rightarrow 0) - \Delta G_{sol}(X \rightarrow 0)$$

ΔG_{gas} сокращается и получается:

$$\Delta G_{bind} = \Delta G_{sol}(L \rightarrow 0) - \Delta G_{sol}(LR \rightarrow R)$$

Замечание: моделирование белков в газовой фазе не самый лучший вариант расчета.

Расчёт изменения энтальпии и энтропии

Изменение свободной энергии может быть рассчитано достаточно точно. В хороших случаях ошибка всего 1 ккал/моль. Изменение энтальпии можно было бы посчитать, сравнив потенциальные энергии двух систем, но там достаточно большие

значения с большими ошибками. На сегодняшний момент расчёт энтальпии и энтропии в моделирование даёт ошибки на порядок большие, чем расчёт свободной энергии.

Компоненты свободной энергии

При использовании термодинамического интегрирования:

$$\Delta A = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial H(p^n, r^N}{\partial \lambda} \right\rangle_\lambda \partial\lambda$$

$$\left\langle \frac{\partial H}{\partial \lambda} \right\rangle = \left\langle \frac{\partial H_{bond}}{\partial \lambda} \right\rangle + \left\langle \frac{\partial H_{angle}}{\partial \lambda} \right\rangle \dots$$

$$\Delta A = \Delta A_{bond} + \Delta A_{angle} \quad (8.12)$$

Необходимо отметить, что только сумма компонент является осмысленной. Интересное применение этого подхода получило в исследовании связывания биотина со стрептавидином.

Подводные камни:

Существует два основных источника ошибок:

Неточность расчёта гамилтиана, ошибки в силовом поле, неправильный расчёт взаимодействий

Недостаточная выборка из фазового пространства.

К сожалению, нет рецепта для определения достаточности выборки. Сравнение прямой и обратной выборки может указывать на гистерезис. Если выборка маленькая гистерезис будет стремится к 0. Это явный признак малой выборки.

Особенности применения методов

- МК используют для малых жестких молекул.
- МД используют для крупных информационно подвижных молекул.
- Метод медленного роста почти не используют, так как считается, что молекула не успевает адаптироваться к изменению λ.
- Преимущество интегрирования и пертурбации – это возможность уточнить некий диапазон λ без пересчёта остальных значений.
- Возможно динамическое изменение λ.
- Иногда используют модифицированные потенциалы.

Потенциал средней силы (PMF)

Мы рассмотрели изменение свободной энергии при “мутации” вещества. Изменение свободной энергии вдоль какой-либо координаты (расстояние, торсионный угол и т.д.) называют потенциал средней силы (PMF). Данный процесс выглядит гораздо более “физичным”.

Опять проблема: МД и МК “не хотят уходить” из областей с низкой энергией. Для решения этой проблемы используют *umbrella sampling:*

Umbrella sampling (US) модифицирует потенциал, что позволяет эффективно исследовать области с высокой энергией. Модификация потенциала записывается как пертурбация:

$U\,(r^N)\,=\,U\,(r^N)\,-\,W\,(r^N)$, где часто: $W\,(r^N)\,=\,k_w(r^N-r_0^N)^2$

Естественно, что распределение не Больцмановское, но можно поправить:

$$\langle A\rangle = \frac{\langle A(r^N)e^{+\frac{W(r^N)}{k_BT}}\rangle_W}{\langle e^{+\frac{W(r^N)}{k_BT}}\rangle_W} \tag{8.13}$$

Пример:

Рассмотрим диссоциацию протофибриллы и одной молекулы аммилоидного пептида:

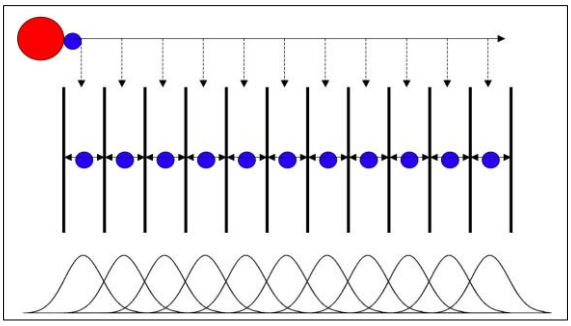


Рисунок 8.5. Схематичное изображение диссоциации протофибриллы

Смотрим как диссоциирует пептид в динамике и зафиксируем несколько состояний, как на рисунке 8.5. При запуске молекулярной динамики мы запрещаем пептиду сильно отклонятся от стартового состояния. В итоге имеет набор гистограмм с пересечением. Далее запустим:

grompp -f umbrella.mdp -c conf0.gro -p topol.top -o umbrella0.tpr

grompp -f umbrella.mdp -c conf450.gro -p topol.top -o umbrella22.tpr

Обработаем:

g_wham -it tpr-files.dat -if pullf-files.dat -o -hist -unit kCal

Итого:

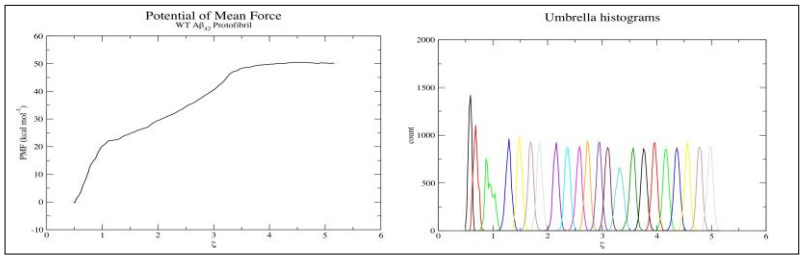


Рисунок 8.6. Пример результата расчета ΔG с помощью PMF

Очевидно, что расчёт свободной энергии — это ресурсоёмкий процесс. Один из путей, как λ-динамика, это получение информации о наборе молекул за один расчёт. Другой путь – это ограничение количества запусков для получения результата.

λ-динамика:

Основная суть — это изменение λ в ходе моделирования, причём изменение не только от одного вещества к другому, но к множеству других. Например, исследуем, как меняется λ при исследовании многих заместителей:

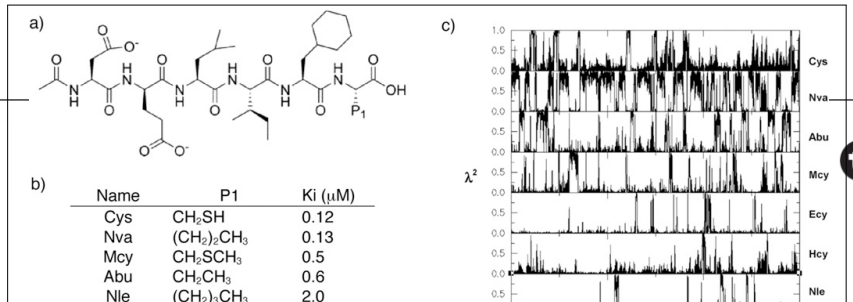


Рисунок 8.7. Пример изменения λ при исследовании многих заместителей

Linear response (LR):

Суть идеи состоит в проведении всего двух запусков: комплекса и лиганда в воде.

$$\Delta G = \beta(\langle U_{LR}^{el}\rangle - \langle U_{LS}^{el}\rangle + \alpha(\langle U_{LR}^{VdW}\rangle - \langle U_{LS}^{VdW}\rangle)$$

$$\Delta G_{hygr} = \beta\langle U_{hygr}^{el}\rangle + \alpha\langle U_{hydr}^{VdW}\rangle - \gamma SASA \quad (8.14)$$

Коэффициенты находятся либо аналитически, либо, как во втором случае, эмпирически, подгонкой.

Лекция 9. Свойства лигандов, построение лигандов, QSAR

Параметры молекул — это некие свойства в количественном описании, которые можно легко рассчитать, зная только формулу молекулы.

Примеры:

- Молекулярная масса
- Количество атомов
- Распределение вода/октан
- Электротопологические индексы
- Молярная поляризуемость
- Топологические двугранные углы

Описание молекул:

1D, брутто формула:

- Молекулярная масса
- Количество атомов

2D, структура:

- Распределение вода/октан
- Fingerprints
- Топологические двугранные углы

3D, пространственная структура:

- Дипольный момент
- Объем и поверхность молекулы
- Частичные заряды на атомах

Распределение октанол/вода

Любое распределение описывается соотношением. Ожидается, что гидрофильное вещество легко растворяется в воде (образование водородных связей и малые потери энтропии для воды), а гидрофобное в органическом растворителе.

Введем обозначение: $log\ P$, где P степень распределения вещества в системе октанол/вода. Для некоторых веществ трудно определить экспериментально. Кроме того, можно использовать метод пертурбации свободной энергии, но есть ряд неудобств.

Для решения данных неудобств используют фрагментарный подход. Распределение для молекулы равно сумме распределения компонентов с коэффициентами.

Реализация: CLOGP. Предложено разбивать молекулу по одинарным связям на фрагменты:

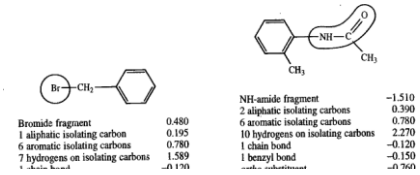


Рисунок 9.1. Пример разбивания молекул по одинарным связям

Молярная рефрактивность

Молярная рефрактивность отражает способность газа преломлять свет, что так же отражает поляризуемость молекулы:

$$MR = \frac{(n^2 - 1)MW}{(n^2 - 1)d} \quad (9.1)$$

где d - плотность, n – некий индекс, не сильно изменяющийся для органических соединений. На основании этого значения часто судят о размере молекулы и собственно плотности вещества. Применяют разбиение на фрагменты, как в CLOGP.

Топологические индексы

Хол и Кир предложили множество индексов позволяющих судить о молекуле в одной цифре:

$$\upsilon = \frac{(p_i - h_i)}{(Z_i - p_i - 1)} \quad (9.2)$$

Где p – количество s и p электронов; h – количество атомов водорода при атоме; d = p – h.

При сравнении, например, третбутилбутана и бензола, использование топологических индексов будет гораздо информативнее, чем MR. Чем больше разветвленность молекулы, тем больше топологический индекс.

Электронный эффект

Константа δ отражает влияние заместителя на смещение электронной плотности. Эффект можно увидеть при отрыве протона:

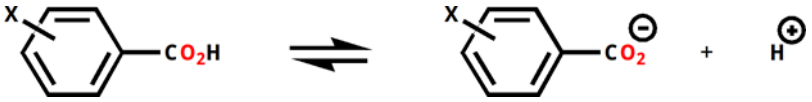


Рисунок 9.2. Диссоциация гомологов бензойной кислоты

$$K_h = K_{diss} = \frac{[PhCO_2^-]}{[PhCO_2H]}$$

Фармакофорные индексы – это попытка описать как в пространстве друг относительно друга находятся группы, которые могли бы взаимодействовать с белком. Если есть описанный каталитический центр белка, то можно предсказать фармакофор. Под этот фармакафор можно проверять все гипотетические соединения с трехмерной структурой. Но есть нюанс – лиганды могут менять трехмерную структуру под белок в более невыгодную конформацию. Как определяют фармакофорные индексы:

- Как правило выделяют три точки.
- Определяют расстояния между точками для разных конформеров.

Создание выборки

Для поиска заготовки для лекарства важно создать хорошую стартовую выборку соединений. Используют рассчитанные параметры молекул для оценки качества выборки, но надо добиться нормального распределения параметров. Кроме того, избегают высокой корреляции для исключения *перепредставленности* выборки.

Оценка подобия соединений

- Dice:

$$S_{a,b} = \frac{2\sum_i^N x_{ia}x_{ib}}{\sum_i^N x_{ia}^2 + \sum_i^N x_{ib}^2} \quad (9.3)$$

Или:

$$S_{a,b} = 2\frac{|x_a \cap x_b|}{|x_a| + |x_b|} \quad (9.3.1)$$

- Tanimoto коофицент.
- Cosine коофицент.
- Евклидово растояние.

QSAR, количественные соотношения структура/ активность

QSAR — построение математической модели для описания соотношений структура-активность. Задача предсказать свойства молекулы *in vivo*. В целом такое соотношение можно записать так:

$$v = f(p)$$

где $v -$ активность, а p $-$ свойства из структуры.

$$\log\left(\frac{1}{C}\right) = k_1 logP - k_2(logP)^2 + k_3\sigma + k_4 \quad (9.4)$$

Введем $\pi = \log\left(\frac{P_X}{P_H}\right)$:

$$\log\left(\frac{1}{C}\right) = k_1\pi - k_2\pi^2 + k_3\sigma + k_4 \quad (9.4.1)$$

На сегодняшний день существует множество модификаций этой простой формулы, которые учитывают множество параметров молекулы. Для поиска коэффициентов надо синтезировать набор соединений с разными P. Желательно равномерное распределение выбранных соединений в шкале P. Возможна также вариация соединений при разных pH и температуре.

Получение уравнения

Простейший подход – это определение зависимости активности от параметра как линейной регрессии. Тогда определение коэффициентов это просто метод наименьших квадратов. Линейная регрессия расширяется до множественной линейной регрессии, где более чем одна независимая переменная. Считается, что для статистически значимого определения необходимо не менее 5 соединений на каждый параметр. Существуют генетические алгоритмы для поиска коэффициентов регрессии. Используется *кросс-валидация, которая является р*аспространённым способом проверки. Кросс-валидация – это выборочное удаление данных из выборки и сравнение результатов.

Использование QSAR

- Главная задача – это предсказать оптимальную структуру
- Часто бывает, что QSAR хорошо работает при интерполяции, а не при экстраполяции.
- Бывает необходимо использовать не линейную зависимость, применяют параболическую.
- Существует билинейная модель:

$$log(\frac{1}{C}) = k_1 logP - k_2(log(\beta P + 1)) + k_3$$

- Отсутствие корреляций для параметра скорее всего означает, что этот параметр не задействован в механизме.

ADME

Кроме того, данный метод позволяет оценить не только взаимодействие с белком, но и другие свойства: Absorption, Distribution, Metabolism, Elimination (“Fail early, fail fast, fail cheap”).

Фрагментарное построение лиганда:

- Сканирование по базам данных удобно, так как позволяет тут же проверить молекулу.
- Сканирование не предполагает создание лиганда de novo.
- Суть фрагментарного построения лиганда состоит в поиске мест в активном центре белка, где связываются небольшие фрагменты молекул
- Соединяя фрагменты связями при сохранении места связывания, добиваются высоких констант.

Графическое отображение данного процесса выглядит так:

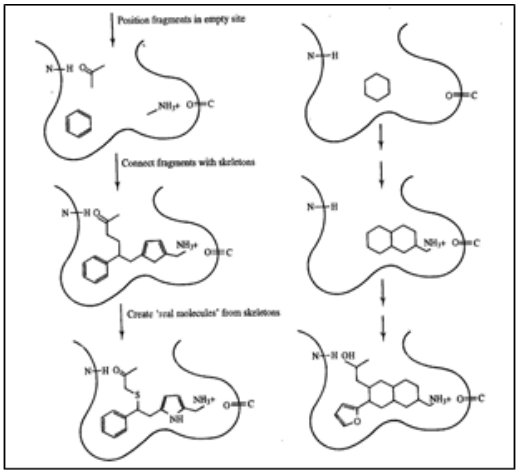


Рисунок 9.3. Реализация фрагментарного построения лиганда.

Реализация:

- GRID аналог докинга.
- MCSS сайт наполняется фрагментами и с помощью EM вычленяется место, где фрагмент наиболее предпочтителен. Взаимодействия между фрагментами не учитываются.

- LUDI использует информацию из банка PDB для задания фрагментов образующих водородные связи и т.д.
- Необязательно всё моделировать, можно использовать PCA и ЯМР для определения места связывания фрагмента.
- Если у Вас есть два и более фрагмента, то можно искать способ их соединения по базам данных.
- Реализовано в CAVEAT.
- Можно строить автоматически строить скелеты. Главный критерий — это сохранение взаимного положения фрагментов.
- Переход от скелета к молекуле сложен, так как надо реализовать возможность синтеза молекулы.

Лекция 10. Предсказание 3D структуры белков.

Область предсказания 3D структуры белков имеет бурный рост, связанный с методами машинного обучения, которые позволяют находить новые способы более точно предсказывать структуры белков.

Основные проблемы:

- Монте-Карло: 100 а.о. 3N степеней свободы, получаем 1048 конформаций.
- Парадокс Левинталя: Промежуток времени, за который полипептид приходит к своему скрученному состоянию, на много порядков меньше, чем если бы полипептид просто перебирал все возможные конфигурации. Причины парадокса Левинталя:

1. Теоретические модели, не соответствуют тому, что природа старается оптимизировать;
2. В ходе эволюции были отобраны только те белки, которые легко сворачиваются;
3. Белки могут сворачиваться разными путями, не обязательно следуя глобально оптимальному пути.
4. Считается, что структура определяется последовательностью, но иногда нужны другие факторы.
5. Структура более консервативна чем последовательность

Для решения разумно использовать накопленные знания для моделирования.

Сравнительное моделирование

Зачем искать конформации, если можно представить, что при подобии последовательностей подобны и структуры. Надо оценить, насколько вероятно, что отличие в последовательности может привести изменению способа укладки цепи. Надо отфильтровать ошибки полученные при определении структуры.

Известные структуры и последовательности. Сейчас известно порядка 105 структур. Примерно 10% это уникальные белки. Только 30% из первого пункта имеют разрешение лучше 3.0 ангстрем. Примерно 25% известных последовательностей можно использовать для сравнительного моделирования. Для 50% последовательностей можно предсказать способ укладки.

Степень идентичности и сравнительное моделирование

Встает вопрос насколько сильно должна быть похожа последовательность, для которой хотим построить структуру, на ту последовательность белка, структура которого нам известна. Основная цель – сравнить последовательности и на этом основании этого сравнения и 3D структуры, предположить структуру белка. Оказывается, что если наши последовательности совпадают от 60 до 100%, то вероятность того, что выбранные белки крайне схожи, велика. Результаты гомологичного моделирования можно использовать вплоть до разработки новых лекарств, улучшения экспериментальных данных и т.д. Если последовательности совпадают от 30 до 60 %, тогда есть вероятность, что у двух структур довольно похожие укладки:

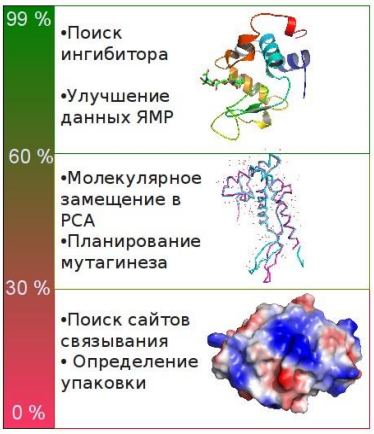


Рисунок 10.1. Использование сравнительного моделирования

Как это реализовать?

1. Надо найти белок заготовку с известной структурой.
2. Построить первичное выравнивание.
3. Улучшить выравнивание.
4. Построить ход основной цепи.
5. Моделирование петель.
6. Достроить/моделировать положение боковых радикалов.
7. Проверка модели.

Поиск белка заготовки:

- Поиск по PDB с помощью:
  1. Blast
  2. Psi-Blast
  3. Методов распознавания упаковки
- Используя биологическую информацию.
- Функциональное аннотирование в базах данных.
- Используя информацию об активных сайтах, или мотивы.

Улучшение выравнивания, пример:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL	CYS
PHE	ASN	VAL	CYS	ARG	THR	PRO	---	---	---	GLU	ALA	ILE	CYS
PHE	ASN	VAL	CYS	ARG	---	---	---	THR	PRO	GLU	ALA	ILE	CYS

Таблица 10.1. Выравнивание структур

В таблице 10.1. представлены выравнивание двух структур. Белок заготовка представлен в первой строке. В двух других показано различное выравнивание одной структуры относительно белка заготовки. В данном примере наблюдается делеция в 3 аминокислотных остатка. И данная делеция может происходить разными способами. В первом случае данная делеция приводит к большому расстоянию между C_α – атомами,

а во втором случае C_α −атомы находятся на расстоянии 1 аминокислотного остатка. Очевидно, что второй способ выравнивания гораздо лучше. Оптимизация выравнивания под известную структуру белка можно легко и непринуждённо. Замечание: с делециями все просто, но что делать, если присутствуют вставки? В таком случае нельзя грамотно распределить аминокислотные остатки и нужно внимательно использовать средства моделирования структуры.

Качество белка заготовки

- Выбор качественного белка заготовки очень важен.
- Лучший вариант не обязательно обладает лучшей степенью идентичности:

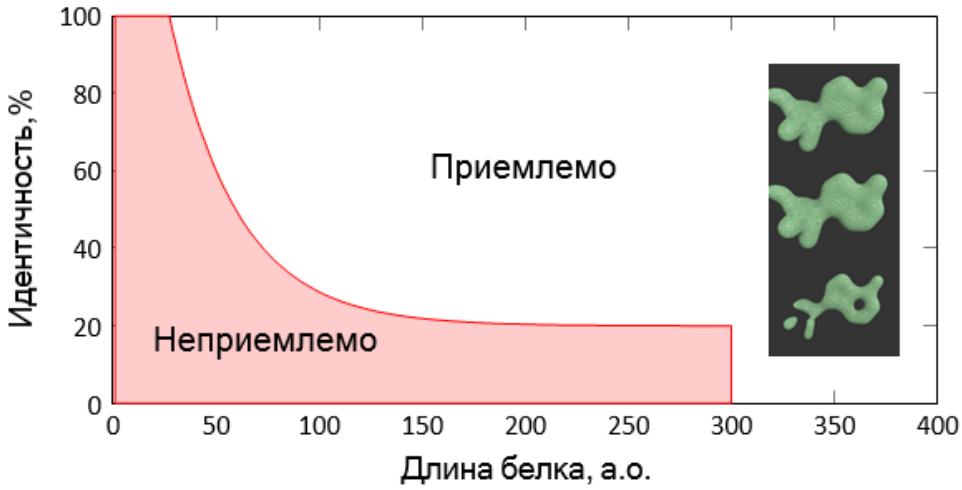


Рисунок 10.2. Критерии выбора белка-заготовки

Пример:

Белок 1: ID 93%, 3.5 ангстрема разрешение. Хуже.

Белок 2: ID 90%, 1.5 ангстрема разрешение. Лучше!

Отдельно отметим структуры, полученные методом ЯМР. Если структура белка заготовки получена ЯМР, то можно получить динамику белка в растворе (отличие от РСА). Как работает выравнивание с использованием структуры ЯМР:

- Определимся какие области определены лучше.
- Соотнесём с выравниванием.
- Если низкая гомология выпадает на “подвижные” области, то структура подходит.

Хорошим способом проверки качества заготовки является сравнение с картами Рамачандран:

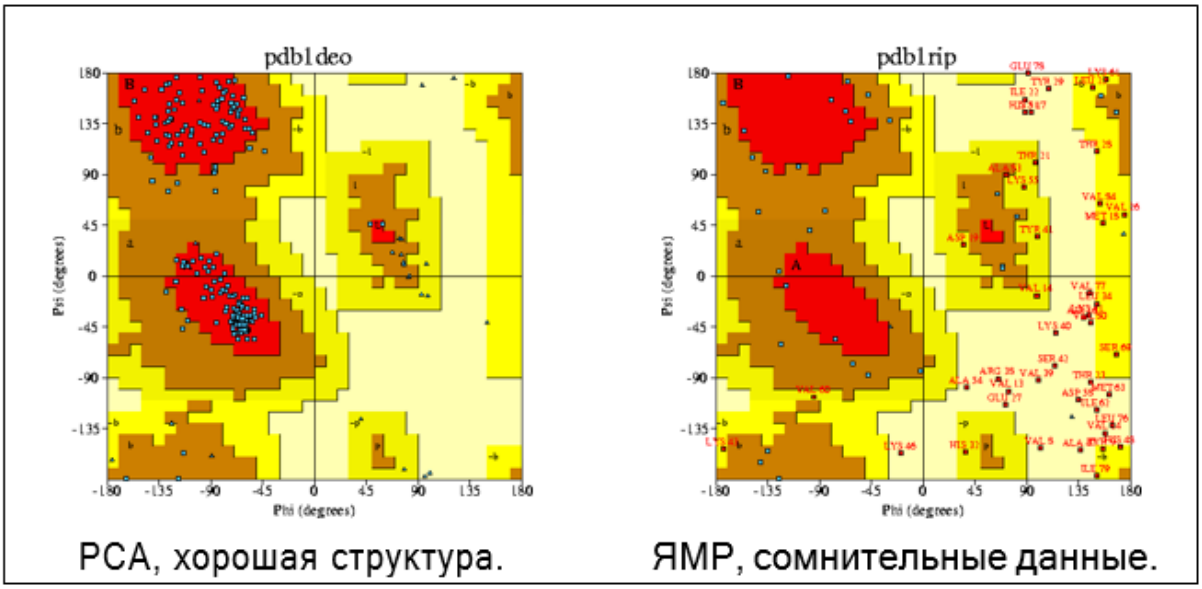


Рисунок 10.3. Сравнение карт Рамачандран для структур полученных двумя разными методами.

Построение остова:

1. Генерируем координаты остова моделируемого белка для остатков из выравненных областей.

2. Не обязательно использовать координаты, могут подойти дистанционные ограничения.

3. Большинство исследователей предпочитают Modeller. Modeller использует дистанционные ограничения.

Моделирование петель:

Эмпирическое моделирование:

1. Поиск подходящего фрагмента по PDB.
2. Использовать базы данных (LIP, etc..).
3. Молекулярная механика.
4. Монте-Карло.

Rosseta:

1. Поиск фрагментов близких по последовательности.
2. Комбинирование результатов поиска с помощью Монте-Карло.

Кроме того, можно использовать комбинации вышеперечисленных.

Если идентичность последовательностей высока, то можно ожидать высокую консервативность третичных контактов. Если же анализ показывает, что важные контакты консервативны то:

- Лучше оставить конформацию боковых радикалов из заготовки чем моделировать.
- Конформация боковых радикалов зависит от конформации основной цепи.
- Существуют базы данных ротамеров.

- Некоторые исследователи считают, что SCWRL метод самый удачный. Это эмпирический метод на основе теории графов. http://dunbrack.fccc.edu/SCWRL3.php

Точность моделирования боковых радикалов:

- Высокая точность моделирования достигается для боковых радикалов внутри глобулы.
- Причина: в экспериментах остатки на поверхности более подвижны.
- Вычислительное проще упаковать гидрофобные остатки, чем учесть полярные контакты и водородные связи с водой или с участием воды.

Улучшение модели:

- Методы минимизации энергии.
- Моделирование молекулярной динамики (оптимизация гидрофобики).
- Моделирование Монте-Карло.
- Любой известный подход для оптимизации структуры.

Ошибки:

- Обычно ошибки не исправляются на последующих этапах моделирования.
- Хорошее выравнивание не исправит плохой выбор белка заготовки.
- Хорошее моделирование петель не исправит плохое выравнивание.
- При обнаружении ошибки необходимо повторять некоторые этапы.

Проверка:

- Большинство программ для моделирования по гомологии выдают правильные значения для связей и валентных углов.
- Карта Рамачандрана в большинстве случаев для модели выглядит также, как для белка-заготовки
- Проверка на ориентацию или положение заряженных остатков может быть полезна.
- Использование любых экспериментальных данных:
- Остатки активного центра.
- Места модификаций.
- Места контактов.

ProQ сервер оптимизирован на поиск правильной модели, а не нативной структуры.

Ресурсы для гомологичного моделирования

1. Modeller
2. SwissModel
3. Eva-CM
4. Nest И т.д.

Предсказание структуры белка *Ab initio*

- Теоретически можно использовать молекулярную динамику.
- Моделирование отжига, как в МД, так и в Монте-Карло.
- На основе фрагментов, *Rosseta*

Ab initio, Rosseta

Метод использует информацию о предсказании вторичной структуры. Сравниваем фрагменты от 3 до 9 остатков с библиотекой известных структур. Строим эти фрагменты. Соединяем эти фрагменты и используем Монте-Карло для оптимизации третичной структуры. Для определения хорошей конформации использую специальные потенциалы, которые делают модель похожей на нативную. Что можно использовать:

- Потенциалы для третичных контактов
- Гидрофобные потенциалы
- Потенциал для уменьшения радиуса вращения молекулы
- Водородные связи и т.д.
- Можно добавить информацию об дисульфидных мостиках, местах связывания катионов металлов и т.д.

Далее сравниваем последовательность со всеми известными способами укладки. Используем потенциалы для определения тенденций в известных способах укладки. Каждую аминокислоту из модели помещаем в позиции белков разных укладок. Определяем, как хорошо эта аминокислота подходит белку заготовке на основе парных взаимодействий. Но основе суммарного результата определяем белок-заготовку.

Threading —протягивание нити

Сравниванием последовательность со всеми известными способами укладки. Используем потенциалы для определения тенденций в известных способах укладки. Каждый аминокислотный остаток из модели помещаем в позиции белков разных укладок. Определяем, как хорошо эта аминокислота, подходит белку-заготовке на основе парных взаимодействий. На основе суммарного результата определяем белок-заготовку:

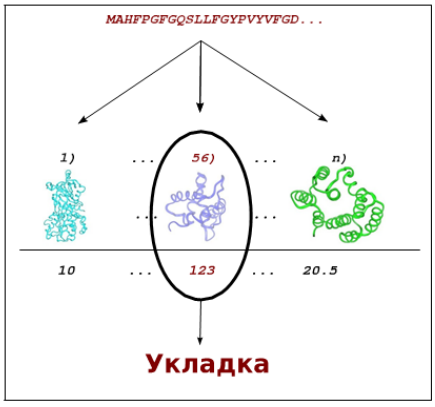


Рисунок 10.4. Сканирование способов укладки белков при помощи Threading.

Основные ошибки:

1. Взаимодействия в белке не всегда описываются парными контактами.
2. Потенциалы часто основываются на профилях последовательностей.

Есть гибридные методы Rosseta/Threading: I-Tasser.

Распознавание укладки, Phyre2

Еще один вариант работы со специфичными весовыми матрицами – распознавание укладки с помощью Phyre2. На основании PSI-Blast находим последовательности, которые данному белку родственны. Из этого можно построить HMM – профиль. Далее сравниваем данные профили с уже известными белками. Данный метод позволяет найти далекие гомологи и предсказывать хорошие модели при идентичности менее 15%:

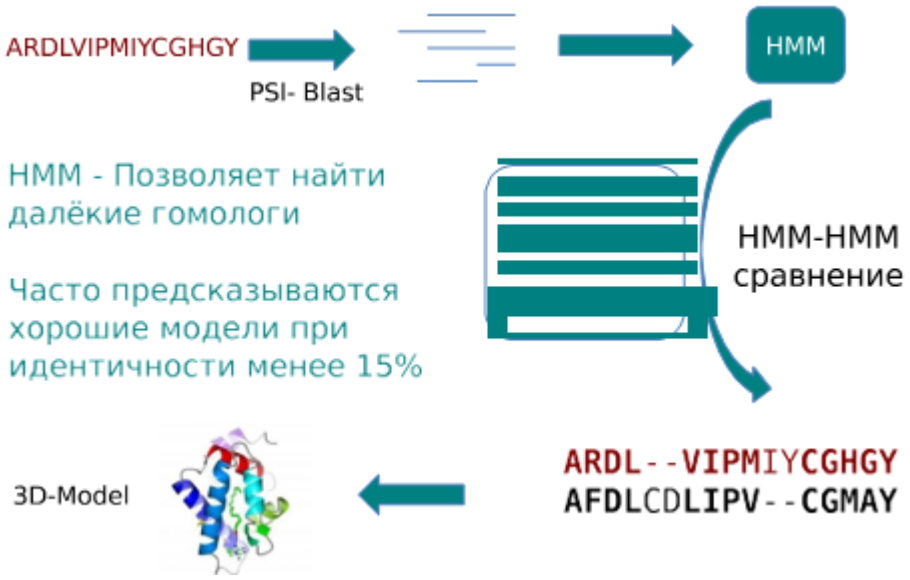


Рисунок 10.5. Схематичное изображение протокола Phyre2

Мета серверы:

1. Сравнение разных методов.
2. Большинство методов предсказывают правильную укладку в первых 10-20 результатах.
3. Удаление структур с высоким значением параметров модели, но с единственной укладкой.
4. Суперпозиция результатов, взвешивание.
5. Часто выдают только позиции атомов остова.

Заключение

- Суть современного моделирования белков – эмпирическая.
- Чем больше известной информации используется при моделировании, тем точнее модель.
- Каждый метод имеет недостатки.
- Критический анализ модели позволяет выявить ошибки и улучшить модель.

Лекция 11. Поиск новых биоактивных молекул и докинг

Докинг белок-лиганд

Докинг – самая распространенная методика по оценке взаимодействий низкомолекулярных лигандов с белками и поиску конформации лигандов в комплексе с белками. Существуют подходы на основе молекулярной динамики, метода Монте-Карло и т.д.

Наиболее используемый вариант – метод Монте-Карло в решеточном потенциале. Решеточный потенциал – попытка разбить пространство, описывающее активный сайт белка, на решетку. В узлах решетки находится информация о том, какого типа остатки белка присутствуют рядом с узлами. В ходе расчета выставляются *бинарные очки,* например, на основе того, есть там водородные связи или нет. Шаг решетки часто составляет порядка 0.5 Å. Метод является очень быстрым и в ходе итераций Монте-Карло чем больше очков набирается, тем лучше. Важно отметить, что такой принцип построения решетки *запрещает движение белка*. В результате мы можем узнать положение лиганда в комплексе с белком и оценить их константу связывания.

Естественно, методология не подразумевает расчет пути попадания лиганда в белок. Современные мощности позволяют оценить связывание лиганда со всей поверхностью белка, от размера которого зависит сложность и длительность расчета. Для этого поверхность белка разбивают на взаимоперекрывающиеся ячейки и вычисления производят итеративно. В рамках модели Монте-Карло можно учитывать или не учитывать подвижность атомов лиганда.

Сайт связывания – место связывания лиганда.

Геометрия связывания – место связывания, ориентация и конформация лиганда.

Основные цели докинга:

- виртуальный поиск лигандов.
- определение геометрии связывания лиганда.

Если известно, как связывается лиганд, то:

- возможно узнать, какие части важны для связывания.
- можно предложить изменения для улучшения константы связывания.
- можем избежать ошибок.

Виртуальный поиск лигандов представляет собой комплексную задачу, поскольку рассматривают миллионы различных лигандов по отношению к одному сайту связывания, при этом не учитывают возможность его расширения. Обычно после подбора малых молекул, потенциально связывающихся в сайте, проводят контрольный эксперимент, в ходе которого рассматривают *связывание лиганда с остальной поверхностью белка.*

В свою очередь, определение геометрии связывания конкретного лиганда с белком является более цельной задачей и дает результаты, хорошо согласующиеся с реальностью. На их основе можем предложить мутантные варианты белка для доказательства наличия конкретного связывания в предложенном сайте.

Два основных компонента программ для докинга:

- алгоритм поиска
    - установление места связывания
    - установление геометрии связывания
- алгоритм расчета константы связывания областей с низкой энергией.

На данный момент существует большое количество программ для реализации докинга белок-лиганд:

- AutoDock, DOCK, e-Hits, FlexX, FRED, Glide, GOLD, LigandFit, QXP, Surflex-Dock и т.д.
- разные алгоритмы оценки аффинности и разные алгоритмы поиска
- важно не путать докинг лиганд-белок и докинг белок-белок

На данный момент существуют программы для докинга (Plans и др.) на основе алгоритма, подсмотренных у муравьев. В отличие от классических программ, приведенных в списке выше, вместо кубической ячейки используется *сферическая*, а также *иным способом трактуются атомы водорода*. Другое отличие состоит в том, что семейство AutoDock направлено на эффективный скрининг малых молекул, тогда как программы, аналогичные Plans, используют для анализа взаимодействии одиночных лигандов, в т.ч. пептидов.

При расчете в определенных программах не следует брать молекулы с большим количеством вращательных связей, поскольку это сильно увеличивает вариативность вычислений. Это может стать преградой при расчете, например, пептидов, состоящих более, чем из 4 аминокислотных остатков.

Практические аспекты докинга белок-лиганд:

- часто PDB-структура содержит молекулы воды и прочие вспомогательные частицы, почти всегда их надо убирать
- надо добавлять протоны к структуре
- часто в PDB неточно определена ориентация некоторых групп, что сказывается на паттерне водородных связей
- протонирование лиганда и его таутомерные формы.

Добавление протонов в структуре может вызвать проблемы в случае наличия *гистидинов в сайте связывания* – необходимо провести расчет как с протонированных, так и с депротонированным состоянием аминокислотного остатка. Если в сайте несколько остатков гистидина, придется осуществить комбинаторный перебор всех состояний всех способов протонирования. Также стоит учитывать то, что *аминогруппы* чаще всего *протонированы*, а *карбоксильные группы – депротонированы*.

Сложности могут возникнуть при расчете *амидной группы* $-C(=O)-NH_2$. Что касается рентгеноструктурного анализа, атом кислорода и аминогруппа имеют одинаковую массу и практически неразличимы данным методом. Итого, для докинга придется рассматривать разную ориентацию данной атомной группировки.

Rigid и Flexible докинг.

- Rigid: лиганд не имеет внутренних степеней свободы, т.е. вращение вокруг связей запрещено

- Flexible: предполагает учет вращения вокруг связей лиганда
- часто белок рассматривается как жесткое тело

Некоторые формы (конформеры) лиганда могут быть неэффективными для связывания белком, что нельзя учесть при расчете, использующем стандартный метод Монте-Карло. Определяют наиболее вероятные конформации лиганда, с помощью которых проводят твердотельный докинг – по отдельности проводят вычисления со всеми жестко зафиксированными конформерами.

В эукариотических клетках белки часто выполняют множественные функции, в большинстве своем заключающимися во взаимодействии с другими белками. Существуют карты, представляющие собой графы, отображающие взаимодействия белков в клетке – интерактом:

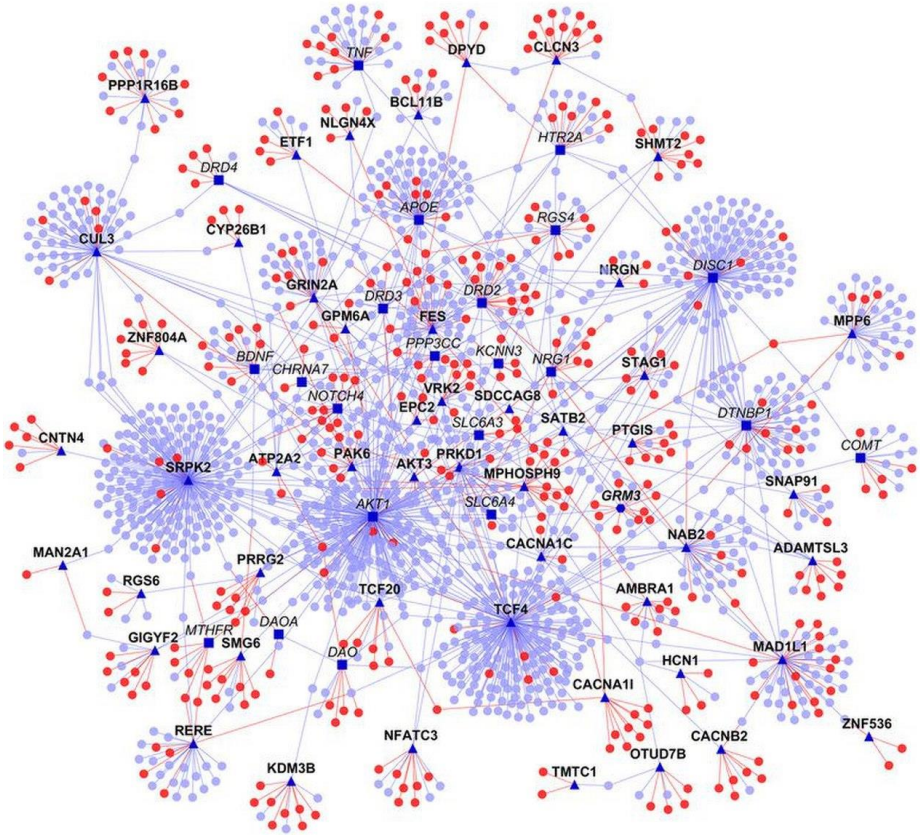


Рисунок 11.1 - Пример интерактома.

Способы предсказания белок-белковых взаимодействий:

- филогенетический профайлинг – поиск пар белковых семейств среди широкого ряда видов (появление и исчезновение пар семейств возможно указывает на взаимодействие)
- предсказание на основе подобия филогенетических деревьев
- методы на основе классификации
- поиск гомологичных мест контакта
- ассоциативные методы – поиск характеристических последовательностей на основе профилей и мотивов

- идентификация структурных паттернов на основе известных структурных данных, построение библиотеки и сканирование по ней
- методы Байеса для анализа экспериментальных результатов с значимым уровнем шума
- методы исключения доменных пар
- моделирование структуры комплекса на основе известной структуры и оценка его качества
- макромолекулярный докинг

Базы данных:

- String – база данных экспериментальных и предсказанных взаимодействий
- IntAct – база данных на основе литературных данных или прямая информация от авторов
- iHOP – информация, слинкованная с другими белками, построена на основе литературных данных, представление в виде фрагментов текста
- BioGRID – в качестве источников используются научные публикации и результаты high-throughput экспериментов
- MIPS – набор отобранных вручную данных о белок-белковых взаимодействиях, собранных из научной литературы (индивидуальных экспериментов).

Макромолекулярный докинг

Поиск наименьшего ΔG. Целью докинга является поиск наименьшего изменения энергии Гиббса реакции комплексообразования между белком и лигандом (в т.ч. с другим белком), что соответствует наиболее энергетически выгодному взаимодействию. Белок представляется неподвижным, а лиганд – подвижной частицей. В сферических координатах производится поиск наиболее оптимальных позиций одного белка относительно другого. Варианты расположения, получаемые в результате расчета, далее ранжируются. Суть метода основывается на поиске соответствия поверхностей для достижения максимальной поверхности контакта.

Рисунок 11.2 - Получение различных вариантов комплексов в ходе докинга (правильный вариант с наибольшей площадью взаимодействия помечен звездочкой).

Самый эффективный метод оптимизации макромолекулярного докинга – методы FFT (Fast Fourier Transform). Для их использования белки описываются в виде решеточных моделей, далее и использованием этих преобразований ищем оптимальное расположение решеток друг относительно друга:

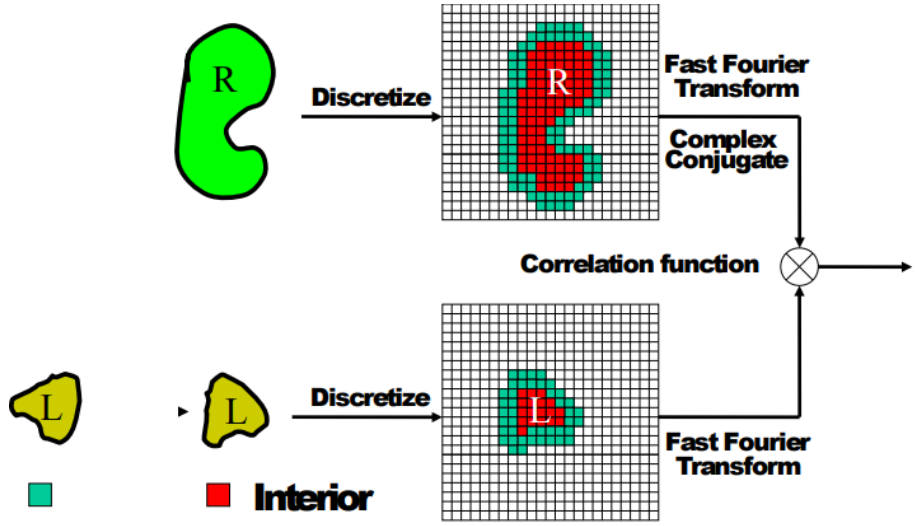


Рисунок 11.3 – Оптимизация макромолекулярного докинга с помощью методов FFT.

Оценка производительности состоит из двух компонентов:

- Success Rate: для некоторого количества предсказаний (N_p) данная величина является долей структур (в %), для которых был найден как минимум один удачный результат
- Hit Count: среднее количество хитов при данном значении N_p.

В сферических координатах ключевой способ изменения конфигурации белка – изменения угла. Значение параметра Success Rate не зависит от выбранного шага вращения:

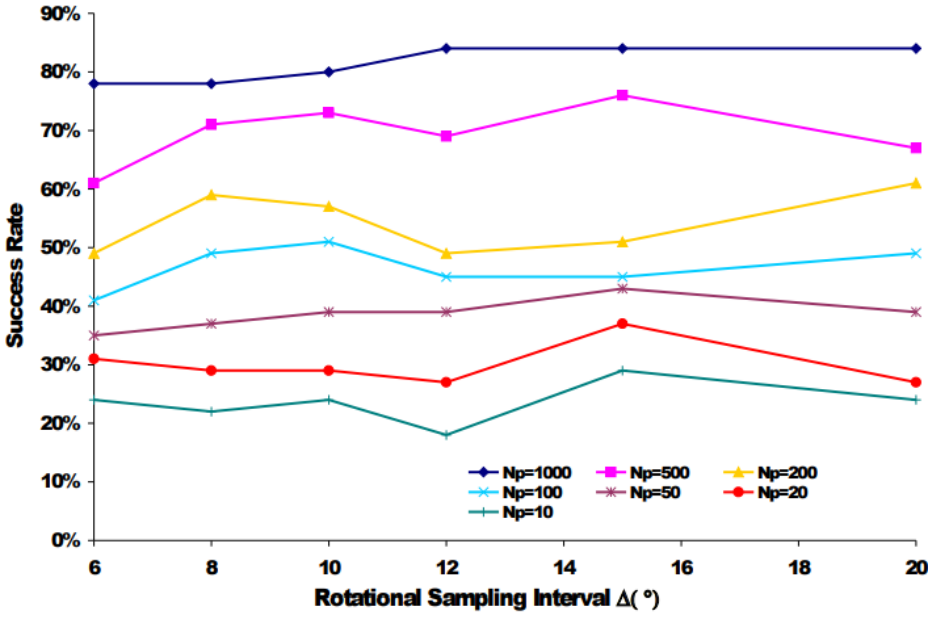


Рисунок 11.4 - Зависимость параметра Success Rate от шага вращения.

Как видно из Рис. 11.5, доля удачных предсказаний зависит от количества запуска алгоритма, однако правильные структуры еще нужно уметь находить.

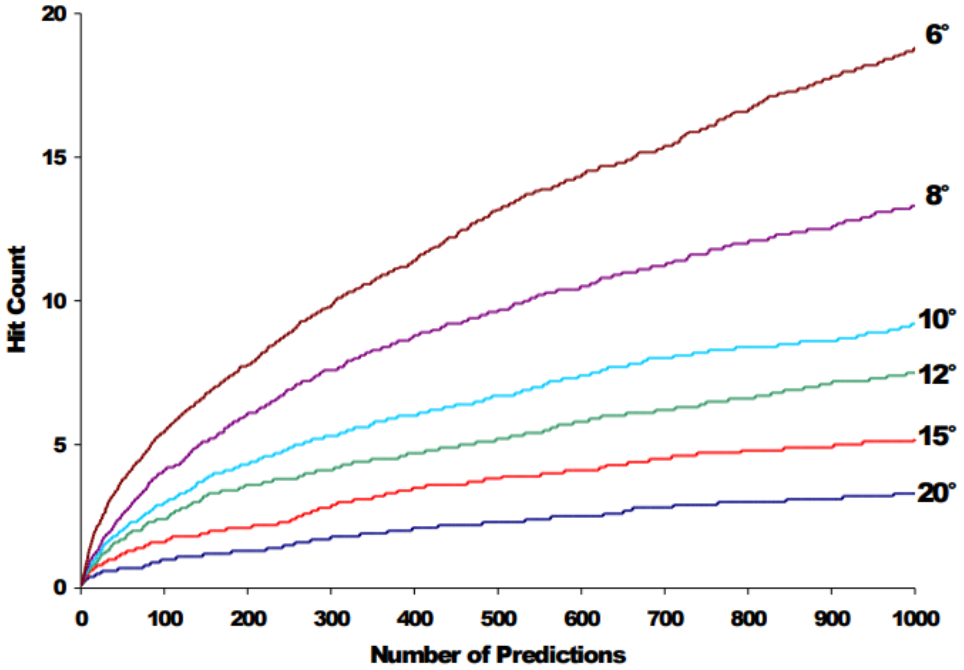


Рисунок 11.5 - Зависимость параметра Hit Count от количества предсказаний.

При взаимодействии белков ожидается, что остатки на поверхности изменяют свою конформацию для обеспечения более эффективного взаимодействия. Поверхность белка представляют в виде решетки, в узлах которых содержатся «флажки». Пересечение серых флажков нежелательно, поскольку в данном случае один белок имеет избыточное перекрывание с другим (белок «въезжает» в другой). Перекрытие флажков «1» означает высокую комплементарность поверхности:

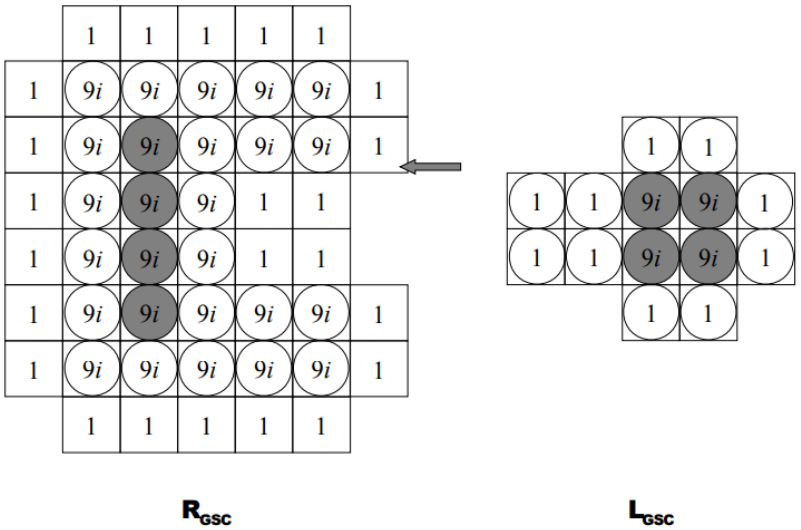


Рисунок 11.6 - Схема метода решеточной комплементарности поверхности.

Улучшенная модель парной комплементарности поверхности учитывает не только совпадения узлов решетки, но и совпадения по соседним узлам решетки (Рис. 11.7).

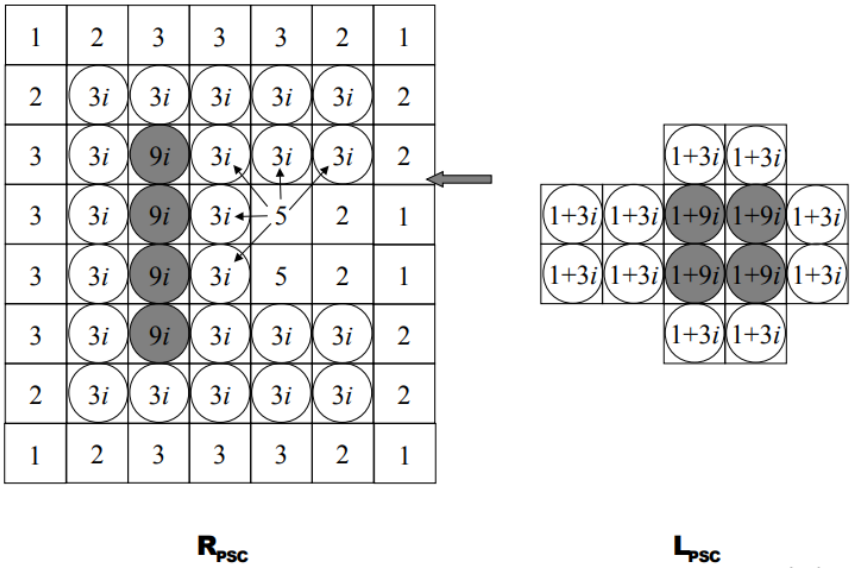


Рисунок 11.7 - Схема метода парной комплементарности поверхности.

Сравнение эффективности двух подходов показывает, что при любом количестве подходов метод парной комплементарной поверхности (PSC) оказывается эффективнее метода решеточной комплементарности поверхности (Рис. 11.8). Важно подчеркнуть, что успешность методов прямо пропорциональна количеству запусков (предсказаний).

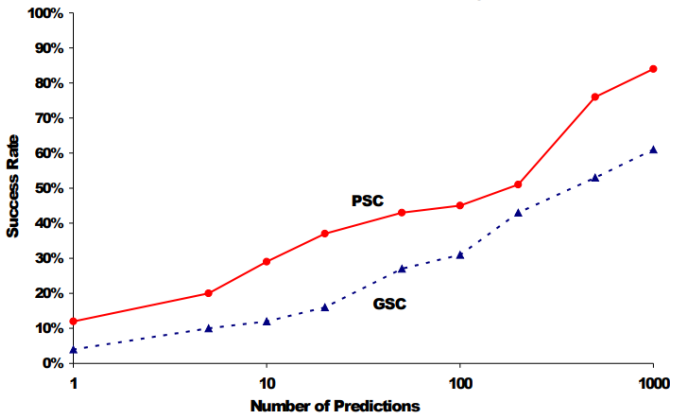


Рисунок 11.8 - Зависимость параметра Success Rate от количества запусков (предсказаний) алгоритма.

Обратите внимание, что параметр Success Rate не достигает 100%, поскольку некоторые белки в ходе образования макромолекулярных комплексов претерпевают слишком сильные изменения конформации, либо детали взаимодействия белок-белок иногда трудно описать с помощью рассматриваемых подходов.

Далее варианты относительного расположения необходимо отсортировать с помощью функции оценки энергии связи:

$$\Delta G = \Delta E_{VdW} + \Delta E_{el.} + \Delta G_{desol} + \Delta G_{const} \quad (11.1)$$

- ΔE_{VdW} – комплементарность поверхности
- ΔG_{desol} –гидрофобные взаимодействия
- $\Delta E_{el.}$ – это электростатические взаимодействия
- ΔG_{const} – изменение вращательной и прочих энтропий.

$$\Delta G_{desol} = \sum_i \sum_j N_{ij} \Delta G_{ij} \quad (11.2)$$

Из нижеприведенного графика видно, что наибольшее влияние на расчет обеспечивает учет электростатических взаимодействий:

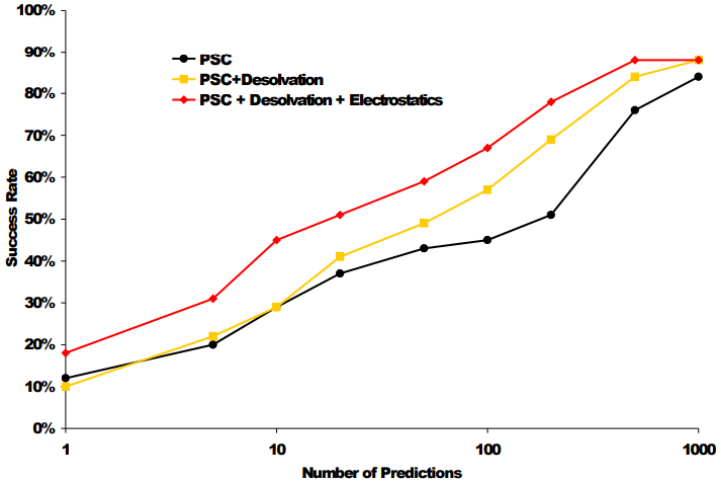


Рисунок 11.9 - Влияние учета гидрофобных и электростатических взаимодействий на величину параметра Success Rate.

На этапе накопления правильных ответов учет гидрофобных взаимодействий имеет больший смысл. Если необходимо найти хотя бы один правильный ответ, то превалирует значение электростатики.

Алгоритм Rosetta

Разрабатываются альтернативные алгоритмы докинга. Алгоритм RosettaDock (Рис. 11.10) представляет собой фреймворк, работающий на основе метода Монте-Карло с белками и низко-/высокомолекулярными лигандами. Большинство библиотек для него написано на языке C++ и метод требует больших вычислительных мощностей.

В начале белки друг относительно друг друга располагаются случайным образом, далее они итеративно изменяют свое положение и параллельно оценивается энергия их взаимодействия.

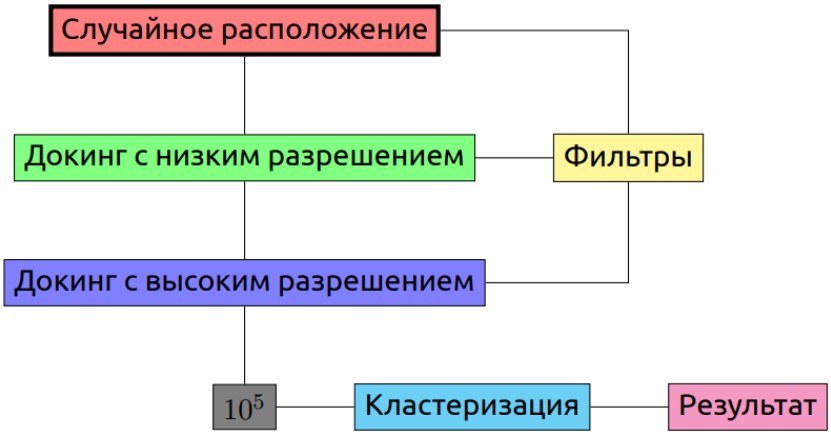


Рисунок 11.10 - Этапы алгоритма RosettaDock.

Особенности метода:

- поиск методом Монте-Карло
- вращение и смещение белка как жесткого тела
- остаток белка представляется как атомы остова и средний центроид (представление боковой цепи белка в виде одной частицы) представляет боковой радикал
- процедура старается воспроизвести физическую диффузию.

После докинга с низким разрешением переходят к оптимизации белок-белкового контакта, которое включает в себя уточнение боковых радикалов, движение остова обоих белков и т.д.

Особенности уточнения с высоким разрешением:

- из библиотеки ротамеров добавляются полноатомные боковые цепи
- используется полноценная оценка энергии (ММ)
- Монте-Карло + оптимизация геометрии
- циклическое использование оптимизации положения как твердого тела и полноатомная оптимизация положения боковых радикалов.

Итого, алгоритм RosettaDock часто оказывается эффективнее классического докинга, но требует огромных затрат вычислительных мощностей и большого опыта специалиста.

Применение экспериментальных данных (полученные с помощью ЯМР-спектроскопии, мутагенеза, SAXS и т.д.) позволяет сильно уменьшить количество получаемых вариантов относительного расположения белков. Примером реализации концепция интегративного моделирования является метод HADDOCK (High-Ambiguity Driven Docking).

Лекция 12. Структура нуклеиновых кислот и хроматин

Нуклеиновые кислоты

Нуклеиновые кислоты (НК) – это высокомолекулярные линейные полярные биополимеры, полинуклеотиды, которые построены из нуклеотидных остатков. Два основных типа НК в клетке:

- дезоксирибонуклеиновая кислота – ДНК
- рибонуклеиновая кислота – РНК

Расшифровка аббревиатуры ДНК:

- ДНК – линейный сополимер ортофосфорной кислоты и дезоксирибозы.
- ДНК – открытие и выделение «нуклеина» из ядер (нуклеус) лейкоцитов Ф. Мишером в 1869 году.
- ДНК – линейный сополимер на основе ортофосфорной кислоты.

В составе циклической формы дезоксирибозы присутствуют две гидроксильные группы, две из которых идут на образование линейного сополимера с трехосновной ортофосфорной кислотой, при этом образуется сахарофосфатный остов:

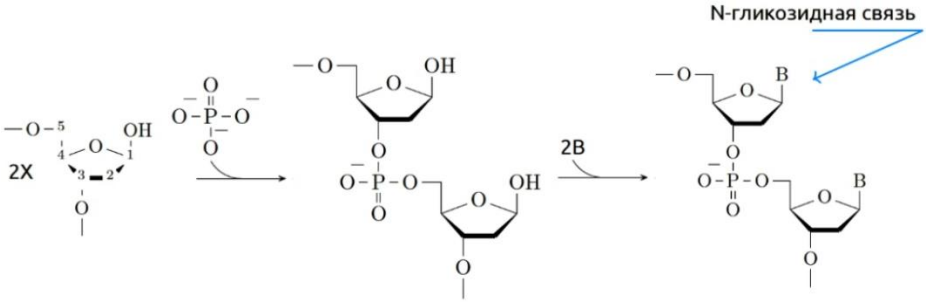


Рисунок 12.1 - Схема образования молекулы ДНК.

Третья гидроксильная группа в ходе образования ДНК заменяется на гетероциклическое (азотистое основание). Азотистые основания делятся на два типа – пуриновые (бициклические) и пиримидиновые (содержат один цикл) (Рис. 12.2), гетероатомы участвуют в создании плоской ароматической системы. Единственная неплоская группа, метильная (CH_3), входит в состав тимина, что влияет на взаимодействие оснований друг с другом.

Рисунок 12.2 - Пуриновые и пиримидиновые гетероциклические основания.

N-гликозидная связь – связь углеводного фрагмента с атомом азота гетероциклического основания.

Нуклеотид в составе ДНК имеет следующее строение:

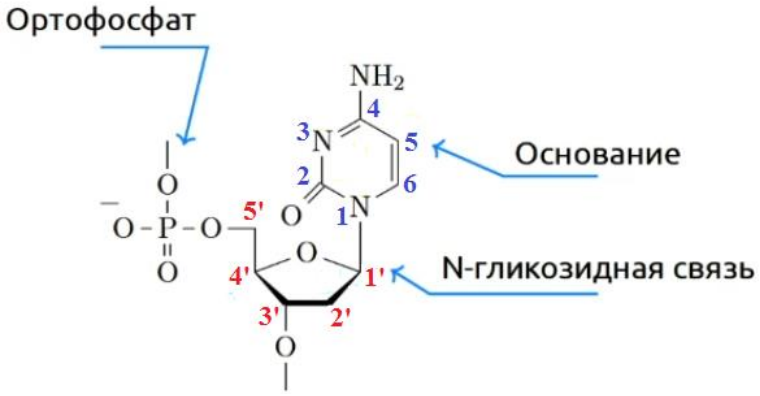


Рисунок 12.3 - Строение и нумерация атомов дезоксирибонуклеотида.

Атомы в углеводном фрагменте, в отличие от атомов азотистых оснований, нумеруются со штрихом.

В линейной цепи ДНК выделяют 5’- и 3’-конец (Рис. 12.4). В природе синтез ДНК (репликация) идет от 5’- к 3’-концу. Интересно отметить, что химический синтез осуществляется в обратном направлении, но строение ДНК от этого не меняется и полученную цепь читают от 5’- к 3’-концу.

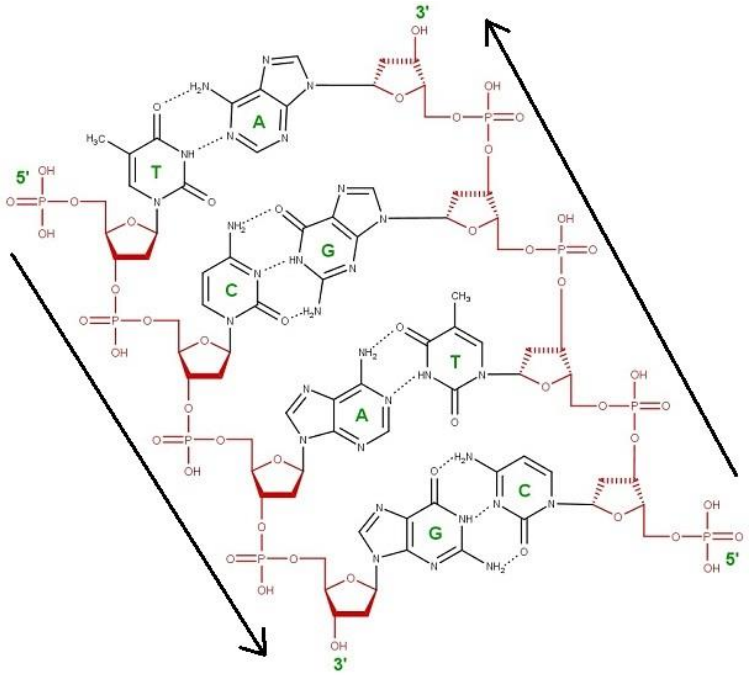


Рисунок 12.4 - Антипараллельная структура двойной спирали ДНК.

ДНК в природе существует в виде двойной спирали, состоящую из двух независимых цепей ДНК, соединенных водородными связями (в паре A=T реализуются две водородные связи, а в паре G≡C – три).

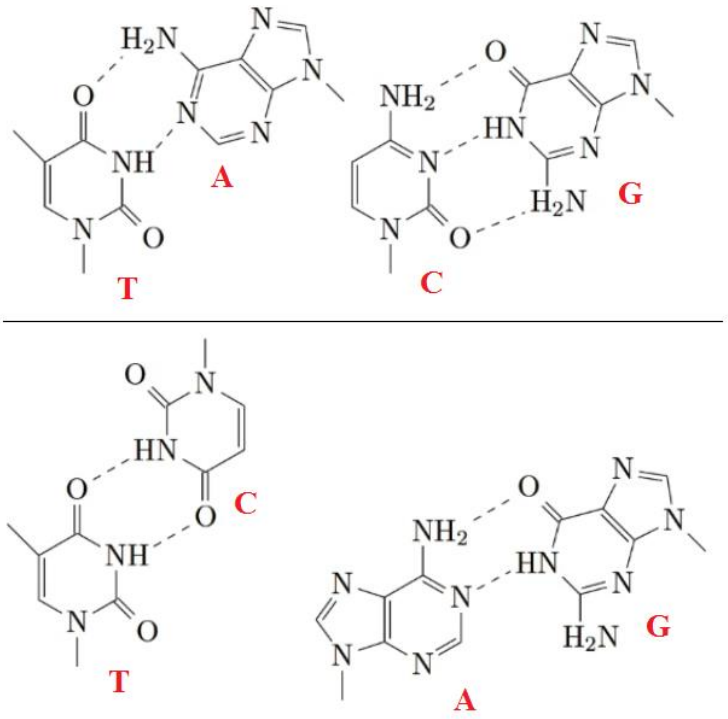


Рисунок 12.5 - Структура канонических AT- и GC-пар (сверху) и пример неканонических CT- и AG-пар (снизу).

Канонические взаимодействия (уотсон-криковские водородные связи) осуществляются в копланарном положении нуклеотидов (оба нуклеотида расположены в одной плоскости).

Структура ДНК

Структура ДНК впервые была предложена Уотсоном и Криком в 1953 году на основе результатов рентгеноструктурного анализа низкого разрешения.

Основные свойства ДНК:

- две антипараллельные цепи (первая цепь 5’-3’, вторая – 3’-5’) (Рис. 12.4)
- ДНК – двойная спираль
- имеет две оси симметрии.

Два типа взаимодействий гетероциклических оснований в ДНК:

- копланарные взаимодействия (в одной плоскости).
- стопочные взаимодействия (стекинг) основаны на Ван-дер-Ваальсовых взаимодействиях

В составе спирали выделяют виток – повторяющийся элемент, его можно поделить на большую и малую бороздку (Рис. 12.6). Большая бороздка от малой мысленно отделяется линией, проходящей вдоль сахарофосфатного остова:

Рисунок 12.6 – Малая и большая бороздка в составе двухцепочечной ДНК.

Часть атомов гетероциклических оснований смотрят в большую бороздку, остальные – в малую (Рис. 12.7). Это имеет значение при узнавании ДНК, например, клеточными белками, которые взаимодействуют с определенными атомными группировками, смотрящими в бороздки – в большой бороздке наблюдается куда большее разнообразие атомов с различными химическими свойствами.

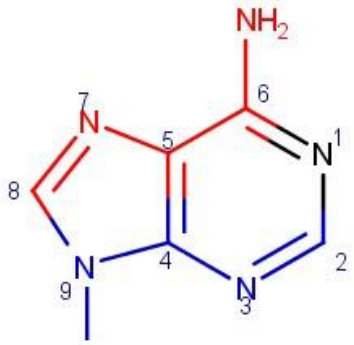


Рисунок 12.7 - Расположение атомов остатка аденина в большой (красные атомы) и малой (синие атомы) бороздке ДНК.

Природные нуклеиновые кислоты в основном существуют в виде двух регулярных A- и B-форм. A-форма ДНК представляет собой «сплющенную» B-ДНК (Рис. 12.8). В составе B-ДНК основания взаимопараллельны, что не является верным для A-формы. Интересно, что A-ДНК условно и является «сплющенной», но из-за этого в ней не образуются новые водородные связи, а количество стекинг-взаимодействий практически не меняется.

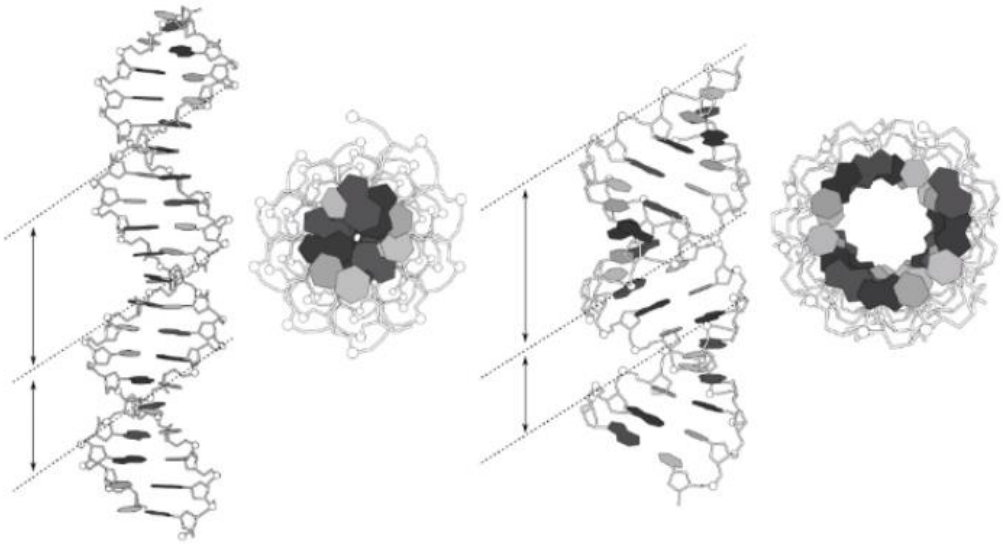
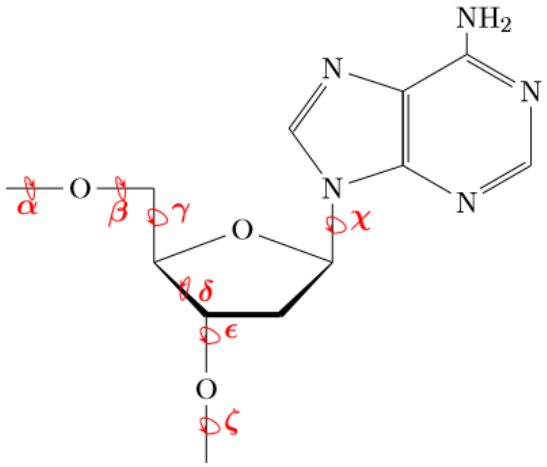


Рисунок 12.8 - B-форма ДНК (слева) и A-форма ДНК (справа).

Введем численное описание относительного расположения атомов нуклеотида в составе ДНК. В первом приближении плоские гетероциклические основания имеют постоянную форму, в отличие от неплоского углеводного фрагмента (так как его атомы являются насыщенными), конформация которого описывается семью торсионными углами:



Форма	α	β	γ	δ	ϵ	ζ	χ
A-ДНК	62	173	52	88/3	178	-50	-160
B-ДНК	63	171	54	123/131	155	-90	-117

Рисунок 12.9 - Торсионные углы, необходимые для численного описания конформации дезоксирибонуклеотида.

Выберем три атома в углеводном остатке, лежащие в одной плоскости: C4', O4' и C1' (можно было выбрать любые три атома, так как три точки всегда лежат в одной плоскости, но данный набор наиболее удобен). В C2'-эндо-конформации углеводного остатка, специфичной для B-формы ДНК, C2'-атом находится по одну сторону от выбранной плоскости с азотистым основанием. В свою очередь, для A-формы ДНК

специфична C2’-экзо конформация, в которой азотистое основание и C2’-атом находятся по разные стороны от плоскости. РНК существует в А-форме из-за наличия 2’-гидроксильной группы, которая отталкивается от гетероциклического основания.

Разные формы ДНК переходят друг в друга при изменении условий внешней среды:

- B-форма стабильна при нормальных физиологических условиях
- дегидратация, понижение относительной влажности до 0.75 инициирует переход B- в A-форму (происходит десольватация гидрофильных фосфатных групп).

Пример: смеси вода-этанол (метанол) при росте доли спирта более 0.75 обеспечивают переход B- в A-форму, при этом ДНК из-за дегидратации можно легко перевести в осадок.

Важно отметить, что сахарофосфатный остов по всей длине заряжен (каждое нуклеотидное звено несет заряд -1 на фосфатной группе), что приводит к сильному отталкиванию внутри молекулы ДНК. Для образования стабильной двухспиральной структуры заряды остова необходимо компенсировать, например, с помощью добавления катионов в раствор.

Вторичная структура РНК

Нуклеотид РНК имеет в своем составе 2’-гидроксильную группу (2’-OH).

В состав РНК часто входят модифицированные основания. Например, псевдоуридин, входящий в состав тРНК, содержит гликозидную связь вместо N-гликозидной, что позволяет данному основанию образовывать водородные связи четырьмя гетероатомами:

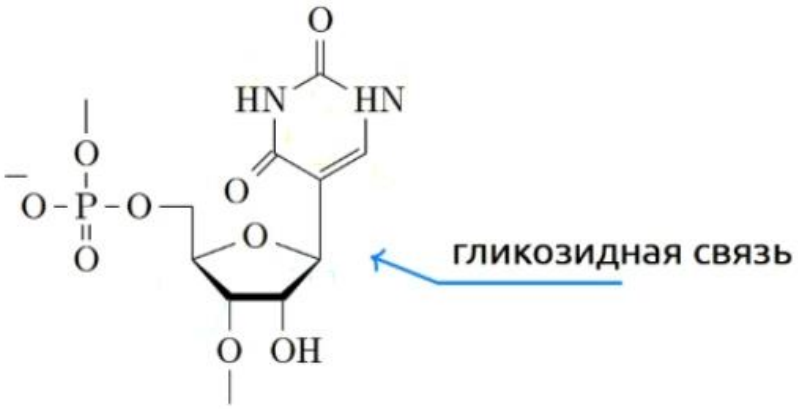


Рисунок 12.10 - Строение неканонического псевдоуридинового рибонуклеотида.

Основные свойства РНК:

- одноцепочечная молекула
- в клетке найдено множество видов РНК и каждый из них имеет специфичную функцию (основная функция у ДНК единственная – хранение и передача генетической информации).

Основные типы: рРНК, мРНК, тРНК.

РНК образует огромное множество вторичных структур:

- петля (loop)

- внутренняя петля (internal loop)
- стебель (stem)
- мультипетля (junction)
- выпетливание (bulge)
- псевдоузел (pseudoknot) и т.д.

Структура РНК определяет ее функцию:

- регуляторная
- структурная
- каталитическая (рибозимы)

Некоторые вирусы имеют РНК геном (ВИЧ, грипп, коронавирус и т.д.).

Представить вторичную структуру РНК можно следующими способами:

- изогнутая первичная структура с нанесенными водородными связями (наиболее часто используется, так как визуально удобно)
- дуги, соединяющие нуклеотиды в линейно изображенной первичной структуре, которые символизируют водородные связи
- dot-bracket notation (компьютерный вариант)

Чем длиннее РНК, тем больше можно представить комбинаций ее вторичной структуры. Существует *алгоритм Зукера*, позволяющий оценить их энергию. Эмпирическим путем множество вторичных структур было исследовано методом плавления, для каждой из них были определены изменения свободной энергии Гиббса (и табулированы), после чего были вычислены средние вклады тех или иных факторов в устойчивость различных элементов вторичной структуры РНК.

РНК, структуру которых трудно предсказать алгоритмом Зукера:

- РНК, связанные с белками (белки могут стабилизировать неустойчивые элементы структуры РНК)
- длинные РНК (большая вариативность структур приводит к тому, что алгоритм дает большое количество вариантов с одинаковым значением энергии)
- псевдоузлы (из-за большой вариативности)

На основе вторичной структуры РНК с помощью простых правил можно предсказать третичную структуру, но это работает только для структурированных РНК, не имеющих протяженные неспаренные участки.

Механические модели ДНК

Моделирование хроматина (ДНК, плотно упакованная в ядре) представляет собой куда более сложную задачу из-за длины этой ДНК (3.2 млрд пар оснований) и ее на данный момент невозможно моделировать на атомном уровне.

Freely Jointed Chain (примитивная модель описания полимеров)

Имеется полимер из N мономеров, соединенных шарнирами. Тогда не упакованная длина $L = N*l$.

- сегменты между собой не взаимодействуют
- полимер флуктуирует и его форма определяется простым распределением
- если полимер образует глобулу:

$$< R^2 \geq Nl^2 = Ll \tag{12.1}$$

$$P(\vec{R}) = \left(\frac{3}{2\pi N l^2}\right)^{\frac{3}{2}} e^{-\frac{3\vec{R}^2}{2Nl^2}} \tag{12.2}$$

$$\sqrt{\langle \overrightarrow{R^2} \rangle} = \sqrt{N}\, l = \sqrt{Ll} \tag{12.3}$$

Однако стоить иметь в виду, что ДНК не склонна образовывать глобулы, так как данная молекула отрицательно заряжена.

Rod Model (модель столбика)

ДНК представляется в виде стержня, способного сгибаться за счет изменения эффективности стекинг-взаимодействий. У данного стержня имеется коэффициент упругости. С помощью атомно-силовой микроскопии для разных коротких последовательностей была оценена их жесткость. В расчетах по данной модели используется выражение следующего вида:

$$\Delta G = \frac{1}{2} EIL\alpha^2 \tag{12.4}$$

где E – коэффициент Юнга, а I – момент инерции (для цилиндра радиусом r момент инерции рассчитывается согласно формуле $I = \pi r^4/4$.

Чем жестче ДНК, тем менее вероятно, что в данном месте можно обнаружить регуляторный элемент, что имеет следующее объяснение. Белки имеют некую форму, представим их в виде сферы. Данные биомолекулы узнают определенные короткие (до 10 нуклеотидов) паттерны в нуклеиновой кислоте. В месте узнавания белка ДНК изгибается для увеличения площади контакта и, следовательно, повышения специфичности взаимодействия.

Worm Like Chain Model (модель типа червячка)

Суть модели – непрерывное описание цепи для решения ряда ограничений:

- энтропийная эластичность ДНК цепи состоит из малых девиаций по оси молекулы из-за температуры
- направление цепи коррелирует с длиной цепи, называемой “the persistence length” (Для ДНК в 10 мМ растворе NaCl P_{DNA} = 150 п.о. или 550 нм)

Силы порядка k_BT/P нужны для выравнивания и направления единиц эластичности вдоль оси полимера.

Для сил, действующих на НК в диапазоне менее 100 фН и более 5 пН малосегментная модель Freely Jointed Chain, хорошо работает:

$$f = \frac{k_B T}{b} \frac{1}{1-\frac{z}{L}} \tag{12.5}$$

Модель Worm Like Chain Model работает во всех диапазонах сил:

$$f = \frac{k_B T}{P}\left[\frac{1}{4\left(1-\frac{z}{L}\right)^2} - \frac{1}{4} + \frac{z}{L}\right] \tag{12.6}$$

Мезомоделирование ДНК

Возможно проводить моделирование ДНК на уровне пар оснований – гетероциклические основания или их пары принимаются за одну частицу (Рис. 12.11). Данное упрощение позволяет перейти к моделированию структур больших ДНК.

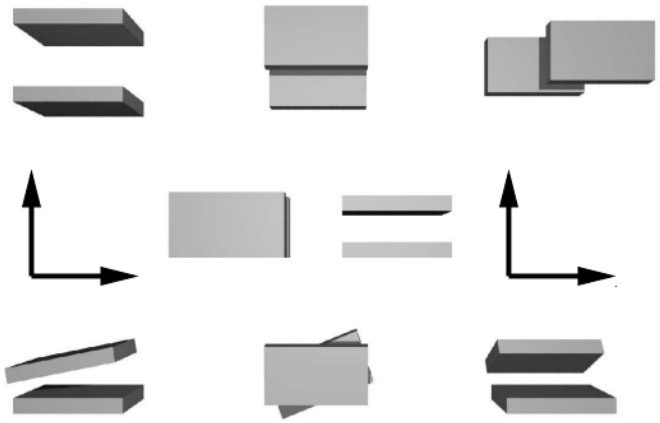


Рисунок 12.11 - Основные типы движений пар оснований относительно друг друга: rise (Ri), slide (Sl), shift (Sh), twist (Tw), roll (Ro), tilt (Ti).

При расчете пользуются модифицированным уравнением Леннарда-Джонса:

$$U(A_1, A_2, r_{12}) = U_r \eta_{12} \chi_{12}$$

$$U_r = 4\epsilon \left(\left(\frac{\sigma}{h+\gamma\sigma} \right)^{12} - \left(\frac{\sigma}{h+\gamma\sigma} \right)^{6} \right) \quad (12.7)$$

В уравнении (12.7) параметры η_{12} и χ_{12} отвечают за силу взаимодействия в зависимости от ориентации эллипсов, которые описывают пару оснований. Эллипсы далее соединяются обычными связями для представления ДНК.

В эксперименте возможно растянуть ДНК за разные концы. После достижения определенной критической силы стекинг-взаимодействия нарушатся, при этом значения теоретически рассчитанных сил совпадают с экспериментальными, а локальные деформации реалистичны. Необходимо понимать, какие силы к каким изменениям ДНК приводят, так как в реальности ДНК упаковано в хроматин и, например, для транскрипции определенного участка ДНК его необходимо растянуть (декомпактизовать), для чего необходима определенная сила. Она обеспечивается белками, а точнее, их модификациями (ацетилирование/метилирование гистонов и т.д.).

Хроматин

В ядре ДНК упакована в хроматин, который представляет собой тяжи ДНК, намотанные на белки-гистоны. Образование такого комплекса обеспечивает высокую плотность упаковки. При моделировании хроматина рассматривают его минимальную единицу – нуклеосому (два витка ДНК, намотанные на гистоны):

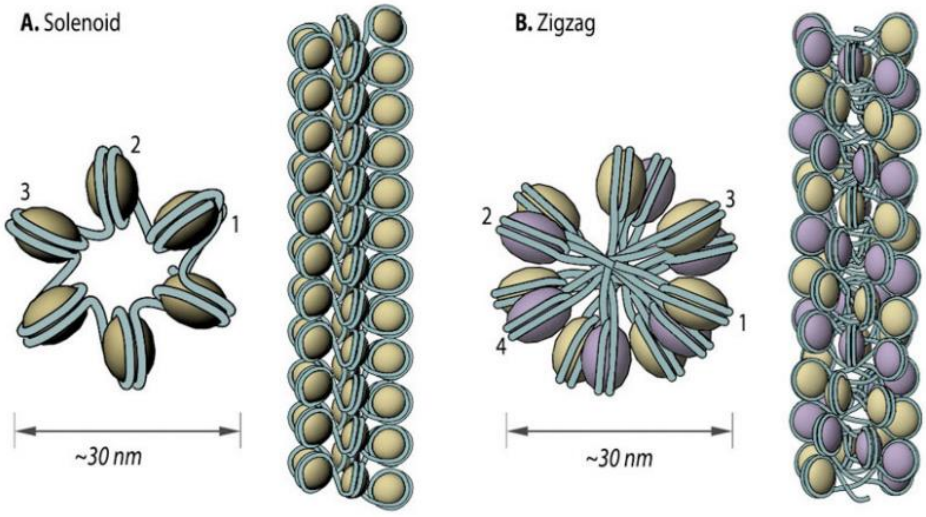


Рисунок 12.12 - Схематичное представление структуры хроматина.

Модель блока для моделирования хроматина описывается следующими параметрами: расстояние между нуклеосомами (r), обыкновенный угол (θ) и торсионный угол (ϕ) (Рис. 12.13). Экспериментальные исследования показали, что у хроматина есть большое количество запрещенных зон (конформаций, в которых он не может находиться).

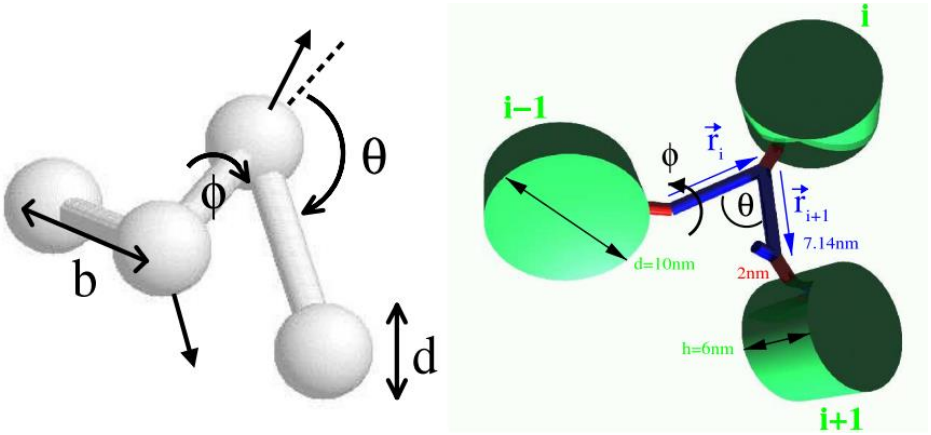


Рисунок 12.13 - Блок для моделирования пространственной структуры хроматина.

Стоит отметить, что с помощью атомно-силовой микроскопии на основании данной модели показали, насколько сильно меняется структура хроматина при растягивании.

Метод HiC позволяет с помощью химических модификаций узнать, какие участки ДНК сближены между собой внутри хроматина. В результате получают «тепловые карты», содержащие участки разной интенсивности. Чем темнее участок, тем более сближены данные фрагменты ДНК. Исходя из данных, получаемых этим методом, обычно пытаются подобрать параметры хроматина, соответствующие имеющимся картам, либо на основании карт пытаемся воспроизвести структуру. В обоих случаях результатом является модель пространственной структуры хроматина.

ФАКУЛЬТЕТ
БИОИНЖЕНЕРИИ И
БИОИНФОРМАТИКИ
МГУ ИМЕНИ
М.В. ЛОМОНОСОВА

teach-in
ЛЕКЦИИ УЧЕНЫХ МГУ