Биомедицинская химия, 2007, том 53, вып. 6, с. 713-728.

ЛЕКЦИЯ

УДК 577.1

ОСНОВНЫЕ ПРИНЦИПЫ МОЛЕКУЛЯРНОГО КОНФОРМАЦИОННОГО АНАЛИЗА ДЛЯ МЕДИКО-БИОЛОГОВ

А.С. Иванов

ГУ НИИ биомедицинской химии им. В.Н. Ореховича РАМН, Погодинская ул., 10, 119121, Москва; факс: (095) 245-0857; эл. почта: alexei.ivanov@ibmc.msk.ru

Кратко изложены основные принципы анализа и оптимизации молекулярной конформации, которые лежат в основе методов молекулярного моделирования, используемых для решения задач в области биоинформатики. Рассмотрены основные подходы в поиске энергетических минимумов модельной молекулярной системы. Данная лекция входит в теоретический цикл "Биоинформатика и компьютерное конструирование лекарств" для студентов 4 курса Медико-биологического факультета РГМУ (специальности – биохимия, биофизика, медицинская кибернетика), может быть также рекомендована для других студентов и аспирантов медико-биологических специальностей.

Ключевые слова: лекция, вычислительная химия, молекулярное моделирование, молекулярная механика, молекулярная конформация, энергетическая минимизация

1. Основные положения.

Первоначально необходимо напомнить следующие общеизвестные постулаты:

1) молекулы, за редким исключением, являются гибкими структурами, т.е. их конформация может меняться;
2) энергия молекулы зависит от ее конформации;
3) молекула с минимальной энергией имеет более стабильную конформацию (в соответствии с законами термодинамики молекулярная система стремится к минимуму энергии).

2. Конформационный анализ.

Одной из задач молекулярного моделирования является нахождение стабильных конформаций молекул. Такой поиск осуществляется методами конформационного анализа, при этом в качестве основного критерия используется величина потенциальной энергии молекулы.

2.1. Потенциальная поверхность.

Зависимость энергии молекулы от ее конформации может быть представлена в виде многомерной потенциальной поверхности. Данная поверхность имеет возвышенности и ямы, соответствующие максимумам и минимумам энергии молекулы при ее перемещении по доступному для нее конформационному пространству. На рисунке 1 показан условный график, иллюстрирующий сечение данной потенциальной поверхности. Самый глубокий энергетический минимум называется глобальным, в то время как другие минимумы называются локальными.

Рисунок 1.
Условный график сечения потенциальной поверхности молекулярной системы.

Первоначально, после создания молекулярной модели, молекула может находиться в любой точке данной потенциальной поверхности. В соответствии с законами термодинамики стабильное состояние молекулярной системы соответствует минимуму ее энергии, и, следовательно, после построения модели необходимо выполнение процедуры оптимизации молекулярной геометрии для минимизации энергии системы. Данная процедура соответствует перемещению положения молекулы на потенциальной поверхности в энергетический минимум. Какой минимум будет при этом найден, зависит как от начальной геометрии, так и от алгоритма процедуры оптимизации.

В связи с этим возникают два вопроса:

1) Каким образом мы можем найти глобальный или локальный минимумы?
2) В какой конформации молекула находится наибольшее время?

Ответы на данные вопросы будут даны в материалах данной лекции.

Существует много различных методов поиска энергетического минимума молекулярных систем. Все используемые подходы можно систематизировать в виде следующих основных групп:

1) Систематический поиск - полный анализ конформационного пространства путем систематической проверки всех вариантов конформаций.
2) Случайный поиск - значения торсионных углов молекулярной модели задаются случайным образом. При этом нет гарантии, что будут найдены все минимумы.
3) Эволюционный поиск (генетический алгоритм и стратегия эволюции) - торсионные углы изменяются по определенным алгоритмам, основанным на генетических принципах.

2.2. Систематический поиск.

С помощью систематического поиска теоретически можно получить полную картину потенциальной поверхности путем систематического перебора всех возможных величин торсионных углов (вращение различных фрагментов молекулы вокруг всех разрешенных для вращения связей). Рассмотрим в качестве простейшего примера молекулу *н*-гексана (C_6H_{14}). Как правило, при конформационном анализе связи типа X-H игнорируются как мало функциональные. X обозначает так называемый “тяжелый” атом, которым может быть любой атом кроме водорода. Такое упрощение выглядит вполне оправданно, так как ковалентная связь между водородом и “тяжелым” атомом заметно короче, чем сумма их радиусов Ван-дер-Ваальса (VDW). В результате водород оказывается сильно “заглубленным” в VDW сферу тяжелого атома, что делает маловыраженной геометрию такой группировки, а значение торсионного угла равновероятным с точки зрения энергии системы. Таким образом в случае примера с *н*-гексаном мы имеем всего три торсионных угла для вращения (рис. 2).

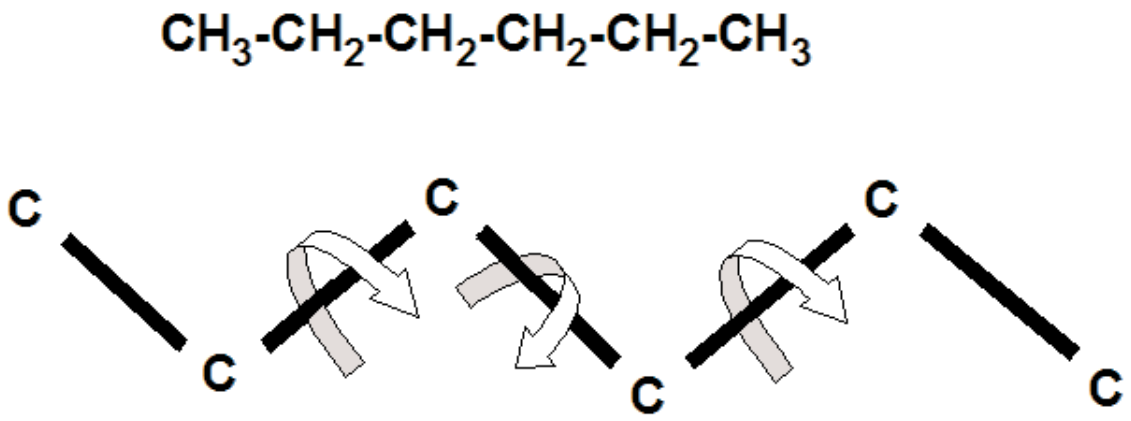

Рисунок 2.
Конформационный анализ простой молекулы на примере н-гексана (C_6H_{14}).
Стрелками показаны три торсионных угла, вращение которых обеспечит выполнение систематического конформационного анализа.

При выполнении конформационного анализа осуществляется дискретное вращение фрагментов молекулы вокруг связей (круговое вращение с определенным угловым шагом или инкриментом) и расчетом энергии системы в каждой точке (после каждого шага). Такой метод конформационного анализа называется *систематическим поиском* или *поиском по решётке*. Последний термин обусловлен тем, что при дискретных изменениях торсионных углов с выбранным угловым шагом потенциальная поверхность покрывается сеткой (в многомерном случае - решеткой) с анализируемыми точками в её узлах.

Для анализа всего конформационного пространства требуются очень большие вычислительные ресурсы и много вычислительного времени. Это следует из самой природы систематического поиска, относящегося к математическим проблемам из области комбинаторики, когда число вариантов крайне быстро возрастает, что может привести к “комбинаторному взрыву”. В общем случае число анализируемых конформеров (***N***) зависит от числа вращаемых связей (***n***) и углового инкремента (***x***):

$$N = \left(\frac{360}{x}\right)^n \qquad (1)$$

В нашем примере с *н*-гексаном $n = 3$.

Число генерируемых конформеров (***N***), а следовательно и число энергетических расчетов, быстро растет с уменьшением величины углового инкремента (***x***) (табл. 1, А).

Таблица 1. Зависимость числа возможных конформеров *н*-гексана от величины углового инкремента (А) и числа вращаемых связей (Б).

	Число вращаемых связей (n)	Угловой инкремент (x)	Число конформеров (N)
А	3	30°	1728
	3	15°	13824
	3	7,5°	110592
Б	4	30°	20736
	5	30°	248832
	6	30°	2985984

При $x = 7{,}5°$, что может обеспечить только грубый конформационный анализ, значение ***N*** для такой простой молекулы как н-гексан превышает 10^5. В случае более сложных молекул число вращаемых связей (***n***) растет и число конформеров может увеличиться драматически (табл. 1, Б).

Основная стратегия систематического поиска представлена на рисунке 3 в виде упрощенной схемы. Возможные варианты конформаций представлены в виде перевернутого дерева (рис. 3А). Каждый узел соответствует одной конформации. Для простоты рассмотрения число возможных значений торсионных углов для вращаемых связей 1 и 3 (ВС1 и ВС3) приравнено 3, а для ВС2 – всего 2. Обычный подход заключается в том, что первоначально ВС1, ВС2 и ВС3 устанавливаются в свои начальные значения и систематически перебираются все значения ВС3. Затем ВС2 изменяется на следующее значение и повторяется перебор всех значений ВС3 и т.д. После исчерпания всех значений ВС2, производят изменение ВС1 на следующее значение и все повторяется для ВС2 и ВС3, и т.д. Такой тип поиска известен как "поиск в глубину" ("depth-first search"), так как наружные ветви сканируются более быстро.

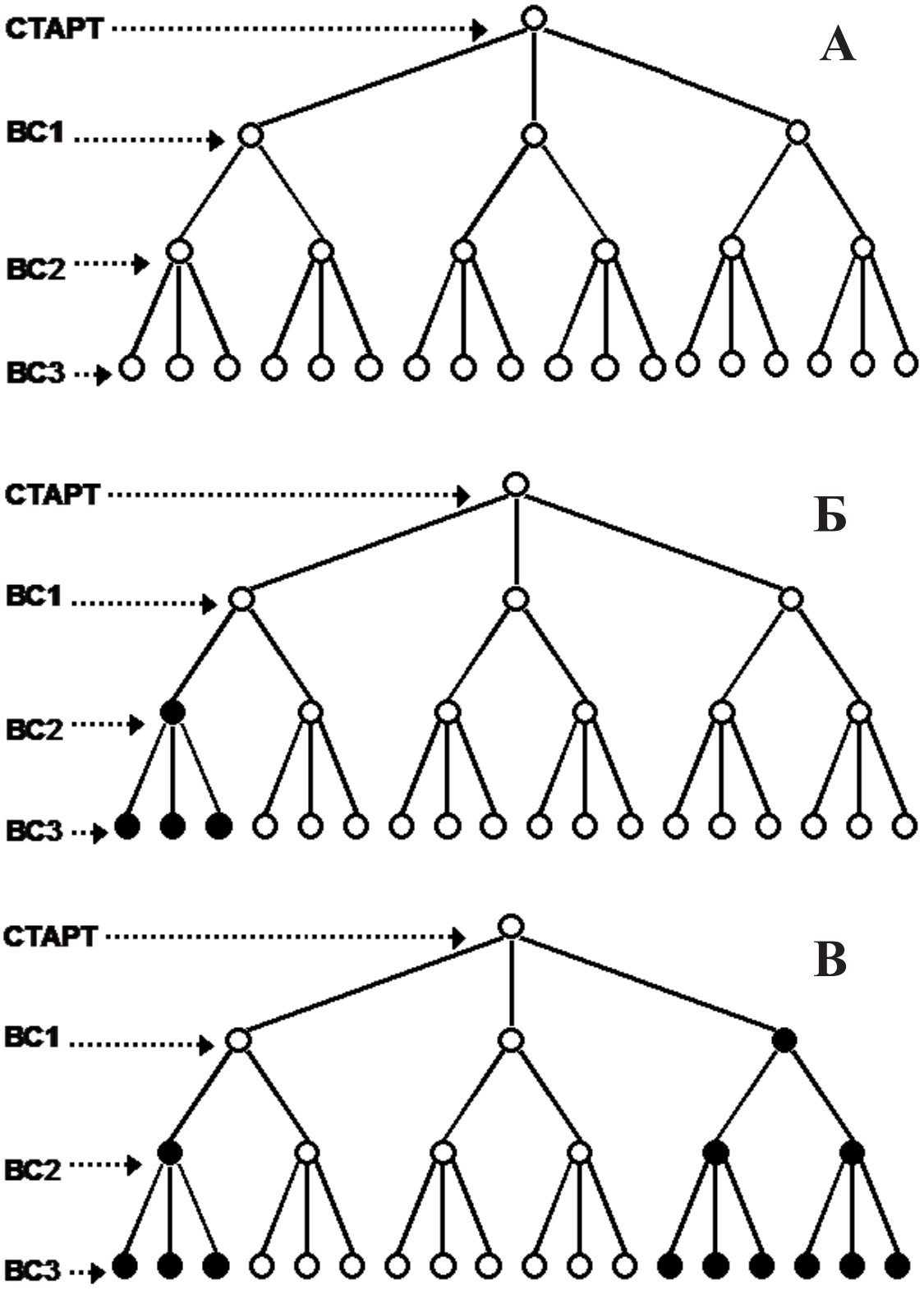

Рисунок 3.
Упрощенная схема стратегии систематического конформационного анализа молекулярной системы.
А – полное дерево всех возможных вариантов конформаций (см. подробности в тексте).
Б и В – отсечение из-за стерических ограничений ветвей ВС2-ВС3 и ВС1-ВС2-ВС3, соответственно.

Число проверяемых конформаций может быть снижено путем простой проверки на стерические конфликты (пересечение VDW сфер) и исследования свойств дерева. Если вариант BC2 приводит к конфликту, то все варианты BC3 - не актуальны, так как они не могут ликвидировать стерический конфликт. Таким образом сканирование BC3 может не проводиться и эта ветвь дерева может быть "отсечена" (рис. 3Б). Таким же образом, если значение BC1 приводит к конфликту, то более крупная ветвь может быть отсечена (рис. 3В). Следовательно, число конформеров, для которых необходимо рассчитывать энергию, может быть быстро и значительно сокращено.

Другой проблемой является выбор величины углового инкремента (***x***). Как было показано выше, при уменьшении значения углового инкремента происходит очень быстрый рост числа конформеров (комбинаторный взрыв). С другой стороны, при относительно больших значениях инкремента может быть пропущен энергетический минимум (рис. 4). Данная проблема (ограничение величины углового инкремента как сверху, так и снизу) носит название "тирании решетки".

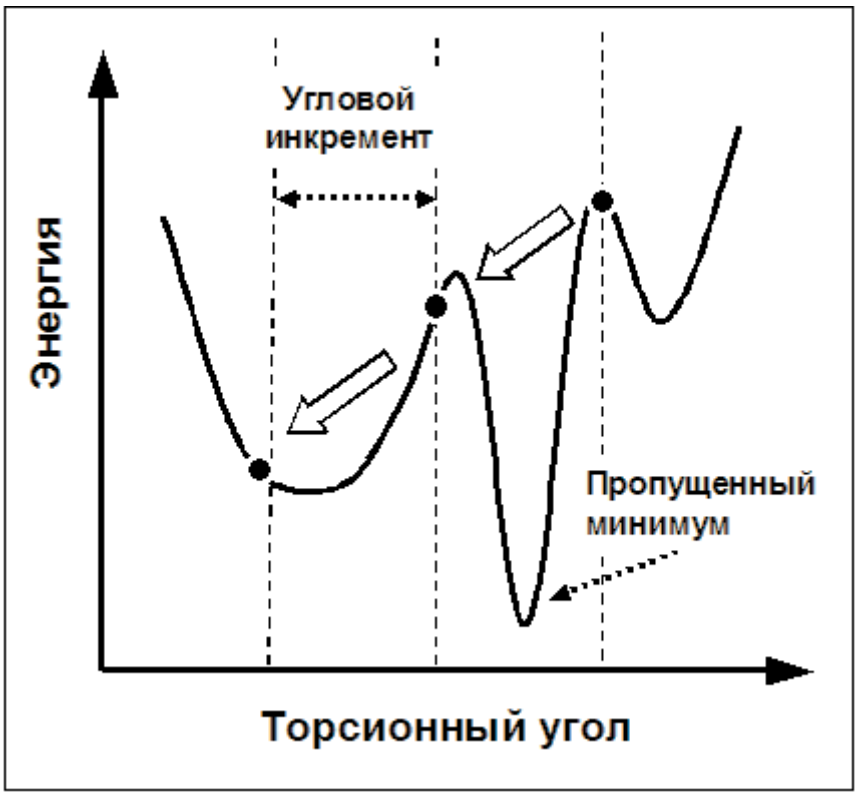

Рисунок 4.
Иллюстрация возможного пропуска энергетического минимума при систематическом поиске в случае использования относительно больших значений углового инкремента.

2.3. Случайный поиск.

2.3.1. Метод Монте-Карло.

Данный подход базируется на многократном повторном (итерационном) выполнении следующих основных шагов:

1) Выполняется расчет энергии для текущей конформации молекулы (E_i).
2) В молекулярную систему вносится небольшое случайное возмущение в виде смещения позиции атома или изменения торсионного угла.
3) Производится расчет энергии для новой конформации системы (E_{i+1}).
4) Новая конформация принимается, если:

$$\Delta E \leq 0 \qquad (2)$$

или

$$e^{\frac{\Delta E}{kT}} > x \qquad (3)$$

где: $\Delta E = E_{i+1} - E_i$

k - константа Больцмана

T - температура

x - случайное число между 0 и 1.

Путём многократного повтора данной процедуры (в течение достаточно длительного периода времени) постепенно получается корректный набор взвешенных по распределению Больцмана структур.

Моделирование Монте-Карло создает относительно большие изменения в системе и проверяет соответствие измененной структуры энергетическим требованиям при заданной температуре. Система может "перепрыгивать" из одной конформации в другую не замечая существующих барьеров.

Так как моделирование Монте-Карло анализирует конформационное пространство вне реального времени и траекторий конформационных изменений, оно не дает информацию о реальной динамике молекулярной структуры во времени. Однако метод Монте-Карло может быть более эффективным по сравнению с моделированием молекулярной динамики в оценке средних термодинамических параметров.

Метод Монте-Карло использует Больцмановские вероятности, а не силы, приводящие к изменению системы.

Существуют различные варианты реализации метода Монте-Карло. Из них наиболее известным является моделирование по алгоритму Метрополиса.

2.3.2. Моделирование молекулярной динамики.

Моделирование молекулярной динамики (МД) представляет собой компьютерное моделирование реальных движений в молекулярной системе во времени. Каждому атому присваивают начальные координаты в пространстве и вектор движения (направление и скорость). Интегрирование второго закона Ньютона ($F = m \cdot a$) позволяет предсказать новые значения координат атомов и новые вектора, которые будет иметь молекулярная система через определенный период времени (Δt, временной шаг МД).

Для того чтобы начать МД необходимо выбрать температуру моделирования. Начальные скорости для всех элементов системы выбираются случайно (с использованием распределения Максвелла):

$$E_{кин} = \frac{1}{2}\sum mv^2 = \frac{3}{2}NkT \qquad (4)$$

Затем могут быть рассчитаны многие свойства системы (энтальпия, энтропия и т.д.) с использованием методов статистической механики.

Временной шаг МД (Δt)

Для корректного моделирования динамики движений в молекулярной системе, шаг МД во времени (Δt) должен быть короче, чем самые быстрые движения в молекуле, которыми являются колебания длины связи атома водорода с тяжелым атомом (**H-X**) с периодом примерно 1 фс (фемтосекунда, 10^{-15} с). Все другие молекулярные движения имеют более длинные периоды колебаний, то есть движения более медленные. Например:

- движения белковых петель имеют период порядка 10^{-11}-10^{-7} с;
- движение боковых цепей аминокислотных остатков на поверхности белковой глобулы - порядка 10^{-11}-10^{-10} с;
- в глубине белковой глобулы боковые цепи аминокислотных остатков движутся гораздо медленнее – период порядка 10^{-4}-10^{-1} с.

Таким образом, временной шаг в моделировании МД должен быть крайне малым, что влечет за собой выполнение огромного объема вычислений, что требует использования мощных вычислительных систем (суперкомпьютеров, многопроцессорных серверов или кластерных систем) и большого расхода машинного времени (от нескольких дней до нескольких месяцев). Недели подобных вычислений на мощных компьютерах позволяют моделировать всего лишь наносекунды (10^{-9} сек) реального времени жизни простейшей молекулярной системы типа “одна молекула белка в капле воды”.

Граничные условия МД

При моделировании МД в определенном объеме трехмерного пространства (как правило, в прямоугольном боксе) возникают так называемые “граничные эффекты”, представляющие собой отклонения свойств молекул в граничной области от свойств молекул в центре бокса. Существует два подхода для решения данной проблемы:

1) Стохастические граничные условия. Ограничение движения атомов в пограничной области путем окружения системы отражающими стенками.

2) Периодически граничные условия. Моделирование непрерывной (бесконечной) молекулярной системы путем размещения копий моделируемого бокса во всех направлениях.

Расчёт электростатических взаимодействий в МД

Так как электростатические взаимодействия являются дальнодействующими, то в расчеты должны включаться заряды, находящиеся на значительных расстояниях. В тоже время сила взаимодействия в соответствии с законом Кулона линейно убывает с расстоянием и для сокращения объема вычислений в МД вводится так называемый уровень “отсечки” (cutoff) – максимальное расстояние между зарядами, в пределах которого производится расчет электростатики (на более дальних расстояниях электростатическими взаимодействиями пренебрегают). В зависимости от типа моделируемой молекулярной системы может быть использована разная геометрия отсечки. Например, при моделировании структуры фрагмента бислойной липидной мембраны целесообразно использовать *цилиндрическую отсечку*, в то время как при моделировании МД белковой молекулы в водной фазе – *сферическую отсечку*. В ряде случаев для дополнительного сокращения объема вычислений при моделировании МД очень больших молекулярных систем используется *сферическая двойная отсечка* - внутри малой сферы (~ 1 нм) расчет электростатики осуществляется на каждом шаге МД, в то время как внутри большой сферы (~ 1,5–2,0 нм) – только на каждом 10 шаге МД.

Наборы параметров МД

Моделирование МД позволяет контролировать ряд параметров молекулярной системы, таких как: число частиц (N), объем (V), температура (T), давление (P), энергия системы (E) и др. В зависимости от задач могут моделироваться различные термодинамические процессы (например, изотермический, изобарический, и т.д.) для чего ряд параметров системы делают постоянными, что описывается так называемым *набором параметров* системы. В качестве примера можно назвать два набора: *канонический набор* (NVT) – постоянны параметры N (число частиц), V (объем) и T (температура); и *микроканонический набор* (NVE) – постоянны параметры N, V и E (энергия системы).

Моделирование "отжига" в МД

При попытке исследования с помощью МД всего конформационного пространства такой сложной молекулярной системы как структура белка, потребуется практически бесконечное время вычислений (даже при использовании мощных суперкомпьютеров). Поэтому для быстрого преодоления энергетических барьеров в МД задают большие значения скоростей атомов, что соответствует нагреву системы. Так как вычислительный эксперимент является виртуальным и молекулярная система не может взорваться или сгореть, то её температуру первоначально поднимают до очень высоких значений (примерно до 10000 K°),

что поднимает энергию системы выше всех возможных энергетических барьеров. В таком состоянии система может легко перемещаться по всему конформационному пространству. Затем температуру системы снижают по определенной траектории (процедура "отжига"), в результате чего система "скатывается" в один из энергетических минимумов. Циклы нагрева и охлаждения повторяют до тех пор, пока перестают обнаруживаться новые энергетические минимумы.

2.4. Эволюционный поиск.

2.4.1. Генетический алгоритм.

Данный эвристический подход основан на применении генетических принципов при расчете изменений торсионных углов в конформационном анализе молекулярной системы. Термины и понятия, используемые в этом подходе, соответствуют оригинальным терминам из генетики: *индивидуумы, мутации, кроссовер, воспроизводство*. Конформационное пространство исследуется с помощью повторяющихся шагов кроссовера и мутаций. При этом могут быть найдены не все минимумы, однако, как правило, быстро находится если не глобальный минимум, то большинство минимумов с очень низкой величиной энергии.

Все индивидуумы одного поколения проверяются с помощью *функции соответствия*. В зависимости от способа смены поколений в следующий цикл воспроизводства включается определенная выборка родителей и потомства. После ряда итераций популяция будет состоять из индивидуумов, которые хорошо адаптированы по оценке функции соответствия.

Основная схема генетического алгоритма:

1) Инициация популяции индивидуумов. Это может быть сделано на основе случайного выбора или на основе дополнительных данных. Индивидуумы представляются в виде последовательности битов, что не ограничивает тип решаемых проблем, так как любые данные (числа, последовательности, структуры) могут быть закодированы в виде последовательностей битов.
2) Оценка всех индивидуумов исходной популяции.
3) Генерация новых индивидуумов. Вероятность воспроизводства для индивидуума пропорциональна относительному соответствию внутри текущего поколения. Для получения новых индивидуумов используются следующие операции:
   - *Мутация* – случайная замена одного или более битов индивидуума на новые значения (0 или 1).
   - *Изменение* – изменение битов таким образом, чтобы число, кодируемое ими, слегка увеличилось или уменьшилось.
   - *Кроссовер* – обмен части (один бит или последовательность битов) одного индивидуума с соответствующей частью другого индивидуума. Результирующие гибридные индивидуумы рассматриваются как индивидуумы нового поколения.
4) Селекция индивидуумов для нового поколения родителей.
   - В оригинальном генетическом алгоритме просто выбираются все потомки, а все родители отбрасываются. Этот подход скопирован с биологической модели и называется "полной заменой поколения".
   - Более новые варианты смены поколений сравнивают индивидуумы настоящих родителей и потомков, которые затем ранжируются по их значениям соответствия. Только ***n*** лучших индивидуумов (***n*** - размер популяции, т.е. число индивидуумов в одном поколении) выбирается для генерации следующего поколения. Этот метод называется "элитарной заменой поколения". Он гарантирует, что хорошие индивидуумы не будут потеряны. В случае полной замены поколения может получиться, что хорошие индивидуумы "вымрут", так как они производят только слабых потомков с точки зрения функции соответствия.

- Другой вариант смены поколений – “стационарная замена” (*steady state*). Из текущей популяции случайным образом выбираются два индивидуума. Применяются генетические операторы и потомки используются для замены родителей в популяции. Стационарная замена часто подвержена конвергенции, так как в среднем она требует меньшей проверки соответствия, чем элитная или полная замена поколения.

5) Новая итерация (повтор всех процедур, начиная с этапа 2).

Процесс повторяется до тех пор, пока не будет достигнуто желаемое значение величины соответствия или пока не будет выполнено заданное число итераций.

2.4.2. Стратегия эволюции.

Наряду с генетическим алгоритмом был разработан другой подход, основанный на аналогичных принципах. Центральной идеей вычислительной стратегии эволюции является получение от μ индивидуумов родителей λ индивидуумов потомства, которые были модифицированы с помощью мутаций. Значения μ и λ обычно представляют собой небольшие целые числа. Так же как и в генетическом алгоритме, родители и потомки кодируются набором числовых параметров. Только индивидуумы потомков (или потомков и родителей) конкурируют за выживание в следующем поколении.

Скорость мутаций может регулироваться для получения удачных поколений и оптимизации процесса поиска. Так называемое “правило 1/5” (примерно 20% всех мутаций в одной популяции дает жизнеспособное потомство) позволяет получить оптимальное время процесса.

Существует, по крайней мере, пять различий между генетическим алгоритмом и стратегией эволюции:

1) Стратегия эволюции была создана как оптимизатор функции, в то время как генетический алгоритм был первоначально разработан для демонстрации выгоды кроссовера в моделируемой эволюции.
2) В генетическом алгоритме воспроизводство пропорционально функции соответствия, а в стратегии эволюции – нет.
3) В генетическом алгоритме делается различие между генотипом и фенотипом индивидуума, в то время как в стратегии эволюции эти понятия совпадают.
4) В эволюционной стратегии родители и потомки могут конкурировать за выживание в следующем поколении, а в генетическом алгоритме – нет.
5) Основной движущей силой стратегии эволюции является мутация, в то время как в генетическом алгоритме – кроссовер.

Существуют также гибридные системы, в которых используются одновременно оба подхода - генетический алгоритм и стратегия эволюции.

2.5. Оптимизация структуры молекулы (минимизация энергии).

2.5.1. Принципы минимизации энергии.

Так как энергия молекулы является функцией координат атомов, то оптимизация компьютерной молекулярной модели сводится к поиску таких координат атомов, которые соответствуют структуре с минимальной энергией. Это достигается с помощью процедуры минимизации энергии, входящей в большинство программ молекулярной механики. Все используемые для этой цели методы называются методами последовательного спуска и представляют собой итерационные процедуры. При каждой итерации изменяются координаты атомов с целью последовательного снижения величины энергии (спуск по склону энергетической поверхности до локального минимума).

С математической точки зрения процедуры минимизации энергии относятся к области оптимизации функции, что часто используется в самых различных вычислениях. В случае биологических макромолекул, функцией, подлежащей оптимизации (минимизации), является энергия молекулярной системы.

Цель энергетической минимизации очень проста - найти локальный энергетический минимум. Так как энергия в этом энергетическом минимуме может быть значительно выше, чем в глобальном минимуме, то часто хотят найти именно глобальный минимум. К сожалению, пока не существует надежных методов, которые гарантировали бы нахождение глобального минимума. Исключением являются тривиальные случаи с очень простыми молекулами, а также применение ряда нелокальных методов, таких как моделирование отжига в МД, частично решающих эту проблему.

Таким образом, методы минимизации энергии обычно приводят систему в ближайший локальный минимум.

Рассмотрим простейший случай функции одной переменной. В максимумах и минимумах градиент функции (первая производная) равен нулю:

$$\frac{df(x)}{dx} = 0 \qquad (5)$$

В максимуме ее кривизна (вторая производная) отрицательна:

$$\frac{d^2 f(x)}{dx^2} < 0 \qquad (6)$$

а в минимуме кривизна, соответственно, положительна:

$$\frac{d^2 f(x)}{dx^2} > 0 \qquad (7)$$

Для многомерного случая (функции многих переменных) рассматриваются соответственно частные производные, но принцип определения максимумов и минимумов остается тем же.

Минимизация в случае функции одной переменной

Нахождение минимума простой функции не является сложной проблемой, так как для решения задачи достаточно построить график функции и выбрать минимум визуально. Однако математическое решение данной задачи лежит в основе решения задачи для функции многих переменных, так как алгоритм математического поиска минимума простой функции используется для линейного поиска в направлении одной из переменных.

Методы поиска локального минимума могут быть разделены на 3 группы:

1) основанные на анализе первой производной энергетической функции;
2) основанные на анализе первой и второй производных энергетической функции;
3) не основанные на анализе производных энергетической функции (например, деление интервала).

Часто в программах используются комбинации этих методов, так как на разных этапах оптимизации каждый из подходов имеет свои преимущества. Например, в сильно напряженных молекулярных структурах энергетическая функция и, соответственно, ее производные могут быть прерывистыми. В этих случаях будет работать только метод, не использующий производных.

2.5.2. Деление интервала.

Данный подход представляет простейший алгоритм нахождения энергетического минимума энергетической функции. Рассмотрим случай, когда определены значения функции в трех точках: x_1, x_2 и x_3 (рис. 5А). При этом значения $f(x_1) > f(x_2) < f(x_3)$.

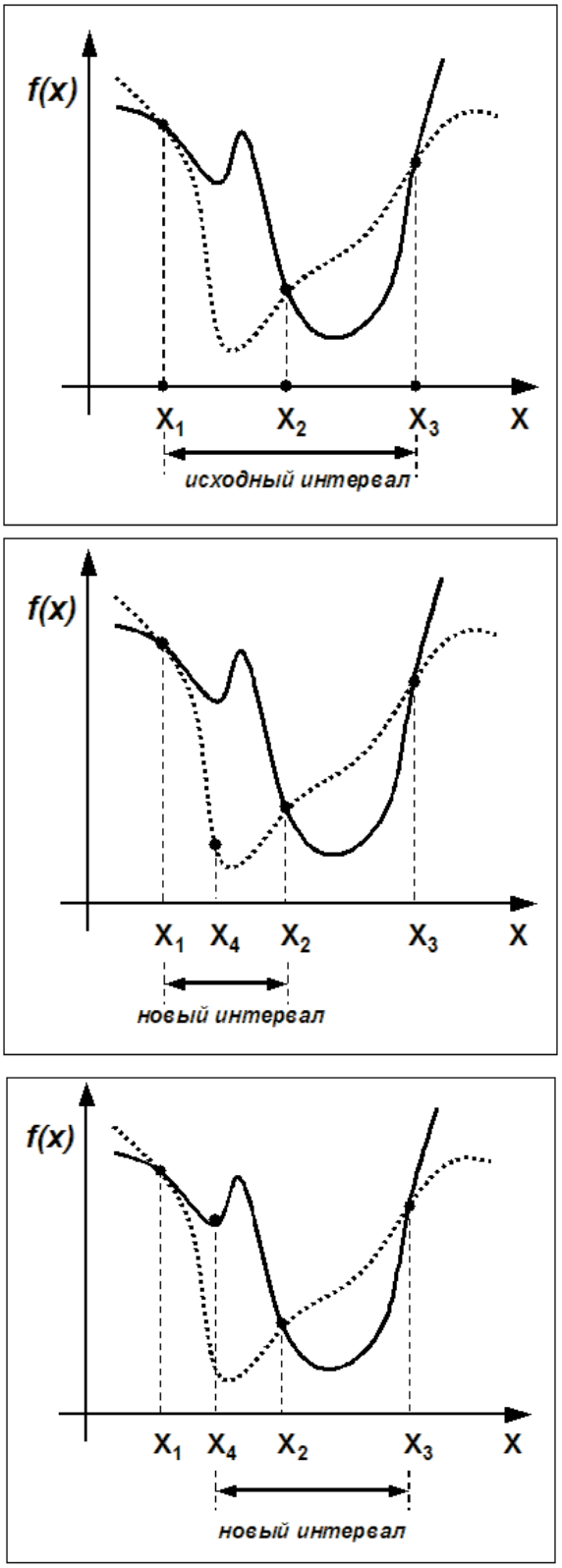

Рисунок 5.

Иллюстрация принципа поиска минимума функции методом деления интервала.

А - определены значения функции в трех точках: x_1, x_2 и x_3. Значения $f(x_1) > f(x_2) < f(x_3)$.

Б – деление интервала x_1 - x_3 путем выбора дополнительной точки x_4 и выбор нового интервала $(x_1 – x_4 – x_2)$ в случае $f(x_4) < f(x_2)$.

В – выбор нового интервала $(x_4 – x_2 – x_3)$ в случае $f(x_4) > f(x_2)$.

Если предположить, что функция непрерывна, то очевидно, что в диапазоне $\mathbf{x_1 - x_3}$ имеется по крайней мере один минимум. Если мы выберем в этом интервале дополнительную точку $\mathbf{x_4}$ и вычислим в ней функцию, то далее возможны 2 варианта:

1) $f(x_4) < f(x_2)$ и для следующего шага выбирается новый интервал $x_1 - x_4 - x_2$ (рис. 5Б);

2) $f(x_4) > f(x_2)$ и тогда выбирается новый интервал $x_4 - x_2 - x_3$ (рис. 5В).

Можно видеть, что в любом случае новый интервал меньше первоначального. Путем повторения процедуры сужения интервала локализуется минимум функции по данной переменной, пока не будет достигнута требуемая точность.

Существует много вариантов данного подхода в зависимости от способа деления интервала.

"Золотое сечение" - деление линейного отрезка (в соотношении 1: τ) таким образом, что отношение большей части ко всему отрезку равно отношению меньшей части к большей (рис. 6):

$$\frac{a\tau}{1+\tau} / a = 1/\tau \qquad (8)$$

где a – длина всего отрезка.

Преобразовав уравнение 8 можно определить значение τ :

$$\tau^2 - \tau - 1 = 0 \text{ ; или } \quad \tau = \frac{1+\sqrt{5}}{2} = 1{,}618$$

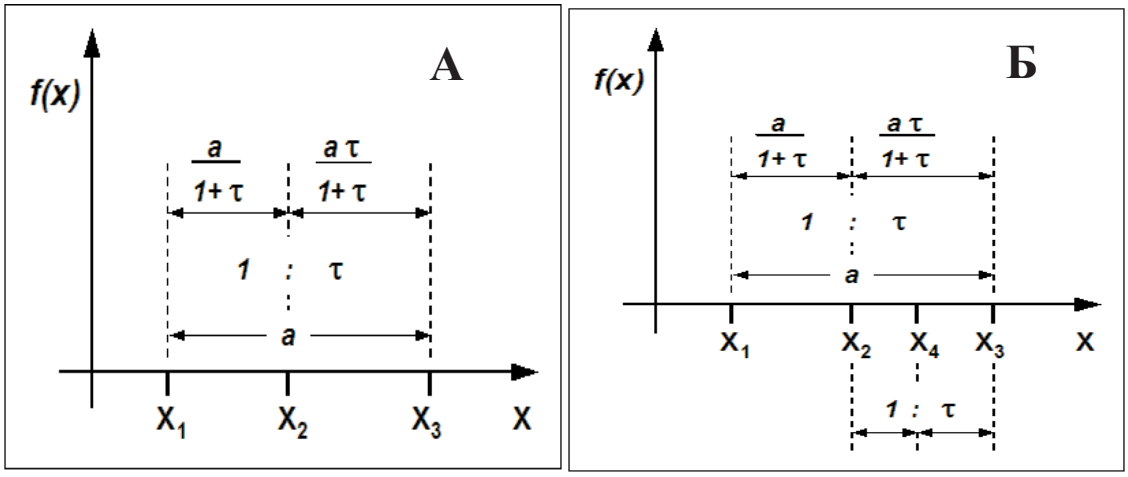

Рисунок 6.
Минимизация энергетической функции по алгоритму деления интервала с использованием "золотого сечения".
А – первая итерация - деление линейного отрезка $\mathbf{x_1 - x_3}$ (в соотношении 1 : τ).
Б – один из вариантов второй итерации – деление отрезка $\mathbf{x_2 - x_3}$ (в соотношении 1 : τ).

Данный подход является строгим ("пессимистическим"), так как не основан на каких-либо предположениях о виде функции. Минимизация функции осуществляется по описанному выше алгоритму деления интервала (см. рис. 5) с использованием "золотого сечения".

Интересно отметить, что "золотое сечение" известно с давних времен как линейное соотношение размеров, удовлетворяющее эстетическим требованиям зрительного восприятия человека, и поэтому оно давно используется в самых различных областях жизни (например, в архитектуре).

Параболическая интерполяция - более “оптимистический” подход, так как в отличие от метода “золотого сечения”, в нем делается предположение о виде функции.

Так как имеются три точки (x_1, x_2 и x_3), то функция может быть аппроксимирована в данном интервале параболой (рис. 7).

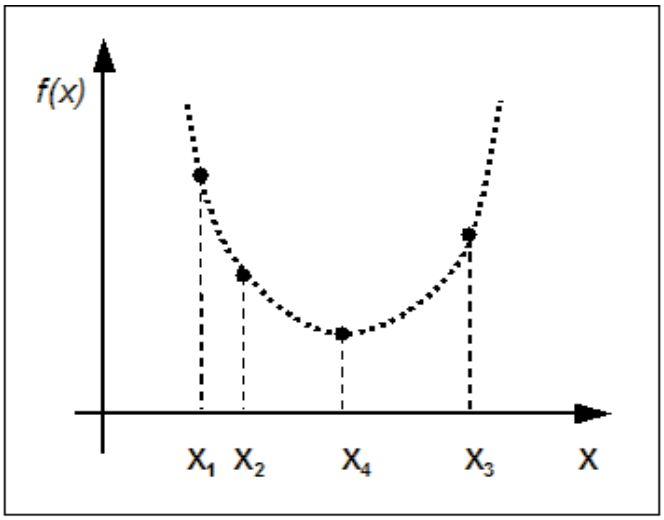

Рисунок 7.
Минимизация энергетической функции по алгоритму деления интервала с использованием параболической интерполяции.

Новая точка (x_4) для деления интервала выбирается в точке минимума параболической функции. Этот метод хорошо работает в том случае, когда система находится вблизи искомого минимума, так как практически любая функция вблизи минимума выглядит как квадратичная.

Данный подход позволяет находить минимум функции значительно быстрее, чем линейное приближение. Однако как часто бывает “за все нужно платить” и преимущество параболической интерполяции имеет свою цену – возможны ошибки в определении минимума. На рис. 8 показана такая возможная ситуация, когда точка x_2 оказывается минимумом параболической функции, построенной по трем известным точкам. В тоже время значение реальной функции не является минимальным в данной точке.

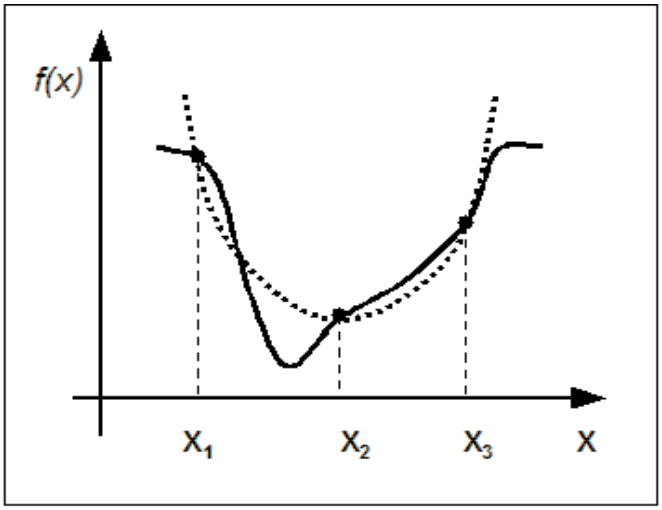

Рисунок 8.
Иллюстрация возможной ошибки в определении минимума энергетической функции методом параболической интерполяции.

2.5.3. Алгоритмы расчета новых координат атомов.

Как было уже отмечено выше, все без исключения методы минимизации являются итерационными, т.е. цикличными. Для их работы необходимы входные данные в виде оценки возможной позиции минимума. В результате одиночного цикла вычислений (итерации) генерируется новая, более точная оценка позиции минимума. Полученная скорректированная оценка используется как входные данные в следующей итерации и т.д. Процесс повторяется до тех пор, пока не будет достигнут удовлетворительный результат в определении позиции минимума.

Большинство методов используют производные энергетической функции и относятся к методам последовательного спуска, т.е. результатом каждой последующей итерации является геометрия молекулы с меньшим (или равным) значением энергетической функции.

Как следствие, эти методы могут находить только ближайший к старту минимум и никогда не могут перейти к другому минимуму (даже если он более глубокий), если он отделен от старта даже небольшим максимумом (энергетическим барьером). Как следствие, различная стартовая геометрия молекулы обычно приводит к разным минимальным энергетическим состояниям. Только для самых простых молекул (тривиальные случаи) возможно нахождение глобального минимума с помощью обычных методов минимизации. Оптимизация геометрии сложной молекулярной системы, например, структуры белка, зависит от стартовой позиции и включает одновременную оптимизацию по многим тысячам переменных.

Большинство методов неспособно находить глобальный минимум даже в случае очень простых молекул. Единственно возможный путь нахождения глобального минимума состоит в систематическом обследовании всего конформационного пространства путем многократного запуска алгоритма минимизации с различными стартовыми координатами атомов.

Нахождение минимума энергетической функции для системы из большого числа переменных само по себе является проблемой. Существует ряд различных методов оптимизации. Некоторые из них используют технику “от атома к атому”, где три координаты каждого атома оптимизируются одновременно. Другие методы оперируют ориентацией целых атомных группировок, внутренние координаты в которых остаются неизменными.

Метод скорейшего спуска – линейный поиск на основе первой производной функции в текущей точке на энергетической поверхности. Информация, полученная в предыдущих итерациях, не используется. Данная процедура хорошо работает в деформированных молекулярных системах, где направление максимального изменения энергии сильно варьирует при каждой итерации. Необходимый резерв оперативной памяти компьютера для реализации данного метода пропорционален числу переменных. Еще одним недостатком данного подхода является его плохая сходимость, т.е. при повторе процедуры минимизации могут получаться близкие, но неидентичные результаты.

Конъюгированный градиент – аккумулирует информацию об энергетической функции от итерации к итерации (используется как первая, так и вторая производная энергетической функции). Сходимость у данного метода выше, чем у скорейшего спуска. Требуемая оперативная память компьютера больше, чем в случае метода скорейшего спуска, но также линейно зависит от числа переменных.

Метод Пауэлла (Powell) – относится к семейству методов минимизации типа конъюгированный градиент. В нем используются расширенные правила определения направления спуска. Он также более устойчив к неточностям линейного поиска. В результате - он как минимум в 3 раза быстрее, чем конъюгированный градиент и пригоден для решения широкого круга проблем.

Метод Ньютона-Рапсона (Newton-Raphson) – матрица вторых производных рассчитывается аналитически или аппроксимируется цифровыми методами (так называемая Гессенская матрица). Для реализации данного подхода требуется

очень большая оперативная память компьютера (пропорциональна квадрату числа переменных). Кроме того метод Ньютона-Рапсона требует, чтобы матрица инвертировалась при каждой итерации. Время, необходимое на такую инверсию матрицы пропорционально кубу числа переменных.

3. РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. *Дашевский В.Г.* (1974) Конформации органических молекул, Химия, М.
2. *Эберт К., Эдерер Х.* (1988) Компьютеры. Применение в химии (пер.с англ.), Мир, М.
3. *Кларк Т.* (1990) Компьютерная химия (пер.с англ.), Мир, М.
4. *Leach A.R.* (1991) Reviews in Computational Chemistry, **2**, 1-55.
5. *Schlick T.* (1992) Reviews in Computational Chemistry, **3**, 1-71.
6. *Doucet Jean-Pierre, Weber Jacques* (1996) Computer-Aided Molecular Design: Theory and Applications, Academic Press, London.
7. *Pettersson I., Liljefors T.* (1996) Reviews in Computational Chemistry, **9**, 167-189.
8. *Степанов Н.Ф.* (2001) Квантовая механика и квантовая химия, Мир, М.
9. *Финкельштейн А.В., Птицын О.Б.* (2002) Физика белка: Курс лекций с цветными и стереоскопическими иллюстрациями, Книжный дом "Университет", М.
10. *Шайтан К.В., Терёшкина К.Б.* Молекулярная динамика белков и пептидов. Методическое пособие (http://www.moldyn.ru/library/manual/).

Поступила: 26. 09. 2007.

BASIC PRINCIPLES OF MOLECULAR CONFORMATION ANALYSIS FOR MEDICAL BIOLOGISTS

A.S. Ivanov

V.N. Orekhovich Institute of Biomedical Chemistry RAMS, Pogodinskaya ul., 10, Moscow, 119121 Russia; fax: +007(495)245 0857; e-mail: alexei.ivanov@ibmc.msk.ru

The main principles of the analysis and optimization of molecular conformation, which is the bases for molecular modeling methods in the field of bioinformatics, are briefly described. The basic approaches to molecular models energy minimization are considered. The given lecture is included into a theoretical cycle “Bioinformatics and computer aided drug design ” for students of Medical and biologic faculty RGMU (biochemists, biophysics and medical cybernetics). It can be also recommended for other medical and biological students, as well as for PhD students.

Key words: lection, computational chemistry, molecular modeling, molecular mechanics, molecular mechanics, molecular conformation, energy-minimization.