

УДК 519.6;54:681.3;54:51

АЛГОРИТМ РАССТАНОВКИ СИЛОВЫХ ПАРАМЕТРОВ НА АТОМАХ ОРГАНИЧЕСКИХ МОЛЕКУЛ И БЕЛКОВ В РАМКАХ СИЛОВОГО ПОЛЯ MMFF94

Ф. В. Григорьев¹, А. Н. Романов¹, О. А. Кондакова¹, С. В. Луцкекина², В. Б. Сулимов¹

Рассмотрен алгоритм FARS (Fire Algorithm for Rings Search) расстановки силовых параметров на атомах органических молекул и белков в рамках силового поля MMFF94. Алгоритм предназначен для использования при решении задачи виртуального перебора молекул из баз данных с целью отбора тех из них, которые наилучшим образом взаимодействуют с белком-мишенью. Сравниваются результаты расстановки параметров алгоритмом FARS и программным комплексом CHARMM на молекулах из базы данных NCI Diversity. Программа FARS, написанная на основе алгоритма, включена в состав веб-ориентированной платформы Keenbase, предназначенной для решения задач виртуального перебора. Программа FARS успешно использовалась при проведении компьютерного эксперимента по поиску ингибиторов к заданным белкам с применением технологий распределенных вычислений. В ходе эксперимента были задействованы вычислительные мощности НИВЦ МГУ и Южно-Уральского государственного университета.

Ключевые слова: силовые поля, силовые параметры, белки-мишени, базы данных, скрининг.

1. Введение. Задача о расчете энергии внутри- и межмолекулярного взаимодействия органических макромолекул в настоящее время решается, как правило, с использованием классических силовых полей. В рамках такого подхода потенциальная энергия системы представляется в виде суммы нескольких слагаемых, каждое из которых определяется расстоянием между силовыми центрами (в роли которых, как правило, выступают атомы) и величинами силовых параметров, которыми являются парциальные заряды, радиусы Ван-дер-Ваальса, постоянные упругости и др. Величины силовых параметров, приписываемых тому или иному атому, зависят от порядкового номера соответствующего элемента в таблице Менделеева и его химического окружения в данной молекуле и определяются правилами, принятыми в рамках конкретного силового поля.

Таким образом, для использования классического силового поля при моделировании свойств макромолекул, требующих расчета потенциальной энергии внутри- и межмолекулярного взаимодействия, необходимо на основе их структуры расставить силовые параметры. В случае моделирования единичных молекул такая процедура может быть проведена вручную, однако в задачах, где требуется в относительно короткое время провести расчеты большого количества различных молекул, возникает проблема автоматизации расстановки параметров. Одной из таких задач является виртуальный перебор (скрининг) органических молекул малого и среднего размера (лигандов) из баз данных с целью отбора тех из них, которые наилучшим образом взаимодействуют с заданной макромолекулой-мишенью (как правило, белком). Число таких соединений, для которых в ходе скрининга рассчитывается потенциальная энергия взаимодействия, может достигать десятков тысяч. Этим обусловлена необходимость использования для проведения скрининга суперкомпьютерных вычислительных ресурсов и технологий распределенных вычислений, позволяющих задействовать для решения задач большое число удаленных компьютеров. Эффективное использование удаленных компьютеров для проведения вычислений возможно при наличии надежных алгоритмов, в том числе и для расстановки силовых параметров.

В настоящей статье представлен алгоритм FARS (Fire Algorithm for Rings Search) расстановки силовых параметров, созданный для решения задачи скрининга на основе силового поля MMFF94 [1–5]. Выбор MMFF94 для решения задачи скрининга обусловлен наличием единых правил для расстановки параметров на органических молекулах любого вида, проведением параметризации на основе *ab initio* квантовохимического моделирования с учетом электронной корреляции, а также функциональной формы для потенциальной энергии взаимодействия, позволяющей учесть особенности строения различных

¹ Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, 119992, Москва; e-mail: fedor.grigoriev@gmail.com; Alexey.Romanov@list.ru; olga.kondakova@srcc.msu.ru; vladimir.sulimov@gmail.com

² Научно-исследовательский институт биомедицинской химии им. В. Н. Ореховича, ул. Погодинская, 10, 119121, Москва; e-mail: sofia.lushekina@gmail.com

молекулярных групп. Эти черты позволяют широко и успешно использовать MMFF94 для решения различных задач, связанных с моделированием образования межмолекулярных комплексов, конформационного поиска, и др. [6–9]. В рамках алгоритма FARS предусмотрена возможность расстановки параметров для белков, построенных из стандартных аминокислотных остатков, для ДНК и РНК, построенных из стандартных нуклеотидов, и для лигандов. Программная реализация алгоритма была протестирована с использованием базы данных NCI Diversity [10] и включена в состав веб-ориентированной платформы Keenbase, предназначенной для поиска в коллекции химических структур набора соединений, избирательно взаимодействующих с активным центром белковой макромолекулы [11].

2. Силовые параметры в рамках MMFF94. Количество и виды силовых параметров в рамках силового поля MMFF94 определяются функциональным выражением для полной потенциальной энергии системы (в ккал/моль):

$$U_{\text{MMFF}} = \sum \text{EB}_{ij} + \sum \text{EA}_{ijk} + \sum \text{ESB}_{ijk} + \sum \text{EOOP}_{ijk;l} + \sum \text{ET}_{ijk;l} + \sum \text{EvdW}_{ijk;l} + \sum \text{EQ}_{ij}, \quad (1)$$

где суммирование проводится по группам атомов. Слагаемые в выражении (1) имеют следующий смысл.

Энергия электростатического взаимодействия между атомами с парциальными зарядами q_i и q_j (в единицах элементарного заряда):

$$\text{EQ}_{ij} = 332.0716 \frac{q_i q_j}{R_{ij}}, \quad (2)$$

где R_{ij} — расстояние между атомами (здесь и далее единицей длины является ангстрем).

Энергия взаимодействия Ван-дер-Ваальса:

$$\text{EvdW}_{ij} = \varepsilon_{ij} \left(\frac{1.07 R_{ij}^*}{R_{ij} + 0.07 R_{ij}^*} \right)^7 \left(\frac{1.12 R_{ij}^{*7}}{R_{ij}^7 + 0.12 R_{ij}^{*7}} - 2 \right), \quad (3)$$

где ε_{ij} , R_{ij}^* — параметры взаимодействия, $[\varepsilon_{ij}] = \text{ккал/моль}$, $[R_{ij}^*] = \text{\AA}$.

Энергия растяжения связи:

$$\text{EB}_{ij} = 143.9325 \frac{kb_{ij}}{2} \Delta r_{ij}^2 \left(1 + cs \Delta r_{ij} + \frac{7}{12} cs^2 \Delta r_{ij}^2 \right), \quad (4)$$

где $\Delta r_{ij} = r_{ij} - r_{ij}^{(0)}$, $cs = -2 \text{\AA}^{-1}$, r_{ij} , $r_{ij}^{(0)}$ — рассчитанная и равновесная длины связи, kb_{ij} — силовая постоянная растяжения связи, $[kb_{ij}] = \text{\AA}^{-2} \text{ккал/моль}$.

Энергия деформации валентного угла:

$$\text{EA}_{ijk} = 0.043844 \frac{ka_{ijk}}{2} \Delta \vartheta_{ijk}^2 (1 + cb \Delta \vartheta_{ijk}), \quad (5)$$

где $\Delta \vartheta_{ijk} = \vartheta_{ijk} - \vartheta_{ijk}^{(0)}$, где ϑ_{ijk} , $\vartheta_{ijk}^{(0)}$ — рассчитанный и равновесный валентные углы (здесь и далее единицей угла является градус), $cb = -0.0070 \text{ град}^{-1}$, ka_{ijk} — силовая постоянная деформации валентного угла. Если равновесное значение валентного угла равно 180° , то $\text{EA}_{ijk} = 143.9325 ka_{ijk} (1 + \cos \Delta \vartheta_{ijk})$.

Энергия перекрестного взаимодействия, обусловленная одновременным растяжением связи и деформацией валентного угла:

$$\text{ESB}_{ijk} = 2.51210 (kba_{ijk} \Delta r_{ij} + kba_{kji} \Delta r_{kj}) \Delta \vartheta_{ijk}, \quad (6)$$

где kba_{kij} , kba_{ijk} — силовые постоянные для одновременного растяжения связи и деформации валентного угла, $[kba_{kij}, kba_{ijk}] = \text{град}^{-1} \text{\AA}^{-1} \text{ккал/моль}$.

Энергия вращения вокруг связи:

$$\text{ET}_{ijk;l} = 0.5 (V_1 (1 + \cos \phi) + V_2 (1 + \cos 2\phi) + V_3 (1 + \cos 3\phi)), \quad (7)$$

где V_1 , V_2 , V_3 — силовые постоянные вращения вокруг связи, $[V_1] = [V_2] = [V_3] = \text{ккал/моль}$, ϕ — угол вращения.

Энергия взаимодействия, обусловленная выходом атома из плоскости:

$$\text{EOOP}_{ijk;l} = 0.043844 \frac{\text{koop}_{ijk;l}}{2} \chi_{ijk;l}^2, \quad (8)$$

где χ — угол между (jl) и плоскостью (jkl) (угол Вильсона), $\text{коор}_{ijk;l}$ — силовая постоянная деформации угла Вильсона, $[\text{коор}_{ijk;l}] = \text{град}^{-2} \text{ ккал/моль}$.

Слагаемые (4)–(8) в (1) описывают энергию валентных взаимодействий, (2), (3) — невалентных. При расчете энергии валентных взаимодействий используются естественные координаты, к которым в рамках MMFF94 относятся длины связей, валентные и торсионные углы и угол Вильсона. Естественные координаты рассчитываются по декартовым координатам группы атомов, состоящей из двух атомов для длины связи, из трех — для валентного угла и из четырех — для торсионного угла и угла Вильсона. Соответственно валентные параметры приписываются не отдельным атомам, а сразу группе атомов.

В рамках MMFF94 для расстановки силовых параметров необходимо выполнить два этапа. Первый этап заключается в типизации (т.е. в расстановке типов) всех атомов, входящих в данную молекулу. Каждому атому приписывается целое число от 1 до 99, которое называется типом атома. Тип атома зависит от порядкового номера соответствующего элемента в таблице Менделеева и локального химического окружения атома в молекуле. Второй этап заключается в генерации всех групп атомов, для которых будут рассчитываться естественные координаты. Для валентных взаимодействий — это группы, образующие длины связей, валентные и торсионные углы и угол Вильсона. Группы, генерируемые для расчета длин связи, используются также для расстановки парциальных зарядов на атомах. Величины силовых параметров, приписываемых группам, расставляются с учетом типов атомов, входящих в состав группы, а также типа химической связи (одинарная, двойная, тройная) между атомами в группе. Исключение составляют параметры, используемые для расчета энергии взаимодействия Ван-дер-Ваальса, которые определяются только типом атома. Далее для каждой группы осуществляется поиск соответствующего ей параметра в файлах с параметрами. Если требуемый параметр отсутствует, то он вычисляется по определенным правилам [5].

Ниже алгоритм выполнения этих этапов рассмотрен более подробно.

3. Типизация атомов. В рамках силового поля MMFF94 предусмотрено 99 типов атомов для 19 элементов из таблицы Менделеева: водород, углерод, азот, кислород, сера, фосфор, кремний, фтор, бром, йод, хлор, литий, калий, натрий, кальций, цинк, железо, магний и медь. Наиболее детально типы разработаны для самых распространенных в органических соединениях элементов — водорода, углерода, кислорода и азота. Помимо номера элемента для определения типа атома необходима информация о его химическом окружении, которая целиком содержится в матрице связности структуры. Матрица связности наряду с номером элемента является необходимой входной информацией для проведения типизации, генерации групп для валентных взаимодействий и строится следующим образом. Для любого i -го атома, имеющего n химических связей, задается n троек чисел, первое из которых одинаково для всех троек и равно i , второе равно номеру атома j , с которым i -й атом имеет химическую связь, и третье число задает тип связи и равно 1, 2 или 3 в случае одинарной, двойной или тройной связи соответственно. При записи матрицы связности в файл структуры молекулы для исключения повторений записываются только те тройки, для которых $i < j$. Отметим, что программа FARS строит матрицу связности только для белков (алгоритм построения рассмотрен в разделе 3.2), а для лиганда матрица связности должна быть задана пользователем, т.е. должна присутствовать в файле, задающем структуру лиганда.

Правила, по которым проводится типизация, детально изложены в [1–5]. В нашей работе мы будем упоминать лишь те их особенности, которые существенны для построения алгоритма на основе этих правил.

В рамках MMFF94 не существует отдельных типов для атомов, входящих в состав стандартных аминокислотных остатков, из которых строятся белки. Таким образом, правила, по которым приписываются типы атомам лигандов и белков, совершенно одинаковы. Однако с практической точки зрения удобно на основе этих правил построить различные алгоритмы типизации для белков и лигандов. Основная особенность молекулы белка с точки зрения типизации заключается в том, что она состоит из большого числа (до нескольких десятков тысяч) атомов и в то же время каждый из этих атомов входит в состав того или иного аминокислотного остатка. Поэтому оказывается удобным заранее (один раз) провести типизацию всех аминокислотных остатков, составить библиотеки, содержащие всю необходимую информацию, и проводить типизацию произвольного белка на их основе. Аналогичный алгоритм используется и для типизации молекул ДНК и РНК.

В то же время, локальное химическое окружение атомов в лигандах гораздо более разнообразно по сравнению с тем, которое встречается в аминокислотных остатках. В этом случае нет смысла пытаться заранее предусмотреть все возможные структуры и типизация должна проводиться на основе общего алгоритма с учетом правил MMFF94. Однако число атомов N в лигандах не превышает нескольких сотен, что снижает требования к сохранению быстродействия алгоритма с ростом N по сравнению со

случае его применения к белкам.

В силу сказанного алгоритмы типизации атомов белка (и молекул ДНК и РНК) и лиганда различны и рассмотрены ниже отдельно.

3.1. Типизация атомов лигандов. В случае лигандов типизация атомов проводится на трех уровнях. На первом уровне типы атомов определяются только номером Z соответствующих элементов в таблице Менделеева со следующими исключениями. По два типа приходятся на атомы фтора, хлора и брома, которые могут находиться как в составе молекулы, так и виде ионов F^- , Cl^- и Br^- . Кроме того, два типа предусмотрены для ионов Fe^{2+} , Fe^{3+} и Cu^+ , Cu^{2+} . Таким образом, алгоритм типизации первого уровня использует величину Z для всех атомов, проверяет по матрице связности наличие или отсутствие химической связи для атомов фтора, хлора и брома и считывает заряд для атомов железа и меди. Для атомов всех элементов, за исключением H, C, N, O, P и S, типизация первого уровня является окончательной.

Типизация второго уровня проводится с учетом типа и числа химических связей, образуемых данным атомом (одинарная, двойная, тройная), с ближайшими соседями. К примеру, для каждой из возможных для атома углерода гибридизаций предусмотрен отдельный тип. Тип атома водорода зависит от того, с атомом какого элемента он имеет химическую связь.

Алгоритм типизации второго уровня выполняет следующие действия.

Если атом в молекуле не является атомом одного из элементов H, C, N, O, P или S, то тип, присвоенный алгоритмом первого уровня, остается без изменений.

Если атом в молекуле является атомом водорода, то для него проверяется, с каким из элементов (C, N, O, P или S) он имеет химическую связь и ему присваивается соответствующий тип.

Если атом в молекуле является атомом одного из элементов C, N, O, P или S, то тип присваивается в зависимости от числа и вида химических связей (одинарная, двойная, тройная), которые он образует с ближайшими соседями (устанавливается по матрице связности).

Наиболее подробной является типизация третьего уровня, в рамках которой учитывается принадлежность атомов к различным специфическим группам (пептиды, аминокислоты, оксиды азота и т. д.), а также проверяется свойство ароматичности.

Алгоритм типизации третьего уровня выполняет следующие действия.

Если атом в молекуле является атомом одного из элементов C, N, O, P или S, то для него проверяется возможность принадлежности к специфическим группам. Для этого по матрице связности проверяются не только ближайшие, но и вторые и третьи соседи данного атома и виды связей между ними. Если они соответствуют тем, которые установлены для специфической группы, то атому присваивается соответствующий тип.

Если атом в молекуле является атомом одного из элементов C, N, O, P или S, то для него проверяется возможность принадлежности к ароматической подсистеме. Соответствующий алгоритм описан ниже.

Если атом в молекуле является атомом водорода, то для него уточняется тип в соответствии с изменением типа тяжелого атома (на предыдущих двух этапах типизации третьего уровня), с которым он имеет химическую связь.

Проверка принадлежности атома к ароматической подсистеме проводится в два этапа. На первом проверяется, входит ли данный атом в состав циклической структуры, а на втором рассчитывается общее число электронов в системе сопряжения и на основе правила Хюккеля делается вывод о том, является ли данная система ароматической. Поиск циклических структур осуществляется с помощью оригинального алгоритма Fire, основанного на построении разветвленных цепей, начальным элементом которых является тяжелый атом, имеющий два тяжелых атома в качестве ближайших соседей. Построение цепей осуществляется на основе матрицы связности, причем каждая связь после использования "сгорает", т.е. цепь может вернуться к начальному атому только "с другой стороны", что и означает принадлежность атома к кольцу. Цепи строятся только по тяжелым атомам, поэтому цепь обрывается на атоме, который имеет среди соседей только один тяжелый (и через который он и попал в цепь). Ветвление цепей возникает вследствие возможности существования атомов (как правило, углерода), имеющих среди соседей три и четыре тяжелых. В ходе работы программы отслеживается каждая вновь возникающая цепь. Если атом принадлежит одновременно различным циклам, отслеживание каждой цепи обеспечивает поиск цикла минимального размера, что существенно для правильного присваивания ароматических типов. Соединение циклов по связи, т.е. когда два цикла имеют общими два связанных друг с другом атома (нафталин), также обеспечивает дополнительные возможности для возникновения обладающей свойством ароматичности замкнутой π -системы, поэтому в этом случае атом считается принадлежащим двум циклам. Аналогичное правило заложено в алгоритм и для трех циклов, имеющих два соединения по связи (антрацен).

Блок-схема алгоритма Fire показана на рис. 1.

Если атом принадлежит циклу, проводится расчет общего числа электронов в системе сопряжения в соответствии с выражением

$$N = n_C + n_{N1} + 2n_{N2} + 2n_O + 2n_S + n_{P1} + n_{P2}, \quad (9)$$

где n_C — число атомов углерода в sp^2 гибридизации, имеющих двойную связь с атомом из цикла; n_{N1} — число атомов азота в sp^2 гибридизации, имеющих двойную связь с атомом из цикла; n_{N2} — число атомов азота в sp^3 гибридизации; n_O — число атомов кислорода в sp^3 гибридизации; n_S — число атомов серы в sp^3 гибридизации; n_{P1} — число атомов фосфора в sp^2 гибридизации, имеющих двойную связь с атомом из цикла; n_{P2} — число атомов фосфора в sp^3 гибридизации.

Выражение (9) получено нами на основе обобщения результатов расчета числа электронов в замкнутой системе сопряжения для ароматических углеводородных циклов различной топологии и гетероциклов с одним и несколькими гетероатомами (фуран, тиофен, пиридин, пиримидин, оксазол и др.). При вычислении числа N по формуле (9) учитывается, что атом может входить в состав нескольких циклов различного размера.

Для того чтобы система сопряжения являлась ароматической, полное число электронов N в ней должно удовлетворять правилу Хюккеля (см. [12])

$$N = 4n + 2, \quad n = 1, 2, 3, \dots \quad (10)$$

Однако даже если число электронов в системе сопряжения удовлетворяет правилу (10), она не является ароматической в случае, если ее непрерывность нарушается. Такое нарушение имеет место, если в составе цикла есть атомы:

- углерода в sp^3 гибридизации;
- углерода в sp^2 гибридизации, который имеет двойную связь с атомом, не входящим в состав цикла;
- четырехкоординированный атом серы.

В этих случаях, даже если число электронов удовлетворяет правилу (10), цикл не является ароматическим. Если система сопряжения непрерывна и число электронов удовлетворяет правилу (10), типизация всех атомов цикла проводится с учетом ароматичности.

Типизация третьего уровня является окончательной для всех атомов.

3.2. Типизация атомов белка. Типизация аминокислотных остатков проводится отдельно для каждого аминокислотного остатка по общему алгоритму, описанному в разделе 3.1. Далее, по результатам применения общего алгоритма составляется библиотека, содержащая всю необходимую информацию для типизации и генерации групп атомов для каждого из двадцати стандартных аминокислотных остатков, которые могут встречаться в белках. Возможность нахождения остатков в различных зарядовых состояниях принимается во внимание следующим образом. Если остаток является донором протона, соответствующий элемент матрицы связности (одинарная связь между атомом кислорода и водорода) убирается; если остаток является акцептором протона, соответствующий элемент матрицы связности (одинарная связь между атомом азота и водорода) добавляется.

Библиотека представляет собой ряд записанных последовательно блоков, каждый из которых имеет следующую структуру: название аминокислотного остатка, число атомов и число химических связей,

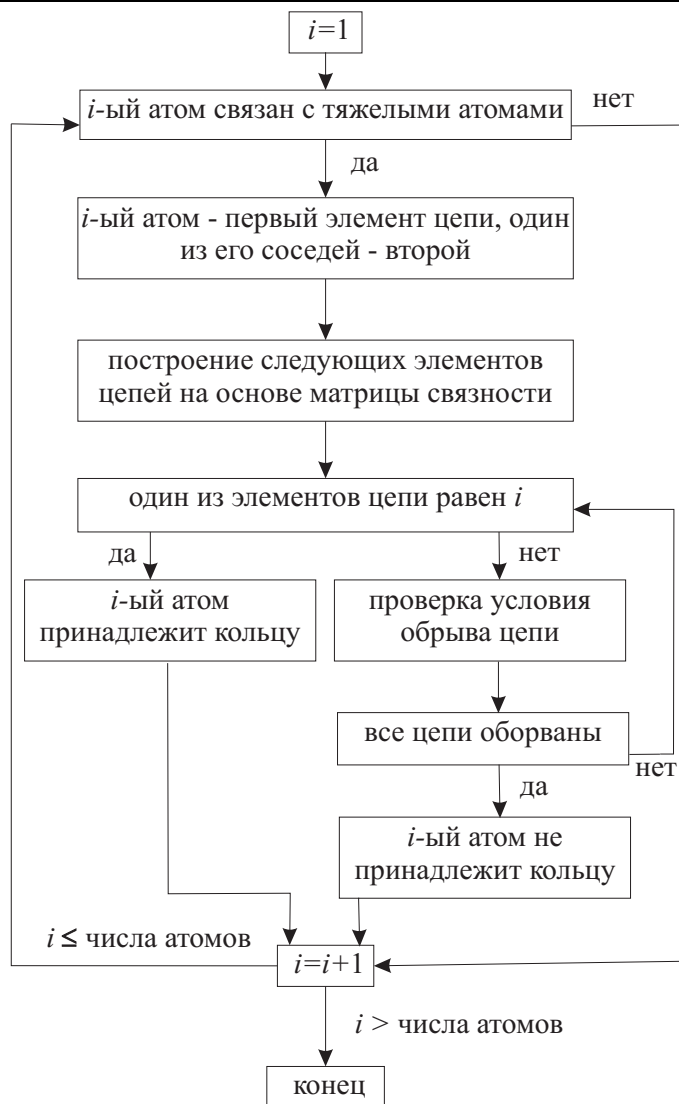


Рис. 1. Блок-схема алгоритма Fire

имя атома в соответствии с классификацией PDB [13], имена атомов, с которыми данный атом имеет химическую связь, и тип этой связи (одинарная или двойная). При загрузке в программу файла белка в формате pdb для каждого из аминокислотных остатков белка в соответствии с его именем должна быть считана матрица связности из библиотеки и для каждого атома считан его тип. Далее, при формировании полной матрицы связности белка необходимо учесть существование химической связи между С-атомом предыдущего аминокислотного остатка и N-атомом последующего. Кроме того, в белке существуют химические связи, называемые мостиковыми, между атомами серы, относящимися к различным аминокислотным остаткам. Такая связь считается существующей, если два атома серы, принадлежащие различным аминокислотным остаткам, являются однокоординированными и расстояние между ними не превышает 2.4 \AA .

Результатом выполнения всех этих действий является формирование матрицы связности белка и расстановка типов атомов.

3.3. Типизация атомов РНК и ДНК. Типизация молекул РНК и ДНК проводится по алгоритму, сходному с описанным в разделе 3.2. Заранее проводится типизация всех нуклеотидов, встречающихся в РНК и ДНК, — аденозина (A), гуанозина (G), цитидина (C), тимидина (T) и уридина (U) по общему алгоритму. После проведения типизации составляется библиотека, представляющая собой ряд записанных последовательно блоков, каждый из которых имеет следующую структуру: название нуклеотида, число атомов и число химических связей, имя атома в соответствии с классификацией PDB [9] и имена атомов, с которыми данный атом имеет химическую связь, и тип этой связи (одинарная или двойная). Если при загрузке файла формата pdb встретится имя одного из нуклеотидов (A, G, C, T, U), то происходит обращение к соответствующему блоку библиотеки, из которого считывается матрица связности. Далее строится общая матрица связности молекулы ДНК (РНК) с учетом химических связей между нуклеотидами. После этого проводится типизация по общему алгоритму.

4. Расстановка силовых параметров. Типизация является необходимым и достаточным условием для расстановки параметров, используемых при расчете энергии взаимодействия Ван-дер-Ваальса. Каждому типу атомов соответствуют индивидуальные параметры, на основе которых вычисляются парные параметры ε_{ij} , R_{ij}^* , входящие в (3) по формулам, приведенным в [1].

После типизации для расстановки остальных силовых параметров прежде всего необходимо сгенерировать все группы атомов, которые необходимы для описания валентных взаимодействий.

Группы, по которым вычисляются длины связей r_{ij} , генерируются следующим образом. Для каждого атома j , имеющего n_j соседей, с которыми он связан ковалентной связью, составляется n_j двоек, первое число в которых одинаково и равно j , а второе равно номеру атома i , с которым он связан ковалентной связью, записанной в матрице связности. Во избежание повторений учитываются только двойки с $j > i$. Отметим, что группы, сгенерированные для расчета длин связей, используются для расстановки парциальных зарядов q_i на атомах. Это обусловлено использованием в MMFF94 метода зарядовых инкрементов для расчета q_i : $q_i = \sum w_{ij}$, где сумма берется по всем атомам, с которым атом i имеет ковалентную связь, и w_{ij} — зарядовый инкремент, определяемый типами атомов i и j .

При генерации групп, по которым вычисляются валентные углы ϑ_{ijk} , вначале осуществляется поиск всех атомов, имеющих не менее двух соседей. Для каждого из таких атомов j генерируются тройки атомов $i - j - k$, где i, k — номера атомов, с которыми атом j связан ковалентной связью. Во избежание повторений учитываются только те тройки, для которых $i \leq k$ (равенство $i = k$ возможно в случае присутствия трехатомного цикла в составе молекулы). Генерация четверок атомов, по которым вычисляются торсионные углы ϕ , осуществляется на основе двоек, сгенерированных для длин связей r_{ij} . Выбираются все двойки $i - j$, в которых атом i имеет связь с атомами $n_i \neq j$ и атом j имеет связь с атомами $n_j \neq i$. Далее генерируются все четверки вида $n_i - i - j - n_j$, для которых $n_i \leq n_j$ (равенство $n_i = n_j$ возможно в случае присутствия четырехатомного цикла в составе молекулы).

Генерация четверок, по которым будет рассчитываться угол Вильсона $\chi_{ijk;l}$, осуществляется следующим образом. Вначале составляется список всех атомов молекулы, имеющих три ковалентных связи. Предположим, что атом i имеет связи с атомами jkl . В каждом таком случае генерируется четыре четверки $i, jkl; l, ijk; k, ijl; j, ikl$, для каждой из которых будет вычислен свой угол Вильсона. Силовой параметр $\text{коор}_{ijk;l}$ будет одним и тем же для всех четырех углов в соответствии с правилами MMFF94.

После того как все группы сгенерированы, необходимо проанализировать виды связей между атомами внутри группы. Необходимость такого анализа поясним на примере определения равновесной длины связи $r_{ij}^{(0)}$ и силовой постоянной растяжения этой связи kb_{ij} для двух атомов углерода в sp^2 гибридизации. Связь между этими атомами может быть как одинарной, так и двойной, при этом тип атомов один и тот же (рис. 2), однако величины $r_{ij}^{(0)}$ и kb_{ij} будут различны.

Аналогичные ситуации могут возникать и в других группах.

По результатам анализа связей внутри групп каждой группе присывается переменная, называемая индексом связи, правила вычисления которой изложены в [1–5]. Таким образом, присвоение параметров группе происходит как с учетом типов атомов, входящих в нее, так и с учетом величины индекса связи.

Далее для каждой группы происходит поиск силового параметра, соответствующего данной группе с данными типами атомов и данным значением индекса связи в файлах с параметрами. Приведем пример фрагмента такого файла, содержащего величины $r_{ij}^{(0)}$ и kb_{ij} :

```
.....
0      2      2      9.505      1.333
1      2      2      5.310      1.430
.....
```

В первом столбце записано значение индекса связи, во втором и третьем — типы атомов i и j соответственно, в четвертом и пятом — величины kb_{ij} и $r_{ij}^{(0)}$ соответственно. В приведенном примере значение индекса связи 1 соответствует одинарной связи между атомами углерода в sp^2 гибридизации, 0 соответствует двойной связи (см. рис. 2).

В случае если параметр для группы с данными характеристиками отсутствует в файлах с параметрами, то он вычисляется по определенным правилам [1]. Такие правила предусмотрены для вычисления всех видов связевых параметров.

Этот порядок действий сохраняется как в случае белка, так и в случае лиганда, однако для белка анализ видов связи между атомами в группе существенно проще вследствие ограниченного разнообразия возможных структур. Кроме того, для белков не нужно использовать процедуру вычисления параметров, поскольку для всех возможных структур в аминокислотных остатках они содержатся в файлах с параметрами. Блок-схема изложенного выше алгоритма приведена на рис. 3.

5. Программная реализация алгоритма расстановки силовых параметров. Тестирование программы. Описанный выше алгоритм расстановки силовых параметров реализован в программе FARS, входящей в состав веб-ориентированного программного комплекса Keenbase [11].

Тестирование программы FARS проводилось в два этапа. На первом этапе проверялся результат типизации программой FARS тестового набора небольших молекул, в который входили стандартные аминокислоты, гетероциклы, молекулы с функциональными группами (аминогруппы, пептиды, сульфогруппы и др.). Для каждой молекулы вручную проверялось соответствие расставленных типов правилам MMFF94.

На втором этапе тестирование программы FARS для лигандов осуществлялось с использованием базы данных NCI Diversity [10], из которой были выбраны 1998 молекул с различными структурными мотивами. Для каждой молекулы из базы данных сравнивались типы, присвоенные атомам программой FARS и программным комплексом CHARMM [14], в котором также имеется алгоритм расстановки силовых параметров в соответствии с силовым полем MMFF94. Кроме того, почленно сравнивались все слагаемые потенциальной энергии, вычисленные FARS и CHARMM, входящие в выражение (1). В случае расхождения между CHARMM и FARS результаты анализировались вручную на соответствие правилам MMFF94. Наиболее существенной ошибкой CHARMM в расстановке силовых параметров, выявленной в результате такого анализа, является неверная величина суммарного заряда атомов двух гуанидиновых групп, имеющих общий атом азота (рис. 4): $q = 1.66667$, в то время как правильное значение $q = 2.0$, что совпадает с результатом FARS. Кроме того, выявлена неверная расстановка типов для ароматических

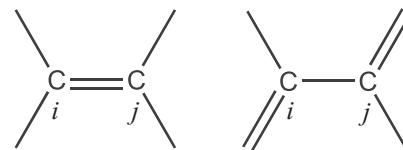


Рис. 2. Одинарная и двойная связь между двумя атомами углерода в sp^2 гибридизации

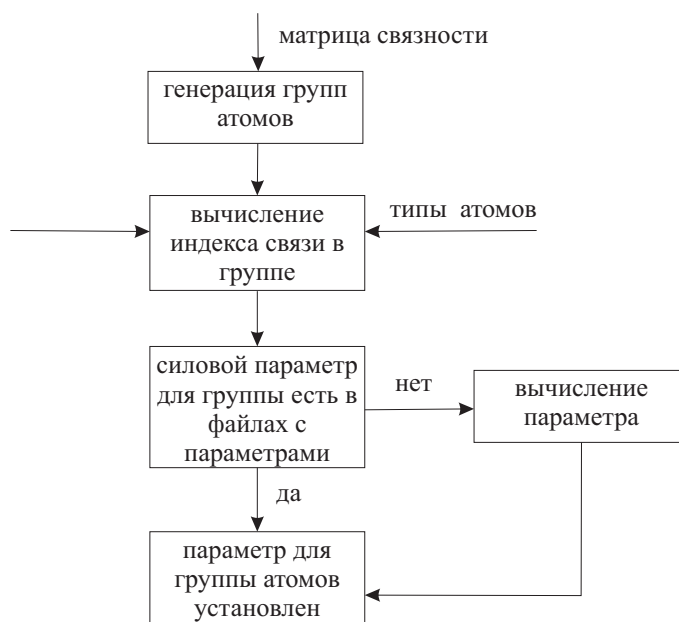


Рис. 3. Блок-схема расстановки параметров после проведения типизации атомов

структур, состоящих из двух соединенных по связи циклов с гетероатомами, содержащими более одного атома азота (рис. 4). В соответствии с правилом Хюккеля все тяжелые атомы в циклических структурах, показанных на рис. 4, должны быть ароматическими, в то время как CHARMM присваивает четырем атомам шестиатомного цикла — тем из них, которые не являются общими с пятиатомным циклом, — неароматические типы.

Тестирование программы FARS

для белков проводилось с использованием 59 комплексов белок–лиганд, взятых из базы данных PDB [15]. Атомы водорода были расставлены с использованием программы Reduce [16]. Кроме того, как и для лигандов, для каждого комплекса сравнивались типы атомов и почленно величина потенциальной энергии, рассчитанной FARS и CHARMM. Не было выявлено ни одного случая расхождения. Отметим также, что алгоритм расстановки силовых параметров, реализованный в CHARMM, прерывает выполнение программы в случае обнаружения ошибок в матрице связности лиганда, что требует вмешательства пользователя для проведения процедуры скрининга для данного лиганда. FARS выдает сообщение об ошибке, однако присваивает атомам лиганда параметры первого уровня и проводит скрининг с этими параметрами. После окончания скрининга пользователь, если сочтет необходимым, может исправить ошибку в матрице связности и повторить скрининг для этих лигандов с уточненными параметрами. Такой способ обработки ошибок матрицы связности предпочтительнее в случае расчета большого числа молекул.

Время работы алгоритма FARS для белков составляет несколько секунд, для лигандов — несколько десятков секунд. Увеличение времени работы FARS для лигандов связано с тщательным поиском всех возможных циклов алгоритмом Fire. Для задачи виртуального скрининга такое время работы является допустимым, так как основные временные затраты в этом случае связаны с расчетом потенциальной энергии взаимодействия для большого числа взаимных ориентаций белка и лиганда. Проведенный анализ показал, что оптимизация алгоритма Fire, связанная с предварительным отсечением крайне маловероятных структур, приводит к уменьшению времени работы на порядок и делает возможным его применение к молекулам лигандов с числом атомов до 1000. Кроме того, возможна дальнейшая оптимизация алгоритма Fire.

6. Применение программы FARS при решении задач скрининга с использованием технологий распределенных вычислений. Программа FARS была задействована при проведении компьютерного эксперимента в период с 31 декабря 2005 г. по 11 января 2006 г. по поиску ингибиторов для заданных белков-мишеней через систему распределенных вычислений X-Com. Поиск ингибиторов осуществлялся с использованием программы докинга SOL [11], в которую была встроена FARS в качестве отдельной подпрограммы. В скрининге участвовали два белка — тромбин, играющий важную роль в процессе свертываемости крови, и ВИЧ-протеаза, с помощью которой происходит воспроизводство вируса иммунодефицита человека, а также 2123 органических молекул — лигандов. С использованием программы FARS были расставлены силовые параметры на белках и лигандах. При проведении расчетов в течение 12 дней были задействованы кластерные и компьютерные мощности НИВЦ МГУ [17] и кластерные мощности Южно-Уральского государственного университета [18]. Общее процессорное время составило 42774 часа, общее число работавших процессоров — 273 с пиковой производительностью более 1 Tflops. За это время было обработано 10615 заданий со средним временем обработки 3–4 часа. На основании проведенных расчетов для тромбина были выбраны наиболее перспективные кандидаты в ингибиторы, из них 30 были заказаны в базе данных Национального института рака (США) NCI [19] и получены для проведения экспериментов. Измерения всех этих 30 соединений в Гематологическом научном центре РАМН (лаборатория Ф.И. Атауллаханова) выявили среди них восемь новых ингибиторов тромбина, два из которых имели высокую ингибирующую активность. В ходе компьютерного эксперимента не было зафиксировано ни одного случая отказа программы FARS.

7. Заключение. В настоящей работе описан алгоритм FARS (Fire Algorithm for Rings Search) расстановки силовых параметров на атомах белков и органических молекул малого и среднего размера, выполненный в рамках силового поля MMFF94 [1–5]. С использованием базы данных NCI Diversity [10] проведено сравнительное тестирование алгоритма FARS и аналогичного алгоритма, реализованного в

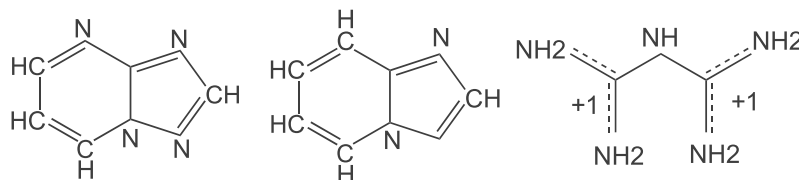


Рис. 4. Структуры, для которых выявлены ошибки типизации и расстановки параметров в CHARMM

программном комплексе CHARMM [14], широко используемом в настоящее время для целей молекулярного моделирования. Программа FARS использовалась при проведении компьютерного эксперимента по поиску ингибиторов для заданных белков-мишеней через систему распределенных вычислений X-Com. В ходе эксперимента были задействованы кластерные и компьютерные мощности НИВЦ МГУ и кластерные мощности Южно-Уральского государственного университета. В течение 12 дней было обработано 10615 заданий со средним времени обработки 3–4 часа. Отказов программы FARS в ходе проведения экспериментов зафиксировано не было.

Программа FARS включена в состав веб-ориентированной платформы Keenbase [11], разработанной в лаборатории вычислительных систем и прикладных технологий программирования НИВЦ МГУ в рамках Государственного контракта НИВЦ МГУ с Федеральным агентством по науке и инновациям № 02.435.11.1008 от 01 августа 2005 г. Работа поддержана РФФИ (код проекта 06-03-33171).

СПИСОК ЛИТЕРАТУРЫ

1. *Halgren T.A.* Merck Molecular Force Field I: Basis, form, scope, parameterization and performance of MMFF94 // *J. of Comput. Chem.* 1996. **17**, N 5–6. 490–519.
2. *Halgren T.A.* Merck Molecular Force Field II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions // *J. of Comput. Chem.* 1996. **17**, N 5–6. 520–552.
3. *Halgren T.A.* Merck Molecular Force Field III. Molecular geometries and vibrational frequencies for MMFF94 // *J. of Comput. Chem.* 1996. **17**, N 5–6. 553–586.
4. *Halgren T.A.* Merck Molecular Force Field IV. Conformational energies and geometries for MMFF94 // *J. of Comput. Chem.* 1996. **17**, N 5–6. 587–615.
5. *Halgren T.A.* Merck Molecular Force Field V. Extension of MMFF94 using experimental data, additional computational data and empirical rules // *J. Comput. Chem.* 1996. **17**, N 5–6. 616–641.
6. *Li X., Volkov A.V., Szalewicz K., Coppens P.* Interaction energies between glycopeptide antibiotics and substrates in complexes determined by X-ray crystallography: application of a theoretical databank of aspherical atoms and a symmetry-adapted perturbation theory-based set of interatomic potentials // *Acta Cryst.* 2006. **D62**. 639–647.
7. *Lane J.W., Estevez A., Mortara K., Callan O., Spencer J.R., Williams R.M.* Antitumor activity of tetrahydroisoquinoline analogues 3-epi-jorumycin and 3-epi-renieramycin G // *Bioorganic & Medicinal Chemistry Letters*. 2006. **16**, N 12. 3180–3183.
8. *Cini N., Danieli E., Menchi G., Trabocchi A., Bottoncetti A., Raspanti S., Pupi A., Guarna A.* 3-Aza-6,8-dioxabicyclo[3.2.1]octanes as new enantiopure heteroatom-rich tropane-like ligands of human dopamine transporter // *Bioorganic & Medicinal Chemistry Letters*. 2006. **14**, N 15. 5110–5120.
9. *Kasyan O., Thondorf I., Bolte M., Kalchenko V., Bohmer V.* Unusual conformations of 1,3-dialkoxythiacalix[4]arenes in the solid state // *Acta Cryst.* 2006. **C62**. 289–294.
10. <http://dtp.nci.nih.gov/>
11. <http://www.keenbase.ru>
12. *Ким А.М.* Органическая химия. Новосибирск: Сиб. универ. изд-во, 2002.
13. http://www.rcsb.org/pdb/docs/format/pdbguide2.2/guide2.2_frame.html
14. *Brooks B.R., Brucoleri R., Olafson B., States D., Swaminathan S., Karplus M.* CHARMM: A program for macromolecular energy, minimization, and dynamics calculations // *J. Comp. Chem.* 1983. **4**, N 5. 187–217.
15. <http://www.rcsb.org/pdb/Welcome.do>
16. *Word J.M.* Asparagine and glutamine: Using hydrogen atom contacts in the choice of side-chain amide orientation // *J. Mol. Biol.* 1999. **285**. 1733–1747.
17. <http://parallel.ru/cluster/>
18. <http://cluster.susu.ru/>
19. <http://www.cancer.gov/>

Поступила в редакцию
28.09.2006