УДК 536.75; 538.9

КОМПЬЮТЕРНЫЙ ДИЗАЙН ЛЕКАРСТВЕННЫХ СРЕДСТВ: ПРОГРАММА ДОКИНГА SOL

А. Н. Романов[1], О. А. Кондакова[2], Ф. В. Григорьев[1], А. В. Сулимов[2], С. В. Лущекина[1], Я. Б. Мартынов[1], В. Б. Сулимов[1]

Описан принцип функционирования и первоначальное тестирование программы позиционирования (докинга) низкомолекулярных лигандов в активном центре протеинов и других биологических объектов. Подобные процедуры широко используются в ходе структурно-ориентированного процесса разработки новых лекарственных средств. Для оптимизации положения лиганда используется генетический алгоритм, при этом оптимизируются внутренние степени свободы лиганда (разрешенные вращения вокруг одинарных химических связей), а также степени свободы, связанные с положением лиганда как целого в активном центре биологической макромолекулы. Оптимизация положения лиганда производится по результатам оценки энергии взаимодействия лиганда с макромолекулой, а также оценки внутренней энергии лиганда. Вычисление энергии осуществляется в рамках модели силового поля MMFF94. Эффекты десольватации в процессе образования комплекса учитываются с использованием обобщенной борновской модели. Программа также оценивает свободную энергию образования комплекса лиганда с макромолекулой и осуществляет кластеризацию найденных решений, основываясь на сходстве соответствующих им пространственных положений лиганда. Первоначальное тестирование программы на примере докинга набора лигандов в несколько популярных биологических мишеней выявило ее способность корректно располагать лиганды в активном центре протеинов, а также производить виртуальный скрининг — отбирать активные соединения из содержащего их набора и соединения, не проявляющие активности на данной биомишени. Работа выполнена при финансовой поддержке РФФИ (код проекта 06–03–33171).

Ключевые слова: компьютерный дизайн лекарственных средств, докинг лигандов, генетический алгоритм, биологическая мишень, глобальная оптимизация, силовое поле MMFF94.

1. Введение. В настоящее время для разработки одного нового лекарственного вещества требуется несколько сотен миллионов долларов США, а период его создания, как правило, занимает не менее 10 – 15 лет. Основной принцип создания лекарственных веществ заключается в следующем. Для многих болезней известны белки, определяющие развитие заболевания. Эти белки могут служить мишенями лекарственного вещества, т.е. молекула лекарственного соединения связывается с определенными участками мишени и изменяет его работу, что положительным образом сказывается на ходе заболевания. Разработка новых органических соединений, способных избирательно связываться с выбранными биологическими мишенями, является ключевым этапом для всего долгого и дорогого процесса создания лекарства.

С конца 80-х годов прошлого века для поиска новых химических соединений, являющихся потенциальными лекарствами, активно используются методы экспериментального высокопроизводительного скрининга (High-Throughput Screening — HTS) и комбинаторного синтеза. Современные роботизированные системы способны синтезировать до миллиона соединений в месяц. Предполагалось, что внедрение подобных технологий приведет к лавинообразному росту количества соединений-лидеров и, соответственно, к существенному увеличению количества новых лекарств, поступающих на рынок. Однако реальная ситуация опровергла подобное предположение. Многие найденные соединения не могли быть в дальнейшем оптимизированы для того, чтобы служить основной лекарственного средства; кроме того, возможности комбинаторного синтеза ограничены весьма узким классом органических соединений. Наряду с другими причинами это привело к тому, что количество оригинальных лекарств, выпускаемых на рынок в последние несколько лет, стало уменьшаться. Данная ситуация заставляет фармацевтические компании и исследовательские группы разрабатывать альтернативные методы поиска новых соединений-лидеров.

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, 119991, Ленинские горы, Москва; e-mail: romanovsuperstar@gmail.com, fedor.grigoriev@gmail.com, vladimir.sulimov@gmail.com

[2] ООО “Димонта”, ул. Нагорная, д. 15, корп. 8, 117186, Москва; e-mail: vladimir.sulimov@gmail.com

В результате на передний план выходят методы молекулярного моделирования, позволяющие сократить материальные и временны́е затраты на начальной стадии создания новых лекарственных средств и повысить эффективность разработок [1, 2].

В настоящее время существует множество программ для моделирования лекарственных соединений, которые создаются как в недрах больших фармацевтических корпораций, так и в лабораториях университетов. В то же время в среде медицинских химиков растет потребность в универсальных платформах, позволяющих не только проводить конкретные расчеты взаимодействия нового соединения с биомишенью, но и одновременно анализировать и модернизировать отобранные ранее разными способами соединения. На сегодняшний день существует только несколько подобных платформ, способных предоставить исследователям необходимые возможности. Примерами таких систем являются программные комплексы Sybyl (продукция компании Tripos) и Discovery Studio (продукция компании Accelrys). Однако эти и подобные им продукты, созданные в крупных фармацевтических компаниях, обладают рядом недостатков, в частности высокой стоимостью, которая не позволяет приобретать их небольшим исследовательским коллективам. Кроме того, многие из подобных программных продуктов делаются максимально универсальными и удобными для пользователя в ущерб научной наполненности составляющих их компонентов. Таким образом, на сегодняшний день актуальной остается задача создания высокотехнологичной компьютерной платформы для применения в области молекулярного моделирования при разработке новых лекарств.

Существует целый ряд обзоров, посвященных подробному описанию различных методов применения вычислительных технологий для поиска молекул-ингибиторов для различных белков-мишеней [3 – 6]. Настоящая статья посвящена одному из таких методов, а именно докингу — поиску положения молекулы-ингибитора, или лиганда (термин “лиганд” происходит от латинского слова ligare — связываться, т.е. лиганд — это молекула, которая должна связаться) в активном центре белка-мишени, соответствующего наибольшему значению свободной энергии связывания ингибитора с белком. С математической точки зрения эта задача сводится к поиску глобального минимума функции свободной энергии, заданной на многомерном гиперпространстве. С физико-химической точки зрения должно проводиться моделирование межмолекулярного взаимодействия в водных растворах с учетом как энтальпийной, так и энтропийной составляющих свободной энергии связывания. Сущность вычислительного аспекта проблемы заключается в необходимости вычислять энергию связывания лиганд–белок с высокой точностью порядка 1 – 2 ккал/моль или 0.05 – 0.1 эВ и с высокой производительностью, позволяющей оперативно проводить расчеты по крайней мере десятков тысяч соединений.

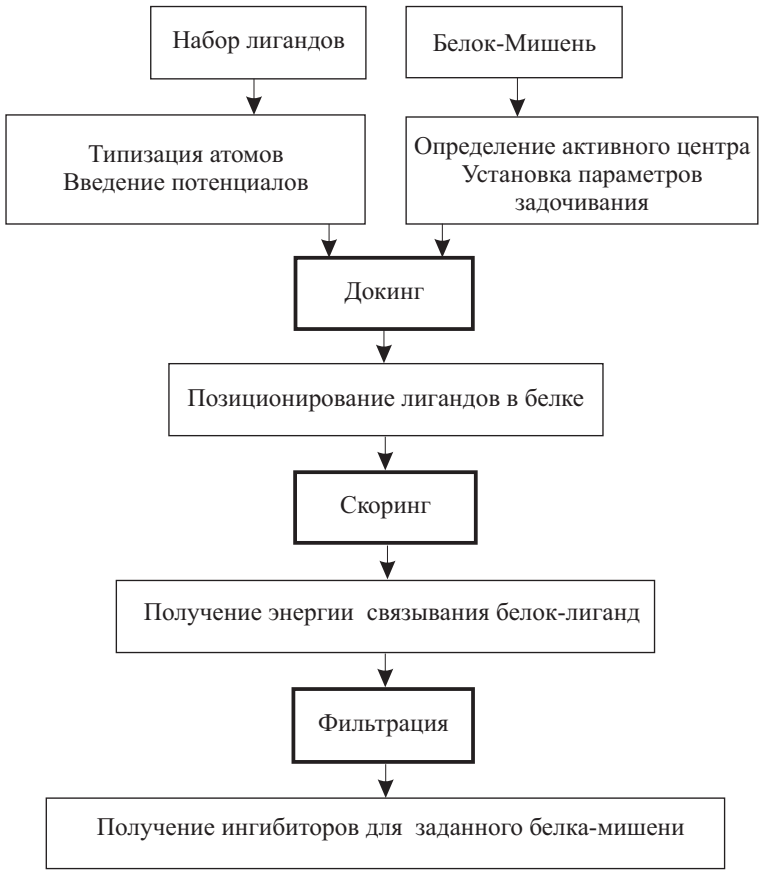

Рис. 1. Общая схема работы системы KeenBASE

Процедура докинга разбивается на две составляющие: поиск наилучшего положения лиганда и оценка свободной энергии связывания белка с лигандом в найденном оптимальном положении, т.е. расчет целевой (скоринг) функции. Одним из наиболее перспективных на сегодняшний день алгоритмов докинга является генетический алгоритм, поскольку по данным [5] наибольшей популярностью в сфере компьютерной разработки лекарств пользуется программа AutoDock [7 – 9], в которой используется именно генетический алгоритм поиска глобального минимума. Несмотря на то что начало работ по генетическим алгоритмам оптимизации относится к 60-м годам прошлого века, они получили популярность только в начале 70-х

после опубликования работы Джона Холланда [10]. Однако вплоть до середины 80-х годов они применялись лишь для решения модельных задач без значительного практического применения. Повышению интереса к данным методам способствовало широкое распространение персональных компьютеров. К этому времени относится ряд работ, посвященных объяснению поведения генетического алгоритма при успешном нахождении глобального минимума [11 – 13]). Чуть позднее генетический алгоритм начал применяться в программах докинга для компьютерной разработки лекарственных веществ [14 – 17]. В настоящее время генетический алгоритм является, как уже упоминалось ранее, наиболее успешным методом поиска положения лиганда в активном центре протеина. Поэтому при создании программного комплекса KeenBASE [18 – 20] для решения задач молекулярного моделирования именно генетический алгоритм был выбран в качестве основы при создании соответствующей программы докинга.

Программный комплекс KeenBASE был разработан для того, чтобы на начальном этапе разработки находить ингибиторы для заданных белков-мишеней с помощью компьютерного моделирования. Общая схема работы программного комплекса KeenBASE приведена на рис. 1. Легкость в перенастройках создаваемой системы позволяет быстро переключаться на новые белки-мишени и значительно расширить круг потенциальных пользователей. Схема работы системы в общем виде соответствует общепринятой схеме виртуального скрининга. Сложность и точность используемых в системе алгоритмов приводит к тому, что время расчета одного лиганда может составлять от 1 до 100 часов на одном процессоре. Для того чтобы иметь возможность анализировать базы данных, содержащие тысячи молекул, в комплекс была встроена система распределенных вычислений X-Com (разработка лаборатории чл.-корр. РАН В. В. Воеводина, НИВЦ МГУ [21 – 23]). Эта система позволяет одновременно производить расчет группы лигандов на различных кластерных платформах и персональных компьютерах независимо от того, какая операционная система на них установлена.

В настоящей статье рассмотрена программа докинга SOL программного комплекса KeenBASE, ее тестирование (валидация) и результаты применения этих программных средств для разработки нового класса ингибиторов тромбина.

2. Программа докинга.

2.1. Представление протеина в виде набора сеток потенциалов. Для расчета свободной энергии образования комплекса белок–лиганд необходимо провести термодинамическое усреднение по всем возможным микросостояниям комплекса с учетом растворителя. Аналогичное усреднение необходимо провести и для белка, и для лиганда в несвязанном состоянии, однако наиболее сложная часть задачи связана именно с комплексом. Поскольку такое усреднение требует очень значительных вычислительных ресурсов, в рамках существующих алгоритмов докинга предполагается, что в комплексе белок–лиганд имеется выделенное микросостояние, дающее основной вклад в свободную энергию связывания. По существу, генетический алгоритм используется для поиска этого микросостояния, при этом энергия такого микросостояния может быть представлена как функция относительных координат белка и лиганда. В этом случае соответствующий вклад в энтальпийную составляющую свободной энергии образования комплекса белок–лиганд определяется потенциальной энергией взаимодействия белка и лиганда с учетом влияния растворителя в найденном микросостоянии. Энтропийная компонента при расчете свободной энергии связывания учитывается путем добавления положительного слагаемого, величина которого пропорциональна числу вращательных степеней свободы лиганда.

Таким образом, примем, что свободная энергия взаимодействия протеин–лиганд может быть получена путем вычисления неких потенциалов атомов лиганда в поле протеина, растворителя и самого лиганда. Для конструирования таких потенциалов на практике используется несколько подходов. В некоторых случаях потенциалы получают путем анализа частоты встречаемости определенных атом-атомных контактов в протеин-лигандных комплексах с известной структурой. В других случаях используют эмпирические потенциалы. В программе SOL мы опирались на так называемые “физические потенциалы”, т.е. на потенциалы взаимодействия, полученные на основании физических представлений о межмолекулярных и внутримолекулярных взаимодействиях. Следует отметить, что наиболее последовательным способом моделировать такие взаимодействия (Ван-дер-Ваальсовы взаимодействия, электростатические взаимодействия, энергия внутреннего напряжения в молекуле лиганда) можно при помощи методов квантовой химии. Однако огромный объем вычислений, необходимый для этого, не позволяет напрямую использовать даже наиболее простые из них при проведении докинга.

По этим причинам часто пользуются моделями “силового поля”, надлежащим образом откалиброванными по результатам экспериментальных наблюдений или квантово-химических расчетов. Подобные модели предполагают наличие “типизации” атомов в органическом соединении, причем каждому типу атомов приписываются некоторые параметры силовых взаимодействий как с ближайшими соседями, с которыми

атом соединен химическими связями (энергии деформации связей, простых и торсионных углов молекулы), так и с дальним окружением (так называемые несвязанные взаимодействия — электростатические и Ван-дер-Ваальсовы). Используя такие потенциалы, мы можем представить свободную энергию взаимодействия протеин–лиганд в виде слагаемых, обусловленных электростатическим и Ван-дер-Ваальсовым взаимодействиями между протеином и лигандом, энергией внутреннего напряжения лиганда, а также потенциалами, возникающими при вытеснении молекул воды из окружения протеина и лиганда в процессе образования комплекса между ними. В программе SOL мы использовали потенциалы в рамках модели силового поля MMFF94, созданной Т. Халгреном в процессе его работы в фирме Merck [24 – 28]. Данная модель сочетает в себе достаточно качественную параметризацию (основанную на результатах квантовохимического моделирования широкого набора простых органических молекул), гибкость (позволяющую применять ее для разнообразнейших соединений) и удобную процедуру типизации атомов, позволяющую автоматизировать этот процесс, что важно для создания автономного программного продукта.

В процессе докинга протеин представлен в виде жесткой неподвижной конструкции, что обычно оправдано, так как небольшие смещения атомов протеина в процессе связывания лиганда учитываются процедурой так называемого "смягчения" потенциала, описываемой ниже. Атомы протеина, получившие соответствующие "типы" в соответствии с моделью MMFF94, создают вокруг протеина поля (электростатическое, Ван-дер-Ваальсово, поле потенциалов эффектов сольватации). Когда лиганд занимает некое фиксированное положение относительно протеина, атомы лиганда приобретают в каждом из этих полей некую энергию (определяемую величиной полей в месте расположения атома и типом атома), причем будем считать, что энергия лиганда представляется суммой вкладов отдельных атомов. В свою очередь, вклад отдельных атомов вычисляется как линейная комбинация энергии атома в каждом из перечисленных выше полей. Коэффициенты перед каждым энергетическим членом в этой линейной комбинации подбираются из соображений физического здравого смысла и наилучшего соответствия экспериментальным результатам в процессе проверки (валидации) процедуры докинга. Помимо этого, в энергию лиганда входит упрощенное представление его внутренней энергии напряжения в соответствии с моделью MMFF94.

Такая аддитивная модель является приближением, однако она позволяет построить эффективный алгоритм, в котором взаимодействие лиганда с протеином может быть быстро рассчитано при помощи вычисленных заранее трехмерных сеток потенциалов для каждого типа атома, который может встретиться в лиганде. Впервые такой подход для представления белка-мишени при описании взаимодействия протеин–лиганд был применен в работе [29], при этом вся необходимая информация для последующих расчетов энергии взаимодействия протеина с лигандом хранится в узлах потенциальной сетки, построенной на основании структуры белка-мишени. Это позволяет ускорить работу программы докинга, поскольку заменяет вычисление всех парных несвязанных взаимодействий атома лиганда с атомами протеина (которых обычно несколько тысяч) трилинейной интерполяцией по восьми ближайшим узлам трехмерной сетки. Хотя в этом случае расчет самих сеток потенциалов занимает достаточно продолжительное время, эта процедура проводится только один раз перед началом процедуры докинга лиганда или даже целой серии лигандов.

Сетки потенциалов должны быть рассчитаны для каждого возможного типа атома лиганда и для каждого потенциала. Они строятся с помощью специального модуля SOL_GRID (написан на языке Fortran 77) путем помещения по очереди пробного атома каждого типа (в дальнейшем — пробы) в узлы сетки и расчета энергии его взаимодействия со всем протеином. Модуль начинает расчет с чтения файла, содержащего данные о структуре белка в формате mrk. В результате чтения mrk-файла формируются массивы данных, содержащие информацию о декартовых координатах каждого атома протеина, его порядковом номере (в периодической системе элементов), принадлежности к какой-либо аминокислоте, типе этого атома в соответствии с моделью MMFF94, а также о зарядах атомов.

Затем вычисляются борновские радиусы [30] атомов протеина и пробы при различных положениях пробы на узлах сетки. Размер кубической сетки: $N_{\mathrm{GRID}} \times N_{\mathrm{GRID}} \times N_{\mathrm{GRID}}$. Обычно используется сетка с характерным количеством $N_{\mathrm{GRID}} = 100$ и физическим размером ребра куба порядка 22 Å (достаточно для покрытия активных центров большинства биологических мишеней), что соответствует шагу сетки, равному 0.22 Å. Соответственно формируется два массива:

1) четырехмерный массив $\mathrm{BP}(i, j, k, l)$ размерностью $N_{\mathrm{ES}} \times N_{\mathrm{GRID}} \times N_{\mathrm{GRID}} \times N_{\mathrm{GRID}}$, содержащий борновский радиус i-й пробы при ее расположении на узле сетки, который определяется тремя целыми числами j, k, l;

2) одномерный массив $\mathrm{BE}(i)$ размерностью N_{P} (N_{P} — количество атомов в протеине), содержащий борновский радиус i-го атома протеина.

Несколько адаптированная нами модель сольватации для поля MMFF94 [31] подразумевает нали-

чие $N_{\mathrm{ES}} = 9$ разных Ван-дер-Ваальсовых радиусов у атомов пробы различного типа. Далее вычисляется электростатическая энергия E_{ES} пробы на узлах сетки: $E_{\mathrm{ES}}(i,j,k) = \sum\limits_{n=1}^{N_{\mathrm{P}}} \dfrac{q_n}{|\overline{r}_n - \overline{r}_{i,j,k}|}$, где q_n — заряд n-го атома протеина, $\overline{r}_n$ — радиус-вектор n-го атома протеина, $\overline{r}_{i,j,k}$ — радиус-вектор узла сетки, определяемого тремя целыми числами i, j, k, а суммирование проводится по всем атомам протеина. Результаты расчета записываются в массив $E_{\mathrm{ES}}(i,j,k)$, содержащий электростатическую энергию пробы с единичным зарядом, расположенной на узле трехмерной сетки с координатами i, j, k.

Борновская энергия экранирования зарядов E_{BE} имеет знак, противоположный электростатической энергии, что отражает тот факт, что в присутствии растворителя заряды взаимодействуют слабее. Энергия E_{BE} для m-й пробы вычисляется по формуле [30]

$$E_{\mathrm{BE}}(m,i,j,k) = -\sum_{n=1}^{N_{\mathrm{P}}} \frac{q_n}{\sqrt{\left(\overline{r}_n - \overline{r}_{i,j,k}\right)^2 + \mathrm{BE}(n)\mathrm{BP}(m,i,j,k)\exp\left(-\dfrac{\left(\overline{r}_n - \overline{r}_{i,j,k}\right)^2}{8\,\mathrm{BE}(n)\mathrm{BP}(m,i,j,k)}\right)}} \tag{1}$$

и записывается в четырехмерный массив $E_{\mathrm{BE}}(m,i,j,k)$ размерностью $N_{\mathrm{ES}} \times N_{\mathrm{GRID}} \times N_{\mathrm{GRID}} \times N_{\mathrm{GRID}}$.

Далее вычисляется энергия десольвации протеина E_{PDS}, возникающая при вытеснении атомами пробы воды из области вблизи протеина. Эта энергия полагается пропорциональной трехмерному интегралу по области внутри Ван-дер-Ваальсовой сферы атома пробы (так называемое "кулоновское" приближение [32]): $E_{\mathrm{PDS}} = \left(1 - \dfrac{1}{\varepsilon}\right) \iiint\limits_V E^2\,dV$, где ε — диэлектрическая проницаемость воды и E — напряженность электростатического поля, создаваемая всеми парциальными зарядами атомов белка в данной точке пространства внутри Ван-дер-Ваальсовой сферы пробного атома. Интеграл вычисляется численно на трехмерной кубической сетке, шаг которой значительно меньше радиуса атома пробы.

Энергия десольватации лиганда вычисляется в виде квадратичной формы по зарядам на атомах лиганда:

$$E_{\mathrm{BE}}(m,i,j,k) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{q_i q_j}{\sqrt{\left(\overline{r}_i - \overline{r}_j\right)^2 + a_i a_j \exp\left(-\dfrac{\left(\overline{r}_i - \overline{r}_j\right)^2}{8 a_i a_j}\right)}}.$$

Здесь двойное суммирование осуществляется по атомам лиганда, q_i, q_j, $\overline{r}_i$, $\overline{r}_j$ — соответственно заряды и радиус-векторы атомов лиганда, а a_i — борновские радиусы атомов лиганда, записанные в массиве $\mathrm{BP}(i,j,k,l)$.

После этого рассчитываются Ван-дер-Ваальсовы потенциалы для пробных атомов. Количество пробных атомов равно 27 — это меньше, чем полное количество (99) типов атомов в модели MMFF94. Не учитывались типы атомов, не встречающиеся в лекарственных веществах. Кроме того, несколько типов атомов, близких по параметрам Ван-дер-Ваальсова взаимодействия, были объединены, что позволило сократить размер файла, содержащего сетки потенциалов. Принцип объединения типов проиллюстрирован в таблице приложения (табл. 4). Соответственно создаются 27 сеток, объединенных в четырехмерный массив $E_{\mathrm{VdW}}(m,i,j,k)$ размерностью $27 \times N_{\mathrm{GRID}} \times N_{\mathrm{GRID}} \times N_{\mathrm{GRID}}$. Потенциалы вычисляются в соответствии с зависимостями Ван-дер-Ваальсова взаимодействия, принятыми в MMFF94:

$$E_{\mathrm{VdW}}(m,i,j,k) = \sum_{n=1}^{N_{\mathrm{P}}} \varepsilon_{mn}\left(\frac{1.07R^*_{mn}}{R_{mn} + R^*_{mn}}\right)^7 \left(\frac{1.12\left(R^*_{mn}\right)^7}{1.12R^7_{mn} + 0.12\left(R^*_{mn}\right)^7} - 2\right).$$

Здесь R_{mn} — расстояние между точкой, в которой находится пробный атом m с координатами, соответствующими узлу сетки с нумерацией i, j, k, и атомом протеина с номером n: $R_{mn} = \left|\overline{r}_n - \overline{r}_{i,j,k}\right|$. R^*_{mn} — расстояние между атомами m и n, соответствующее минимуму Ван-дер-Ваальсова потенциала, которое рассчитывается по формуле $R^*_{mn} = \dfrac{1}{2}\left(R^*_{mn} + R^*_{nn}\right)\Big(1 + B\big(1 - \exp(-\beta\chi^2_{mn})\big)\Big)$, причем $B = 0.2$, $\beta = 12$ и $\chi_{mn} = \dfrac{R^*_{mn} - R^*_{nn}}{R^*_{mn} + R^*_{nn}}$; ε_{mn} — расстояние между атомами m и n, соответствующее минимуму Ван-дер-

Ваальсова потенциала, которое рассчитывается по формуле

$$\varepsilon_{mn} = \frac{181.16\, G_m G_n \alpha_m \alpha_n}{\left(\frac{\alpha_m}{N_m}\right)^{1/2} + \left(\frac{\alpha_n}{N_n}\right)^{1/2}} \frac{1}{\left(R^*_{mn}\right)^6},$$

где, в свою очередь, N_m и N_n — эффективное число валентных электронов для атомов, соответствующих типам m и n. Здесь G_m и G_n — масштабирующие факторы, а α_m и α_n — поляризуемости атомов типа m и n в соответствии с типизацией MMFF94. Значения этих факторов приводятся для всех типов атомов и формируются в виде массивов.

Для того чтобы косвенным образом учесть подвижность атомов протеина в процессе докинга, программный модуль вычисления Ван-дер-Ваальсова потенциала на сетке предусматривает введение так называемого "смягчения" потенциала. Процедура смягчения состоит в том, что имеющееся значение расстояния между атомами R_{mn} заменяется другим, вычисляемым по правилу

$$R_{mn} = \begin{cases} R_{mn} + \delta, & \text{если} \quad R_{mn} < R^*_{mn} - \delta, \\ R^*_{mn}, & \text{если} \quad R^*_{mn} - \delta < R_{mn} < R^*_{mn} + \delta, \\ R_{mn} - \delta, & \text{если} \quad R_{mn} > R^*_{mn} + \delta, \end{cases}$$

где R^*_{mn} — расстояние между атомами, соответствующее минимуму Ван-дер-Ваальсова потенциала их взаимодействия.

Физический смысл "смягчения" заключается в уширении потенциальной ямы Ван-дер-Ваальсова взаимодействия, отражающим тот факт, что атомы протеина обладают ограниченной подвижностью и способны подстраиваться под конкретный лиганд. Параметр δ называется радиусом уширения и задается во входных данных (рис. 2).

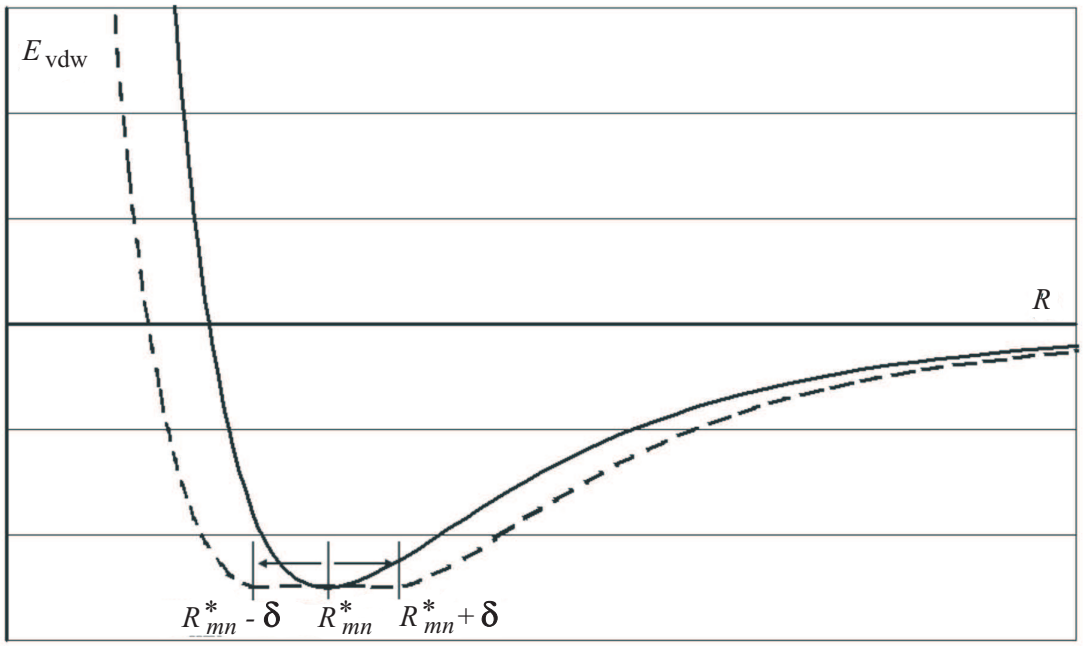

Рис. 2. Ван-дер-Ваальсов потенциал до (сплошная линия) и после (пунктир) процедуры "смягчения потенциала"

Затем осуществляется суммирование отдельных потенциалов по признаку зависимости взаимодействий, которые они описывают, от парциального заряда атома, находящегося в таком потенциале. Так, Ван-дер-Ваальсова энергия и энергия десольвации протеина лигандом не зависят от заряда на атомах лиганда. Электростатическая энергия и энергия экранирования зарядов в потенциале (1) зависят от заряда атомов лиганда, а энергия десольвации атомов лиганда определяется квадратичной формой от зарядов на лиганде. Суммирование дополняется умножением на эмпирические масштабирующие множители. Как

уже сказано выше, значения множителей выбираются на основе здравого смысла и наилучшего соответствия результатов докинга и экспериментальных структур протеин–лиганд, а также воспроизведения экспериментальных констант связывания лигандов с протеинами. По сути дела, выбором этих множителей осуществляется параметризация модели свободной энергии образования комплекса протеин–лиганд как линейной комбинации отдельных энергетических взаимодействий между ними. Такого рода приближения, как указывалось выше, всегда лежат в основе процедур докинга [6]. В случае программы SOL этот подход реализуется в наиболее последовательной форме.

По окончании работы модуля построения сеток потенциалов формируется файл, содержащий три четырехмерных массива данных для потенциалов взаимодействий, зависящих от заряда пробного атома в степени 0, 1, 2. В этот же файл записываются данные о положении построенной сетки относительно протеина, размере ребра куба сетки и другая вспомогательная информация.

2.2. Позиционирование лигандов: модуль докинга SOL. Основным вычислительным модулем программного комплекса KeenBASE является модуль докинга SOL (написан на языке Fortran 77), в основе которого лежит генетический алгоритм поиска глобального минимума функции, заданной на многомерном пространстве.

Входными файлами для программного модуля SOL являются файлы лигандов, т.е. файлы, определяющие трехмерную структуру органических молекул в формате mrk или hin. Эти файлы прежде всего обрабатываются встроенной в модуль SOL программой FARS [33], выполняющей типизацию атомов лиганда в соответствии с типами атомов в силовом поле MMFF94 [24 – 28].

Перед генетическим алгоритмом глобальной оптимизации стоит задача нахождения глобального минимума общей энергии в пространстве размерностью $N + 6$, где N — количество внутренних степеней свободы лиганда. Под общей энергией мы понимаем сумму внутренней энергии напряжения лиганда и энергии лиганда в потенциальном поле протеина. Цифра 6 учитывает сумму трех вращательных и трех поступательных степеней свободы при движении лиганда как целого. Понятно, что даже при умеренной сложности молекулы лиганда (значение N равно 5 или 6) проблема поиска глобального минимума функции в пространстве размерностью более 10 является непростой задачей.

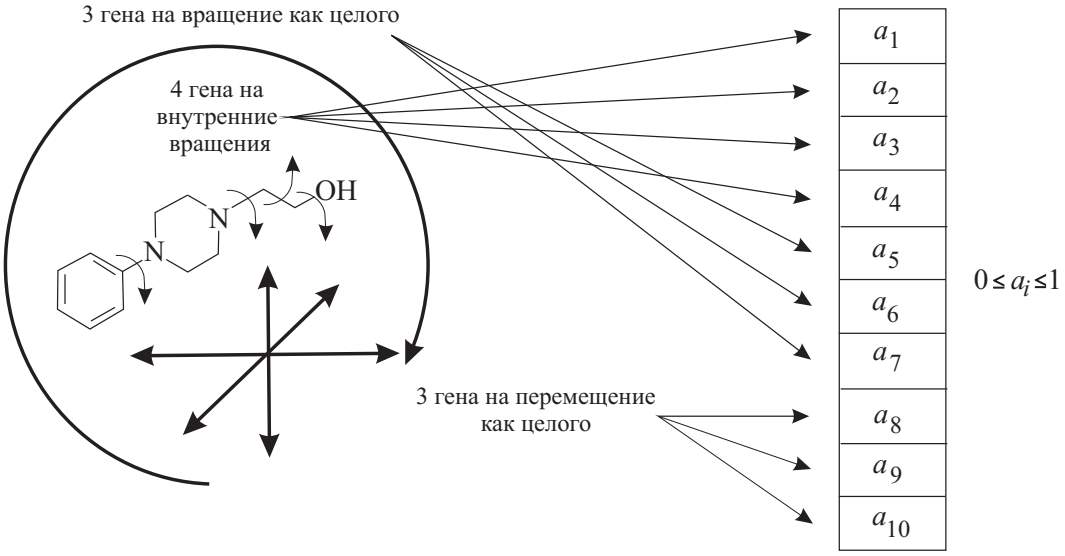

Рис. 3. Определение понятий фенотипа, генотипа и хромосомы в случае лиганда с 4 внутренними степенями свободы

Суть генетического алгоритма заключается в моделировании процессов эволюции по Дарвину некой популяции, составленной из отдельных особей. Естественный отбор осуществляется в соответствии с функцией фитнесса — особи с максимальным фитнессом дают потомство, закрепляющее их наследственность (генотип). В случае докинга под фитнессом мы будем понимать полную энергию лиганда, умноженную на -1. Таким образом, идет поиск особей с максимальным фитнессом, т.е. с минимальной полной энергией лиганда. Под генотипом каждой особи будем понимать одномерный массив a_i вещественных чисел (генов) в диапазоне от 0 до 1 — хромосому. Размер массива соответствует количеству степеней свободы лиганда и, тем самым, размерности пространства, на котором задана функция полной энергии лиганда. Таким образом, хромосома однозначно определяет положение лиганда в пространстве (в терминах генетического алгоритма это положение будет называться фенотипом), при этом величина каждого гена соответству-

ет нормированной на 1 величине соответствующей степени свободы лиганда (например, это показано на рис. 3 для лиганда с четырьмя внутренними степенями свободы).

Процедура перевода генотипа в фенотип можно назвать проецированием векторного пространства генотипа на пространство фенотипа, т.е. на пространство декартовых координат атомов лиганда. Это проецирование производится по следующей схеме.

1. Вычисляются координаты центра тяжести лиганда. При вычислении центра тяжести полагается, что все атомы лиганда имеют одинаковый вес. Координаты лиганда приводятся к центру тяжести таким образом, что центр тяжести оказывается в начале новой системы отсчета.

2. Осуществляются последовательные повороты фрагментов лиганда вокруг химических связей на угол $\varphi = 2\pi\, a_i$, $1 \leqslant i \leqslant N$, которые определяются N генами, кодирующими значения торсионных углов лиганда. Связь, вокруг которой может происходить поворот, делит молекулу на две части, и при ликвидации данной связи эти части становятся не связанными друг с другом. При повороте часть молекулы, содержащая большее число атомов, остается неподвижной, в то время как другая часть испытывает преобразование вращения вокруг оси, совпадающей с направлением связи, на угол φ. Связи, ликвидация которых не приводит к разобщению молекулы на несвязанные фрагменты, находятся в циклических структурах. Свободное вращение вокруг таких связей невозможно и не производится. Кроме того, не производится вращение вокруг двойных химических связей, так как энергетический барьер такого вращения очень высок и при обычной температуре ротамеры с различными значениями торсионных углов поворота вокруг таких связей существуют как разные вещества (цис- и трансизомеры).

3. Получившаяся структура подвергается преобразованию трехмерного вращения как целого. Трехмерное вращение задается четырьмя генами $a_{N+1}, \ldots, a_{N+4}$, которые представляют собой четыре компоненты q_x, q_y, q_z, q_w кватерниона, задающего трехмерное вращение, нормированное на 1: $q_x^2+q_y^2+q_z^2+q_w^2 = 1$. Данный кватернион определяет матрицу вращения:

$$\begin{pmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y+q_wq_z) & 2(q_xq_z-q_wq_y) \\ 2(q_xq_y+q_wq_z) & 1-2(q_x^2+q_z^2) & 2(q_yq_z+q_wq_x) \\ 2(q_xq_z+q_wq_y) & 2(q_yq_z-q_wq_x) & 1-2(q_x^2+q_y^2) \end{pmatrix},$$

в соответствии с которой и производится вращение структуры лиганда как целого. Выше указывалось, что для задания трехмерного вращения лиганда как целого необходимо три параметра. Описание вращения четырьмя компонентами кватерниона является избыточным, и для однозначности генотипа на кватернион налагается дополнительное условие нормировки на 1.

4. В заключение структура подвергается трансляции, параметры которой определяются тремя генами a_{N+5}, a_{N+6}, a_{N+7} таким образом, что координаты каждого атома лиганда подвергаются преобразованию $\overline{x} = x + 2r(a_{N+5}-1)$, $\overline{y} = y + 2r(a_{N+6}-1)$, $\overline{z} = z + 2r(a_{N+7}-1)$, где r — радиус активного центра протеина или (что равносильно) половина ребра куба, в который вписаны сетки потенциалов.

После того как по генотипу вычисляется фенотип, т.е. декартовы координаты атомов лиганда, для каждого фенотипа может быть вычислена полная энергия.

Полная энергия лиганда вычисляется путем сложения внутренней энергии E_{inner} напряжения лиганда и энергии лиганда $E_{\text{lig–prot}}$ в поле протеина.

Внутренняя энергия напряжения лиганда вычисляется по формуле $E_{\text{inner}} = E_{\text{lig–tors}} + E_{\text{lig–VdW}} + E_{\text{lig–ES}}$, где $E_{\text{lig–tors}}$ — энергия торсионного напряжения конформера, $E_{\text{lig–VdW}}$ и $E_{\text{lig–ES}}$ — энергии несвязанных Ван-дер-Ваальсовых и электростатических взаимодействий соответственно. Энергия торсионного напряжения $E_{\text{lig–tors}}$ молекулы лиганда вычисляется по соответствующей формуле силового поля MMFF94:

$$E_{\text{lig–tors}} = \frac{1}{2}\sum_{n=1}^{N_{\text{tors}}} \bigl(V_{1n}(1+\cos\,\Phi_n) + V_{2n}(1+\cos\,2\Phi_n) + V_{3n}(1+\cos\,3\Phi_n)\bigr).$$

Здесь N_{tors} — количество возможных четверок атомов i–j–k–l, связанных цепочкой связей в приведенном порядке, причем вокруг связи j–k в молекуле возможно внутреннее торсионное вращение, Φ_n — двугранный угол, образуемый плоскостями i–j–k и j–k–l, образованными вышеупомянутой четверкой атомов, V_{1n}, V_{2n}, V_{3n} — энергетические параметры для торсионного вращения, определяемые типизацией четверки i–j–k–l. Эти параметры определены в соответствии с силовым полем MMFF94.

Внутренняя Ван-дер-Ваальсова энергия лиганда $E_{\text{lig–VdW}}$ в рамках силового поля MMFF94 вычисля-

ется по формуле

$$E_{\text{lig–VdW}} = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\varepsilon_{mn}\left(\frac{1.07\,R^*_{mn}}{R_{mn}+R^*_{mn}}\right)^7\left(\frac{1.12\big(R^*_{mn}\big)^7}{1.12\,R^7_{mn}+0.12\big(R^*_{mn}\big)^7}-2\right),$$

где N — число атомов в лиганде.

Суммирование производится по всем парам атомов кроме тех, которые разделены химической связью (1–2 пара) или валентным углом (1–3 пара). Исключение 1–2 и 1–3 пар является стандартной процедурой при вычислении энергии несвязанного взаимодействия при использовании моделей силовых полей.

Внутренняя электростатическая энергия лиганда $E_{\text{lig–ES}}$ в рамках силового поля MMFF94 вычисляется по формуле $E_{\text{lig–ES}} = \dfrac{1}{2\varepsilon_{\text{in}}}\displaystyle\sum_{n=1}^{N}\sum_{m=1}^{N}\sigma_{mn}\,\frac{q_m q_n}{R_{mn}}$, где q_m и q_n — парциальные заряды на атомах лиганда m и n, ε_{in} — коэффициент диэлектрической проницаемости для вычисления внутренней электростатической энергии лиганда. Он определяется окружением лиганда — растворителем и протеином. Рекомендованное значение, принимаемое по умолчанию, составляет $\varepsilon_{\text{in}} = 2$.

Коэффициент σ_{mn} определяет величину вклада электростатической энергии пары атомов m–n. Таким образом, для 1–2 и 1–3 пар имеем $\sigma = 0$, для 1–4 пар (разделенных торсионным углом) — $\sigma = 0.75$, а для остальных пар атомов — $\sigma = 1$.

Таким образом, внутренняя энергия лиганда представлена исчерпывающим образом в рамках модели силового поля в отличие от большинства программ докинга (например AutoDock), где внутренняя энергия оценивается весьма условно.

Энергия лиганда $E_{\text{lig–prot}}$ в поле белка определяется как сумма энергий каждого атома лиганда, которые оцениваются при помощи заранее насчитанных трехмерных сеток потенциалов в соответствии с типизацией атома: $E_{\text{lig–prot}} = \displaystyle\sum_{i=1}^{N} E_i$, где N — количество атомов в молекуле лиганда и E_i — энергия атома: $E_i = E0_i + E1_i + E2_i$; здесь $E0_i$, $E1_i$, $E2_i$ — составляющие энергии атома в поле белка, которые соответственно не зависят от заряда атома, пропорциональны заряду атома и квадрату заряда атома.

Поскольку сетка дискретна, а атомы могут занимать любые положения, энергия атома в потенциале белка определяется при помощи процедуры трилинейной интерполяции значений энергии в узлах сетки, образующих куб вокруг данного положения атома.

Первоначальная популяция особей создается заполнением генов всех хромосом при помощи генератора случайных чисел. После этого происходит подсчет полной энергии для каждой особи; лучшие особи с минимальной энергией отбираются в так называемый mating pool, где они участвуют в создании нового поколения особей. Оно формируется из особей mating pool при помощи трех генетических операторов: кроссинговера, прямого переноса и мутации.

Кроссинговер представляет собой модель полового размножения — создание новой хромосомы из двух родительских хромосом путем случайного объединения их генов. Вероятность кроссинговера задается во входных параметрах программы. Реализованы несколько разновидностей кроссинговера: одноточечный кроссинговер, двухточечный кроссинговер и однородный кроссинговер.

Одноточечный кроссинговер подразумевает случайный выбор на хромосоме некоторой точки излома. Все гены новой дочерней особи, расположенные до этой точки, берутся из первой родительской особи, тогда как гены, расположенные после этой точки, берутся от второго родителя. Двухточечный кроссинговер реализуется случайным выбором двух точек излома, при этом гены, расположенные между точками излома, берутся от одной родительской особи, а остальные гены — от другой. При однородном кроссинговере для каждого гена в дочерней хромосоме существует вероятность P быть взятым от одной родительской особи и вероятность $1 - P$ быть взятым от другой родительской особи. Тип и параметры кроссинговера могут быть заданы на входе программы.

Все остальные особи создаются путем прямого переноса из mating pool — аналог вегетативного размножения. Затем на образованную таким образом популяцию действует оператор мутации, который с некоторой вероятностью подвергает случайному изменению значение гена в хромосоме. Вероятность мутации и плотность вероятности отклонения мутированной величины от исходной являются параметрами программы и могут задаваться пользователем. Учитывая тот факт, что гены, представляющие внутренние степени свободы, и гены, кодирующие положение лиганда как целого, влияют на координаты лиганда различным образом, введена опция раздельного задания вероятности мутации и плотности вероятности отклонения мутированной величины от исходной для трех групп генов: кодирующих торсионные вращения, кодирующих вращение лиганда как целого и кодирующих трансляцию лиганда как целого.

Наряду с описанными операторами осуществляется также прямой перенос элиты (особей с наименьшей общей энергией) из предыдущего в последующее поколение. Это производится во избежание утери уже найденных в процессе оптимизации лучших решений. Описанная процедура смены поколений повторяется заданное пользователем число раз. Размер популяции и размер mating pool также являются параметрами программы и задаются пользователем.

Генетический алгоритм является сбалансированной процедурой поиска глобального минимума, так как наряду с локальным улучшением уже найденных решений путем небольших мутаций соответствующих генов осуществляются процедуры кроссинговера, а также сильные мутации, что способствует исследованию областей многомерной поверхности, далеко отстоящих от текущего найденного минимума. Для того чтобы еще более стимулировать исследование разнообразных областей энергетической поверхности и предотвратить преждевременную концентрацию процесса поиска вокруг уже найденного и, возможно, локального минимума, в алгоритме реализована так называемая концепция нишинга [11]. Она моделирует наблюдаемое в природе ослабление конкуренции между особями, занимающими различные экологические ниши, т.е. сильно отличающимися генотипами. Такое ослабление конкуренции стимулирует разнообразие особей в популяциях и предотвращает преждевременное попадание поиска в локальный минимум. Программно нишинг реализуется следующим образом: на этапе выбора особей в mating pool после каждой очередной выбранной особи вычисляется некое расстояние между ее генотипом и генотипом особей текущего поколения: $D = \sqrt{\sum_{i=1}^{N}(b_i - a_i)^2}$, где N — количество генов в хромосоме, a_i, b_i — значения генов двух особей, между которыми вычисляется расстояние.

Затем к значению общей энергии особей текущего поколения прибавляется величина $\dfrac{\text{Niching\_factor}}{D}$. Таким образом, особи, близкие к уже отобранным в mating pool, будут иметь большие, а значит худшие значения общей энергии и меньше шансов быть отобранными в mating pool. Такая процедура приводит к большему разнообразию особей в mating pool, причем величину эволюционного давления в сторону увеличения разнообразия можно регулировать заданием входного параметра Niching_factor.

2.3. Анализ результатов докинга. В ходе работы программы процедуру вызова генетического алгоритма можно совершать несколько раз (число вызовов определяется пользователем во входных параметрах программы), получая несколько решений. Анализ взаимного положения этих решений важен для выяснения качества работы программы и корректности полученных результатов. С этой целью программа SOL, как и другие программы докинга [8], производит кластеризацию найденных решений (поз лиганда) соответственно их расположению по отношению к протеину-мишени. Для этого вычисляется матрица RMSD среднеквадратичного расстояния между химически эквивалентными атомами лиганда в двух различных позах:

$$\text{RMSD}_{ij} = \sqrt{\sum_{k=1}^{N}\left((x_{ik} - x_{jk})^2 + (y_{ik} - y_{jk})^2 + (z_{ik} - z_{jk})^2\right)},$$

где N — количество атомов в молекуле, x_{ik}, y_{ik}, z_{ik} — соответствующие координаты k-го атома i-й позы.

Для вычисления матрицы RMSD необходимо правильное химическое сопоставление атомов между отдельными позами, так как в молекуле может быть несколько химически эквивалентных атомов и, соответственно, много вариантов сопоставления. Для выявления всех возможных вариантов сопоставления необходимо проанализировать структуру молекулы и найти химически эквивалентные атомы. Эта процедура выполняется в отдельном блоке программы.

После получения матрицы RMSD происходит процедура кластеризации: решения, RMSD-расстояния между которыми не превышают некоторой заданной пользователем величины (обычно 1 ангстрем), собираются в один кластер. Кластеризация выявляет качество докинга. Если имеется небольшое количество многонаселенных кластеров, то можно считать, что докинг произведен успешно. Если кластеров много и они содержат небольшое количество решений, то, возможно, глобальный минимум не найден и необходим новый запуск программы с другими параметрами (больше особей в популяции, большее количество поколений, большее количество отдельных вызовов генетического алгоритма). Для лучшей ориентировки пользователя программа выдает исчерпывающую информацию о кластеризации решений:

1) количество кластеров,
2) населенность каждого кластера,
3) минимальная энергия решений кластера,
4) минимальное RMSD-расстояние между особями кластера,

5) среднее RMSD-расстояние между особями кластера,

6) максимальное RMSD-расстояние между особями кластера.

Кроме того, для каждого решения сообщается его среднеквадратичное отклонение от начальной геометрии лиганда, общая энергия, внутренняя энергия, энергия в поле протеина и ее распределение по трем составляющим: энергия, не зависящая от заряда атомов, энергия, пропорциональная заряду атомов, и энергия, пропорциональная квадрату заряда атомов.

Энергия лиганда в поле протеина дает указание на прочность протеин-лигандного комплекса. Чем ниже эта энергия, тем интереснее данный лиганд. Однако при оценке прочности связывания необходимо принимать во внимание и энтропийные факторы, возникающие из-за ограничений степеней свободы лиганда при его связывании с протеином; поэтому для окончательной оценки прочности связывания лиганда с протеином пользователю предоставляется так называемая скоринг-функция (SOL Score, SS), вычисленная для каждого решения, найденного генетическим алгоритмом. SS представляет собой линейную форму от энергии лиганда в поле протеина и количества топологически возможных торсионных поворотов в молекуле лиганда: $\mathrm{SS} = \alpha E_{\mathrm{lig–prot}} + \beta N_{\mathrm{tors}}$, где α, β — задаваемые пользователем коэффициенты.

Учет топологически возможных поворотов в молекуле лиганда (предполагается, что эти повороты замораживаются в процессе связывания с протеином) позволяет в простой форме учесть влияние энтропийных процессов при связывании протеина и лиганда в комплекс. Этот энтропийный вклад может быть значительным: так, например, для лиганда с 10 вращательными степенями свободы он составляет примерно 3 ккал/моль. Значения SS также выдаются для каждого полученного генетическим алгоритмом решения.

3. Тестирование (валидация) программы докинга SOL. Для подтверждения корректности физических и математических закономерностей, заложенных в программу докинга, было проведено специальное исследование, называемое валидацией программы. Это исследование проводилось по двум направлениям.

1. Первое направление — выявление программой докинга набора активных ингибиторов для данного белка из массива неактивных.

2. Второе направление — определение точности позиционирования лигандов в белках-мишенях программой докинга на основе расчета среднеквадратичного отклонения положения позиционированных лигандов от положения нативных лигандов (нативными лигандами обычно называют те лиганды, которые закристаллизованы в комплексе с белком; их положения в белке определены экспериментально) в выбранном наборе белков-мишеней. По данному направлению расчеты проводились также программой AutoDock 3.05 с целью сравнения результатов, полученных данной программой с результатами нашей программы докинга.

3.1. Выявление активных ингибиторов в массиве неактивных молекул. Основным результатом работы программы докинга является позиция лиганда в активном центре белка-мишени и соответствующее ей значение целевой функции (скоринг-функции), характеризующее наибольшую свободную энергию связывания белок–лиганд. Чем меньше значение этой функции, тем лучше лиганд позиционируется в белке. Тем самым качественно работающая программа докинга должна хорошо отличать активные для данного протеина лиганды от неактивных, выдавая существенно различные значения скоринг-функции для обоих типов соединений. Иными словами, активные лиганды должны обладать значительно меньшим значением целевой функции по сравнению с неактивными (или “мусором”). Таким образом, ранжируя весь набор позиционированных лигандов, программа докинга должна выводить в верхнюю часть списка лиганды, активные для данного белка и обладающие более отрицательным значением скоринг-функции. Чем больше активных лигандов попадет на самый верх списка, тем более качественно работает программа докинга, позволяя более точно определять потенциально активные лиганды.

Для проведения валидации по данному направлению был выбран следующий набор из четырех белков-мишеней, взятых из базы данных Protein Data Bank [34]: тромбин (PDB 1o2g, разрешение 1.58 Å), p38 MAP киназа (PDB 1a9u, разрешение 2.50 Å), фактор Xa (PDB 1lqd, разрешение 2.70 Å) и рецептор эстрогена (PDB 1xpc, разрешение 1.60 Å). Выбор этих белков-мишеней был обусловлен их подробной исследованностью, существованием большого набора экспериментально известных ингибиторов для данных протеинов и большим практическим интересом.

Подготовка исходных данных, структур лигандов и протеинов имеет свои особенности, обусловленные спецификой нашей программы. Для того чтобы избежать ошибок при позиционировании, которые могут возникнуть из-за пропущенных аминокислотных остатков, для приведенных выше четырех белков была выполнена специальная работа по восстановлению недостающих фрагментов их структуры. Затем про-

водилась расстановка атомов водорода на аминокислотных остатках с помощью программы Reduce [35], после чего лиганд и молекулы растворителя были удалены из кристаллической структуры. Следующим этапом подготовки явилась типизация атомов протеина в соответствии с типами силового поля MMFF94. Подробно процедура типизации описана в [33]. На следующем этапе подготовки определялись координаты центра активного сайта белка, вокруг которого впоследствии будет размещаться активная область, в которую проводится позиционирование лиганда. В нашем случае область выбиралась кубической формы с величиной ребра, равной 22 $\mathring{\text{A}}$, или 101 точке сетки. Центр куба соответствовал геометрическому центру лиганда в исходной кристаллической структуре из PDB. В результате были получены файлы соответствующих белков в формате mrk, которые были использованы при построении сетки (процедура построения описана в разделе 2.1).

Для создания набора неактивных лигандов "мусора", среди которых программа докинга должна выявлять примешанные активные лиганды, была выбрана база данных NCI Diversity [36]. Для проведения валидации было проверено 1894 соединения из этой базы данных и у ряда соединений изменены зарядовые состояния (т.е. добавлены или удалены протоны) в соответствии с кислотностью среды pH = 7, так как выбранные для валидации белки работают при данной кислотности. Активные соединения для белков-мишеней: тромбин, p38 MAP киназа и рецептор эстрогена были взяты из материалов работы [37] в формате SMILES. Эти соединения трансформировались в трехмерную структуру при помощи программы Corina [38]. Кроме того, набор активных лигандов для некоторых белков был дополнен структурами, взятыми из следующих источников: тромбин [39–48], p38 MAP киназа [49] и фактор Xa [50].

Обработка результатов валидации заключалась в ранжировании соединений в соответствии с лучшими оценками энергии связывания лигандов с белком-мишенью, полученными при помощи скоринг-функции программы докинга. Для каждого из четырех белков-мишений в соответствии с полученными результатами проводилось построение кривой обогащения (enrichment plot, EP) — кривой, представляющей относительное число известных реальных ингибиторов (нормализованное на полное число известных реальных ингибиторов, присутствующих в полном наборе лигандов (1894 + активные)) и построенной как функция от числа соединений с наилучшими скорами, находящихся в верхней части всех отранжированных лигандов, которые содержат в себе эти ингибиторы. Например, если восемь известных реальных ингибиторов находятся в тестовом наборе из 2000 лигандов и если среди двухсот лучших лигандов (т.е. 200 лигандов, находящихся в верхней части отранжированного по скору списка всех лигандов тестового набора) находятся четыре известных реальных ингибитора, то соответствующая точка графика EP имеет координаты $x = 10$ и $y = 50$ в процентах, поскольку $\dfrac{200}{2000} = 0.1$ и $\dfrac{4}{8} = 0.5$. Этот график позволяет вычислить коэффициент обогащения (enrichment value, EV) — значение площади под графиком, являющееся собственно результатом и позволяющее оценить возможность программы докинга выявить активные соединения из массива неактивных. В случае отсутствия обогащения EV принимает значение равное 0.5, в случае идеального обогащения (теоретического предела) EV должен быть равен 1. При расчете значений EV мы будем ориентироваться на следующие критерии качества работы программы докинга: значение площади (EV) находится в пределах 0.6–0.7 — удовлетворительное качество, 0.7–0.9 — хорошее качество, больше 0.9 — отличное качество докинга.

На рис. 4 приведены графики кривых обогащения EP. Значения коэффициентов обогащения EV и теоретических пределов коэффициентов обогащения для всех четырех белков-мишеней на полном наборе лигандов приведены в табл. 1.

Таблица 1

Значения коэффициентов обогащения для четырех отобранных протеинов

Название протеина	Коэффициент обогащения (EV)	Коэффициент обогащения — теоретический предел
Фактор Xa	0.939988	0.984887
p38 MAP киназа	0.637652	0.993226
Рецептор эстрогена	0.954927	0.986814
Тромбин	0.788898	0.993473

Коэффициент обогащения p38 MAP киназы попадает в интервал 0.6–0.7, что является удовлетворительным качеством докинга. Коэффициент обогащения для тромбина попадает в интервал 0.7–0.9, что

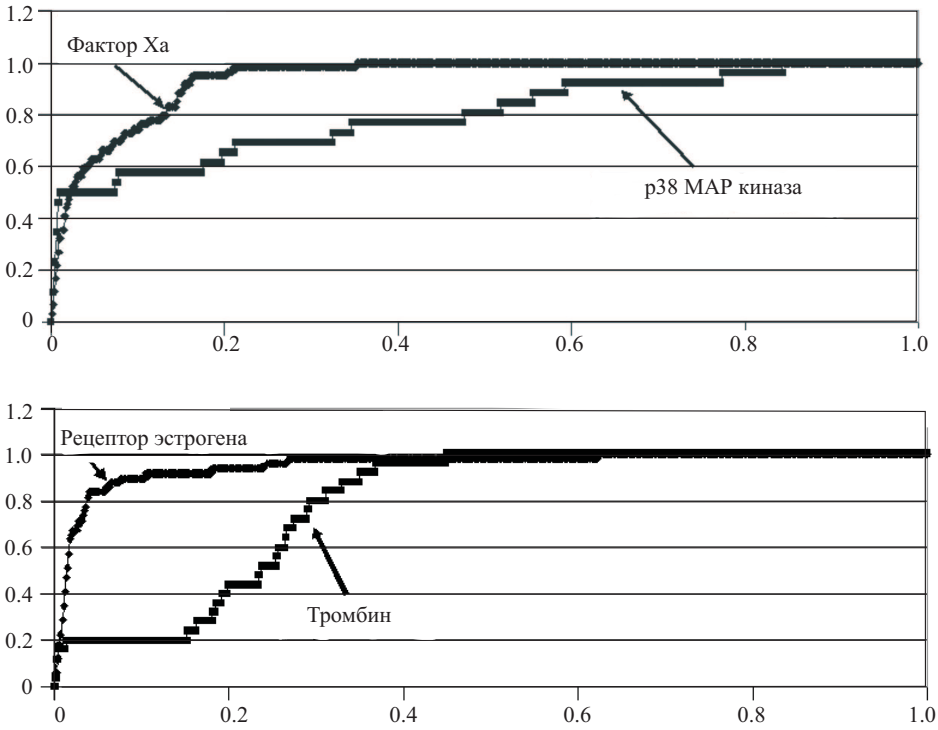

Рис. 4. Кривые обогащения для четырех отобранных протеинов

является хорошими качеством докинга. Для двух других белков (фактор Xa и рецептор эстрогена) коэффициент обогащения превышает 0.9, что является отличным качеством докинга.

При анализе этих результатов надо принимать во внимание, что все приведенные выше значения EP занижены. Это обусловлено тем, что a priori нельзя быть уверенным, что все химические соединения из базы данных NCI-Diversity являются "мусором" для того или иного белка-мишени. Среди них вполне могут находиться соединения, хорошо связывающиеся с тем или иным белком-мишенью. Очевидно, такие соединения нельзя отнести к "мусору", и их присутствие в топе ранжированного валидационного набора, например для тромбина, снизило значение EP. Этот вывод подтверждается и тем, что при использовании в качестве "мусора" не всего набора NCI-Diversity, а только 1000 соединений, имеющих наименее отрицательный скор, и тех же 25 активных лигандов коэффициент обогащения EV оказался равным 0.986, что является отличным качеством докинга и значительно отличается от значения EV=0.788898, приведенного на рис. 4.

3.2. Определение точности позиционирования лигандов в белках-мишенях. Основная трудность при проведении валидации данного типа — это выбор валидационного набора белков-мишеней. Белки в валидационный набор отбирались на основе следующих критериев: хорошее разрешение рентгеноструктурного анализа при определении структуры белка, чтобы исключить отсутствие в данной структуре каких-либо аминокислотных остатков или атомов, и разнообразие структур лигандов в выбранных белках, от маленьких (несколько десятков атомов) до больших (около ста атомов). Перед нами не стояла задача создания оригинального валидационного набора, поэтому часть белков была взята из наборов, описанных в литературе [6, 51]. Из валидационного набора были исключены металлопротеины, поскольку для таких белков процесс докинга в рамках созданной программы не может быть проведен корректно вследствие сложности описания взаимодействия лиганда с атомом металла. Кроме того, были исключены протеины, содержащие различные кофакторы, такие как, например, гем (heme), АТФ, НАДФ и др. Таким образом, было отобрано 80 комплексов, структуры которых были взяты из базы PDB [34]. Процесс подготовки данных комплексов для проведения докинга в точности соответствовал методике, описанной выше. Предварительные исследования показали, что из всех 20 параметров, управляющих процедурой докинга программы SOL, наиболее существенно на качество докинга влияют три: число независимых запусков программы SOL (NUMBER OF RUNS), размер популяции (число особей, POPULATION SIZE) и число поколений, участвующих в процессе глобальной оптимизации (NUMBER OF GENERATIONS). При проведении валидации были выбраны следующие значения этих параметров: NUMBER OF RUNS =

Таблица 2

Имя PDB рассчитанных комплексов и полученные для них результаты с помощью программы SOL и AutoDock

Имя PDB	RMSD по SOL, Å	RMSD по AutoDock, Å	N торсионных степеней свободы тяжелых атомов	Имя PDB	RMSD по SOL, Å	RMSD по AutoDock, Å	N торсионных степеней свободы тяжелых атомов
1pax	0.31	0.57	0	1mor	1.18	1.92	3
3pax	0.31	0.8	2	1qkq	1.22	4.83	1
1fh7	0.35	3.47	2	3kiv	1.45	1.51	5
1a28	0.48	0.75	1	1k1j	1.63	3.85	10
1ju4	0.5	1.38	1	1icm	1.69	1.79	12
1exa	0.5	0.78	5	1br6	1.85	2.76	4
1h70	0.51	0.64	6	1br5	1.85	1.13	3
1mq6	0.51	1.21	10	1lif	1.87	1.12	16
1c83	0.55	0.66	4	1ezq	1.87	2.34	11
4rsk	0.55	4.26	4	2sak	1.93	1.78	3
1abe	0.57	1.5	0	2hmb	2.04	4.06	14
1j01	0.59	3.63	2	1mq5	2.53	1.31	8
1ifu	0.63	8.87	1	1l8g	2.87	1.62	7
1jd3	0.64	0.56	1	1ppa	2.93	6.07	0
2pax	0.69	0.41	0	1art	2.94	0.88	3
2dri	0.7	0.57	0	1ifs	3.14	1.05	0
1fm6	0.73	3.93	7	1nli	3.14	2.09	0
1h1s	0.73	1.49	6	1f4g	3.17	2.61	14
1tsy	0.74	6.49	4	3jdw	3.43	4.08	4
1qpe	0.74	0.51	2	1mrg	3.44	0.78	0
3eng	0.8	1.18	4	2cmd	3.44	8.0	5
1ydr	0.83	1.45	2	2enb	3.48	2.34	6
1i7z	0.86	3.44	5	1hi3	3.83	3.76	6
1fut	0.88	1.17	4	1ikg	4.13	0.82	15
1b9v	0.92	1.37	8	1fao	4.2	0.83	8
1hpv	0.94	0.99	13	1rob	4.39	4.64	4
1efy	0.95	0.76	3	4dfr	4.57	1.13	10
1h52	0.95	11.01	2	1oxp	4.61	1.33	8
1mai	0.96	1.09	6	1d6v	4.62	3.59	6
1pot	0.98	0.63	7	1pph	4.73	3.49	8
1jgi	0.99	1.64	5	1jj0	5.18	1.88	5
1lqd	0.99	0.41	7	1akb	5.75	1.69	9
1ane	1.01	0.36	1	1a4k	5.75	1.69	6
1afq	1.02	3.06	11	1h1p	6.06	4.12	3
2cgr	1.05	1.12	10	1gor	6.48	3.75	2
3ert	1.09	1.6	10	1htf	6.85	2.74	15
1flz	1.1	7.56	0	1lzg	7.73	4.47	8
1fkg	1.14	1.27	12	1gc5	9.27	5.44	6
2ifb	1.17	1.31	14	2ovw	10.06	1.5	4
1ppc	1.17	5.45	11	1lr4	13.64	9.14	1

50, POPULATION SIZE = 30000 и NUMBER OF GENERATIONS = 500. При использовании программы AutoDock число независимых запусков совпадало с данным параметром для программы SOL и также равнялось 50 (RUNS = 50). В данном типе валидации критериями качества докинга являются значения среднеквадратичных отклонений положения задоченных лигандов от положения нативных лигандов (RMSD). Можно выделить четыре критерия качества: RMSD < 1 Å — отличное качество, 1 Å $<$ RMSD < 2 Å — хорошее качество, 2 Å $<$ RMSD < 3 Å — удовлетворительное качество и 3 Å $<$ RMSD — плохое качество.

В табл. 2 приведены результаты по 80 рассчитанным комплексам. Результаты представлены в виде

значения среднеквадратичного отклонения координат позиционированных лигандов от их координат в нативном положении соответствующего PDB-комплекса (RMSD, Å) и числа торсионных степеней свободы тяжелых атомов каждого лиганда. Данные отсортированы по возрастанию среднеквадратичных отклонений положений лигандов (RMSD), полученных по программе SOL.

Из таблицы видно, что для 50 комплексов из 80 среднеквадратичные отклонения положений, полученные по программе SOL, не превышают 2 Å (RMSD < 2 Å), в случае программы AutoDock количество таких комплексов равно 48.

Таблица 3 содержит количество комплексов, соответствующих каждому из критериев качества, и их отношение к общему количеству комплексов, выраженному в процентах.

Таблица 3

Значение среднеквадратичного отклонения, Å (критерии качества)	Количество комплексов по каждому из критериев качества		Суммарное количество комплексов, разделенных границей в 3 Å в процентном отношении к общему числу комплексов	
	SOL	AutoDock	SOL	AutoDock
< 1	32	19	55 (68.7 %)	54 (67.5 %)
1–2	18	29		
2–3	5	6		
> 3	25	26	25 (31.3 %)	26 (32.5 %)

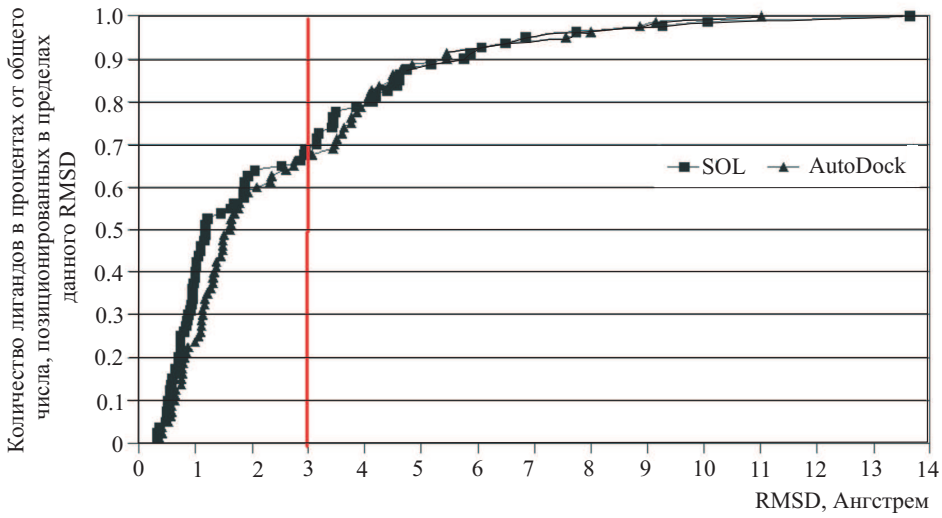

Рис. 5. Кривые, отражающие качество позиционирования лигандов из валидационного набора

Более наглядно результаты сравнения качества докинга обеих программ представлены на рис. 5.

Приведенные данные показывают, что на границе в 3 Å результаты, полученные по SOL и AutoDock, по числу комплексов практически совпадают. Однако количество комплексов с RMSD < 1 Å при расчетах программой SOL почти в два раза превышает число комплексов с таким же RMSD по результатам AutoDock, а для интервала от 1 до 2 Å ситуация обратная. В то же время не наблюдается явной корреляции между обеими программами в отношении качества докинга для одних и тех же комплексов, т.е. для каждого отдельного комплекса качество докинга по SOL и по AutoDock может быть диаметрально противоположным. Например, для комплекса 1ifu RMSD по SOL составляет 0.63 Å, а по AutoDock — 8.87 Å. В целом качество докинга у программы SOL не хуже, чем у AutoDock, а в интервале RMSD до

1 Å существенно превосходит последнюю.

Следует упомянуть также об еще одном преимуществе программы SOL по сравнению с AutoDock — это незначительное количество параметров, задаваемых пользователем для запуска программы. Так, для запуска SOL достаточно просто указать имя входного файла, содержащего структуру молекулы лиганда, тогда как в случае AutoDock исследователь должен сам определить количество торсионных степеней свободы и другие параметры ингибитора.

4. Практическое применение программы SOL для разработки новых ингибиторов тромбина. Нарушения процесса свертываемости крови — одна из распространенных проблем, с которыми сталкивается современная медицина. Образование нежелательных тромбов угрожает, например, пассажирам, совершающим длительные авиаперелеты (так называемый синдром экономического класса), а также пациентам, проходящим процедуру гемодиализа или возмещения большой кровопотери. Кроме того, в группу риска входят все люди старше 60 лет. Ключевым ферментом в сложной системе свертываемости крови является тромбин. Именно он производит превращение белка фибриногена в фибрин, который и образует основу тромба. Казалось бы, наиболее очевидный путь подавления нежелательного образования тромба — блокировать работу тромбина, являющегося основным ферментом системы свертываемости крови, т.е. найти прямой ингибитор этого фермента.

В настоящее время в клинической практике для предотвращения избыточного тромбообразования используется гепарин, который является всего лишь кофактором природного ингибитора тромбина — антитромбина (ATIII). При недостатке в крови ATIII введение гепарина не подавляет активность тромбина, а при некоторых заболеваниях крови введение гепарина невозможно; единственный выход в таких случаях — это добавить прямой ингибитор тромбина. Понятен интерес, который проявляют ведущие мировые фармакологические фирмы (Merck, Boehringer-Mannheim и др.) к разработке лекарственного средства — прямого ингибитора тромбина. На сегодняшний день существует только один прямой низкомолекулярный синтетический ингибитор тромбина — аргатробан, разрешенный к клиническому применению [52].

В течение полутора лет НИВЦ МГУ совместно с Гематологическим научным центром РАМН (лаборатория профессора Ф. И. Атауллаханова) и Институтом органической химии РАН проводил разработку новых низкомолекулярных ингибиторов тромбина. Начальный этап разработки — поиск соединений-лидеров — проводился в рамках системы KeenBASE с помощью программы докинга SOL. В ходе разработки было проведено несколько больших компьютерных экспериментов как на кластере НИВЦ МГУ, так и с участием удаленных компьютерных центров: Южноуральского (г. Челябинск) и Башкирского университетов (г. Уфа), университета Дубны и центра “Скиф-Siberia” Томского госуниверситета. В результате в рекордно короткие сроки (за полтора года) был разработан и запатентован новый класс низкомолекулярных синтетических прямых ингибиторов тромбина, по своим ингибирующим свойствам (IC50 = 2 наномоля) значительно превосходящих единственный применяемый в клиниках аналог — аргатробан (IC50 = 100 наномолей) [53 – 56]. Кроме того, уже экспериментально показано, что эти новые ингибиторы тромбина частично подавляют гиперкоагуляцию, возникающую при разбавлении плазмы крови искусственными плазмозамещающими растворами [57]. Этот эффект открывает возможности для создания нового класса кровезаменителей на основе новых ингибиторов тромбина.

Применение молекулярного дизайна ингибиторов на основе суперкомпьютерных расчетов с помощью программы SOL, работающей в рамках программного комплекса KeenBASE, привело к существенной экономии затрат на разработку: было рассмотрено около шести тысяч потенциальных ингибиторов — для них проведена процедура докинга и дана оценка энергии связывания. При этом первый ингибитор тромбина с константой связывания в наномолярном диапазоне был синтезирован с двадцатой попытки, т.е. для решения задачи поиска ингибитора, превосходящего по своим характеристикам зарубежный аналог, потребовалось произвести синтез всего 20 соединений вместо 6000, входящих в рассмотренные виртуальные комбинаторные библиотеки.

5. Заключение. С использованием генетического алгоритма была создана программа докинга SOL, позволяющая проводить позиционирование малых органических молекул в активные центры различных белков-мишеней. Приведены основные идеи и методы, использованные при реализации этой программы, а также результаты ее валидации.

Полученные результаты демонстрируют, что с помощью данной программы с хорошей точностью можно проводить докинг лигандов, число внутренних степеней свободы которых превышает 10. Сравнение SOL с широко распространенной программой AutoDock показывает, что точность докинга нашей программы не хуже, а иногда даже и лучше упомянутой программы на всем массиве валидационного набора. Кратко приведены результаты практического применения программы SOL для разработки новых ингибиторов для одного заданного белка-мишени: быстро (за 1.5 года) открыт и запатентован новый

класс низкомолекулярных синтетических ингибиторов тромбина, по своим характеристикам существенно превосходящий существующие аналоги.

6. Приложение. В приложении представлена табл. 4 типизации атомов в рамках модели силового поля MMFF и соответствия между типами MMFF и типами, принятыми в программе SOL для атомов лигандов.

СПИСОК ЛИТЕРАТУРЫ

1. *Gani O.A.B.S.M.* Signposts of docking and scoring in drug design // Chem. Biol. Drug Des. 2007. **70**. 360–365.
2. *Klebe G.* Virtual ligand screening: strategies, perspectives and limitations // Drug Discovery Today. 2006. **11**. 580–594.
3. *Kitchen D.B., Decornez H., Furr J.R., Bajorath J.* Docking and scoring in virtual screening for drug discovery: methods and applications // Nat. Rev. Drug Discov. 2004. **3**. 935–949.
4. *Lyne P.D.* Structure-based virtual screening: an overview // Drug Discovery Today. 2002. **7**. 1047–1055.
5. *Sousa S.F., Fernandes P.A., Ramos M.J.* Protein-ligand docking: current status and future challenges // Proteins: Struct., Funct. and Bioinf. 2006. **65**. 15–26.
6. Virtual screening in drug discovery // Alvarez J., Shoichet B. (Eds.). Boca Raton: Taylor & Francis, 2005.
7. *Goodsell D.S., Olson A.J.* Automated docking of substrates to proteins by simulated annealing // Proteins: Structure, Function and Genetics. 1990. **8**. 195–202.
8. *Goodsell D.S., Morris G.M., Olson A.J.* Automated docking of flexible ligands: applications of AutoDock // J. Mol. Recognition. 1996. **9**. 1–5.
9. *Morris G.M., Goodsell D.S., Halliday R.S., Huey R., Hart W.E., Belew R.K., Olson A.J.* Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function // J. Comp. Chem. 1998. **19**. 1639–1662.
10. *Holland J.H.* Adaptation in natural and artificial systems. Ann Arbor: University of Michigan Press, 1975.
11. *Goldberg D.E.* Genetic algorithms in search, optimization, and machine learning. Reading: Addison-Wesley, 1989.
12. *Goldberg D.E.* Real-coded genetic algorithm, virtual alphabets, and blocking // Complex Systems. 1991. **5**. 139–167.
13. *Goldberg D.E., Deb K.* A comparative analysis of selection schemes used in genetic algorithms // Foundations of Genetic Algorithms. Rawlins G.J.E. (Ed.). San Mateo: Morgan Kaufmann, 1991. 69–93.
14. *Oshiro C.M., Kuntz I.D., Dixon J.S.* Flexible ligand docking using a genetic algorithm // J. Comput.-Aided Mol. Design. 1995. **9**. 113–130.
15. *Westhead D.R., Clark D.E., Frenkel D., Li J., Murray C.W., Robson B., Waszkowycz B.* PRO-LIGAND: An approach to de novo molecular design. 3. A genetic algorithm for structure refinement // J. Comput.-Aided Mol. Design. 1995. **9**. 139–148.
16. *Clark D.E., Westhead D.R.* Evolutionary algorithms in computer-aided molecular design // J. Comput.-Aided Mol. Design. 1996. **10**. 337–358.
17. *Pegg S.C.-H., Haresco J.J., Kuntz I.D.* A genetic algorithm for structure-based de novo design // J. Comput.-Aided Mol. Design. 2001. **15**. 911–933.
18. *Sulimov V., Romanov A., Grigoriev F., Kondakova O., Sulimov A., Bryzgalov P., Zhabin S., Chernobrovkin A., Sobolev S.* Web-oriented system Keenbase for virtual screening and design of new ligands for biological macromolecules. Application for new drug searches // Proc. of the St. Petersburg Int. Workshop on NanoBiotechnologies, 27 – 29, November, 2006. St. Petersburg, 2006. 33–34.
19. *Sulimov A.V., Sulimov V.B., Romanov A.N., Grigoriev F.V., Kondakova O.A., Bryzgalov P.A., Ostapenko D.A.* Web-oriented system Keenbase for new drugs design // Proc. of the Fourth Int. Symposium on Computational Methods in Toxicology and Pharmacology Integrating Internet Resourses (CMPTI-2007, Moscow, Russia, September, 1 – 5, 2007). Moscow, 2007. 158.
20. *Сулимов В.Б., Романов А.Н., Григорьев Ф.В., Кондакова О.А., Сулимов А.В., Жабин С.Н., Соболев С.И.* Веб-ориентированная система молекулярного моделирования Keenbase для разработки новых лекарств // Тр. Всероссийской научной конференции "Научный сервис в сети Интернет: технологии параллельного программирования", 18 – 23 сентября, 2006, Новороссийск. М.: Изд-во Моск. ун-та, 170–172.
21. *Воеводин Вл.В., Филамофитский М.П.* X-Com — проект организации распределенных вычислений // Тр. Всероссийской научной конференции "Научный сервис в сети Интернет". М.: Изд-во Моск. ун-та, 2001. 11–13.
22. *Филамофитский М.П.* Система поддержки метакомпьютерных расчетов X-Com: архитектура и технология работы // Вычислительные методы и программирование. 2004. **5**, № 1. 128–136.
23. http://www.parallel.ru
24. *Halgren T.A.* Merck molecular force field. I. Basis, form, scope, parametrization and performance of MMFF94 // J. of Comp. Chem. 1996. **5 & 6**. 490–519.
25. *Halgren T.A.* Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions // J. of Comp. Chem. 1996. **5 & 6**. 520–552.
26. *Halgren T.A.* Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94 // J. of Comp. Chem. 1996. **5 & 6**. 553–586.

Таблица 4

Тип атома в модели MMFF94 1	Описание 2	Тип атома лиганда в программе SOL 3
1	Алкильный углерод	1
2	Винильный углерод, углерод гуанидиновой группы, общий тип углерода с гибридизацией sp^2	2
3	Общий тип углерода карбонильной группы, углерод группы C=N, углерод кетонной, альдегидной, амидной, карбоксильной, сложноэфирной, карбаматной групп. Углерод тиокарбоновых кислот, тиоэфиров и тиоамидов	3
4	Ацетиленовый и алленовый углероды	4
5	Водороды при атомах углерода, кремния, фосфора	5
6	Кислород при углероде sp^3 гибридизации, кислород воды, двухкоординированный кислород сложных эфиров и карбоновых кислот (в нейтральном состоянии), кислород енолов, фенолов. Двухкоординированный кислород сульфатов, сульфитов, фосфатов, фосфитов, кислород группы O—C=N, общий тип двухкоординированного кислорода	6
7	Общий тип кислорода карбонильной группы, кислород амидов, кетонов, альдегидов, кислород при двойной связи карбоновых кислот, сложных эфиров, сульфоксидов, кислород нитрозогруппы	7
8	Азот аминов	8
9	Азот иминов и азосоединений	9
10	Азот амидов, тиоамидов, азидной группы, группы N—N=C	9
11	Фтор	10
12	Хлор	11
13	Бром	12
14	Йод	13
15	Сера тиолов и сульфидов	14
16	Сера, при двойной связи с углеродом	14
17	Сера в сульфоксидах	15
18	Сера в сульфонах, сульфонамидах, сульфонатах, азотных аналогах сульфонов	15
19	Кремний	14
20	Углерод циклобутильной группы	1
21	Водород при кислороде в спиртах, общий тип водорода при кислороде	16
22	Углерод циклопропильной группы	3
23	Водород при азоте в аминах, аммиаке, водород при азоте пиррола, общий тип водорода при азоте	17
24	Водород при кислороде в карбоновых и фосфоновых кислотах	17
25	Фосфор в фосфорной кислоте и эфирах фосфорной кислоты, общий тип четырехкоординированного фосфора	18
26	Трехкоординированный фосфор	14

27. *Halgren T.A., Nachbar R.B.* Merck molecular force field. IV. Conformational energies and geometries for MMFF94 // J. of Comp. Chem. 1996. **5 & 6**. 587–615.
28. *Halgren T.A.* Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data and empirical rules // J. of Comp. Chem. 1996. **5 & 6**. 616–641.

Продолжение таблицы 4

1	2	3
27	Водород при азоте иминогруппы, азогруппы	17
28	Водород при азоте амидов, енаминов, тиоамидов, водород в группах HN–C=N, HN–N=C, HN–N=N, общий тип водорода при sp^2 азоте	16
29	Водород енолов и фенолов, водород группы HO–C=N	16
30	Углерод при двойной связи в четырехчленном кольце	2
31	Водород в воде	16
32	Кислород карбоксилатного аниона, кислород в N-оксидах, кислород в нитрогруппах и нитратах, монокоординированный кислород при тетракоординированной сере, кислород в сульфонах, сульфонамидах и сульфонатах, кислород сульфат-иона, кислород при двойной связи с атомом фосфора	19
33	Водород сульфиновых и сульфоновых кислот	16
34	Кватернизованный азот	20
35	Оксидный кислород на sp^2 и sp^3 углероде	19
36	Водород при положительно заряженном атоме азота, водород имидазолия, гуанидиния, протонированной группы HN+=C–N	17
37	Ароматический атом углерода	21
38	Ароматический азот пиридина	22
39	Ароматический азот пиррола	23
40	Азот групп N–C=C и N–C=N	9
41	Углерод в карбоксилатном ионе	3
42	Азот при тройной связи	9
43	Азот в сульфонамидах	9
44	Сера в тиофене	14
45	Азот нитрогруппы и нитратов	24
46	Азот нитрозогруппы	24
47	Терминальный азот азидогруппы	9
48	Азот при двойной связи, на другом конце которой — сера	24
49	Трехкоординированный кислород (кислород оксония)	19
50	Водород при кислороде оксония	16
51	Кислород оксения	7
52	Водород при кислороде оксения	16
53	Азот в группе =N+=	20
54	Азот протонированных имино- и азогрупп (иминия и азония)	24
55	Азот группы N+=C–N (формальный заряд Q=1/2)	25
56	Азот протонированного гуанидина (гуанидиния)	25
57	Углерод гуанидиния	26
58	Азот иона пиридиния	25
59	Кислород в ароматическом кольце (фуран)	7
60	Углерод изонитрильной группы	2
61	Азот изонитрильной группы	22
62	Азот сульфонамидов	8
63	Альфа-углеродный атом пятичленных ароматических колец	21
64	Бета-углеродный атом пятичленных ароматических колец	21
65	Альфа-атом азота пятичленных ароматических колец	9

Продолжение таблицы 4

1	2	3
66	Бета-атом азота пятичленных ароматических колец	22
67	Трехкоординированный азот в N-оксидах	9
68	Четырехкоординированный азот в N-оксидах	9
69	Ароматический азот шестичленных колец в N-оксидах	9
70	Кислород в воде	19
71	Водород в тиолах	16
72	Сера в тиокарбоксилатах, монокоординированная сера при двойной связи с фосфором, углеродом, монокоординированная сера в тиосульфинатах	27
73	Двухкоординированная сера в сульфинатах и тиосульфинатах	27
74	Сера в сульфинилах (C=S=O)	27
75	Фосфор при двойной связи с углеродом	27
76	Отрицательно заряженный азот тетразольного кольца	24
77	Хлор в хлорат-ионе	—
78	Общий тип углерода в 5-членном ароматическом кольце	21
79	Общий тип азота в 5-членном ароматическом кольце	9
80	Углерод фрагмента N–C–N заряженного кольца имидазола (имидазолия)	26
81	Азот фрагмента N–C–N заряженного кольца имидазола (имидазолия)	25
82	Азот ароматических пятичленных колец в N-оксидах	9
87	Fe^{+2}	—
88	Fe^{+3}	—
89	F^{-}	—
90	Cl^{-}	—
91	Br^{-}	—
92	Li^{+}	—
93	Na^{+}	—
94	K^{+}	—
95	Zn^{+}	—
96	Ca^{+2}	—
97	Cu^{+}	—
98	Cu^{+2}	—
99	Mg^{+2}	—

29. *Goodford P.J.* A computational procedure for determining energetically favorable binding sites on biologically important macromolecules // J. Med. Chem. 1985. **28**. 849–857.
30. *Romanov A.N., Jabin S.N., Martynov Y.B., Sulimov A.V., Grigoriev F.V., Sulimov V.B.* Surface generalized Born method: a simple, fast and precise implicit solvent model beyond the Coulomb approximation // J. Phys. Chem. A. 2004. **108**. 9323–9327.
31. *Bordner A.J., Cavasotto C.N., Abagyan R.A.* Accurate transferable model for water, *n*-octanol, and *n*-hexadecane solvation free energies // J. Phys. Chem. B. 2002. **106**. 11009–11015.
32. *Ghosh A., Rapp C.S., Friesner R.A.* Generalized Born model based on a surface integral formulation // J. Phys. Chem. B. 1998. **102**. 10983–10990.
33. *Григорьев Ф.В., Романов А.Н., Кондакова О.А., Лущекина С.В., Сулимов В.Б.* Алгоритм расстановки силовых параметров на атомах органических молекул и белков в рамках силового поля MMFF 94 // Вычислительные методы и программирование. 2006. **7**. 128–136.
34. *Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissing H., Shindyalov I.N., Bourne P.E.* The protein data bank // Nucleic Acids Res. 2000. **28**. 235–242.
35. *Word J.M., Lovell S.C., Richardson J.S., Richardson D.C.* Asparagine and glutamine: using hydrogen atom contacts in the choice of sidechain amide orientation // J. Mol. Biol. 1999. **285**. 1733–1745.

36. NCI (National cancer institute) diversity data base (http://dtp.nci.nih.gov/docs/3d_database/Structural_information/structural_data.html).
37. *Stahl M., Rarey M.* Detailed analysis of scoring functions for virtual screening // J. Med. Chem. 2001. **44**. 1035–1042.
38. *Gasteiger J., Rudolph C., Sadowski J.* Automatic generation of 3D-atomic coordinates for organic molecules // Tetrahedron Comput. Methodol. 1990. **3**. 537–547.
39. *Maruyama I.* // Jpn. J. Clin. Hematol. 1990. **31**. 776–781.
40. *Kikumoto R., Tamao Y., Tezuka T., Tonomura A., Hara H., Ninomiya K., Hijikata A., Okamoto S.* Selective inhibition of thrombin by (2R,4R)-4-methyl-1-[N2-[1,2,3,4-tetrahydro-8-quinolinyl)sulfonyl]-L-arginyl]-2-piperidine-carboxylic acid // Biochemistry. 1984. **23**. 85–90.
41. *Okamoto S., Hijikata A.* Potent inhibition of thrombin by the newly synthesized arginine derivative No. 805. The importance of stereo-structure of its hydrophobic carboxamide portion // Biochemical and Biophysical Research Communications. 1981. **101**. 440–446.
42. *Linder R., Frebelius S., Jansson K., Swedwnborg J.* Inhibition of endothelial cell-mediated generation of activated protein C by direct and antithrombin-dependent thrombin inhibitors // Blood Coagulation and Fibrinolysis. 2003. **14**. 139–146.
43. *Okamoto S., Hijikata-Okunomiya A.* Synthetic selective inhibitors of thrombin // Methods in Enzymology. 1993. **222**. 328–340.
44. *Hijikata-Okunomiya A., Okamoto S.* A strategy for a rational approach to designing synthetic selective inhibitors // Seminars in Thrombosis and Hemostasis. 1992. **18**. 135–140.
45. *Vacca J.* New advances in the discovery of thrombin and factor Xa inhibitors // Current Opinion in Chemical Biology. 2000. **4**. 394–400.
46. *Steinmetzer T., Hauptmann J., Sturzebecher J.* Advances in the development of thrombin inhibitors // Exp. Opin. Invest. Drugs. 2001. **10**. 845–864.
47. *Shafer J.A.* Cardiovascular chemotherapy: anticoagulants // Current Opinion in Chemical Biology. 1998. **2**. 458–465.
48. *Hauptmann J., Sturzebecher J.* Synthetic inhibitors of thrombin and factor Xa: from bench to bedside // Thrombosis Research. 1999. **93**. 203–241.
49. *Varma S.* // 3d US Catalyst User's Group meeting Boehringer-Ingelheim pharmaceuticals, Inc. May 11, 2001.
50. *Pauls H.W., Ewing W.R., Choi-Sledeski Y.M.* The design of competitive, small-molecule inhibitors of coagulation factor Xa // Frontiers Med. Chem. 2004. **1**. 129–152.
51. http://www.ccdc.cam.ac.uk/products/life_sciences/validate
52. *Frenkel E.P., Shen Y.M., Haley B.B.* The direct thrombin inhibitors: their role and use for rational anticoagulation // Hematol. Oncol. Clin. N. Am. 2005. **19**. 119–145.
53. *Синауридзе Е.И., Сулимов В.Б., Семенов В.В., Грибкова И.В., Горбатенко А.С., Боголюбов А.А., Романов А.Н., Кондакова О.А., Атауллаханов Ф.И.* Новый ряд ингибиторов тромбина // Тр. IV Международного конгресса "Биотехнология: состояние и перспективы развития", 12 – 16 марта 2007, Москва. Часть 1. М.: 2007, 103.
54. *Романов А.Н., Сулимов В.Б., Кондакова О.А., Синауаридзе Е.И., Кузнецов Ю.В., Воеводин В.В., Атауллаханов Ф.И.* Новые ингибиторы тромбина: молекулярный дизайн с использованием суперкомпьютеров и экспериментальное подтверждение активности // Тр. IV Сибирской школы-семинара по параллельным и высокопроизводительным вычислениям. Томск, 2007. 18–19.
55. *Грибкова И.В., Синауридзе Е.И., Сулимов В.Б., Горбатенко А.С., Кузнецов Ю.В., Монаков М.Ю., Боголюбов А.А., Романов А.Н., Кондакова О.А., Атауллаханов Ф.И.* Поиск новых ингибиторов тромбина // Тр. XI Международной Пущинской школы-конференции молодых ученых 29 октября – 2 ноября 2007. Пущино, 2007. 240–241.
56. *Sulimov V.B., Romanov A.N., Kondakova O.A., Sinauridze E.I., Butylin A.A., Gribkova I.V., Gorbatenko A.S., Bogoliubov A.A., Titov I.Yu., Polunin E.V., Kuznetsov Yu.V., Taidakov I.V., Voevodin V.V., Sobolev S.I., Ataullakhanov F.I.* New thrombin inhibitors: molecular design and experimental discovery // (IDDST), BIT's 5th Anniversary Congress of International Drug Discovery Science & Technology, Serial II: Advances and Challenges Toward Major Diseases. Theme: Extension on New Hope, November 7 – 13, 2007. Xi'an & Beijing, China.
57. *Синауридзе Е.И., Горбатенко А.С., Грибкова И.В., Сулимов В.Б., Романов А.Н., Кондакова О.А., Ажигирова М.А., Дереза Т.Л., Кузнецов Ю.В., Боголюбов А.А., Атауллаханов Ф.И.* Гиперкоагуляция, вызываемая разбавлением плазмы искусственными плазмозамещающими растворами // Технологии живых систем. 2008. **5**, № 1. 3–14.

Поступила в редакцию
29.05.2008