УДК 004.051; 615.011.3; 538.9

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПАРАМЕТРОВ ГЕНЕТИЧЕСКОГО АЛГОРИТМА НА ЭФФЕКТИВНОСТЬ ДОКИНГА С ПОМОЩЬЮ ПРОГРАММЫ SOL

Е. В. Каткова[1]

Проведен анализ ключевых параметров генетического алгоритма и изучено их влияние на эффективность процедуры докинга (позиционирования низкомолекулярных веществ в активном центре молекулы белка), реализованного в многопроцессорной версии программы SOL. Тестирование проводилось на наборе структур, представляющих собой комплексы белка урокиназы, закристаллизованной с различными низкомолекулярными ингибиторами (лигандами), и было ориентировано на оптимизацию дальнейшей работы по поиску новых ингибиторов урокиназы как потенциальных противоопухолевых препаратов. Оптимальные значения параметров генетического алгоритма получены как для режима, в котором проводится докинг одной молекулы, так и для режима виртуального скрининга, т.е. докинга больших библиотек химических соединений. Тестирование проводилось на кластерных суперкомпьютерах Московского государственного университета “Чебышев” и “Ломоносов”.

Ключевые слова: докинг, разработка лекарственных средств, генетический алгоритм, глобальная оптимизация, параметризация.

1. Введение. Методы молекулярного моделирования в настоящее время играют важную роль при решении задач, связанных с поиском новых лекарственных средств. Зачастую необходимо ингибировать некоторые белки в терапевтических целях. Молекулы-ингибиторы обычно представляют собой низкомолекулярные органические соединения, которые избирательно связываются с активным центром белка-мишени и подавляют его активность. Для того чтобы предсказать структуру таких молекул, используются методы молекулярного моделирования [1–3].

Одним из инструментов подобного моделирования, позволяющим быстро проводить поиск новых структур на начальном этапе разработки лекарственного средства, является докинг. Докинг — это позиционирование молекул-кандидатов в ингибиторы (лигандов) в активном центре белка-мишени и оценка их энергии связывания. Чем сильнее молекула связывается с белком, тем лучше ингибитор и эффективнее новое лекарственное средство на его основе. Этот процесс в некотором смысле является виртуальным аналогом процедуры высокопроизводительного скрининга и позволяет сократить число соединений-кандидатов, проверяемых экспериментально, на несколько порядков.

С математической точки зрения докинг представляет собой поиск глобального минимума функции свободной энергии, заданной на многомерном пространстве всех возможных способов связывания лиганда с белком. Известные алгоритмы поиска наилучшего связывания могут быть разделены на следующие категории: систематические методы, случайные или стохастические эвристические методы, методы молекулярной динамики и термодинамические методы [4].

Методами, гарантирующими нахождение глобального минимума за конечное число шагов, являются систематические методы, т.е. методы последовательного перебора всех возможных положений лиганда в активном центре белка-мишени. Однако ввиду большого количества требуемых вычислений этот метод требует введения значительных упрощений.

Широко распространены другие методы глобальной оптимизации, которые не гарантируют нахождение глобального минимума за конечное число шагов программы, однако на практике оказывается, что они способны отыскивать такие минимумы гораздо быстрее, чем методы систематического перебора. Цена, которую приходится платить за скорость, — это отсутствие уверенности, что найденный минимум глобальный. Подобные методы можно разделить на две большие группы: эвристические и термодинамические.

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М.В. Ломоносова, Ленинские горы, д. 1, 119992, стр. 4, Москва, техник; ООО “Димонта”, ул. Нагорная, д. 15, корп. 8, 117186, Москва, мл. науч. сотр.; Московский государственный университет им. М.В. Ломоносова, физический факультет, Ленинские горы, д. 1, стр. 2, 119991, Москва, аспирант; e-mail: katkova@dimonta.com

Эвристические методы используют некие эмпирические стратегии поиска глобального минимума, ускоряющие процедуру по сравнению с простым сканированием гиперповерхности. На сегодняшний день наиболее известны и популярны следующие эвристические методы:

— эволюционные и генетические алгоритмы (evolutionary and genetic algorithms) [5, 6];

— алгоритмы поиска с табу (taboo search), основанные на работах Ф. Гловера в области поиска глобального минимума дискретных функций [7];

— стайные алгоритмы оптимизации (particle swarm optimization, PSO) [8].

К термодинамическим методам относят моделирование отжига (simulated annealing) [9]. Глобальная оптимизация методом моделирования отжига может использоваться, например, в программе AutoDock [10], хотя и не является ее основным методом.

Воспроизведение эволюции системы лиганд–белок при помощи метода молекулярной динамики считается самым реалистичным методом из существующих, однако это и самый медленный метод; обычно его применяют для уточнения решений, найденных другими способами [11–13].

На сегодняшний день наибольшее распространение в задачах позиционирования малых молекул в активных центрах белков и оценки их энергии связывания с белками нашли генетические алгоритмы. Наиболее известные программы докинга, использующие эти методы: AutoDock [14], GOLD [15, 16], MolDock [17], DARWIN [18], DIVALI [19], PSI-Dock [20], FLIP-Dock [21], Lead Finder [22].

Программа докинга SOL [23, 24] имеет в своей основе генетический алгоритм поиска глобального минимума энергии. Эта программа хорошо зарекомендовала себя при разработке нового ингибитора тромбина [25]. Работой алгоритма управляет большое число параметров, изменение которых позволяет настроить программу для более эффективного решения конкретных задач. Поскольку различные реализации алгоритма имеют свои особенности, то стандартных параметров генетического алгоритма, которые были бы лучшими для всех задач оптимизации, не существует. Настоящая статья посвящена исследованию функционирования генетического алгоритма, реализованного в многопроцессорной версии программы SOL. Предпосылкой для нашей работы стала задача поиска новых ингибиторов урокиназы как потенциальных противоопухолевых препаратов [26]. Урокиназа представляет собой сериновую протеазу, активирующую плазмин, а также играющую роль в делении клеток и их миграции. При нормальных условиях в организме этот белок ответственен за прорастание сосудов, однако в случае патологий все эти факторы приводят к росту опухоли и появлению метастазов. Этот факт был доказан экспериментально, и было показано, что блокирование урокиназы замедляет эти процессы. Поэтому выбор урокиназы в качестве мишени для противоопухолевого препарата и поиск ингибиторов ее протеолитического центра представляется довольно перспективным направлением [27].

Таким образом, эффективность работы программы SOL проверялась на структурах урокиназы и закристаллизованных с ними лигандах, имеющих известные энергии связывания (или константы ингибирования). В результате были получены оптимальные значения параметров генетического алгоритма, лежащего в основе программы докинга SOL. Повышение качества работы программы позволяет предполагать, что поставленная задача — задача поиска новых лекарственных средств методами компьютерного моделирования — будет решаться более эффективно.

2. Описание генетического алгоритма, реализованного в программе докинга SOL. Программа докинга SOL осуществляет позиционирование лиганда в активном центре заданного белка-мишени и дает оценку энергии связывания. Для ускорения вычислений и упрощения поиска наилучшего положения лиганда делаются некоторые приближения.

1. Белок предполагается жестким, а для имитации подвижности атомов белка вводится уширение потенциалов [23], составляющее обычно 0.3–0.4 Å. Докинг лигандов при этом осуществляется в куб докинга (обычно он имеет длину ребра 22 Å), охватывающий (с большим запасом) активный центр белка-мишени. В этом кубе в узлах трехмерной сетки размером $101 \times 101 \times 101$ записаны потенциалы взаимодействия атомов лиганда с атомами всего белка, включая те атомы белка, которые находятся вне куба докинга. В каждом узле записаны значения различных потенциалов взаимодействия каждого типа пробных атомов лиганда: электростатическое взаимодействие, вандерваальсово взаимодействие, потенциалы десольватации, а типы пробных атомов лиганда относятся к типам, определяемым силовым полем MMFF94.

2. Упрощенная типизация атомов: некоторые типы атомов, являющиеся различными в соответствии с типизацией силового поля MMFF94, объединяются в единые типы на основе некоторых интуитивных соображений и вместо 99 возможных сеток потенциалов создается всего 27 сеток потенциалов для 27 различных упрощенных типов пробных атомов [23].

3. При вычислении энергии десольватации влияние растворителя учитывается при помощи упрощенной обобщенной модели Борна [28], в которой упрощения вводятся, чтобы нелокальную задачу электро-

статического экранирования свести к локальным потенциалам на сетке.

4. При позиционировании лиганда в процессе работы программы докинга локальная оптимизация энергии по положениям атомов лиганда не проводится, что, вообще говоря, может приводить к завышению энергии системы лиганд–белок.

5. Длины связей и валентные углы между связями лиганда в процессе докинга остаются неизменными, а лиганд может изменять свое положение при помощи торсионных вращений вокруг одинарных связей, а также за счет вращения и трансляции его как целого.

6. Атомы белка создают вокруг него поля (электростатическое, вандерваальсово, поле эффектов десольватации). Когда лиганд занимает некоторое фиксированное положение относительно белка, атомы лиганда приобретают в каждом из этих полей определенную энергию, а весь лиганд приобретает энергию в виде суммы энергий составляющих его атомов и добавки, отвечающей за внутреннюю энергию напряжений лиганда, рассчитанную относительно его начального положения. При этом вклад в энергию отдельных атомов в поле белка вычисляется как линейная комбинация энергии атома в каждом из перечисленных выше полей.

Функция, которая минимизируется в процессе докинга, представляет собой эту линейную комбинацию, в которой все члены (электростатическое, вандерваальсово, десольватационное взаимодействия, а также энергия внутренних напряжений лиганда) берутся с единичными коэффициентами. Однако целевой функцией, т.е. рассчитанной энергией взаимодействия лиганда и белка, является другая функция, так называемая скоринг-функция, которая представляет собой ту же линейную комбинацию (за исключением энергии внутренних напряжений лиганда), взятую с некоторыми коэффициентами перед каждым энергетическим членом. Эти коэффициенты подбираются из соображений наилучшего соответствия экспериментальным результатам по константам связывания и положению нативных лигандов (т.е. таких лигандов, которые были закристаллизованы вместе с белком и которые, таким образом, имеют известные координаты в кристаллической структуре комплекса) в процессе валидации программы докинга. Помимо этого в скоринг-функцию входит дополнительный член, отвечающий за энтропийную составляющую энергии связывания белка и лиганда.

В дальнейшем функцию, минимизацию которой и осуществляет генетический алгоритм, будем обозначать как Docked energy (она представляет собой, по сути, энтальпию образования комплекса лиганд–белок), а целевую функцию — как Score (с ее помощью оценивается свободная энергия Гиббса связывания лиганд–белок).

Реализованный в программе докинга SOL генетический алгоритм (ГА) является математическим методом нахождения глобального минимума, в основе которого лежит процесс моделирования эволюции некой популяции особей по Дарвину с учетом генетических механизмов (мутаций, кроссинговера и т.п.): выживают сильнейшие. Особи — это положения лиганда в активном центре белка-мишени. Критерием выживаемости особи является энергия связывания лиганда с белком-мишенью: чем выше эта энергия, тем лучше выживает особь. В свою очередь, энергия связывания лиганд–белок выше тогда, когда полная энергия системы лиганд–белок имеет бóльшую по модулю отрицательную величину по сравнению с состоянием, когда белок и лиганд находятся далеко друг от друга и не взаимодействуют. Тем самым, положению лиганда в глобальном минимуме энергии системы лиганд–белок соответствует наибольшая энтальпия связывания лиганд–белок (Docked energy).

С точки зрения удобства реализации этого алгоритма особи, т.е. отдельные положения лиганда в активном центре белка-мишени, кодируются совокупностью генов, “генотипом” $\{a_i\}$, представляющим собой массив безразмерных чисел от 0 до 1, определяющих степени свободы лиганда. Совокупность этих генов (хромосома) полностью определяет положение линанда в активном центре белка-мишени. Таким образом, с помощью данного метода возможна эффективная генерация произвольных изменений положения лиганда в пространстве путем изменения его генотипа — совокупности значений генов. В программе SOL генотип особи определяет координаты геометрического центра лиганда, вращения лиганда как целого и углы вращения частей лиганда вокруг одинарных валентных связей.

Существует однозначное соответствие между генотипом особи (совокупность значения генов a_i) и ее фенотипом — конкретным положением лиганда в пространстве. После того как по генотипу вычисляется фенотип, т.е. декартовы координаты атомов лиганда, для каждого фенотипа может быть вычислена энергия взаимодействия протеина с лигандом и произведен естественный отбор. Чем ниже энергия комплекса лиганд–белок, тем особь “успешнее”.

Одна итерация алгоритма состоит из следующих шагов (рис. 1).

1. Начальная популяция, состоящая из некоторого числа особей, равного размеру популяции POPULATION SIZE, инициализируется путем заполнения вектора хромосомы случайными числами.

2. Рассчитывается значение оптимизируемой функции (энергии взаимодействия лиганд–белок — Docked energy) для каждой особи (т.е. для каждой конформации лиганда), а значения этой функции ранжируются.

3. В программе докинга SOL существуют три варианта выбора особей в так называемый пул родителей (MATING POOL), из которого формируется следующая популяция: пошаговый отбор с применением нишинга, метод, основанный на принципе колеса рулетки, и отбор особей с наименьшей энергией. В первом случае, после ранжирования всех особей из POPULATION SIZE по значению их энергий взаимодействия лиганд–белок и отбора в пул родителей первой особи с минимальной энергией, остальным особям присваивается некий "штраф" энергии, величина которого зависит от того, насколько особь близка по генотипу к особи, уже отобранной в пул родителей. Далее, особи вновь ранжируются уже с учетом наложенного штрафа, и процедура повторяется. Таким образом, подобный метод позволяет избежать попадания в пул родителей идентичных особей, а значит, и "вырождения" популяции, когда все особи задают примерно одно и то же положение лиганда. С математической точки зрения подобная процедура предотвращает преждевременное схождение алгоритма к одному из локальных минимумов энергии. Во втором случае отбор в пул родителей носит случайный характер и осуществляется с помощью некоторого числа запусков колеса рулетки, равного значению MATING POOL. Рулетка разделена на секторы, каждому из которых соответствует одна хромосома, причем величина каждого сектора устанавливается пропорциональной значению функции взаимодействия лиганд–белок данной хромосомы, поэтому чем больше по модулю отрицательное значение этой функции, тем больше сектор на колесе рулетки, а следовательно, тем выше шанс, что будет выбрана именно эта хромосома. В третьем случае в пул родителей отбираются особи с наибольшими по модулю отрицательными энергиями взаимодействия лиганд–белок.

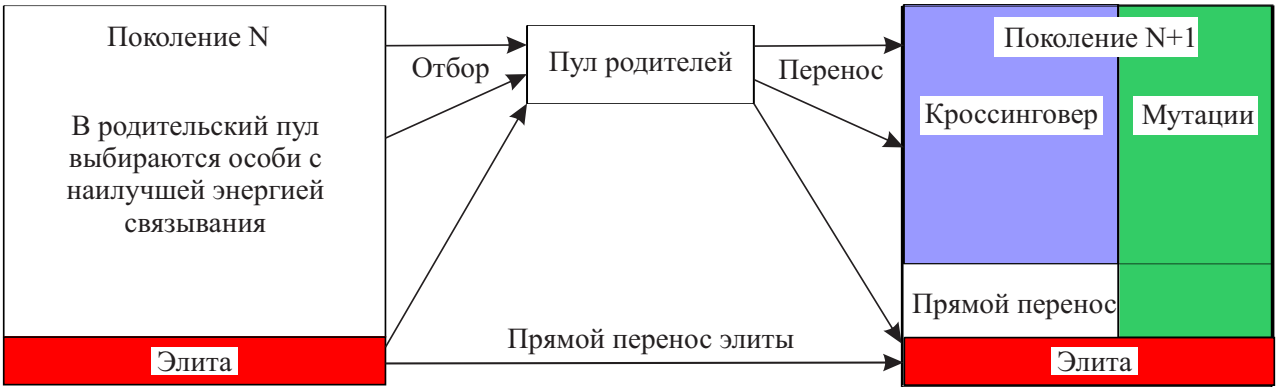

Рис. 1. Схема формирования нового поколения в ходе выполнения генетического алгоритма

4. После того как некоторое число особей, равное значению MATING POOL SIZE, переносятся в пул родителей, из него формируется следующая популяция. Небольшое количество самых "успешных" особей — элита — переносятся в следующее поколение без изменения. Другие же особи следующего поколения получаются при помощи комбинации генов родителей и случайного изменения генов родителей (путем применения операторов мутации и кроссинговера).

5. Если число поколений не достигло максимального значения, то происходит переход к п. 2.

Процедура смены поколений повторяется некоторое количество раз (количество поколений, NUMBER OF GENERATION), причем минимальное значение оптимизируемой функции в каждом поколении уменьшается с ростом номера поколения. Процедура прекращается по достижении заданного количества поколений NUMBER OF GENERATION.

Так как ГА не может гарантировать нахождение глобального минимума Docked energy и соответствующего лучшего положения лиганда, то для контроля достоверности полученных результатов проводится несколько независимых запусков ГА. Результаты каждого запуска запоминаются, и после проведения всех независимых запусков ГА определяется, насколько похожи их результаты между собой. При достаточно близких результатах, полученных в разных независимых запусках ГА, они считаются достоверными.

Работа генетического алгоритма контролируется рядом параметров, значения которых подбираются в зависимости от поставленной задачи и влияют на качество докинга. Оценка качества работы генетического алгоритма во время исследования может быть проведена по нескольким критериям.

1) Оценка функции минимизации (энтальпии образования комплекса лиганд–белок) Docked energy и оценка целевой скоринг-функции Score. Поскольку эти две функции связаны между собой, как описано

выше, то можно ожидать, что с уменьшением значения Docked energy будет уменьшаться и значение Score.

2) Оценка качества позиционирования, характеризующаяся среднеквадратичным отклонением положения лиганда после докинга от его нативного положения в белке — RMSD. Предполагается, что система лиганд–белок была закристаллизована в состоянии с минимальной свободной энергией взаимодействия. Однако в случае молекулярного моделирования в силу неточного соответствия минимизируемой функции реальной свободной энергии связывания, а также в силу погрешностей при кристаллизации это условие может не выполняться, а при минимальном значении Docked energy положение лиганда не всегда будет точно соответствовать его нативному положению в кристаллической структуре комплекса. Таким образом, это один из критериев, по которому можно определить качество докинга, но не эффективность работы генетического алгоритма.

3) Оценка кластеризации, показывающая, объединяются ли решения для разных независимых запусков программы в один или несколько кластеров близких друг к другу поз лиганда или же все они отличаются друг от друга, что говорит о том, что глобальный минимум, вероятно, не был найден. В программе докинга SOL решения, RMSD-расстояния между которыми не превышают 1 Å, собираются в один кластер. Первым обозначается кластер, в который после докинга входит решение, обладающее наименьшей энергией Docked energy. Населенность первого кластера, обозначаемая через N1, и характеризует успешность нахождения глобального минимума: чем больше решений в этом кластере, тем чаще алгоритм сходился к данному минимуму.

4) Оценка затрат ресурсов на проведение вычислений.

Цель настоящей статьи — поиск таких параметров генетического алгоритма, при которых задача докинга низкомолекулярных соединений в белки программой SOL решается наиболее эффективно.

3. Параметры генетического алгоритма. Для анализа эффективности работы программы докинга SOL были взяты комплексы урокиназы лиганд–белок из базы данных PDB [29]: всего 20 комплексов. Все комплексы были подготовлены для докинга в программе SOL. Лиганды имели разное количество внутренних вращательных степеней свободы (от 6 до 19).

Докинг проводился с помощью многопроцессорной версии программы SOL [24], на вычислительных кластерах МГУ “Ломоносов” и “Чебышев”, что позволило значительно увеличить скорость расчетов и сократить время проведенных исследований.

Такие параметры, как количество независимых запусков программы, размер популяции и количество вычислений энергии в одном запуске (число поколений), часто служат предметом исследования в случае каждой новой реализации генетического алгоритма, а также для различных применений алгоритма (например, докинг небольших лигандов и докинг пептидов) [30–32].

Здесь мы рассмотрим следующие параметры генетического алгоритма, величина которых оказывает влияние на эффективность работы программы SOL: количество независимых запусков программы; размер популяции; число поколений; способ отбора в пул родителей; способ кроссинговера; параметры, характеризующие кроссинговер и мутации.

3.1. Количество независимых запусков программы. NUMBER OF RUNS (NOR) — параметр, определяющий количество выполняемых независимых запусков генетического алгоритма. Очевидно, что увеличение количества независимых запусков программы повысит вероятность нахождения глобального минимума, однако время, требуемое на расчеты, будет расти пропорционально количеству запусков.

Было выбрано 3 точки для проверки: 20, 50, 99 независимых запусков. Результаты докинга, включающие в себя энергии взаимодействия белка и лиганда, скоринг-функции, значения среднеквадратичных отклонений найденных при докинге положений от нативных, а также оценку кластеризации (населенность первого кластера), приведены в табл. ДМ1 в Дополнительных материалах [33]. В результате докинга для 20 комплексов лиганд–белок было показано, что при 20 независимых запусках 4 комплекса из 20 не достигают минимального значения энергии взаимодействия, которое находится при других значениях параметров. При этом нативные лиганды в трех из этих четырех комплексов имели большое число внутренних торсионных степеней свободы (16, 17 и 19), что очевидно затрудняло работу генетического алгоритма. При 50 запусках минимального значения не достигает один комплекс, а также для одного из комплексов наименьшая энергия взаимодействия наблюдается при 20 независимых запусках, а при 50 и 99 — это минимальное значение, найденное для 20 независимых запусков, не достигается. Таким образом, и при 50, и при 99 независимых запусках программы алгоритм эффективно находит глобальный минимум, но во втором случае требуется в два раза больше временны́х затрат, что очевидно является излишним. С другой стороны, даже при 20 запусках для лигандов с небольшим числом внутренних торсионных степеней свободы (меньше 14) глобальный минимум обычно достигается, что можно использовать при скрининге

больших баз данных.

3.2. Размер популяции и количество поколений. Параметр POPULATION SIZE (PS) определяет количество особей в популяции (количество особей, создаваемых на одном шаге эволюции). Параметр NUMBER OF GENERATIONS (NG) определяет, сколько раз нужно провести смену поколений в одном запуске генетического алгоритма. Эти параметры варьировались вместе, поскольку оба значительно влияют как на эффективность работы генетического алгоритма, так и на временны́е затраты.

Помимо непосредственно параметров генетического алгоритма, перечисленных выше, программа докинга SOL содержит также параметры расчета энергии взаимодействия лиганд–белок. К ним относится уже упомянутый параметр уширения потенциала. Физический смысл этого параметра заключается в уширении потенциальной ямы вандерваальсова взаимодействия, отражающем тот факт, что атомы белка обладают ограниченной подвижностью и способны подстраиваться под конкретный лиганд. Введением этого уширения косвенным образом учитывается подвижность атомов белка в используемой модели жесткого белка. Можно предположить, что чем шире окажется потенциальная яма энергии взаимодействия лиганда и белка, тем проще будет генетическому алгоритму находить минимум этой ямы.

На этом этапе работы мы использовали два значения параметра уширения потенциала: 0.3 Å и 0.4 Å, наиболее часто используемые в ходе докинга (при этих значениях при докинге нативных лигандов наблюдаются минимальные среднеквадратичные отклонения положения молекулы после докинга от ее экспериментального положения в кристаллическом комплексе с белком). Как и было замечено, оптимальные значения размера популяции и количества поколений зависели от этого параметра.

Для начала было взято 4 комплекса лиганд–белок, выбранные на основе того, что при небольших значениях параметров числа поколений и размера популяции скоринг-функции не достигали глобального минимума, тогда как при этих же значениях параметров глобальный минимум достигался для многих других комплексов. Три из этих четырех комплексов имели лиганды с большим числом внутренних вращательных степеней свободы (16, 17 и 19). Еще для одного комплекса в результате докинга программа находила два кластера решений с высоким числом особей в каждом, причем в зависимости от значений параметров населенности этих кластеров различались.

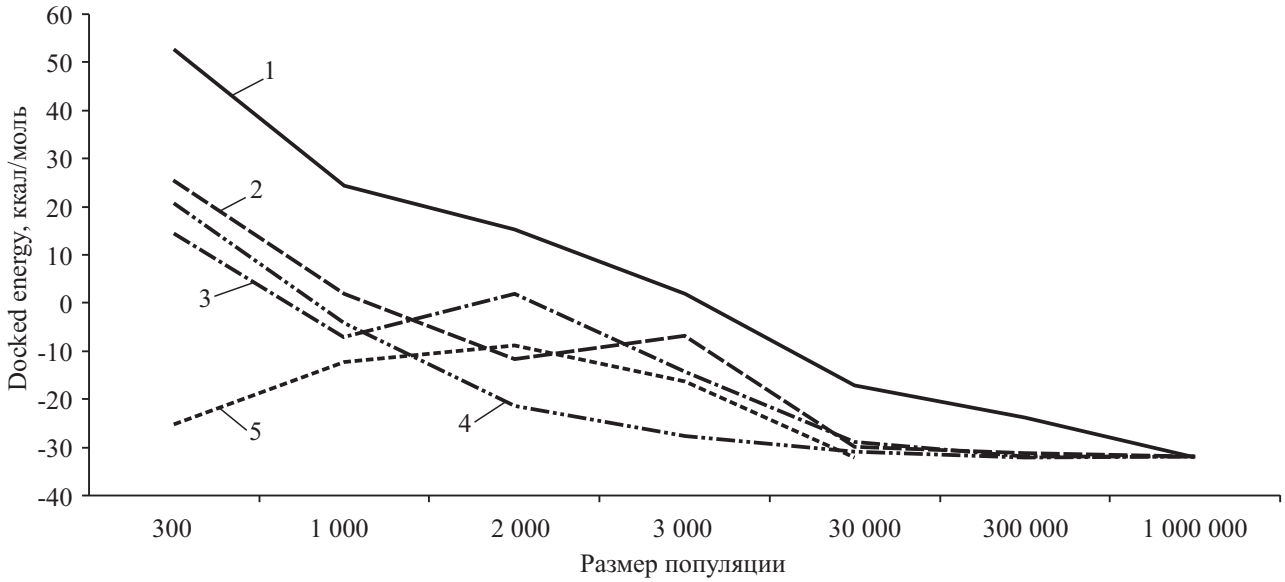

Рис. 2. Зависимость Docked energy от размера популяции и числа поколений для комплекса 1f92. NG — Number of generation — число поколений: 1) NG 200, 2) NG 1000, 3) NG 2000, 4) NG 20000, 5) NG 200000

Значения размера популяции варьировались в пределах от 300 до 1000000, а значения числа поколений — от 200 до 200 000. Соответствующие результаты приведены в табл. ДМ2а и ДМ2б в Дополнительных материалах [34]. При определенных наборах параметров значения энергий связывания выходят на плато (рис. 2 и 3) и не дают дальнейших улучшений при увеличении размера популяции или числа поколений (можно полагать, что в этом случае программа находит глобальный минимум энергии). Подобное увеличение нецелесообразно также с точки зрения требуемых вычислительных ресурсов (табл. 1). На рис. 2 и 3 приведены зависимости Docked energy от размера популяции и числа поколений для одного из обозначенных комплексов (1f92) для уширения потенциала 0.3 Å. Из этих рисунков следует, ято Docked energy

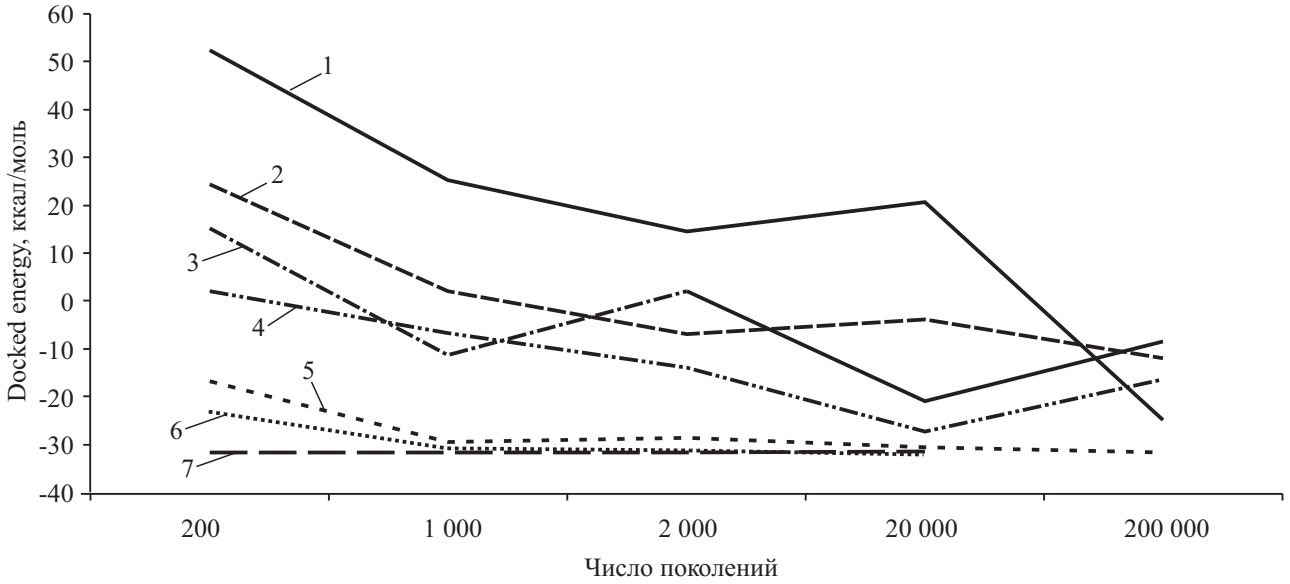

Рис. 3. Зависимость Docked energy от размера популяции и числа поколений для комплекса 1f92. PS — Population size — размер популяции: 1) PS 300, 2) PS 1000, 3) PS 2000, 4) PS 3000, 5) PS 30000, 6) PS 300000, 7) PS 1000000

не опускается ниже некоторого минимального значения. В некоторых случаях (например, при размере популяции 1000 или 2000 и числе поколений 200 000 для комплекса 1f92) наблюдается отклонение: энергия не убывает относительно размера популяции, равного 300, а наоборот возрастает. Это вероятнее всего обусловлено случайностью поиска глобального минимума, которая при небольших значениях параметров играет более значительную роль (некоторые участки пространства остаются не исследованными). При дальнейшем увеличении параметров Docked energy снова начинает убывать, пока не достигает минимального значения.

Таблица 1

Наборы параметров (число поколений — размер популяции), при которых значение Docked energy достигает минимума, и требуемые для расчетов при этих параметров ресурсы

Уширение 0.3 Å				Уширение 0.4 Å			
Число поколений	Размер популяции	Время счета	Количество процессоров	Число поколений	Размер популяции	Время счета	Количество процессоров
200	1 000 000	20–90 мин	128	200	1 000 000	20–90 мин	128
1000	300 000	30–90 мин.	32	1000	300 000	30–90 мин	32
2000	30 000	10–90 мин.	32	2000	30 0000	2–6 часов	64
20 000	3000	2–15 часов	32	20 000	3000	10–90 мин	32

Как видно из табл. 1, наборы параметров “число поколений 2000, размер популяции 30 000”, “число поколений 20 000, размер популяции 3000” и “число поколений 1000, размер популяции 300 000” требуют меньших вычислительных ресурсов, чем при использовании других наборов. Кроме того, следует отметить, что время счета зависит не только от выбранных параметров генетического алгоритма, но и от гибкости лиганда (количества степеней свободы), что обусловливает разброс времен в каждой из ячеек таблицы.

При более тщательном исследовании выбранных параметров на расширенном наборе из 20 комплексов лиганд–белок для наборов “число поколений 2000, размер популяции 30 000” и “число поколений 20 000, размер популяции 3000” наблюдались единичные непопадания в глобальный минимум. Примеры приведены в табл. 2. Жирным шрифтом в таблице отмечены варианты, для которых наблюдаются наилучшие значения Docked energy. Принятые в таблице обозначения: ID — номер структуры в базе данных PDB, DE — Docked energy — энтальпия образования комплекса лиганд–белок (минимизируемая функция), скоринг-функция — целевая функция, RMSD — среднеквадратичное отклонение положения лиганда после докинга от нативного положения, N1 — число особей в первом кластере.

Из табл. 2 видно, что для всех приведенных примеров в случае выбора числа поколений 1000 и размера популяции 300 000 алгоритм находит более глубокий минимум энергии (большие по модулю отрицательные значения Docked energy), а целевая функция (Score) при этом также принимает лучшие значения. К тому же можно заметить, что населенность первого кластера во всех случаях выше для набора “1000–300 000”, что дает бóльшую уверенность в том, что найденный минимум и есть искомый глобальный минимум.

Таблица 2

Результаты докинга для некоторых комплексов при различных наборах параметров “число поколений–размер популяции”

Уширение 0.3 Å	ID = 3ig6			
Число поколений–размер популяции	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
1000–300 000	-5.102	-4.937	10.355	7
2000–30 000	-4.242	2.056	10.273	3
Уширение 0.3 Å	ID = 1w13			
Число поколений–размер популяции	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
1000–300 000	-4.607	-134.748	1.986	6
2000–30 000	-3.716	-131.079	1.719	2
Уширение 0.4 Å	ID = 1w0z			
Число поколений–размер популяции	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
1000–300 000	-5.200	-98.669	2.178	3
2000–30 000	-4.431	-93.001	2.043	1
Уширение 0.4 Å	ID = 1w13			
Число поколений–размер популяции	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
1000–300 000	-4.532	-134.220	2.034	1
2000–30 000	-3.605	-129.570	2.143	1
Уширение 0.4 Å	ID = 3ig6			
Число поколений–размер популяции	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
1000–300 000	-5.094	-5.144	10.380	6
2000–30 000	-4.347	4.618	2.319	1

Приведенные в таблице значения среднеквадратичных отклонений положения лиганда после докинга от нативного положения (RMSD) хотя и являются критериями качества докинга, но, вообще говоря, не связаны напрямую со значением минимизируемой функции (Docked energy). Тем не менее, мы считаем, что нативное положение лиганда соответствует минимальному значению реальной (например, определенной экспериментально) свободной энергии взаимодействия лиганд–белок. Следует также заметить, что в случае комплекса 3ig6 мы можем наблюдать значение среднеквадратичного отклонения положения лиганда после докинга относительно нативного положения равное 10.3 Å при параметрах 1000–300 000, что выходит далеко за пределы хорошего докинга, и 2.3 Å при параметрах 20 000–3000.

Лиганд (рис. 4) обладает большим числом степеней свободы (11 внутренних и 17, включая вращение и перемещение лиганда как целого); это одна из причин, которые затрудняют докинг. Другая причина состоит в том, что функция энергии взаимодействия лиганда и белка вполне может иметь несколько близких минимумов энергии, что правомерно и для реальной системы лиганд–белок (т.е. для реальных

значений свободной энергии взаимодействия лиганд–белок), и нативное положение лиганда в кристаллической структуре может быть одним из минимумов, тогда как программа находит другой. Не следует также исключать того факта, что минимизируемая функция (Docked energy) имеет свои недостатки и неточности, и поэтому минимум реальной свободной энергии взаимодействия белка и лиганда не соответствует минимуму Docked energy. Если мы предполагаем, что нативное положение лиганда соответствует глобальному минимуму реальной свободной энергии, то алгоритм, находя глобальный минимум своей минимизируемой функции, находит соответственно и другое положение лиганда.

Таким образом, были определены следующие оптимальные значения размера популяции и числа поколений для двух значений параметра уширения потенциала: число поколений равно 1000, число особей в популяции равно 300 000.

Поскольку в большинстве случаев среднеквадратичное отклонение положения лиганда, найденного после докинга, от нативного имеет меньшее значение при уширении потенциала, равном 0.3 Å, то далее все результаты будут приведены для этого значения уширения. К тому же процедура уширения потенциалов представляет собой в некотором роде искусственное завышение энергии взаимодействия, которое тем больше, чем больше значение данного параметра.

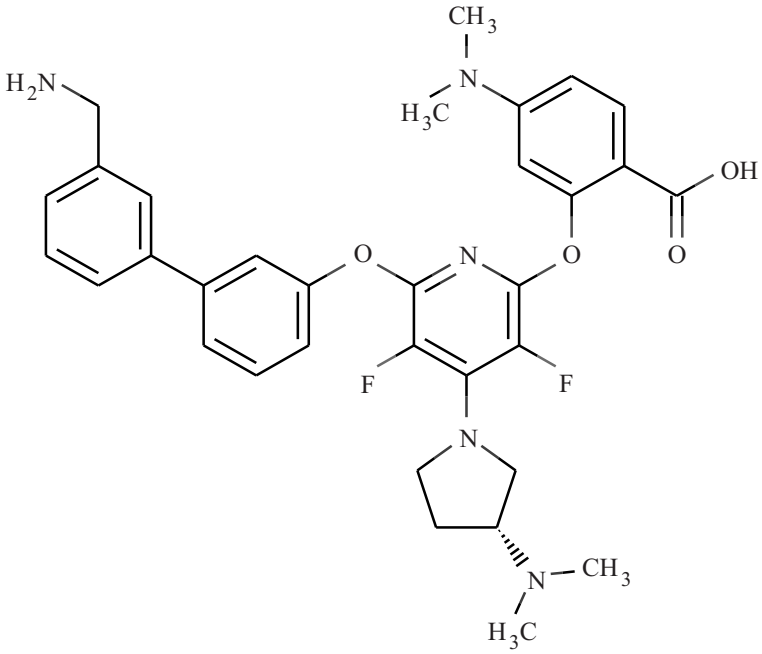

Рис. 4. Лиганд комплекса 3ig6

3.3. Размер пула родителей. Параметр MATING POOL SIZE определяет количество родителей следующего поколения. Эти родители набираются из популяции, и при помощи их попарных комбинаций и прямого копирования создается новая популяция. Значения параметра варьировались от 10 особей до 30% от числа особей в популяции, равного 100 000 (табл. ДМ3 в Дополнительных материалах [35]). При больших размерах пула родителей возникает нехватка места в буфере памяти, предоставленном пользователю, и программа прерывает свое выполнение. Это исследование проводилось также только на 4 “проблемных” комплексах. Варьирование размера пула родителей дало разброс приемлемых значений от 50 до 100 особей, при которых алгоритм достаточно стабильно находит глобальный минимум. При этом для MATING POOL SIZE = 100 в большинстве случаев наблюдались лучшие результаты. Время, требуемое на докинг одного лиганда, также не зависело от этого параметра.

3.4. Параметры выбора одного из трех вариантов алгоритма выбора родителей. Как уже упоминалось выше, в программе SOL реализованы следующие варианты:

STEP FUNCTION SELECTION WITH NICHING — последовательный выбор родителей с учетом нишинга, который строится на основании их энергии (должна быть низка) и генетического (т.е. геометрического) разнообразия в пуле отобранных родителей (должно быть высоко);

ROULETTE WHEEL SELECTION — выбор родителей осуществляется случайным образом с вероятностью, пропорциональной значениям энергии связывания;

STEP FUNCTION SELECTION — отбираются родители с самыми низкими энергиями связывания.

В первом случае дополнительный параметр NICHING DIVERSITY FACTOR определяет, насколько разные родители должны попадать в пул родителей (чем он больше, тем больше разнообразия). Для этого после каждой выбранной в пул родителей особи вычисляется 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 и b_i — значения генов двух особей, между которыми вычисляется расстояние, а к значению общей энергии особей текущего поколения прибавляется величина NICHING DIVERSITY FACTOR$/D$.

Величины значений NICHING DIVERSITY FACTOR варьировались от 0 до 10 000 [ккал/моль] (см. табл. ДМ4 в Дополнительных материалах [35]); в интервале от 0.8 до 3 наблюдался некоторый оптимум

значений, при которых для всех четырех комплексов довольно уверенно находится глобальный минимум.

В случае выбора алгоритма родителей, основанного на запуске колеса рулетки, используется параметр FITNESS NORMALIZATION COEFFICIENT, который определяет, насколько сильно зависит от энергии лиганда его плотность вероятности попасть в пул родителей. Если E_{best} — лучшая энергия в популяции, то для особи с энергией E выделяется сектор круга рулетки с площадью

$$S = \exp\big(-(\text{FITNESS NORMALIZATION COEFFICIENT}) \times (E - E_{\text{best}})\big).$$

После запуска рулетки и выбора ее случайного участка особь, связанная с этим участком, переносится в пул родителей и снимается с рулетки.

При варьировании параметра FITNESS NORMALIZATION COEFFICIENT в диапазоне значений от 0.1 до 10 000 [моль/ккал] этот коэффициент ни на значения скоринг-функций, ни на временны́е затраты при вычислениях не влияет. Однако при низких значениях этого коэффициента (< 0.001) программа SOL не находит минимум энергии, определенный при значениях > 0.001; следовательно, эффективность работы алгоритма падает (табл. ДМ5 в Дополнительных материалах [35]).

Сравнение эффективности алгоритмов выбора в пул родителей проведено на 20 комплексах. Лучшие результаты достигались при использовании алгоритма с нишингом (STEP FUNCTION SELECTION WITH NICHING) или при использовании алгоритма, основанного на запуске колеса рулетки (ROULETTE WHEEL SELECTION): во всех 20 случаях докинг дал схожие минимальные энергии. В случае отбора в пул родителей особей с минимальными энергиями (STEP FUNCTION SELECTION) для 6 комплексов из 20 минимум энергии не был достигнут.

3.5. Количество особей элиты. ELITISM — определяет, сколько наилучших особей должно заведомо переноситься в следующее поколение, чтобы найденный результат не был потерян.

Параметр варьировался от 0 до числа особей в Mating pool (было выбрано 70); в этих пределах количество особей в элите мало влияло на значение скоринг-функции (табл. ДМ6 в Дополнительных материалах [35]). Это можно объяснить тем, что число особей в популяции велико (300 000) и этого разнообразия оказывается достаточно, чтобы не только не потерять, но даже улучшить результат, полученный в предыдущем поколении. Очевидно, что при небольшом значении числа особей в популяции этот параметр будет играть более значительную роль. Тем не менее, при любых значениях размера популяции введение этого параметра дает некую гарантию того, что найденные лучшие особи не будут потеряны в следующих поколениях в результате применения операторов мутации и кроссинговера.

3.6. Параметры выбора одного из трех вариантов алгоритма кроссинговера. Кроссинговер представляет собой модель полового размножения — создание новой хромосомы из двух родительских хромосом путем случайного объединения их генов. В программе докинга SOL реализованы три вида кроссинговера:

UNIFORM CROSSOVER — однородный кроссинговер строится на случайном выборе каждого i-го гена потомка из i-го гена первого родителя (с вероятностью P) или i-го гена второго родителя (с вероятностью $1 - P$);

ONE POINT CROSSOVER — одноточечный кроссинговер: в геном потомка включаются первые гены от первого родителя и последние гены от второго родителя; точка разделения “первых” и “последних” генов выбирается случайным образом;

TWO POINT CROSSOVER — двухточечный кроссинговер: выбираются две точки разделения генома потомка, средняя часть заполняется соответствующим куском генома первого родителя, а первая и последняя части — соответствующими кусками генома от второго родителя.

Вариация параметра проводилась на 20 комплексах. За исключением четырех случаев, приведенных в табл. 3, различные варианты кроссинговера дали одинаковые результаты. Жирным шрифтом в таблице отмечены варианты кроссинговера, для которых наблюдаются наилучшие значения Docked energy. Принятые в таблице обозначения: ID — номер структуры в базе данных PDB, DE (Docked Energy) — энтальпия образования комплекса лиганд–белок (минимизируемая функция), Скоринг-функция — целевая функция, RMSD — среднеквадратичное отклонение положения лиганда после докинга от нативного положения, N1 — число особей в первом кластере.

Из табл. 3 можно заметить, что большее число непопаданий в минимум наблюдается для одноточечного кроссинговера (ONE POINT CROSSOVER), тогда как двухточечный кроссинговер (TWO POINT CROSSOVER) для всех случаев дает хороший результат.

3.7. Параметры для операторов мутаций и кроссинговера. Параметр CROSSOVER RATE определяет, какая доля от популяции следующего поколения будет получена путем попарных комбинаций родительских особей. Остальная часть популяции получается непосредственным переносом геномов

Таблица 3

Различные результаты докинга при использовании различных вариантов алгоритма кроссинговера

Тип кроссинговера	ID = 1ejn			
	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
Одноточечный	-5.967	-172.207	1.405	1
Двухточечный	**-6.578**	**-177.139**	**0.851**	**1**
Однородный	**-6.637**	**-178.591**	**0.829**	**1**
Тип кроссинговера	ID = 1f92			
	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
Одноточечный	-5.094	-30.307	1.819	2
Двухточечный	**-5.359**	**-31.474**	**1.423**	**2**
Однородный	**-5.421**	**-31.996**	**1.44**8	**6**
Тип кроссинговера	ID = 1vj9			
	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
Одноточечный	-3.217	-41.7812	1.359	1
Двухточечный	**-4.488**	**-48.824**	**1.113**	**2**
Однородный	**-4.551**	**-50.403**	**1.309**	**1**
Тип кроссинговера	ID = 1vja			
	Скоринг-функция, ккал/моль	DE, ккал/моль	RMSD, Å	N1
Одноточечный	**-5.09**2	**-70.396**	**0.934**	**1**
Двухточечный	**-4.690**	**-68.62**2	**0.869**	**2**
Однородный	-3.940	-60.312	5.633	1

родителей целиком.

Практически для всех значений CROSSOVER RATE > 0.3 алгоритм успешно находит глобальный минимум, поэтому оптимальное значение выбиралось на основе лучшей кластеризации результатов и было принято равным 0.9 (табл. ДМ7 в Дополнительных материалах [35]).

Параметр MUTATION PROBABILITY задается тремя вещественными числами в одной строчке, разделенными пробелом. Первое число определяет, насколько вероятна мутация в гене, ответственном за вращение вокруг торсионной степени свободы. Второе число определяет вероятность мутации в генах, ответственных за поворот лиганда как целого. Третье число определяет вероятность мутации в трансляционных генах (перенос лиганда как целого).

Вероятность мутации генов, ответственных за торсионы, варьировалась от 0 до 1. При значениях от 0.08 до 0.2 для всех 4 комплексов программа докинга находила глобальный минимум. Оптимальным было выбрано значение 0.1.

Вероятность мутации в генах, ответственных за поворот лиганда как целого, и вероятность мутации в трансляционных генах варьировались в тех же пределах. Для второго параметра алгоритм устойчиво находил глобальный минимум при значениях больше 0.6, а для третьего параметра — при значениях больше 0.4 (табл. ДМ8, ДМ9 и ДМ10 в Дополнительных материалах [35]). Оптимальными в обоих случаях были выбраны значения 0.9.

Для параметра MUTATION WINDOW также задаются три вещественных числа в одной строчке, разделенные пробелом. Эти числа определяют форму кривой плотности вероятности отклонения значения мутированного гена от исходного для генов, контролирующих мутации, ответственные за вращение вокруг торсионных степеней свободы (MUTATION WINDOW 1), мутации, ответственные за вращение лиганда как целого (MUTATION WINDOW 2), и мутации трансляционных генов (MUTATION WINDOW 3).

Например, для мутаций в торсионах параметр MUTATION WINDOW 1 входит в уравнение, согласно которому определяется случайная величина отклонения значения каждого гена от его текущей величины в процессе мутации: $\triangle = 0.5 \times \big[(\text{MUTATION WINDOW 1})x^3 + (1 - \text{MUTATION WINDOW 1})x\big]$, где x — случайная величина с равномерным распределением на отрезке [-1,1].

Таким образом, параметр MUTATION WINDOW эффективно контролирует форму распределения случайной величины отклонения мутированной величины гена от исходного значения, причем в случае MUTATION WINDOW = 1 плотность вероятности отклонения имеет ярко выраженный пик при $\triangle = 0$, в то время как при MUTATION WINDOW = 0 плотность распределения равномерная.

При варьировании первого параметра — кривизны распределения в мутациях торсионных генов — в пределах от 0 до 10 оптимальным значением было выбрано 1. Другие два параметра также варьировались в пределах от 0 до 10; для MUTATION WINDOW 2 было также выбрано значение 1 (в целом параметр мало влияет на скоринг-функцию), а для MUTATION WINDOW 3 — значение 0.9 (табл. ДМ11, ДМ12 и ДМ13 в Дополнительных материалах [35]).

4. Результаты и выводы. В ходе работы были выявлены параметры генетического алгоритма, оказывающие наиболее сильное влияние на эффективность работы программы докинга, и получены оптимальные значения этих параметров. Их значения приведены в табл. 4, в которой приняты названия параметров генетического алгоритма такими, как они были обозначены выше: число независимых запусков программы — NUMBER OF RUNS, размер популяции — POPULATION SIZE, число поколений — NUMBER OF GENERATIONS, размер пула родителей — MATING POOL SIZE, количество особей элиты — ELITISM, алгоритм кроссинговера — CROSSOVER ALGORITHM, алгоритм выбора родителей — SELECTION ALGORITHM, нишинг (параметр разнообразия родителей) — NICHING DIVERSITY FACTOR, вероятность кроссинговера — CROSSOVER RATE, вероятность мутации 1 (мутации в торсионных степенях свободы) — MUTATION PROBABILITY 1, вероятность мутации 2 (мутации при вращении лиганда как целого) — MUTATION PROBABILITY 2, вероятность мутации 3 (мутации при трансляции лиганда как целого) — MUTATION PROBABILITY 3, форма кривой плотности вероятности мутации торсионов — MUTATION WINDOW 1, форма кривой плотности вероятности мутации вращения лиганда — MUTATION WINDOW 2, форма кривой плотности вероятности мутации трансляции лиганда — MUTATION WINDOW 3.

Таблица 4

Оптимальные значения параметров генетического алгоритма

NUMBER OF RUNS	50
POPULATION SIZE	300000
NUMBER OF GENERATIONS	1000
MATING POOL SIZE	100
ELITISM	4
CROSSOVER ALGORITHM	TWO POINT CROSSOVER
SELECTION ALGORITHM	STEP FUNCTION SELECTION WITH NICHING
NICHING DIVERSITY FACTOR	1
CROSSOVER RATE	0.9
MUTATION PROBABILITY 1	0.1
MUTATION PROBABILITY 2	0.9
MUTATION PROBABILITY 3	0.9
MUTATION WINDOW 1	1
MUTATION WINDOW 2	1
MUTATION WINDOW 3	0.9

Заметное влияние на время расчетов оказывают такие параметры, как количество независимых запусков, размер популяции и количество поколений. Время счета при значениях этих параметров, приведенных выше, составляет в среднем от 20 до 90 минут на 64 процессорах. Ускорение и эффективность расчетов в зависимости от числа вычислительных ядер исследованы в работе [24]. От остальных пара-

Таблица 5
Сравнение рекомендованных параметров генетического алгоритма для программ SOL, AutoDock, GOLD и MolDock

Параметр	SOL	AutoDock	GOLD	MolDock
NUMBER OF RUNS	50	50	50	10
POPULATION SIZE	300000	150	100	50
NUMBER OF GENERATIONS	1000	27000	100000	2000
CROSSOVER RATE	0.9	0.8	0.95	0.9
MUTATION PROBABILITY	0.1 0.9 0.9	0.2	0.95	—

метров временны́е характеристики алгоритма не зависят.

Для сравнения в табл. 5 приведены параметры генетического алгоритма, используемые по умолчанию программами AutoDock, Gold и MolDock. Различия эти обусловлены отличиями в реализациях генетического алгоритма в конкретных программах, а также собственными соображениями разработчиков относительно соотношения между быстродействием и качеством работы программы.

Следует отметить, что хотя поиск глобального минимума является основной задачей докинга, знание значений среднеквадратичных отклонений для различных независимых запусков программы, а также знание кластеризации решений позволяет определить наличие и других минимумов функции энергии взаимодействия лиганда и белка, которые находит программа в ходе генетического алгоритма. Таким образом, можно получить не одну конфигурацию лиганд–белок, а наборы таких конфигураций, и реальная система может существовать с некоторой вероятностью в нескольких состояниях.

Подобная задача обычно решается с помощью методов молекулярной динамики, однако видно, что программа докинга также может находить различные минимумы энергии.

Можно заметить, что при использовании уширения, равного 0.4 Å для комплекса 1F5L при меньших значениях числа особей и числа популяций, программа докинга находила минимум со значением скоринг-функции ~ -3.9 ккал/моль и хорошим значением RMSD — ~ 0.69 Å. При увеличении параметров программа находит более глубокий минимум ~ -4.7 ккал/моль, но с RMSD уже 3.1 Å. Очевидно, что первое положение более близко к нативному положению лиганда. Такое положение вещей можно объяснить как неточностью расчета скоринг-функции программой SOL, так и тем, что, вероятно, комплекс оказался закристаллизован не в самом выгодном энергетическом состоянии, а в некотором соседнем локальном минимуме энергии.

Похожая ситуация наблюдается и при докинге комплекса 1VJ9. При этом значения скоринг-функций мало отличаются по энергии (значения обоих минимумов ~ -4.5 ккал/моль), однако по значениям среднеквадратичных отклонений от нативного положения лиганда видно, что это различные конфигурации лиганд–белок (соответствующие значения RMSD ~ 1 Å и ~ 3.2 Å).

Варьирование параметров и оценка эффективности работы программы SOL проводилась в данной работе только на комплексах урокиназа–лиганд; тем не менее, есть основания полагать, что и для других белков алгоритм будет устойчиво находить глобальный минимум энергии взаимодействия протеин–лиганд с приведенными выше значениями параметров. Для примера мы провели докинг 148 комплексов лиганд–белок (не имеющих отношение к урокиназе) из базы данных PDB (табл. ДМ14 в Дополнительных материалах [35]) как для приведенных выше параметров, так и для менее строгих параметров (NUMBER OF RUNS = 20, POPULATION SIZE = 30 000, NUMBER OF GENERATIONS = 500). При этом в 66 случаях из 148 (45%) более строгие параметры давали улучшение энергии Docked energy хотя бы на 1 ккал/моль, что говорит о нахождении более глубокого минимума. Кроме того, в 29 случаях улучшалось среднеквадратичное отклонение от нативного положения (RMDS) хотя бы на 1 Å. Неудачный докинг (RMSD > 3 Å, Score > -5 ккал/моль) в большинстве случаев наблюдался для лигандов с большим числом степеней свободы (> 10 торсионов).

С помощью программы SOL был также выполнен докинг комплексов, предоставленных для конкурса ресурсом CSAR (Community Structure-Activity Resource) [36] в 2009 г. База данных состояла из 345 комплексов, из которых мы исключили те, которые имели в активном центре атомы металлов, оставив для рассмотрения 281 комплекс. Закономерности, прослеживаемые при варьировании размера популяции в пределах от 30 000 до 100 000, не противоречили результатам, полученным в статье.

Расчеты были проведены для значений числа независимых запусков = 50, числа поколений = 1000 и размера популяции = 30 000, в результате чего 67% лигандов, для которых был проведен докинг, имели RMSD менее 3 Å по отношению к нативному положению, а 59% — RMSD менее 2 Å.

Следует также заметить, что рекомендованные в работе параметры выбраны на основе результатов докинга лигандов, обладающих большим числом степеней свободы. Однако при проведении объемного скрининга многих соединений — кандидатов в лекарственные средства, имеющих зачастую небольшие размеры и меньшее число внутренних вращательных степеней свободы, разумнее будет уменьшить значения таких параметров, как число независимых запусков программы, размер популяции и число поколений.

Пример выбора параметров для скрининга больших баз данных лигандов приведен в табл. 6. При этих значениях докинг одного лиганда на 32 вычислительных ядрах занимает от нескольких секунд до нескольких минут, а на одном вычислительном ядре — от получаса до двух часов.

Таблица 6
Пример выбора параметров генетического алгоритма для проведения скрининга больших баз данных

Параметр	SOL
NUMBER OF RUNS	20
POPULATION SIZE	30 000
NUMBER OF GENERATIONS	500

5. Заключение. Генетический алгоритм представляет собой удобный инструмент для решения задачи поиска глобального минимума в многомерном пространстве. В данной работе рассмотрены условия наиболее эффективной работы генетического алгоритма, реализованного для поиска минимума функции свободной энергии взаимодействия лиганд–белок в программе докинга SOL. Приведены описания различных параметров генетического алгоритма, а также различных режимов работы программы; рассмотрено, каким образом они влияют на поиск глобального минимума. Полученные значения оптимальных параметров генетического алгоритма являются рекомендуемыми и могут варьироваться в зависимости от задачи (докинг единичной молекулы или скрининг большой базы данных химических соединений) или имеющихся ресурсов.

Работа выполнена при финансовой поддержке ООО “Димонта”, г. Москва. Автор признателен В. Б. Сулимову, А. В. Сулимову, И. В. Офёркину и Д. К. Кутову за поддержку в работе и за предоставленные результаты тестирования по базе CSAR2009.

СПИСОК ЛИТЕРАТУРЫ

1. *Садовничий В.А., Сулимов В.Б.* Суперкомпьютерные технологии в медицине // Суперкомпьютерные технологии в науке, образовании и промышленности / Под ред. В.А. Садовничего, Г.И. Савина, Вл.В. Воеводина. М: Изд-во Моск. ун-та, 2009. 16–23.
2. *Klebe G.* Virtual ligand screening: strategies, perspectives and limitations // Drug Disc. Today. 2006. **11**. 580–594.
3. *Alvarez J., Shoichet B.* Virtual screening in drug discovery. Boca Raton: CRC Press, 2005.
4. *Kitchen D.B., Decornez H., Furr J.R., Bajorath J.* Docking and scoring in virtual screening for drug discovery: methods and applications // Nature Reviews Drug Discovery. 2004. N 3. 935–949.
5. *Pegga S.C.-H., Harescob J.J., Kuntza Ir.D.* A genetic algorithm for structure-based de novo design // J. Computer-Aided Molecular Design. 2001. **15**, N 10. 911–933.
6. *Oshiro C.M., Kuntz I.D., Dixon J.S.* Flexible ligand docking using a genetic algorithm // J. Computer-Aided Molecular Design. 1995. N 9. 113–130.
7. *Baxter C.A., Murray C.W., Clark D.E., Westhead D.R., Eldridge M.D.* Flexible docking using tabu search and an empirical estimate of binding affinity // Proteins: Structure, Function, and Bioinformatics. 1999. N 33. 367–382.
8. *Namasivayam V., Günther R.* Pso@autodock: a fast flexible molecular docking program based on swarm intelligence // Chemical Biology & Drug Design. 2007. **70**, N 6. 475–484.
9. *Goodsell D.S., Olson A.J.* Automated docking of substrates to proteins by simulated annealing // Proteins. 1990. **8**, N 3. 195–202.
10. *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 an empirical binding free energy function // J. Comput. Chem. 1998. **19**. 1639–1662.
11. *Deng Y., Roux B.* Computations of standard binding free energies with molecular dynamics simulations // J. Phys. Chem. B. 2009. **113**. 2234–2246.
12. *Frenkel D., Smit B.* Understanding molecular simulation: from algorithms to applications. New York: Academic Press, 2007.
13. *Zwanzig R.W.* High-temperature equation of state by a perturbation method. I. Nonpolar gases // J. Chem. Phys. 1954. **22**. 1420–1426.
14. *Morris G.M., Huey R., Lindstrom W., Sanner M.F., Belew R.K., Goodsell D.S., Olson A.J.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility // J. Comput. Chem. 2009. **30**. 2785–2791.

15. *Jones G., Willett P., Glen R.C., Leach A.R., Taylor R.* Development and validation of a genetic algorithm for flexible docking // J. Mol. Biol. 1997. **267**. 727–748.
16. *Verdonk M.L., Cole J.C., Hartshorn M.J., Murray C.W., Taylor R.D.* Improved protein-ligand docking using GOLD // Proteins. 2003. **52**. 609–623.
17. *Thomsen R., Christensen M.H.* MolDock: a new technique for high-accuracy molecular docking // J. Med. Chem. 2006. **49**. 3315–3321.
18. *Taylor J.S., Burnett R.M.* DARWIN: a program for docking flexible molecules // Proteins. 2000. **41**. 173–191.
19. *Clark K.P., Jain A.N.* Flexible ligand docking without parameter adjustment across four ligand receptor complexes // J. Comput. Chem. 1995. **16**. 1210–1226.
20. *Pei J.F., Wang Q., Liu Z.M., Li Q.L., Yang K., Lai L.* PSI-DOCK: towards highly efficient and accurate flexible ligand docking // Proteins-Structure Function and Bioinformatics. 2006. **62**, N 4. 934–946.
21. *Zhao Y., Sanner M.F.* FLIPDock: docking flexible ligands into flexible receptors // Proteins. 2007. **68**. 726–737.
22. *Stroganov O.V., Novikov F.N., Stroylov V.S., Kulkov V., Chilov G.G.* Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening // J. Chem. Inf. Model. 2008. **48**. 2371–2385.
23. *Романов А.Н., Кондакова О.А., Григорьев Ф.В., Сулимов А.В., Лущекина С.В., Мартынов Я.Б., Сулимов В.Б.* Компьютерный дизайн лекарственных средств: программа докинга SOL // Вычислительные методы и программирование. 2008. **9**. 213–233.
24. *Оферкин И.В., Сулимов А.В., Кондакова О.А., Сулимов В.Б.* Реализация поддержки параллельных вычислений в программах докинга SOLGRID и SOL // Вычислительные методы и программирование. 2011. **12**, № 1. 205–219.
25. *Sinauridze E.I., Romanov A.N., Gribkova I.V., Kondakova O.A., Surov S.S., et al.* New synthetic thrombin inhibitors: molecular design and experimental verification // PLoS ONE. 2011. **6**, N 5. e19969 (doi:10.1371/journal.pone.-0019969).
26. *Садовничий В.А., Сулимов В.Б., Каткова Е.В., Романов А.Н., Сулимов А.В., Оферкин И.В., Стамбольский Д.В., Белоглазова И.Б., Ткачук В.А.* Молекулярное моделирование для разработки новых лекарств на основе ингибиторов урокиназы // Постгеномные исследования и технологии / Под ред. С.Д. Варфоломеева. М.: Изд-во Моск. ун-та, 2011. 103–140.
27. *Choong P.F., Nadesapillai A.P.* Urokinase plasminogen activator system: a multifunctional role in tumor progression and metastasis // Clin. Orthop. Relat. Res. 2003. **415S**. S46–S58.
28. *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. 2004. **A 108**. 9323–9327.
29. The Protein Data Bank (http://www.rcsb.org/pdb/home/home.do).
30. *Hetenyi C., van der Spoel D.* Efficient docking of peptides to proteins without prior knowledge of the binding site // Protein Sci. 2002. **11**. 1729–1737.
31. *Taylor J.S., Burnett R.M.* DARWIN: a program for docking flexible molecules // Proteins. 2000. **41**. 173–191.
32. *Douglet D., Thoreau E., Grassy G.* A genetic algorithm for the automated generation of small organic molecules: drug design using an evolutionary algorithm // J. Computer-Aided Molecular Design. 2000. **14**. 449–466.
33. http://keenbase.ru/file/solparam_add_1.pdf
34. http://keenbase.ru/file/solparam_add_2.pdf
35. http://keenbase.ru/file/solparam_add_3.pdf
36. The Community Structure-Activity Resource (CSAR) (http://www.csardock.org/index.jsp).

Поступила в редакцию
11.09.2012