

УДК 541.69:541.572+615.781.6.
©Коллектив авторов

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ МОЛЕКУЛЯРНОГО 3D ДИЗАЙНА БИОЛОГИЧЕСКИ АКТИВНЫХ СОЕДИНЕНИЙ

В.А. Потемкин¹, М.А. Гришина¹, Г.В. Пожиленкова¹, К.М. Микушина², С. Лауфер²

¹Челябинский государственный университет 454021 Челябинск, ул. Бр. Кашириных 129,
ЧелГУ, Химический факультет; тел.: (3512) 419 01,
факс: (3512) 420 925; эл. почта: pva@csu.ru

²Университет Тюбингена

Предложен алгоритм DesPot для молекулярного 3D дизайна потенциальных биологически активных соединений. Алгоритм включает в себя процедуры генетической генерации новых молекулярных структур, оптимизации их геометрии с поиском глобального минимума и оценки биологической активности в рамках 3D-QSAR метода BiS. Для генерации молекулярных структур используется генетический алгоритм. Подход приводит к более быстрому достижению оптимума в ряде задач дискретной оптимизации. В качестве критерия оптимальности принята вероятность проявления соединением биологической активности. Для дизайна новых соединений используется выборка (популяция) соединений с известными показателями активности. Геометрия вновь полученных структур оптимизируется в силовом поле ММ3. Производится ориентация молекул по отношению к модельному рецептору, в рамках алгоритма BiS, рассчитываются характеристики взаимодействия и вероятность проявления биологической активности. Алгоритм протестирован на 40 видах биологической активности.

Ключевые слова: биологическая активность, молекулярный дизайн, 3D-QSAR.

ВВЕДЕНИЕ. Направленный поиск перспективных лекарственных средств является одной из важнейших проблем химии, медицины и биологии. Эта задача зачастую решается на интуитивном уровне и эффективность такого поиска, как правило, не высока. В то же время, финансовые и временные затраты на синтез и биологическое тестирование соединений весьма велики. Поэтому развиваются различные компьютерные технологии оценки активности гипотетических соединений и молекулярного дизайна соединений с заданной активностью. Это разнообразные варианты виртуального скрининга с использованием соединений из известных баз данных, а также предположительно активных структур с дальнейшей оценкой их биологической активности с использованием тех или иных методов QSAR (например, [1-3]). Такая оценка способна существенно снизить затраты на синтез и биологическое тестирование соединений при повышении вероятности выявления действительно активных соединений. Известен ряд алгоритмов молекулярного дизайна соединений с заданным свойством или уровнем биологической активности, использующих молекулярно-топологические представления и процедуры перечисления графов (см., например, [4,5]). Благодаря таким методам, наблюдается серьезный прогресс в направленном поиске новых лекарственных средств, создании комбинаторных библиотек, планировании направленного синтеза. Однако очевидно, что биологическая активность соединений определяется не только молекулярно-топологической формой, не только наличием тех или иных фармакофорных фрагментов, но и их сочетанием, взаимным расположением, то есть геометрическим строением, конформационным и таутомерным состоянием соединений, поэтому перспективным направлением является молекулярный дизайн соединений на уровне трехмерной структуры. В данной работе предложен новый алгоритм DesPot для молекулярного 3D дизайна потенциальных биологически активных соединений.

МЕТОДИКА. Принципиальная блок-схема алгоритма представлена на рисунке 1.

Алгоритм молекулярного 3D дизайна соединений с заданным уровнем биологической активности должен включать блоки генерации молекулярных структур, оптимизации их геометрии и прогноза их биологической активности. Предварительно, на выборке соединений с известной биологической активностью создается решающее правило в том или ином виде (регрессионное, полиномиальное описание, кластер-анализ, методы распознавания, нейронные сети и т.д.), устанавливающее связь характеристик молекулярной структуры с биологической активностью.

В данной работе для прогноза биологической активности соединений использован 3D QSAR алгоритм BiS [6-12], сочетающий возможности определения ориентации соединений относительно рецептора, достаточно надежного прогноза биологической активности при относительно малом времени расчета. В отличие от работ [6-12], в данной работе использована не количественная оценка активности, а оценка вероятности проявления биологической активности данного вида.

Для решения задачи генерации молекулярных структур имеется ряд подходов, описанных в [4,5,13-15]. Как правило, данные подходы включают комбинаторные процедуры сочетания фармакофорных фрагментов и (или) перечисления графов. В данной работе предлагается генетический алгоритм дизайна новых молекулярных структур. Данный подход приводит к более быстрому достижению оптимума во многих задачах дискретной оптимизации [16-19].

В качестве критерия оптимальности молекулярной структуры в данной задаче принята вероятность проявления соединением конкретного вида биологической активности. Для дизайна новых перспективных соединений в качестве исходной

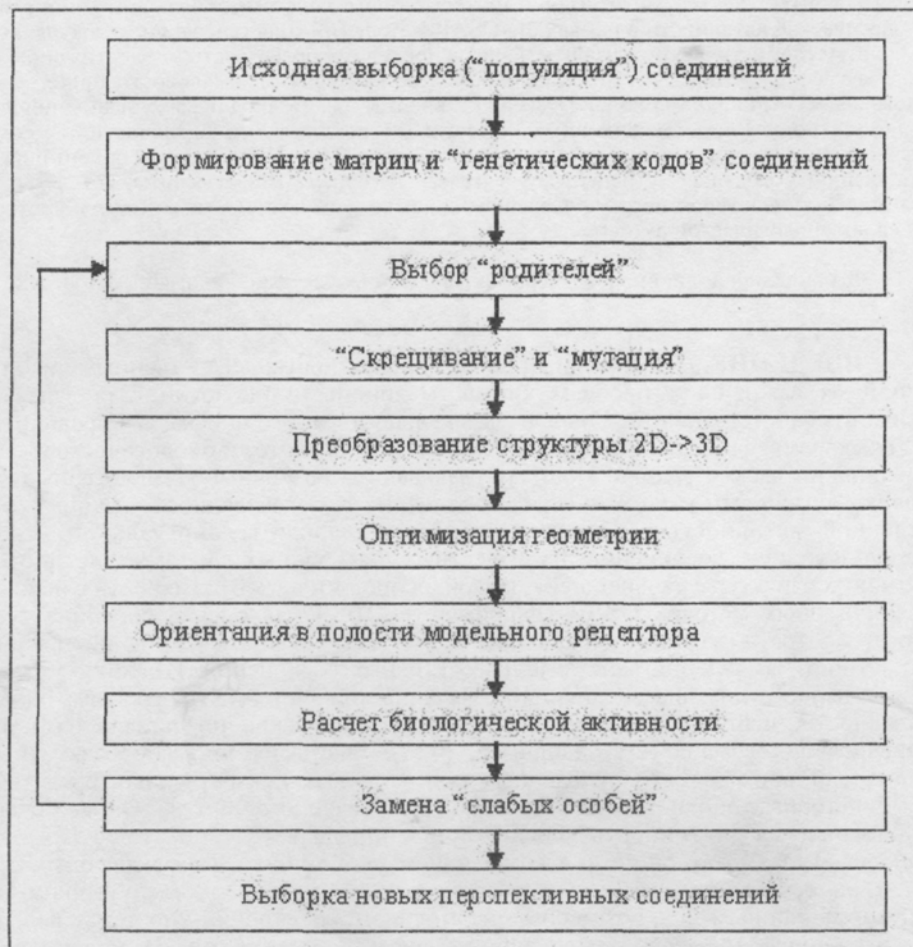


Рисунок 1.
Принципиальная блок-схема алгоритма DesPot

используется выборка известных биологически активных веществ. При этом для уверенного определения фармакофорных, нейтральных и антифармакофорных частей, важно, чтобы в выборке были представлены как соединения с высокой биологической активностью, так и малоактивные соединения. В терминологии генетических алгоритмов данная выборка называется популяцией. В рамках алгоритма BiS для соединений выборки строится псевдо-атомная модель рецептора, определяется расположение молекул в полости модельного рецептора, строится зависимость биологической активности от характеристик взаимодействия в комплексе "рецептор-лиганд". Для каждого из соединений исходной популяции определяется матрица, аналогичная топологической матрице смежности. Недиагональные элементы данной матрицы равны порядку связи для смежных вершин и нулю во всех остальных случаях; диагональным элементам присваивается значение атомного номера элемента в Периодической системе. Атомы водорода в матрице не указываются. Примеры таких матриц приведены на рисунке 2. В отличие от обычных топологических матриц смежности, нумерация атомов при составлении матрицы не является произвольной. Атомы нумеруются в порядке удаления от "донной" части псевдо-атомной модели рецептора, построенного в рамках алгоритма BiS. Таким образом, фармакофорные, нейтральные и антифармакофорные части молекул отражаются в примерно одинаковых блоках матриц. Очевидно, что матрицы соединений имеют различный размер в зависимости от числа атомов (как видно на рис. 2).

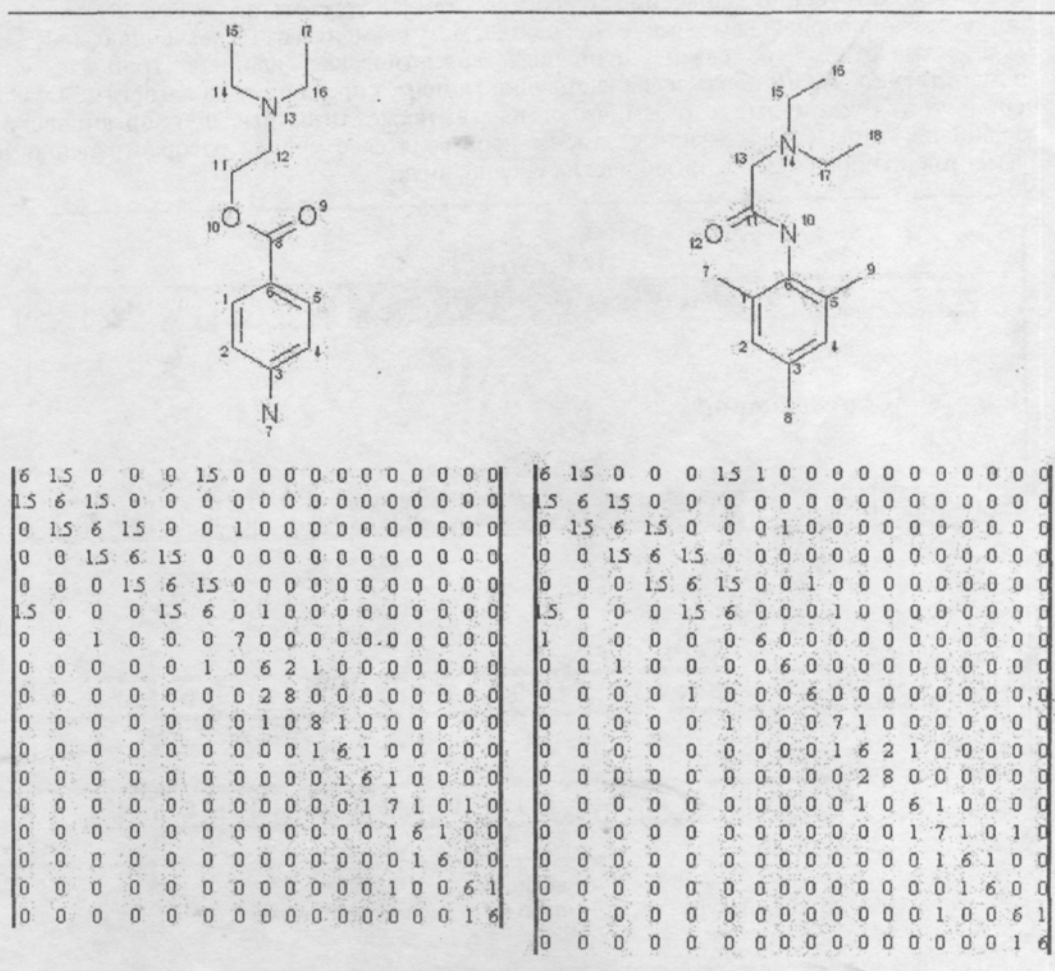


Рисунок 2.
Формирование матриц соединений

Матрицы всех соединений приводятся к единой размерности путем дополнения строками и столбцами с нулевыми элементами до $(1,5N_{\max}) \times (1,5N_{\max})$. Таким образом,

матрицы соединений имеют размерность в полтора раза большую, чем число атомов самой большой молекулы выборки. Данная операция производится для того, чтобы впоследствии у системы генерации новых молекул была возможность создавать молекулы большие, чем молекулы обучающей выборки. Верхний треугольник матрицы каждого соединения преобразуется к одномерному массиву, который однозначно определяет топологическую схему молекулы. Данный массив в терминологии генетических алгоритмов является генетическим кодом соединения. Таким образом, исходная выборка (популяция) представляется набором одномерных массивов (генетических кодов). Из популяции случайным образом с учетом вероятности проявления биологической активности выбирается пара молекул, в терминологии генетических алгоритмов называемых родителями. Случайным образом выбирается фрагмент генетического кода, который заменяется у одного из родителей на такой же фрагмент другого родителя, как показано на рисунке 3. Данная процедура носит название скрещивания. Результатом данной процедуры являются 2 новых генетических кода новых соединений, которые в терминологии генетических алгоритмов носят название детей. С точки зрения химической структуры, в результате скрещивания происходит обмен фрагментами молекул, как это показано на рисунке 4. Кроме того, для каждого из детей случайным образом производится мутация, то есть случайная замена элемента одномерного массива на другое значение. Элементы, отвечающие за номер атома в Периодической системе, могут заменяться на значения 6, 7, 8, 9, 15, 16, 17, 35, 53, 0, то есть на атомы C, N, O, F, P, S, Cl, Br, I, соответственно. Значение 0 соответствует отсутствию атома. Элементы, соответствующие порядкам химических связей, могут заменяться на значения 0, 1, 1.5, 2, 3, то есть "отсутствие связи", одинарная, ароматическая, двойная, тройная связи соответственно. Таким образом, на настоящий момент в предлагаемом алгоритме DesPot используются типичные атомы-органогены, а также типичные для органических соединений типы связей. Данные замены производятся с учетом распространенности элементов и типов связей в органических соединениях.

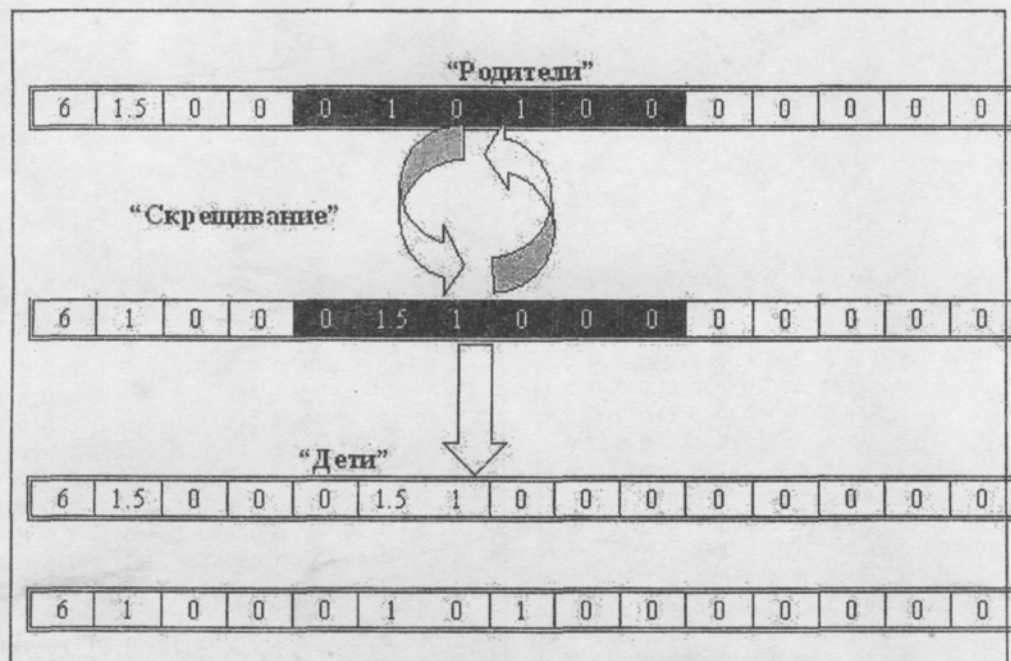


Рисунок 3.
Процедура скрещивания в генетическом алгоритме

Полученные после скрещивания и мутации одномерные массивы с генетическими кодами пары детей преобразуются к двумерным. При операциях скрещивания и мутации возможно нарушение правил валентности для химических соединений. Поэтому производится проверка соответствия числа валентных связей в строке или столбце двумерного массива типу атома, указанного в качестве диагонального элемента. В

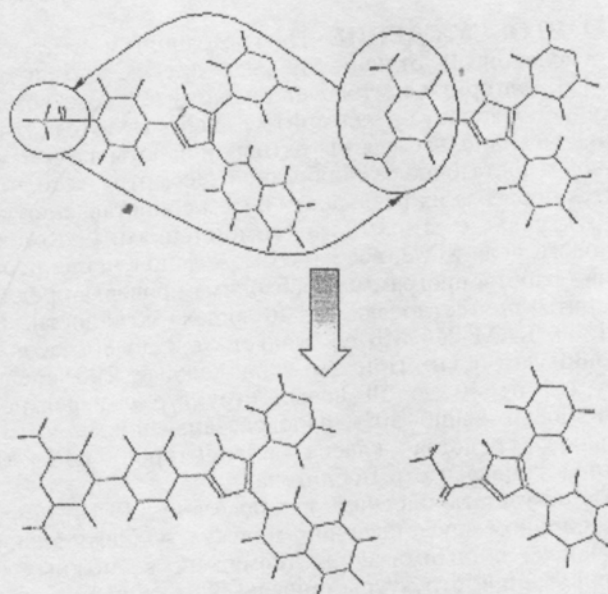


Рисунок 4.

Обмен фрагментов молекул в результате скрещивания

настоящей версии алгоритма DesPot рассматриваются только молекулярные формы соединений, поэтому число валентных связей (сумма недиагональных элементов строки или столбца) для атома углерода не должна превышать 4, азота - 3, кислорода - 2, галогенов - 1, фосфора - 5, серы - 6. Для атома азота допускается значение 4, если он содержит две ароматические связи (нитрогруппа). В случае, если сумма элементов строки превышает предельную валентность элемента, производится уменьшение порядка случайно выбранной связи. В случае, если сумма порядков связей атома принимает нецелочисленное значение, то производится изменение порядка случайно выбранной одинарной, двойной или полуторной связи этого атома на 0,5 (увеличение порядка одинарной связи, уменьшение порядка двойной связи, увеличение или уменьшение порядка ароматической связи). В полученной матрице отыскивается максимальный связный граф и строится его 2D модель. Для избежания повторений структур производится проверка изоморфизма данного графа с уже имеющимися структурами. Затем к каждому из атомов добавляются атомы водорода в соответствии с недостающей валентностью. Очевидно, что в результате операций скрещивания и мутации могут получаться нереальные структуры. При этом достаточно сложно формализовать многие эмпирические правила структурных ограничений, существующие в органической химии (например, редко встречаются гем-диольные формы, тройная связь в ароматическом цикле и т.п.). Для снижения числа генерируемых нереальных структур производится сопоставление комбинаций атомов в полученных структурах с типами и комбинациями атомов, имеющихся в параметризации силового поля MM3.

После добавления атомов водорода производится построение 3D структуры молекулы, в соответствии с методом [20]. Производится оптимизация геометрических характеристик молекулы с использованием молекулярно-механического силового поля MM3 [21,22] с применением алгоритмов, описанных в работах [23,24]. Далее производится ориентация молекулы в полости модельного рецептора в рамках алгоритма BiS [6-12], рассчитываются энергетические и силовые характеристики взаимодействия, а также вероятность проявления биологической активности. Если данная вероятность выше, чем хотя бы у одного из соединений популяции, то в популяции производится замена одного из соединений с меньшей вероятностью (слабой особи) на вновь сконструированную молекулу. Данный цикл процедур от выбора родителей до замены в популяции в терминологии генетических алгоритмов носит название поколения. Очевидно, что в обновленной популяции вновь можно выбрать пару родителей, произвести процедуры скрещивания, мутации, преобразования к 3D структуре, прогноза активности и замены слабых особей, то есть получить следующее поколение. Обычно в

генетических алгоритмах используется от нескольких сотен до нескольких сотен тысяч поколений.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ. На настоящий момент в алгоритме DesPot реализуется 1000 поколений. В отличие от ряда других алгоритмов молекулярного дизайна (например, [5]), алгоритм не требует предварительного создания базы данных фрагментов молекул, поскольку в алгоритме используются готовые фрагменты соединений выборки. Очевидно, что для эффективной работы алгоритма выборка должна быть представительной и достаточно разнообразной. Несмотря на то, что в ходе генерации структур производится проверка их реальности путем сопоставления комбинаций атомов в полученных структурах с типами и комбинациями атомов, имеющихся в параметризации силового поля ММЗ, все же невозможно получение нереальных структур. Поэтому по окончании работы программы необходима проверка предложенных структур специалистом. Алгоритм протестирован на 40 видах активности. Время расчета на процессоре Athlon-1200, RAM 256 МБ составляет от 8 до 80 часов в зависимости от специфики соединений выборки. При этом на каждые 100 соединений исходной популяции генерируется от 10 до 50 новых структур с вероятностью проявления биологической активности выше 50%. Предсказаны новые антиревматические и противовоспалительные средства класса ингибиторов р38 MAP-киназы для последующего синтеза в университете Тюбингена.

Более надежные результаты прогноза, как показано в [9,10], могут быть получены при мультиконформационном представлении молекул. Однако включение процедуры конформационного поиска с оптимизацией геометрии возможных конформеров для каждой вновь сгенерированной структуры привело бы к существенному увеличению (в десятки раз) времени расчета. Тем не менее, процедуры конформационного поиска и оптимизации геометрии каждого конформера являются независимыми и удобными для распараллеливания, поэтому начата работа по созданию мультиконформационной версии алгоритма DesPot в суперкомпьютерном варианте для Т-системы на кластере СКИФ Союзного Государства России и Белоруссии.

ЗАКЛЮЧЕНИЕ. Таким образом, предложен новый алгоритм DesPot для молекулярного 3D дизайна потенциальных биологически активных соединений. Алгоритм включает в себя процедуры генетической генерации новых молекулярных структур, оптимизации их геометрии с поиском глобального минимума и оценки биологической активности в рамках 3D-QSAR метода BiS.

Работа выполнена при финансовой поддержке суперкомпьютерной программы СКИФ Союзного Государства России и Белоруссии.

ЛИТЕРАТУРА

1. Walters W.P., Stahl M.T., Murcko M.A. (1998) Drug Discovery Today, 1998, 3, 160 - 178.
2. Filimonov D.A., Poroikov V.V. (1996) in: Bioactive Compound Design: Possibilities for Industrial Use. BIOS Scientific Publishers, Oxford, pp. 47 - 56.
3. Bohm H.J. (2001) in: Rational Approaches to Drug Design. Proceedings of the 13th European Symp. on Quantitative Structure-Activity Relationships, QSAR 2000 (H.-D. Holtje and W.Sippl eds.) Prous Science Publishers, Barcelona, pp. 367 - 371.
4. Kubinyi H. (1995) Pharmazie, 50, 647 - 662.
5. Van Drie J.H., Lajiness M.S. (1998) Drug Discovery Today, 3, 274 - 283.
6. Potemkin V. A., Bartashevich E. V., Grishina M. A., Guccione S. (2001) in: Rational Approaches to Drug Design. Proceedings of the 13th European Symp. on Quantitative Structure-Activity Relationships, QSAR 2000 (H.-D. Holtje and W.Sippl eds.) Prous Science Publishers, Barcelona, pp. 349 - 353.
7. Потемкин В.А., Гришина М.А., Белик А.В., Чупахин О.Н. (2002) Хим.-фарм. журн., 36(1), 22 - 25.
8. Grishina M. A., Potemkin V. A., Rusinov G.L., Bartashevich E. V., Guccione S., Perspicace S., Chupakhin O.N. (2002) Comparing a BiS (Biological Substrate Search) alignment of anticancer DNA and DNA/RNA antimetabolites into their active sites with .pdb co-crystals, "From Genes to Drugs via Crystallography", 33rd crystallographic course at the E. Majorana Centre, Erice, Italy, 23 May - 2 June 2002. Book of Abstracts. Poster Abstracts, 48.

9. Гришина М.А., Потемкин В.А., Арсламбеков Р.М., Белик А.В. (2002) Первая Национальная Конференция "Информационно-вычислительные технологии в решении фундаментальных научных проблем и прикладных задач химии, биологии, фармацевтики, медицины". Тез. Докл. - Научный и учебный методический центр, М., с. 22 - 23.
10. Потемкин В.А., Арсламбеков Р.М., Барташевич Е.В., Гришина М.А., Белик А.В., Перспека С., Гуччионе С. (2002) Журн. структ. химии, **43**, 1134 - 1138.
11. Mikuchina K., Potemkin V., Grishina M., Laufer S. (2002) Arch. Pharm. Pharm. Med. Chem., **335**(1), C74.
12. Потемкин В.А., Гришина М.А., Федорова О.В., Русинов Г.Л., Овчинникова И.Г., Ишметова Р.И. (2003) Хим.-фарм. журн, **37**(9), 17 - 21.
13. Golender V.E., Rosenblit A.B. (1983) Logical and Combinatorial Algorithms for Drug Design, Wiley&Sons, L.
14. Murcko M.A., Caron P.R., Charifson P.S. (1999) Ann. Rep. Med. Chem., **34**, 297-306.
15. Bohm H.J., Stahl M. (2000) Curr. Op. Chem. Biol., **4**, 283 - 286.
16. Hasegawa K., Fumatsu K. (1998) J. Mol. Struct. (Theochem), **425**, 255 - 262.
17. Zacharias C.R., Lemes M.R., Dal Pino Jr A. (1998) J. Mol. Struct. (Theochem), **430**, 29-39.
18. Keser M., Stupp S.I. (1998) Computers Chem., **22**, 345 - 351.
19. Гришина М.А., Барташевич Е.В., Потемкин В.А., Белик А.В. (2002) Журн. структ. химии, **43**, 1128 - 1133.
20. Gasteiger J., Rudolph C., Sadowski J. (1990) Tetrahedron Comp. Method, **3**, 537-547.
21. Allinger N.L., Yuh Y.H., Lii J.-H. (1989) J. Am. Chem. Soc., **111**, 8551-8566.
22. Lii J.-H., Allinger N.L. (1989) J. Am. Chem. Soc., **111**, 8566-8575.
23. Nakamura S., Hirose H., Ikeguchi M., Doi J. (1995) J. Phys. Chem., **99**, 8374-8378.
24. Ponder J.W., Richards F.M. (1987) J. Comp. Chem., **8**, 1016-1024.

GENETIC ALGORITHM FOR MOLECULAR 3D DRUG DESIGN

V.A. Potemkin¹, M.A. Grishina¹, G.V. Pozhilenkova¹, K.M. Mikuchina², S. Laufer²

¹Chelyabinsk State University

²Universitat Tübingen

ul. Br. Kashirinykh 129, Chelyabinsk, 454021, Russia; tel.: +7 (3512) 419 012;

fax: +7 (3512) 420 925; e-mail: pva@csu.ru

A new algorithm DesPot for 3D design of new prospective biological by active compounds has been proposed. The algorithm includes the following procedures: a genetic construction of new molecular structures, their geometry optimization using the global minimum search along the Hessian modes and evaluation of their biological activity by 3D-QSAR algorithm BiS. For the design of new structures it has been suggested to use the genetic algorithm. This algorithm takes less time in order to achieve the optimum in a number of discrete optimization problems. The probability of biological activity has been used as criteria of optimum. The design of new molecules is carried out using the set of the known drugs. The geometry of new designed molecules is optimized in MM3 force field. Then the orientation of each new molecule in the model receptor is defined using BiS algorithm. The characteristics of "model receptor - ligand" interactions and the probability of biological activity are computed. Algorithm DesPot has been used for design of new compounds, tested by 40 types of biological activity.

Key words: biological activity, molecular design, 3D-QSAR