Эволюционные алгоритмы оптимизации в задаче предсказания пространственной структуры пептидов

С. В. Полуян[1,а], Н. М. Ершов[2,б]

[1] ГБОУ ВО МО «Университет «Дубна», институт системного анализа и управления,
141980, Московская область, г. Дубна, ул. Университетская, 19

[2] Московский государственный университет им. М. В. Ломоносова,
факультет вычислительной математики и кибернетики,
119991, г. Москва, ГСП-1, Ленинские горы, 1-52

E-mail: [а] svpoluyan@gmail.com, [б] ershovnm@gmail.com

Настоящая работа посвящена исследованию области применимости стохастических эволюционных алгоритмов оптимизации в задаче предсказания вторичной структуры пептидов. Рассматривается одна из задач структурной биоинформатики – предсказание трехмерной структуры пептида по аминокислотной последовательности. Ставиться задача поиска двух основных регулярных вторичных структур у различных пептидов разной длины в непрерывном пространстве торсионных углов главной и боковых цепей. При этом рассматриваются следующие пептиды с известной нативной структурой: модельные, сконструированные синтетические, природного происхождения. Продемонстрированы основные предположения сводящую задачу предсказания пространственной структуры пептида к задаче непрерывной глобальной оптимизации. Важно отметить, что в рассматриваемой постановке не присутствует ограничение пространства поиска с использованием статистической информации о приемлемых значениях углов в боковых цепях и библиотек фрагментов полученных из известных на сегодняшний день структур. Проводится анализ существующих на сегодняшний день подходов решения вышеописанной задачи. Рассматриваются основные особенности силового поля, используемого для вычисления энергии пептида. Разработан подход к изменению одного из параметров силового поля, который характеризует нековалентные взаимодействия, в процессе поиска оптимальной структуры. Отличительная особенность предложенного подхода заключается в том, что в предлагаемом методе выделения параметра общая постановка задачи остается однокритериальной. Представлены результаты разнообразных численных экспериментов с использованием различных эволюционных алгоритмов оптимизации, а также итоги сравнения основных эволюционных операторов. Указаны наиболее эффективные операторы. Описана схема распараллеливания рассматриваемых алгоритмов и приводятся результаты использования параллельных вычислений. Показаны некоторые недостатки использования методов оптимизации при решении поставленной задачи. Выполнено сравнение найденных структур со структурами полученными с использованием актуальных методов решения поставленной задачи. Практическая значимость проведенного исследования заключается в выявлении границ области применения эволюционных методов оптимизации, которые позволяют оценить перспективу использования эволюционных алгоритмов в актуальных задачах структурной биоинформатики (например, в задаче поиска оптимального положения пептида на белке).

Ключевые слова: вторичная структура, конформационный поиск, эволюционные вычисления, глобальная оптимизация.

1. Введение

Белки являются макромолекулами состоящими из α-аминокислот, соединенных в цепочку пептидной связью, тем самым образуя полипептидную цепь. Предсказание структуры белка – предсказание по аминокислотной последовательности трехмерной структуры белка, которая определяет нативное, т.е. функционально активное, состояние (выделяют вторичную, третичную и четвертичную). Короткие белки называют пептидами. В настоящей работе рассматривается задача поиска двух основных регулярных вторичных структур встречающихся у пептидов: α-спирали и β-листа.

Наиболее широко принимаемая гипотеза, объясняющая процесс самоорганизации белковых молекул была сформулирована Анфинсеном [Anfinsen, 1973]. Основные идеи предложенной им «термодинамической гипотезы» следующие: нативное состояние белка уникально; нативное состояние белка находится в глобальном минимуме свободной энергии. Таким образом, процесс сворачивания полипептидной цепи можно представить как процесс минимизации свободной энергии белка, тогда задача предсказывания структуры сводится к задаче глобальной оптимизации.

Если поставить задачу классификации методов предсказывания структуры белка, то можно выделить два основных подхода. Первый состоит в использовании информации известных белковых структур. Такие методы предсказывания называют моделированием по гомологии. Второй подход называют ab initio, то есть процесс сворачивания цепи рассматривается без привлечения каких-либо дополнительных эмпирических предположений, только естественные законы природы.

При численном исследовании алгоритмов не будет использоваться никаких статистических известных низкоэнергетических «шаблонных» структур и какой-либо другой вспомогательной информации, поскольку авторы ставят перед собой целью выяснение потенциала алгоритмов в рамках рассматриваемых целевых функциях. В дальнейшем планируется рассматривать задачу взаимодействия вида пептид-белок. При таких взаимодействиях в структуре белка и пептида происходят сильные взаимосвязанные неспецифичные конформационные изменения, как правило, слабо поддающиеся статистическому анализу.

2. Силовое поле

В численных экспериментах использовалось силовое поле ROSETTA [O'Meara, Leaver-Fay, 2015]. Отличительной особенностью данного силового поля является использование, при вычислении энергии пептида, неявного растворителя, различных потенциалов и статистически полученных дынных.

Целевая функция (именуемая также скоринг-функцией) представляет собой сумму так называемых термов, которые входят в состав суммы с определенным весом. Веса термов калибруются на определенной выборке белков. Термы описывают межатомные взаимодействия с использованием классической механики (силы отталкивания и притяжения Леннарда-Джонса, электростатические взаимодействия), так и эмпирически известные данные (планарность торсионного угла ω главной цепи и водорода в гидроксильной группе). Водородные связи разбиты на четыре группы: взаимодействия между атомами основной цепи в зависимости от положения в первичной структуре (близкие и дальние); взаимодействия между атомами главной цепи и боковыми цепями; взаимодействия между боковыми цепями. В рассматриваемых скоринг-функциях использовалось приблизительно 15 термов. В связи с тем, что при вычислении целевой функции используются эмпирические термы и все веса термов откалиброваны невозможно говорить о получаемой энергии пептида как о потенциальной энергии выражаемой в килокалориях на моль. Вместо этого рассматривается просто получаемое значение скоринг-функции.

В качестве целевых функций использованы две скоринг-функции – *score12* и *talaris2014*, соответствующие предыдущему и текущему стандарту скоринг-функции у силового поля ROSETTA. Принципиальное различие *score12* и *talaris2014* заключается в способе вычисления электростатических взаимодействий. В первом случае используется терм описывающий статистически полученные данные из PDB [O'Meara, Leaver-Fay, …, 2015], во втором случае в явном виде вычисляется кулоновский потенциал.

Выбор рассматриваемого силового поля обусловлен широкой распространенностью, быстродействием и ориентированностью к проблеме предсказания пространственной структуры белков.

3. Результаты численных экспериментов

Для поиска оптимальной структуры использовались следующие эволюционные алгоритмы: адаптивная дифференциальная эволюция JADE [Zhang, Sanderson, 2009], эволюционная стратегия ESCH [Silva-Santos, Goncalves, Hernandez-Figueroa, 2010], метод роя частиц PSO [Kennedy, Eberhart, 1995] с локальным поиском SW [Solis, Wets, 1981], алгоритм бактериального поиска с адаптивным изменение шага SABFO [Полуян, Рейнгард, Ершов, 2014], алгоритм роевой оптимизации со стратегией соревнования особей CSO [Cheng, Jin, 2015], неоднородный клеточный генетический алгоритм NCGA [Ершов, 2015], эволюционная стратегии с адаптацией матрицы ковариаций CMAES [Hansen, Ostermeier, 1996], гибрид дифференциальной эволюции с CMAES для локальной оптимизации JDE-CMAES [Brest, Zamuda, …, 2010]. Выбор рассматриваемых алгоритмов обусловлен хорошими результатами при решении различных практических задач оптимизации [Cheng, Jin, 2015], а также разнообразием стратегий у различных операторов.

На первом этапе вычислительных экспериментов ставилась задача нахождения оптимальной структуры модельного пептида длинной 10 аминокислотных остатков A10 [Sung, 1994], с искомой структурой – α-спираль. Задача поиска структуры ставилась в непрерывном пространстве: торсионных углов главной цепи пептида (углы φ и ψ, пространство поиска $[-\pi, \pi]$); торсионных углов главной цепи ω (стремится быть планарным, поэтому $[\pi-\delta, \pi+\delta]$, где $\delta = 0.2$ рад.). Размерность задачи составила 27 параметров. Количество вызовов целевой функции ограничено одним миллионом. Для каждого алгоритма выполнено 25 независимых запусков.

Важно отметить, что сходимость у эволюционных алгоритмов в значительной степени зависит от используемых параметров, причем их число варьируется от двух (CSO) до 12 (SABFO). В проводимых экспериментах часть параметров подбирались с учетом размерности, границ рассматриваемой задачи и рекомендаций авторов. Так как некоторые из рассматриваемых алгоритмов адаптивно меняют в процессе поиска часть параметров (например, с целью увеличения скорости сходимости) выбрано довольно большое число вызовов целевой функции. Для максимальной объективности сравнения размер популяции для всех алгоритмов составлял 400 особей. Чувствительность алгоритмов к размеру популяции в данном случае нивелируется большим числом итераций.

Оптимальная структура для рассматриваемого пептида получена с помощью сервера PEP-FOLD [Shen, Maupetit, …, 2014], предсказывающим структуру пептида с использованием статистической информации (низкоэнергетических фрагментов) для получения пула крупнозернистых моделей и последующей полноатомной минимизацей.

На рис. 1 показано среднеквадратичное отклонение координат атомов α-углерода главной цепи получаемых после оптимизации пептидов относительно найденной с помощью PEP-FOLD структуры. Так как вторичную структуру определяет конформационное расположение атомов главной цепи такой способ сравнения наиболее объективен. Силовое поле используемое в методе PEP-FOLD отличается от ROSETTA, поэтому перед сравнением здесь и далее получаемая PEP-FOLD структура проходит процедуру релаксации в ROSETTA стандартными средствами пакета. Следует отметить, что первичная цепочка для оптимизации алгоритмами порождалась

средствами ROSETTA с идеализированными значениями валентных углов и длин ковалентных связей. В рассматриваемом случае при оптимизации эти значения не изменялись и конечные структуры несколько отличаются от получаемой с помощью PEP-FOLD.

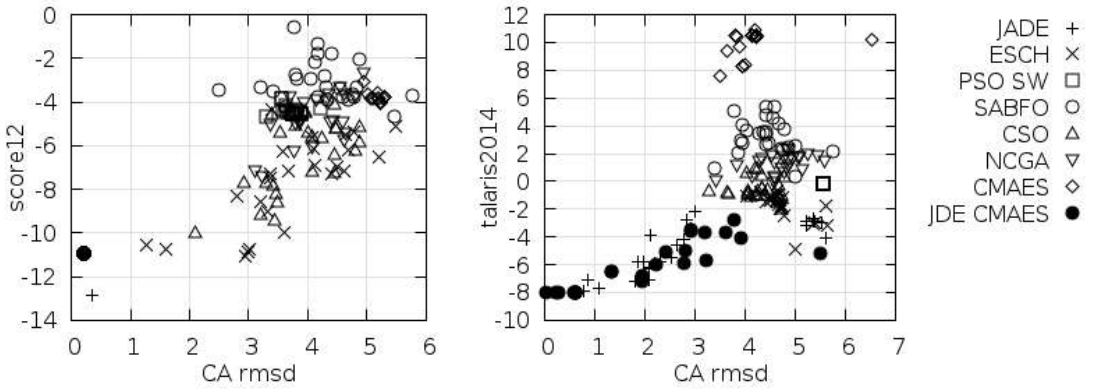

Рис. 1. Результаты 25 независимых запусков для рассматриваемых алгоритмов

Можно заметить, что меньшее значение целевой функции соответствуют большему отклонению атомов главной цепи. Здесь необходимо учесть, что проводимая пакетом релаксация носит локальный характер.

На рис. 1 видно, что лучшие результаты демонстрируют алгоритм адаптивной дифференциальной эволюции JADE и гибрид JDE-CMAES. Оба алгоритма отыскали оптимальную структуру для каждого из 25 запусков, в то время как ни один другой не показал результата менее одного ангстрема.

Рассматриваемые алгоритмы имеют схожую структуру, включающую в себя три основных оператора – отбор, скрещивание, мутация. Причем значительного различия в операторах отбора не наблюдается. Операторы скрещивания у JADE, JDE и NCGA одинаковы, разница только в вероятности выполнения оператора. Этому оператору соответствует шаг репродукции у алгоритма SABFO, строящийся по более локальному принципу, что подтверждают результаты. Однако операторы мутации во всех приведенных алгоритмах разные: в случае с JADE используется стратегия *current-to-best*; в JDE – классическая для дифференциальной эволюции стратегия *rand*; в NCGA – классическая стратегия для генетического алгоритма. Отдельно следует отметить алгоритм CMAES, который показывает в начале оптимизации самую высокую скорость сходимости среди всех алгоритмов, однако дает один из худших результатов, показывая, тем самым, свою только локальную эффективность.

На основании представленных результатов и перечисленных выше аргументов можно сделать вывод, что при решении поставленной задачи принципиальными оператором является оператор мутации, причем со стратегией *current-to-best*.

На втором этапе вычислительных экспериментов ставилась задача нахождения оптимальной структуры модельного пептида V4GGV4 [Sung, 1999] (с искомой структурой β-лист) и вышеописанной спирали в полноатомном разрешении. Задача поиска оптимальной структуры ставится аналогично для торсионных углов главной цепи, с добавлением основных торсионных углов для каждой боковой цепи $\chi1$–$\chi4$ (пространство поиска $[-\pi, \pi]$), длин ковалентных связей ($\delta1 = 0.05$ Å), валентных углов ($\delta2 = 0.1$ рад.) для каждого атома пептида, неосновных торсионных углов боковой цепи каждого атома ($\delta3 = 0.1$ рад.). Границы с $\delta1$–$\delta3$ рассчитывались относительно идеализированных значений используемых в ROSETTA (аналогично методу CONCOORD [de Groot, van Aalten, …, 1997]), пространство поиска непрерывно. Таким образом, размерность задач для α-спирали и β-листа составила 302 и 428 параметров соответственно.

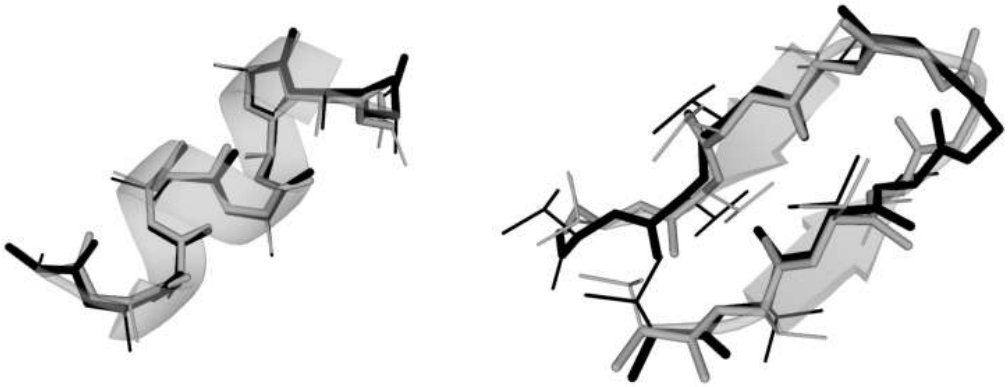

Рис. 2. Суперпозиция главных цепей пептидов, PEP-FOLD (черный) и JADE (серый)

На рис. 2 показаны результаты оптимизации с помощью алгоритма JADE при 10 миллионах вызовах целевой функции. Полученное отклонение атомов главной цепи составило меньше половины ангстрема относительно структуры найденной с помощью PEP-FOLD. Суперпозиции на рис. 2 получены с использованием 3DSS [Sumathi, Ananthalakshmi, …, 2006].

4. Имитация отжига для кулоновского потенциала

Численные эксперименты и результаты, представленные в разделе выше, показывают, что наибольшее усложнение целевой функции порождает кулоновский (электростатический) потенциал. Алгоритмы дифференциальной эволюции JADE и JDE-CMAES способны достичь минимума лишь в нескольких случаях.

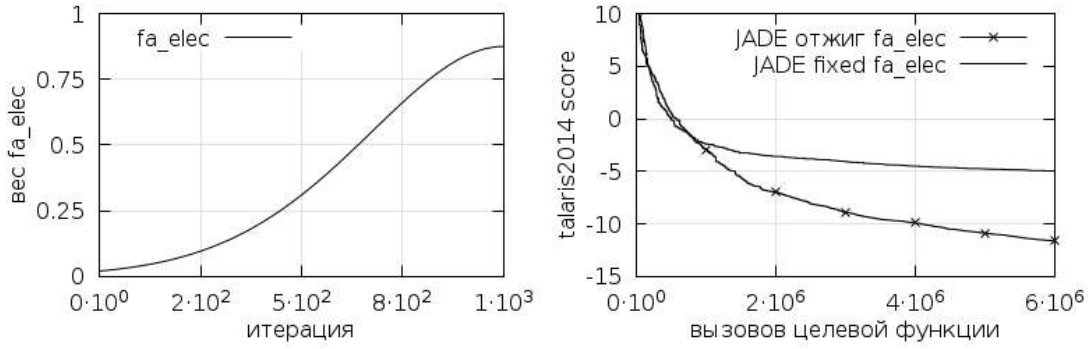

Рис. 3. Функция изменения кулоновского потенциала и сходимость для α-спирали

Поскольку данный потенциал (*fa_elec*) определяет нековалентные взаимодействия, было предложено использовать метод имитации отжига [Kirkpatrick, Gelatt, Vecchi, 1983] для изменения соответствующего веса потенциала в процессе оптимизации. Такой подход был использован для пептидов YMEARAMEARA (α-спираль) и Ace-ITVNGKTY-Nme (β-лист) [Galzitskaya, Higo, Finkelstein, 2002], размерность задач составила 501 и 395 параметров соответственно. На рис. 3 представлены результаты с использованием отжига и без для α-спирали. На рис. 4 представлены получаемые структуры.

Поскольку у эволюционных алгоритмов операция вычисления целевой функции для каждого члена популяции обычно выносится из основных операторов, большинство из них довольно просто и масштабируемо распараллеливаются. В результате выполненной работы произведена параллельная реализация данного этапа у алгоритма JADE с использованием технологии параллельных вычислений OpenMP.

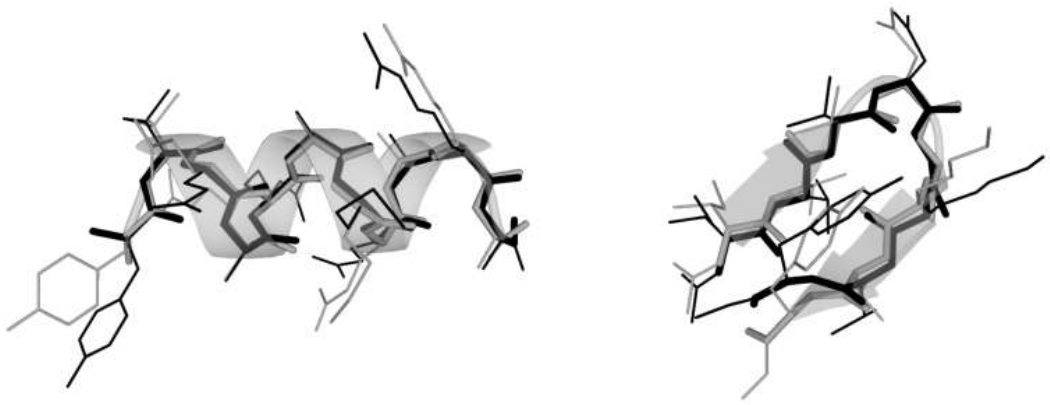

Рис. 4. Суперпозиция главных цепей пептидов, PEP-FOLD (черный) и JADE (серый)

В настоящих экспериментах использовался один вычислительный узел, содержащий два 12-ядерных процессора Intel Xeon. В среднем, для исследуемого пептида длинной 11 аминокислотных остатков с использованием скоринг-функции *talaris2014* запуск алгоритма в один поток с ограничением вызовов целевой функции в миллион составлял около 668 секунд, а при использовании 24 потоков порядка 63 секунд. Тем самым получено ускорение приблизительно в десять раз. Важно отметить, что некоторые операторы алгоритма также поддаются параллелизации, в связи с этим можно добиться более заметных результатов. Все вычисления выполнены на кластере ОИЯИ HybriLIT [HybriLIT, 2016].

5. Заключение

В результате выполненной работы выявлены границы области применения эволюционных алгоритмов оптимизации к задаче предсказания структуры пептидов. Показаны какие операторы существенны и какие стратегии показывают наилучший результат. Предложена схема к изменению веса кулоновского потенциала у силового поля в процессе поиска оптимальной структуры и показана ее эффективность. Проведено численное исследование алгоритмов с использованием предложенной схемы на различных пептидах. Для некоторых алгоритмов произведена параллельная реализация. Результаты проведенных в данной работе исследований демонстрируют, что стратегия мутации в алгоритме дифференциальной эволюции в значительной степени определяет эффективность сходимости.

На основании проведенных исследований можно заключить, что, с использованием предложенной схемы, эволюционные алгоритмы оптимизации способны находить оптимальную структуру коротких пептидов длинной порядка десяти аминокислотных остатков в полноатомном разрешении. Целью дальнейшей работы является расширение схемы разбиения весов силового поля (в том числе разбиение задачи на несколько критериев), рассмотрение пептидов большей длины, а также применение эволюционных алгоритмов в задаче поиска оптимального положения пептида на белке.

Список литературы

Anfinsen C. Principles that Govern the Folding of Protein Chains // Science. — Jul. 1973. — Vol. 181, Issue 4096. — P. 330–331.

O'Meara M.J., Leaver-Fay A., Tyka M.D., Stein A., Houlihan K., DiMaio F., Bradley P., Kortemme T., Baker D., Snoeyink J., Kuhlman B. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta // Journal of Chemical Theory and Computation. — 2015. — Vol. 11, Issue 2. — P. 609–622.

Zhang J., Sanderson A. JADE: Adaptive differential evolution with optional external archive // IEEE Transactions on Evolutionary Computation. — 2009. — Vol. 13, Issue 5. — P. 945-958.

Silva-Santos C.H., Goncalves M.S., Hernandez-Figueroa H.E. Designing Novel Photonic Devices by Bio-Inspired Computing // IEEE Photonics Technology Letters. — 2010. — Vol. 22, Issue 10. — P. 1177–1179.

Kennedy J., Eberhart R. Particle swarm optimization // Proceedings of IEEE International Conference on Neural Networks. — 1995. — Vol. 4. — P. 1942-1948.

Solis F.J., Wets R.J-B. Minimization by random search techniques // Mathematics of Operation Research. — 1981. — Vol. 6, Issue 1. — P. 19–30.

Полуян С.В., Рейнгард Н.М., Ершов Н.М. Самоадаптация в алгоритмах роевой оптимизации // Вестник Российского университета дружбы народов: Серия Математика, информатика, физика. — 2014. — № 2. — С. 415–418.

Poluyan S.V., Reinhard N.M., Ershov N.M. Samoadaptaciya v algoritmah roevoi optimizacii [Self-Adaptation in swarm optimization algorithms] // Vestnik RUDN ser. Mathematics, Informatics, Physics. — 2014. — No. 2. — P. 159–173 (in Russian).

Cheng R., Jin Y. A Competitive Swarm Optimizer for Large Scale Optimization // IEEE Transactions on Cybernetics. — 2015. — Vol. 45, Issue 2. — P. 191–204.

Ершов Н.М. Неоднородные клеточные генетические алгоритмы // Компьютерные исследования и моделирование. — 2015. — Т. 7, № 3. — С. 775–780.

Ershov N.M. Neodnorodnye kletochnye geneticheskie algoritmy [Non-uniform cellular genetic algorithms] // Computer Research and Modeling. — 2015. — Vol. 7, No. 3. — P. 775–780 (in Russian).

Hansen N., Ostermeier A. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation // Proceedings of the 1996 IEEE International Conference on Evolutionary Computation. — 1996. — P. 312–317.

Brest J., Zamuda A., Fister I., Maucec M.S. Large scale global optimization using self-adaptive differential evolution algorithm // IEEE World Congress on Computational Intelligence. — 2010. — P. 1–8.

Sung S.S. Helix Folding Simulations with Various Initial Conformations // Biophysical Journal. — Jan. 1994. — Vol. 66, Issue 6. — P. 1796–1803.

Shen Y., Maupetit J., Derreumaux P., Tuffery P. Improved PEP-FOLD approach for peptide and miniprotein structure prediction // Journal of Chemical Theory and Computation. — 2014. — Vol. 10. — P. 4745–4758.

Sung S.S. Monte Carlo Simulations of β-Hairpin Folding at Constant Temperature // Biophysical Journal. — Jan. 1999. — Vol. 76. — P. 164–175.

de Groot B.L., van Aalten D.M., Scheek R.M., Amadei A., Vriend G., Berendsen H.J. Prediction of protein conformational freedom from distance constraints // Proteins. — Oct. 1997. — Vol. 29, Issue 2. — P. 240–251.

Sumathi K., Ananthalakshmi P., Roshan M.N., Sekar K. 3dSS: 3D structural superposition // Nucleic Acids Research. — Jul. 2006. — Vol. 34, Issue suppl. 2. — P. 128–132.

Kirkpatrick S., Gelatt C.D., Vecchi M.P. Optimization by Simulated Annealing // Science. — 1983. — Vol. 220, Issue 4598. — P. 671–680.

Galzitskaya O.V., Higo J., Finkelstein A.V. α-Helix and β-Hairpin Folding from Experiment, Analytical Theory and Molecular Dynamics Simulations // Current Protein and Peptide Science. — Apr. 2002. — Vol. 2, Issue 3. — P. 191–200.

HybriLIT — Heterogeneous Computing Cluster. [Electronic resource]. URL: http://hybrilit.jinr.ru/en/ (дата обращения: 30.10.2016).

Evolutionary optimization algorithms in peptide structure prediction

S. V. Poluyan[1,a], N. M. Ershov[2,b]

[1] Dubna State University, 19, University st., Dubna, 141980, Russia

[2] Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics, 1-52, Leninskiye Gory, GSP-1, Moscow, 119991, Russia

E-mail: [a] svpoluyan@gmail.com, [b] ershovnm@gmail.com

The paper presents an exploration of an area of application of the stochastic evolutionary optimization algorithms in the problem of peptide secondary structure prediction. The paper considers one of the problems in structural bioinformatics – prediction of three-dimensional peptide structure from amino acid sequence. The task of finding two main regular secondary structures of various peptides with different length in continuous space of main-chain and side-chain torsion is formulated. It includes the following peptides with known native structure: model, designed, and naturally occurring. This paper presents the main assumptions that reduces the task of peptide spatial structure prediction to continuous global optimization problem. It is important to note that in the considered statement was not used a restriction of the search space by using statistical information of preferable values of angles in side-chains and libraries of fragments extracted from known protein structures. The analysis of presently existing methods for solving the above-described task was carried out. The main features of the force field used for calculation of energy of peptide are considered. The paper proposes a scheme for changing the force-field parameter, which characterizes non-covalent interactions, during optimal structure search. The feature of this approach is that the general formulation of the task as a single-objective problem remains in force. The paper presents the results of various numerical experiments of different evolutionary algorithms with comparison between basic evolutionary operators. The most effective operators are indicated. The paper presents an approach of parallelization of the considered algorithms. Some limitations of using optimization methods for solving the above-described task are shown. The comparison between found structures and structures obtained by relevant methods of solving the above-described task was carried out. The practical importance of the present study consists in identification of limits of applicability of evolutionary optimization methods, which allows to estimate the prospect of using evolutionary algorithms in contemporary issues of structural bioinformatics (for example, in the peptide-protein docking problem).

Keywords: secondary structure, conformational search, evolutionary computation, global optimization.