УДК 519.6

С. В. Полуян, Н. М. Ершов

Применение эволюционных методов оптимизации для предсказания пространственной структуры пептидов

Рассматривается вопрос применимости стохастических эволюционных алгоритмов оптимизации к задаче предсказания вторичной структуры пептидов и описанию подхода к изменению одного из параметров силового поля в процессе поиска оптимальной структуры.

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

Об авторах

Ершов Николай Михайлович — кандидат физико-математических наук, старший научный сотрудник факультета вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова, доцент кафедры прикладной математики и информатики Государственного университета «Дубна».

Полуян Сергей Владимирович — аспирант и ассистент кафедры прикладной математики и информатики факультета естественных и инженерных наук Государственного университета «Дубна».

В работе рассматривается одна из основных задач структурной биоинформатики — предсказание трёхмерной структуры белка по аминокислотной последовательности. Белки являются макромолекулами, состоящими из α-аминокислот, соединённых в цепочку пептидной связью, тем самым образуя полипептидную цепь. Предсказание структуры белка — предсказание по аминокислотной последовательности трёхмерной 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.

Важно отметить, что основы молекулярной биологии, относящиеся к белкам, базируются на экспериментах и впоследствии выдвинутых гипотез, которые, вообще говоря, далеко не всегда являются до конца выясненным вопросами. Поэтому исследователи оперируют гипотезами, которые, спустя некоторое время и подтверждения, превращаются в постулаты.

Наиболее широко принимаемая гипотеза, уже превратившаяся на сегодняшний день в догму, объясняющая процесс самоорганизации белка, была сформулирована Анфинсеном (Anfinsen) в 1973 г. В получившем Нобелевскую премию эксперименте [3] было показано, что после денатуризации (с химическим разрушением дисульфидных связей) развёрнутая цепь рибонуклеазы при переводе в химическую среду, близкую к физиологическим условиям, ренатурирует, т.е. восстанавливает свою нативную пространственную структуру. В своих работах Анфинсен пришёл к выводу, что при сворачивании не используется никакой информации кроме той, что уже содержится в последовательности белка. В результате Анфинсеном была представлена так называемая «термодинамическая гипотеза», которую можно представить в виде трёх положений: нативное состояние белка уникально; у белка не существует никаких других конфигураций с аналогичной свободной энергией; термодинамическое равновесие; небольшие изменения в окружающей среде не могут дать начало изменениям в нативной конфигурации белка, т.е. нативное состояние стабильно; доступность минимума свободной энергии; сворачивание белковой цепи не должно включать очень сложные конформационные изменения. Необходимо отметить, что в проведённом Анфинсеном эксперименте

участвовал белок длиной 124 остатка, в то время как средняя длина аминокислотной последовательности, по данным [12], равна порядка 317. Таким образом, гипотеза утверждает, что для малых глобулярных белков нативная структура определяется только последовательностью аминокислот в белке.

Здесь также необходимо упомянуть парадокс Левинталя (Levinthal) [12], сформулированный в 1969 г.: промежуток времени, за который полипептидная цепь приходит к своему свёрнутому состоянию, на много порядков меньше, чем если бы просто перебирались все возможные конфигурации цепи. Действительно, если рассматривать все возможные степени свободы у развёрнутой полипептидной цепи, можно получить астрономическое число конформаций. Этот парадокс приводит к интересному выводу — самоорганизация белка является направленным процессом, и существует некий путь (не обязательно один) от растянутой полипептидной цепи к нативному состоянию. С учётом термодинамической гипотезы процесс сворачивания полипептидной цепи представим как процесс минимизации свободной энергии белка.

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*, т.е. процесс сворачивания цепи рассматривается без привлечения каких-либо дополнительных эмпирических предположений, только естественные законы природы.

Настоящая работа посвящена вопросу применимости стохастических эволюционных алгоритмов оптимизации к вышеописанной задаче и описанию подхода к изменению одного из параметров силового поля в процессе поиска оптимальной структуры пептида.

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.

Силовое поле

В численных экспериментах использовалось силовое поле *ROSETTA* [13]. 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 определённым весом. Веса термов калибруются на определённой выборке белков. Термы описывают межатомные взаимодействия с использованием классической механики (силы отталкивания и притяжения Леннарда-Джонса, электростатические взаимодействия) 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 — *score*12 и *talaris*2014, соответствующие предыдущему и текущему стандарту скоринг-функции у силового поля *ROSETTA*. Принципиальное различие *score*12 и *talaris*2014 заключается в способе вычисления электростатических

взаимодействий. В первом случае используется терм, описывающий статистически полученные данные из *PDB* [13], во втором случае в явном виде вычисляется кулоновский потенциал.

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

Результаты численных экспериментов

Для поиска оптимальной структуры использовались следующие эволюционные алгоритмы: адаптивная дифференциальная эволюция *JADE* [21], эволюционная стратегия *ESCH* [15], метод роя частиц *PSO* [10] с локальным поиском *SW* [16], алгоритм бактериального поиска с адаптивным изменение шага *SABFO* [2], алгоритм роевой оптимизации со стратегией соревнования особей *CSO* [5], неоднородный клеточный генетический алгоритм *NCGA* [1], эволюционная стратегии с адаптацией матрицы ковариаций *CMAES* [8], гибрид дифференциальной эволюции с *CMAES* для локальной оптимизации *JDE-CMAES* [4]. Выбор рассматриваемых алгоритмов обусловлен хорошими результатами при решении различных практических задач оптимизации [5], а также разнообразием стратегий у различных операторов.

На первом этапе вычислительных экспериментов ставилась задача нахождения оптимальной структуры модельного пептида длинной 10 аминокислотных остатков A10 [18], 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 10^6 для *score*12 и *talaris*2014. Для каждого алгоритма выполнено 25 независимых запусков.

Важно отметить, что сходимость у эволюционных алгоритмов в значительной степени зависит от используемых параметров, причём их число варьируется от двух (CSO) до 12 (SABFO). В проводимых экспериментах часть параметров подбирались с учётом 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* [14; 20], предсказывающим структуру пептида с использованием статистической информации (низкоэнергетических фрагментов) для получения пула крупнозернистых моделей и последующей полноатомной минимизацией.

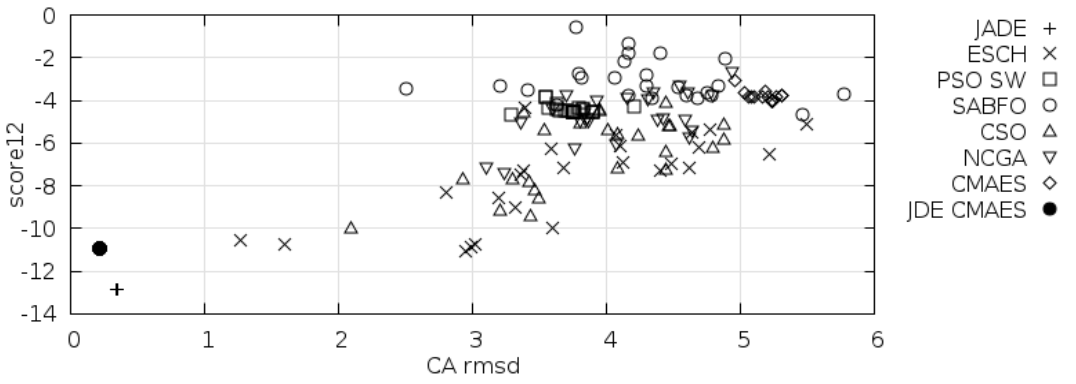

Рис. 1. Результаты 25 независимых запусков для рассматриваемых алгоритмов

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.

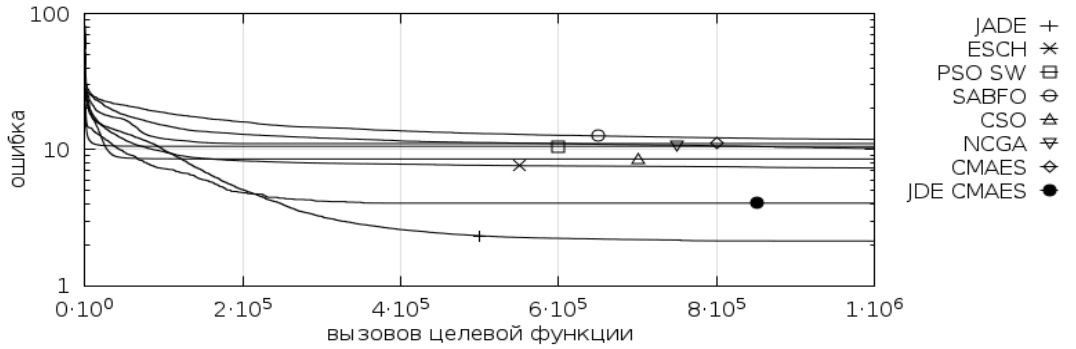

Рис. 2. Усреднённая по 25 запускам сходимость алгоритмов (скоринг-функция *score*12)

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. 2 показана усреднённая сходимость для каждого алгоритма. Для построения логарифмической шкалы использовалась ошибка относительно глобального минимума со значением –15.

На рис. 1 и 2 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.

На рис. 3 и 4 представлены результаты при скоринг-функции *talaris*2014.

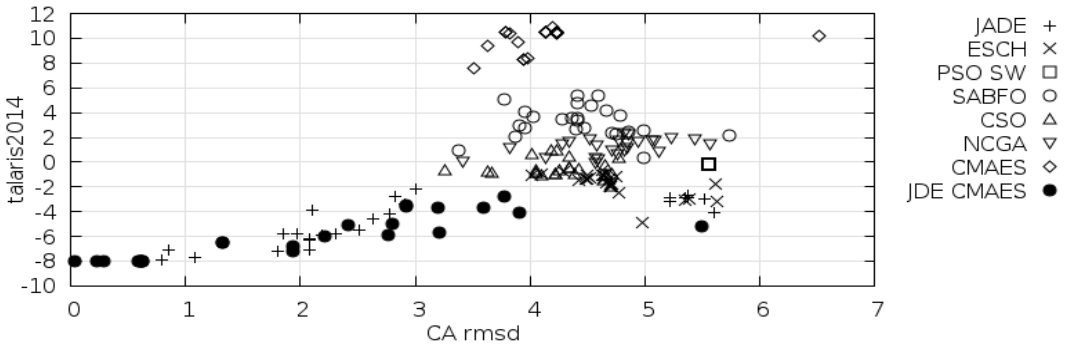

Рис. 3. Результаты 25 независимых запусков для рассматриваемых алгоритмов

Рассматриваемые эволюционные алгоритмы имеют схожую структуру, включающую в себя три основных оператора — отбор, скрещивание, мутация. Причём 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 приведённых 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 даёт 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 принципиальным оператором является оператор мутации, причём со стратегией *current-to-best*.

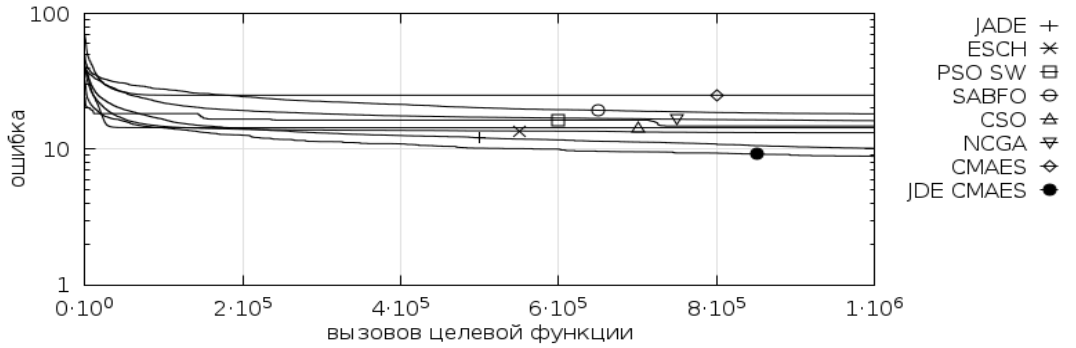

Рис. 4. Усреднённая по 25 запускам сходимость алгоритмов (скоринг-функция *talaris*2014)

На втором этапе вычислительных экспериментов ставилась задача нахождения оптимальной структуры модельного пептида V4GGV4 [19] (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 χ_{1-4} (пространство поиска $[-\pi, \pi]$), длин ковалентных связей (δ_1 = 0,05 Å), валентных углов (δ_2 = 0,1 рад.) для каждого атома пептида, неосновных торсионных углов боковой цепи каждого атома (δ_3 = 0,1 рад.). Границы с δ_{1-3} рассчитывались относительно идеализированных значений используемых в ROSETTA (аналогично методу CONCOORD [6]), пространство поиска непрерывно. Таким образом, размерность задач для α-спирали и β-листа составила 302 и 428 параметров соответственно.

На рис. 5 показаны результаты оптимизации с помощью алгоритма JADE при 10^7 вызовах целевой функции. Полученное отклонение атомов главной цепи составило меньше половины ангстрема относительно структуры найденной с помощью PEP-FOLD. Суперпозиции на рис. 5 получены с использованием 3DSS [17].

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.

Поскольку данный потенциал (*fa_elec*) определяет нековалентные взаимодействия, было предложено использовать метод имитации отжига [11] для изменения соответ-

ствующего веса потенциала в процессе оптимизации. Такой подход был использован для пептидов YMEARAMEARA (α-спираль) и Ace-ITVNGKTY-Nme (β-лист) [7], размерность задач составила 501 и 395 параметров соответственно. На рис. 6 представлены результаты с использованием отжига и без для α-спирали. На рис. 7 показаны получаемые структуры.

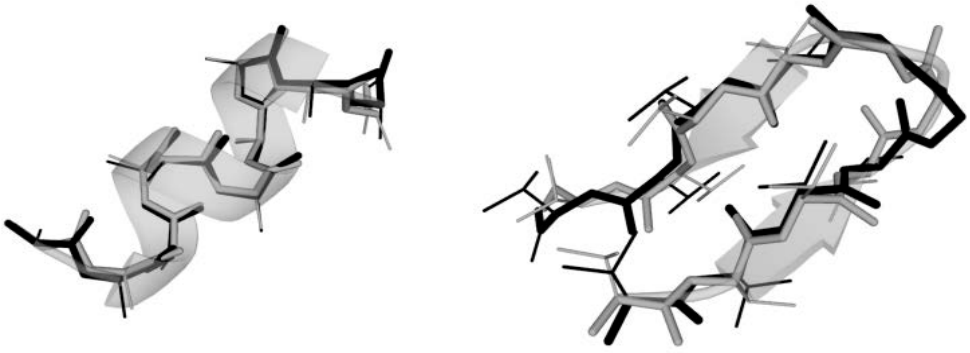

Рис. 5. Суперпозиция главных цепей пептидов, PEP-FOLD (чёрный) и JADE (серый)

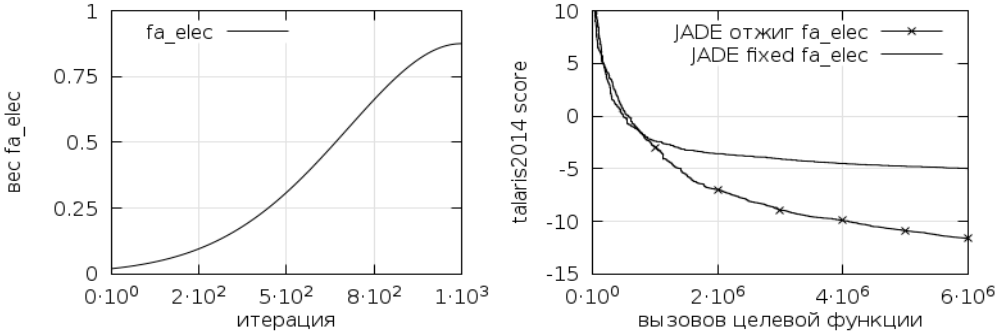

Рис. 6. Функция изменения кулоновского потенциала и сходимость для α-спирали

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*. 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 10 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* [9].

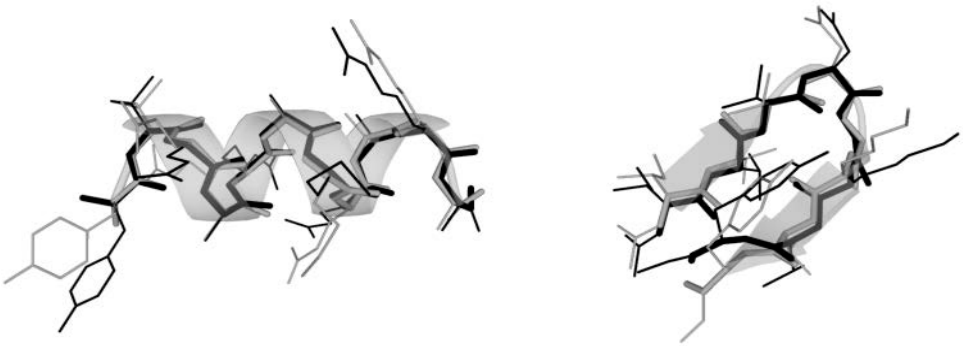

Рис. 7. Суперпозиция главных цепей пептидов, PEP-FOLD (чёрный) и JADE (серый)

Выводы

В результате выполненной работы проведено исследование применимости эволюционных алгоритмов оптимизации к задаче предсказания структуры пептидов. Показано, 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 её эффективность. Проведено численное исследование алгоритмов с использованием предложенной схемы на модельных и реальных пептидах. Для некоторых рассматриваемых алгоритмов произведена параллельная реализация. Результаты проведённых в данной работе исследований демонстрируют, что стратегия мутации в алгоритме дифференциальной эволюции в значительной степени определяет эффективность сходимости.

Исследование гибридов различных эволюционных алгоритмов и разработка критериев для изменения параметров и применение эволюционных алгоритмов или других методов теории искусственного интеллекта к решению задачи настройки параметров базового алгоритма (мета-оптимизации) у эволюционных алгоритмов могут быть целью дальнейшей работы.

На основании проведённых исследований можно заключить, что с использованием предложенной схемы эволюционные алгоритмы оптимизации способны находить оптимальную структуру коротких пептидов длинной порядка 10 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.

Библиографический список

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

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

3. Anfinsen C. Principles that Govern the Folding of Protein Chains / C. Anfinsen // Science. – 1973. – V. 181(4096). – P. 330–331.

4. Brest, J. Large scale global optimization using self-adaptive differential evolution algorithm / J. Brest [et al.] // IEEE World Congress on Computational Intelligence. – 2010. – P. 1–8.

5. Cheng, R. Competitive Swarm Optimizer for Large Scale Optimization / R. Cheng, Y.A Jin // IEEE Transactions on Cybernetics. – 2015. – V. 45(2). – P. 191–204.

6. de Groot, B. L. Prediction of protein conformational freedom from distance constraints / B.L. de Groot [et al.] // Proteins. – 1997. – V. 29(2). – P. 240–251.

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

8. Hansen, N. Adapting arbitrary normal mutation distributions in evolution strategies: The

covariance matrix adaptation / N. Hansen, A. Ostermeier // Proceedings of the 1996 IEEE International Conference on Evolutionary Computation. – 1996. – P. 312–317.

9. Heterogeneous Computing Cluster HybriLIT, 2015. – URL: http://hybrilit.jinr.ru/en/.

10. Kennedy, J. Particle swarm optimization / J. Kennedy, R. Eberhart // Proceedings of IEEE International Conference on Neural Networks. – 1995. – V. 4. – P. 1942–1948.

11. Kirkpatrick, S. Optimization by Simulated Annealing / S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi // Science. – 1983. – V. 220(4598). – P. 671–680.

12. Levinthal, C. How to Fold Graciously / C. Levinthal // Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, University of Illinois Press. – 1969. – Jan. – P. 22–24.

13. O'Meara, M. J. Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta / M.J. O'Meara [et al.] // Journal of Chemical Theory and Computation. – 2015. – V. 11(2). – P. 609–622.

14. Shen, Y. Improved PEP-FOLD approach for peptide and miniprotein structure prediction / Y. Shen [et al.] // Journal of Chemical Theory and Computation. – 2014. – V. 10. – P. 4745–4758.

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

16. Solis, F. J. Minimization by random search techniques / F.J. Solis, R.J-B. Wets // Mathematics of Operation Research. – 1981. – V. 6(1). – P. 19–30.

17. Sumathi, K. 3dSS: 3D structural superposition / K. Sumathi [et al.] // Nucleic Acids Research. – 2006. – V. 34. – P. 128–132.

18. Sung, S. S. Helix Folding Simulations with Various Initial Conformations / S.S. Sung // Biophysical Journal. – 1994. – V. 66. – P. 1796–1803.

19. Sung, S. S. Monte Carlo Simulations of β-Hairpin Folding at Constant Temperature / S.S. Sung // Biophysical Journal. – 1999. – V. 76. – P. 164–175.

20. Thevenet, P. PEP-FOLD: an updated de novo structure prediction server for both linear and disulfide bonded cyclic peptides / P. Thevenet [et al.] // Nucleic Acids Research. – 2012. – V. 40. – P. 288–293.

21. Zhang, J. JADE: Adaptive differential evolution with optional external archive / J. Zhang, A. Sanderson // IEEE Transactions on Evolutionary Computation. – 2009. – V. 13(5). – P. 945–958.

Поступила в редакцию
15.08.2016