ПРИМЕНЕНИЕ ЭВОЛЮЦИОННЫХ И РОЕВЫХ АЛГОРИТМОВ ОПТИМИЗАЦИИ ДЛЯ РЕШЕНИЯ МОДЕЛЬНОЙ ЗАДАЧИ ПРЕДСКАЗАНИЯ СТРУКТУРЫ БЕЛКА

Быстров М.А. [1, a], Ершов Н.М. [2, b]

[1] *ГБОУ ВО МО «Университет «Дубна», Институт системного анализа и управления, ул. Университетская 19, Дубна, Московская область, 141982, Россия*

[2] *МГУ имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, ул. Ленинские Горы 1, стр.52, Москва, 119234, Россия*

E-mail: [a] max-zdec@mail.ru, [b] ershov@cs.msu.ru

Работа посвящена проблемам прогнозирования пространственной структуры белковых молекул, полипептидов и их комплексов. Предлагаемый нами метод основан на решении задачи оптимизации, в которой целевой функцией является потенциальная энергия молекулы, а параметрами оптимизации – длины связей между атомами и углы вращения. Главными особенностями таких задач является большая размерность и высокая вычислительная сложность. Проведенные предварительные исследования показали, что множество существующих алгоритмов оптимизации решают такую задачу неудовлетворительно. Типичное время вычисления задачи занимает от нескольких часов до нескольких дней. Кроме того, для корректного вычисления целевой функции требуется серьезное программное обеспечение. Эти факторы существенно усложняют процесс разработки эффективных алгоритмов для решения задачи предсказания структуры белков и их комплексов. В настоящей работе предлагается упрощенная, модельная задача укладки графа на плоскости, которая позволяет проводить расчеты быстрее и без использования специального программного обеспечения, разрабатывать новые и улучшать существующие алгоритмы оптимизации. В работе приводятся результаты численного исследования ряда алгоритмов роевой и эволюционной оптимизации при решении поставленной модельной задачи.

Ключевые слова: биоинформатика, алгоритмы роевой оптимизации, эволюционные алгоритмы, алгоритмы роевого интеллекта, укладка графа, предсказание структуры белка.

1. Введение

Важной задачей структурной биоинформатики является предсказание трехмерной структуры белков и их комплексов. Существующие методы моделирования пространственной структуры белковой молекулы по ее линейной последовательности аминокислот являются недостаточно быстрыми и точными, в настоящий момент не существует таких алгоритмов, с помощью которых можно было бы за разумное время предсказать с высокой точностью пространственную структуру белка. Один из подходов к решению задач данного класса заключается в их сведение к задачам непрерывной оптимизации. Параметрами оптимизации при этом являются длины межатомных связей и углы их вращения. В качестве целевой функции берется потенциальная энергия всей последовательности аминокислот. Глобальный минимум потенциальной энергии молекулы будет соответствовать искомой пространственной структуре. Данную задачу выделяет высокая размерность и, как следствие, длительное время вычисления целевой функции.

В силу указанных причин было предложено вести разработку эффективных алгоритмов предсказания структуры на более простой модельной задаче, обладающей следующими свойствами:

- она должна быть аналогичной рассматриваемым задачам биоинформатики;
- целевая функция должна вычисляться быстро;
- вычисление целевой функции не требует специализированного программного обеспечения [1].

В качестве такой модельной задачи нами была выбрана задача укладки графа на плоскости.

2. Модельная задача укладки графов на плоскости

Задается граф (дерево), длины ребер которого являются одинаковыми и фиксированными. Переменными параметрами являются углы отклонения ребер графа от ребер-родителей (рис. 1).

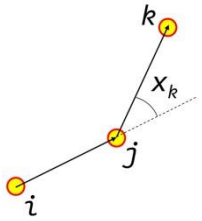

Рисунок 1. Параметры оптимизации

Для графа, состоящего из n вершин, свободными параметрами будут являться (n – 2) углов. Положение первых двух вершин является фиксированным. Каждая вершин графа имеет заданный тип, всего имеется десять типов, меткой типа служит символ. Каждый тип вершины характеризуется двумя параметрами: массой w и зарядом q. Целевая функция — потенциальная энергия укладки графа – складывается из попарных потенциалов для всех пар вершин и представляется следующей формулой:

$$E(x) = \alpha \cdot \sum_{i \neq j} \frac{w_i w_j}{r_{ij}^G} + \beta \cdot \sum_{i \neq j} \frac{q_i q_j}{r_{ij}^E} + \gamma \cdot \sum_{i \neq j} \frac{p_{ij}}{r_{ij}^P} \qquad (1)$$

Первое слагаемое соответствует «гравитационному» притяжению вершин, второе – «электростатическому» взаимодействию, третье – взаимному отталкиванию вершин. Неизвестными параметрами являются набор углов x. Параметры α, β и γ — это весовые коэффициенты; G, E, P — параметры потенциалов; pij — штраф за слишком близкое расположение двух вершин друг другу.

3. Программная реализация

Для решения поставленной модельной задачи предполагается использовать алгоритмы роевой оптимизации, которые хорошо зарекомендовали себя при решении задач глобальной оптимизации [2]. Нами разработан комплекс программ, предназначенный для автоматизации процесса разработки и анализа подобных алгоритмов. Отдельная программа (C++) делает заданное количество запусков алгоритма, вычисляет и сохраняет в файле статистику выполнения алгоритма. Эти данные затем обрабатываются специальным скриптом (Python), на основе чего строятся изображения (в том числе анимированные) графов и графики сходимости текущего алгоритма.

Основная программа выполняет загрузку и сохранение данных, сбор статистики и визуализацию. Алгоритмы оптимизации включены в программный комплекс, их взаимодействие с основной программой ограничивается запросами на вычисление целевой функции. Следует подчеркнуть, что сходимость алгоритмов зависит от используемых параметров. В проводимых экспериментах часть параметров подбирались с учетом рекомендаций авторов алгоритмов и собственных исследований. Для объективности сравнения эксперименты проводились на одинаковом размере популяций решений и одинаковом ограничении на максимальное количество обращений к целевой функции. Число частиц для всех алгоритмов было выбрано равным 1000, maxefs — максимальное число обращений к целевой функции, efs — счетчик обращений к целевой функции. В настоящий момент в систему включены следующие пять алгоритмов.

1. *Particle Swarm Optimization (PSO)* – метод роя частиц. В нашей работе мы использовали *canonical PSO* [3]. Коэффициент инерции $\chi = 0.75$. Коэффициенты ускорения: $c_1 = 1.8$, $c_2 = 2.0$.
2. *A Competitive Swarm Optimizer (CSO)* – метод конкурирующего роя [4]. Коэффициент социальной составляющей $\varphi = 0.3$.
3. *Bacterial Foraging Optimization (BFO)* – алгоритм бактериального поиска [5]. Параметры: длина цикла отбора и размножения $l = 32$, длина шага бактерии $step = 1.0 \cdot \frac{efs}{maxefs}$; число шагов за каждую итерацию цикла отбора и размножения случайно, но не более ста; рассеяние для каждой бактерии берется с вероятностью $P_{BFO} = 0.05$.
4. *Genetic algorithm (GA)* – генетический алгоритм [6]. Для выбора более приспособленных особей использовался метод рулетки. Вероятность мутации каждого гена P_{GA}=0.004.
5. *Differential evolution (DE)* – алгоритм дифференциальной эволюции [7]. Сила мутации $F = 0.66$. Вероятность замещения мутированного вектора исходным вектором: $0.2 \le P_{DE} \le 0.5$, при $\frac{efs}{maxefs} < 0.5$; $0 \le P_{DE} \le 0.2$, при $\frac{efs}{maxefs} \ge 0.5$.

4. Численное исследование

На первом этапе вычислительных экспериментов ставилась задача нахождения оптимальной структуры модельного белка, с искомой структурой, аналогичной структурам α-спирали и β-листа. Поиск укладок проводился на цепочках **AEEAAEEAAE** и **AAAACCEEEE**. Все вершины имеют массу $m = 2$, значения зарядов: $q_A = +2$, $q_C = 0$, $q_E = -2$. Для всех анализируемых алгоритмов выполнялось по 50 независимых запусков для каждого из двух графов. Результаты сравнения эффективности данных пяти алгоритмов представлены на рис. 2. Видно, что абсолютным победителем в проведенном сравнении является алгоритм CSO, а наиболее слабые результаты показал генетический алгоритм. Интересно, что алгоритм бактериального поиска оказался по-разному эффективен на двух рассматриваемых задачах: β-лист он укладывает намного лучше, чем α-спираль.

На рис. 3 построены графики, показывающие динамику сходимости алгоритма CSO для лучшего запуска (с абсолютным минимумом целевой функции) на обоих графах, а также показаны найденные им укладки.

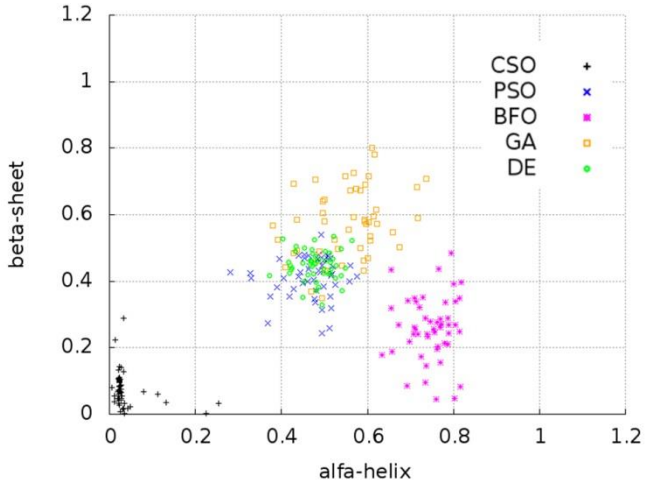

Рисунок 2. Сравнительный анализ работы пяти алгоритмов оптимизации

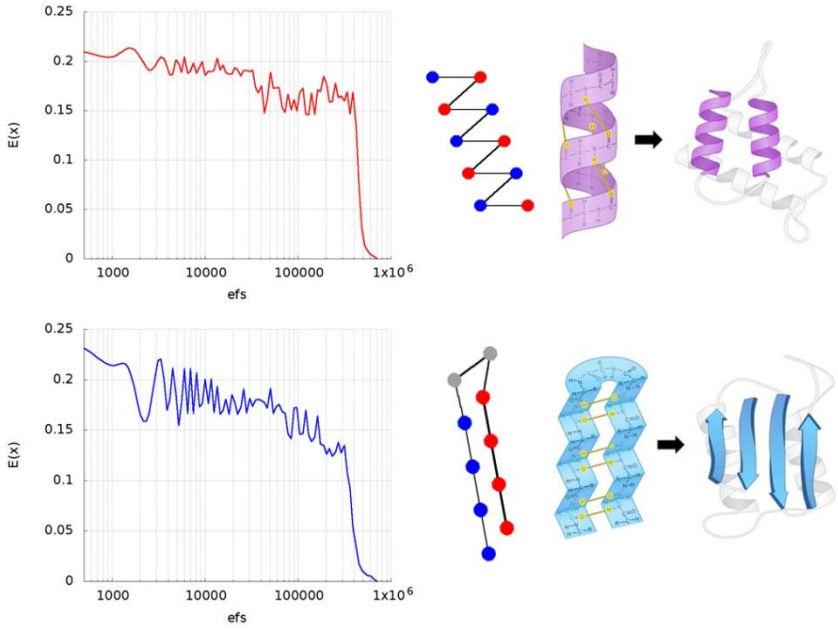

Рисунок 3. Графики сходимости алгоритма *CSO*, справа представлены найденные укладки графов

5. Заключение

В результате выполненной работы были получены следующие результаты:

- поставлена модельная задача укладки графа на плоскости, аналогичная задаче предсказания трехмерной структуры белка;
- разработана программная система для тестирования популяционных алгоритмов непрерывной оптимизации и визуализации результатов их работы;
- проведено численное исследование пяти наиболее популярных алгоритмов для двух типов графов, со структурой аналогичной α-спиралям и β-листам.

Результаты проведенных численных экспериментов показали, что наиболее эффективным для решения задачи укладки графа является алгоритм CSO. Однако полученные результаты не дают пока достаточной информации о преимуществах того или иного алгоритма. Сходимость сильно зависит от большого набора параметров каждого алгоритма, от числа частиц роя и от ограничения на число вычислений целевой функции. Целью дальнейшей

работы является настройка оптимальных параметров для каждого алгоритма при решении модельной задачи, а также применение настроенных таким образом алгоритмов к решению задач предсказания трехмерной структуры белков и их комплексов.

Литература

[1] Полуян С. В., Ершов Н. М. Разработка эффективных алгоритмов биоинформатики на основе решения модельной задачи укладки графов // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем: материалы Всероссийской конференции с международным участием. Москва, РУДН, 24-28 апреля 2017 г. — РУДН Москва, 2017. — С. 333–335.

[2] Карпенко А.П. Современные алгоритмы поисковой оптимизации. М.: Изд-во МГТУ им. Н.Э. Баумана, 2014.

[3] Clerc J. K. M. The particle swarm – explosion, stability, and convergence in a multidimensional complex space, IEEE Transactions on Evolutionary Computation, 2002, V. 6, № 1, pp. 58-73.

[4] Cheng R., Jin Y. A Competitive Swarm Optimizer for Large Scale Optimization, in IEEE Transactions on Cybernetics, Feb. 2015, Vol. 45, No. 2, pp. 191-204.

[5] Passino K. M. Biomimicry of bacterial foraging for distributed optimization and control, IEEE Control Systems Magazine, 2002, 22, pp. 52–67.

[6] Whitley D. A Genetic Algorithm Tutorial, Statistics and Computing, 1994, 4, pp. 65-85.

[7] Price K., Storn R., Lampinen J. Differential Evolution: A Practical Approach to Global Optimization, Springer, 2005.

APPLICATION OF EVOLUTIONARY AND SWARM OPTIMIZATION ALGORITHMS FOR SOLVING THE MODEL PROBLEM OF PREDICTING THE PROTEIN STRUCTURE

Bystrov M.A.[1, a], Ershov N.M.[2, b]

[1]*Dubna State University, the Institute of System Analysis and Management, Universitetskaya str.19, Dubna, Moscow region, 141982, Russia*

[2] *Lomonosov Moscow State University, the Faculty of Computational Mathematics and Cybernetics, Leninskie Gory str. 1, bdg. 52, Moscow, 119234, Russia*

E-mail: [a] max-zdec@mail.ru, [b] ershov@cs.msu.ru

The article is devoted to the problems of predicting the space structure of protein molecules, polypeptides and their complexes. The proposed method is based on solving the optimization problem, in which the objective function is the potential energy of the molecule, and the optimization parameters are the length of interatomic bindings and the rotation angles. The main features of such tasks are high dimensionality and high computational complexity. Preliminary studies have shown that many existing optimization algorithms solve this problem unsatisfactorily. Typical calculation time takes from several hours to several days. In addition, the correct calculation of the objective function requires serious software. These factors significantly complicate the process of developing efficient algorithms for solving the problem of predicting the structure of proteins and their complexes. A simplified model problem of laying a graph on a plane, which allows you to perform calculations faster and without using special software, develop new and improve existing optimization algorithms is proposed in this paper. The results of a numerical study concerning a number of algorithms for swarm and evolutionary optimization in solving a set model problem are presented.

Keywords: bioinformatics, swarm optimization algorithms, evolutionary algorithms, swarm intelligence algorithms, graph laying, protein structure prediction.