

Суперкомпьютерные дни в России

Конференция молодых ученых

Москва, 25–26 сентября 2017 г.

Применение параллельных эволюционных алгоритмов оптимизации в задачах структурной биоинформатики

С.В. Полуян

svpoluyan@gmail.com

Государственный университет «Дубна»

Институт системного анализа и управления

Исследование производительности эволюционных алгоритмов:

- ▶ в задаче однокритериальной глобальной оптимизации;
- ▶ в непрерывном пространстве поиска;
- ▶ вне синтетических тестов.

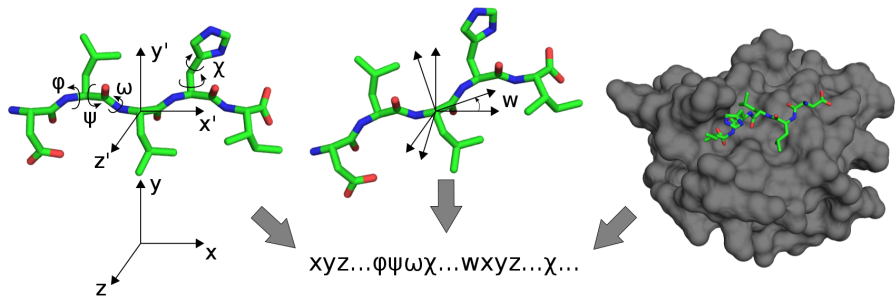
Нахождение глобального минимума энергии:

- ▶ вторичной структуры пептида по первичной структуре;
- ▶ комплекса белок-пептид (задача докинга).

Постановка задачи

Структура белка и пептида:

- ▶ линейная последовательность аминокислот;
- ▶ 20 стандартных аминокислот.

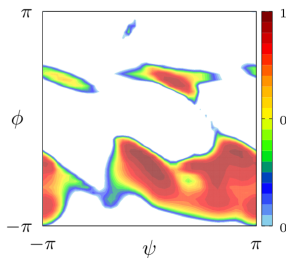


Непрерывное пространство поиска:

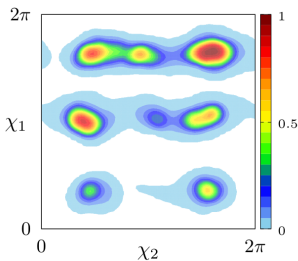
- ▶ углы главной цепи пептида: $\phi, \psi \in [-\pi, \pi]$ и $\omega \in [\pi - \delta, \pi + \delta]$;
- ▶ углы боковых цепей пептида и белка: $\chi_{1-4} [-\pi, \pi]$;
- ▶ вектор для смещения пептида относительно белка и поворот пептида в трехмерном пространстве.

Использование статистической информации

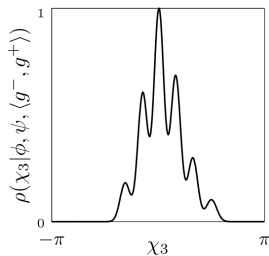
- ▶ Основная цель – исключение невозможных конформаций.
- ▶ Главная цепь – карты Рамачандрана с зависимостью от соседних остатков.
- ▶ Боковые цепи – центрируются sp^3 - sp^3 и sp^3 - sp^2 гибридизацией атомных орбиталей.



Карта Рамачандрана
(аспарагин)



Данные Top8000
(гистидин)



Библиотека Dunbrack 2010
(глутамин)

- ▶ Реализовано построение функций распределения.
- ▶ Непрерывность достигается линейной интерполяцией.

Rosetta All-Atom Force-Field:

- ▶ потенциал Леннарда-Джонса;
- ▶ неявный растворитель – Lazaridis-Karplus;
- ▶ электростатический потенциал;
- ▶ использование статистических данных;
- ▶ более 15 компонент.

Целевые функции:

- ▶ скоринг-функции *score12* и *talaris2014*;
- ▶ энергия комплекса

$$E_{complex} = E_{protein} + E_{peptide}$$

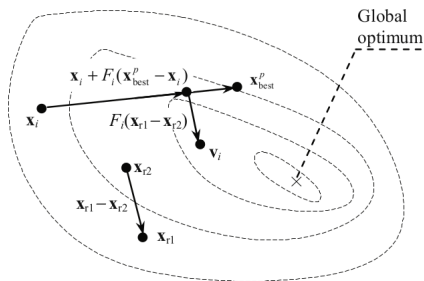
- ▶ энергия связывания

$$E_{interface} = E_{complex} - E_{protein} - E_{peptide}$$

Методы оптимизации

Эволюционные вычисления:

- ▶ стохастическая оптимизация;
- ▶ эвристические подходы;
- ▶ эволюция популяции решений;
- ▶ поддаются параллелизации.



Дифференциальная эволюция:

- ▶ мутация

$$v_i = x_i + F_i(x_{best}^p - x_i) + F_i(x_{r1} - \tilde{x}_{r2})$$

- ▶ скрещивание

$$u_{i,j}(t) = \begin{cases} v_{j,i} & \text{если } U_j(0, 1) \leq CR_i \\ x_{j,i} & \text{иначе} \end{cases}$$

- ▶ формирование нового поколения;
- ▶ адаптация F_i и CR_i .

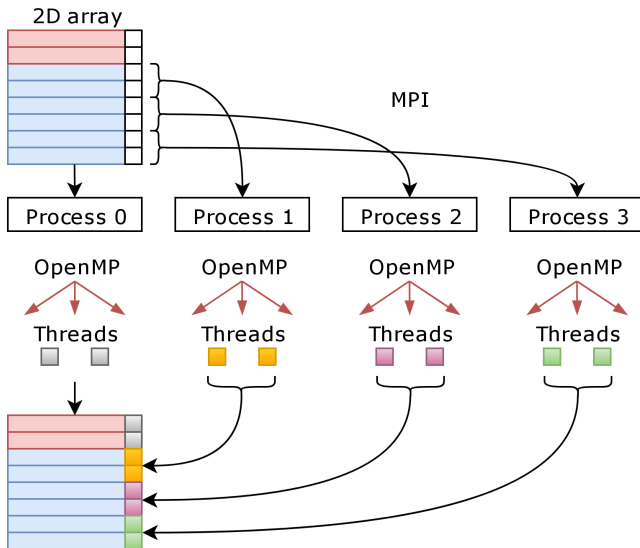
Выбранные алгоритмы и стратегии

- ▶ Эволюционная стратегия с адаптацией матрицы ковариаций CMAES.
- ▶ Эволюционная стратегия ESCH.
- ▶ Метод роя частиц PSO с локальным поиском SW.
- ▶ Алгоритм бактериального поиска SABFO.
- ▶ Алгоритм роевой оптимизации CSO.
- ▶ Неоднородный клеточный генетический алгоритм NCGA.
- ▶ Адаптивная дифференциальная эволюция JADE.
- ▶ Гибрид дифференциальной эволюции (JDE) с CMAES.

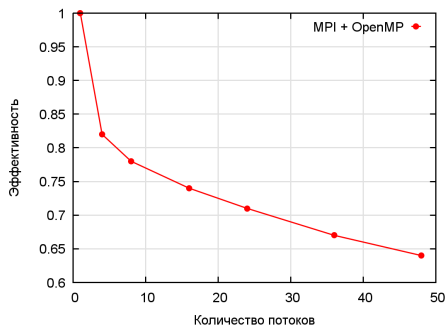
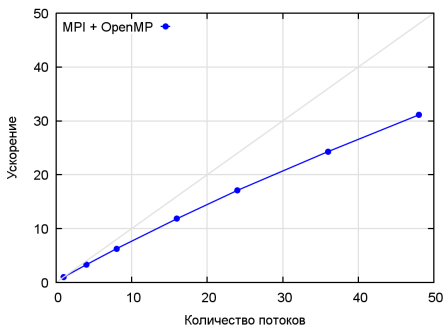
Особенности реализации:

- ▶ большинство алгоритмов реализовано на C++ с применением технологий MPI/OpenMP;
- ▶ все вычисления произведены на кластере ОИЯИ HybriLIT с двумя 12-ядерными процессорами Intel Xeon на узел.

Параллельная реализация

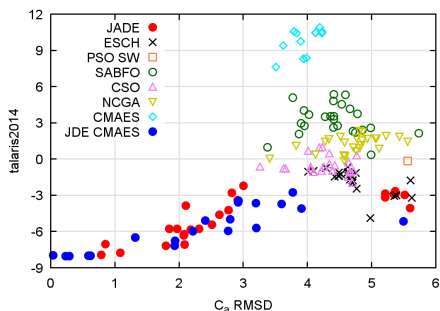
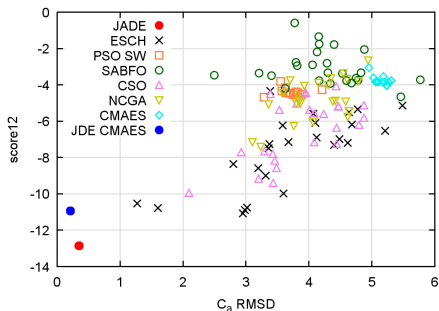


Результаты параллельной реализации

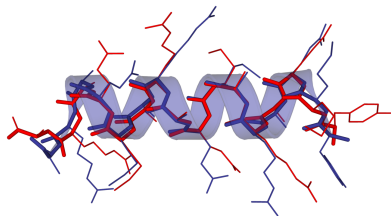


Количество потоков / узлов	4 / 1	8 / 1	16 / 1	24 / 1	36 / 2	48 / 2
Ускорение	3.31	6.28	11.84	17.1	24.3	31.1
Эффективность	0.82	0.78	0.74	0.71	0.67	0.64

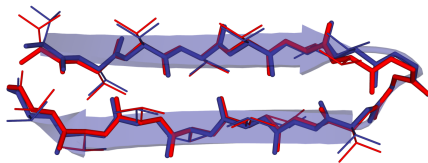
Предсказание структуры пептидов



Результаты 25 независимых запусков для модельного декапептида A₁₀

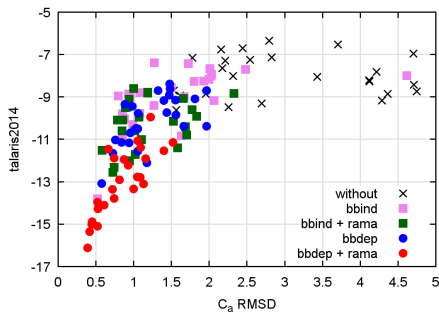
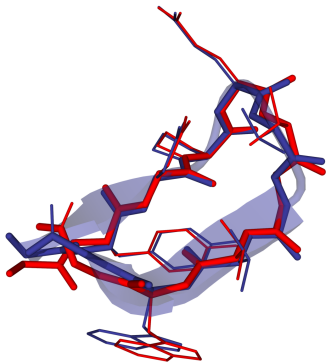


α -спираль Y(MEARA)₃



β -лист V₇G₂V₇

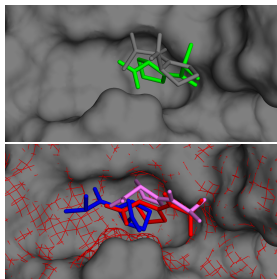
Предсказание структуры пептидов



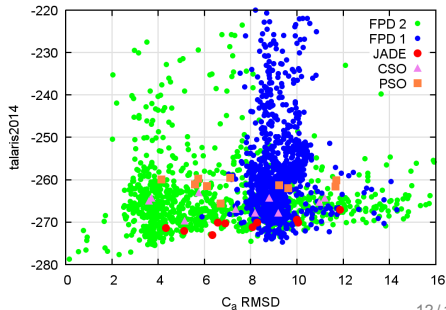
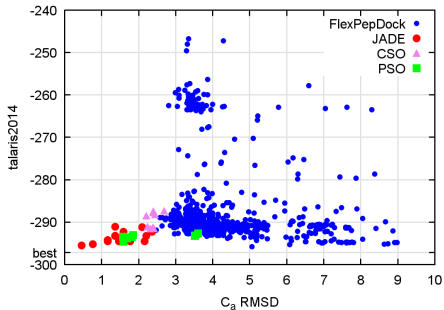
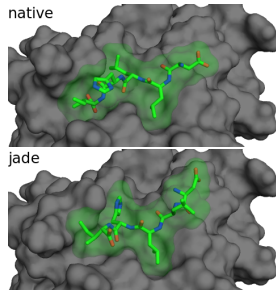
Результаты для пептида 1UAO (PDB id)

Результаты прямого докинга

2CYH [A:B] (PDB id): AP, $R = 10 \text{ \AA}$, $D = 25$.



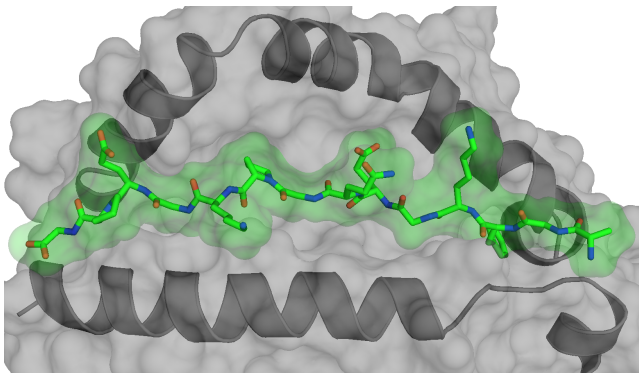
1JWG [B:D] (PDB id): DLLHI, $R = 5 \text{ \AA}$, $D = 54$.



- ▶ Только короткие пептиды, 10–15 остатков.
- ▶ Докинг с пределом в 30–50 параметров.
- ▶ Алгоритм дифференциальной эволюции.
- ▶ Плохая производительность схем адаптации.
- ▶ Вырождение популяции.
- ▶ Неспособность работать с локальными оптимизаторами.

Цели дальнейшей работы

- ▶ Алгоритмы оценки распределений.
- ▶ Алгоритмы Байесовской оптимизации.
- ▶ Многокритериальность.
- ▶ Алгоритм Бройдена–Флетчера–Гольдфарба–Шанно (BFGS).
- ▶ Актуальные задачи докинга с известной структурой пептида при связывании.



- ▶ Научному руководителю *Н.М. Ершову*, к.ф.-м.н., с.н.с. ВМК МГУ им. Ломоносова, за полезные обсуждения и замечания.
- ▶ *А.В. Головину*, д.х.м., с.н.с. НИИ ФХБ им. А.Н. Белозерского, преподавателю ФББ МГУ им. Ломоносова за ценные консультации, указание актуальных задач в области докинга и подбор силового поля.
- ▶ *Коллективу* гетерогенного кластера HybriLIT ОИЯИ, за поддержку и предоставление вычислительных ресурсов.

Спасибо за
внимание!