Компьютерные методы в фармакологии

В.Б.Сулимов

НИВЦ МГУ

Лекция № 5

Генетический алгоритм программы SOL, программа докинга FLM

Программа докинга SOL – классический докинг: 10 000 лигандов/неделя на Ломоносов-2 МГУ

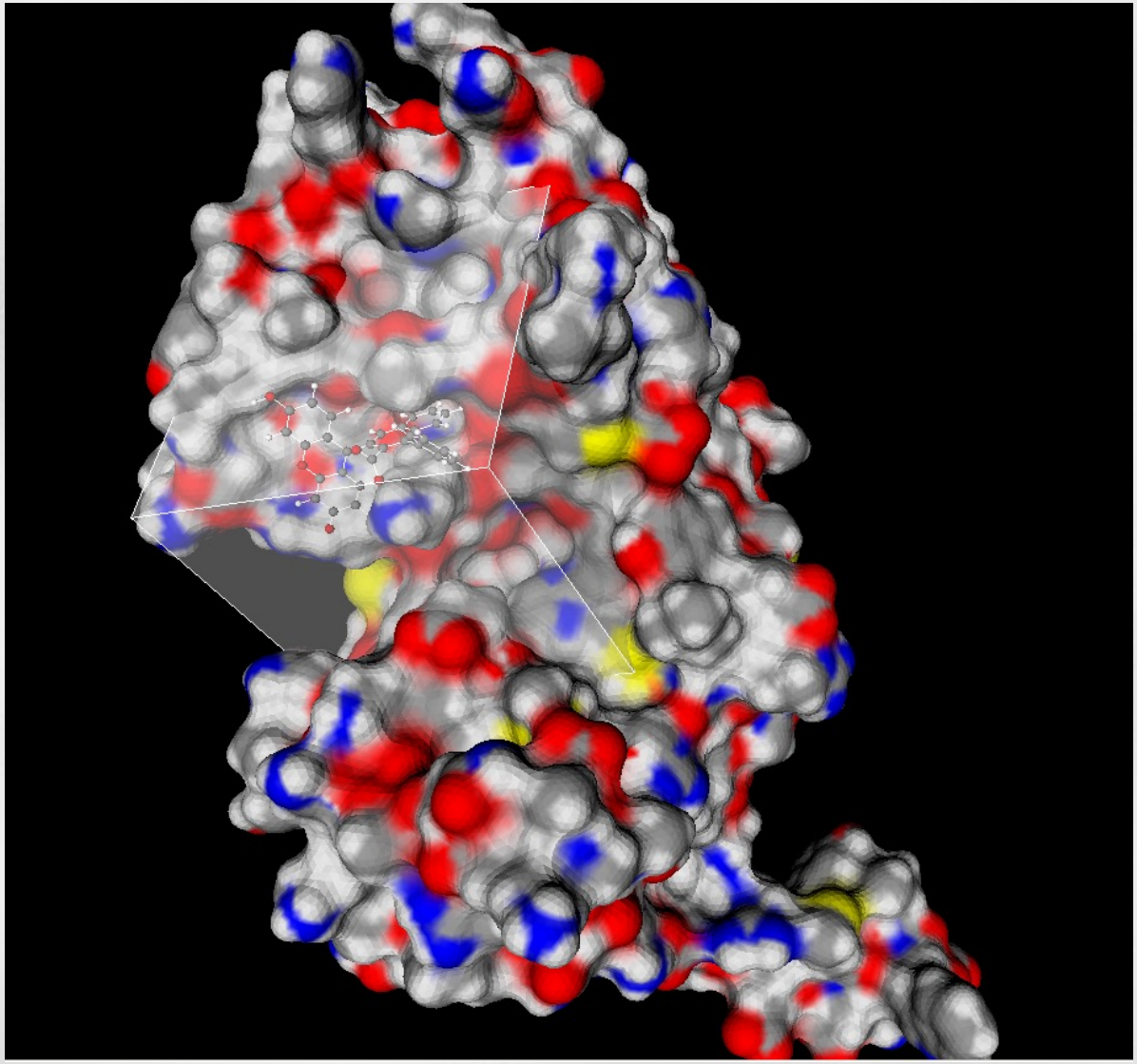
Позиционирование лиганда в активном центре заданного белка-мишени

- Жёсткий белок и гибкий лиганд
- Генетический алгоритм поиска положения лиганда в белке, соответствующего самой низкой энергии системы белок-лиганд
- Универсальное силовое поле MMFF94 фирмы Merck
- Активный центр белка в виде набора сеток различных потенциалов – жесткий белок
- Учет влияния растворителя - воды
- Гибкий лиганд и учет его внутренних напряжений

Модель белка – программа SOLGRID

- Белок жесткий
- Белок представлен сеткой потенциалов:
  - Кулоновские взаимодействия
  - Ван-дер-ваальсовы взаимодействия
  - Взаимодействие с растворителем
- Сетка потенциалов вычисляется заранее
- Сетка – это набор «сеток» потенциалов для разных типов атомов и разных типов взаимодействий
- Область докинга: куб с ребром 22 Ангстрема
  - 101 x 101 x 101 точек
  - длину ребра куба можно менять

Программа SOL-Grid



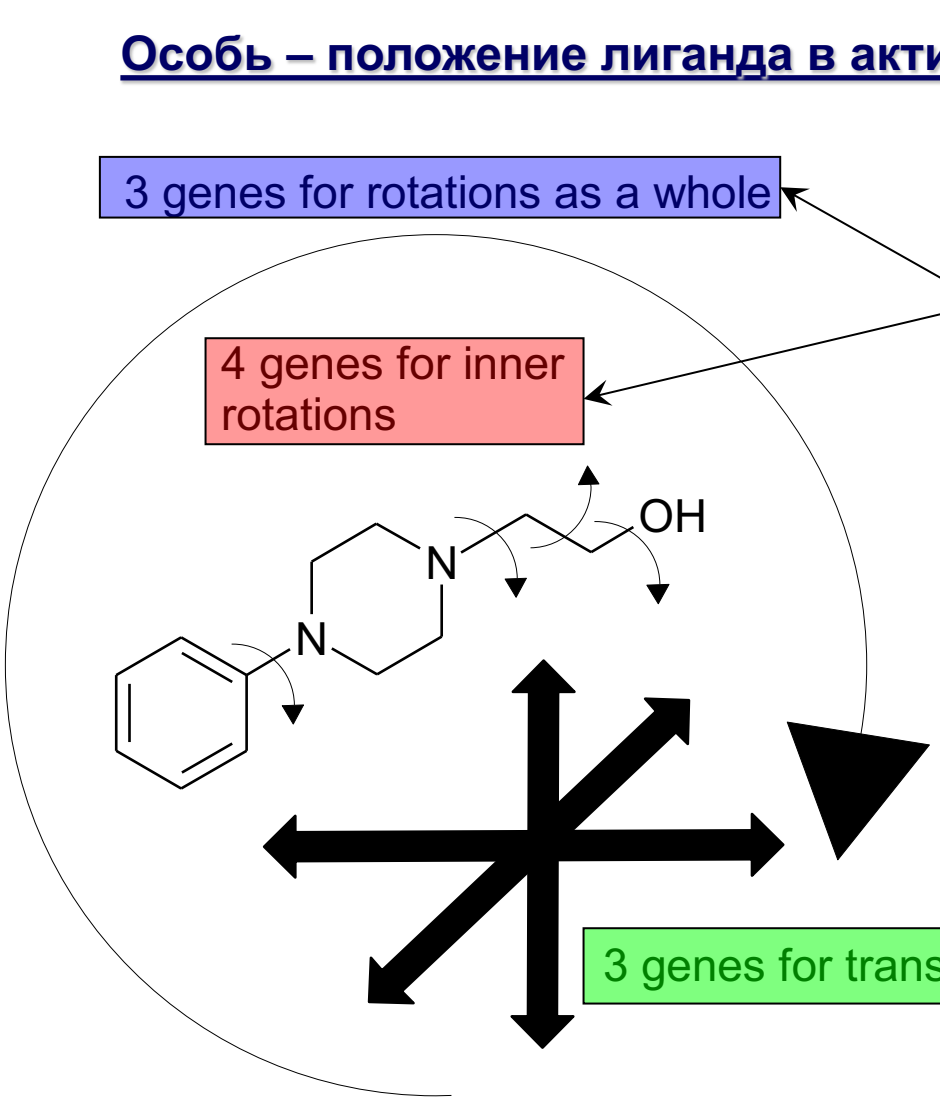
Программа SOL: кодирование положения лиганда в белке

ФЕНОТИН ГЕНОТИН

Эволюция популяции особей

$0 \le a_i \le 1$

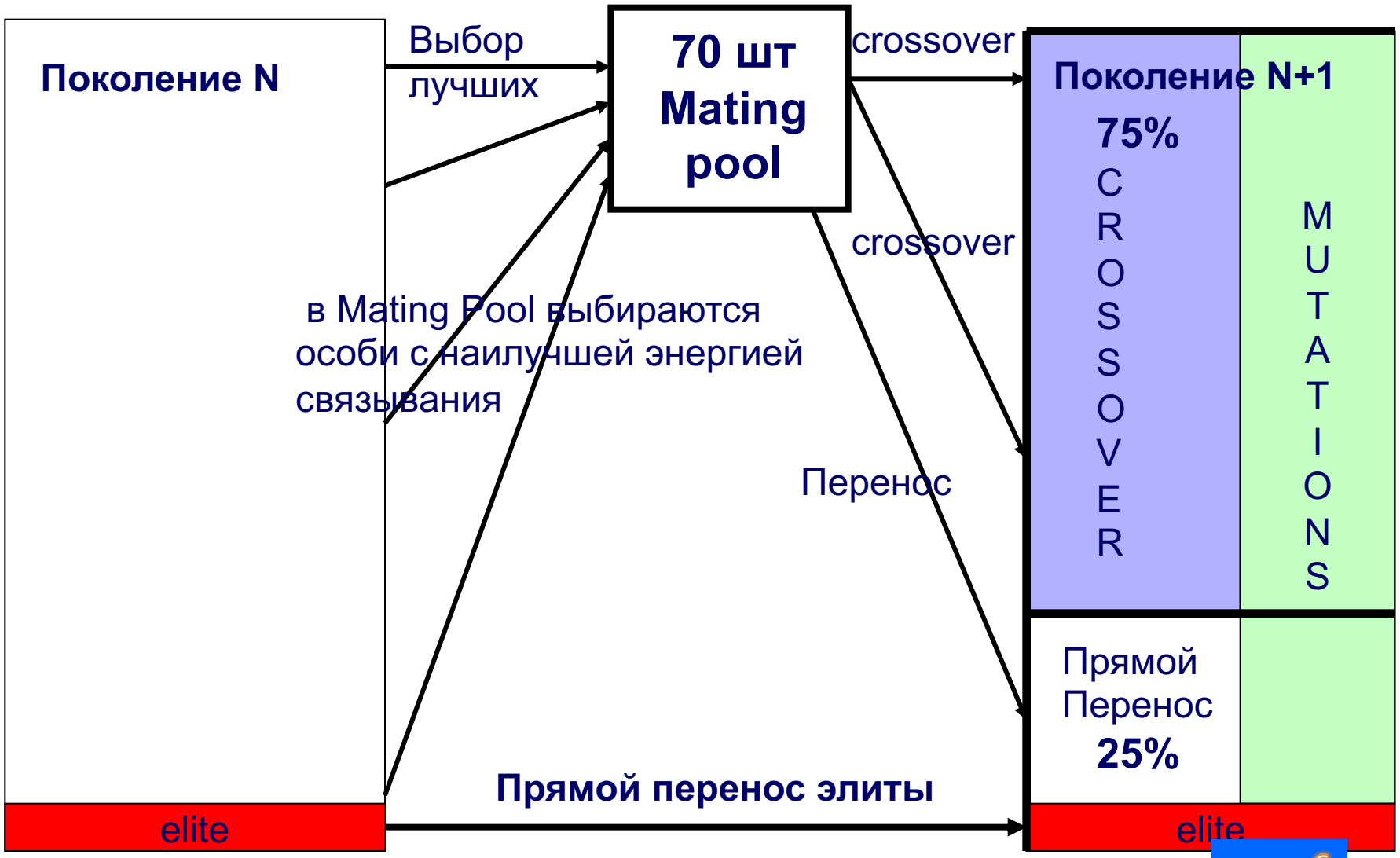
Особь – положение лиганда в активном центре белка



Хромосома

a_1
a_2
a_3
a_4
a_5
a_6
a_7
a_8
a_9
a_{10}

Генетический алгоритм глобальной оптимизации



Поколение N → Поколение N+1

- **Прямой перенос элиты**
- **Отбор в Mating pool: 70 лучших по энергии особей с учетом нишинга**
- **Случайно выбранная в Mating pool пара особей с помощью crossover'а создает 1 особь в новом поколении – 75% всего нового поколения**
- **25% нового поколения – прямой перенос случайно выбранной в Mating pool особи**
- **Все особи нового поколения подвергаются мутациям**

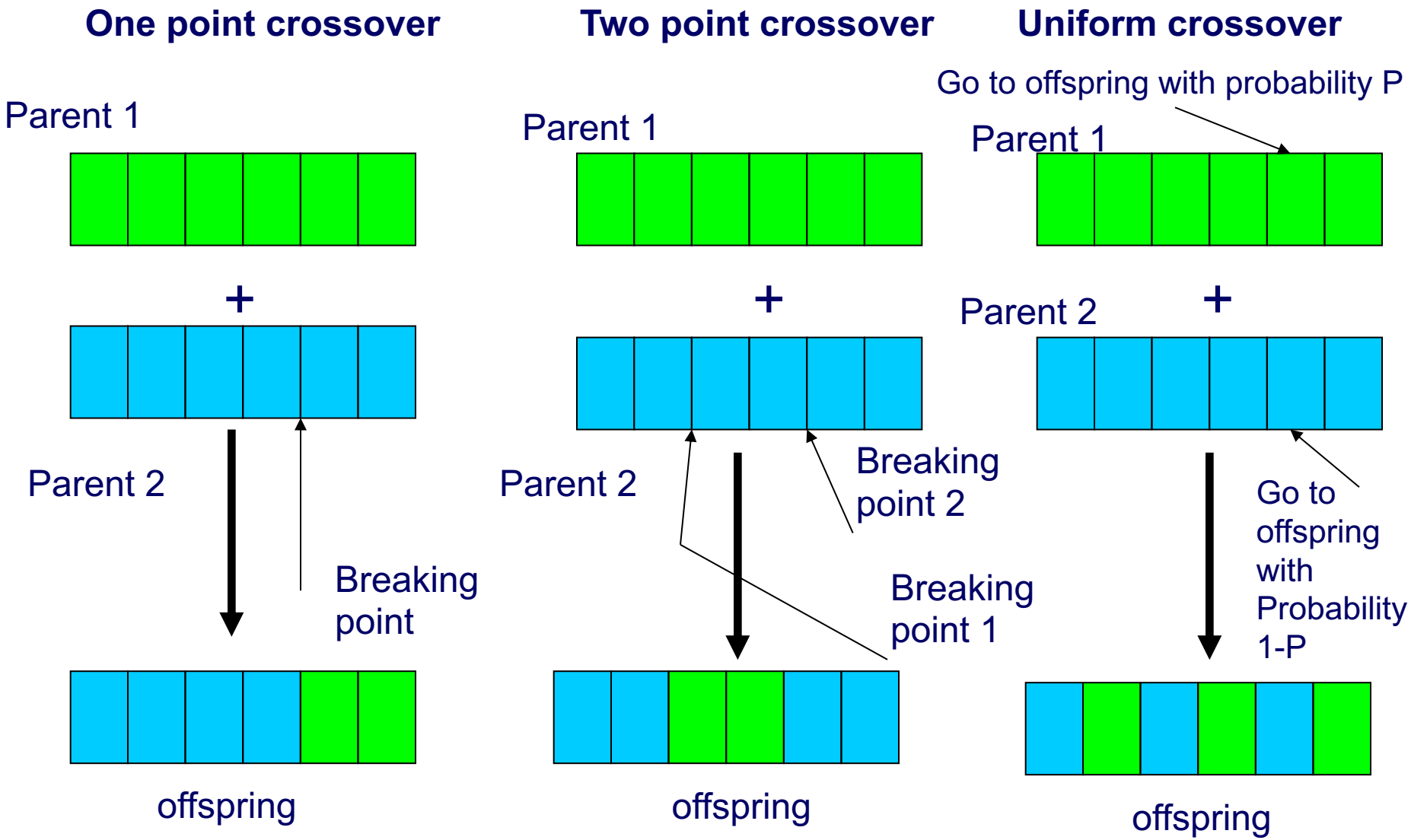
Нишинг (Niching)

Увеличивает популяционное разнообразие и предотвращает преждевременную сходимость генетического алгоритма

После выбора очередной особи в Mating Pool вычисляется «расстояние» **D** между ее генотипом и генотипом оставшихся особей. К энергии оставшихся особей добавляется положительный **штраф ≈ 1/D**, т.е. чем **ближе** особи друг к другу, тем **больше штраф**

$$D = \sqrt{\sum_{i=1}^{N} \left(b_i - a_i \right)^2}$$

CROSSOVER


One point crossover
Two point crossover
Uniform crossover
Parent 1
Parent 2
Breaking point
offspring
Parent 1
Parent 2
Breaking point 2
Breaking point 1
offspring
Go to offspring with probability P
Parent 1
Parent 2
Go to offspring with Probability 1-P
offspring

Выбор параметров докинга

- **Генетический алгоритм докинга – Вероятностный процесс**
  - **Число попыток поиска минимума: 50**
  - **Размер популяции: 30000 – число начальных позиций**
  - **Размер Mating Pool: 70**
  - **Число поколений: 500-1500**

- **Объем памяти под все переменные: 512 MB**

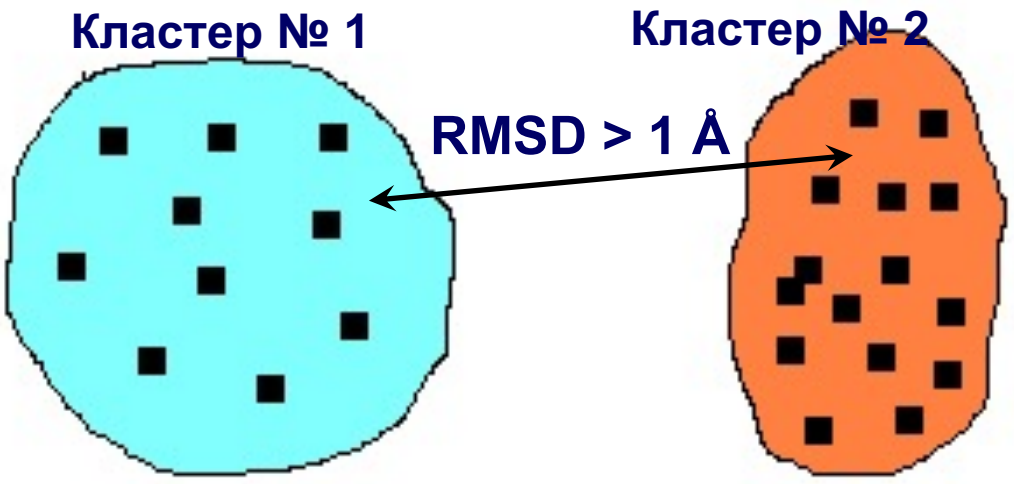
Решения задачи докинга

- Приготовление сетки потенциалов, имитирующих белок-мишень – 3-5 часов – однократно.
- Докинг: на каждый лиганд надо 1-5 часов (50 запусков GA)
- Поиск наиболее энергетически выгодных поз лиганда: энергия лиганда в поле белка + энергия внутренних напряжений лиганда
- Решения задачи докинга: позы лиганда в белке
- 50 независимых запусков Генетического Алгоритма дают 50 решений - 50 лучших поз лиганда: энергия наиболее отрицательная
- Совпадение решений - кластеризация

Кластеризация решений : 50 решений после 50 независимых запусков генетического алгоритма

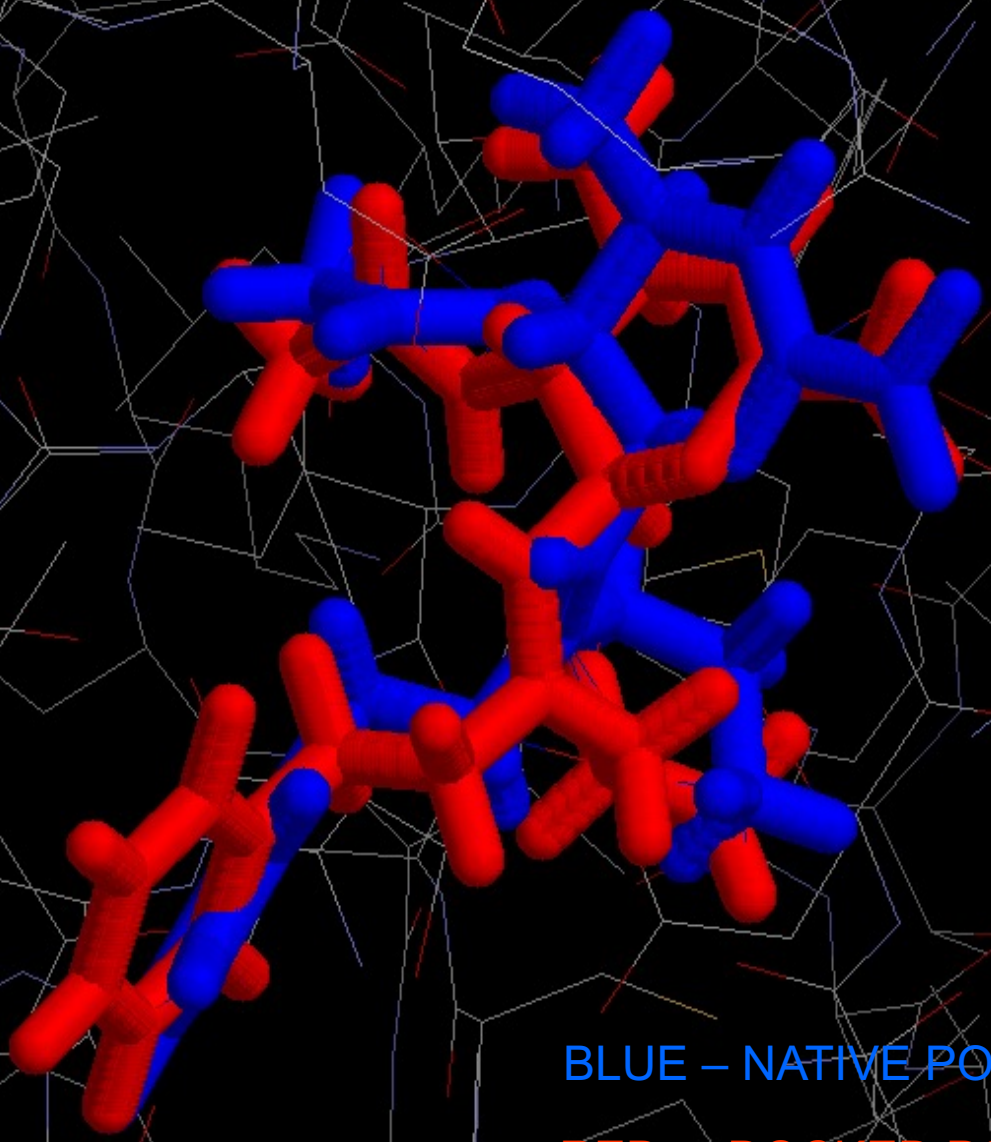
- Вычисление средне-квадратичного отклонения (RMSD) решений – поз лиганда, друг от друга
- При вычислении RMSD берутся разности декартовых координат одних и тех же атомов лиганда в двух позах
- Все решения разбиваются на группы (кластеры): внутри одного кластера RMSD < 1 Å между любыми двумя позами
- Кластеры нумеруются по возрастанию энергии лучшей позы лиганда в кластере: кластер №1 содержит позу с самой низкой энергией

Если в кластер №1 попадает большой процент из 50 решений, то задача глобальной оптимизации – **задача докинга**, решена с высокой надежностью



DOCKING RESULT vs. PDB POSITION (SIALIDASE B, PDB:1A4Q

RMSD:1.6Å



BLUE – NATIVE POSITION FROM PDB

RED – DOCKED POSITION

Валидация системы разработки ингибиторов урокиназы

- **Докинг 20 нативных соединений из базы данных PDB (http://www.rcsb.org)**

95% - RMSD < 2Å (хорошее качество позиционирования)

- **Докинг известных ингибиторов: из 25:**

  выбрана граница скора, разделяющая ингибиторы и не-ингибиторы.

- **Из 2000 неактивных + 8 активных соединений ранжирование по скору выявило все 8 реальных ингибиторов**

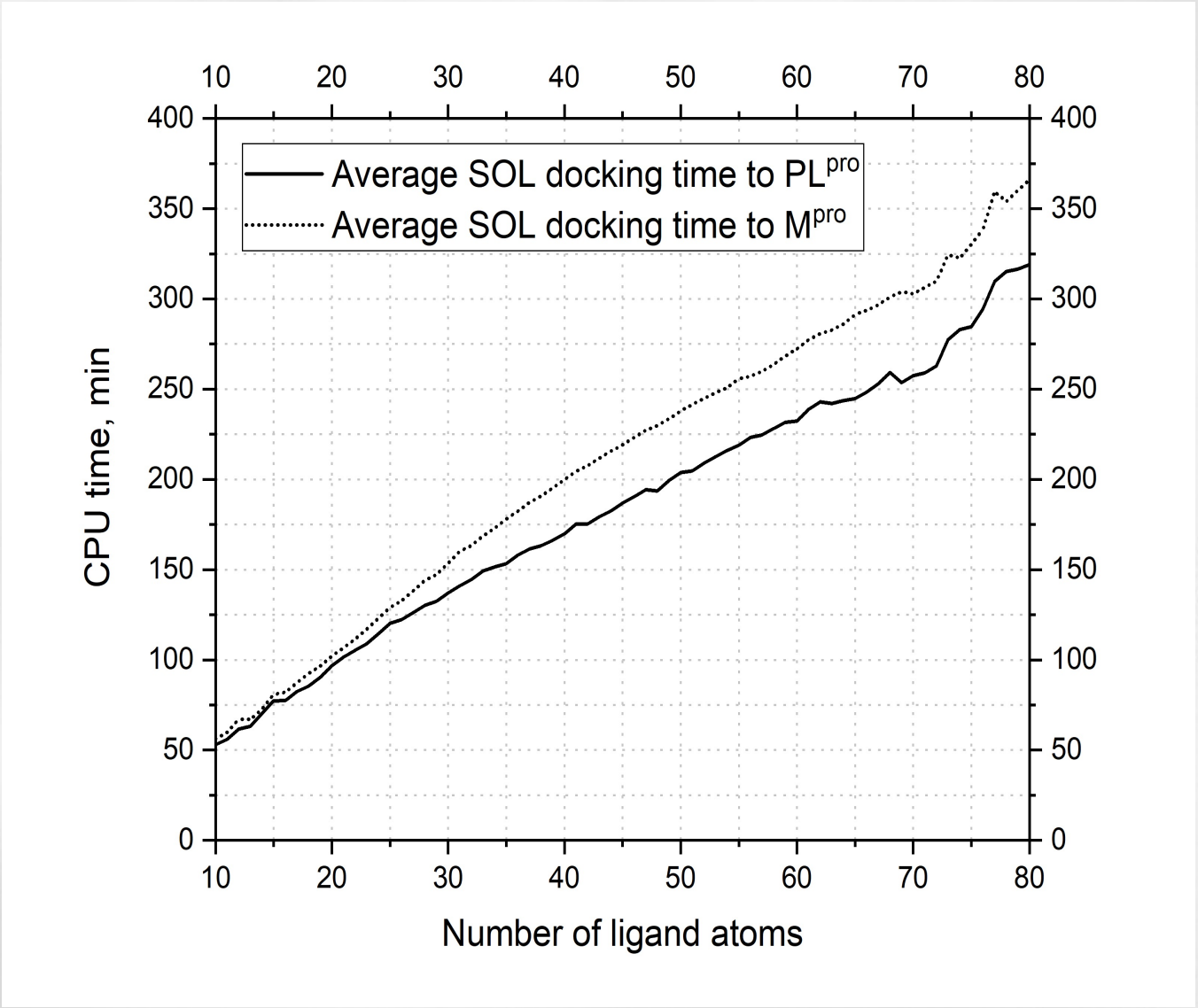
Кривая обогащения для урокиназы

2000 неактивных лигандов

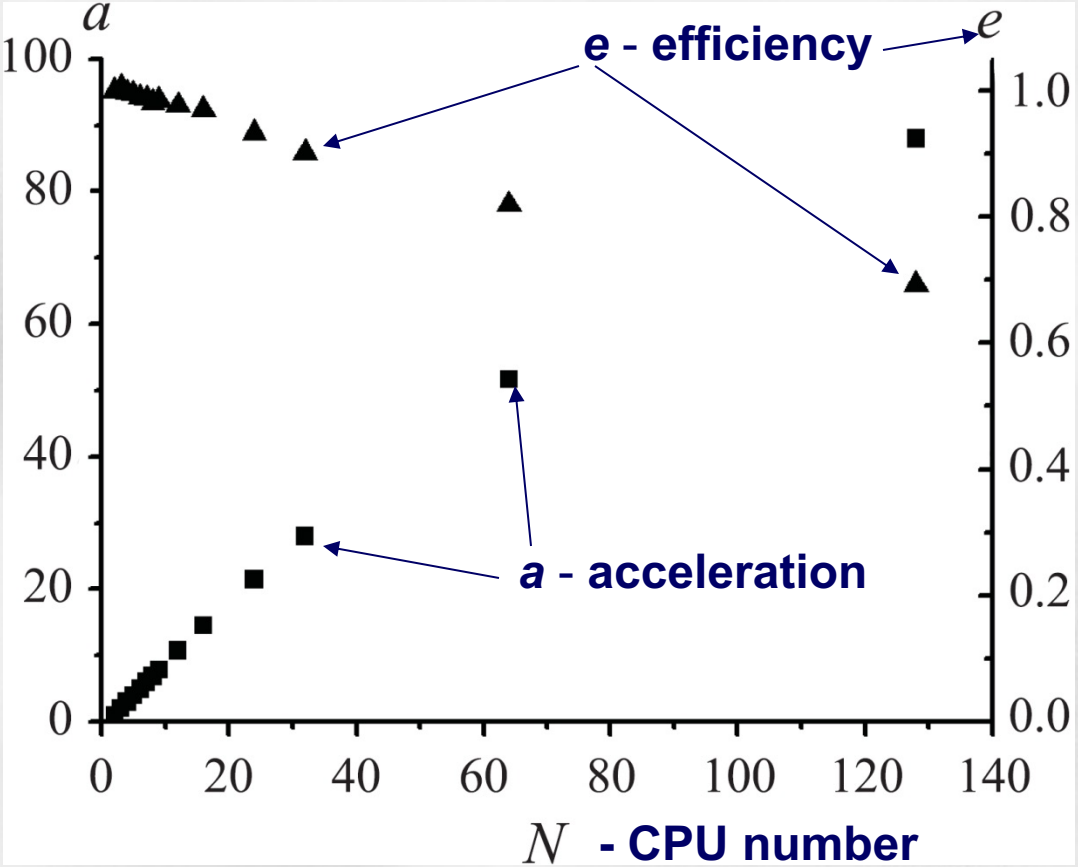
8 известных ингибиторов

Площадь под кривой обогащения = 0,98

Время CPU SOL на одном ядре зависит от числа атомов лиганда



Parallel docking grid generation SOLGRID



▲ -Efficiency: $e(N) = a(N)/N$

■ - Acceleration: $a(N) = t_1/t_N$

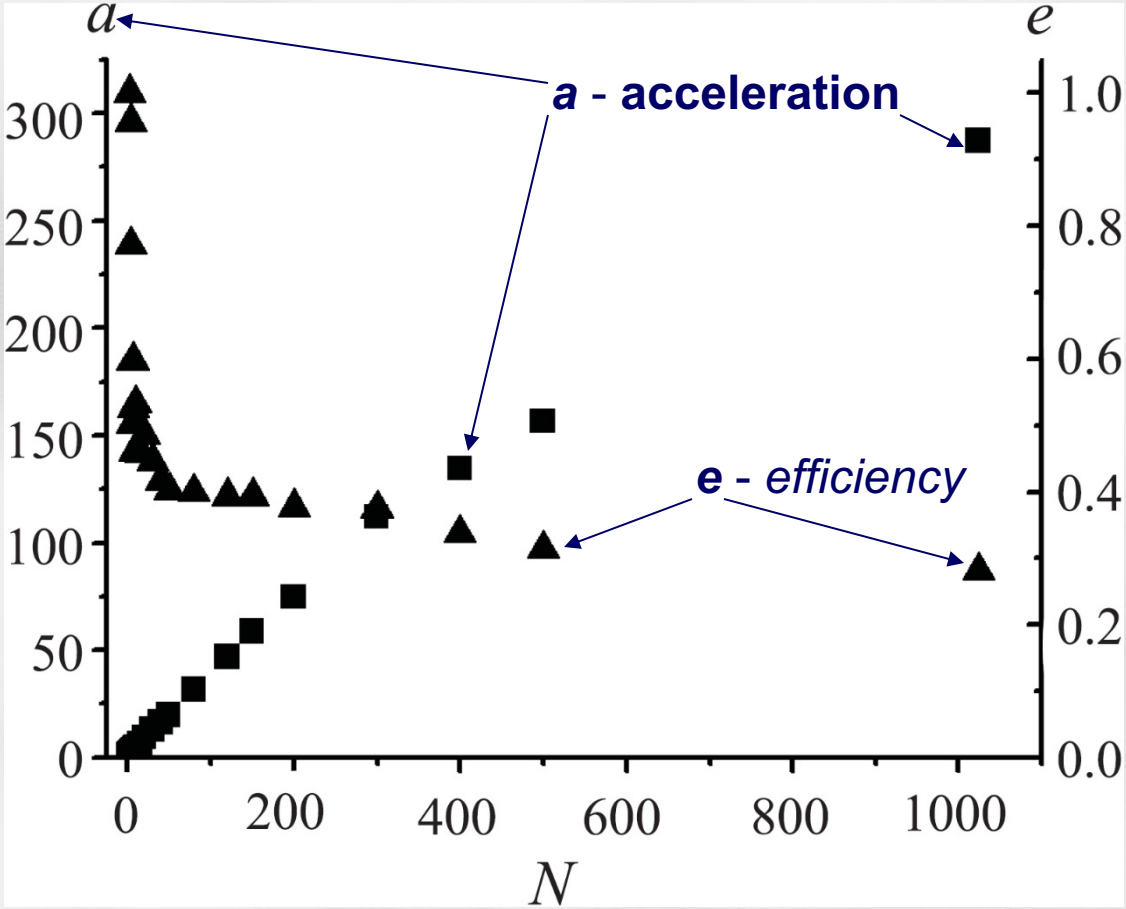
t_1 – 1 CPU

t_N – N CPU

1 minute the grid generation

Parallel ligand docking SOL

Cluster-type supercomputer



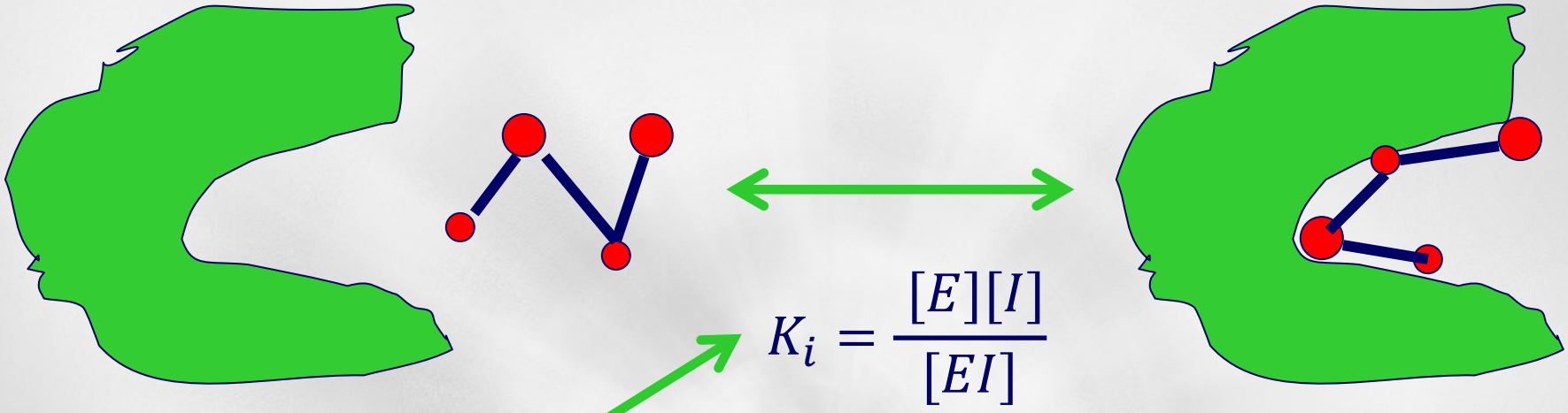
▲ -Efficiency: $e(N) = a(N)/N$ ■ - Acceleration: $a(N) = t_1/t_N$

t_1 – 1 CPU

t_N – N CPU

Docking of Large Databases – serial program

Задача: найти ингибитор с низкой Ki



$$K_i = \frac{[E][I]}{[EI]}$$

константа ингибирования

свободная энергия связывания белка с ингибитором

$$RT * Ln(K_i) = \Delta G = G_{EI} - G_E - G_I$$

газовая постоянная (8.31 Дж/К*моль)

температура (310 К)

энтальпия

энтропия

$$G = H - TS$$

ΔG = -10 ккал/моль лучше, чем ΔG = -5 ккал/моль

Программы докинга стали популярны. Десятки программ докинга востребованы при разработке лекарств. Каждая программа – уникальное сочетание моделей и приближений, скорости

Возможно ли увеличить точность докинга?

- **Точность позиционирования – *удовлетворительна***
- **Точность расчетов энергии связывания белок-лиганд ΔG_{bind} – *плохая***

Точность докинга → **Эффективность разработки лекарств**

Обзор методов и программ докинга:
Vladimir B. Sulimov, et al. Docking Paradigm in Drug Design, Current Topics in Medicinal Chemistry, 2021, 21, 507-546

Задача обобщенного докинга: увеличить точность

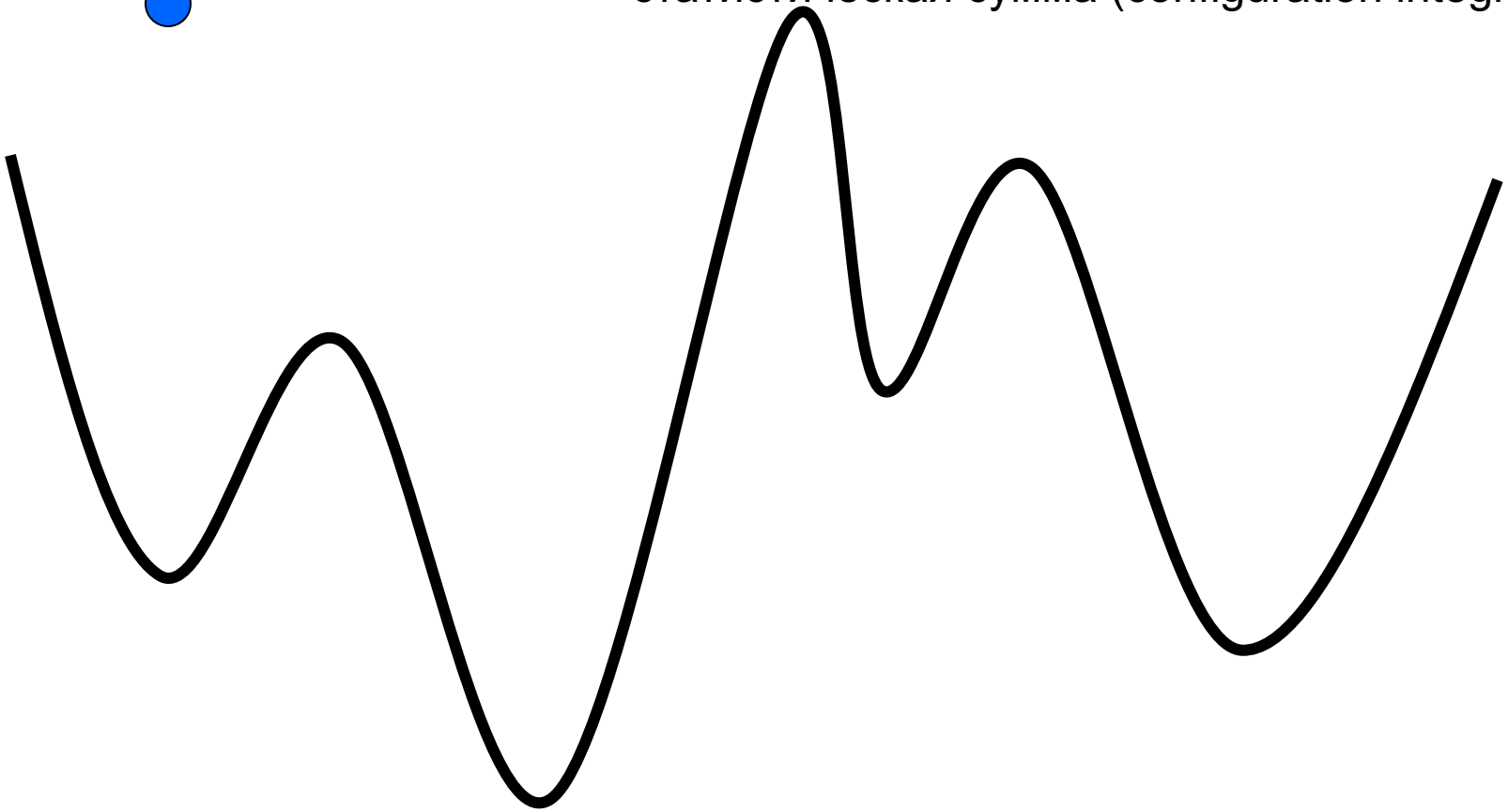
- **Гибкий лиганд + Гибкий белок**: Подвижность атомов белка и лиганда учитывается в процессе докинга одновременно и на одинаковой основе
- *Генетический алгоритм (один из самых популярных алгоритмов докинга) не справляется с глобальной оптимизации при размерности энергетической поверхности* **> 20 – нужен новый алгоритм докинга**
- Энергия комплекса белок-лиганд вычисляется в рамках силового поля **MMFF94** без упрощений
- Нет заранее вычисляемой сетки потенциалов взаимодействия атомов лиганда с белком; можно точно учитывать растворитель
- Докинг осуществляется путем поиска целого спектра низкоэнергетических минимумов системы белок-лиганд
- Точности позиционирования определяет точность расчета энергии связывания белок-лиганд

Жёсткий белок → **Белок с подвижными атомами**

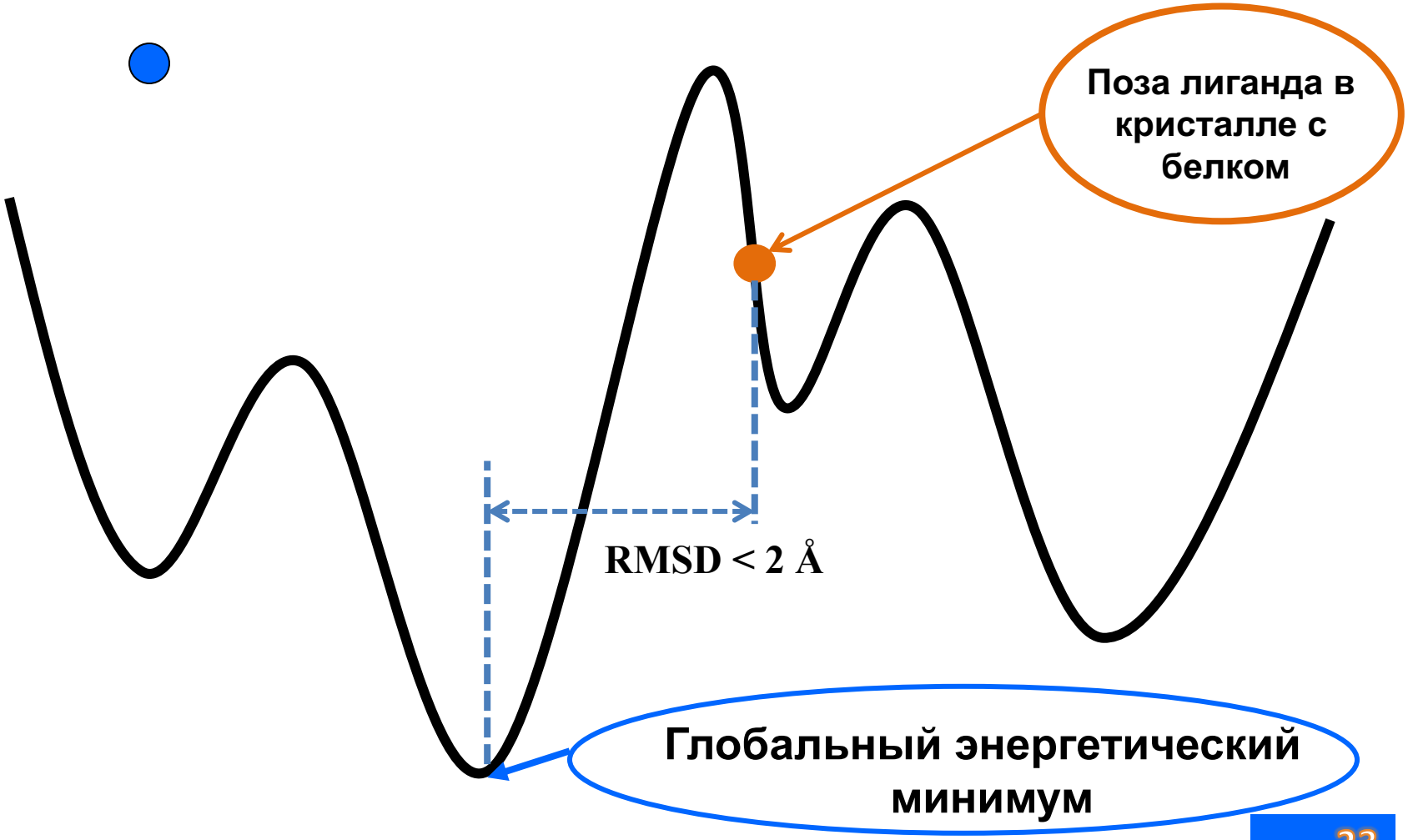
Движение лиганда в белке-мишени

$$G = -RT * Ln(Z) \qquad Z = \frac{1}{(2\pi\hbar)^{3n}} \int e^{-(U+W)/kT} \, dx_1 \ldots dx_{3n} dp_1 \ldots dp_{3n}$$

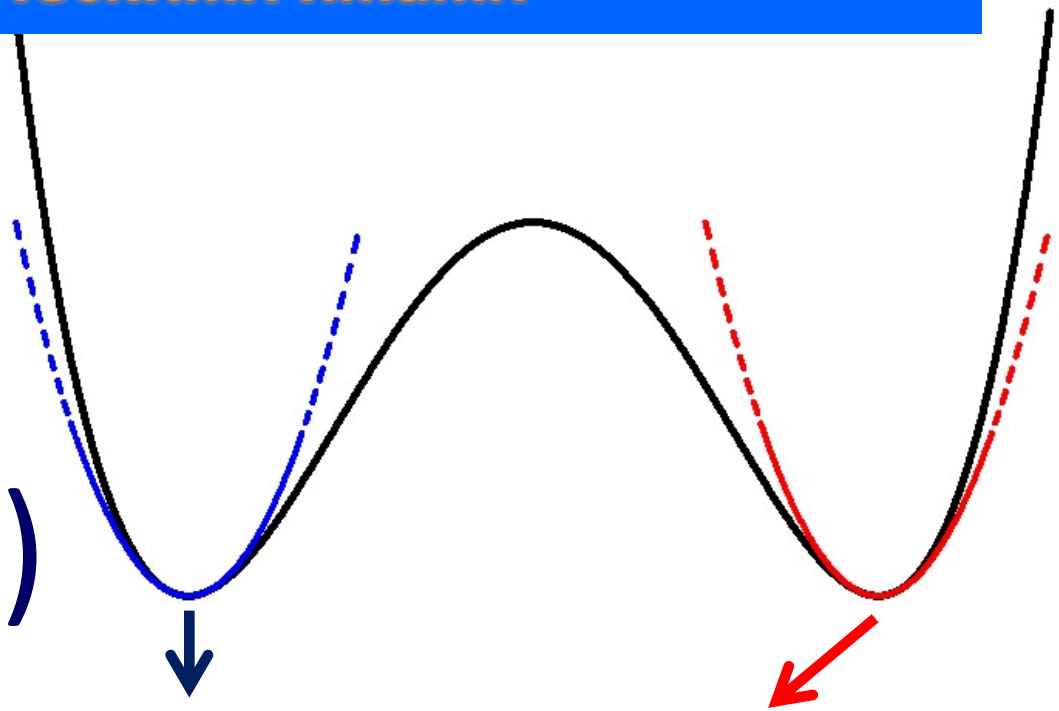
статистическая сумма (configuration integral)



На многомерной энергетической поверхности системы белок-лиганд очень много локальных минимумов



Аппроксимация U(x) независимыми гармоническими ямами



$$G = -kT*\ln(Z)$$

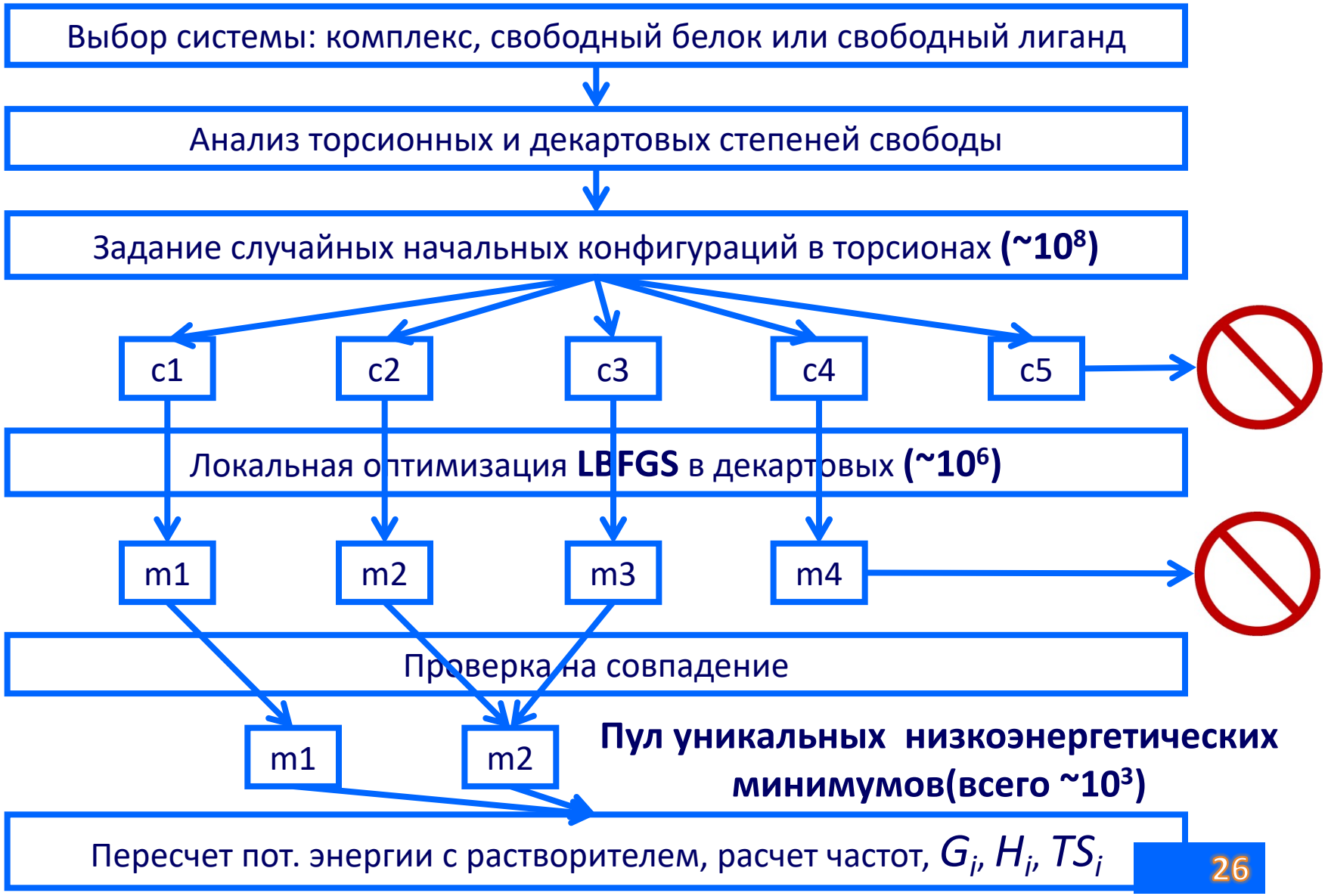
$$Z = Z_1 + Z_2$$

$$Z_i = e^{-\frac{E_0^{\,i}}{kT}} * \prod_j \frac{e^{-\hbar\omega_j^{\,i}/2kT}}{1-e^{-\hbar\omega_j^{\,i}/kT}}$$

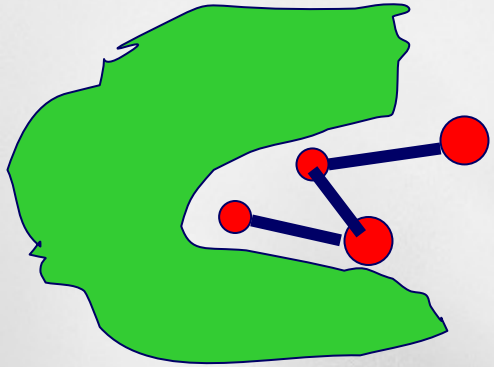
Программа FLM

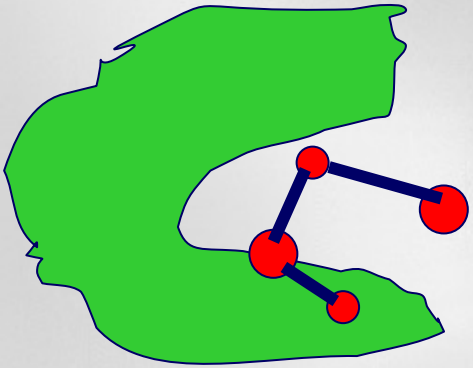
- FLM не использует заранее рассчитанную сетку потенциалов
- Жесткий белок, случайное «бросание» гибкого лиганда
- Локальная оптимизация энергии системы белок-лиганд из случайной позы лиганда: варьируются положения всех атомов лиганда
- В вакууме или с учетом растворителя в континуальной модели
- Силовое поле MMFF94, без упрощений, без подгоночных коэффициентов
- Исчерпывающий поиск спектра низкоэнергетических минимумов комплекса белок-лиганд
- Многие тысячи минимумов – отбор уникальных «на лету»
- Параллельные вычисления - докинг 1 лиганда: на **8191 ядрах** занимает несколько часов Ломоносова ≈10 000 CPU·часов

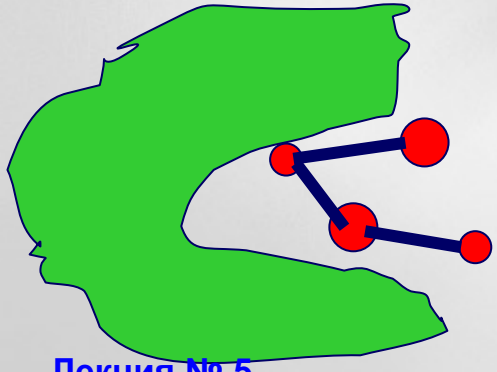
Программа FLM

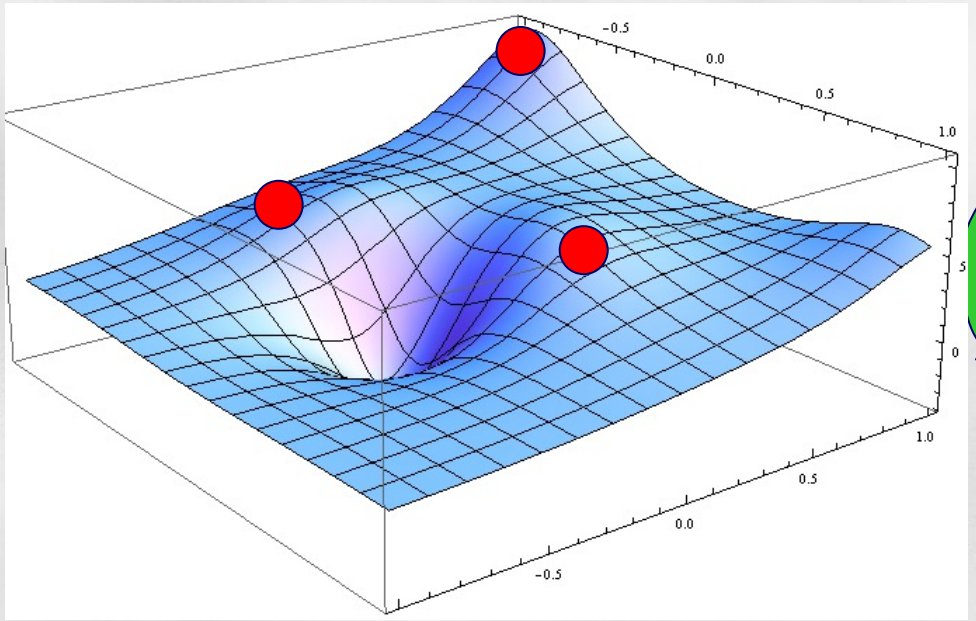


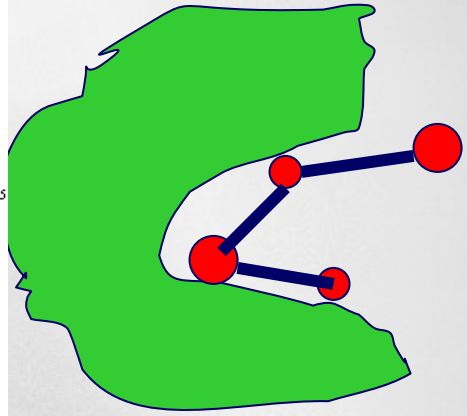
Программа FLM



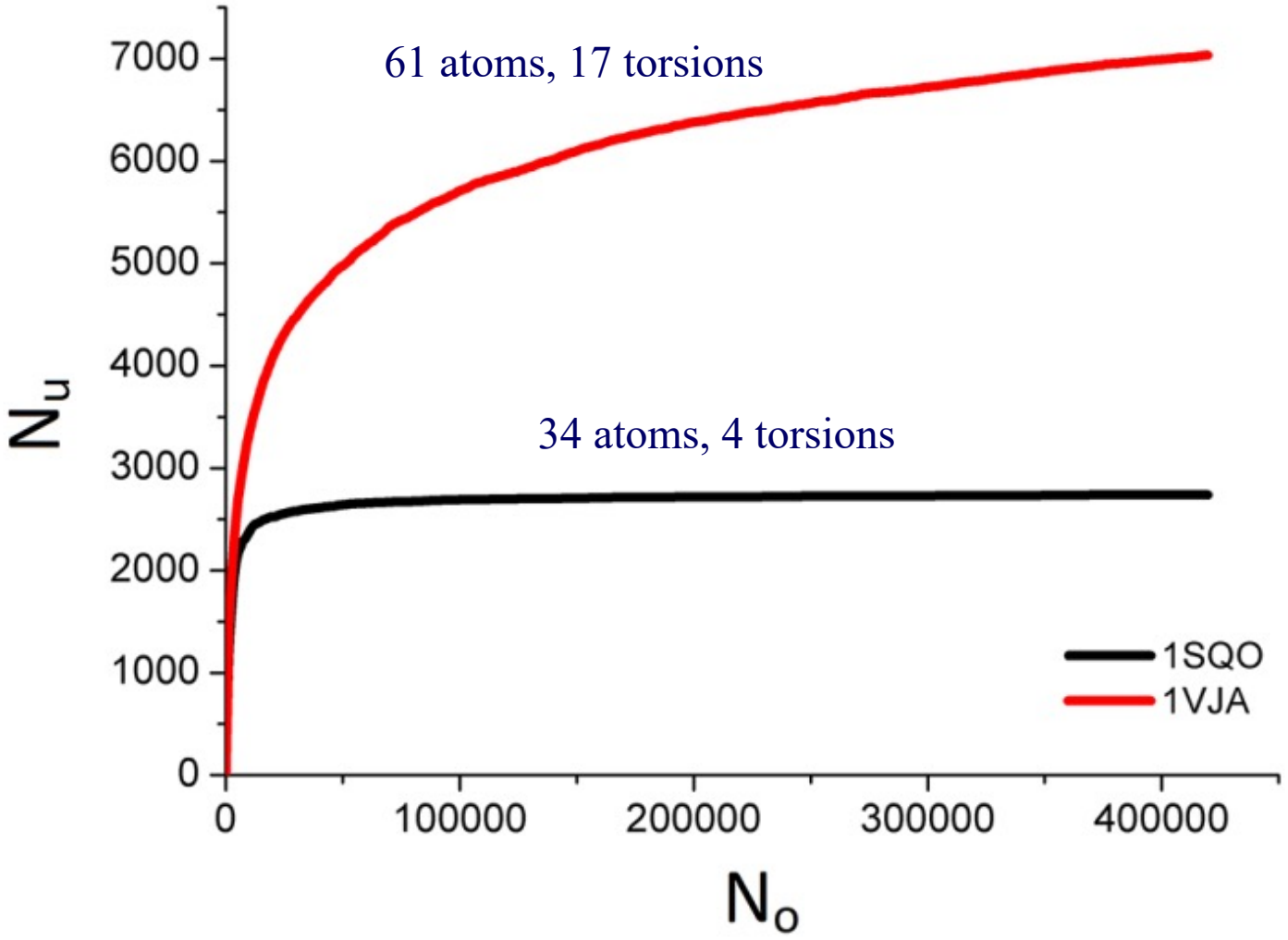




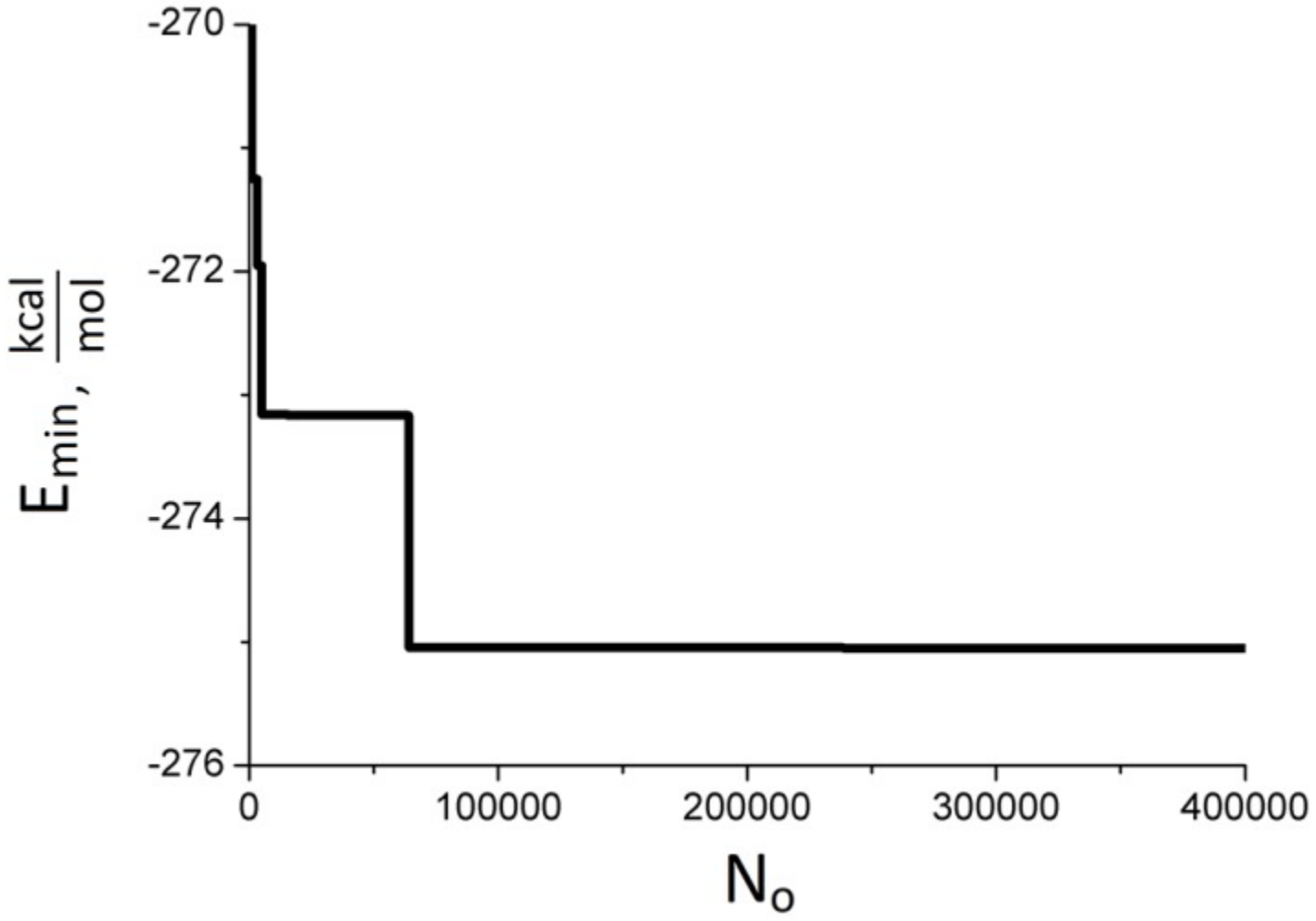




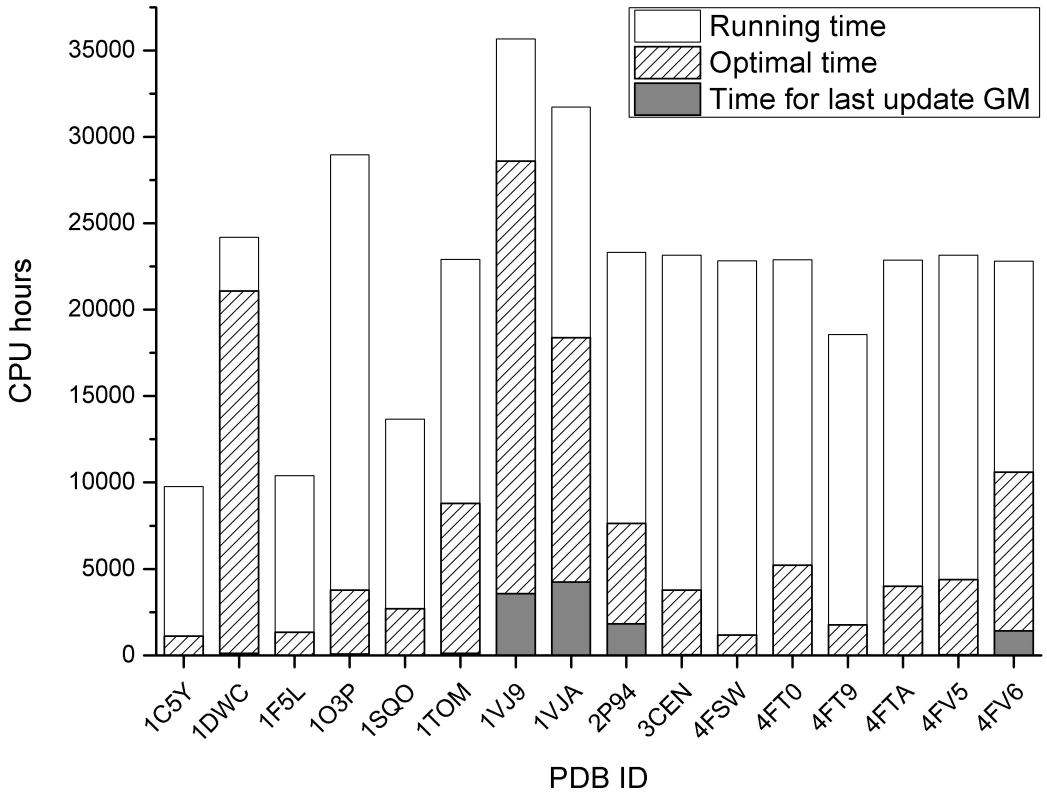
Nu – число обновлений пула уникальных низкоэнергетических минимумов
No – число оптимизаций



E_{min} – глобальный минимум энергии
N_o – число оптимизаций, комплекс 1VJA



Время счета FLM



- CPU зависит от числа атомов лиганда и от числа внутренних вращательных степеней свободы
- Глобальный минимум находится гораздо быстрее, чем весь пул низкоэнергетических минимумов

Докинг с помощью программы FLM

- FLM опирается на парадигму докинга: лиганд связывается с белком вблизи глобального минимума системы белок-лиганд
- FLM – это инструмент для тестирования применимости энергетических функций системы белок-лиганд для докинга
- FLM – это инструмент для тестирования эффективности алгоритмов глобальной оптимизации энергии системы белок-лиганд
- FLM нужно около 10 000 CPU*часов для докинга 1 лиганда
  - **FLM не может быть использована для скрининга больших баз данных – докинг только отдельных лигандов**
  - **FLM не может быть использована для докинга с учетом подвижности атомов белка**
  - **FLM для проведения квантово-химического докинга - квазидокинг**

Спасибо за внимание

- *…Surely every medicine is an innovation; and he that will not apply new remedies, must expect new evils…*
- *…Каждый медицинский метод есть инновация; а кто не хочет применять новые средства, должен ждать новых бед…*

Sir Francis Bacon (1561-1626)

OF INNOVATIONS