ОПТИМИЗАЦИЯ ВЫЧИСЛИТЕЛЬНОГО ПРОЦЕССА В DESKTOP GRID НА БАЗЕ BOINC ДЛЯ ВИРТУАЛЬНОГО СКРИНИНГА ЛЕКАРСТВ

ЕВГЕНИЙ ИВАШКО, НАТАЛИЯ НИКИТИНА

ИНСТИТУТ ПРИКЛАДНЫХ МАТЕМАТИЧЕСКИХ ИССЛЕДОВАНИЙ

КАРЕЛЬСКОГО НАУЧНОГО ЦЕНТРА РАН, ПЕТРОЗАВОДСК

Desktop Grid на базе BOINC

Мировой потенциал Desktop Grid оценивается в сотни экзафлопс (D. Anderson, 2020, DOI: 10.1007/s10723-019-09497-9)

Rosetta@home: ~25 TeraFlops

- Исследования белков, дизайн белков
- Точно предсказали трехмерную молекулярную структуру шиповидного белка SARS-CoV-2 за несколько недель до ее описания с помощью криоэлектронной микроскопии

- Folding@home: ~128 PetaFlops
  - Важные результаты в области сворачивания белков (“фолдинга”)
- World Community Grid: ~1 PetaFlops
  - Важные результаты в лечении рака
  - Ряд открытий в области экологии
- LHC@home (ЦЕРН): ~31 TeraFlops
  - Исследования физики частиц в рамках БАК
- Einstein@Home: ~3.7 PetaFlops
  - Исследования в области астрофизики, открытия пульсаров

Системы типа Desktop Grid составляют существенную часть отрасли HPC/HTC

Desktop Grid на базе BOINC

BOINC (Berkeley Open Infrastructure for Network Computing): промежуточное программное обеспечение (middleware) с открытым исходным кодом, де-факто стандарт для организации Desktop Grid

Отдельные вычислительные проекты

Вычислительные ресурсы предоставляются сообществом на добровольной основе для решения научных вычислительноемких задач

“Зонтичные” проекты

Ряд научных проектов объединяются и совместно используют вычислительные ресурсы, предоставляемые сообществом на добровольной основе

Enterprise Desktop Grid

Вычислительные ресурсы предоставляются организацией или группой организаций для решения вычислительноемкой задачи

Desktop Grid на базе BOINC

- Низкая стоимость (используются доступные вычислительные ресурсы)
- Простота в разворачивании и использовании
- Высокая масштабируемость и производительность

- Отсутствие определенности во времени доступности вычислительных ресурсов
- Разнородность и ненадежность вычислительных ресурсов
- Относительно невысокая производительность отдельных вычислительных узлов
- Низкая пропускная способность каналов передачи данных

- Использование специализированных алгоритмов управления вычислительными ресурсами, оптимизированных и учитывающих свойства решаемых вычислительноемких задач, может кратно повысить пропускную способность, вычислительную мощность и масштабируемость Desktop Grid.

BOINC-проект SiDock@home

BOINC-проект SiDock@home: введение

Команда проекта SiDock@home:

- Marko Jukić и Črtomir Podlipnik
  - University of Ljublana, Slovenia
  - University of Maribor, Slovenia
- Natalia Nikitina и Evgeny Ivashko
  - Karelian Research Center of the Russian Academy of Sciences, Petrozavodsk, Russia
- Maxim Manzyuk
  - Internet portal BOINC.Ru, Moscow, Russia
- Ilya Kurochkin и Alexander Albertian
  - Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

BOINC-проект SiDock@home: введение

Разработка лекарства – времяемкий и ресурсоемкий процесс, занимающий до 12-17 лет.

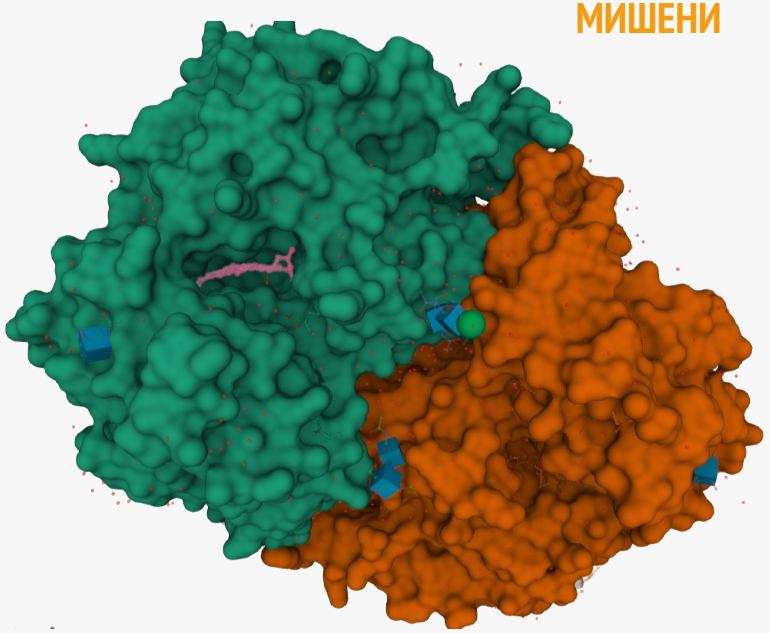


Рис. 1: пример мишени и лиганда. Мишень - циклооксигеназа-2, фермент, участвующий в воспалительном процессе в организме человека. Лиганд аспирин необратимо связывается с мишенью и препятствует развитию воспалительного процесса.

Мишень - биологическая макромолекула, нарушение функции которой приводит к заболеванию.

Лекарство - химическое соединение, способное взаимодействовать (связываться) со своей мишенью и влиять на протекание заболевания.

Лиганды – малые молекулы, потенциально способные стать лекарством.

Хиты - лиганды, которые оказались потенциально способны связываться с мишенью и влиять на протекание заболевания. В дальнейшем хиты подлежат проверке *in vitro* (в лаборатории).

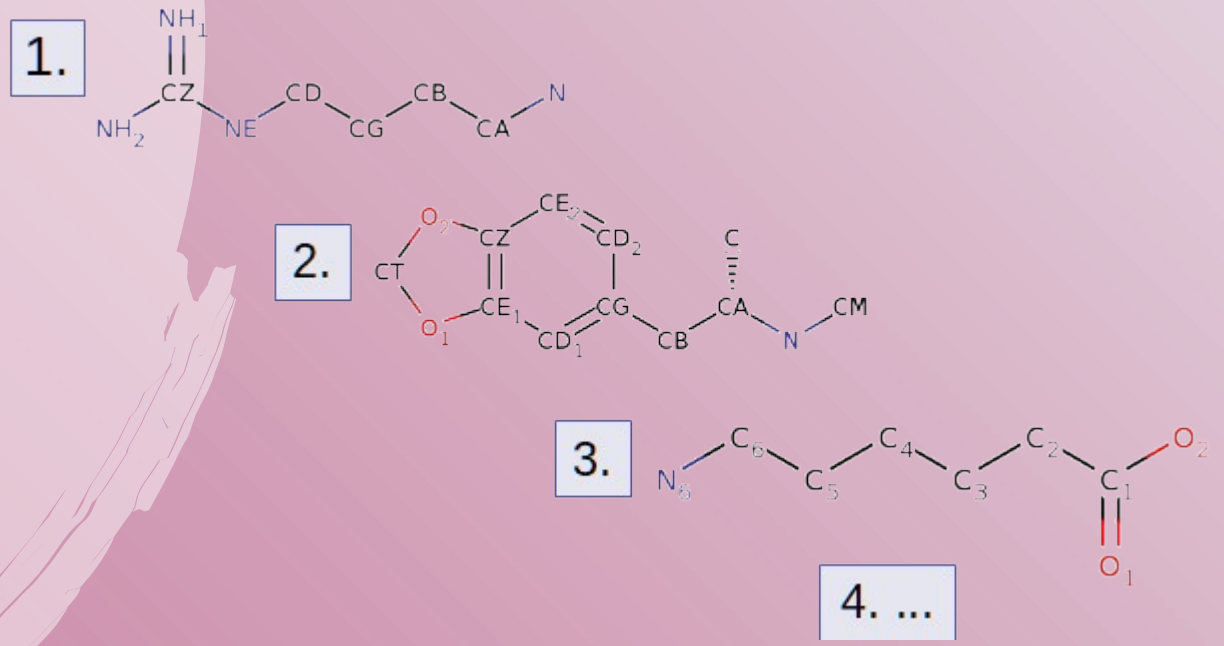
BOINC-проект SiDock@home: введение

Виртуальный скрининг – вычислительноемкая процедура, включающая в себя компьютерное моделирование связывания лиганда с мишенью (молекулярный докинг) на множестве независимых моделей молекул.

Библиотеки лигандов имеют большой объем (сотни миллионов, миллиарды молекул; потенциально ~10^{60})

Результат виртуального скрининга – список хитов, упорядоченных по предсказанной способности связывания с мишенью.

Цель виртуального скрининга - сократить время и стоимость начального этапа разработки лекарства.



BOINC-проект SiDock@home: введение

BOINC-проект SiDock@home направлен на разработку лекарств.

Проводится виртуальный скрининг библиотеки из ~1 миллиарда химических соединений.

Найденные хиты подлежат дальнейшему отбору, оптимизации и тестированию в лаборатории.

Вычислительные эксперименты в проекте:

- Март 2020 г. - сентябрь 2022 г.: виртуальный скрининг для 20 мишеней коронавируса SARS-CoV-2.
- Сентябрь 2022 г. и далее: продолжение виртуального скрининга для других мишеней различных вирусов.

Рис. 2: разнообразие потенциальных мишеней для терапевтического воздействия на коронавирус SARS-CoV-2.

BOINC-проект SiDock@home: введение

Статистика проекта (22 сентября 2022 г.):

- 9326 зарегистрированных пользователей
- 2462 активных пользователя
- 40419 зарегистрированных компьютеров
- 6126 активных компьютеров
- Производительность: 100 TeraFlops

- Проведен полный виртуальный скрининг библиотеки лигандов для 11 мишеней
- Ряд соединений протестированы в лаборатории
- Найдены новые микромолярные ингибиторы шиповидного белка SARS-CoV-2

BOINC-проект SiDock@home: введение

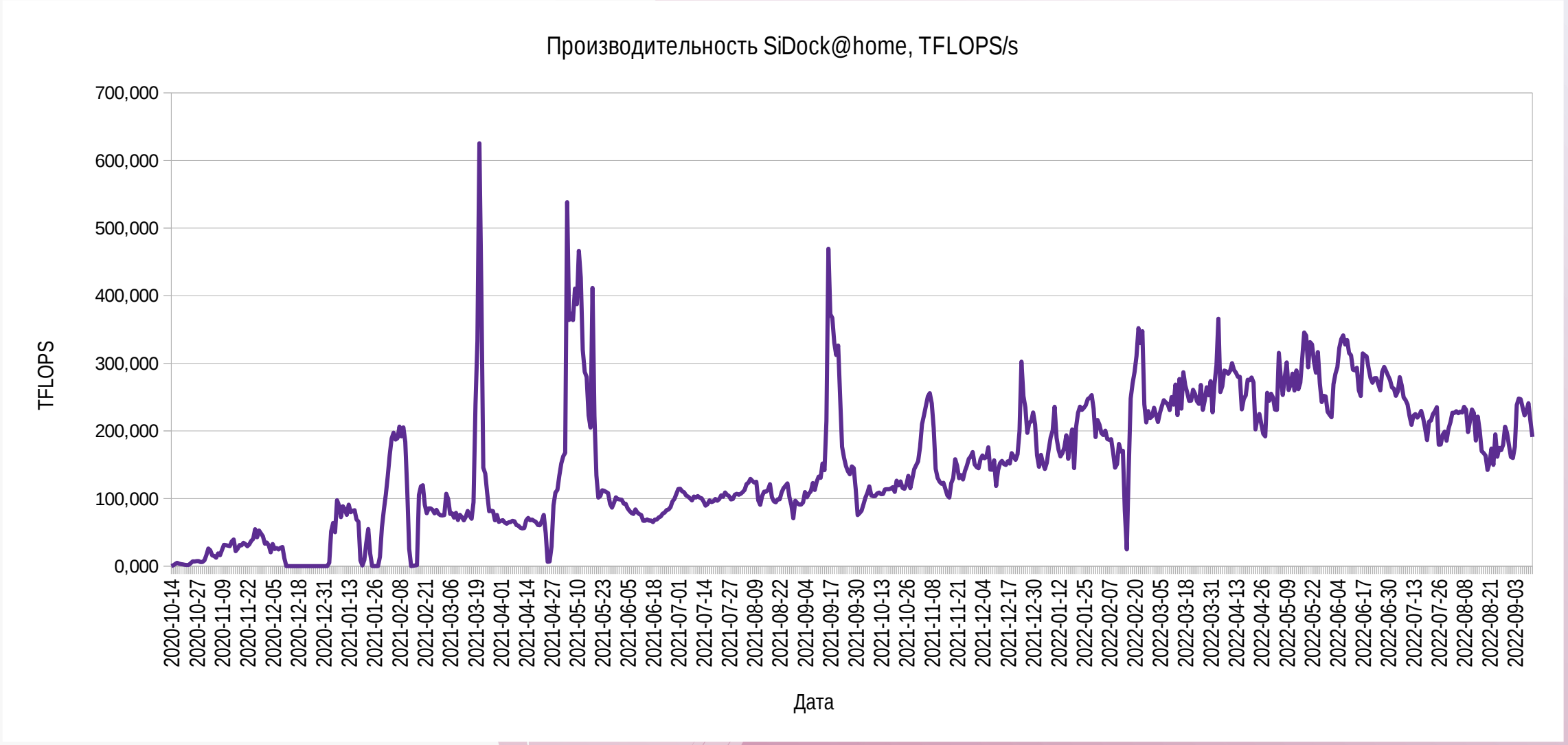


Рис. 3: динамика производительности проекта, TeraFlops. Пики производительности соответствуют соревнованиям, проводимым в сообществе BOINC. Снижения производительности соответствуют периодам технических работ и временным оттокам вычислительных мощностей в другие BOINC-проекты.

BOINC-проект SiDock@home: анализ вычислительного процесса

Вычислительный эксперимент: мишень + библиотека молекул + протокол виртуального скрининга

Первоначальная организация вычислительного процесса:

- 1,964,803 заданий по 500 лигандов
- Задание содержит фиксированный список лигандов
- Среднее время выполнения задания на настольном компьютере: ~1-2 часа
- Кворум Q = 2, валидация по формату выходных файлов, сравнение результатов по спискам лигандов
- За все валидные результаты начисляются BOINC-кредиты
- Итоговый результат задания выбирается среди полученных реплик, остальные отбрасываются

BOINC-проект SiDock@home: анализ вычислительного процесса

- ✓ Проведен виртуальный скрининг для 11 мишеней за 17 месяцев
- ✓ Получено множество хитов для тестирования в лаборатории и дальнейшей оптимизации
- ✓ В частично автоматизированном режиме выявлен и исправлен ряд технических ошибок на стороне BOINC-сервера, библиотеки лигандов и приложения для молекулярного докинга

BOINC-проект SiDock@home: анализ вычислительного процесса

✓ Проведен виртуальный скрининг для 19 мишеней за 23 месяца

✓ Получено множество хитов для тестирования в лаборатории и дальнейшей оптимизации

✓ В частично автоматизированном режиме выявлен и исправлен ряд технических ошибок на стороне BOINC-сервера, библиотеки лигандов и приложения для молекулярного докинга

✘ Более половины результатов отброшены из-за избыточности (требовались только для достижения кворума), позднего завершения (кворум был достигнут раньше их получения) или технических ошибок на стороне клиента

✘ Валидация сравнением двух результатов затруднительна из-за случайного характера моделирования

✘ В случае ошибки в расчете одного из 500 лигандов результат задания не засчитывался

✘ Лиганды, не ставшие хитами (не прошедшие фильтрацию), отбрасывались на стороне клиента

✘ В случае установки слишком высокого порога для фильтрации требовался пересчет всей библиотеки

✘ Время выполнения задания могло сильно меняться в зависимости от мишени и протокола

BOINC-проект SiDock@home: анализ вычислительного процесса

Основные механизмы оптимизации вычислительного процесса в BOINC:

- **Репликация и кворум**
  - Для сверки результатов от ненадежных вычислительных узлов.
    RakeSearch, Q = 2. По достижении кворума задание считается выполненным, если результаты совпали.
  - Для определения надежности вычислительных узлов.
    World Community Grid, Q = 2 (подпроект Help Conquer Cancer). Если задание досталось надежному узлу, то Q = 1. Иначе по достижении кворума Q = 2 результаты реплик сравниваются, и если они совпали, то ненадежный узел становится надежным.
  - Для разделения заданий на подзадания.
    World Community Grid, Q = 15 (подпроект Human Proteome Folding - Phase 2); Q = 10 (подпроект Nutritious Rice for the World). По достижении кворума задание считается выполненным, и результат складывается из 15(10) частей.
- **Дедлайн**
- Для выявления узлов, вышедших из состава Desktop Grid.
- Для определения приоритетов заданий (быстрее получить результаты с высоким приоритетом).

BOINC-проект SiDock@home: оптимизация вычислительного процесса

Новая организация вычислительного процесса в SiDock@home:

- Таблица **<target_id, ligand_id, score, task_id>**

  score – оценка энергии связывания лиганда **ligand_id** с мишенью **target_id**, найденная в задании **task_id**
- Задание генерируется динамически и содержит список лигандов
- Результат содержит оценки связывания каждого лиганда с мишенью
- Кворум Q = 1, валидация результата по списку идентификаторов лигандов

✓ Не требуется избыточных вычислений для достижения кворума

✓ Сохраняются результаты молекулярного докинга по всей библиотеке лигандов

✓ Можно применять различные варианты отбора хитов среди результатов

✓ Можно генерировать задания требуемого размера или по избранным спискам лигандов

✓ Валидация по уникальному списку лигандов в конкретном задании

BOINC-проект SiDock@home: оптимизация вычислительного процесса

Реализация вычислительного процесса:

- Новые таблицы в БД проекта: сводная по результатам молекулярного докинга и ряд вспомогательных.
- Дополнения в программном коде валидатора и ассимилятора результатов.
- На стороне клиента изменений не требуется.
- Возможна реализация адаптивной репликации для ненадежных клиентов с любым кворумом.
- Выходной файл содержит результаты докинга всех лигандов задания. Поэтому требуется балансировка нагрузки на сервер проекта: генерация заданий оптимального размера и/или задействование дополнительного выделенного файлового сервера.

- К настоящему моменту предложенные изменения частично реализованы и апробируются в проекте.

СПАСИБО ЗА ВНИМАНИЕ!

Евгений Евгеньевич Ивашко
ivashko@krc.Karelia.ru