



МГУ имени М.В.Ломоносова

Установление взаимосвязи между структурой и функциональными свойствами белков/ферментов с использованием методов компьютерной биологии и разработка предсказательных алгоритмов

Суплатов Дмитрий Андреевич

к.х.н., с.н.с. НИИ ФХБ имени А.Н.Белозерского МГУ

Воеводин Владимир Валентинович

д.ф.-м.н., чл.-корр. РАН, проф. ВМК МГУ, зам. директора НИВЦ МГУ

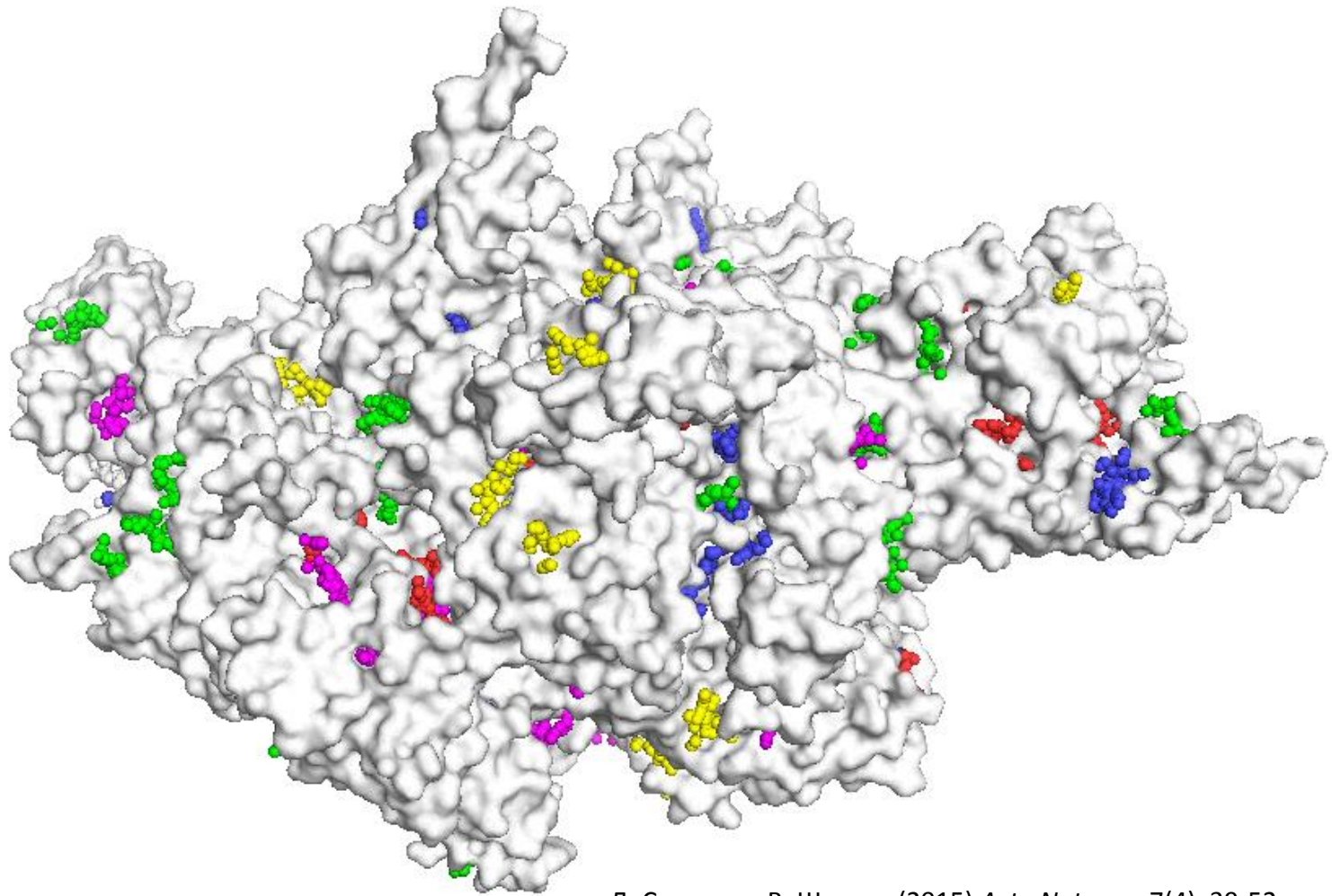
Швядас Витас

д.х.н., проф. ФББ МГУ, гл.н.с. НИИ ФХБ имени А.Н.Белозерского МГУ

Семинар “Суперкомпьютерные технологии в науке, образовании и промышленности”

4 июня 2019 г., Москва

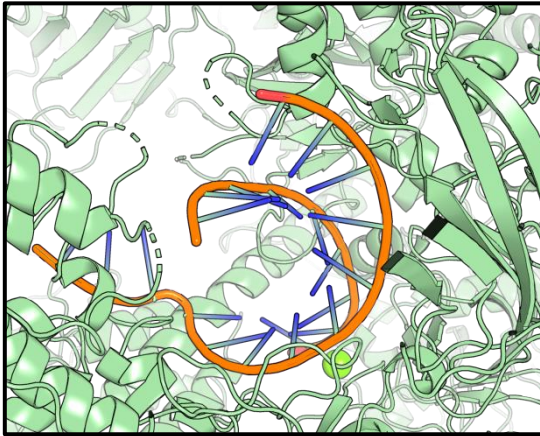
Структура ДНК-зависимой РНК-полимеразы



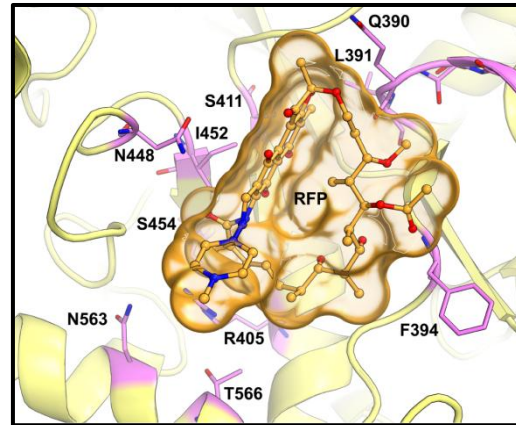
Д. Суплатов, В. Швядас (2015) *Acta Naturae*, 7(4), 39-52.

Поверхность этого крупного полисубъединичного фермента покрыта большим количеством полостей – потенциальных сайтов связывания

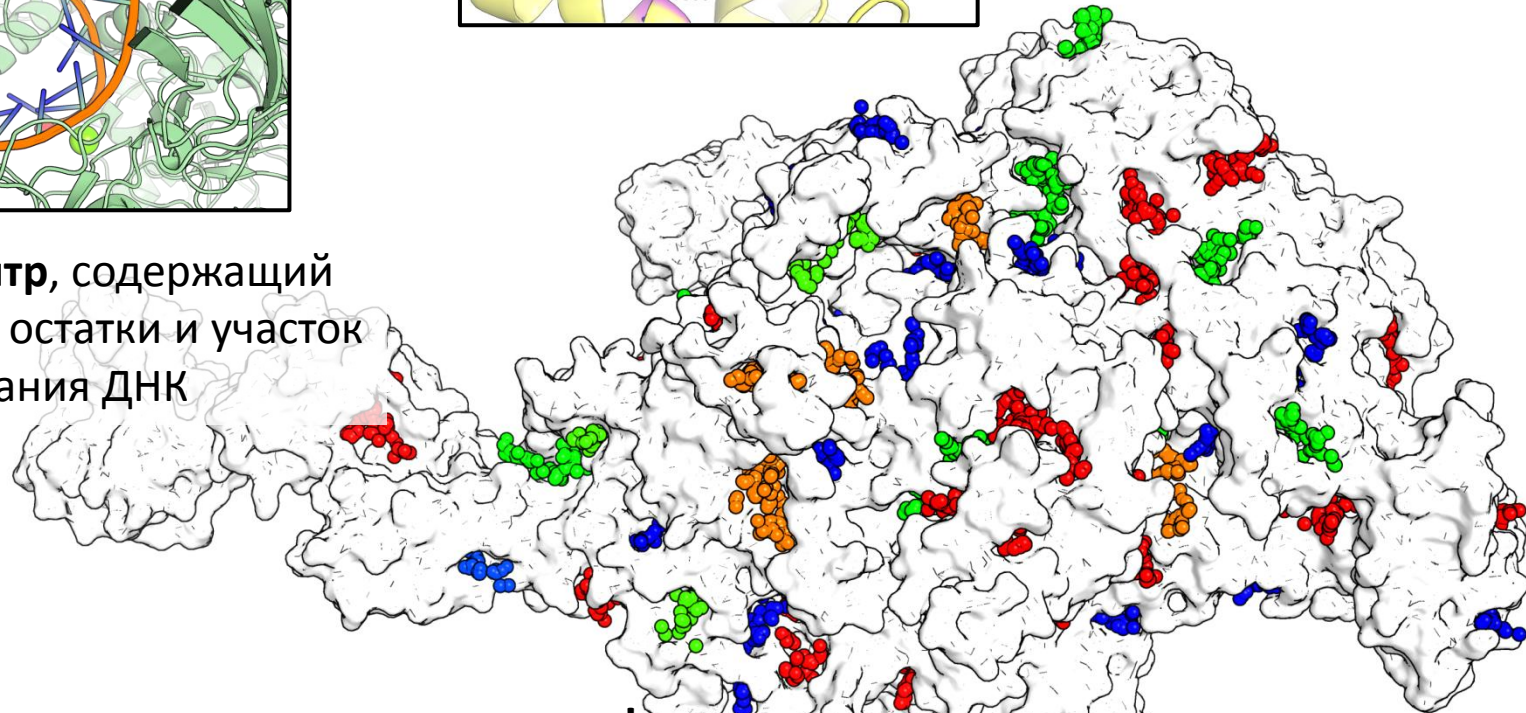
Структура ДНК-зависимой РНК-полимеразы



Активный центр, содержащий
каталитические остатки и участок
связывания ДНК



Несколько известных
аллостерических сайтов,
способных связывать
регуляторные лиганды

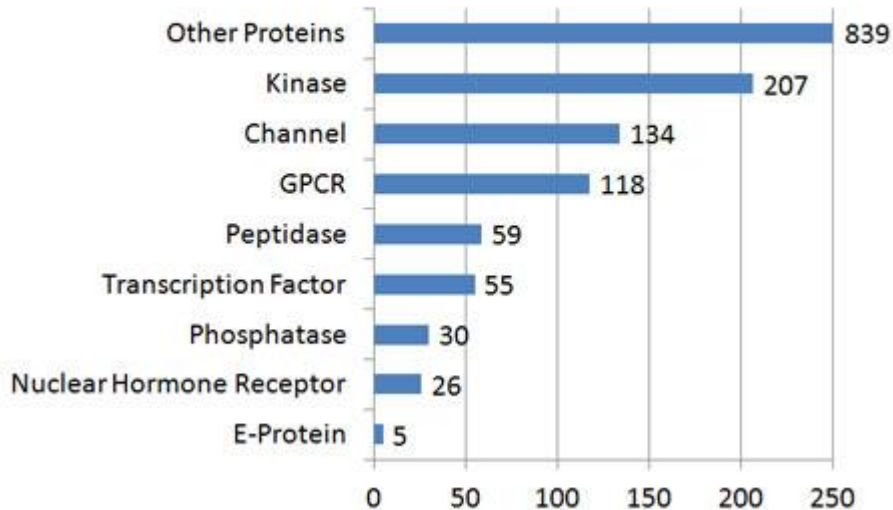


Роль большинства других центров неизвестна!

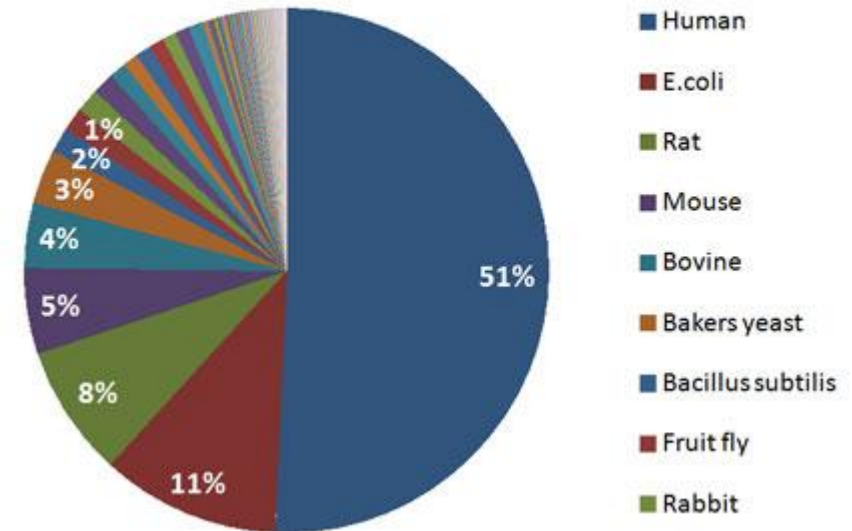
Д. Суплатов, В. Швядас (2015) *Acta Naturae*, 7(4), 39-52.

Аллостерические белки

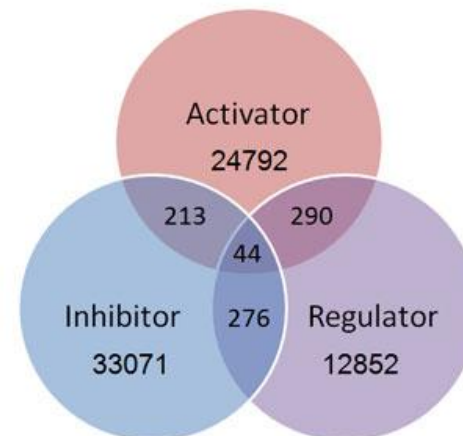
ASD Macromolecules in ASD v3.0



Allosteric Targets — Species Distribution



Allosteric Modulators — Category Distribution



Allosteric Target Entries	1473
Allosteric Sites	1930
Allosteric Modulator Entries	71538

Явление мунлайтинга

Наличие нескольких (вплоть до двух десятков!) функциональных активностей в рамках одной белковой структуры

- MoonProt (<http://moonlightingproteins.org/>):
 - 399 записей (347 записей Uniprot)
- MoonDB (<http://tagc.univ-mrs.fr/MoonDB/>):
 - 469 записей
- Multitasking Proteins DataBase (<http://wallace.uab.es/multitaskII/>)
 - 694 записи

Изучение структурно-функциональных взаимосвязей в белках/ферментах

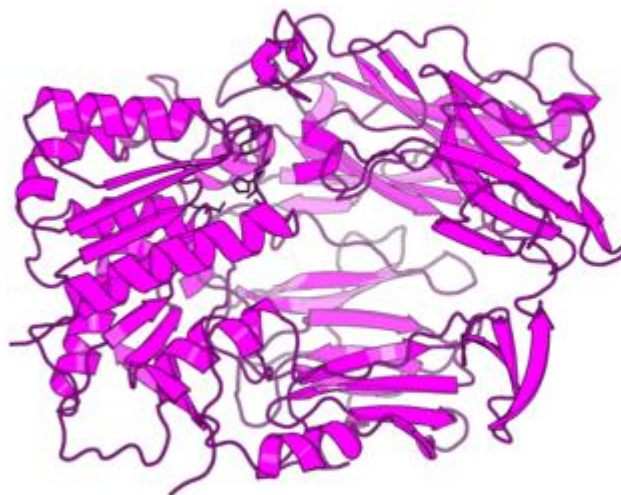
- Как найти все центры в структуре белка и изучить их функциональную/регуляторную роль?

Липаза и амидаза суперсемейства α/β -гидролаз

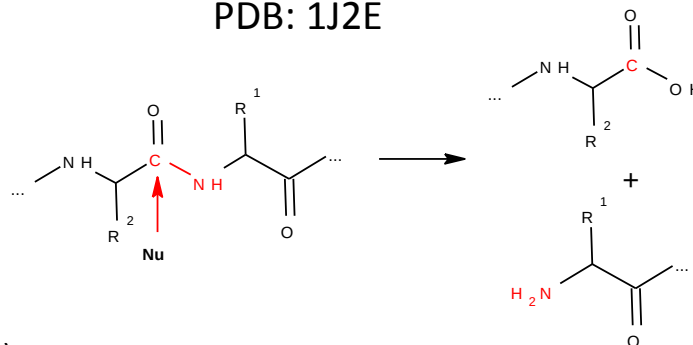
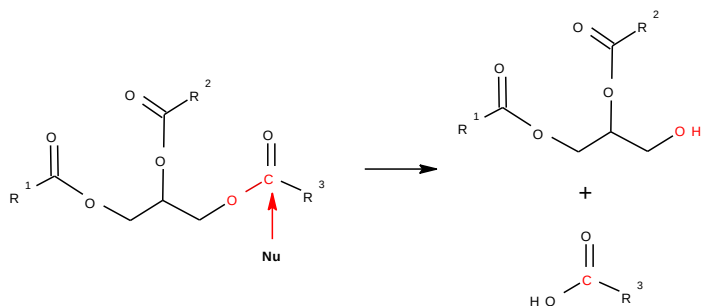
Эволюционно родственные ферменты с разными каталитическими активностями в рамках общей структурной организации



Candida antarctica
Липаза Б, PDB:1TCB



Дипептидил пептидаза IV
PDB: 1J2E



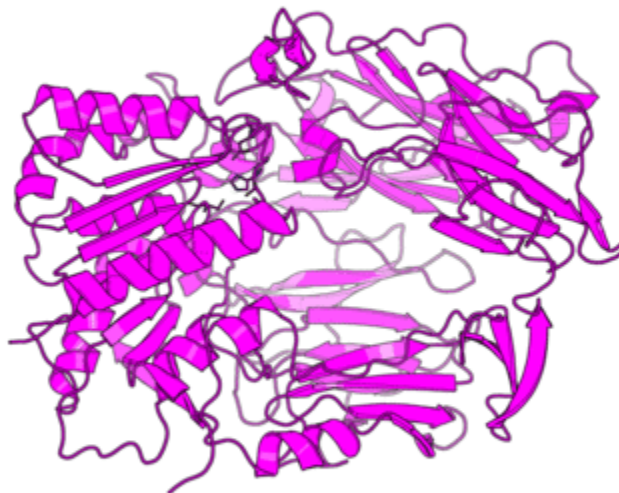
Suplatov, D. A., Besenmatter, W., Švedas, V. K., & Svendsen, A. (2012).
Protein Engineering, Design & Selection, 25(11), 689-697.

Липаза и амидаза суперсемейства α/β -гидролаз

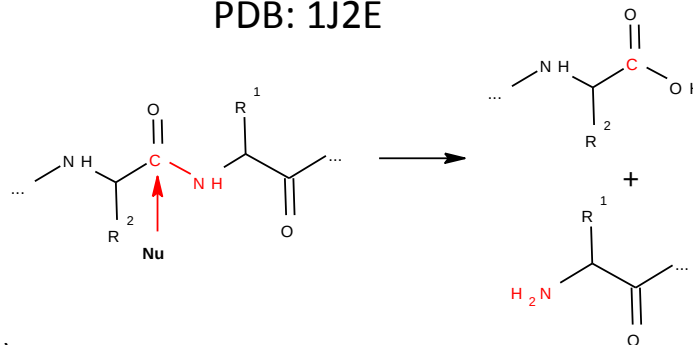
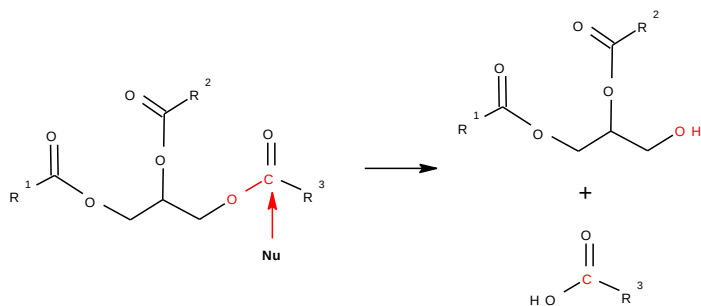
Эволюционно родственные ферменты с разными каталитическими активностями в рамках общей структурной организации



Candida antarctica
Липаза Б, PDB:1TCB



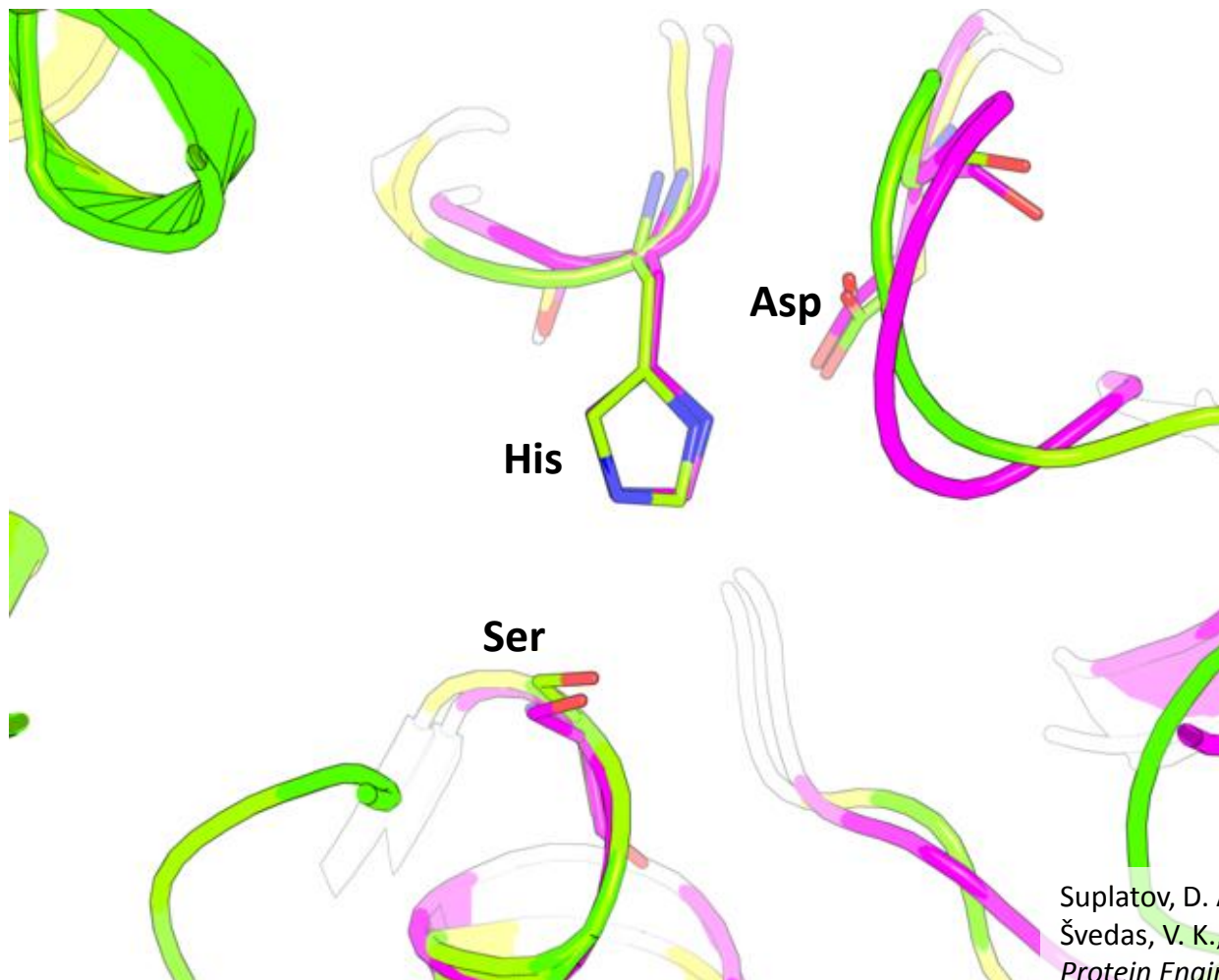
Дипептидил пептидаза IV
PDB: 1J2E



Suplatov, D. A., Besenmatter, W., Švedas, V. K., & Svendsen, A. (2012).
Protein Engineering, Design & Selection, 25(11), 689-697.

Липаза и амидаза суперсемейства α/β -гидролаз

Эволюционно родственные ферменты с разными каталитическими активностями в рамках общей структурной организации



Каталитическая триада консервативна
как по последовательности, так и по структуре

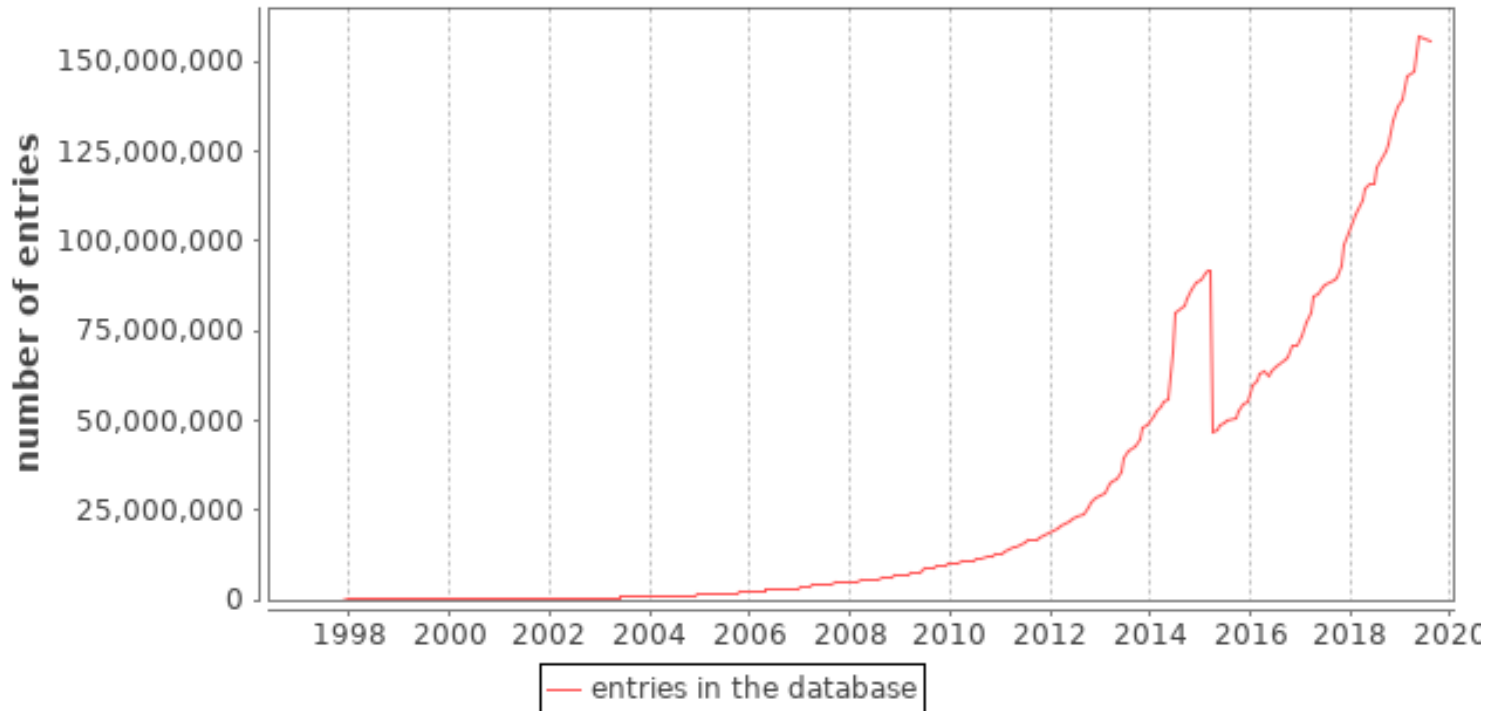
Suplatov, D. A., Besenmatter, W.,
Švedas, V. K., & Svendsen, A. (2012).
Protein Engineering, Design & Selection, 25(11), 689-697.

Изучение структурно-функциональных взаимосвязей в белках/ферментах

- Почему структурно **похожие** активные центры в гомологичных ферментах катализируют **различные** химические превращения?
- Как изучать структурно-функциональные взаимосвязи и предсказывать **изменения структуры**, которые приводят к **функциональному разнообразию**?
- Как определять «**горячие точки**» в структуре для того, чтобы **изменять функцию** ферментов и создавать более эффективные биокатализаторы?

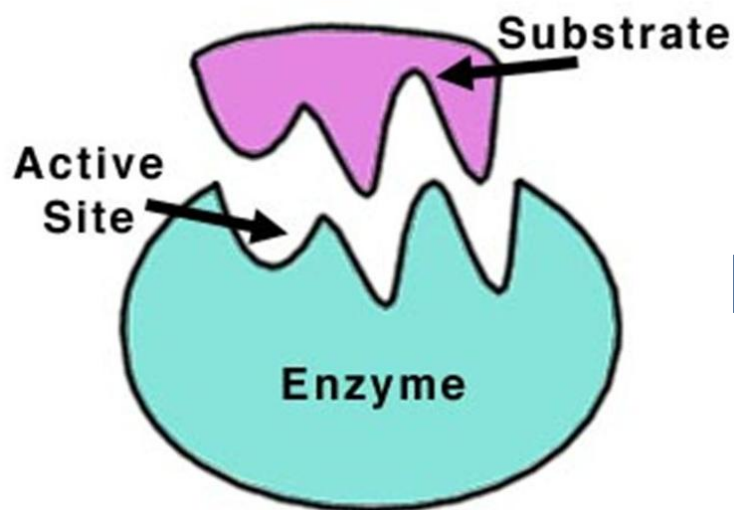
Экспоненциальный рост числа аминокислотных последовательностей белков в базе данных UniProtKB

Number of entries in UniProtKB/TrEMBL over time

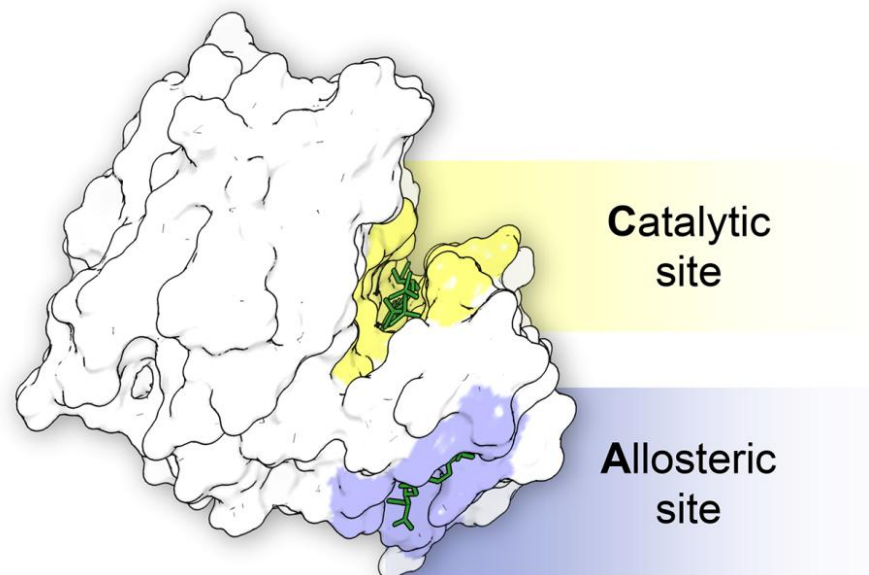


- На сегодняшний день менее 0.3% от ~ 156 000 000 известных белков были изучены экспериментально (UniProtKB, май 2019)
- Свойства оставшихся 99.7% никто и никогда не изучал

На сегодняшний день менее 0.3% известных белков изучены экспериментально, однако уже сейчас понятно, что взаимосвязь «последовательность-структура-функция» организована значительно более сложно, чем казалось еще недавно



«один фермент – один сайт – одна функция»



«один центр – много функций» (ферменты с разнородными активностями),

«одна функция – много центров» (аллостерически регулируемые белки)

«один белок – много функций» (мунлайтинговые белки)

Три уровня компьютерного моделирования для установления взаимосвязи между структурой и функциональными свойствами белков/ферментов

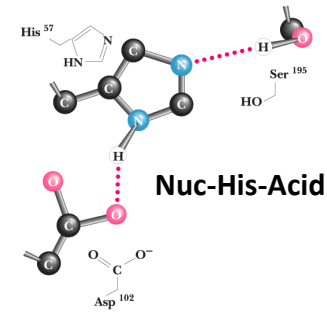
- 1. Биоинформатический анализ** (в том числе, методы машинного обучения) всех доступных структур и последовательностей белков/ферментов функционально разнообразных суперсемейств для изучения особенностей организации и поиска характеристических признаков в каталитических и регуляторных центрах гомологичных и аналогичных белков;
- 2. Методы молекулярного моделирования** (докинга, молекулярной динамики, гибридных квантово-химических подходов) для уточнения механизмов и сравнительного анализа биохимической основы процессов, реализуемых в конкретных активных центрах;
- 3. Автоматизированный интеллектуальный анализ информационных ресурсов**, таких как статьи, патенты, абстракты конференций и веб-сайты научных лабораторий, для экстракции и систематизации разрозненной экспериментальной информации о функциях и свойствах белков.

Уровень I: Биоинформатический анализ всех доступных структур и последовательностей белков/ферментов функционально разнообразных суперсемейств

Изучение белков/ферментов на уровне суперсемейств

Суперсемейство

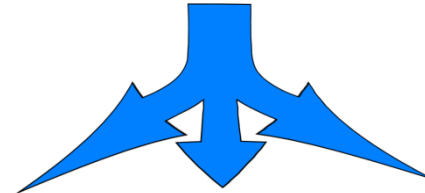
- **Общая структурная организация и сходство базовых функций** дают основание предполагать общее эволюционное происхождение
- Белки одного суперсемейства произошли от очень далекого общего предка и характеризуются **значительным функциональным разнообразием**



α/β -гидролазы

Семейство

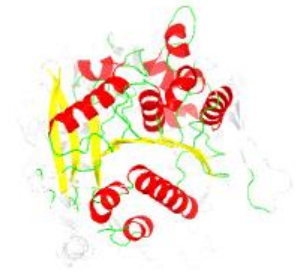
- **Эволюционно близкие белки** с высокой степенью сходства по аминокислотной последовательности, что подразумевает наличие близкого общего предка
- **Менее выраженное разнообразие свойств:** как правило обладают общим механизмом реакции, но могут превращать различные субстраты



Липазы



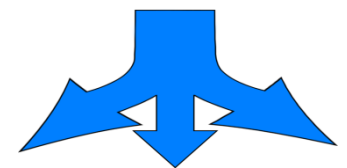
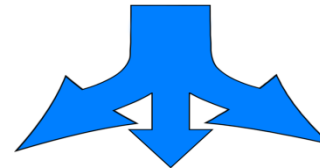
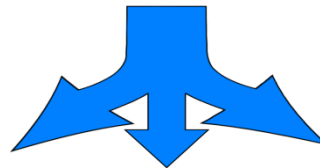
Эпоксид-гидролазы



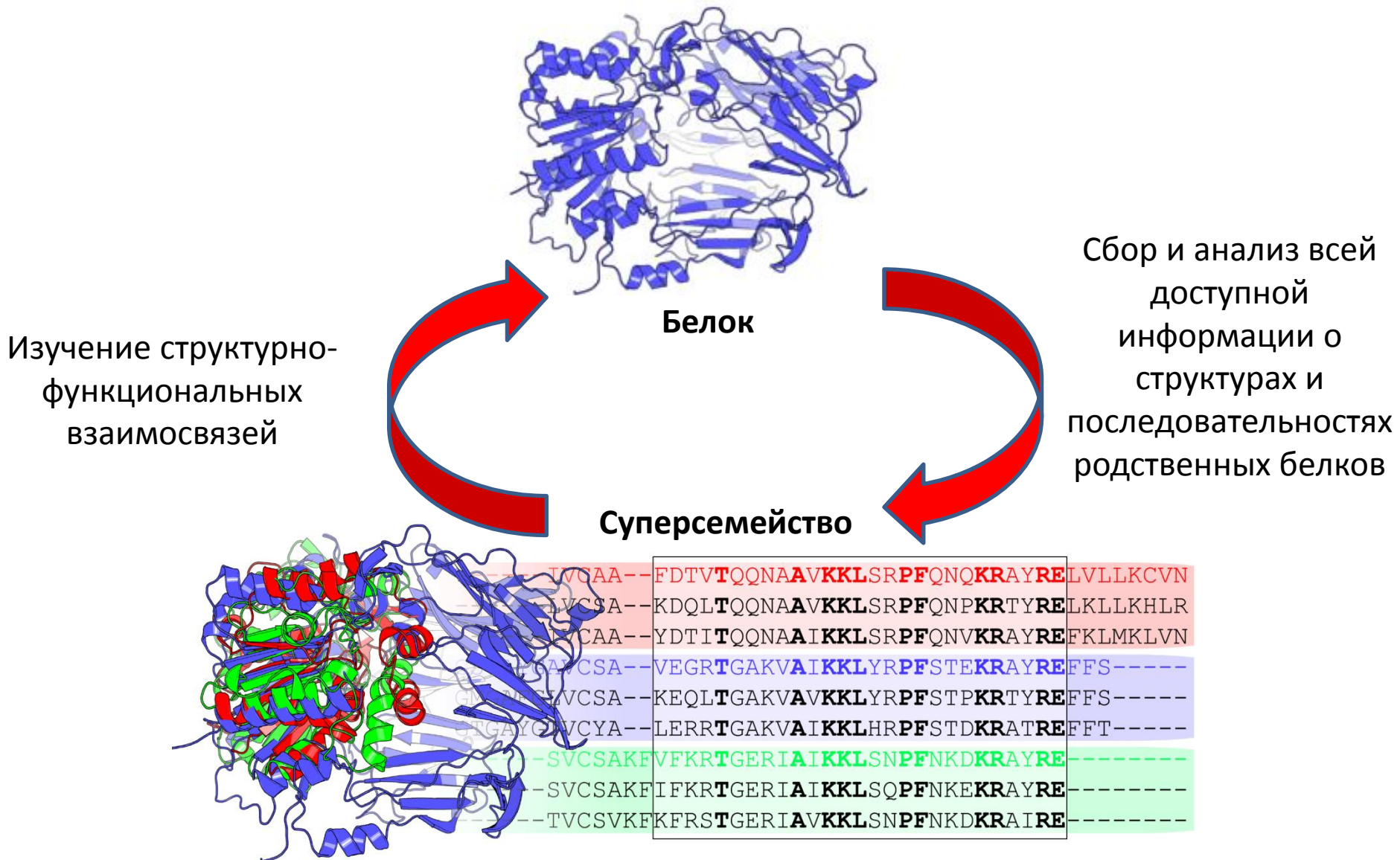
Пептидазы

Под-семейство

- **Еще более близкие белки**, характеризующиеся высокой степенью сходства по первичной и третичной структуре, и **гомогенностью функции**



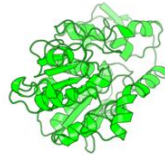
Изучение белков/ферментов на уровне суперсемейств



Построение множественного выравнивания суперсемейства белков

«Затравка»

(например, наиболее изученный белок суперсемейства)



> **protein1: ...AWSQGLP...**

Построение множественного выравнивания суперсемейства белков

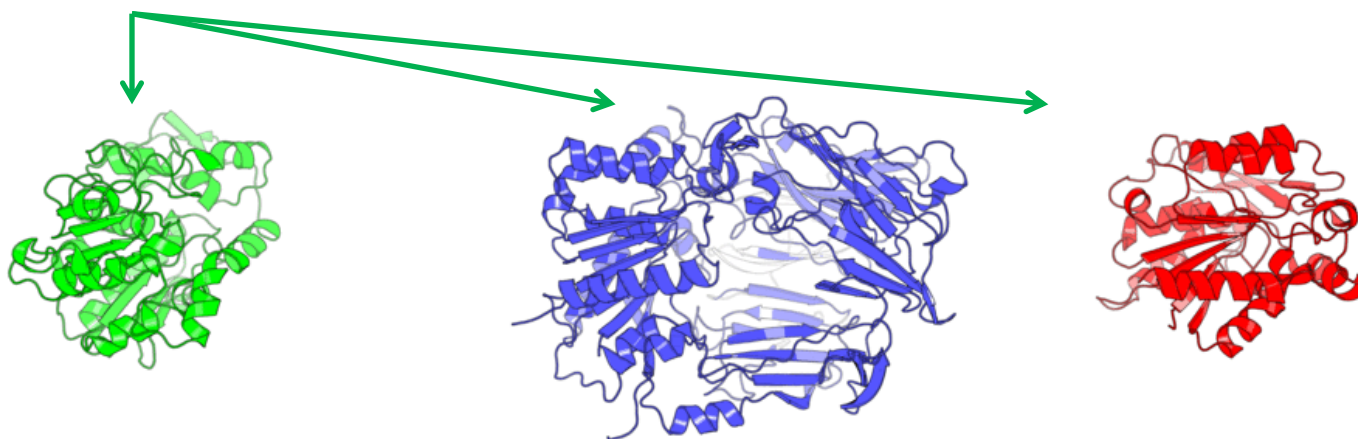
«Затравка»

(например, наиболее изученный белок суперсемейства)



Поиск по структурному сходству

Построение выборки *эволюционно удаленных* родственных белков на основании поиска в публичных базах данных



Построение множественного выравнивания суперсемейства белков

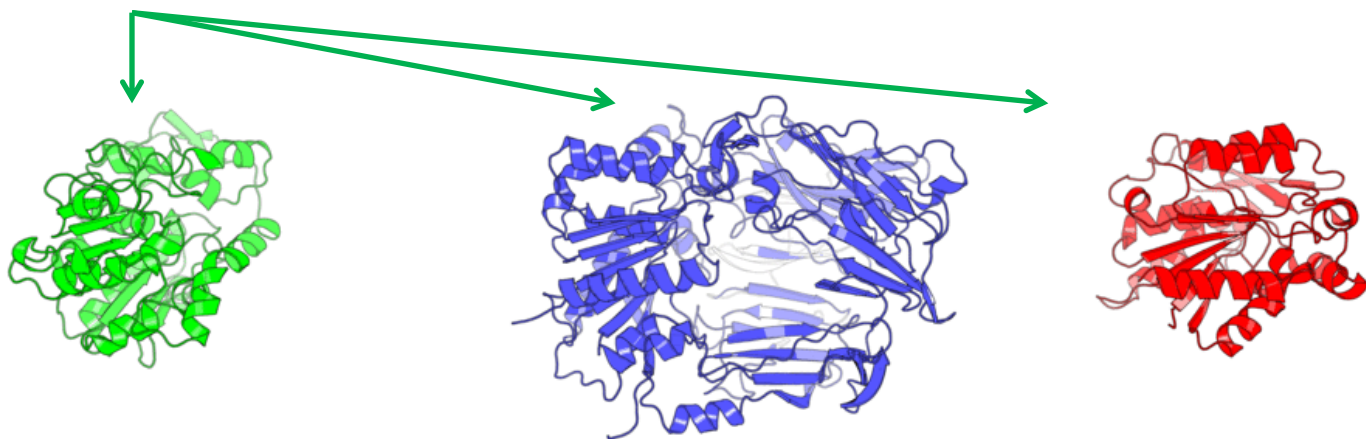
«Затравка»

(например, наиболее изученный белок суперсемейства)



Поиск по структурному сходству

Построение выборки *эволюционно удаленных* родственных белков на основании поиска в публичных базах данных



Поиск по сходству последовательностей

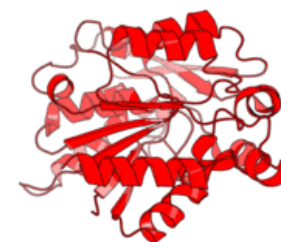
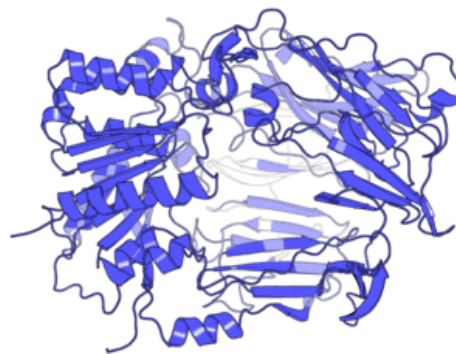
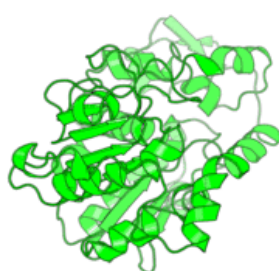
Построение выборки *эволюционно близких* белков на основании поиска в публичных базах данных

> **protein1: ...AWSQGLP...**
> **protein2: ...ASSQGAP...**
> **protein3: ...LSNLDP...**
> **protein4: ...WGSNAP...**
> **protein5: ...LASNLP...**

> **protein1: ...DGTMFDP...**
> **protein2: ...NATDP...**
> **protein3: ...DATMYDP...**
> **protein4: ...DGTEP...**
> **protein5: ...NATMFDP...**

> **protein1: ...KCDP...**
> **protein2: ...RCDP...**
> **protein3: ...RCDP...**
> **protein4: ...KCEP...**
> **protein5: ...KCDP...**

Построение множественного выравнивания суперсемейства белков



Выравнивание

последовательностей

Сравнение «алфавита»
аминокислот с
использованием
эмпирических моделей
аминокислотных замен

```
> protein1: ...AWSQGLP...
> protein2: ...ASSQGAP...
> protein3: ...L-SNLDP...
> protein4: ...WGSNA-P...
> protein5: ...LASNL-P...
```



```
> protein1: ...DGTMFDP...
> protein2: ...NAT--DP...
> protein3: ...DATMYDP...
> protein4: ...DGT--EP...
> protein5: ...NATMFDP...
```

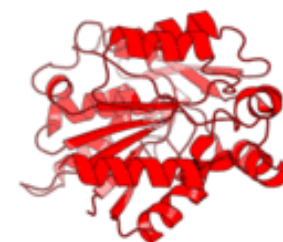
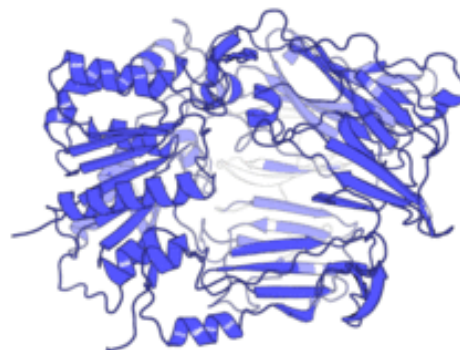
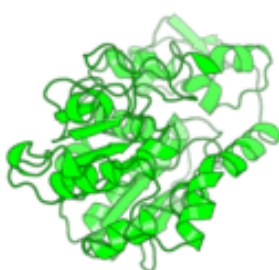


```
> protein1: ...KCDEP...
> protein2: ...RCDEP...
> protein3: ...RCDEP...
> protein4: ...KCEP...
> protein5: ...KCDEP...
```

Построение множественного выравнивания суперсемейства белков

Выравнивание структур

Поиск оптимального наложения координат атомов в трехмерном пространстве



Выравнивание последовательностей

Сравнение «алфавита» аминокислот с использованием эмпирических моделей аминокислотных замен

```
> protein1: ...AWSQGLP...
> protein2: ...ASSQGAP...
> protein3: ...L-SNLDP...
> protein4: ...WGSNA-P...
> protein5: ...LASNL-P...
```

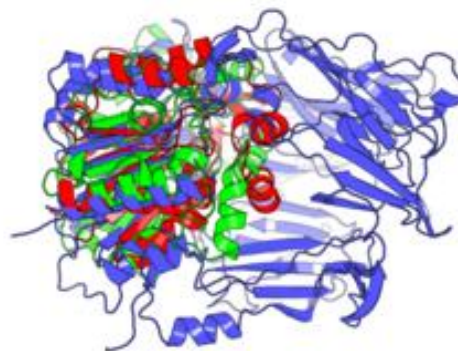
```
> protein1: ...DGTMFDP...
> protein2: ...NAT--DP...
> protein3: ...DATMYDP...
> protein4: ...DGT--EP...
> protein5: ...NATMFDP...
```

```
> protein1: ...KCDEP...
> protein2: ...RCDEP...
> protein3: ...RCDEP...
> protein4: ...KCEP...
> protein5: ...KCDEP...
```

Построение множественного выравнивания суперсемейства белков

Выравнивание структур

Поиск оптимального наложения координат атомов в трехмерном пространстве



Выравнивание последовательностей

Сравнение «алфавита» аминокислот с использованием эмпирических моделей аминокислотных замен

```
> protein1: ...AWSQGLP...
> protein2: ...ASSQGAP...
> protein3: ...L-SNLDP...
> protein4: ...WGSNA-P...
> protein5: ...LASNL-P...
```

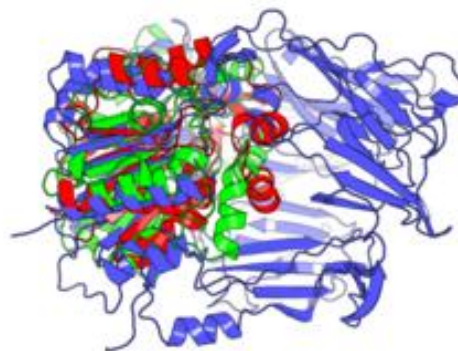
```
> protein1: ...DGTMFDP...
> protein2: ...NAT--DP...
> protein3: ...DATMYDP...
> protein4: ...DGT--EP...
> protein5: ...NATMFDP...
```

```
> protein1: ...KC DP...
> protein2: ...RC DP...
> protein3: ...RC DP...
> protein4: ...KC EP...
> protein5: ...KC DP...
```

Построение множественного выравнивания суперсемейства белков

Структурно-опосредованное множественное выравнивание аминокислотных последовательностей

Эволюционно удаленные белки, которые потеряли сходство по последовательностям в процессе естественного отбора и специализации от общего предка, сравниваются посредством выравнивания структур, в то время как выравнивание аминокислотных последовательностей дает полезный результат в случае близких гомологов



```
> protein1: ...AWSQGLP...
> protein2: ...ASSQGAP...
> protein3: ...L-SNLDP...
> protein4: ...WGSNA-P...
> protein5: ...LASNL-P...
```

```
> protein1: ...DGTMFDP...
> protein2: ...NAT--DP...
> protein3: ...DATMYDP...
> protein4: ...DGT--EP...
> protein5: ...NATMFDP...
```

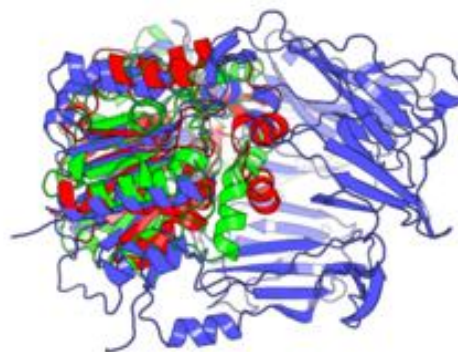
```
> protein1: ...KCDP...
> protein2: ...RCDP...
> protein3: ...RCDP...
> protein4: ...KCEP...
> protein5: ...KCDP...
```

Построение множественного выравнивания суперсемейства белков

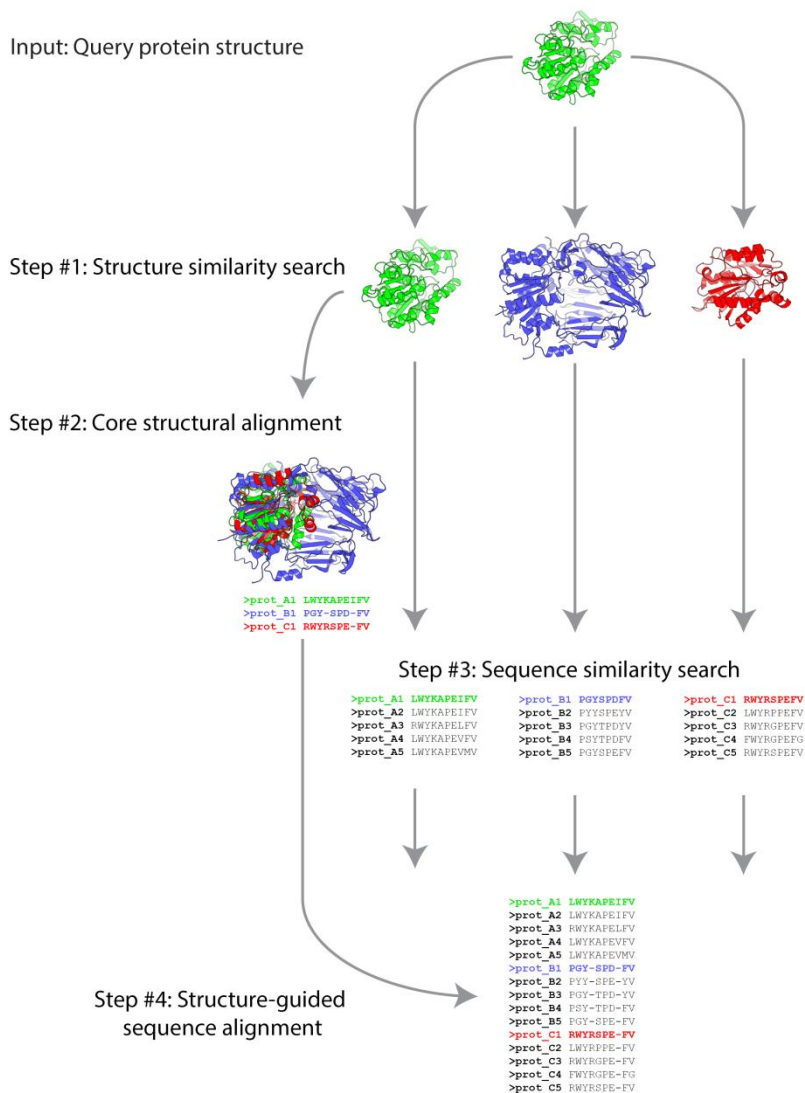
Структурно-опосредованное множественное выравнивание

Эволюционно удаленные белки, которые потеряли сходство по последовательностям в процессе естественного отбора и специализации от общего предка, сравниваются посредством выравнивания структур, в то время как выравнивание аминокислотных последовательностей дает полезный результат в случае близких гомологов

```
> protein1: ...A-WSQGLP...
> protein2: ...A-SSQGAP...
> protein3: ...L--SNLDP...
> protein4: ...W-GSNA-P...
> protein5: ...L-ASNL-P...
> protein1: ...DG-TMFDP...
> protein2: ...NA-T--DP...
> protein3: ...DA-TMYDP...
> protein4: ...DG-T--EP...
> protein5: ...NA-TMFDP...
> protein1: ...K--C--DP...
> protein2: ...R--C--DP...
> protein3: ...R--C--DP...
> protein4: ...K--C--EP...
> protein5: ...K--C--DP...
```



Построение множественного выравнивания суперсемейства белков

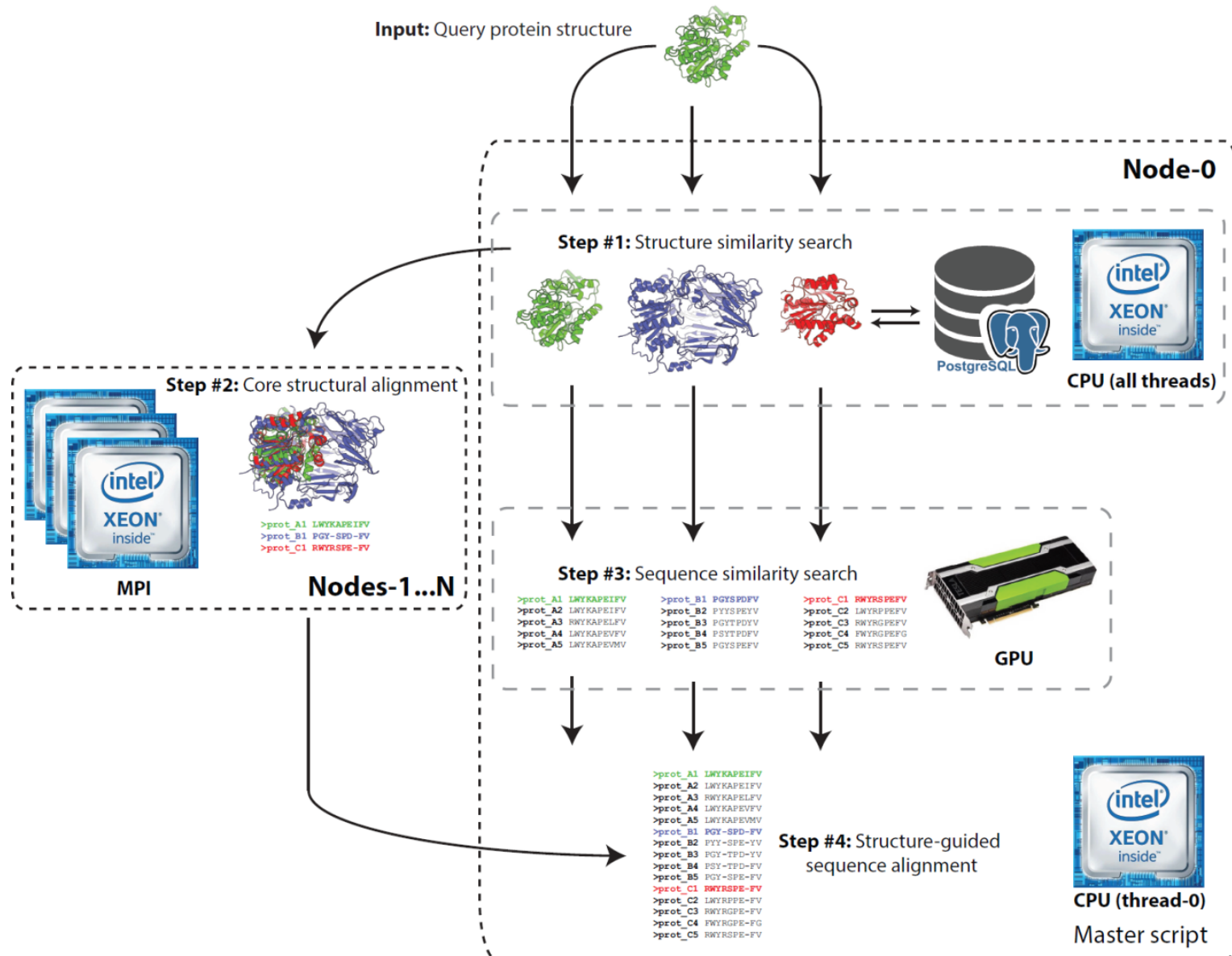


Конвейер из последовательных и независимых в программном отношении этапов, которые примерно равнозначных по вычислительной трудоемкости;

Для эффективной работы нужно оптимизировать все стадии единого комплексного решения;

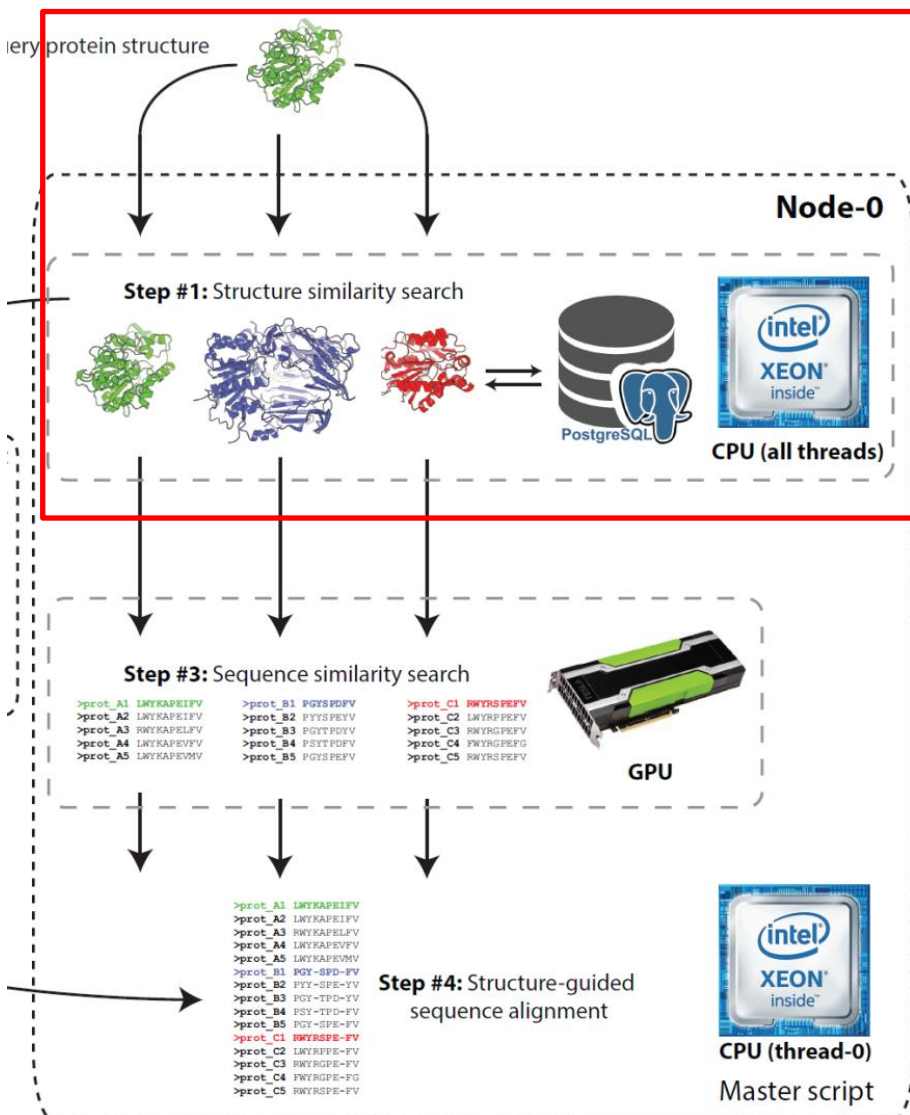
Suplatov D.A., Kopylov K.E., Popova N.N., Voevodin V.V., Švedas V.K. (2018) Mustguseal: a Server for Multiple Structure-Guided Sequence Alignment of Protein Families. *Bioinformatics*, 34(9), 1583-1585

Гибридный подход к построению множественного выравнивания суперсемейства белков



Построение множественного выравнивания суперсемейства белков

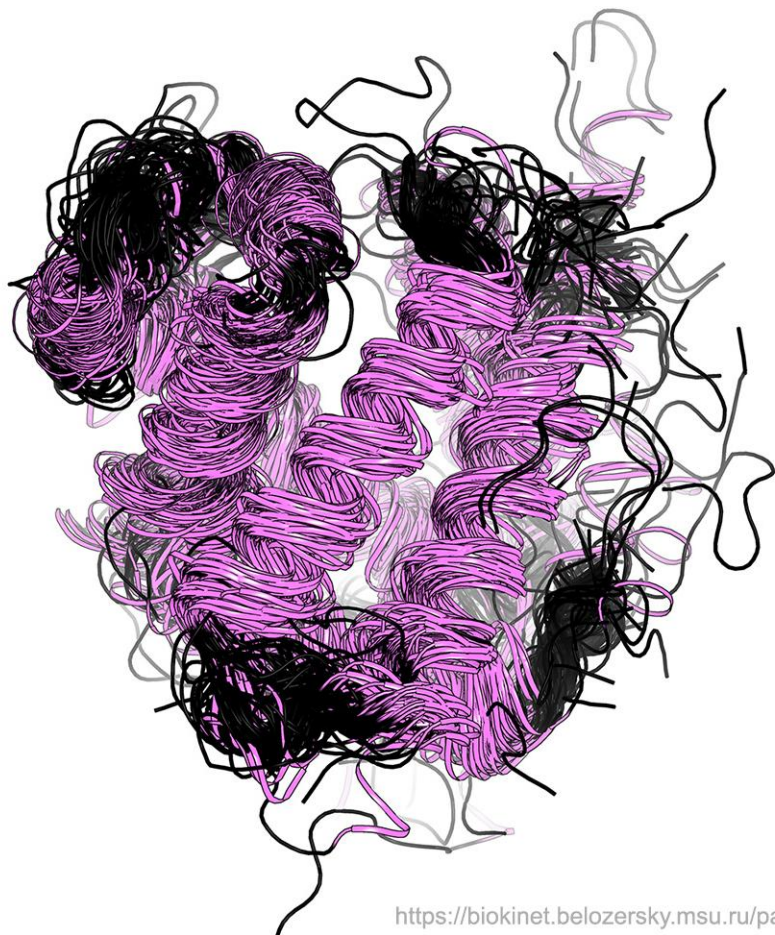
Этап 1: Поиск по структурному сходству



- Поиск в базе данных PDB ~330 тыс. структур отдельных цепей;
- Задача поиска в базе разделяется на ~30000 – 330000 подзадач, которые полностью независимы друг от друга;
- Решается через распараллеливание в рамках общей памяти. За счет препроцессинга базы данных и удаления заведомо непохожих структур из поиска задача может быть решена за ~10-60 минут на Intel Xeon Gold 6126 CPU,
- Однажды посчитанные парные выравнивания сохраняются в базе данных на основе PostgreSQL для повторного использования. Повторное использование ранее посчитанных результатов может сократить время этапа до нескольких секунд.

Построение множественного выравнивания суперсемейства белков

Этап 2: Множественное структурное выравнивание



<https://biokinet.belozersky.msu.ru/parmatt>



Построение множественного выравнивания суперсемейства белков

Этап 2: Множественное структурное выравнивание

- Все существующие программы для построения множественного 3D-выравнивания белков способны использовать только одно ядро CPU (в лучшем случае – частично ускоряются в режиме OpenMP);
- Это связано с устаревшим представлением о том, что информации о 3D-структур белков «мало».

Построение множественного выравнивания суперсемейства белков

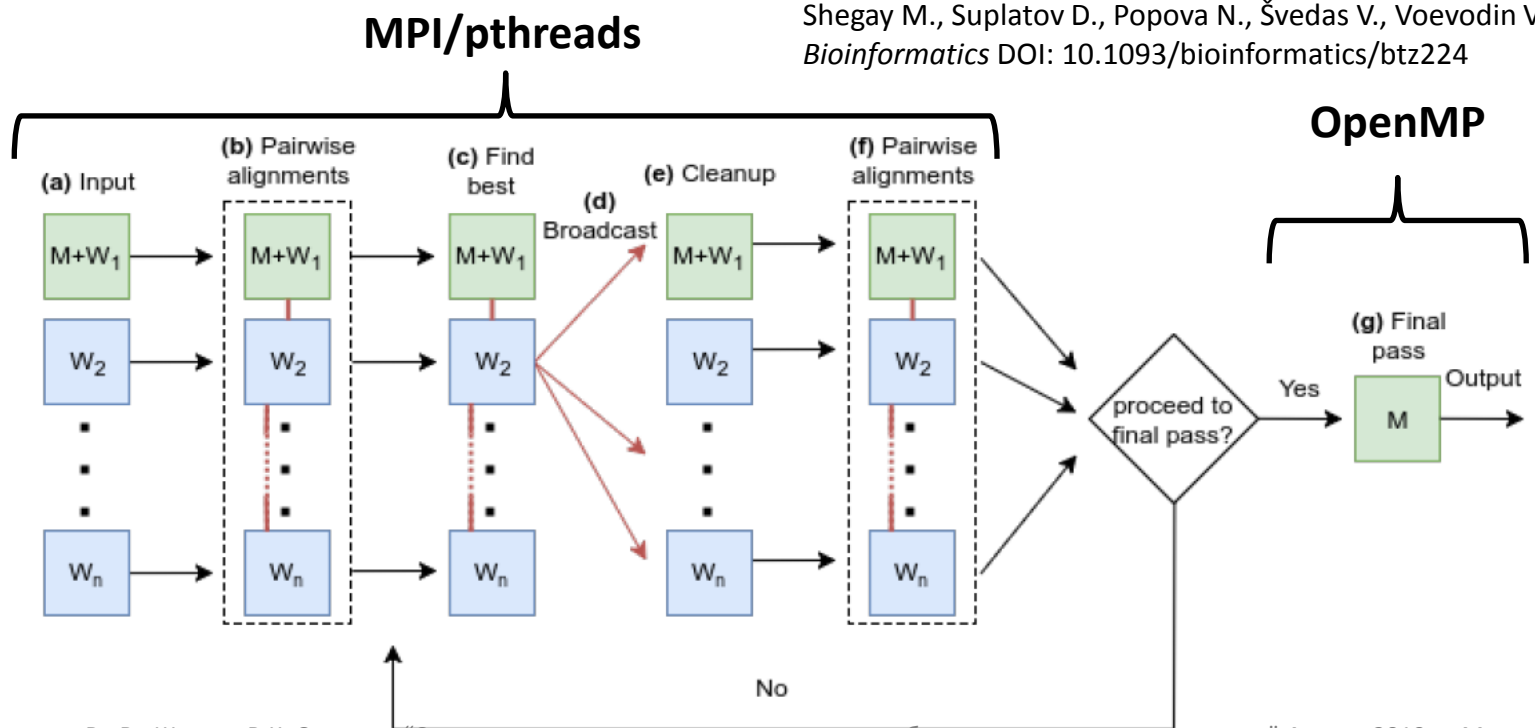
Этап 2: Множественное структурное выравнивание

- Когда-то информации о 3D-структурах белков действительно было мало, но теперь ее доступность в базе PDB растет почти в геометрической прогрессии;
- Сегодня суперсемейства белков могут быть представлены тысячами структур, и сравнительный анализ такого объема 3D-данных на одном CPU не всегда возможен;
- Очевидно, что уже в ближайшем будущем использование одного CPU для множественного 3D-выравнивания структур белков станет невозможным.



parMATT – первая программа для построения множественных 3D-выравниваний белков на суперкомпьютере

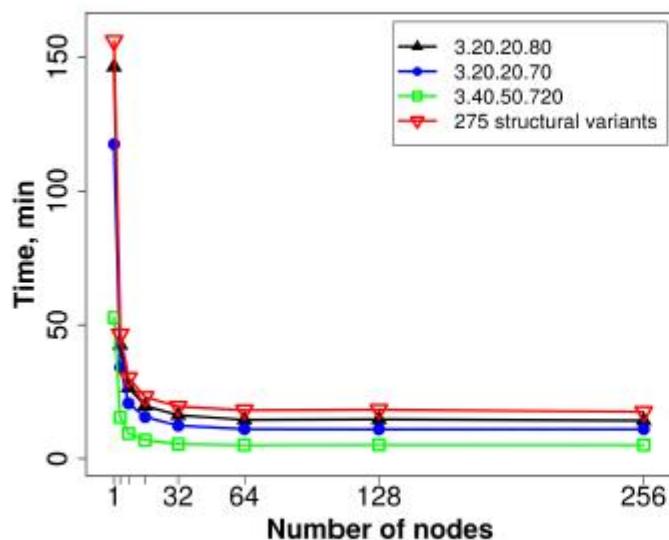
- parMATT** - это гибридная параллельная ре-имплементация алгоритма MATT с использованием MPI/threads/OpenMP; Наиболее сложные в вычислительном отношении этапы алгоритма MATT - первоначальное построение парных выравниваний между всеми входными структурами и последующее прогрессивное множественное выравнивание – были распараллелены с использованием MPI и pthreads, а заключительный этап уточнения был оптимизирован путем введения поддержки OpenMP;



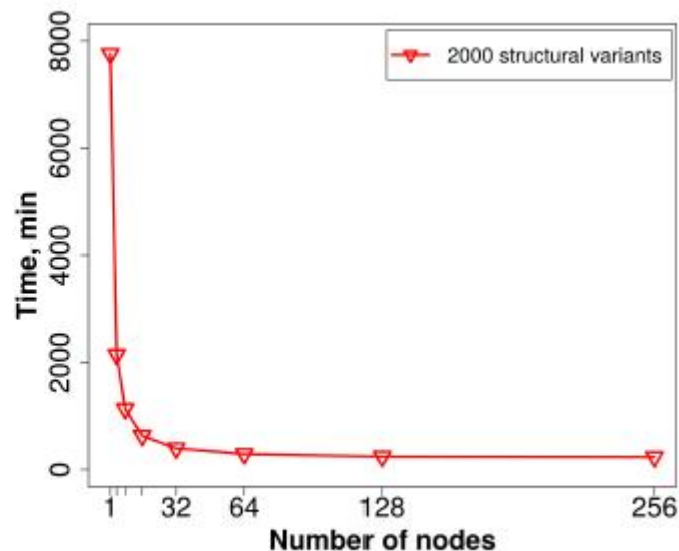
parMATT – первая программа для построения множественных 3D-выравниваний белков на суперкомпьютере

- parMATT может работать быстрее, чем исходный MATT, даже на одном многоядерном процессоре, но главное свойство новой программы – принципиальное ускорение при использовании классических кластеров и суперкомпьютеров).
- Ускорение и «точка насыщения» зависят от свойств входных данных – в первую очередь от числа и размера 3D-структур;
- Использование parMATT позволяет ускорить выравнивание выборки из 2000 структур в ~32 раза на 128 узлах суперкомпьютера «Ломоносов-1» (с 6 дней до 4 часов)

Shegay M., Suplatov D., Popova N., Švedas V., Voevodin V. (2019)
Bioinformatics DOI: 10.1093/bioinformatics/btz224



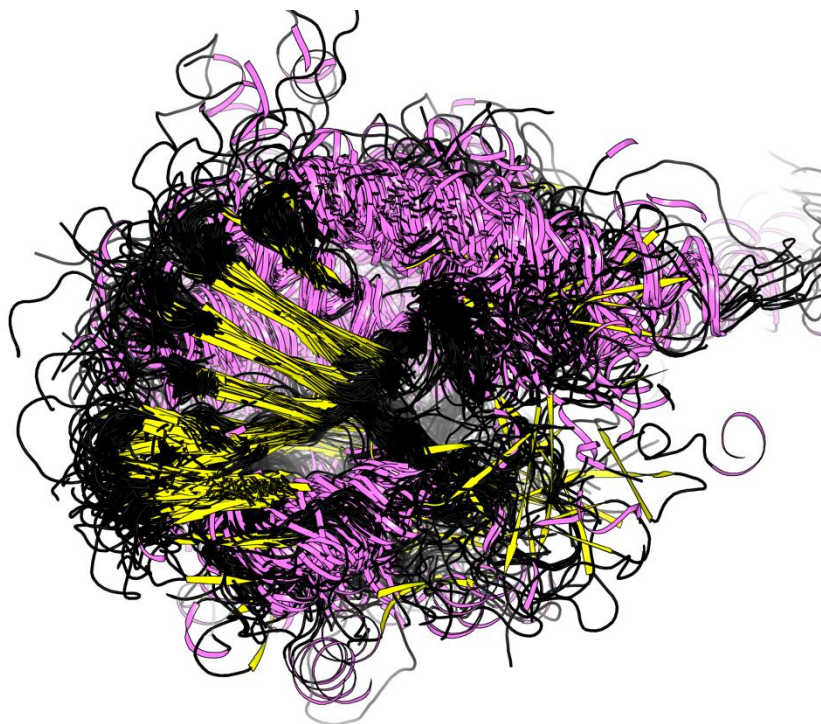
(A) Larger CATH sets



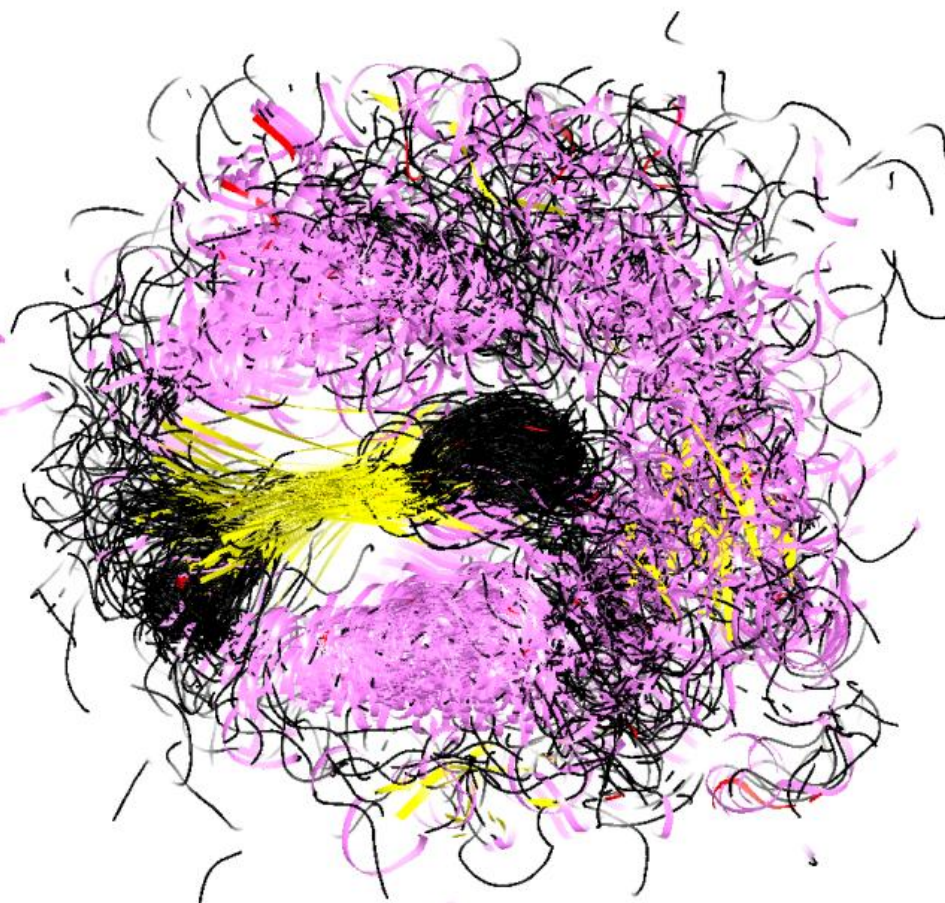
(B) MD-generated ensemble of structural variants

Построение множественного выравнивания суперсемейства белков

Этап 2: Множественное структурное выравнивание



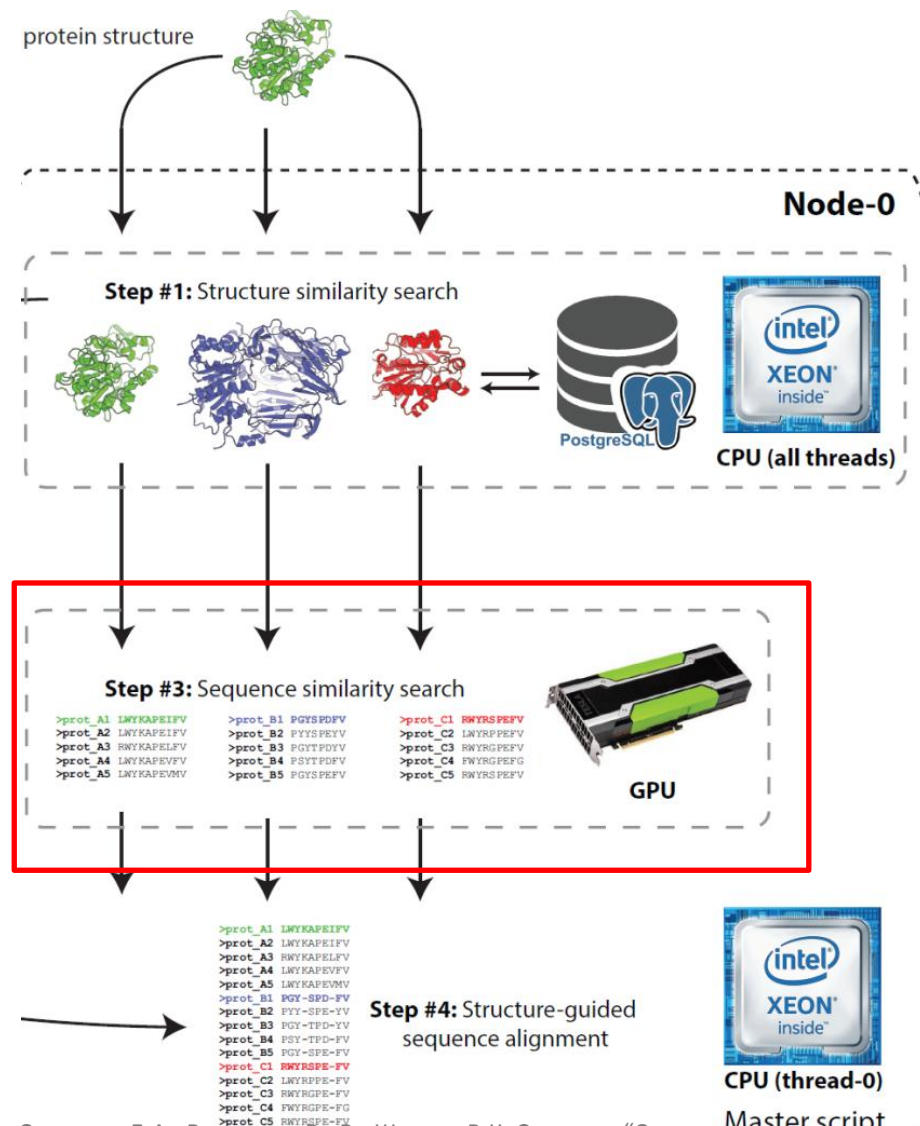
Выравнивание структур родственных
ГАФД и ГАФД-подобных белков



Выравнивание структур родственных
белков суперсемейства α/β -гидролаз

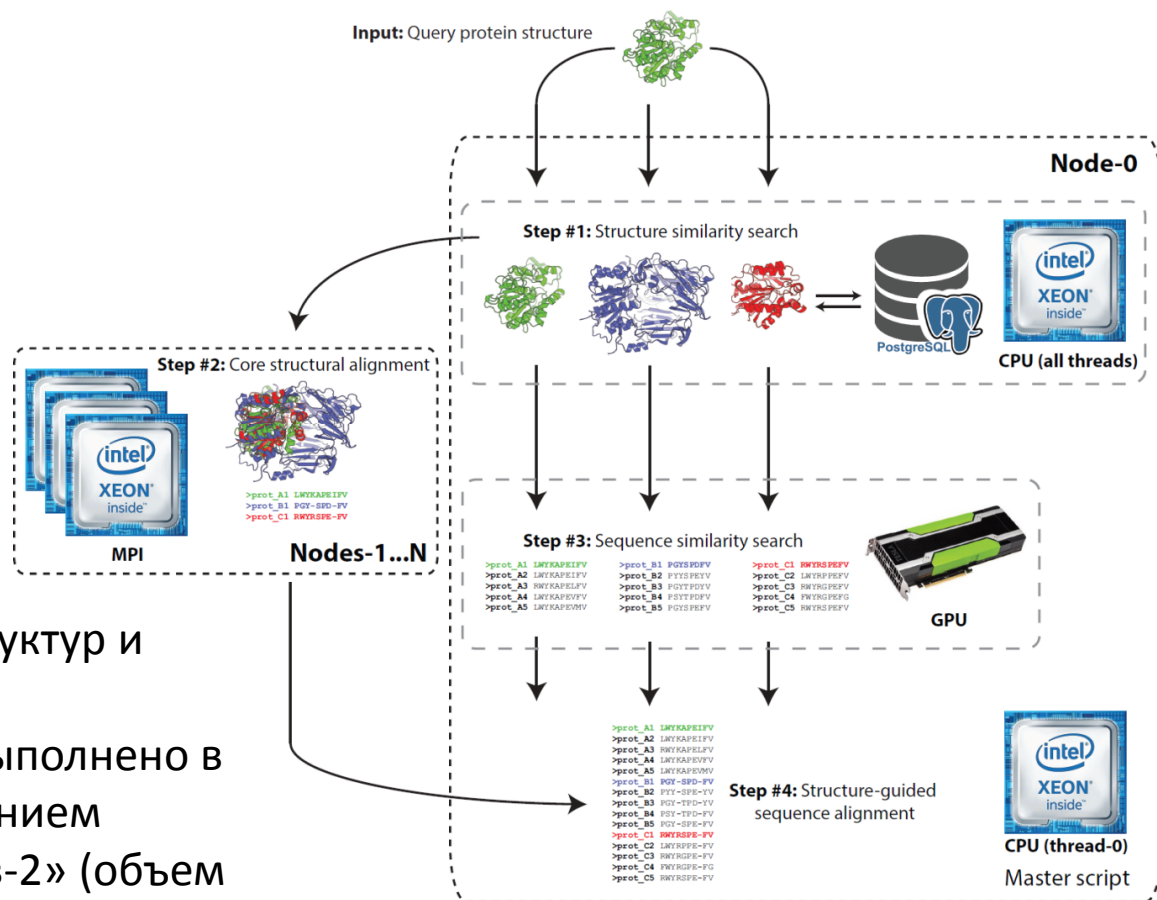
Построение множественного выравнивания суперсемейства белков

Этап 3: Поиск по сходству и выравнивание аминокислотных последовательностей



- Размер банка данных UniProtKB в мае 2019 года составляет ~156'000'000 последовательностей (увеличился в два раза за два года);
- Используется **GPU-версия** алгоритма BLAST;
- На одном узле раздела “Pascal” (12-ядерный CPU и 2 GPU) может быть одновременно запущено 12 параллельных BLAST поисков без потери скорости расчета.

Гибридный подход к построению множественного выравнивания суперсемейства белков



Выравнивание сотен тысяч структур и последовательностей белков суперсемейства может быть выполнено в течение 12 часов с использованием суперкомпьютера «Ломоносов-2» (объем ресурсов зависит от характеристик входных данных)

Dmitry Suplatov, Yana Sharapova, Maxim Shegay, Nina Popova, Kateryna Fesko, Vladimir Voevodin and Vytas Švedas (2019) High-Performance Hybrid Computing for Bioinformatic Analysis of Protein Superfamilies – включено в предварительную программу конференции RuSCDays'19

Построение множественного выравнивания суперсемейства белков

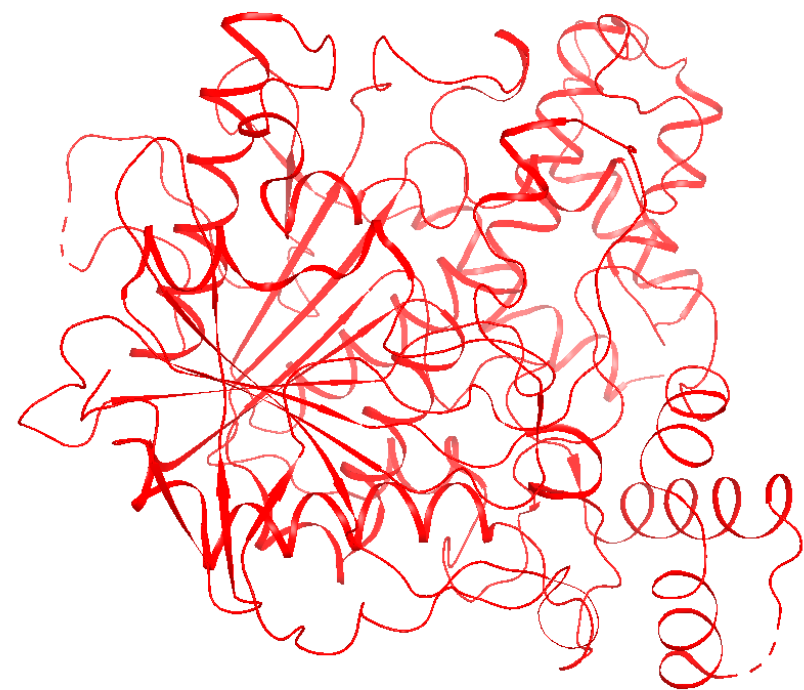


Структура
аспартоацилазы человека

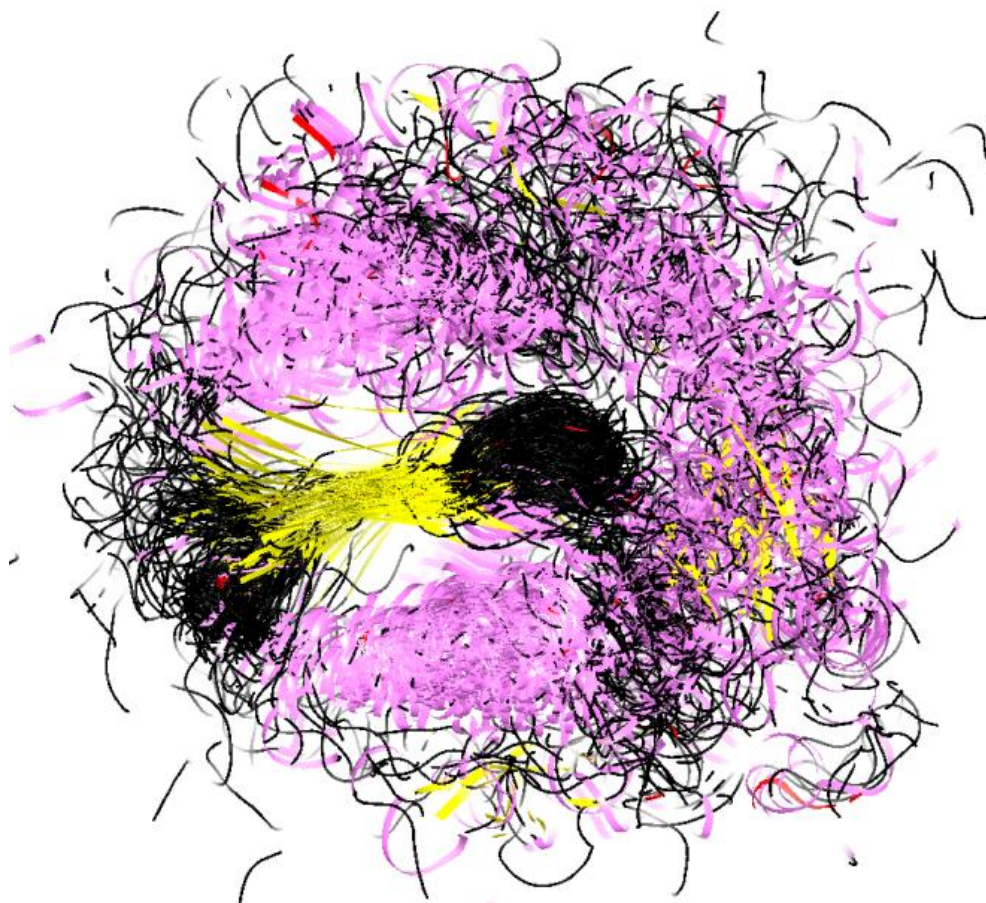


С использованием суперкомпьютера в публичных базах данных обнаружено **78'341** последовательностей и структур гомологов аспартоацилазы из суперсемейства **Zn-зависимых пептидаз**

Построение множественного выравнивания суперсемейства белков



Структура
ацетилхолинэстеразы
человека



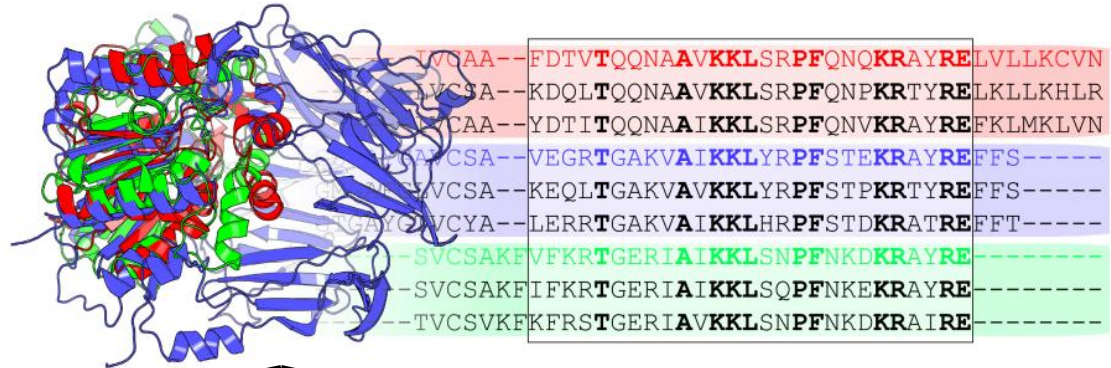
С использованием суперкомпьютера в публичных
базах данных обнаружено
139'157 гомологов ацетилхолинэстеразы
из суперсемейства α/β -гидролаз

Открытая он-лайн платформа для изучения структуры, функции и регуляции белков методами биоинформатики

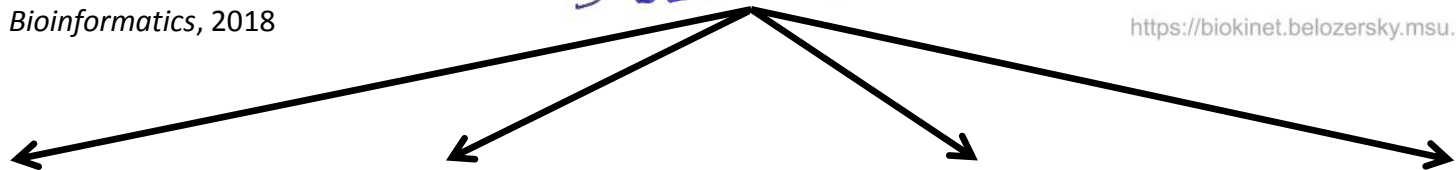
Mustguseal

Сбор всей доступной информации о последовательностях и структурах родственных белков

Bioinformatics, 2018



<https://biokinet.belozersky.msu.ru/mustguseal>



Zebra

Анализ механизмов функционального разнообразия в родственных белках

J Biomol Struct Dyn., 2014

pocketZebra

Поиск ранее неизвестных сайтов связывания субстратов и регуляторных лигандов в структурах белков

Nucleic Acids Research, 2014

visualCMAT

Поиск ко-эволюционирующих позиций в структурах родственных белков, изучение путей аллостерической коммуникации

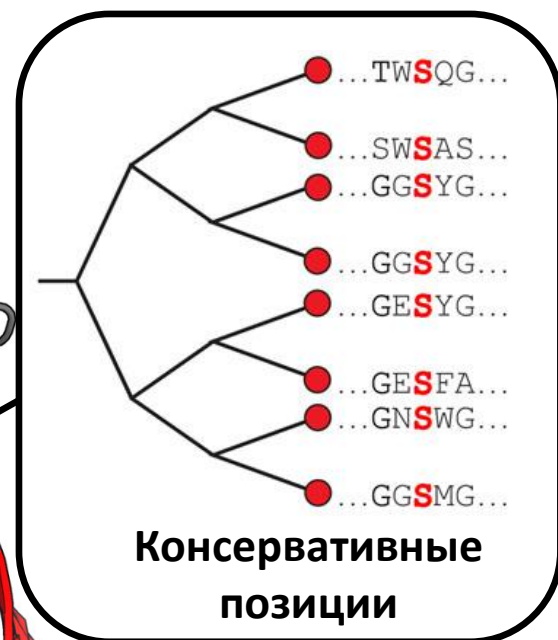
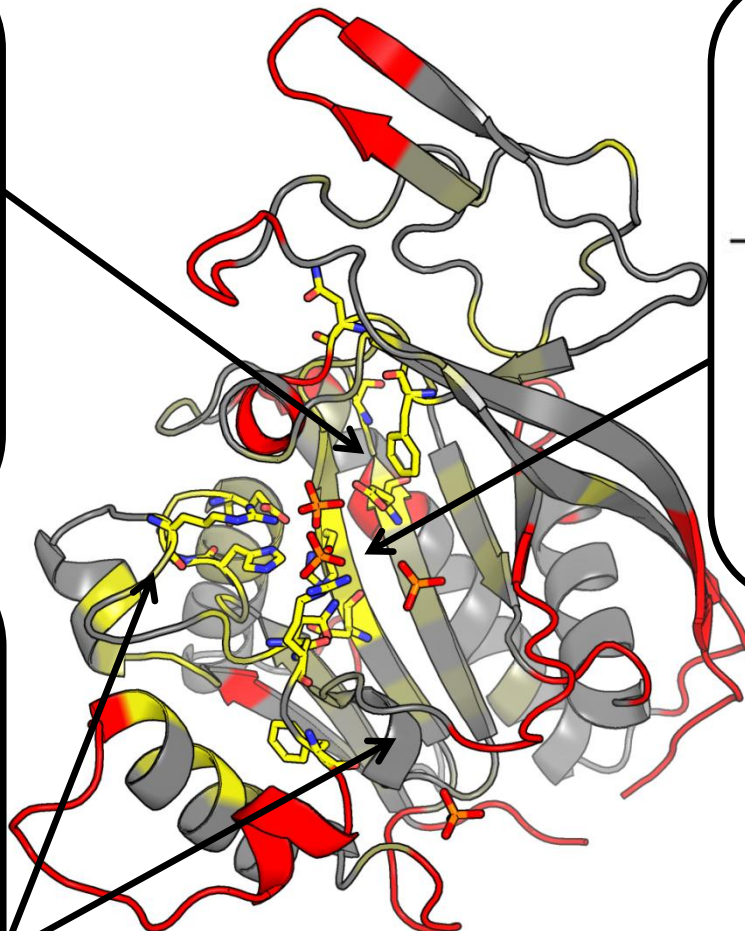
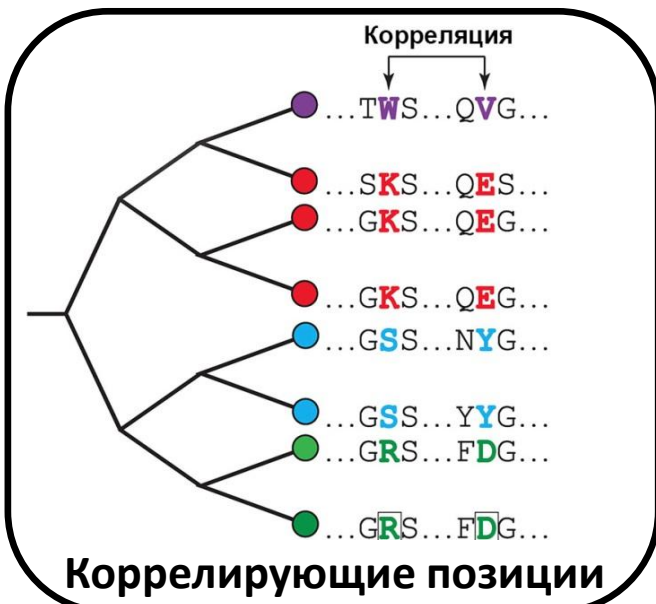
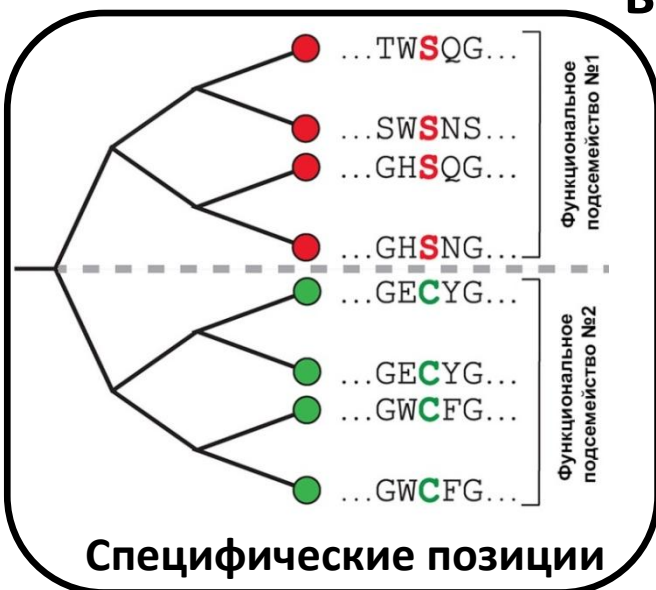
J Bioinform Comput Biol., 2018

Yosshi

Изучение роли дисульфидных связей в функции и регуляции родственных белков

Nucleic Acids Research, 2019

Вариабельность аминокислот в участке структуры фермента и его гомологов в суперсемействе помогает понять роль остатков в функции и регуляции



Биоинформатический анализ

Список публично-доступных web-ресурсов

- Веб-сервис для построения множественных выравниваний
<https://biokinet.belozersky.msu.ru/mustguseal>
- Программа parMATT для построения множественных 3D-выравниваний на суперкомпьютере
<https://biokinet.belozersky.msu.ru/parmatt>
- Веб-сервисы для биоинформатического анализа множественных выравниваний, поиска и классификации функционально важных аминокислотных остатков, идентификации новых центров связывания и изучения дисульфидных связей в суперсемействах белков
<https://biokinet.belozersky.msu.ru/pocketzebra>
<https://biokinet.belozersky.msu.ru/zebra>
<https://biokinet.belozersky.msu.ru/visualCMAT>
<https://biokinet.belozersky.msu.ru/yosshi>

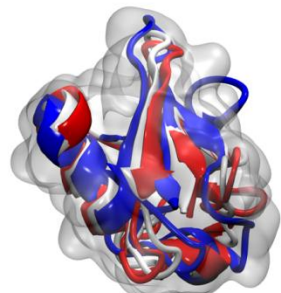
Уровень II: Методы молекулярного моделирования для уточнения механизмов и сравнительного анализа биохимической основы процессов, реализуемых в конкретных активных центрах

Уровень I + II

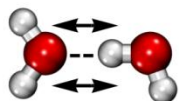
- **Биоинформатический анализ** позволяет
 - собрать и систематизировать информацию о суперсемейства белков/ферментов;
 - выбрать и классифицировать функционально важные аминокислотные остатки;
 - идентифицировать новые центры связывания;
- **Молекулярное моделирование** позволяет уточнить результаты биоинформатического анализа:
 - изучить роль ключевых остатков;
 - найти (создать) лиганды (субстраты, регуляторные лиганды/ингибиторы), комплементарные выбранным сайтам в структуре белка;
 - Изучить/уточнить механизмы взаимодействия белков/ферментов с субстратами и регуляторными лигандами.

Молекулярная динамика: метод для изучения структуры белка в движении

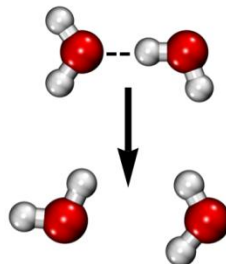
domain vibrations



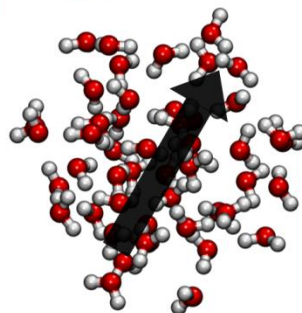
HB vibrations



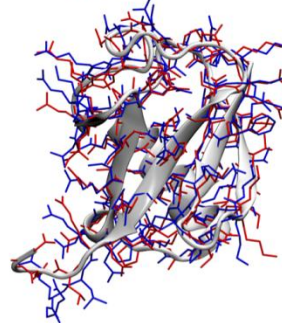
HB breaking
rotational relaxation
translational diffusion



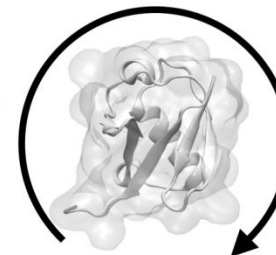
collective water
dipole relaxation



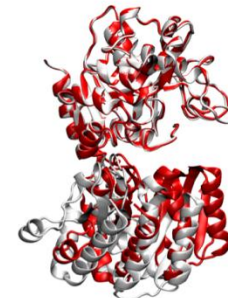
sidechain
fluctuations



protein tumbling



conformational
transitions



0.1–1 ps

1–10 ps

10–100 ps

1–10 ns

10–100 ns

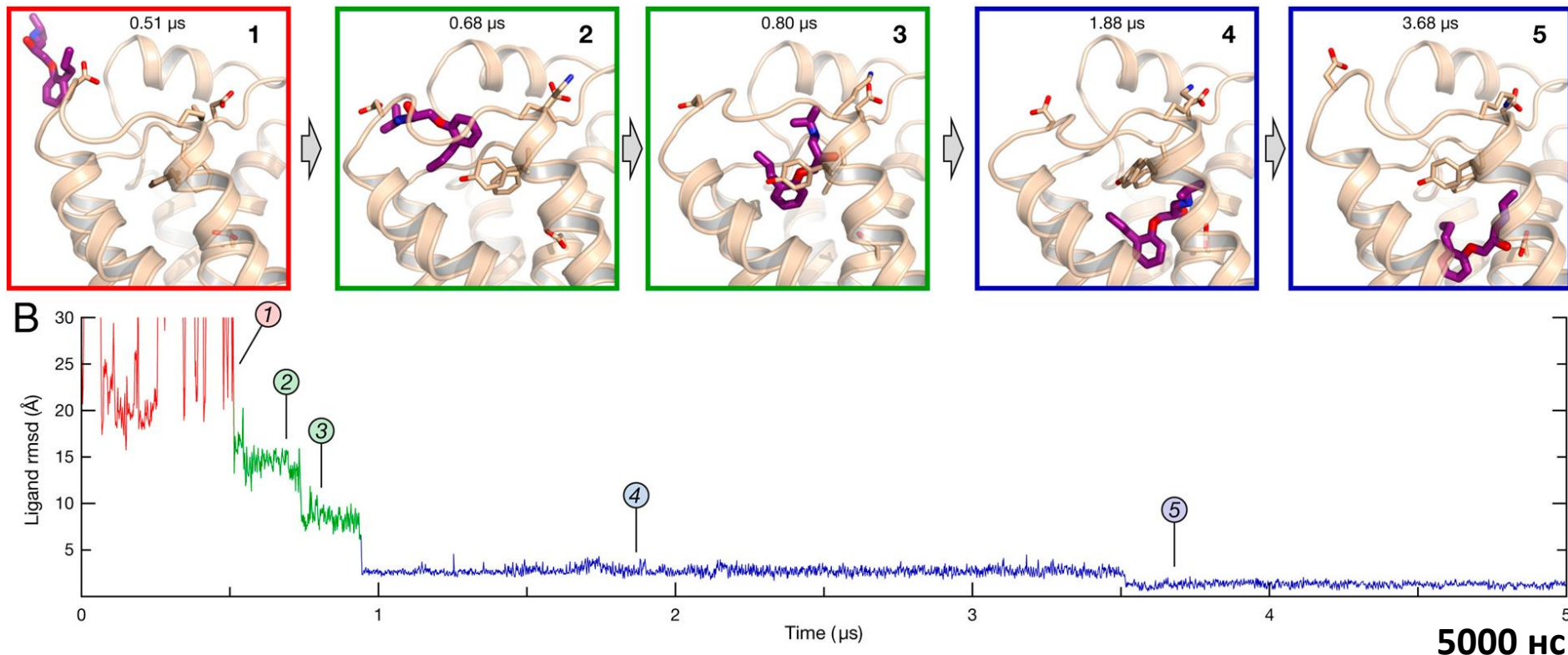
μs

<https://isearch.asu.edu/profile/3160407>

© M. Heyden

Для изучения более сложных конформационных изменений в структуре белка необходимо рассчитывать более длинные траектории молекулярной динамики

Примеры работы с использованием молекулярной динамики (2011 год)



Dror R. O. et al. Pathway and mechanism of drug binding to G-protein-coupled receptors //Proceedings of the National Academy of Sciences. – 2011. – T. 108. – №. 32. – C. 13118-13123.

Специализированные суперкомпьютеры Anton/Anton-2 для молекулярной динамики



(a)



(b)



(c)

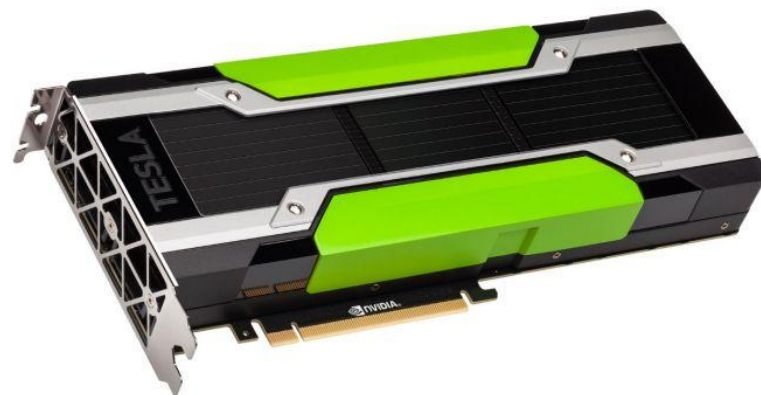
An Anton 2 ASIC. (b) An Anton 2 node board, with one ASIC underneath a phase-change heat sink. (c) A 512-node Anton 2 machine, w

- Продукт частной биотехнологической компании «DE Shaw Research»;
- Результат глубокой переработки и ко-дизайна программного обеспечения и аппаратных средств для его исполнения;
- Предназначен для исполнения только специального программного обеспечения;
- Очень быстрое, но редкое (штучное) оборудование, доступное только в США

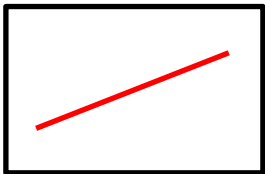
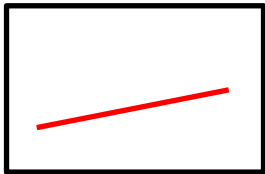
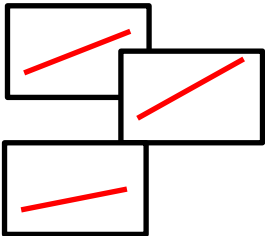
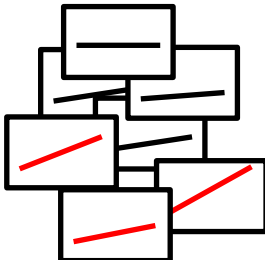
Shaw D. E. et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer //Proceedings of the international conference for high performance computing, networking, storage and analysis. – IEEE Press, 2014. – С. 41-53.

Развитие компьютерных технологий: ускорение

- Увеличение мощности **классических** процессоров (CPUs);
- Разработка специализированных высокопроизводительных вычислительных архитектур (например, **GPU-ускорители**);
- Появление мощных многоузловых вычислительных систем – **кластеров и суперкомпьютеров** (в марте 2018 года пиковая производительность суперкомпьютера Ломоносов-2 увеличилась до 5 Пфлопс).



Объем вычислений молекулярной динамики в повседневной лабораторной практике

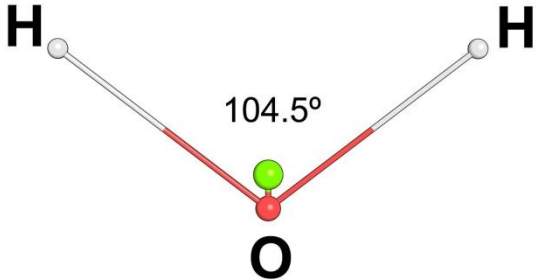
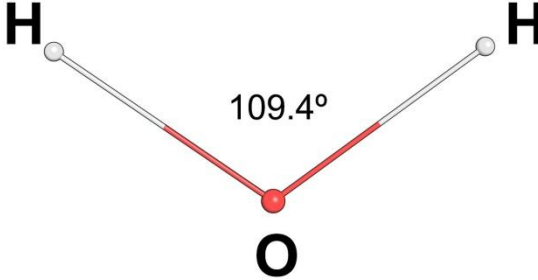
До 2000	2000 – 2008	2008 – 2014	2014-2017	2018
< 1 нс	10 нс	10 – 100 нс	100 нс	?
				
		Чебышев Ломоносов-1	Ломоносов-2 	

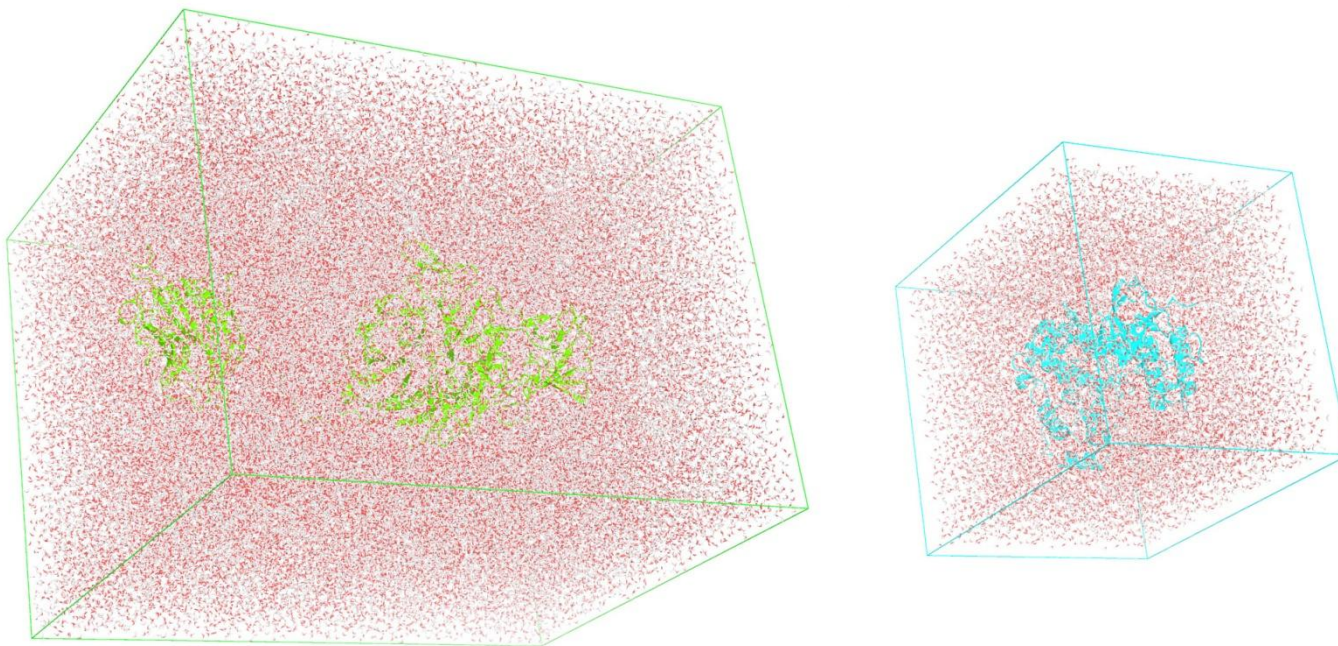
Скорость расчета молекулярной динамики определяется:

- оборудованием, его конфигурацией;
- выбранных параметров алгоритма;
- свойств входных данных

Использование различных силовых полей и моделей воды

AMBER FF14SB	AMBER FF15IPQ
«Классическое» силовое поле, основное рекомендуемое силовое поле для молекулярной динамики белков в AMBER	Новое «улучшенное» силовое поле, предназначенное для длительных молекулярно-динамических симуляций [Ahmed & Papaleo, 2018; Koes & Vries, 2017; Debiec et al., 2016]

TIP4P-Ew	SPC/Eb
	



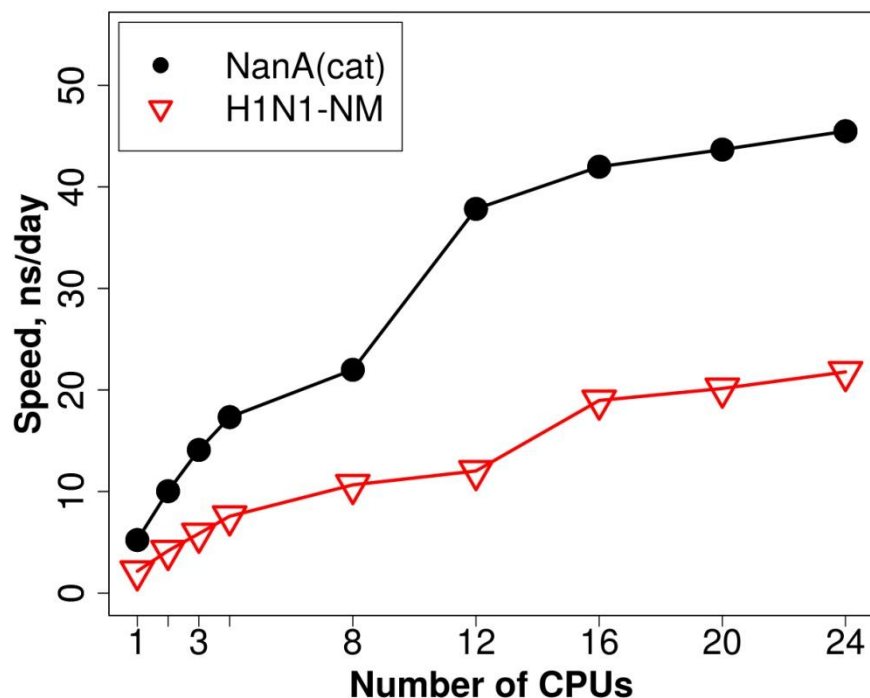
Использовались системы от **65 000 до 350 000 атомов** на основе пяти белков разных размеров, представляющих интерес как мишеней для дизайна новых лекарств от заболеваний человека:

- **MAPK** — митоген-активируемая протеинкиназа человека;
- **NanC** — нейраминидаза C из *Streptococcus pneumoniae*;
- **LDH** — лактатдегидрогеназа из *Staphylococcus aureus*;
- **H1N1-NM** — нейраминидаза вируса гриппа H1N1;
- **NanA** — нейраминидаза A из *Streptococcus pneumoniae*.

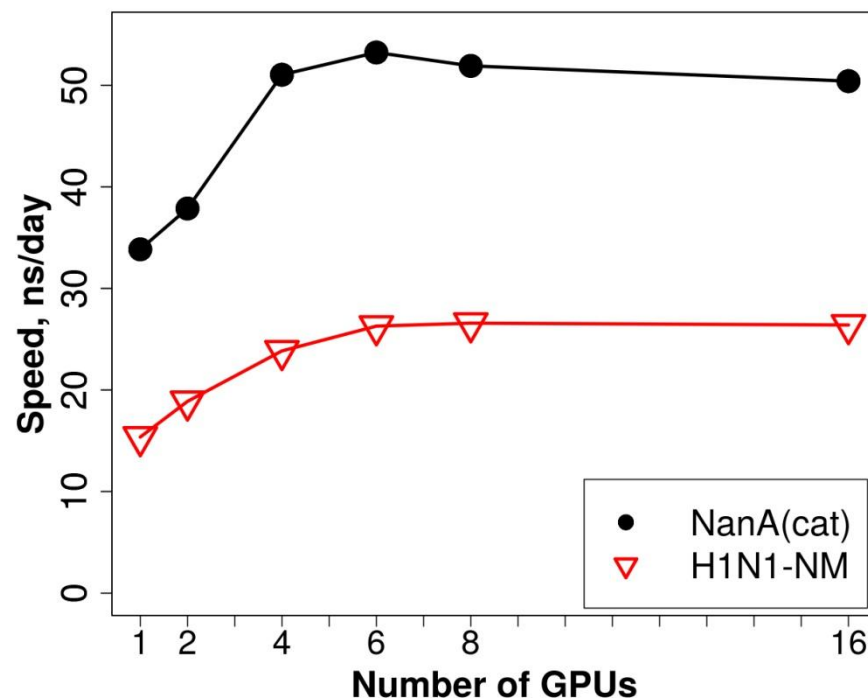
Суплатов, Д. А., Шарапова, Я. А., Попова, Н. Н., Копылов, К. Е., Воеводин, В. В., & Швядас, В. К. (2019). Молекулярная динамика в силовом поле FF14SB в воде TIP4P-EW, и в силовом поле FF15IPQ в воде SPC/Eb: сравнительный анализ на GPU и CPU. Вычислительная математика и информатика, 8(1), 71-88.

Молекулярная динамика на CPU vs GPU

Intel Xeon Gold 6126



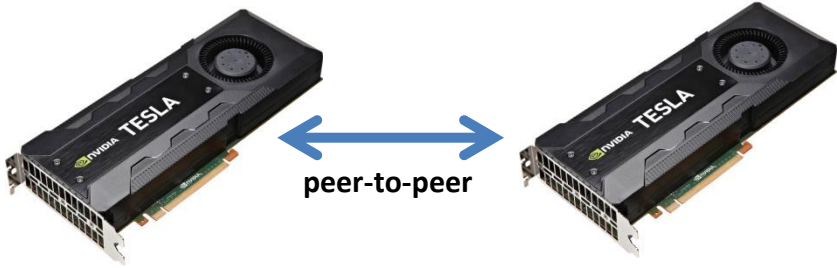
Tesla K40



Ускорение МД на CPU возможно, но не эффективно

Sharapova, Y. A., Suplatov, D. A., & Švedas, V. K. (2018). Simulating the Long-timescale Structural Behavior of Bacterial and Influenza Neuraminidases with Different HPC Resources. *Supercomputing Frontiers and Innovations*, 5(3), 30-33.

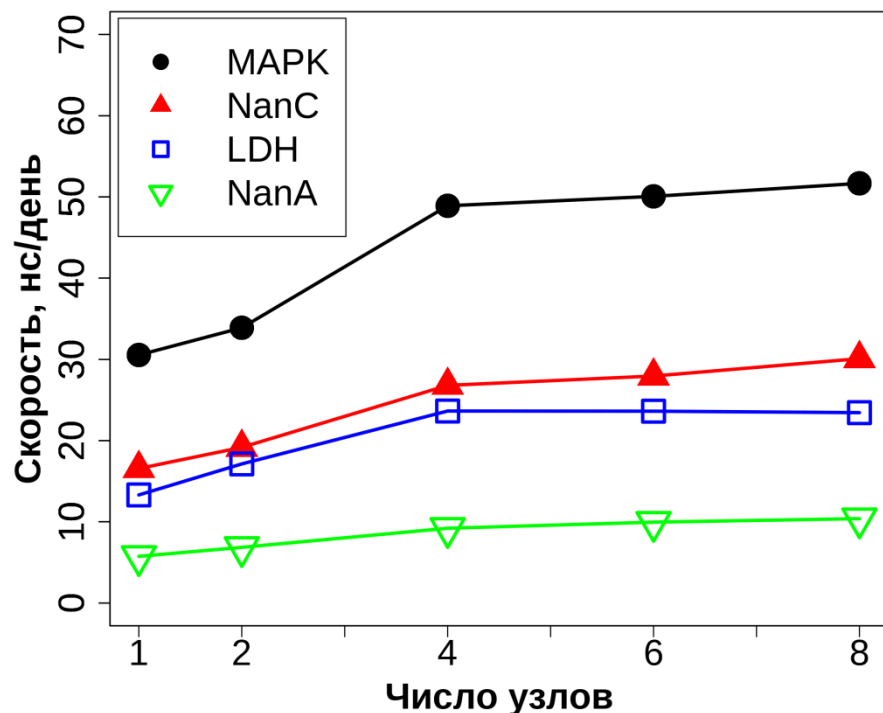
Доступное GPU-оборудование

Ломоносов-2 «Compute» (2016)	Ломоносов-2 «Pascal» (март 2018)
Intel Xeon E5-2697 v3 1 × Tesla K40  A single NVIDIA Tesla K40 GPU card, black with a fan, shown from a three-quarter perspective.	Intel Xeon Gold 6126 2 × Tesla P100  Two NVIDIA Tesla P100 GPU cards are shown side-by-side. A blue double-headed arrow connects them, with the text "peer-to-peer" below it, indicating direct communication between the two GPUs.

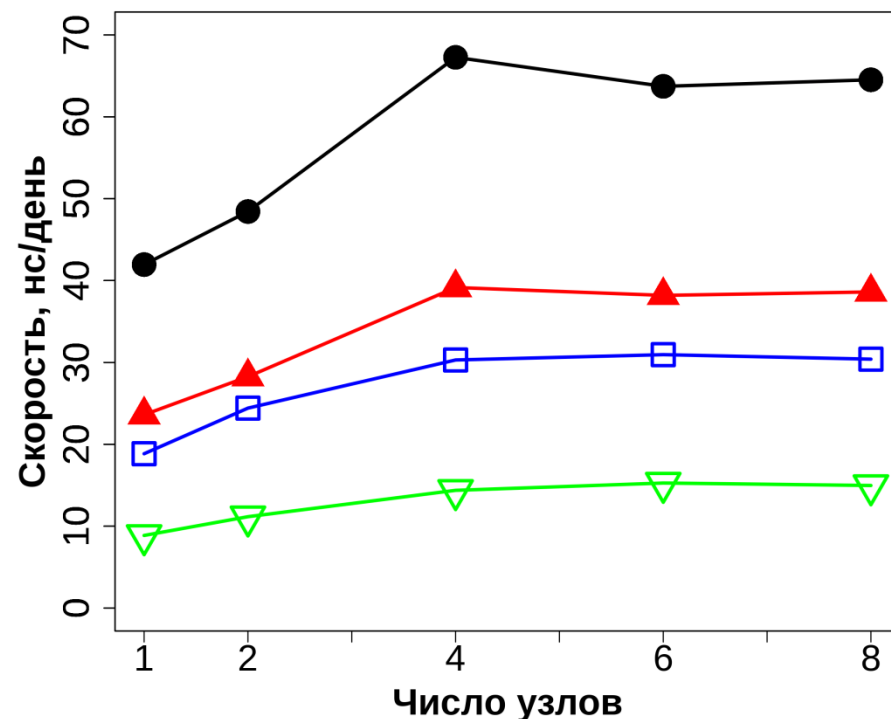
Молекулярная динамика на Tesla K40 GPU

биологических систем разных размеров и плотности

FF14SB + TIP4P-Ew



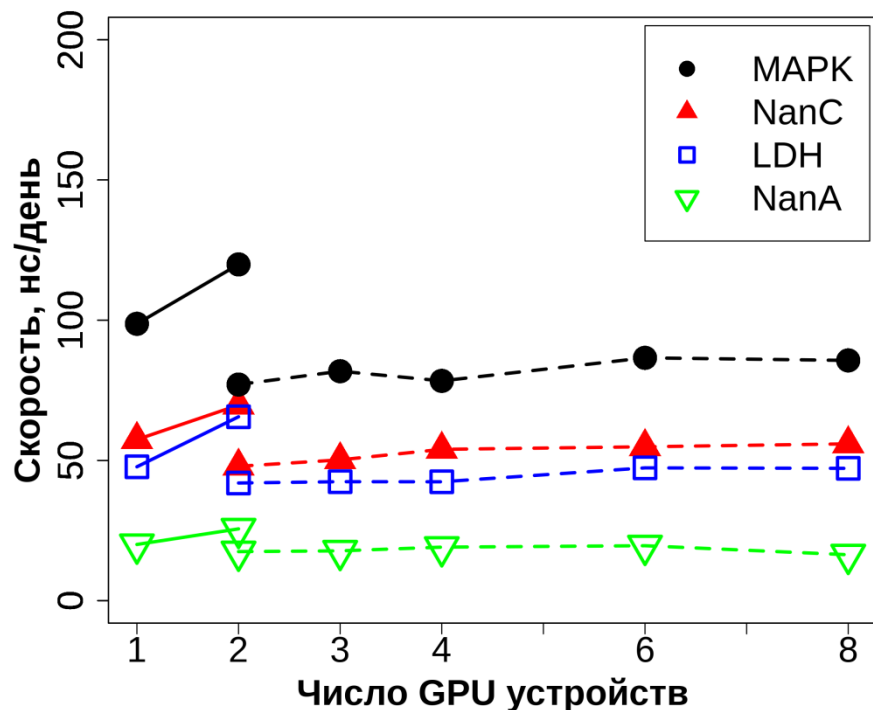
FF15IPQ + SPC/Eb



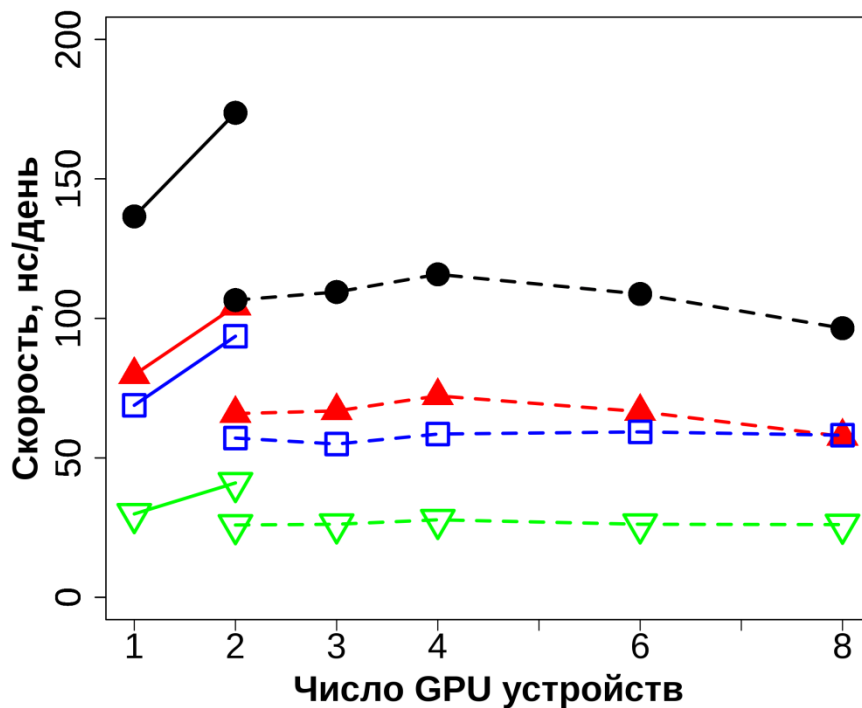
- Уменьшение количества атомов в ячейке в **×1,30** раза при одинаковой плотности ускоряет расчет МД в **×1.30** раза;
- Уменьшение количества атомов в ячейке в **×1,30** раза в результате использования 3-центральной модели растворителя (т.е. с уменьшением плотности) ускоряет расчет МД в **×1.60** раза;
- Использование 2-4 узлов с Tesla K40 ускоряет расчет по сравнению с одним узлом.

Молекулярная динамика на Tesla P100 GPU биологических систем разных размеров и плотности

FF14SB + TIP4P-Ew



FF15IPQ + SPC/Eb



Ускорение на двух Tesla P100 одного узла в режиме peer-to-peer по сравнению с одной Tesla P100 составило $\times 1,21 - \times 1,37$, однако производительность на двух и более узлах во всех случаях была существенно ниже производительности на одной видеокарте

Объем вычислений молекулярной динамики в повседневной лабораторной практике

Скорость расчет МД биологических систем наиболее распространенных размеров (65 000 – 150 000 атомов) на двух видеокартах Tesla P100 **одного узла** в режиме peer-to-peer составляет 73 – 150 нс/день, т.е. **микросекундная шкала может быть достигнута за 1-2 недели.**

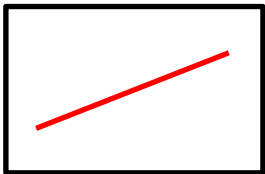
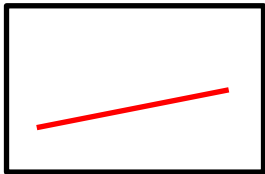
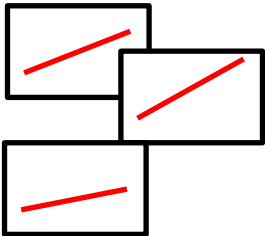
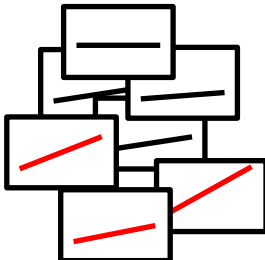
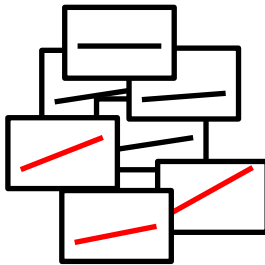
Это позволяет:

- считать существенно **более длинные траектории**;
- считать **больше количество независимых траекторий** (запущенных из разных начальных состояний или с разными начальными условиями), что критически важно, поскольку позволяет собрать больше данных для последующего анализа

Sharapova, Y. A., Suplatov, D. A., & Švedas, V. K. (2018). Simulating the Long-timescale Structural Behavior of Bacterial and Influenza Neuraminidases with Different HPC Resources. *Supercomputing Frontiers and Innovations*, 5(3), 30-33.

Суплатов Д.А., Шарапова Я.А., Попова Н.Н., Копылов К.Е., Воеводин Вл.В., Швядас В.К. Молекулярная динамика в силовом поле FF14SB в воде TIP4P-Ew, и в силовом поле FF15IPQ в воде SPC/Eb: сравнительный анализ на GPU и CPU // *Вестник ЮУрГУ. Серия: Вычислительная математика и информатика*. 2019 — №8 (1) – С. 71-88.

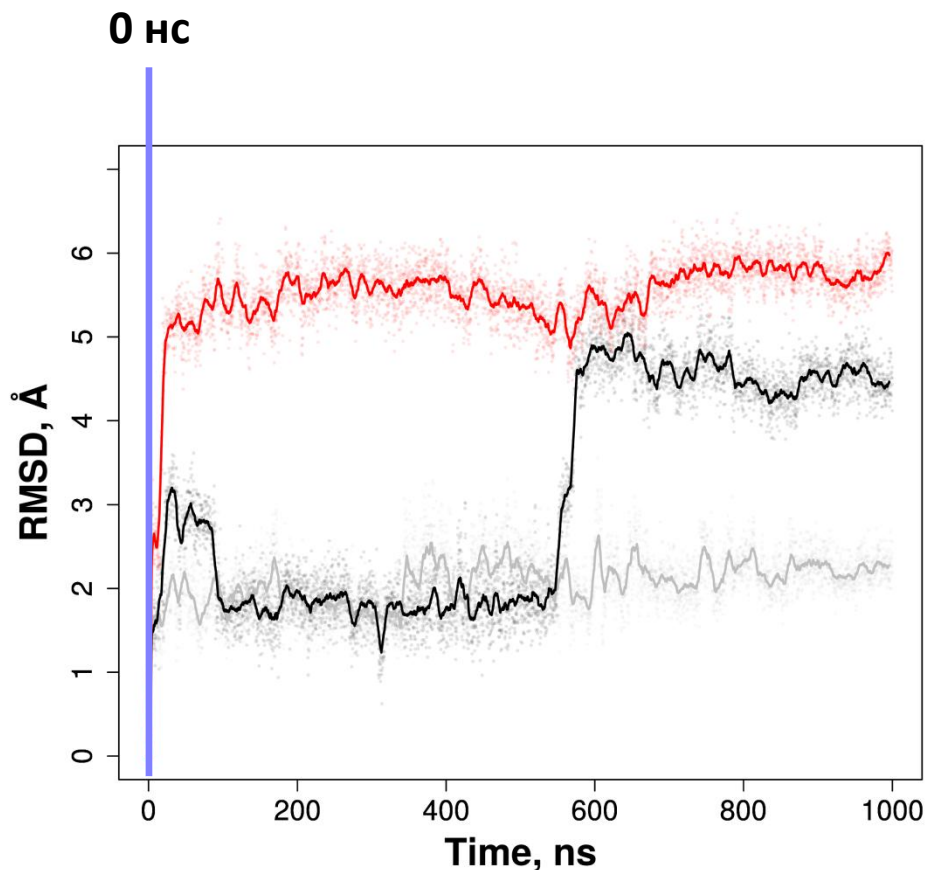
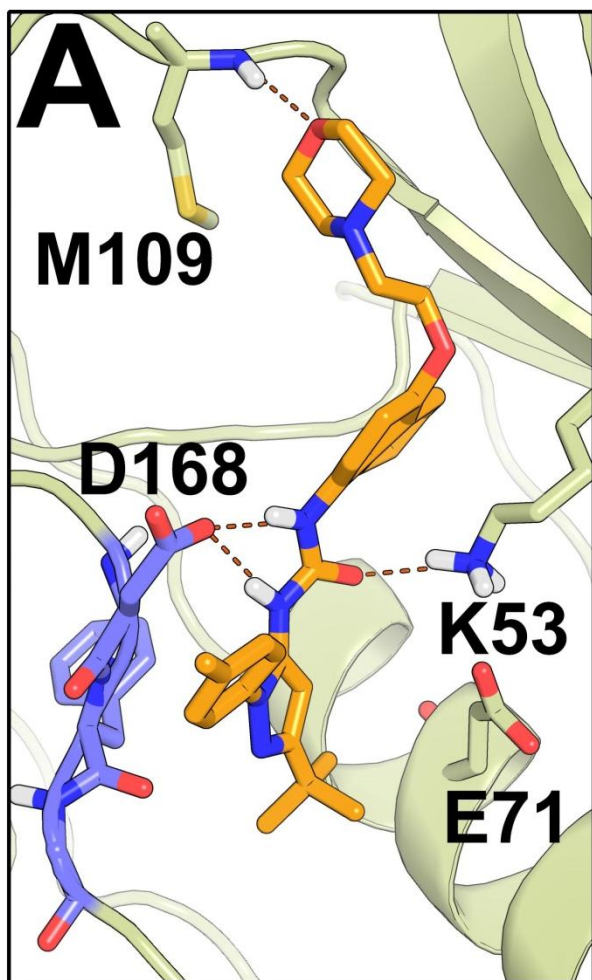
Объем вычислений молекулярной динамики в повседневной лабораторной практике

До 2000	2000 – 2008	2008 – 2014	2014-2017	2018
< 1 нс	10 нс	10 – 100 нс	100 нс	1000 нс
				
		Чебышев Ломоносов-1	Ломоносов-2 	Ломоносов-2 

Объем вычислений молекулярной динамики в повседневной лабораторной практике

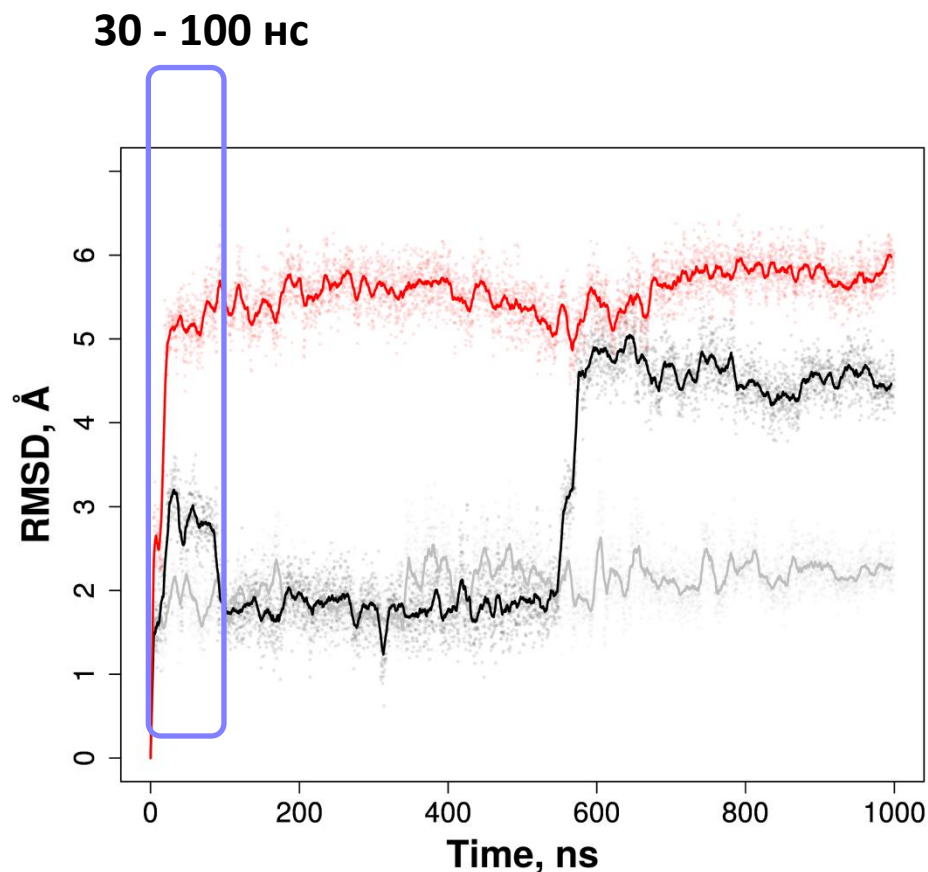
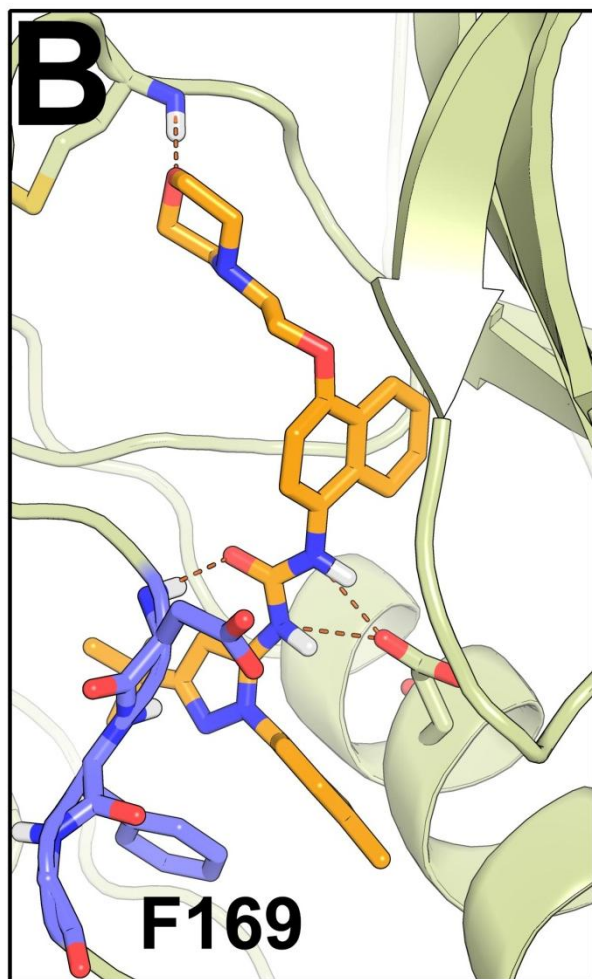
- Зачем считать длинные траектории?
 - Чтобы получить возможность увидеть и изучить конформационные перестройки, возникающие на больших временных шкалах;

Симуляция структурной подвижности комплекса р38α МАРК человека с ингибитором дорамапимодом



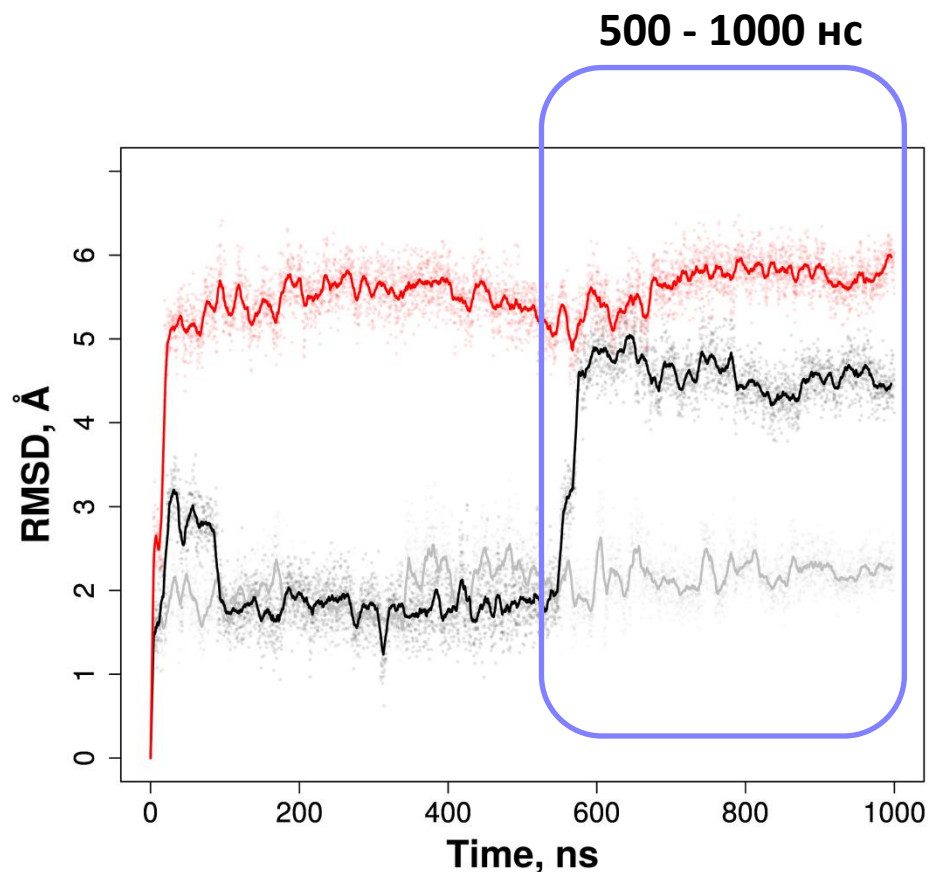
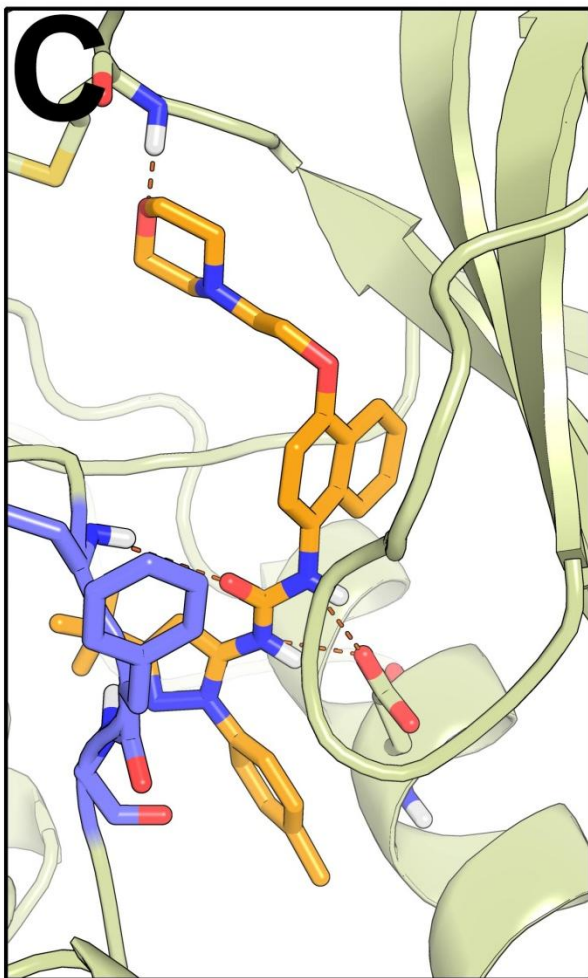
Suplatov, D. A., Kopylov, K. E., Sharapova Y.A., Švedas V. (2019) *J. Biomol. Struct. Dyn.*, 37(8), 2049-2060

Симуляция структурной подвижности комплекса р38α МАРК человека с ингибитором дорамапимодом



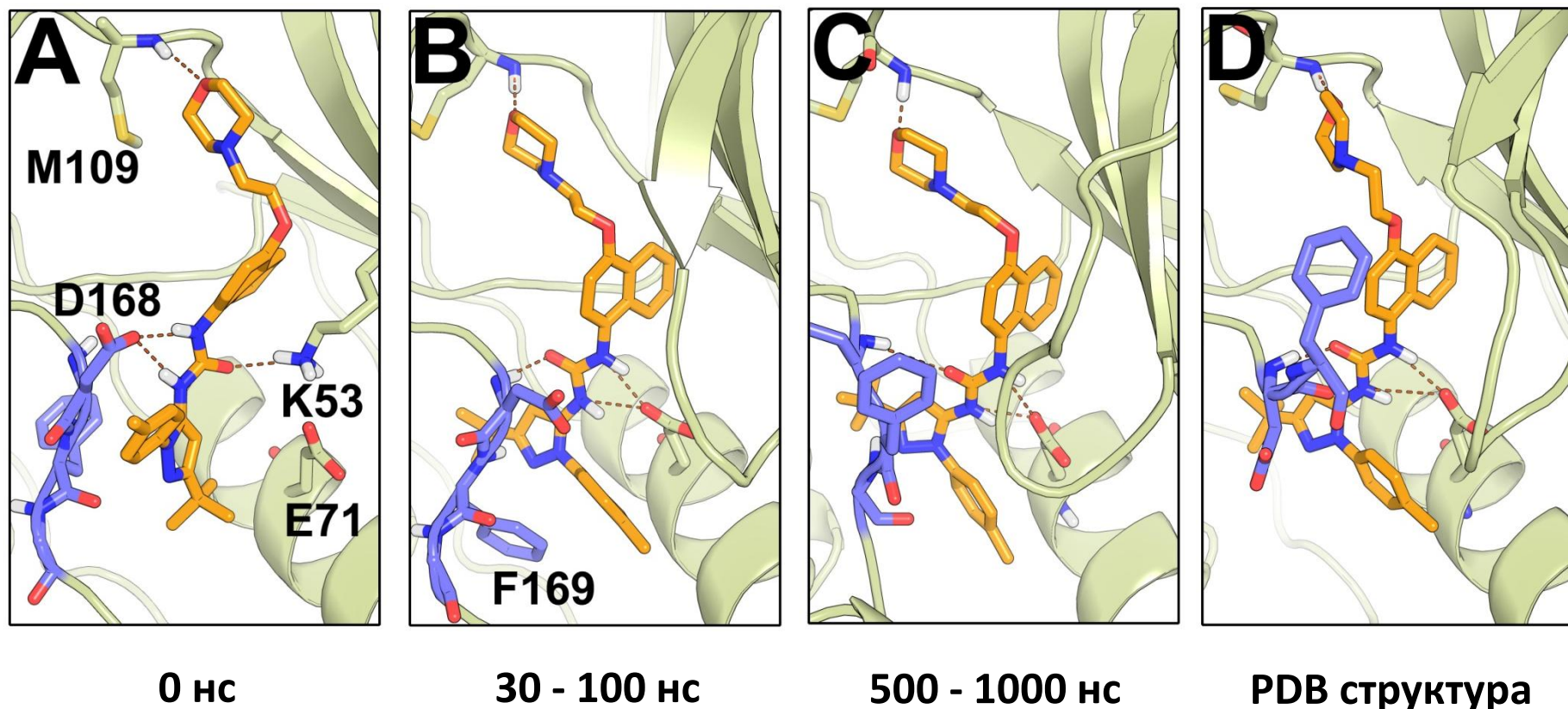
Suplatov, D. A., Kopylov, K. E., Sharapova Y.A., Švedas V. (2019) *J. Biomol. Struct. Dyn.*, 37(8), 2049-2060

Симуляция структурной подвижности комплекса р38α МАРК человека с ингибитором дорамапимодом



Suplatov, D. A., Kopylov, K. E., Sharapova Y.A., Švedas V. (2019) *J. Biomol. Struct. Dyn.*, 37(8), 2049-2060

Длительная симуляция структурной подвижности p38a MAPK человека позволила изучить последовательные стадии механизма связывания дорамапимода

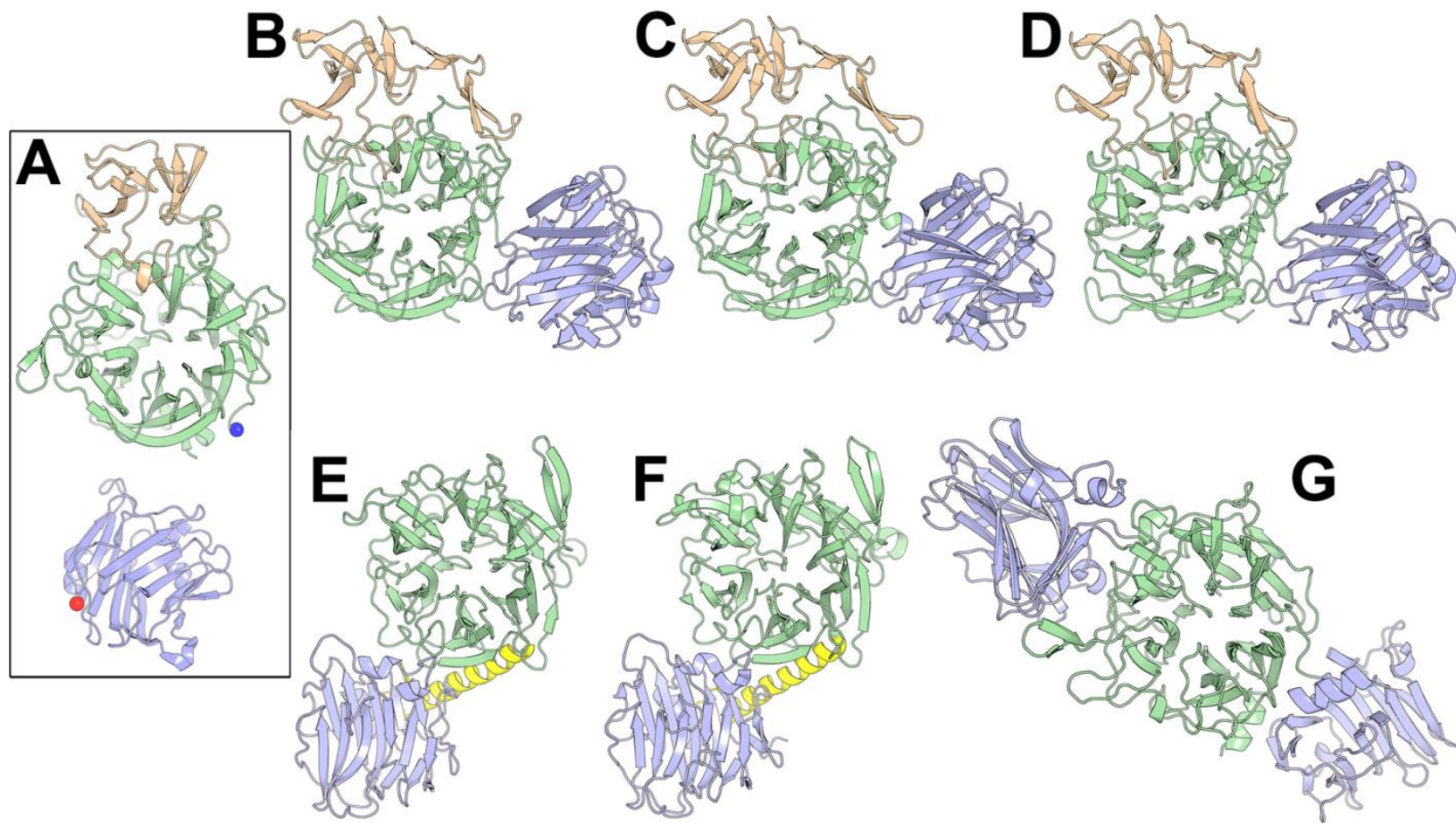


Suplatov, D. A., Kopylov, K. E., Sharapova Y.A., Švedas V. Human p38 α Mitogen-Activated Protein Kinase in the Asp168-Phe169-Gly170-in (DFG-in) state can bind allosteric inhibitor Doramapimod (2019) *J. Biomol. Struct. Dyn.*, 37(8), 2049-2060

Объем вычислений молекулярной динамики в повседневной лабораторной практике

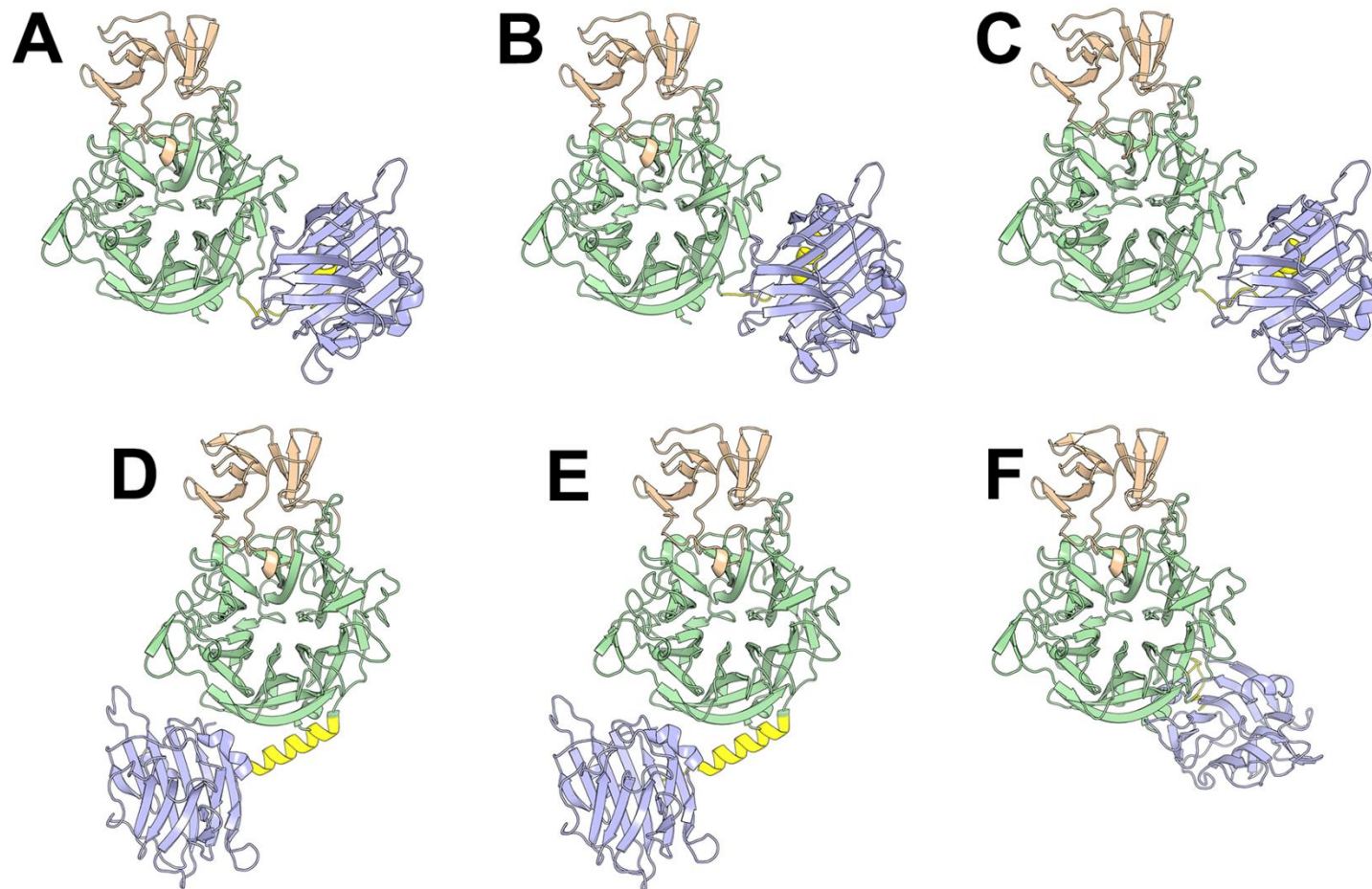
- Зачем считать несколько параллельных траекторий?
 - Для запуска траекторий с разными начальными условиями с целью собрать больше информации о конкретном объекте для последующего анализа;
 - Для запуска траекторий с разными параметрами/настройками молекулярной динамики (например, при разных температурах);
 - Для одновременного изучения разных объектов/моделей, что позволяет ускорить исследование в целом;

Симуляция структурной подвижности для изучения доменной организации бактериальных нейраминдаз



Sharapova Y., Suplatov D., Švedas V. Neuraminidase A from *Streptococcus pneumoniae* Has a Modular Organization of Catalytic and Lectin Domains Separated by a Flexible Linker. *The FEBS Journal*. 2018. vol. 285, no. 13. pp. 2428–2445. DOI: 10.1111/febs.14486.

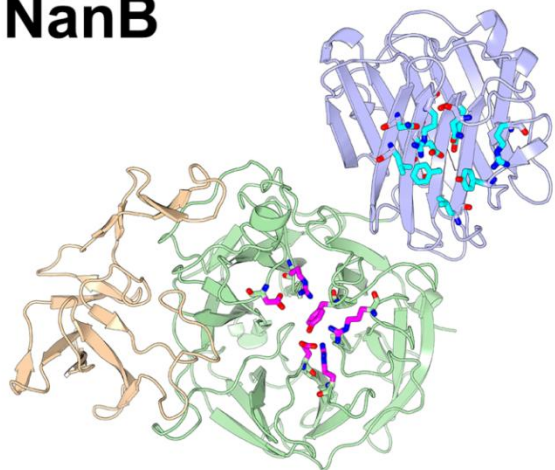
Симуляция структурной подвижности для изучения доменной организации бактериальных нейраминдаз



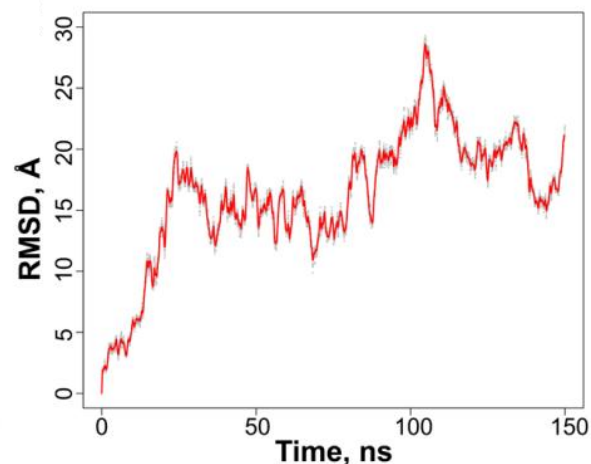
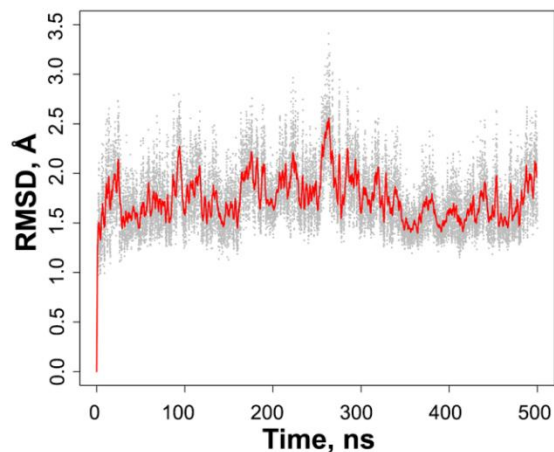
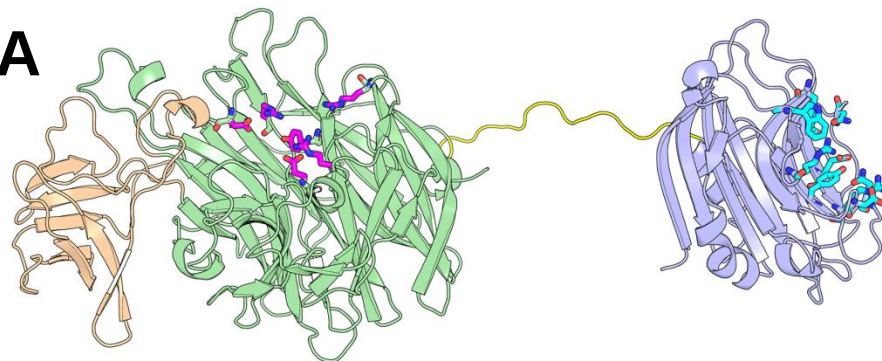
Sharapova Y., Suplatov D., Švedas V. Neuraminidase A from *Streptococcus Pneumoniae* Has a Modular Organization of Catalytic and Lectin Domains Separated by a Flexible Linker. *The FEBS Journal*. 2018. vol. 285, no. 13. pp. 2428–2445. DOI: 10.1111/febs.14486.

Симуляция структурной подвижности для изучения доменной организации бактериальных нейраминдаз

NanB

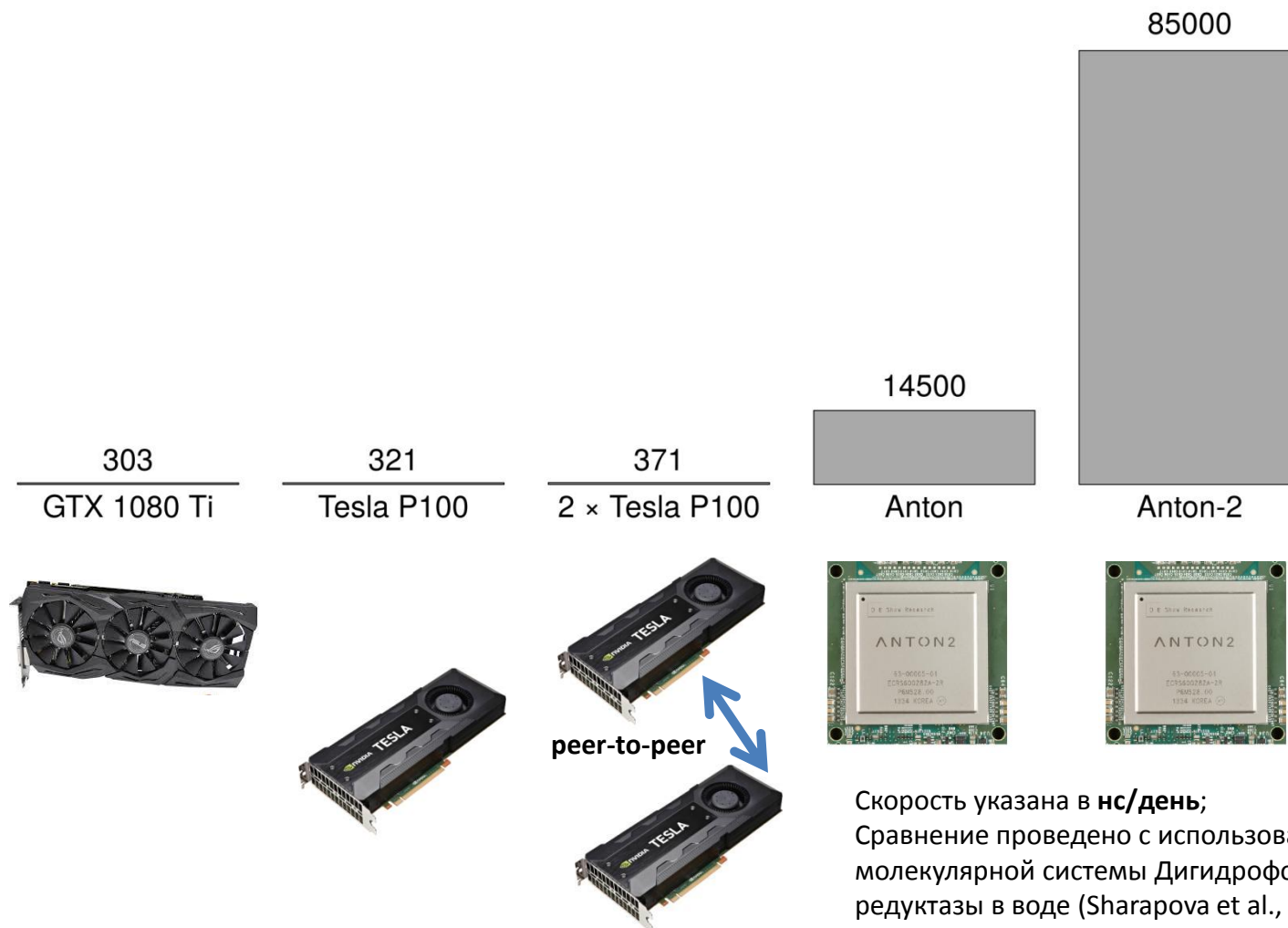


NanA

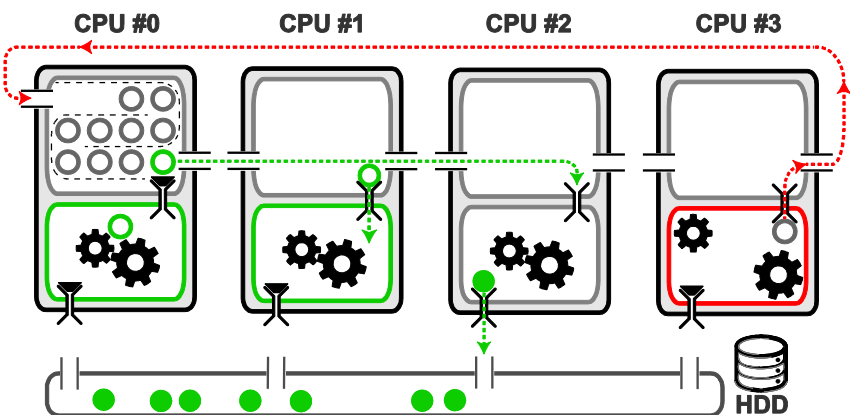


Sharapova Y., Suplatov D., Švedas V. Neuraminidase A from *Streptococcus Pneumoniae* Has a Modular Organization of Catalytic and Lectin Domains Separated by a Flexible Linker. *The FEBS Journal*. 2018. vol. 285, no. 13. pp. 2428–2445. DOI: 10.1111/febs.14486.

Сравнение скорости молекулярной динамики на Pascal GPU и Anton (2008) /Anton-2 (2014)



Высокопроизводительный *in silico* скрининг (докинг)



Программа **mpiWrapper** для запуска докинга программами **AutoDock** и **Leadfinder**:

- Докинг 1 лиганда в 1 сайт белка рассматривается как одна подзадача;
- Скрининг 1 сайта белка библиотекой из n лигандов рассматривается как набор из n независимых подзадач;

Функции программы **mpiWrapper** и комплекса вспомогательных подпрограмм:

- Равномерное распределение задач по узлам;
- Контроль технических ошибок на узлах (перестановка задач на другие узлы);
- Хранение и систематизация данных.

Скринирование больших библиотек рецепторов и лигандов:

- Использование программ **mpiWrapper** и **Leadfinder** позволяет вычислять десятки миллионов докингов в течение нескольких дней на 10-100 узлах суперкомпьютера «Ломоносов-2» (в зависимости от выбранных параметров).

Suplatov, D., Popova, N., Zhumatiy, S., Voevodin, V., & Švedas, V. (2016). Parallel workflow manager for non-parallel bioinformatic applications to solve large-scale biological problems on a supercomputer. *Journal of bioinformatics and computational biology*, 14(02), 1641008.

<https://biokinet.belozersky.msu.ru/mpiwrapper>

Уровень I + II

- **Биоинформатический анализ** позволяет
 - собрать и систематизировать информацию о суперсемейства белков/ферментов;
 - выбрать и классифицировать функционально важные аминокислотные остатки;
 - идентифицировать новые центры связывания;
- **Молекулярное моделирование** позволяет уточнить результаты биоинформатического анализа:
 - изучить роль ключевых остатков;
 - найти (создать) лиганды (субстраты, регуляторные лиганды/ингибиторы), комплементарные выбранным сайтам в структуре белка;
 - Изучить/уточнить механизмы взаимодействия белков/ферментов с субстратами и регуляторными лигандами.

Суплатов Д.А., Шмальгаузен Е.В., Муронец В.И., Швядас В.К. Селективные ингибиторы глицеральдегид-3-фосфатдегидрогеназы микобактерий.

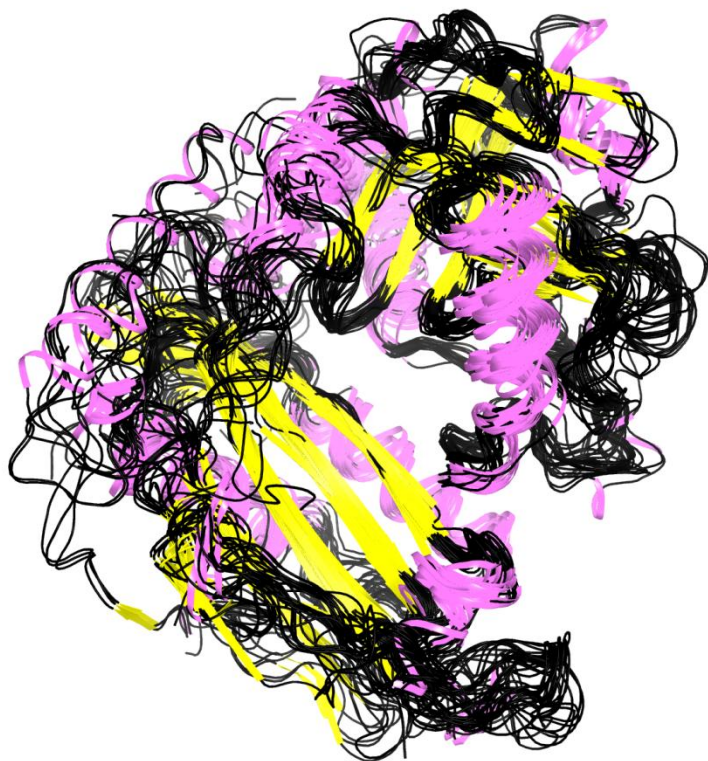
Патент № RU 2 661 151 от 12 июля 2018 г.

- Построено и проанализировано множественное выравнивание репрезентативной выборки из >10000 глицеральдегид-3-фосфатдегидрогеназ (ГАФД);
- С использованием биоинформатического анализа охарактеризован **новый аллостерический сайт** в структуре ГАФД микобактерий;
- С использованием молекулярного моделирования предложены селективные ингибиторы ГАФД микобактерий, которые связываются в новом аллостерическом сайте, и описан механизм их действия;
- Экспериментально подтверждена способность отобранных соединений **селективно ингибировать ГАФД микобактерий**, при этом не оказывая влияния на активность гомологичного фермента человека;
- Экспериментально подтверждена способность отобранных соединений **подавлять рост возбудителя туберкулеза** *Mycobacterium tuberculosis* штамм H37Rv

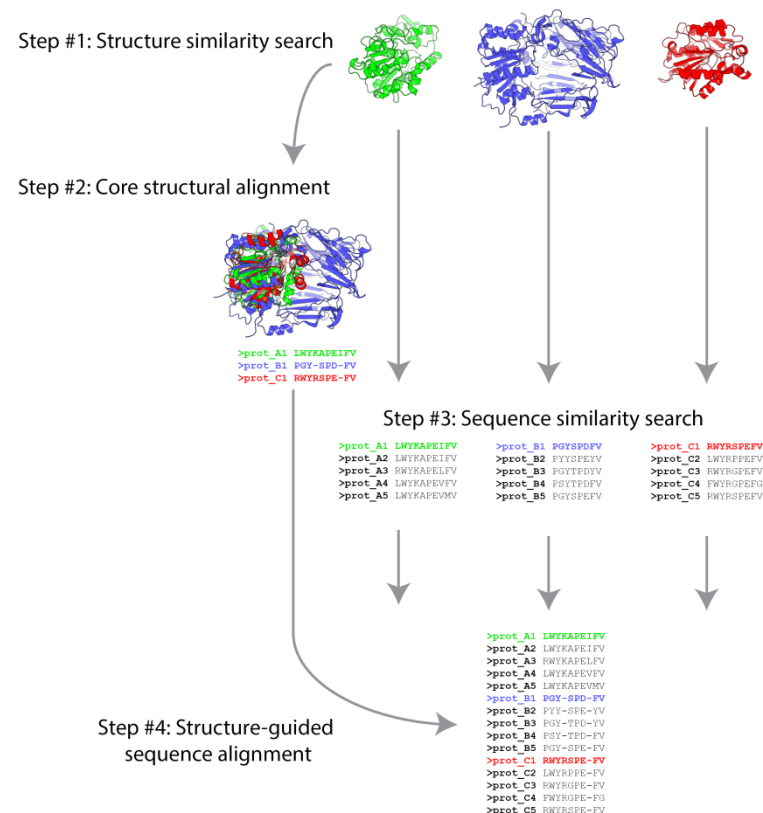
Fesko K., Suplatov D., Švedas V. Bioinformatic analysis of the fold type i plp-dependent enzymes reveals determinants of reaction specificity in l-threonine aldolase from aeromonas jandaei // *FEBS open bio.* — 2018. — Vol. 8. — no. 6. — P. 1013–1028

- Построено и проанализировано множественное выравнивание репрезентативной выборки из 9000 PLP-зависимых ферментов I фолд-типа;
- С использованием биоинформатического анализа предсказаны аминокислоты, влияющие на «реакционную специфичность» (тип катализируемой реакции) в суперсемействе;
- Роль отобранных позиций в катализе изучена сайт-направленным мутагенезом;
- Получены препараты (мутанты) **треонинальдолазы** с **рацемазной** активностью за счет внесения мутаций в позиции, предсказанные биоинформатическим анализом.

Дизайн дисульфидных связей в PLP-зависимых ферментах



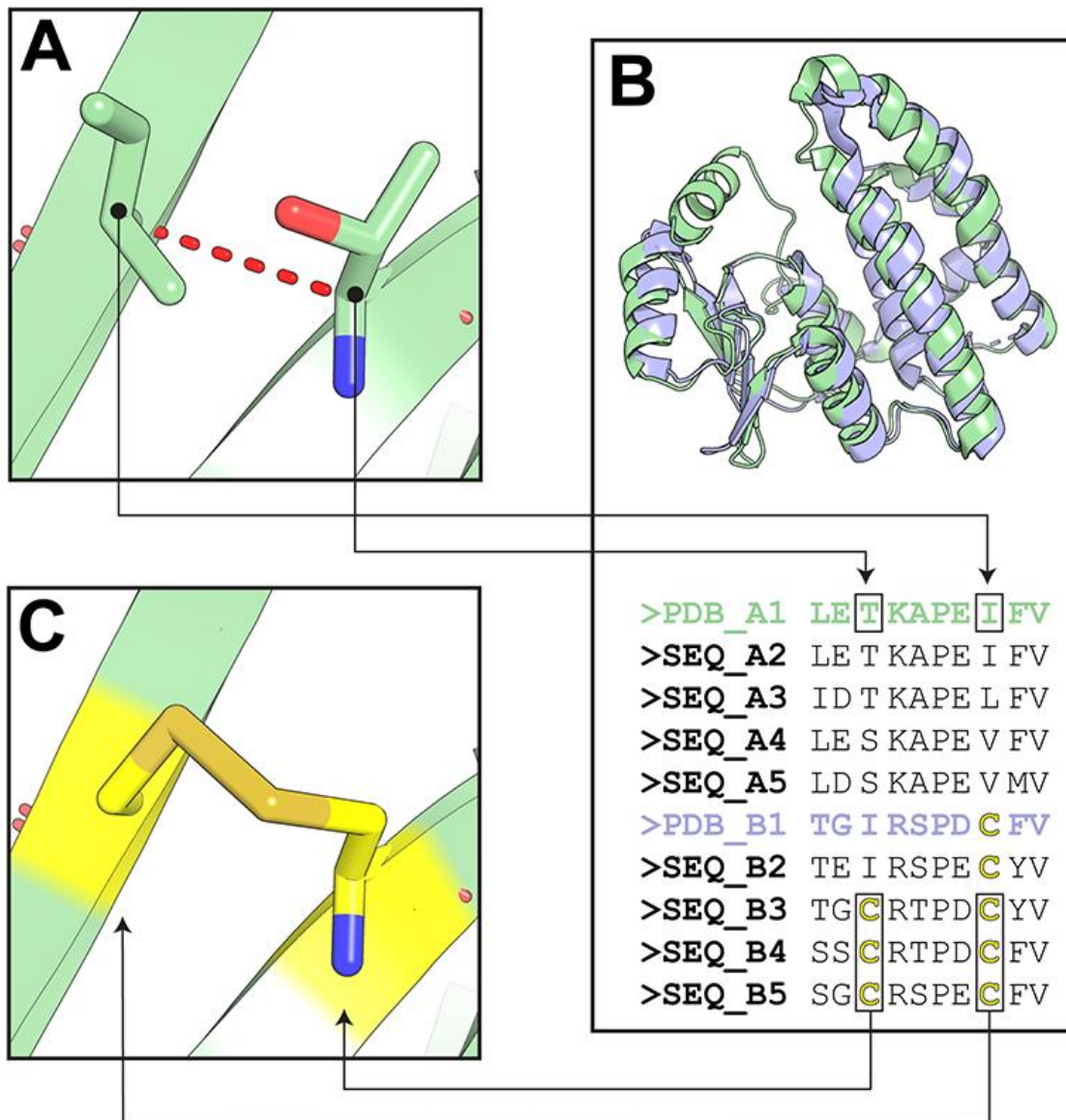
Структурное выравнивание
репрезентативной выборки из
186 представителя суперсемейства
PLP ферментов



<https://biokinet.belozersky.msu.ru/mustguseal>

Структурно-опосредованное
множественное выравнивание
299'976 последовательностей и структур
PLP ферментов

Дизайн дисульфидных связей в RLP-зависимых ферментах



- Введение дисульфидной связи, которая отсутствует в треонинальдолазе из *Aeromonas jandaei*, но присутствует в структурах ее гомологов в суперсемействе, привела к **значительному увеличению активности и стабильности мутантного препарата** по сравнению с ферментом дикого типа

Использованный метод:
 Suplatov D.A., Timonina D.S., Sharapova Y.A., Švedas V.K. (2019) Yosshi: a web-server for disulfide engineering by bioinformatic analysis of diverse protein families. *Nucleic Acids Res.*, DOI:10.1093/nar/gkz385

Уровень III: Автоматизированный интеллектуальный анализ информационных ресурсов, таких как статьи, патенты, абстракты конференций и веб-сайты научных лабораторий, для экстракции и систематизации разрозненной экспериментальной информации о функциях и свойствах белков

Интеллектуальный анализ информационных ресурсов

- Автоматизированное извлечение информации из доступных текстов (статей, патентов, баз данных);
- Полученная информация будет полезна для анализа и интерпретации результатов уровней I и II.

Интеллектуальный анализ информационных ресурсов

- Объем доступных статей по естественно-научной тематике ~ 50 TB;
- Объем доступных статей по отдельному направлению ~ 50 GB
 - «disulfide bonds»: 37968 статей
 - «neuraminidases/sialidases»: 21787 статей
 - «beta-lactamases»: 34653 статей
- Текущая задача – автоматизированное извлечение информации об экспериментальном эффекте точечных мутаций

Благодарности

Участники междисциплинарного проекта

НИИ ФХБ имени А.Н.Белозерского МГУ, ФББ МГУ,

- Швядас Витас
- Суплатов Дмитрий Андреевич
- Шарапова Яна Александровна
- Тимонина Дарья Сергеевна
- Панин Николай Владимирович
- Дробот Виктор Валерьевич
- Никулин Максим Владимирович
- Муронец Владимир Израилевич
- Шмальгаузен Елена Викторовна

НИВЦ МГУ, ВМК МГУ

- Воеводин Владимир Валентинович
- Попова Нина Николаевна
- Лукашевич Наталья Валентиновна
- Шегай Максим Викторович

**Institute of Organic Chemistry,
Graz University of Technology (Австрия)
Kateryna Fesko**

Благодарности

Финансовая поддержка

РФФИ №18-29-13060 «Установление взаимосвязи между структурой и функциональными свойствами белков/ферментов на основе интеллектуального анализа информационных ресурсов и разработка предсказательных алгоритмов»

РФФИ №17-07-00751 «Поиск новых решений ресурсоемких задач биоинформатики и молекулярного моделирования с использованием GPU-ускорителей»

РНФ №15-14-00069 «Изучение особенностей структурной организации центров связывания в ферментах патогенов и разработка новых антибиотиков для лечения заболеваний нижних дыхательных путей с использованием системных подходов биоинформатики и молекулярного моделирования»

РФФИ №14-07-00437 «Технологии суперкомпьютерного кодизайна для решения ресурсоемких вычислительных задач биокатализа»