Факультет биоинженерии и биоинформатики,
Московский государственный университет имени М. В. Ломоносова

Функция. Эволюция

Лекция 5, биоинформатика, 4 курс ФББ МГУ, осенний семестр
Злобин А. С., alexander.zlobin@fbb.msu.ru

Пара слов об RMSD и прочем

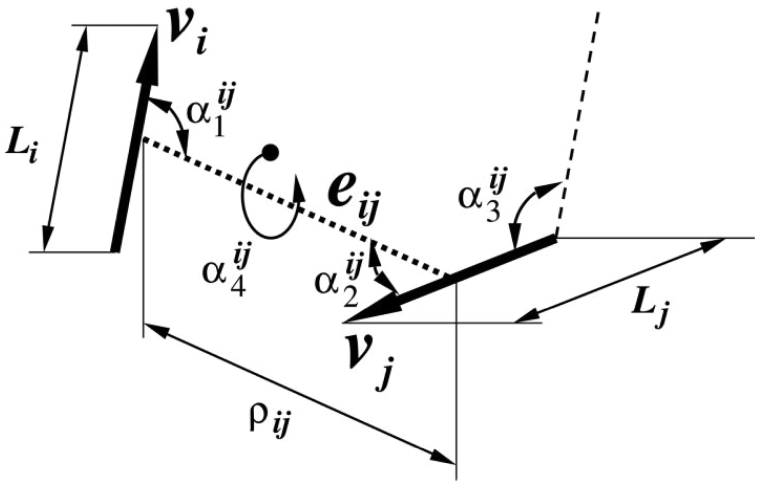
RMSD, eRMSD, TM-score, GDT_TS и прочие могут быть применены в двух основных сценариях:

- Для численной оценки “похожести” разных конформаций одного и того же белка
- Для численной оценки “похожести” разных белков как результат выбора оптимального структурного выравнивания

Алгоритм PDBeFold (SSM)

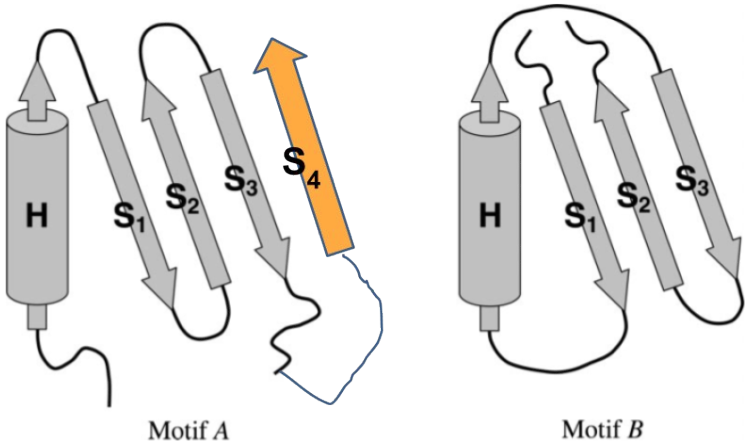
Сравниваем между собой взаиморасположения пар элементов вторичной структуры

- Отдаленно напоминает DALI – там мы сравнивали взаиморасположения пар гексамеров
- Однако в DALI мы описывали их как матрицы расстояний
- А в SSM мы присвоим каждому элементу вторичной структуры вектор из начала в конец и зададим их взаиморасположение набором чисел



Будем использовать эти наборы, чтобы решать, сопоставляется ли пара элементов из белка А с парой элементов из белка В, т.е. похожее ли между элементами взаиморасположение в двух структурах

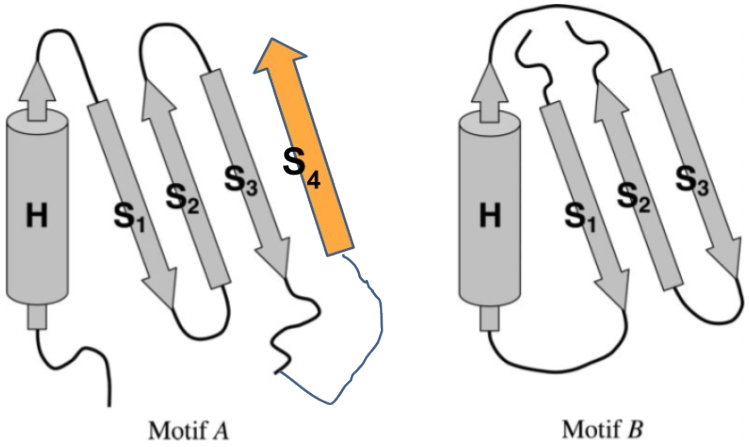
Алгоритм PDBeFold (SSM)



Это модифицированный пример из оригинальной статьи. Я добавил к первой структуре тяж S_4. Вопрос: какие элементы вторичной структуры тут расположены примерно одинаково и в A, и в B?

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1)

Алгоритм PDBeFold (SSM)



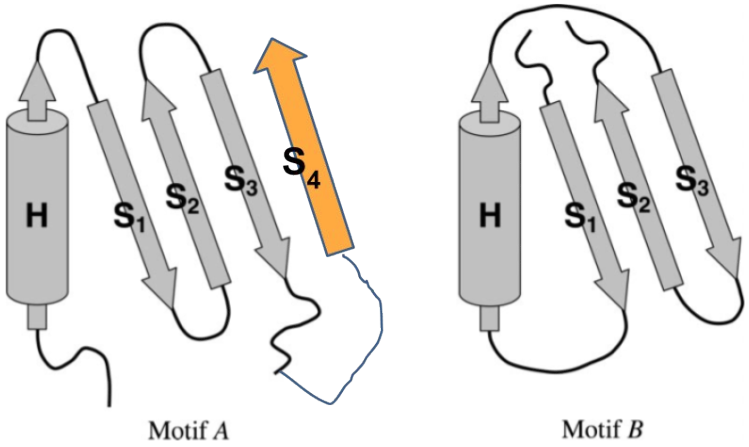
Это модифицированный пример из оригинальной статьи. Я добавил к первой структуре тяж S_4. Вопрос: какие элементы вторичной структуры тут расположены примерно одинаково и в A, и в B?

Ответ: H, S_1, S_2, S_3.

1. Рассмотрим все пары элементов из первой структуры:
   (H, S_1), (H, S_2), (H, S_3), (H, S_4), (S_1, S_2), (S_1, S_3), (S_1, S_4),
   (S_2, S_3), (S_2, S_4), (S_3, S_4)
   + еще 10 (S_1, H), (S_2, H) и т.д.
2. Рассмотрим все пары элементов из второй структуры:
   (H', S_1'), (H', S_2'), (H', S_3'), (S_1', S_2'), (S_1', S_3'),
   (S_2', S_3')
   + еще 6 (S_1, H), (S_2, H) и т.д.
3. Рассмотрим двойки пар из п. 1 и 2 (например, (H, S_1) + (H', S_3') и проч.)
   Вопрос: сколько всего таких двоек?
   Ответ: 20 x 12 = 240.
   Вопрос: Все ли такие двойки пар соответствуют парам элементов, которые одновременно могут быть выровнены?

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1)

Алгоритм PDBeFold (SSM)



Рассмотрим двойки пар (например, (H, S_1) + (H', S_3') и проч.) – это гипотезы об одновременном выравнивании пар элементов

Вопрос: Все ли такие двойки пар соответствуют парам элементов, которые одновременно могут быть выровнены?

(H, S_1) + (H', S_3')
Может, т.к. спираль соответствует спирали, а тяж - тяжу

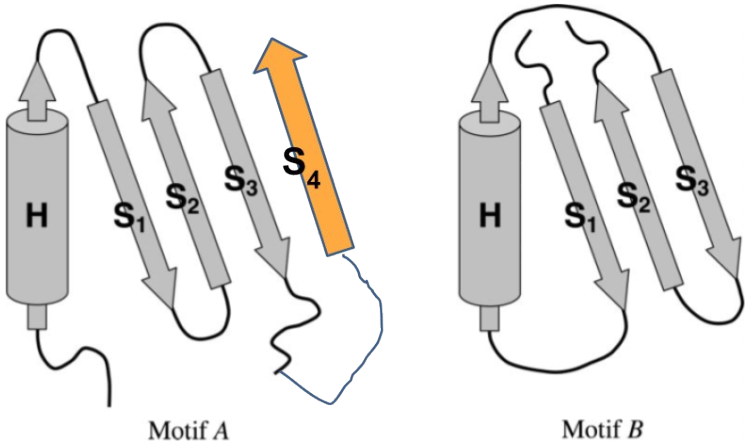
(H, S_1) + (S_1', S_3')
Не может, т.к. спираль не может быть выровнена с β-тяжем

(S_1, S_2) + (S_1', S_3')
Может

Из рисунка видно, что не все эти двойки реально выровнены. Но алгоритм этого пока не знает – это надлежит установить.

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1)

Алгоритм PDBeFold (SSM)



$(H, S_1) + (H', S_1')$

$(H, S_2) + (H', S_2')$

$(H, S_3) + (H', S_3')$

$(S_1, S_2) + (S'_1, S'_2)$

$(S_1, S_3) + (S'_1, S'_3)$

$(S_2, S_4) + (S'_1, S'_3)$

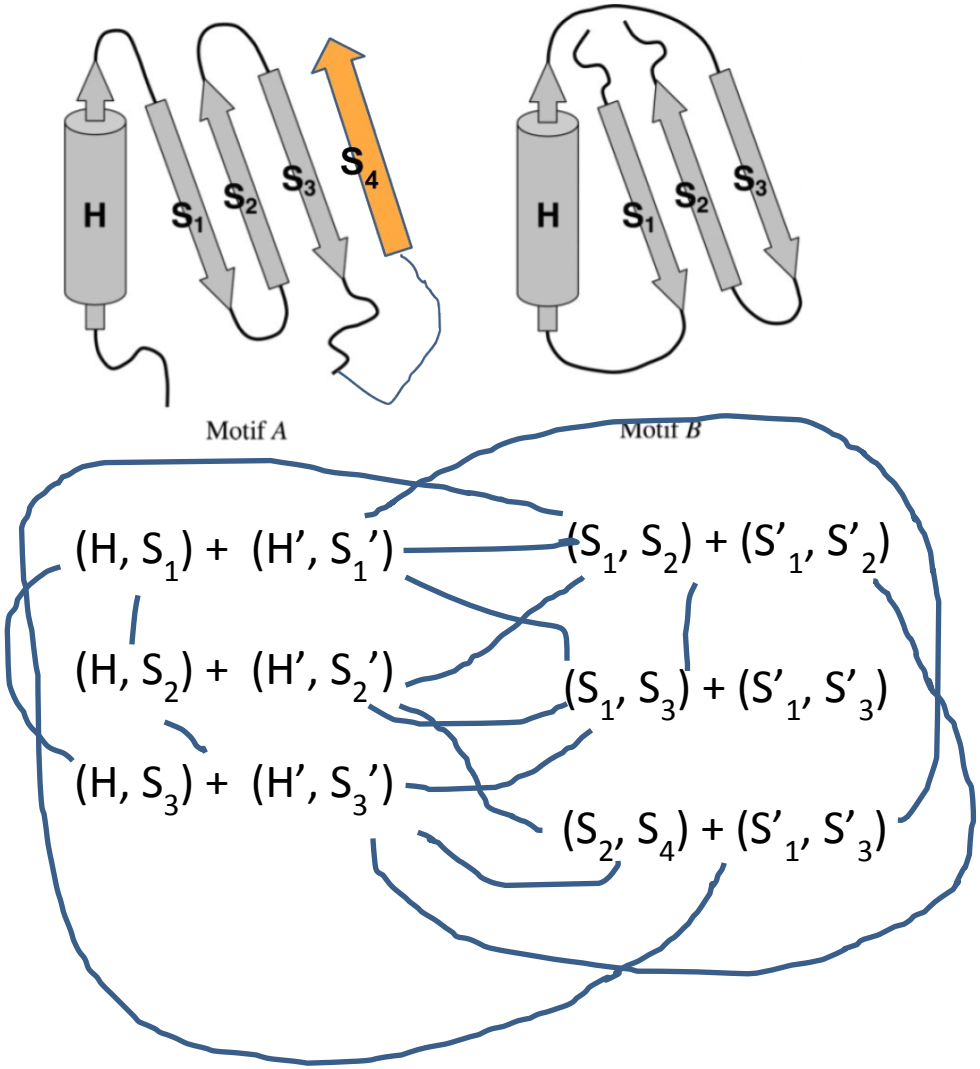
Отберем из всех двоек вида
те, которые потенциально могут быть выровнены. Это будут вершины графа. (На рисунке ниже показаны далеко не все вершины.)

Часть вершин соответствует правильному выравниванию. Часть – нет, но мы этого пока не знаем. Такова, например, двойка $(S_2, S_4) + (S'_1, S'_3)$. Как видно из рисунка в правильном выравнивании такой двойки нет. Но вообще говоря, это две пары тяжей, идущих в одном направлении, примерно с одинаковым расстоянием между ними.

Теперь соединим ребрами те вершины, между которыми нет противоречия.

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1)

Алгоритм PDBeFold (SSM)



Теперь соединим ребрами те вершины, между которыми нет противоречия.

Например, двойки
$(H, S_1) + (H', S_1')$ и
$(H, S_3) + (H', S_3')$
вполне могут быть в одном выравнивании.

А двойки
$(S_1, S_3) + (S'_1, S'_3)$ и
$(S_2, S_4) + (S'_1, S'_3)$
- нет. Действительно, S'_1 не может быть одновременно выровнен
и с S_1, и с S_2.

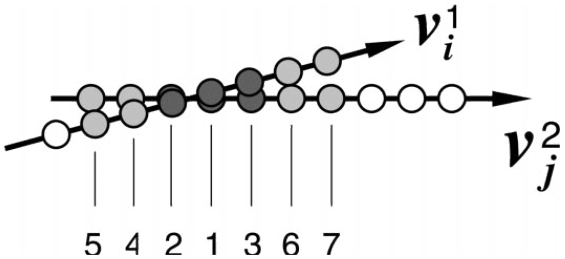
Клика в таком графе соответствует максимальному множеству выровненных элементов.

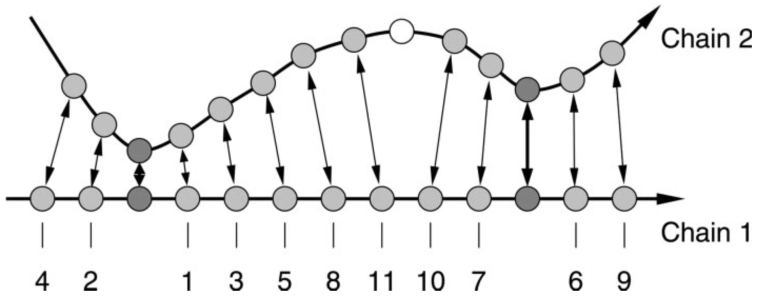
Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1)

Алгоритм PDBeFold (SSM)

Далее, заменим каждый SSE на точку (его центр). Поскольку у нас уже есть списки совпадающих элементов, эти точки можно совместить. Получится черновое совмещение структур.

Теперь надо перейти от выравнивания элементов вторичной структуры к выравниванию последовательностей.





Для совпадающих SSE (см. предыдущий этап) выбирает 3-4 аминокислоты, наиболее близких друг к другу. Затем выравнивание расширяется на весь тяж или спираль.

Для всех остальных алгоритм находит взаимно наиболее близкие C_α атомы. Соседние с ними пары атомов также считаются выровненными.

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1)

Алгоритм PDBeFold (SSM)

Далее, заменим каждый SSE на точку (его центр). Поскольку у нас уже есть списки совпадающих элементов, эти точки можно совместить. Получится черновое совмещение структур.

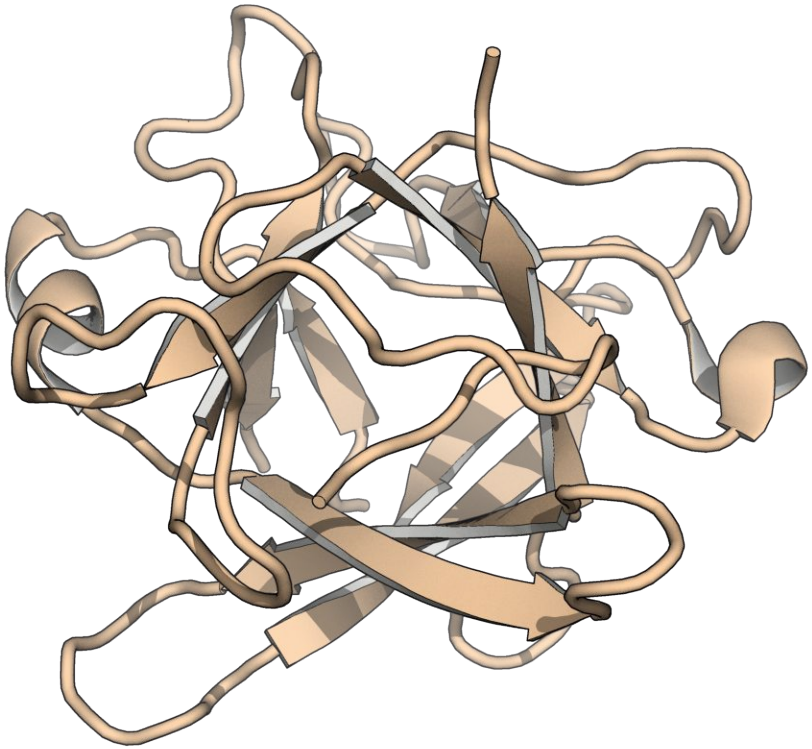
Теперь можно перейти от выравнивания элементов вторичной структуры к выравниванию последовательностей. При этом близко расположенные C_α атомы считаются выровненными и это выравнивание распространяется на их соседей по полипептидной цепочке.

Теперь у нас есть **черновое выравнивание** последовательностей. Ему соответствует некое совмещение структур, длина выравнивания, RMSD и Q-score. Его можно **оптимизировать**: менять разными способами, строить новое совмещение и добиваться увеличения Q. И так много-много раз, пока **решение не стабилизируется**.

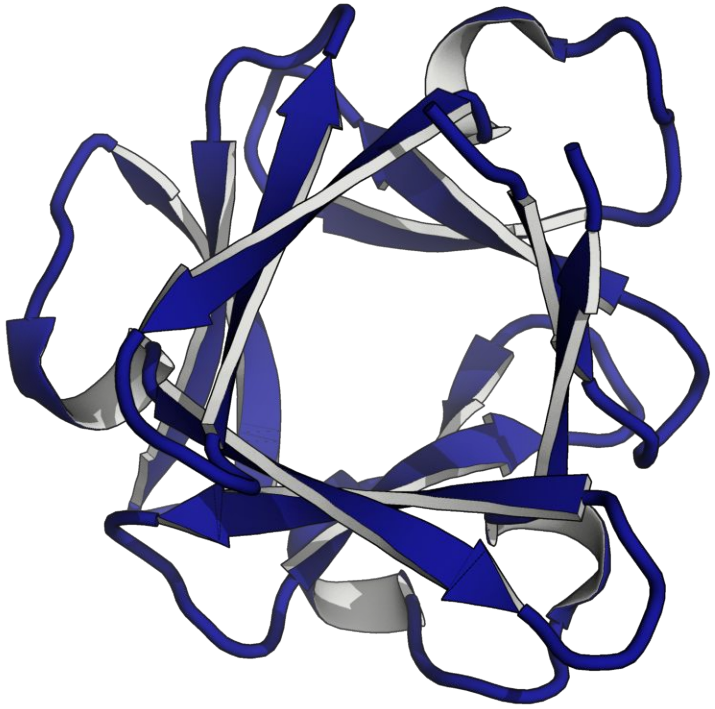
$$Q = \frac{N_{align}^2}{(1 + \frac{RMSD}{R_0})N_1 N_2)}$$

Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr. 2004 Dec;60(Pt 12 Pt 1)

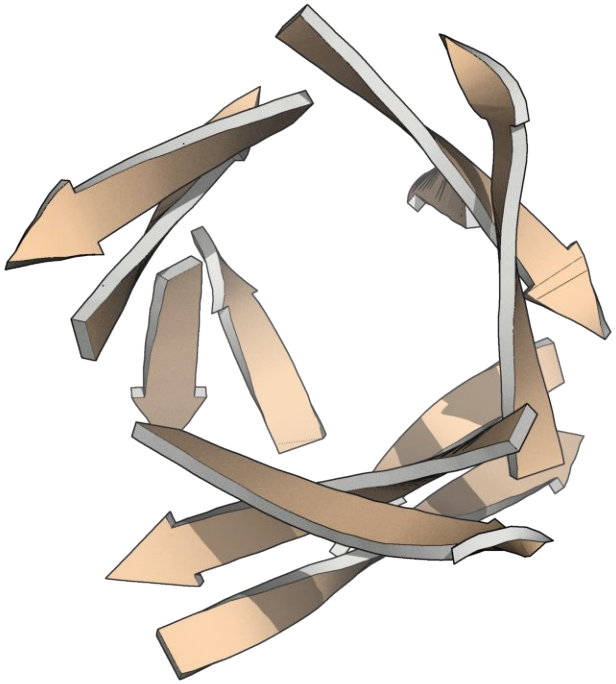
Структурный поиск



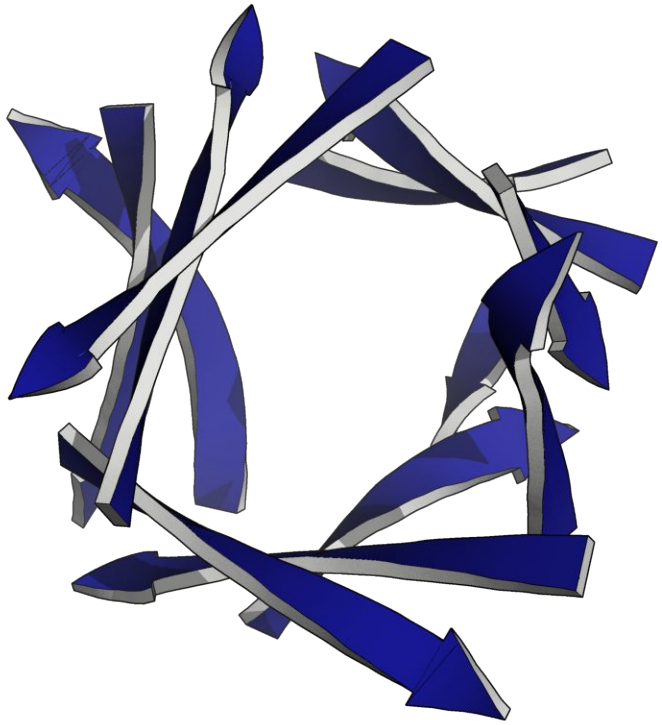
1TIE



4FGF

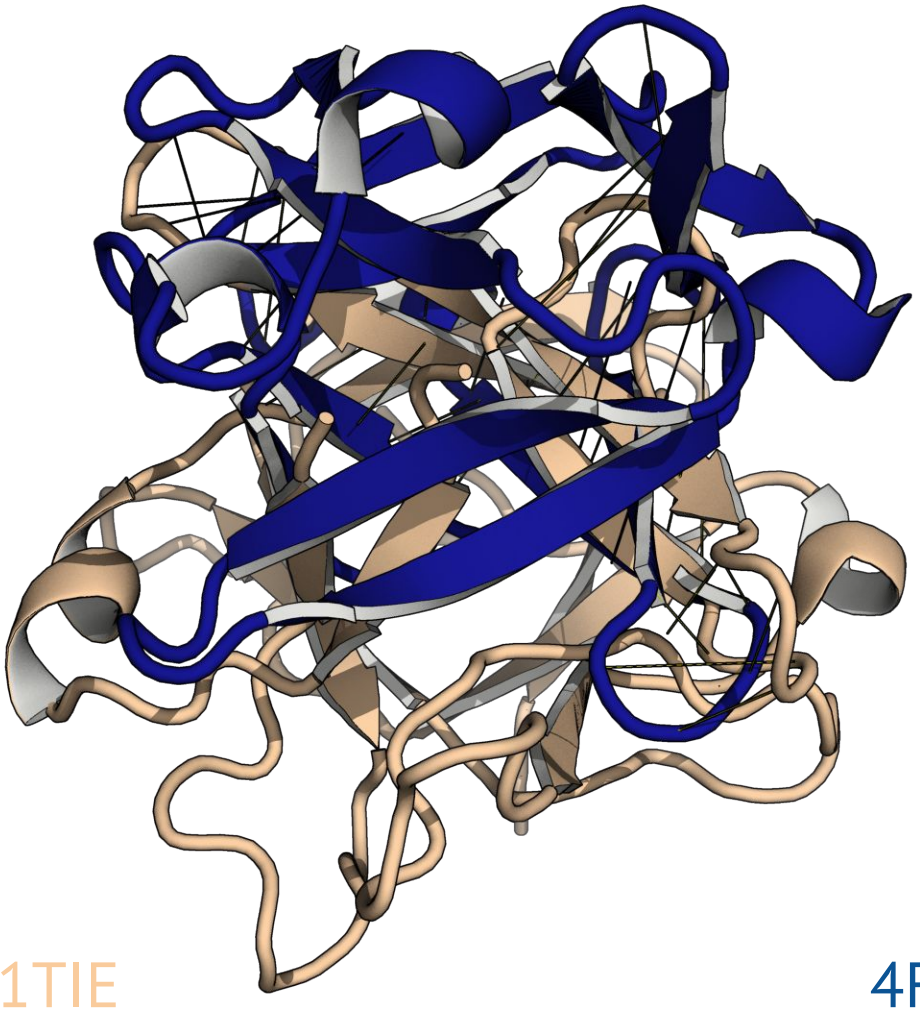


1TIE



4FGF

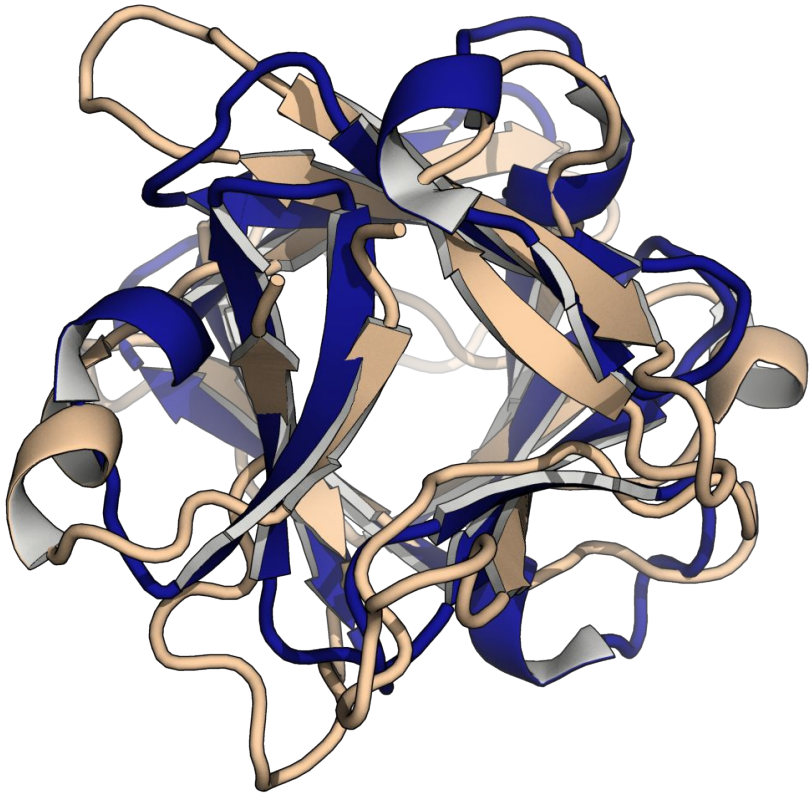
Align

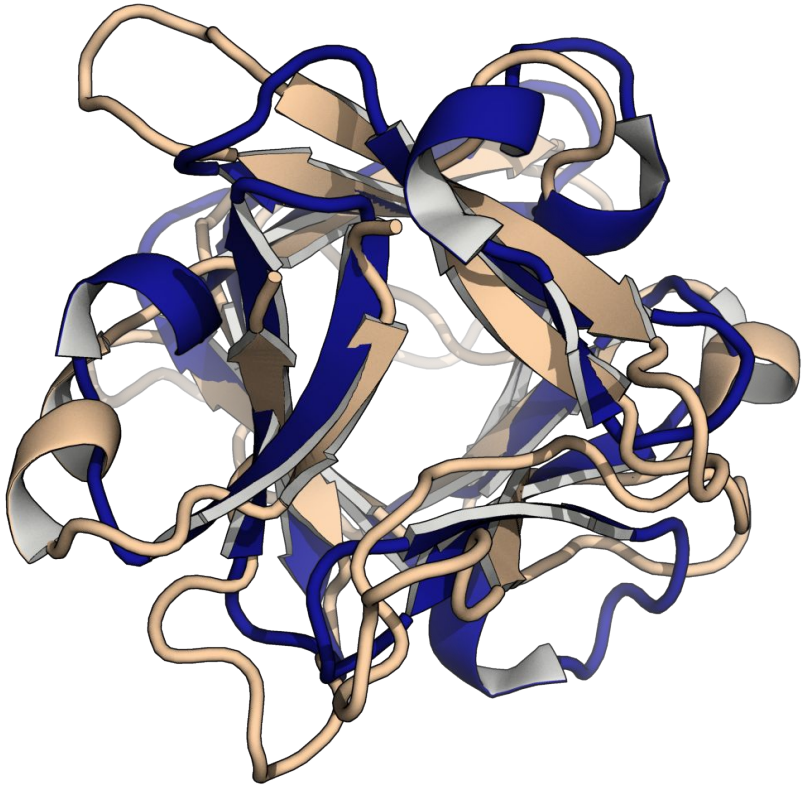


1TIE

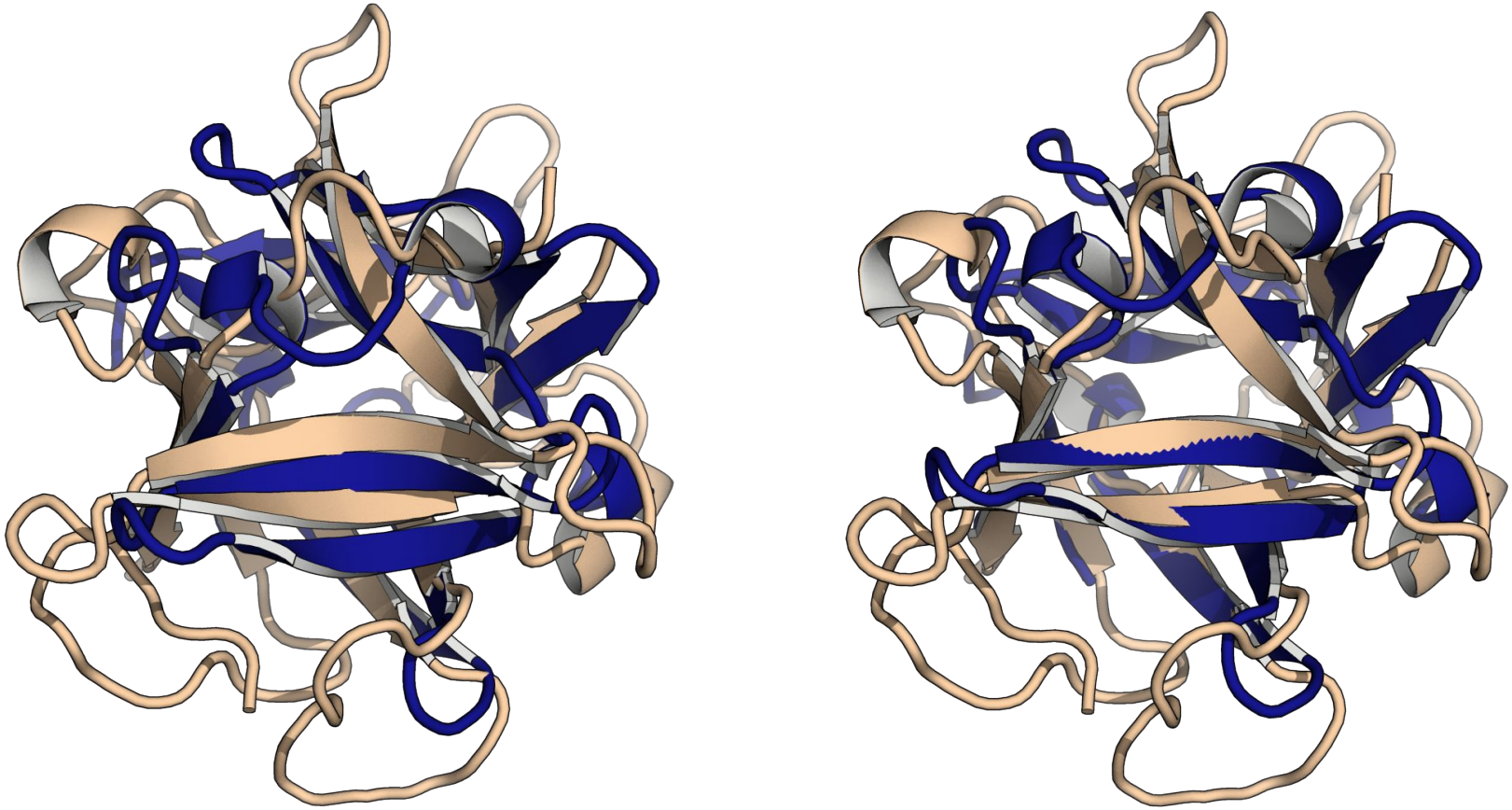
4FGF

Super

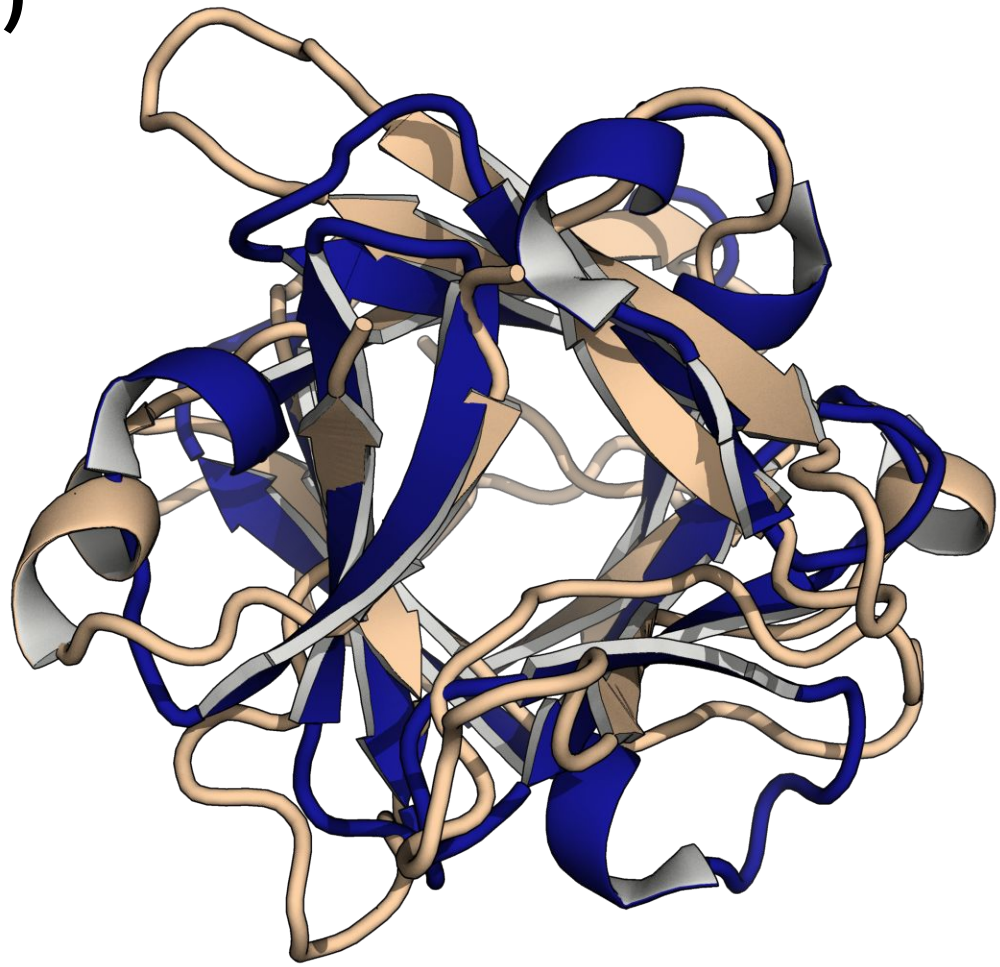




Super



CE (cealign)

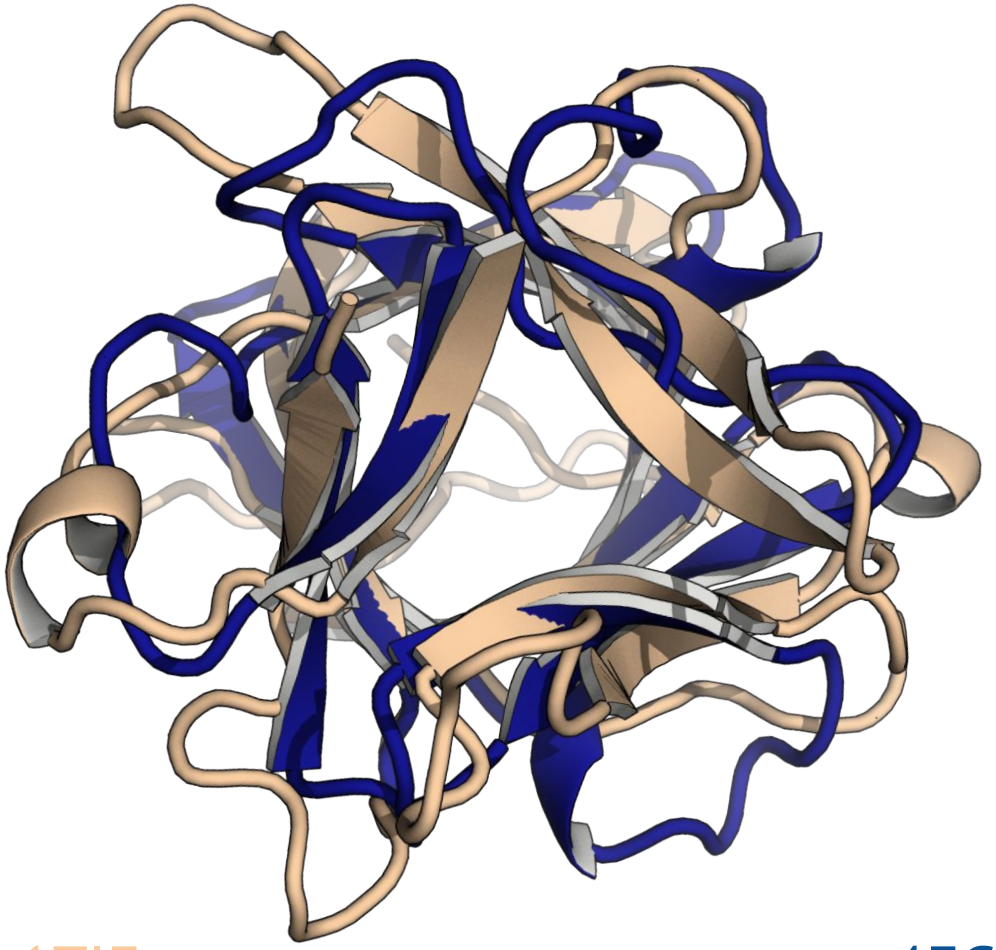


1TIE

4FGF

TM-align

(нет в PyMol)



1TIE

4FGF

Иногда имеет смысл более прицельное выравнивание

Если вы изучаете различия в окружении

Структура и функция

Структура и функция

Взгляд на текущий момент во времени:

- Имеем пространство аннотированных функций
- Имеем пространство аннотированных структур

Как они соотносятся?

Тривиальные отношения:

- Для каждой функции есть своя структура
- Для каждой структуры есть своя функция

Структура и функция

Взгляд на текущий момент во времени:

- Имеем пространство аннотированных функций
- Имеем пространство аннотированных структур

Как они соотносятся?

Тривиальные отношения:

Интуиция подсказывает, что все не может быть настолько просто

- Для каждой функции есть своя структура
- Для каждой структуры есть своя функция

Структура и функция

Взгляд на текущий момент во времени:

- Имеем пространство аннотированных функций
- Имеем пространство аннотированных структур

Как они соотносятся?

Прежде чем отвечать на этот вопрос, надо определиться с тем

- А что такое вообще функция?
- А что такое вообще структура?

Структура и функция

Взгляд на текущий момент во времени:

- Имеем пространство аннотированных функций
- Имеем пространство аннотированных структур

Как они соотносятся?

Прежде чем отвечать на этот вопрос, надо определиться с тем

- А что такое вообще функция?
  - То, что волнует каждого исследователя в отдельности
  - Фенотип
  - EC код
  - GO терм
  - Фитнесс
  - Kcat
  - Аффинность
  - Селективность
  - Любой лейбл или число, имеющие смысл
- А что такое вообще структура?
  - Взаиморасположение всех атомов
  - Взаиморасположение каких-то атомов
  - Взаиморасположение более высокоуровневых паттернов

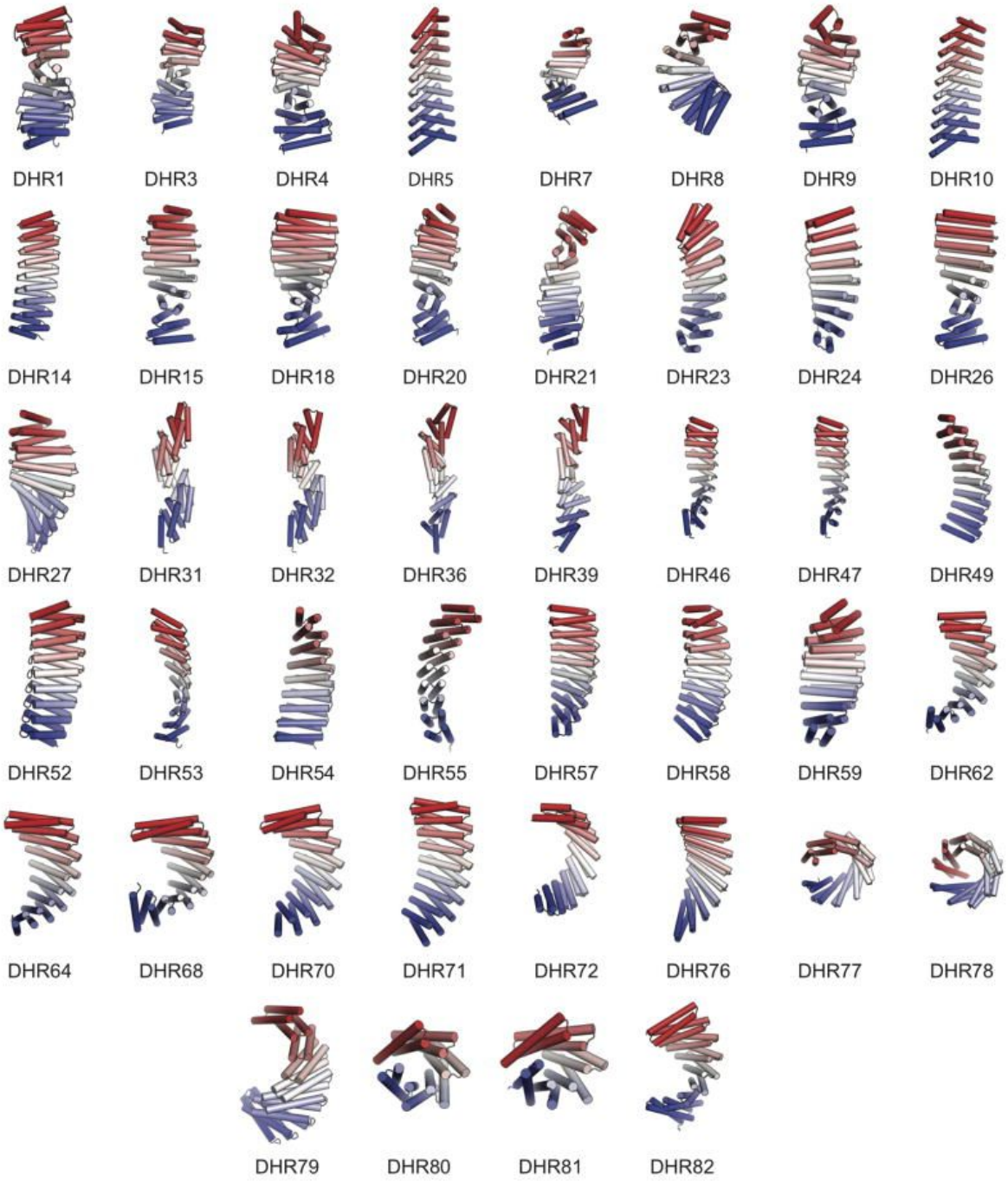
Обзор пространства укладок

Даже просто информация о том, какие укладки из теоретически возможных мы наблюдаем чаще, дает некоторое косвенное понимание структурно-функциональных отношений.

CATH, 2019 год. 1390 топологий

Оценки: всего в природе до 10000 топологий

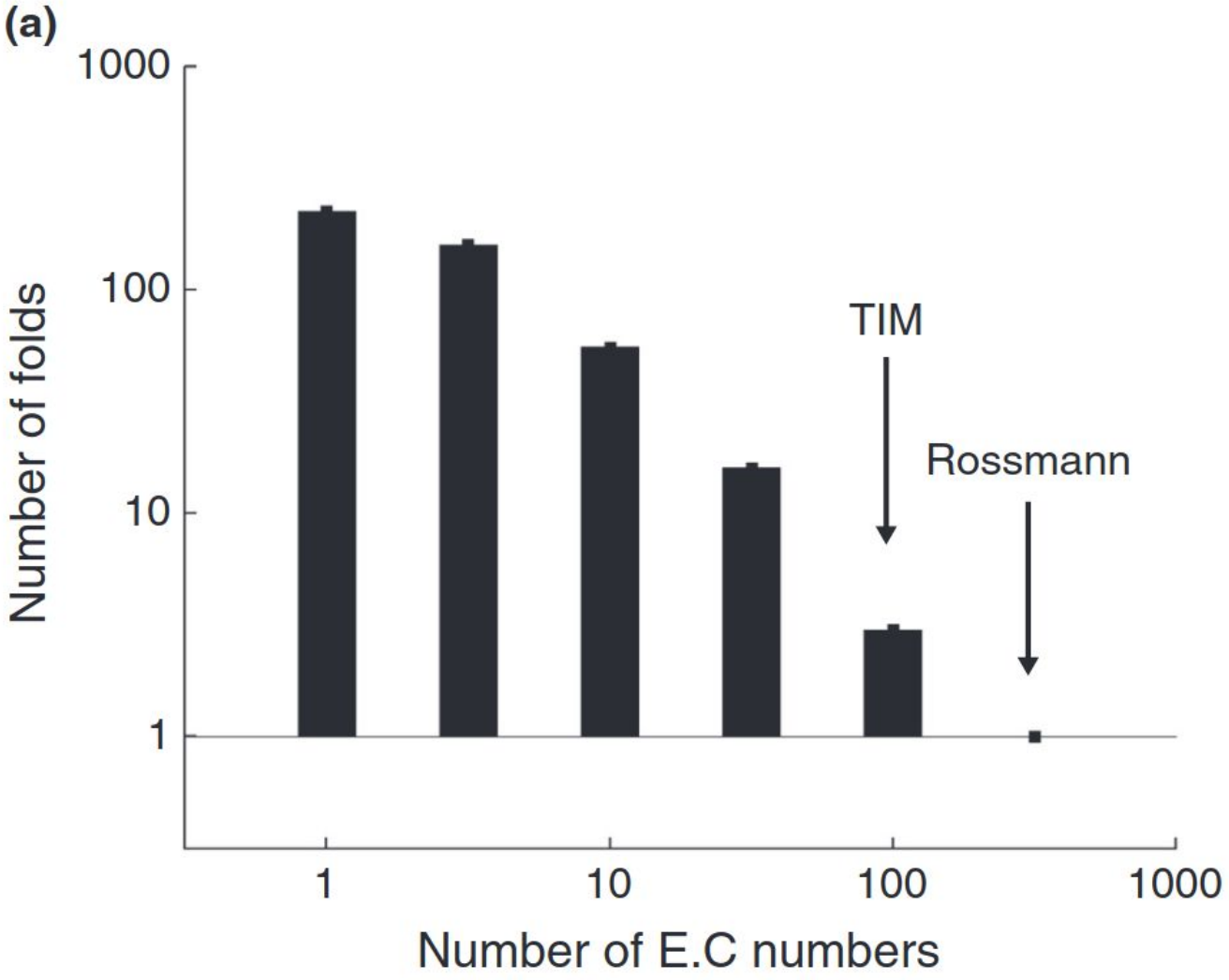
А сколько в теории?



https://www.nature.com/articles/nature16162

Уже не единичны исследования по получению неприродных укладок. Например, тут ученые получили 82 новые альфа-спиральные укладки.

Структура и функция



Функциональные мотивы

- Фолд не задает однозначно функцию (как бы мы ее не определили)
- Функция не задает однозначно фолд
- Внутри фолда не все работает на функцию напрямую, что-то работает на стабильность структуры

Как быть?

Попытаемся сопоставить взаиморасположения типов атомов в пространстве в составе какой-то подструктуры с функцией. Назовем эту подструктуру **функциональным мотивом**.

Мотивы

Паттерны взаиморасположения атомов остатков в пространстве

Структурные	Функциональные
Первоочередна роль остов-остовных взаимодействий	Все взаимодействия могут играть роль
Высокая ёмкость	Низкая ёмкость, зачастую можно сопоставить с паттерном в последовательности
Независимость структуры от окружения (способность ее поддерживать в отсутствие остального белка)	В отсутствие окружения структура изменится

Walker A motif, a.k.a P-loop

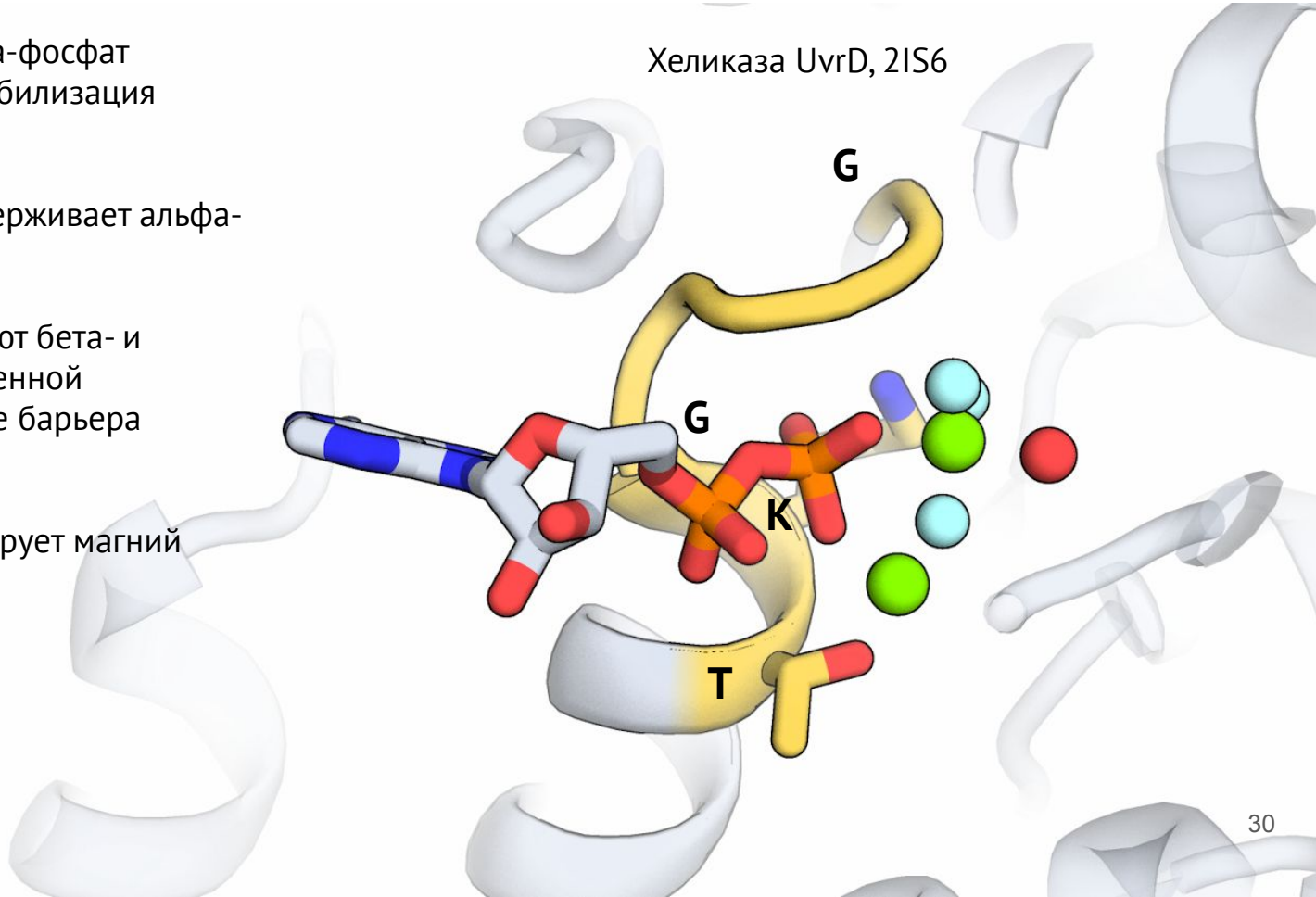
G-x(4)-GK-[TS] – связывание фосфатов

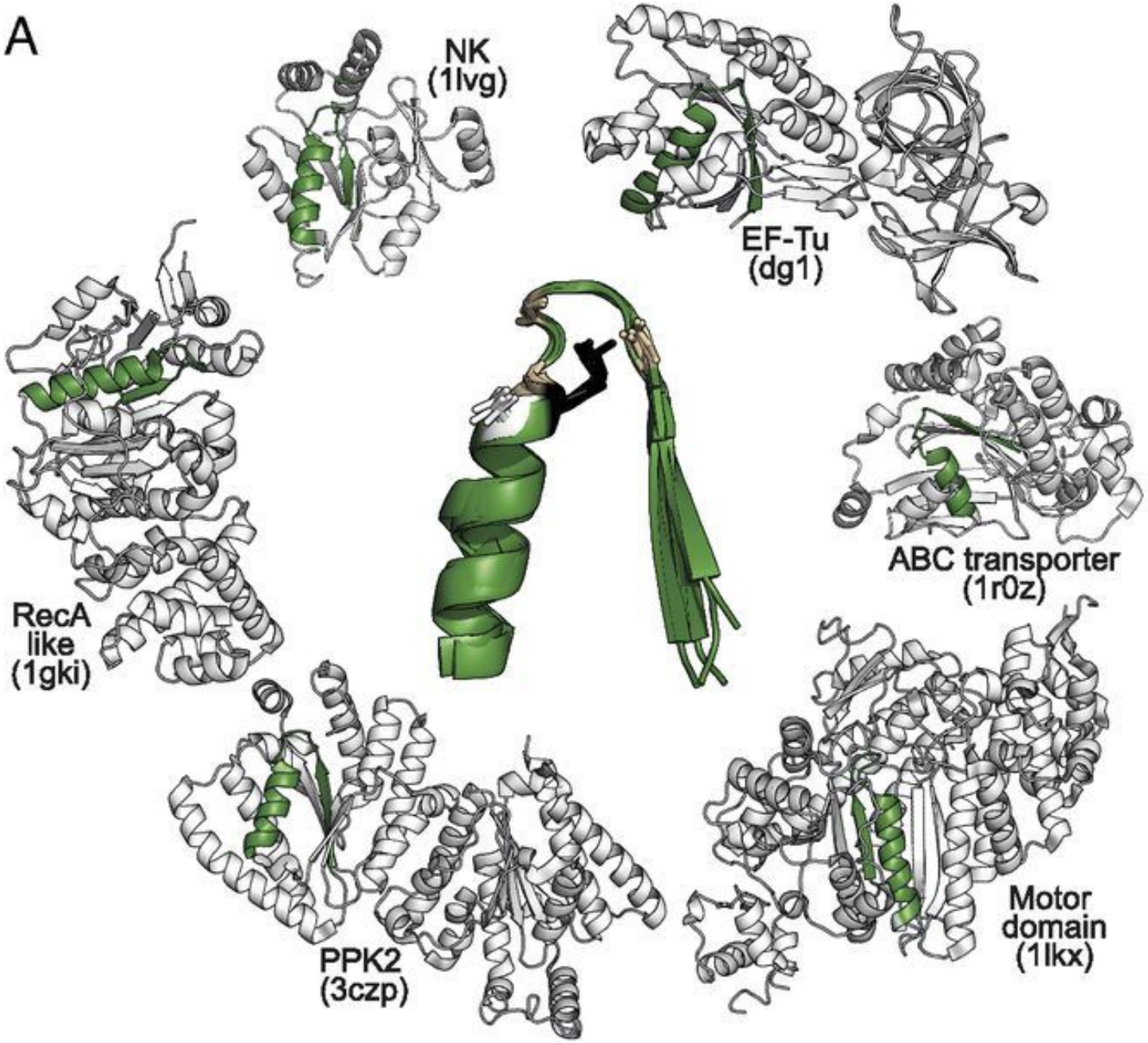
Остов стабилизирует бета-фосфат водородной связью > стабилизация продукта

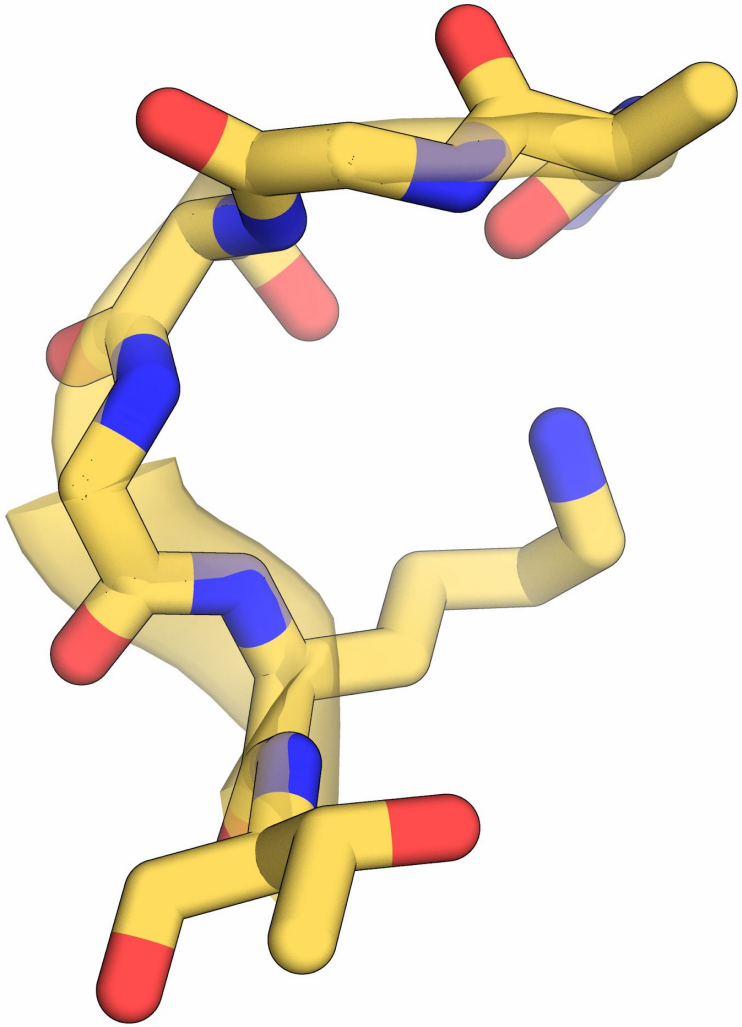
Конец альфа-спирали удерживает альфа- и бета-фосфаты на месте

Лизин и магний фиксируют бета- и гамма-фосфаты в заслоненной конформации > снижение барьера активации

Серин/треонин координирует магний




A
NK
(1lvg)
EF-Tu
(dg1)
ABC transporter
(1r0z)
RecA
like
(1gki)
PPK2
(3czp)
Motor
domain
(1lkx)



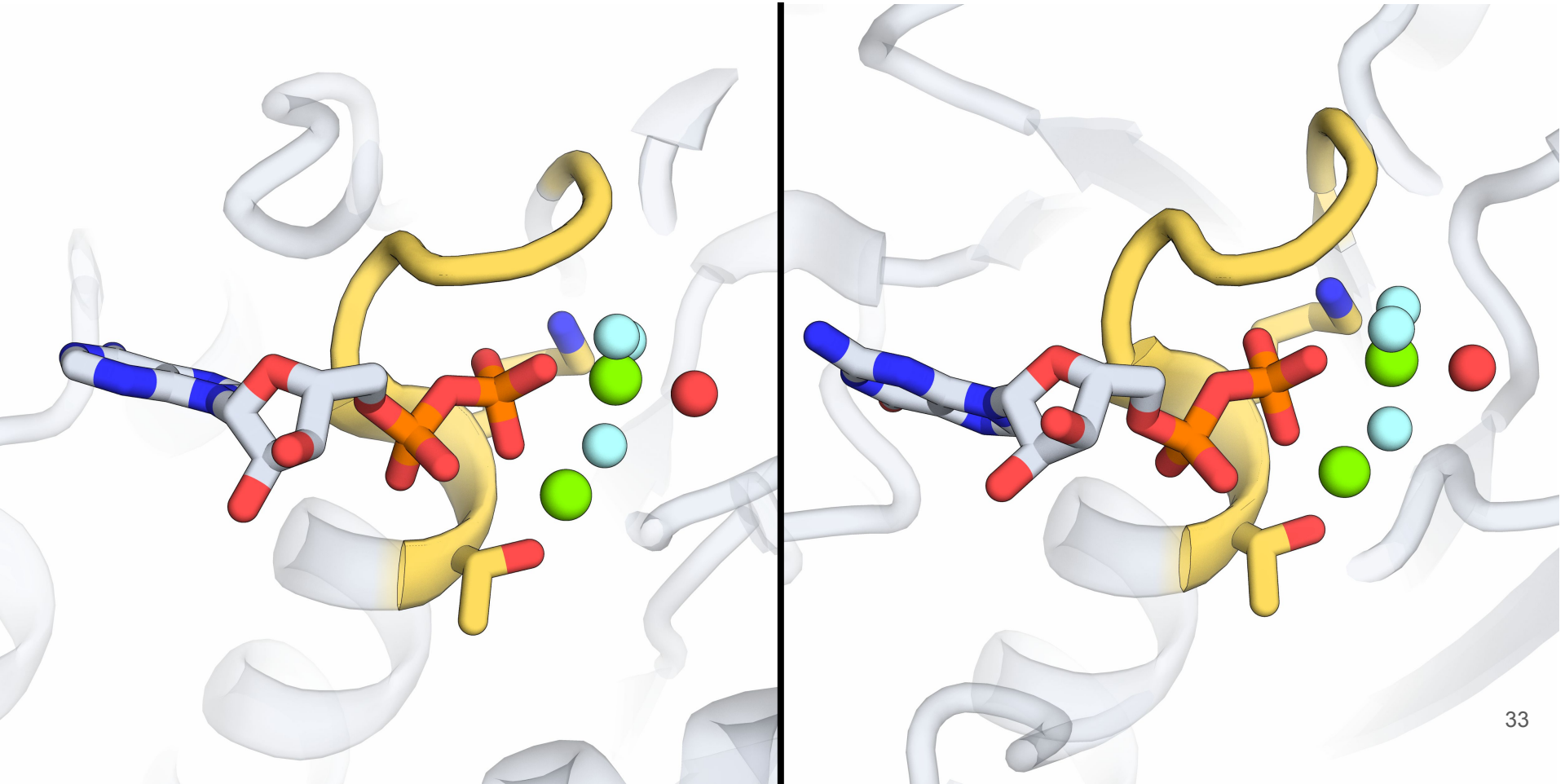
Функциональные мотивы имеют определенную структуру не потому, что это глобальный минимум энергии для этого фрагмента, а потому что такая структура нужна для функции.

Стабилизация структуры достигается за счет окружения, и глобальный минимум, соответствующий этой конформации, существует уже только для всего домена.

Найдите 10 отличий

Хеликаза UvrD, 2IS6

ГТФаза Rho, 1OW3



Структуры точно не назовешь идентичными

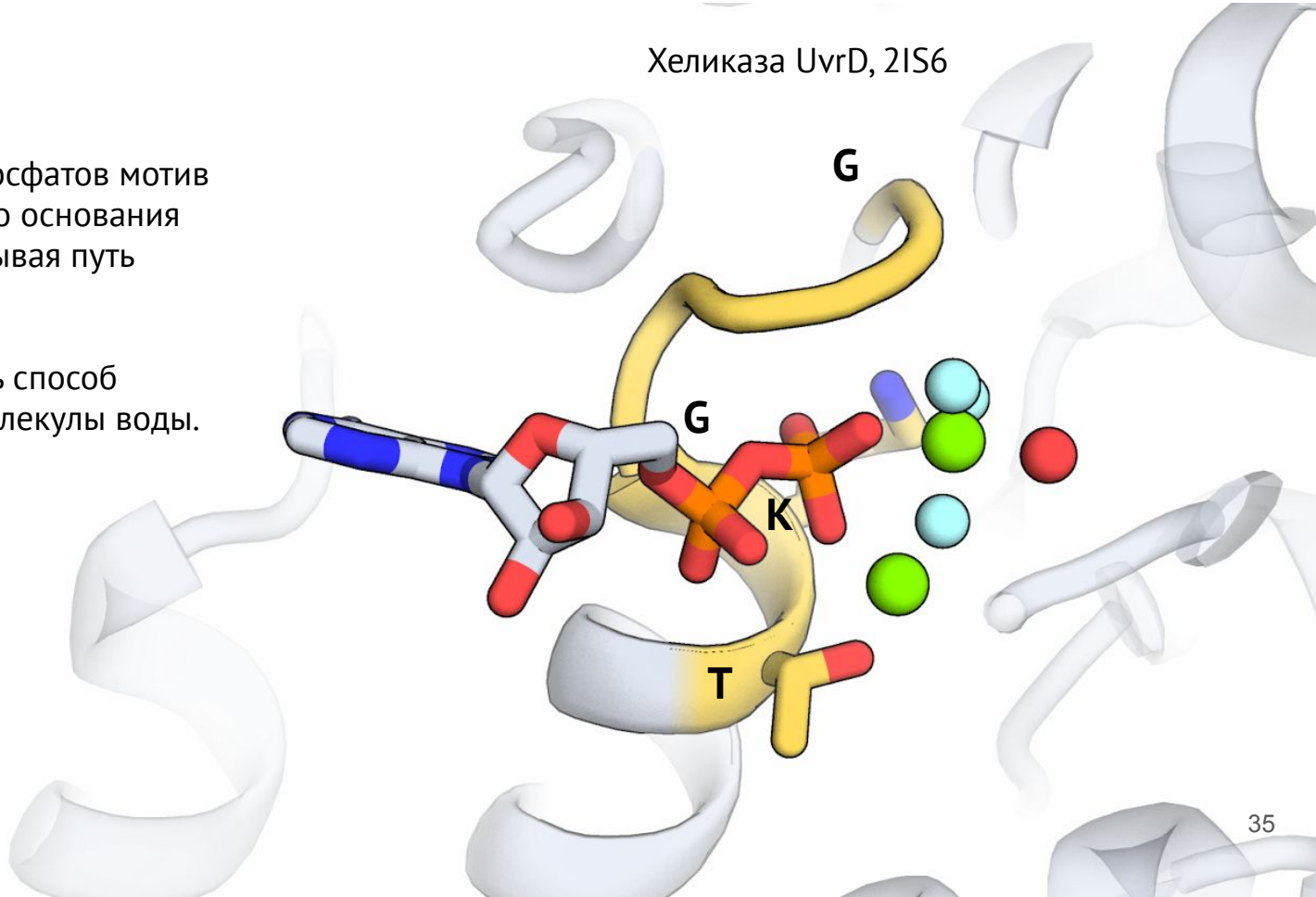
Walker A motif, a.k.a P-loop

G-x(4)-GK-[TS] – связывание фосфатов

Модульность:

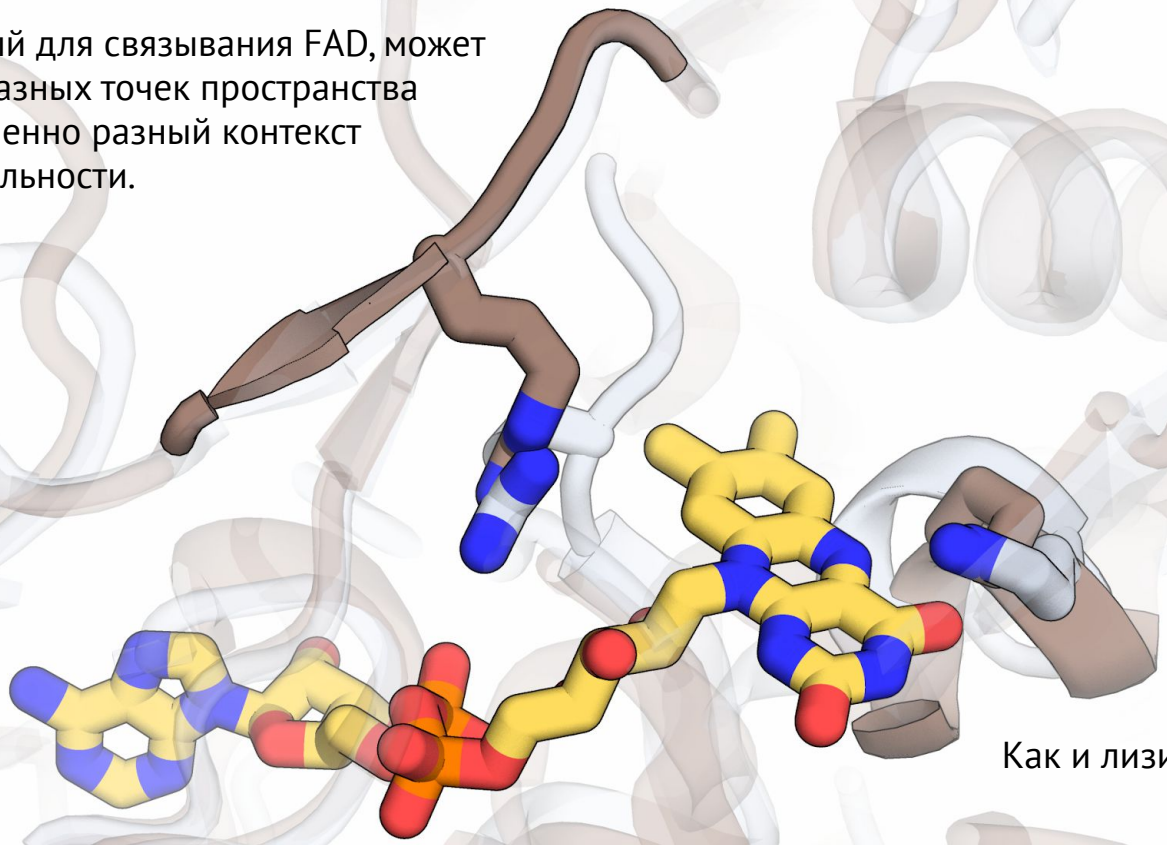
При данном мотиве для процессирования полифосфатов мотив для удержания азотистого основания может варьировать, открывая путь к разной специфичности.

Также может варьировать способ активации атакующей молекулы воды.



Функциональный мотив – не обязательно паттерн в последовательности

Аргинин, нужный для связывания FAD, может приходить из разных точек пространства и иметь совершенно разный контекст по последовательности.



Как и лизин

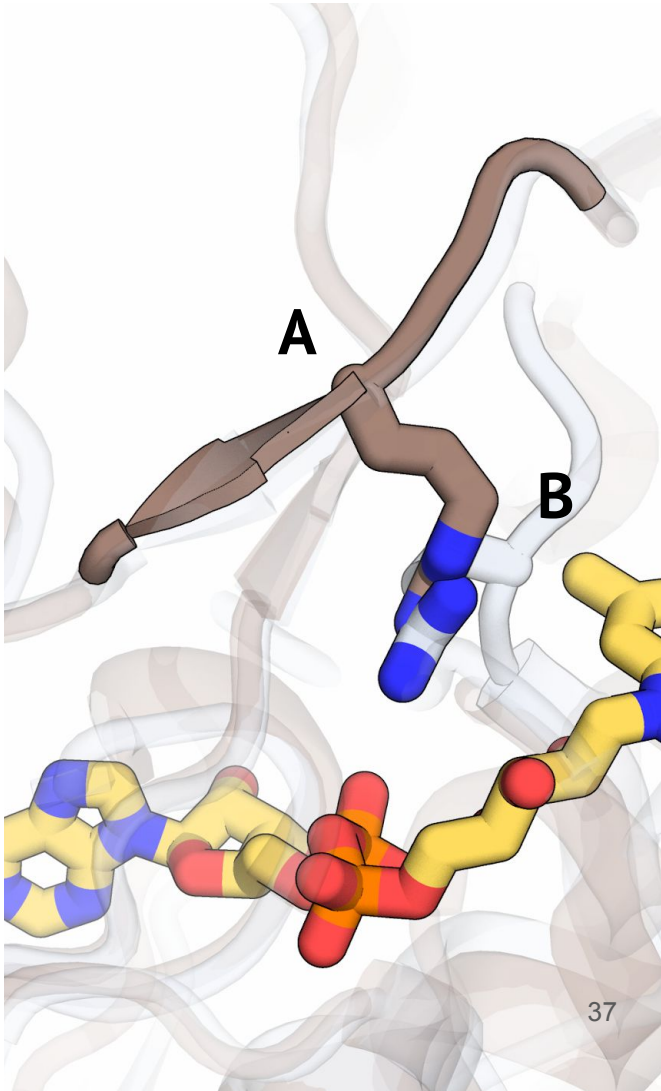
Функциональный мотив – не обязательно паттерн в последовательности

Аргинин, нужный для связывания FAD, может приходить из разных точек пространства и иметь совершенно разный контекст по последовательности.

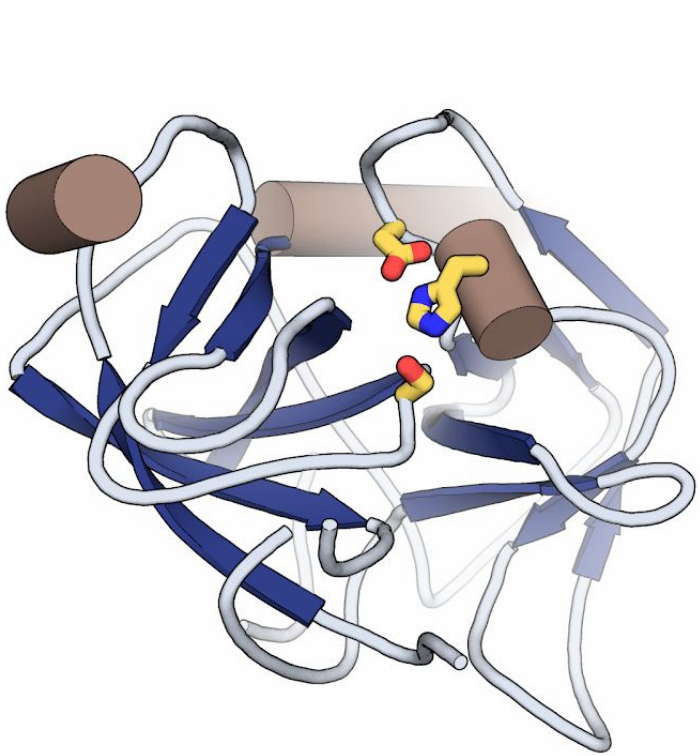
Однако наличие аргинина и на позиции A, и на позиции B недопустимо. Имеем дело с коэволюционирующими позициями!

Степень коэволюции можно вычислять из выравниваний последовательности, например, с помощью Mutual information – как много мы можем сказать о позиции A, зная, что находится на позиции B.

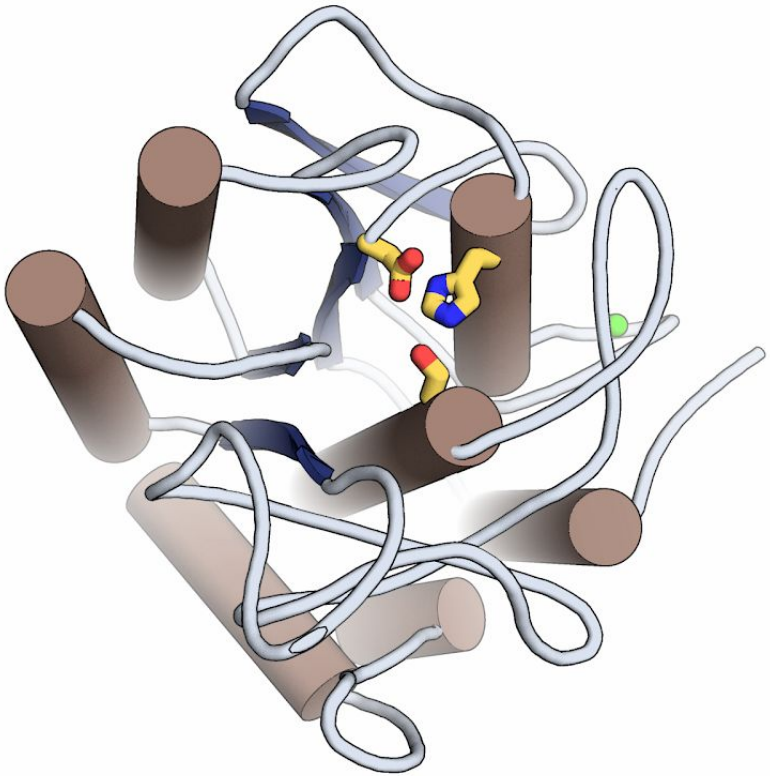
Чаще всего коэволюция указывает на контакт в пространстве, но бывают и такие случаи.



Очевидный пример функциональных мотивов – каталитические сайты

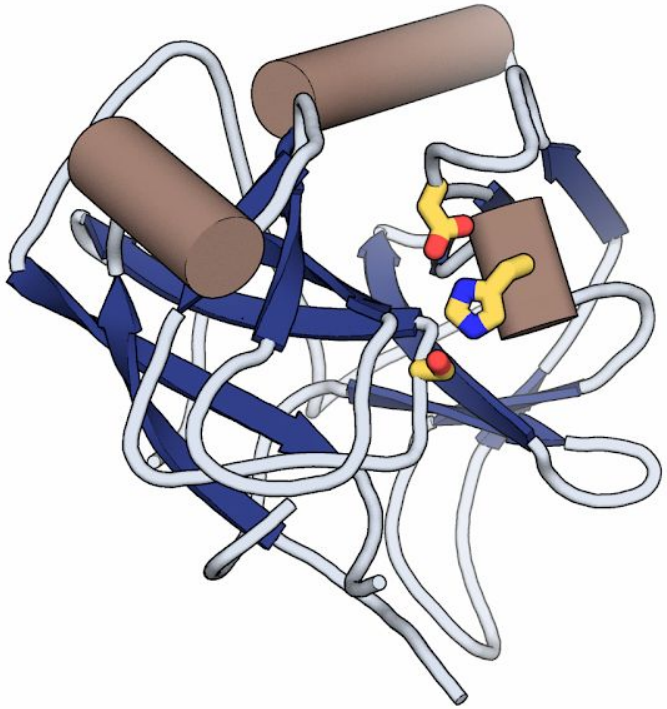


Химотрипсин

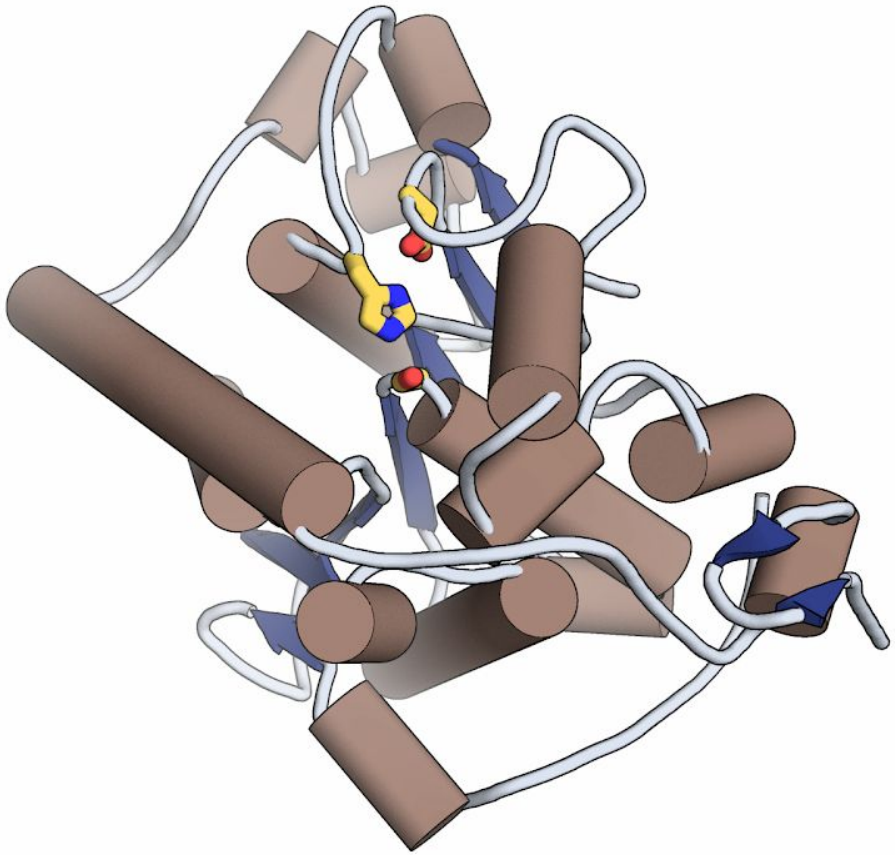


Субтилизин

Очевидный пример функциональных мотивов – каталитические сайты



Химотрипсин



Липаза B

Поиск по мотивам

Суперфолды: взгляд из “сейчас”

TIM barrel и Rossmann фолд являются примерами **суперфолдов**: укладок, допускающих большое количество реализаций в виде последовательности.

Ёмкость (**Sequence capacity**): число последовательностей, способных стабильно сворачиваться в заданный фолд. Также иногда называется **designability**.

Чем больше ёмкость фолда, тем он быстрее способен эволюционировать, т.е. обладает высокой **эволюционируемостью.**

Суперфолды: взгляд во времени

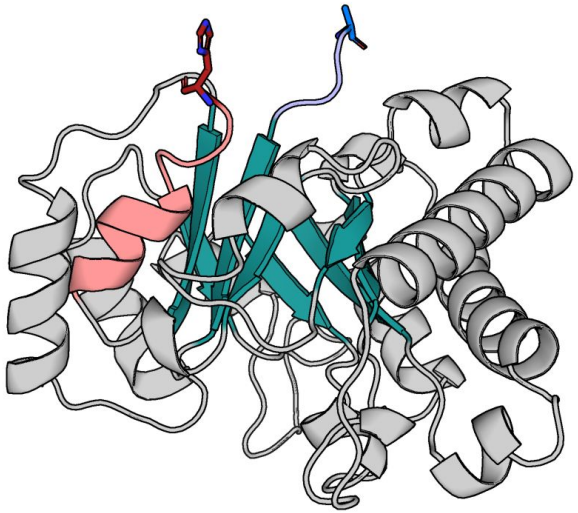
Какие-то структурные особенности суперфолдов позволили им так широко эволюционировать к текущему моменту, когда мы смогли их собрать и изучить. Какие? Что коррелирует с эволюционируемостью?

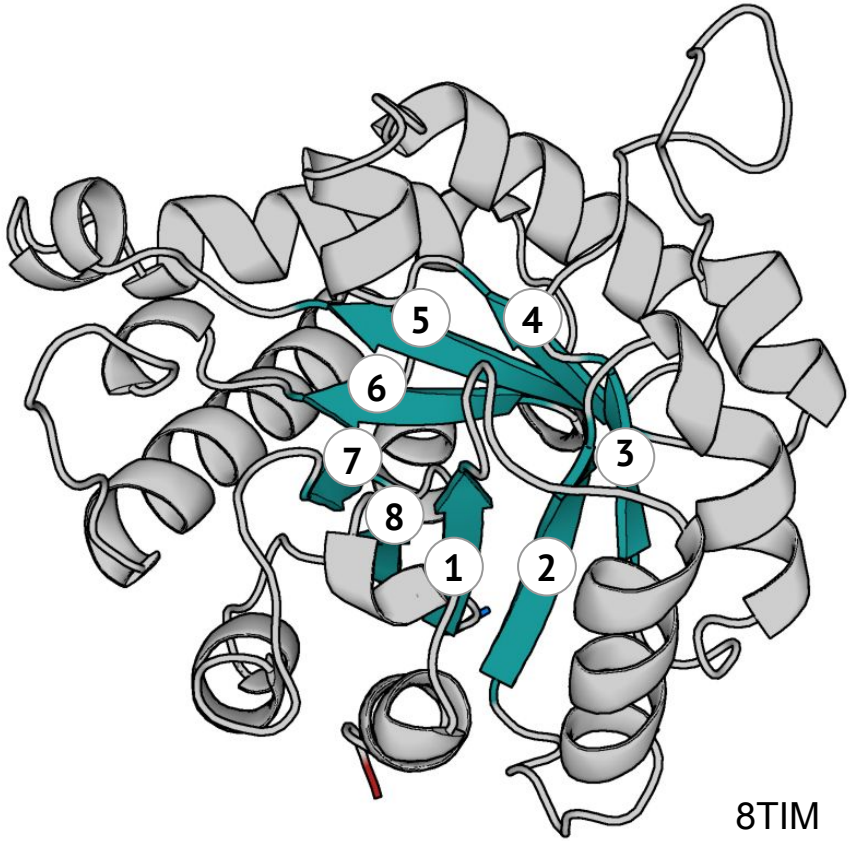
Эволюционируемость (**evolvability**): способность со временем принимать изменения в последовательности и функции.

Робастность (**robustness**): способность сохранять фенотип при изменении генотипа.

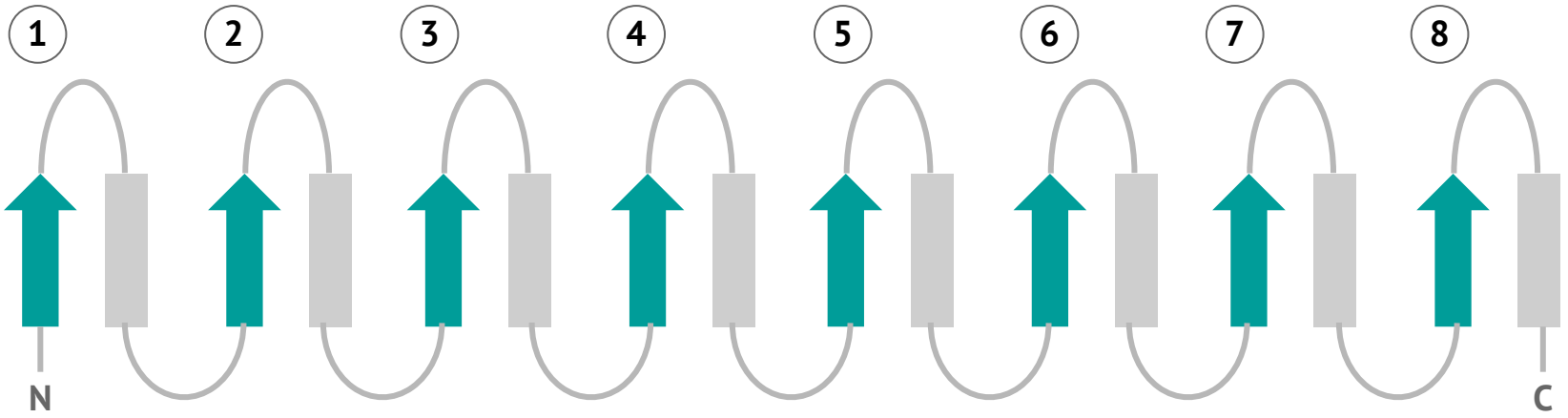
Инновабельность (**innovability**): способность приобретать новые функции путем сравнительно небольших изменений в последовательности.

TIM barrel

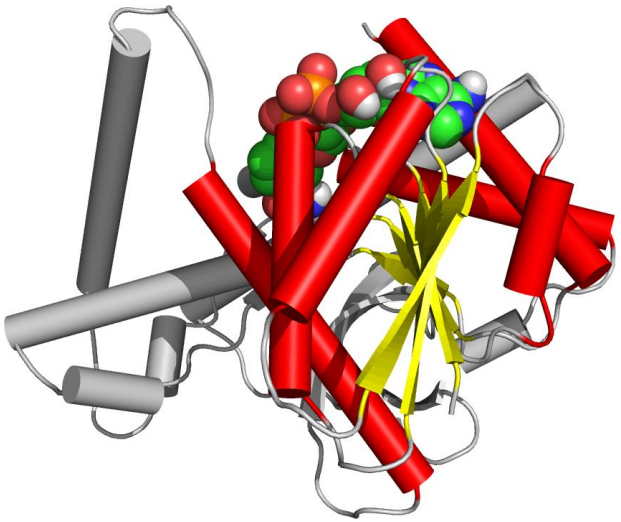


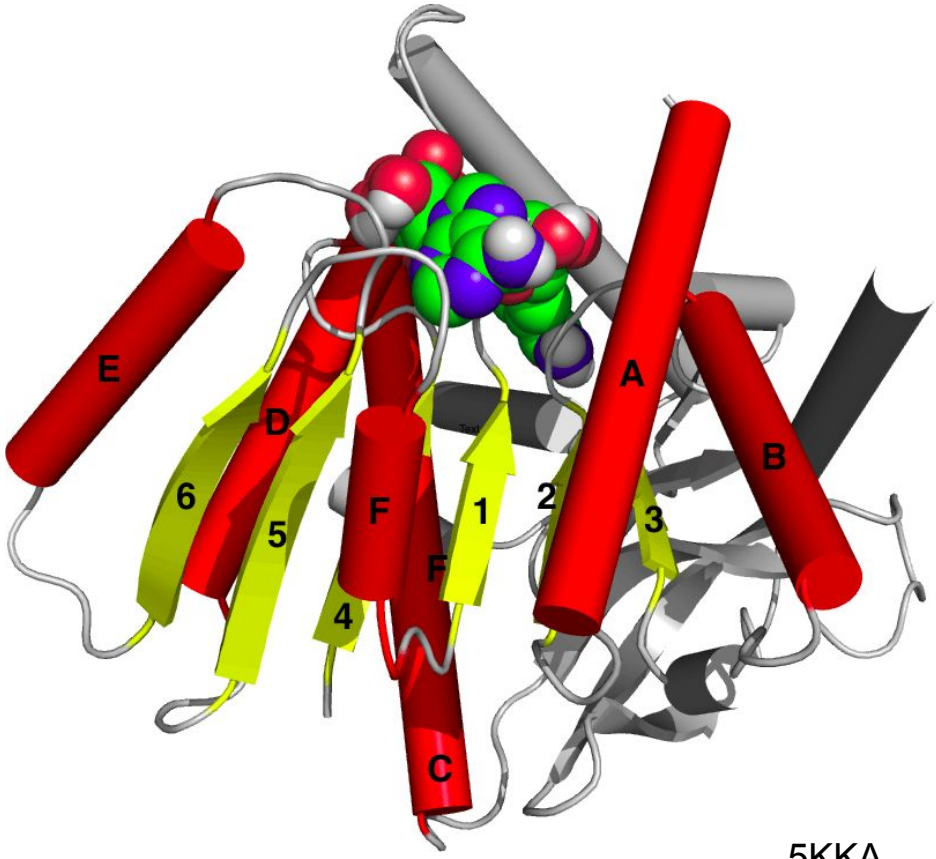


8TIM

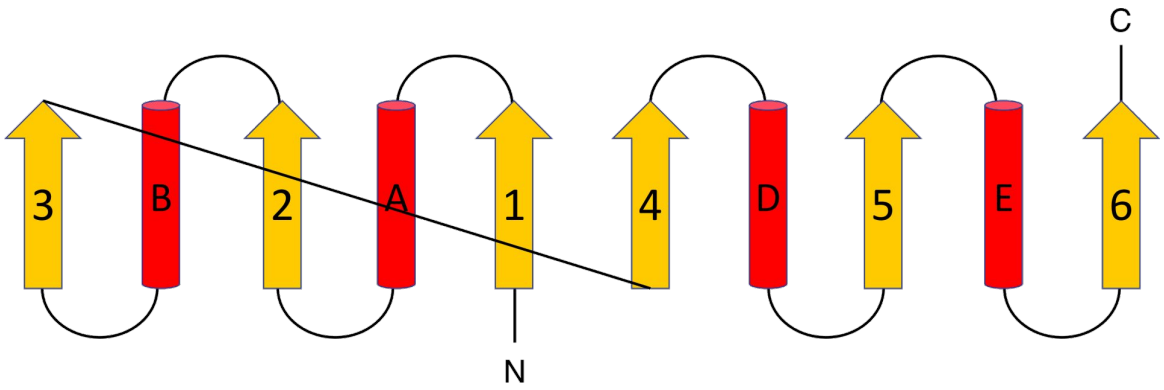


Rossman фолд





5KKA



TIM barrel и Rossmann fold

TIM barrel и Rossmann fold составлены из элементов надвторичной структуры, которые обладают

- высокой склонностью к самостоятельному фолдингу
- высокой ёмкостью (так как в основном укладка определяется остовом, задача сайдчейнов – не навредить)
- низкой гибкостью (так как это вторичная структура) – меньше вероятность изменить ход остова при замене, меньше вероятность нарушить работоспособную геометрию активного центра

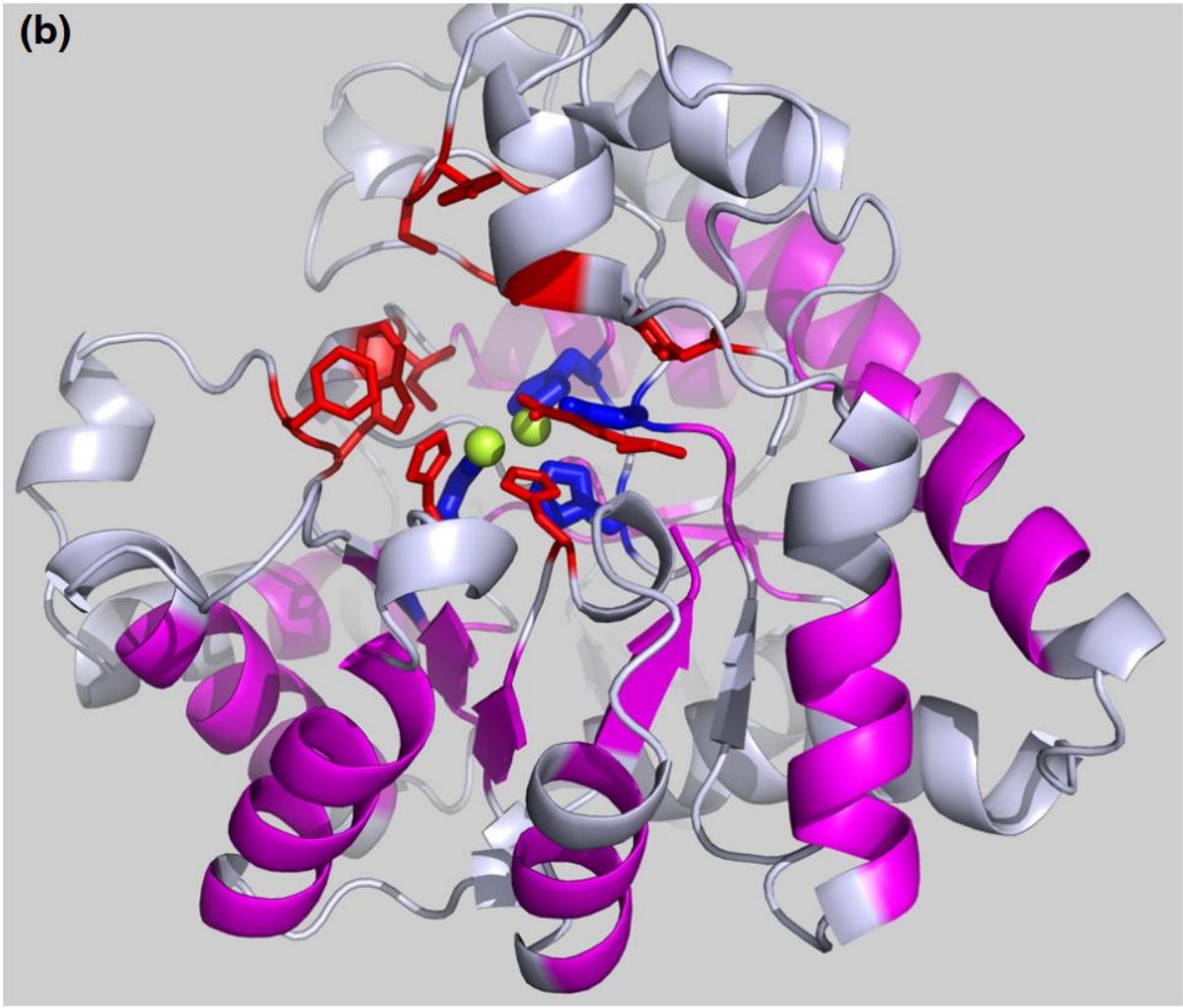
Отсюда высокая **робастность**

Откуда можем взять **инновабельность?**

Активный центр vs каркас

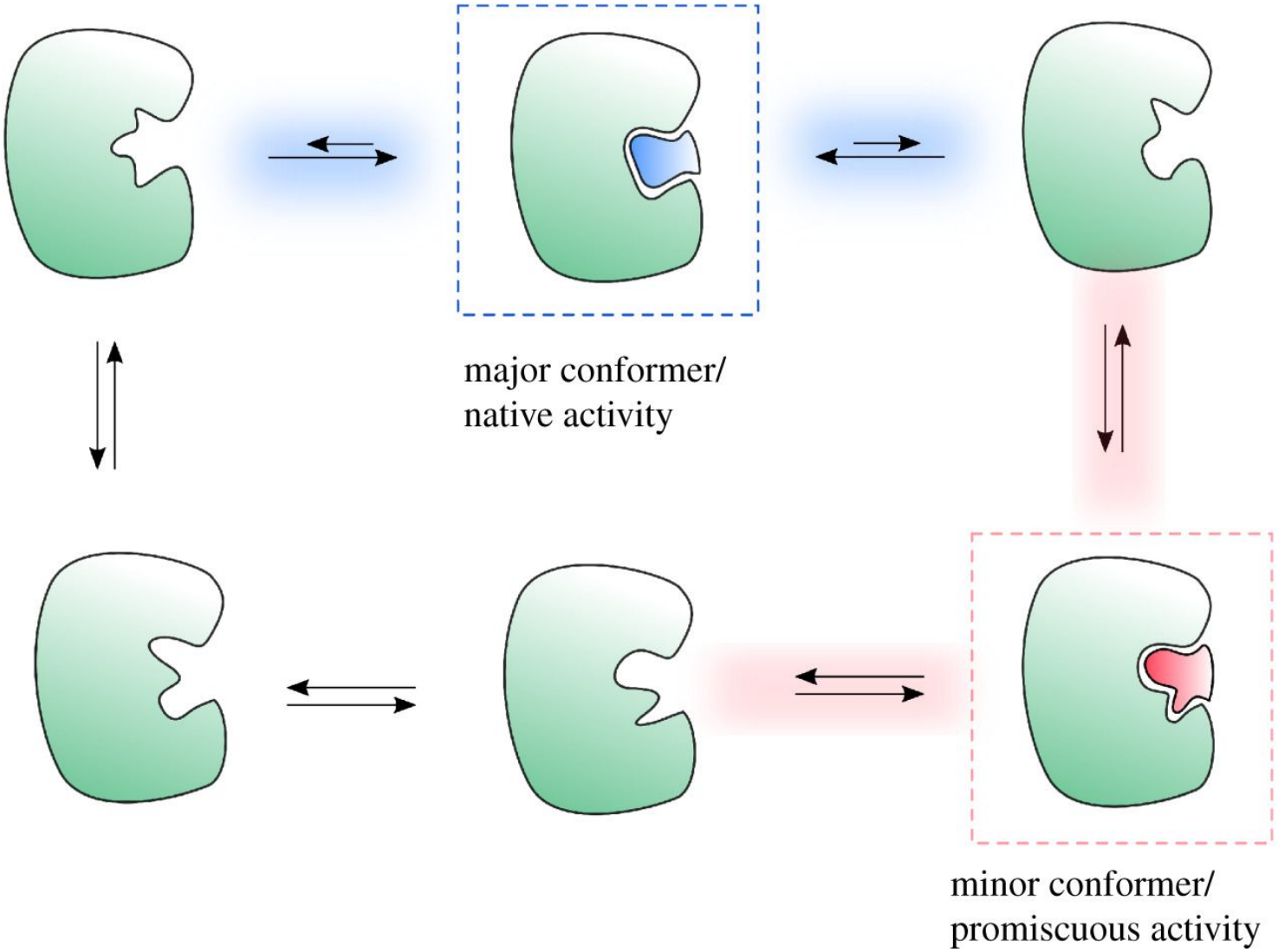
Подвижные участки способствуют **инновабельности**, например в типе процессируемого субстрата (показаны красным).

Почему?



Current Opinion in Structural Biology

Подвижность и инновабельность



Подвижность и робастность

Подвижные участки вблизи каталитической машинерии снижают робастность.

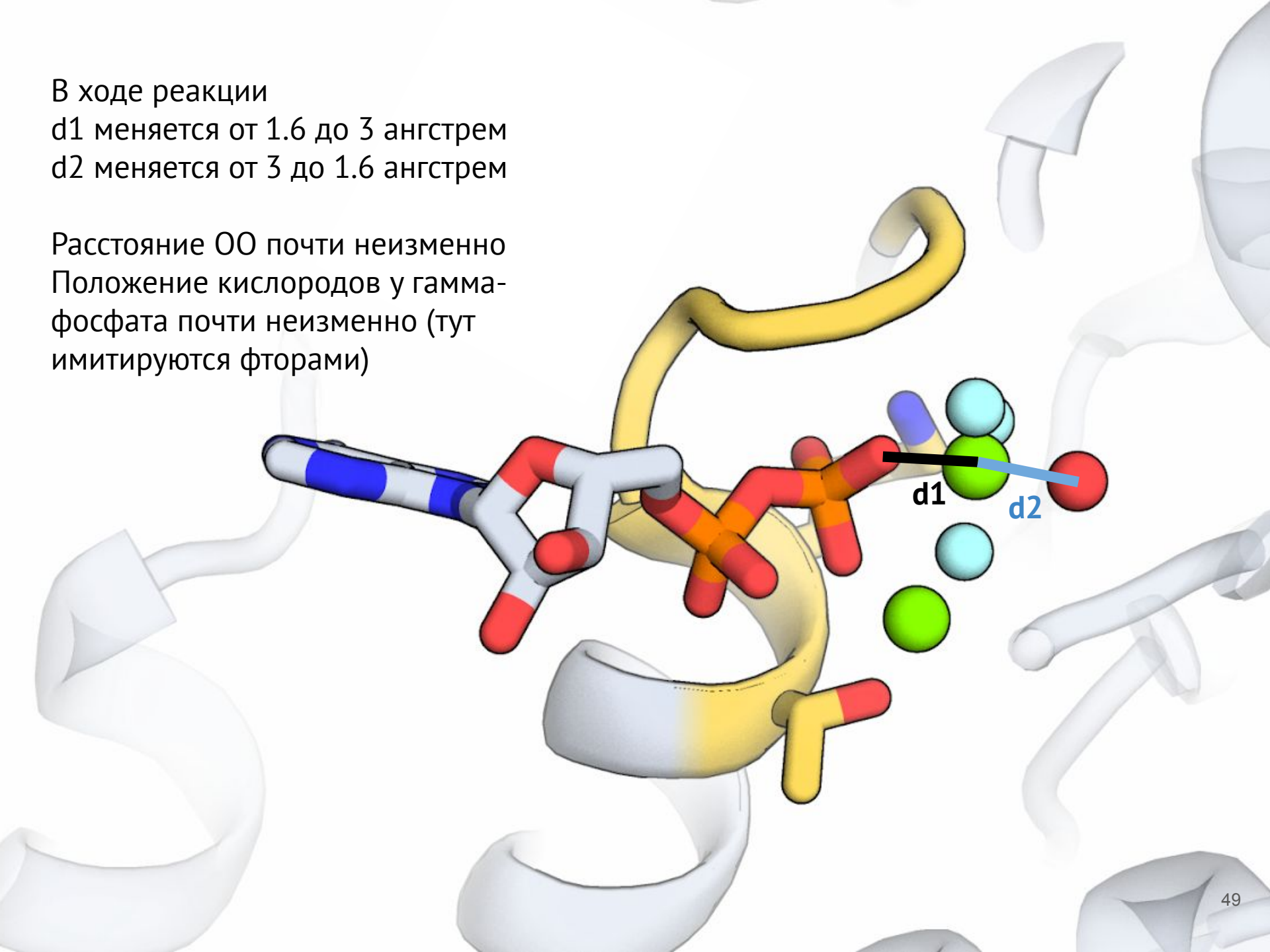
В идеальном ферменте:

- Структурные изменения активного центра вдоль направления координаты реакции минимальны (чем они больше, тем менее вероятна реакция)
- Масштабы перестроек в ходе реакции минимальны (изменение положений отдельных атомов субстрата в пределах 2-3 ангстрем)
- Активный центр фиксирован в конформации, максимизирующей вероятность такой перестройки (иными словами снижающей энергию переходного состояния)

Подвижность разбалтывает активный центр и снижает его вероятность принять эффективную конформацию

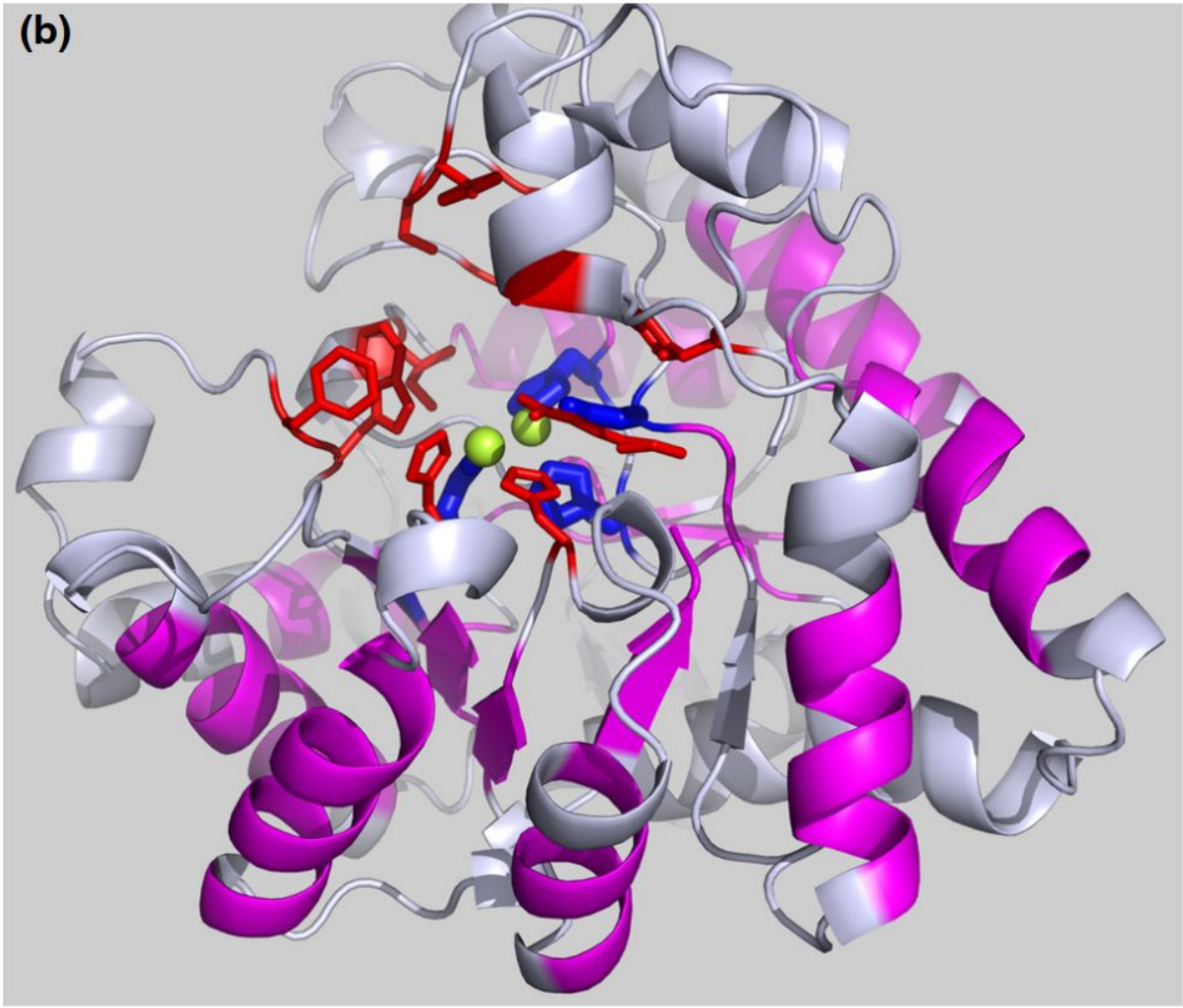
В ходе реакции
d1 меняется от 1.6 до 3 ангстрем
d2 меняется от 3 до 1.6 ангстрем

Расстояние ОО почти неизменно
Положение кислородов у гамма-фосфата почти неизменно (тут имитируются фторами)



Активный центр vs каркас

Большая степень отделенности остатков активного центра от остатков структурного каркаса и от подвижного модуля способствует и **инновабельности**, и **робастности**.



Current Opinion in Structural Biology

Эволюция структур – движущие силы

1. Более робастные фолды эволюционируют быстрее (больше мутаций могут быть нейтральными, больше вероятность закрепиться для нейтральной мутации)
2. Белки эволюционируют против вероятности мисфолдинга. Чем более высоко экспрессируется белок, тем медленнее он эволюционирует (так как выше абсолютное число неверно уложенных цепочек)
3. Белки эволюционируют против вероятности неспецифических взаимодействий
4. Белки эволюционируют в сторону оптимизации функции (metabolic flux). Если это фермент, то играет роль не только эффективность процессирования своего субстрата, но и неэффективность процессирования других субстратов.

Как изучать эволюцию

Горизонтальный подход

Изучаем следствия эволюции, данные нам на текущий момент: наборы структур гомологичных белков с различной функцией, например, специфичностью. Можем выделить консервативные, специфические, вариабельные остатки, коэволюционирующие остатки. Не можем ответить на вопрос, какая функция была раньше, как именно шла эволюция, как выглядели древние белки, как нам сымитировать эволюцию, чтобы изменить функцию дальше.

Вертикальный подход

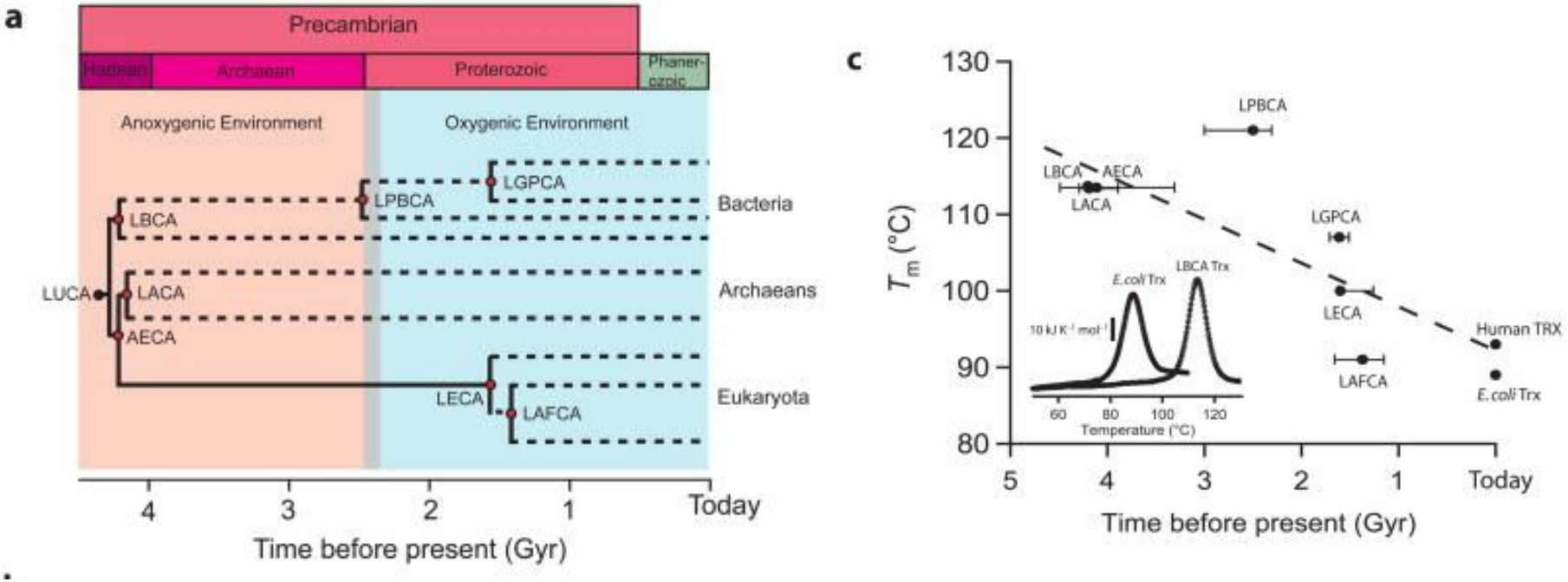
Пытаемся восстановить ход эволюции и предсказать последовательности предковых белков.

Вертикальный подход: ancestral reconstruction

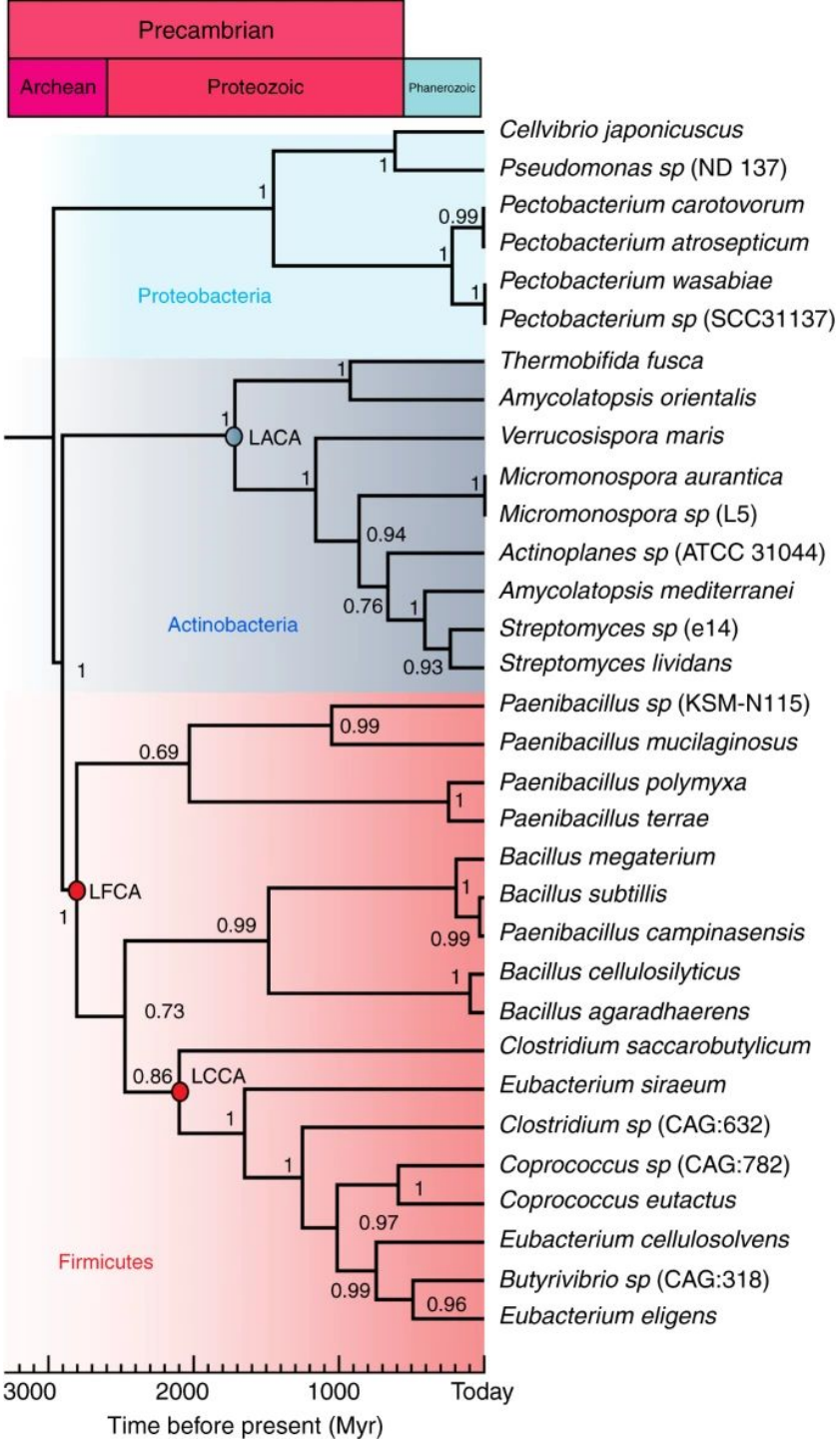
Что нам нужно:

- Способ построения филогенетического дерева
  Топология является ключевой для метода – в каких местах находятся узлы?
- Статистическая модель возникновения замен
  Каждому узлу нужно сопоставить синтезированную последовательность белка
- Способ получения информации о структуре предковых белков
  Эксперимент или моделирование

Древние ферменты: палеоэнзимология



“Воскрешенные” тиоредоксины возрастом 4 миллиарда лет заметно стабильнее в кислой среде и при высокой температуре, при этом сохраняя сопоставимую активность



https://www.nature.com/articles/s42004-019-0176-6

“Воскрешенные” целлюлазы возрастом 2 миллиарда лет также заметно стабильнее при высокой температуре, имеют широкий диапазон pH, при которых сохраняют активность.

Древние ферменты и эволюция

Применение методов предковой реконструкции на текущий момент дало нам следующие знания:

- Специализированные родственные ферменты происходят от предков-генералистов, демонстрирующих функциональную промискуитетность
- Они достигали этого за счет большей гибкости, подвижности структуры. Как следствие предковые ферменты обладали большей термостабильностью.
- На пути к специализации замены стабилизировали подвижные участки
- Древнейшие ферменты вероятно были плохими байндерами, но обладали хоть какой-то каталитической функцией, поставившей их под влияние отбора
- Древние фолды – привлекательная стартовая точка для дизайна ферментов