



Компьютерная биология в изучении структуры, функции и регуляции белков: методы, задачи и решения

**в рамках курса «Суперкомпьютерное моделирование»
для магистров 2 года обучения ВМК МГУ**

Дмитрий Суплатов

К.Х.Н., СТ.Н.С.

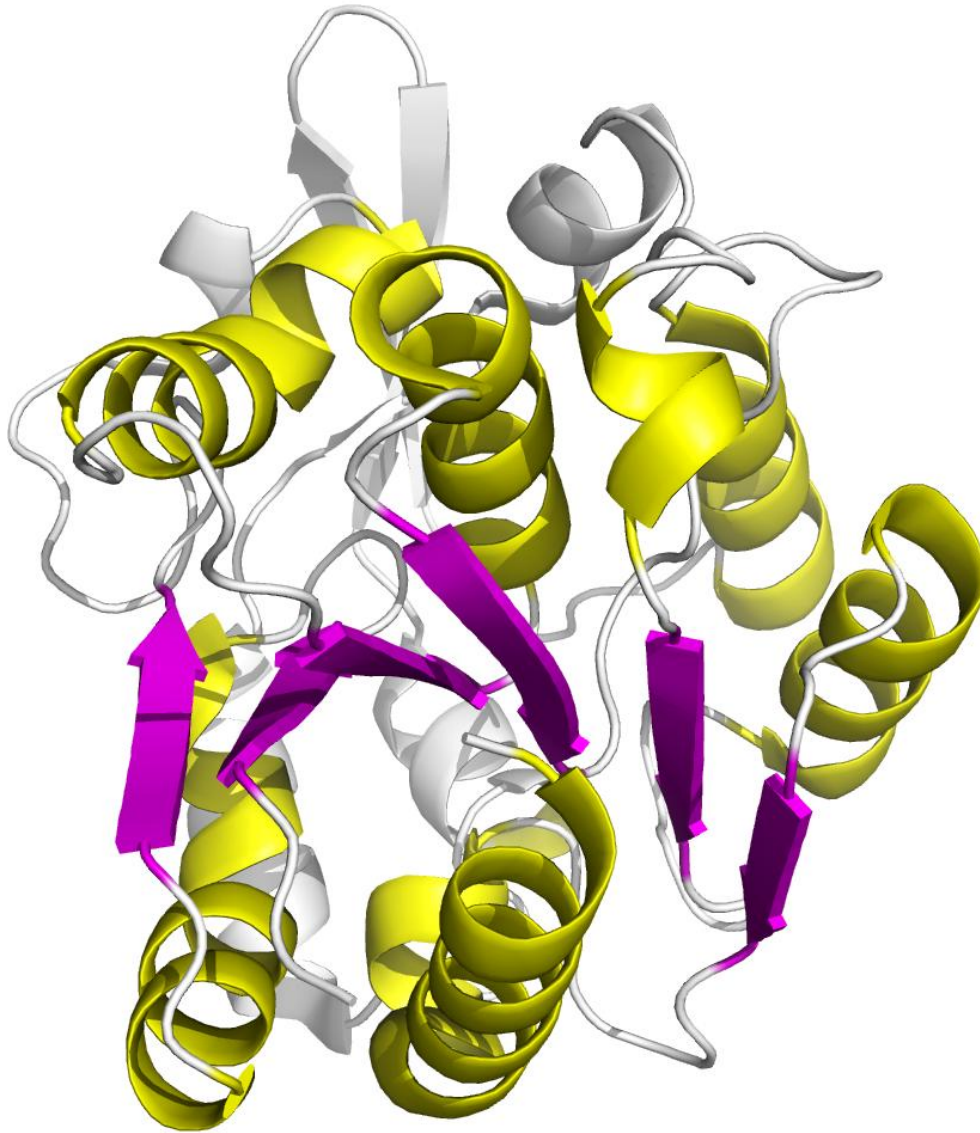
НИИ ФХБ имени А.Н.Белозерского

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

логин: FBB2017
пароль: fbb2017

Введение

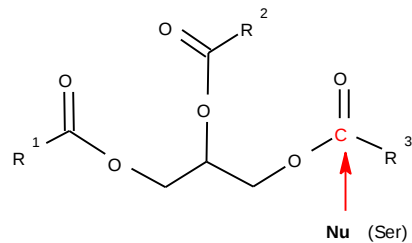
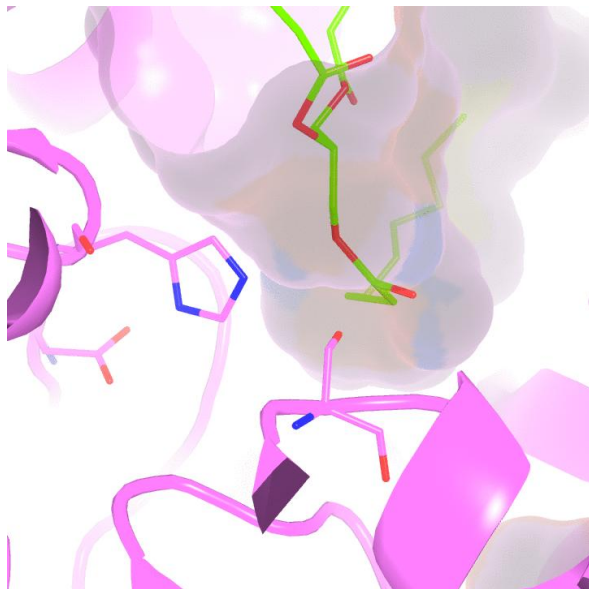
Суперсемейство α/β Гидролаз



- Более 20000 ферментов, разделенных на семейства с различными каталитическими свойствами: **Липазы и эстеразы, Оксинитрилазы, Карбоксипептидазы** и др.
- α/β слои (восемь β -тяжей соединенных α -спиралями)
- В каталитической триаде **гистидин** полностью консервативен, в то время как два других остатка варьируются.

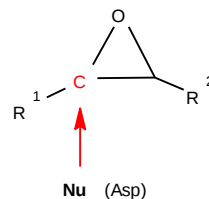
Изучение структурно-функциональных взаимосвязей в ферментах

**Липазы
(ЕС 3.1.1.3)**



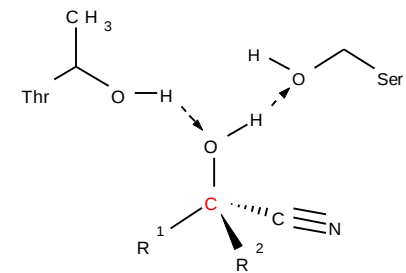
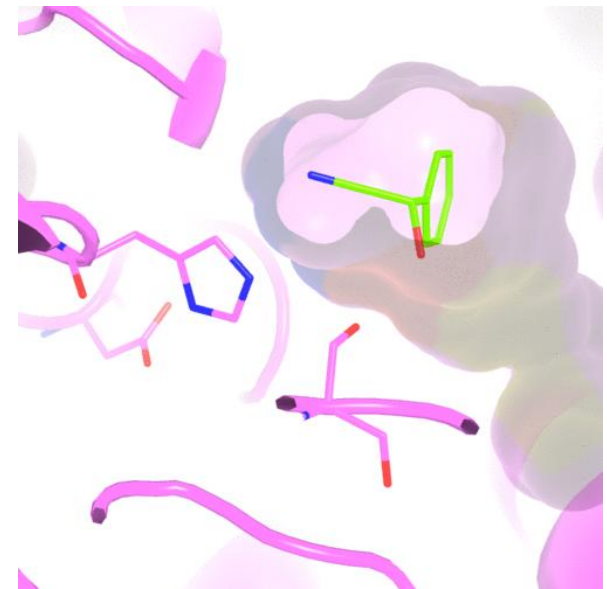
Триацилглицерол

**Эпоксид-гидролазы
(ЕС 3.3.2.-)**



Эпоксид

**Оксинитрилазы
(ЕС 4.1.2.47)**



Циангидрин

Какие структурные факторы
определяют каталитические
свойства ферментов?

Почему ферменты катализируют
различные химические
превращения в рамках **общей**
структурной организации?

Постановка задачи

MKNRNRMIVN CVTASLMYYW SLPALAEQSS SEIKIVRDEY GMPHIYANDT WHLFYGYGYV
VAQDRLFQME MARRSTQGTV AEVLGKDFVK FDKDIRRNYW PDAIRAQIAA LSPEDMSILQ
GYADGMNAWI DKVNTNPETL LPKQFNTFGF TPKRWEPFDV AMIFVGTMAN RFSDSTSEID
NLALLTALKD KYGVSQGMAY FNQLKWLVPN SAPTTIAVQE SNYPLKFNQQ NSQTAALLPR
YDLPAPMLDR PAKGADGALL ALTAGKNRET IAAQFAQGGA NGLAGYPTTS NMWVIGKSKA
QDAKAIMVNG PQFGWYAPAY TYGIGLHGAG YDVTGNTPFA YPGLVFGHNG VISWGSTAGF
GDDVDIFAER LSAEKPGYYL HNGKWKMLS REETITVKNG QAETFTVWRT VHGNILQTDQ
TTQTAYAKSR AWDGKEVASL LAWTHQMKAK NWQEWTOQAA KQALTINWYY ADVNGNIGYV
HTGAYPDRQS GHDPRLPVPV TGKWDWKGLL PFEMNPKVYN PQSGYIANWN NSPQKDYPAS
DLFAFLWGGA DRVTEIDRLL EQKPRLTADQ AWDVIRQTSR QDLNLRLFLP TLQAATSGLT
QSDPRRQLVE TLTRWDGINL LNDDGKTWQQ PGSAILNVWL TSMLKRTVVA AVPMPFDKWY
SASGYETTQD GPTGSLNISV GAKILYEAVQ GDKSPIPQAV DLFAGKPQQE VVLAALEDTW
ETLSKRYGNN VSNWKTPAMA LTFRANNFFG VPQAAAEETR HQAEYQNRGT ENDMIVFSPT
TSDRPVLAWD VVAPGQSGFI APDGTVDKHY EDQLKMYENF GRKSLWLTQ DVEAHKESQE
VLHVQR

1. Какие функции и свойства характерны для этого фермента?
2. Чем обусловлены эти свойства на структурном уровне?
3. Как мы можем их изменить?

Что такое «Компьютерная биология»?

- Биоинформатика
- Молекулярное моделирование

Что такое биоинформатика?

- Может быть определена как «приложение компьютерных методов анализа для *систематизации* и *изучения* большого массива данных о биологических макромолекулах» (Luscombe, 2001)
- Цели биоинформатики :
 - Обеспечить хранение и доступ к существующей информации
 - Разработка методов и ресурсов для анализа данных
 - Анализ и биологически осмысленная интерпретация данных
- Подходы биоинформатики:
 - Сравнение и группировка данных на основании биологически значимого сходства
 - Анализ данных одного типа для прогнозирования свойств данных другого типа

Зачем нужны базы данных в биоинформатике?

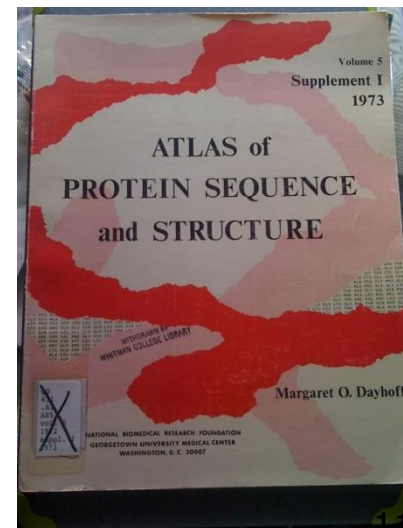
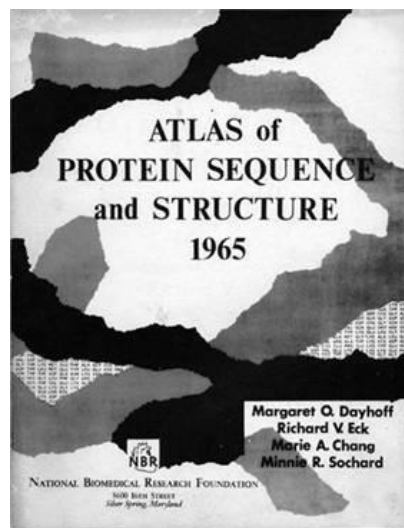
Публично доступные базы данных – основа как теоретической биоинформатики, так и исследований с использованием методов биоинформатики

- Сбор информации из различных источников
- Распространение информации
- Предоставление доступа для автоматизированного анализа

«Atlas of Protein Sequences and Structures»

Первая база данных в биоинформатике

- Исторически, первая база данных в биоинформатике содержала аминокислотные последовательности белков;
- В 1960-е года Маргарет Дэйхоф (Margaret Dayhoff) с коллегами собрали всю информацию об известных на тот момент белковых последовательностях (могли бы уместиться на одной дискете);
- Результаты работы были *изданы в виде книги* в 1965;
- Новые издания книги выходили до 1978



Публичные базы данных в биоинформатике

- **первичные** базы данных последовательностей и структур белков и нуклеиновых кислот;
- **вторичные** базы данных: анализ и систематизация исходной информации;
- **консолидированные** базы данных о структуре и функции ферментов (MACIE, CSA, BRENDA)

Первичные базы данных нуклеотидных последовательностей

Построены на основании экспериментально полученной информации, например последовательностей и структур белков и аминокислот

Основные первичные базы данных нуклеотидных последовательностей :



EMBL (European Bioinformatics Institute) : www.ebi.ac.uk/embl



GenBank (National Center for Biotechnology Information):
www.ncbi.nlm.nih.gov/GenBank



DDBJ (DNA Data Bank of Japan): www.ddbj.nig.ac.jp

Особенности:

- Единая схема маркировки и обновления данных (International Nucleotide Sequence Database Collaboration, 1988)
- Ежедневно синхронизируются
- Без ограничений на использование данных (некоторые данные запатентованы)

Первичные базы данных нуклеотидных последовательностей

Пример записи в базе данных белковых последовательностей

```
LOCUS AF111785 5925 bp mRNA PRI 01-SEP-1999
DEFINITION Homo sapiens myosin heavy chain IIx/d mRNA, complete cds.
ACCESSION AF111785
VERSION AF111785.1 GI:4808814
KEYWORDS .
SOURCE human.
ORGANISM Homo sapiens
... ..
sig peptide 160..231
CDS 160..>2301
/codon start=1
/product="EGF-receptor"
/protein id="CAA42219.1"
/db xref="GI:50804"
/db xref="MGD:MGI:95294"
/db xref="SWISS-PROT:Q01279"
/translation="MRPSGTARTLLVLLTALCAAGGALEEKKVCQGTSNRLTQLGTF
EDHFLSLQRMYNNCSEVVLGNLEITYVQRNYDLSFLKTIQEVAGYVLIALNTVERIPLE
NLQIIRGNALYENTYALAILSNYGTNRTGLRELPMRNLQELIGAVRFSNNPILCNMD
... ..
RSIKEISDGDVIISGNRNL CYANTINWKKLFGTPNQTKIMNNRAEKDCKAVNHVCNP
LCSSEGCWGPPEPRDCVSCQNVSRGRECVKWNILEGEPREFVENSECIQCHPECLPQA
MNITCTGRGPDNCIQCAHYIDGPHCVKTCPAGIMGENTLVWKYADANNVCHLCHANC
TYGCAGPGLQGCEVWPSPGPKIPSIATGIVGGLLFIVVVALGIGLFMRRRHIVRKRTL
RLQERELVEPLTPSGEAPNQAHRLILKETEF"
mat peptide 232..>2301
/product="EGF-receptor"
... ..
ORIGIN
1 atgagttctg actctgagat ggccattttt ggggaggctg ctcctttcct ccgaaagtct
61 gaaagggagc gaattgaagc ccagaacaag ccttttgatg ccaagacatc agtctttgtg
121 gtggacccta aggagtcctt tgtgaaagca acagtgcaga gcagggaagg ggggaagggtg
... ..
5701 cggaggatcc agcacgagct ggaggaggcc gaggaaaggg ctgacattgc tgagtccag
5761 gtaacaagc tgagggtgaa gagcaggag gttcacacaa aatcataag tgaagagtaa
5821 ttatctaac tgctgaaagg tgaccaaaga aatgcacaaa atgtgaaaat ctttgtcact
5881 ccattttgta cttatgactt ttggagataa aaaatttatc tgcca
//
```

Первичные базы данных белковых последовательностей

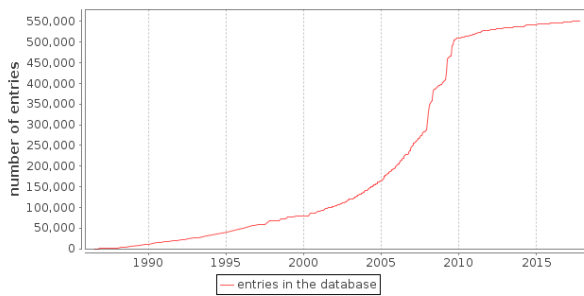
Swiss-Prot

(UniProtKB/Swiss-Prot)

База данных аннотированных белков

- Высокий уровень достоверности аннотации (данные тщательно анализируются и проверяются специалистами)
- Интегрированность с другими базами данных
- Минимум повторений
- Создан в 1986 году и курируется EBI и SBI
- **01-Мая-13** в UniProtKB/Swiss-Prot содержалось **540.052 записей**, на **14-Ноя-2017** – **556,006 записей**.

Number of entries in UniProtKB/Swiss-Prot over time



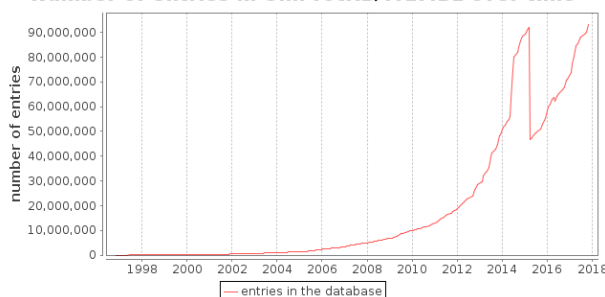
TrEMBL

(UniProtKB/TrEMBL)

Translation of EMBL Nucleotide Sequence Databases

- Позволяет быстро получать информацию о новых последовательностях
- Компьютерная аннотация транслированных CDS
- По состоянию
- на **01-Мая-13** в UniProtKB/TrEMBL содержится **33.995.348 записей**,
- на **17-Ноя-2014** - **86.536.393 записей**,
- на **14-Ноя-2017** – **93,236,986 записей**.

Number of entries in UniProtKB/TrEMBL over time



Первичные базы данных структур белков и нуклеиновых кислот

PDB



Protein Data Bank

Единый репозиторий для сбора и распространения информации о трехмерной структуре биологических макромолекул

NDB



Nucleic Acid Database

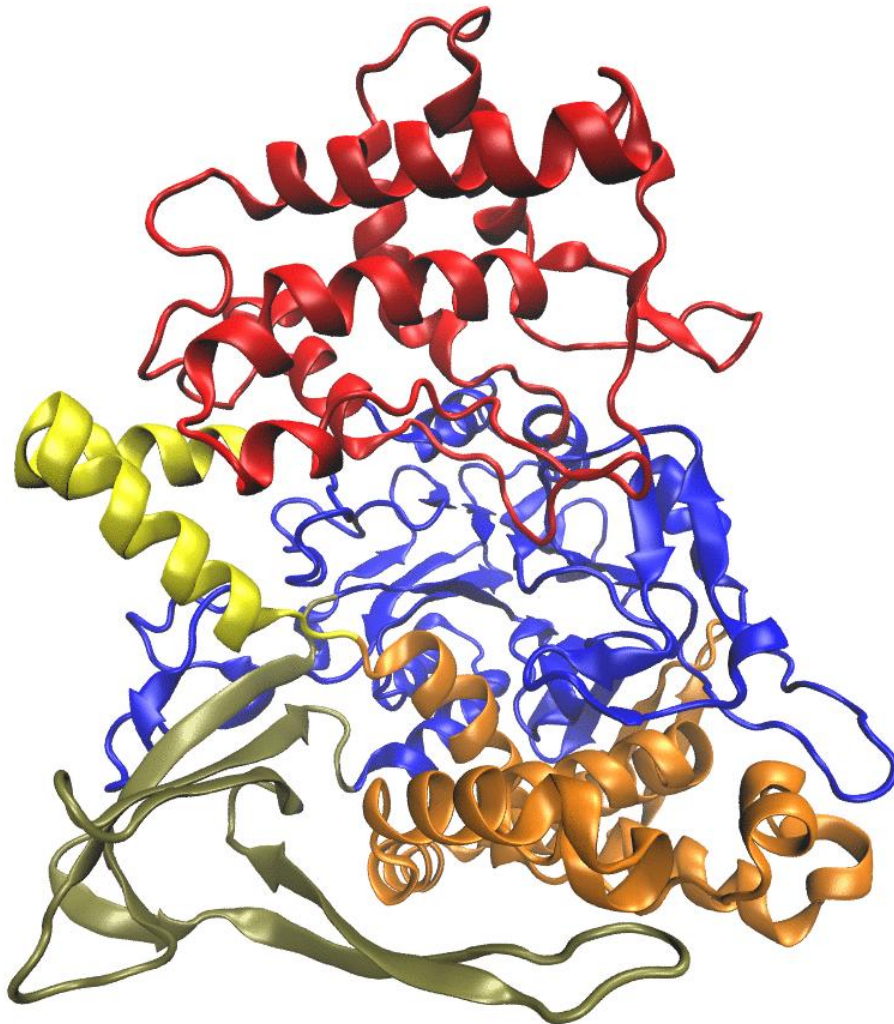
Репозиторий трехмерной информации о нуклеиновых кислотах

Первичные базы данных структур белков и нуклеиновых кислот

ATOM	1731	N	GLY	B	7	29.049	12.423	-10.592	1.00	11.39	N
ATOM	1732	CA	GLY	B	7	30.352	13.003	-10.347	1.00	11.61	C
ATOM	1733	C	GLY	B	7	30.715	14.015	-11.482	1.00	13.95	C
ATOM	1734	O	GLY	B	7	29.965	14.215	-12.439	1.00	14.60	O
ATOM	1735	N	LYS	B	8	31.884	14.589	-11.255	1.00	14.78	N
ATOM	1736	CA	LYS	B	8	32.381	15.672	-12.112	1.00	16.17	C
ATOM	1737	C	LYS	B	8	32.260	15.444	-13.588	1.00	18.47	C
ATOM	1738	O	LYS	B	8	31.785	16.290	-14.370	1.00	18.69	O
ATOM	1739	CB	LYS	B	8	33.822	16.001	-11.646	1.00	20.19	C
ATOM	1740	CG	LYS	B	8	34.319	17.295	-12.298	1.00	22.94	C
ATOM	1741	CD	LYS	B	8	35.651	17.667	-11.668	1.00	23.91	C
ATOM	1742	CE	LYS	B	8	36.255	18.918	-12.242	1.00	25.20	C
ATOM	1743	NZ	LYS	B	8	37.654	19.136	-11.747	1.00	26.74	N
ATOM	1744	N	SER	B	9	32.551	14.209	-14.072	1.00	16.53	N
ATOM	1745	CA	SER	B	9	32.554	13.907	-15.488	1.00	18.52	C
ATOM	1746	C	SER	B	9	31.186	13.686	-16.046	1.00	17.69	C
ATOM	1747	O	SER	B	9	30.984	13.622	-17.281	1.00	18.78	O
ATOM	1748	CB	SER	B	9	33.454	12.662	-15.736	1.00	22.10	C
ATOM	1749	OG	SER	B	9	32.743	11.528	-15.151	1.00	20.23	O
ATOM	1750	N	LYS	B	10	30.130	13.645	-15.220	1.00	16.88	N
ATOM	1751	CA	LYS	B	10	28.797	13.436	-15.672	1.00	13.54	C
ATOM	1752	C	LYS	B	10	27.804	14.550	-15.348	1.00	12.50	C
ATOM	1753	O	LYS	B	10	26.625	14.515	-15.723	1.00	15.99	O
ATOM	1754	CB	LYS	B	10	28.228	12.118	-15.159	1.00	18.14	C

Вторичные базы данных структур белков

CATH структурная классификация доменов



A01 (α 3-149)

A02 (α 150-179)

B01

- B01-01 (β 1-72)
- B01-02 (β 146-290)
- B01-03 (β 452-536)

B02 (β 73-145)

B03 (β 291-451)

Методы

Ключевые классы методов в биоинформатике

Сравнительный анализ биологических последовательностей и структур

- «Выравнивания» белковых **последовательностей**
 - **поиск** по сходству последовательностей;
 - парные и множественные **выравнивания**;
- «Выравнивания» белковых **структур**
 - **поиск** по сходству структур;
 - парные и множественные **выравнивание** структур белков;

Множественное выравнивание белковых последовательностей

	10	20	30	40	50	60	70	80																																																																	
P08819.2=CBP2_WHE	--VANVLF	LDSPAGV	GFSYNT	---	SSDIY	TSGDN	RTHADS	YAF	LAKWFERFPHKYRDFYIAGESYAG-HYVPELSQLVHRSKN----																																																																
AAD01263.1=_gluco	WT	KVAN	IIFID	QPA	G	TGYS	YANT	---	SEAYNCN-DTL	SVTLT	YDF	L	R	K	W	L	M	D	H	P	E	Y	L	N	N	P	L	Y	V	G	G	S	Y	S	G	-	I	F	V	A	L	L	T	R	K	I	Y	D	G	I	E	V	G	D	K																		
AAD01265.1=_gluco	WT	KVAN	IIFID	QPA	G	TGYS	YANT	---	SEAYNCN-DTL	SVTLT	YDF	L	R	K	W	L	M	D	H	P	E	Y	L	N	N	P	L	Y	V	G	G	S	Y	S	G	-	I	F	V	A	L	L	T	R	K	I	Y	D	G	I	E	V	G	D	K																		
AAF64227.1=AF248	WT	KVAN	IIFID	QPA	G	TGYS	YANT	---	SEAYNCN-DTL	SVTLT	YDF	L	R	K	W	L	M	D	H	P	E	Y	L	N	N	P	L	Y	V	G	G	S	Y	S	G	-	I	F	V	A	L	L	T	R	K	I	Y	D	G	I	E	V	G	D	R																		
AAD01264.1=_gluco	WT	KVAN	IIFID	QPA	G	TGYS	YAKT	---	SEAYNCN-DTL	SVTLT	YDF	L	R	K	W	L	M	D	H	P	E	Y	L	N	N	P	L	Y	V	G	G	S	Y	S	G	-	I	F	V	A	L	L	T	R	K	I	Y	D	G	I	E	V	G	D	R																		
BAF99696.1=_1Oacy	WT	KEAS	IIFVD	SPV	G	TGYS	YNT	---	FEGYH	ST	DHKA	DDL	YAF	L	R	K	W	L	K	H	P	E	Y	L	N	N	P	L	Y	V	G	G	S	Y	S	G	-	K	F	V	A	L	V	W	R	I	S	Q	I	D	A	G	H	E																			
BAF99699.1=_1Oacy	L	TKVAS	IIFID	SPV	K	A	G	F	S	Y	A	T	---	YEGYN	M	S	D	T	K	A	K	E	A	T	T	F	L	K	K	W	L	L	E	H	P	E	F	D	K	N	P	L	Y	I	A	G	D	S	Y	A	G	-	L	I	V	P	M	V	F	H	V	S	N	A	I	E	A	G	Q	M			
XP_002269439.1=_PWNT	KVAS	IIFID	APV	G	T	G	F	S	Y	A	E	---	SYGYN	V	S	D	T	S	A	A	Q	T	Y	Q	F	L	R	K	W	L	T	F	H	N	F	A	G	N	P	L	Y	I	G	G	S	Y	S	G	-	I	V	A	P	I	L	I	K	D	I	L	H	G	-	L	E	V	G	L					
XP_002522133.1=_s	WT	KVAS	IIFID	APV	G	T	G	F	S	Y	A	T	---	NEG	A	N	T	S	D	T	L	A	H	Q	S	I	F	L	R	K	W	L	M	N	H	P	K	F	L	G	S	Q	I	Y	I	G	G	D	S	Y	S	G	-	I	I	V	P	L	L	V	Q	N	I	L	E	G	-	I	E	S	C	L	
BAF56655.1=_serine	WT	KVAN	IIFID	QPA	G	T	G	F	S	Y	A	T	---	AAA	Y	N	S	S	D	T	V	A	A	A	Q	V	K	F	L	R	K	W	L	M	Y	N	P	K	F	A	N	P	L	Y	V	G	G	D	L	Y	S	G	-	I	I	V	P	L	L	V	Q	T	I	L	D	G	-	I	G	S	G		
CAN74923.1=_hypot	WT	KVAS	IIFLD	APV	G	T	G	F	S	Y	A	T	---	PDD	Y	A	S	-	D	T	I	A	R	D	N	I	F	I	R	K	W	L	I	D	H	P	R	F	L	N	P	L	Y	I	G	G	D	S	Y	S	G	-	I	I	V	P	I	L	T	L	E	I	A	N	G	-	I	Q	M	C	L		
CB12424.1=_unnam	WT	KVAS	IIFLD	APV	G	T	G	F	S	Y	A	T	---	PDD	Y	A	S	-	D	T	I	A	R	D	N	I	F	I	R	K	W	L	I	D	H	P	R	F	L	N	P	L	Y	I	G	G	D	S	Y	S	G	-	I	I	V	P	I	L	T	L	E	I	A	N	G	-	I	Q	M	C	L		
XP_002522137.1=_s	WT	KVAN	IIFLD	APV	G	T	G	F	S	Y	A	T	---	QEG	Y	T	S	-	D	T	S	Q	A	I	Y	Q	F	L	R	K	W	L	V	D	H	P	K	F	K	N	P	L	Y	V	A	G	D	S	Y	S	G	-	M	I	V	P	T	V	I	E	I	A	K	G	-	N	K	A	K	H			
CB12420.1=_unnam	WT	KLAS	IIFLD	APV	G	T	G	F	S	Y	A	T	---	AEG	Y	N	M	-	D	T	L	S	A	Q	I	Y	A	F	L	R	K	W	L	I	N	H	P	K	F	K	N	P	L	Y	V	S	G	D	S	Y	S	G	-	I	I	I	P	M	V	Q	E	I	S	N	G	-	N	D	E	C	K		
XP_002266354.1=_PWNT	KLAS	IIFLD	APV	G	T	G	F	S	Y	A	T	---	AEG	Y	N	M	-	D	T	L	S	A	Q	I	Y	A	F	L	R	K	W	L	I	N	H	P	K	F	K	N	P	L	Y	V	S	G	D	S	Y	S	G	-	I	I	I	P	M	V	Q	E	I	S	N	G	-	N	D	E	C	K			
CB12421.1=_unnam	WT	KVAS	IIFLD	SPV	G	T	G	F	S	Y	A	S	---	SEG	Y	R	T	-	D	S	L	A	A	A	Q	Y	D	F	L	K	K	W	L	I	D	H	P	E	F	L	R	N	R	L	Y	I	A	G	D	S	Y	S	G	-	L	F	V	P	I	I	A	K	I	S	O	G	-	N	E	A	Q	K	
XP_002298786.1=_p	WT	KIAN	IIFVD	APV	G	T	G	F	S	Y	A	T	---	WEG	Y	Q	S	-	D	T	L	A	A	A	Q	Y	E	F	L	R	K	W	L	V	D	H	P	R	F	L	T	N	P	L	Y	V	A	G	D	S	Y	S	G	-	I	V	A	P	I	I	V	Q	E	I	S	O	G	-	N	E	V	G	R
XP_002300280.1=_p	WT	KIAN	IIFVD	APV	G	T	G	F	S	Y	A	T	---	WEG	Y	Q	S	-	D	T	L	A	A	A	Q	Y	E	F	L	R	K	W	L	V	D	H	P	R	F	L	T	N	P	L	Y	V	A	G	D	S	Y	S	G	-	I	V	A	P	I	I	V	Q	E	I	S	O	G	-	N	E	V	G	R
XP_002522135.1=_s	WT	KVAN	IIFLD	APV	G	T	G	F	S	Y	A	T	---	GEA	Y	H	N	-	D	T	I	A	A	E	I	Y	D	F	L	R	K	W	L	V	S	N	P	K	F	L	A	N	P	L	Y	I	G	G	D	S	Y	S	G	-	I	I	V	P	I	V	Q	E	I	L	N	G	-	N	E	M	C	L	
CB117848.1=_unnam	WT	KLIN	IIFVD	APV	G	T	G	F	S	Y	A	T	---	QEG	Y	I	M	-	D	L	K	Y	A	A	Q	T	Y	E	F	L	K	K	W	L	V	D	H	P	E	F	L	K	N	E	L	Y	V	G	G	D	S	Y	S	G	-	I	P	V	P	M	V	Q	E	I	Y	G	S	-	---	---			
XP_002268517.1=_PWNT	S	GLNI	IIFVD	TPV	G	A	G	F	S	Y	A	T	---	QEG	Y	Y	S	-	D	Y	K	S	T	H	T	Y	E	F	L	N	K	W	L	D	H	P	E	F	L	K	N	N	L	Y	V	G	G	D	S	Y	S	G	-	I	V	L	P	M	I	T	E	K	I	Y	Y	G	N	I	G	T	F		
XP_002268642.1=_PWNT	E	TLNI	IIFLD	APV	G	T	G	F	S	Y	A	T	---	QEG	Y	T	S	-	D	Y	K	S	A	A	Q	I	Y	E	F	L	K	K	W	L	I	Q	H	P	E	F	L	K	N	N	L	Y	I	G	G	D	S	Y	S	G	-	I	P	V	P	M	I	V	Q	E	I	Y	S	E	R	G	S		
CB124085.1=_unnam	WT	QGLN	IIFID	APV	G	T	G	F	S	Y	A	T	---	SGY	Y	V	D	-	D	A	E	N	A	A	Q	T	Y	E	F	L	R	K	W	L	V	Q	H	P	D	F	L	G	N	E	L	Y	I	A	G	V	S	Y	S	G	-	I	P	V	P	M	I	V	N	E	I	E	C	N	I	L	G	S	
BAH89272.1=_putati	WT	KTAS	IIFLD	QPV	G	T	G	F	S	Y	A	T	---	QEG	W	P	S	-	D	T	K	S	E	Q	S	Y	E	F	L	K	K	W	L	E	E	N	P	Y	L	K	V	Q	L	F	V	G	G	D	S	Y	A	G	-	K	I	V	P	L	V	R	L	I	A	D	G	N	K	N	G	T			
XP_002272116.1=_PWNT	KTAS	IIFLD	APV	G	T	G	F	S	Y	A	T	---	ADG	W	S	S	-	D	T	D	S	A	L	E	T	Y	E	F	L	R	K	W	L	I	E	H	P	K	Y	L	P	L	Q	L	Y	V	G	G	D	S	Y	S	G	-	I	I	V	P	L	V	K	H	I	V	A	I	D	E	H	T	V		
CB125147.1=_unnam	WT	KTAN	IIFVD	APV	G	T	G	F	S	Y	A	T	---	SGA	Y	T	T	-	D	T	L	S	V	I	Q	T	L	G	F	L	R	N	W	L	N	D	H	P	D	F	K	L	N	P	L	F	I	A	G	D	S	Y	S	G	-	L	I	A	P	I	I	A	G	E	I	M	D	G	N	G	V	G	E
XP_002266867.1=_PWNT	KTAN	IIFVD	APV	G	T	G	F	S	Y	A	T	---	SGA	Y	T	T	-	D	T	L	S	V	I	Q	T	L	G	F	L	R	N	W	L	N	D	H	P	D	F	K	L	N	P	L	F	I	A	G	D	S	Y	S	G	-	L	I	A	P	I	I	A	G	E	I	M	D	G	N	G	V	G	E	
XP_002300184.1=_p	WT	KTAS	IIFID	QPV	G	T	G	F	S	Y	A	T	---	SEG	Y	H	M	-	D	T	G	S	A	K	L	V	H	Q	F	L	R	T	W	L	I	D	H	P	E	T	K	I	P	F	V	A	S	D	T	Y	A	G	-	I	I	T	P	I	V	A	K	E	I	F	D	G	N	E	A	L	D		
AAG51076.1=AC069	WT	KVAN	IIFLD	QPV	G	T	G	F	S	Y	A	T	---	PFA	D	I	P	S	-	D	T	G	S	V	K	R	N	E	F	V	R	K	W	L	A	K	H	P	E	Y	F	S	N	P	F	Y	V	T	G	N	S	Y	S	G	-	K	V	I	P	A	I	V	Q	E	I	S	N	G	N	Y	I	C	K
Q9C7D2.2=SCP15	AWNT	KVAN	IIFLD	QPV	G	T	G	F	S	Y	A	T	---	PFA	D	I	P	S	-	D	T	G	S	V	K	R	N	E	F	V	R	K	W	L	A	K	H	P	E	Y	F	S	N	P	F	Y	V	T	G	N	S	Y	S	G	-	K	V	I	P	A	I	V	Q	E	I	S	N	G	N	Y	I	C	K
BA803132.1=_serine	WT	KVAN	IIFLD	QPV	G	T	G	F	S	Y	A	T	---	PFA	D	R	P	S	-	D	T	G	S	A	K	L	N	E	F	V	R	K	W	L	A	K	H	P	E	Y	F	S	N	P	F	Y	V	T	G	N	S	Y	S	G	-	K	V	I	P	A	I	V	Q	E	I	S	N	G	N	Y	I	C	K
Q9C7D3.1=SCP14	AWNT	KVAN	IIFLD	QPV	G	T	G	F	S	Y	A	T	---	PFA	D	R	P	S	-	D	T	G	S	A	K	L	N	E	F	V	R	K	W	L	A	K	H	P	E	Y	F	S	N	P	F	Y	V	T	G	N	S	Y	S	G	-	K	V	I	P	A	I	V	Q	E	I	S	N	G	N	Y	I	C	K
Q9C7D4.1=SCP16	AWNT	KVAN	IIFLD	QPV	G	T	G	F	S	Y	A	T	---	PLA	D	I	P	S	-	D	T	G	S	A	K	R	V	D	E	F	L	H																																									

Основные критерии для построения множественного выравнивания

Структурный критерий:

Аминокислоты, играющие одну и ту же роль в структуре, окажутся **в одной колонке**.

Эволюционный критерий:

Аминокислоты, «произшедшие» от одной и той же аминокислоты в предковой последовательности, окажутся **в одной колонке**.

Функциональный критерий:

Аминокислоты, аналогичные по выполняемой функции, окажутся **в одной колонке**.

Схожесть алфавита:

Расположение аминокислот по колонкам должно приводить к выравниванию максимального сходства

Основные критерии для построения множественного выравнивания

✓ Структурный критерий:

Аминокислоты, играющие одну и ту же роль в структуре, окажутся **в одной колонке**.

Эволюционный критерий:

Аминокислоты, «произшедшие» от одной и той же аминокислоты в предковой последовательности, окажутся **в одной колонке**.

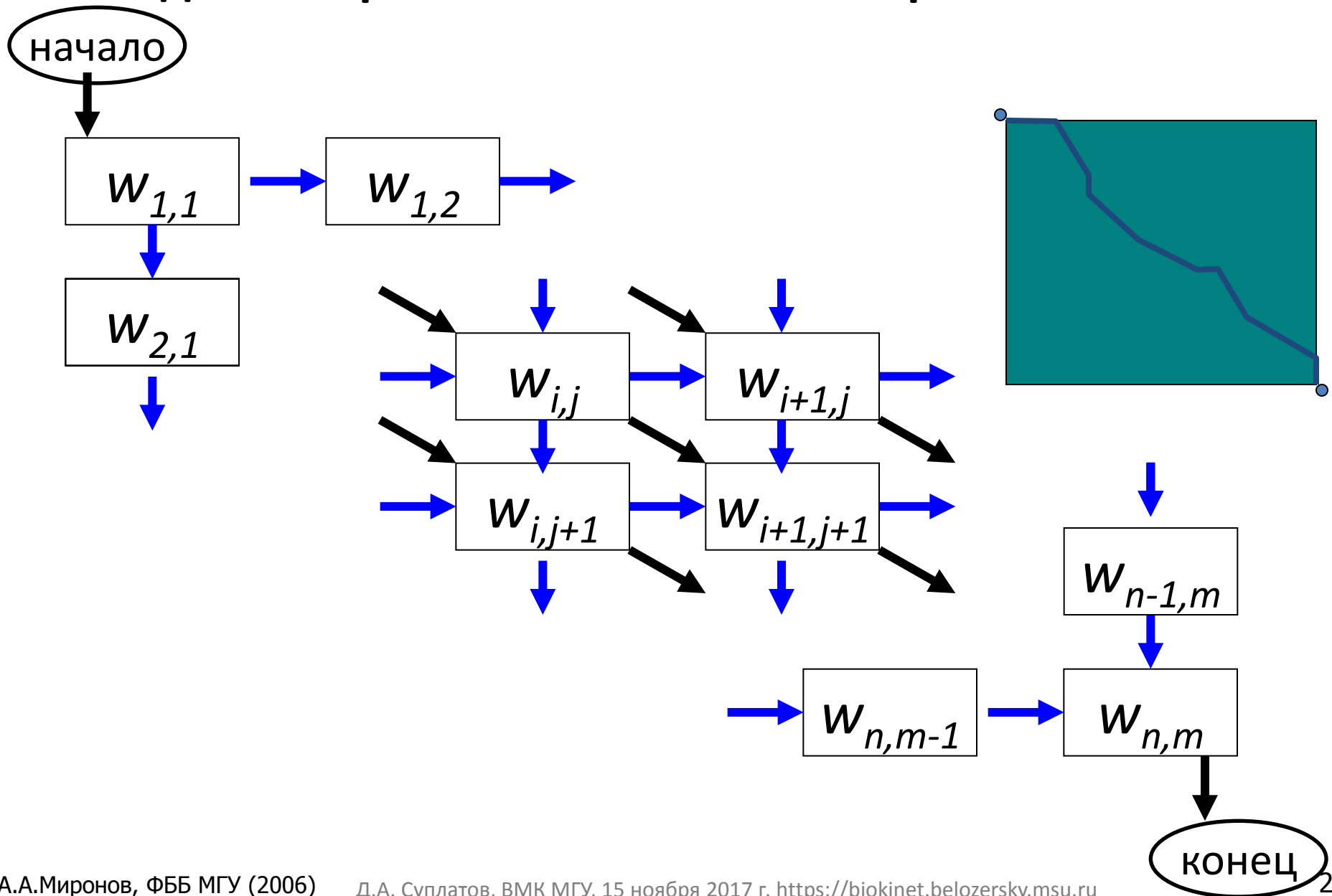
Функциональный критерий:

Аминокислоты, аналогичные по выполняемой функции, окажутся **в одной колонке**.

✓ Схожесть алфавита:

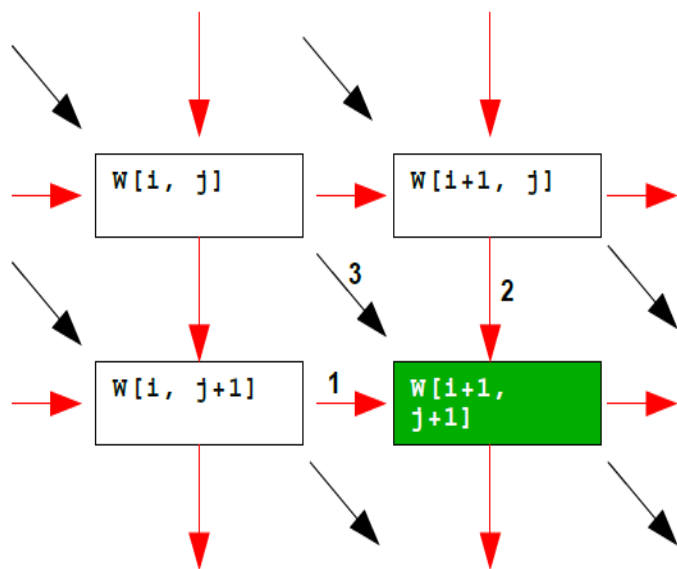
Расположение аминокислот по колонкам должно приводить к выравниванию максимального сходства

Алгоритм Нидлмана — Вунша для построения глобального выравнивания

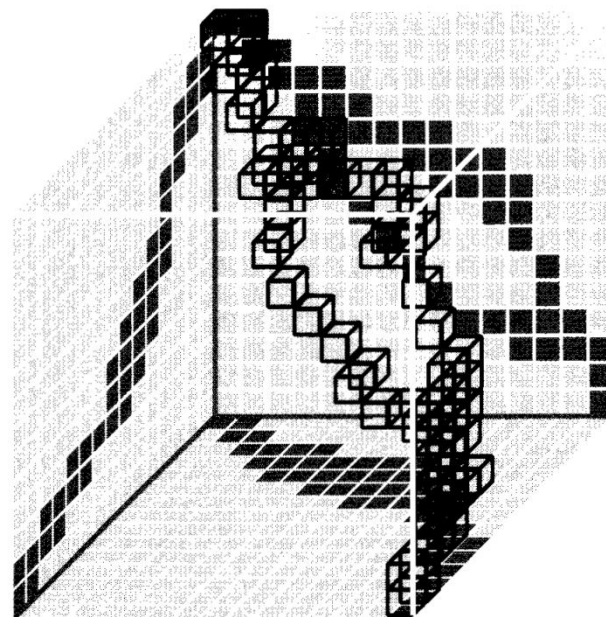


Множественное выравнивание

2D:



n-D:



Durbin R., Biological Sequence Analysis,
Cambridge University Press, 2002

HOW CAN I ALIGN *MANY* SEQUENCES



2 Globins => 1 Min

HOW CAN I ALIGN *MANY* SEQUENCES



3 Globins => 2 hours

HOW CAN I ALIGN *MANY* SEQUENCES



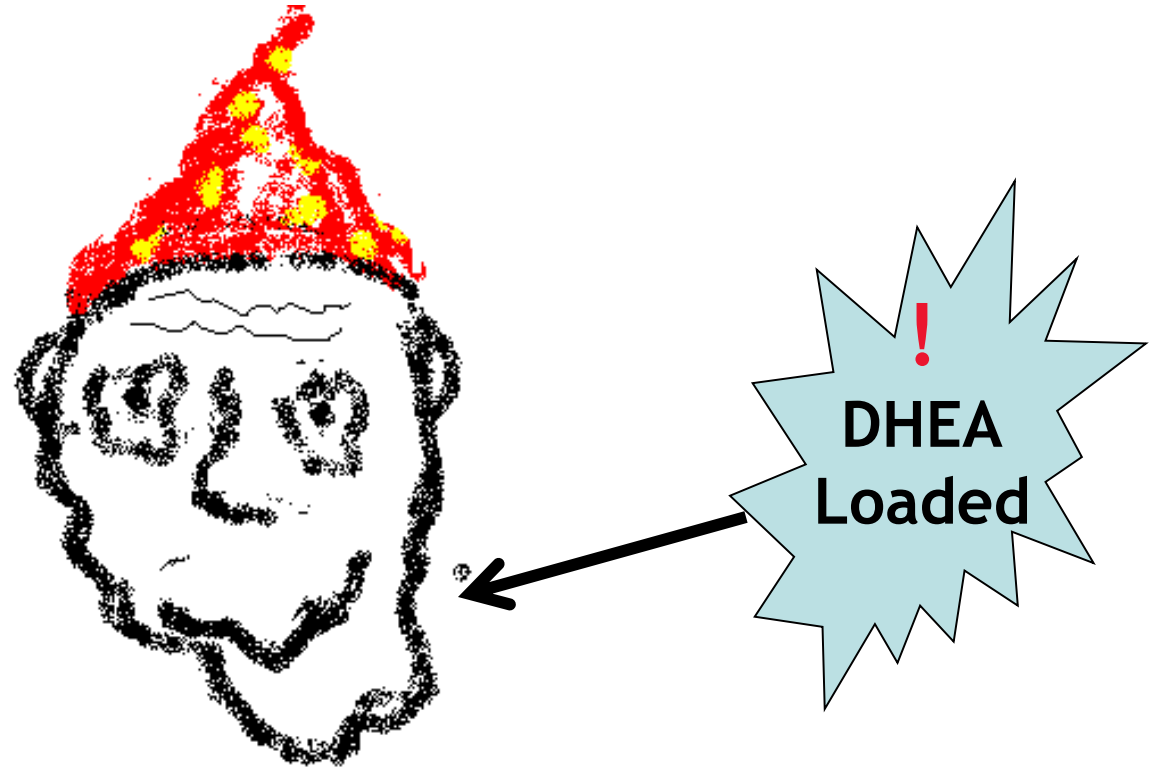
4 Globins => 10 days

HOW CAN I ALIGN *MANY* SEQUENCES



5 Globins => 3 years

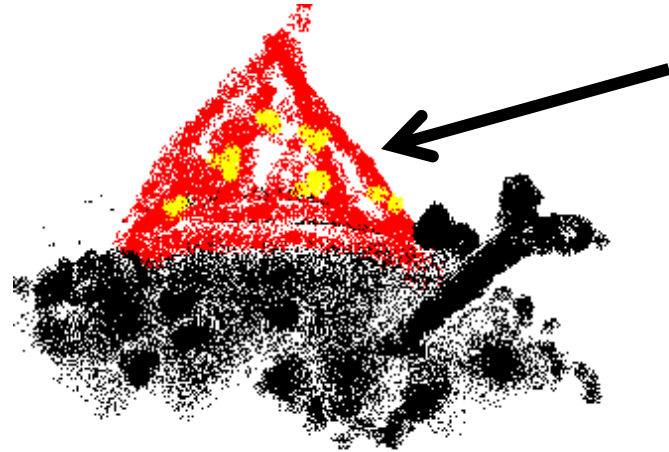
HOW CAN I ALIGN *MANY* SEQUENCES



6 Globins => 300 years

HOW CAN I ALIGN *MANY* SEQUENCES

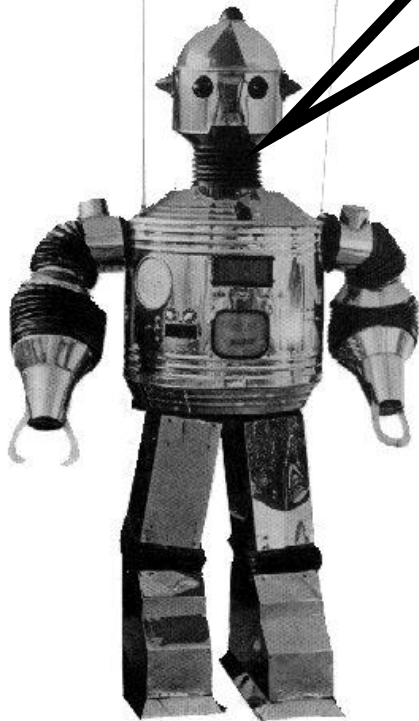
7 Globins => **30. 000** years



Solidified Fossil,
Old stuff

HOW CAN I ALIGN *MANY* SEQUENCES

HEMO-WHAT?



8 Globins => **3 Million** years

MSA: программы

Recent progress in multiple sequence alignment: a survey. C.Notredame,
Pharmacogenomics (2002), 3(1), 131-144

■ Точные алгоритмы

- MSA <http://www.ibc.wustl.edu/ibc/msa.html>
- DCA <http://bibiserv.techfak.uni-bielefeld.de/dca>

■ Прогрессивные алгоритмы

- ClustalW <ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW> or clustalX
- MultAlign <http://www.toulouse.inra.fr/multalin.html>

■ Итеративные алгоритмы

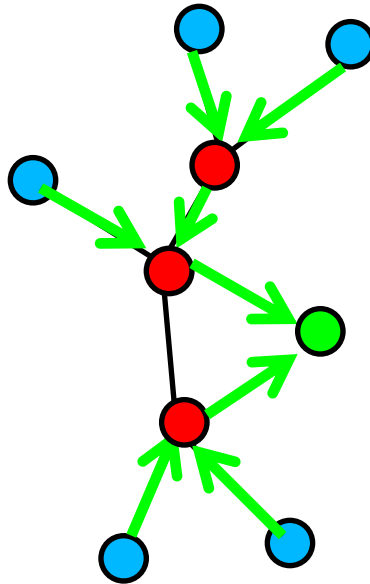
- Prrp <ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/>
- SAM rph@cse.ucsc.edu
- HMMER <http://hmmer.wustl.edu/>
- SAGA <http://igs-server.cnrs-mrs.fr/~cnotred>

■ Consistency-based

- T-Coffee <http://www.tcoffee.org>

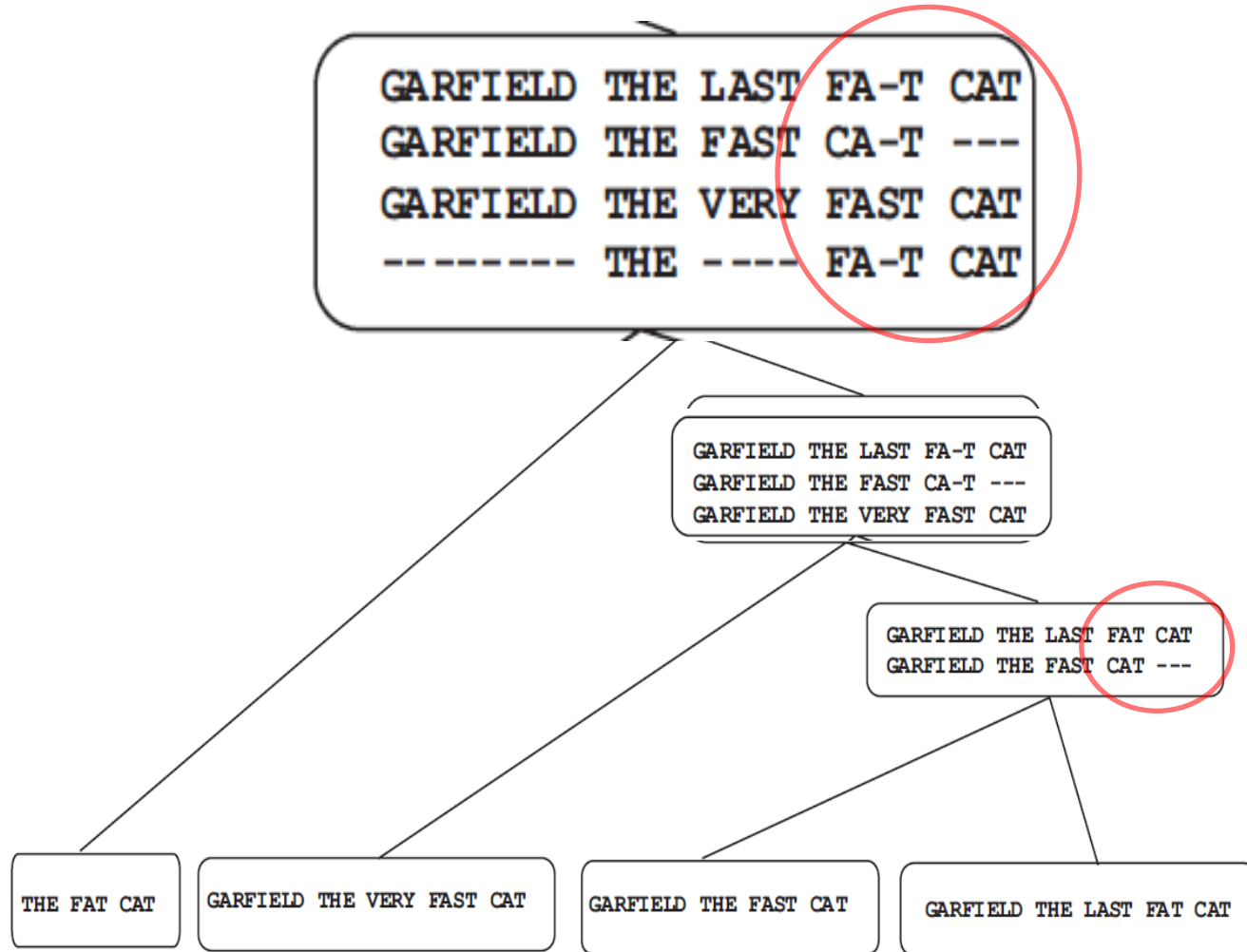
Прогрессивное выравнивание

Алгоритм ClustalW



Недостатки прогрессивного подхода

Однажды построенный профиль не перестраивается
(однажды сделанная ошибка влечет за собой другие)



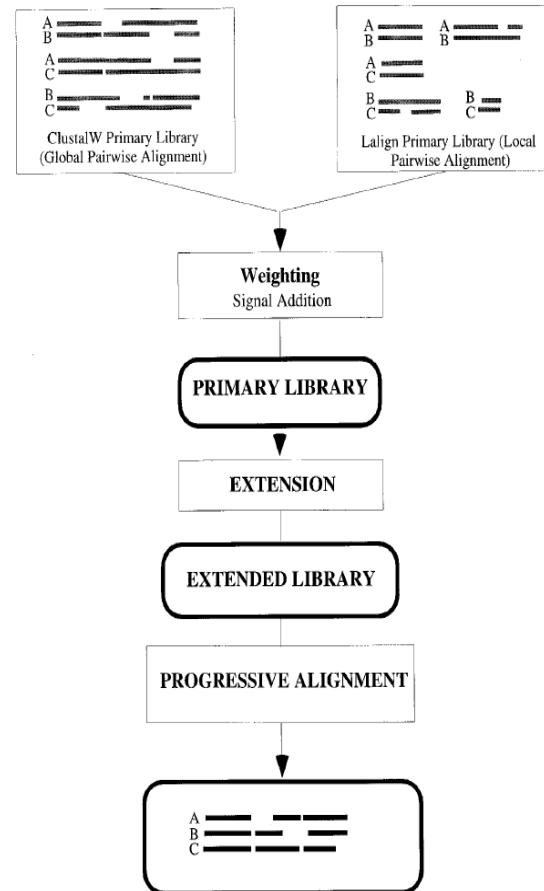
Recent progress in multiple sequence
alignment: a survey. C. Notredame,
Pharmacogenomics (2002), 3(1), 131-144

T-Coffee

Tree-based Consistency Objective Function for alignment Evaluation

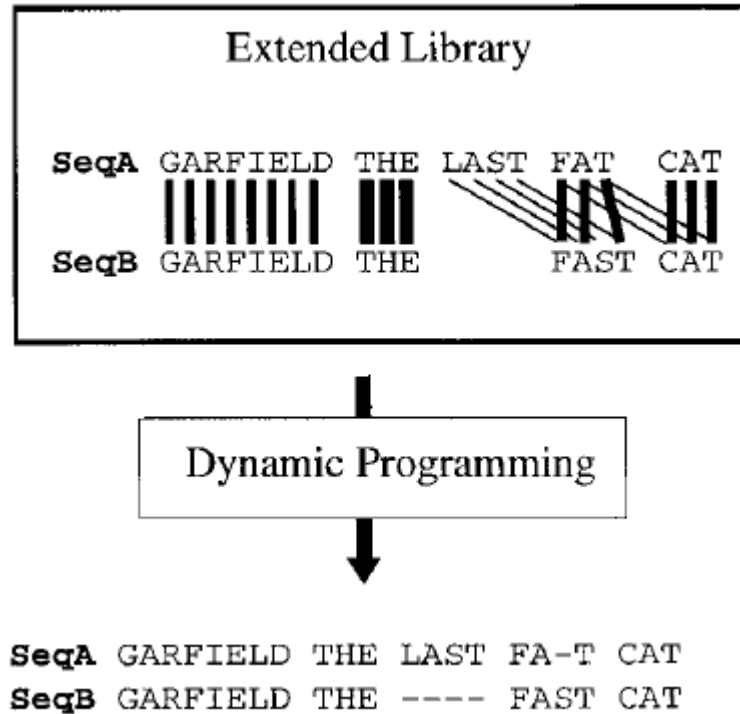
Основа метода:

- **комбинируем различные алгоритмы** (или программы) выравнивания последовательностей
- в качестве результата **выбираем выравнивание, которое наилучшим образом согласуется со всеми остальными**
- По сравнению с другими программами в меньшей степени зависит от конкретной характеристической функции и параметров выравнивания (штрафы за делеции, матрица аминокислотных замен)



T-Coffee

Расширение первичных библиотек

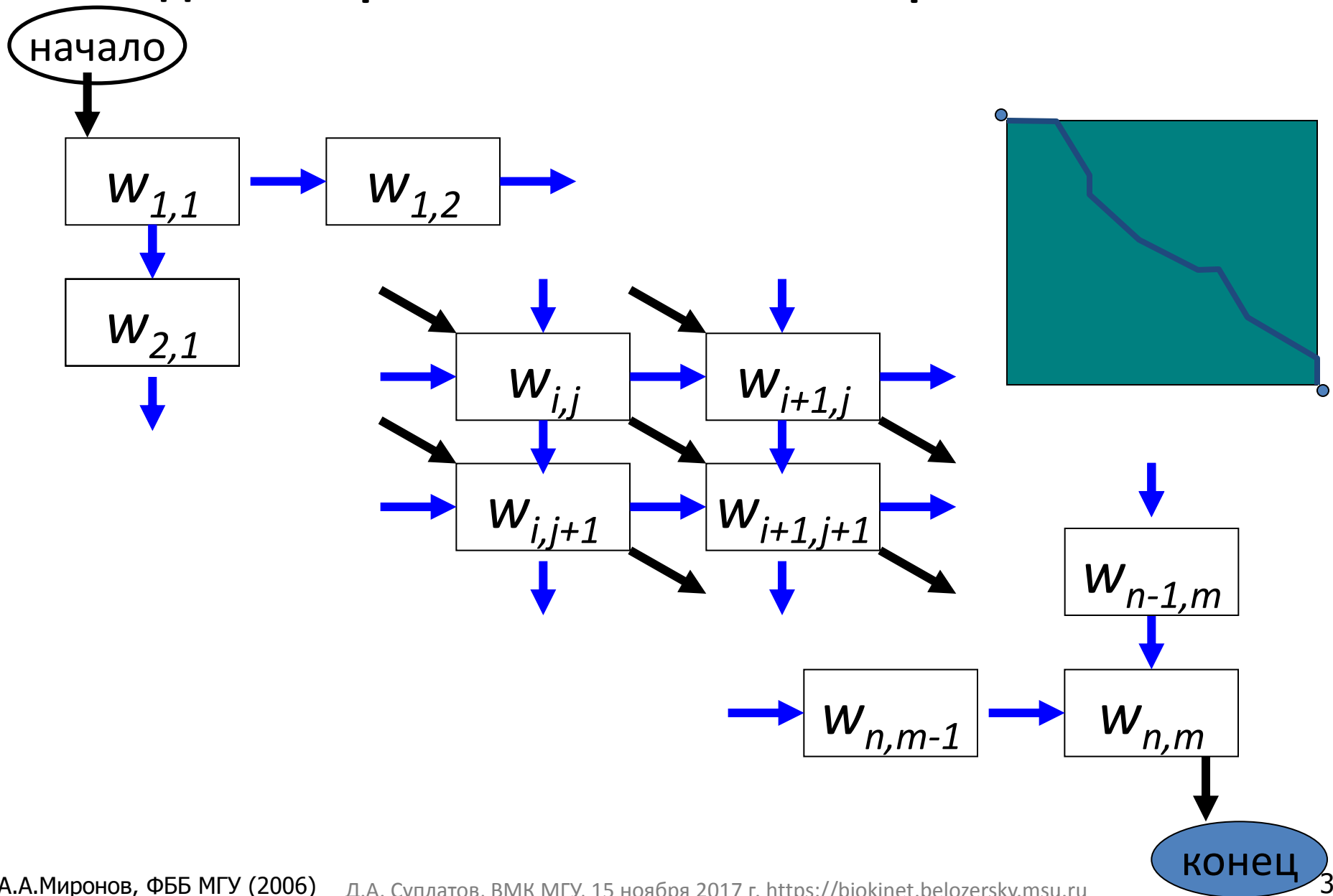


Очень медленно
работает, очень
требовательно к памяти,
пригодно только для
выравнивания
небольших выборок
белков

Проблема выравнивания аминокислотных последовательностей

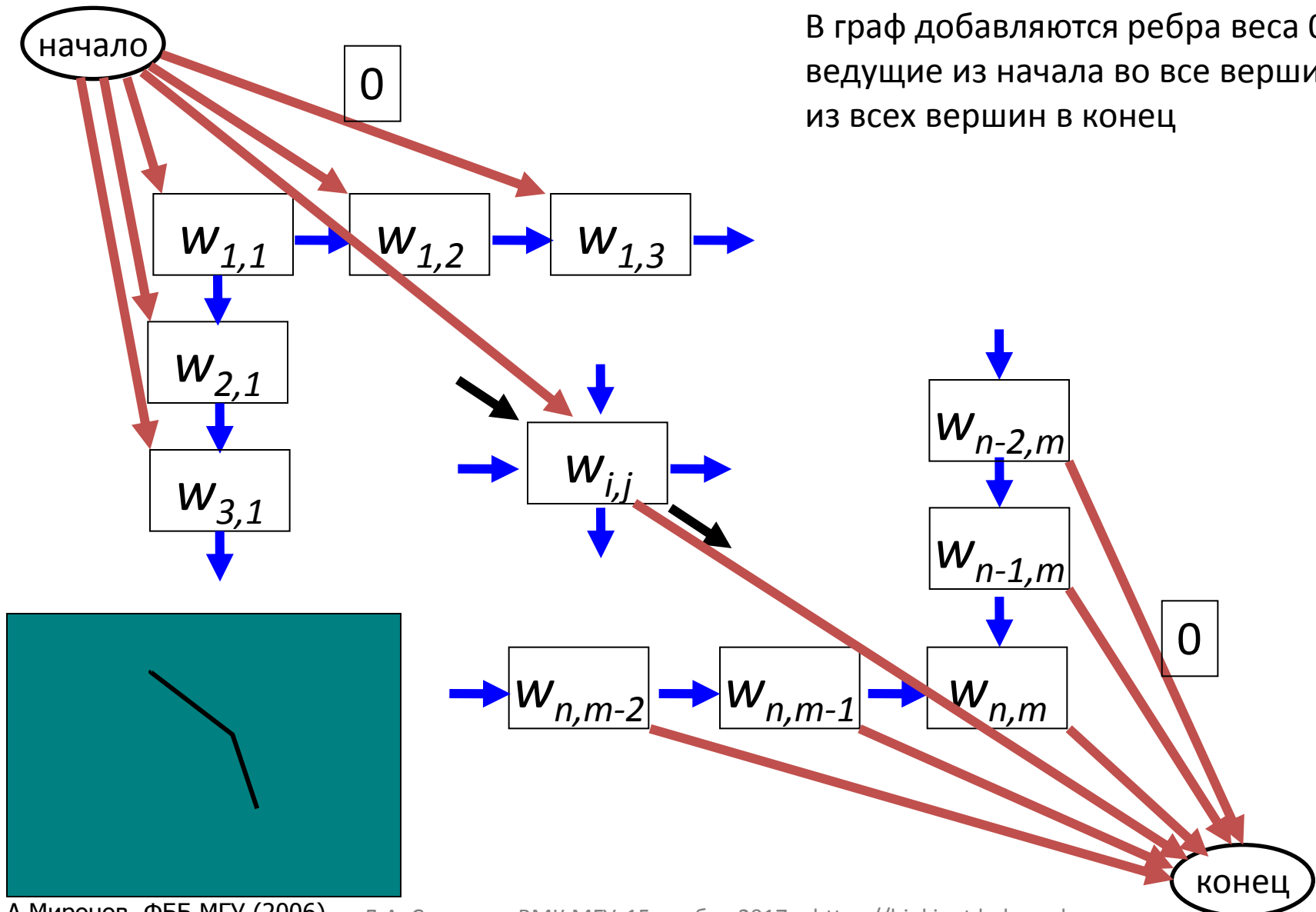
- Алгоритмы, которые гарантируют «правильное» решение не решают задачу за разумное время;
- «Быстрые алгоритмы», которые решают задачу за разумное время не гарантируют правильное решение;
- Решение: «быстрые алгоритмы» выравнивания аминокислотных последовательностей нужно использовать только для сравнения близких гомологов

Алгоритм Нидлмана — Вунша для построения глобального выравнивания



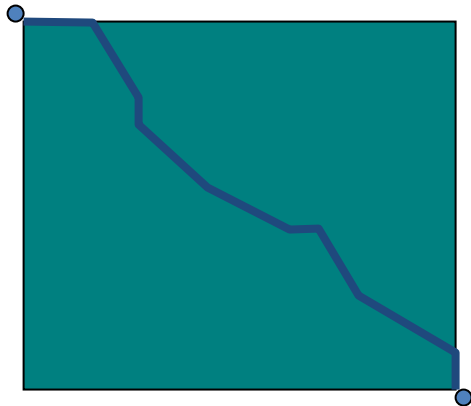
Алгоритм Смита — Ватермана для построения глобального выравнивания

В граф добавляются ребра веса 0, ведущие из начала во все вершины и из всех вершин в конец



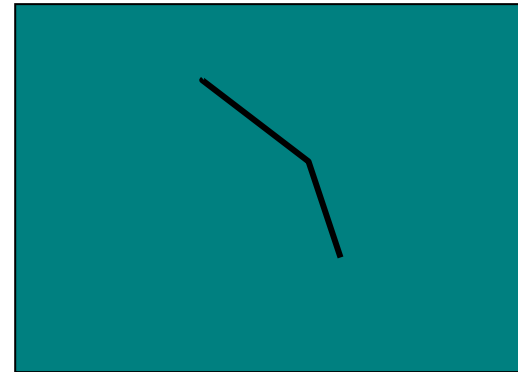
Для чего применяется каждый алгоритм (или зачем нужно локальное выравнивание если можно построить глобальное) ?

Алгоритм Нидлмана — Вунша



Алгоритмы построения глобальных выравниваний: интересны для задач, целью которых является **построение выравниваний**, т.е. дальше планируется использовать это выравнивание для решение какой-то проблемы (сравнение белков с заданными свойствами)

Алгоритм Смита — Ватермана

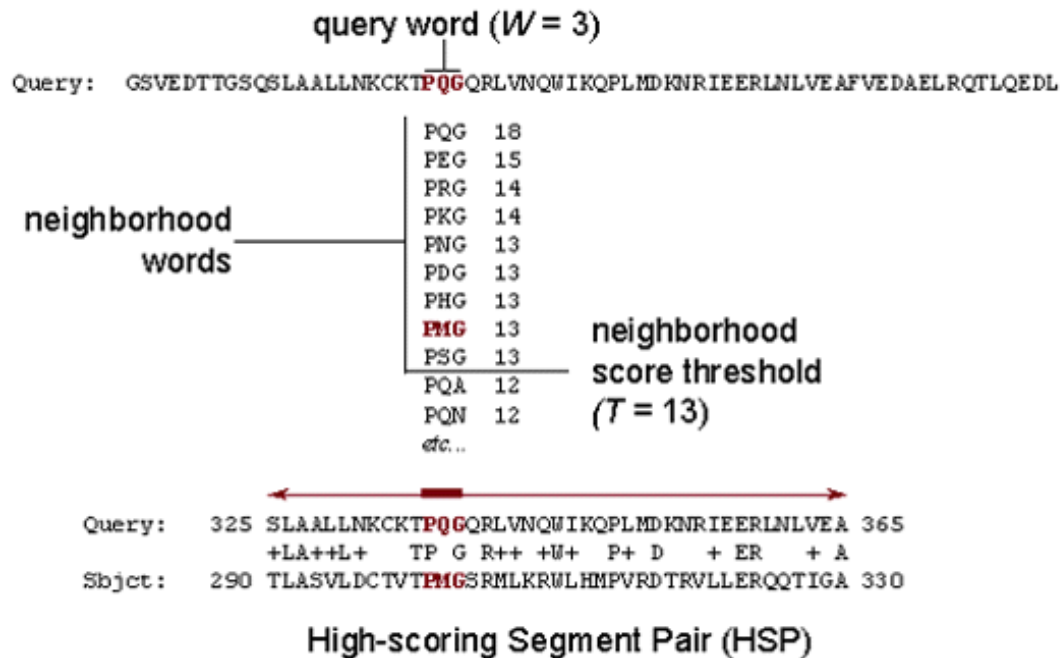


Алгоритмы построения локальных выравниваний: Для задач, целью которых является **поиск сходства** в больших базах данных, т.е. само выравнивание не нужно, интересует именно набор похожих белков

Подходы, основанные на сравнении последовательностей

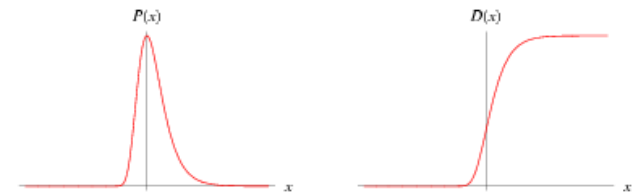
Best Local Alignment Search Tool (BLAST)

The BLAST Search Algorithm



Результаты поиска имеют распределение экстремального значения (extreme value distribution)

$$P(x) = \frac{e^{(a-x)/b} - e^{(a-x)/b}}{b} \quad D(x) = e^{-e^{(a-x)/b}}$$



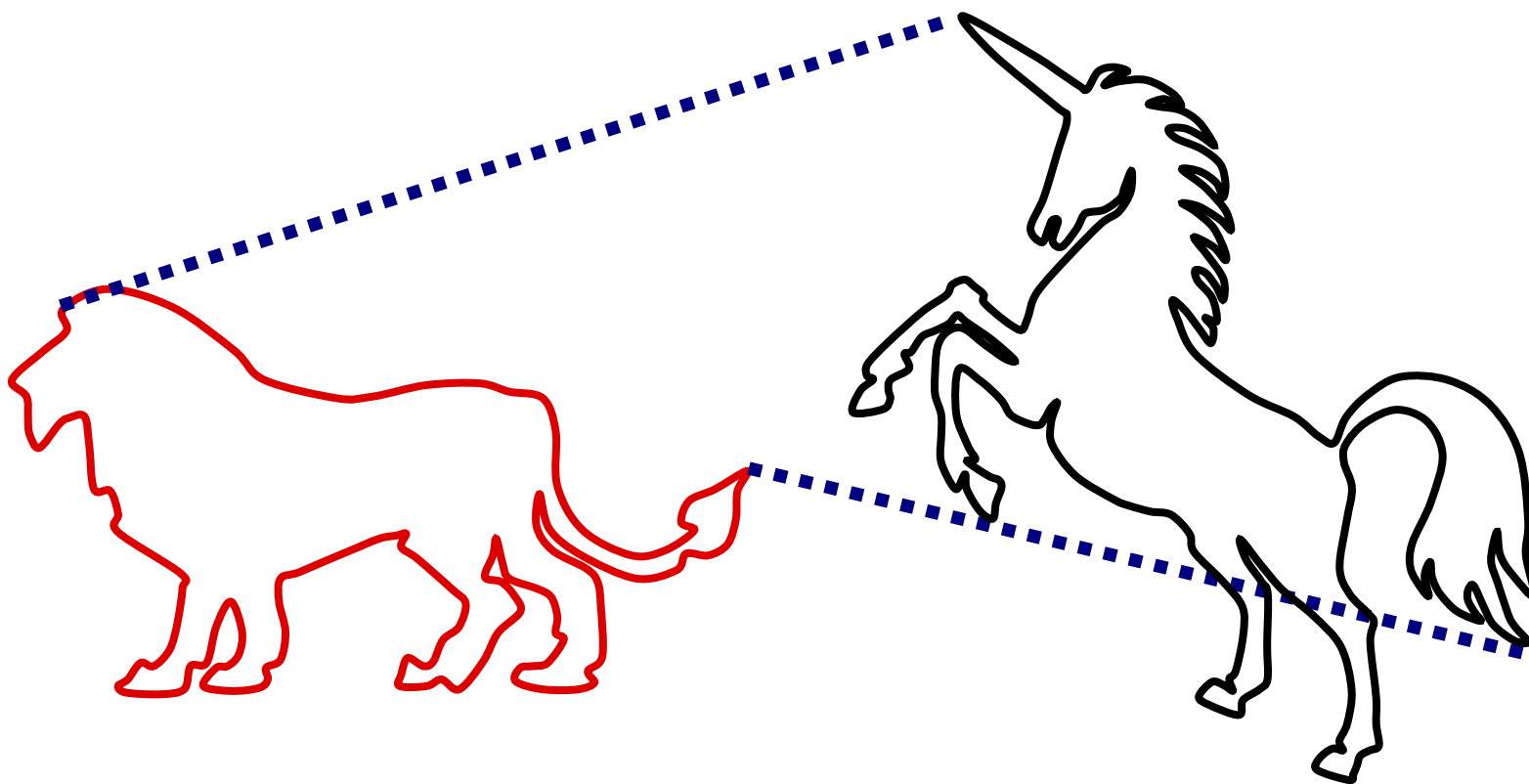
$$E(S) = Kmn e^{-\lambda S}$$

- e-value: $E(S)$ - ожидаемое количество случайных выравниваний с заданным весом
- p-value: $p(x > S) = 1 - e^{-E(S)}$
Вероятность встретить случайное выравнивание с таким или большим весом

Зачем делать структурное выравнивание?

- Структурное сходство позволяет выявить удаленных гомологов, претерпевших значительные изменения в процессе специализации и эволюции от общего предка
- Консервативные структурные мотивы могут иметь схожую функцию у гомологичных белков и позволяют изучать структурно-функциональные взаимосвязи

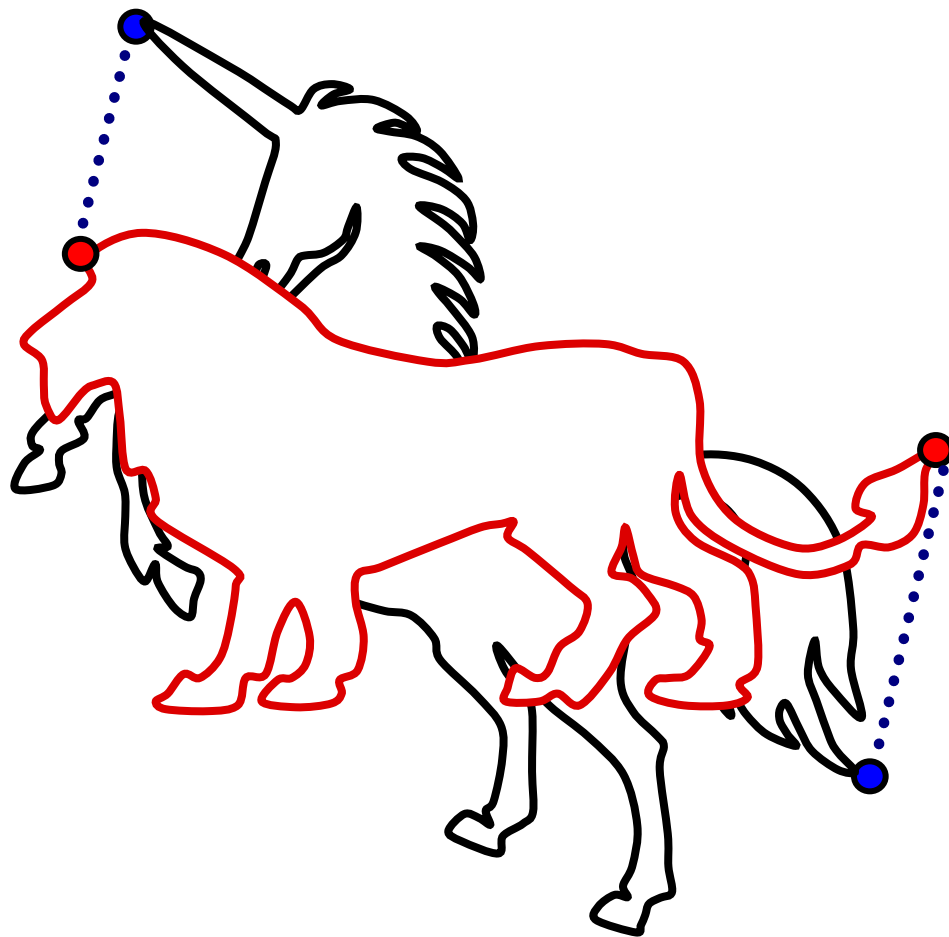
Как наилучшим способом превратить льва на единорога?



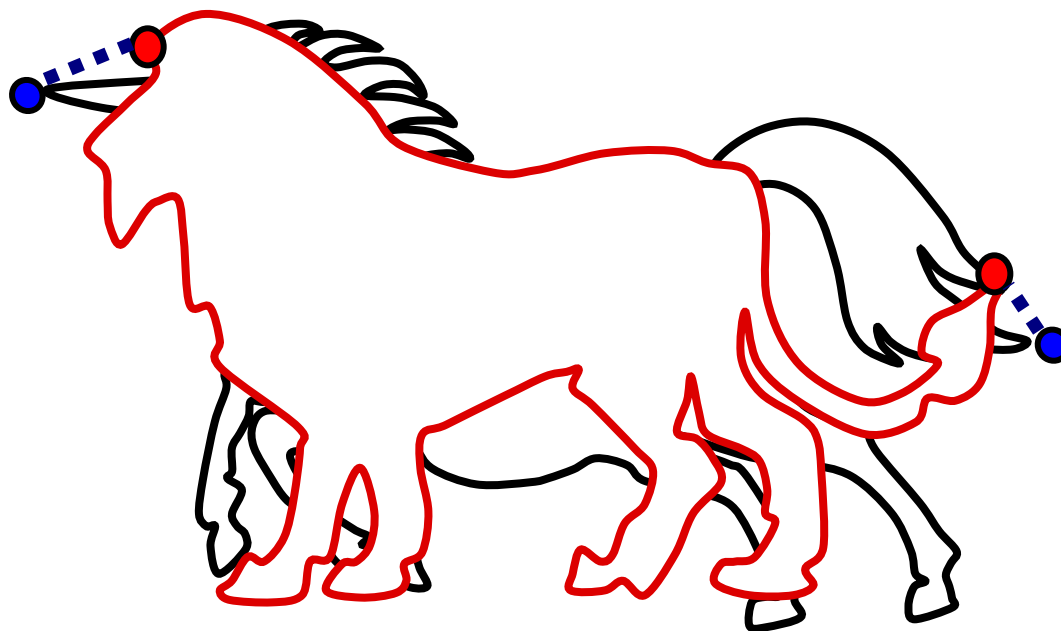
Допустимые действия

- Повороты
- Трансляции

Используем трансляцию



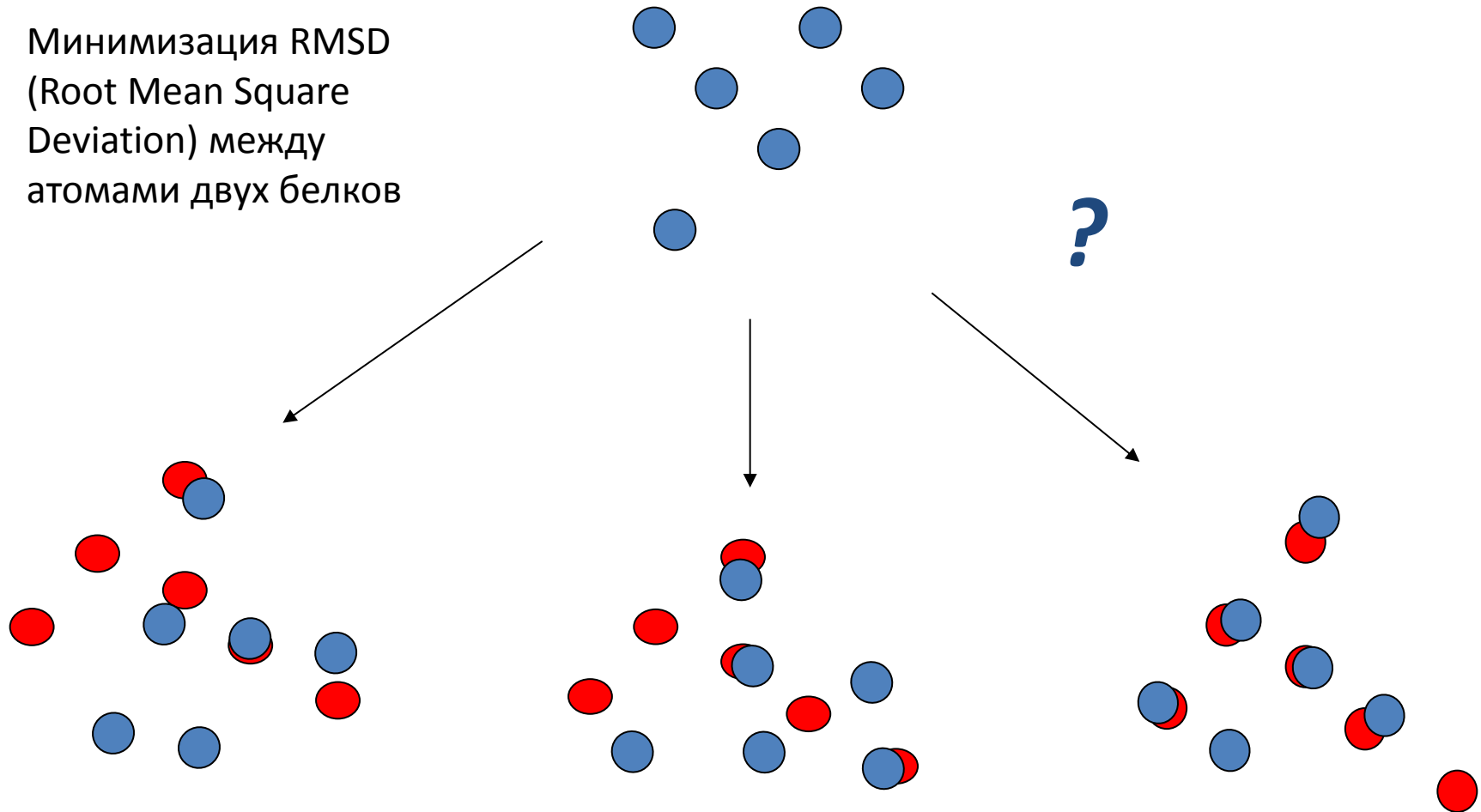
Используем трансляцию + поворот



Задача парного структурного выравнивания

Найти путь оптимальной трансформации между двумя структурами белков в пространстве

Минимизация RMSD
(Root Mean Square
Deviation) между
атомами двух белков



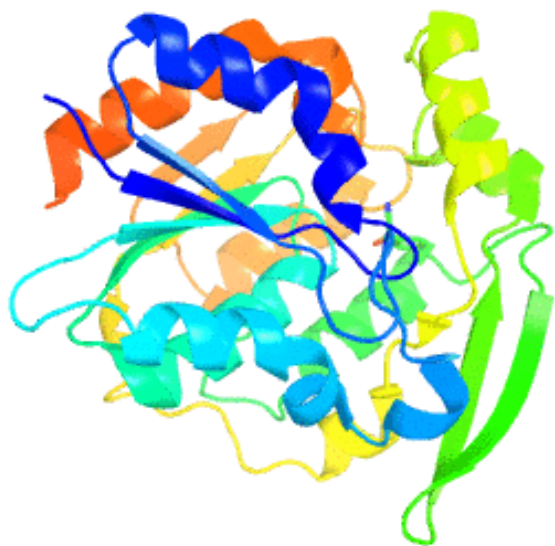
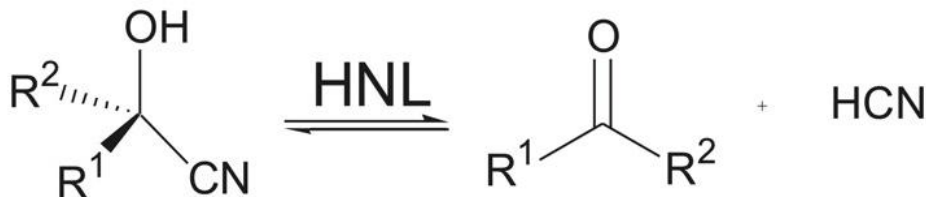
Что такое молекулярное моделирование?

Класс методов, основанных на теоретической химии и физике, для изучения механизмов действия белков/нуклеиновых кислот на молекулярном уровне, в т.ч.:

- Молекулярный докинг (изучение связывания лигандов в сайтах на поверхности структур белков);
- Молекулярная динамика (изучение конформационной variability и структурной подвижности);
- Квантово-химические методы (моделирование химических реакций).

От методов – к задачам

Функциональное сходство в рамках различной структурной организации

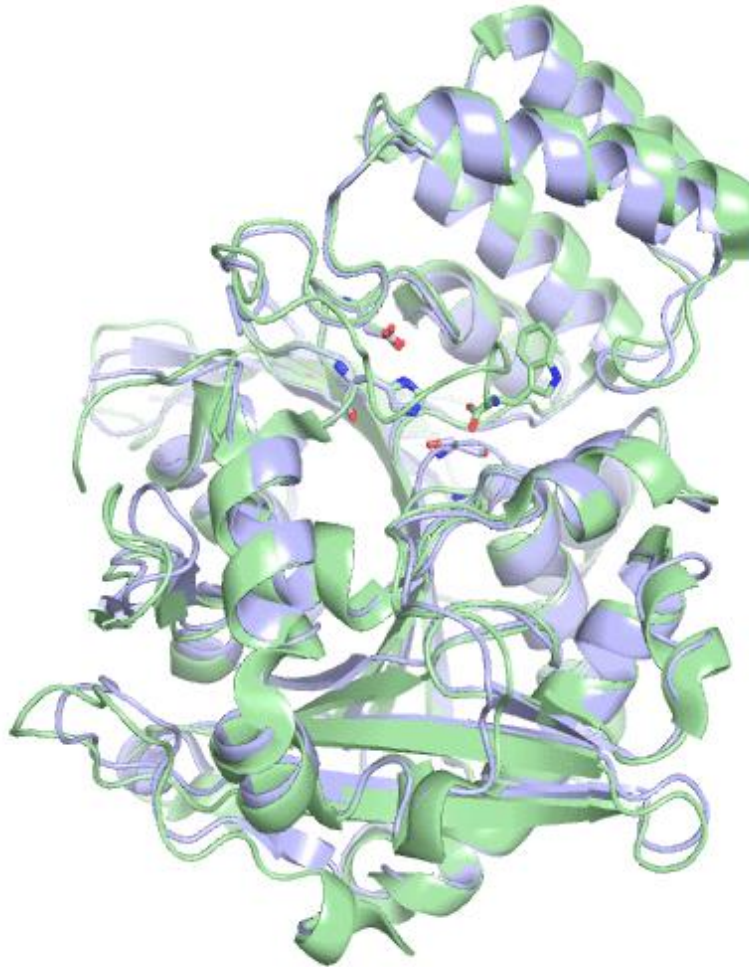


Hydroxynitrile lyase from
H.brasiliensis (4.1.2.47)



Hydroxynitrile lyase from
S.bicolor (4.1.2.11)

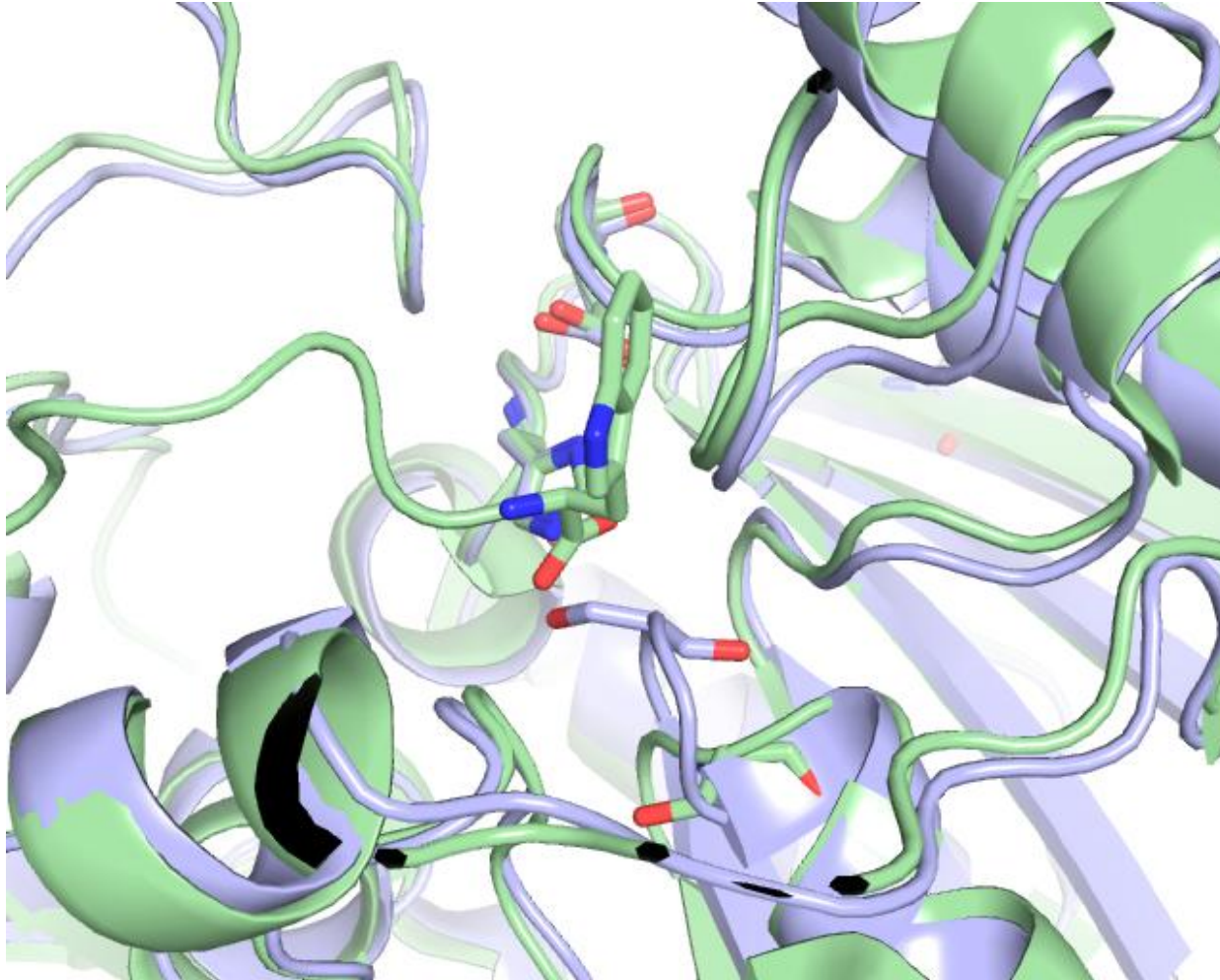
Функциональное разнообразие в рамках общей структурной организации



Wheat serine ***carboxypeptidase*** from
Triticum aestivum (3.4.16.6)

Hydroxynitrile lyase from
S.bicolor (4.1.2.11)

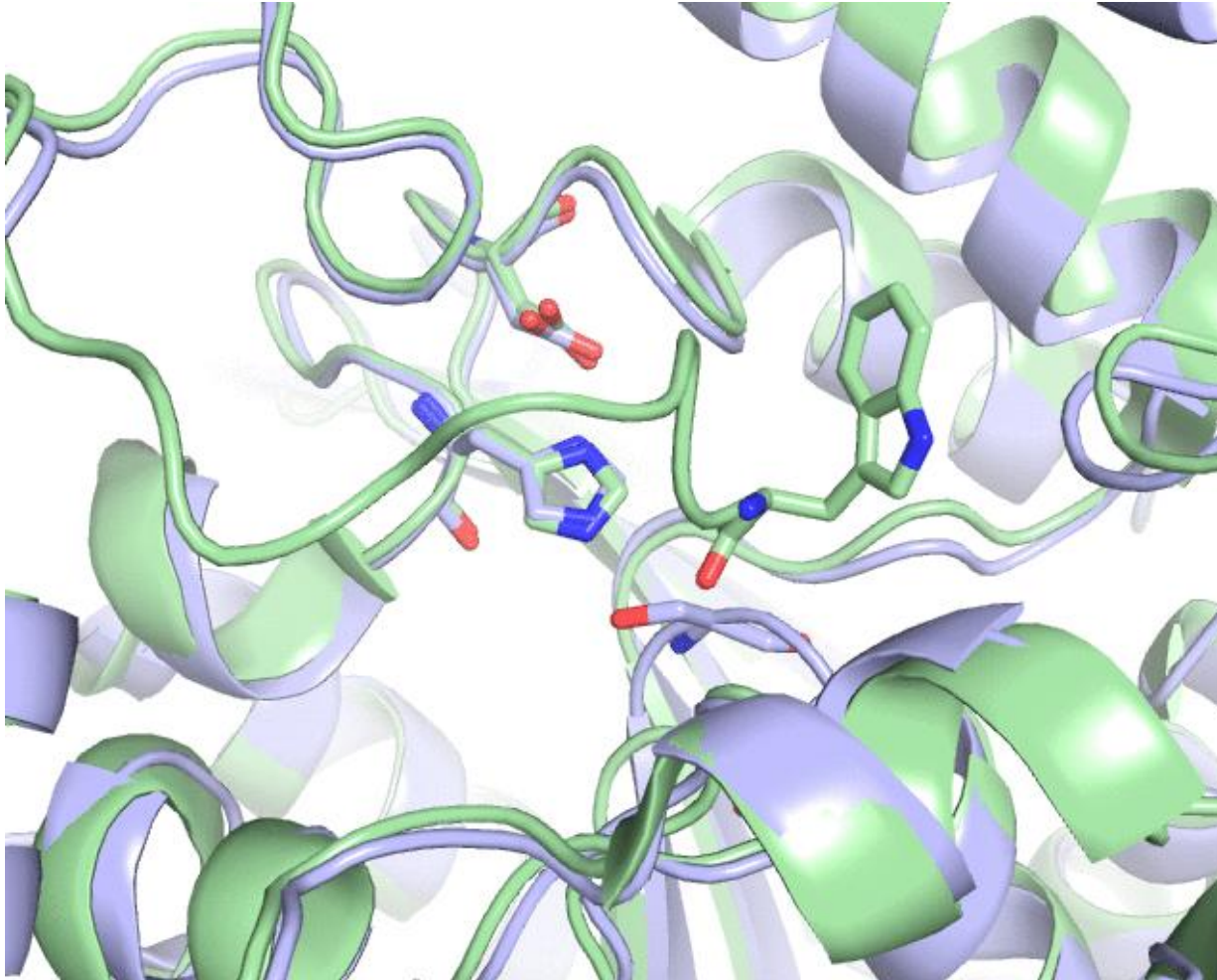
Функциональное разнообразие в рамках общей структурной организации



Wheat serine *carboxypeptidase* from
Triticum aestivum (3.4.16.6)

Hydroxynitrile lyase from
S.bicolor (4.1.2.11)

Функциональное разнообразие в рамках общей структурной организации



Wheat serine ***carboxypeptidase*** from
Triticum aestivum (3.4.16.6)

Hydroxynitrile lyase from
S.bicolor (4.1.2.11)

Функциональное разнообразие в рамках общей структурной организации

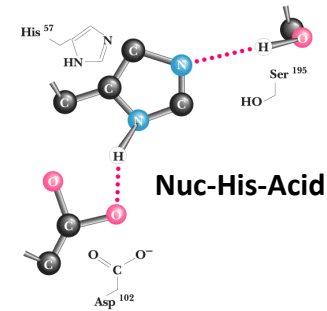
1. Схожесть структуры не означает тождественность функций – при общей укладке полипептидной цепи белки могут обладать принципиально разными свойствами;
2. Одна и та же функция может выполняться белками с различной структурой.

Суперсемейство – набор родственных
(гомологичных) белков – ключевое
понятие в биоинформатике

Структурно-функциональная классификация ферментов

Суперсемейство

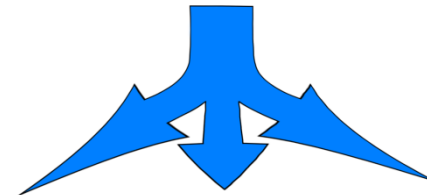
- **Общая структурная организация и сходство базовых функций** дают основание предполагать общее эволюционное происхождение
- Белки одного суперсемейства произошли от очень далекого общего предка и характеризуются **значительным функциональным разнообразием**



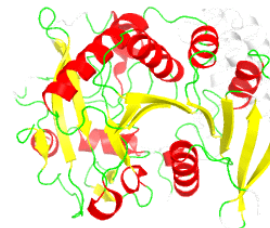
α/β -гидролазы

Семейство

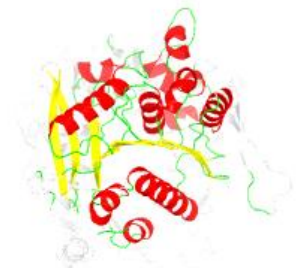
- **Эволюционно близкие белки** с высокой степенью сходства по аминокислотной последовательности, что подразумевает наличие близкого общего предка
- **Менее выраженное разнообразие свойств:** как правило обладают общим механизмом реакции, но могут превращать различные субстраты



Липазы



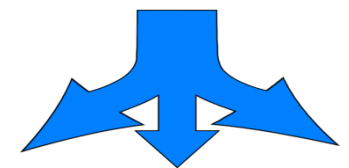
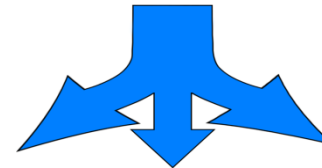
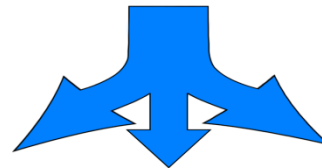
Эпоксид-гидролазы



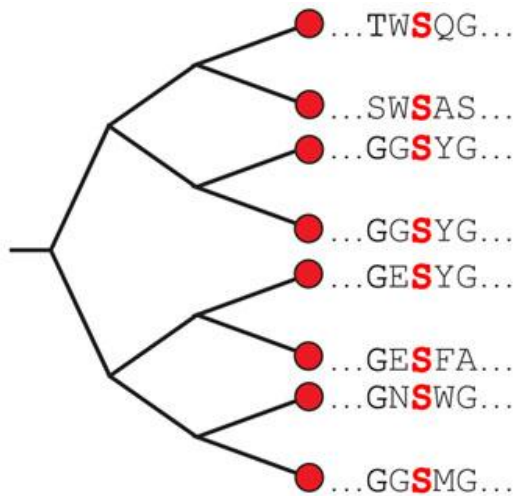
Пептидазы

Под-семейство

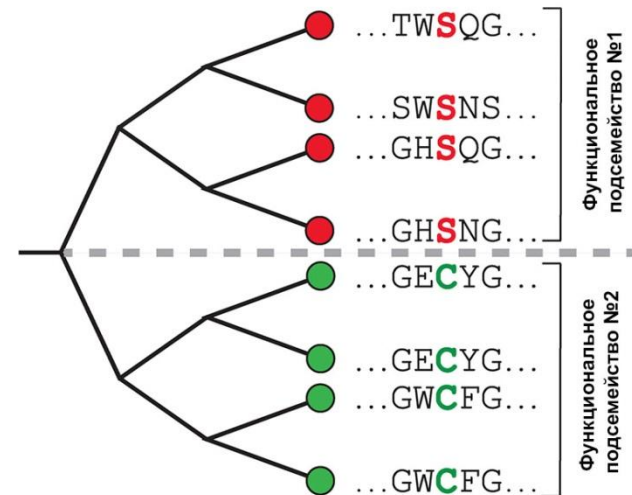
- **Еще более близкие белки**, характеризующиеся высокой степенью сходства по первичной и третичной структуре, и **гомогенностью функции**



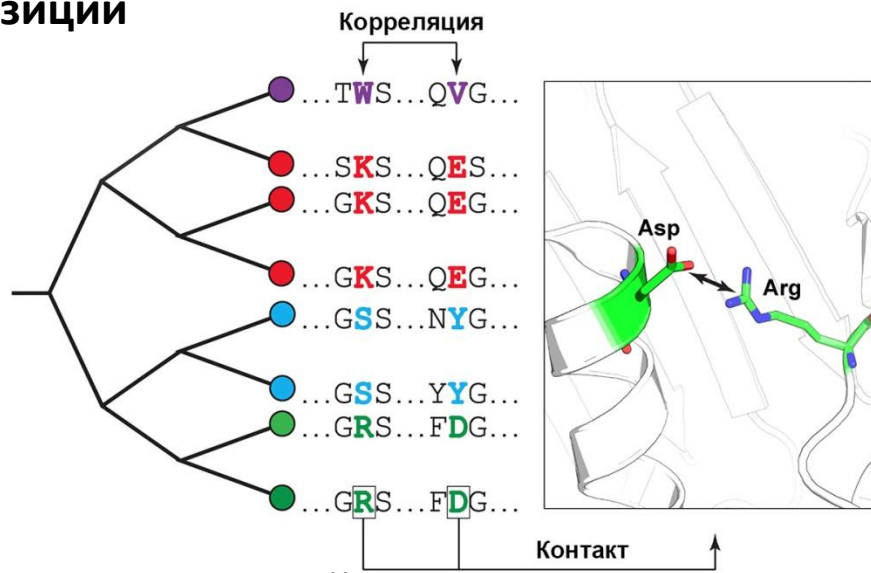
Систематический анализ выравнивания



**Консервативные
позиции**



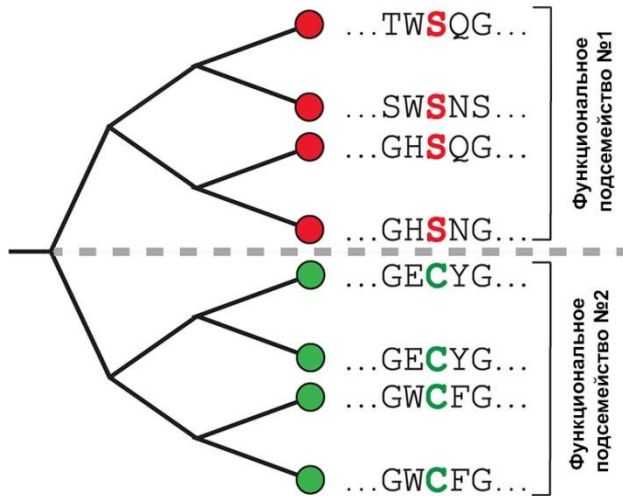
**Специфические позиции
подсемейств**



**Коррелирующие
(ко-эволюционирующие)
позиции**

Систематический анализ выравнивания

Специфические позиции



**Специфические позиции
подсемейств**

Как **интерпретировать** специфические позиции?

- Могут объяснить разнообразие свойств/функций в выборке (в суперсемействе)

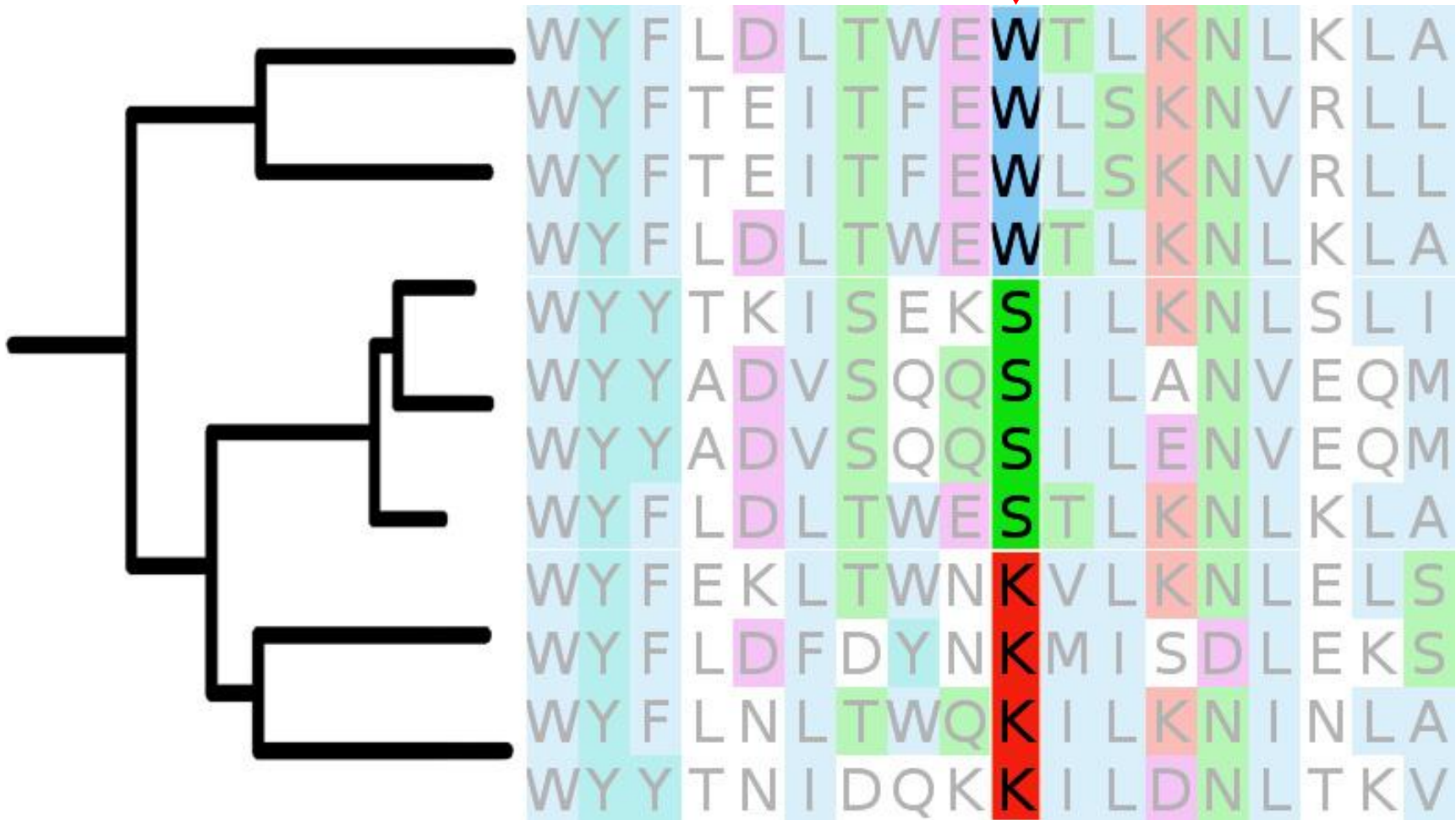
Что нужно для расчета специфических позиций:

- Выравнивание функционально разнообразных гомологов;
- Классификация на семейства/подсемейства;
- Функция оценки специфичности;
- Анализ статистической значимости.

Систематический анализ выравнивания

Специфические позиции

Специфическая позиция подсемейства (СПП)



Оценка специфичности позиции

G	G	G	G	G	G	G	G	G
G	G	G	G	G	G	G	G	G
G	G	G	G	G	G	G	G	G
G	G	G	G	G	G	G	G	G
G	G	G	G	G	G	G	G	G
G	G	G	G	G	G	G	G	G
G	G	G	G	G	G	G	G	G
M	M	M	M	M	M	M	M	M
M	M	M	M	M	M	M	M	M
M	M	M	M	M	M	M	M	M
M	M	M	M	M	M	M	M	M
M	M	M	M	M	M	M	M	M
M	M	M	M	M	M	M	M	M
M	M	M	M	M	M	M	M	M
M	M	M	M	M	M	M	M	M

$$S_i = \frac{\left[\sum_G \sum_{AB} M(AB) \times q_i(AB, G) \right] \times \left[\sum_G \sum_A q_i(A, G) \times \log \frac{q_i(A, G)}{q_i(A)} \right]}{n_G \times \sum_G \log \frac{N}{N_G}}$$

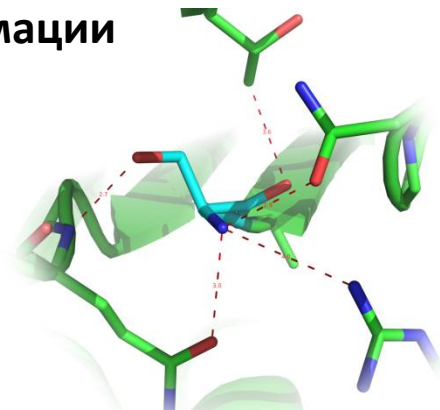
Физико-химическое сходство

3. Коррекция

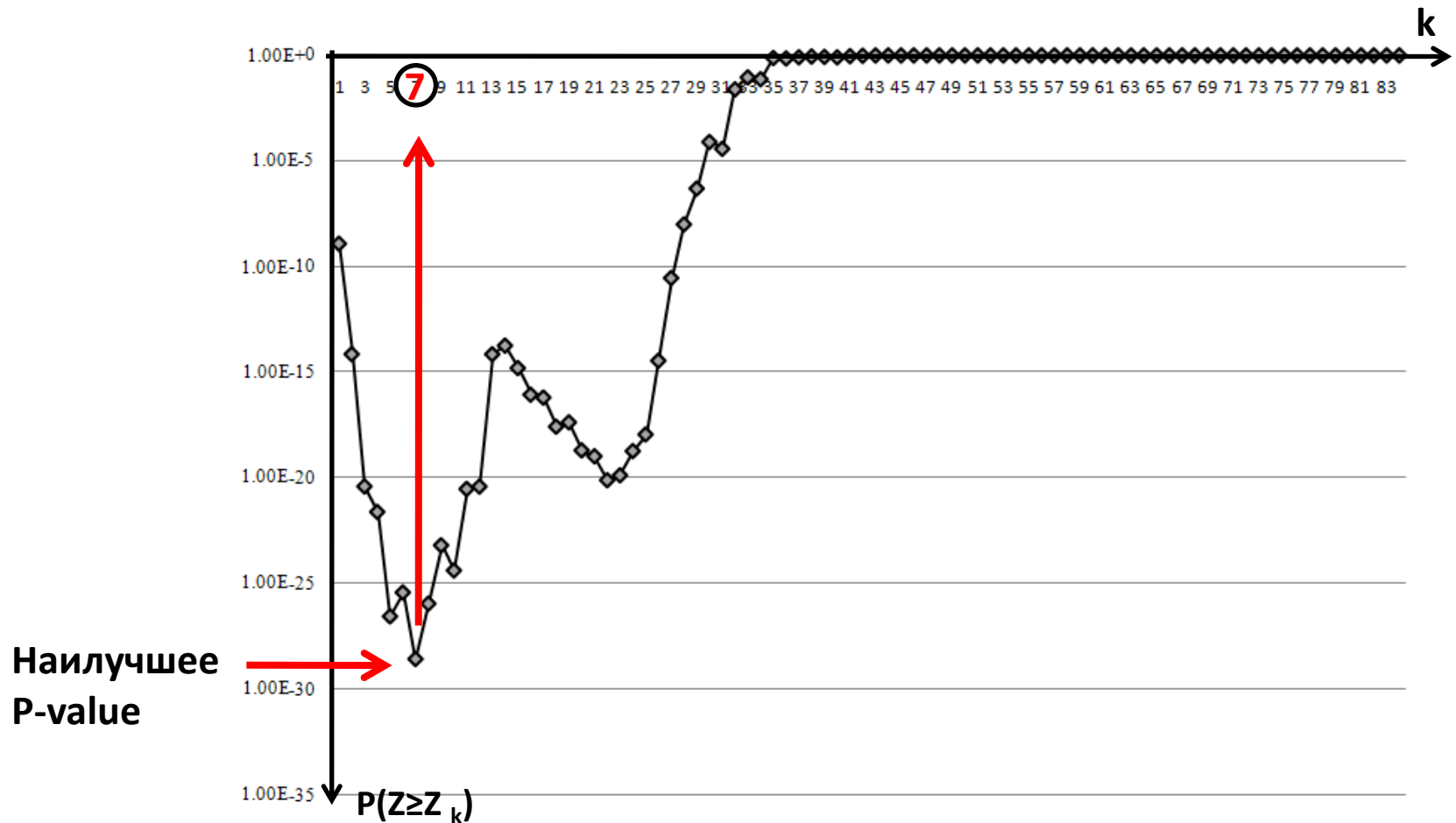
«филогенетической»
специфичности

$$Z_i = \frac{S_i - \langle S_i^{rnd} \rangle}{\sigma(S_i^{rnd})}$$

4. Использование структурной информации



Анализ статистической значимости



$$k = \arg_k \min P(\text{there are at least } k \text{ Z-scores } Z \geq Z_k) = \arg_k \min \left(1 - \sum_{i=n-k+1}^n c_n^i q^i p^{n-i} \right),$$

where n is the total number of Z-scores, and $p = P(Z \geq Z_k) = \int_{Z_k}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-Z^2) dZ$, $q = 1 - p$.

Систематический анализ выравнивания

Специфические позиции

- . G .
 - . G .
 - . G .
 - . G .
 - . G .
 - . G .
 - . M .
 - . M .
 - . M .
 - . M .
 - . M .
 - . M .
1. Информацию о переменных, эволюционно «гибких» **позициях**, которые отвечают за функциональное разнообразие между гомологами
 2. Предложения о **заменах**, необходимых для переключения функции между подсемействами ($G \leftrightarrow M$)
- Биоинформатический анализ дает нам **воспроизводимую и экспериментально проверяемую** гипотезу для последующего изучения

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями

Почему **липазы** являются плохими катализаторами гидролиза **амидов** несмотря на широкое распространение протеаз в суперсемействе α/β -гидролаз?



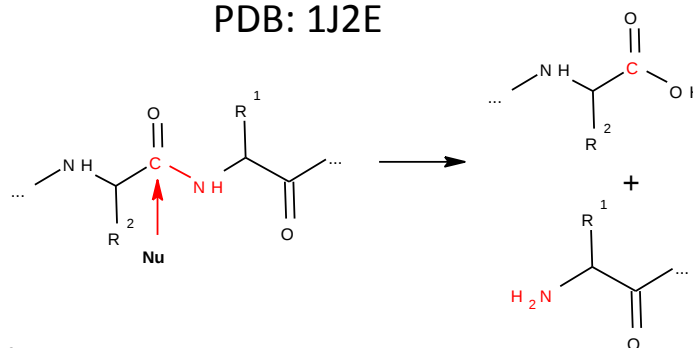
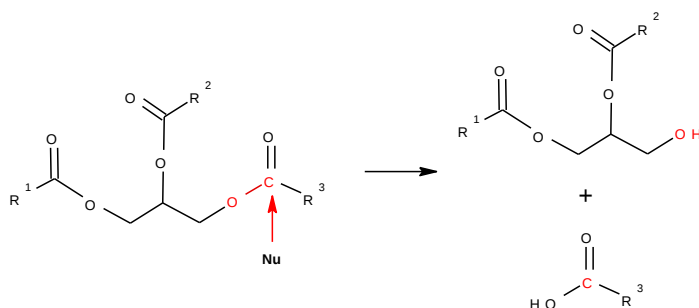
Candida antarctica

Липаза Б, PDB:1TCB



Дипептидил пептидаза IV

PDB: 1J2E



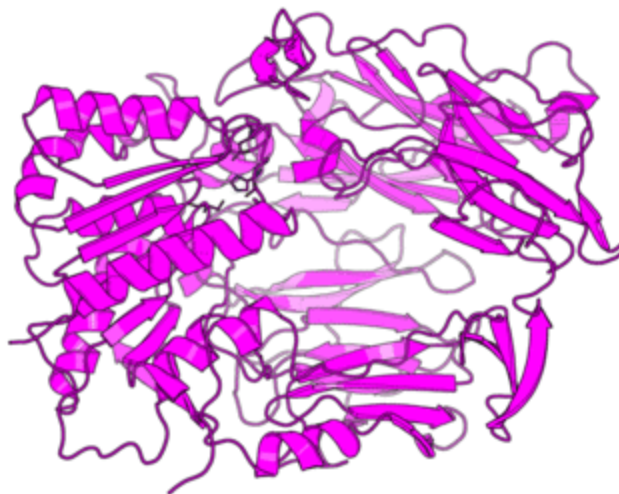
Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями

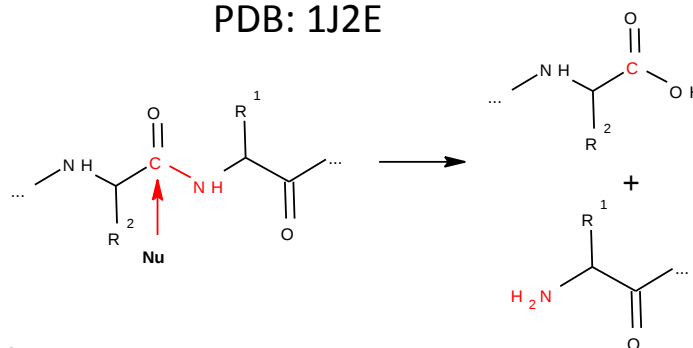
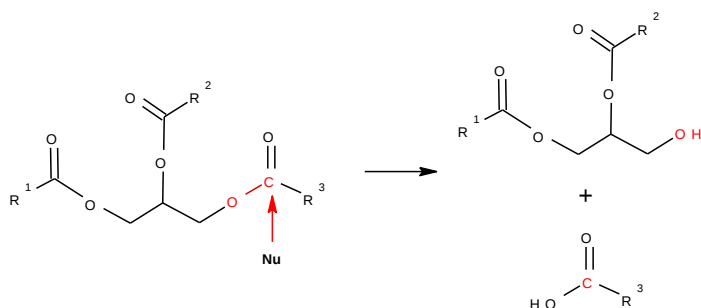
Почему липазы являются плохими катализаторами гидролиза амидов несмотря на широкое распространение протеаз в суперсемействе α/β -гидролаз?



Candida antarctica
Липаза Б, PDB:1TCB

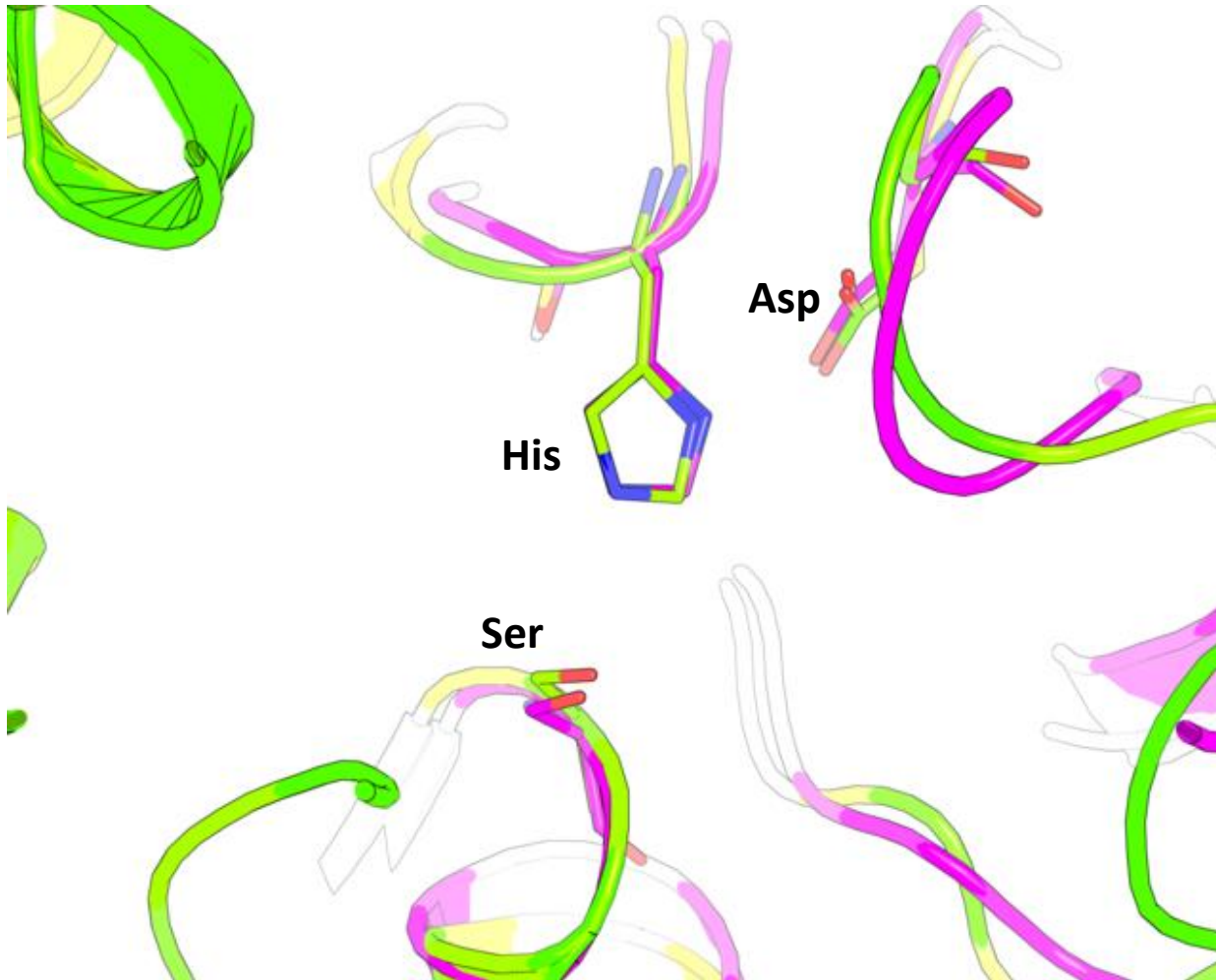


Дипептидил пептидаза IV
PDB: 1J2E



Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями

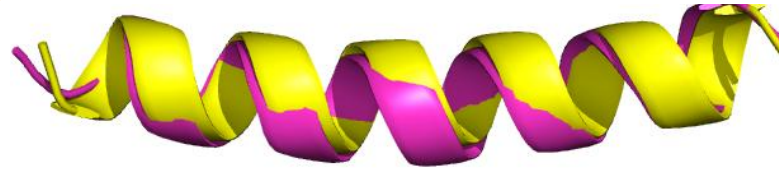


Каталитическая триада консервативна
как по последовательности, так и по структуре

Биоинформатический анализ информации о последовательностях, структурах и функциях α/β -гидролаз

- Структурное выравнивание липаз и пептидаз
 - Липаза Б из *Candida antarctica* (1TCB)
 - Сериновая карбоксипептидаза пшеницы (1WHS)
 - Дипептидил пептидазы человека 4 и 7 (1J2E и 3N0T)
 - Пролиновая иминопептидаза (1AZW)
- Структурно-опосредованное множественное выравнивание аминокислотных последовательностей

ABC72084.1	GPCSR - LGT FEVED - -WGWSYGG - Y - IAVAPVSRW
ABQ63099.1	GPCSR - LGTYEVED - -WGWSYGG - Y - IAVAPVSRW
YP_827462.1	G - FDA - KKQNVFND - -MGGSNGG - L - VAQVGVMMDM
YP_862986.1	GPGSE - LGKYEVED - -WGWSYGG - F - IAVAPVTSW
ZP_00949910.1	GPGSE - LGKYEVED - -WGWSYGG - F - IAVAPVTSW
ZP_01062505.1	G - FNM - QKQNVFDD - -RGSNGG - L - LPAVGVMMDM
ZP_01255422.1	GPGSE - LGKYELED - -WGWSYGG - F - IAVAPVTSW
ZP_01612295.1	G - FNM - QKQNVFDD - -RGSNGG - L - LPAVGVLDM
1j2e	GPCSR - LGT FEVED - -WGWSYGG - Y - IAVAPVSRW



1tcB	GTGT - NDTQVNT EYMV - TWSQGG - LV - -FAPDYK G
ZP_06264471.1	GTNF - GD I RVS AKAVS - GWSQGGG PL - -LAPSNHG
YP_003336117.1	GTFE - GD I AQSAGQLS - GHSQGG - MM - -I VPSNHG
ZP_05477113.1	GTTA - GPVAGTAGQLA - GHSQGG - MN - -LSPSNYG
ZP_04387428.1	GTDT - GD I VGSATAFG - GYSQGA - TV - -LASPTYG
ZP_06044588.1	GTTS - GD I QVSAEYVV - GHSQGG - MV - -LSSSNHG
ZP_03394192.1	GMNS - TPLDTALAE LT - GHSAGG - AL - -LAGVLHG
YP_119822.1	GTDS - EDLLASAAQIG - GFSQGA - TV - -LASPTYG

Использование специфических позиций подсемейства как «горячих точек замен»

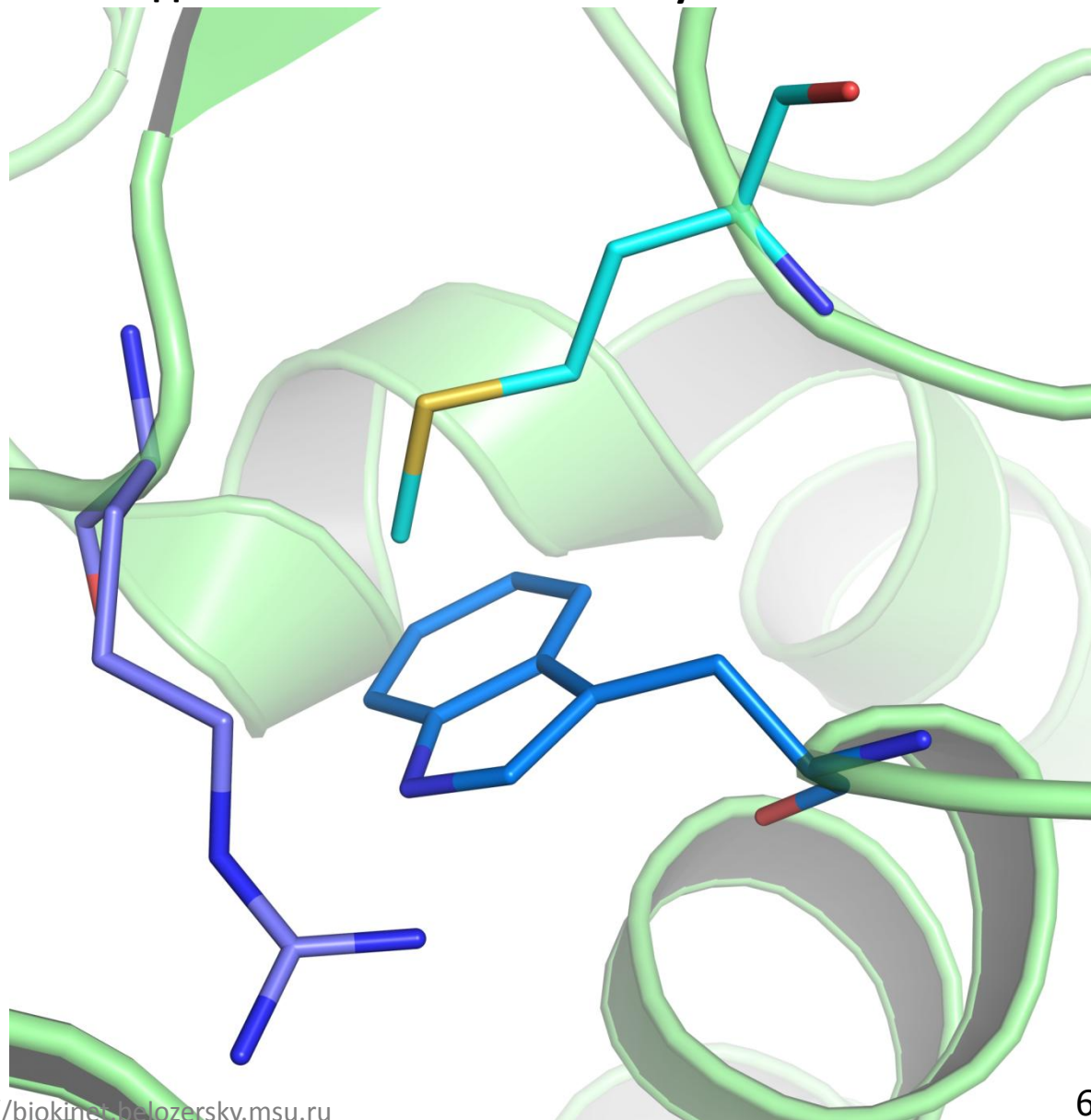
Для внесения амидазной активности в липазу

Липазы



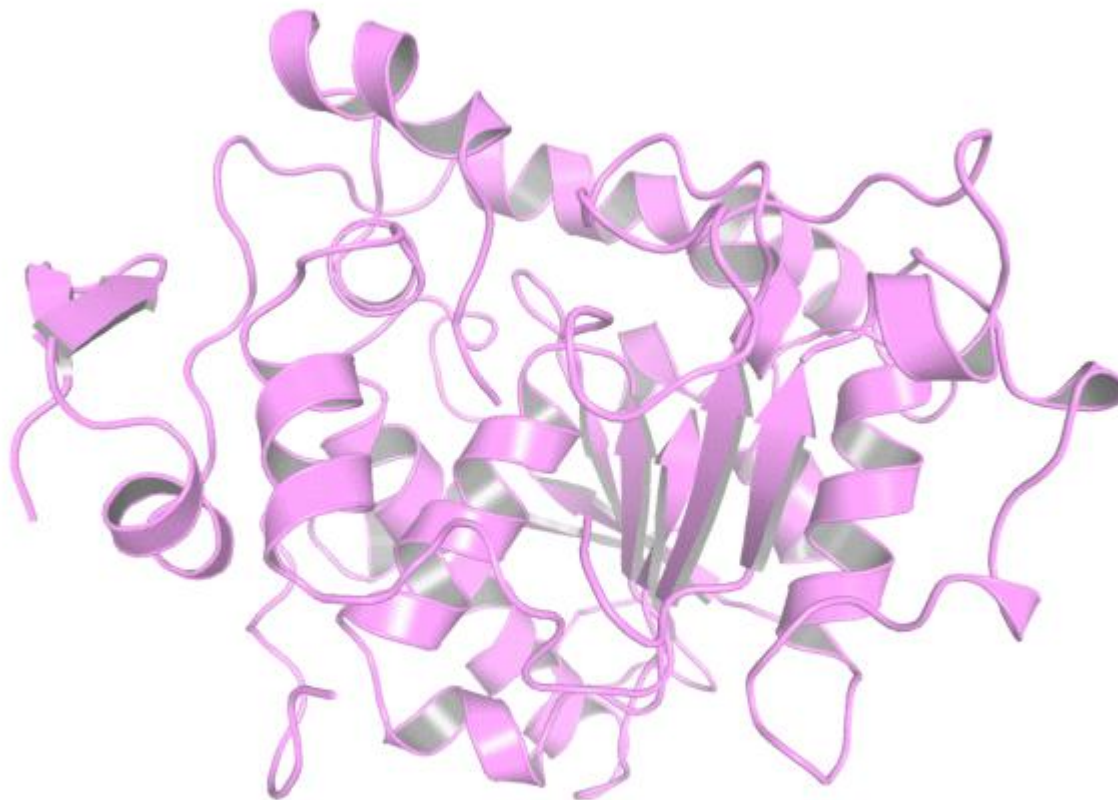
Амидазы

I	L	C
I	L	C
I	L	C
I	L	C
I	L	C
I	L	C
M	R	W
M	R	W
M	R	W
M	R	W
M	R	W
M	R	W



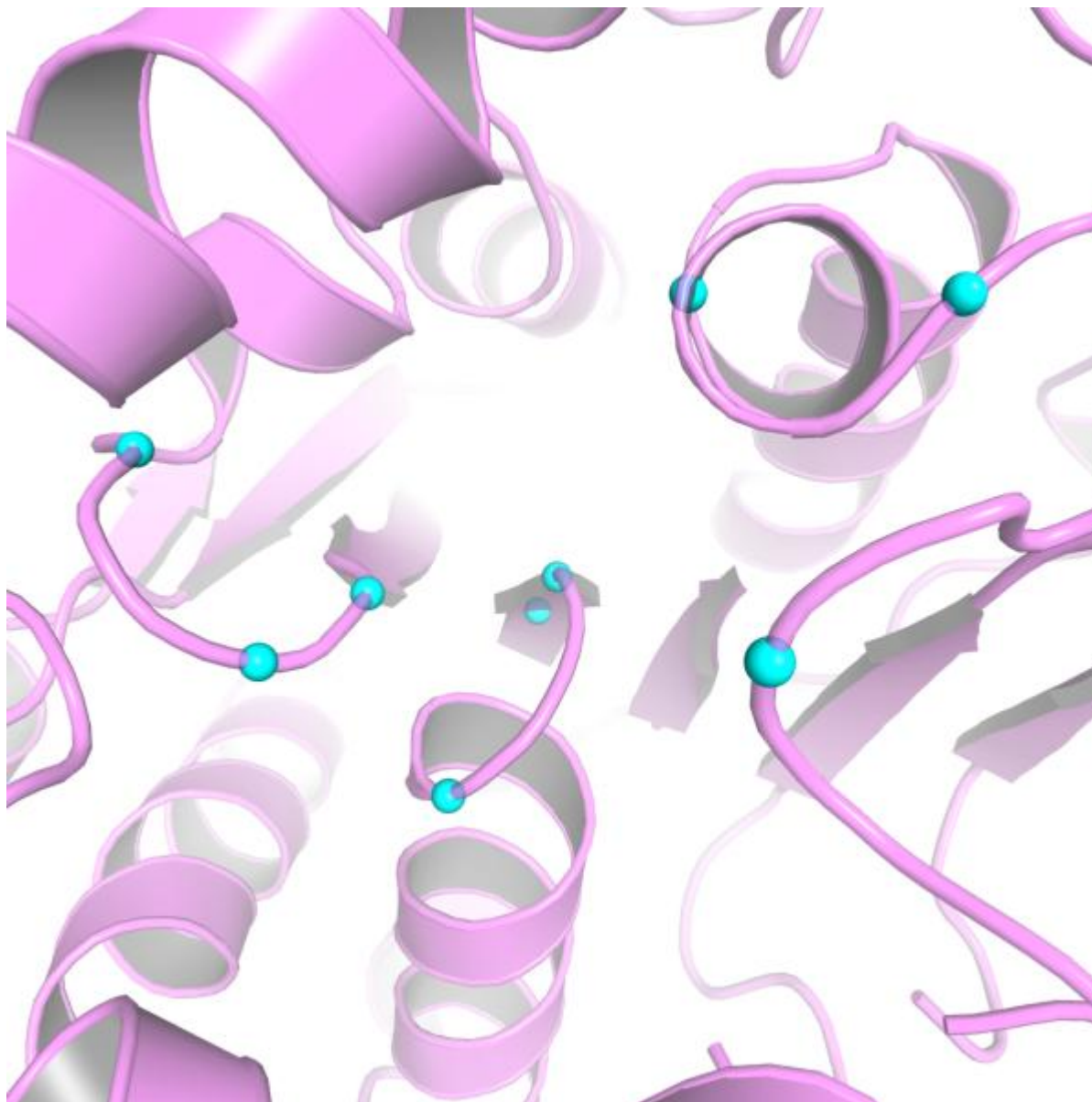
Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

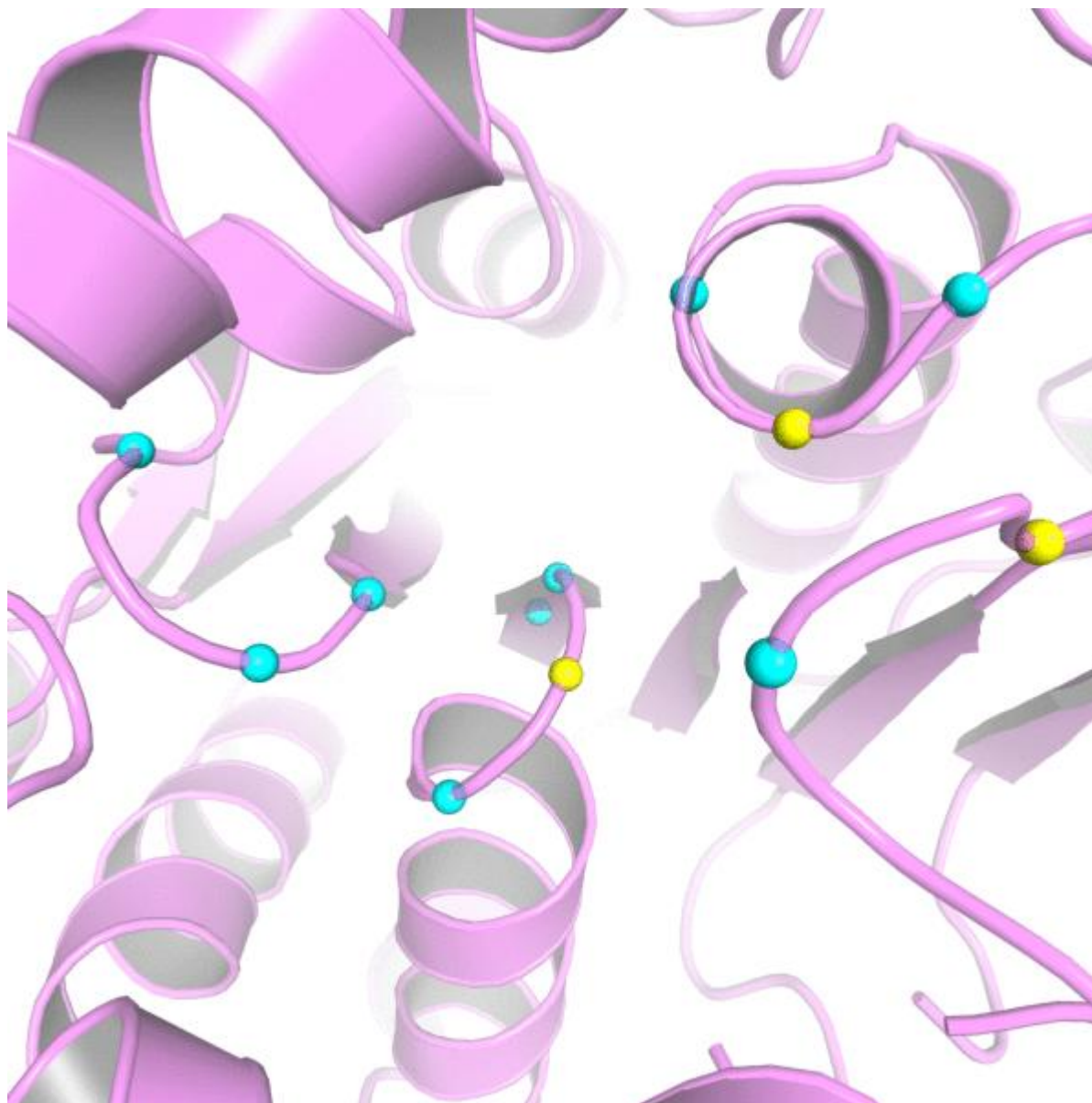
Asp223

Ala225

Leu278

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

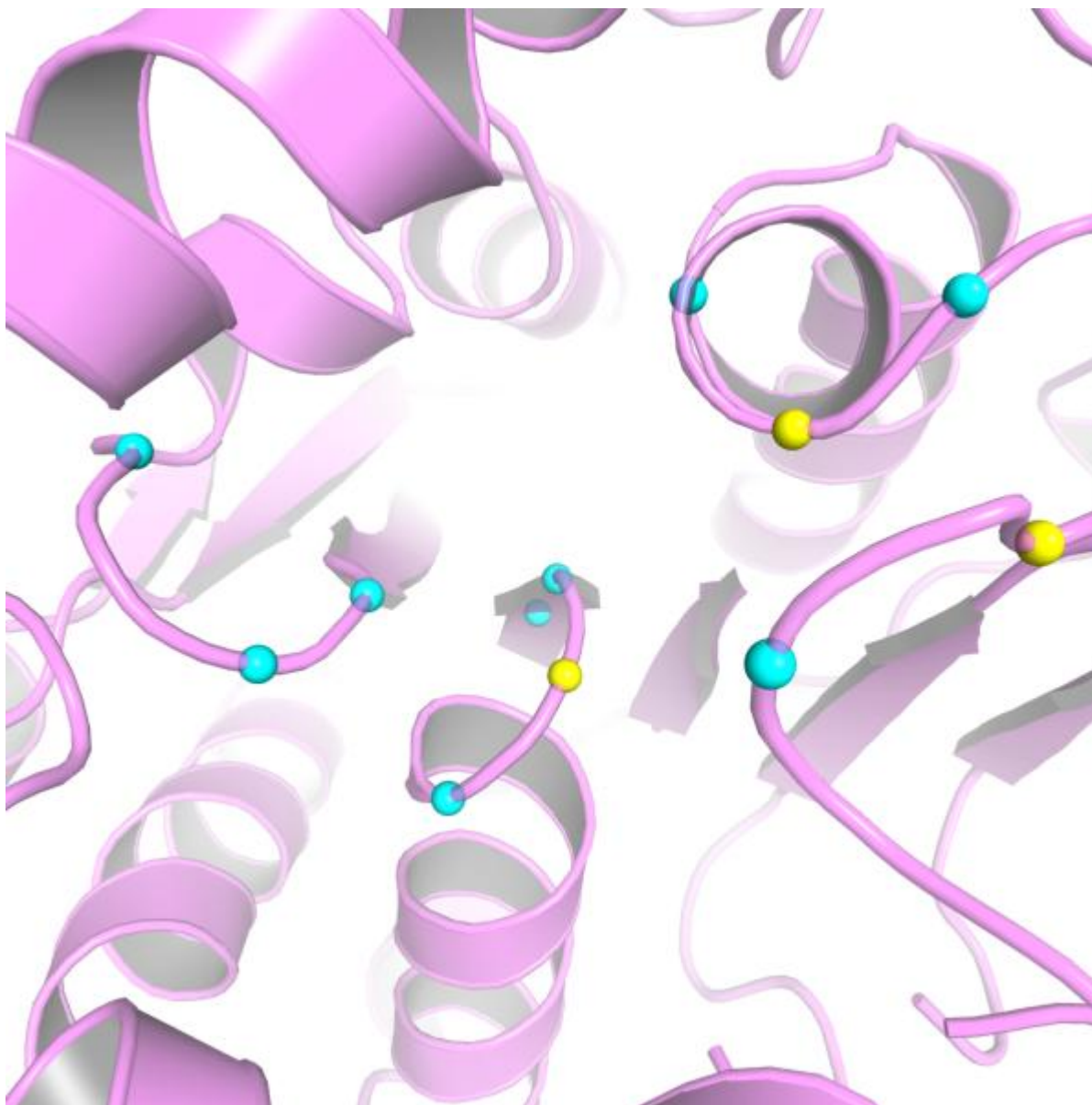
Asp223

Ala225

Leu278

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Catalytic
triad

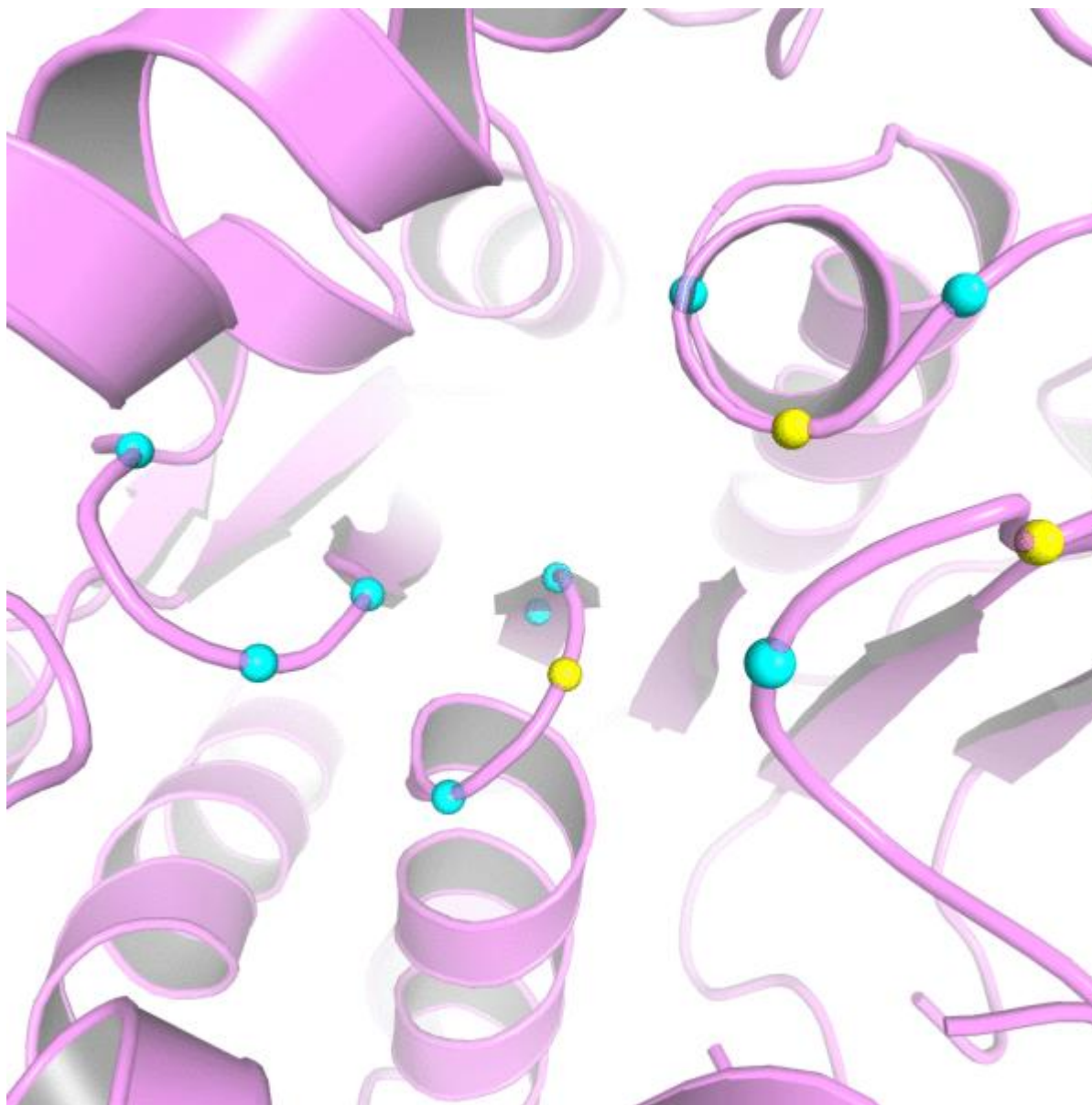
Ser105

His224

Asp187

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Catalytic
triad

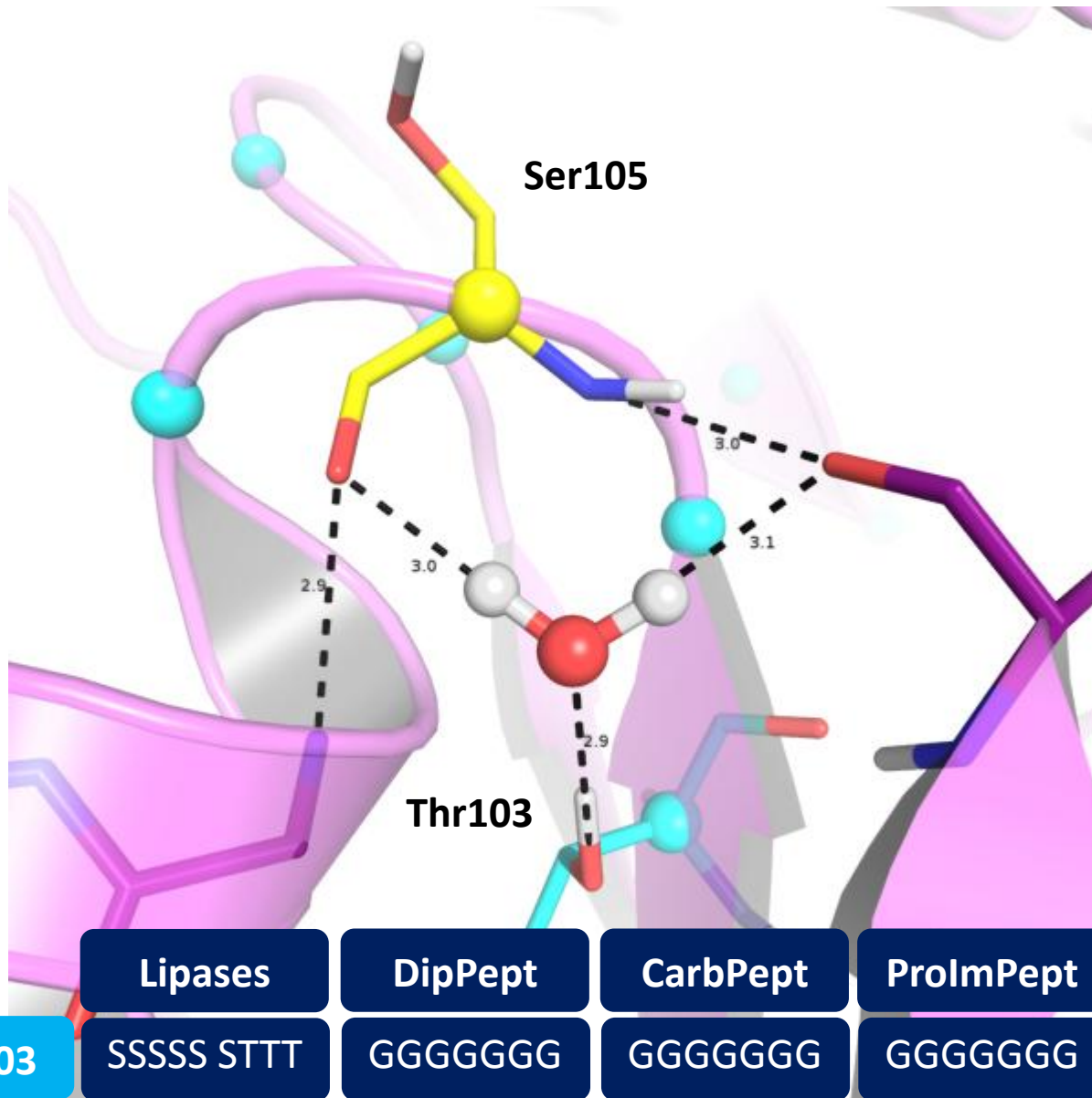
Ser105

His224

Asp187

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Catalytic
triad

Ser105

His224

Asp187

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями

Ser105

Гипотеза:
Thr103 → Gly

Стабилизирует
предреакционное положение
субстрата в активном центре

Thr103

SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Catalytic
triad

Ser105

His224

Asp187

Lipases

DipPept

CarbPept

ProImPept

T103

SSSSS STTT

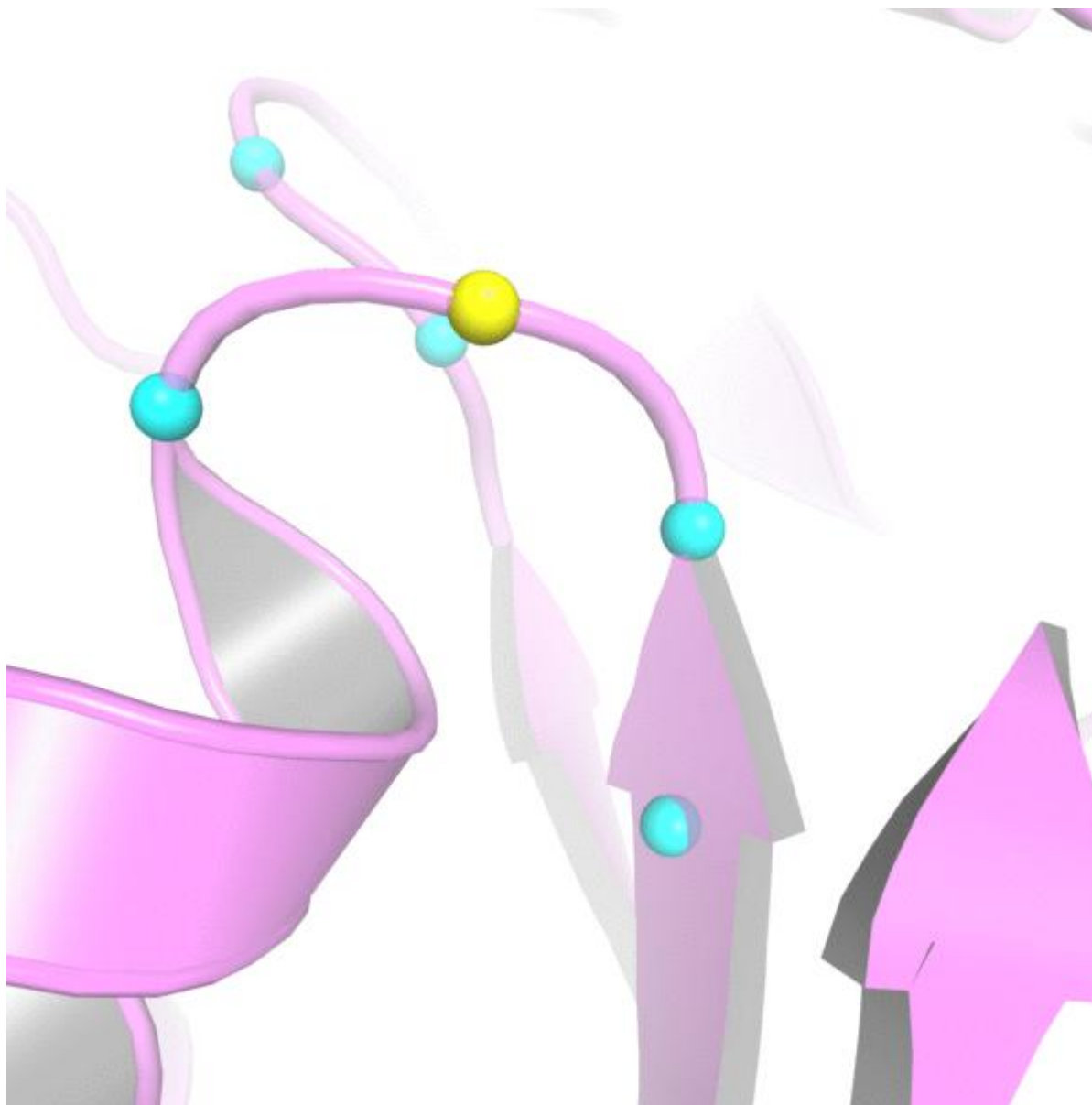
GGGGGGG

GGGGGGG

GGGGGGG

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Catalytic
triad

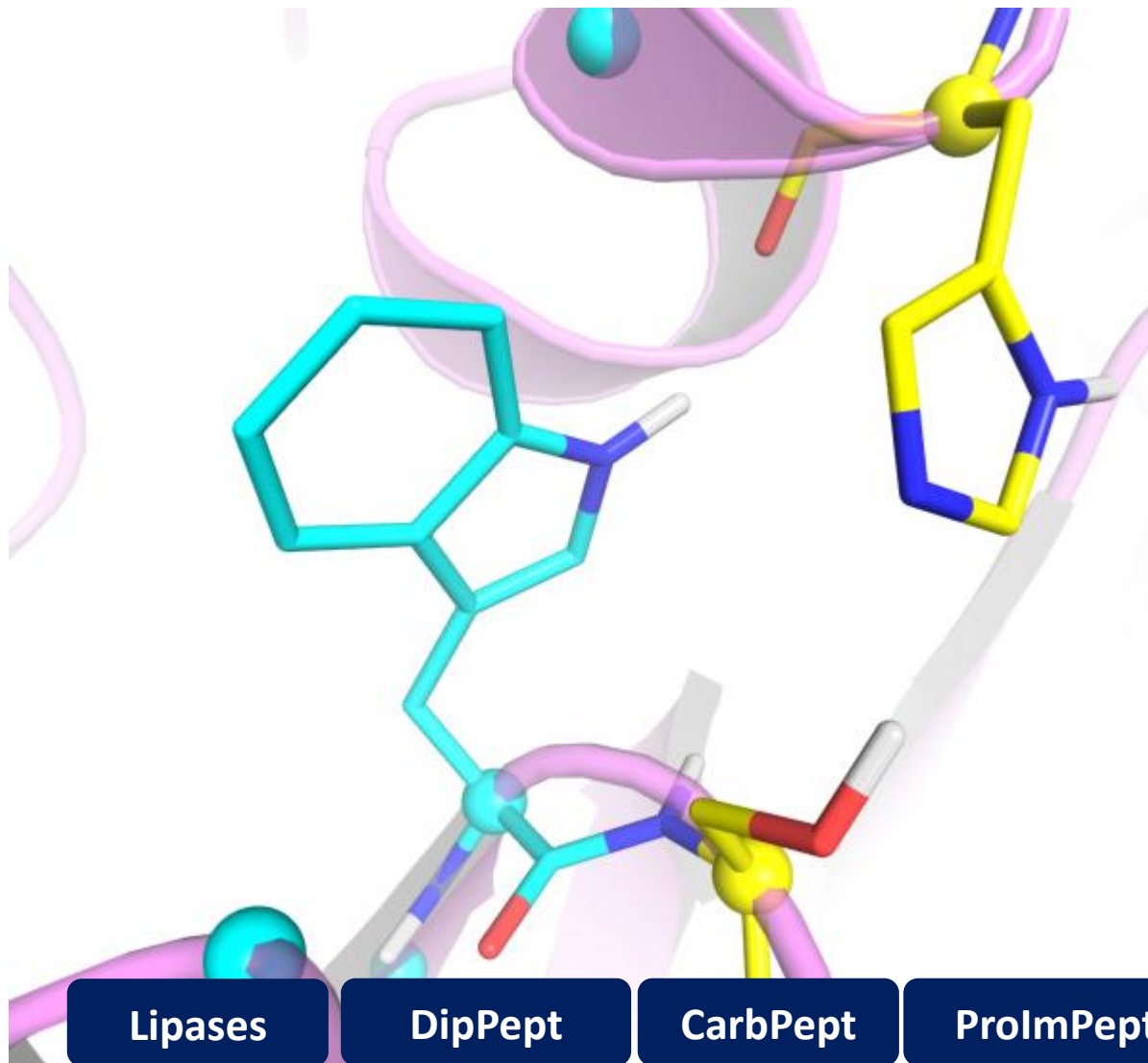
Ser105

His224

Asp187

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Catalytic
triad

Gly39

Ser105

Thr40

His224

Thr42

Asp187

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Lipases

DipPept

CarbPept

ProImPept

W104

WWWWWW

GGG

W*W*W*

EEEEEEEE

NNNNGGQQH

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями

Hypothesis:

Trp104 → [FYGHQNG]

(smaller and less hydrophobic residues)

Увеличение доступного объема в подсайте связывания уходящей группы

SSPs

Catalytic triad

Ser105

His224

Asp187

Ile189

Asp223

Ala225

Leu278

Lipases

DipPept

CarbPept

ProImPept

W104

WWWWWW

GGG

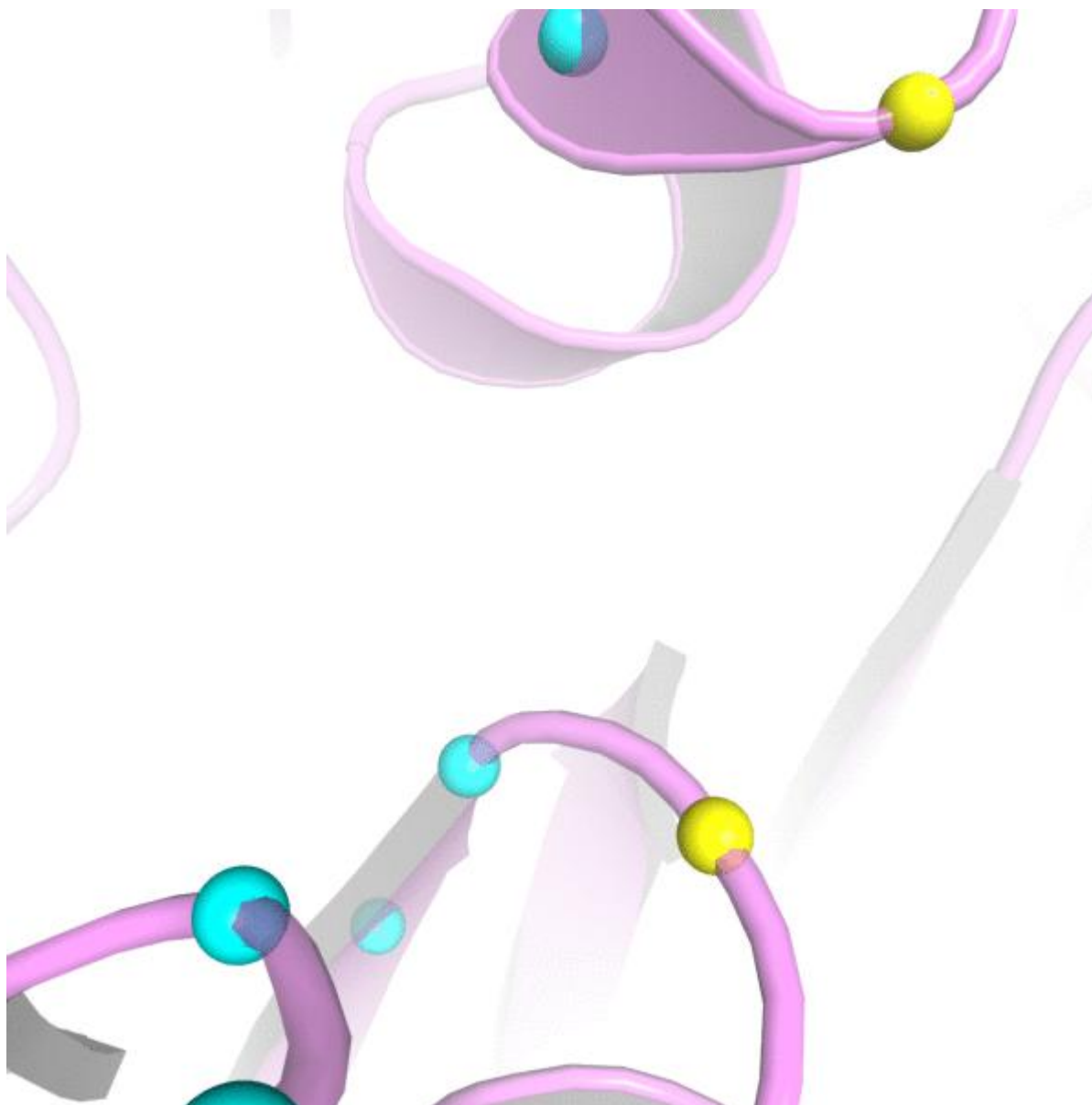
W*W*W*

EEEEEEEE

NNNNGGQQH

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Gly39

Thr40

Thr42

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Catalytic
triad

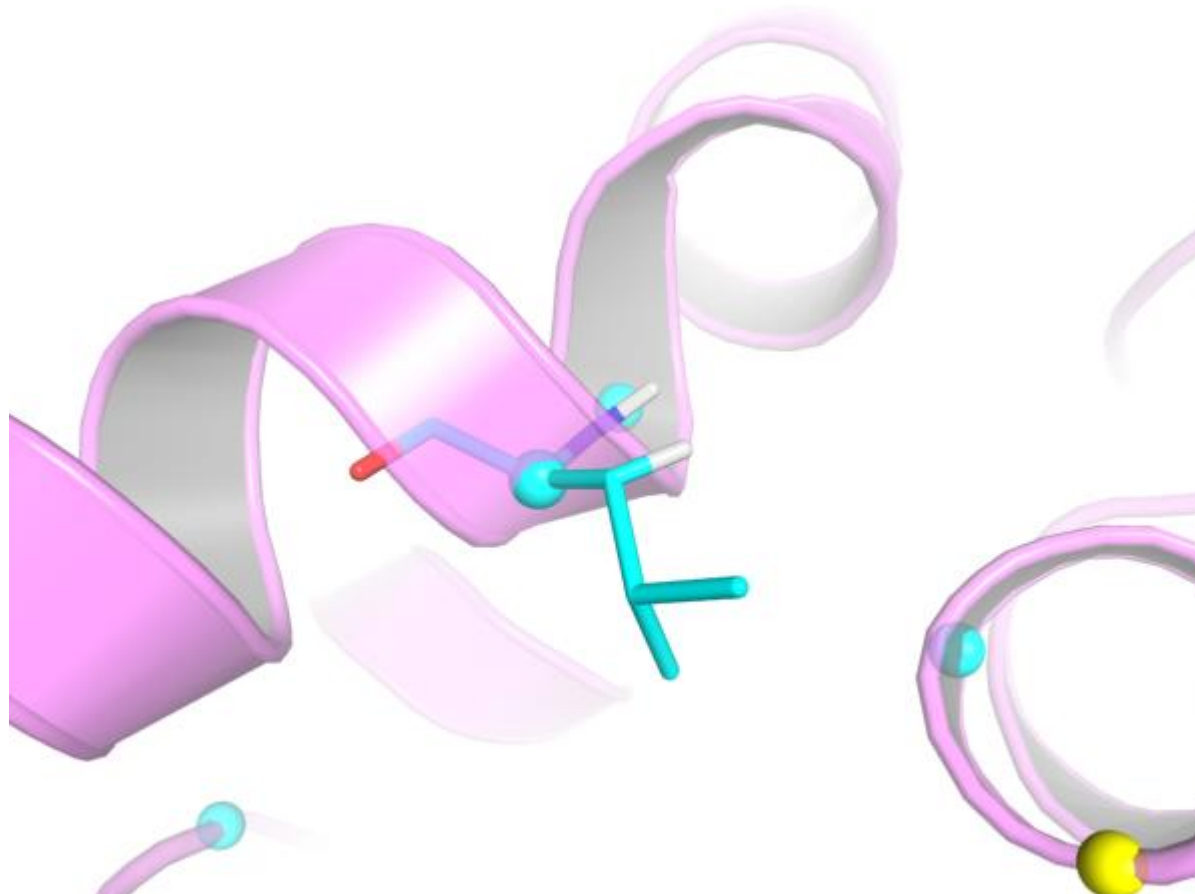
Ser105

His224

Asp187

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями



SSPs

Catalytic
triad

Gly39

Ser105

Thr40

His224

Thr42

Asp187

Thr103

Trp104

Gln106

Ile189

Asp223

Ala225

Leu278

Lipases

DipPept

CarbPept

ProlmPept

L278

LLLLLLLLIIIV

Биоинформатический анализ α/β -гидролаз

С липазной и амидазной активностями

Гипотеза:

Leu278 → [AG]

Симуляция делеции (больше места в центре связывания субстрата)

SSPs

Catalytic
triad

Ser105

His224

Asp187

Gln106

Ile189

Asp223

Ala225

Leu278

Lipases

DipPept

CarbPept

ProImPept

L278

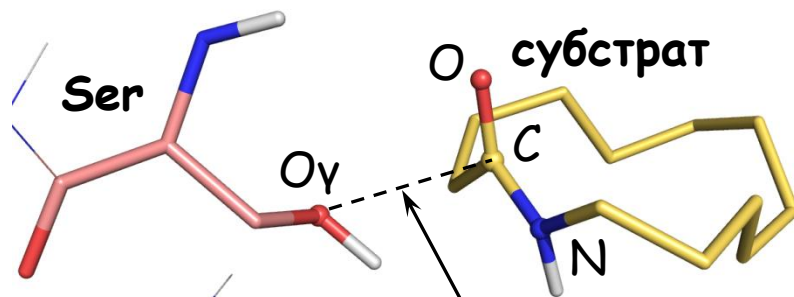
LLLLLLLLLIIV

Молекулярное моделирование фермент-субстратных комплексов

Использование геометрических критериев амидазной каталитической активности для отбора наиболее перспективных образцов фермента

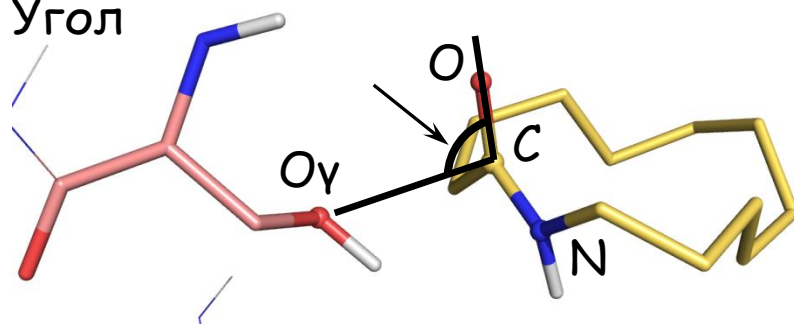
In silico библиотека:
 10^3 - 10^5 мутантов

1. Расстояние



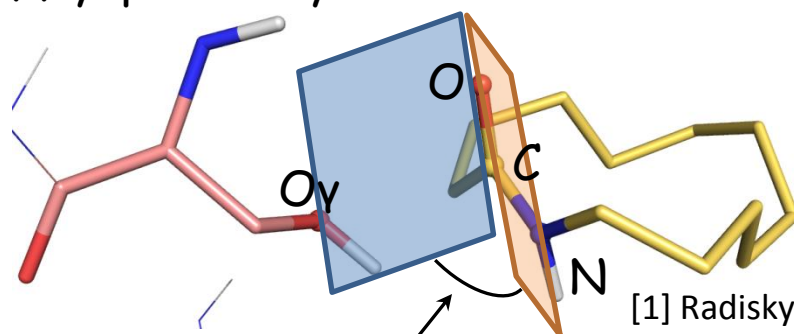
2.72 - 2.9 Å

2. Угол



89-92 °

3. Двугранный угол

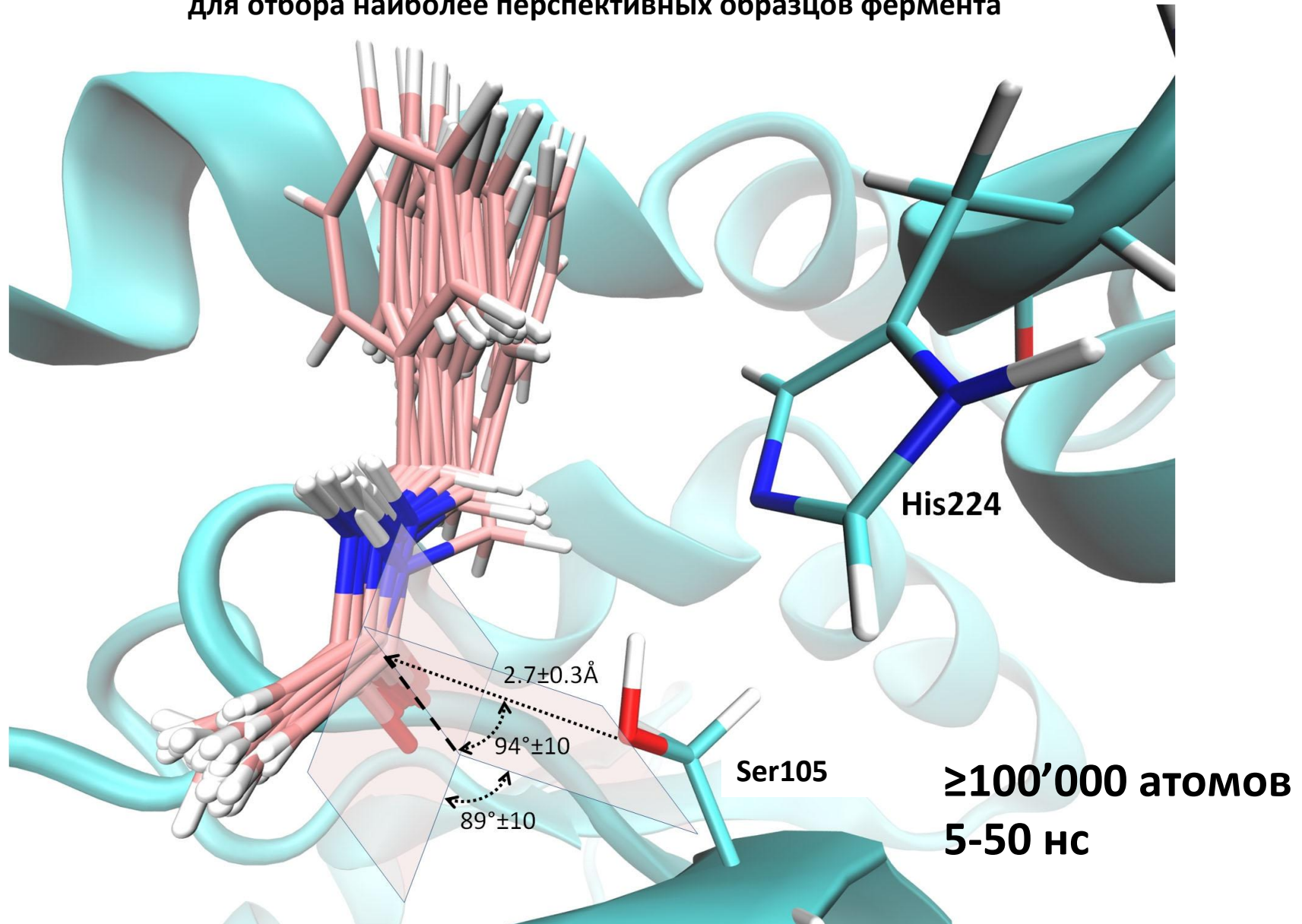


84-94 °

[1] Radisky E.S. & Koshland D.E. Jr., Proc Natl Acad Sci USA, 2002

Молекулярное моделирование фермент-субстратных комплексов

Использование геометрических критериев амидазной каталитической активности для отбора наиболее перспективных образцов фермента



Амидазная активность мутантов липазы Б

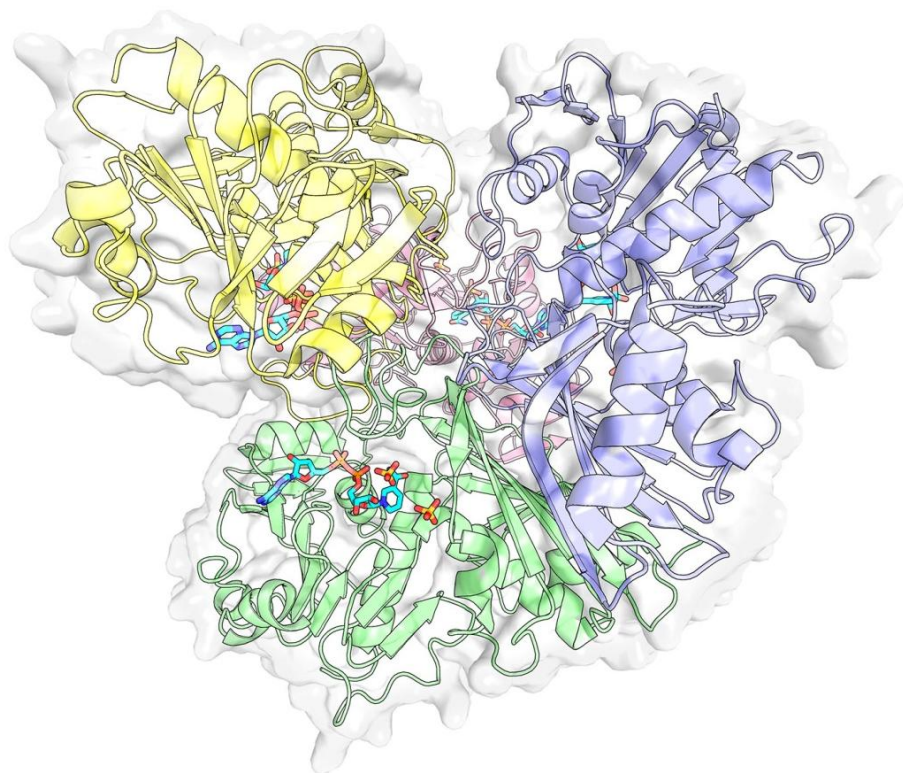
Измеренная *in silico* и экспериментальными методами

Мутант	Моделирование							Эксперимент
	Субстрат C(N)=O RMSD, Å	Расстояние, Å				Угол, °		
		Ser105 γO ... C	Thr40 N ... O	Thr40 γ1O ... O	Gln106 N ... O	A _{O...C=O}	D _{O...C(N)=O}	
G39A T103G W104F L278A	0.85±0.33	3.11±0.23	3.34±0.61	2.79± 0.22	3.79± 0.76	94.99±10.75	94.84±14.73	x 11.2
G39A T103G W104Y L278A	0.89±0.26	3.28±0.23	3.07±0.22	2.78± 0.20	3.57± 0.32	91.07±8.06	111.73±11.00	x 2.8
G39A T103G W104Q L278A	0.93±0.37	3.19±0.25	3.26±0.35	3.49±0.63	3.90±0.44	79.40±10.53	106.16±12.14	x 1.9
G39A W104F L278A	1.01±0.30	3.27±0.22	3.01±0.19	2.88± 0.24	3.46± 0.29	95.72±8.70	91.40±28.18	x 6.3
G39A W104F	1.19±0.49	3.26± 0.27	4.73±1.53	5.72± 1.00	5.18± 1.22	108.53± 27.84	91.76± 29.41	x 4.2
G39A T103G L278A	1.88±0.55	3.48±0.34	5.39±0.82	4.12± 0.70	6.39± 0.89	133.55±22.79	89.10±22.43	x 3.8
G39A L278A	1.91±0.44	3.34± 0.28	5.15± 0.64	3.91± 0.60	6.11± 0.66	126.06± 20.27	88.87± 13.79	x 3.3
G39A T103G	2.29±0.79	3.45±0.56	5.46±1.09	4.10± 0.92	6.19± 1.01	123.60±25.19	78.94±46.04	x 0.8
Фермент дикого типа	2.67±1.20	4.52±0.96	6.92±1.47	5.18±1.19	7.40±1.29	137.23±23.82	68.67±136.28	1.0
G39A W104F A225K L278A	3.13±0.96	5.34± 1.04	6.46± 1.03	4.79± 1.01	8.10± 1.36	148.14± 21.18	70.15±61.19	x 0.8

Suplatov, D. A., Besenmatter, W., Švedas, V. K., & Svendsen, A. (2012). **Bioinformatic analysis of alpha/beta-hydrolase fold enzymes reveals subfamily-specific positions responsible for discrimination of amidase and lipase activities.** *Protein Engineering Design and Selection*, 25(11), 689-697.

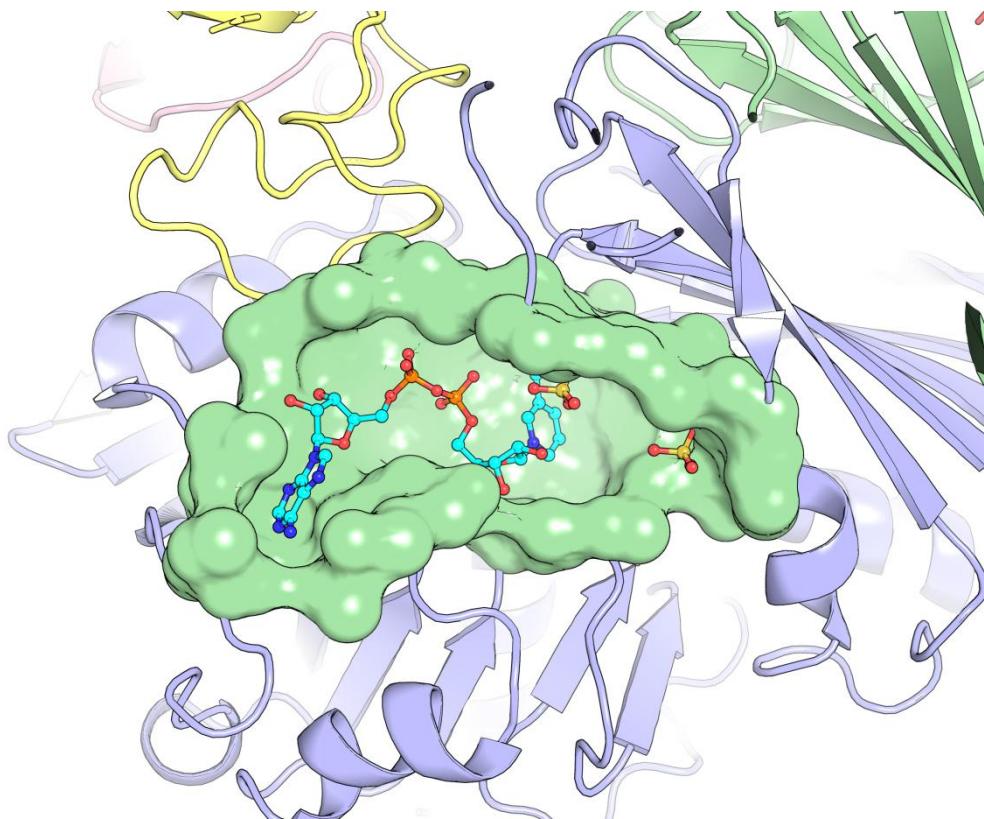
От задач – к решению

Поиск ингибиторов глицеральдегид-3-фосфат дегидрогеназы (ГАФД) с новым механизмом действия



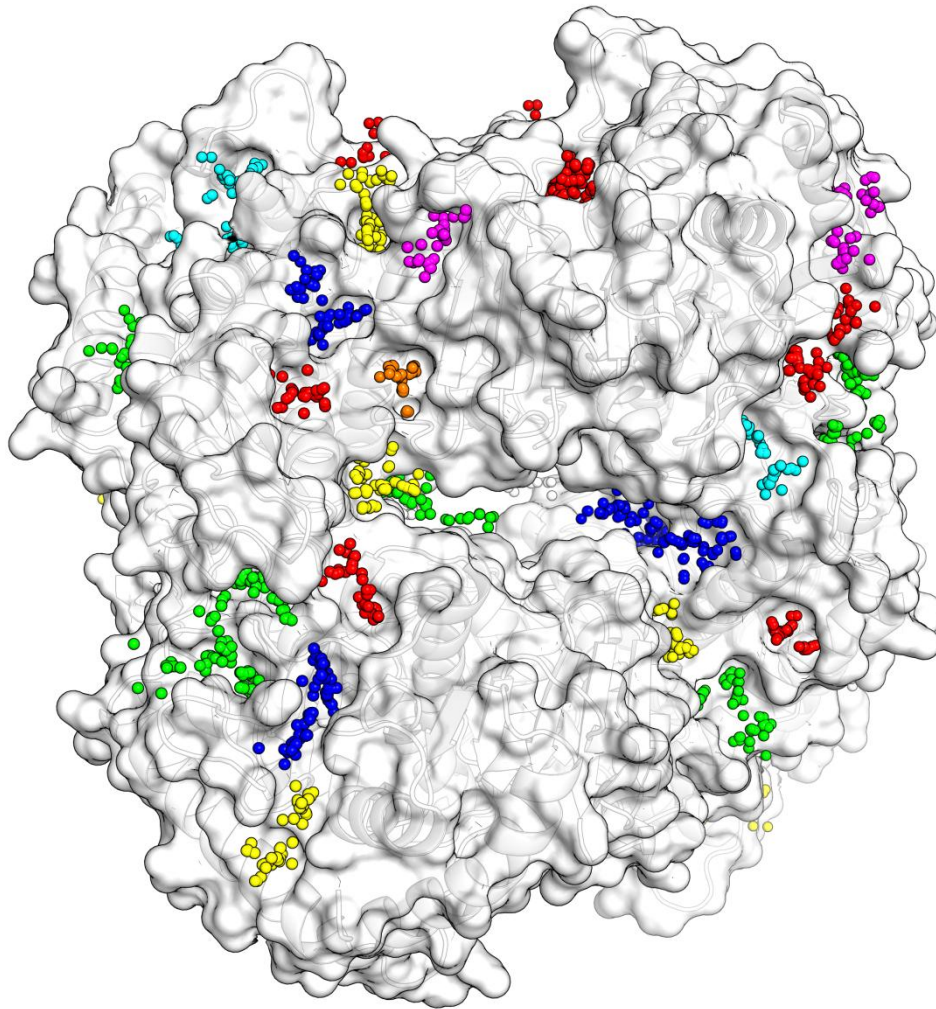
- ГАФД – ключевой метаболический фермент в клетке;
- Рассматривается как перспективная мишень для создания новых антибиотиков;

Проблема ГАФД как мишени



- Активный центр фермента очень похож (консервативен) у патогенных бактерий и человека;
- Трудно создать селективные ингибиторы;

Задача поиска новых путей регуляции ГАФД



- В последние годы появляется все больше примеров явления аллостерии (т.е. существование нескольких важных сайтов в белках);
- Структура ГАФД содержит большое количество полостей – потенциальных сайтов связывания;
- Задача: найти новый регуляторный сайт, строение которого отличается в ГАФД из человека и бактерий;

Оригинальный комплексный подход для поиска новых путей регуляции ферментов

1. **Биоинформатический анализ** семейства ГАФД и ГАФД-подобных белков для поиска и характеристики новых центров связывания;
2. ***In silico* скрининг (докинг)** для поиска лигандов, комплементарных найденному центру в ГАФД из бактерий, но не человека;
3. **Молекулярное моделирование** механизма действия предложенных ингибиторов;
4. **Экспериментальная проверка** предложенных соединений.

Проблема

- Решение задачи представляет собой конвейер из последовательных этапов, исполняемых **различными программами**, которые разрабатываются несвязанными между собой коллективами;
- Наибольшее распространение в биоинформатике получили такие языки как **Java, Perl и Python**, соответственно многие программы и скрепляющие их скрипты написаны на этих языках;
- Только отдельные алгоритмы способны использовать параллельный ресурс: филогенетический анализ и молекулярная динамики (MPI), структурные выравнивания (OpenMP), некоторые другие.

Задача поиска новых путей регуляции ферментов

- Требуется **анализировать данные большого объема** – о структурах (135 тыс.) и последовательностях (90 млн.) белков в публичных базах данных;
- Требуется **увеличивать точность вычислений** – за счет перенастройки параметров и усложнения протокола;
- Необходима **автоматизация основных процедур** для того, чтобы применять оригинальный комплексный подход к различным объектам (а не только ГАФД);
- Решение должно быть **«практичным»**.

1. Биоинформатический анализ для поиска и характеристики новых центров связывания

1.1 Поиск родственных ГАФД и ГАФД-подобных белков в базах данных на основании сходства структур и последовательностей;

1.2 Построение множественного выравнивания ГАФД и ГАФД-подобных белков;

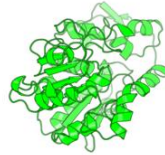
1.3 Анализ выравнивания и аннотация новых сайтов в структурах белков.

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

«Затравка»

(например, наиболее изученный белок суперсемейства)



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

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

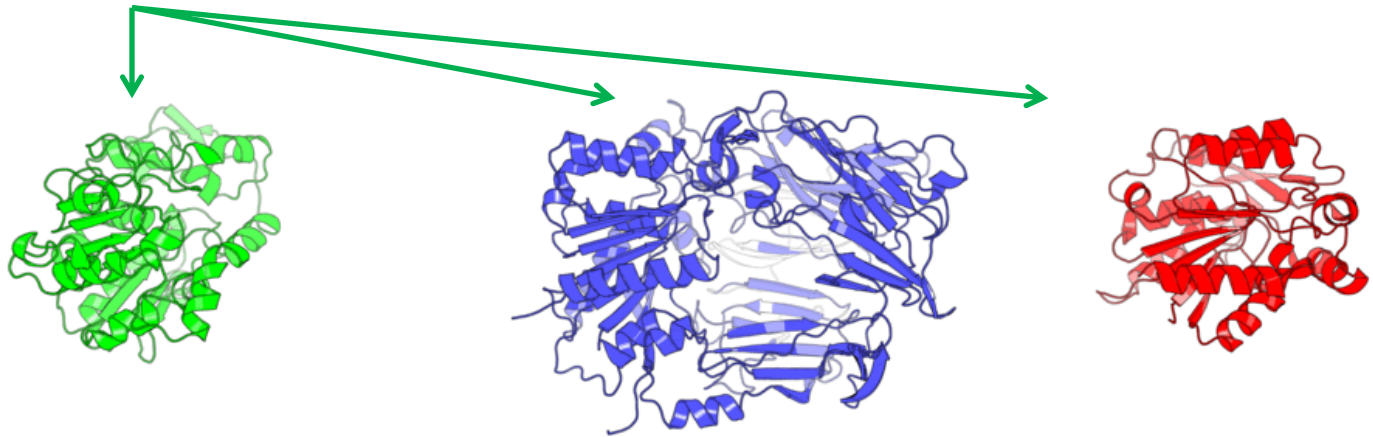
«Затравка»

(например, наиболее изученный белок суперсемейства)



Поиск по структурному схождению

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



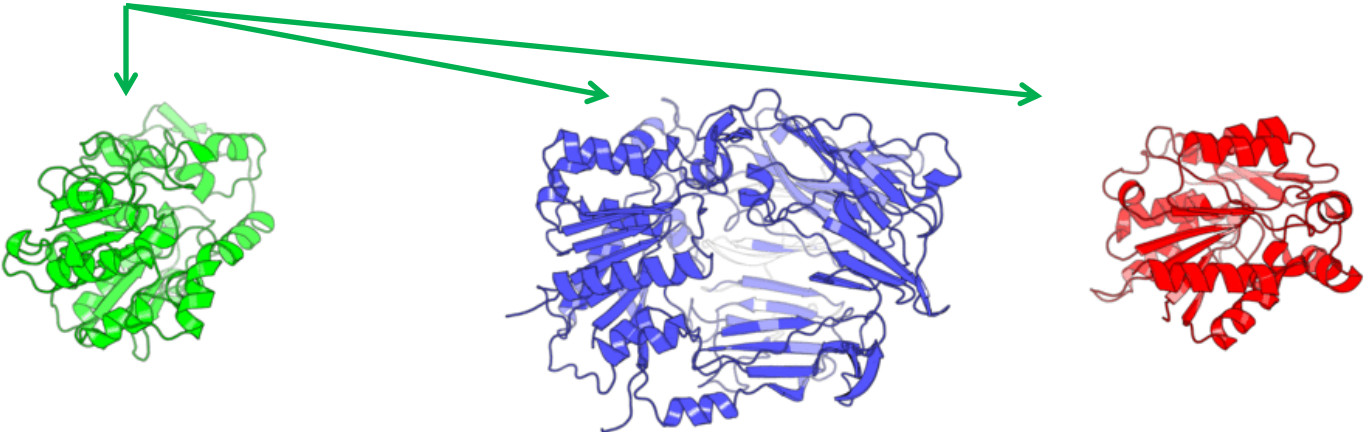
«Затравка»

(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

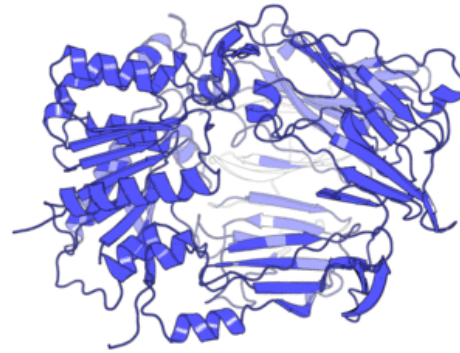
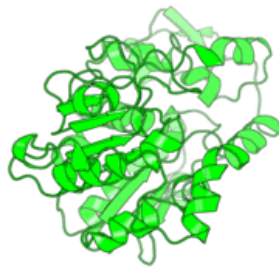
- > **protein1: ...AWSQGLP...**
- > **protein2: ...ASSQGAP...**
- > **protein3: ...LSNLDP...**
- > **protein4: ...WGSNAP...**
- > **protein5: ...LASNLP...**

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

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

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания



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...
```



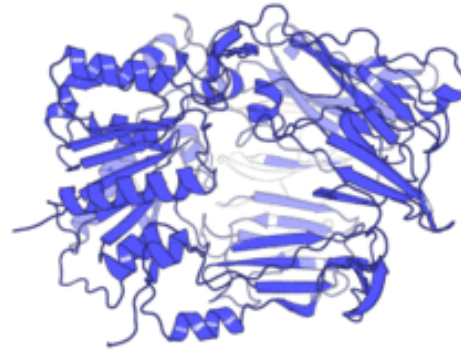
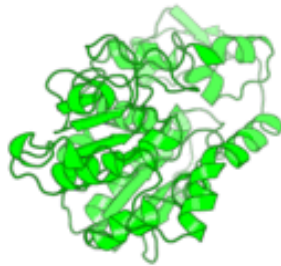
```
> protein1: ...KCDEP...
> protein2: ...RCDEP...
> protein3: ...RCDEP...
> protein4: ...KCEP...
> protein5: ...KCDEP...
```

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

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...
```

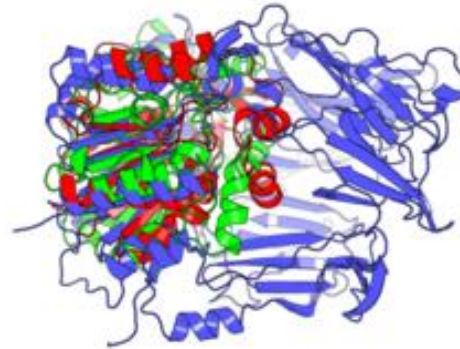
```
> protein1: ...KCDEP...
> protein2: ...RCDEP...
> protein3: ...RCDEP...
> protein4: ...KCEP...
> protein5: ...KCDEP...
```

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

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...
```

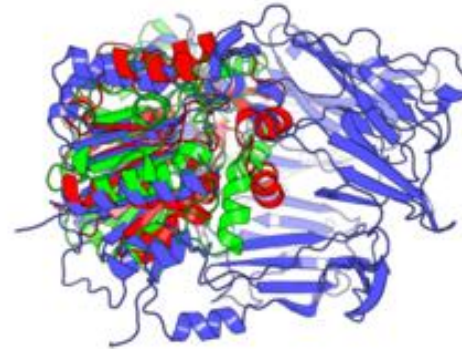
```
> protein1: ...KC DP...  
> protein2: ...RC DP...  
> protein3: ...RC DP...  
> protein4: ...KCEP...  
> protein5: ...KC DP...
```

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

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

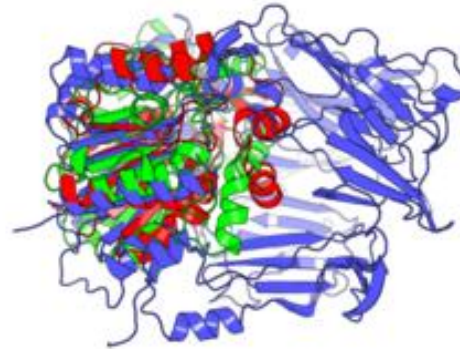
1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

Структурно-опосредованное множественное выравнивание

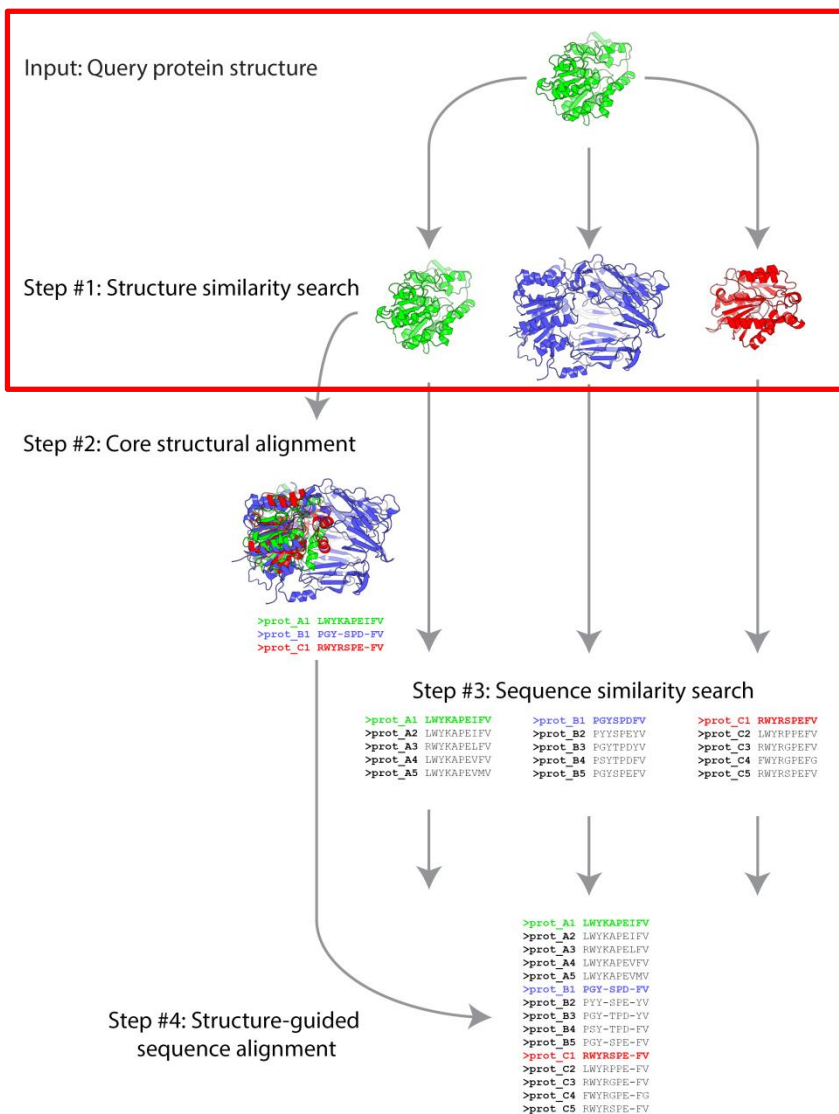
Эволюционно удаленные белки, которые потеряли сходство по последовательностям в процессе естественного отбора и специализации от общего предка, сравниваются посредством выравнивания структур, в то время как выравнивание аминокислотных последовательностей дает полезный результат в случае близких гомологов

```
> 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...
```



1.1 Поиск родственных белков

1.2 Построение множественного выравнивания



Этап 1: Поиск по структурному сходству

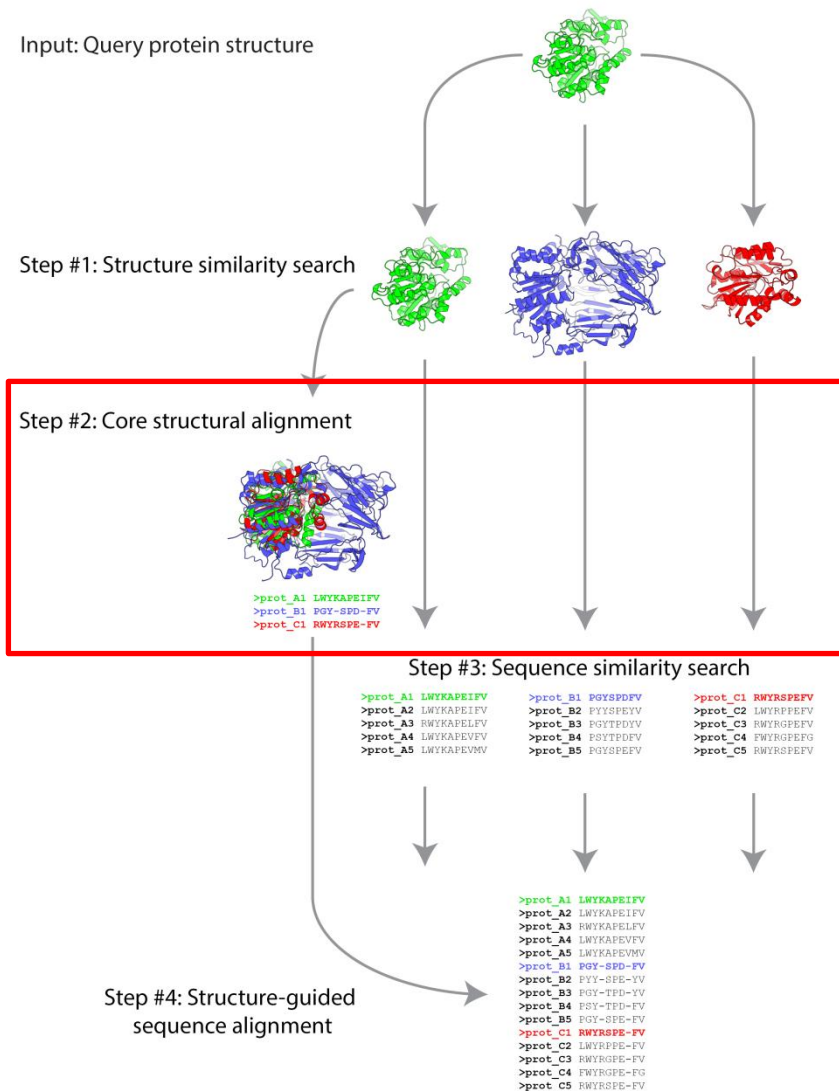
- Поиск в базе данных PDB ~135 тыс. структур белков или ~300 тыс. структур отдельных цепей – **несколько часов на 4 ядрах 1 CPU**;
- Уменьшаем объем базы для сравнения с конкретной затравкой:
 - Группируем структуры в базе по сходству и сравниваем только с одним представителем из каждой группы;
 - Не сравниваем с заведомо непохожими белками;
 - От **нескольких минут до 1 часа** на 4 ядрах 1 CPU
- Ускоряем повторный поиск:
 - Однажды посчитанные парные сравнения сохраняются в базе данных на основе PostgreSQL;
 - База хранится на SSD диске;
 - При повторном запросе поиск составляет **несколько секунд**;

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

Этап 2: Структурное выравнивание

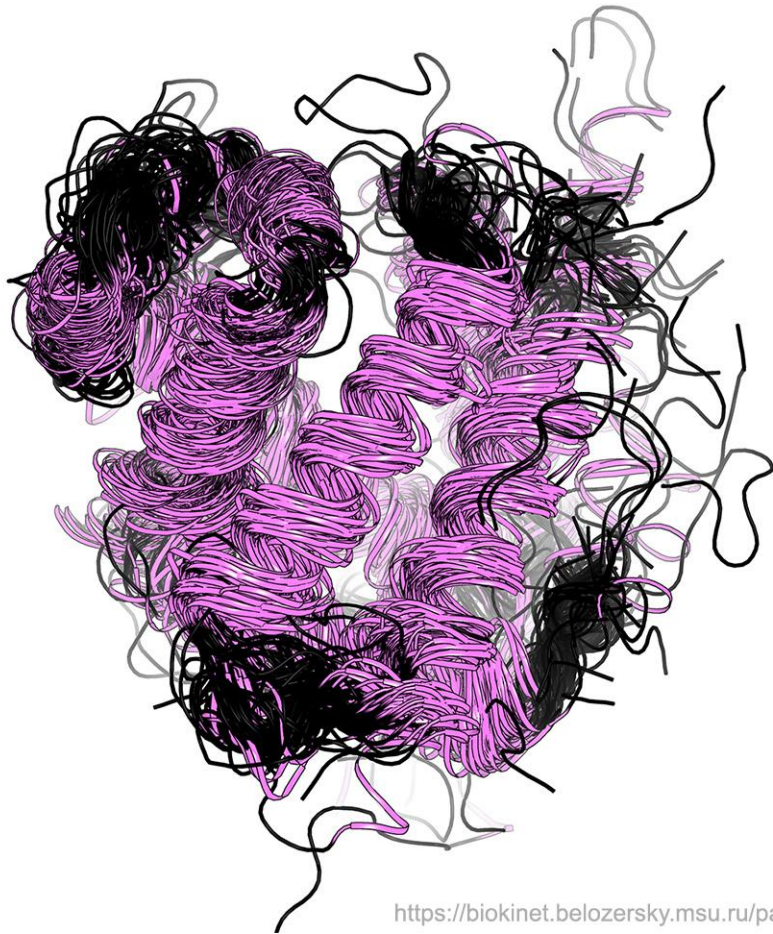
- От нескольких минут (2-32 структур) до 24 часов (<2000 структур) на одном CPU в режиме openMP (программа MATT);



1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

Этап 2: Структурное выравнивание



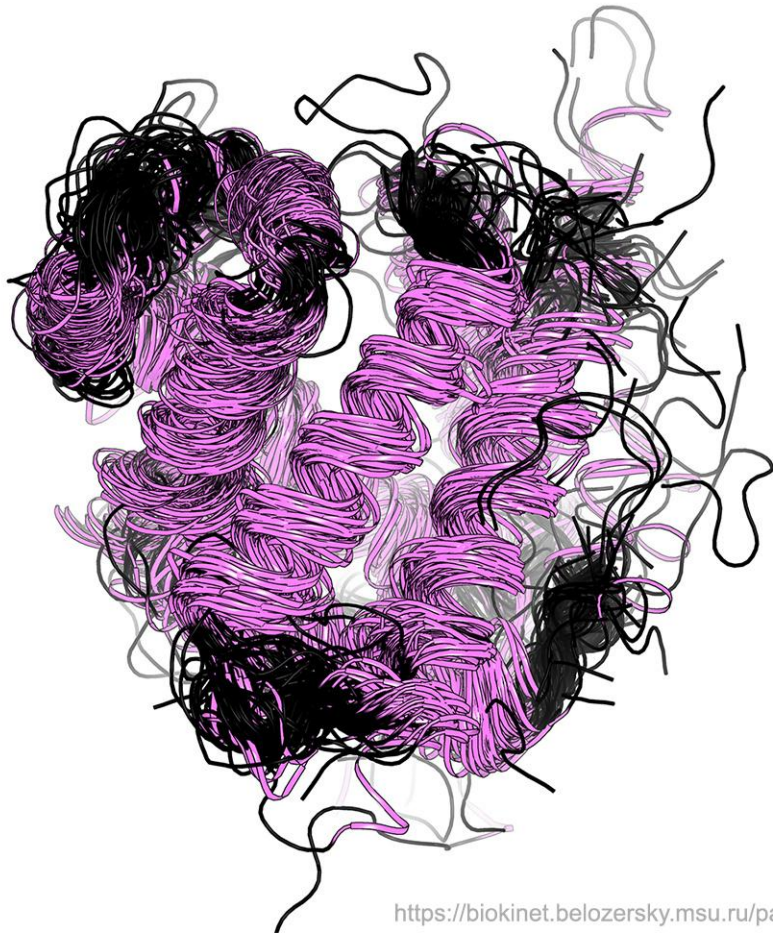
<https://biokinet.belozersky.msu.ru/parmatt>

- **От нескольких минут** (2-32 структур) **до 24 часов** (<2000 структур) на одном CPU в режиме openMP (программа MATT);

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

Этап 2: Структурное выравнивание



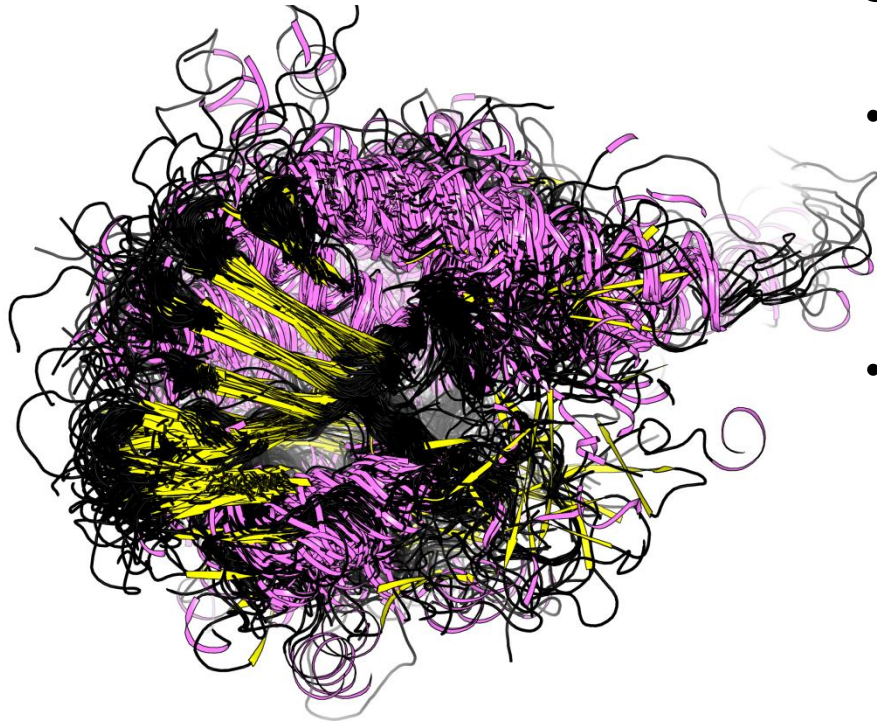
<https://biokinet.belozersky.msu.ru/parmatt>

- От нескольких минут (2-32 структур) до **24 часов** (<2000 структур) на одном CPU в режиме openMP (программа МАТТ);
- Разработана оригинальная программа **parМАТТ**:
 - Запуск в режиме MPI;
 - Ускорение на n узлах пропорционально размеру выборки и длине (количеству аминокислот) в выравниваемых белках;
 - Лучшее ускорение в тестах на Ломоносов-2:
 - До **x1.93** на **2** узлах;
 - До **x3.34** на **4** узлах;
 - До **x5.56** на **8** узлах;
 - До **x7.42** на **16** узлах;

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания

Этап 2: Структурное выравнивание

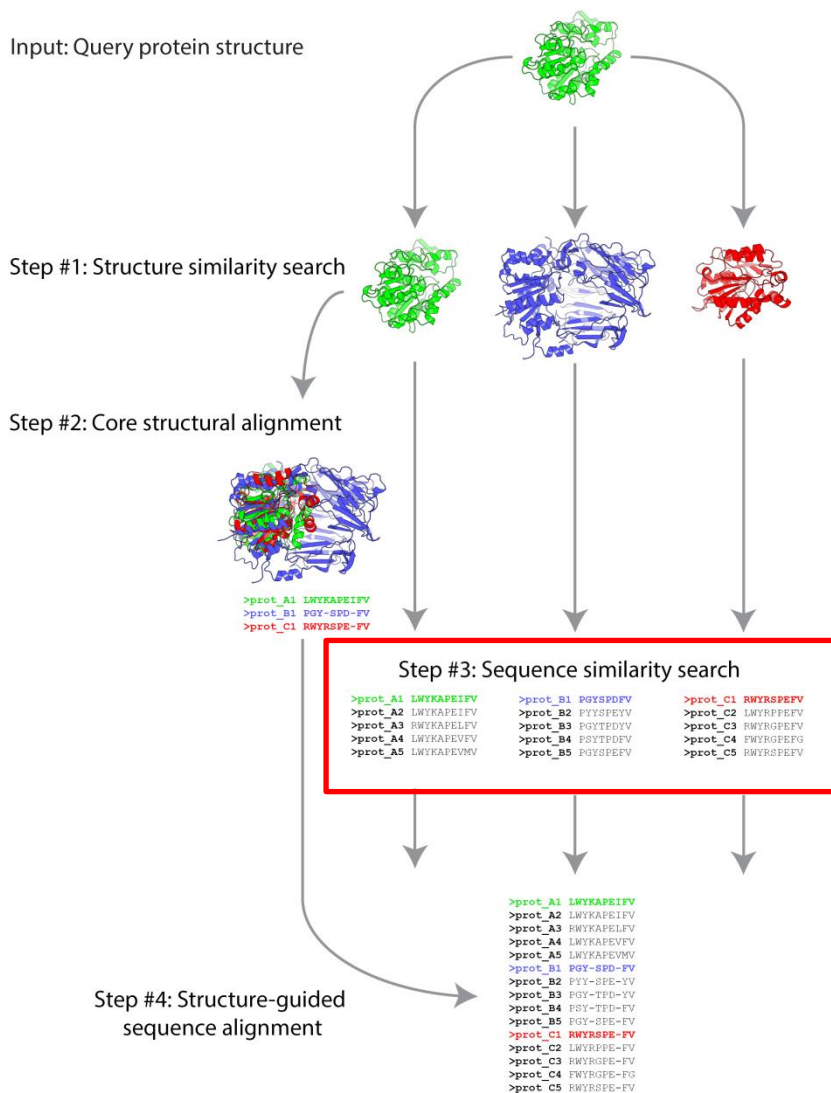


Выравнивание структур
родственных ГАФД и ГАФД-
подобных белков

- От нескольких минут (2-32 структур) до **24 часов** (<2000 структур) на одном CPU в режиме openMP (программа MATТ);
- Разработана программа **parMATТ** на основе алгоритма множественного структурного выравнивания белков MATТ (<https://biokinet.belozersky.msu.ru/parmatt>):
 - Запуск в режиме MPI;
 - Ускорение на n узлах пропорционально размеру выборки и длине (количеству аминокислот) в выравниваемых белках;
 - Лучшее ускорение в тестах на Ломоносов-2:
 - До **x1.93** на **2** узлах;
 - До **x3.34** на **4** узлах;
 - До **x5.56** на **8** узлах;
 - До **x7.42** на **16** узлах;

1.1 Поиск родственных белков

1.2 Построение множественного выравнивания



Этап 3: Поиск по сходству и выравнивание аминокислотных последовательностей

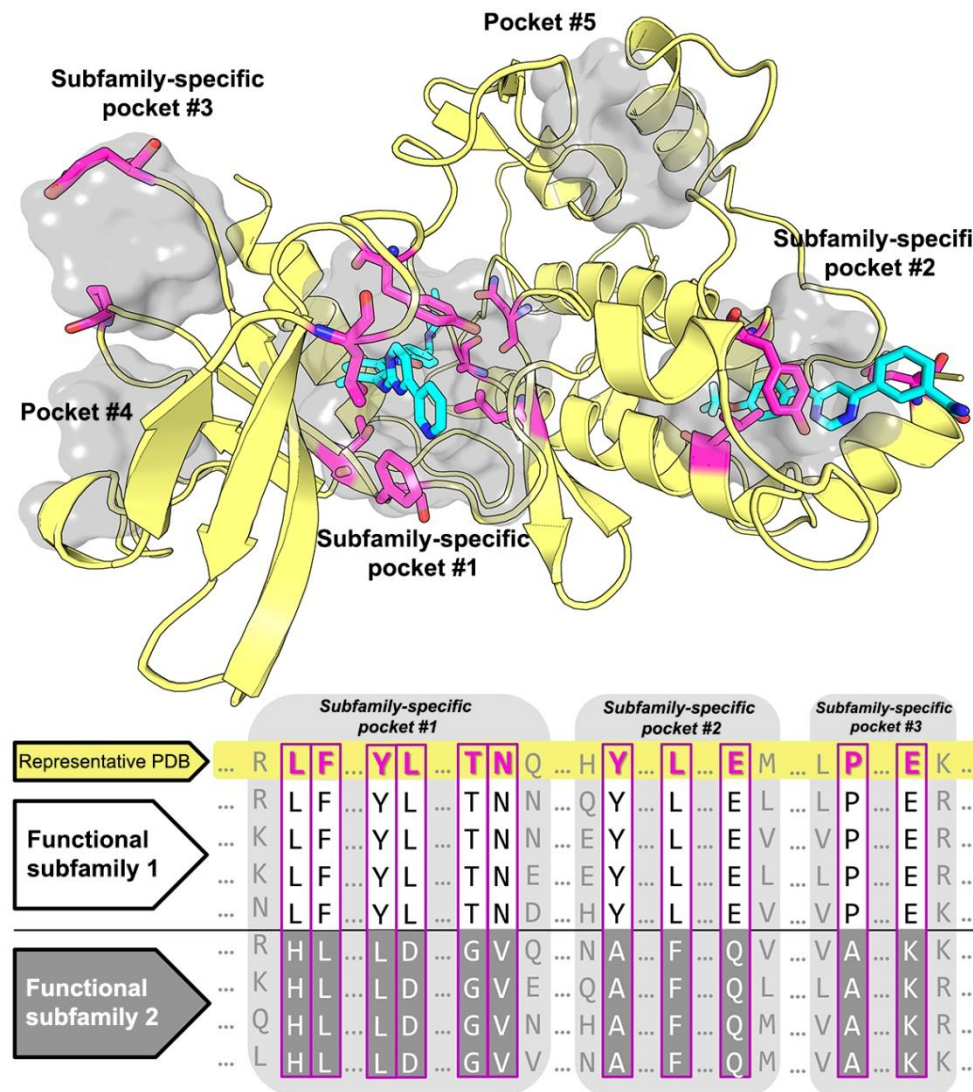
- Используется **GPU-версия** алгоритма BLAST
 - Поиск по базе Swiss-Prot (100 тыс.)
 - несколько секунд;
 - Поиск по базе TrEMBL (100 млн.)
 - до 10 минут;
 - до 8 поисков одновременно на одной видеокарте GTX 980 Ti;
- Выравнивание каждой собранной выборки последовательностей
 - Несколько секунд - до 1 минуты алгоритмом MAFFT на одном ядре.

1. Биоинформатический анализ для поиска и характеристики новых центров связывания

- Методы выравнивания аминокислотных **последовательностей** используются для сравнения эволюционно **близких** белков;
- Методы выравнивания **структур** используются для сравнения **далеких** гомологов;

1.3 Анализ выравнивания и предсказание сайтов

Биоинформатический анализ методом Zebra/pocketZebra



- Предсказание специфических позиций в колонках множественного выравнивания
 - параллельный анализ каждой отдельной колонки;
- Аннотация (ранжирование) сайтов в структуре по специфичности;

Suplatov, D., Kirilin, E., & Švedas, V. (2016). Bioinformatic Analysis of Protein Families to Select Function-Related Variable Positions. In *Understanding Enzymes: Function, Design, Engineering, and Analysis* (pp. 351-385) Ed. Allan Svendsen. Pan Stanford.

Suplatov, D., Voevodin, V., & Švedas, V. (2015). Robust enzyme design: Bioinformatic tools for improved protein stability. *Biotechnology journal*, 10(3), 344-355.

Suplatov, D., & Švedas, V. (2015). Study of functional and allosteric sites in protein superfamilies. *Acta Naturae*, 7(4), 27, 34-45.

Suplatov, D., Kirilin, E., Arbatsky, M., Takhaviev, V., & Švedas, V. (2014). pocketZebra: a web-server for automated selection and classification of subfamily-specific binding sites by bioinformatic analysis of diverse protein families. *Nucleic acids research*, 42(W1), W344-W349.

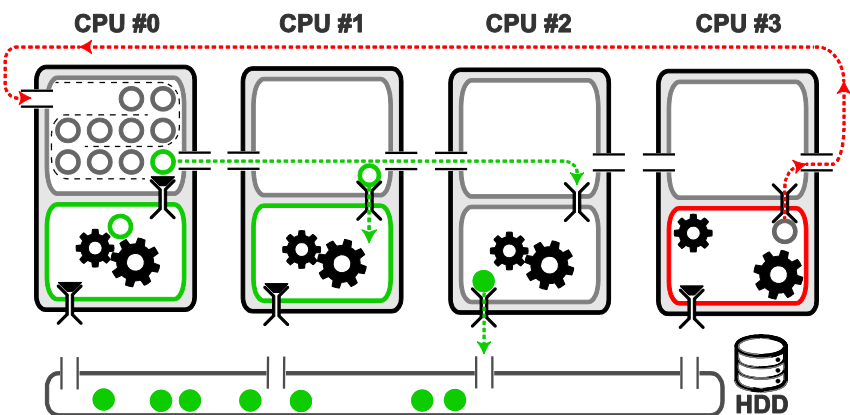
Suplatov, D., Kirilin, E., Takhaviev, V., & Švedas, V. (2014). Zebra: a web server for bioinformatic analysis of diverse protein families. *Journal of Biomolecular Structure and Dynamics*, 32(11), 1752-1758.

Suplatov, D., Shalaeva, D., Kirilin, E., Arzhanik, V., & Švedas, V. (2014). Bioinformatic analysis of protein families for identification of variable amino acid residues responsible for functional diversity. *Journal of Biomolecular Structure and Dynamics*, 32(1), 75-87.

1. Биоинформатический анализ для поиска и характеристики новых центров связывания

- Может быть эффективно исполнен **на одном узле**, содержащем
 - Мощный многоядерный процессор;
 - Мощный GPU ускоритель (для GPU-BLAST);
 - SSD накопитель (для хранения баз данных);
- Платформа Mustguseal (Multiple Structure-Guided Sequence Alignment of Protein Families) и интегрированные сервисы для анализа выравниваний:
 - <https://biokinet.belozersky.msu.ru/mustguseal>
 - **От нескольких секунд до 60 минут** на одно выравнивание размером до 10000 структур и последовательностей белков одного суперсемейства;
 - Общее число суперсемейств в базе данных CATH – 2738 записей.

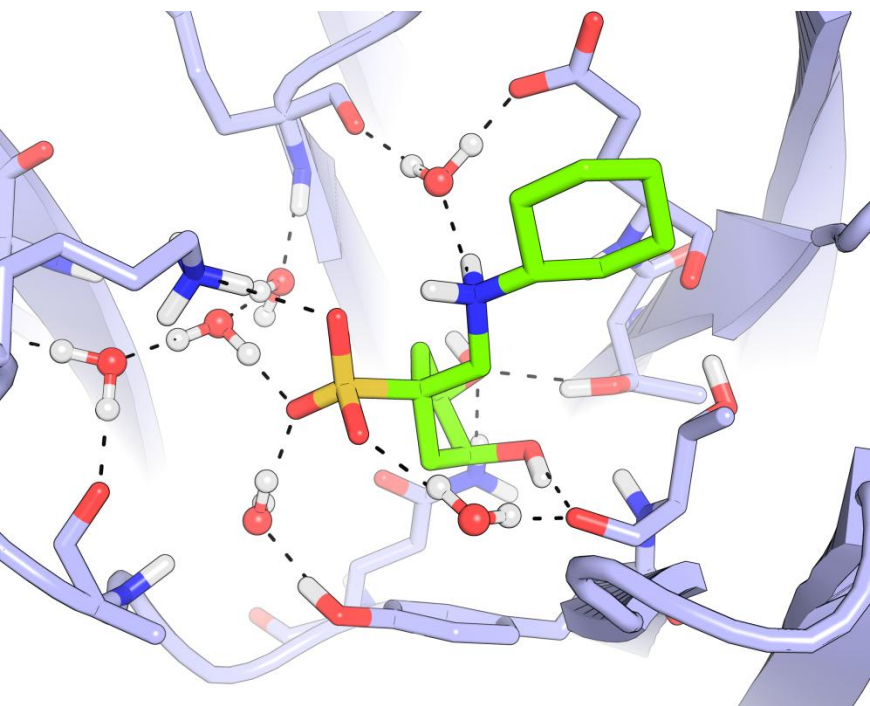
2. Скрининг ингибиторов (докинг)



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.

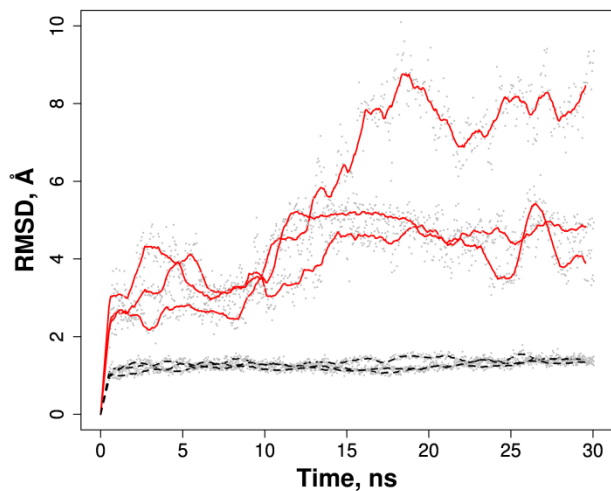
- **Программа mpiWrapper для запуска докинга программой Leadfinder:**
 - Докинг 1 лиганда в 1 сайт белка рассматривается как одна подзадача;
 - Скрининг 1 сайта белка библиотекой из n лигандов рассматривается как набор независимых подзадач;
- **Скринирование больших библиотек рецепторов и лигандов:**
 - Библиотека 10 конформеров сайта ГАФД из *Mycobacterium tuberculosis*;
 - Библиотека 2 млн. лигандов;
 - 20 млн. докингов за 12 часов на 100 узлах «Ломоносов-2».

3. Молекулярное моделирование распознавания ингибиторов в новой сайте в структуре ГАФД



GPU-версия программы AMBER:

- Позволяет анализировать большие системы (система ГАФД – более 100 тыс. атомов);
- Позволяет считать более длинные траектории (50 нс – 100 нс – 1 μ с);
- Позволяет усложнять процедуру моделирования:
 - Удлинять и усложнять количество стадий релаксации, чередовать канонические ансамбли (NVT, NPT);
 - Использовать более сложные 4-атомные молекулы воды;
 - Подготовительные этапы можно исполнять на локальном видеоускорителе (GTX 980 Ti);



4. Экспериментальная проверка

- Найдено **три селективных ингибитора** (X1, X2, X3) ГАФД из микобактерии – т.е. угнетают активность ГАФД из микобактерии и при этом не оказывают влияния на ГАФД из человека;
- Ингибитор X1 **подавляет рост патогенной культуры** *Mycobacterium tuberculosis* на 92%.

Выводы

- Изучение структуры, функции и регуляции белков с использованием биоинформатики и молекулярного моделирования представляет собой **конвейер из последовательных этапов**, исполняемых различными программами, предъявляющими свои требования к вычислительным ресурсам;
- Гибридные вычислительные кластеры – системы, обладающие существенной мощностью и разнообразием аппаратных возможностей – необходимы для того, чтобы **оптимально исполнить каждую отдельную стадию** единого комплексного решения;
- **GPU-ускорители** открывают новые возможности для поиска эффективных решений ресурсоемких задач биоинформатики и молекулярного моделирования.