

НММ-профили

Для поиска доменов и не только

Какие методы были в предыдущих сериях. Знать и уметь использовать

- 1) **PWM** для поиска коротких сигналов в ДНК
- 2) **MEME** и **FIMO** для поиск мотивов *de novo* в ДНК, РНК и белках
- 3) **PSSM** и **PSI-BLAST** для поиска семейств белков
- 4) Запись **паттернов** в ДНК, в JalView и в PROSITE
- 5) **Пакет HMMER**: создание профиля, калибровка (в HMMER2), поиск по профилю

1. PWM, вес и информационное содержание

TTATGCC
 ATCTTCA
 GTATTA

	1	2	3	4	5	6	7
G	0.26	-1.3	-1.3	-1.3	0.26	-1.3	-1.3
A	0.26	-1.3	0.74	-1.3	-1.3	0.26	0.74
T	0.26	1.18	-1.3	1.18	0.74	-1.3	-1.3
C	-1.3	-1.3	0.26	-1.3	-1.3	0.74	0.26

выравнивание



PWM для данного выравнивания

Элементы PWM: S_{ki} для основания i в позиции k ,
 p_i — фоновая частота основания i
 f_{ki} — частота основания i в позиции k
 (с учётом псевдоотсчётов)
 λ — любое число (для удобства)

$$I_k = \sum_i f_{ki} \log_2 \frac{f_{ki}}{p_i}$$

$$I = \sum_k I_k$$

$$S_{ki} = \frac{1}{\lambda} \log \frac{f_{ki}}{p_i}$$

Информационное содержание (I) позволяет понять, как много похожих на мотив последовательностей мы найдем в наших данных по случайным причинам.

Применение PWM

Приложив позиционную весовую матрицу (PWM) к последовательности той же длины, можно понять, содержит ли последовательность сигнал, описываемый этой PWM.

Чем выше вес, тем более вероятно, что последовательность содержит сигнал.

можно искать вероятные вхождения мотива в длинную последовательность (например, геном), считая вес всех возможных отрезков нужной длины: где вес выше порога, там предсказывается мотив. Выбор порога — отдельная задача.

2. MEME и FIMO

- Ищет мотивы заданной длины во входных последовательностях
- Мотив – набор участков заданной длины со сходными последовательностями в нескольких последовательностях.
- Результат PFM - частотная матрица букв в каждой колонке выравнивания (без гэпов)
- FIMO по PFM в формате MEME ищет мотивы в заданном множестве последовательностей

3. PSSM — position-specific scoring matrix

PSSM строится по выравниванию белков.

По той же формуле, что и PWM для сигналов ДНК.

Гэп используется как 21я буква

$$S_{ki} = \frac{1}{\lambda} \log \frac{f_{ki}}{p_i}$$

S_{ki} — элемент позиционной весовой матрицы - вес буквы i в позиции k

p_i — фоновая частота остатка i

f_{ki} — частота остатка i в позиции k
(с учётом псевдоотсчётов)

PSSM применяется для поиска последовательностей,
новых представителей семейства

Если вес последовательности белка относительно
PSSM выше порога, предсказывается принадлежность
белка семейству.

В вычислении частоты остатка в позиции
учитываются веса последовательности

Вес (weight) последовательностей, имеющих много
родственников, маленький, а у «одиноких»
последовательностей — большой.

При расчете частоты остатка i в позиции k используются веса
последовательностей

$$f_{ki} = \frac{\sum_{s:a_{sk}=i} w_s + \psi_i}{\sum_s w_s + \sum_i \psi_i}$$

Здесь a_{sk} — буква последовательности s в позиции k ,
 ψ_i — псевдоотсчёт для остатка i .

w_s - вес последовательности s

Если все веса последовательностей равны, то получится
обычная частота.

PSI-BLAST (Position-Specific Iterative BLAST) — разновидность BLASTP

- PSI-BLAST основан на использовании PSSM
- Работает интерактивно.
- Запускается BLASTP
- Находки выравниваются
- По выравниванию строится PSSM
- На второй итерации PSSM (вместо входной последовательности) выравнивается с белковыми последовательностями из банка и отбираются находки
- По находкам строится новая PSSM
- Итерации повторяются пока список находок не стабилизируется.

Благодаря использованию PSSM, PSI-BLAST способен находить более дальних родственников входного белка.

4. Паттерны что такое

Запись выравнивания в виде регулярного выражения

Правила записи:

<https://myhits.sib.swiss/cgi-bin/help?doc=pattern.html>

Пример паттерна

< A-x-[ST](2)-x(0,1)-{RK}-V

Поиск по паттерну

- PROSITE
- MyHits <https://myhits.sib.swiss/>
- fuzzpro из пакета EMBOSS (на kodomo стоит)

Паттерн для цинкового пальца

Prosite

Паттерн для цинкового пальца типа C2H2:

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

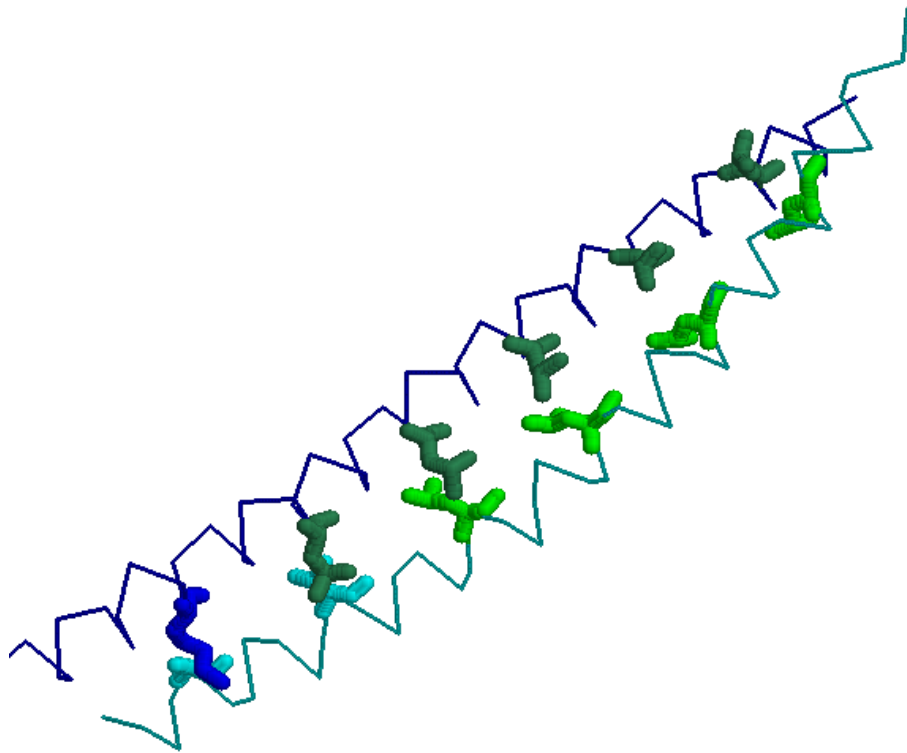
- [a-zAZ] – все возможные аминокислоты в данной позиции
- X(2,4) – любая аминокислота от 2 до 4 раз
- X(3) – любая аминокислота ровно 3 раза
- {P} – любая аминокислота, кроме пролина

Паттерны (fingerprints) для белков и средства поиска по паттерну есть в ProSite, myHits, пакете EMBOSS

Пример мотива: Лейциновая молния (Leucine zipper)

LEUCINE_ZIPPER, [PS00029](#); Leucine zipper pattern (PATTERN with a high probability of occurrence!)

L-x(6)-L-x(6)-L-x(6)-L



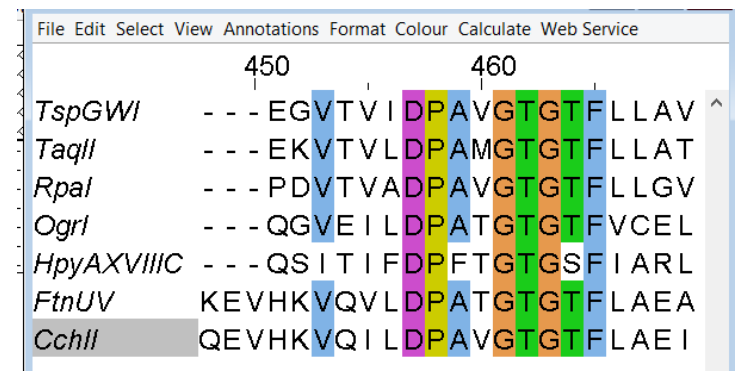
Показаны каждый 7й остаток цепей А и В;
Leu - зеленые (А) и темнозеленые (В)

PDB код 1ci6

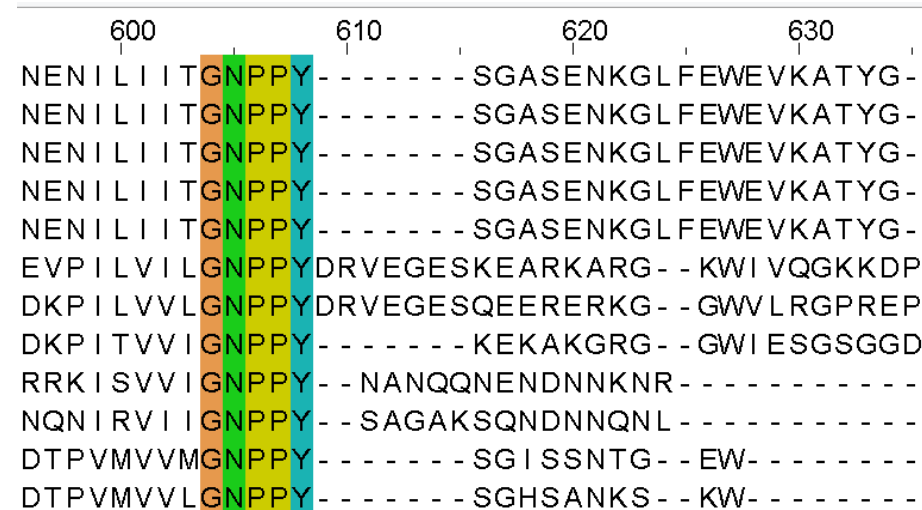
Паттерны в Jalview (Select => find)

Regular Expression Element	Effect
.	Matches any single character
[]	Matches any one of the characters in the brackets
^	Matches at the start of an ID or sequence
\$	Matches at the end of an ID or sequence
*	Matches if the preceding element matches zero or more times
?	Matches if the preceding element matched once or not at all
+	Matches if the preceding element matched at least once
{count}	Matches if the preceding element matches a specified number of times
{min,}	Matches if the preceding element matched at least the specified number of times
{min,max}	Matches if the preceding element matches min or at most max number of times

Поиск мотивов в выравнивании белков



- Короткие консервативные последовательности в гомологичных (иногда и не гомологичных) белках
 - Активные центры ферментов
GNPPY у одного семейства ДНК метилтрансфераз
 - Участки связывания лигандов
D...GTG[ST]F связывание SAM – источника метильной группы у того же семейства
 - Участки взаимодействия с другими белками, ДНК, РНК
 - Поддержание 3D структуры белка
 - Другие



II. Технология профилей

База данных Pfam

<http://pfam-legacy.xfam.org/>

Поглощена БД INTERPRO в 2022

ПРОФИЛЬ – описание выравнивания, вроде PWM и PSSM, но другая теория

Разрешаются и обоснованно штрафуются индели (=INsertions and DEletions) в выравнивании. Этим профили отличаются от PWM и PSSM

ПРОФИЛИ применяются для поиска **доменов** в последовательностях белков и новых представителей семейств гомологичных белков

НММ (Hidden Markov Model) профиль

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	m->m	m->i	m->d	i->m	i->i	d->m	d->d	b->m	m->e											
	-415	*	-2000																	
1	-791	-1639	2523	-46	-1622	-1478	-559	-1172	-464	-1286	3030	-325	-1789	-271	-936	-789	-810	-1041	-1997	-1392
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	-415	*											
2	-736	-652	-2436	-1882	1566	-2201	-1008	349	-1593	1464	629	-1652	-2226	-1291	-1596	-1297	1715	233	-906	-449
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
3	-859	-1565	-395	2243	-1354	-1667	-572	-808	-308	1279	-469	-504	-1886	-286	-662	-909	-833	-789	-1827	-1297
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
4	-827	-2402	1673	1893	-2690	-1247	-316	-2490	-203	-2436	-1614	126	-1606	73	-830	1567	-829	-2029	-2632	-1831
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
5	-570	-1337	-812	-211	-1531	-1573	-150	-1058	508	-1233	-515	-382	-1659	2039	1408	-610	-491	1401	-1595	-1138
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
6	-1612	-1300	-3691	-3208	-436	-3362	-2409	754	-2880	2446	670	-2994	-3170	-2428	-2764	-2592	-1571	1569	-1833	-1642
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
7	-425	-1013	-1661	-1792	-2558	-1187	-1802	-2565	-1897	-2800	-2098	-1367	-1878	-1746	-1993	3270	-814	-1808	-2742	-2368
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
8	-947	-2237	2611	156	-2593	-1441	-415	-2401	-37	-2357	-1575	-157	-1753	-36	2126	-793	-935	-1995	-2445	-1818
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
9	-918	-2246	23	2288	-2630	-1481	-246	-2321	2042	-2231	-1436	-134	-1719	159	70	-755	-856	-1933	-2330	-1753
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											
10	-1267	-3033	2464	2571	-3246	-1297	-591	-3134	-741	-3040	-2335	96	-1794	-250	-1516	-962	-1317	-2641	-3214	-2285
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-249
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											

Разберем где что позже.

Порядок действий при создании профиля.

1. Эксперт составляет выравнивание seed.
Одним из источников новых доменов служат автоматически собираемые сходные фрагменты из разных белков. Ранее они хранились в Pfam-B секции. Записи из Pfam-B ныне переформатированы в DUF.
2. Строит HMM профиль с помощью пакета HMMER.
Программа hmtbuild
3. Калибрует профиль на случайных последовательностях для нормализации веса и E-value последовательностей (в HMMER3 входит в hmtbuild)
4. Проверяет профиль на перепредсказания (белки, в которых не должно быть находок) и недопредсказания (белки, в которых можно ожидать наличие находок)
5. С помощью профиля находит домены в всех последовательностях из БД (Uniprot и др.)
6. Готовит запись в банк Pfam

Ссылки на описания пакета HMMER

- <https://rothlab.ucdavis.edu/genhelp/hmmerbuild.html>
Простая, но старая (2005). Годится для HMMER2
- Eddy и команда разработчиков, 2023
chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://eddylab.org/software/hmmer/Userguide.pdf
Описывает HMMER3 (про HMMER2 тоже есть)
Недостаток – очень много всего. Надо искать нужное поиском. Текст написан б.м. понятно

Домен HPPK. Выравнивание SEED для профиля

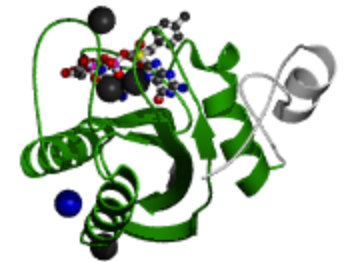
Seed sequence alignment for PF01288

Family: *HPPK* (PF01288)

Q02AG5_SOLUE/5-132	YLSLGSNI	G	D	R	H	A	N	L	RAAI	EAL	D	AG
S0EYB2_CHTCT/11-144	YLGGLGSSL	G	D	R	L	Q	N	L	QKAL	QRL		
C0ZID7_BREBN/6-134	YLALGSN	L	D	R	A	Q	N	L	RRAI	QRL	NE	QP
Q5WLU7_BACSK/5-133	YIALGSN	V	D	R	E	N	Y	L	QEAM	KLL	DA	DA
E6TSF5_BACCJ/10-138	YLSLGSN	I	E	S	R	Y	D	L	TFAL	KKL	RE	NP
G2THL3_BACCO/6-134	YLALGSN	I	E	P	R	F	D	L	QHAI	RLL	RN	NP
K0J162_AMPXN/5-133	YIALGSN	I	N	P	R	N	E	F	L	EQAI	NEI	EQ
I0JH59_HALH3/5-133	YIALGSN	I	S	K	R	E	E	F	L	ENAV	ASI	DD
Q8EU11_OCEIH/5-133	YVALGTN	I	E	P	R	E	N	F	I	NQAL	QFL	DD
B7GFK5_ANOFW/6-134	YIALGSN	I	G	D	R	F	E	Y	L	CKAV	IAL	RD
Q5L443_GEOKA/6-134	YLALGSN	L	G	D	R	V	S	Y	L	RSAL	EAL	HH
C5D399_GEOSW/6-134	YIALGSN	I	G	D	R	L	Y	Y	L	REAV	KML	DR
Q65PE2_BACLD/6-134	YIALGSN	I	G	R	R	E	E	Y	L	KKAV	SLL	HQ
HPPK_BACSU/6-134	YIALGSN	I	G	D	R	E	T	Y	L	RQAV	ALL	HQ
A8F946_BACP2/6-134	YIALGSN	I	G	K	K	E	T	Y	L	KEAV	KKL	HE
Q81VW6_BACAN/6-134	YIALGSN	I	G	E	R	Y	T	Y	L	TEAI	QFL	NK
Q9KGG7_BACHD/6-134	YIALGSN	I	G	D	R	S	R	F	L	EEAI	QQL	AE
D3FR36_BACPE/6-134	YIALGSN	I	G	D	R	A	A	Y	L	EEAI	DRL	DK
N0ATU2_9BACI/7-135	YLSIGSN	M	G	D	R	F	Y	Y	L	KNAI	QLL	TN
U5L4K3_9BACI/6-134	FIALGSN	M	G	D	R	A	A	N	L	KEAI	QML	SE
H6NSD7_9BACL/18-146	YIIGLSN	L	G	D	R	E	Q	Y	L	KEAL	RML	EE
L0EIN8_THECK/16-144	YIALGSN	L	G	D	R	E	A	Q	L	AEAL	RRL	HA
D3E785_GEOS4/13-141	YIALGAN	L	G	D	R	E	G	N	L	MEAL	ERL	DE
E3EET6_PAEPS/13-141	YIALGAN	L	G	E	R	E	H	T	L	YEAI	TAL	DE
X4ZBV9_9BACL/13-141	YIALGAN	L	G	D	R	E	Q	S	L	KEAL	TLL	NA
C6CRP5_PAESJ/14-142	YIALGSN	L	N	D	R	E	E	L	L	QQAV	EHL	RQ
C4KZT0_EXISA/3-130	YIALGANI	G	D	R	A	G	Q	L	SAAI		DE	ME
B1YGR6_EXIS2/5-133	YIALGSNI	G	D	K	A	G	H	L	RAAI	EA	MR	
E6U3M2_ETHHY/10-137	YIALGSN	M	G	D	R	A	G	Y	L	EAAR	KKI	AE
I0IE19_PHYMF/13-147	HWALGSNL	G	D	R	G	A	H	L	LAAC	RRLA	AAPG	
C9RLK0_FIBSS/8-134	YIALGSNL	P	D	R	S	A	H	L	KAGR	DML	HR	
K4LLB0_THEPS/7-135	FLSLGSN	L	G	N	R	S	A	Y	L	EAAC	REL	AA
L7VQA6_CLOSH/5-133	ILSLGSNI	G	D	R	E	K	N	L	KTAL	YHI	IQ	NP
A3DIK4_CLOTH/6-134	FLSLGSN	I	E	D	R	E	K	Y	L	LDAL	DNI	SA
G8LSW4_CLOCD/5-133	FLSLGSN	L	G	D	R	E	K	Y	L	FEAV	DEI	SK
D9QRZ5_ACEAZ/5-133	YLSLGSN	K	E	S	R	E	E	Y	L	QRAL	KKL	QD
E4RM72_HALHG/5-133	FLGLGSNI	E	P	R	S	E	Y	L	KKAA	AEL		
F0SWA2_SYNGF/4-132	FLGLGSN	L	G	D	R	R	S	Y	L	KKAV	RML	KE
F4LQD8_TREBD/64-196	VLGLGSNR	S	F	G	L	L	S	S	A	E	I	L
F2NVX4_TRES6/5-137	VLGLGSNK	S	F	G	A	F	S	S	L	E	L	L
F8F3E4_TRECH/9-138	VLGLGSNQ	G	E	S	R	T	I	L	QHAI	TDLESRIODL		
F5YC59_TREAZ/5-134	VLGLGSNQ	G	D	S	L	R	I	L	EKAV	EVLGII	LSL	
F2F163_SOLSS/6-134	YLSIGTN	I	G	E	R	E	Q	N	L	QDAV	KLL	TA
Q8YAC0_LISMO/5-133	FLSIGTN	I	G	E	R	L	E	N	L	NDAL	RGL	AA
Q2G0Q5_STAA8/5-133	YIIGLSN	I	G	D	R	E	S	Q	L	NDAL	KIL	NE
Q2G0Q5_STAA8/5-133 (SS)	EEEEEE	S	S	S	I	I	H	I	I	H	H	H
Q5HRN8_STAEQ/5-133	YIIGLSN	I	G	N	R	E	L	Q	L	NEAI	KIL	HD
Q18BX4_PEPD6/5-133	YIIGITN	M	G	D	R	F	D	N	L	SRAC	ELL	KN

Seed
(1006)

7,8-dihydro-6-hydroxymethylpterin-pyrophosphokinase (HPPK)



Строки с именем, помеченным (ss) содержат вторичную структуру белка с известной 3D структурой

Take home message

Выравнивания сотен и тысяч последовательностей белков почти ВСЕГДА СОДЕРЖАТ ОШИБКИ

Проблема: исправление ошибок возможно, но нет программ, которые сделают это за вас автоматически.

Есть соображения как создать такую. Ал

НММ Профиль. Немножко теории

- По выравниванию создается автомат для генерации последовательностей
 - Этот автомат умеет генерировать случайные последовательности конечной (но не фиксированной!) длины
 - Он настроен так, чтобы создавать последовательности, “похожие” на выравнивание, с бóльшей вероятностью
- Для каждой входной последовательности можно (т.е. существуют алгоритмы) определить вероятность её сгенерировать этим автоматом.
- Если эта вероятность превышает порог, то последовательность считается соответствующей профилю.

Автомат выглядит так:

Выравнивание

A	C	A	-	-	-	A	T	G
T	C	A	A	C	T	A	T	C
A	C	A	C	-	-	A	G	C
A	G	A	-	-	-	A	T	C
A	C	C	G	-	-	A	T	C

Вероятности в квадратиках называются

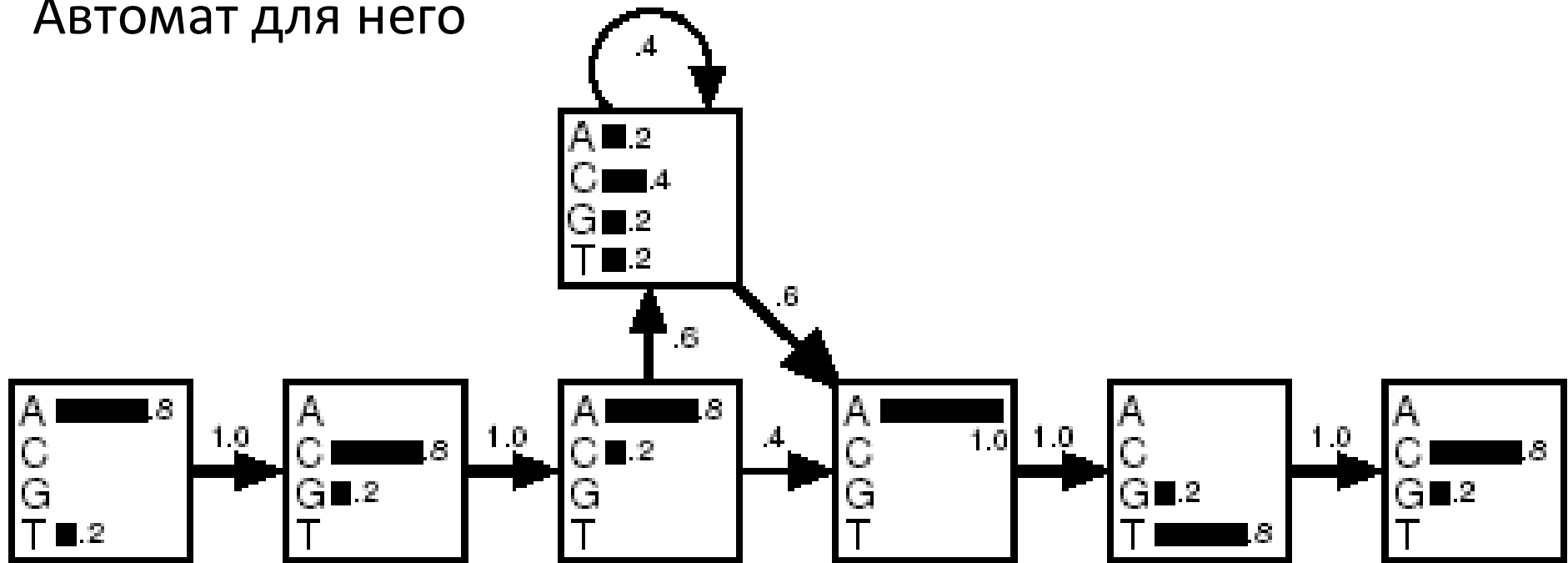
эмиссионными (emission)

Вероятности на стрелочках -

вероятностями перехода (transition)

Вероятности вычисляются по частотам

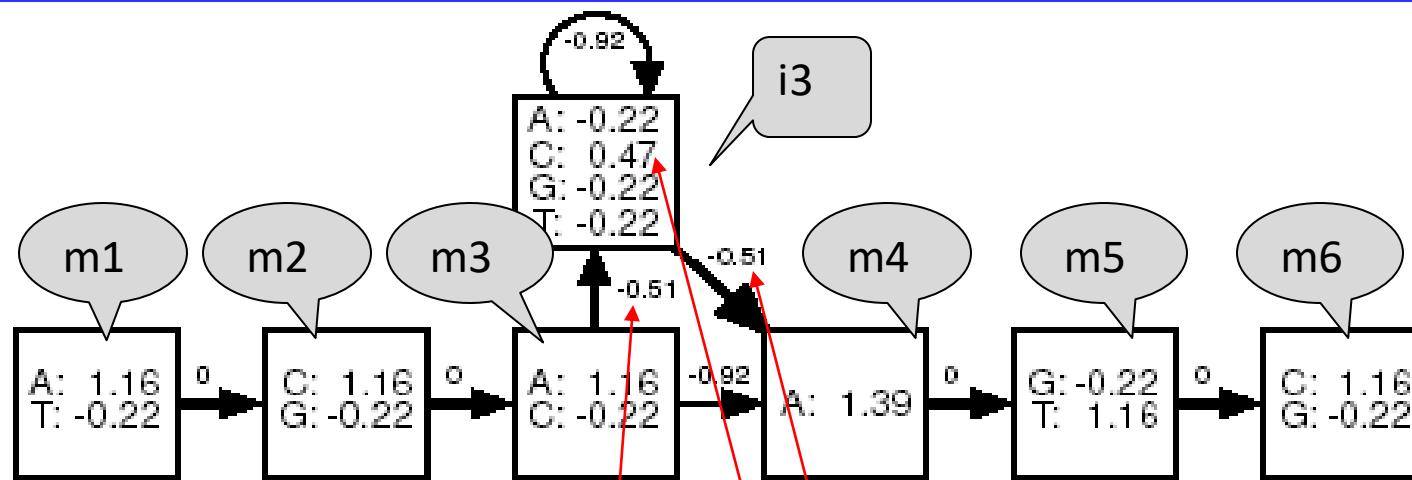
Автомат для него



Вес выравнивания последовательности ACACATC с профилем

Вместо вероятностей в профиле *используют логарифмы отношения правдоподобия*
 $\log_2(\text{частота буквы в колонке}/\text{базовая частота буквы})$

	Sequence	Probability $\times 100$	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original	A C A - - - A T G	3.3	4.9
sequences	T C A A C T A T C	0.0075	3.0
	A C A C - - A G C	1.2	5.3
	A G A - - - A T C	3.3	4.9
	A C C G - - A T C	0.59	4.6
Exceptional	T G C T - - A G G	0.0023	-0.97



$$-0.51 + 0.47 - 0.51$$

$$\text{Log odds} = 1.16 + 0 + 1.16 + 0 + 1.16 + 0 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.64$$

Мы нашли

- Оптимальное выравнивание
 - **A C A C A T C**
 - **m1 m2 m3 i3 m4 m5 m6**
- Его вес $1.16 + 0 + 1.16 + 0 + 1.16 - 0.51 + 0.47 - 0.51 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.64$

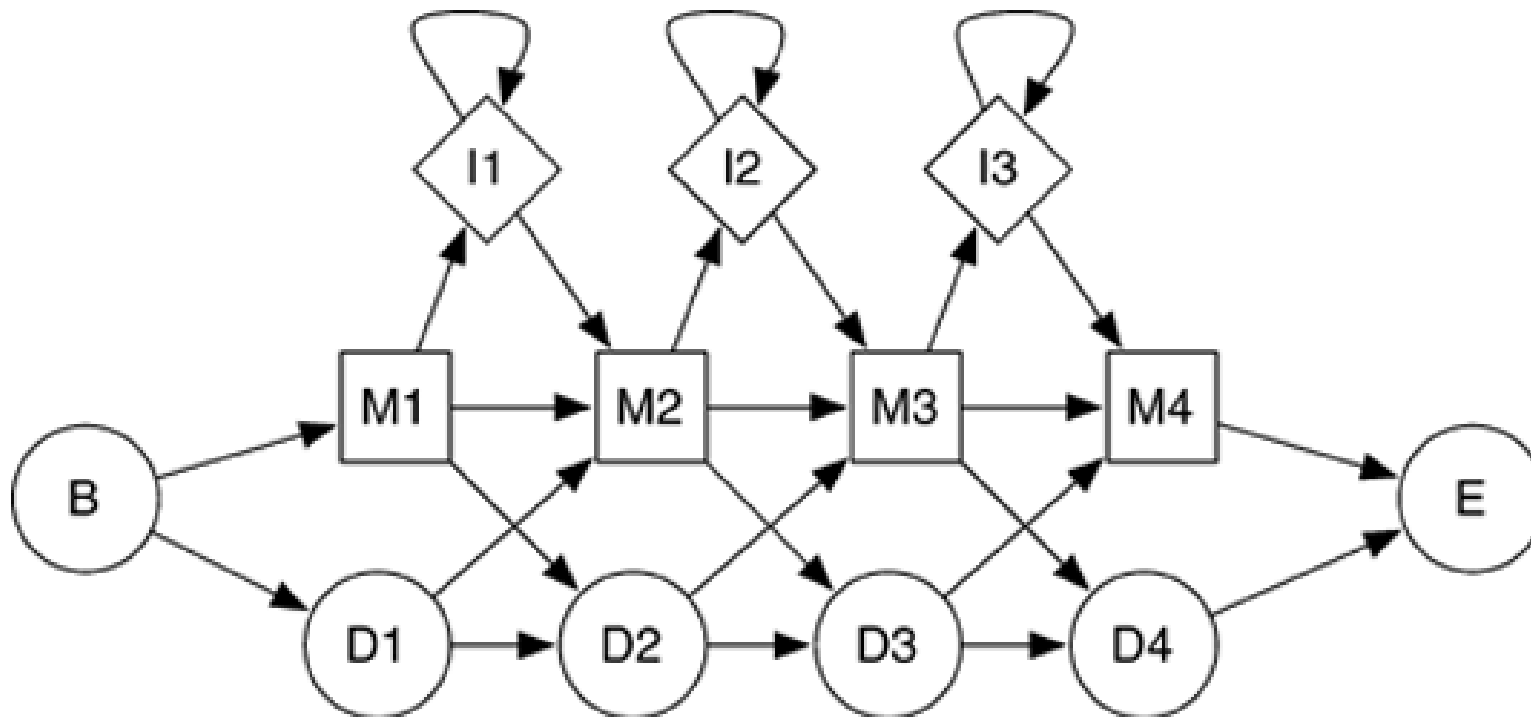
Задачу нахождения лучшего по весу выравнивания входной последовательности и НММ профиля решает алгоритм Viterbi

Семь типов транзиций (не считая начала и конца) Замены (M), Вставки (I), делеции (D)

На КАЖДОЙ стрелке стоит число – вероятность перехода (transition)

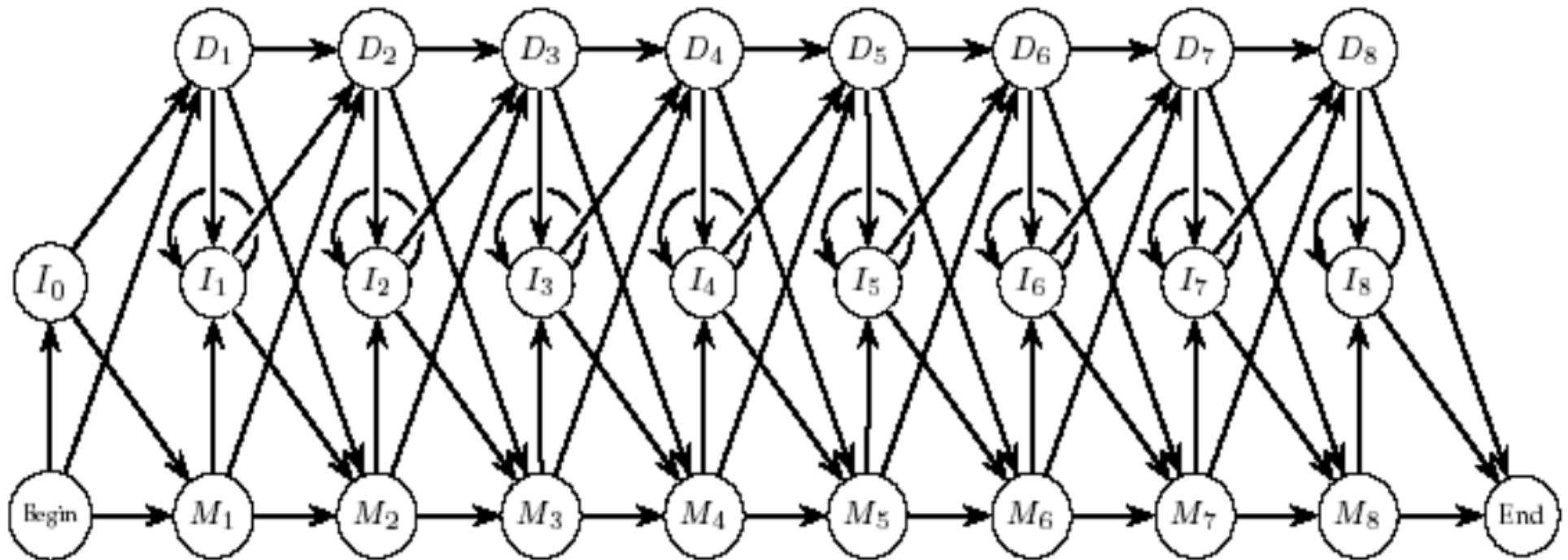
В КАЖДОЙ клеточке [M] для каждой буквы с стоит её вероятность (emission)

В КАЖДОЙ клеточке [M] для стрелки в ромбик <I> для каждой буквы стоит вероятность её вставки (emission)



В профиле хранятся не вероятности,
а логарифмы отношения правдоподобия

Граф НММ для выравнивания, в котором восемь колонок без гэпов, вставки и делеции разрешены в любом месте, но штрафуются



Из презентации безымянного сотрудника ИППИ)

НММ Профиль – описание автомата НММ в файле.

- НММ профиль имеет строгий формат
- По НММ-profile можно нарисовать автомат, по автомату можно создать НММ-profile
- РАСШИРЕНИЕ файла .hmm
- два основных пакета для работы с hmm-профилями:
 - HMMER2 старый – существует с конца 1990х, развивается, актуален и сейчас
 - HMMER3 новый - существует с 2010х, развивается, превосходит HMMER2, но не во всем
 - Pftools в Швейцарии и др.
- HMMER2 и HMMER3 полностью независимы
- Есть различия в используемых математических моделях описания выравнивания в профиле
- Форматы сопоставимы, но числа в них разные
- Есть и другие различия

HMMER2

Математическая модель в рамках
обсуждённого на лекциях

Частоты заменяются весами - логарифмами отношения правдоподобия (log-odds)

- Пусть базовые частоты всех букв одинаковы и, следовательно, равны 0.25. Пример на слайде -6 от этого.
- Отношение правдоподобия для буквы А в первой позиции примера равно $0.8/0.25 = 3.2$. Логарифм $\ln 3.2 = 1.16$
- Log-odds $\gg 0$ – за то, что буква А не случайно похожа на колонку выравнивания
- Log-odds ≈ 0 – за то, что буква А соответствует случайному выбору
- Log-odds $\ll 0$ – за то, что буква А избегается в колонке выравнивания
- Вероятности перехода заменяются логарифмами:
 $\ln(0.6) = -0.51$ Это как бы штраф за открытие гэпа
 $\ln(0.4) = -0.92$ Это как бы штраф за продолжение гэпа. Он большой, т.к. в примере только одна длинная вставка (пример на слайде -6 от этого)

	A	C	D	E	F	G	H
	m->m	m->i	m->d	i->m	i->i	d->m	d->d
	-50	*	-4862				
1	1250	-3656	-2029	-1481	1058	-3156	-1815
-	-149	-500	233	43	-381	399	106
-	-3	-9972	-11014	-894	-1115	-701	-1378
2	-2220	-3693	-283	-1518	-521	-1048	-1852
-	-149	-500	233	43	-381	399	106
-	-3	-10017	-11060	-894	-1115	-701	-1378
3	-174	-3763	-2136	-1588	-4084	-3263	-1922

Рис. Фрагмент файла bah-pf00145.hmm, построенного по выравниванию bah-pf00145-revised.fasta белков с двумя доменами: bah => pf00145

- Номера в первой колонке соответствуют клеточкам M автомата (M от Match).
- В строке с номером (1, 2 и далее) стоят веса за каждую букву в этой колонке (в выравнивании последовательности с профилем)
- В следующей строке стоят веса каждой буквы при начале вставки в последовательности по сравнению с профилем (см. пример на слайде -5). В данном файле эти строки идентичны для всех позиций
- В третьей строке стоят веса за все 7 типов transitions. Они разнятся в позициях

НММ профиль, построенный НМMer2'ом

log-odds(эмиссионных вероятностей для m)

log(вероятностей переходов

log-odds(эмиссионных вероятностей для i)

	A	C	D	E	F	G	H	I	K	L	M
	m->m	m->i	m->d	i->m	i->I	d->m	d->d	b->m	m->e		
1	-126	*	-3585								
-	-3610	-3114	-6053	-5506	2082	-5684	-4554	1759	-5277	2345	-632
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	-126	*		
2	604	2386	-4230	-3967	-3020	-2605	-3120	685	-3662	-2921	-2216
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
3	595	-2622	-4509	-4862	-5190	3595	-4388	-5082	-4974	-5307	-4405
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
4	-4592	-3891	-6106	-6010	4096	-5830	-2943	-1896	-5700	1283	-1205
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
5	403	-1180	-3654	-3023	2363	-2897	-1771	922	-2629	268	-383
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
6	-3348	-5115	3925	-1340	-5451	-3081	-2608	-5586	-3075	-5406	-4883
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		
7	2841	-2218	-4381	-4396	-4354	1529	-3793	-4064	-4191	-4344	1956
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720
-	-6	-8606	-9649	-894	-1115	-701	-1378	*	*		

Основные программы HMMER2 (на kodomo)

- Пользователь строит множественное выравнивание с инделями `my_alignment.fasta`
- `hmm2build -f --cpu 1 my_alignment.hmm my_alignment.fasta`
выходной файл `my_alignment.hmm`
опции `-f` строит локальные профили, а не один глобальный
`--cpu 1` – использовать один процессор.
ТРЕБОВАНИЕ И.Русинова
STDOUT содержит мин, макс, средн веса входных последовательностей относительно профиля
- `hmm2calibrate --cpu 1 my_alignment.hmm`
перезаписывает этот файл, уточняя некоторые константы модели
- `hmm2search --cpu 1 -T 200 my_alignment.hmm sequences.fasta`
ищет находки профиля среди `sequences`
`-T 200` – порог веса находки 200
Результат выдаётся в `stdout`. Важна 2я таблица с указанием `E-value`, веса, координат находки в профиле и в последовательности

HMMER3

Математическая модель

- Такова, что ВСЕ веса в профиле ≥ 0 .
- Нечто вроде информационного содержания.
- Точнее не объясняем, т.к. не можем просто объяснить математическую теорию, на которой основано вычисление весов эмиссии и транзиции
- В публикации “Sean R. Eddy and the HMMER development team. HMMER User’s Guide, 2023” Eddy, один из главных основателей технологии профилей, объясняет, что приписывание весов буквам в позиции и построение оптимального выравнивания последовательности с профилем недостаточно обоснованная идея. В HMMER3 рассматриваются “ансамбли выравниваний”. Ссылки на математическую теорию 20-25 летней давности есть. Но в них ещё разбираться.
- Вывод. Числа в hmm профиле HMMER3 не суммируются для получения веса выравнивания последовательности с профилем HMMER3
- Достоинство HMMER3
 - работает на два порядка быстрее HMMER2
 - Есть удобства в интерфейсе и выходных данных
- Недостатки
 - Трудно использовать, не понимая что и как происходит
 - Входное выравнивание должно быть в формате Stockholm .sto. Jalview умеет сохранять выравнивания в этом формате

MM	A	C	D	E	F	G	H
	m->m	m->i	m->d	i->m	i->i	d->m	d->d
COMPO	2.56202	4.09469	2.93753	2.62076	3.29263	2.91609	3.55988
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494
	0.26132	4.89778	1.50282	0.61958	0.77255	0.00000	*
1	2.12526	4.24902	4.56867	3.97633	2.65134	4.01181	4.34659
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494
	0.01430	4.65075	5.37310	0.61958	0.77255	0.85760	0.55196
2	2.75356	4.94348	2.96010	2.60289	3.79393	3.33778	3.79215
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494
	0.01398	4.67328	5.39563	0.61958	0.77255	0.75334	0.63637
3	2.56960	4.33401	4.77660	4.18394	3.42447	4.15874	4.50657
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494
	0.01334	4.71967	5.44202	0.61958	0.77255	0.83313	0.57037

Рис. Фрагмент файла bah-pf00145-hmm3.hmm, построенного по выравниванию bah-pf00145-revised.sto белков с двумя доменами: bah => pf00145

Основные программы HMMER3 (на kodomo)

- Пользователь строит множественное выравнивание с инделями
my_alignment.sto
- hmmbuild -n my_alignment-hmm3.hmm --cpu 1 my_alignment-hmm3.hmm
my_alignment.sto

Опции: -n -- name the HMM

-o <f> -- direct summary output to file <f>, not stdout

-O <f> : resave annotated, possibly modified MSA to file <f>

Проверил, немножко меняет выравнивание локально
Калибровка выполняется самой программой hmmbuild

- hmmsearch --cpu 1 -T 200 -A hmm-hits.sto --domtblout hit_table-hmm3
my_alignment.hmm sequences.fasta

ищет находки профиля среди sequences

-o <f> -- direct output to file <f>, not stdout

-A <f> -- save multiple alignment of all hits to file <f> Полезная штука

кажется.

--domtblout <f> -- save parseable table of per-domain hits to file <f>

содержит

координаты выравнивания в профиле и в
последовательности

--acc -- prefer accessions over names in output

-T <x> -- report sequences \geq this score threshold in output

Базовые задачи поиска в базах последовательностей белков

1. Найти белки, гомологичные данному
А что такое гомологичные белки?
2. Найти белки имеющие гомологичные участки
А могут быть гомологичные участки у негомологичных белков?
3. Найти консервативные мотивы связанные с функцией белков
Гомологичных: белков? участков?
Или любых, в том числе негомологичных белков?

Вспомним. Гомологию мы выводим из сходства последовательностей, которую нельзя объяснить случайностью

III. Домены

База данных Pfam

<http://pfam-legacy.xfam.org/>

Поглощена БД INTERPRO в 2022

ДОМЕН - Домены – единицы
непрерывной эволюции белков

Непрерывная эволюция это замены
остатков, небольшие делеции и вставки.

Домены можно обнаружить с помощью
выравнивания

Кроме непрерывной эволюции бывают
единовременные крупные изменения в
последовательностях белков

Так выглядит выравнивание белков, содержащих два домена:

гомеодомен (PF00046) и OAR(PF03826), не гомологичных по всей длине

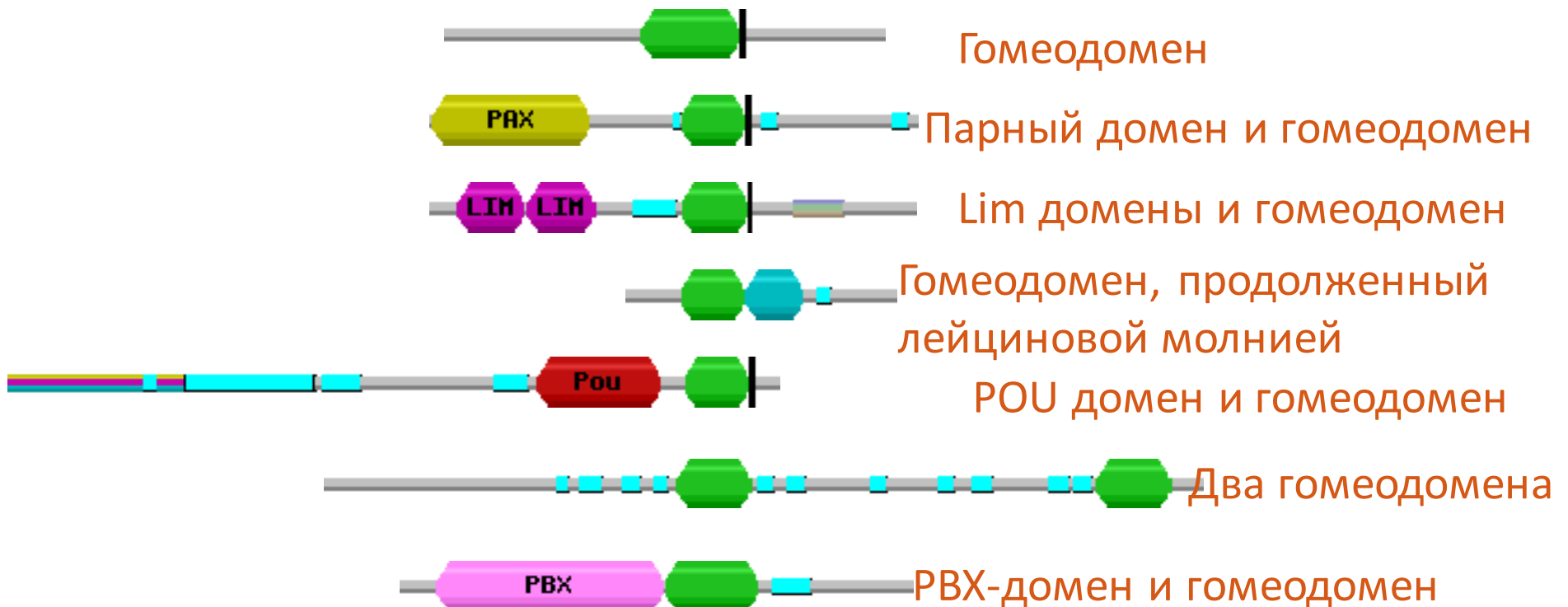
Выравнивание доменов в выравнивании негомологичных белков получается не всегда

		*	20	*	40	*	60	*	80	*	100	*	120	*	140	*	1
SW: PMX1_CHICK/1	:-----																: 80
SW: PMX2_HUMAN/1	:-----																: 86
SW: PMX1_HUMAN/1	:-----																: 80
SW: ARX_BRARE/1	: ISQAPQVISRSKSYREN-APFQS---D-ECQSP--EHMAQELVELST---																: 84
SW: ARX_MOUSE/1	: ISQAPQVISRSKSYRENCAPFVPPPPALD-ELSGPGVAHPERLSAASGPGSAPAACGGTGAEDDEEELLEDEDEEELLEDDEELLEDDARALLKPRRCVATTGTVAAAAAATAAATAAATGCGELSPKEELLHPEDAEKDCDSVCLS																: 157
SW: AL_DROME/1-1	:-----																: 72
SW: ALX4_MOUSE/1	: -TFLSAGAKQCGFCDAKSRARYGACQQDLAAPLESSGARGSFNMFQPQPPTPQP-----PPAPPAPAHLYLQRCACKTTPDGSNLKQBGSGGHHMAALQVPYAKESNLCGPELPDPSPVCGMDSYLVEKTKAGKQQDRASAEIPLS																: 145
SW: ALX4_HUMAN/1	: -TFLSAAAQAQCGFCDAKSRARYGACQQDLATPLESCAGARGSFNMFQPQPSTPPQPQPQPQPQPQPPAPHLYLQRCACKTTPDGSNLKQBGSGGSHAAALQVPYAKESNLCEPELPDPDSTVGMHDSYSLVKEACVKQPDRASSDLPSP																: 157
SW: RX2_CHICK/1	:-----																: 83
SW: RX2_BRARE/1	:-----																: 92
SW: RX1_XENLA/1	:-----																: 91
SW: RX_HUMAN/1-1	:-----																: 97
SW: FIX2_BRARE/1	:-----																: 66
SW: FIX2_HUMAN/1	:-----																: 68
SW: FIX1_HUMAN/1	:-----																: 64
SW: OTP_MOUSE/1	:-----																: 90
<hr/>																	
	60	*	180	*	200	*	220	*	240	*	260	*	280	*	300	*	
SW: PMX1_CHICK/1	: MDQLNSEE-----KRRRQRRRTTFNSSQALRRVRRERHYPDVFRVRLARRVNLRRVRRVWVFNRRRAKFRNRERLAMLASKNASLLKSYSQDVAVTAVQPIVPRPARPRTDYLWGTASPYSAMATYSTTCTCNAS-----																: 213
SW: PMX2_HUMAN/1	: GECPSPGRGS-----AAKRKRQRRRTTFNSSQALRRVRRERHYPDVFRVRELARRVNLRRVRRVWVFNRRRAKFRNRERLAMLASRSASLLKSYSQEA-AATIEQPVARPTALSPDYLWSTASSPYSTVPPYSPGSSGP-----																: 221
SW: PMX1_HUMAN/1	: MDQLNSEE-----KRRRQRRRTTFNSSQALRRVRRERHYPDVFRVRELARVNLRRVRRVWVFNRRRAKFRNRERLAMLANKNASLLKSYSQDVAVTAVQPIVPRPARPRTDYLWGTASPYSAMATYSTCTCNAS-----																: 213
SW: ARX_BRARE/1	: AGSDSEEG-----MLKRRQRRTTFTFTSYQEELELRARQRTHYPDVFTREELAMRLDLELRARVWVFNRRRAKFRNRERACVQAHPTGLFPPCPPLAAHPLSHYLEGCGFPFPHHPALESAMTAAAAAAAAPGLAPPNSSALPP-ATPLC																: 230
SW: ARX_MOUSE/1	: AGSDSEEG-----LLKRRQRRTTFTFTSYQEELELRARQRTHYPDVFTREELAMRLDLELRARVWVFNRRRAKFRNRERCAQTHPPGLFPPCPPLSAATHPLSPYLDASFPFPHHPALDSAMTAAAAAAAAPGLAPPNSSALPP-ATPLC																: 303
SW: AL_DROME/1-1	: DCEADBYA-----PKRRQRRTTFTFTSYQEELELRARQRTHYPDVFTREELAMKLGLELRARVWVFNRRRAKFRNRERFCVQAHPTGLFPPCPPLAAHPLSHYLEGCGFPFPHHPALESAMTAAAAAAAAPGLAPPNSSALPP-ATPLC																: 212
SW: ALX4_MOUSE/1	: EKDTSSEN-----KCRKRQRRTTFTFTSYQEELELRARQRTHYPDVYARQLAMRDLLELRARVWVFNRRRAKFRNRERFCGMQVQVRFTHFSTAYEPLLLTRAEINYAIQNPISWLGNMNCAASVPPACVVPDPPVACMSPHAPPSCASSVT																: 290
SW: ALX4_HUMAN/1	: EKADSEN-----KCRKRQRRTTFTFTSYQEELELRARQRTHYPDVYARQLAMRDLLELRARVWVFNRRRAKFRNRERFCGMQVQVRFTHFSTAYEPLLLTRAEINYAIQNPISWLGNMNCAASVPPACVVPDPPVACMSPHAPPSCASSVT																: 302
SW: RX2_CHICK/1	: KPSDEEQ-----PKKRQRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERLEVSSMKLQDSPILSFSRSPQAAPVGCALG-----CSLPLETGLCPVPPCG--AALQSLPGFAAPPQC																: 215
SW: RX2_BRARE/1	: PDIPDEDQ-----PKKRQRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERMDTCTMKLHDSPIRSFNRPPMAPNVGPMSS-----NSLPLDPLWSSPLSSA--TMMHSIPCFMCPQC																: 225
SW: RX1_XENLA/1	: KLSDEEQ-----PKKRQRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERMDTCTMKLHDSPIRSFNRPPMAPNVGPMSS-----NSLPLDPLWSSPLSSA--TMMHSIPCFMCPQC																: 224
SW: RX_HUMAN/1-1	: KLSDEEQ-----PKKRQRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERMDTCTMKLHDSPIRSFNRPPMAPNVGPMSS-----NSLPLDPLWSSPLSSA--TMMHSIPCFMCPQC																: 242
SW: FIX2_BRARE/1	: SKNEDSN-----DDPSKRRQRRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERMDTCTMKLHDSPIRSFNRPPMAPNVGPMSS-----NSLPLDPLWSSPLSSA--TMMHSIPCFMCPQC																: 212
SW: FIX2_HUMAN/1	: GKNEDVCA-----EDPSKRRQRRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERMDTCTMKLHDSPIRSFNRPPMAPNVGPMSS-----NSLPLDPLWSSPLSSA--TMMHSIPCFMCPQC																: 215
SW: FIX1_HUMAN/1	: KCPEDSGAGGTCCCGADDPAKRRQRRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERMDTCTMKLHDSPIRSFNRPPMAPNVGPMSS-----NSLPLDPLWSSPLSSA--TMMHSIPCFMCPQC																: 218
SW: OTP_MOUSE/1	: NPSQACQQ-----QCQQRRRRTTFTFTSYQEELELRARQRTHYPDVYSREELAMKVMLELRARVWVFNRRRAKFRNRERMDTCTMKLHDSPIRSFNRPPMAPNVGPMSS-----NSLPLDPLWSSPLSSA--TMMHSIPCFMCPQC																: 236
k 4 4R RT Ft QL eLE F 4 hYPD RE 6A L E R6qVWFqNRRAK544 e4																	
<hr/>																	
	320	*	340	*	360	*	380	*	400	*	420						
SW: PMX1_CHICK/1	:-----																: 245
SW: PMX2_HUMAN/1	:-----																: 253
SW: PMX1_HUMAN/1	:-----																: 245
SW: ARX_BRARE/1	: LGTFGLTAMFRHPAFICPTFGLRFSMCPPLTSAATAAALLRQTPAPPVESPVQAALPEPPSSSSSTADRRASIAALRRKREHSA-QLTQLNLLPSTACKEVC-----																: 336
SW: ARX_MOUSE/1	: LSTFLCAAFFRHPAFISPAFCRLFTMAPLTSAATAAALLRQTPAVECAVASGALADP-----ATAAADRRASIAALRRKREHSAQAQLTQLNLLPCTSTCKEVC-----																: 404
SW: AL_DROME/1-1	: PPTSASGHAXPQLVGIALTQQASSLPT---QTSFVALTSHSPQRQLPPSHQAQPPPPRAATTTPEDRRTSIAALRRKREHSLKLELRLQMGCMQDV-----VS																: 313
SW: ALX4_MOUSE/1	: DFL-----SVSCAGSHVCGQTHMCSLFGAAGISPCNGLYEMNCPDRKTSIAALRRKREHSAATSWAT-----																: 354
SW: ALX4_HUMAN/1	: DFL-----SVSCAGSHVCGQTHMCSLFGAAGISPCNGLYEMNCPDRKTSIAALRRKREHSAATSWAT-----																: 366
SW: RX2_CHICK/1	: LPASYTPPPFFL-----NSPAVTHALQPLGAMCPPPPYQCCAAAFVQKFLDCEGDPNTSIAALRRKREHISTQIGKWPQIT-----																: 290
SW: RX2_BRARE/1	: LQPTYTAHPGFL-----NTSPGMQMNIQPM---PPPPYQCPVFNDKYLEVD---RSSIAALRRKREHISTQIGKWPQIT-----																: 297
SW: RX1_XENLA/1	: LQPSYTPPPFI-----NPPSVCHALQPLGAMCPPPPYQCCAAAFVQKFLDCEGDPNTSIAALRRKREHISTQIGKWPQIT-----																: 296
SW: RX_HUMAN/1-1	: LPASYTPPPPPFFL-----NSPPLGPGQLPL---APPVSYPCGCGFDKFLDCEGDPNTSIAALRRKREHISTQIGKWPQIT-----																: 319
SW: FIX2_BRARE/1	: SSSMSMSSSMVPSAVTGVPGSSL-----NSLNNLNNLSMPSLNSGVPTACPYAPPTPPY-VYRDTCNSSLASLELRARQHSDFCYASVQNPASNLACQYAVDRPV																: 314
SW: FIX2_HUMAN/1	: SSSMSMSSSMVPSAVTGVPGSSL-----NSLNNLNNLSMPSLNSGVPTACPYAPPTPPY-VYRDTCNSSLASLELRARQHSDFCYASVQNPASNLACQYAVDRPV																: 317
SW: FIX1_HUMAN/1	: SSSMTPSSMCPAVPGMPNSGL-----NNIN---MLTSSLNSMSPGACPYCTPASPYSVYRDTCNSSLASLELRARQHSDFCYASVQNPASNLACQYAVDRPV																: 314
SW: OTP_MOUSE/1	: SQCSLAACPPPNMCLNSLACSNAGLQ---SHLYQPAFPGMVPSLPCPSNVSCSGLCSSPSSDVWRCTSLASLELRARQHSDFCYASVQNPASNLACQYAVDRPV																: 325

В эволюции гомеодомены *Homeodomain* (PF00046)

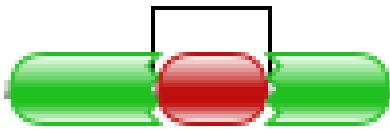
включались в разные архитектуры

Об этом можно судить по 1618 различным доменным архитектурам гомеобелков, представленным в банке Pfam



Типы объектов кроме доменов в Pfam

Domains of unknown function (DUFs)

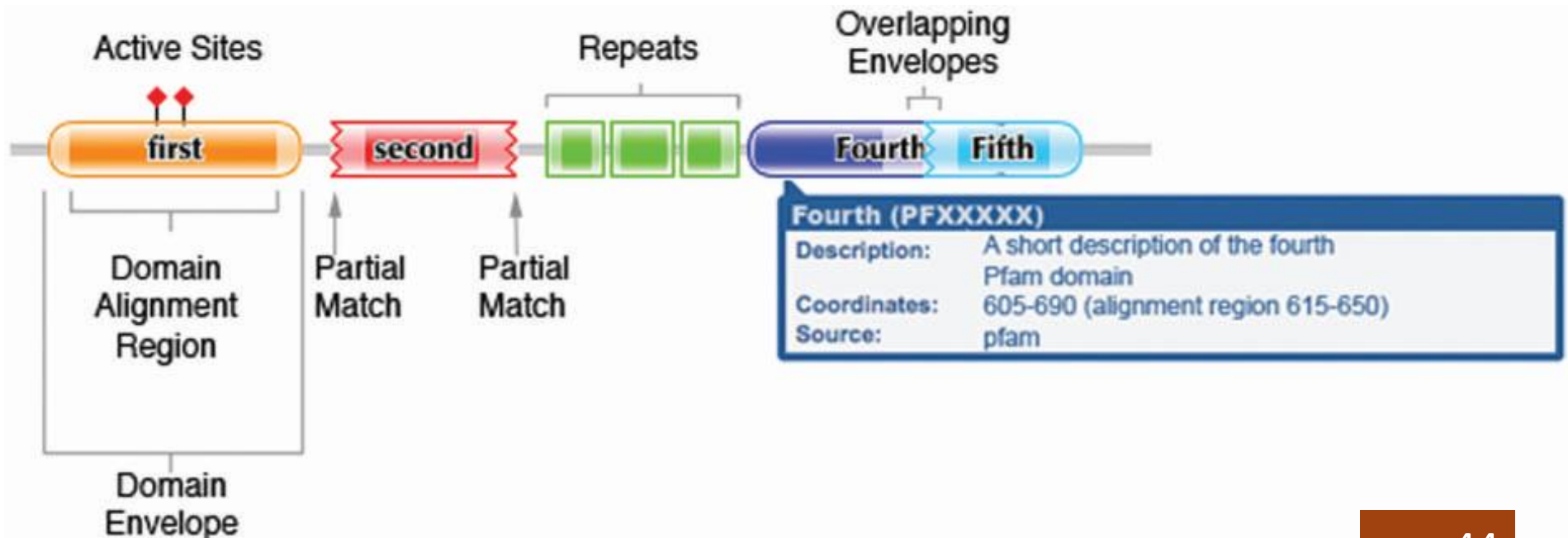


Язык Pfam :

Семейство – коллекция гомологичных доменов из разных белков.

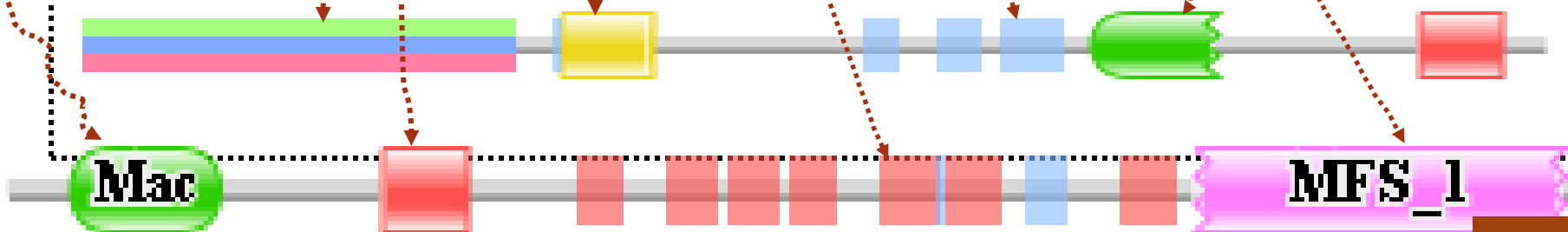
Домен – структурная единица, которую можно найти во множественном выравнении.

Повтор – короткая единица, нестабильная сама по себе, но образует стабильные структуры, если есть много копий.



Какая информация закодирована в картинке из Pfam, изображающей доменную архитектуру белка

- Прямоугольники с гладкими краями – найден домен целиком.
- Край прямоугольника зубчатый – найден только фрагмент домена, за зубчиками домен не продолжается, хотя должен был быть.
- Прямоугольник с острыми краями – мотив, трансмембранный участок, участок малой сложности (например, десять остатков A) и т.п. – не является эволюционным доменом!
- Домен, имеющий ID вида DUF... с номером – Domain of Unknown Function



БД Pfam

- Единица хранения – семейство гомологичных доменов. Говорят «домен», отождествляя его с семейством
- Идентификаторы ID (напр. Pterin_bind), AC (PF00809), название домена (Pterin binding enzyme)
- Описание функции домена (не всегда), ссылки на литературу
- Ссылки на 3D структуры домена, если есть расшифровки
- Множества последовательностей содержащих домен, их последовательности
- Seed alignment – это выравнивание, по которому составлен профиль домена.
- Профиль домена
- Доменные архитектуры, в которых встречается домен
- Распределение белков с доменом по таксонам разного уровня

Сервис Pfam позволяет показать доменную архитектуру последовательности, скачать многие файлы, составляющие базу данных

Задание на дом:

Создать HMM-профиль подсемейства семейства белков с выбранным доменом и проверить его работу на всех белках семейства.

Выделить подсемейство можно

- По доменной архитектуре (рекомендуется)
- По таксономии
- Как кладу в выравнивании SEED
- Ещё как-нибудь

Конец презентации

Далее задание для работы в классе

Задание на «занятии»

1. Выберите домен домен Pfam и подсемейство семейства белков с этим доменом. Рекомендуется выбрать белки с определенной доменной архитектурой.
2. Заполните форму , ссылка стоит на сайте под практикумом 11

Если не хватило времени – выполните задание сегодня до вечера. Возможно редактирование формы после заполнения

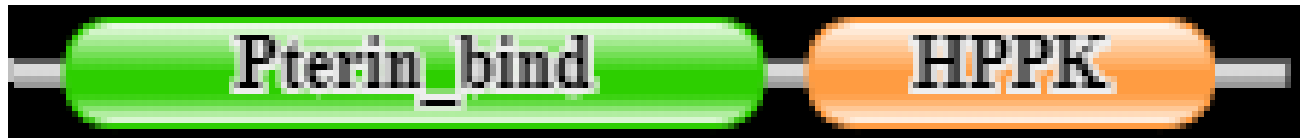
Пример крупной перестройки в эволюции.

Гомологичны ли эти 41 + 9 белков?

There are 41 sequences with the following architecture:

Pterin_bind, HPPK

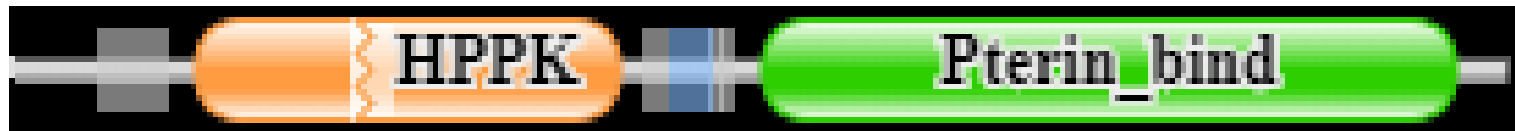
[R9KWZ5_9ACTN](#) [Enterorhabdus caecimuris B7] Dihydropteroate synthase {ECO:0000313|EMBL:EOS50736.1} (437 residues)



There are 9 sequences with the following architecture:

HPPK x 2, Pterin_bind

[G2XU66_BOTF4](#) [Botryotinia fuckeliana (strain T4) (Noble rot fungus) (Botrytis cinerea)] Similar to folic acid synthesis protein CO:0000313|EMBL:CCD44036.1} (541 residues)



Выравнивание гомологичных доменов из разных белков. Пример из БД PFAM семейств доменов (фрагмент)

Seed sequence alignment for PF00809

Family: *Pterin_bind* (PF00809)

```

Q9X8H8_STRCO/24-269      MGVMNVT PDSFSDGGGRF.FDTTAAIKHGLDLVAQGAADLVVGGESTRPGA..TRVDEEELRRVVPVVRGLAS.
DHPS1_MYCLE/9-255       IGVLNVT DNFSDGGGRY.LDPDDAVQHGLAMVAEGAIVDVGGESTRPGA..IRTPRVELSRIVPVVKELAA.
DHPS1_MYCTU/9-255       MGVLNVT DDFSDGGGCY.LLDDAVKHGLAMAAAGAGIVDVGGESSRPGA..TRVDPAVETSRIVPVKELAA.
DHPS1_MYCTU/9-255 (SS)  EEEEE-S--TT-SS---S-#####TT-SEEEE-----#####
DHPS_STRR6/13-284       CGIINVT PDSFSDGGGF.FALEQALQQARKLIAEGASMLDIGGESTRPGS..SYVEIEEIIQRVVPVIKAIK.
DHPS_STRR6/13-284 (SS)  EEEEE-----#####CT-SEEEE-----#####
DHPS2_MYCTU/45-289      MAIVNRT PDSFYDKGAT.FSDAAARDAVHRAVDGADVVDVGGVKAQPG...ERVDVDEITRLVPIEILRG.
DHPS2_MYCTU/45-289 (SS)  EEEEE-----#####TT-SEEEE-----#####
Q2G0Q7_STAA8/7-241      MGILNVT PDSFSDGGKF.NNVESAINRVKAMIDEGADIIDVGGVSTRPGH..EMVSLSEEMNRVLPVVEAIVG.
FOLM_ARATH/276-531      MGILNLT PDSFSDGGKF.QSIDSASRVRSMISEGADIIDIGAQSTRPMA..SRISSQEELDRLLPVLEAVRGM
FOLKP_CHLTR/183-431     MGIVNIT DNFSDTGLF.LEARRAAHAERLFAEGASIIDLGAQATNPRV.KDLGSVEQEWERLEPVLRLLAER
M4R6K4_BIBTR/79-320     FGIVNIT SDSFDGGRY.LAPDAAIAQARKLMAEGADVIDLGPASSNPDA..APVSSDTEIARIAPVLDALKA.
Q6NFE5_CORDI/9-252      GERINGMFGDIKRAIQE.RDPAPVQEWARRQEEGGARALDLNVGPA.....VQDKVSAMENLVEVTQ.....
Q2RJ78_MOOTA/5-228     GERLNAT GSKRFREMLFARDLEGLALAREQVEEGAHALDLSVAWT.....GRDELEDLRWLLPHLA.....
Q5SKM5_THET8/372-605   GERLNAT GSKRFREMLFARDLEGLALAREQVEEGAHALDLSVAWT.....GRDELEDLRWLLPHLA.....
Q5SKM5_THET8/372-605 (SS)  EEEEETTT-#####TT-#####TT-SEEEE---T.....TS-#####
METH_CAEEL/364-602     GERCNVA GSRRFCNLIKNENYDTAIDVARVQVDSGAQILDVNMDG.....LLDGPYAMSKFLRLISSEPD.
METH_RAT/363-601       GERCNVA GSKKFAKLIMAGNYEEALSVAKVQVEMGAQVLDINMDG.....MLDGPAMTKFCNFIASEPD.
METH_ECOLI/360-598     GERTNVT GSAKFKRLIKEEKYSEALDVARQVENGQAQIIDINMDEG.....MLDAEAMVRFNLNLIAGEPD.
Q9RVQ6_DEIRA/372-610   GERTNVT GSPKFSKAILAGDYDAGLKIARQVNTNGAQIVDINFDEG.....MLDGEGAMVKFLNLLAGEPD.
METH_MYCLE/354-590     GERTNANGSKVFREAMIADYQKCLDIAKDQTRGGAHLLDLCVDYV.....GRNGVADMKALAGRLA.....
METH_SYNY3/344-576     GERLNAS GSKKCRDLLNAEDWDSLVS LAKSQVKEGAQILDVNVDYV.....GRDGVDRMKELASRLV.....
Q9RX6_DEIRA/36-273     MGILNAT PDSFSDGGQH.LQLDAALATARRMRDVGFIIDIGGESTRPGA..EPVDAATELDRVLPILIRALRG.
DHPS_NEIMB/21-266      MGIVNLT PDSFSDGGVYSONAQTALAHAEQLLKEGADILDIGGESTRPGA..DYVSPPEEWARVEPVLAEVAG.
DHPS_HAEIN/18-257     MGILNFT PDSFSDGGQF.FSLDKALFQVEKMLEEGATIIDIGGESTRPGA..DEVSEQUELHRVVPVEAVRN.
DHPS_ECOLI/18-257     MGILNVT PDSFSDGGTH.NSLIDAVKXANLMINAGATIIDVGGESTRPGA..AEVSEEEELQRVIPVVEAIAQ.
DHPS_ECOLI/18-257 (SS)  EEEEE--TTTIIIIIS.T#####T-SEEEESS--STT-#####
Q9WXP7_THEMA/19-258    MGIINVT PDSFFADSRK.QSVLEAVETAKMIEEGADIIDVGGMSTRPGS..DPVDEEEELNRVIPVIRAIRS.
DHPS_HELPY/122-361     MAVLNLT PDSFYEKSRF..DSKKALEEIQWLEKGITLIDIGAASSRPES..EIIDPKIEQDRLKEILLEIKSQ
O67448_AQUAE/129-378   MGVLNVT PDSFSDGGGF.LEPKKAVERAVKMAQEGAEIIDIGGESTRPGS..KRISAEELNRVLPALKEVRR.
FOL1_SCHPO/468-714     MGILNVT PDSFSDGGKV..SQNNILEKAKSMVGDGASILDIGGSTKPGA..DPVSVEEELRRVPMISLLRS.
B6KBG5_TOXGV/447-710   MGILNVSPDSFTD..HFSASVDEAVAAAEAMVTDGADVVDVGGVATNPFVAGGEVPLAVERERVVQKILD.
DHPS_SYNY3/31-272     MGILNTP PDSFSDGGGF.NSLPTAIHQAKTMVQGGAHIIDIGGSTRPGA..ETVSLKEELERTIPPIIQLRQ.
DHPS_BACSU/28-261     MGILNVT PDSFSDGGKY.DSLDKALLHAKEMIDDGAHIIDIGGESTRPGA..ECVSEDEEMSRVPIVIERITK.
C5B125_METEA/26-262    MGILNVT PDSFSDGGRF.EGVDAARAQAAALTEGAHILDIGGESTRPGH..TPVPAEEQARVLPVIEAVAP.
    
```

**Seed
(30)**

Pterin binding enzyme

This family includes a variety of pterin binding enzymes that all adopt a TIM barrel fold. The family includes

