HMM-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
GTATTAA

выравнивание

	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

PWM для данного выравнивания

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

$$I = \sum_k I_k$$

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

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)

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

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

$$f_{ki} = \frac{\sum\limits_{s:a_{sk}=i} w_s + \psi_i}{\sum\limits_{s} w_s + \sum\limits_{i} \psi_i}$$

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

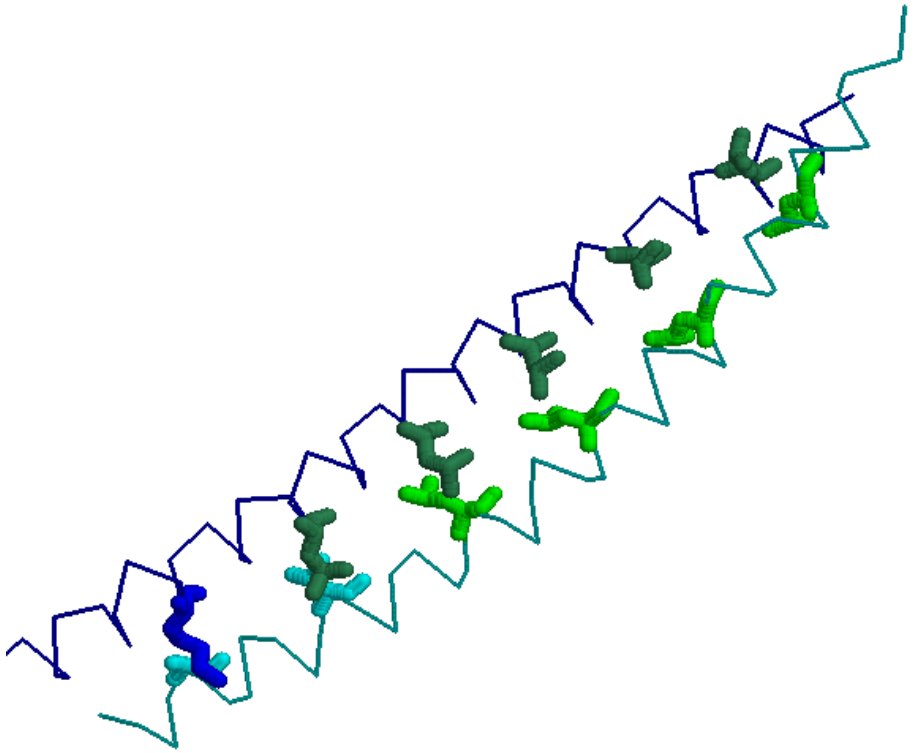
- [a-zAZ] –все возможные аминокислоты в данной позиции
- X(2,4) – любая аминокислота от 2 до 4 раз
- X(3) – любая аминокислота ровно 3 раза
- {P} – любая аминокислота, кроме пролина

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



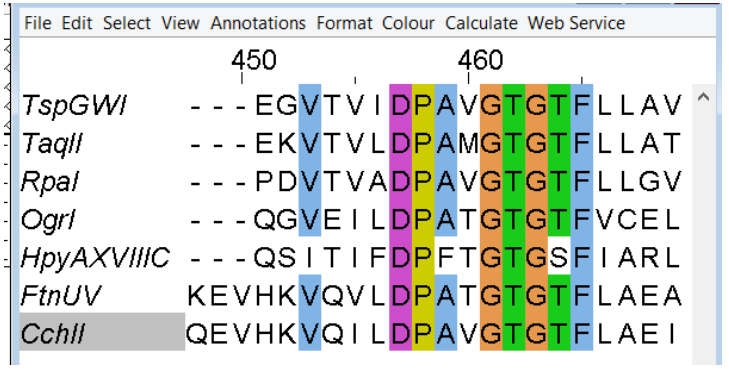
Показаны каждый 7й остаток цепей A и B; Leu - зеленые (A) и темнозеленые (B)

PDB код 1ci6

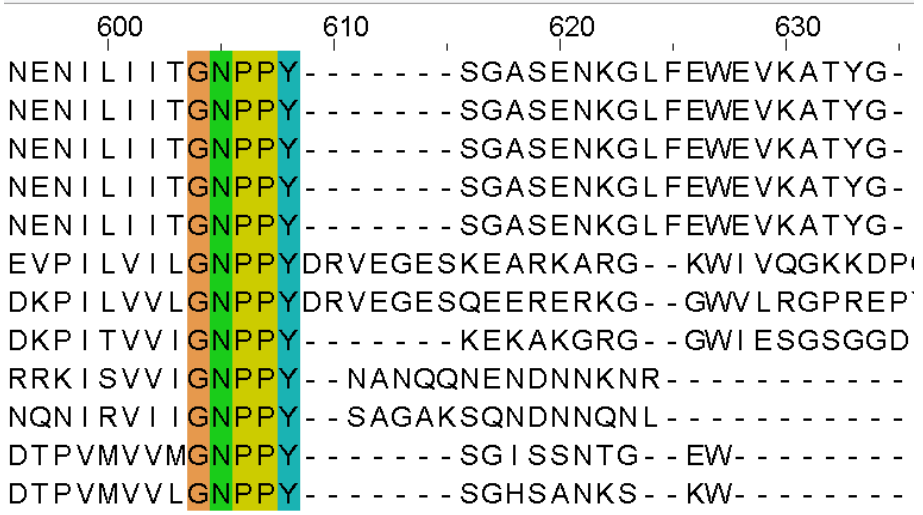
Паттерны в Jalview (Select => find)

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 of 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 DELitions) в выравнивании. Этим профили отличаются от PWM и PSSM

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

HMM (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 нынче переформатированы в DUF.
2. Строит HMM профиль с помощью пакета HMMER. Программа hmmbuild
3. Калибрует профиль на случайных последовательностях для нормализации веса и E-value последовтельностей (в HMMER3 входит в hmmbuild)
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

- ttps://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.AN....L...RAAI....EAL....D..............AG.
S0EYB2_CHTCT/11-144         YLGLGSSL.........G..........D.R...L.QN....L...QKAL....QRL.....................
C0ZID7_BREBN/6-134          YLALGSN....L.....G..........D.R...A.QN....L...RRAI....QRL....NE..............QP
Q5WLU7_BACSK/5-133          YIALGSN....V.....G..........D.R...E.NY....L...QEAM....KLL....DA..............DA
E6TSF5_BACCJ/10-138         YLSLGSN....I.....E..........S.R...Y.DY....L...TFAL....KKL....RE..............NP
G2THL3_BACCO/6-134          YLALGSN....I.....E..........P.R...F.DY....L...QHAI....RLL....RN..............NP
K0J162_AMPXN/5-133          YIALGSN....I.....N..........P.R...N.EF....L...EQAI....NEI....EQ..............IK
I0JH59_HALH3/5-133          YIALGSN....I.....S..........K.R...E.EF....L...ENAV....ASI....DD..............HP
Q8EU11_OCEIH/5-133          YVALGTN....I.....E..........P.R...E.NF....I...NQAL....QFL....DD..............HP
B7GFK5_ANOFW/6-134          YIALGSN....I.....G..........D.R...F.EY....L...CKAV....IAL....RD..............HT
Q5L443_GEOKA/6-134          YLALGSN....L.....G..........D.R...V.SY....L...RSAI....EAL....HH..............HQ
C5D399_GEOSW/6-134          YIALGSN....I.....G..........D.R...L.YY....L...REAV....KML....DR..............HE
Q65PE2_BACLD/6-134          YIALGSN....I.....G..........R.R...E.EY....L...KKAV....SLL....HQ..............HP
HPPK_BACSU/6-134            YIALGSN....I.....G..........D.R...E.TY....L...RQAV....ALL....HQ..............HA
A8F946_BACP2/6-134          YIALGSN....I.....G..........K.K...E.TY....L...KEAV....KKL....HE..............HP
Q81VW6_BACAN/6-134          YIALGSN....I.....G..........E.R...Y.TY....L...TEAI....QFL....NK..............NP
Q9KGG7_BACHD/6-134          YIALGSN....I.....G..........D.R...S.RF....L...EEAI....QQL....AE..............HD
D3FR36_BACPE/6-134          YIALGSN....I.....G..........D.R...A.AY....L...EEAI....DRL....DK..............EE
N0ATU2_9BACI/7-135          YLSIGSN....M.....G..........D.R...F.YY....L...KNAI....QLL....TN..............EK
U5L4K3_9BACI/6-134          FIALGSN....M.....G..........D.R...A.AN....L...KEAI....QML....SE..............HP
H6NSD7_9BACL/18-146         YIGLGSN....L.....G..........D.R...E.QY....L...KEAL....RML....EE..............HP
L0EIN8_THECK/16-144         YIALGSN....L.....G..........D.R...E.AQ....L...AEAL....RRL....HA..............RD
D3E785_GEOS4/13-141         YIALGAN....L.....G..........D.R...E.GN....L...MEAL....ERL....DE..............VP
E3EET6_PAEPS/13-141         YIALGAN....L.....G..........E.R...E.HT....L...YEAI....TAL....DE..............HP
X4ZBV9_9BACL/13-141         YIALGAN....L.....G..........D.R...E.QS....L...KEAL....TLL....NA..............HE
C6CRP5_PAESJ/14-142         YIALGSN....L.....N..........D.R...E.EL....L...QQAV....EHL....RQ..............QS
C4KZT0_EXISA/3-130          YIALGANI.........G..........D.R...A.GQ....L...SAAI.....DE....ME..............RT.
B1YGR6_EXIS2/5-133          YIALGSNI.........G..........D.K...A.GH....L...RAAI.....EA....MR.....................
E6U3M2_ETHHY/10-137         YIALGSN....M.....G..........D.R...A.GY....L...EAAR....KKI....AE..............S
I0IE19_PHYMF/13-147         WVALGSNL.........G..........D.R...G.AHL.......LAAC....RRLA.AAPG.....................
C9RLK0_FIBSS/8-134          YIALGSNL.........P..........D.R...S.AH....L...KAGR....DML....HR.....................
K4LLB0_THEPS/7-135          FLSLGSN....L.....G..........N.R...S.AY....L...EAAC....REL....AA..............HP
L7VQA6_CLOSH/5-133          ILSLGSNI.........G..........D.R...E.KN....L...KTAL....YHI....IQ..............NP.
A3DIK4_CLOTH/6-134          FLSLGSN....I.....E..........D.R...E.KY....L...LDAI....DNI....SA..............VS
G8LSW4_CLOCD/5-133          FLSLGSN....L.....G..........D.R...E.KY....L...FEAV....DEI....SK..............IP
D9QRZ5_ACEAZ/5-133          YLSLGSN....K.....E..........S.R...E.EY....L...QRAI....KKL....QD..............HS
E4RM72_HALHG/5-133          FLGLGSNI.........E..........P.R...S.EY....L...KKAA....AEL.....................
F0SWA2_SYNGF/4-132          FLGLGSN....L.....G..........D.R...R.SY....L...KKAV....RML....KE..............RS
F4LQD8_TREBD/64-196         VLGLGSNR.........S.............FGLLS...SAEIL...RDAC....AQLSGRISSL.....................
F2NVX4_TRES6/5-137          VLGLGSNK.........S.............FGAFS...SLELL...KRAC....SCLADFIHGL.....................
F8F3E4_TRECH/9-138          VLGLGSNQ.........G.................E...SRTIL...QHAI....TDLESRIQDL.....................
F5YC59_TREAZ/5-134          VLGLGSNQ.........G.................D...SLRIL...EKAV....EVLGIILGSL.....................
F2F163_SOLSS/6-134          YLSIGTN....I.....G..........E.R...E.QN....L...QDAV....KLL....TA..............HE
Q8YAC0_LISMO/5-133          FLSIGTN....I.....G..........E.R...L.EN....L...NDAL....RGL....AA..............SNQ
Q2G0Q5_STAA8/5-133          YLGLGSN....I.....G..........D.R...E.SQ....L...NDAI....KIL....NE..............YD
Q2G0Q5_STAA8/5-133 (SS)     EEEEEE-....S.....S..........S.I...I.HH....I...HHHH....HHH....HH..............ST
Q5HRN8_STAEQ/5-133          YLGLGSN....I.....G..........N.R...E.LQ....L...NEAI....KIL....HD..............YQ
Q18BX4_PEPD6/5-133          YLGIGTN....M.....G..........D.R...F.DN....L...SRAC....ELL....KN..............SD
```

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. AAл

HMM 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

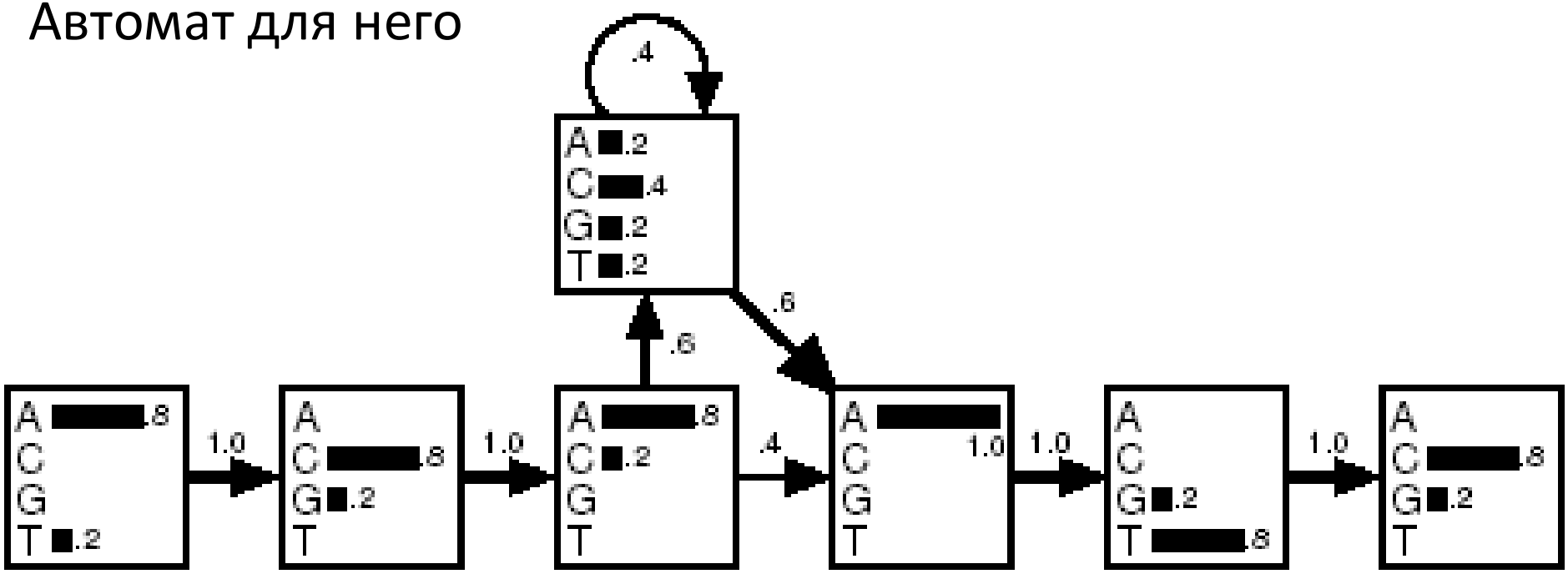


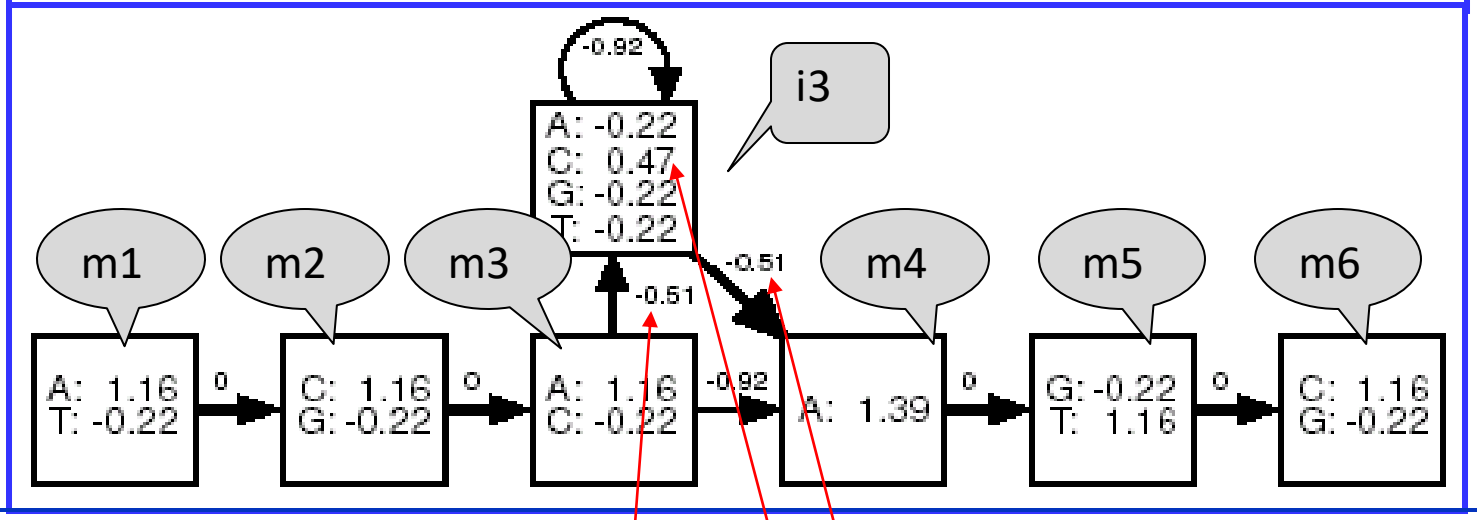
Fig. from Krogh, "Computational Methods in molecular biology, pages 45-63, Elsevier, 1998.

Вес выравнивания последовательности ACACATC с профилем

Вместо вероятностей в профиле используют логарифмы отношения правдоподобия

log2(частота буквы в колонке/базовая частота буквы)

	Sequence	Probability ×100	Log odds
Consensus	A C A C - - A T C	4.7	6.7
Original sequences	A C A - - - A T G	3.3	4.9
	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

Log odds = 1.16 + 0 + 1.16 + 0 + 1.16 + +1.39 + 0 +1.116 + 0 + 1.16 = **6.64**

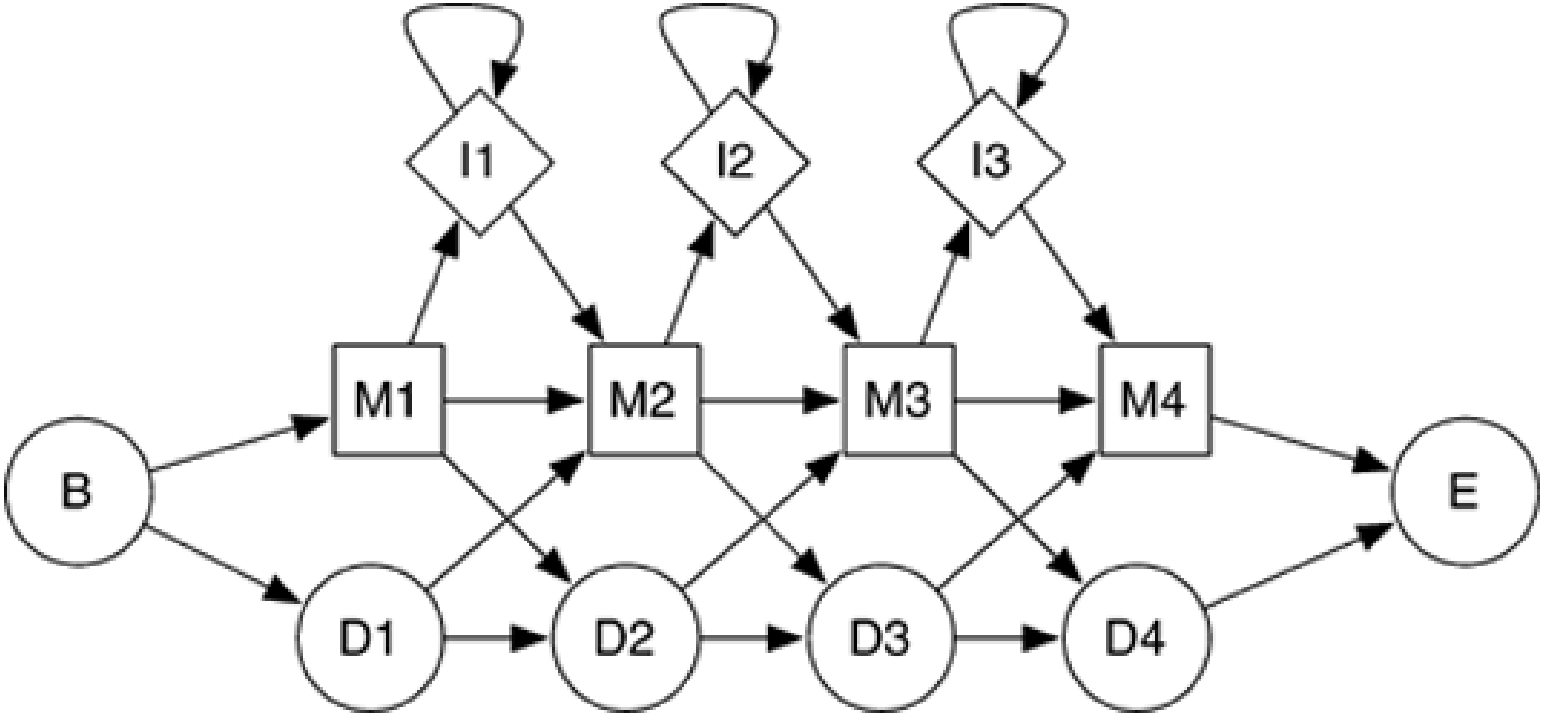
Мы нашли

- 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 HMM профиля решает **алгоритм Viterbi**

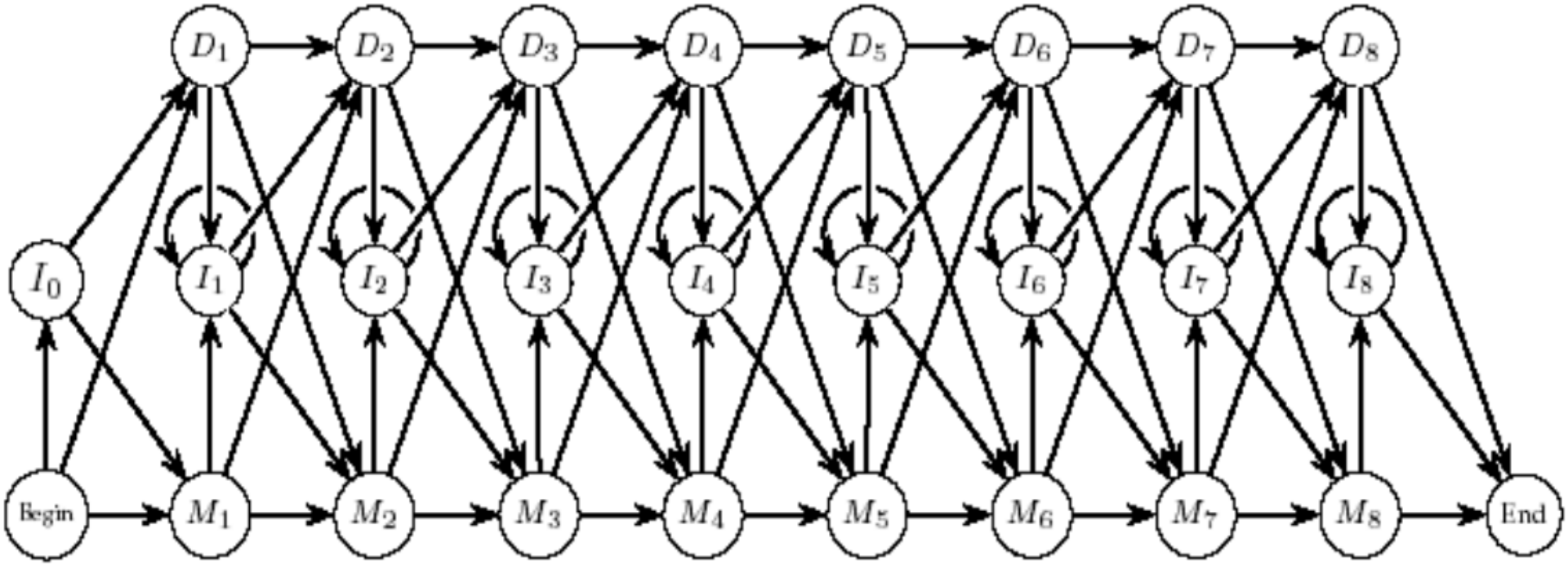
Семь типов транзиций (не считая начала и конца)
Замены (M), Вставки (I), делеции (D)

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



Из презентации безымянного сотрудника ИППИ)

HMM Профиль – описание автомата HMM в файле.

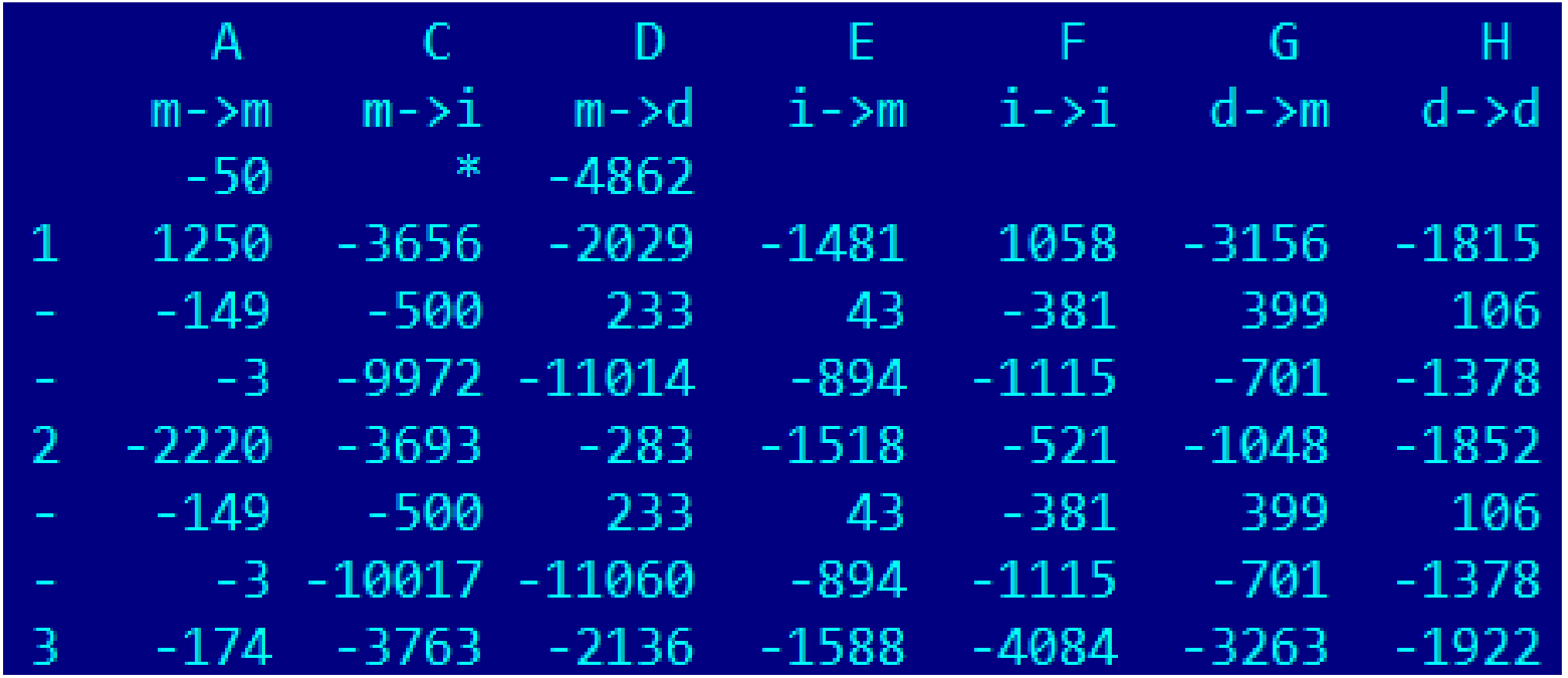
- HMM профиль имеет строгий формат
- По HMM-profile можно нарисовать автомат, по автомату можно создать HMM-profile
- РАСШИРЕНИЕ файла .hmm
- два основных пакета для работы с hmm-профилями:
  - HMMER2 старый – существует с конца 1990х, развивается, актуален и сейчас
  - HMMER3 новый - существует с 2010х, 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 A в первой позиции примера равно 0.8/0.25 = 3.2. Логарифм ln 3.2 = 1.16
- Log-odds >> 0 – за то, что буква A не случайно похожа на колонку выравнивания
- Log-odds ≈ 0 – за то, что буква A соответствует случайному выбору
- Log-odds << 0 – за то, что буква A 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 третей строке стоят веса за все 7 типов transitions. Они разнятся в позициях

HMM профиль, построенный HMMer2'ом

log-odds(эмиссионных вероятностей для **m**)

log-odds(эмиссионных вероятностей для **i)**

log(вероятностей переходов

	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**		
	-126	*	-3585								
1	**-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**	*****	*****		

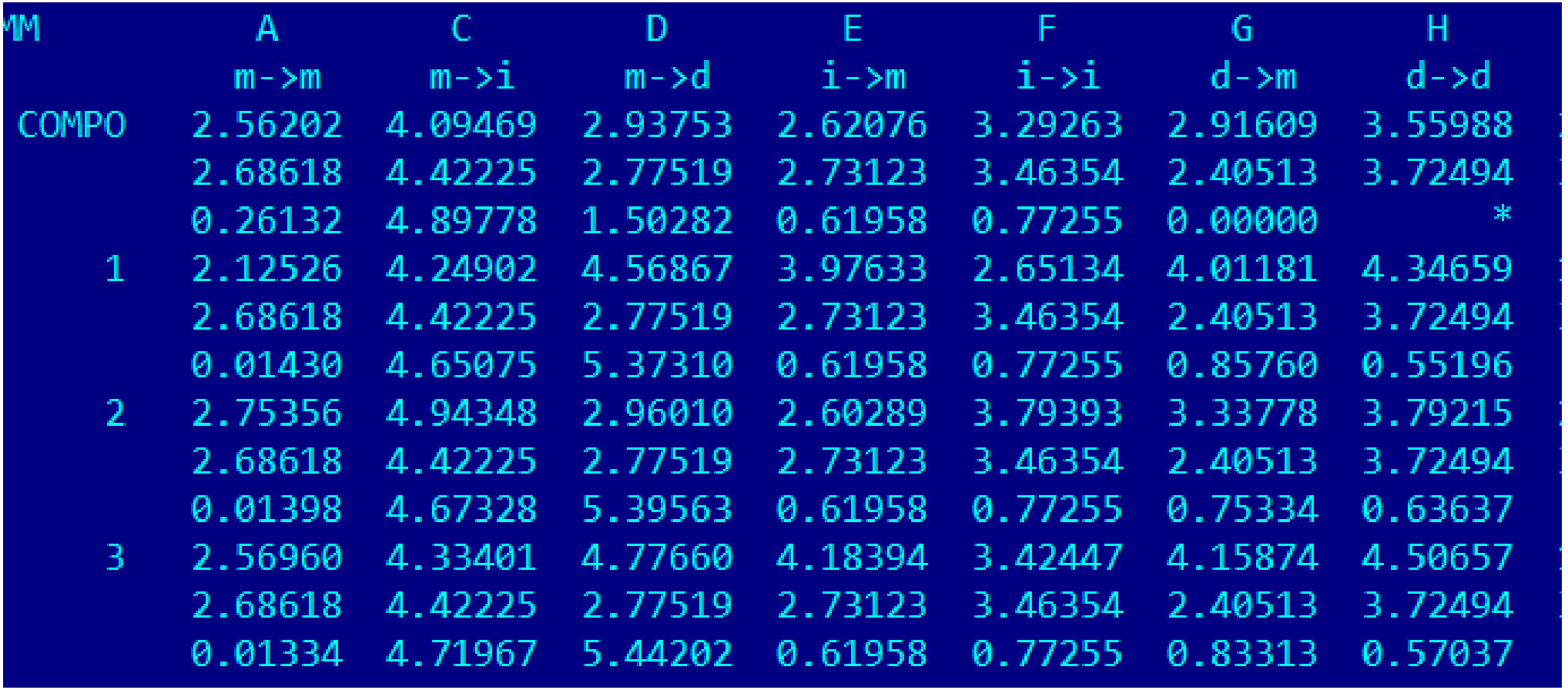
Основные программы HMMER2 (на kodomo)

- 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 – использовть 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 выравнивния 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 Stokholm .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 >= 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

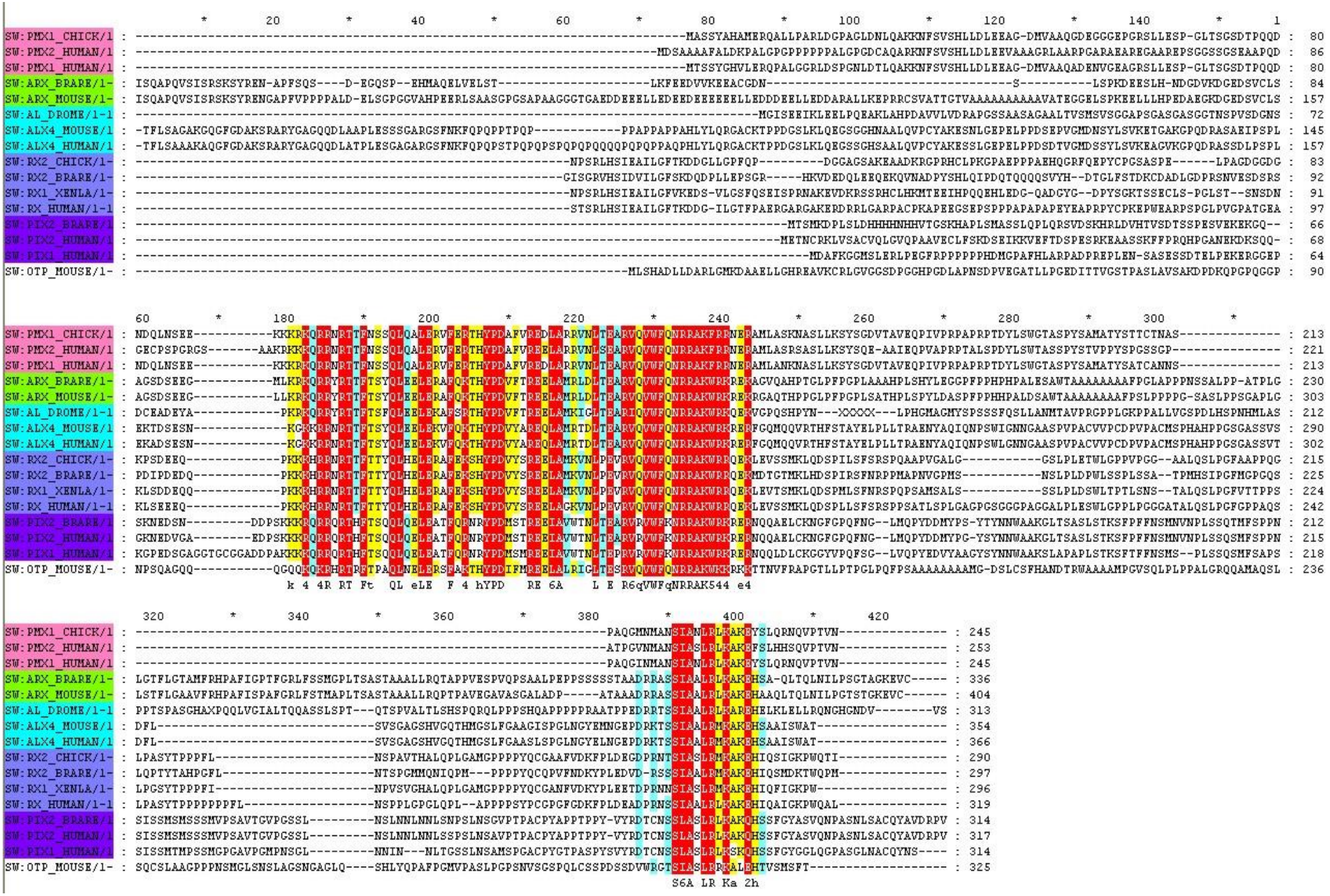
Непрерывная эволюция это замены остатков, небольшие делеции и вставки.

Домены можно обнаружить с помощью выравнивания

Кроме непрерывной эволюции бывают единовременные крупные изменения в последовательностях белков

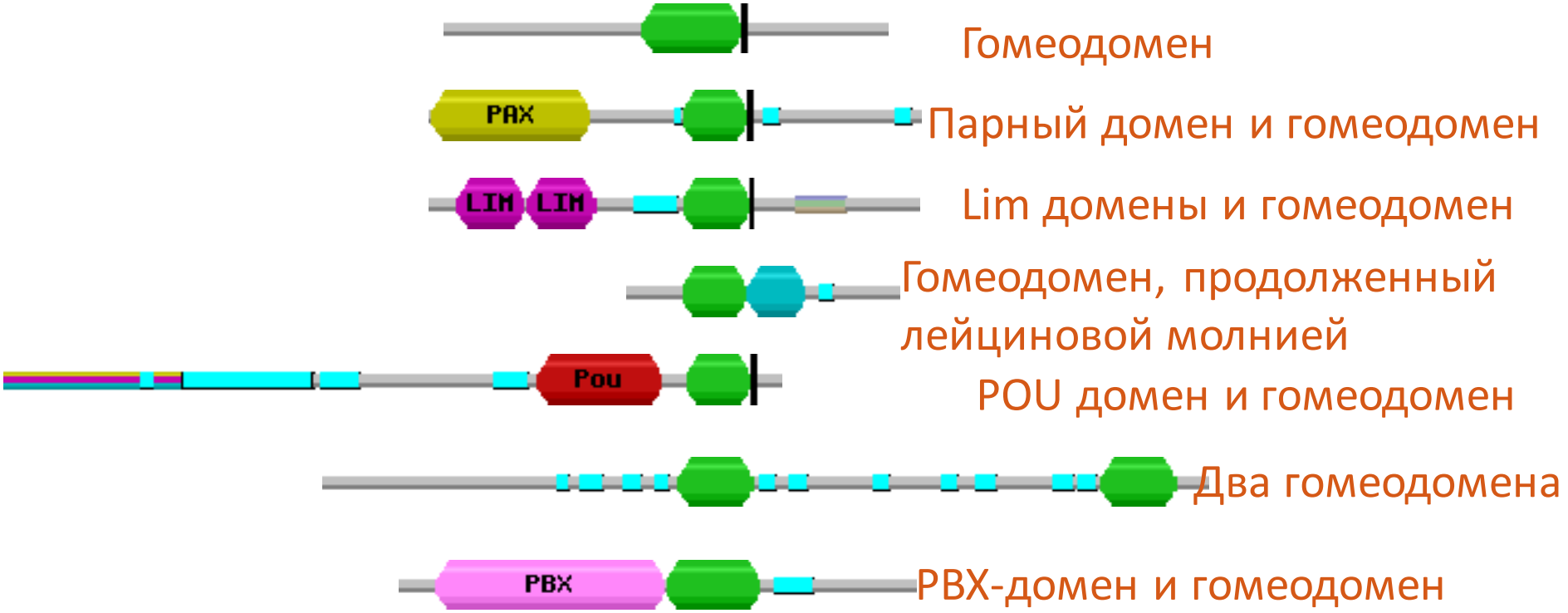
Так выглядит выравнивание белков, содержащих два домена: гомеодомен (PF00046) и OAR(PF03826), не гомологичных по всей длине

Выравнивание доменов в выравнивании негомологичных белков получается не всегда



В эволюции гомеодомены ***Homeodomain*** **(PF00046)** включались в разные архитектуры

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



Гомеодомен

Парный домен и гомеодомен

Lim домены и гомеодомен

Гомеодомен, продолженный лейциновой молнией

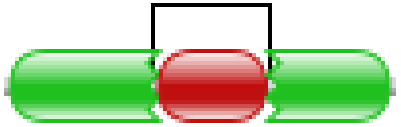
POU домен и гомеодомен

Два гомеодомена

PBX-домен и гомеодомен

Типы объектов кроме доменов в Pfam

Domains of unknown function (DUFs)

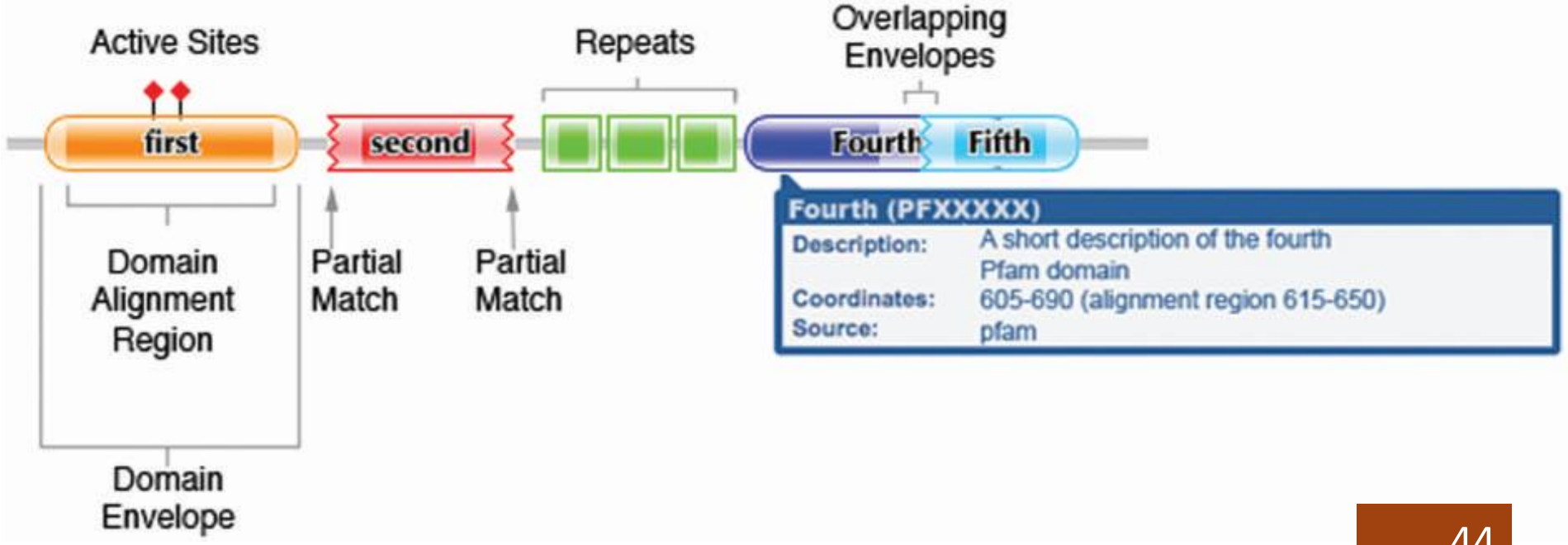


Язык Pfam :

Семейство – коллекция гомологичных доменов из разных белков.

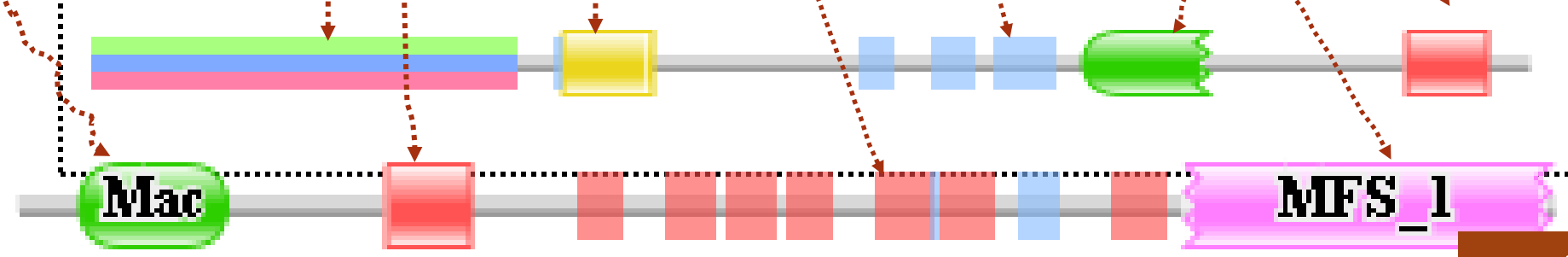
Домен – структурная единица, которую можно найти во множественном выравнивании.

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 бы 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»

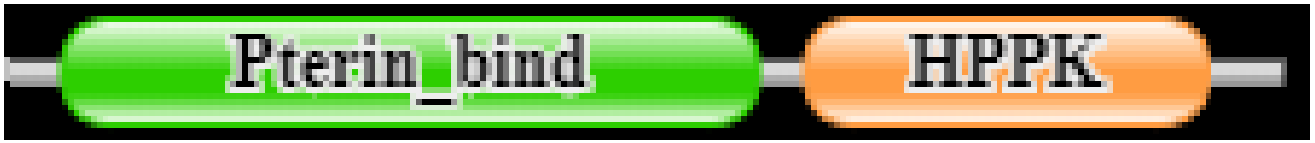
1. Выберите домен домен Pfam и подсемейство семейства белков с этим доменом. Рекомендуется выбрать белки с определенной доменной архитектурой.
2. Заполните форму , ссылка стоит на сайте под практикумом 11

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

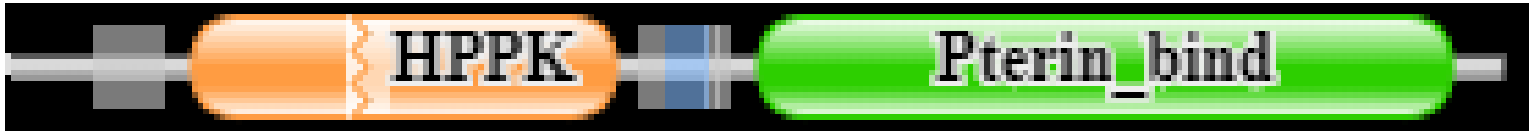
Пример крупной перестройки в эволюции.

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)



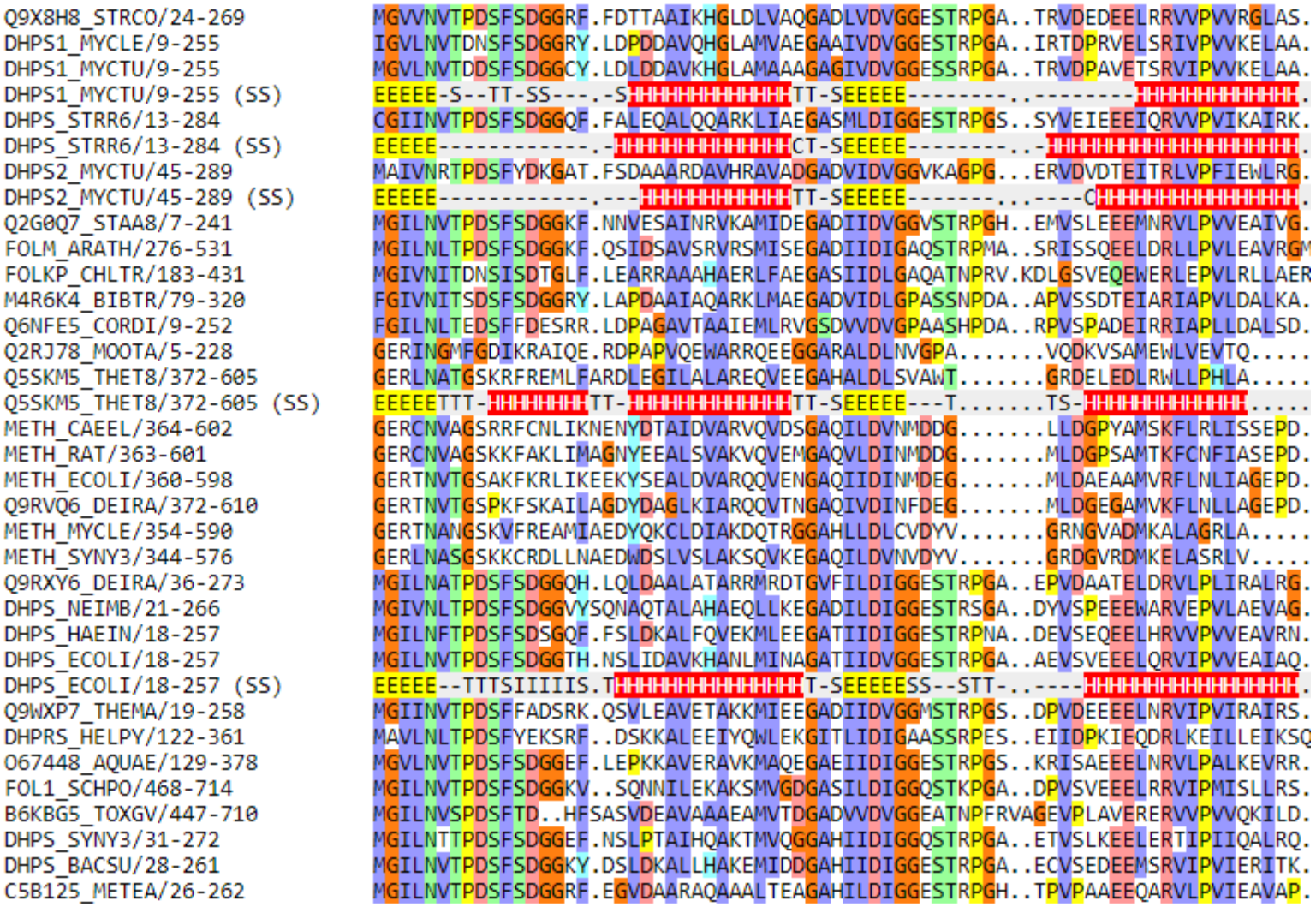
There are 9 sequences with the following architecture:
HPPK x 2, Pterin_bind
G2XU66_BOTF4 [Botryotinia fuckeliana (strain T4) (Noble rot fungus) (Botrytis cinerea)] Similar to folic acid synthesis protein CO:0000313|EMBL:CCD44036.1} (541 residues)



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

Seed sequence alignment for PF00809

Family: *Pterin_bind* (PF00809)



Seed (30)

Pterin binding enzyme

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