HMM-профили

Для поиска доменов и не только

- HMM-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

2

	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	*	*											

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 последовательностей (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

Где почитать про пакет HMMER

https://rothlab.ucdavis.edu/genhelp/hmmerbuild.html

Простая, но старая (2005). Годится для HMMER2

Eddy и команда разработчиков, 2023

http://eddylab.org/software/hmmer/Userguide.pdf

Описывает HMMER3 (про HMMER2 тоже есть) Недостаток – очень много всего. Надо искать нужное поиском.

Домен HPPK. Выравнивание SEED для профиля

Seed sequence alignment for 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....H...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

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)*

Автомат для него

Вероятности вычисляются по частотам

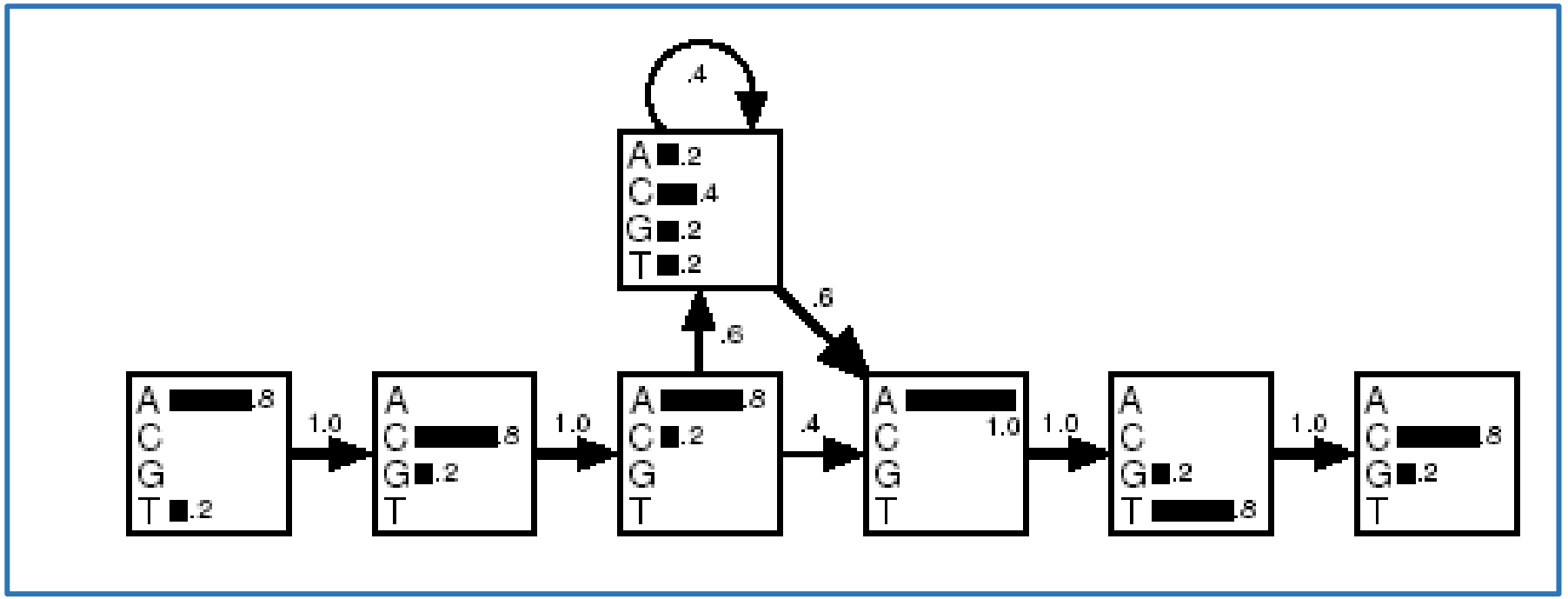
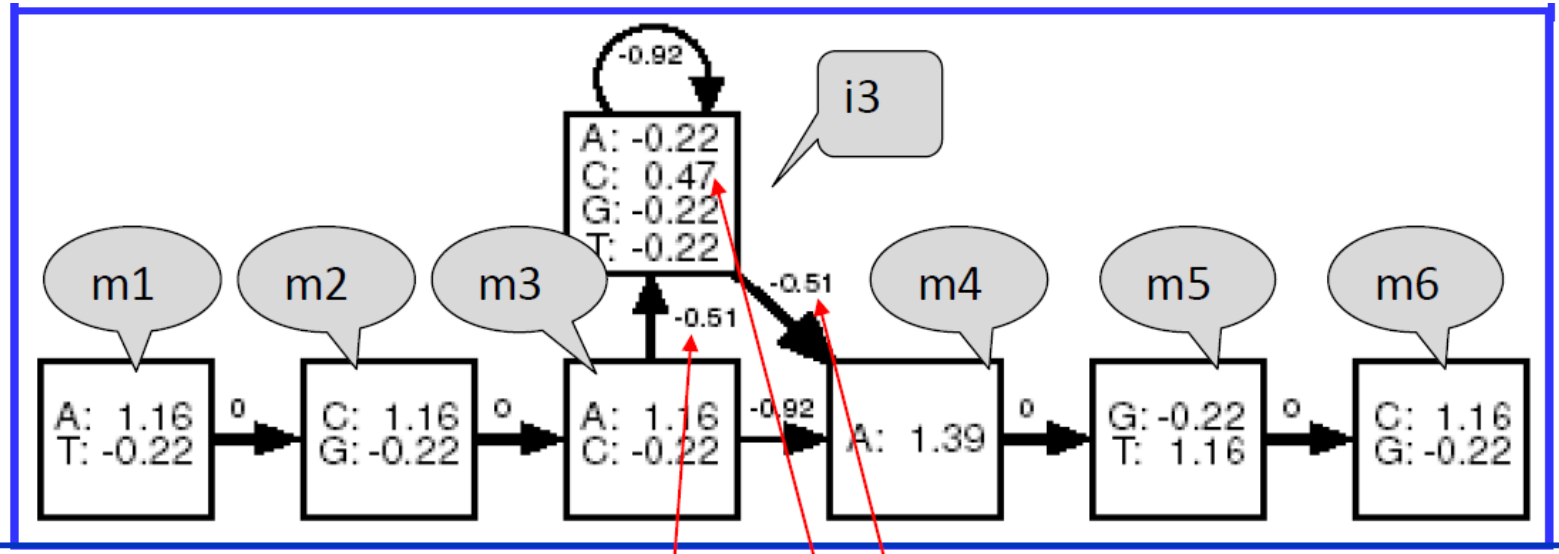


Fig. from Krogh, “Computational Methods in molecular biology, pages 45-63, Elsevier, 1998.

Вес выравнивания последовательности ACACATC с профилем

	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

Вместо вероятностей в профиле используют логарифмы отношения правдоподобия log2(частота буквы в колонке/базовая частота буквы)



- 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

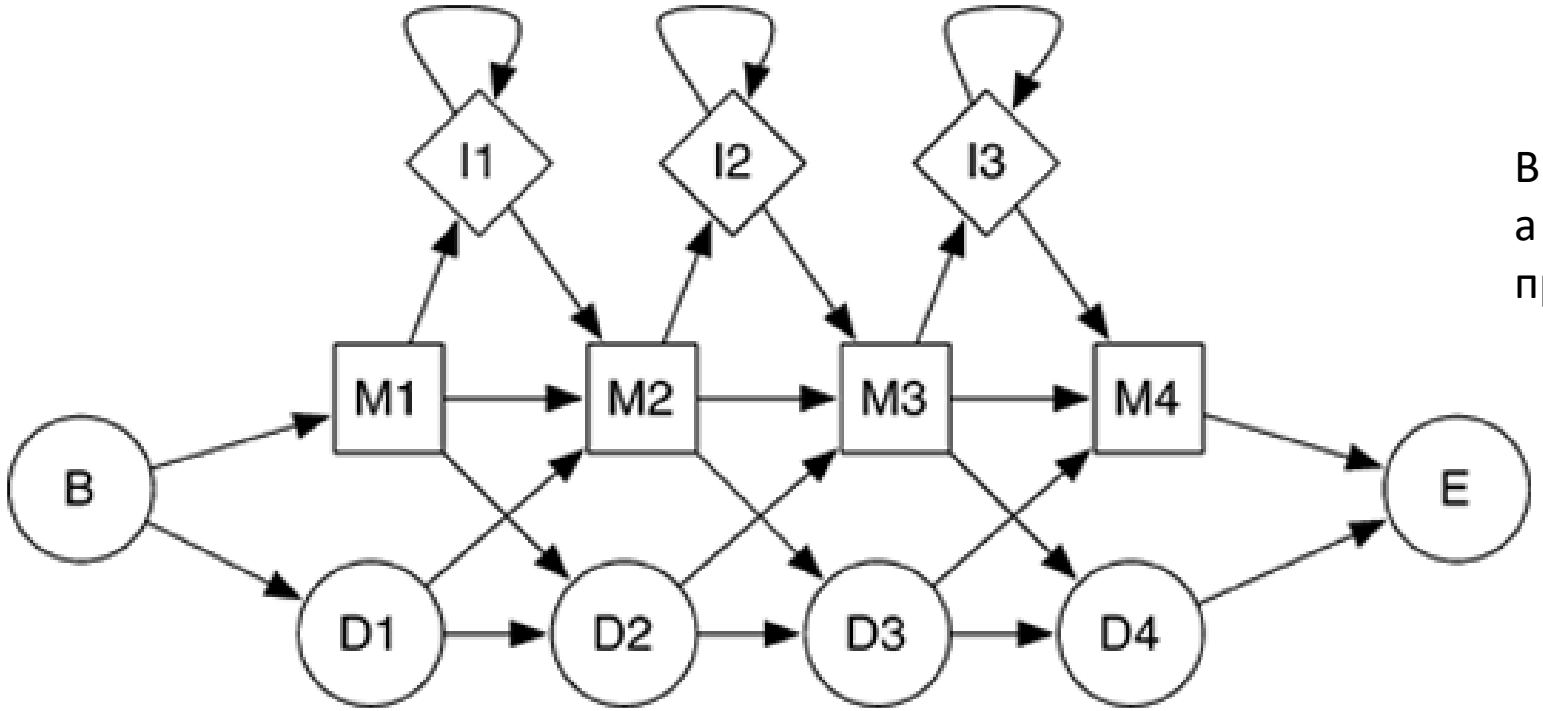
m1 m2 m3 i3 m4 m5 m6

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**

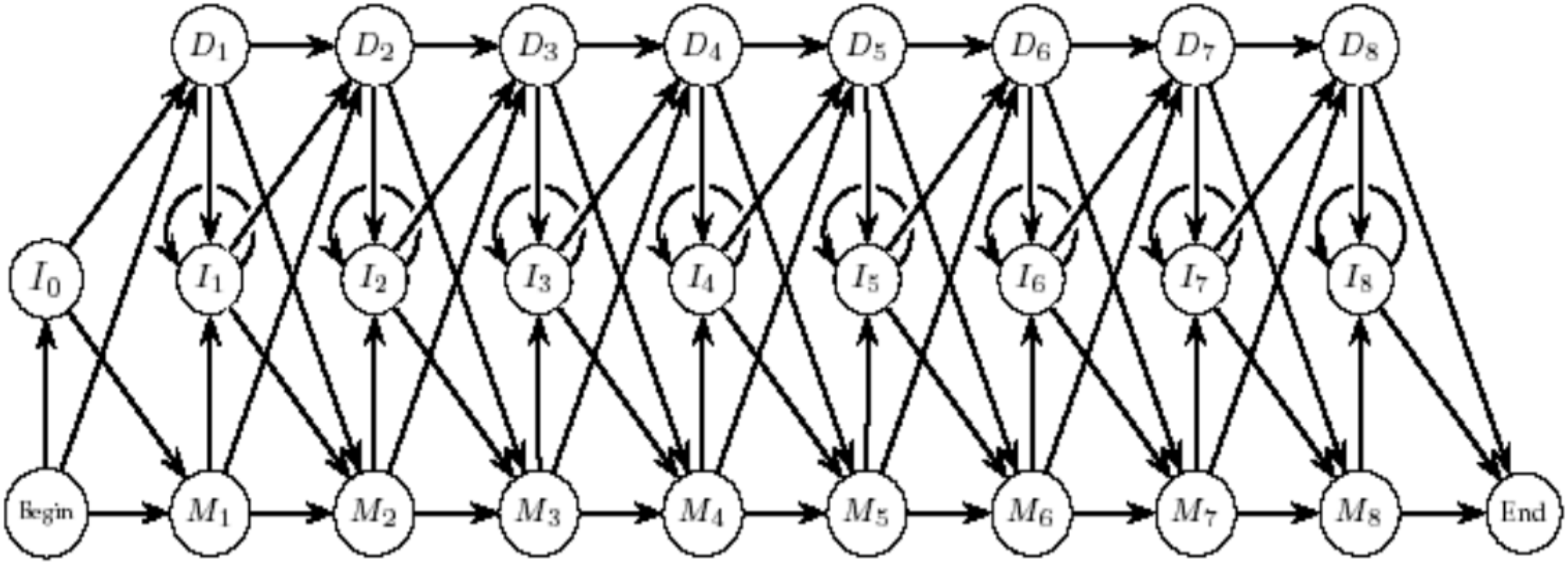
Семь типов транзиций (не считая начала и конца) Замены (M), Вставки (I), делеции(D)

- 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

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 >> 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 << 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

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

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		
	-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	32399	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 --cpu1 my_alignment.hmm my_alignment.fasta

Выходной файл:

my_alignment.hmm

Опции:

-f строит локальные профили, а не один глобальный

--cpu1–использовать **один** процессор. ТРЕБОВАНИЕ И.Русинова

STDOUT содержит мин, макс, средн. веса входных последовательностей относительно профиля

hmm2calibrate --cpu1 my_alignment.hmm

перезаписывает этот файл, уточняя некоторые константы модели

hmm2search --cpu1 -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. Eddyand 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

Достоинства HMMER3

- работает на два порядка быстрее HMMER2
- есть удобства в интерфейсе и выходных данных

Недостатки

- Трудно использовать, не понимая что и как происходит

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

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 --cpu1 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--cpu1 -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 parseabletable 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

Домены

База данных Pfam

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

Поглощена БД INTERPRO в 2022

(ЭВОЛЮЦИОННЫЙ) ДОМЕН

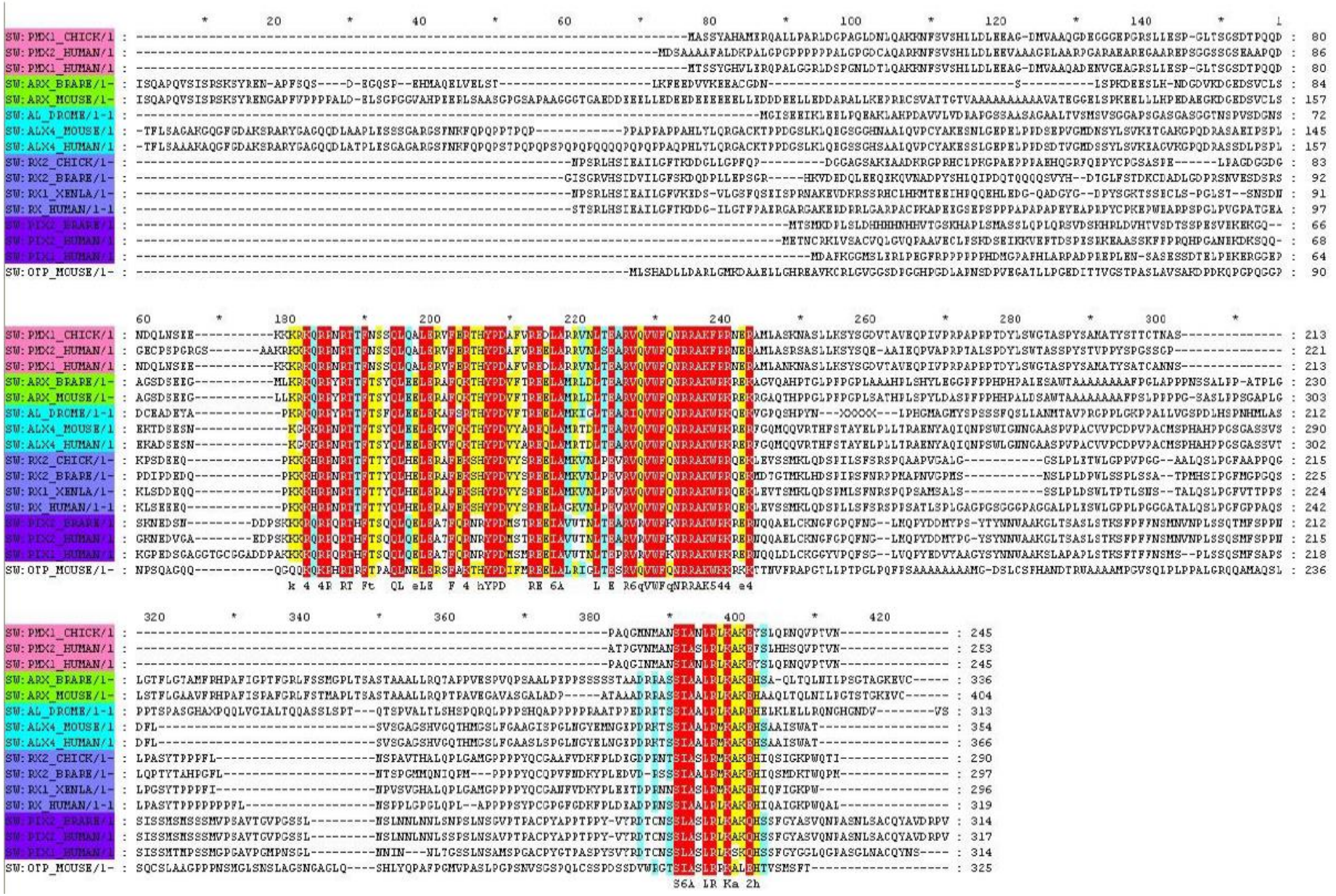
Домены – единицы непрерывной эволюции белков

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

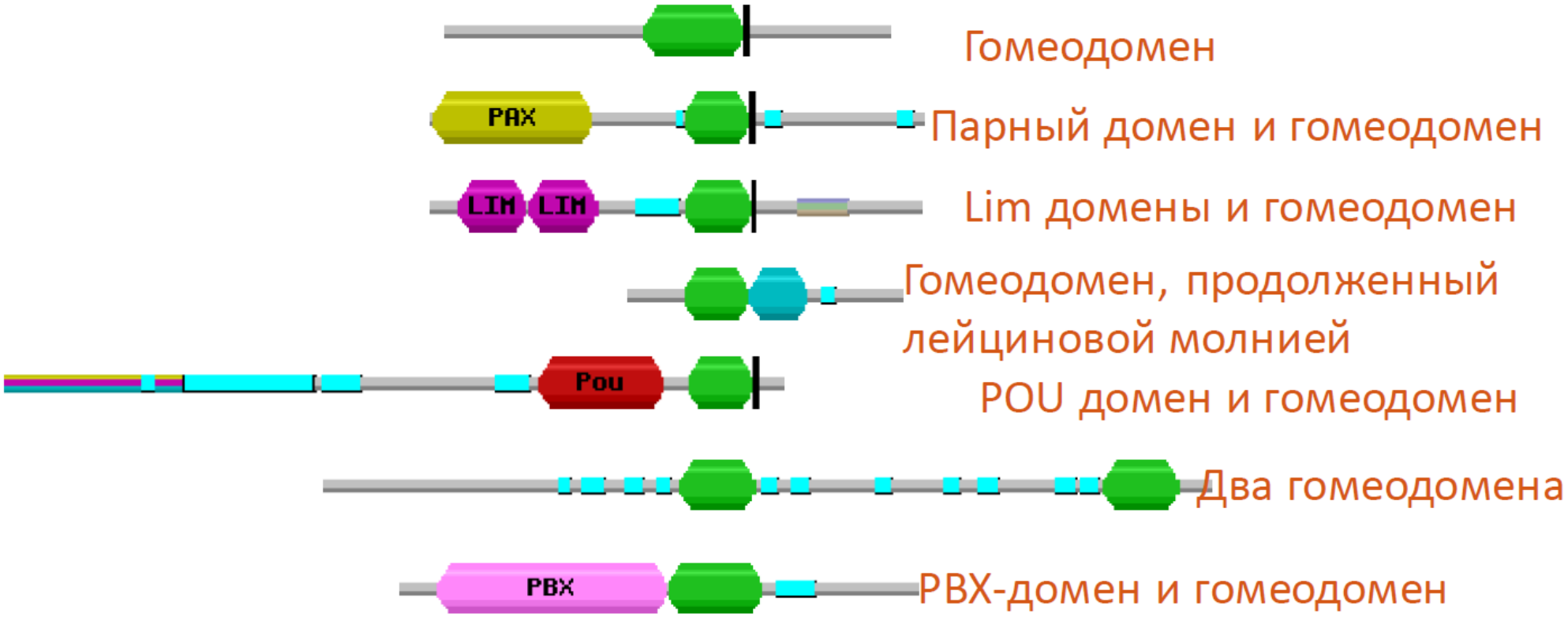
Так выглядит выравнивание белков, содержащих два домена гомеодомен PF00046) и OAR(PF03826)), не гомологичных по всей длине



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



БД Pfam

Единица хранения – семейство гомологичных доменов. Говорят «домен», отождествляя его с семейством

Идентификаторы ID (напр.Pterin_bind), AC (PF00809), название домена (Pterinbinding 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

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