БЕЛКИ:
Мотивы => паттерны
Семейства => PSI BLAST
Домены => профили

Сначала то, что не было сказано в прошлой лекции

I. 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

* Service MyHits https://myhits.sib.swiss/

* fuzzpro из пакета EMBOSS (на kodomo стоит)

Паттерн для цинкового пальца

Prosite

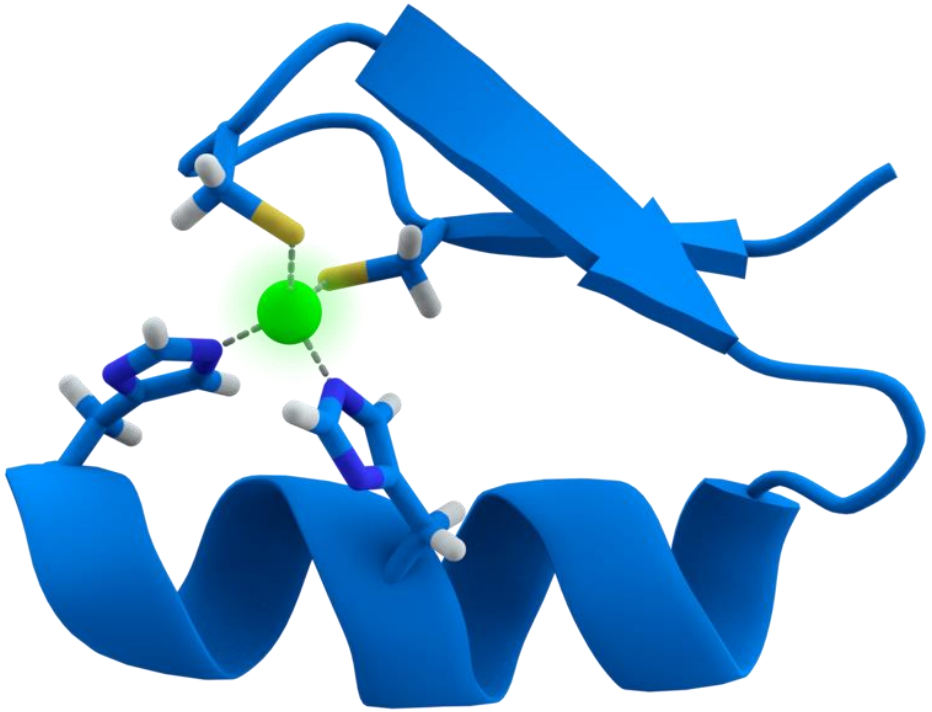
Паттерн для цинкового пальца типа C2H2:
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

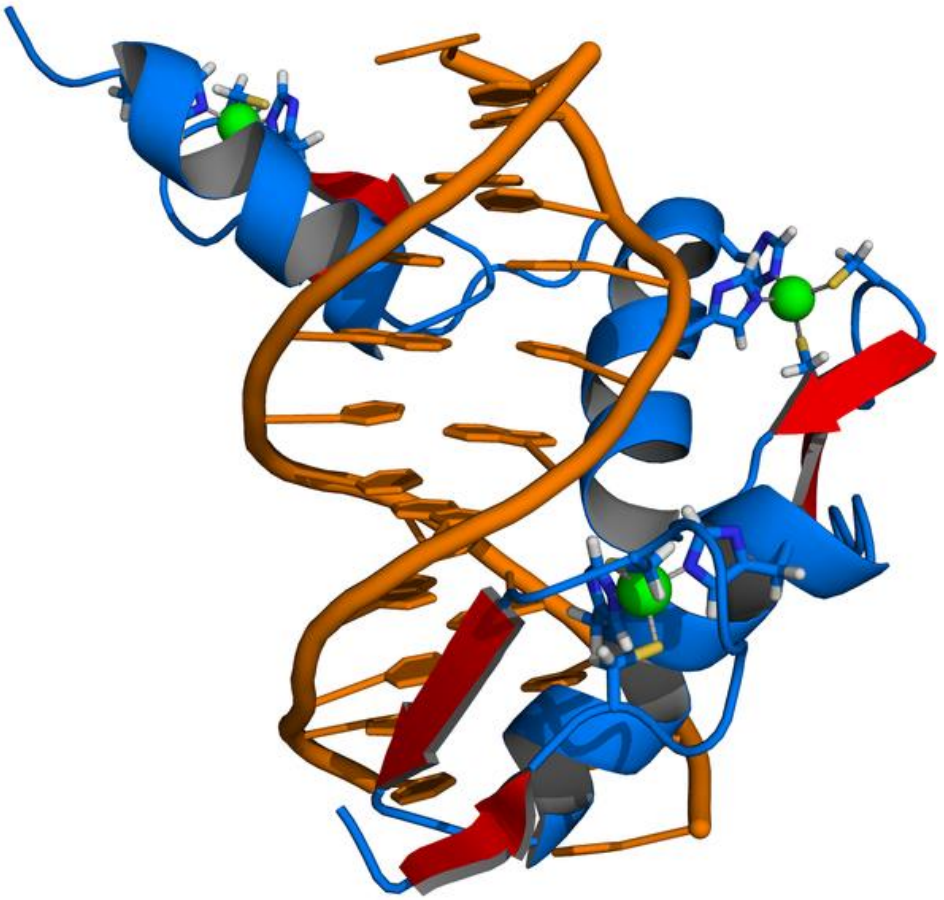
- [a-zAZ] –все возможные аминокислоты в данной позиции
- X(2,4) – любая аминокислота от 2 до 4 раз
- X(3) – любая аминокислота ровно 3 раза
- {P} – любая аминокислота, кроме пролина

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

ПРОПУСТИТЬ

Цинковые пальцы C2H2

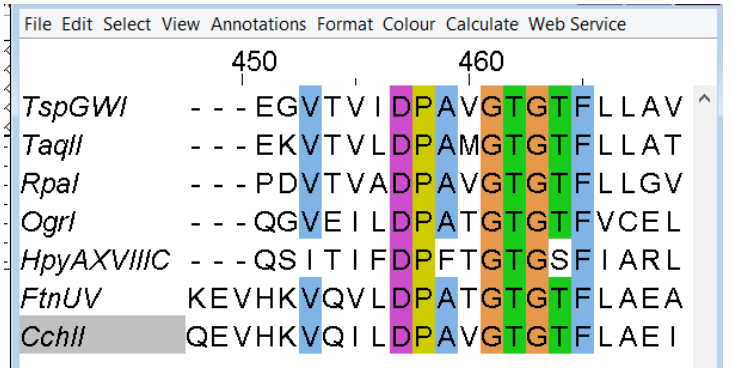




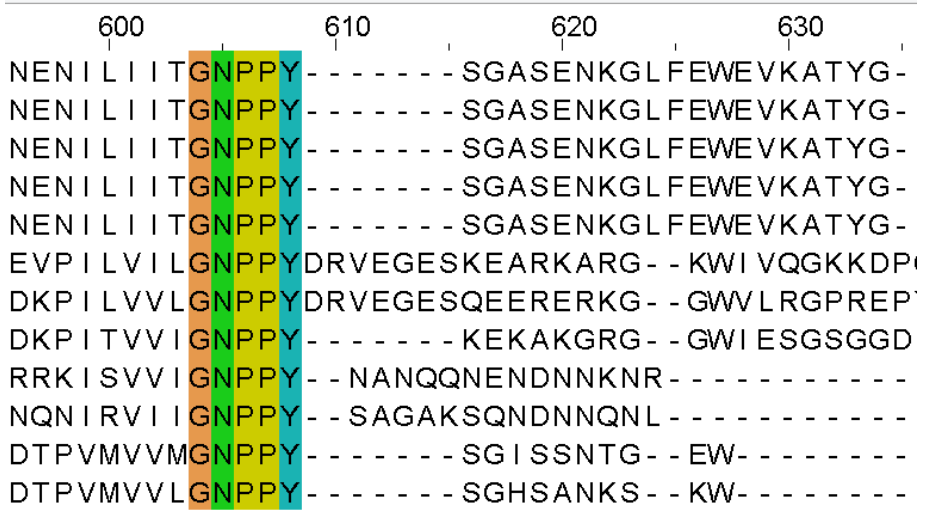
Паттерны в Jalview (Select => find)

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

I. Мотивы в белках



- 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



«Сигналы» в белках vs сигналы в ДНК

- **Промотор** можно вставить в вектор перед нужным геном и он будет работать
- Последовательность каталитического мотива P**D**……[**DE**]X**K** эндонуклеазы теоретически можно вставить в другой белок - путем вставки фрагмента ДНК, кодирующего мотив. **НИКТО ТАК НЕ ПОСТУПАЕТ**, т.к. все знают, что с полученного гена белок с эндонуклеазной активностью не получится со 100% гарантией
- БЕЛКОВАЯ ИНЖЕНЕРИЯ требует совсем других методов

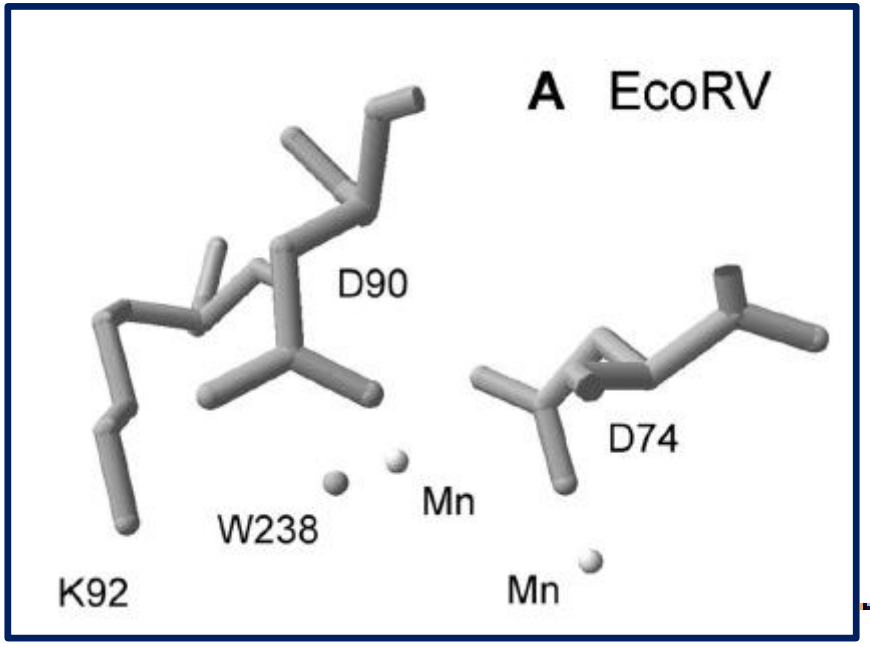
Каталитический мотив одного семейства эндонуклеаз рестрикции типа II

PD.....(D/E)XK

Многоточие – линкер переменной длины

A.Pingoud et al.

CMLS, Cell. Mol. Life Sci. 62 (2005) 685–707



Не показаны

- остаток 73, место пролина
- Участок 75 – 89
- Остаток 91 не показан

Молекула воды W238 и два иона марганца необходимы для реакции расщепления ДНК

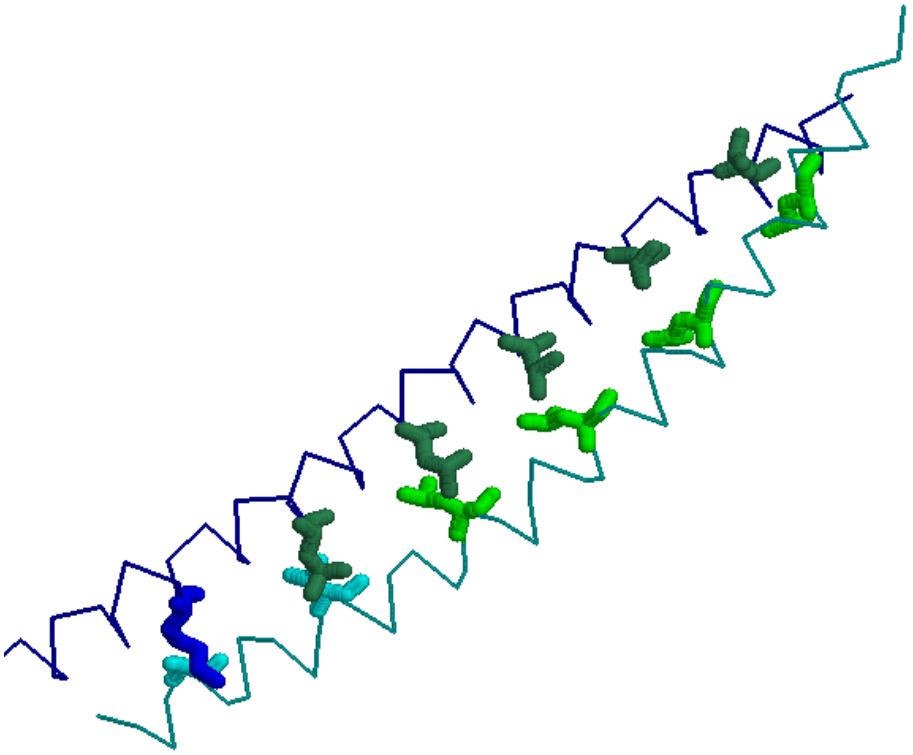
Xie et al.

J Inorg Biochem. 2010 June ; 104(6): 665–672.

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

II. PSSM и PSI BLAST для белков

С.А.Спирин, А.С.Ершова

Вспомним 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.

PSSM — position-specific scoring matrix

По смыслу PSSM — это то же, что PWM, но термин PWM используется для мотивов в ДНК, а PSSM для мотивов в белках или для описания семейств родственных белков.

Можно использовать гэпы, учитываются как 21 буква.

PSSM применяется так же, как PWM:

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

PSSM — position-specific scoring matrix

Как создать PSSM по выравниванию?

Базовая идея — та же, что для PWM:

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

где 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)

Для PSSM актуальна проблема почти идентичных последовательностей в выравнивании

- Если в выравнивании много очень похожих последовательностей, то частоты букв из них во многих колонках будут велики по сравнению с частотами букв из остальных последовательностей
- Проблема решается введением веса каждой последовательности.
  - Вес w_s последовательности s маленький если много похожих последовательностей.
  - Вес w_s последовательности s большой, если она значительно отличается от остальных последовательностей в выравнивании.
  - Предложено и используется много разных систем приписывания веса последовательности в выравнивании

Оценка частоты остатка в позиции с учетом веса последовательности

Придумали такой способ: присвоить последовательности вес (weight) так, чтобы у последовательностей, имеющих много родственников, он был маленьким, а у «одиноких» последовательностей — большой. При расчете частоты остатка i в позиции k используются веса последовательностей w_s:

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

Если все веса последовательностей равны, то получится обычная частота. Здесь a_{sk} — буква последовательности s в позиции k, ψ_i — псевдоотсчёт для остатка i.

Внимание: слово «вес» имеет два разных значения

- Вес = Score, вес выравнивания двух последовательностей или последовательности относительно профиля (PWM или PSSM или HMM), обычно обозначается S.
- Вес = Weight, вес последовательности, используемый при построении PSSM по множественному выравниванию, обычно обозначается w.

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.

Алгоритм PSI-BLAST

На входе — последовательность и порог по e-value, на выходе — набор найденных последовательностей и построенная по ним PSSM.

1. На первом этапе запускается обычный BLASTP входной последовательности против выбранного банка последовательностей
2. Для находок со значениями e-value лучше заданного порога строится множественное выравнивание.
3. Это выравнивание используется для получения PSSM.
4. На следующем шаге опять происходит запуск BLAST для исходной последовательности против того же банка последовательностей, но вместо матрицы замен остатков используется PSSM, полученная на предыдущем шаге.
5. Повторяем шаги 2-4, пока не перестанут добавляться новые последовательности.

Дополнительные возможности PSI-BLAST

- Можно вручную включать/исключать последовательности, которые используются для построения PSSM
- Можно использовать PSSM, созданную на основе поиска в одном банке, для поиска в другом банке.

III. Домены и профили

База данных Pfam

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

Поглощена БД INTERPRO в 2022

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

Разрешаются индели в выравнивании. Этим профили отличаются от PWM и PSSM

ПРОФИЛИ применяются для поиска **доменов** в последовательностях белков (и не только)

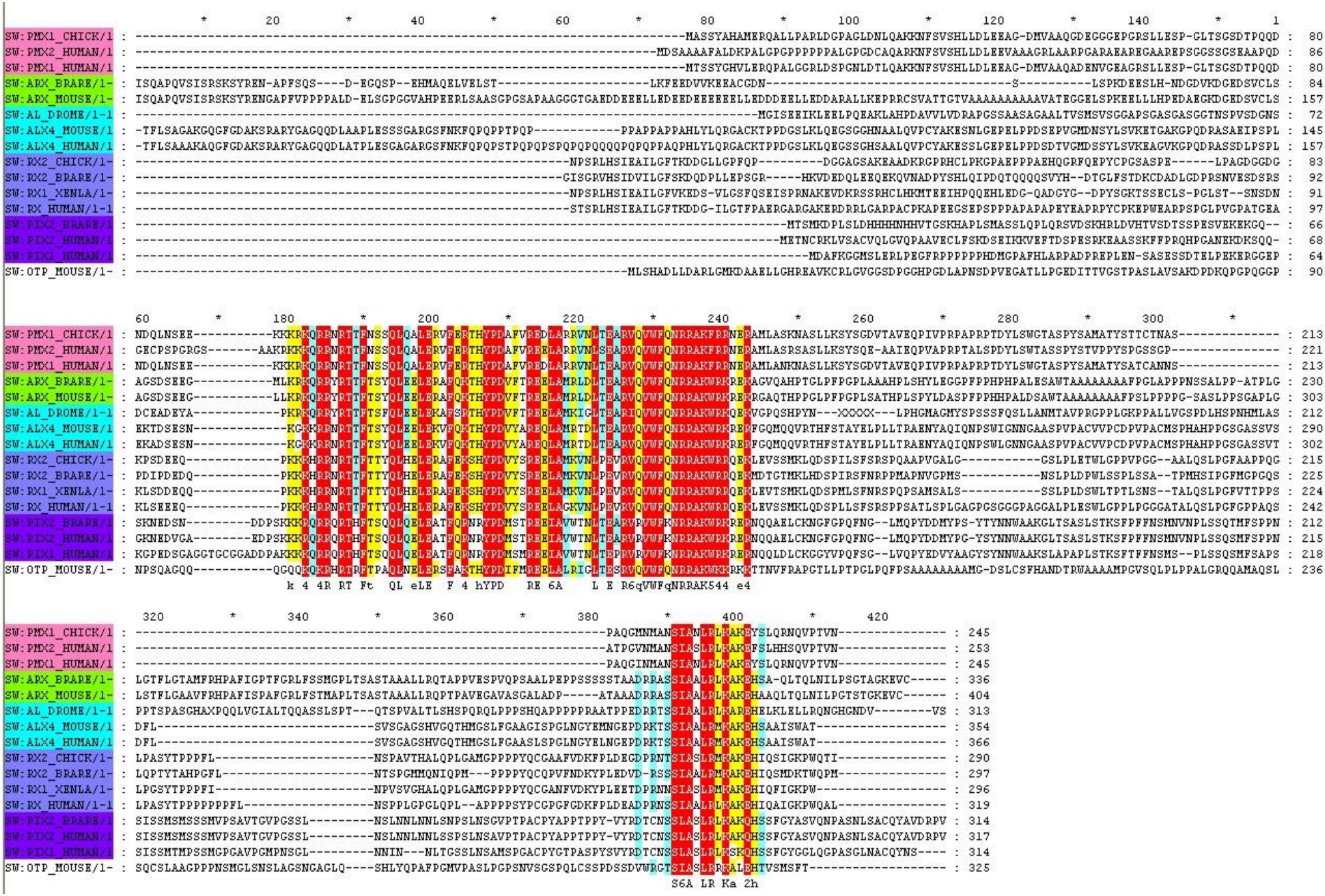
ДОМЕН - Домены – единицы непрерывной эволюции белков

Непрерывная эволюция это замены остатков, небольшие делеции и вставки.

Домены можно обнаружить с помощью выравнивания

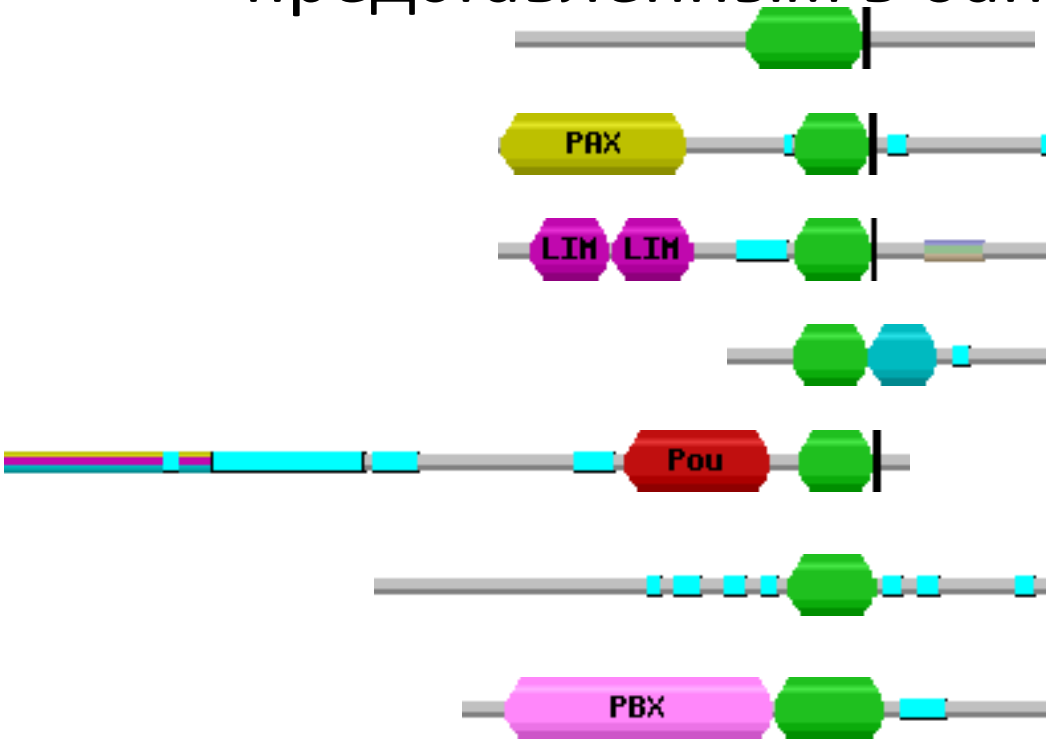
Кроме непрерывной эволюции бывают единовременные крупные изменения в последовательностях белков

Гомеобелки (подсемейство с OAR доменом)



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

- Об этом можно судить по 898 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

А про этот домен что скажете? *zf-C2H2* (PF00096)

См. слайд 4.

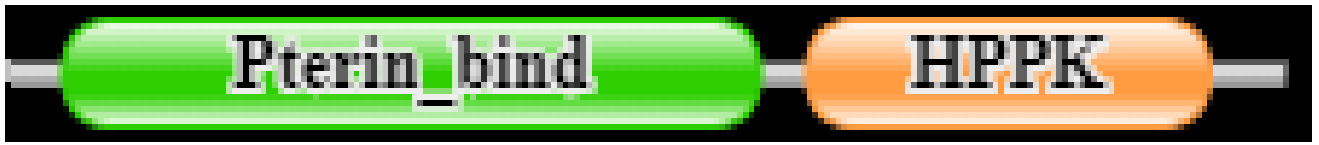
Обнаружен в **1 201 914** sequences

Белках с 14182-я доменными архитектурами

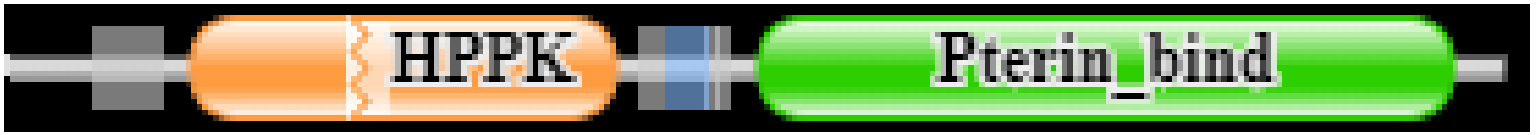
Пример крупной перестройки в эволюции.

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)



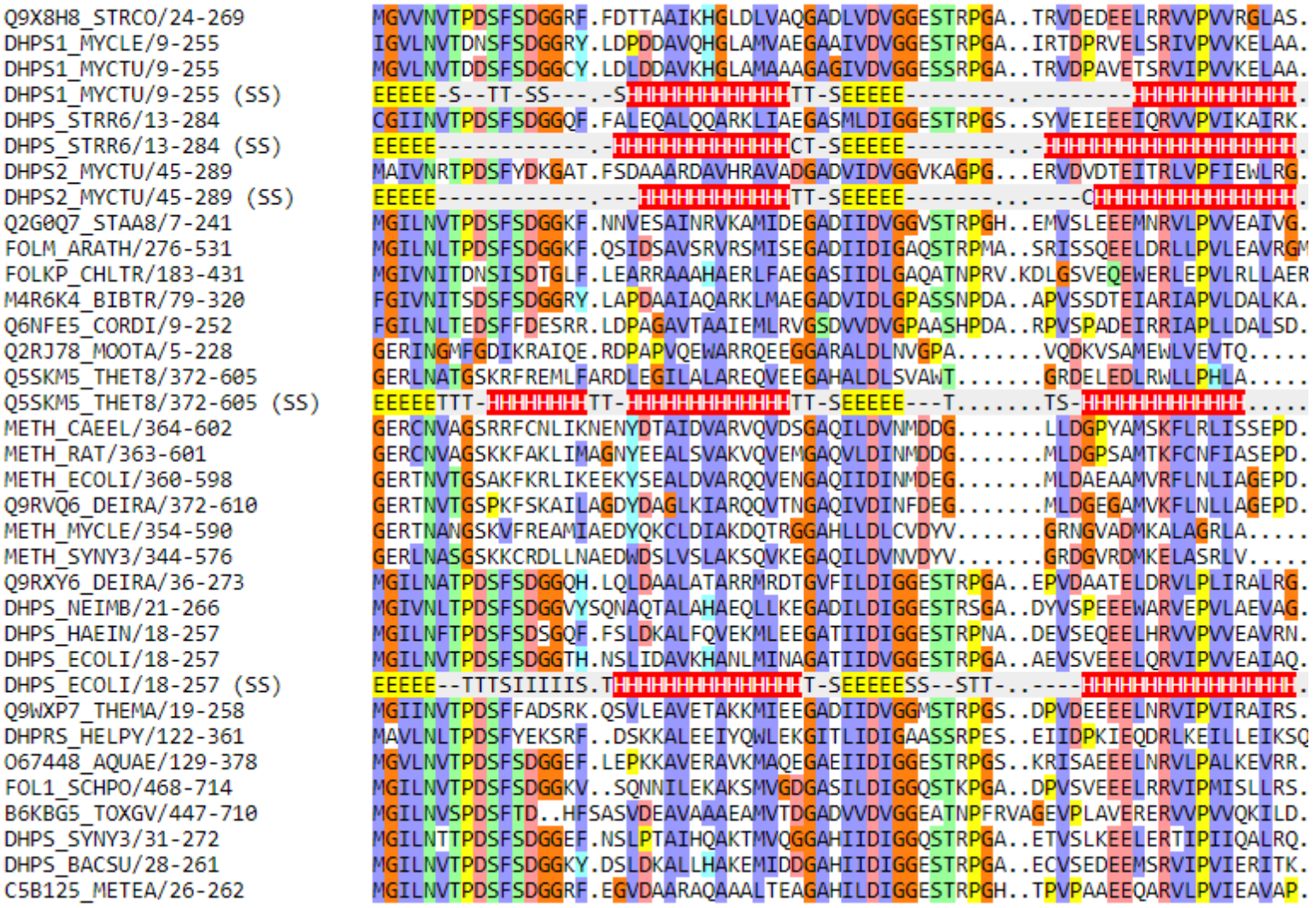
There are 9 sequences with the following architecture:
HPPK x 2, Pterin_bind
G2XU66_BOTF4 [Botryotinia fuckeliana (strain T4) (Noble rot fungus) (Botrytis cinerea)] Similar to folic acid synthesis protein CO:0000313|EMBL:CCD44036.1} (541 residues)



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 ……..

HMM 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	*	*											

Домены HPPK

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)

Take home message

Выравнивания сотен и тысяч последовательностей белков всегда содержат ошибки.

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

Выравнивания seed – входные данные для построения ПРОФИЛЯ

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 случайном банке для подбора порога веса и E-value
4. С помощью профиля находит все домены в базовых множествах последовательностей Pfam (основа Uniprot с отставанием на пару лет)
5. Готовит запись в банк Pfam

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

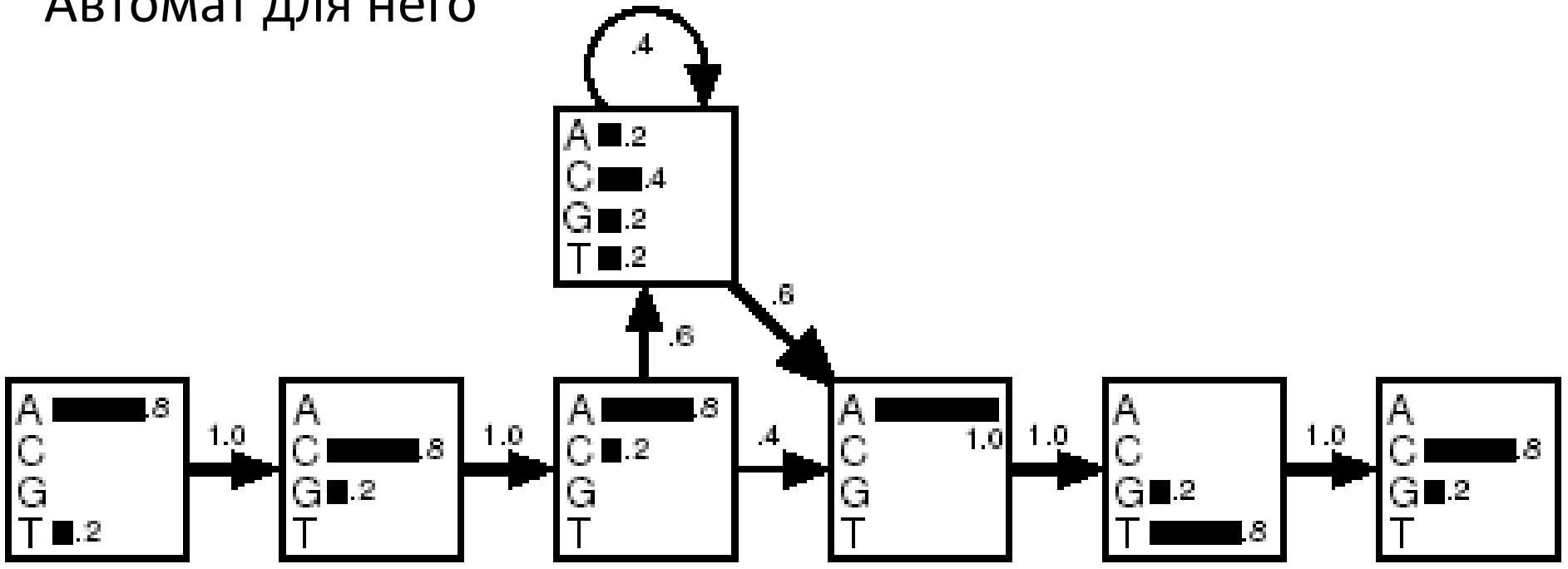
```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

Вероятности в квадратиках называются *эмиссионными вероятностями*

Вероятности на стрелочках - *вероятностями перехода*

Вероятности вычисляются по частотам

Автомат для него



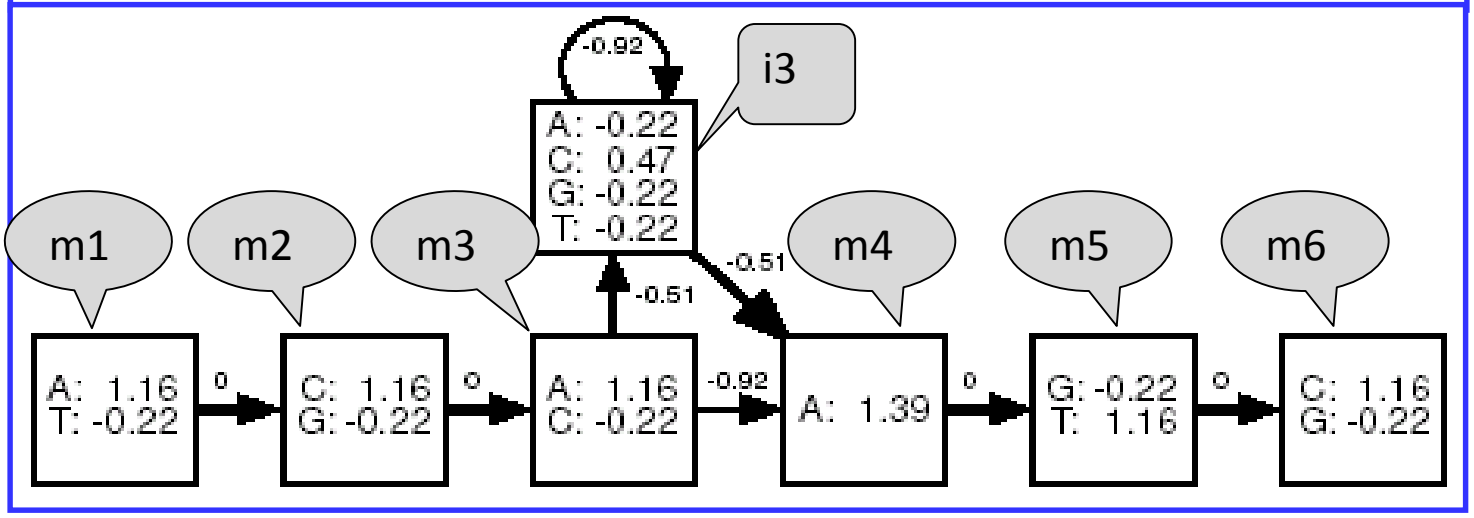
from Krogh, “Computational Methods in molecular biology, pages 45-63, Elsevier, 1998.

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

- Пусть базовые частоты всех букв одинаковы и, следовательно, равны 0.25
- 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

Определим 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



$$\begin{aligned}\text{log-odds}(\mathbf{ACACATC}) &= 1.16+0+1.16+0+1.16-0.51+\\ &\quad 0.47-0.51+1.39+01.16+0+1.16\\ &= 6.64.\end{aligned}$$

Мы нашли

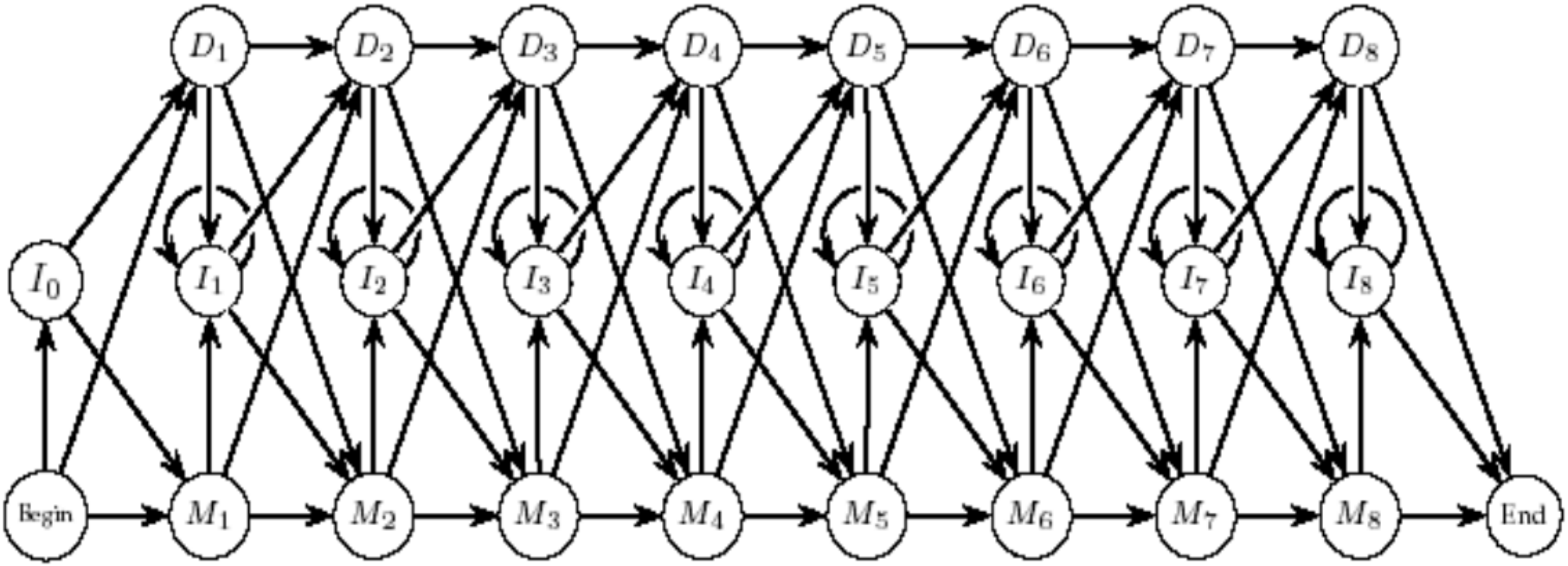
- Оптимальное выравнивание
  - **A C A C A T C**
  - **m1 m2 m3 i3 m4 m5 m6**
- Его вес 1.16 + 0 + 1.16 + 0 + 1.16 – 0.51 + 0.47 – 0.51 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.64

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

Более сложная ситуация

- Возможны вставки (i) в любом месте
- Возможны делеции (d) в любом месте
- Разрешены все возможные переходы между вершинами b (begin), m(match), i(insertion), d(delelion), e(end):
  - b => m1, b => d1, b => i1
  - m => следующую m, m => i, m=>d, m => e
  - i => i, i => m, i => d, i => e
  - d => d, d => m, d => i, d => e

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



Из презентации безымянного сотрудника ИППИ)

Профили

- На вход подается выравнивание с инделями
- По нему строится т.н. профиль HMM (Hidden Markov Model)
- Профиль HMM можно выровнять с последовательностью и получить вес выравнивания. Локальное и глобальное выравнивание.
- Профиль калибруется по случайному банку для нормализации веса и расчета E-value
- При наличии множества последовательностей, про которые известен ответ – есть в них домен или нет, - можно уточнить порог нормализованного веса для находки
- С помощию профиля в базе последовательностей (Uniprot) находятся участки с весом больше порога, следовательно, белки, содержащие домен.
- Важное отличие профиля от PWM:
  профиль может быть построен по выравниванию с инделями

HMM профиль, построенный HMMer'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		
	-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**	*****	*****		

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

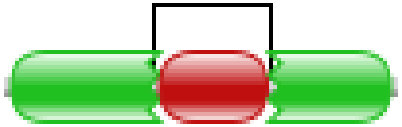
БД Pfam

- 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 выравнивания
- Seed alignment – это выравнивание, по которому составлен профиль домена. Дерево этого выравнивания
- 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

Типы объектов кроме доменов в Pfam

Domains of unknown function (DUFs)



Язык Pfam :

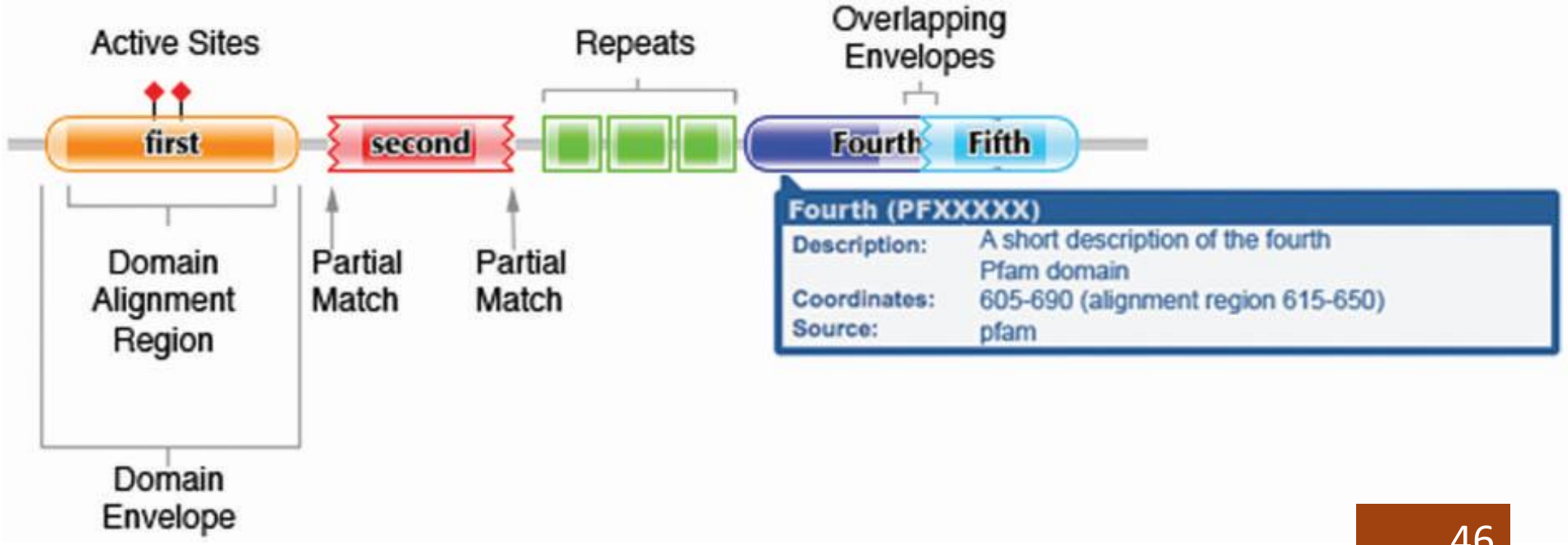
Семейство – коллекция гомологичных доменов из разных белков.

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

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

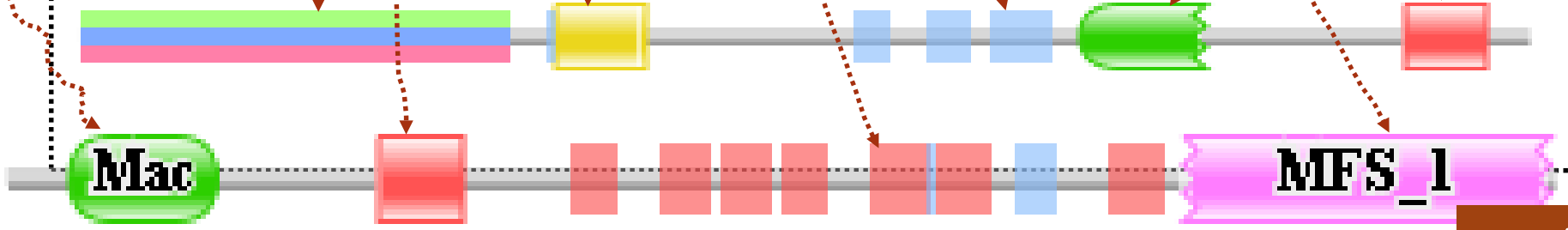
Мотив – короткая единица структуры вне глобулярных доменов.

Клан – группа родственных записей.



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



Задание на «занятии» (до …)

- **Задание 2.1 Выберите домен и доменную архитектуру, в которую входит домен**

  Составьте список белков Uniprot с выбранной доменной архитектурой (табл.1)

- c. (*) Определите интервал типичных длин белков от - до (мода на гистограмме длин)
- d. (*) Составьте выборку из 40 – 60 последовательностей характерной длины. Чтобы получить представительную выборку, из нескольких семейств выбирайте по несколько последовательностей, принадлежащих разным семействам.

Конец презентации

Сигналы в ДНК vs сигналы в белках

Применимы ли технологии сигналов: PWM, IC, MEME и FIMO для последовательностей белков?

СИГНАЛ в последовательности белка? Бывает? Я задумался … сайты протеолиза, разве что, и то…

НЕТ: «сигналы» на поверхности белковой глобулы: активные центры, сайты связывания ко-факторов, поверхности белок-белкового взаимодействия. Консервативные структурные мотивы.

ДА: аналогичные технологии используются для поиска

1. гомологичных участков в белках (доменов)
2. консервативных мотивов

Какие сайты расщепляет фитаспаза из риса (*Oryza sativa)*

Фитаспазы - аспартат-специфические протеазы растений. Необходимые белки апоптоза у растений. Расщепляют белки после аспартата в тетрапептиде с общим паттерном XXXD (!)

Помощью комбинаторных библиотек показано влияние всех остатков в позициях XXX на эффективность гидролиза