Домены и профили

База данных Pfam

ПРОФИЛЬ – описание выравнивания, вроде PWM, 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.

Кроме непрерывной эволюции бывают единовременные крупные изменения в последовательностях белков

HMM профиль

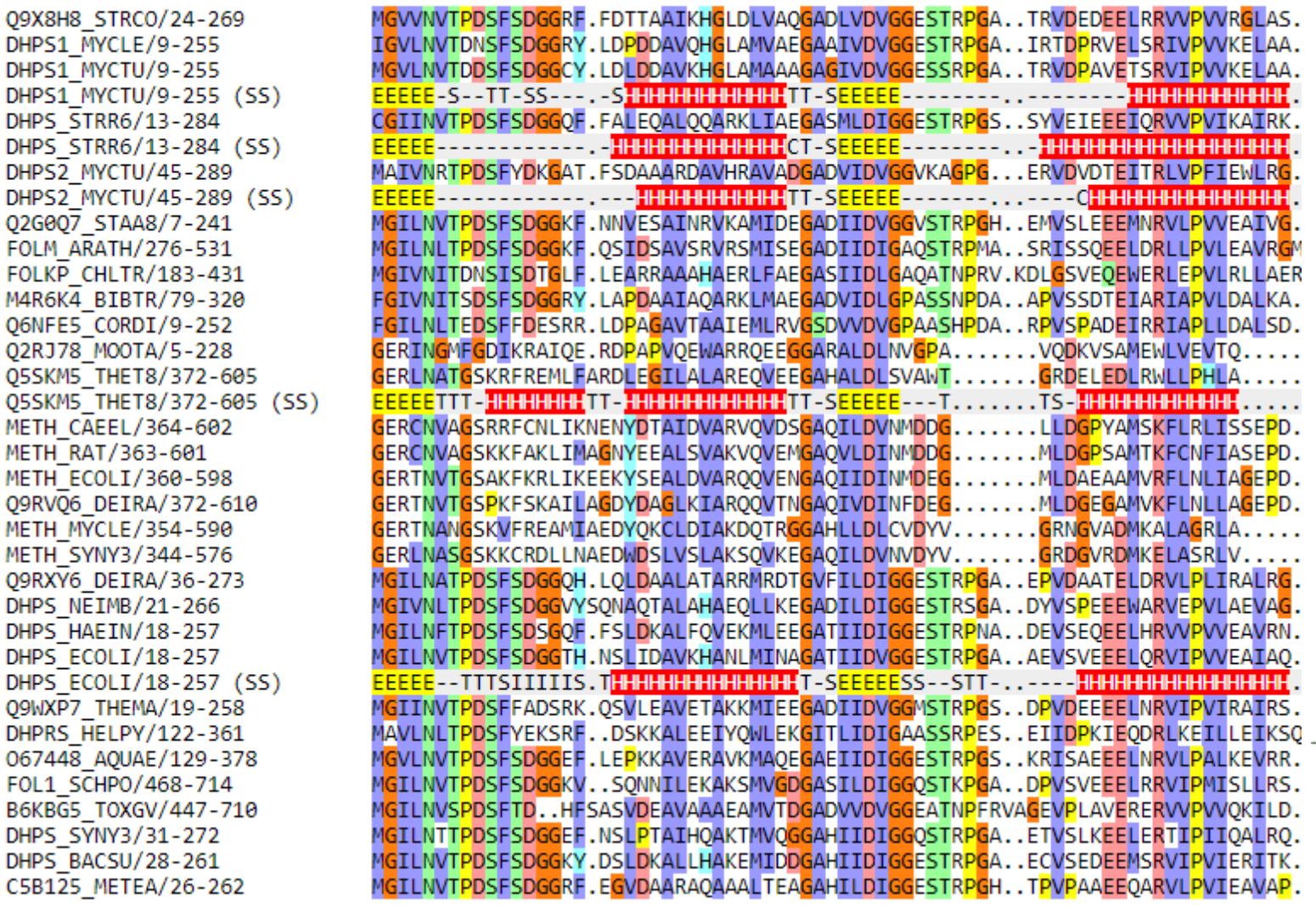
20 аминокислотных остатков

позиции выравнивания

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	
	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	-1
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-1
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-1
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-1
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-1
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-2
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-1
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-1
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-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	-2
-	-149	-500	233	43	-381	399	106	-626	210	-466	-720	275	394	45	96	359	117	-369	-294	-
-	-21	-6672	-7714	-894	-1115	-701	-1378	*	*											

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.

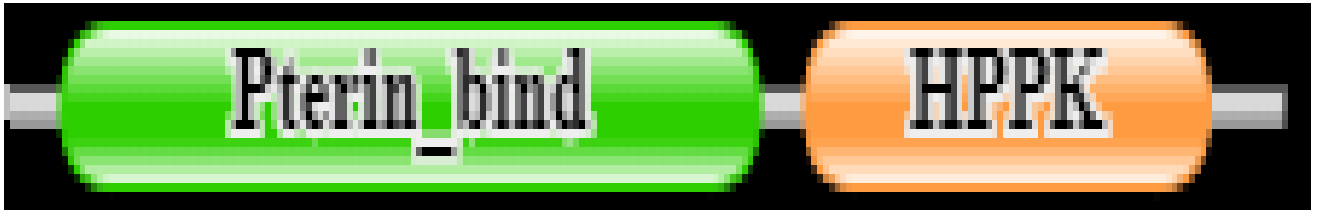
The family includes ……..

Пример крупной перестройки в эволюции.

Гомологичны ли эти 41 + 9 белков?

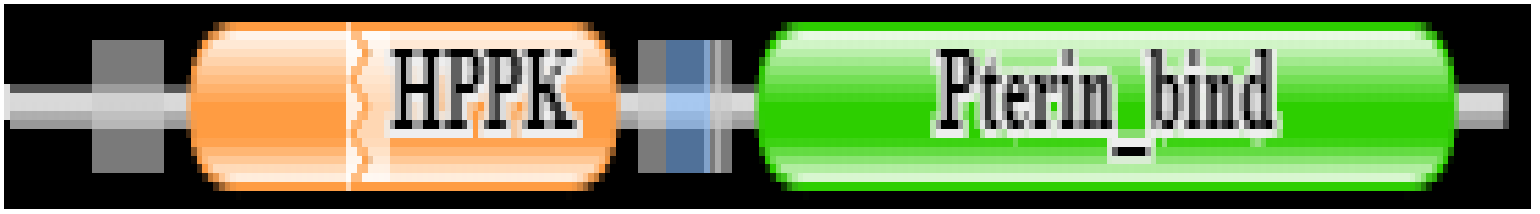
There are 41 sequences with the following architecture: Pterin_bind, HPPK

{ECR9KWZ5_9ACTN [Enterorhabdus caecimuris B7] Dihydropteroate synthase (437 residues)



There are 9 sequences with the following architecture: HPPK x 2, Pterin_bind

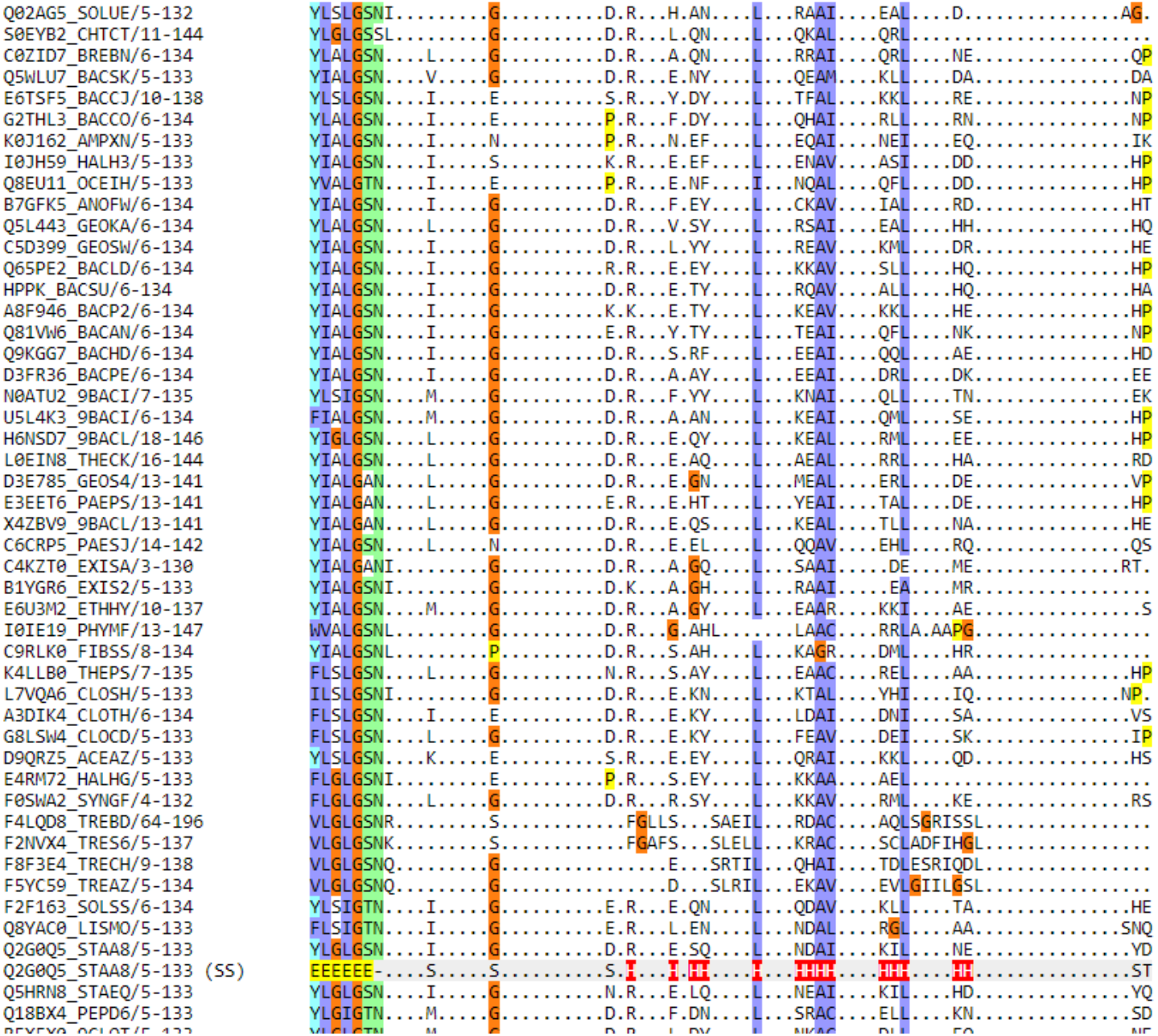
G2XU66_BOTF4 [Botryotinia fuckeliana (strain T4) (Botrytis cinerea)] Similar to folic acid synthesis protein (541 residues)



Домены HPPK

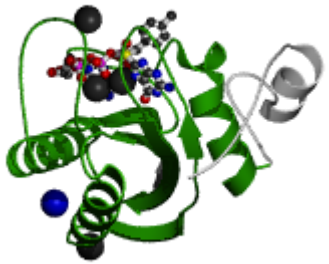
Family: *HPPK* (PF01288)

Seed sequence alignment for PF01288



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

A Rascal?

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

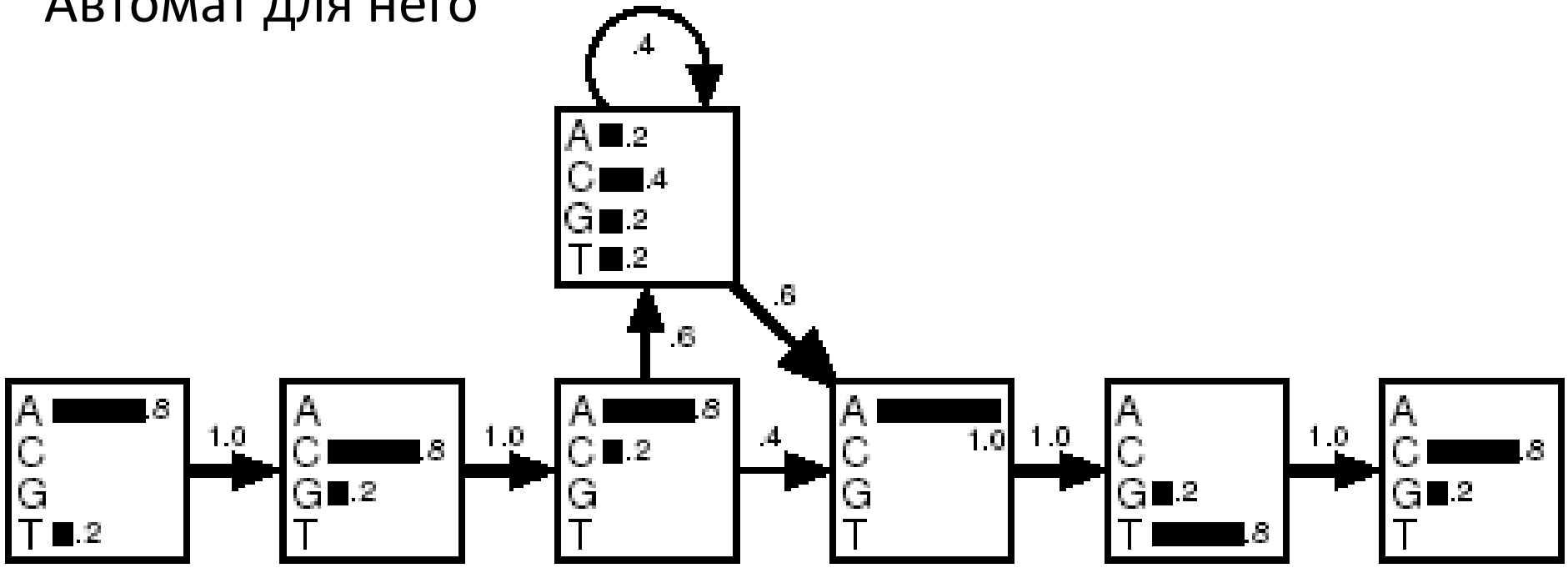
```
A C A - - - A T G
T C A A C T A T C
A C A C - - A G C
A G A - - - A T C
A C C G - - A T C
```

Вероятности в квадратиках называются *эмиссионными вероятностями*

Вероятности на стрелочках - *вероятностями перехода*

Вероятности вычисляются по частотам

Автомат для него



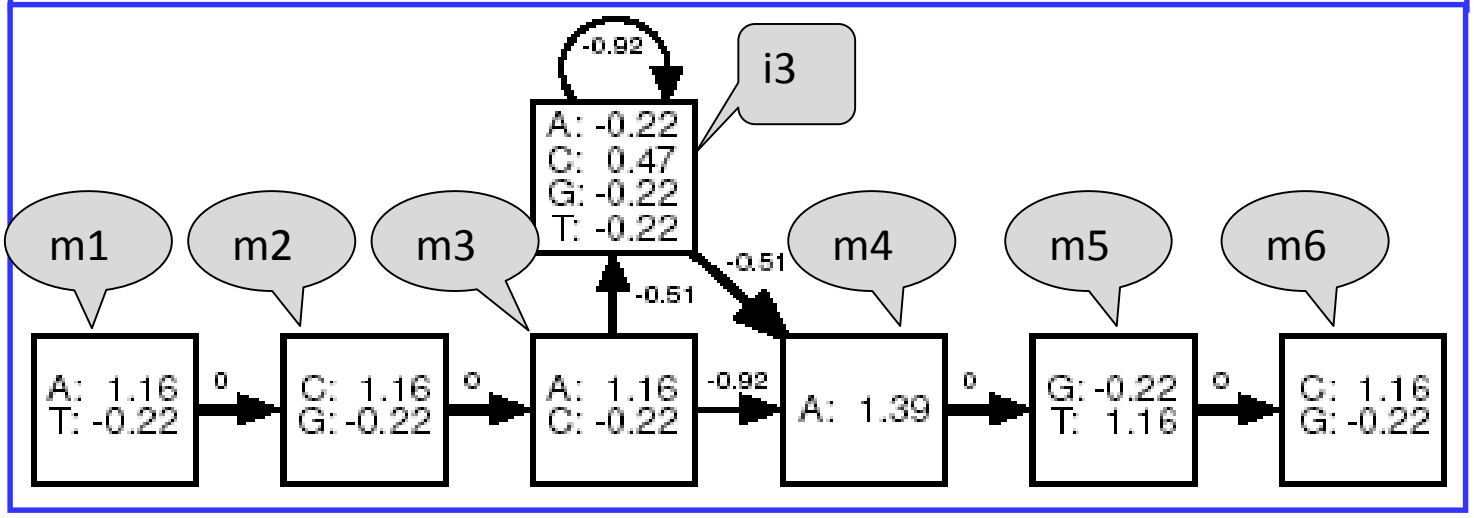
from Krogh, “Computational Methods in molecular biology, pages 45-63, Elsevier, 1998.

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

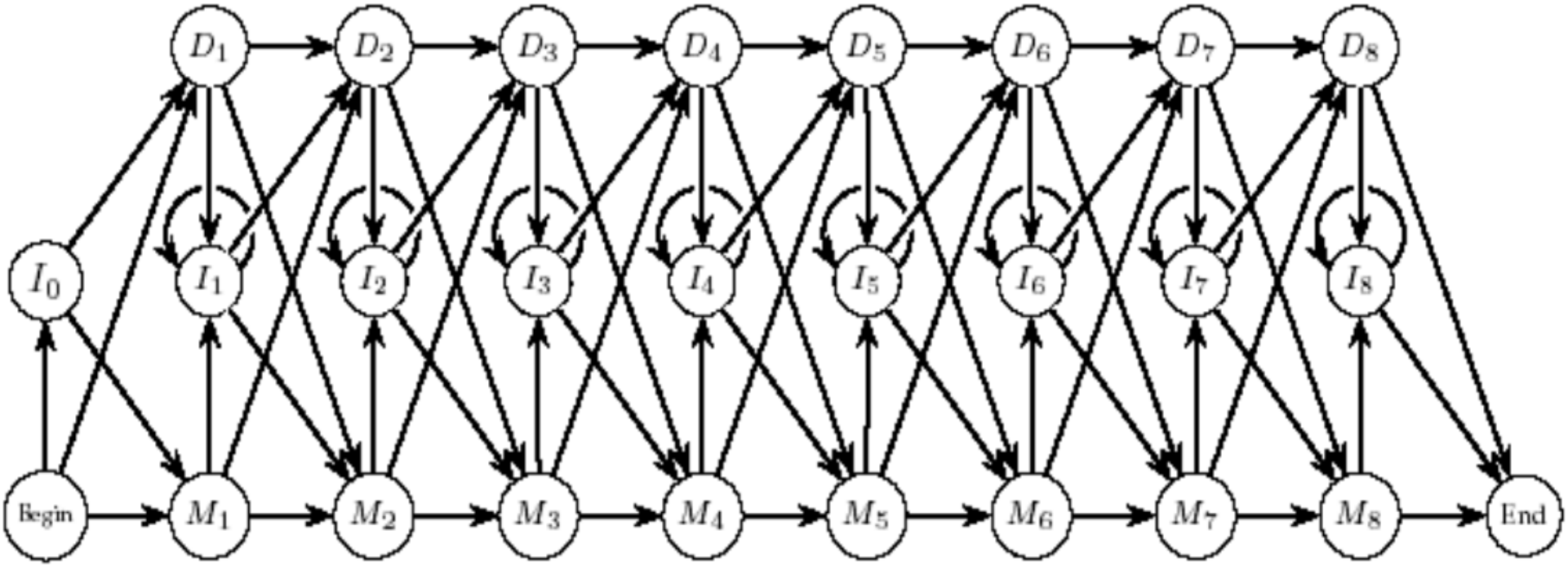
- Оптимальное выравнивание
  - **A C A C A T C**
  - **m1 m2 m3 i3 m4 m5 m6**
- Его вес 1.16 + 0 + 1.16 + 0 + 1.16 – 0.51 + 0.47 – 0.51 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.64

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

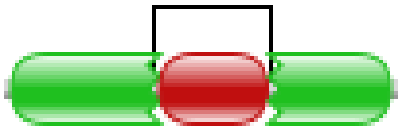
БД Pfam

- 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)



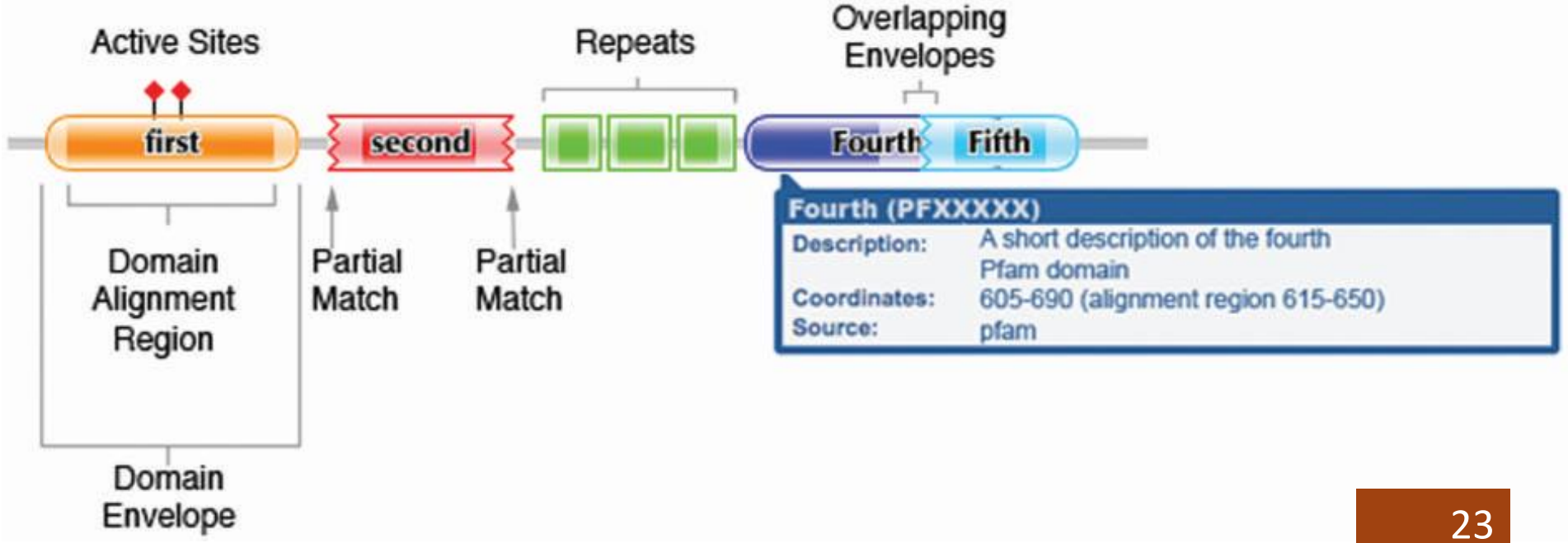
Язык Pfam :

Семейство – коллекция гомологичных доменов из разных белков.

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

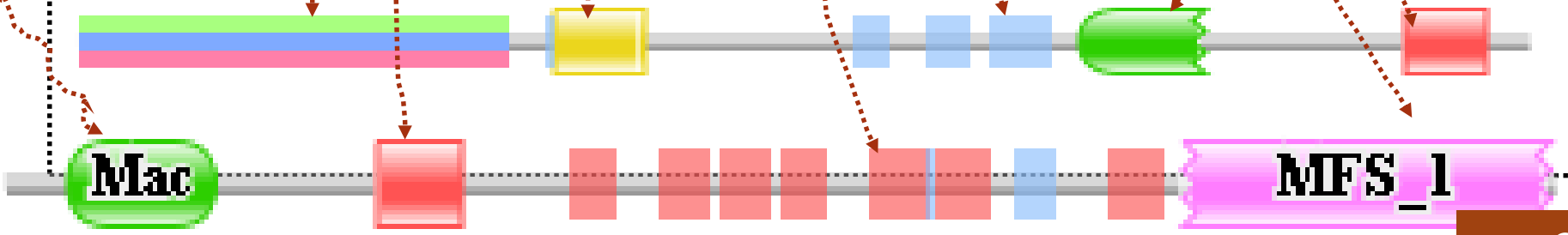
Повтор – короткая единица, нестабильная сама по себе, но образует стабильные структуры, если есть много копий.

Мотив – короткая единица структуры вне



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 А) и т.п. – *не является эволюционным доменом!*
- Домен, имеющий ID вида DUF… с номером – Domain of Unknown Function



Конец презентации

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.

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.

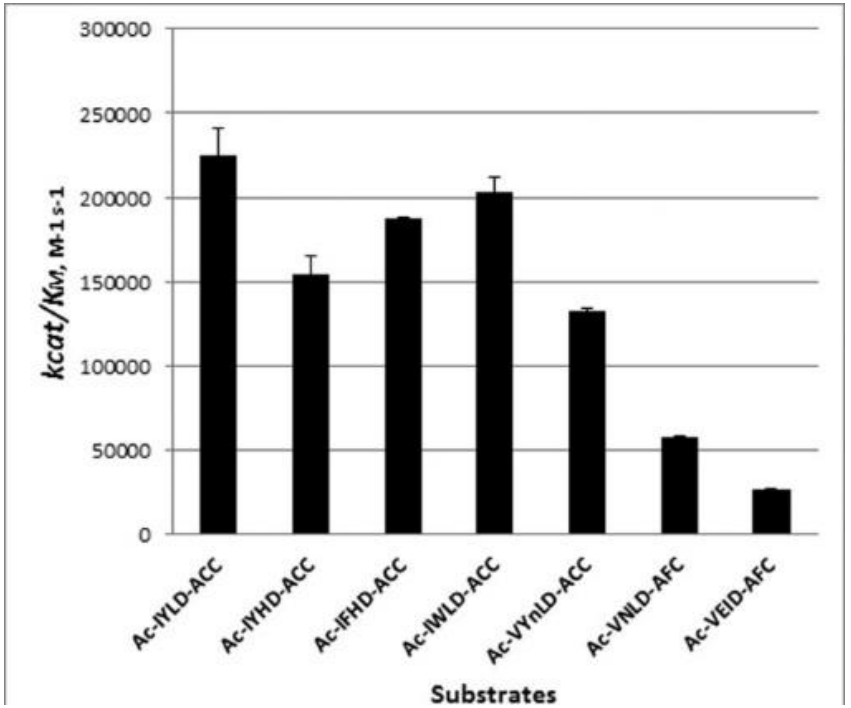
ДА: аналогичные технологии используются для поиска

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



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

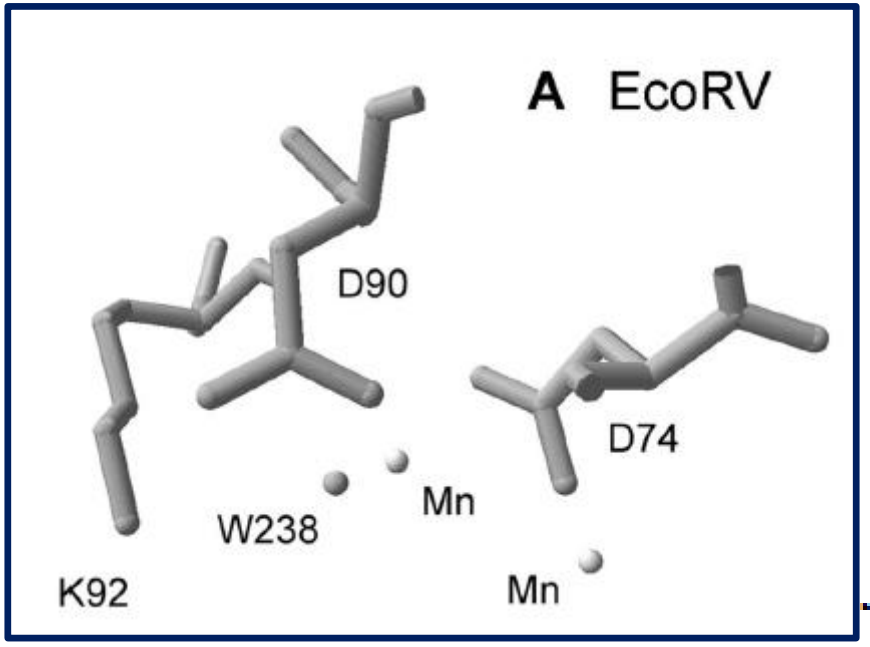
Каталитический мотив одного семейства эндонуклеаз рестрикции типа II

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

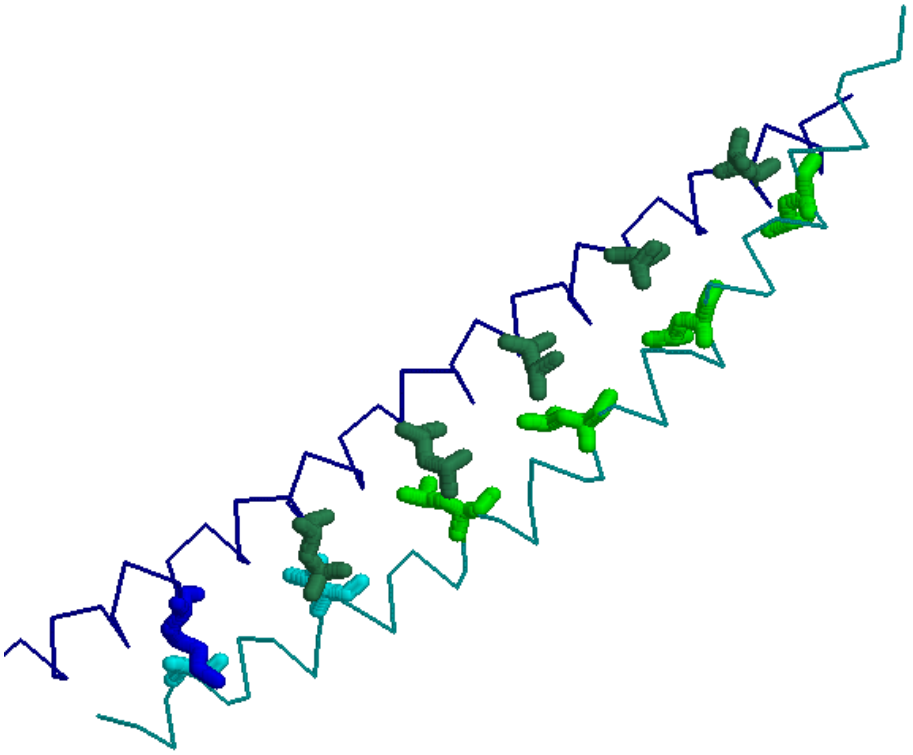
Xie et al.

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

Пример мотива 2: 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