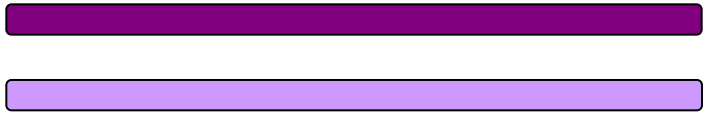
Семейства белков.

Мотив и распознающее правило.

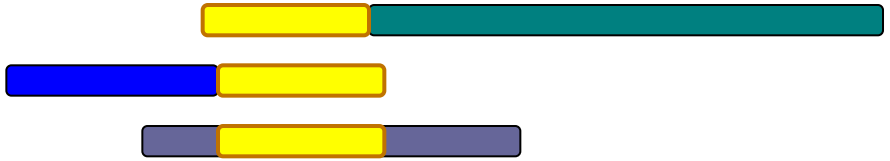
БД Pfam, InterPro.

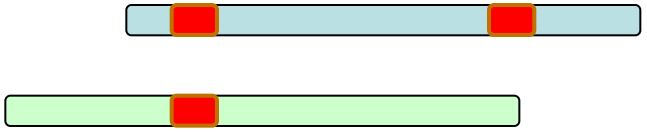
Первые же попытки классификации коллекций аминокислотных последовательностей показали существование

семейств гомологичных белков



консервативных мотивов в негомологичных белках





Мотив ?

- **Мотив в аминокислотной последовательности – набор консервативных остатков, важных для функции белка и расположенных на определенном (обычно коротком) расстоянии друг от друга в последовательности.**

- **Мотив структуры (структурный мотив) – часто встречающийся в белках элемент пространственной структуры (α-спираль, β-шпилька, β-поворот, четырёхспиральный пучок, ТИМ-баррель).**

Не в любом выравнивании легко найти мотив!

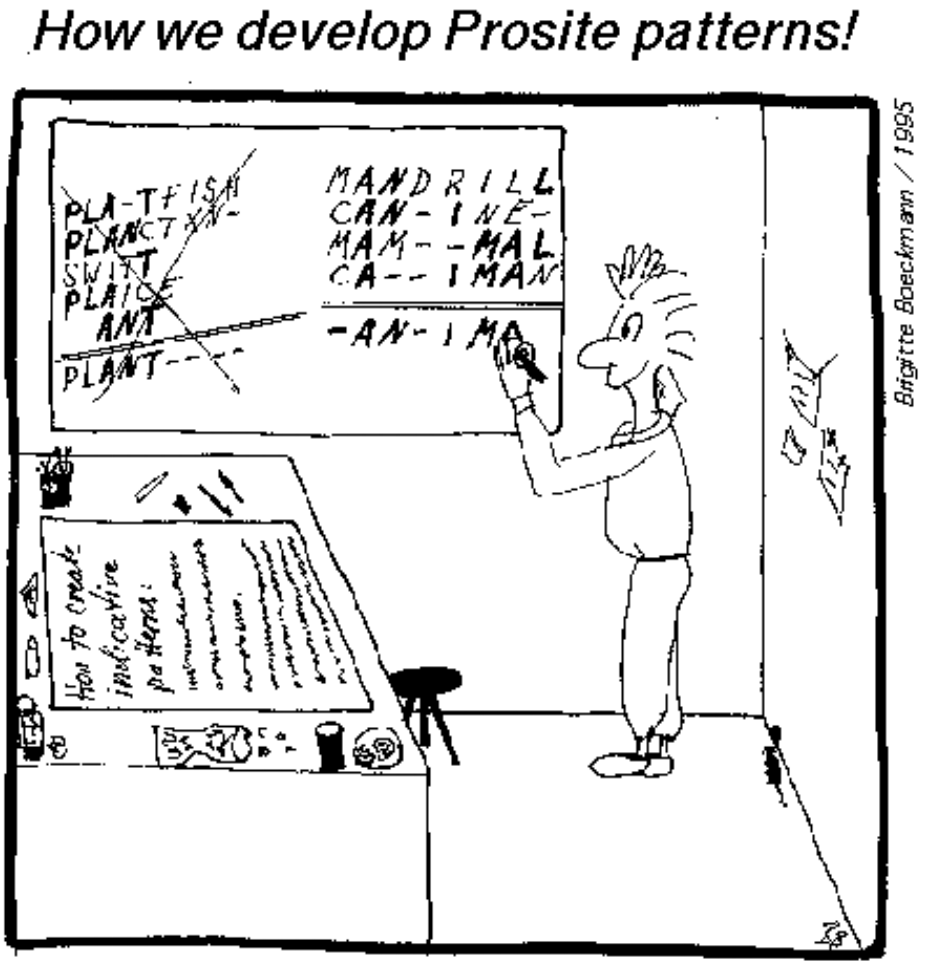
Словарик

Типы мотивов	Типы подписей (signature)
Сайт (site) ***Мотив (motif)*** **Повтор *(repeat)*** **Домен (domain)** **…..**	**Паттерн (pattern)** **Профиль–PSSM** **Профиль–HMM** **…..**

Простой пример:	
CCHC- цинксвязывающий мотив	**Подпись типа паттерн –** **C-X(2)-C-X(4)-H-X(4)-C**

Вспоминаем БД PROSITE

PROSITE — БД белковых доменов, семейств и функциональных сайтов. Содержит описание объектов + описание паттернов, профилей и правил для их обнаружения.



Релиз 20-Apr-2010
1577 документов,
1308 паттернов, 886 профилей
+ 883 ProRule

БД белковых доменов, семейств и функциональных сайтов. Содержит описание объектов + **описание паттернов**, профилей и правил для их обнаружения.

Паттерн – регулярное выражение UNIX'а:

[AC]-x-V-x(4)-{ED}

Ala или Cys- x-Val- x- x- x - x- (любой, но не Glu и не Asp)

— БД белковых доменов, семейств и функциональных сайтов. Содержит описание объектов + описание паттернов, **профилей** и правил для их обнаружения.

Позиция	1	2	3	4	5	6
Sequence 1	A	T	G	T	C	G
Sequence 2	A	A	G	A	C	T
Sequence 3	T	A	C	T	C	A
Sequence 4	C	G	G	A	G	G
Sequence 5	A	A	C	C	T	G

Pos	1	2	3	4	5	6	Сред. частота
A	0.6	0.6	-	0.4	-	0.2	0.3
T	0.2	0.2	-	0.4	0.2	0.2	0.2
G	-	0.2	0.6	-	0.2	0.6	0.27
C	0.2	-	0.4	0.2	0.6	-	0.23

Наблюдаемые частоты по позициям

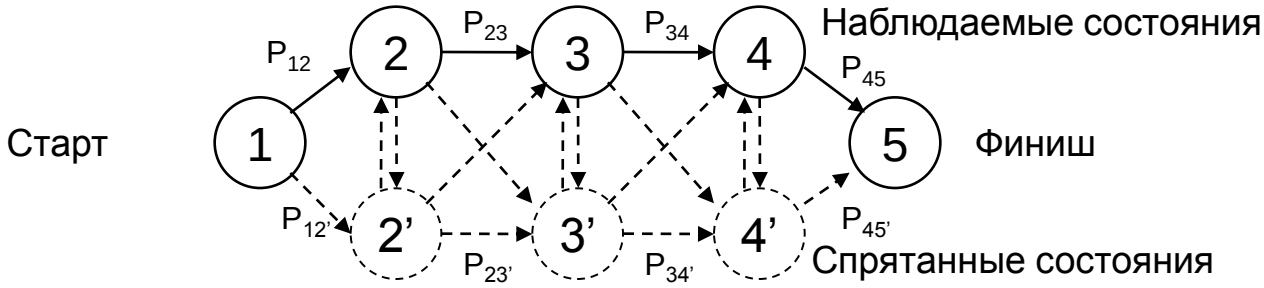
Pos	1	2	3	4	5	6
A	2.0	2.0	-	1.33	-	0.67
T	1.0	1.0	-	2.0	1.0	1.0
G	-	0.74	2.22	-	0.74	2.22
C	0.87	-	1.74	0.87	2.61	-

Нормализированные значения

Pos	1	2	3	4	5	6
A	1.0	1.0	-	0.41	-	-0.58
T	0.0	0.0	-	1.0	0.0	0.0
G	-	-0.43	1.15	-	-0.43	1.15
C	-0.2	-	0.8	-0.2	1.38	-

Переведённые в $\log_2$

Профиль HMM



HMM – вероятностная модель.
Используется
для предсказания генов и промоторов,
для предсказания вторичной структуры и трансмембранных участков белков….

HMM-профиль – обобщение PSSM, допускающая возможность вставок и делеций (в белке относительно профиля).

HMM-профиль для C2H2 из Prosite

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=28;
/DISJOINT: DEFINITION=PROTECT; N1=3; N2=26;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.6689; R2=0.02078310; TEXT='-LogE';
/CUT_OFF: LEVEL=0; SCORE=441; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=344; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: D=-20; I=-20; B1=-50; E1=-50; MI=-105; MD=-105; IM=-105; DM=-105;

                A   B   C   D   E   F   G   H   I   K   L   M   N   P   Q   R   S   T   V   W   Y   Z
/I:         B1=0; BI=-105; BD=-105;
.............
/M: SY='C'; M=-10,-20,118,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30;
/M: SY='E'; M= -5,  3,-24,  3,  6,-22,-11, -6,-20,  1,-21,-14,  4, -1,  1, -3,  5,  2,-18,-29,-15,  3;
/I:         I=-12; MI=0; MD=-30; IM=0; DM=-30;
/M: SY='E'; M= -9, -2,-26,  1, 14,-18,-17, -4,-13, -1,-11, -8, -5,-12,  4, -5, -5, -8,-12,-24, -9,  8;
/M: SY='C'; M=-10,-20,119,-30,-30,-20,-30,-30,-30,-30,-20,-20,-20,-40,-30,-30,-10,-10,-10,-50,-29,-30;
/M: SY='G'; M= -3, -1,-28, -1, -7,-28, 36,-11,-33,-11,-27,-18,  4,-15,-10,-12,  1,-13,-27,-24,-23, -9;
/M: SY='K'; M=-10, -2,-28, -3,  8,-25,-19, -7,-26, 36,-24, -8, -1,-12, 10, 27, -9, -9,-18,-19, -8,  8;
/M: SY='A'; M=  8, -7, -9,-11, -7,-17, -7,-14,-16, -6,-16,-11, -4,-15, -6, -5,  8,  4, -7,-27,-15, -7;
/M: SY='F'; M=-19,-29,-19,-37,-28, 71,-29,-17,  0,-28,  9,  0,-20,-30,-36,-19,-19, -9, -1,  9, 31,-28;
.................
/M: SY='H'; M=-20,  0,-30,  0,  0,-20,-20, 99,-30,-10,-20,  0, 10,-20, 10,  0,-10,-20,-30,-30, 20,  0;
/M: SY='Q'; M=-10,-10,-25,-12,  1,-16,-22, -2, -6,  1, -3,  6, -9,-17, 13,  3, -9, -8, -9,-19, -4,  6;
/M: SY='R'; M=-13, -8,-26, -9,  0,-19,-19, -4,-21, 20,-16, -6, -2,-17,  6, 35, -8, -7,-14,-21, -9,  0;
/I:         I=-12; MI=0; MD=-29; IM=0; DM=-29;
/M: SY='V'; M= -3,-16,-17,-21,-17, -6,-25,-20, 11,-15,  2,  3,-12,-18,-14,-14, -2,  9, 13,-25, -7,-17;
/M: SY='H'; M=-20,  0,-30,  0,  0,-20,-20, 97,-30,-10,-20,  0, 10,-20, 10,  0,-10,-20,-30,-30, 19,  0;
...................
/I:         E1=0;
```

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

Ср. паттерн

Домен

Домен – единица эволюции, структуры и функции белков.

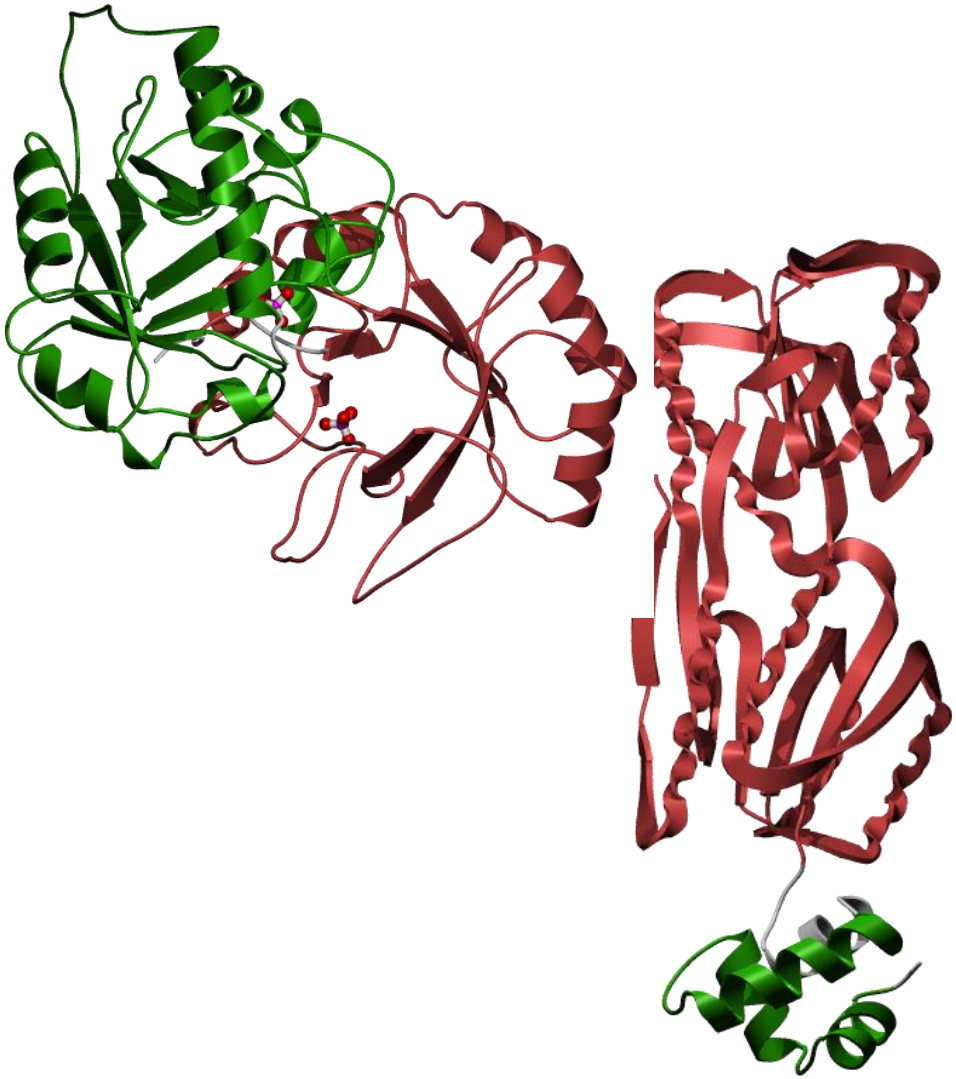
Структурный домен – компактная, относительно независимо сворачивающаяся структура.

Домен в последовательности – относительно консервативная в процессе эволюции последовательность.

Белки могут состоять из одного или многих доменов.
Один домен может содержать один или несколько мотивов в аминокислотной последовательности.
Малые мотивы могут и не входить в домены.

nitrogen fixation positive activator protein

Домены, найденные в последовательностях, часто, но далеко не всегда совпадают со структурными доменами.



Почему это интересно? Примеры доменных перестроек

223 белка 

EC 4.1.2.25

243 белка 

EC 2.7.6.3

507 белков 

EC 2.5.1.15

- **25 белков** 

- **9 белков** 

- **2 белка** 

- **12 белков** 

Pfam

- http://pfam.xfam.org/
- Большая коллекция семейств доменов (16306 на июнь 2016)

  Для каждого семейства есть множественное выравнивание и профиль-HMM .
- Удобна для анализа доменной структуры белков.

Язык Pfam :

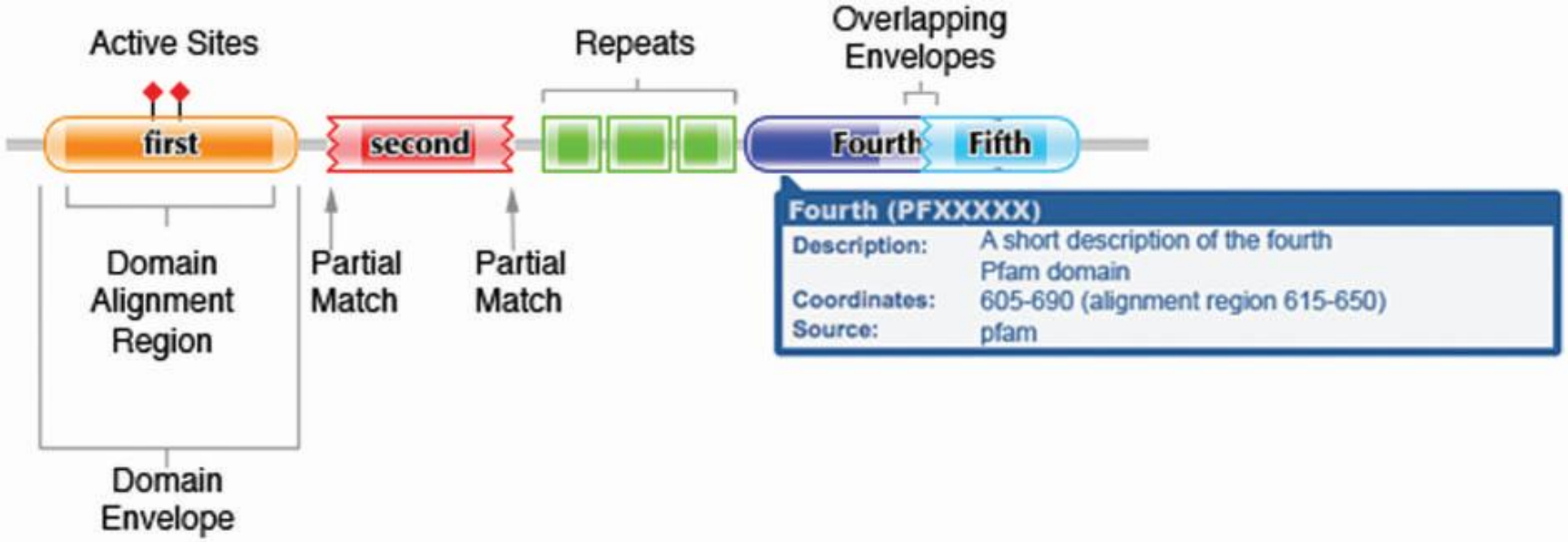
Семейство – коллекция гомологичных участков белков.

Домен – структурная единица (предположительно может самостоятельно свернуться в 3D структуру).

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

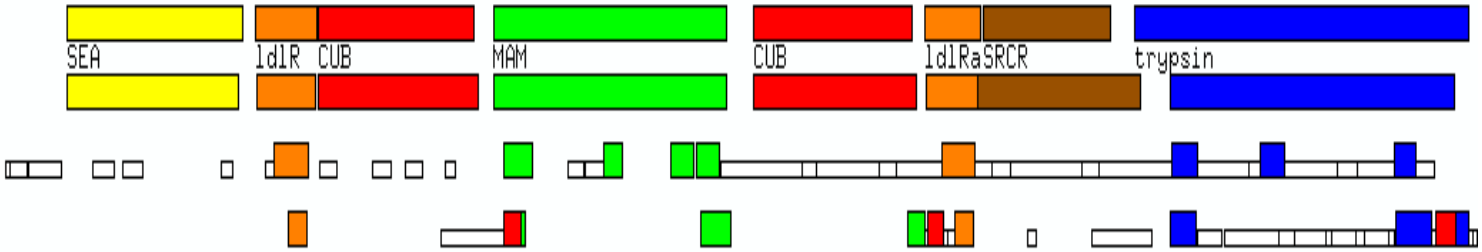
Мотив – короткая единица структуры вне глобулярных доменов.

Клан – группа родственных (в каком-нибудь смысле) записей.

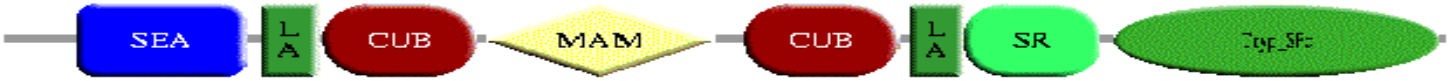


Сравнение

Pfam
Prosite
Prints
Blocks



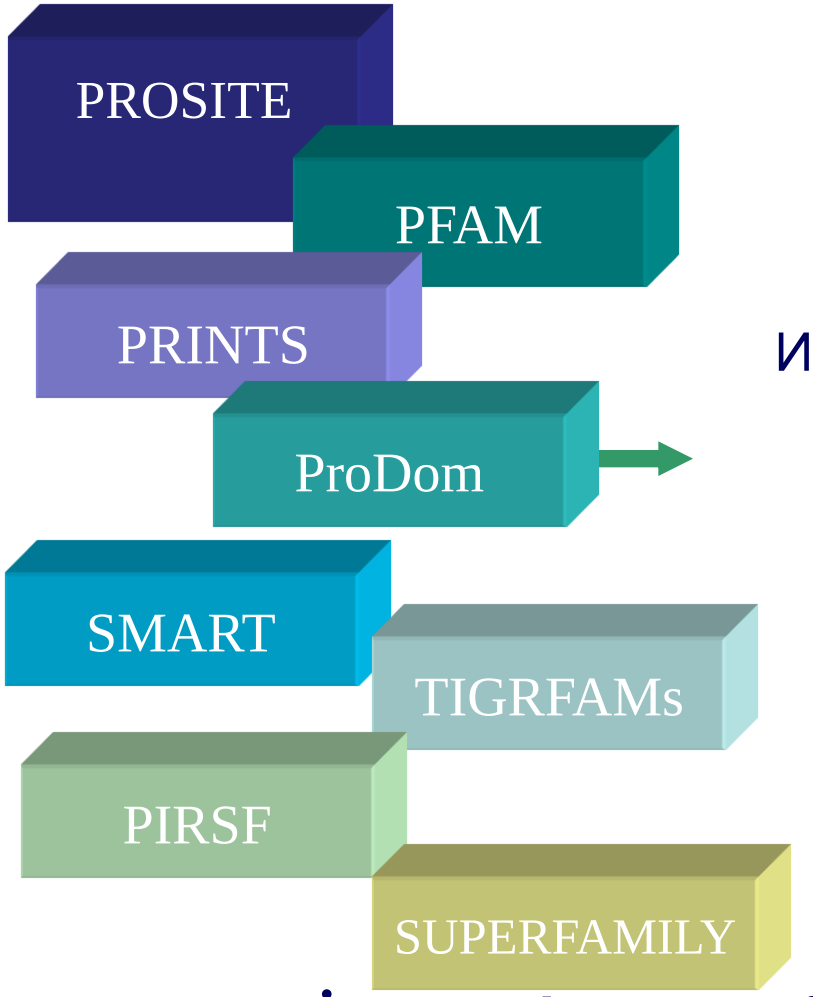
Smart



(ProDom, PIRaln, ProClass, Systers, Picasso etc. not shown)

Example: ENTK_HUMAN (Enteropeptidase precursor)

Создание интегрированной базы данных InterPro



Интегрирование родственных подписей «вручную»

InterPro entries

IPR000001-

IPR011000

InterPro- an integrated resource of protein families, domains and functional sites.

InterPro

Entry types in InterPro

- **Family** – group of evolutionarily related proteins, that share one or more domains/repeats in common.
- **Domain** – independent structural unit which can be found alone or in conjunction with other domains or repeats.
- **Repeat** – region occurring more than once that is not expected to fold into a globular domain on its own.
- **PTM** (post-translational modification) – The sequence motif is defined by the molecular recognition of this region in a cell.
- **Active site** – catalytic pockets of enzymes where the catalytic residues are known.
- **Binding site** – binds compounds but is not necessarily involved in catalysis.

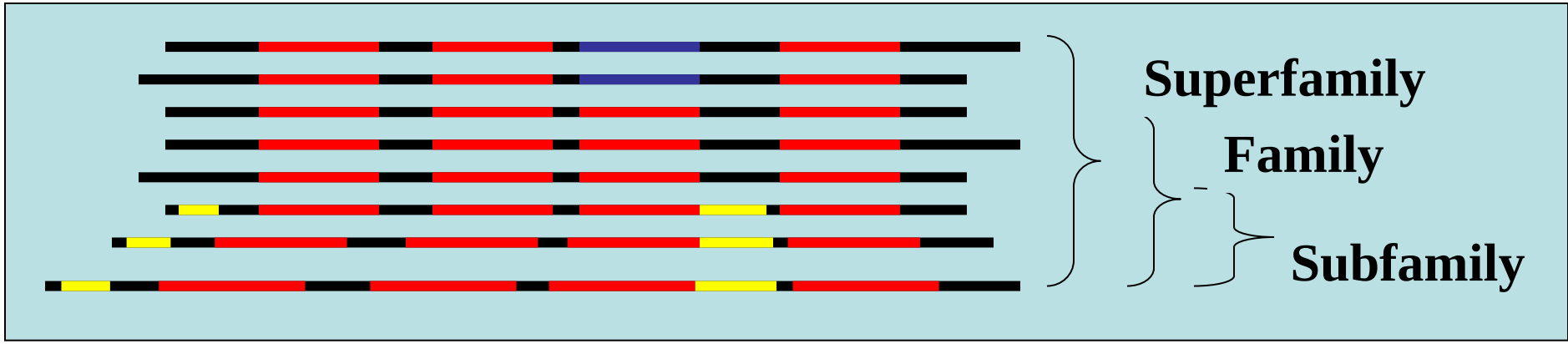
Осторожно: белковое семейство…..

Интуитивно понятно:

Семейство – группа белков, имеющая общее происхождение.
Аминокислотные последовательности выравниваются по всей длине со значимым весом и имеют сходную доменную структуру.

Мнения расходятся, когда речь идет о критериях:

- насколько должны быть похожи белки одного семейства (id ≥ 30%, id ≥ 50%) ???
- должны ли белки одного семейства иметь в точности одну и ту же доменную структуру?



Не корректнее ли говорить о семействах доменов?