

Анализ аминокислотной последовательности:

паттерны, домены, семейства

...

или

что, где и как искать?

Что будем искать ?

НАД-связывающий
сайт/центр

Сайты возможной
посттрансляционной
модификации (PTM)

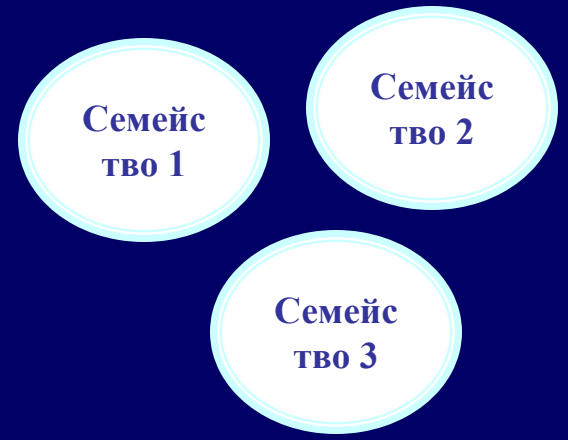


Ортологическое

семейство:

особенности
последовательностей,
характерный тип
структуры,
функции, таксономия и т.п.

«Похожие»
семейства



Основные понятия и термины

- Место, сайт(site) -
- *Мотив (motif)* –
- Домен (domain) –
- *Семейство* –
- *Суперсемейство* -
- Паттерн (pattern) –
- Позиционно специфическая матрица весов (PSSM) –
- Профиль–PSSM –
- Профиль–HMM -
- *Подпись (signature)* –
- «Отпечатки пальцев» (fingerprints) –
- Кластер -

?



Эволюции, структуры и функции белков.

Домен – компактная, относительно независимо сворачивающаяся структура, относительно консервативная в процессе эволюции.

Белки могут состоять из одного или

мног



nitrogen fixation positive activator protein

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).

В общем случае, структурные мотивы не обязательно соответствуют мотивам в аминокислотных последовательностях.

Один домен может содержать один или несколько мотивов в аминокислотной последовательности. Мотив может не входить в домены.

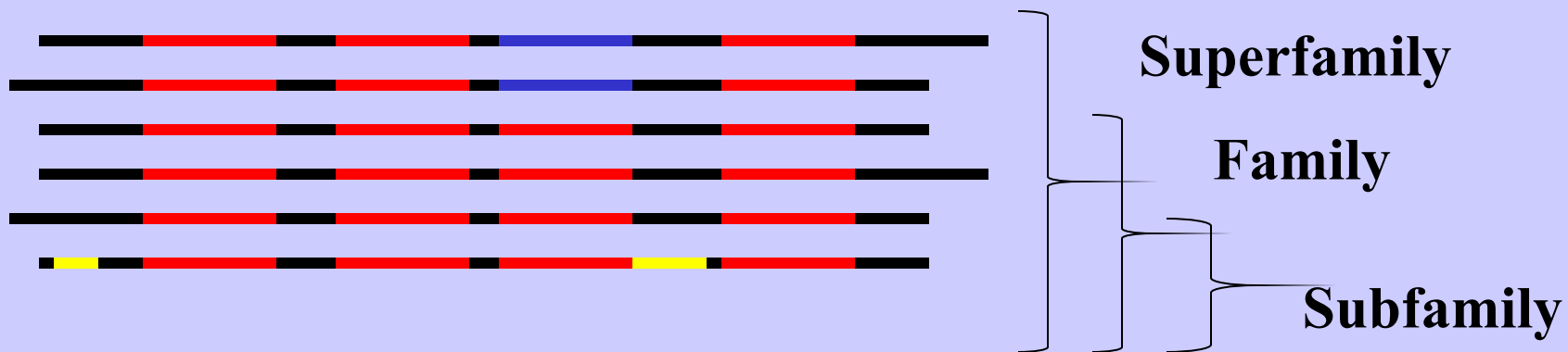
Не в любом выравнивании легко найти мотив.

Интуитивно понятно:

- Семейство - группа белков, имеющая общее происхождение, их 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 ($id \geq 30\%$, $id \geq 50\%$) ???
должны белки одного семейства выполнять одну и ту же функцию??



No comments

Число	Интегрированные базы данных , невырожденный набор последовательностей PIR, SwissProt/TrEMBL		Классификация структур (PDB)		Автоматическая кластеризация	
	InterPro, 7.2	iProClass	Scop	Cath	ProDom	
Типы доменов	2 415	7 310			391 935	
Семейств	8 035	145 300	2 327	4 023		
Суперсемейств		36 300	1 294	1 459		

Основные понятия и термины

- Место, сайт(site) -
- *Мотив (motif)* –
- Домен (domain) –
- *Семейство* –
- *Суперсемейство* -
- Паттерн (pattern) –
- Позиционно специфическая матрица весов (PSSM) –
- Профиль–PSSM –
- Профиль–HMM -
- *Подпись (signature)* –
- «Отпечатки пальцев» (fingerprints) -

?



InterPro

PRINTS

Protein Fingerprint Database

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH
tigr fams



Superfamily



ProtoMap

prosite

PIR
SuperFamily

The Protein Domain Database

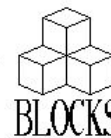
ProDom

SMART

Protein domain database logo with sequence: KMICKHKNIINLLGACTQ... VIV... KGNL... RPPGL... SHND... 137
TQL-RHSNLVQLGIVIV... 93
TQL... VIV... 93
KGF... VVSE...
QEV-SHPNVIKLLGACTS... SLR... RI... AD... 52



Blocks WWW Server



Банки белковых семейств и доменов, производные от банков аминокислотных последовательностей



Коллекции мотивов

Коллекции доменов

PROSITE , 1989

BLOCKS

PRINTS

Pfam

SMART

ProDom, 1995

SUPERFAMILY



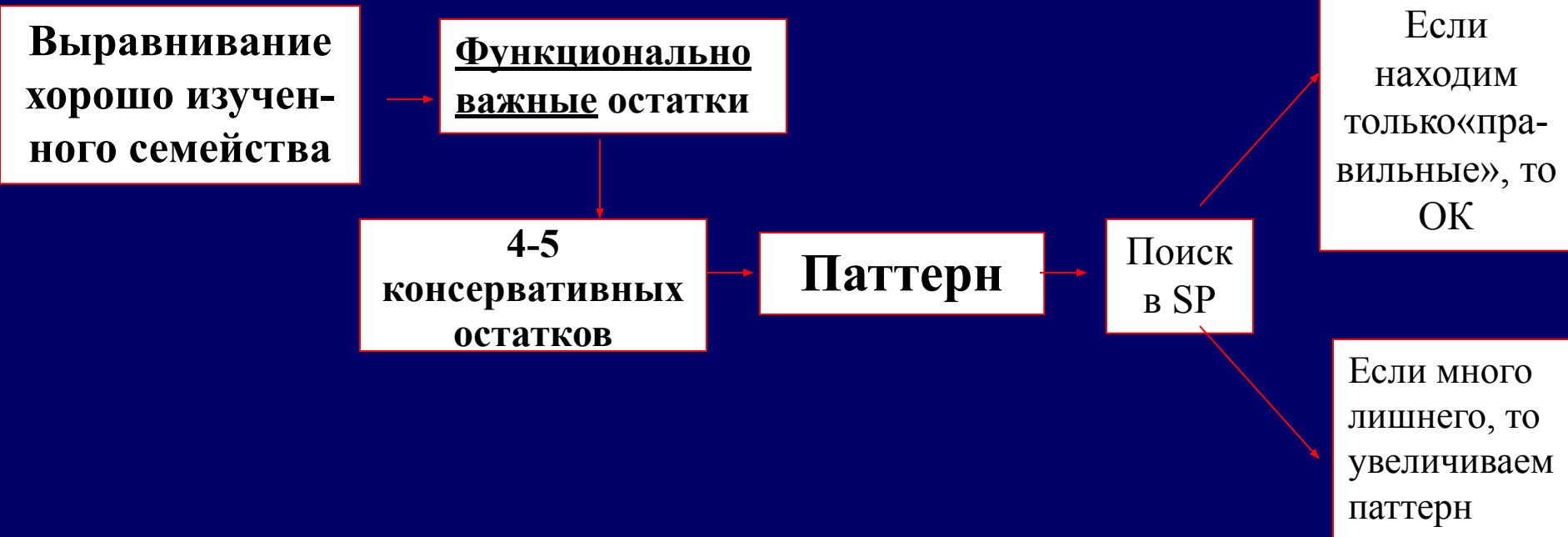
InterPro, 1999

(Integrated Resource of Protein Families)

PROSITE - биологически значимые сайты, паттерны и профили



<http://www.expasy.ch/prosite/>

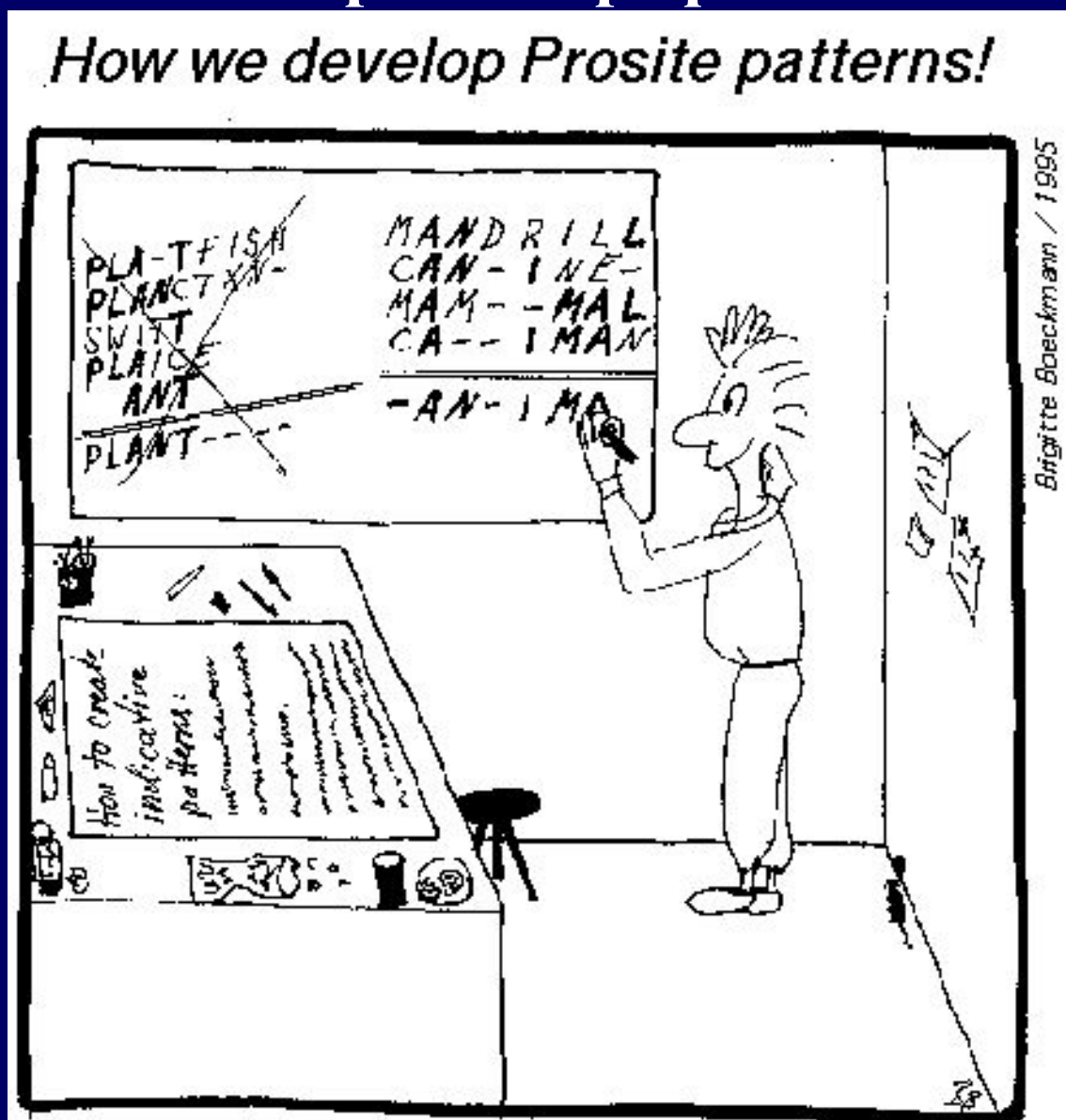


Паттерн – регулярное выражение UNIX'a:

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

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

PROSITE - биологически значимые сайты, паттерны и профили



F	K	L	L	S	H	C	L	L	V
F	K	A	F	G	Q	T	M	F	Q
Y	P	I	V	G	Q	E	L	L	G
F	P	V	V	K	E	A	I	L	K
F	K	V	L	A	A	V	I	A	D
L	E	F	I	S	E	C	I	I	Q
F	K	L	L	G	N				

Релиз 18.25,

14.04 2004

1257 документов,

1706 разных

паттернов, правил и профилей.



Профиль или весовая матрица

A	-18	-10	-1	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	10	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

Pfam



- <http://www.sanger.ac.uk/Software/Pfam/index.shtml>
- Большая коллекция множественных выравниваний, доменов, семейств и профилей-НММ для них.
- Состоит из 2-х частей:
 - PfamA – курируемая часть, покрывает 73% SWISS-Prot+TrEMBL
 - PfamB – большое число маленьких семейств из автоматически сгенерированной базы доменов ProDom, не вошедших в PfamA.
- Удобна для анализа доменной структуры белков.



Pfam



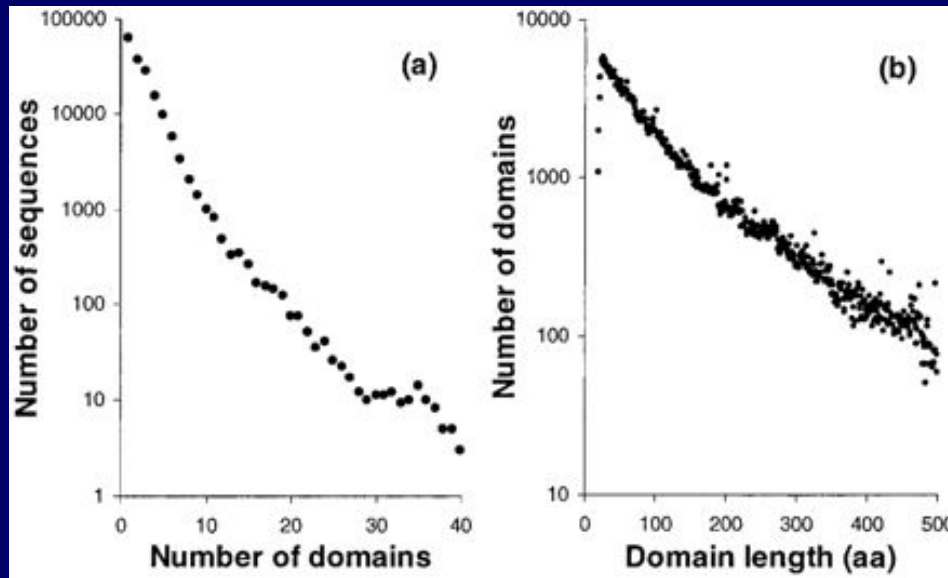
1. Множественное выравнивание (ClustalX) некоторого семейства или кластера.
2. Экспертиза и корректировка выравнивания-затравки.
3. Построение профиля-НММ для затравки.
4. Поиск в базе данных а.к.последовательностей новых членов данной группы.

ProDom

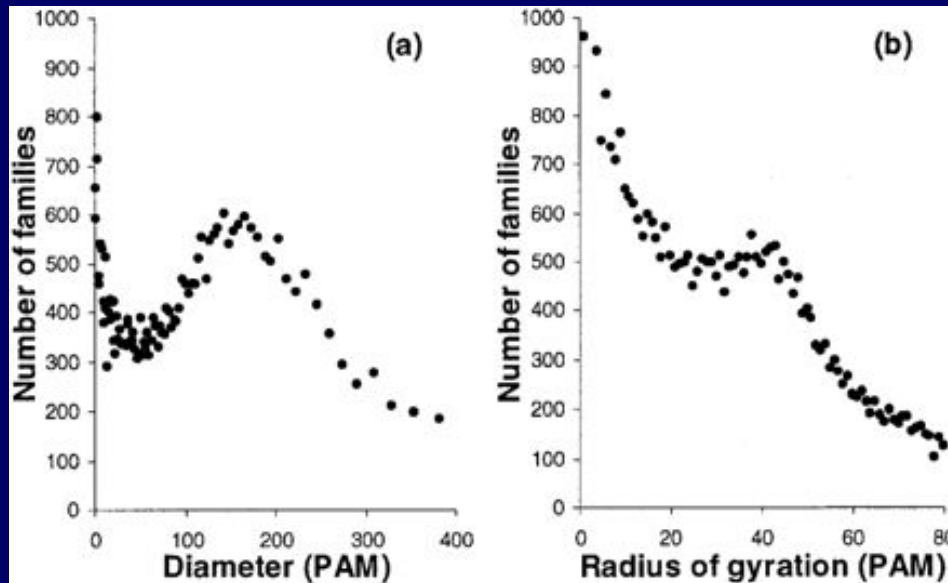


- <http://www.toulouse.inra.fr/prodom.html>
- Рассматриваются все последовательности в SWISS-Prot+TrEMBL.
- Автоматическое выделение доменов (программа DOMAINER: сначала локальное попарное выравнивание (blastp) всех против всех, затем кластеризация)
- Коллекция доменов - >150 000 семейств.
- Некоторые семейства выделены на основе выравниваний из PfamA.
- Гомогенность семейства оценивается с помощью диаметра (max расстояния между 2 доменами в семействе) и радиуса (ср.кв. расстояние между доменами и консенсусом семейства). Оба параметра измеряются в РAМ

Статистика ProDom

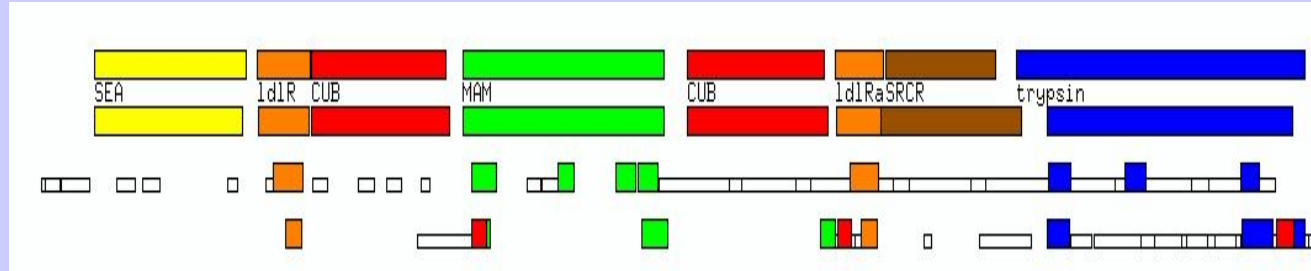


Всего – 157 167 семейств.
43 965 из них содержат
более 2 последовательностей.
Среднее число доменов в
последовательности – 2.8
Средняя длина – ~ 130
а.к. остатков

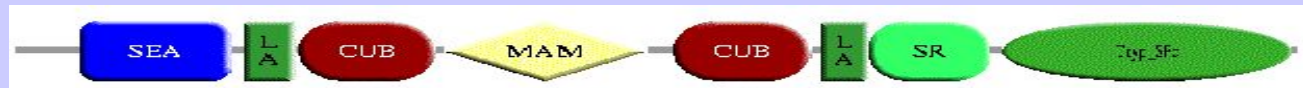


Comparison of protein family databases: an example

Pfam
Prosite
Prints
Blocks



Smart

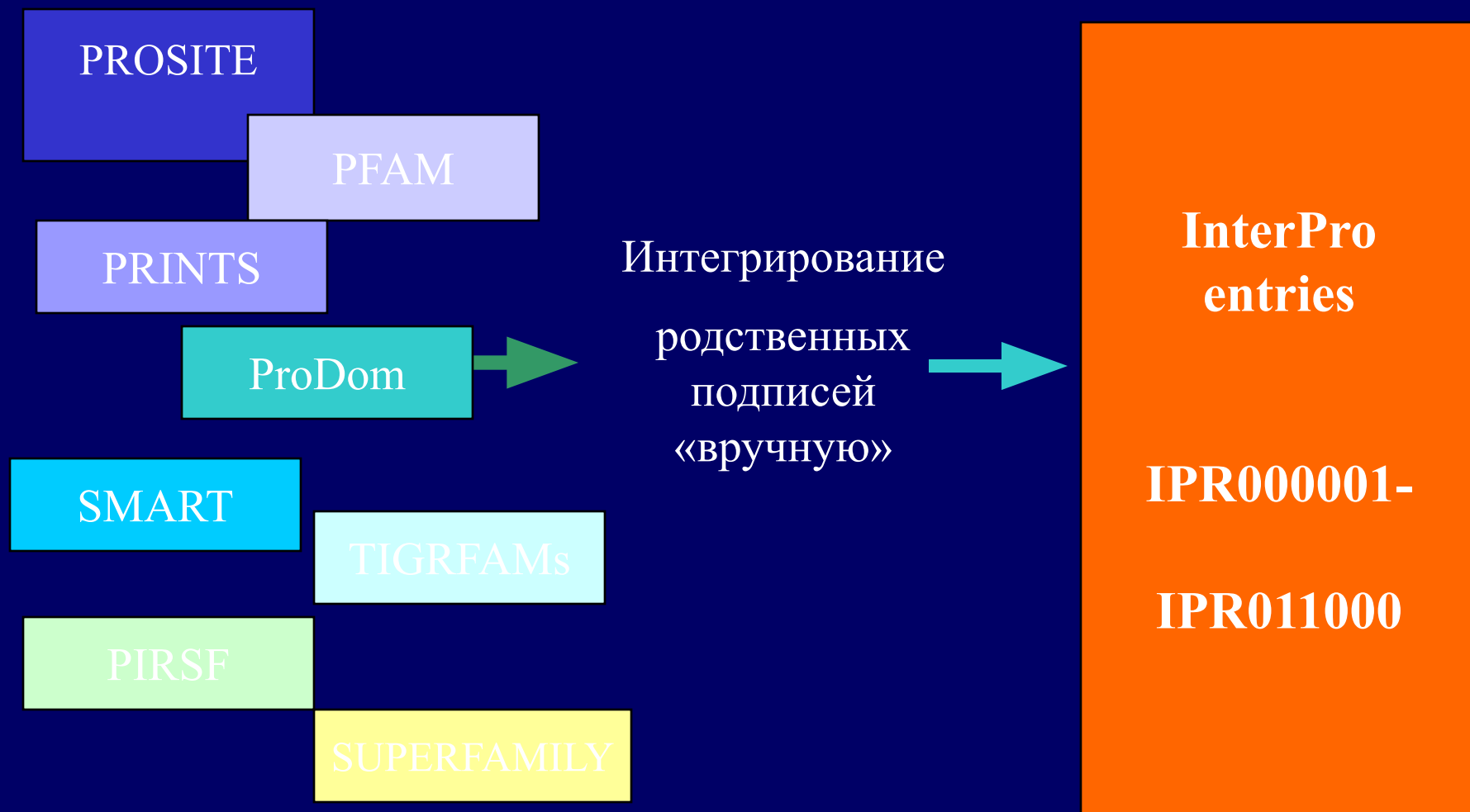


(ProDom, PIRaln, ProClass, Systers, Picasso etc. not shown)

Example: ENTK_HUMAN (Enteropeptidase precursor)

Создание интегрированной базы данных InterPro

InterPro



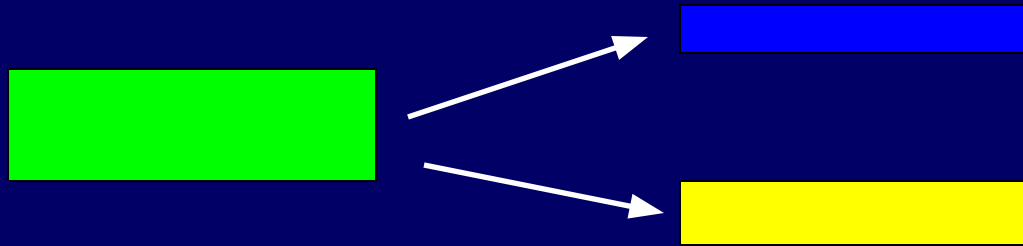
InterPro- an integrated resource of protein families, domains and functional sites.

Entry types in InterPro

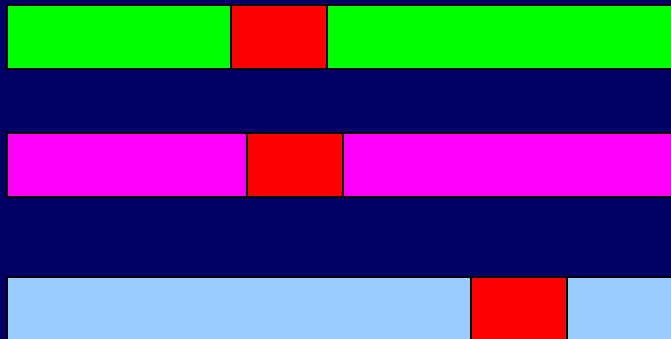
- **Family** - group of evolutionarily related proteins, that share one or more domains/repeats in common.
- **Domain** -independent structural unit which can be found alone or in conjunction with other domains or repeats.
- **Repeat** -region occurring more than once that is not expected to fold into 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.

Взаимосвязи подписей в InterPro

- **Parent/child** \longrightarrow уровень семейства

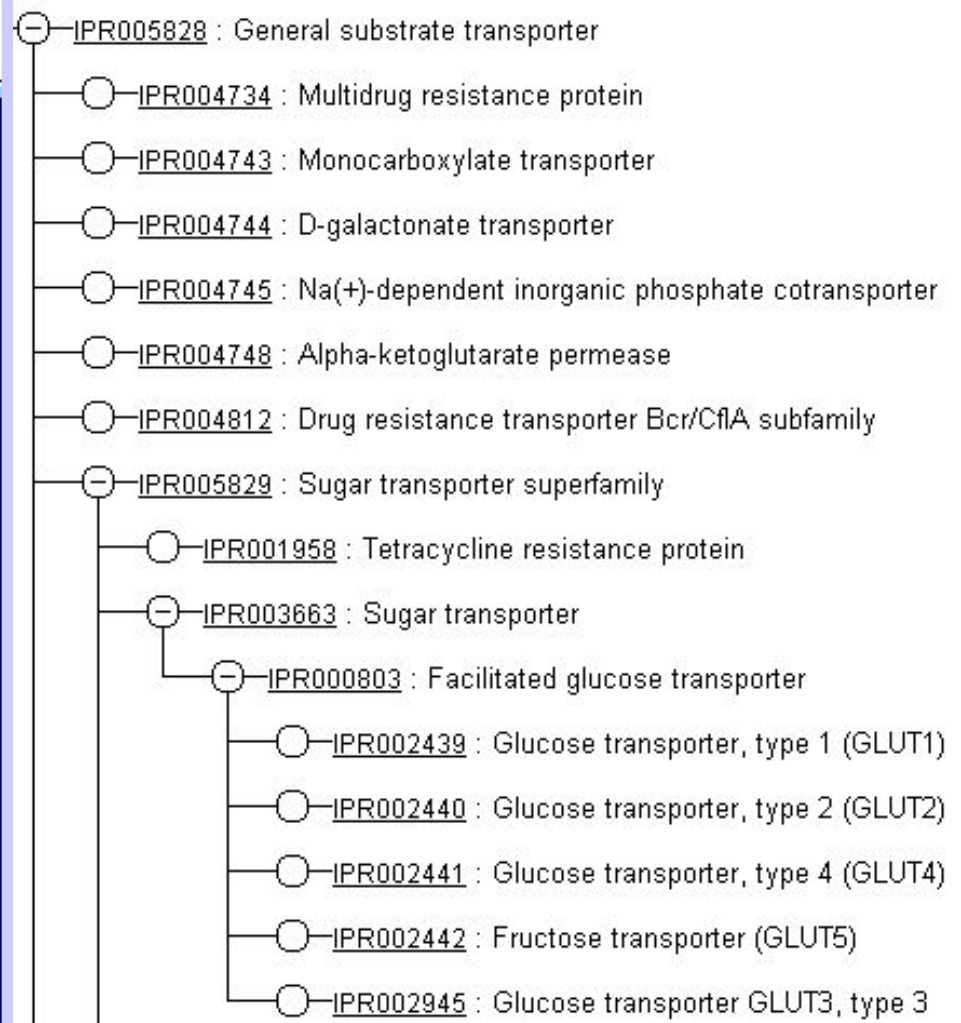


- **Contains/found in** \longrightarrow состав домена



Parent/child- family level

Children [?] [tree]	<p>IPR002439; Glucose transporter, type 1 (GLUT1)</p> <p>IPR002440; Glucose transporter, type 2 (GLUT2)</p> <p>IPR002441; Glucose transporter, type 4 (GLUT4)</p> <p>IPR002442; Fructose transporter (GLUT5)</p> <p>IPR002945; Glucose transporter GLUT3, type 3</p>
Parent [?] [tree]	<p>IPR003663; Sugar transporter</p>



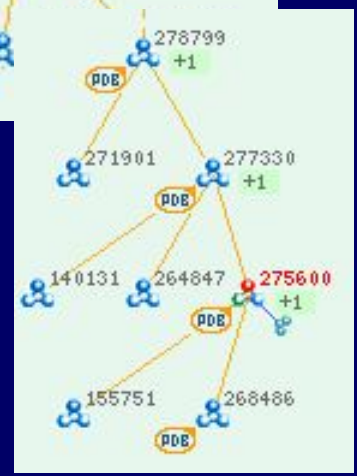
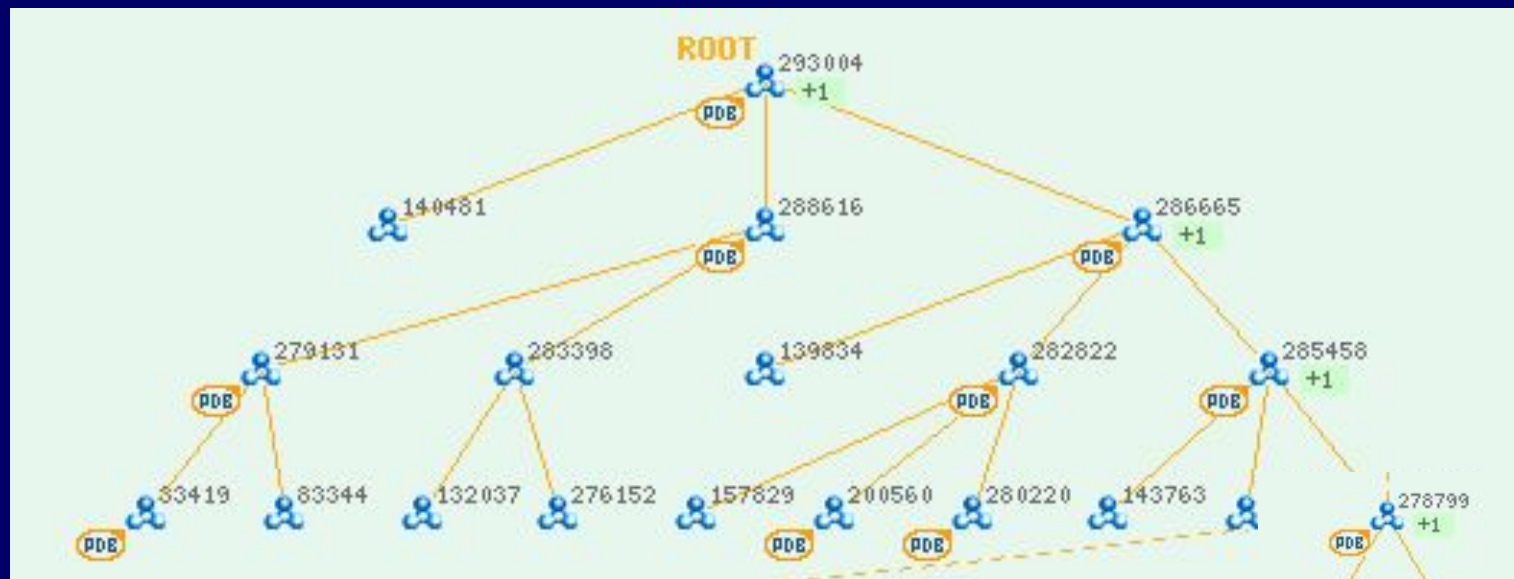
Contains/found in

UniProt-SwissProt STHA_PSEFL Q05139 GO! Scale: 10aa	IPR000205: PS50205		NAD_BINDING
	IPR000815: PR00945		HGRDTASE
	IPR001100: PR00411		PNRDRTASEI
	IPR001327: PD000139		FAD_pyr_redox
	IPR001327: PF00070		Pyr_redox
	IPR001327: PR00368		FADPNR
	IPR004099: PF02852		Pyr_redox_dim
UniProt-SwissProt GLD2_MYCTU Q07168 GO! Scale: 10aa	IPR000205: PS50205		NAD_BINDING
	IPR000447: PR01001		FADG3PDH
	IPR000447: PS00977		FAD_G3PDH_1
	IPR000447: PS00978		FAD_G3PDH_2
	IPR001100: PR00411		PNRDRTASEI
	IPR006076: PF01266		DAO
UniProt-SwissProt TRKA_MYCTU Q07194 GO! Scale: 10aa	IPR000205: PS50205		NAD_BINDING
	IPR003148: PF02254		TrkA_N
	IPR006036: PR00335		KUPTAKETRKA
	IPR006037: PF02080		TrkA_C

PROTOMAP



- <http://www.protomap.cs.huji.ac.il>
- Automatic classification of all SWISS-PROT proteins into groups of related proteins (also including TrEMBL now)
- Based on pairwise similarities
- Has hierarchical organisation for sub- and super-family distinctions
- 13 354 clusters, 5869 \geq 2 proteins, 1403 \geq 10
- Keeps SP annotation eg description, keywords
- Can search with a sequence -classify it into existing clusters



4207252

	41P4_IPUB4	3TP4_ARFT4	41P4_PEP	11TD_KLL17	41HU_YEPL1	11TD_AGRBI	41HU_MELUX	11TD_HUPON	3TFD_BOLIN	3TFD_KAT	3TFD_CREE_	3TFD_P3G
A207252												
ATP4_IPUB4	■											
ATP4_ARAT1	■											
ATP4_PEP												
ATPD_KLUUA												
ATPD_YEAGT												
ATPD_AGR3C												
ATPD_MEUCR												
H_HU_NUMHR												
ATPD_BOVIN												
ATPD_PAT												
ATPD_CACIL												
ATPD_PTR												

ATP4_IPUB4 & ATP4_ARAT
-< 11-11