HMM-профили

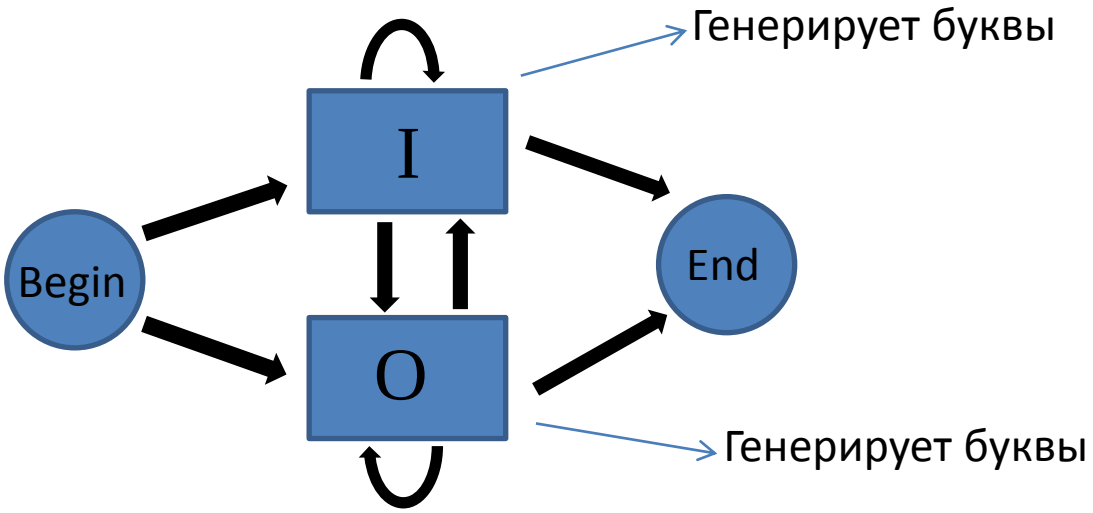
17 апреля 2026

Что такое HMM

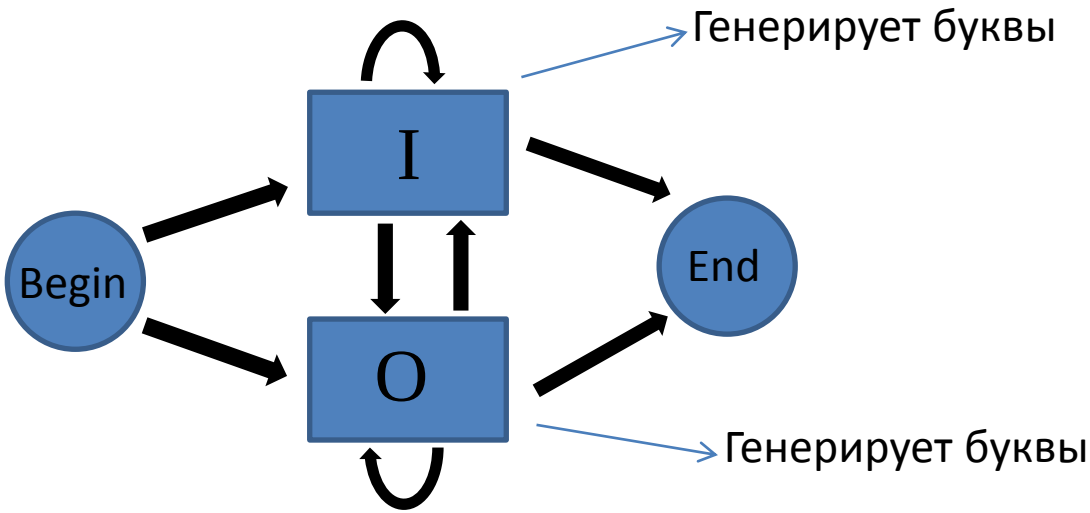
HMM = Hidden Markov model = скрытая марковская модель

Она состоит из нескольких **скрытых состояний**, между которыми заданы **вероятности переходов**. Кроме того, часть состояний **эмиссионные**, для них заданы «эмиссионные вероятности», то есть вероятности генерации символов (для биоинформатических HMM символы — либо нуклеотиды A, C, G, T, либо аминокислоты).

Пример: скрытые состояния — внутримембранное (I) и внемембранное (O). Эмиссионные вероятности для состояния I соответствуют частотам остатков в трансмембранных сегментах, а для состояния O — вне их.



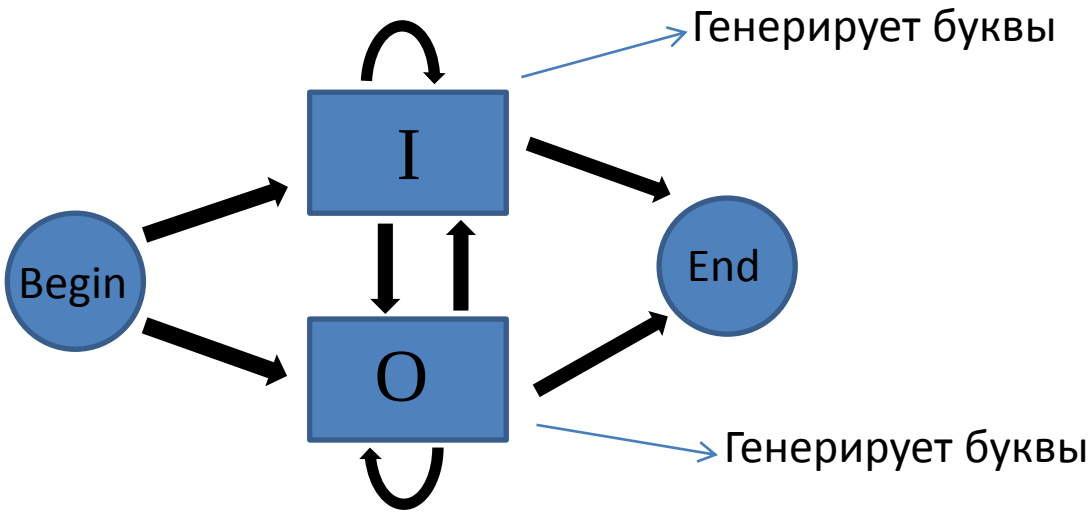
Как применить HMM?



Можно применять HMM для генерации случайных последовательностей
(в данном случае имитирующих последовательности белков с трансмембранными сегментами)

- Начинаем с Begin,
- на каждом шаге разыгрываем, по какой стрелке идти
(в соответствии с вероятностями переходов)
- когда попадаем в эмиссионное состояние, «печатаем» букву

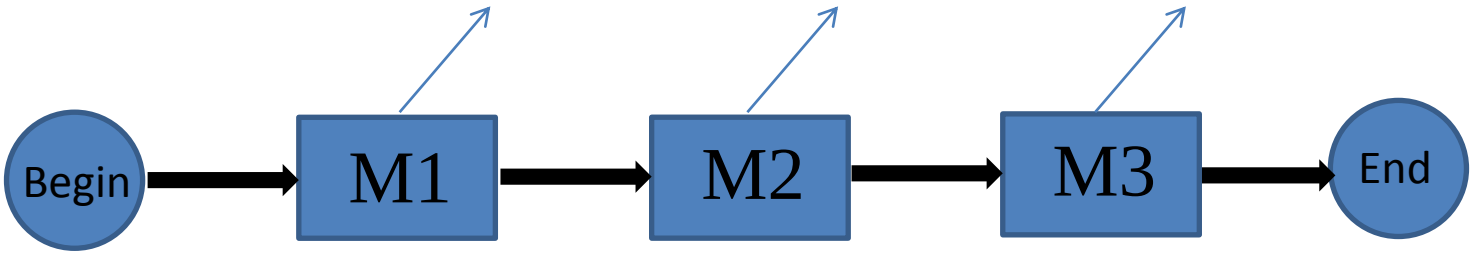
Как применить HMM?



Можно применять HMM для генерации случайных последовательностей
(в данном случае имитирующих последовательности белков с трансмембранными сегментами)

… но гораздо чаще их используют для **разметки** реальных последовательностей: находим такой путь по состояниям, при котором вероятность генерации данной последовательности максимальна (для этого имеются эффективные алгоритмы)
Например, данная HMM может быть использована для предсказания трансмембранных участков.

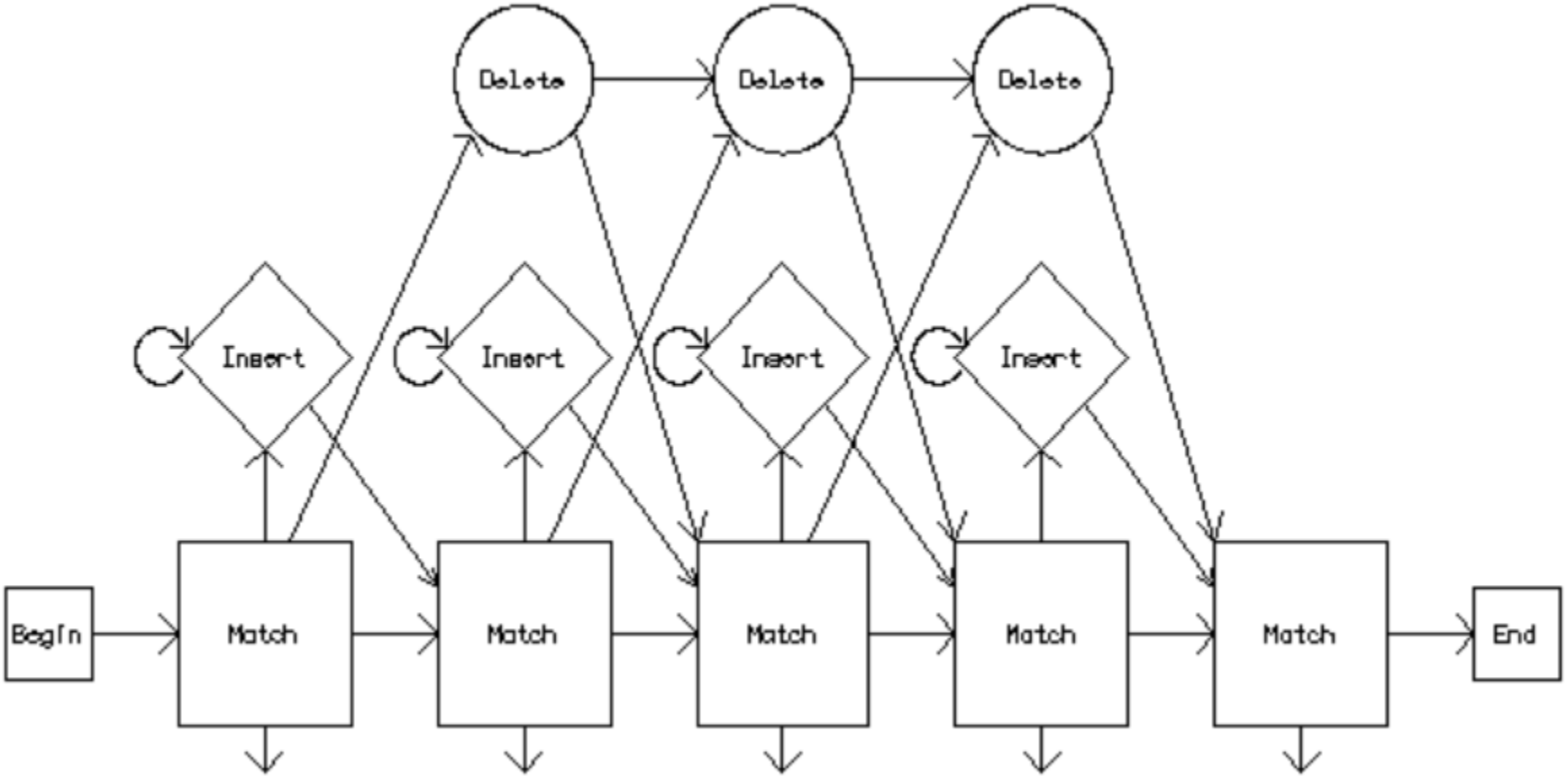
PWM в форме HMM



Если все вероятности переходов сделать равными 1, а эмиссионные вероятности — пропорциональными частотам букв в позициях мотива (с псевдоотсчётами!), то получится полный эквивалент PWM!

Только зачем?

Схема HMM для профилей, допускающих вставки и делеции



https://bip.weizmann.ac.il/education/materials/gcg/hmmanalysis.html

Пакеты

Имеется два популярных пакета работы с HMM-профилями:

PFtools https://github.com/sib-swiss/pftools3

HMMer http://hmmer.org/ , https://github.com/EddyRivasLab/hmmer

HMMer установлен на kodomo

Форматы хранения профилей различаются, но суть одна (и существуют программы-конверторы для переформатирования)

Профили в формате PFtools используются в банке Prosite

Профили в формате HMMer используются в банке Pfam

Во всех форматах вместо вероятностей приведены их логарифмы либо (для эмиссий) логарифмы отношений правдоподобия.

Можно даже забыть про вероятности, и считать, что у каждого варианта прохождения модели есть **вес**, складывающийся из весов переходов и весов эмиссий.

Обучение HMM-профиля

Выравнивание  HMM-профиль

Стадии:

• Взвешивание последовательностей (!)

• Позиции выравнивания → скрытые состояния типа Match

• Определение эмиссионных весов для состояний типа Match (логарифмы отношений частот букв в колонке к фоновым частотам)

• Определение вероятностей переходов (Match→Insert, Match→Delete, и т.д.)

• Калибровка профиля на случайных последовательностях

В пакете HMMer это делает программа hmmbuild

Взвешивание

Из help'а программы hmmsearch

```
Alternative relative sequence weighting strategies:
  --wpb     : Henikoff position-based weights  [default]
  --wgsc    : Gerstein/Sonnhammer/Chothia tree weights
  --wblosum : Henikoff simple filter weights
  --wnone   : don't do any relative weighting; set all to 1
  --wgiven  : use weights as given in MSA file
  --wid <x> : for --wblosum: set identity cutoff  [0.62]  (0<=x<=1)
```

Применение HMM-профиля

… в принципе такое же, как PWM и PSSM:

Профиль + последовательность → **вес** (score) «выравнивания» профиля на последовательность

При поиске в базе данных ставится **порог** на вес или на E-value

Находки выше порога предсказываются как принадлежащие семейству

Профиль PFtools 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

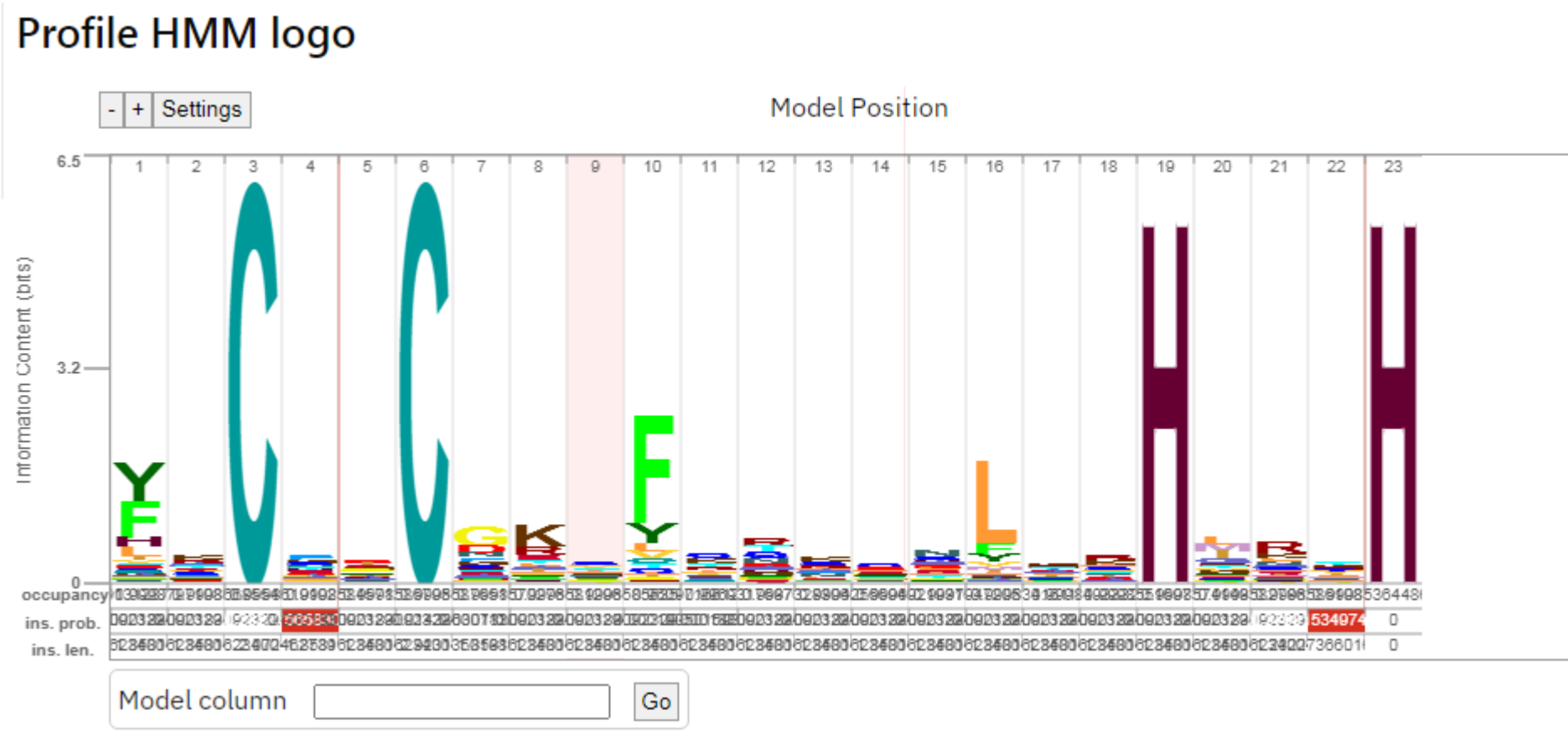
Профиль HMMer для C2H2 из Pfam

```
HMMER3/f [3.3 | Nov 2019]
NAME  zf-C2H2
ACC   PF00096.33
DESC  Zinc finger, C2H2 type
LENG  23
ALPH  amino
RF    no
MM    no
CONS  yes
CS    no
MAP   yes
DATE  Wed May 28 15:05:26 2025
NSEQ  151
EFFN  151.000000
CKSUM 1962597926
GA    25.2 15.8;
TC    25.2 15.8;
NC    25.1 15.7;
BM    hmmbuild HMM.ann SEED.ann
SM    hmmsearch -Z 90746521 --cpu 8 -E 1000 HMM pfamseq
STATS LOCAL MSV       -6.9231  0.73169
STATS LOCAL VITERBI   -7.1290  0.73169
STATS LOCAL FORWARD   -3.3875  0.73169
```

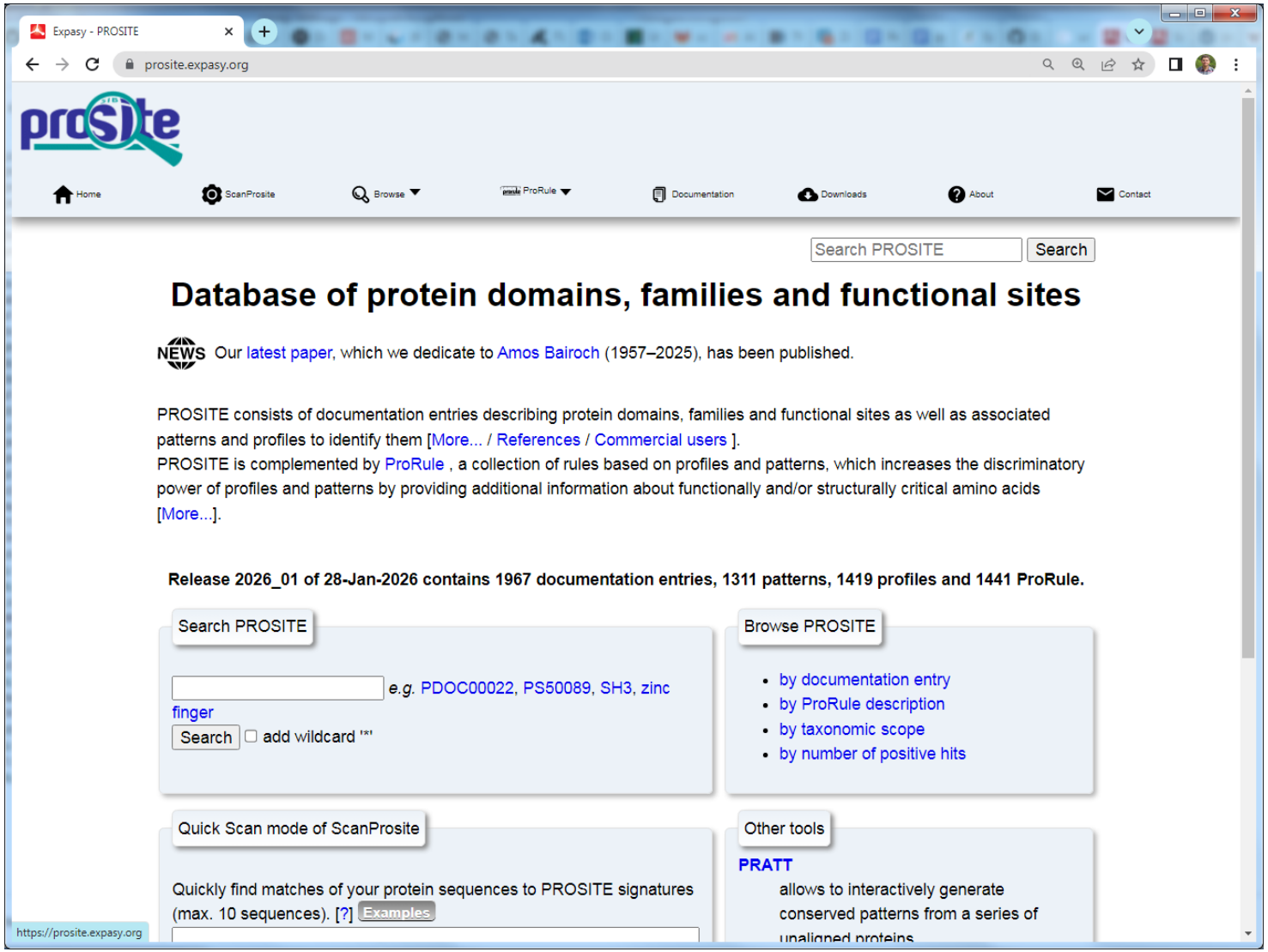
HMM	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
	m->m	m->i	m->d	i->m	i->i	d->m	d->d													
COMPO	3.26868	2.37928	3.33162	3.16806	2.84668	3.29992	2.19881	3.41335	2.49788	2.65516	3.91895	3.18686	3.78164	3.21073	2.68459	2.73141	2.84619	3.26916	5.01512	3.22441
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89801	2.37887	2.77519	2.98518	4.58477	3.61503
	0.00027	8.60552	9.32786	0.61958	0.77255	0.00000	*													
1	5.84261	3.85403	3.76302	7.18583	1.17711	4.48788	2.54686	3.44314	4.12799	2.56472	4.71694	4.95208	7.37507	4.36689	6.98197	3.78243	3.72879	3.70623	4.82606	1.14088
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89801	2.37887	2.77519	2.98518	4.58477	3.61503
	0.00027	8.60552	9.32786	0.61958	0.77255	0.48576	0.95510													
2	3.19269	8.14813	3.61663	2.39341	3.94421	6.41199	4.02608	3.23731	1.81209	4.05473	3.86561	3.53163	3.09854	2.32596	2.30638	2.58786	2.34317	2.47591	8.58097	3.39401
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89801	2.37887	2.77519	2.98518	4.58477	3.61503
	0.00027	8.60552	9.32786	0.61958	0.77255	0.48576	0.95510													
3	10.09845	0.00066	10.01335	10.10669	10.65829	9.07075	10.62028	11.10576	10.29984	10.11699	11.50026	10.45947	9.70347	10.61742	10.03071	10.54064	10.48005	10.75798	10.90209	10.91128
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	3.29354	2.67741	2.69355	4.24690	2.90347	2.73739	3.18146	2.89801	2.37887	2.77519	2.98518	4.58477	3.61503
	0.00027	8.60552	9.32786	0.61958	0.77255	0.48576	0.95510													
4	3.32328	4.88587	2.72603	2.60509	3.84314	3.04565	2.57977	6.98384	2.20565	3.44462	4.74520	2.49681	1.89415	3.21094	3.48069	2.39213	2.55238	3.69157	5.01568	4.15127
	2.68527	4.42954	2.77377	2.73070	3.45292	2.41027	3.72060	3.29326	2.67445	2.70083	4.25418	2.91075	2.72863	3.18130	2.90032	2.38303	2.77949	2.98462	4.53449	3.60197
	0.26457	1.45940	9.32786	0.72874	0.65878	0.48576	0.95510													

*Все числа положительные — это **не** логарифмы отношений правдоподобия!*

LOGO профиля
(из Pfam)



Банк Prosite



https://prosite.expasy.org/

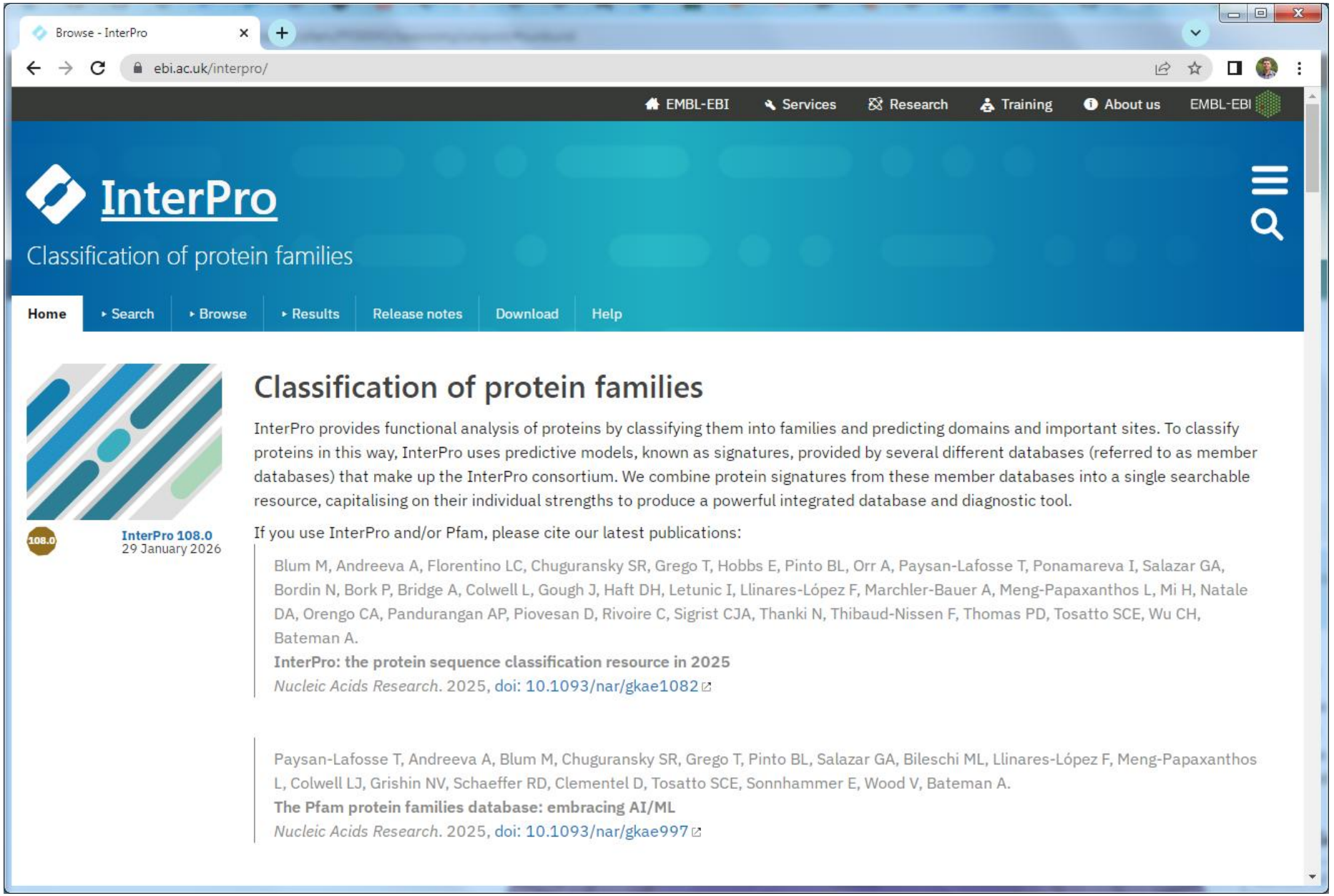
InterPro

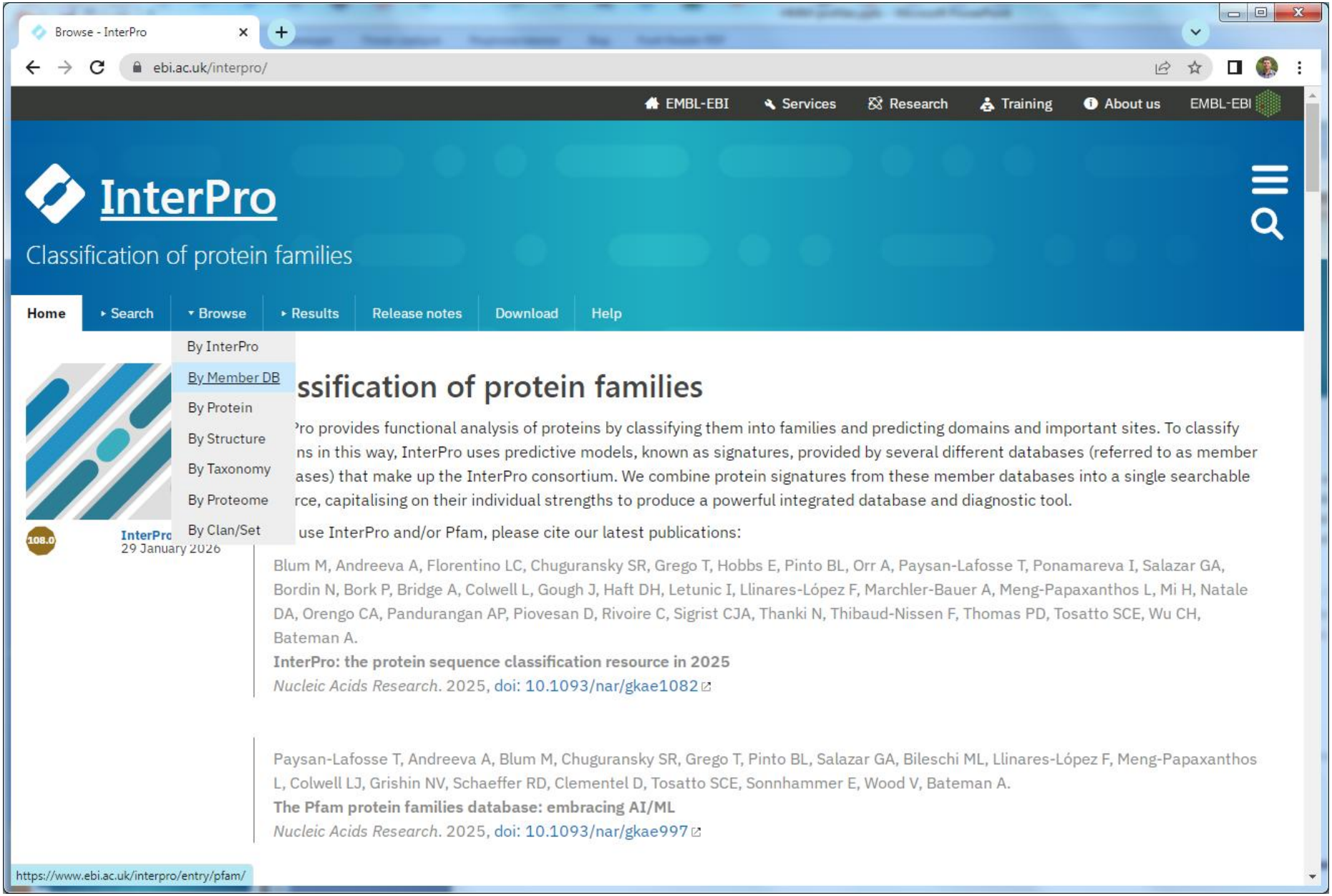
Prosite, Pfam и ещё несколько похожих по замыслу ресурсов интегрированы в InterPro https://www.ebi.ac.uk/interpro/

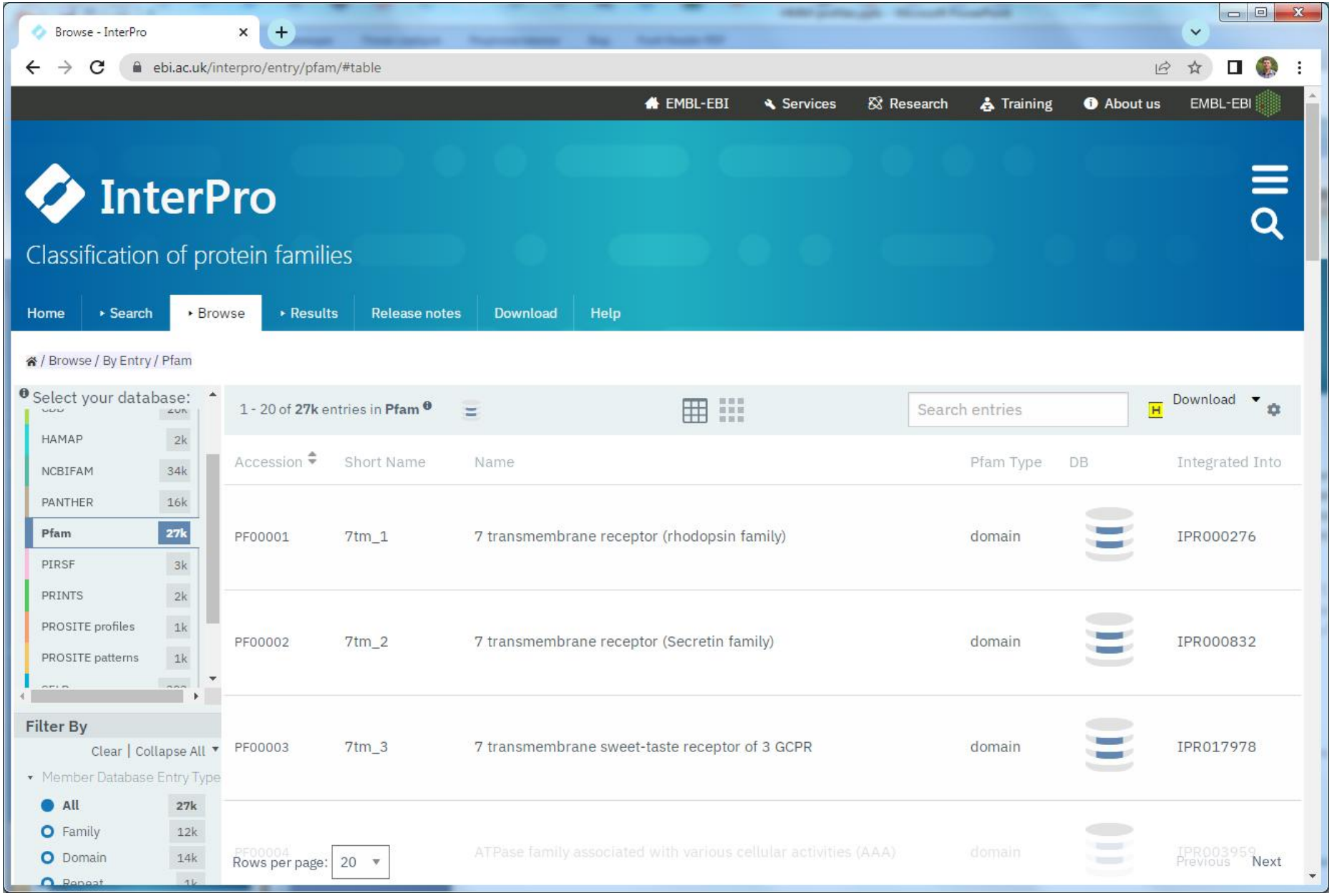
Prosite — коллекция **паттернов** и **HMM-профилей** + описания семейств белков, «правила», характеристики качества паттернов и профилей и др. Записи Prosite могут описывать как набольшие мотивы, так и целые домены

Pfam — прежде всего **коллекция HMM-профилей**, описывающих **семейства эволюционные домены**. Для каждого семейства, помимо профиля, доступны “seed alignment” (то, по которому построен профиль) и “full alignment” (полученный подравниванием всех находок в референсных протеомах с помощью профиля), а также разнообразная информация о семействе

Prosite можно использовать и через веб-интерфейс InterPro, и через собственный, а Pfam несколько лет назад перестал поддерживать свой веб-интерфейс (но продолжает обновлять свою коллекцию семейств)


Browse - InterPro
ebi.ac.uk/interpro/
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
InterPro 108.0
29 January 2026
Classification of protein families
InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.
If you use InterPro and/or Pfam, please cite our latest publications:
Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinares-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A.
InterPro: the protein sequence classification resource in 2025
Nucleic Acids Research. 2025, doi: 10.1093/nar/gkae1082
Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, Salazar GA, Bileschi ML, Llinares-López F, Meng-Papaxanthos L, Colwell LJ, Grishin NV, Schaeffer RD, Clementel D, Tosatto SCE, Sonnhammer E, Wood V, Bateman A.
The Pfam protein families database: embracing AI/ML
Nucleic Acids Research. 2025, doi: 10.1093/nar/gkae997


Browse - InterPro
ebi.ac.uk/interpro/
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
By InterPro
By Member DB
By Protein
By Structure
By Taxonomy
By Proteome
By Clan/Set
ssification of protein families
ro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify
ns in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member
ases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable
rce, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.
use InterPro and/or Pfam, please cite our latest publications:
Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinares-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A.
InterPro: the protein sequence classification resource in 2025
Nucleic Acids Research. 2025, doi: 10.1093/nar/gkae1082
Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, Salazar GA, Bileschi ML, Llinares-López F, Meng-Papaxanthos L, Colwell LJ, Grishin NV, Schaeffer RD, Clementel D, Tosatto SCE, Sonnhammer E, Wood V, Bateman A.
The Pfam protein families database: embracing AI/ML
Nucleic Acids Research. 2025, doi: 10.1093/nar/gkae997
29 January 2026
https://www.ebi.ac.uk/interpro/entry/pfam/


Browse - InterPro
ebi.ac.uk/interpro/entry/pfam/#table
EMBL-EBI
Services
Research
Training
About us
EMBL-EBI
InterPro
Classification of protein families
Home
Search
Browse
Results
Release notes
Download
Help
/ Browse / By Entry / Pfam
Select your database:
HAMAP 2k
NCBIFAM 34k
PANTHER 16k
Pfam 27k
PIRSF 3k
PRINTS 2k
PROSITE profiles 1k
PROSITE patterns 1k
Filter By
Clear | Collapse All
Member Database Entry Type
All 27k
Family 12k
Domain 14k
1 - 20 of 27k entries in Pfam
Search entries
Download
Accession
Short Name
Name
Pfam Type
DB
Integrated Into
PF00001 7tm_1 7 transmembrane receptor (rhodopsin family) domain IPR000276
PF00002 7tm_2 7 transmembrane receptor (Secretin family) domain IPR000832
PF00003 7tm_3 7 transmembrane sweet-taste receptor of 3 GCPR domain IPR017978
PF00004 ATPase family associated with various cellular activities (AAA) domain IPR003959
Rows per page: 20
Previous
Next

Напоминаем про эволюционные домены

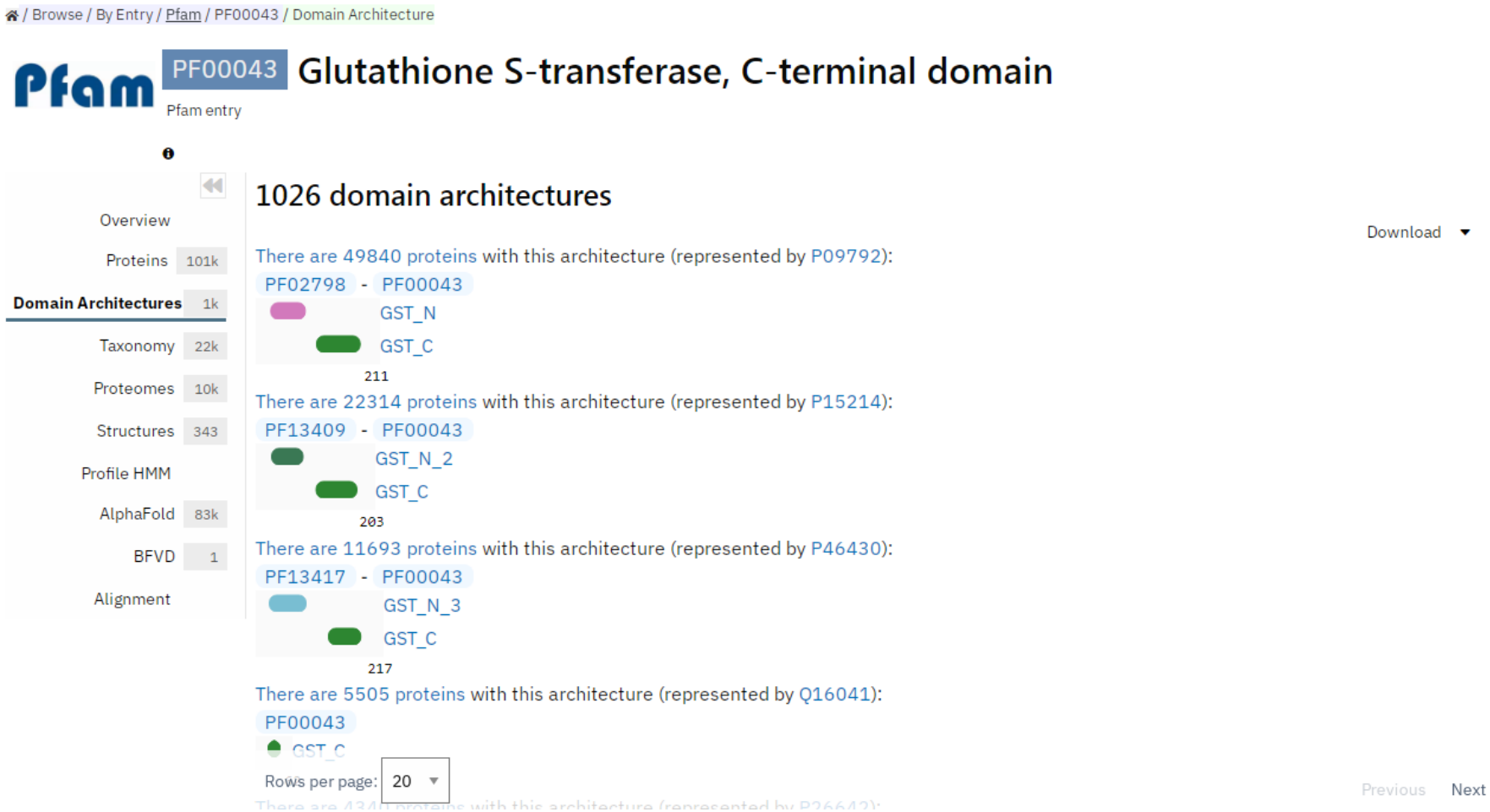
Эволюционный домен – достаточно длинный (более многих десятков а.к.о.) участок предкового белка, который на пути ко всем белкам-потомкам эволюционировал только путём локальных мутаций (замен, небольших вставок и делеций, без крупных перестроек)

При этом белки-потомки могли претерпевать крупные перестройки, не затрагивающие домен, например, добавление или удаление больших участков вне домена, перестановка доменов, приобретение новых доменов, утрата доменов

В белке может быть один домен, два или много.

Доменная архитектура = последовательность доменов в белке.
Домену дают название. Собирают представителей домена из всех белков, в которых их удаётся найти и строят выравнивание участков, соответствующих домену. Семейство доменов – совокупность таких участков.
Понятия «домен» и «семейство доменов» различаются нечётко

Доменные архитектуры



… имеются в виду **семейства** доменов, с учётом их порядка в белках