

# PSI-BLAST. Множественное выравнивание. Профили. Домены

Многие слайды и материалы используемые в презентации взяты из книги Bioinformatics and Functional Genomics by Jonathan Pevsner Copyright © 2009 by John Wiley & Sons, Inc. и соответствующего курса по биоинформатике Johns Hopkins School of Medicine

# BLAST не может решить две проблемы

[1] При использовании человеческого бета-глобина в виде запроса для белков RefSeq, BLASTP не "найдет" миоглобин человека. Потому что эти два белка имеют слишком отдаленное родство. PSI-BLAST в NCBI, а также скрытые Марковские модели легко решают эту проблему.

[2] Нельзя задавать запрос для поиска в виде 10 000 пар оснований или миллионов пар оснований. Многие BLAST подобные инструменты для геномной ДНК имеют такие возможности: PatternHunter, Megablast, BLAT и BLASTZ.

# Position specific iterated BLAST: PSI-BLAST

Цель PSI-BLAST - посмотреть глубже в базу данных в поисках совпадений с вашей последовательностью белка путем использования оценочной матрицы, которая настроена на ваш запрос.

Общая идея : заменяем сиквенс белка вероятностной моделью семейства белков

# **Поиск в PSI-BLAST выполняется в пять шагов**

**[1] Выберите последовательность и запустите поиск в базе данных последовательностей белков**

**[2] PSI-BLAST строит множественное выравнивание последовательностей затем создает «профиль» или специализированную позиционно-специфическую оценочную матрицу (PSSM - position-specific scoring matrix).**

# Проверка вывода BLASTP для выявления эмпирических "правил" в отношении изменчивости аминокислот в каждой позиции

<a href="#">730496</a>	66	FTVDENGQMSATAKGRVRLFNWWDVCA	ADMIGSFTD	TEDPAKF	KMKYWG	VASFLQ	KGNDDH	125
<a href="#">200679</a>	63	FSVDEKGHMSATAKGRVRLLSNWEVCA	DMVGTFTD	TEDPAKF	KMKYWG	VASFLQ	RGNDDH	122
<a href="#">206589</a>	34	FSVDEKGHMSATAKGRVRLLSNWEVCA	DMVGTFTD	TEDPAKF	KMKYWG	VASFLQ	RGNDDH	93
<a href="#">2136812</a>	2	MSATAKGRVRLLSNWWDVCA	DMVGTFTD	TEDPAKF	KMKYWG	VASFLQ	KGNDDH	53
<a href="#">132408</a>	65	FKIEDNGKTTATAKGRVRILDKLELCA	NMVGTFT	ETNDPAK	YRMKYH	GALAIL	ERGLDDH	124
<a href="#">267584</a>	44	FSVDESGKVTATAHGRVILNNWEMCA	NMFGTFED	TPDPAKF	KMRWGA	AAASYL	QTGNDDH	103
<a href="#">267585</a>	44	FSVDGSGKVTATAQGRVILNNWEMCA	NMFGTFED	TPDPAKF	KMRWGA	AAAYLQ	SGNDDH	103
<a href="#">8777608</a>	63	FTIHEDGAMTATAKGRVILNNWEMCA	DMMATFET	TPDPAKF	RMRYWGA	AAASYL	QTGNDDH	122
<a href="#">6687453</a>	60	FKVEEDGTMTATAIGRVILNNWEMCA	NMFGTFED	TEDPAKF	KMKYWG	AAAYLQ	TCYDDH	119
<a href="#">10697027</a>	81	FKVQEDGTMTATATGRVILNNWEMCA	NMFGTFED	TEEPARF	KMKYWG	AAAYLQ	TCYDDH	140
<a href="#">13645517</a>	1		MVGTFTD	TEDPAKF	KMKYWG	VASFLQ	KGNDDH	32
<a href="#">13925316</a>	38	FSVDGSGKMTATAQGRVILNNWEMCA	NMFGTFED	TPDPAKF	KMRWGA	AAAYLQ	SGNDDH	97
<a href="#">131649</a>	65	YTVEEDGTMTASSKGRVKLFGFWVIC	ADMAAQYTD	PTPAKMY	MTYQGL	ASYLSS	GGONY	126

R,I,K

C

D,E,T

K,R,T

N,L,Y,G

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-3	-3	-3	-1	-3	-3	-3	-1	-3	-3	-3	-3	-3	-3	-1	-3	-1	-1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
5	W	3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7	L	2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8	L	1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9	L	1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10	L	2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
14	A	3	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
15	A	2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
16	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
...																					
37	S	2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
38	G	0	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
39	T	0	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
40	W	3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41	Y	2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

20 аминокислот

Все аминокислоты от  
позиции 1 до  
последней позиции  
белковой  
последовательности  
запроса в PSI-BLAST

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	0	0	0	1	1	2	-1	1	0	-3	-2	0
7	L	-2	-2	-4	-4											-3	-3	-1	-2	-1	1
8	L	-1	-3	-3	-4											-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4											-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4											-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2											-1	1	0	-3	-2	0
12	A	5	-2	-2	-2											-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4											-3	-3	-2	7	0	0
14	A	3	-2	-1	-2											-1	1	-1	-3	-3	-1
15	A	2	-1	0	-2											-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2											-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1											-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2											-2	0	-2	-3	-3	-4
39	T	0	-1	0	-1											-1	1	5	-3	-2	0
40	W	-3	-3	-4	-5											-4	-3	-3	12	2	-3
41	Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0

обратите внимание, что данная аминокислота (например, аланин) в последовательности запроса может по-разному оцениваться при совпадении с аланином - в зависимости от положения в белке

# Поиск в PSI-BLAST выполняется в пять шагов

[1] Выберите последовательность и запустите поиск в базе данных последовательностей белков

[2] PSI-BLAST строит множественное выравнивание последовательностей затем создает «профиль» или специализированную позиционно-специфическую оценочную матрицу (PSSM - position-specific scoring matrix).

[3] PSSM используется при запросе для дальнейшего поиска в базе данных

[4] PSI-BLAST оценивает статистическую значимость (E values)



	gi 6978523 ref NP_036909.1	apolipoprotein D [Rattus norvegicus]...	147	4e-35
	gi 1542847 dbj BAA13453.1	(D87752) alpha1-microglobulin/bikunin...	144	6e-34
	gi 619383 gb AAB32200.1	apolipoprotein D, apoD [human, plasma, ...	143	8e-34
	gi 5419892 emb CAB46489.1	(X02824) RBP (aa 101-172) [Homo sapiens]	139	1e-32
	gi 4502163 ref NP_001638.1	apolipoprotein D precursor [Homo sap...	138	4e-32
	gi 584763 sp P37153 APD_RABIT	APOLIPOPROTEIN D PRECURSOR >gi 482...	134	4e-31
	gi 1703341 sp P51909 APD_CAVPO	APOLIPOPROTEIN D PRECURSOR >gi 11...	133	7e-31
	gi 2895204 gb AAC02945.1	(AF025334) mutant retinol binding prot...	80	9e-15
	gi 1246096 gb AAB35919.1	(S80440) apolipoprotein D, apoD (C-ter...	77	8e-14
	gi 2895206 gb AAC02946.1	(AF025335) mutant retinol binding prot...	67	8e-11
NEW	gi 1346419 sp P49291 LAZA_SCHAM	LAZARILLO PROTEIN PRECURSOR >gi ...	63	1e-09
NEW	gi 2506821 sp P00978 AMBP_BOVIN	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
NEW	gi 2497696 sp Q07456 AMBP_MOUSE	AMBP PROTEIN PRECURSOR [CONTAINS...	63	2e-09
NEW	gi 6680684 ref NP_031469.1	alpha 1 microglobulin/bikunin [Mus m...	62	2e-09
NEW	gi 12836446 dbj BAB23659.1	(AK004907) putative [Mus musculus]	62	3e-09
NEW	gi 6978497 ref NP_037033.1	alpha-1 microglobulin/bikunin [Rattu...	62	3e-09
NEW	gi 2507586 sp P04366 AMBP_PIG	AMBP PROTEIN PRECURSOR [CONTAINS: ...	61	8e-09
NEW	gi 1085207 pir  JC2556	alpha-1-microglobulin/inter-alpha-trypsin...	60	1e-08
NEW	gi 2988354 dbj BAA25305.1	(AB006444) alpha-1-microglobulin/biku...	59	2e-08
NEW	gi 108233 pir  S13493	alpha-1-microglobulin - pig	59	2e-08
NEW	gi 1882 emb CAA36306.1	(X52087) precursor codes for two protein...	59	2e-08
NEW	gi 9181923 gb AAF85707.1 AF276505_1	(AF276505) neural Lazarillo ...	59	3e-08
NEW	gi 7296083 gb AAF51378.1	(AE003586) NLaz gene product [Drosophi...	58	3e-08
NEW	gi 117330 sp P80007 CRA2_HOMGA	CRUSTACYANIN A2 SUBUNIT >gi 10275...	57	8e-08
NEW	gi 2497695 sp Q60559 AMBP_MESAU	AMBP PROTEIN PRECURSOR [CONTAINS...	57	1e-07
NEW	gi 102968 pir  S22400	insecticyanin A - tobacco hornworm >gi 971...	56	1e-07
NEW	gi 4502067 ref NP_001624.1	alpha-1-microglobulin/bikunin precu...	56	2e-07
NEW	gi 1146408 gb AAA85089.1	(L41641) gallerin [Galleria mellonella]	56	2e-07
NEW	gi 2497694 sp Q62577 AMBP_MERUN	AMBP PROTEIN PRECURSOR [CONTAINS...	55	3e-07
NEW	gi 1213589 dbj BAA12075.1	(D83712) Prostaglandin D Synthase [Xe...	54	5e-07
	gi 539717 pir  A61233	retinol-binding protein - cat (fragment)	54	8e-07
NEW	gi 266472 sp Q01584 LIPO_BUFMA	LIPOCALIN PRECURSOR >gi 104284 pi...	53	1e-06
	gi 265042 gb AAB25283.1	retinol-binding protein, RBP (N-termina...	52	3e-06
NEW	gi 1079295 pir  S52354	gene cpl-1 protein - African clawed frog ...	52	3e-06
NEW	gi 732003 sp P39281 BLC_ECOLI	OUTER MEMBRANE LIPOPROTEIN BLC PRE...	51	9e-06

# Поиск в PSI-BLAST выполняется в пять шагов

[1] Выберите последовательность и запустите поиск в базе данных последовательностей белков

[2] PSI-BLAST строит множественное выравнивание последовательностей затем создает «профиль» или специализированную позиционно-специфическую оценочную матрицу (PSSM - position-specific scoring matrix).

[3] PSSM используется как запрос для поиска в базе данных

[4] PSI-BLAST оценивает статистическую значимость (E values)

[5] Итеративное повторение шагов [3] и [4], обычно 5 раз.





**При каждом новом поиске, новый профиль используется в качестве запроса.**

# Результаты поиска PSI-BLAST

		Кол. посл.
<u>Итерация</u>		<u>Кол. посл. &gt; threshold</u>
1	104	49
2	173	96
3	236	178
4	301	240
5	344	283
6	342	298
7	378	310
8	382	320




# Поиск PSI-BLAST: RBP4 человека по RefSeq БД, итерация 1

## Sequences with E-value BETTER than threshold

Sequences producing significant alignments:				Score (Bits)	E Value	
NEW	<input checked="" type="checkbox"/>	<a href="#">ref NP_006735.2 </a>	retinol-binding protein 4, plasma precursor [Ho	<u>398</u>	1e-111	
NEW	<input checked="" type="checkbox"/>	<a href="#">ref NP_001638.1 </a>	apolipoprotein D precursor [Homo sapiens]	<u>57.4</u>	7e-09	
NEW	<input checked="" type="checkbox"/>	<a href="#">ref NP_001018059.1 </a>	glycodelin precursor [Homo sapiens] >ref ...	<u>36.2</u>	0.019	
NEW	<input checked="" type="checkbox"/>	<a href="#">ref NP_001624.1 </a>	alpha-1-microglobulin/bikunin precursor [Homo s	<u>35.8</u>	0.021	

Run PSI-Blast iteration 2

## Sequences with E-value WORSE than threshold

<input type="checkbox"/>	<a href="#">ref NP_000597.1 </a>	complement component 8, gamma polypeptide [Homo	<u>33.9</u>	0.077	
<input type="checkbox"/>	<a href="#">ref NP_976222.1 </a>	MSFL2541 [Homo sapiens]	<u>28.5</u>	3.8	
<input type="checkbox"/>	<a href="#">ref NP_066015.2 </a>	hypothetical protein LOC57724 [Homo sapiens]	<u>27.3</u>	7.5	

Run PSI-Blast iteration 2



# Поиск PSI-BLAST: RBP4 человека по RefSeq БД, итерация 2

## Sequences with E-value BETTER than threshold

Sequences producing significant alignments:			Score (Bits)	E Value	
<input checked="" type="checkbox"/>	<a href="#">ref NP_006735.2 </a>	retinol-binding protein 4, plasma precursor [Ho	368	1e-102	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_001638.1 </a>	apolipoprotein D precursor [Homo sapiens]	149	2e-36	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_001018059.1 </a>	glycodelin precursor [Homo sapiens] >ref ...	134	5e-32	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_001624.1 </a>	alpha-1-microglobulin/bikunin precursor [Homo s	125	2e-29	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref XP_001129927.1 </a>	PREDICTED: similar to Glycodelin precursor...	70.0	1e-12	G
NEW <input checked="" type="checkbox"/>	<a href="#">ref XP_944162.1 </a>	PREDICTED: similar to Glycodelin precursor (...)	69.3	2e-12	G
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_000945.3 </a>	prostaglandin H2 D-isomerase [Homo sapiens]	43.5	1e-04	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_976222.1 </a>	MSFL254l [Homo sapiens]	39.6	0.002	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_848564.2 </a>	lipocalin 8 [Homo sapiens]	39.2	0.002	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_001001676.1 </a>	lipocalin 9 [Homo sapiens]	38.5	0.003	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_000597.1 </a>	complement component 8, gamma polypeptide [Homo	36.9	0.010	UG

Run PSI-Blast iteration 3

## Sequences with E-value WORSE than threshold

<input type="checkbox"/>	<a href="#">ref NP_002288.1 </a>	lipocalin 1 precursor [Homo sapiens]	31.5	0.48	UG
<input type="checkbox"/>	<a href="#">ref NP_004534.2 </a>	nebulin [Homo sapiens]	30.4	1.0	UG
<input type="checkbox"/>	<a href="#">ref NP_775903.2 </a>	zinc finger protein 776 [Homo sapiens]	30.0	1.2	UG
<input type="checkbox"/>	<a href="#">ref NP_055983.1 </a>	hypothetical protein LOC23211 [Homo sapiens]	29.2	2.1	G
<input type="checkbox"/>	<a href="#">ref NP_001035982.1 </a>	diaphanous homolog 3 isoform a [Homo sapiens]	28.8	2.5	UG
<input type="checkbox"/>	<a href="#">ref NP_060146.2 </a>	zinc finger, H2C2 domain containing [Homo sapie	28.4	3.4	UG
<input type="checkbox"/>	<a href="#">ref NP_055397.1 </a>	odorant binding protein 2A precursor [Homo sapi	28.1	4.4	UG
<input type="checkbox"/>	<a href="#">ref NP_945184.1 </a>	lipocalin 6 [Homo sapiens]	27.3	9.3	UG

Run PSI-Blast iteration 3

# Поиск PSI-BLAST: RBP4 человека по RefSeq БД, итерация 3

## Sequences with E-value BETTER than threshold

Sequences producing significant alignments:			Score (Bits)	E Value	
<input checked="" type="checkbox"/>	<a href="#">ref NP_006735.2 </a>	retinol-binding protein 4, plasma precursor [Ho	358	2e-99	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_000597.1 </a>	complement component 8, gamma polypeptide [Homo	140	6e-34	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_001638.1 </a>	apolipoprotein D precursor [Homo sapiens]	133	7e-32	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_976222.1 </a>	MSFL2541 [Homo sapiens]	128	2e-30	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_001018059.1 </a>	glycodelin precursor [Homo sapiens] >ref ...	119	1e-27	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_001624.1 </a>	alpha-1-microglobulin/bikunin precursor [Homo s	112	2e-25	UG
<input checked="" type="checkbox"/>	<a href="#">ref XP_001129927.1 </a>	PREDICTED: similar to Glycodelin precursor...	60.8	7e-10	G
<input checked="" type="checkbox"/>	<a href="#">ref XP_944162.1 </a>	PREDICTED: similar to Glycodelin precursor (...)	60.4	8e-10	G
<input checked="" type="checkbox"/>	<a href="#">ref NP_000945.3 </a>	prostaglandin H2 D-isomerase [Homo sapiens]	58.1	4e-09	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_848564.2 </a>	lipocalin 8 [Homo sapiens]	42.7	2e-04	UG
<input checked="" type="checkbox"/>	<a href="#">ref NP_001001676.1 </a>	lipocalin 9 [Homo sapiens]	42.3	3e-04	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_945184.1 </a>	lipocalin 6 [Homo sapiens]	41.5	4e-04	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_055397.1 </a>	odorant binding protein 2A precursor [Homo sapi	38.4	0.003	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_055396.1 </a>	odorant binding protein 2B [Homo sapiens]	36.5	0.016	UG
NEW <input checked="" type="checkbox"/>	<a href="#">ref NP_002288.1 </a>	lipocalin 1 precursor [Homo sapiens]	34.9	0.039	UG

Run PSI-Blast iteration 4

## Sequences with E-value WORSE than threshold

<input type="checkbox"/>	<a href="#">ref NP_848631.2 </a>	lipocalcin 12 [Homo sapiens]	31.1	0.66	UG
<input type="checkbox"/>	<a href="#">ref NP_001001712.2 </a>	lipocalin 10 [Homo sapiens]	30.7	0.82	UG
<input type="checkbox"/>	<a href="#">ref NP_536341.1 </a>	septin 4 isoform 3 [Homo sapiens]	30.3	0.99	UG
<input type="checkbox"/>	<a href="#">ref NP_004565.1 </a>	septin 4 isoform 1 [Homo sapiens]	30.3	0.99	UG
<input type="checkbox"/>	<a href="#">ref NP_246273.2 </a>	phosphodiesterase 5A isoform 3 [Homo sapiens]	27.6	5.9	UG
<input type="checkbox"/>	<a href="#">ref NP_001074.2 </a>	phosphodiesterase 5A isoform 1 [Homo sapiens]	27.6	5.9	UG
<input type="checkbox"/>	<a href="#">ref NP_236914.2 </a>	phosphodiesterase 5A isoform 2 [Homo sapiens]	27.6	5.9	UG
<input type="checkbox"/>	<a href="#">ref NP_003977.1 </a>	gamma-butyrobetaine dioxygenase [Homo sapiens]	27.2	8.5	UG
<input type="checkbox"/>	<a href="#">ref NP_004534.2 </a>	nebulin [Homo sapiens]	27.2	9.6	UG

Run PSI-Blast iteration 4

# Парное выравнивание RBP4 с ApoD, PSI-BLAST итерация 1, E value 3e-07

```
>[ref|NP_001638.1| UG apolipoprotein D precursor [Homo sapiens]
Length=189

Score = 57.4 bits (137), Expect = 3e-07, Method: Composition-based stats.
Identities = 47/151 (31%), Positives = 78/151 (51%), Gaps = 39/151 (25%)

Query   29   VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETQMSATAKGRVRLNNDVC   88
          V+ENFD  ++ G WY + +K P           I A +S+ E G           ++++LN  ++
Sbjct   33   VQENFDVMKYLGRWYEI-EKIPTTFENGRCIQANYSLMENG-----KIKVLNQ-ELR   82

Query   89   ADMVGTFDTDE-----DPAKFKMKY-WGVASFLQKGNDHWIVDTDYDTYAVQYSC   138
          AD  GT   E           +PAK ++K+ W + S           +WI+ TDY+ YA+ YSC
Sbjct   83   AD--GTVNQIEGEATPVNLTPEAKLEVKFSWFMP-----APYWILATDYENYALVYSC   134

Query   139  ----RLLNLDGTCADSYSFVFSRDPNGLPPE   165
          +L ++D           +++++ +R+PN LPPE
Sbjct   135  TCIIQLFHVD-----FAWILARNPN-LPPE   158
```

# Парное выравнивание RBP4 с ApoD, PSI-BLAST итерация 2, E value 1e-42!!!

```
>[ref|NP_001638.1| UG apolipoprotein D precursor [Homo sapiens]
Length=189

Score = 175 bits (443), Expect = 1e-42, Method: Composition-based stats.
Identities = 45/163 (27%), Positives = 77/163 (47%), Gaps = 31/163 (19%)

Query 14  GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA 73
          G+A   + + V+ENFD ++ G WY + +K P           I A +S+ E G++
Sbjct 18  AEGQAFHLGKCPNPPVQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENGKIKV 76

Query 74  TAK-----GRVRLNNDVVCADMVGTFTDTEDPAKFKMKY-WGVASFLQKGNDHWHIVDT 127
          +   G V +               T + +PAK ++K+ W + S           +WI+ T
Sbjct 77  LNQELRADGTVNQIEG-----EATPVNLTEPAKLEVKFSWFMPs-----APYWILAT 123

Query 128 DYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA 166
          DY+ YA+ YSC      L ++D      +++++R+PN LPPE
Sbjct 124 DYENYALVYSCTCIIQLFHVD-----FAWILARNPN-LPPET 159
```



# Парное выравнивание RBP4 с ApoD, PSI-BLAST итерация 3, E value 6e-34

```
>[ref|NP_001638.1| UG apolipoprotein D precursor [Homo sapiens]
Length=189

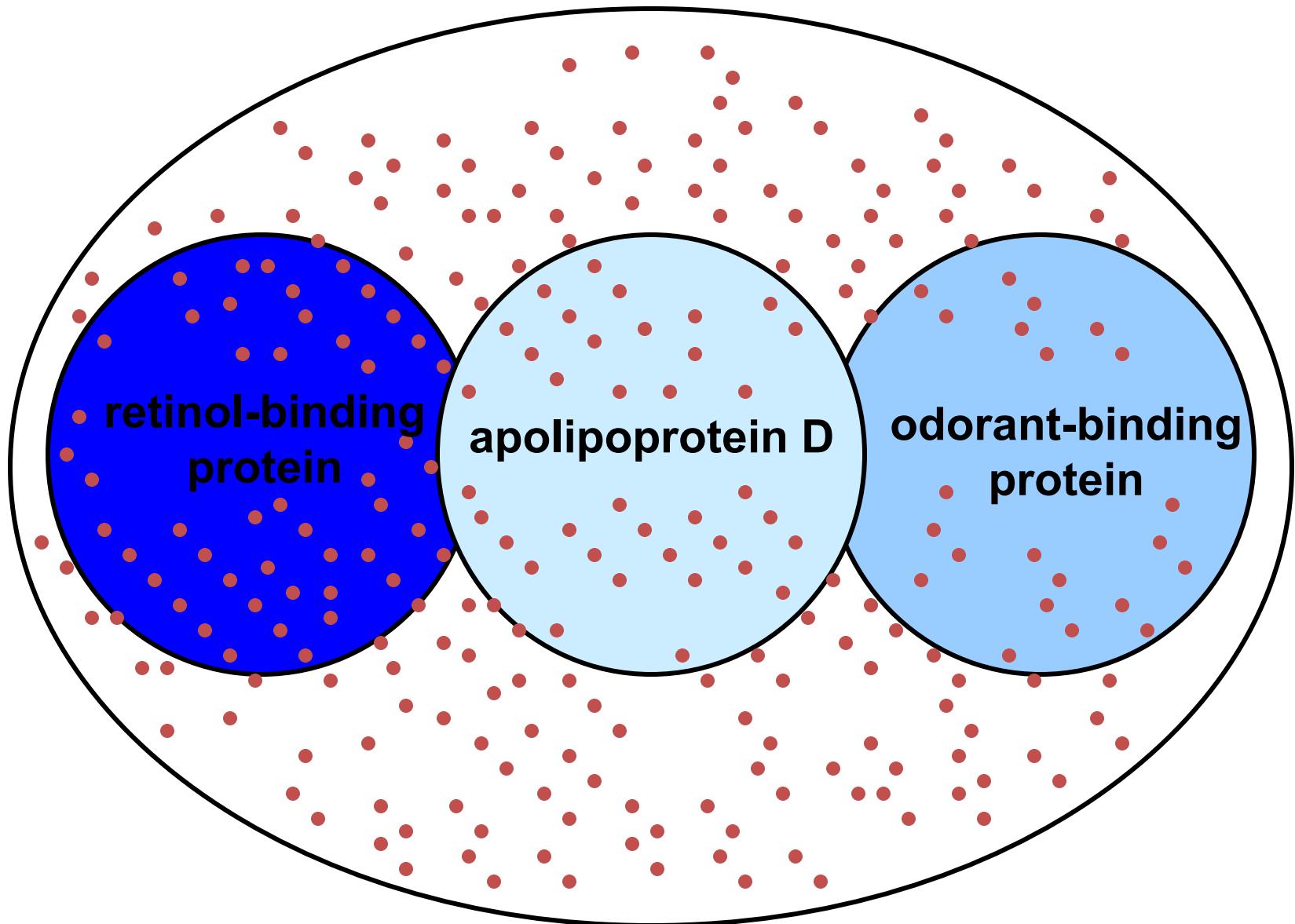
Score = 146 bits (368), Expect = 6e-34, Method: Composition-based stats.
Identities = 41/163 (25%), Positives = 76/163 (46%), Gaps = 20/163 (12%)

Query 14  GSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSA 73
          G+A   + + V+ENFD ++ G WY + +K P           I A +S+ E G++
Sbjct 18  AEGQAFHLGKCPNPPVQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMENGKIKV 76

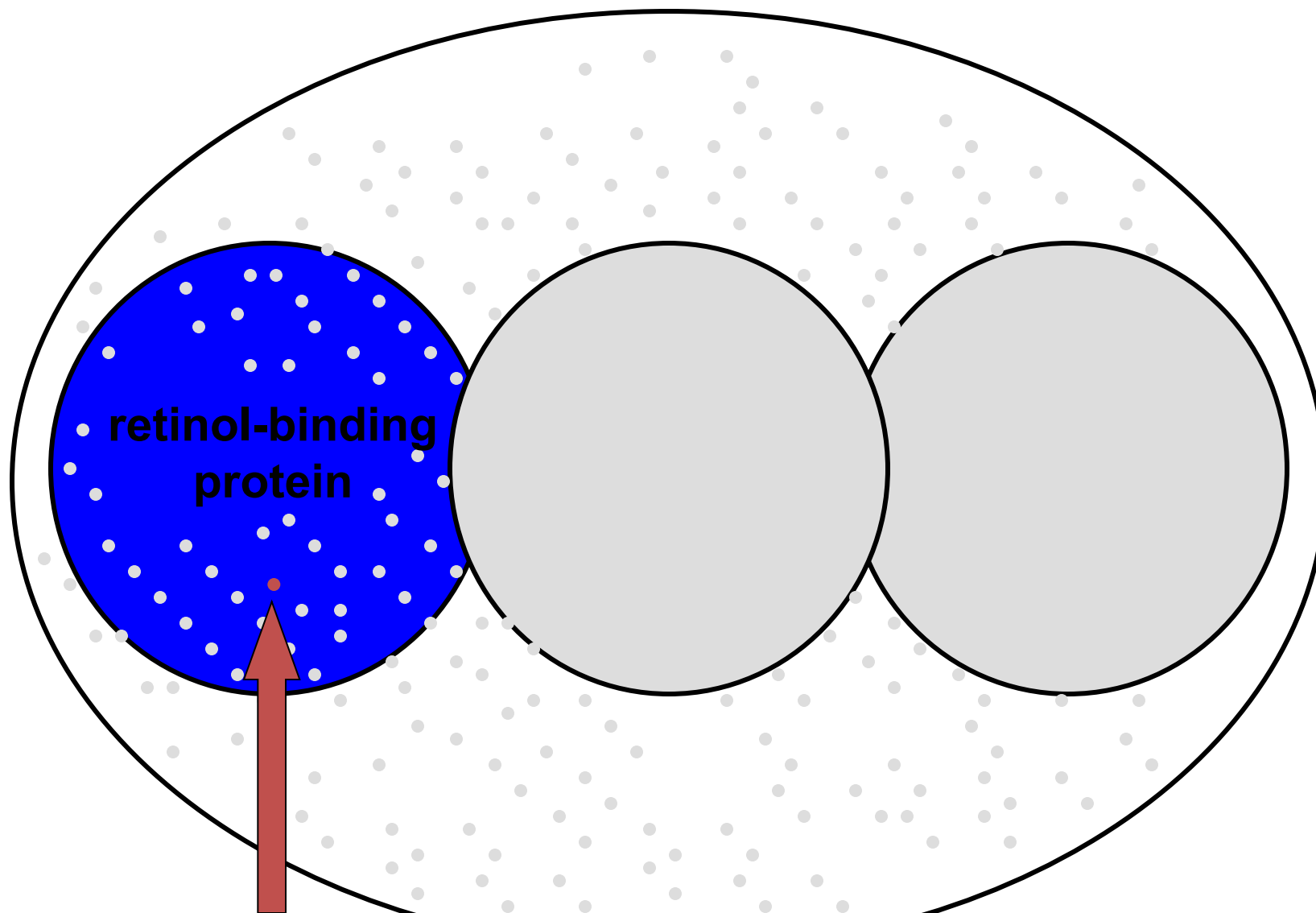
Query 74  TAKGRVRLLNWWDVCADMVGTFTDTEDPAKFKMKY-WGVASFLQKGNDHWDHWDYDITY 132
          + +R   + + T + +PAK ++K+ W + S           +WI+ TDY+ Y
Sbjct 77  LNQ-ELRADGTVNQI-EGEATPVNLTEPAKLEVKFSWFMP-----APYWILATDYENY 128

Query 133 AVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEAQKIVR 171
          A+ YSC      L ++D      ++++ +R+PN P      +
Sbjct 129 ALVYSCTCIIQLFHVD-----FAWILARNPNLPPETVDSLKN 165
```

# Вселенная липокалинов (каждая точка - белок)

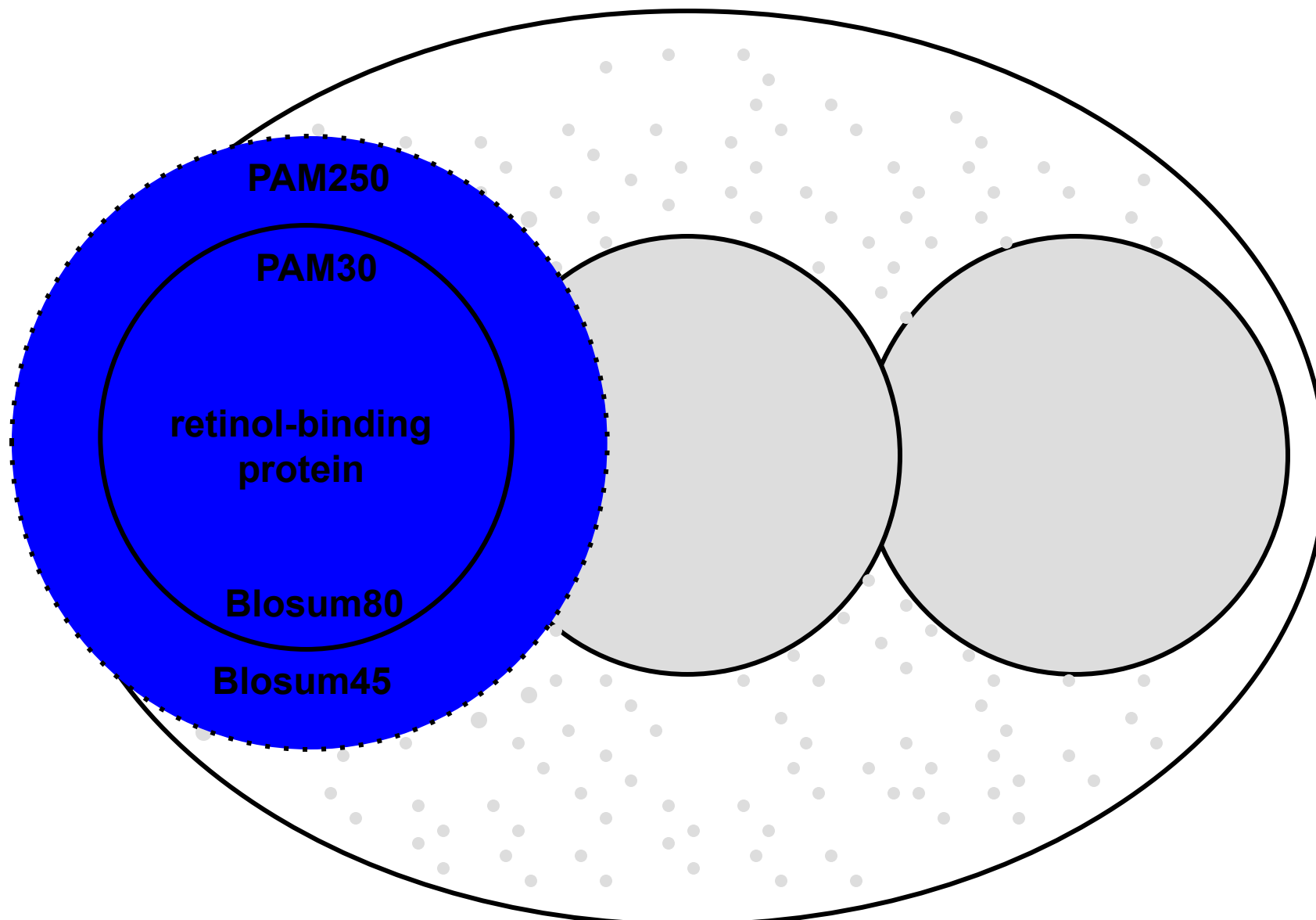


# Скоринг матрицы позволяют сосредоточиться на большой (или маленькой) картине

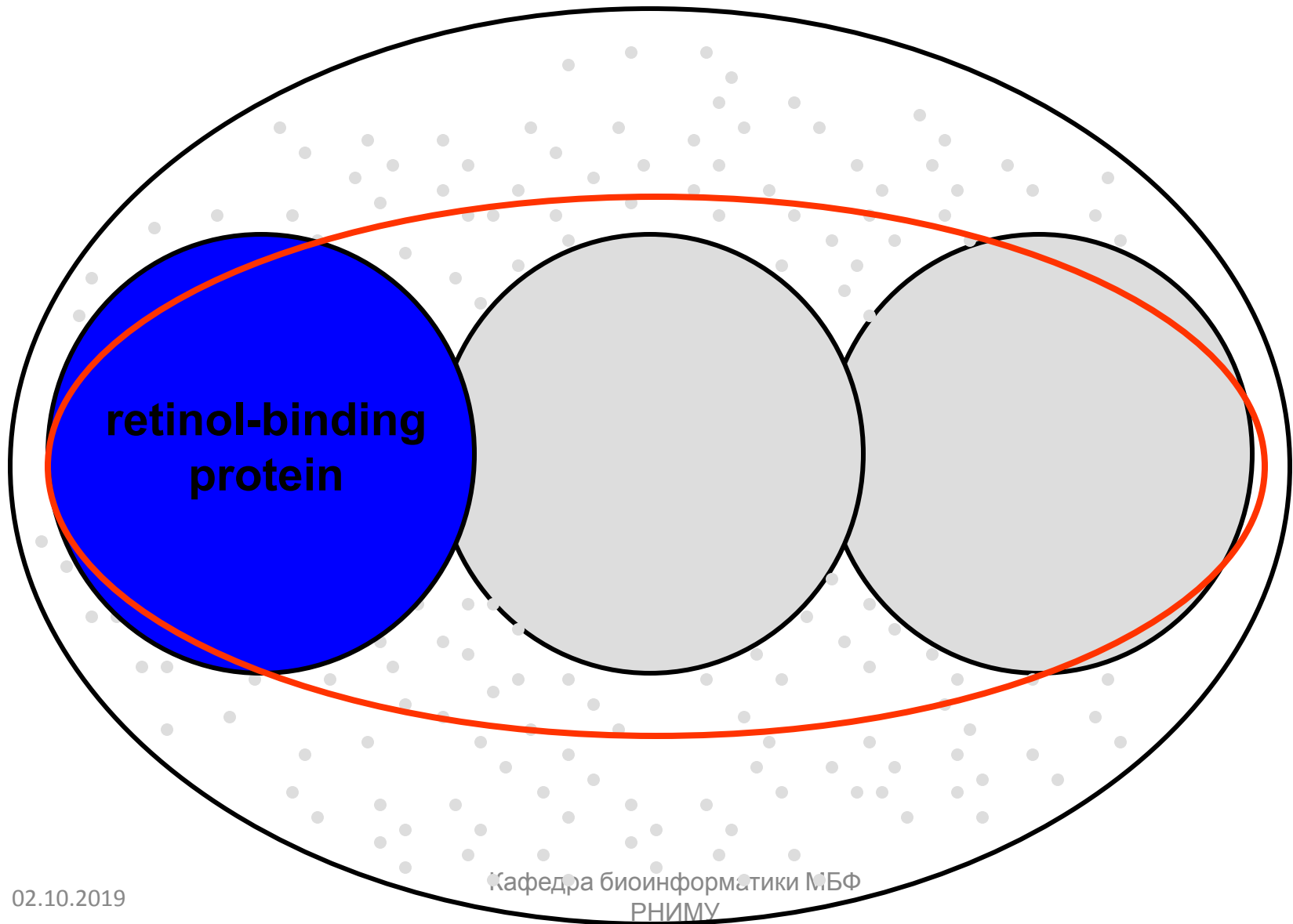


**Запрос RBP**

# Скоринг матрицы позволяют сосредоточиться на большой (или маленькой) картине



# PSI-BLAST создает скоринг матрицы более мощные чем PAM или BLOSUM



# PSI-BLAST: оценка эффективности

PSI-BLAST полезен для обнаружения слабых, но биологически значимых связей между белками (<40% идентичность аминокислот)

Основным источником ложно-положительных оценок является ложное усиление последовательностей, не связанных с запросом. Например, запрос с биспиральным (coiled-coil) мотивом может выявить тысячи других негомологичных белков с этим мотивом.

Даже однажды вошедший выше порога в результат поиска PSI-BLAST ложный белок останется при последующих итерациях – **проблема искажения (corruption)**

# PSI-BLAST: проблема искажения

Искажение определяется как присутствие, по меньшей мере, одного ложно-положительного выравнивания со значением  $E < 10^{-4}$  после пяти итераций.

Три подхода к борьбе с искажением:

- [1] Применить фильтрацию искажающих участков профиля сгенерированного PSI-BLAST (например, программа SEG: <http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Seg.html>)
- [2] Настроить порог  $E$  значения ниже 0,001 (по умолчанию), например  $E = 0,0001$ .
- [3] Просмотреть результаты каждой из итерации. Удалить подозрительные хиты, сняв флажок.

# Множественное выравнивание последовательностей

- **Эволюционный анализ**
  - определение гомологии
  - филогенетические построения
  - эволюционные тестовые модели
- **Функциональный анализ**
  - определить консервативные участки
  - идентификация белковых семейств
- **Структурный анализ**
  - определить последовательность ковариация
  - моделирование гомологии
- **Практическое применение**
  - определить консервативные сайты связывания праймеров
  - конструирование мутагенетических экспериментов
  - анализ мутантов

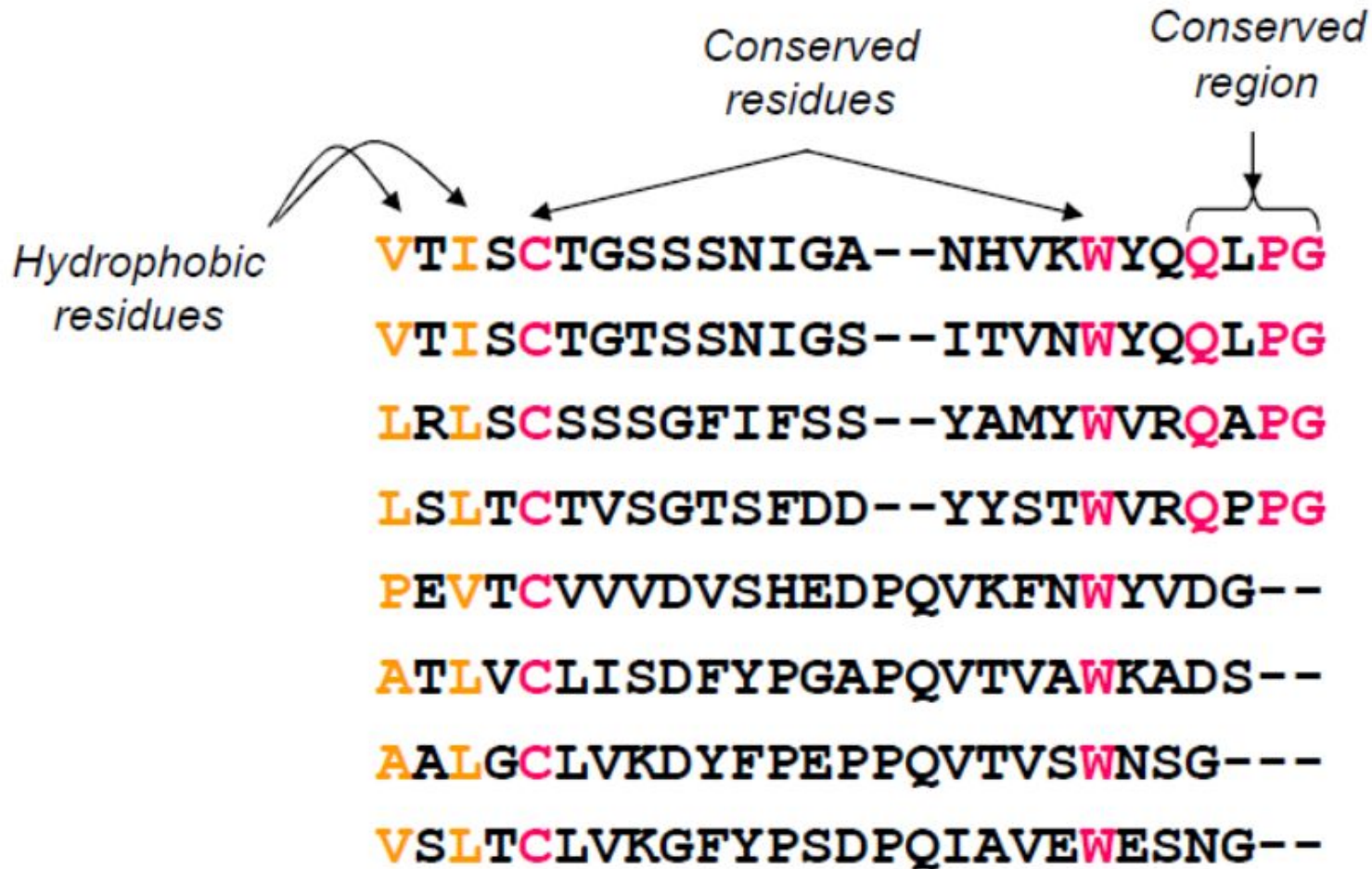
В 1990-х исследователи начали понимать, что выравнивание нескольких последовательностей (профилей) дает гораздо больше информации, чем парные выравнивания.



# Множественное выравнивание последовательностей

- Набор из трех или более белковых (или нуклеотидных) последовательностей, которые частично или полностью выровнены
- Гомологичные остатки выровнены в столбцах по всей длине последовательностей
- Остатки гомологичны в эволюционном смысле
- Остатки гомологичны в структурном смысле

# Множественное выравнивание последовательностей



*N. Provart & D. Guttman. Bioinformatic Methods I. Coursera*

Кафедра биоинформатики МБФ  
РНИМУ

# HomoloGene включает группы эукариотических белков, парные и множественные выравнивания, и много другое

NCBI HomoloGene Discover Homologs

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journ.

Search HomoloGene for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display HomoloGene Show 20 Send to

All: 1 Fungi: 0 Mammals: 0

☐ 1: HomoloGene:116063. Gene conserved in Eukaryota

**Genes**  
*Genes identified as putative homologs of one another during the construction of HomoloGene.*

- zgc:136799, *Danio rerio*  
zgc:136799
- LOC100148385, *Danio rerio*  
hypothetical protein LOC100148385
- ARAC9/AtROP8/ROP8, *Arabidopsis thaliana*  
ARAC9/AtROP8/ROP8 (rho-related protein from plants 8); GTP binding

**Proteins**  
*Proteins used in sequence comparisons and their conserved domain architectures.*

- NP\_001034907.1 192 aa
- XP\_001918572.1 192 aa
- NP\_566024.1 209 aa

**Protein Alignments**  
*Protein multiple alignment, pairwise similarity scores and evolutionary distances.*

Show Multiple Alignment

Show Pairwise Alignment Scores

Pairwise alignments generated using BLAST

Regenerate Alignments

NP\_001034907.1 (Danio rerio)

XP\_001918572.1 (Danio rerio)

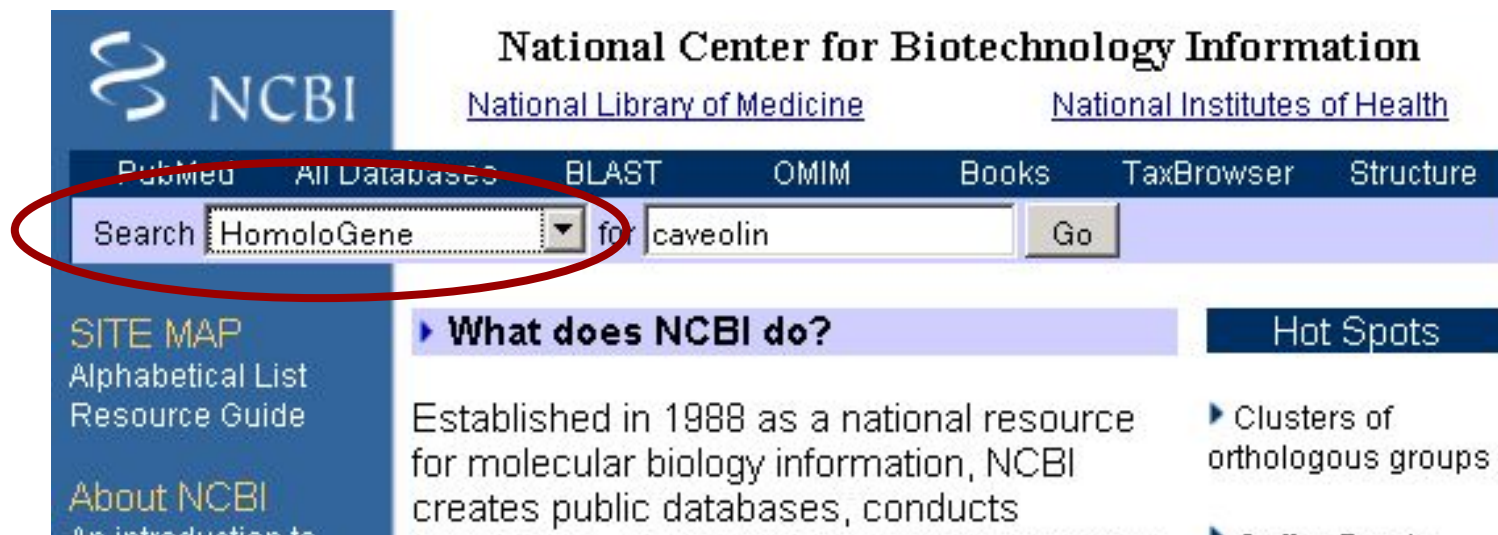
BLAST

**Conserved Domains**  
*Conserved Domains from CDD found in protein sequences by rpsblast searching.*

- Ras\_like\_GTPase (c110444)
  - Ras-like GTPase superfamily. The Ras-like superfamily of small GTPases consists of several families with an extremely high degree of structural and functional similarity. The Ras superfamily is divided into at least four families in eukaryotes: the Ras...

**UniGene**  
*Links to groups of transcribed sequences established by tblastn searching of UniGene.*

# Пример. Шаг 1: в NCBI выберете меню HomoloGene и введите caveolin в поле поиска



**NCBI**  
National Center for Biotechnology Information  
[National Library of Medicine](#) [National Institutes of Health](#)

PubMed All Databases BLAST OMIM Books TaxBrowser Structure

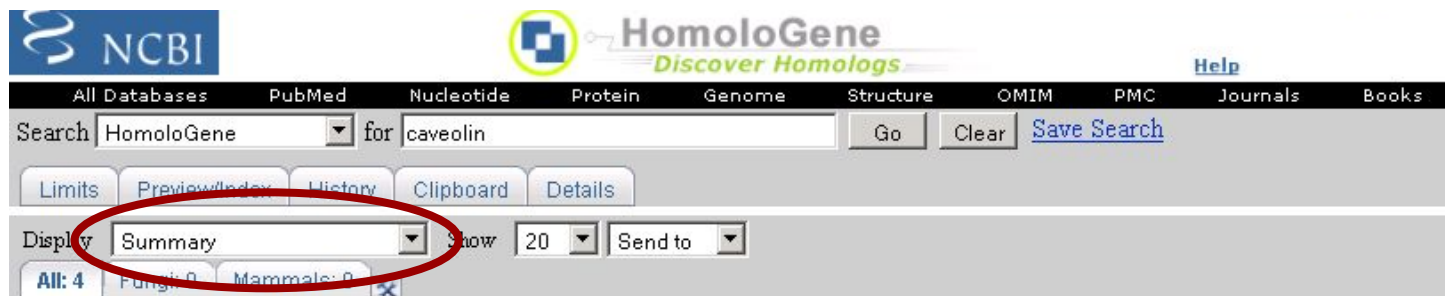
Search **HomoloGene** for caveolin

**SITE MAP**  
Alphabetical List  
Resource Guide  
**About NCBI**  
An introduction to

**What does NCBI do?**  
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts

**Hot Spots**  
Clusters of orthologous groups

# Пример. Шаг 2: проверить результаты. Возьмем первый набор кавеолинов. Изменить Display на Multiple alignment.



NCBI HomoloGene Discover Homologs

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search HomoloGene for caveolin Go Clear Save Search

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Send to

All: 4 Fungi: 0 Mammals: 0

Items 1 - 4 of 4

☐ 1: HomoloGene:1330. Gene conserved in Euteleostomi

[Download](#)

CAV1	caveolin 1, caveolae protein, 22kDa
CAV1	caveolin 1, caveolae protein, 22kDa
CAV1	caveolin 1, caveolae protein, 22kDa
CAV1	caveolin 1, caveolae protein, 22kDa
Cav1	caveolin 1, caveolae protein
Cav1	caveolin 1, caveolae protein
CAV1	caveolin 1, caveolae protein, 22kDa
cav1	caveolin 1

*Homo sapiens*  
*Pan troglodytes*  
*Canis lupus familiaris*  
*Bos taurus*  
*Mus musculus*  
*Rattus norvegicus*  
*Gallus gallus*  
*Danio rerio*

☐ 2: HomoloGene:7255. Gene conserved in Euteleostomi

[Download](#)

CAV3	caveolin 3
CAV3	caveolin 3
CAV3	caveolin 3
CAV3	caveolin 3
Cav3	caveolin 3
Cav3	caveolin 3
CAV3	caveolin 3
cav3	caveolin 3

*Homo sapiens*  
*Pan troglodytes*  
*Canis lupus familiaris*  
*Bos taurus*  
*Mus musculus*  
*Rattus norvegicus*  
*Gallus gallus*  
*Danio rerio*



# Пример. Шаг 3: проверим множественное выравнивание. Восемь белков хорошо выравнены, хотя пробелы также включены.

1: HomoloGene:1330. Gene conserved in Euteleostomi

Download

## Multiple Sequence Alignment

Generated by MUSCLE [\[see reference\]](#) version 3.6 (using option: -maxiters 2).

<a href="#">NP_001744.2</a>	1	MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADELSEKQVYDAHT	46
<a href="#">XP_519325.2</a>	1	MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADELSEKQVYDAHT	46
<a href="#">NP_001003296.1</a>	1	MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-AEEMSEKQVYDAHT	46
<a href="#">NP_776429.1</a>	1	MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-AEEMNEKQVYDAHT	46
<a href="#">NP_031642.1</a>	1	MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADEVTEKQVYDAHT	46
<a href="#">NP_113744.1</a>	1	MSGGKYVDS---EGHLYTVPIREQGNIYKPNNKAM-ADEVNEKQVYDAHT	46
<a href="#">XP_001234148.1</a>	1	---MEYFQ---EAFLYAAPVREQGNIYKPNNKMM-ADELSEKAVHDVHT	42
<a href="#">NP_997816.1</a>	1	MTSG-YKDGTPEEEYAHSPFIRKQGNIYKPNNKEMDNDNINEKTLQDVHT	49
<a href="#">NP_001744.2</a>	47	KEIDLVNRPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK	96
<a href="#">XP_519325.2</a>	47	KEIDLVNRPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK	96
<a href="#">NP_001003296.1</a>	47	KEIDLVNRPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK	96
<a href="#">NP_776429.1</a>	47	KEIDLVNRPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK	96
<a href="#">NP_031642.1</a>	47	KEIDLVNRPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK	96
<a href="#">NP_113744.1</a>	47	KEIDLVNRPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK	96
<a href="#">XP_001234148.1</a>	43	KEIDLVNRPKHLNDDVVKIDFEDVIAEPEGTHSFDGIWKASFTTFTVTK	92
<a href="#">NP_997816.1</a>	50	KEIDLVNRPKHLNDDVVKVDFEDVIAEPAGTYSFDGVWKASFTTFTVTK	99

# Другое множественное выравнивание, Рас:

NP 061485.1	1	-----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
XP 855587.1	1	-----MQAIKCVVVEDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
NP 776588.1	1	-----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
NP 033033.1	1	-----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
NP 599193.1	1	-----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
NP 990348.1	1	-----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
NP 956065.1	1	-----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
NP 648121.1	1	-----MQAIKCVVVG DGAVGKTCLLISYTTNAFPGEYIPTVFDNYSANVM	45
XP 366655.1	1	MAAPGVQSLKCVVTGDGAVGKTCLLISYTTNAFPGEYIPTVFDNYSASVM	50
XP 329350.1	1	MLTGEMLTLD FLLL-----TCLLISYTTNAFPGEYIPTVFDNYSASVM	43
NP 195320.1	1	--MSASRF IKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV	48
NP 179371.1	1	--MSASRF IKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV	48
NP 190698.1	1	--MSASRFVKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVV	48
NP 195228.1	1	--MSASRF IKCVTVGDGAVGKTCLLISYTSNTFPTDYVPTVFDNFSANVI	48
NP 001048639.1	1	--MSASRF IKCVTVGDGAVGKTCMLISYTSNTFPTDYVPTVFDNFSANVV	48

NP 061485.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQTVGETYGKDITSRGKDKPIAD	95
XP 855587.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D	76
NP 776588.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D	76
NP 033033.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D	76
NP 599193.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D	76
NP 990348.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D	76
NP 956065.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D	76
NP 648121.1	46	VDGKPVNLGLWDTAGQEDYDRLRPLSYPQT-----D	76
XP 366655.1	51	VDGKPI SLGLWDTAGQEI	
XP 329350.1	44	VDGKPVSLGLWDTAGQEI	
NP 195320.1	49	VNGATVNLGLWDTAGQEI	
NP 179371.1	49	VNGATVNLGLWDTAGQEI	
NP 190698.1	49	VNGSTVNLGLWDTAGQEI	
NP 195228.1	49	VDGNTINLGLWDTAGQEI	
NP 001048639.1	49	VDGSTVNLGLWDTA--EDYINRLRPLSYRGA-----D	77

Эта вставка может быть  
альтернативным  
сплайсингом

# Пример: 5 выравниваний 5 глобинов

- Давайте посмотрим на множественное выравнивание последовательности (MSA) пяти глобинов белков. Мы будем использовать пять известных программ MSA: ClustalW, Praline, MUSCLE (используется в HomoloGene), ProbCons и TCoffee. Каждая программа имеет уникальные особенности.
- Мы сосредоточимся на остатках гистидина (H), который имеет важную роль в связывании кислорода в глобинах, и должны быть выровнены. Но часто выравнивание не совпадает, и все пять программ дают разные ответы.
- **Вывод:** не существует единственно верного подхода к MSA. Десятки новых программ были разработаны в последние годы.



# ClustalW

CLUSTAL W (1.83) multiple sequence alignment

```

beta globin  -----MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFESFG- 47
myoglobin   -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFK- 48
neuroglobin -----MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR 47
soybean      -----MVAFTEKQDALVSSSFEAFKANIPOYSVVFYTSILEKAPAAKDLFSFLA- 49
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSFLR- 59
              :   :   :   :   .   .   .   :   :   *   *   .
              ▽
beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFFATLS-----ELHCDKLHVDPE 102
myoglobin   HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLA-----QSHATKHKIPVK 103
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLAS---LGRKHRAVGVKLS 104
soybean      --NGVDPT--NPKLTGHAEKLFALVRDSAGQLKASGTVVADAA----LGSVHAQKAVTDP 101
rice         --NSDVPLEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA 117
              .   .   .   *   .   :   :   :   :   :   :   :
              :   :   :   :   :   :   :   :   :   :   :
beta globin  NFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH----- 147
myoglobin   YLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG 154
neuroglobin SFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDGE---- 151
soybean      QFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA----- 144
rice         HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE--- 166
              :   :   :   :   :   :   *   .   .   :
    
```

Обратите внимание как участок консервативного гистидина (▼) изменяется в зависимости от используемого алгоритма

# Praline

## (a) Praline multiple sequence alignment

beta globin	.....MVHLT <b>PEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG</b>	▼
myoglobin	.....MGLS <b>DGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGH</b> PETLEKFDK.FK	
neuroglobin	.....MERPE <b>PELI</b> RQSWRAVSR <b>SPLEHGTVLFARLFALEPDLLPLFQYNCR</b>	
soybean	.....MVA <b>FT</b> E <b>KQDALVSSSFEAFKANIPQYSVVFYTSILEKAPA</b> <b>AKDLFS..FL</b>	
rice	MALVEDNNAVAVSF <b>SEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL</b>	
Consistency	000000000014265438257934573463364343624453686433*35344*50063	
▽		
beta globin	DLST <b>PDAVMGNPKVKAHGKKVLGAFSDG</b> LAHLDNLKGT <b>FATLSEL..HCDKLH....VDP</b>	▼
myoglobin	HLKSEDEMKA <b>SEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV</b>	
neuroglobin	QFSSPEDCLSS <b>PEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL</b>	
soybean	A.NGVDP..TN <b>PKLTGHA</b> EKL <b>FALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD</b>	
rice	R.NSDVPLEKN <b>PKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRD</b> TTLKRLGATH <b>KLKYGVD</b>	
Consistency	3166354224776653*4368635424454451335634333542003335440000922	
▽		
beta globin	<b>ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKV</b> VAGVANALAHKYH.....	
myoglobin	<b>KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG</b>	
neuroglobin	SSFSTVGESLLYM <b>LEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..</b>	
soybean	<b>PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELA</b> AAIKKA.....	
rice	<b>AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...</b>	
Consistency	43744844498258542305336554454*55465426446754322001000	



# Probcons

(c)

## PROBCONS

beta globin M-----VHLT**PEEK**SAVTALWGKVNVD--**EVGGEALGRLLVVY**PWTQRFFES-FG  
 myoglobin M-----GLS**DGEWQLVLNV**WGKVEAD**IPGHGQEV**LIRLFKGHPETLEKFDK-FK  
 neuroglobin M-----ERPE**PELI**RQSWRAVSRS**PLEHGT**VLFA**RLF**ALEPDLLPLFQYNCR  
 soybean M-----VAFTE**EKQDALVSS**SFEAFKAN**IPQYSV**VFYTSILEK**APAAKDLFSF**-LA  
 rice MALVEDNNAVAVSFSE**EEQEALVLK**SWAILKKDSANIALRFFLK**IFEVAPSASQMF**SF-LR  
 \* \* : : : : . . : : \* \*

beta globin DLST**PDAVMGNPKVKAH**GKKVLGAFSDGLAHL**D**---NLK---GTFATLSEL**H**CDKLHVDP  
 myoglobin HLKSEDEMKA**SEDLKKH**GATVLTALGGI---LKKKG**H**HE---AEIK**PLAQSHAT**KKHKIPV  
 neuroglobin QFSSPEDCLSS**PEFLDHIR**KVMLVIDAAVTNVEDLSS**LE**---EY**LASLGRKH**RAV-GVKL  
 soybean **NGVDP**---**TNPKLTGHA**EKL**FALVRDSAGQLKAS**GT**V**---**ADAALGSVHAQK**-AVTD  
 rice NSDVP--LEKN**PKLKT**HAMSVFVMTCEAA**AQLRKAGK**VTVRDTTL**KRLGATH**LKY-GVGD  
 . : . . \* . : : : . \* \*

beta globin **ENFRLLGNVLVCVLA**HHF-GKEFT**PPVQAAYQKV**VAGVANALAHK-----YH  
 myoglobin **KYLEFI**SECIIQVLQSKH-PGDFGADAQ**GAMNKALELFRK**D**MASNYKEL**GFQG  
 neuroglobin SSFSTVGESLLYM**LEKCL**-GPAFT**PATRAAWSQ**LYGAV**VQAM**SRG---W-DGE  
 soybean **PQFVVVKEALLKTI**KA**AV**-GDK**WSELSRAWEVAYDE**LAAAIK-----KA  
 rice **AHFEVVKFALLDTI**KEEVPADM**WSPAMKSAWSEAYDHLVAAIKQE**---MKPAE  
 : : : : : \* . . :

# TCoffee

(d)

CLUSTAL FORMAT for T-COFFEE Version\_5.13

```

beta globin  -----MVHLTPEEKSAVTALWGKVNVDD--EVGGEALGRLLVVYPWTQRFFESFG
myoglobin   -----MGLSDGEWQLVLNVWVGKVEADIPGHGQEVLIIRLFKGH PETLEKFD-KFK
neuroglobin -----MERPEPELIHQSWRAVSRS PLEHGT VLFARLFALEPDLLPLFQYNCR
soybean      -----MVAFT EKQDALVSSSF EAFKANIPQYSVV FYTSILEKAPAAKD LFS-FLA
rice         MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
               :   :   :   :   . . .   .   : :   *   * .

               ▽                               ▾

beta globin  DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVD P
myoglobin   HLKSEDEMKA SEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIEV
neuroglobin QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLA SLGRKH-RAVGVKL
soybean      NGVDP---TNPKLTGHA EKLFALVRDSAGQLKASGT VVAD---AALGSVHAQKAVTDP
rice         NSDVP--LEKNPKLKTHAMSVFVMTCEAA AQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
               .       . . . * . : :           :       * . *

beta globin  ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKV VAGVANALAHKYH-----
myoglobin   KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean      Q-FV VVKEALLKTIKAAV-GDKWSELSRAW EVAYDELA AAIKKA-----
rice         H-FEVVKFALLDTIKEEVPADMWS PAMKSAWSEAYDHLVAAIKQE---MKPAE
               :   :   : :   :           :       * .   .   :
    
```

# Свойства множественного выравнивания последовательностей

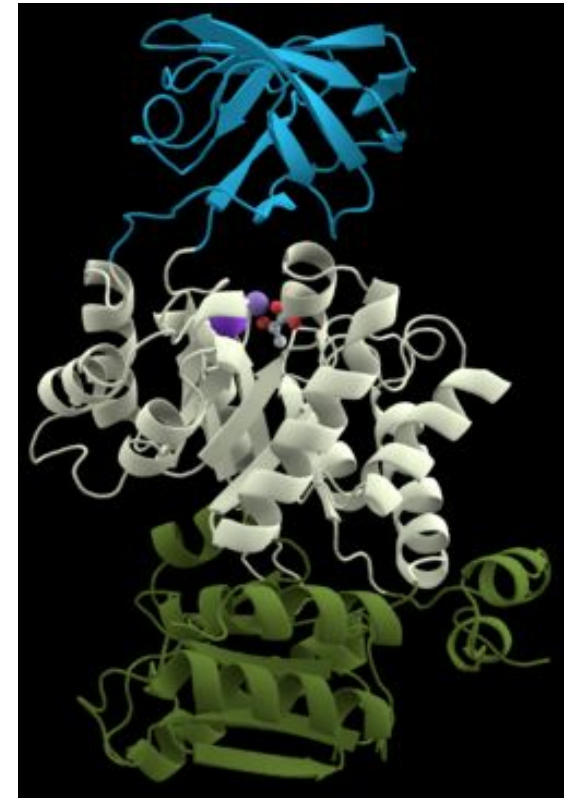
- Не обязательно, что существует одно "правильное" выравнивание семейства белков
- Эволюционируют белковые последовательности ...
- Соответствующие трехмерные структуры белков также эволюционируют...
- может оказаться невозможным идентификации аминокислотных остатков, которые выравниваются должным образом (структурно) в течение множественного выравнивания последовательностей
- Для двух белков, с 30% идентичностью аминокислотной последовательности, совмещается около 50% отдельных аминокислот в двух структурах

# Особенности множественного выравнивания последовательностей

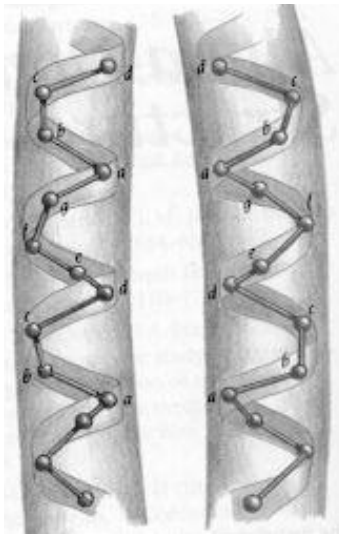
- некоторые выровненные остатки, такие как цистеина, образующие дисульфидные мостики, могут быть высоко консервативны
- может быть консервативные домены, такие как трансмембранный домен
- может быть консервативны особенности вторичной структуры
- может быть участки в последовательностях являются паттернами вставок или делеций

# Домены

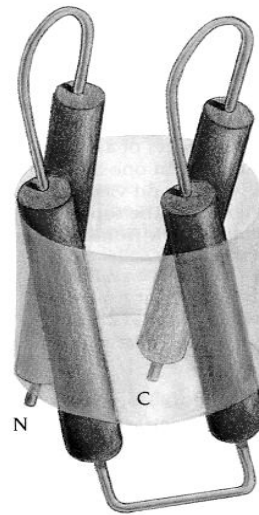
- Домен белка — элемент третичной структуры белка, представляющий собой достаточно стабильную и независимую подструктуру белка, фолдинг которой проходит независимо от остальных частей [wikipedia].
- Домен – это часть полипептидной цепи (или вся цепочка), которая сворачивается **независимо** в **стабильную** третичную структуру [C.Brenden & John Tooze]
- Доменами в белках называют области в третичной структуре, которым свойственна **определенная** автономия структурной организации [Степанов В.М.]



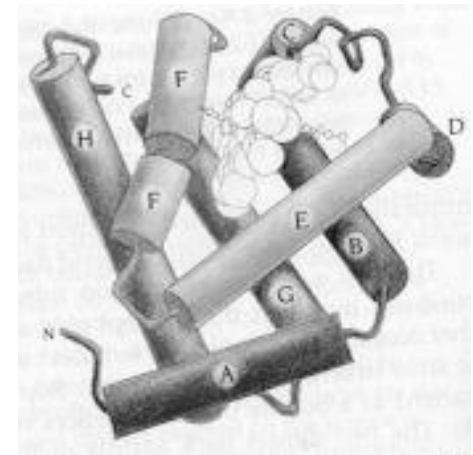
# $\alpha$ -ДОМЕНЫ



**Лейциновая молния**



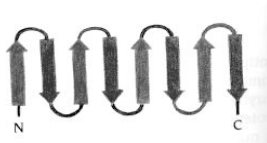
**Связка из 4  
спиралей**



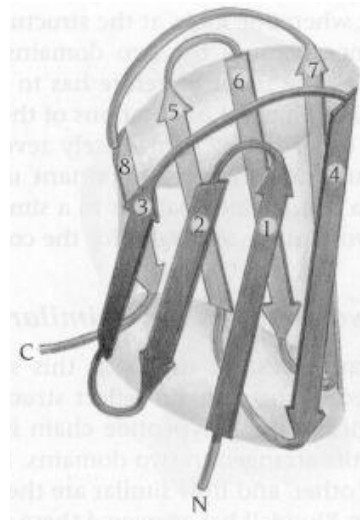
**Глобиновая укладка**



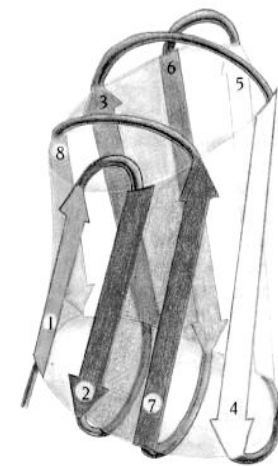
# β-ДОМЕНЫ



**Up and down barrel**

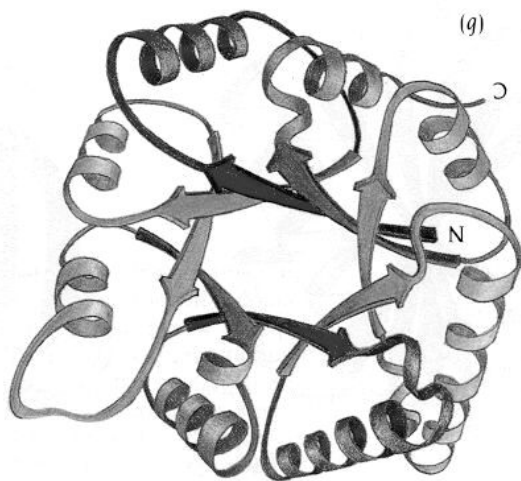


**Баррел на основе греческих ключей**

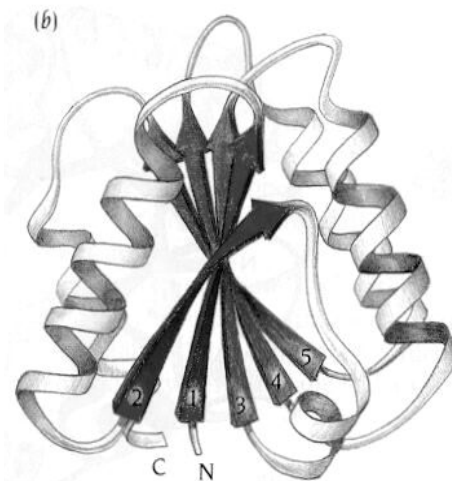


**Jelly roll barrel**

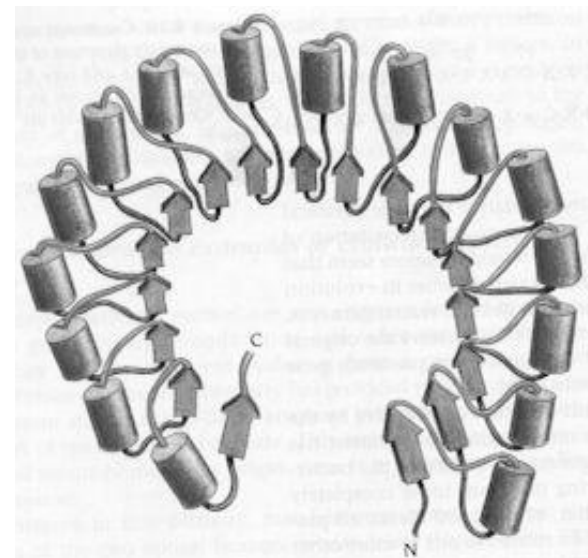
# $\alpha/\beta$ - ДОМЕНЫ



**ТИМ-укладка**



**$\alpha/\beta$ -пропеллер  
Укладка Россмана**



**Подкова**



**Метилмалонил-коА-  
мутаза**

# Использование множественного выравнивания

- Более чувствительно, чем попарное выравнивание для выявления гомологов
- Результат BLAST может принять форму множественного выравнивания, и может раскрыть консервативные остатки или мотивы
- Демографические данные могут быть проанализированы в множественном выравнивании (PopSet)
- Отдельный запрос может быть использован для поиска в базе данных множественных выравниваний (например, PFAM)
- Регуляторные области генов могут быть консенсусными последовательностями идентифицируемыми множественным выравниванием

# Методы множественного выравнивания

- Точные методы
- Прогрессивный (ClustalW)
- Итеративный (MUSCLE)
- Согласованный (ProbCons)
- Основанный на структуре (Expresso)

# Прогрессивный метод (ClustalW)

Прогрессивные методы: используют направляющей дерево (связанное с филогенетическим деревом), чтобы определить, как объединить попарные выравнивания по одному для создания множественного выравнивания.

[1] Сделать ряд глобальных попарных выравниваний (Needleman и Wunsch динамический алгоритм программирования)

[2] Создать направляющее дерево

[3] Постепенно выровнять последовательности

# Шаг 1. Построение попарных выравниваний

(% идентичности)



SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score
=====	=====	=====	=====	=====
1 beta_globin	147	2 myoglobin	154	25
1 beta_globin	147	3 neuroglobin	151	15
1 beta_globin	147	4 soybean	144	13
1 beta_globin	147	5 rice	166	21
2 myoglobin	154	3 neuroglobin	151	16
2 myoglobin	154	4 soybean	144	8
2 myoglobin	154	5 rice	166	12
3 neuroglobin	151	4 soybean	144	17
3 neuroglobin	151	5 rice	166	18
4 soybean	144	5 rice	166	43
=====	=====	=====	=====	=====



**best  
Score**

Для  $n$  последовательностей,  $(n-1)(n) / 2$

Для 5 последовательностей,  $(4)(5) / 2 = 10$

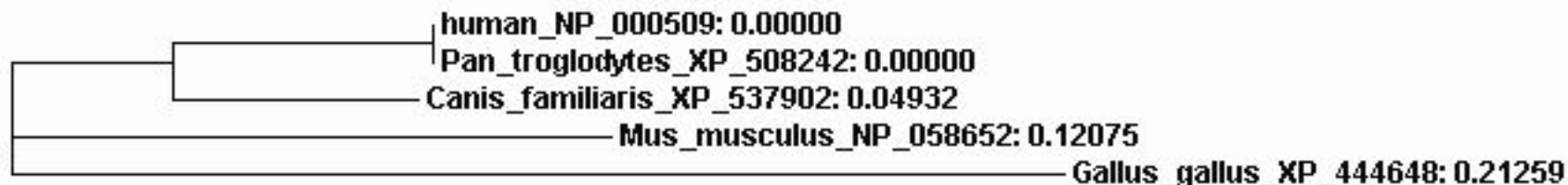
Для 200 последовательностей,  $(199)(200) / 2 = 19,900$

SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score
1 human_NP_000509	147	2 Pan_troglodytes_XP_508242	147	100
1 human_NP_000509	147	3 Canis_familiaris_XP_537902	147	89
1 human_NP_000509	147	4 Mus_musculus_NP_058652	147	80
1 human_NP_000509	147	5 Gallus_gallus_XP_444648	147	69
2 Pan_troglodytes_XP_508242	147	3 Canis_familiaris_XP_537902	147	89
2 Pan_troglodytes_XP_508242	147	4 Mus_musculus_NP_058652	147	80
2 Pan_troglodytes_XP_508242	147	5 Gallus_gallus_XP_444648	147	69
3 Canis_familiaris_XP_537902	147	4 Mus_musculus_NP_058652	147	78
3 Canis_familiaris_XP_537902	147	5 Gallus_gallus_XP_444648	147	71
4 Mus_musculus_NP_058652	147	5 Gallus_gallus_XP_444648	147	66

Конвертация  
баллов  
сходства в  
баллы  
расстояния

```
(
(
(
human_NP_000509:0.00000,
Pan_troglodytes_XP_508242:0.00000)
:0.05272,
Canis_familiaris_XP_537902:0.04932)
:0.03231,
Mus_musculus_NP_058652:0.12075,
Gallus_gallus_XP_444648:0.21259);
```

5 близко  
родственных  
глобинов





# Множественное выравнивание для профилей скрытых Марковских моделей (HMMs - Hidden Markov models)

- Скрытые Марковские модели (HMMs) являются "состояниями", которые описывают вероятность наличия конкретного аминокислотного остатка расположенного в колонке множественного выравнивания последовательностей
- HMMs являются вероятностными моделями
- HMMs может дать более чувствительные выравнивания, чем традиционные методы, такие как прогрессивное выравнивание

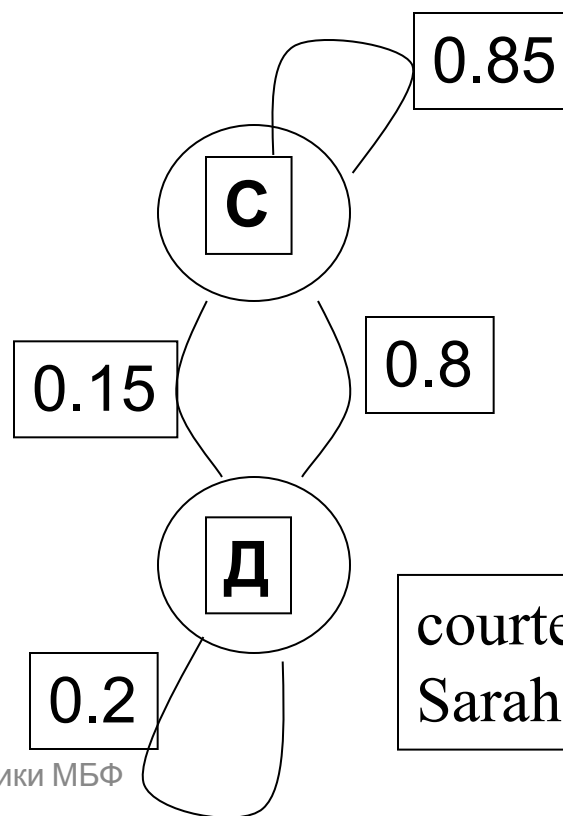
# Простая Марковская модель



Марковское состояние = не зависимость от ближайшего предыдущего состояния ("Без памяти")

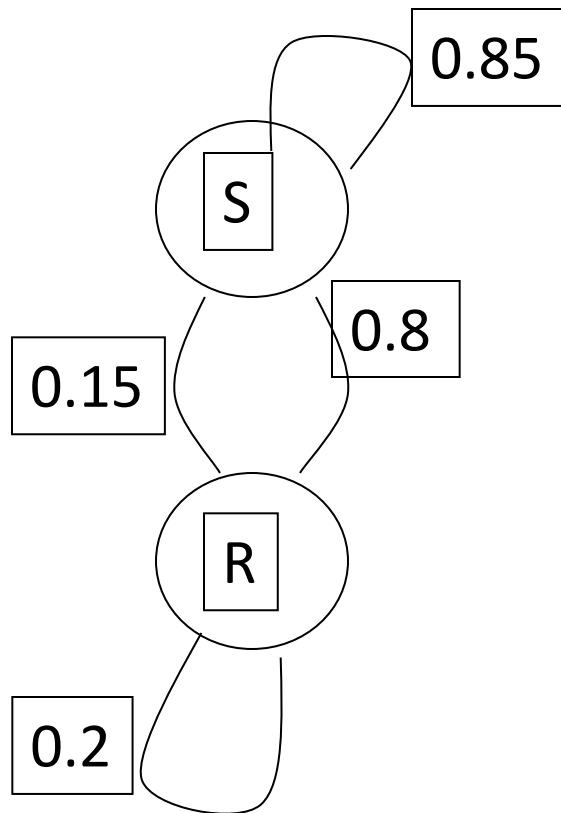
**Дождь** = собака может не захотеть выйти на улицу

**Солнце** = собака, вероятно, выйдет на улицу



courtesy of  
Sarah Wheelan

# Простая скрытая Марковская модель



$P(\text{собака идет в дождь}) = 0.2$

$P(\text{собака идет на солнце}) = 0.85$

Наблюдение: YNNNYNNNNYN

(Y=идет, N=не идет)

Что лежит в основе реальности  
(скрытом состоянием цепи)?

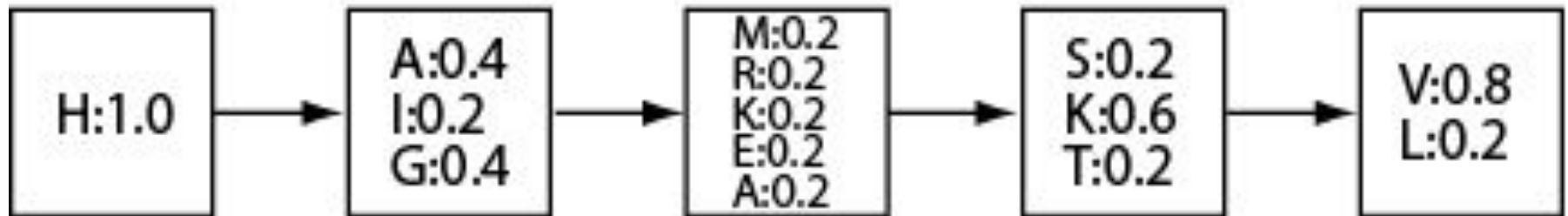
1D8U  
1OJ6A  
2hhbB  
1FSL  
2MM1

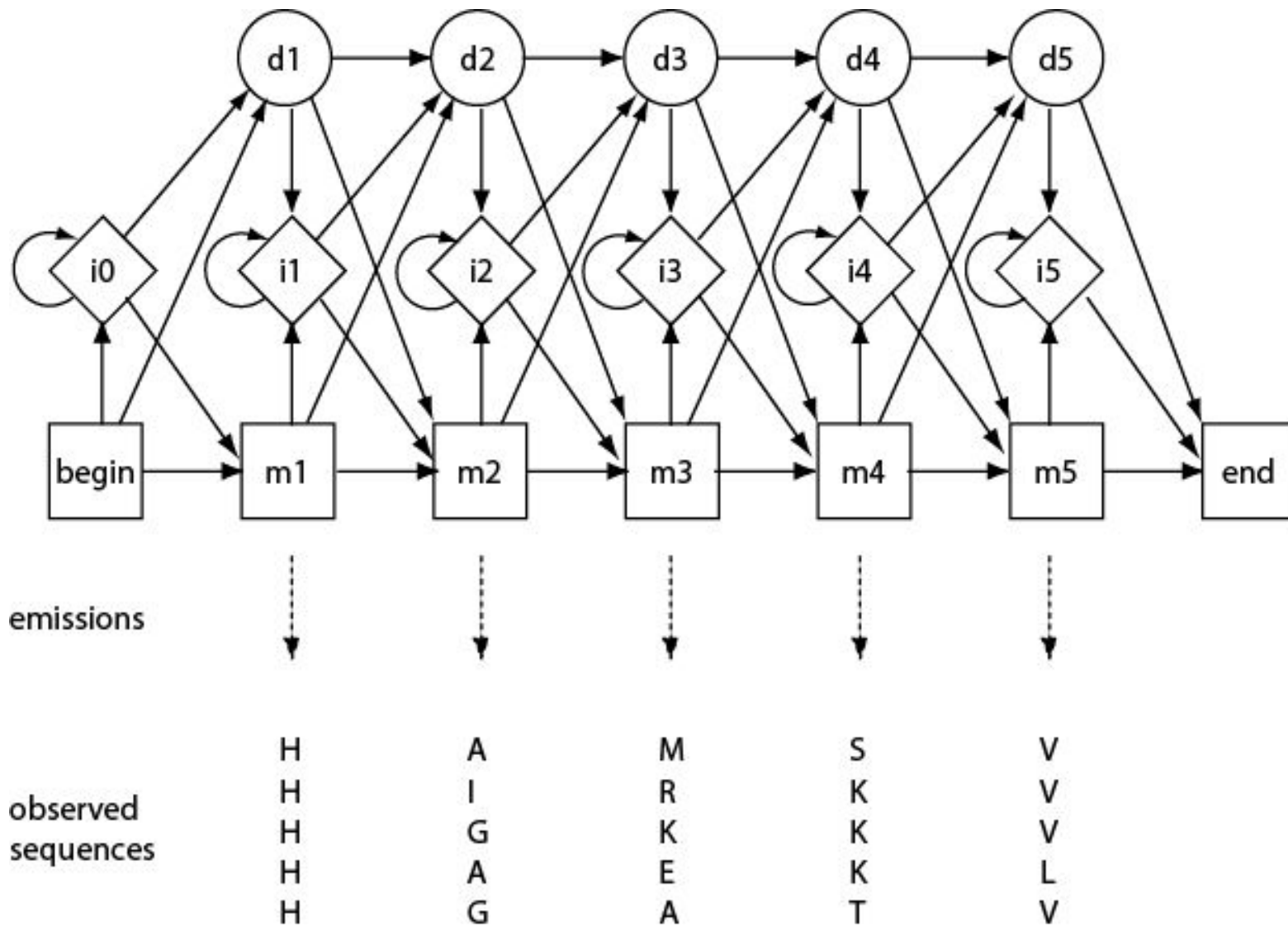
HAMSV  
HIRKV  
HGKKV  
HAEKL  
HGATV

Probability	position				
	1	2	3	4	5
p(H)	1.0				
p(A)		0.4			
p(I)		0.2			
p(G)		0.4			
p(M)			0.2		
p(R)			0.2		
p(K)			0.2		
p(E)			0.2		
p(A)			0.2		
p(S)				0.2	
p(K)				0.6	
p(T)				0.2	
p(V)					0.8
p(L)					0.2

$$p(\text{HARTV}) = (1.0)(0.4)(0.2)(0.2)(0.8) = 0.0128$$

$$\text{Log odds score} = \ln(1.0) + \ln(0.4) + \ln(0.2) + \ln(0.2) + \ln(0.8) =$$





# МОТИВЫ

- Мотив в молекулярной биологии — это характерная последовательность нуклеотидов (в ДНК, РНК) или аминокислот (в белках), которые имеют существенное биологическое значение. Мотивы в белках позволяют найти участки белков, отвечающие за определённые свойства.
- Для обозначения мотива используют стандартные обозначения регулярных выражений



# Регулярные выражения

- Алфавит — совокупность отдельных символов, обозначающих определенную аминокислоту или набор аминокислот.
- Строка из символов алфавита — обозначающая последовательность соответствующих аминокислот.
- $[ABC]$  — любая строка символов, взятых из алфавита в квадратных скобках соответствует любому из соответствующих аминокислот; например  $[ABC]$  соответствует любому из аминокислот, из представленных: или a или b или c.
- $\{ABC\}$  — любая строка символов, взятых из алфавита соответствует любой аминокислоте кроме тех, что находятся в фигурных скобках; например  $\{ABC\}$  соответствует любой аминокислоте, кроме: a, b и c.
- Главная идея, лежащая в этих обозначениях — принцип соответствия: последовательность элементов паттерна совпадает с последовательностью аминокислот, если и только если последнюю последовательность можно разбить на подпоследовательности таким образом, что каждый элемент массива соответствует соответствующий подпоследовательности в свою очередь.
- Например, модель  $[AB] [CDE] F$  соответствует шести последовательности аминокислот: ACF, ADF, AEF, BCF, BDF и BEF.

# PROSITE – база данных для поиска МОТИВОВ в белках ([prosite.expasy.org](http://prosite.expasy.org))

PROSITE дополняет список выражений, описанных выше:

1. «х» — шаблон элемента обозначают любую аминокислоту.
2. '<' — шаблон ограничивается N-концом последовательности.
3. '>' — шаблон ограничивается C-концом последовательности.

Также символ '>' может находиться внутри квадратных скобок, например: S [ T> ] соответствует как " ST " и « S >».

4. Если e — шаблон элемента, и m и n два целых десятичных числа и  $m \leq n$ , то:

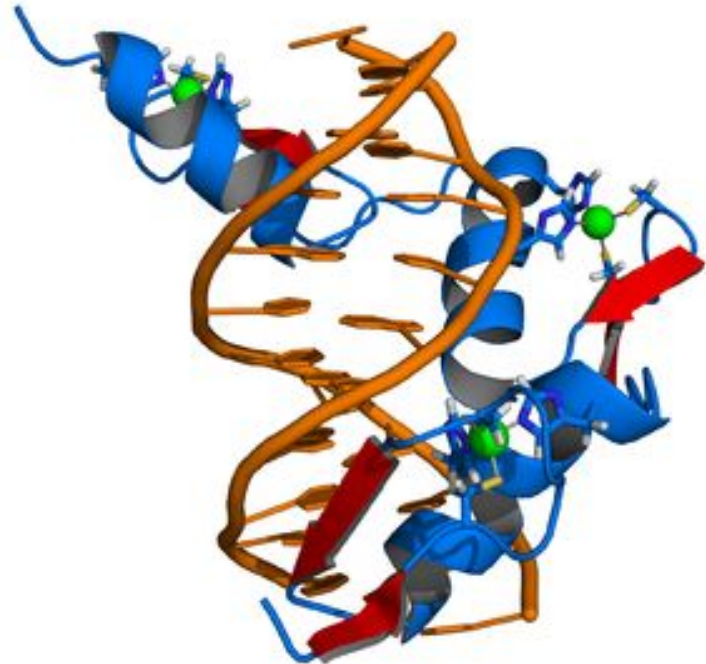
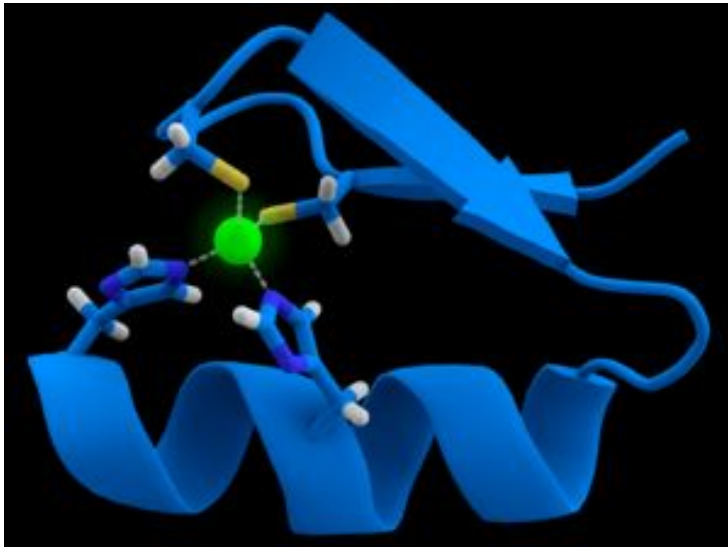
- e (m) эквивалентно повторению e ровно m раз - e ( m, n) эквивалентно повторению e ровно k раз для любого целого k удовлетворяющей :  $m \leq k \leq n$  Например:

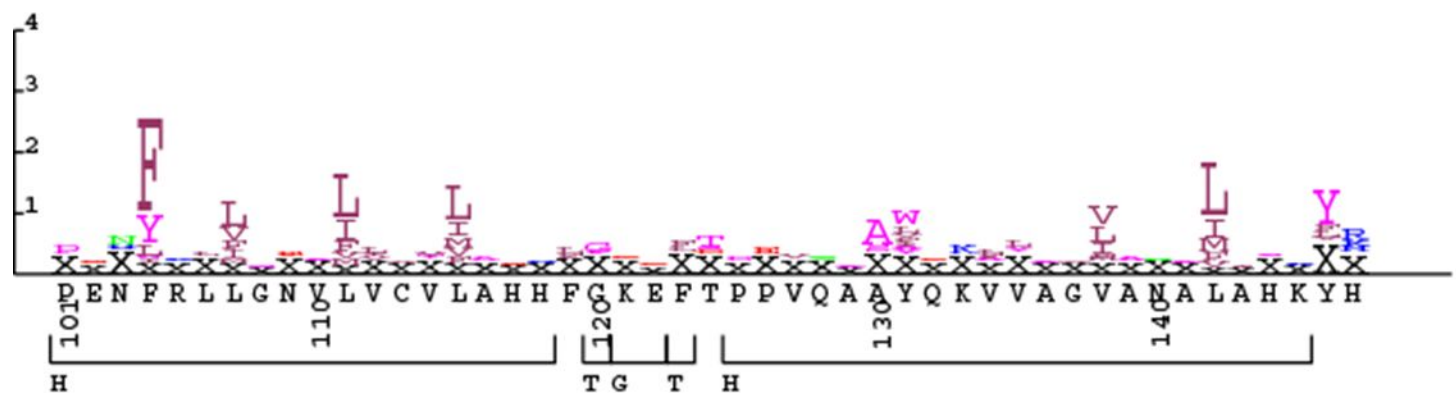
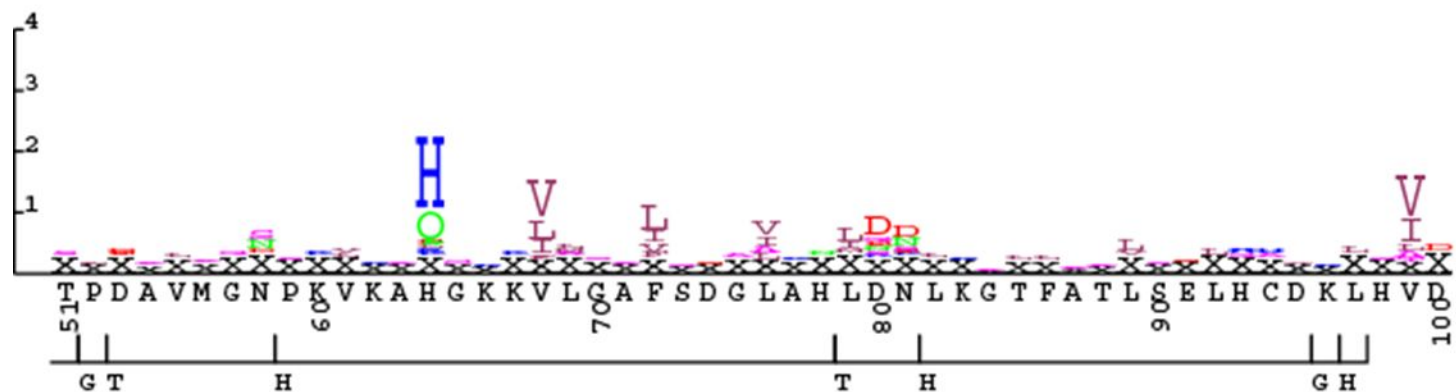
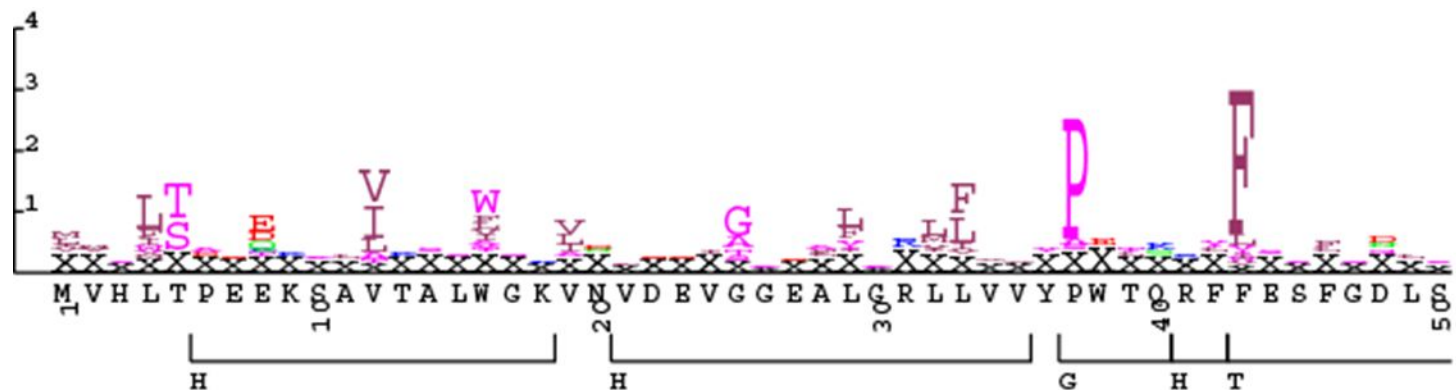
x (3) эквивалентно X-X-X.

x (2,4) соответствует любой последовательности, которая соответствует xx или xxx или xxxx.

# Мотив домена цинковый палец:

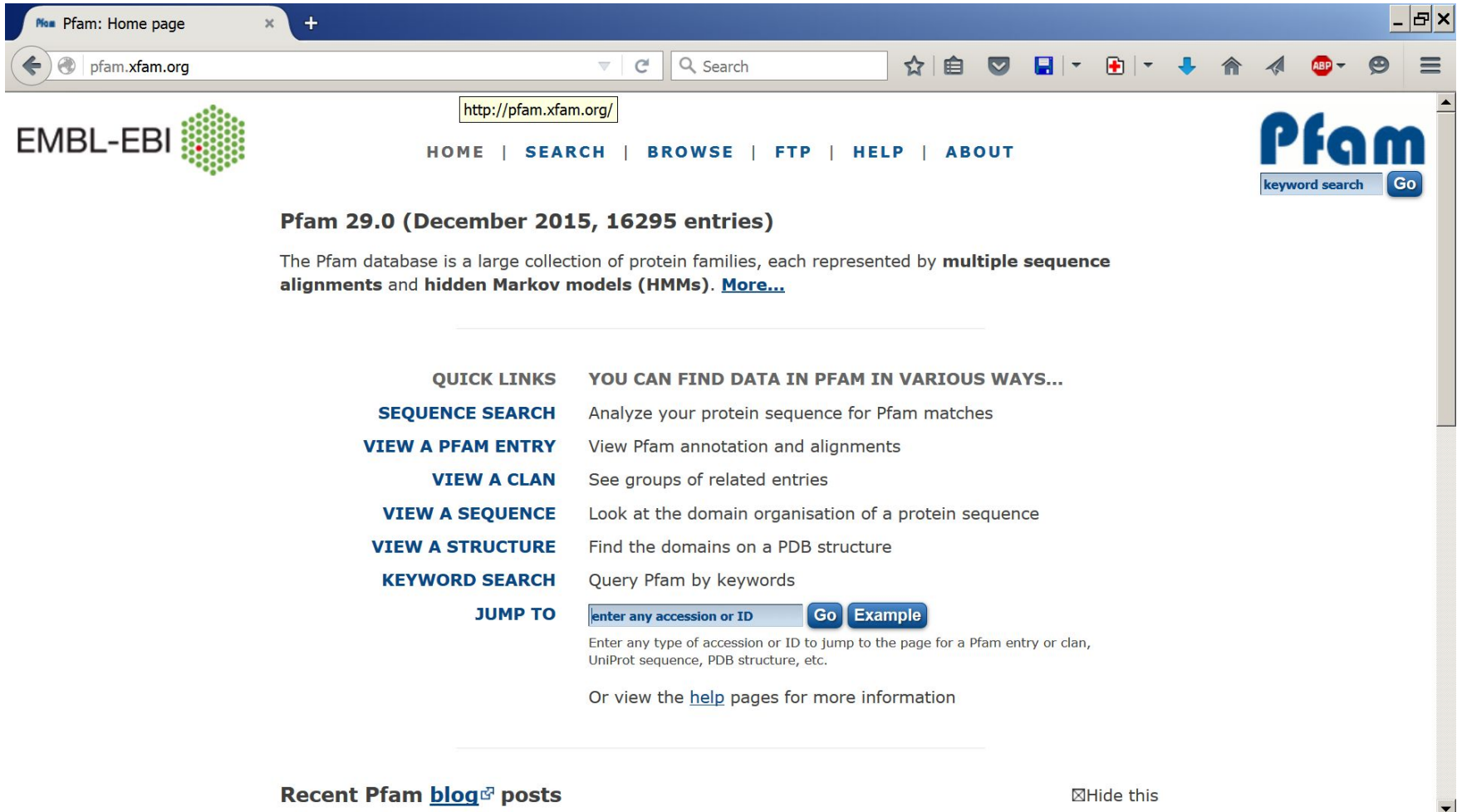
C-x (2,4)-C-x (3)-[LIVMFYWCS]- x(8)-H-x(3,5)-H





# PFAM (protein family) БД – наиболее известный ресурс по анализу белковых семейств

<http://pfam.xfam.org/>



The screenshot shows the Pfam website homepage in a web browser. The browser's address bar displays 'pfam.xfam.org'. The page features the EMBL-EBI logo on the left and the Pfam logo on the right, which includes a 'keyword search' button. A navigation menu at the top center contains links for HOME, SEARCH, BROWSE, FTP, HELP, and ABOUT. The main heading is 'Pfam 29.0 (December 2015, 16295 entries)'. Below this, a paragraph describes the database as a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs), with a 'More...' link. The page is divided into two columns. The left column, titled 'QUICK LINKS', lists: SEQUENCE SEARCH, VIEW A PFAM ENTRY, VIEW A CLAN, VIEW A SEQUENCE, VIEW A STRUCTURE, KEYWORD SEARCH, and JUMP TO. The right column, titled 'YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...', lists: Analyze your protein sequence for Pfam matches, View Pfam annotation and alignments, See groups of related entries, Look at the domain organisation of a protein sequence, Find the domains on a PDB structure, Query Pfam by keywords, and a search box with 'Go' and 'Example' buttons. Below the search box, it says 'Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.' and 'Or view the help pages for more information'. At the bottom, there is a link to 'Recent Pfam blog posts' and a 'Hide this' checkbox.

EMBL-EBI

<http://pfam.xfam.org/>

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

**Pfam 29.0 (December 2015, 16295 entries)**

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

**QUICK LINKS**

- [SEQUENCE SEARCH](#)
- [VIEW A PFAM ENTRY](#)
- [VIEW A CLAN](#)
- [VIEW A SEQUENCE](#)
- [VIEW A STRUCTURE](#)
- [KEYWORD SEARCH](#)
- [JUMP TO](#)

**YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...**

- Analyze your protein sequence for Pfam matches
- View Pfam annotation and alignments
- See groups of related entries
- Look at the domain organisation of a protein sequence
- Find the domains on a PDB structure
- Query Pfam by keywords

[Go](#) [Example](#)

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

Recent Pfam [blog](#) posts [Hide this](#)

# База данных PFAM (protein family)

