

Motifs, Patterns and Profiles



INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Dr. Jaspreet Kaur Dhanjal

Assistant Professor, Department of Computational Biology

Email ID: jaspreet@iiitd.ac.in

September 30, 2025

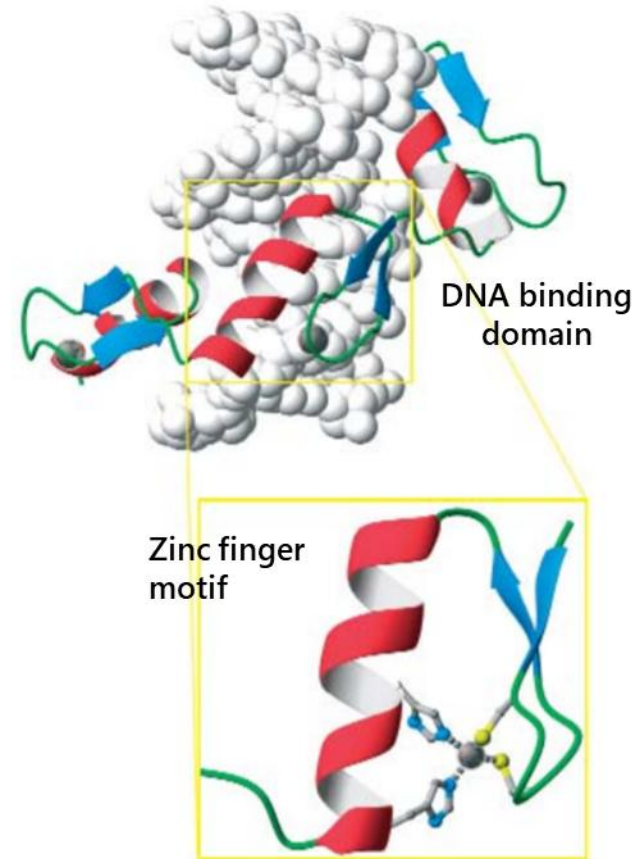
Motifs and Domains

Motifs (supersecondary structure)

- Simple combinations of secondary structure
- Formed from consecutive sequences of primary structure
- Examples: helix-loop helix (EF hand), beta-alpha-beta, Greek key

Domains

- Stable, independently folded, globular units, often consisting of combinations of motifs
- Vary from 25 to 300 amino acids, average length – 100
- Large globular proteins may consist of several domains linked by stretches of polypeptide
- Separate domain may have distinct functions
- In many cases binding site formed by cleft between 2 domains



Common Motifs

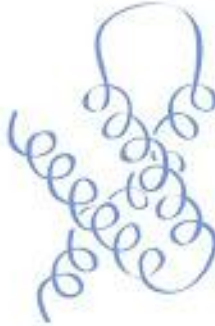
(a) Helix-loop-helix



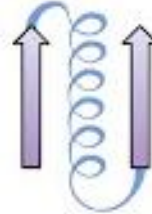
(b) Coiled coil



(c) Helix bundle



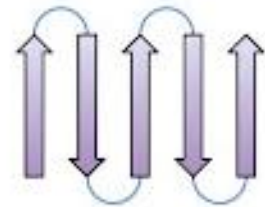
(d) $\beta\alpha\beta$ unit



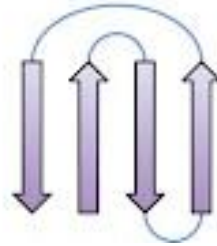
(e) Hairpin



(f) β meander



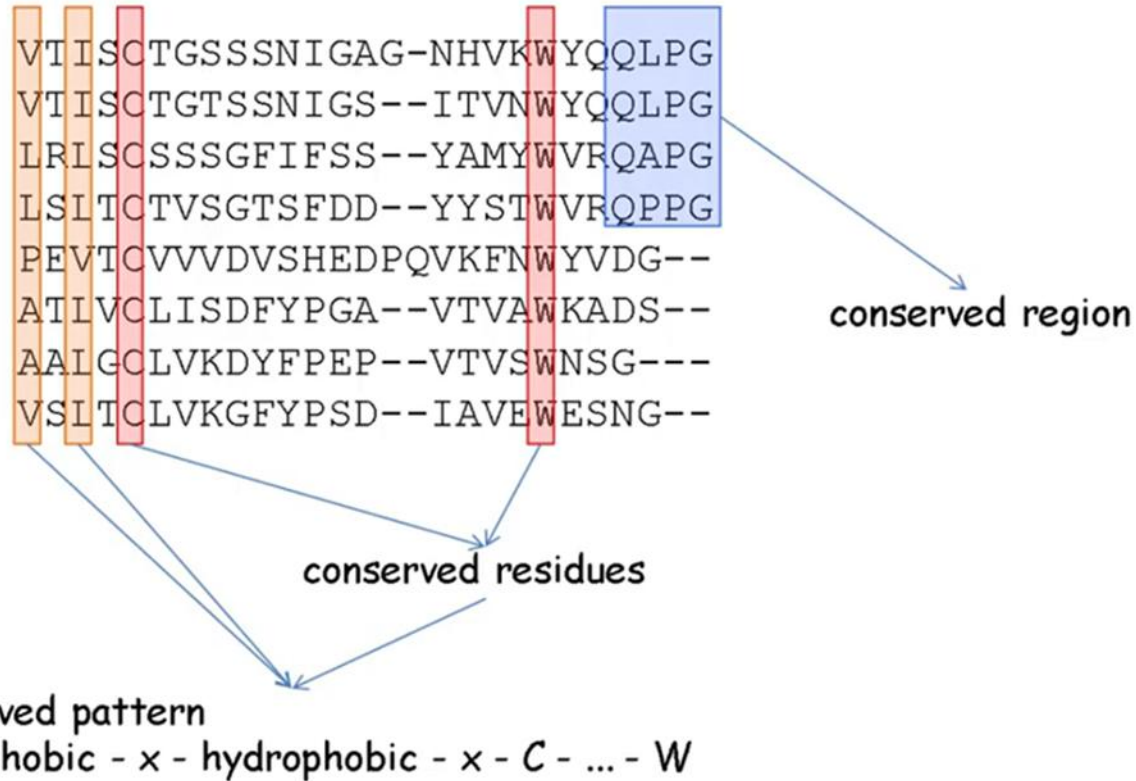
(g) Greek key



(h) β -sandwich

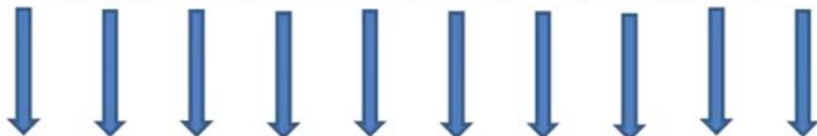


Multiple sequence alignment



Position specific scoring matrices (PSSM)

gene1:	A	A	G	A	G	T	—	—	A	A
gene2:	A	A	G	A	C	—	—	—	T	A
gene3:	G	A	G	A	C	T	G	C	T	A
gene4:	G	A	—	A	C	C	G	C	A	A
gene5:	T	A	G	T	G	C	G	C	T	A



%A:	40	100	0	80	0	0	0	0	40	100
%C:	0	0	0	0	60	40	0	60	0	0
%G:	40	0	80	0	40	0	60	0	0	0
%T:	20	0	0	20	0	40	0	0	60	0

$$f_{u,a} = \frac{n_{u,a}}{N_{seq}}$$

- $n_{u,a}$: number of residues of type a at column u
- N_{seq} : number of sequences

Position specific scoring matrices (PSSM)

protein1:	A	R	S	N	C	P	—	—	A	A
protein2:	A	R	S	N	C	—	—	—	T	A
protein3:	L	R	C	N	C	P	G	C	T	A
protein4:	L	R	—	D	C	C	G	C	A	A
protein5:	I	R	C	D	G	C	G	C	T	A
	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
%A:	40	0	0	0	0	0	0	0	40	100
%R:	0	100	0	0	0	0	0	0	0	0
%N:	0	0	0	60	0	0	0	0	0	0
%D:	0	0	0	40	0	0	0	0	0	0
%C:	0	0	40	0	80	40	0	60	0	0
...

Position specific scoring matrices (PSSM)

Pseudo-counts

protein1:

protein2:

protein3:

protein4:

protein5:

fakeprotein A:

fakeprotein R:

fakeprotein N:

...

A	R	S	N	C	P	—	—	A	A
A	R	S	N	C	—	—	—	T	A
L	R	C	N	C	P	G	C	T	A
L	R	—	D	C	C	G	C	A	A
I	R	C	D	G	C	G	C	T	A
A	A	A	A	A	A	A	A	A	A
R	R	R	R	R	R	R	R	R	R
N	N	N	N	N	N	N	N	N	N
...

Position specific scoring matrices (PSSM)

Correction for lack of or bias in data

- Use pseudocounts: $f'_{u,a} = \frac{n_{u,a} + 1}{N_{seq} + 20}$
- Use pseudocounts by background frequencies: $f'_{u,a} = \frac{n_{u,a} + \beta p_a}{N_{seq} + \beta}$
 - *Lower the contribution of pseudocounts or substitution matrix if there is enough data*
- Use of substitution matrix: $f'_{u,a} = \sum_b f_{u,b} S_{a,b}$
- Weight the sequence contributions
 - *Lower the weights of highly similar sequences*

Representing profiles using logos

- Entropy (uncertainty) in a column: $H_u = - \sum_a f_{u,a} \log_2(f_{u,a})$
- Information: $I_u = \log_2 20 - H_u$
- Contribution of a residue a: $I_{u,a} = f_{u,a} I_u$



Sequence Logo example

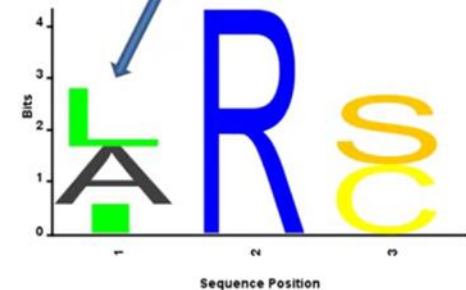
protein1:	A	R	S	N	C	P	—	—	A	A
protein2:	A	R	S	N	C	—	—	—	T	A
protein3:	L	R	C	N	C	P	G	C	T	A
protein4:	L	R	—	D	C	C	G	C	A	A
protein5:	I	R	C	D	G	C	G	C	T	A

%A:	40	0	0	0	0	0	0	0	40	100
%R:	0	100	0	0	0	0	0	0	0	0
%N:	0	0	0	60	0	0	0	0	0	0
%D:	0	0	0	40	0	0	0	0	0	0
%C:	0	0	40	0	80	40	0	60	0	0
...



$$\begin{aligned}
 H_1 &= -(f_A * \log_2 f_A + f_L * \log_2 f_L + f_I * \log_2 f_I) \\
 &= -(0.4 * \log_2 0.4 + 0.4 * \log_2 0.4 + 0.2 * \log_2 0.2) \\
 &= 1.52 \\
 I_1 &= \log_2 20 - H_1 = 2.8
 \end{aligned}$$

$$\begin{aligned}
 I_{1,A} &= 0.4 * 2.8 = 1.12 \\
 I_{1,L} &= 0.4 * 2.8 = 1.12 \\
 I_{1,I} &= 0.2 * 2.8 = 0.56
 \end{aligned}$$

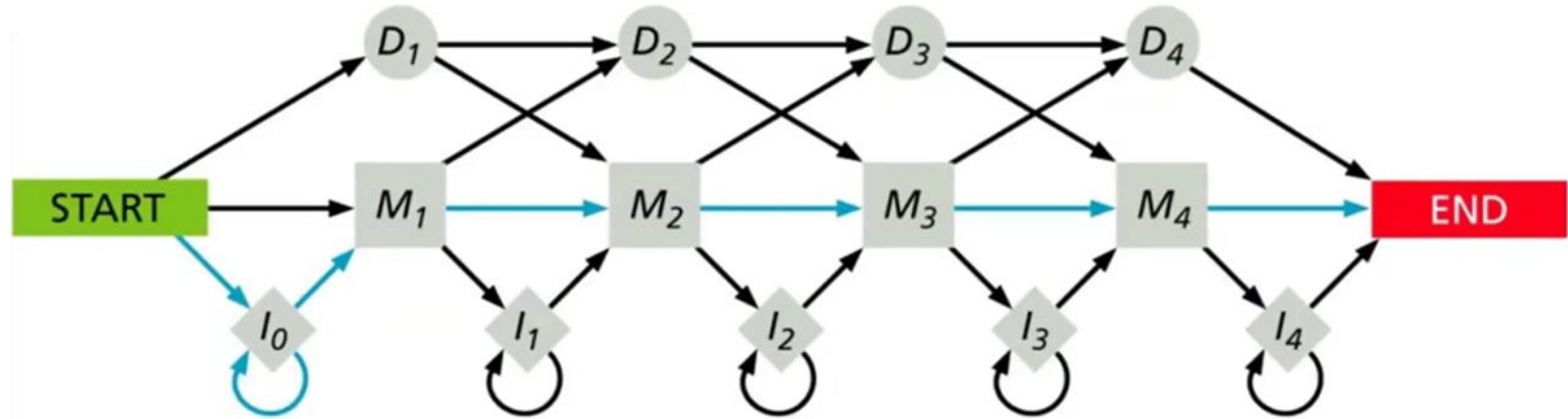


Prosite Patterns

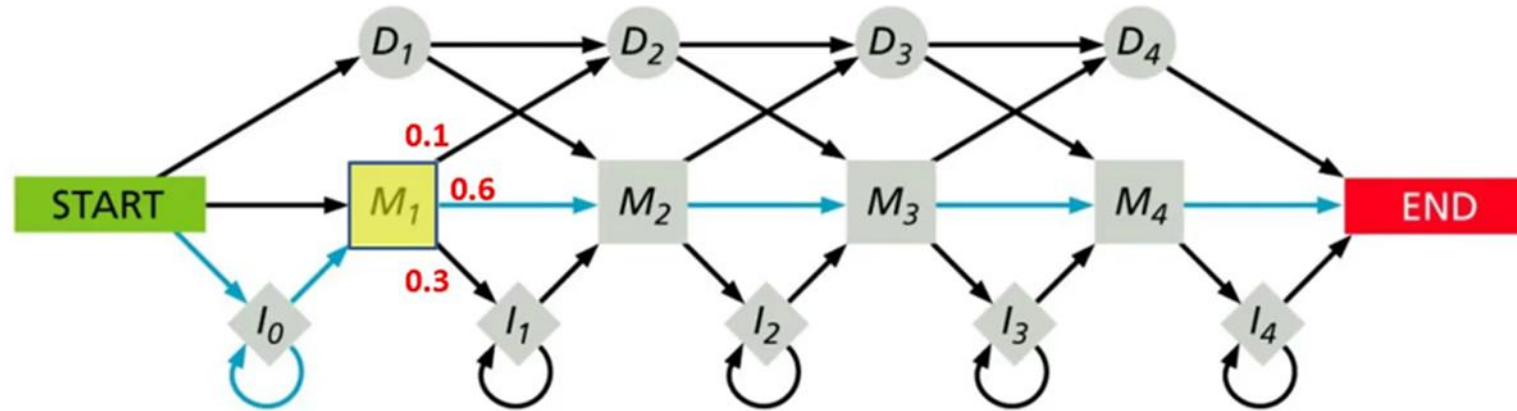
- [AC]-x-V-x(4)-{ED}
 - [Ala or Cys]-any-Val-any-any-any-any-{any but Glu or Asp}
- <A-x-[ST](2)-x(0,1)-V
 - Nterminal Ala-any-[Ser or Thr]-[Ser or Thr]-(any or none)-Val

Prosite	Regular Expression	
x	.	any character
[ALT]	[ALT]	any of A, L, or T
{AM}	[^AM]	anything but A or M
A(3)	A{3}	AAA
A(2,4)	A{2,4}	AA, or AAA, or AAAA
<A	^A	A at the N-terminus
A>	A\$	A at the C-terminus

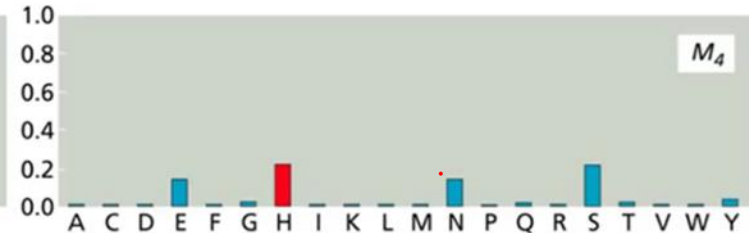
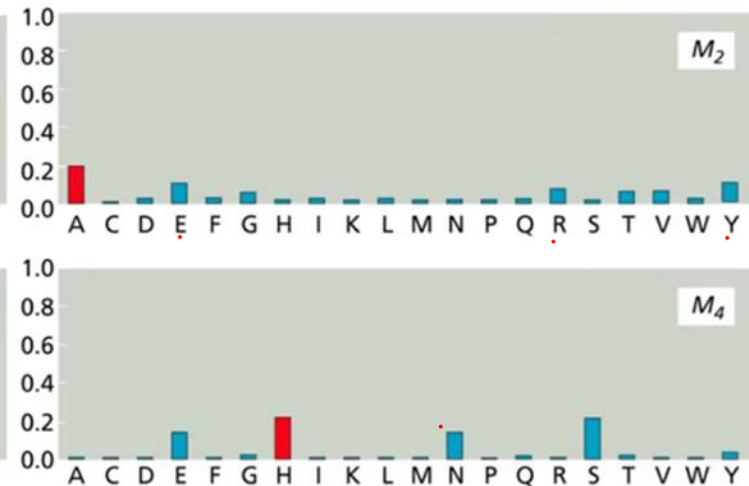
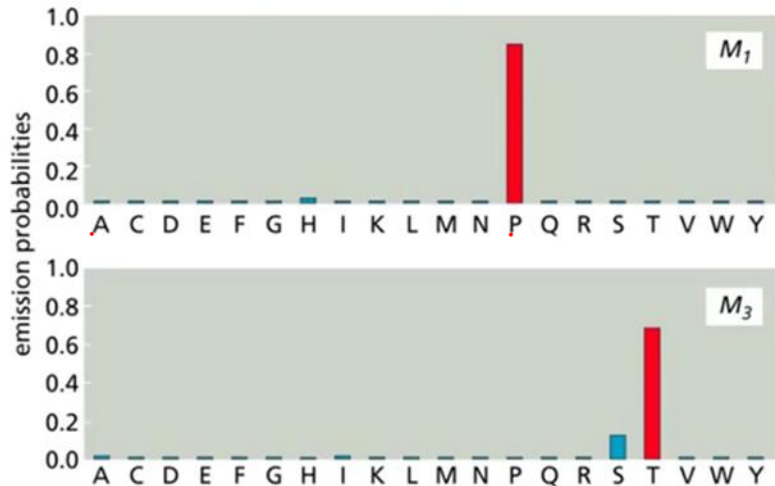
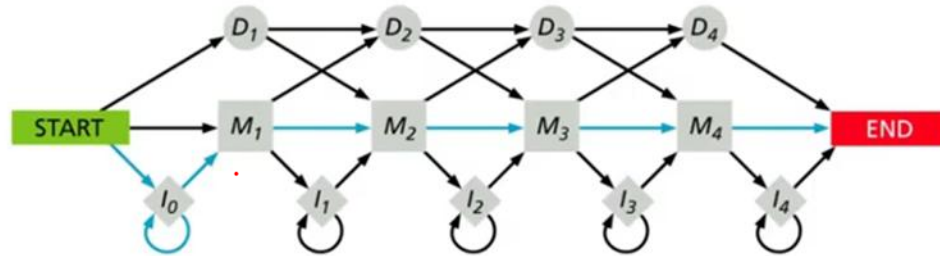
Hidden Markov Model (HMM)



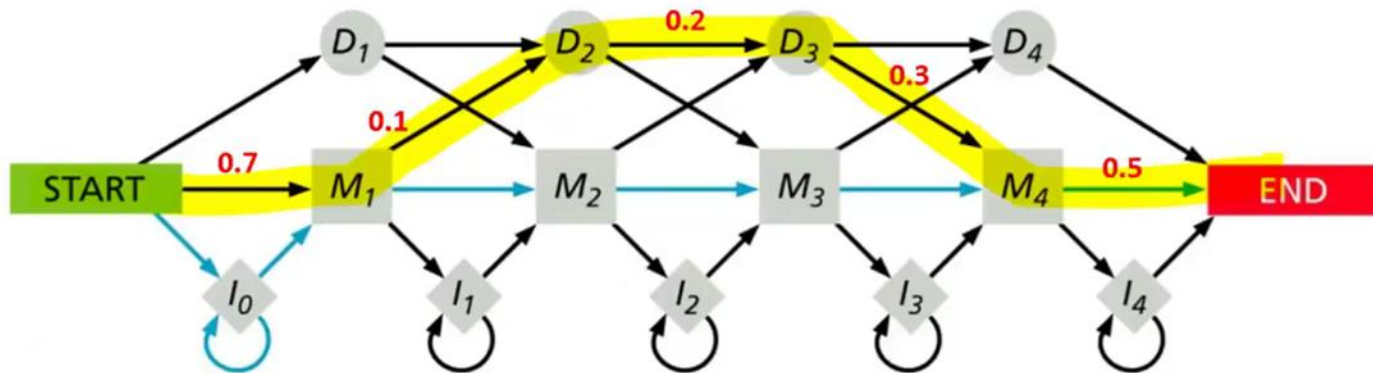
Hidden Markov Model (HMM)



Hidden Markov Model (HMM)



Hidden Markov Model (HMM)



- $M_{1,P} = 0.9$, $M_{4,N} = 0.2$
- Path probability = $0.7 * 0.9 * 0.1 * 0.2 * 0.3 * 0.2 * 0.5 = 3.8e-04$
- Sequence produced: **PN**

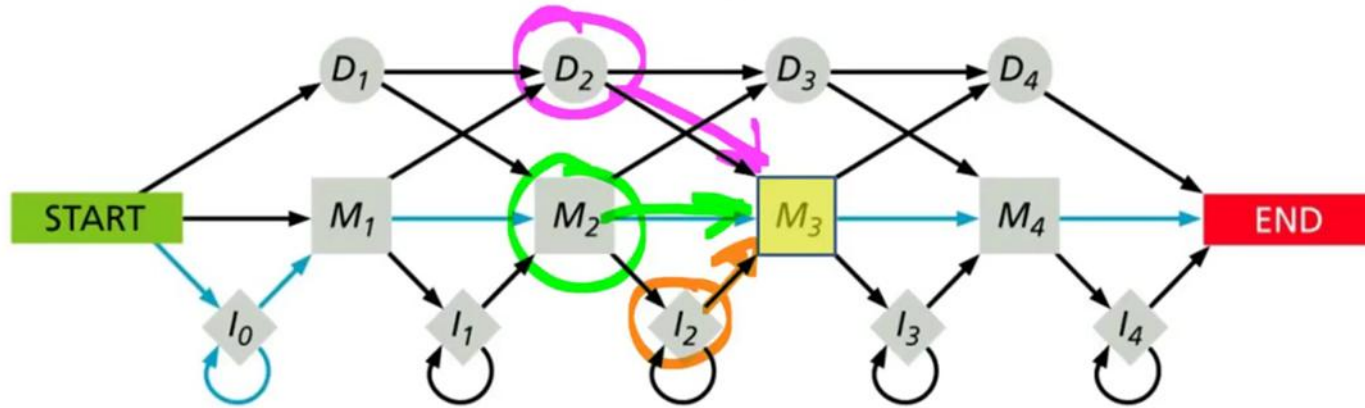
HMM questions

- What is the most likely path?
- What is the probability of a sequence being produced?
- How do we construct the HMM and identify its parameters?

HMM answers

- Dynamic programming: Probability of a node can be decomposed into probabilities of transitioning into it from previous states.
- The most likely path
 - Viterbi algorithm
 - $\max()$ of each previous path

Viterbi algorithm

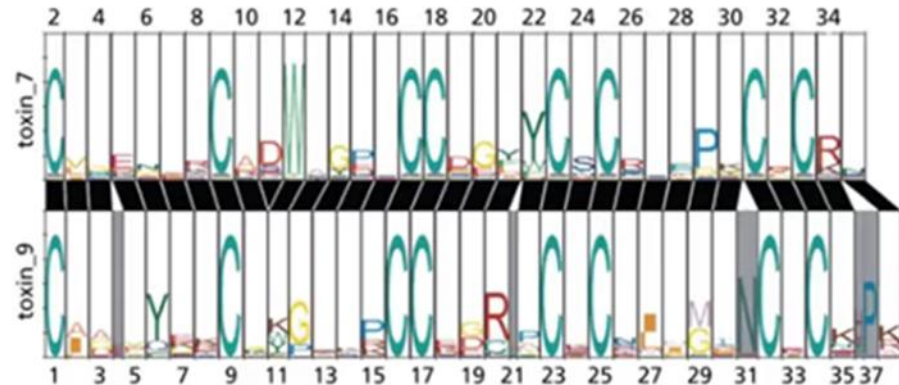
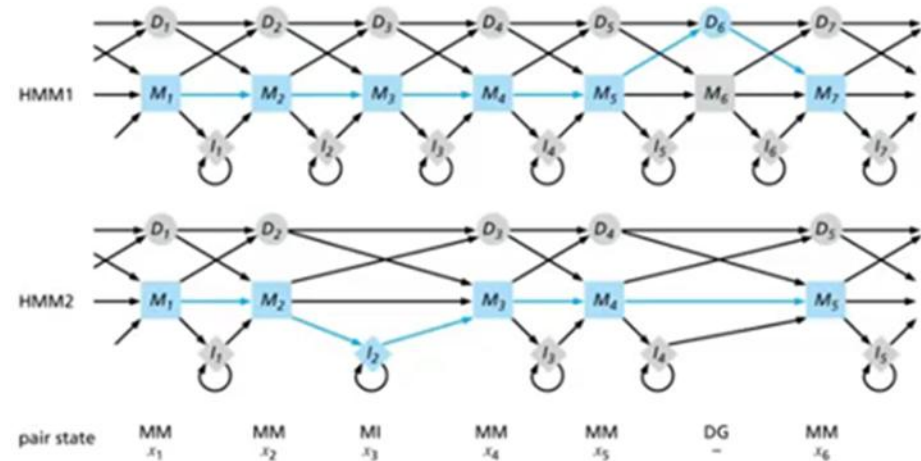


H_{M_3} , most likely path into M_3 :

$$\max \begin{cases} H_{D_2} * T_{D_2 \rightarrow M_3} \\ H_{M_2} * T_{M_2 \rightarrow M_3} \\ H_{I_2} * T_{I_2 \rightarrow M_3} \end{cases}$$

Aligning Families

- Two HMMs can be aligned
 - COACH, HHSEARCH programs

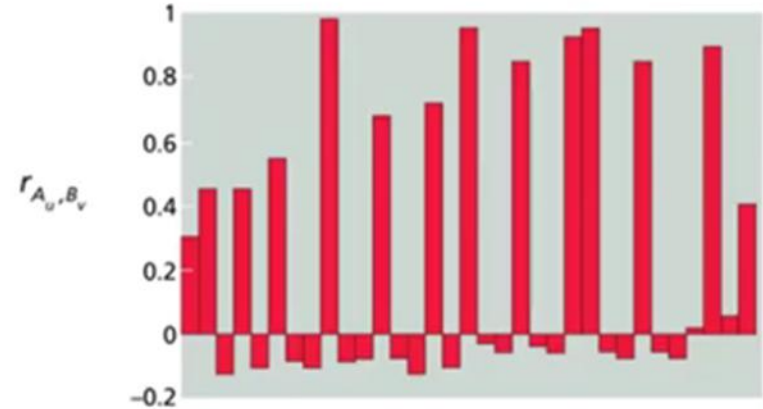


Aligning Families

- Two PSSMs can be compared using Pearson correlation coefficient
 - LAMA program

OXDA_FUSSO	319	LDDETWIV	HNYGHS	GWGYQGSY	GCAENVVQ	LVD	351
OXDD_BOVIN	294	DSRRLPVV	HHYGHGSG	GIAMHW	GTAL	EATR	LVN 326
OXDA_HUMAN	299	GPSNTEVI	HNYGHGGY	GLTIHWGC	AL	EAAK	LFG 331
OXDA_MOUSE	297	GSSSAEVI	HNYGHGGY	GLTIHWGC	AMEAN	LFG	329
OXDA_PIG	299	GSSNTEVI	HNYGHGGY	GLTIHWGC	AL	EAAK	LFG 331
OXDA_RABIT	299	GPSKTEVI	HNYGHGGY	GLTIHWGC	AL	EAAK	LFG 331

DHSA_BACSU	229	GEFIQIHPTAIPGDDKLRL	MS	ESARG	EGGRVWT	261
DHSA_ECOLI	234	QDMEMWQFHPTG	IAGAGV	LVTEG	CRGEGGY	LN 266
FRDA_WOLSU	249	GNMEAVQFHPTPLFPS	GILLTEG	CRGDGGI	LRR	281
DHSA_BOVIN	289	QDLEFVQFHPTGIY	GAGCLITEG	CRGEGGI	LIN	321
DHSA_RICPR	238	QDMEFVQFHPTGIY	GAGCLITEG	ARGEGGY	LVN	270
DHSA_YEAST	279	QDLEFVQFHPSGIY	GSGCLITEG	ARGEGGFLVN	311	
FRDA_ECOLI	224	RDMEFVQYHPTGLP	GSGILMTEG	CRGEGGI	LVN	256
FRDA_PROVU	225	RDMEFVQYHPTGLP	GSGILMTEG	CRGEGGI	LVN	257

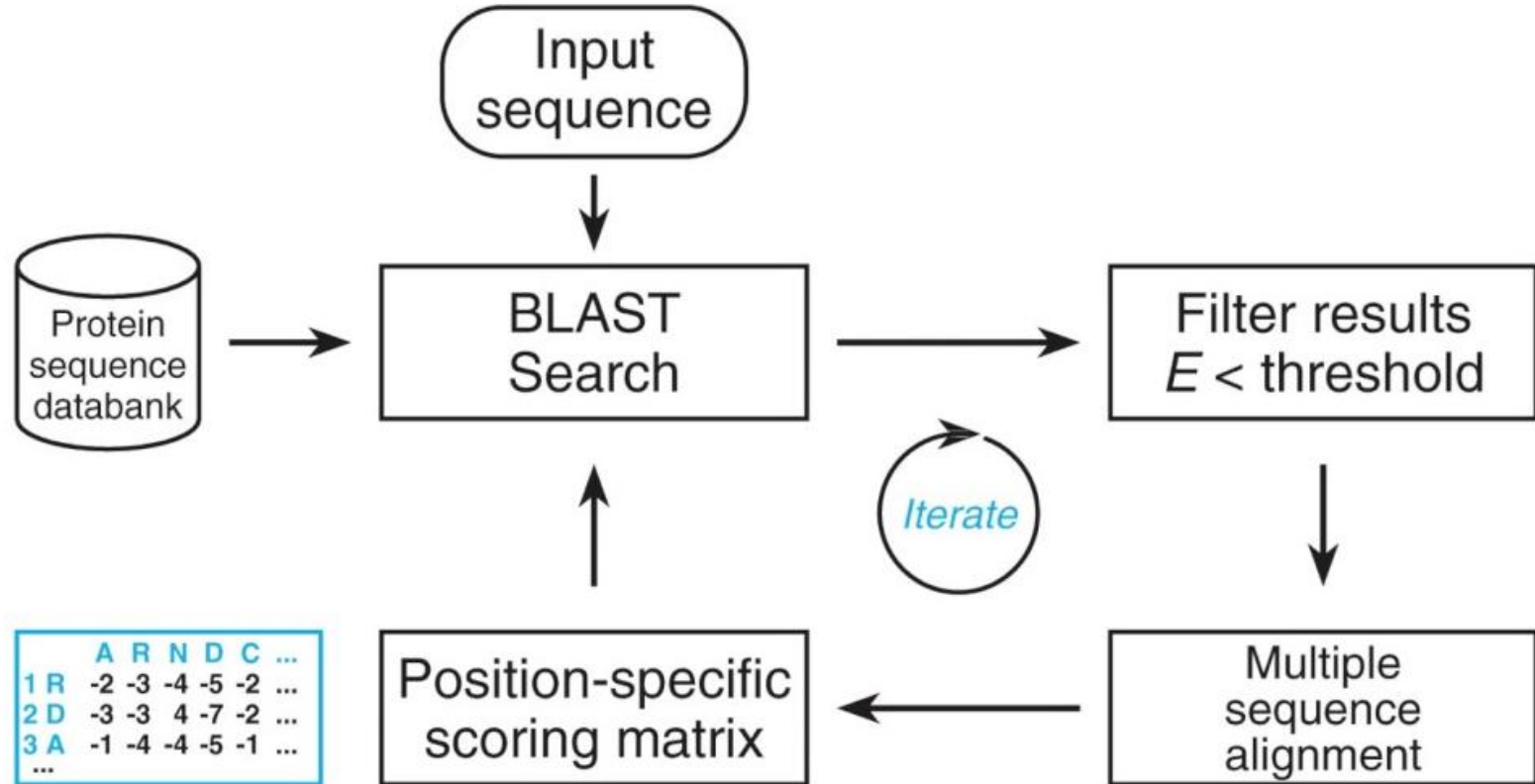


Position-Specific Iterative BLAST (PSI-BLAST)

PSI-BLAST principle

1. A standard BLAST search is performed against a database using a substitution matrix (e.g. BLOSUM62).
2. A PSSM (checkpoint) is constructed automatically from a multiple alignment of the highest scoring hits of the initial BLAST search. High conserved positions receive high scores and weakly conserved positions receive low scores.
3. The PSSM replaces the initial matrix (e.g. BLOSUM62) to perform a second BLAST search.
4. Steps 3 and 4 can be repeated and the new found sequences included to build a new PSSM.
5. We say that the PSI-BLAST has converged if no new sequences are included in the last cycle.

Position-Specific Iterative BLAST (PSI-BLAST)



PSI-BLAST

The purpose of PSI-BLAST is to look deeper into the database for matches to your query protein sequence by employing a scoring matrix that is customized to your query.

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

PSI-BLAST

730496	66	FTVDENGQMSATAKGRVRLFNWWDVCA DMIGSFTD TEDPAKF KMKYWGVASFLQKGNDDH	125
200679	63	FSVDEK GHMSATAKGRVRLLSNWEVCAD MVGTFTD TEDPAKF KMKYWGVASFLQKGNDDH	122
206589	34	FSVDEK GHMSATAKGRVRLLSNWEVCAD MVGTFTD TEDPAKF KMKYWGVASFLQKGNDDH	93
2136812	2	MSATAKGRVRLLSNWDVCA DMVGTFTD TEDPAKF KMKYWGVASFLQKGNDDH	53
132408	65	FKIEDNGKTTATAKGRVRILDKLELCAN MVGTFTIETNDPAKYRMKYHGALAILERGLDDH	124
267584	44	FSVDESGKVTATAHGRVILNNWEMCANMFGTFEDTPDPAKF KMRYWGAAAYLQSGNDDH	103
267585	44	FSVDGSGKVTATAQGRVILNNWEMCANMFGTFEDTPDPAKF KMRYWGAAAYLQSGNDDH	103
8777608	63	FTIHEDGAMTATAKGRVILNNWEMCADMMATFETTPDPAKFRMRYWGAAAYLQSGNDDH	122
6687453	60	FKVEEDGTMTATAIGRVILNNWEMCANMFGTFEDTEDPAKF KMKYWGAAAYLQGYDDH	119
10697027	81	FKVQEDGTMTATATGRVILNNWEMCANMFGTFEDTEEPARFKMRYWGAAAYLQGYDDH	140
13645517	1	MVGTFTD TEDPAKF KMKYWGVASFLQKGNDDH	32
13925316	38	FSVDGSGKMTATAQGRVILNNWEMCANMFGTFEDTPDPAKF KMRYWGAAAYLQSGNDDH	97
131649	65	YTVEEDGTMTASSKGRVKLFGFVVICADMAAOYTDPTTPAKMYMTYOGLASYLSSGGDNY	126

PSI-BLAST

PSI-BLAST is performed in five steps

[1] Select a query and search it against a protein database

[2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)

PSI-BLAST

730496	66	FTVDENGQMSATAKGRVRLFNNWDVCADMIGSFTDTE	PAKFKMKYWG	VASFLQKG	NDDH	125
200679	63	FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTE	PAKFKMKYWG	VASFLQ	RGNDH	122
206589	34	FSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTE	PAKFKMKYWG	VASFLQ	RGNDH	93
2136812	2	MSATAKGRVRLLSNWDVCADMVGTFTDTE	PAKFKMKYWG	VASFLQ	KGNDH	53
132408	65	FKIEDNGKTTATAKGRVRILDKLELCANMVGTFIETND	PAKYRMKYHG	ALAILERGL	DDH	124
267584	44	FSVDESGKVTATAHGRVILNNWEMCANMFGTFEDTPD	PAKFKMRYWG	AASYLQ	TGNDH	103
267585	44	FSVDGSGKVTATAQGRVILNNWEMCANMFGTFEDTPD	PAKFKMRYWG	AAAYLQ	SGNDH	103
8777608	63	FTIHEDGAMTATAKGRVILNNWEMCADMMATFETTPD	PAKFRMRYWG	AASYLQ	TGNDH	122
6687453	60	FKVEEDGTMTATAIGRVILNNWEMCANMFGTFEDTE	PAKFKMKYWG	AAAYLQ	TGYDDH	119
10697027	81	FKVQEDGTMTATATGRVILNNWEMCANMFGTFEDTE	EPARFKMKYWG	AAAYLQ	TGYDDH	140
13645517	1	MVGTFTDTE	PAKFKMKYWG	VASFLQ	KGNDH	32
13925316	38	FSVDGSGKMTATAQGRVILNNWEMCANMFGTFEDTPD	PAKFKMRYWG	AAAYLQ	SGNDH	97
131649	65	YTVEEDGTMTASSKGRVKLFGFWVICADMAAQYTDPT	TPAKMYMTYQ	GLASYLSS	GGDNY	126

R,I,K

C

D,E,T

K,R,T

N,L,Y,G

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	-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	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14 A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15 A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16 A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
...																				
37 S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38 G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39 T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
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

PSI-BLAST

PSI-BLAST is performed in five steps

- [1] Select a query and search it against a protein database
- [2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)
- [3] The PSSM is used as a query against the database
- [4] PSI-BLAST estimates statistical significance (E values)

PSI-BLAST

✓	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

PSI-BLAST

PSI-BLAST is performed in five steps

- [1] Select a query and search it against a protein database
- [2] PSI-BLAST constructs a multiple sequence alignment then creates a “profile” or specialized position-specific scoring matrix (PSSM)
- [3] The PSSM is used as a query against the database
- [4] PSI-BLAST estimates statistical significance (E values)
- [5] Repeat steps [3] and [4] iteratively, typically 5 times. At each new search, a new profile is used as the query.

PSI-BLAST

Results of a PSI-BLAST search

<u>Iteration</u>	<u># hits</u>	<u># hits > 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

PSI-BLAST

PSI-BLAST alignment of RBP and b-lactoglobulin: iteration 1

Score = 46.2 bits (108), Expect = 2e-04

Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27 VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLNNWDVC 86
V+ENFD ++ G WY + +K P + I A +S+ E G + K ++

Sbjct: 33 VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIENVLNK-----ELS 82

Query: 87 ADMVGTF-----TDTEDPAKFKMKYWGVASFLQKGNDHWDVTDYDTYAVQYSCR 137
D GT ++ +PAK +++++ + +WI+ TDY+ YA+ YSC

Sbjct: 83 PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
L ++D + ++ R+P LPPE

Sbjct: 136 TFFWLFHVD-----FFWILGRNPY-LPPE 158

PSI-BLAST

PSI-BLAST alignment of RBP and b-lactoglobulin: iteration 2

Score = 140 bits (353), Expect = 1e-32

Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)

Query: 4 VWALLLLAAWAAAERDCRVSSF-----RVKENFDKARFSGTWYAMAKKDPEGLFLQD 55
V L+ LA A + +F V+ENFD ++ G WY + +K P +

Sbjct: 2 VTMLMFLATLAGLFTTAKGQNFHLGKCPSPVQENFDVKKYLGRWYEI-EKIPASFEKGN 60

Query: 56 NIVAEFSVDETGQMSATAKGRVRLNNWDVCADMV---GTFTDTEDPAKFKMKYWGVASF 112
I A +S+ E G + K + D + V ++ +PAK +++++ +

Sbjct: 61 CIQANYSLMENGNIIEVLNKL-----SPDGTMNQVKGEAKQSNVSEPAKLEVQFFPL--- 112

Query: 113 LQKGNDHWHIVD TDYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA 164
+WI+ TDY+ YA+ YSC L ++D + ++ R+P LPPE

Sbjct: 113 --MPPAPYWILATDYENYALVYSCTTFFWLFHVD-----FFWILGRNPY-LPPET 159

PSI-BLAST

PSI-BLAST alignment of RBP and b-lactoglobulin: iteration 3

Score = 159 bits (404), Expect = 1e-38

Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3 WVWALLLLAAWAAAERD-----CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
 V L+ LA A + S V+ENFD ++ G WY + K

Sbjct: 1 MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55 DNIVAEFSVDETGQMSATAKGRVRLNNDVCDMVGTFDTEDPAKFKMKYWGVASFLQ 114
 + I A +S+ E G + V + ++ +PAK +++++ +

Sbjct: 60 NCIQANYSLMENGNIIEVLNKEKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115 KGNDDHWIVD TDYDTYAVQYSCRLNLDGTCADSYSFVFSRDPNGLPPEA 164
 +WI+ TDY+ YA+ YSC + ++ R+P LPPE

Sbjct: 113 MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159

PSI-BLAST

1

Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

```
Query: 27 VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFVSDETQMSATAKGRVRLNNWDVC 86
          V+ENFD ++ G WY + +K P      + I A +S+ E G +      K      ++
Sbjct: 33 VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIIEVLNK-----ELS 82

Query: 87 ADMVGTF-----TDTEDPAKFKMKYWGVSFLQKGNDDHWIVD TDYDTYAVQYSCR 137
          D  GT              ++ +PAK +++++ +              +WI+ TDY+ YA+ YSC
Sbjct: 83 PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
          L ++D      + ++  R+P  LPPE
Sbjct: 136 TFFWL FHV D-----FFWILGRNPY-LPPE 158
```

3

Score = 159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

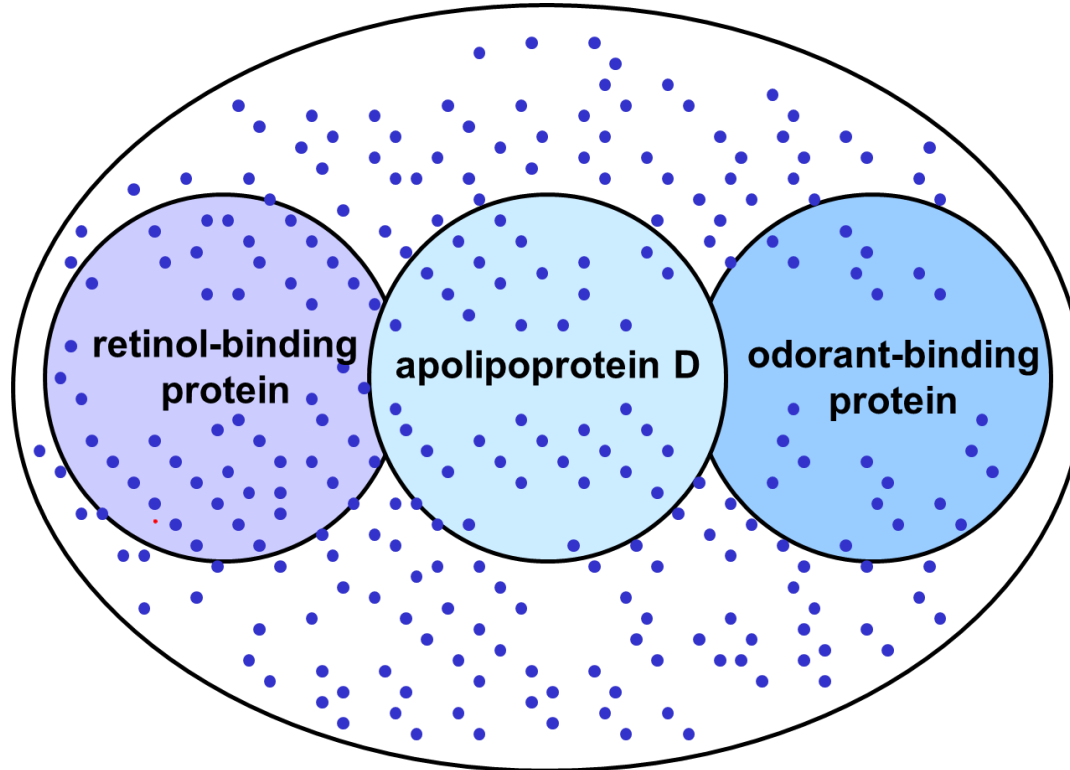
```
Query: 3  WVWALLLLA AAAAAERD-----CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
          V  L+ LA  A      + S  V+ENFD ++ G WY + K
Sbjct: 1  MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYIEIKIPASFE-KG 59

Query: 55 DNIVAEFVSDETQMSATAKGRVRLNNWDVCADMVGTF TDTEDPAKFKMKYWGVSFLQ 114
          + I A +S+ E G +      K      V  +      ++ +PAK +++++ +
Sbjct: 60 NCIQANYSLMENGNIIEVLNKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115 KGNDDHWIVD TDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEA 164
          +WI+ TDY+ YA+ YSC      + ++  R+P  LPPE
Sbjct: 113 MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159
```

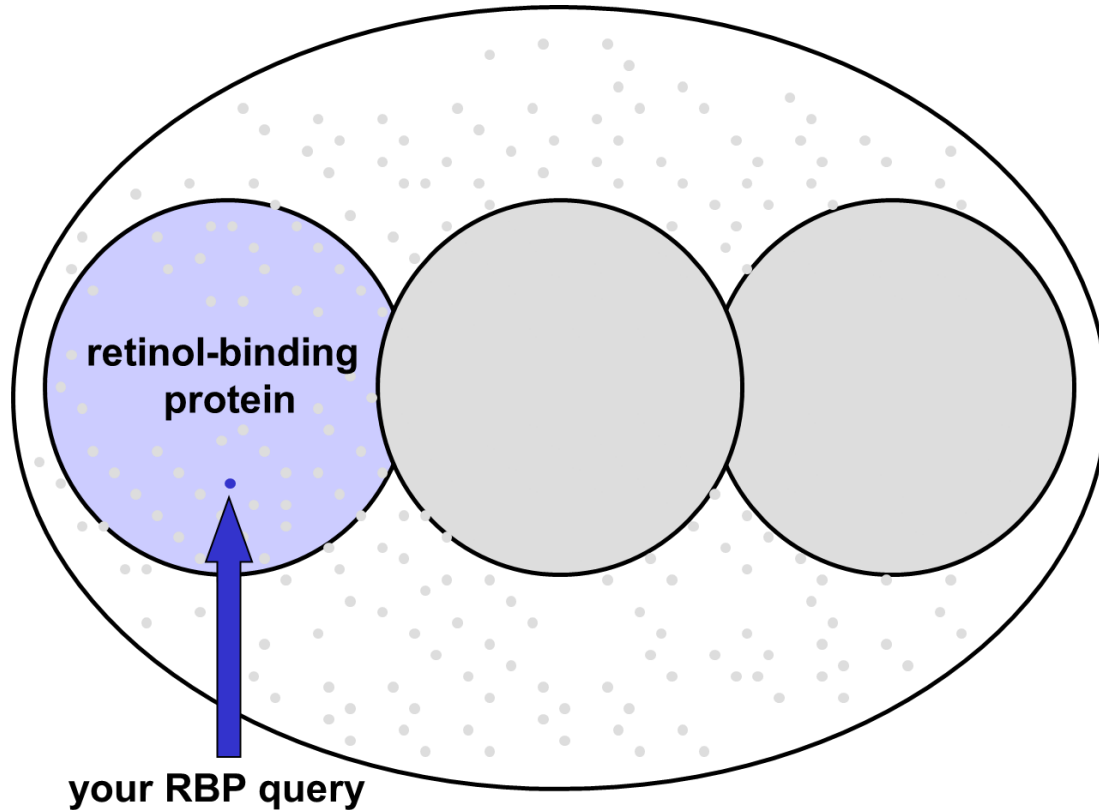
PSI-BLAST

The universe of lipocalins (each dot is a protein)



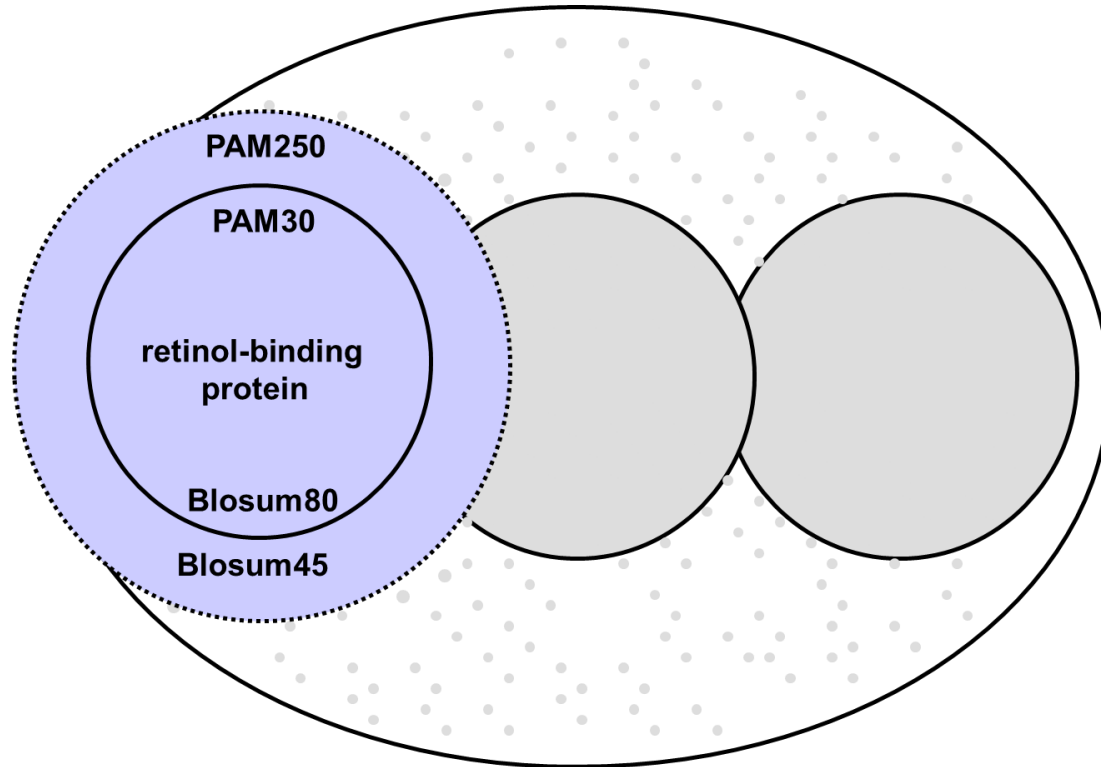
PSI-BLAST

Scoring matrices let you focus on the big (or small) picture



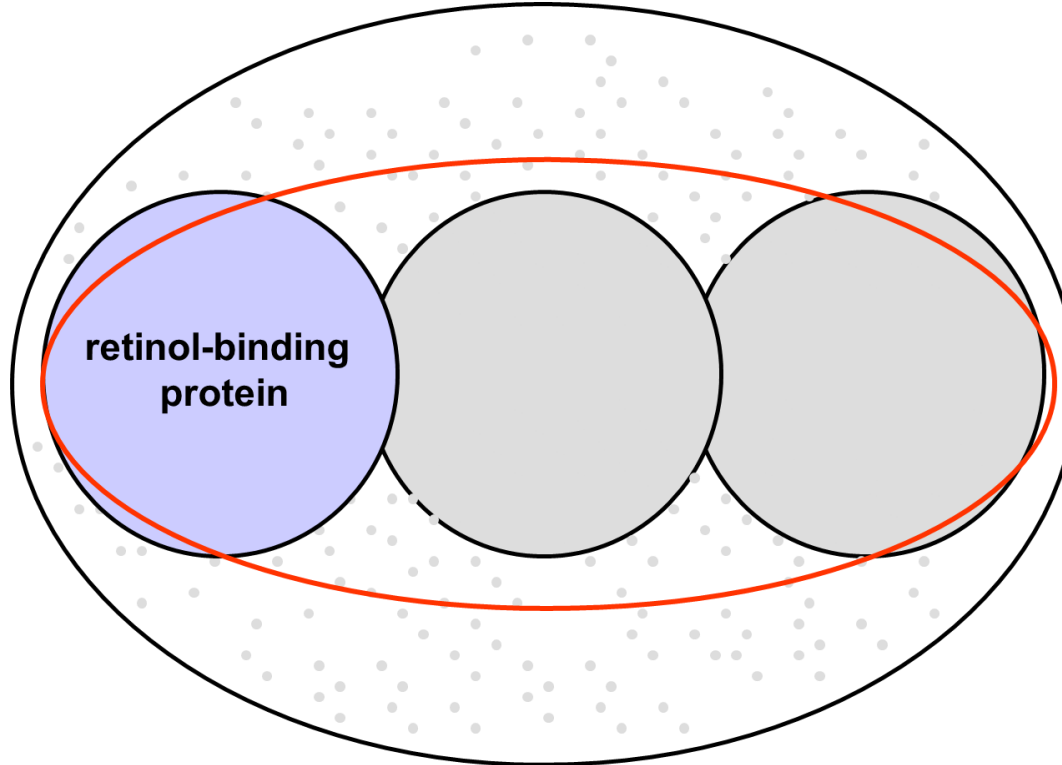
PSI-BLAST

Scoring matrices let you focus on the big (or small) picture



PSI-BLAST

PSI-BLAST generates scoring matrices more powerful than PAM or BLOSUM



Advantages and Disadvantages of PSI-BLAST

Advantages

- [1] Fast (40 times faster than DP)
- [2] Good E-value estimates
- [3] Useful to detect weak but biologically meaningful relationships between proteins

Disadvantages

- [1] Not optimal alignments

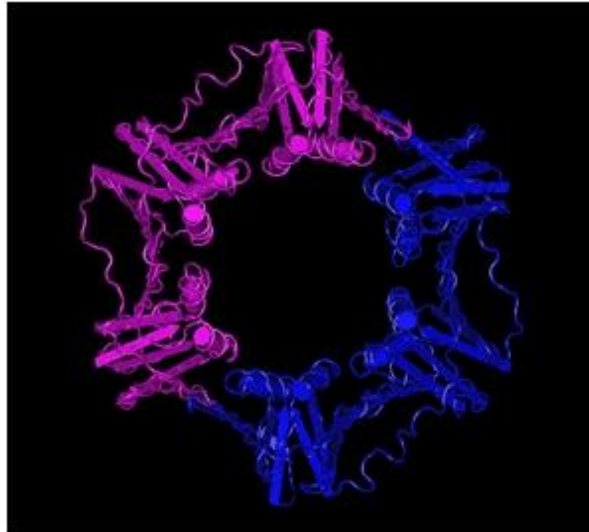
PSI-BLAST: the problem of corruption

- The main source of false positives is the spurious amplification of sequences not related to the query. For instance, a query with a coiled-coil motif may detect thousands of other proteins with this motif that are not homologous.
- Once even a single spurious protein is included in a PSI-BLAST search above threshold, it will not go away.

Example -1

Cellular DNA polymerase enzymes tend to dissociate from DNA after adding a few nucleotides and require an accessory factor to tether them to DNA while elongating the growing DNA chain.

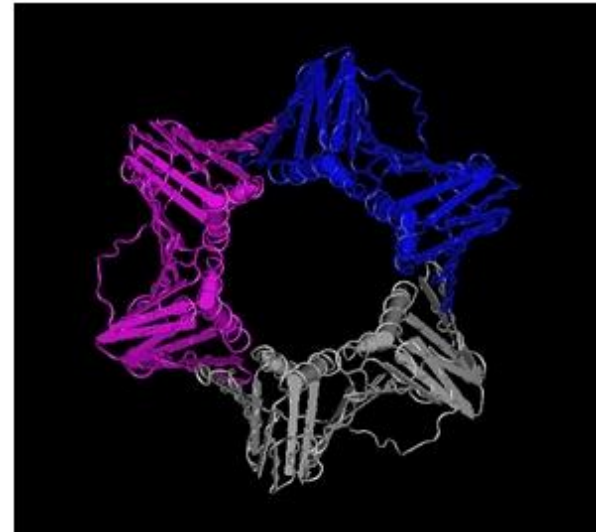
PCNA



Ring structure
called
 β -clamp

Eukaryotes and Archaea

β -subunit of DNA polymerase



Prokaryotes

Example -1

```
Query 5 RLVQGSILKKVLEALKDLINEACWDISSSGVNLQSMDSHVSQTLTLRSEGFDTYRCDR 64
      RL++ + + ++ +N ++ + + D +++ + +
Sbjct 137 RLIEATQFSMAHQDVRYYLNGMLFETEGEELRTVATDGHRLAVCSMPIGQSL-----P 189

Query 65 NLAGVNLTSMSKILKCAGNEDIITLRAEDNADTLALVFEAPNQEKVSD---YEMKLMDL 121
      ++ V + ++++ + + L + + N + KL+D
Sbjct 190 SHSVIVPRKGVIELMRML-----DGGDNPLRVQIGSNNIRAHVGDFIFTSKLVDG 239

Query 122 DVEQL-GIPEQEYSCVVKMPSGEFARICRDLSHIGDA---VVISCAKDGVKFSASGELG 176
      + + ++ + + + + V + +++ +K +A+
Sbjct 240 RFPDYRRVLPKNPDKHLEAGCDLLKQAFARAAILSNEKFRGVRLYVSENQLKITANNPEQ 299

Query 177 NGNIKLSQTSNVDKEEEAVTIEMNEPVQLTFALRYLNFFTKATPLSSTVTLSMSADVPLV 236
      + EE +++ F + Y+ A V + + D
Sbjct 300 E-----EAEEILDVTYSGAEMEIGFNVSYVLDVNLALK-CENVRMML-TDSVSS 346

Query 237 VEYKIADMGHLKYYLAP 253
      V+ + A Y + P
Sbjct 347 VQIEDAASQSAAYVVM 363
```

Hydrophobic amino acids

- Alanine - Ala - A
- Isoleucine - Ile - I
- Leucine - Leu - L
- Methionine - Met - M
- Phenylalanine - Phe - F
- Valine - Val - V
- Proline - Pro - P
- Glycine - Gly - G

Charged amino acids

- Arginine - Arg - R
- Lysine - Lys - K
- Aspartic acid - Asp - D
- Glutamic acid - Glu - E