

Hidden Markov Models and other Multiple-sequence Profile approaches

Mount, Chapter 4, pp. 185-192
Durbin et al, Chapter 5 (also earlier chapters)

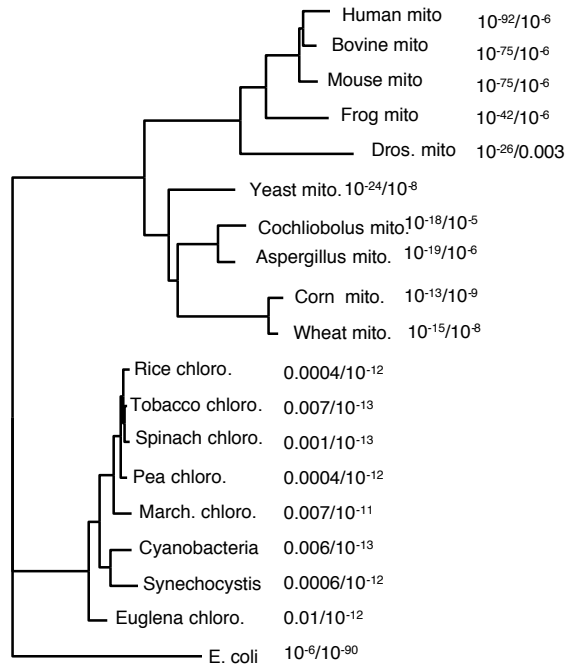
taken from Sean Eddy
Dept. of Genetics
Washington U., St. Louis

HMMs, PSSMs, Motifs, and consensus – representing conserved positions

- HMMs (Hidden Markov Models), PSSMs (Position Specific Scoring Matrices), BLOCKS and Profiles seek to effectively represent conserved positions in *homologous* sequences. Functional conservation after *divergence*
- Some (but not all) motifs and consensus sequences represent conserved positions in *non-homologous* sequences (promoters, modification sites, regulatory sites). Functional acquisition by *convergence*.

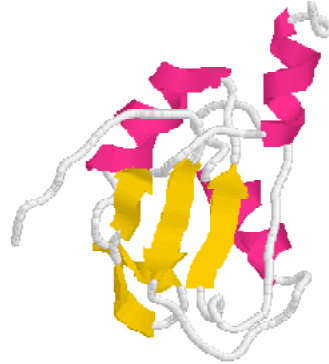
The best scores are:

	s-w bits	E(14548)	% id	alen	
PWHU6 H+-trans. ATP syn. - human mito.	400	326.7	3.3e-90	1.000	226
PWBO6 H+-trans. ATP syn. - cow mito.	157	271.3	1.6e-73	0.779	226
PWMS6 H+-trans. ATP syn. - mouse mito.	118	262.4	7.6e-71	0.757	226
PWXL6 H+-trans. ATP syn. - frog mito.	745	177.3	3.1e-45	0.533	229
PWFF6 H+-trans. ATP syn. - D. melanog.	471	114.8	2.0e-26	0.378	222
PWBY3 H+-trans. ATP syn. - yeast mito.	438	107.3	4.4e-24	0.362	232
PWAS6N H+-trans. ATP syn. - E. nidulans	365	90.6	4.4e-19	0.304	230
PWKQ6 H+-trans. ATP syn. - H. maydis	353	87.9	3.0e-18	0.313	214
PWWT6 H+-trans. ATP syn. - wheat mito.	309	77.8	4.9e-15	0.292	233
PWNT6M H+-trans. ATP syn. - tobacco	309	77.8	5.0e-15	0.283	233
PWZM6M H+-trans. ATP syn. - corn mito.	283	71.9	2.2e-13	0.311	180
LWEC6 H+-trans. ATP syn. - E. coli	178	48.0	3.3e-06	0.237	236
LWRZ6 H+-trans. ATP syn. - rice chloro.	144	40.2	0.00063	0.242	231
PWPMA6 H+-trans. ATP syn. - pea chloro.	143	40.0	0.00074	0.250	232
PWYBAA H+-trans. ATP syn. - Cyano. syn.	142	39.7	0.00099	0.265	170
PWSPA6 H+-trans. ATP syn. - spinach	138	38.9	0.0016	0.238	231
PWYCA6 H+-trans. ATP syn. - Synecho.	127	36.3	0.0099	0.263	167
LWNT6 H+-trans. ATP syn. - tobacco	126	36.1	0.011	0.221	231
LWLV6 H+-trans. ATP syn. - liverwort	126	36.1	0.011	0.244	168
PWEGAC H+-trans. ATP syn. - euglena	123	35.4	0.018	0.257	214
JQ0026 ATP/ADP translocase tlc1 - Ricket	122	35.1	0.045	0.247	154
S17420 ubiquinol--cytochrome-c reductase	113	33.1	0.14	0.228	158
QXBO2M NADH dehydrogenase (ubiquinone)	107	31.7	0.32	0.261	211
S17415 ubiquinol--cytochrome-c reductase	105	31.3	0.49	0.277	137
S17417 ubiquinol--cytochrome-c reductase	104	31.0	0.57	0.277	137
DNHUN2 NADH dehydrogenase (ubiquinone)	103	30.8	0.61	0.201	149
CBHU ubiquinol--cytochrome-c reductase	102	30.6	0.79	0.268	205
QRECA aromatic amino acid trans. prot.	103	30.8	0.82	0.234	111
S17419 ubiquinol--cytochrome-c reductase	101	30.3	0.92	0.234	158





sxl_drome (1sxl)



ru1a_human (u1a)

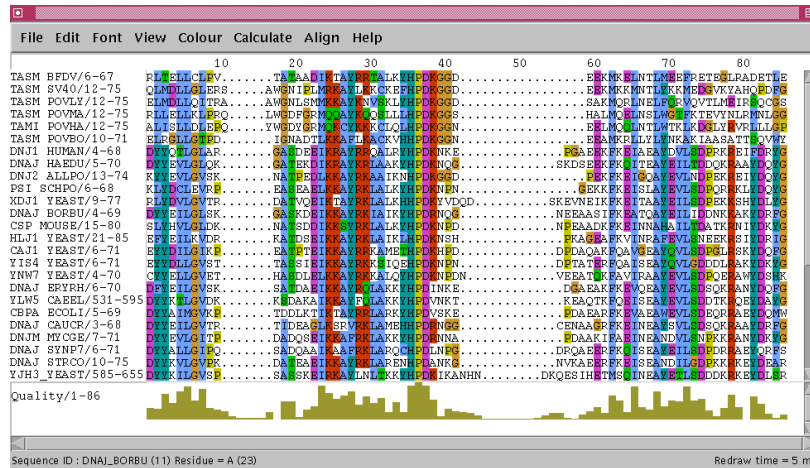
PSI-Blast E()
iteration 1: <7
iteration 2: 10⁻⁸

```

SXL_DROME/127-198 LIVNYL----PQDMTRELIALFRA-IGPINTCRIMRD----YKTGYSFGYAFVDFTSEMSORAIKVLNG-IITVRNKRLLV
U2AF_HUMAN/261-332 LPIGGL----PNYLNDQVKELLS-FGLKAFNLVKD----SATGLSKGYAFCEYVDINVTDOAIALNG-MQLGDKKLLV
U2AF_SCHPO/312-383 IYISNL----PLNLGEDQVVELLKP-FGDLLSFQLIKN----IADGSSKGFCECFKNPDAEVAISGLDG-KDTYGNKLHA
SXL_DROME/213-285 LYVTNL----PRTITDDQLDTIFGK-YGSIVQKNILRD----KLTGRPRGVAFVRYNKRHEAQEAISALNNVIEGGSQPLS
ROC_HUMAN/18-82 VFIGNL----NTLVVKKSDVEAIFSK-YGKIVGCSVHK-----GFAPVQYVNEHNARAAVAGEDG-RMIAGQVLDI
RU1A_HUMAN/12-84 IYINNLNEIKKDELKKSIAIFSQ-FQQLLDILVSR-----SLKMRQAFVIFKEVVSATNALRSMQG-FPFYDKPMRI
RU2B_HUMAN/9-81 IYINNMNDKIKKEELKRSIALFSQ-FGHVVDIVALK-----TMKMRQAFVIFKELASSTNALRQLOG-FPFYDKPMRI
SOD_DROME/138-208 IPVGGI----TTETSDEEIKTYFGQ-FGNIVEVEEMPLD----KQKSORKGFCEITFDSECVVTDLLK-TPK-QKIAGEKVDV
SOD_DROME/138-128 LPVGGI----SWETTEKELRDHFGK-YGEIESINVKTD----PQTGRSRGFVIFVTNTHAIDKVSAA-ADE-HIINSKQVDP
SPR2_CHICK/16-87 LVDNL----TYRTSPDTRRVFEK-YGRVGDVYIPRD----RYTKESRGFVRFHDKHDAEDAMDMDG-AVLDGRELRV
SPR1_HUMAN/17-85 IYVGNL----PPDIRTKDIEDVFK-YGAIRDIDLKNR-----RGGPPFAFVEFDPRDAEDAMVGRDG-YDYDGYRLRV
SR55_DROME/5-68 VYVGGI----PYGVRELDLRFPPK-YGRTRDILIKN-----GYGVFEFEDYRDAADAVYELNG-KELLGERVVV
SPR3_HUMAN/12-78 VYVGNL----GNNKTELEAFPGY-YGRLRSVWVARN-----PPGFAPVEFDPRDAADAVRELDG-RTLCCGRVRY
RNP1_YEAST/37-109 LRVGNL----PKWCRQQLRDLFEPNYKKTINMLKKK----PLKPKLRFAPFIEQGVNLKVEKMBG-KIFMNEKIVI
PES4_YEAST/305-374 IFIXNL----PTITTRDDILNPFSE-VGPIKSIYLSN-----ATKVYLWAFVYKNSDSEKAIKRYNN-FYPRGKLLLV
YHH5_YEAST/315-384 ILVKNL----PSDTTQEEVLDYFST-IGPIKSVFISEK-----QANTPHKAFVYKNEHESKAKCLNK-TIFKNHTIIV
YHC4_YEAST/348-415 IPVGQL----DKETTRELNRFRST-HGKIQDINLIFK-----PTNIFAFIKYTEHAAAAALESENH-AIFLNKTMHV
SPR1_HUMAN/122-186 VVVSGL----PPSGSWQDLKDHMRG-AGDVCIADVYRD-----GTGVVEFVRKDMTYAVRKLDN-TKFRSHEGET
RU1A_HUMAN/210-276 LPLTNL----PEETNELMLSMLFNQ-FPGFKEVRLVPG-----RHDIAFVEFDNEVOAGAARDALQG-FKITQNNAMK
RU2B_HUMAN/153-220 LPLNLL----PEETNELMLSMLFNQ-FPGFKEVRLVPG-----RHDIAFVEFDNEVOAGAARDALQGFKITPSHAMKI
RU1A_YEAST/229-293 LLIQNL----PSGTFEQLLSQILGN-EALV-EIRLVSV-----RNLAFVEYFVADATKIFMQLGS-TRYKQNNVDT
PR24_YEAST/43-111 VLVKNL----PKSVYQKVKYKFKH-CPTIIVQVAD-----SLKKNFRFARIEFARIYGLAALIT-KTH-KVWQNEIIV
PR24_YEAST/212-284 IMIRNL----STELLDENLLRESFEG-FGSIKINIPAG---QKHSFNWCCAFVWFENKHSARALQ-MNR-SLLGNREISV
SSB1_YEAST/39-114 IFVGNV----AHECTEDDLKQLFVEBFGDEVSEIPIK-EHTDGHIPASKHALVKFPTKIDFDNIEKNYDT-KVVKDREIHI
RN12_YEAST/200-267 IYIKFO----GPALTEEEIYSLFRR-YGTI--IDIFP-----PTAANNVAVKRYRSFHGAISAKNCVSG-IEIHNTVLHI
U2AG_HUMAN/67-142 CAVSDVEMQEHYDEFFEEVFTMEEEKYGEVEEMVCDN-----LGDHLVGNVYVYKFRREHDAEKAVIDLNN-RWFNGQPIHA
LA_DROME/151-225 AYAKGF----PLDSQISELLDPTAN-YDKVNLIMRNSYDKPTKSYKFKGSIPLTPTKHQAKAFIE-QEK-IVYKRELLR
LA_HUMAN/113-182 VYIKGF----PTDATLDDIKEWLED-KGQVNIQMR-----TLHKAFKGSIFVVFDSISAKKFEV-TPG-QYKTEDLLI

```

Alignments show conservation



Profiles

- Profiles invented in the late eighties
 - Barton and Sternberg
 - Gribskov
 - Taylor
- Uses a multiple alignment and “rules” to convert frequencies to scores

Profiles – the problems

- Could not compare between two profiles
 - user had to be involved in interpretation
- Gap penalties – again the user had to be involved

HMMs

- Hidden Markov Models have been successfully used for
 - speech recognition
 - passive sonar work
 - other “signal detection” problems

profile-HMMs

- Anders Krogh in David Haussler's group.
- Takes the "standard" profiles and uses HMM based "standard" mathematics to solve two problems
 - Profile-HMM scores are comparable (*)
 - Setting gap costs
- Theoretical framework for what we are doing.
- (* this is not really true. see later)

Pairwise Alignment

```

RU1A_HUMAN rrm2  VOAGAAR
PABP_DROME rrm3  EAAEAAV
  
```

+2 +2 +2

score matrices:
 20x20, 210
 parameters
 position-
independent

Cys	12									
Ser	0	2								
Thr	-2	1	3							
Pro	-1	1	0	6						
Ala	-2	1	1	1	②					
Gly	-3	1	0	-1	1	5				
Asn	-4	1	0	-1	0	0	2			
Asp	-5	0	0	-1	0	1	2	4		
Glu	-5	0	0	-1	0	0	1	3	4	
Gln	-5	-1	-1	0	0	-1	1	2	2	4
	C	S	T	P	A	G	N	D	E	Q

Profile Alignment

RU1A_HUMAN	rrm1	SSATNAL	
RU1A_HUMAN	rrm2	VQAGAAR	query
SFR1_HUMAN	rrm1	RDAEDAV	
SXLF_DROME	rrm1	MDSORAI	
PABP_DROME	rrm3	EAAEAAV	target
		+3 +4	
		0	

profile: 20 scores *per column*
position-*dependent*

Where pairwise scores come from –

$$\text{score}(AA) = \log \frac{P(A|A)}{f(A)}$$

“probability of A given an A”
 the observed probability of seeing an A
 aligned to an A in real alignments

“frequency of A”
 the expected frequency of A in any sequence

$$\text{Sc}(AA) = \log_2 \frac{0.64}{0.04} = +4$$

$$\text{Sc}(AE) = \log_2 \frac{0.01}{0.04} = -2$$

Where profile scores (should) come from

$$\text{score}(A|x) = \log \frac{P(\text{Alposition } x)}{f(A)}$$

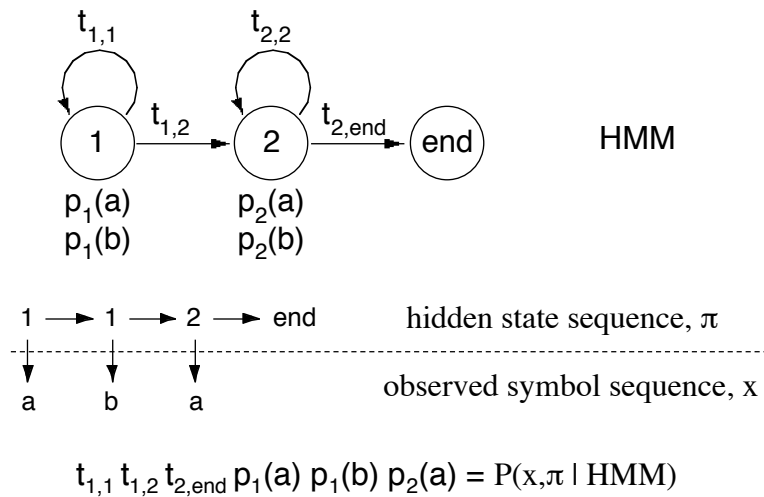
“probability of A at position x”
the observed probability of seeing an A
in the consensus column x

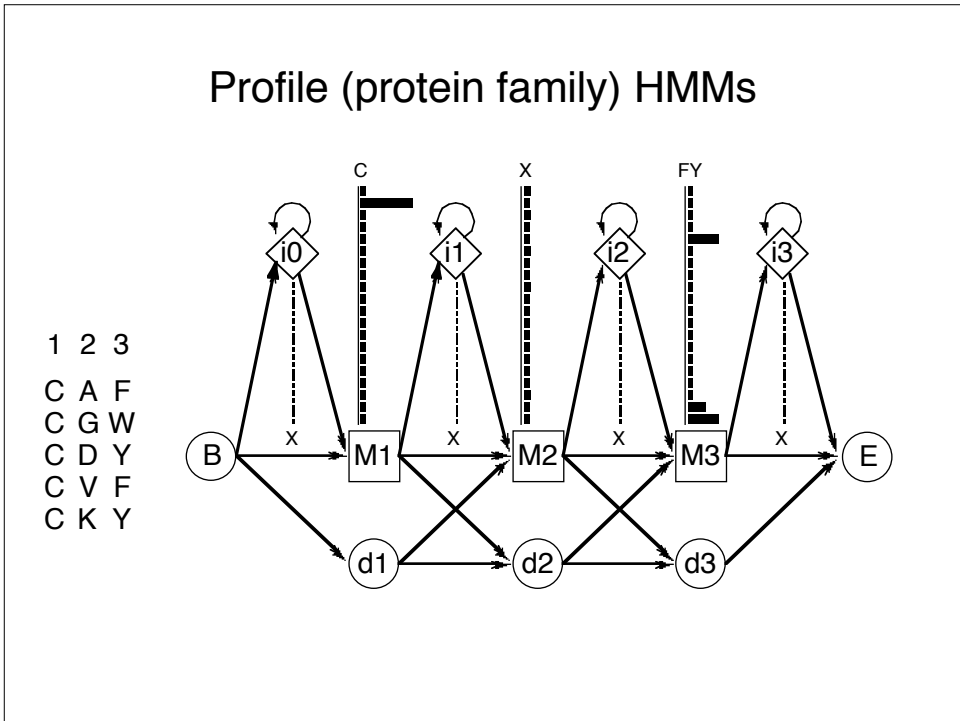
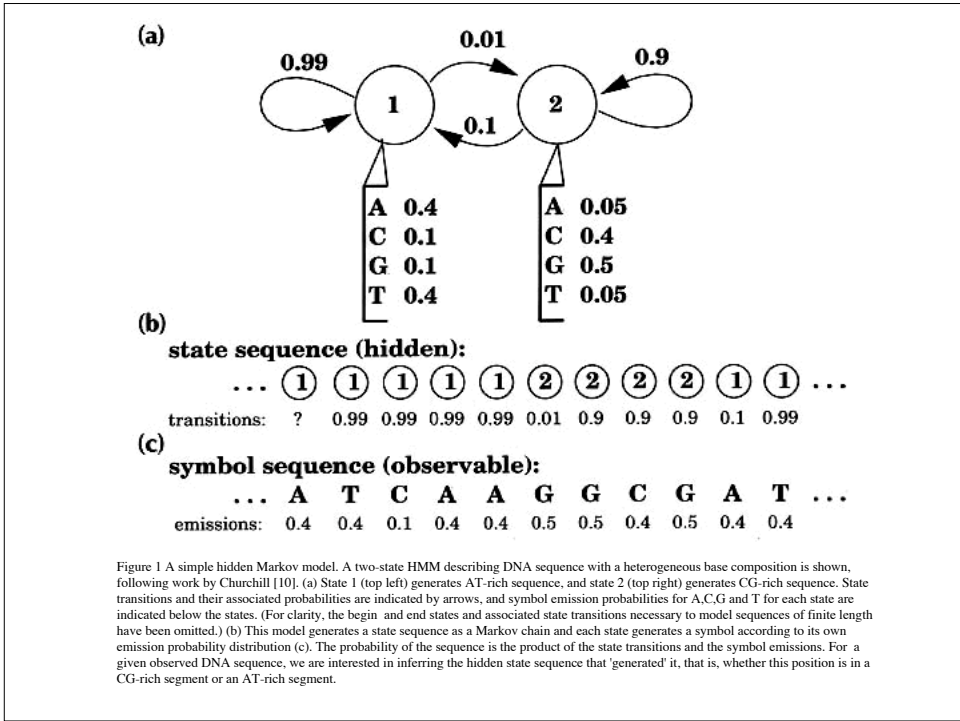
$$\text{Sc}(A|6) = \log_2 \frac{1.00}{0.04} = +4.6 \quad \text{Sc}(A|5) = \log_2 \frac{0.04}{0.04} = 0$$

$$\text{Sc}(N|6) = \log_2 \frac{0.00}{0.06} = -\text{inf} \quad \text{Sc}(N|5) = \log_2 \frac{0.06}{0.06} = 0$$

1. what about position-specific gap penalties?
2. how to estimate parameters from small numbers of observations?

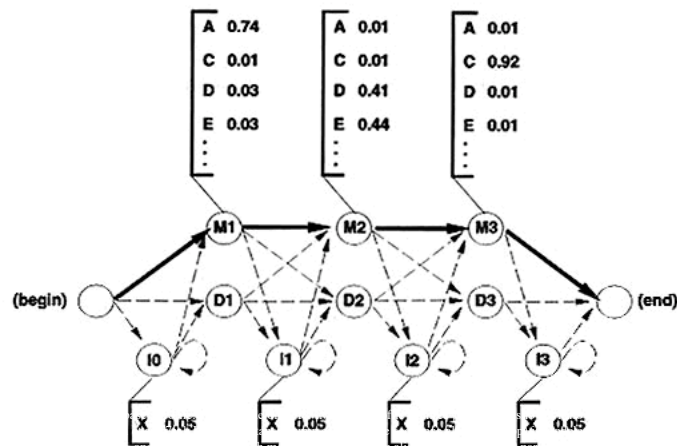
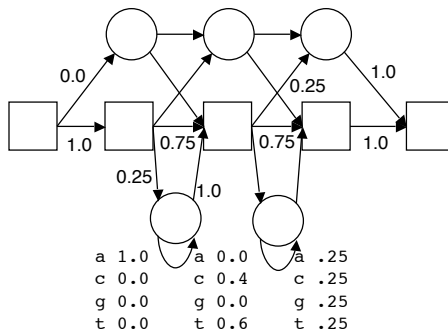
Hidden Markov Models (HMMs)



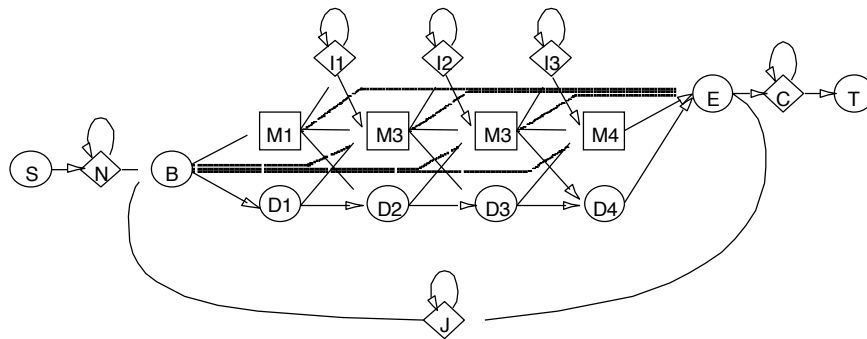


HMM transitions and emissions are probabilities

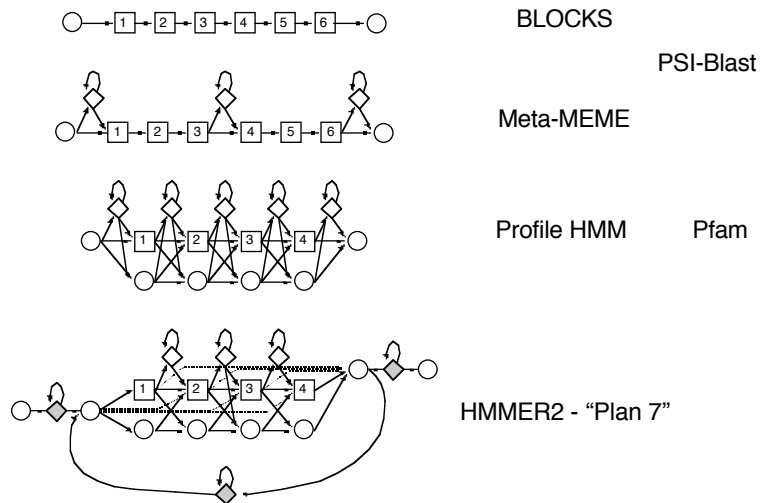
a - c g
 a - t a
 a - c c
 a t t t
 a - c -



HMMER- 'Plan 7' profile HMM



Many approaches are HMM-based (or HMM-like)



HMM Algorithms

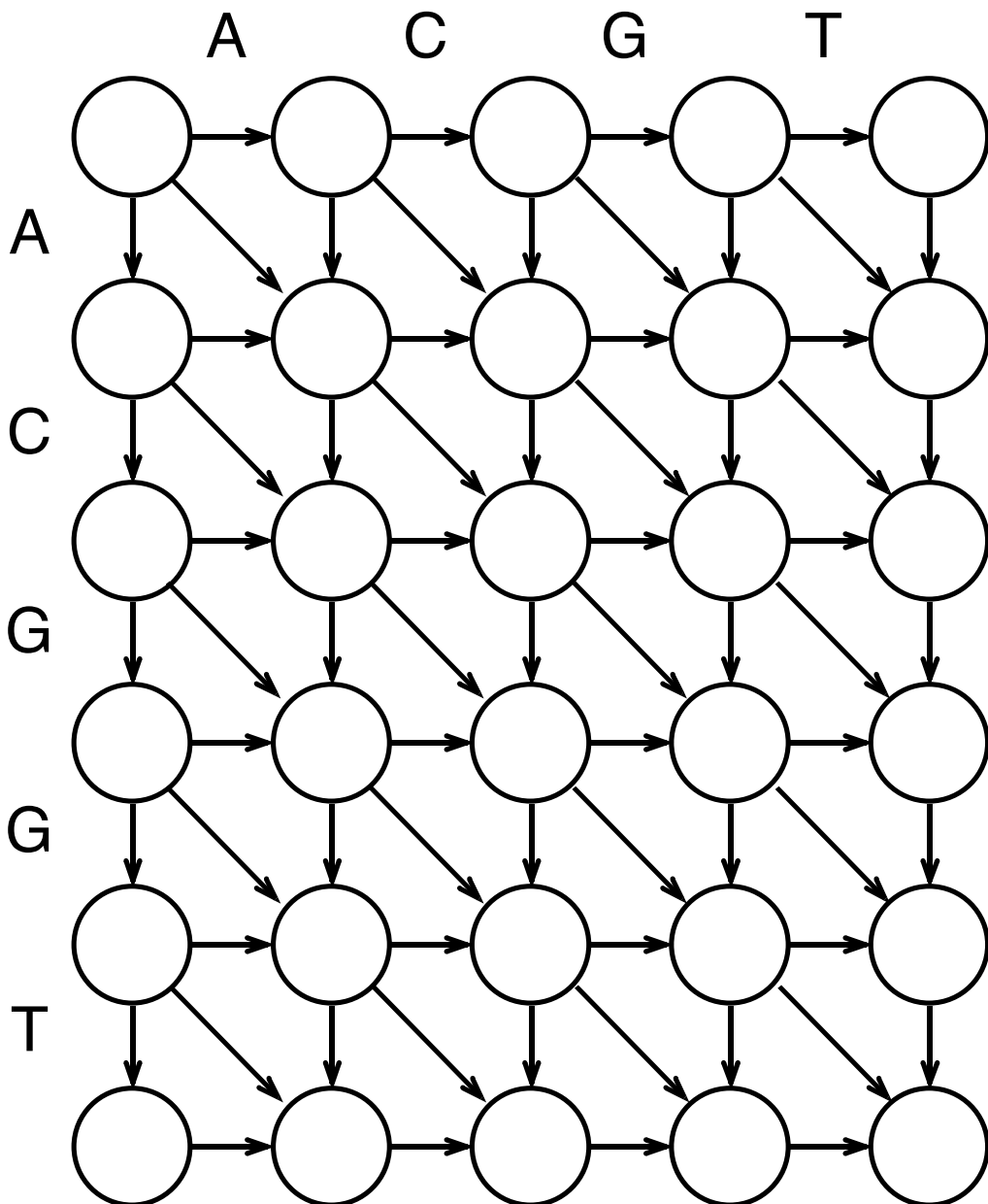
1. The scoring problem: $P(\text{seq} \mid \text{model})$
 "Forward" algorithm
 (sums over all alignments)
2. The alignment problem: $\max P(\text{seq}, \text{statepath} \mid \text{model})$
 "Viterbi" algorithm
3. The training problem:
 "Forward-backward" algorithm and
 Baum-Welch expectation maximization

For profile HMMs, all three algorithms use $O(MN)$ dynamic programming -- same as "standard" Smith/Waterman and Needleman/Wunsch.

Global and Local Alignment Paths

Global	Local																																																																																																																																																																																																																																																																																																																																
<table style="font-family: monospace; border-collapse: collapse;"> <tr><td>A</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td></tr> <tr><td>A</td><td>1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td><td>-1</td></tr> <tr><td>B</td><td>\</td><td>!</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td></tr> <tr><td>B</td><td>-1</td><td>2</td><td>0</td><td>-2</td><td>-2</td><td>-2</td><td>-2</td><td>-2</td><td>-2</td></tr> <tr><td>D</td><td>\</td><td>\</td><td>!</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td></tr> <tr><td>D</td><td>-1</td><td>0</td><td>3</td><td>1</td><td>-1</td><td>-3</td><td>-3</td><td>-3</td><td>-3</td></tr> <tr><td>E</td><td>\</td><td>\</td><td>!</td><td>!</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td></tr> <tr><td>E</td><td>-1</td><td>-2</td><td>1</td><td>2</td><td>2</td><td>0</td><td>-2</td><td>-4</td><td>-4</td></tr> <tr><td>G</td><td>\</td><td>\</td><td>\</td><td>!</td><td>!</td><td>!</td><td>\</td><td>\</td><td>\</td></tr> <tr><td>G</td><td>-1</td><td>-2</td><td>-1</td><td>0</td><td>1</td><td>1</td><td>1</td><td>-1</td><td>-3</td></tr> <tr><td>K</td><td>\</td><td>\</td><td>\</td><td>!</td><td>!</td><td>!</td><td>\</td><td>\</td><td>\</td></tr> <tr><td>K</td><td>-1</td><td>-2</td><td>-3</td><td>-2</td><td>-1</td><td>0</td><td>0</td><td>0</td><td>-2</td></tr> <tr><td>H</td><td>\</td><td>\</td><td>\</td><td>\</td><td>!</td><td>!</td><td>!</td><td>\</td><td>\</td></tr> <tr><td>H</td><td>-1</td><td>-2</td><td>-3</td><td>-4</td><td>-3</td><td>-2</td><td>-1</td><td>1</td><td>-1</td></tr> <tr><td>I</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>!</td><td>!</td><td>!</td><td>!</td></tr> <tr><td>I</td><td>-1</td><td>-2</td><td>-3</td><td>-4</td><td>-5</td><td>-4</td><td>-3</td><td>-1</td><td>2</td></tr> </table>	A	\	\	\	\	\	\	\	\	\	A	1	-1	-1	-1	-1	-1	-1	-1	-1	B	\	!	\	\	\	\	\	\	\	B	-1	2	0	-2	-2	-2	-2	-2	-2	D	\	\	!	\	\	\	\	\	\	D	-1	0	3	1	-1	-3	-3	-3	-3	E	\	\	!	!	\	\	\	\	\	E	-1	-2	1	2	2	0	-2	-4	-4	G	\	\	\	!	!	!	\	\	\	G	-1	-2	-1	0	1	1	1	-1	-3	K	\	\	\	!	!	!	\	\	\	K	-1	-2	-3	-2	-1	0	0	0	-2	H	\	\	\	\	!	!	!	\	\	H	-1	-2	-3	-4	-3	-2	-1	1	-1	I	\	\	\	\	\	!	!	!	!	I	-1	-2	-3	-4	-5	-4	-3	-1	2	<table style="font-family: monospace; border-collapse: collapse;"> <tr><td>A</td><td>\</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr> <tr><td>A</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>B</td><td>\</td><td>\</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr> <tr><td>B</td><td>0</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>D</td><td>\</td><td>!</td><td>\</td><td>\</td><td></td><td></td><td></td><td></td><td></td></tr> <tr><td>D</td><td>0</td><td>0</td><td>3</td><td>1</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>E</td><td>\</td><td>\</td><td>!</td><td>\</td><td>\</td><td></td><td></td><td></td><td></td></tr> <tr><td>E</td><td>0</td><td>0</td><td>1</td><td>2</td><td>2</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>G</td><td>\</td><td>\</td><td>\</td><td>!</td><td>\</td><td>\</td><td>\</td><td></td><td></td></tr> <tr><td>G</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td></tr> <tr><td>K</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td></td></tr> <tr><td>K</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td></tr> <tr><td>H</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td></td></tr> <tr><td>H</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>0</td></tr> <tr><td>I</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td><td>\</td></tr> <tr><td>I</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>2</td></tr> </table>	A	\									A	1	0	0	0	0	0	0	0	0	B	\	\								B	0	2	0	0	0	0	0	0	0	D	\	!	\	\						D	0	0	3	1	0	0	0	0	0	E	\	\	!	\	\					E	0	0	1	2	2	0	0	0	0	G	\	\	\	!	\	\	\			G	0	0	0	0	1	1	1	0	0	K	\	\	\	\	\	\	\	\		K	0	0	0	0	0	0	0	0	0	H	\	\	\	\	\	\	\	\		H	0	0	0	0	0	0	0	1	0	I	\	\	\	\	\	\	\	\	\	I	0	0	0	0	0	0	0	0	2
A	\	\	\	\	\	\	\	\	\																																																																																																																																																																																																																																																																																																																								
A	1	-1	-1	-1	-1	-1	-1	-1	-1																																																																																																																																																																																																																																																																																																																								
B	\	!	\	\	\	\	\	\	\																																																																																																																																																																																																																																																																																																																								
B	-1	2	0	-2	-2	-2	-2	-2	-2																																																																																																																																																																																																																																																																																																																								
D	\	\	!	\	\	\	\	\	\																																																																																																																																																																																																																																																																																																																								
D	-1	0	3	1	-1	-3	-3	-3	-3																																																																																																																																																																																																																																																																																																																								
E	\	\	!	!	\	\	\	\	\																																																																																																																																																																																																																																																																																																																								
E	-1	-2	1	2	2	0	-2	-4	-4																																																																																																																																																																																																																																																																																																																								
G	\	\	\	!	!	!	\	\	\																																																																																																																																																																																																																																																																																																																								
G	-1	-2	-1	0	1	1	1	-1	-3																																																																																																																																																																																																																																																																																																																								
K	\	\	\	!	!	!	\	\	\																																																																																																																																																																																																																																																																																																																								
K	-1	-2	-3	-2	-1	0	0	0	-2																																																																																																																																																																																																																																																																																																																								
H	\	\	\	\	!	!	!	\	\																																																																																																																																																																																																																																																																																																																								
H	-1	-2	-3	-4	-3	-2	-1	1	-1																																																																																																																																																																																																																																																																																																																								
I	\	\	\	\	\	!	!	!	!																																																																																																																																																																																																																																																																																																																								
I	-1	-2	-3	-4	-5	-4	-3	-1	2																																																																																																																																																																																																																																																																																																																								
A	\																																																																																																																																																																																																																																																																																																																																
A	1	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																																																																								
B	\	\																																																																																																																																																																																																																																																																																																																															
B	0	2	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																																																																								
D	\	!	\	\																																																																																																																																																																																																																																																																																																																													
D	0	0	3	1	0	0	0	0	0																																																																																																																																																																																																																																																																																																																								
E	\	\	!	\	\																																																																																																																																																																																																																																																																																																																												
E	0	0	1	2	2	0	0	0	0																																																																																																																																																																																																																																																																																																																								
G	\	\	\	!	\	\	\																																																																																																																																																																																																																																																																																																																										
G	0	0	0	0	1	1	1	0	0																																																																																																																																																																																																																																																																																																																								
K	\	\	\	\	\	\	\	\																																																																																																																																																																																																																																																																																																																									
K	0	0	0	0	0	0	0	0	0																																																																																																																																																																																																																																																																																																																								
H	\	\	\	\	\	\	\	\																																																																																																																																																																																																																																																																																																																									
H	0	0	0	0	0	0	0	1	0																																																																																																																																																																																																																																																																																																																								
I	\	\	\	\	\	\	\	\	\																																																																																																																																																																																																																																																																																																																								
I	0	0	0	0	0	0	0	0	2																																																																																																																																																																																																																																																																																																																								
<p>Optimum global alignment (score: 2)</p> <p style="margin-left: 20px;">A B D D E F G H I (top)</p> <p style="margin-left: 20px;">A B D - E G K H I (side)</p> <p>or A B - D E G K H I</p>	<p>Optimal local alignment (score 3):</p> <p style="margin-left: 20px;">A B D (top)</p> <p style="margin-left: 20px;">A B D (side)</p>																																																																																																																																																																																																																																																																																																																																

+1 : match
-1 : mis-match
-2 : gap



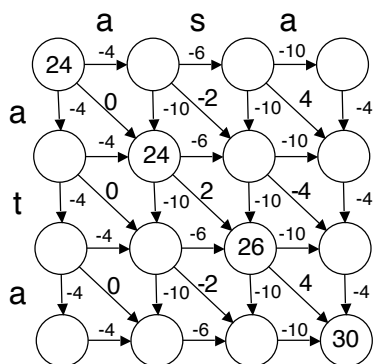
Algorithms for Global and Local Similarity Scores

```

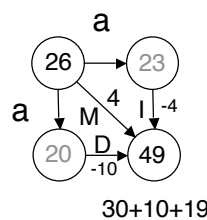
Global:
  S(0,0) ← 0
  for j ← 1 to N do
    S(0,j) ← S(0,j-1) + σ(  $\bar{b}_j$  )
  for i ← 1 to M do
    [ S(i,0) ← S(i-1,0) + σ(  $a_i$  )
      for j ← 1 to N do
        S(i,j) ← max[S(i-1,j-1) + σ(  $\frac{a_i}{b_j}$  ), S(i-1,j) + σ(  $\frac{a_i}{-}$  ), S(i,j-1) + σ(  $\bar{b}_j$  ) ]
      ]
  write "Global similarity score is" S(M,N)

Local:
  best ← 0
  for j ← 1 to N do
    S'(0,j) ← 0
  for i ← 1 to M do
    [ S'(i,0) ← 0
      for j ← 1 to N do
        [ S'(i,j) ← max[0, S'(i-1,j-1) + σ(  $\frac{a_i}{b_j}$  ), S'(i-1,j) + σ(  $\frac{a_i}{-}$  ), S'(i,j-1) + σ(  $\bar{b}_j$  ) ]
          best ← max(S'(i,j), best)
        ]
      ]
  write "Local similarity score is" best
  
```

HMM Alignment

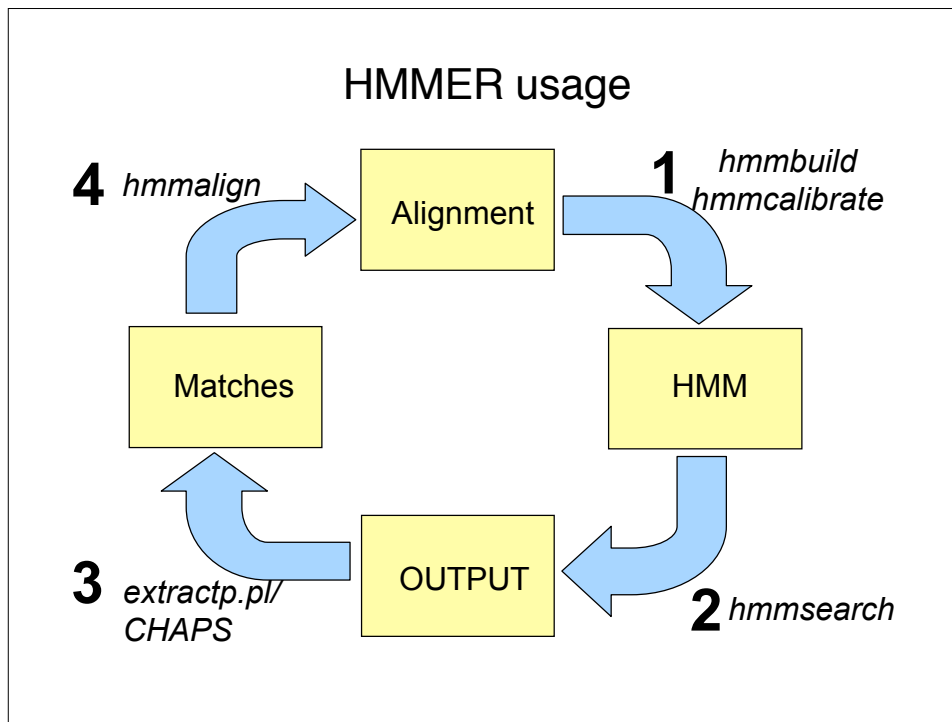


Needleman-Wunsch
max log likelihood
HMM Viterbi alignment



$$F_j^M(i) = \log \frac{e_{M_j}(x_i)}{q_{x_i}} + \log [a_{M_{j-1}M_j} \exp(F_{j-1}^M(i-1)) + a_{I_{j-1}M_j} \exp(F_{j-1}^I(i-1)) + a_{D_{j-1}M_j} \exp(F_{j-1}^D(i-1))] + \log [a_{M_{j-1}M_j} \exp(F_{j-1}^M(i-1)) + a_{I_{j-1}M_j} \exp(F_{j-1}^I(i-1)) + a_{D_{j-1}M_j} \exp(F_{j-1}^D(i-1))]$$

HMM Forward (score)
 \sum probabilities

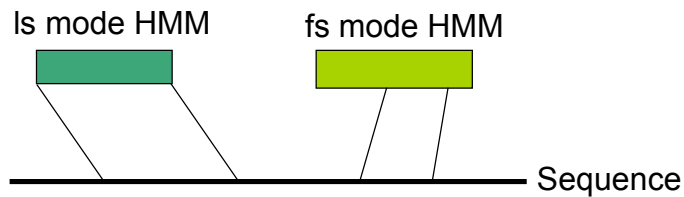


1. Building HMMER models

- Can use HMMER as a black box!
- Usage: *hmmbuild* <hmm> <align>
- Usage: *hmmcalibrate* <hmm>

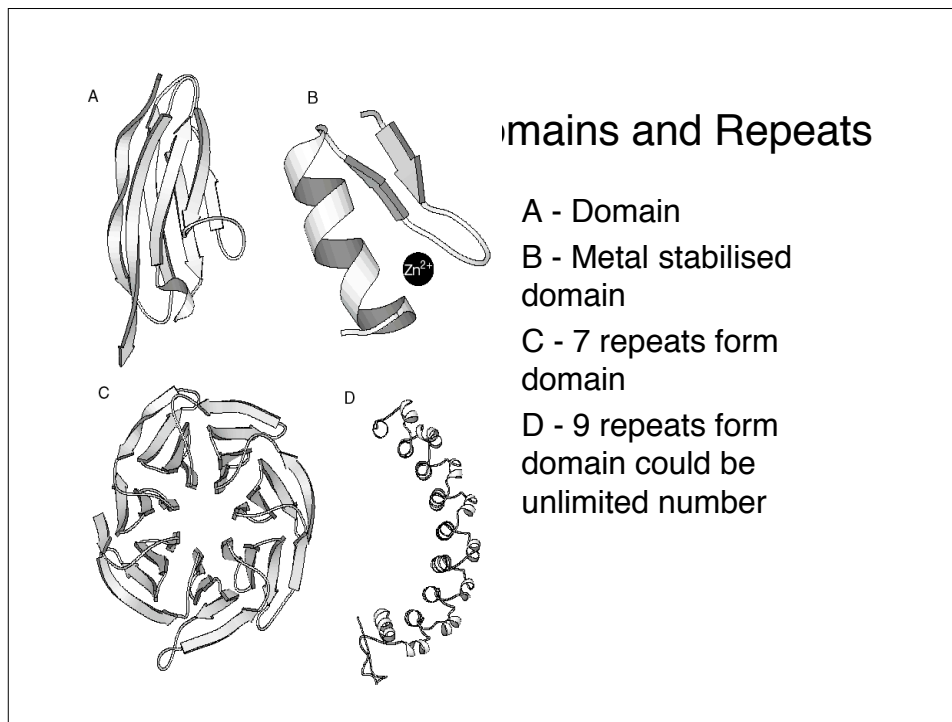
1. Global and Local models

- HMMER provides two main styles of model
- ls - Match whole model within sequence
 - Most sensitive mode. Good for whole domains
 - Default mode
- fs - Match part of model to part of sequence
 - Good when you don't know domain structure



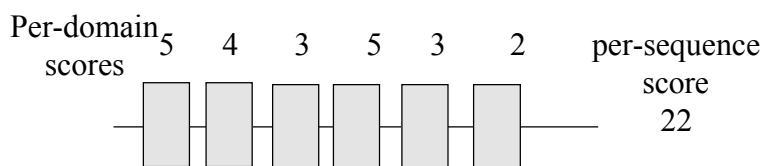
2. HMMER searches

- Search protein database
- Usage: `hmmsearch <hmm> <database>`
 - Option `-A` Use to limit size of output
 - Option `-E` Raise to get more distant matches
- Will take > 20 minutes
- Specialized hardware exists (TimeLogic, Paracel, Compugen)



3. Choosing a threshold

- HMMER has two types of threshold:
 - per-sequence
 - per-domain



3. HMMER output

Header information

```
hmmsearch - search a sequence database with a profile HMM
HMMER 2.1.2 (May 1999)
Copyright (C) 1992-1999 Washington University School of Medicine
HMMER is freely distributed under the GNU General Public License
(GPL).
```

```
-----
-
HMM file:                /nfs/somewhere/CBS/HMM [SEED]
Sequence database:       /nfs/somewhere/pfamseq
per-sequence score cutoff: [none]
per-domain score cutoff:  [none]
per-sequence Eval cutoff: 1e+03
per-domain Eval cutoff:   [none]
-----
```

```
Query HMM: SEED||
[HMM has been calibrated; E-values are empirical estimates]
```

3. HMMER output

Per-sequence scores

```
Scores for complete sequences (score includes all domains):
Sequence  Description                               Score  E-value  N
-----  -
YC25_METJA Q58622 HYPOTHETICAL PROTEIN MJ1225.      210.0  2.3e-58  4
O27291    O27291 INOSINE-5'-MONOPHOSPHATE DEHYDROGEN  198.1   9e-55  4
O27292    O27292 INOSINE-5'-MONOPHOSPHATE DEHYDROGEN  192.6  3.9e-53  4
Q9YDY4    Q9YDY4 HYPOTHETICAL 70.0 KDA PROTEIN APE07    185.5  5.5e-51  7
O29410    O29410 CONSERVED HYPOTHETICAL PROTEIN.       184.7  9.9e-51  4
YE04_METJA Q58799 HYPOTHETICAL PROTEIN MJ1404.             180.1  2.4e-49  4
AAKG_HUMAN P54619 5'-AMP-ACTIVATED PROTEIN KINASE, GA    179.9  2.6e-49  4
YR33_THEPE P15889 HYPOTHETICAL 33.4 KDA PROTEIN IN RI     179.0  5.1e-49  4
AAKG_RAT   P80385 5'-AMP-ACTIVATED PROTEIN KINASE, GA    173.2  2.7e-47  4
Q9V1T3    Q9V1T3 HYPOTHETICAL 32.1 KDA PROTEIN.          170.2  2.2e-46  4
Q9YFL7    Q9YFL7 HYPOTHETICAL 31.7 KDA PROTEIN APE02    168.0  9.9e-46  4
```

3. HMMER output

Per-domain scores

Parsed for domains:

Sequence	Domain	seq-f	seq-t	hmm-f	hmm-t	score	E-value
O26740	2/2	103	156 ..	1	54 []	75.0	1e-17
O26740	1/2	6	59 ..	1	54 []	72.3	6.7e-17
Q9X0M4	2/2	79	132 ..	1	54 []	69.3	5.4e-16
YE26_METJA	2/2	115	168 .]	1	54 []	69.0	6.5e-16
Q9X175	2/2	83	136 ..	1	54 []	68.9	6.8e-16
Q9UY49	2/2	154	207 ..	1	54 []	68.9	6.9e-16
IMDH_PYRFU	2/2	154	207 ..	1	54 []	68.8	7.6e-16
P74477	2/2	518	571 ..	1	54 []	68.7	8.1e-16
IMDH_METKA	2/2	74	127 ..	1	54 []	67.7	1.6e-15
IMDH_PYRHO	2/2	154	207 ..	1	54 []	67.7	1.6e-15
YC32_METJA	2/2	234	287 ..	1	54 []	66.4	3.9e-15
O58317	3/4	134	187 ..	1	54 []	66.3	4.2e-15
O29411	2/2	74	127 ..	1	54 []	66.1	4.7e-15
IMDH_AQUAE	1/2	94	147 ..	1	54 []	65.7	6.5e-15
ACUB_BACSU	1/2	5	58 ..	1	54 []	65.4	8e-15
O29915	2/2	294	346 ..	1	54 []	65.3	8.4e-15
Q9UYR4	3/4	134	187 ..	1	54 []	65.1	9.8e-15

4. Making an alignment

- HMMER can be used to align all the matches
- Usage: *hmmalign* <hmm> <sequences>
- Much faster compared to other methods

Profile-HMM alignments

- It is hard to manually align a large number of sequences (100-10000)
- Solution (The Pfam way):
 - Build a good quality manual alignment of representative members
 - Build profile-HMM
 - Align all members automatically

Pitfalls of HMMER

- More complex to run than PSI-BLAST, you must iterate yourself
- Slow compared to PSI-BLAST

Pitfalls of all profile methods

- **Iterative scoring schemes**
 - Once a false positive is included all its friends come too!
- **Limit to modeling**
 - For large families some members will be missed
- **Profile wander**
 - Matches from earlier rounds can be lost

Multiple Sequence/Profile/HMMs

- Identify larger fraction of protein family with “1” sequence (model)
- Position-specific gap penalties – better alignments
- Rigorous statistical/mathematical model
- fast/automatic (in PSI-Blast)
- Many parameters to estimate (100+ sequences preferred)
- Poor multiple alignments
- Accurate statistical estimates?
- Misled by non-homologs

Conclusions

- Profile-HMMs formalise profile technology
- There are four steps in HMMER use
 - build model
 - search database
 - choose threshold
 - build alignment

Learning more – HMMs

HMMER

<http://hmmmer.wustl.edu/>

Includes links to documentation and other software packages.

The HMMER User's Guide (and tutorial)

<http://hmmmer.wustl.edu/hmmmer-html/>

"Profile hidden Markov models"

S.R. Eddy, *Bioinformatics* 14:755-763, 1998.

"Multiple-alignment and -sequence searches"

S.R. Eddy, *Trends Guide to Bioinformatics*, pp.15-18, 1998.

<http://www.genetics.wustl.edu/eddy/publications/tigs-9808/>

Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids Durbin, Eddy, Krogh, and Mitchison;
Cambridge Univ. Press, 1998.