

# HMMER

## Laboratory of Bioinformatics I Module 2

Emidio Capriotti

<http://biofold.org/>



**Biomolecules  
Folding and  
Disease**

Department of Pharmacy,  
and Biotechnology (FaBiT)  
University of Bologna



# HMMER

[DOWNLOAD](#)[DOCUMENTATION](#)[SEARCH](#)[PUBLICATIONS](#)[BLOG](#)

## HMMER: biosequence analysis using profile hidden Markov models

Get the latest version

**v3.1b2**

[Download \( MacOSX / Intel \)](#)

[Alternative Download Options](#)

HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs).

HMMER is often used together with a profile database, such as [Pfam](#) or many of the databases that participate in [Interpro](#). But HMMER can also work with query *sequences*, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with [phmmer](#), or do an iterative search with [jackhmmmer](#).

HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models. In the past, this strength came at significant computational expense, but as of the new HMMER3 project, HMMER is now essentially as fast as BLAST.

HMMER can be downloaded and installed as a command line tool on your own hardware, and now it is also more widely accessible to the scientific community via [new search servers](#) at the European Bioinformatics Institute.

### PERFORM A SEARCH

An online interactive [search](#) service is available at the European Bioinformatics Institute. Go there to [search](#) against the latest Uniprot databases.

### DOCUMENTATION

The HMMER User's Guide: [\[PDF, 119 pages\]](#).  
[Release notes](#) for the current release.

### NEWS

See the blog [Cryptogenomicon](#) for more information and discussion about HMMER3.

Eddy SR (1998) *Profile hidden Markov models*. **Bioinformatics** 14:755-763  
Eddy SR (2008) *A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation*. **PLoS Comp. Biol.** 4: e1000069  
Eddy SR (2011) *Accelerated profile HMM searches*. **PLoS Comp. Biol.** 7:e1002195

<http://hmmer.org>

# Why HMMER?

HMMER is an Hidden Markov Model based tool used for

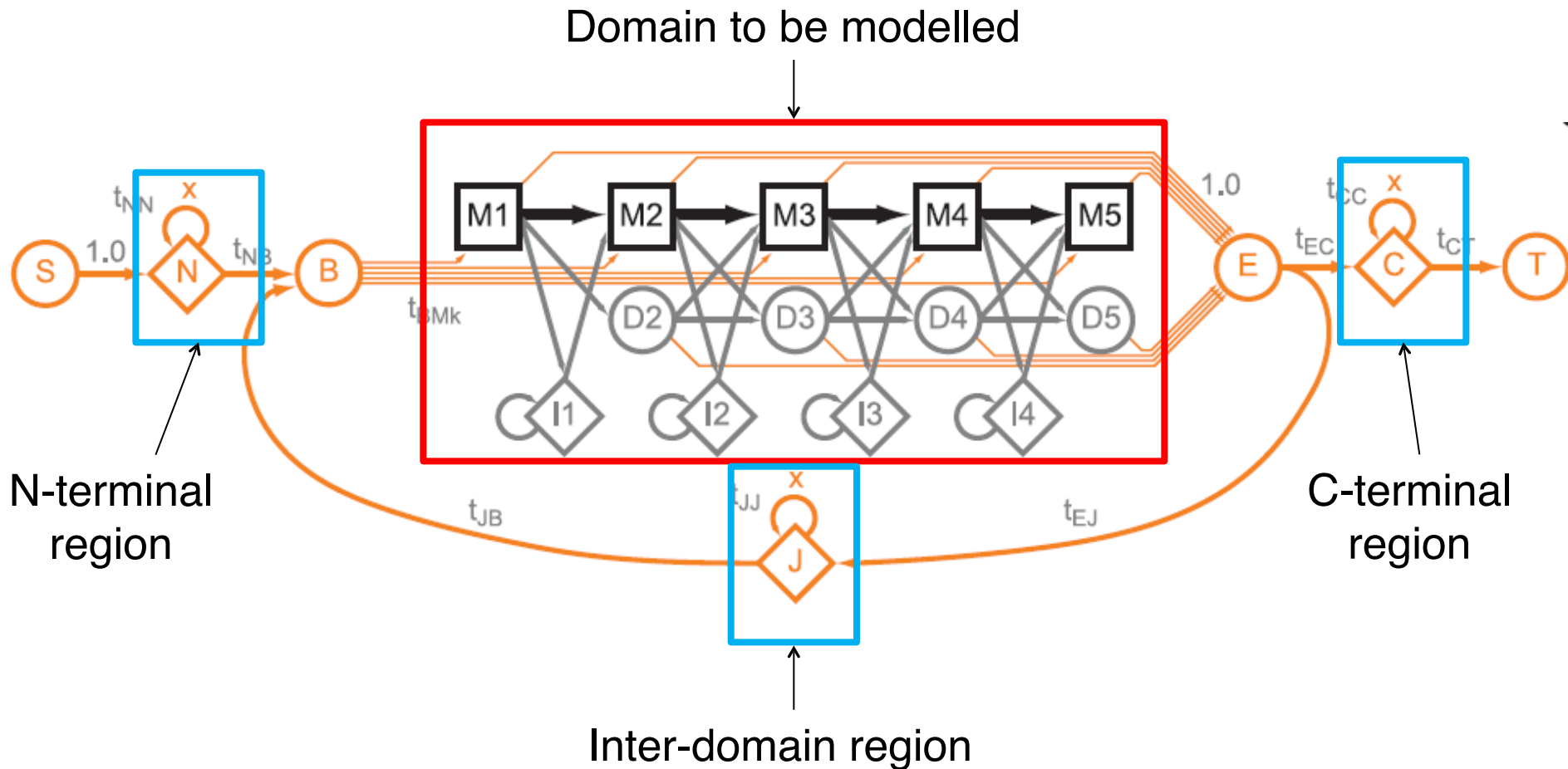
- searching sequence databases for sequence homologs
- make sequence alignments

HMMER is designed to detect remote homologs relying on the strength of its underlying probability models.

The new version of HMMER (HMMER3) is as fast as BLAST.

# HMMER: General Model

The domain model has multiple hits and for each hit insertion and deletion

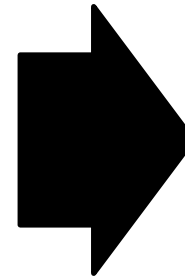


# Aligning a protein family

Takes the aligned sequences, checks for redundancy and sets the emission and the transitions probabilities of a HMM, setting the parameters for the Extreme value distribution

```
10 20 30 40 50 60 70 80
lqb1 pea/1-471 -GPTDKQKQALVNSSE--FKQNLGGYSILFYTIILEKAPAAKGLFSFKI---TAGVDESEPKLQAHAEVYGLVRDSSAQL
lqb1 vicfa/1-471 -GPTDKQKQALVNSSSQLFKQNTSMNSVLFYTIILQKAPTAKAMFSFKI---SAGVVDSEPKLGAHAEKVYGMVRDSSAQL
hbb speci/1-471 VHLSDGSENAISTAMGKV--HAAEVGAELGRLLVVYEWQRFPSDFGDLSSASAVMGNAKVNAHGKKVIDSFSNGLKHL
hbb speto/1-471 VHLSDGSENAISTAMGKV--HAAEVGAELGRLLVVYEWQRFPSDFGDLSSASAVMGNAKVNAHGKKVIDSFSNGLKHL
hbb equhe/1-471 VQLSGEEAAVLAALWDKV--NEEEVGGEALGRLLVVYEWQRFPSDFGDLSSAAVVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb sunnu/1-471 VHLSGEEAACVTGLMGKV--NEDEVGAELGRLLVVYEWQRFPSDFGDLSSASAVMGNAKVNAHGKKVLSHSPGEGVHHL
hbb tupql/1-471 VHLSGEEAAVTLGLMGKV--DLEKVGGOGLGSLIVYEWQRFPSDFGDLSSASAVMGNAKVNAHGKKVLSHSPGEGVHHL
hbb calar/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb mansp/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb rabbit/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb ursma/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb triin/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb ornan/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb tacac/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbe ponpy/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb colli/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb larri/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb1 varex/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb2 xentr/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb1 ranca/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hbb2 tracr/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
hba4 salir/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
myg_escgi/1-471 VHLTGEESASVTALMGKV--NVDEVGGEALGRLLVVYEWQRFPSDFGDLSSADAVMGNPKVNAHGKKVLSHSPGEGVHHL
```

*hmmbuild*



Trained profile-HMM

HMM calibrated with the  
accurate E-value  
statistics

# Scoring the Sequence

After the training, the model  $M$  associates to a sequence  $s$  the probability  $P(s | M)$

This probability answers the question:

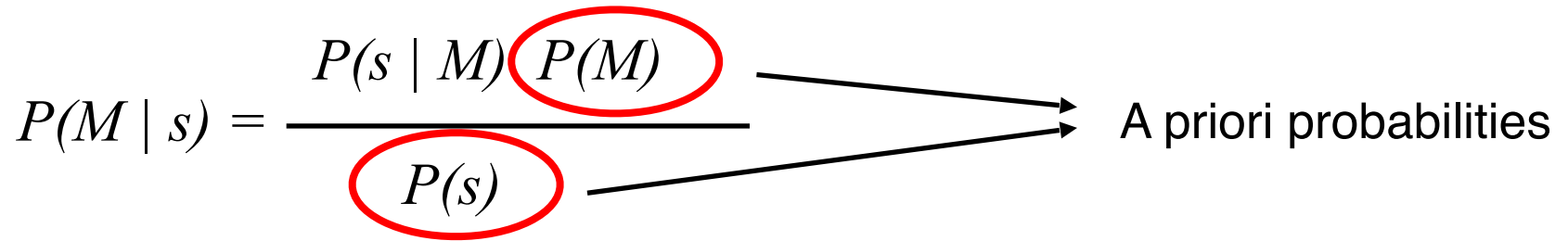
What is the probability for a model  $M$  (e.g. describing the Globins) to generate the sequence  $s$ ?

BUT the question we want to answer to is:

Given a sequence  $s$ , does it belong to the class described by the model  $M$ ? (e.g. is it a Globin?)

We need to compute  $P(M | s)$  !!

# A Priori Probabilities

$$P(M | s) = \frac{P(s | M) P(M)}{P(s)}$$


A priori probabilities

$P(M)$  is the probability of the model (i.e. of the class described by the model) BEFORE we know the sequence:

can be estimated as the abundance of the class

$P(s)$  is the probability of the sequence in the sequence space.

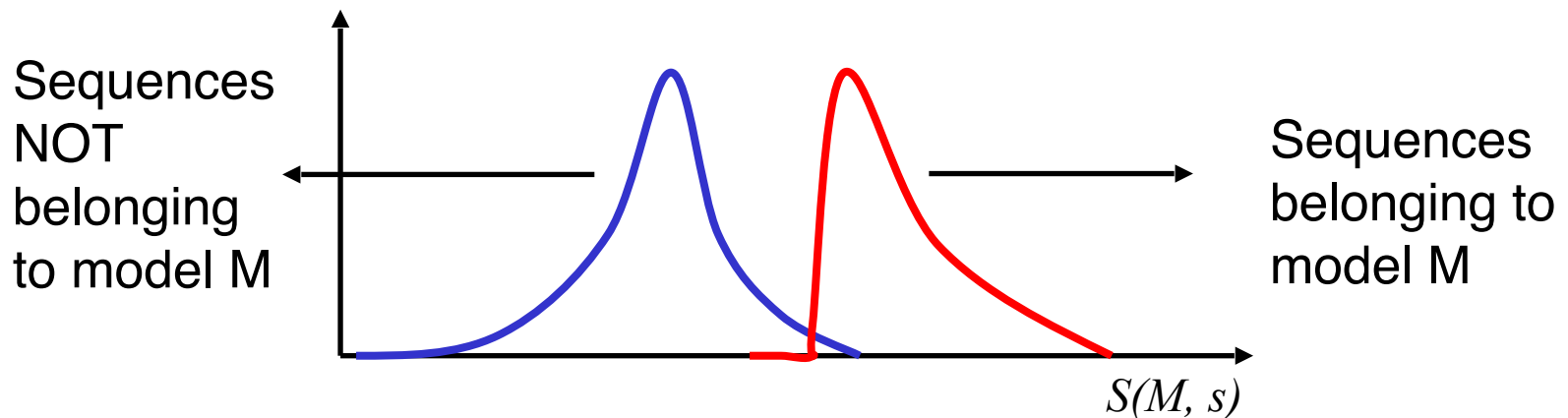
Cannot be reliably estimated!!

# The Null Model

Null Model: a model that generates ALL the possible sequences with probabilities depending ONLY on letter (e.g. residue) statistical abundance (in HMMER3, by default, mean residue frequencies in Swiss-Prot 50.8 (October 2006))

**Log odd score  
(in bits)**

$$S(M, s) = \log_2 \frac{P(s | M)}{P(s | N)}$$

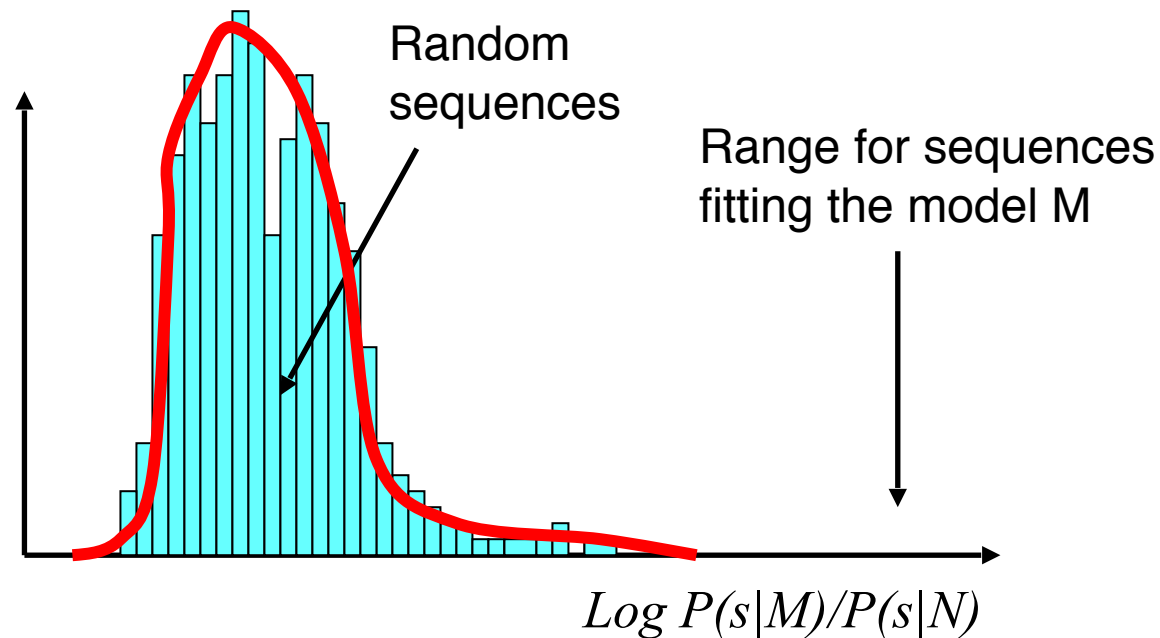


In this case we need a threshold and a statistic for evaluating the significance (E-value, P-value)



# Extreme Value Distribution

Given a trained model  $M$ , a number of  $N$  (default 200 in HMMER3) random sequences are generated and scored with the model.



The random distribution is fitted with a Gumbel distribution, by estimating  $\lambda$  and  $\mu$

$$P(S \geq t) = 1 - \exp[-e^{-\lambda(t-\mu)}]$$

# The E-value

After setting  $\lambda$  and  $\mu$

$$P(S \geq t) = 1 - \exp[-e^{-\lambda(t-\mu)}]$$

gives the probability of finding random matches with score  $> t$ :  
This is by definition the P-value corresponding to the score  $t$

The E-value( $t$ ), namely expected number of random sequences with a score  $> t$ , is obtained with the relation

$$P = 1 - e^{-E}$$

# E-value vs P-value

If  $E$  is the expected (average) number of occurrences of a rare event, we can adopt the Poisson's statistics to estimate the probability of observing  $a$  of such events.

$$p(a) = e^{-E} \frac{E^a}{a!}$$

P-value ( $P$ ) is the probability of observing at least one rare event, that is

$$P = 1 - p(0)$$

$$P = 1 - e^{-E}$$

# HMMER Installation

For Debian/Ubuntu Linux distributions

```
apt-get install hmmer  
apt-get install hmmer-doc
```

Root privileges are needed

User guide <http://eddylib.org/software/hmmer3/3.1b2/Userguide.pdf>

Otherwise:

Download the  
Linux version



The screenshot shows the HMMER website with the following content:

- Header: HMMER logo, DOWNLOAD, DOCUMENTATION, SEARCH, PUBLICATIONS, BLOG
- Section: HMMER: biosequence analysis using profile hidden Markov models
- Text: "Get the latest version" followed by "v3.1b2" and a blue button labeled "Download ( MacOSX / Intel )". Below the button is the text "Alternative Download Options". This section is circled in red.
- Text: "HMMER is used for searching sequence databases for sequence homologs, and for making sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (profile HMMs)."
- Text: "HMMER is often used together with a profile database, such as Pfam or many of the databases that participate in Interpro. But HMMER can also work with query sequences, not just profiles, just like BLAST. For example, you can search a protein query sequence against a database with phmmer, or do an iterative search with jackhmmer."
- Text: "HMMER is designed to detect remote homologs as sensitively as possible, relying on the strength of its underlying probability models. In the past, this strength came at significant computational expense, but as of the new HMMER3 project, HMMER is now essentially as fast as BLAST."
- Text: "HMMER can be downloaded and installed as a command line tool on your own hardware, and now it is also more widely accessible to the scientific community via new search servers at the European Bioinformatics Institute."
- Footer: Three columns: "PERFORM A SEARCH" (An online interactive search service is available at the European Bioinformatics Institute. Go there to search against the latest Uniprot databases.), "DOCUMENTATION" (The HMMER User's Guide: [PDF, 119 pages]. Release notes for the current release.), and "NEWS" (See the blog Cryptogenomicon for more information and discussion about HMMER3.).

<http://hmmer.org>

# Globin Alignment

The multiple sequence alignment is provided in Stockholm format

```
# STOCKHOLM 1.0
```

```
HBB_HUMAN      .....VHLTPEEKSAVTALWGKV....NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKVKAHGKKVL
HBA_HUMAN      .....VLSPADKTNVKAAWGKVGA..HAGEYGAEALERMFSLFPTTKTYFPHF.DLS.....HGSAQVKGHGKKVA
MYG_PHYCA      .....VLSEGEWQLVLHVWAKVEA..DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVL
GLB5_PETMA     PIVDTGSVAPLSAAEKTIRSAPVYS..TYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSADVRWHAERII
```

```
HBB_HUMAN      GAFSDGLAHL...D..NLKGTFFATLSELHCDKL..HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVAGVANAL
HBA_HUMAN      DALTNAVAVH...D..DMPNALSALSDLHAHKL..RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVL
MYG_PHYCA      TALGAILKK...K.GHHEAELKPLAQSHATKH..KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDI
GLB5_PETMA     NAVNDAVASM..DDTEKMSMKLRDLSGKHAKSF..QVDPQYFKVLAAVIADTVAAG.....DAGFEKLMSMICILL
```

```
HBB_HUMAN      AHKYH.....
HBA_HUMAN      TSKYR.....
MYG_PHYCA      AAKYKELGYQG
GLB5_PETMA     RSAY.....
//
```

```
HEADER:      # STOCKHOLM 1.0
END:         //
GAP:         .
```

# Stockholm Format

More information can be added. They can be used to guide the HMM training.

**#=GC SS\_cons**      Secondary structure (consensus)

**#=GC RF**              Reference annotation

Often the consensus RNA or protein sequence is used as a reference

Any non-gap character (e.g. x's) can indicate consensus/conserved/match columns

. or - indicate insert columns

~ indicate unaligned insertions

Upper and lower case can be used to discriminate strong and weakly conserved residues respectively

**#=GC MM**              Model Mask

Indicates which columns in an alignment should be masked, such that the emission probabilities for match states corresponding to those columns will be the background distribution. Masked positions are marked with “m”

# Building the HMM model

## SINTAX:

```
hmmbuild <hmm_file> <msa_file>
```

```
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# -----
# input alignment file:          globins4.sto
# output HMM file:              globins4.hmm
# -----
```

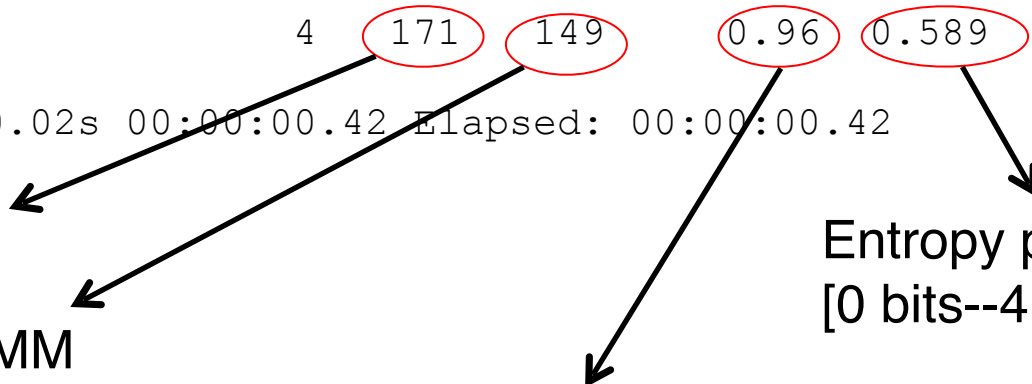
```
# idx name                nseq  alen  mlen  eff_nseq  re/pos  description
#-----
1      globins4            4      171   149    0.96    0.589
# CPU time: 0.40u 0.02s 00:00:00.42 Elapsed: 00:00:00.42
```

Length of the MSA

Length of the HMM  
(number of match states)

Number of “effective” sequences

Entropy per position  
[0 bits--4.32 bits]



# HMM Model Output

HMMER3/f [3.1b2 | February 2015]

## HEADER: General info

NAME globins4

LENG 149

ALPH amino

RF no

MM no

CONS yes

CS no

MAP yes

DATE Sat Mar 8 23:36:44 2014

NSEQ 4

EFFN 0.964844

CKSUM 2027839109

STATS LOCAL MSV -9.9014 0.70957

STATS LOCAL VITERBI -10.7224 0.70957

STATS LOCAL FORWARD -4.1637 0.70957

HMM A C D E F G H ...

m->m m->i m->d i->m i->i d->m d->d

COMPO	2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...	
	2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...	
	0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*		
1	1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	...	9 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...	
	0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405		
2	2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	...	10 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...	
	0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510		

Use information from RF and MM lines in Stockholm

Build a consensus sequence

Use information from SS\_cons line in Stockholm

Map with respect to the alignment file

Statistical parameters needed for  
E-value calculations ( $\mu$ ,  $\lambda$ )

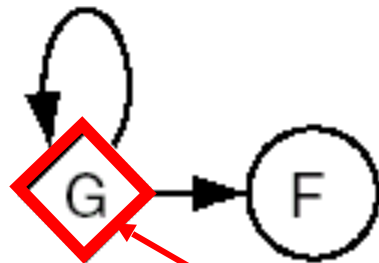


# NULL Model

Score =  $-\ln(p)$  or '\*' if  $p=0$

Where  $p$  = is function of the natural abundance of residues

Swiss-Prot 50.8 (October 2006)

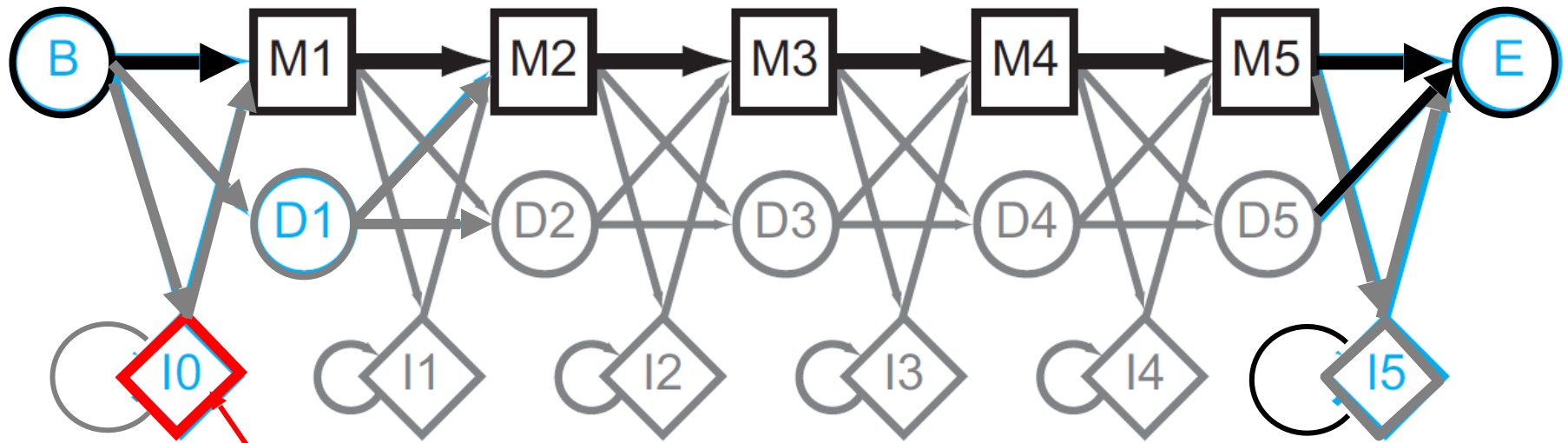


HMM

	A	C	D	E	F	G	H	...
	m->m	m->i	m->d	i->m	i->i	d->m	d->d	
COMPO	2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...
	2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...
	0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*	
1	1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	... 9 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405	
2	2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	... 10 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510	

# 0-states (emissions)

Score =  $-\ln(p)$  or '\*' if  $p=0$

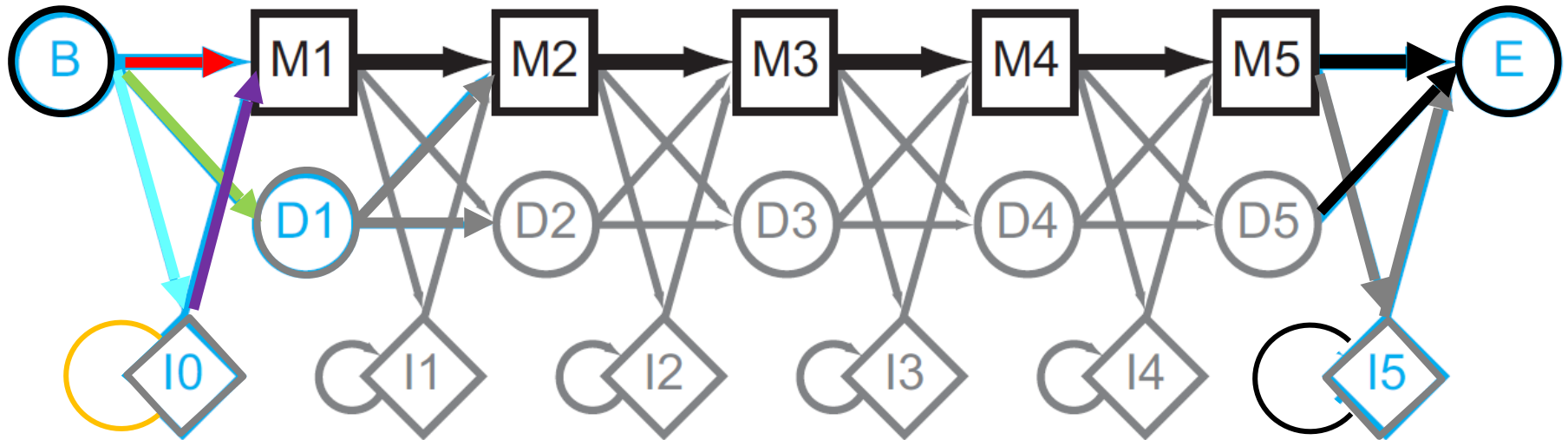


HMM

	A	C	D	E	F	G	H	...
	m->m	m->i	m->d	i->m	i->i	d->m	d->d	
COMPO	2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...
	2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...
	0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*	
1	1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	... 9 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405	
2	2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	... 10 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510	

# 0-states (transitions)

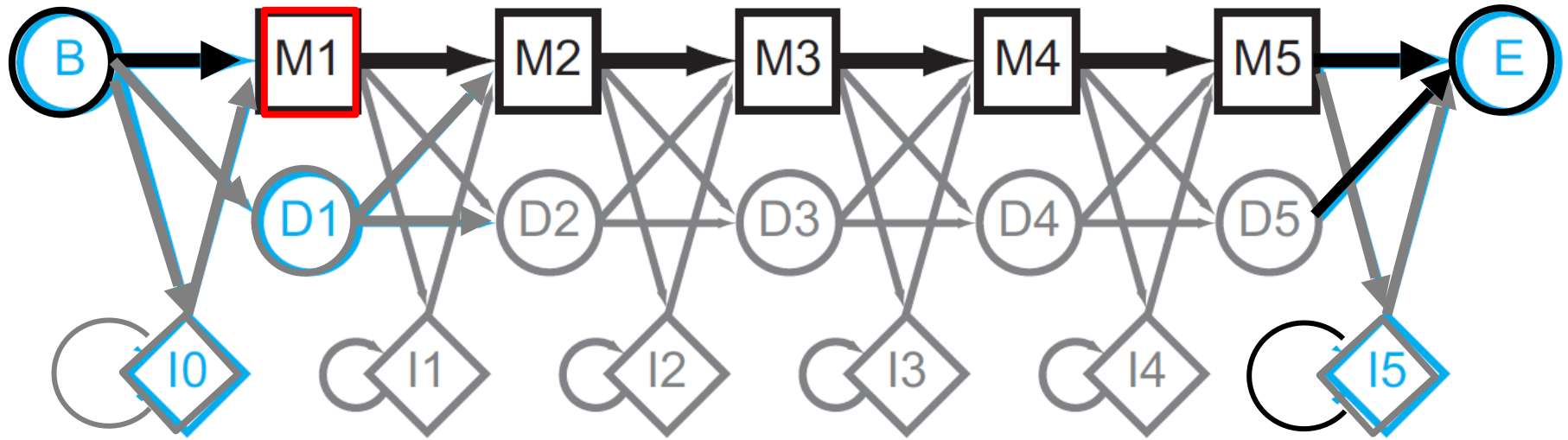
Score =  $-\ln(p)$  or '\*' if  $p=0$



HMM	A	C	D	E	F	G	H	...
	m->m	m->i	m->d	i->m	i->i	d->m	d->d	
COMPO	2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...
	2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...
	0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*	
1	1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	... 9 v - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405	
2	2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	... 10 v - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510	

# 1-states (emissions)

Score =  $-\ln(p)$  or '\*' if  $p=0$



Corresponding position in initial MSA

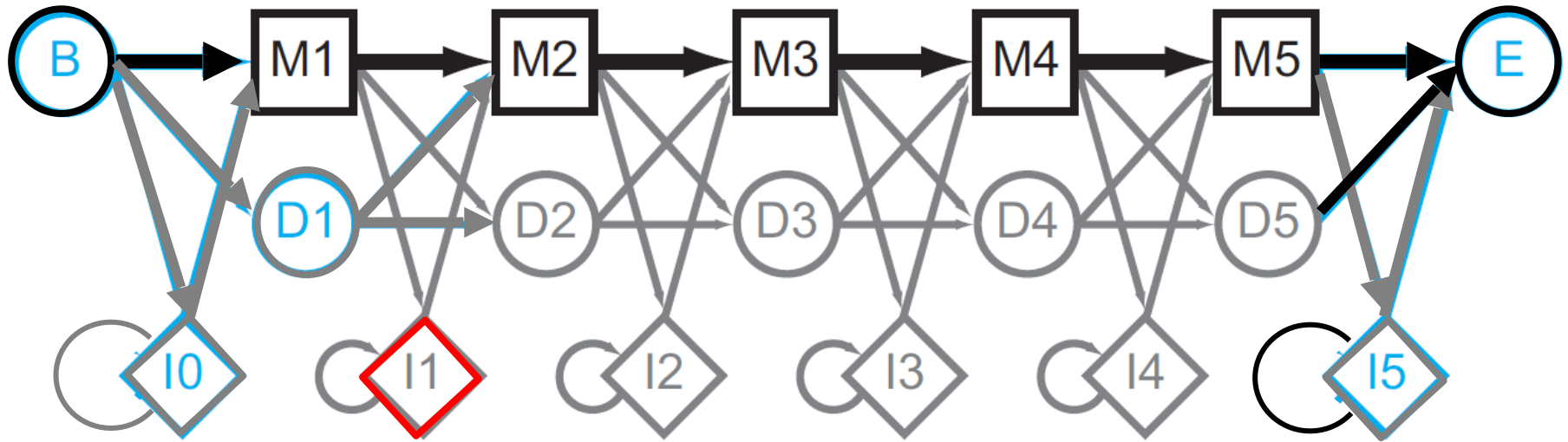
Consensus residue (lowest score)

HMM	A	C	D	E	F	G	H	...	
	m->m	m->i	m->d	i->m	i->i	d->m	d->d		
COMPO	2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...	
	2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...	
	0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*		
1	1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	...	9 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...	
	0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405		
2	2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	...	10 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...	
	0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510		

Column annotation for RF, MM, SS\_cons (if present)

# States of layer 1 (emissions)

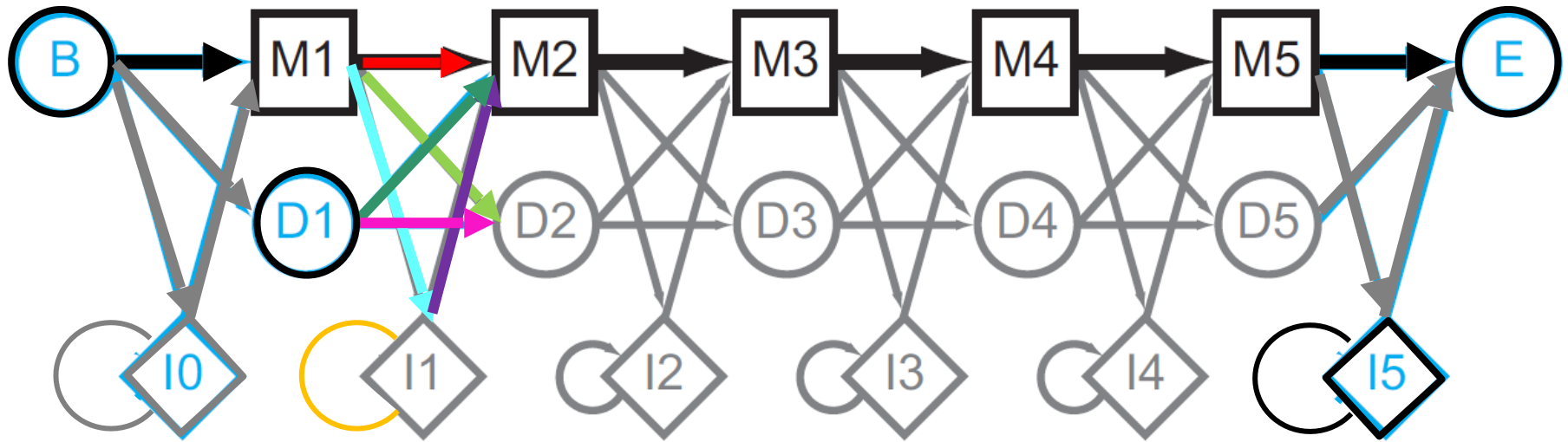
Score =  $-\ln(p)$  or '\*' if  $p=0$



HMM	A	C	D	E	F	G	H	...
	m->m	m->i	m->d	i->m	i->i	d->m	d->d	
COMPO	2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...
	2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...
	0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*	
1	1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	... 9 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405	
2	2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	... 10 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510	

# States of layer 1 (transitions)

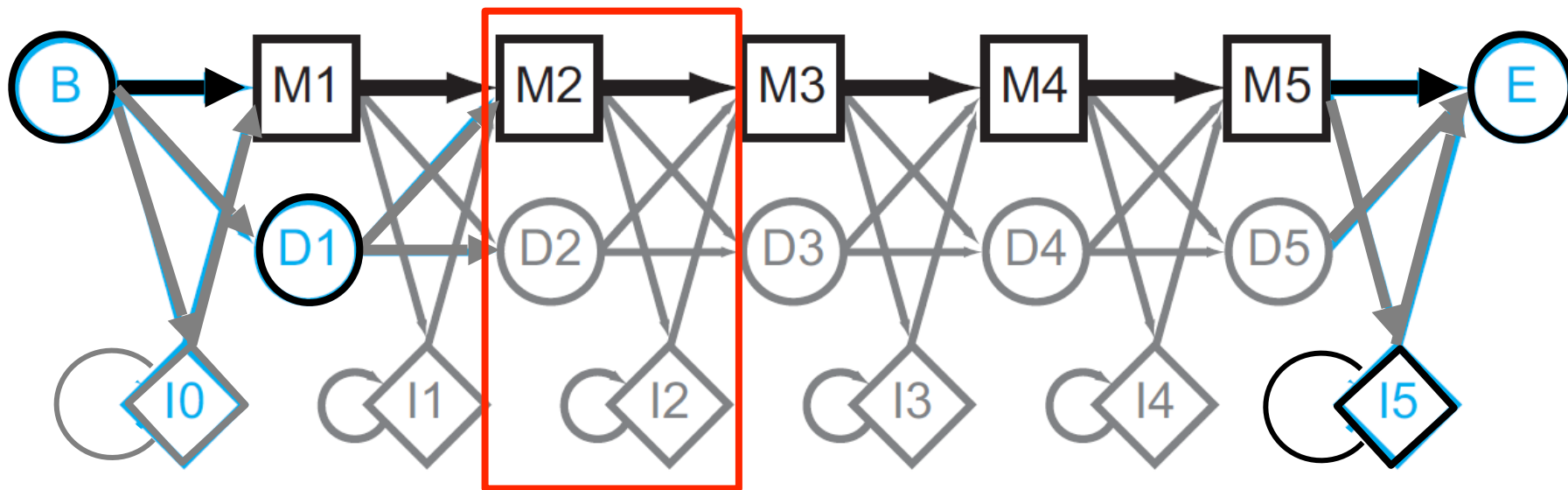
Score =  $-\ln(p)$  or '\*' if  $p=0$



HMM	A	C	D	E	F	G	H	...
	m->m	m->i	m->d	i->m	i->i	d->m	d->d	
COMPO	2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...
	2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...
	0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*	...
1	1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	... 9 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405	...
2	2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	... 10 v - - -
	2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...
	0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510	...

# States of Layer 2

Score =  $-\ln(p)$  or '\*' if  $p=0$



HMM

		A	C	D	E	F	G	H	...	
		m->m	m->i	m->d	i->m	i->i	d->m	d->d		
COMPO		2.36553	4.52577	2.96709	2.70473	3.20818	3.02239	3.41069	...	
		2.68640	4.42247	2.77497	2.73145	3.46376	2.40504	3.72516	...	
		0.57544	1.78073	1.31293	1.75577	0.18968	0.00000	*		
1		1.70038	4.17733	3.76164	3.36686	3.72281	3.29583	4.27570	...	9 v - - -
		2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...	
		0.03156	3.86736	4.58970	0.61958	0.77255	0.34406	1.23405		
2		2.62748	4.47174	3.31917	2.82619	3.63815	3.49607	2.75382	...	10 v - - -
		2.68618	4.42225	2.77519	2.73123	3.46354	2.40513	3.72494	...	
		0.02321	4.17053	4.89288	0.61958	0.77255	0.48576	0.95510		

# Skyalign

## Interactive logos for alignments and profile HMMs

Skyalign is a tool for creating logos representing both sequence alignments and profile hidden Markov models. Submit to the form on the right in order to produce (i) interactive logos for inclusion in webpages, or (ii) static logos for use in documents.

[See an example](#)

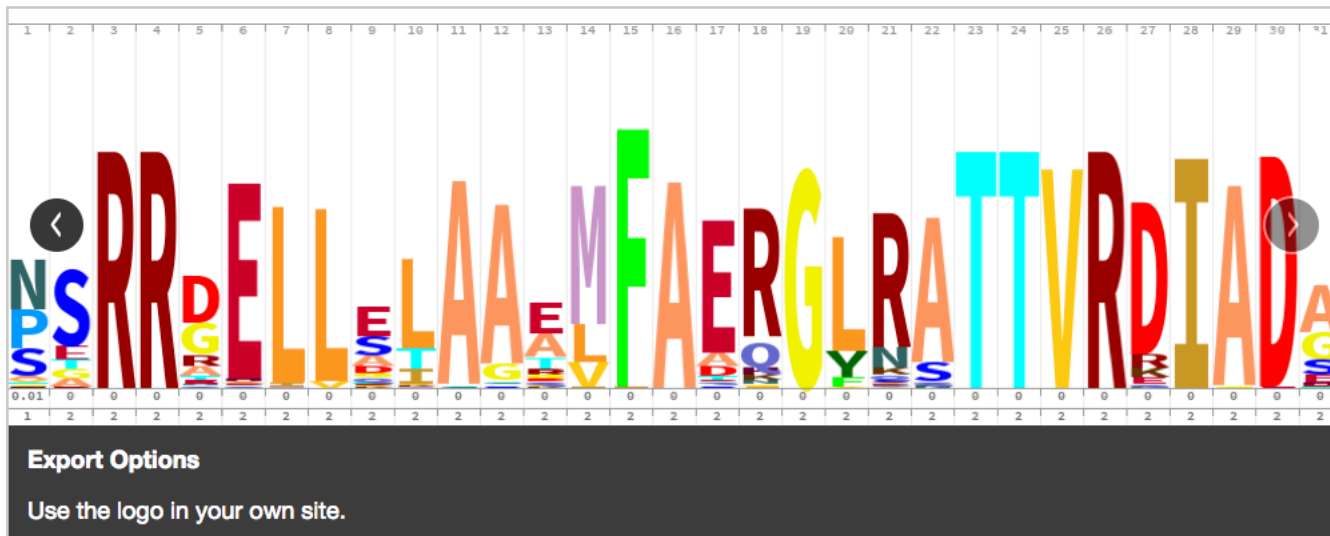
## Create your logo

Upload an HMM or Multiple sequence alignment ?

No file chosen

### Letter Height

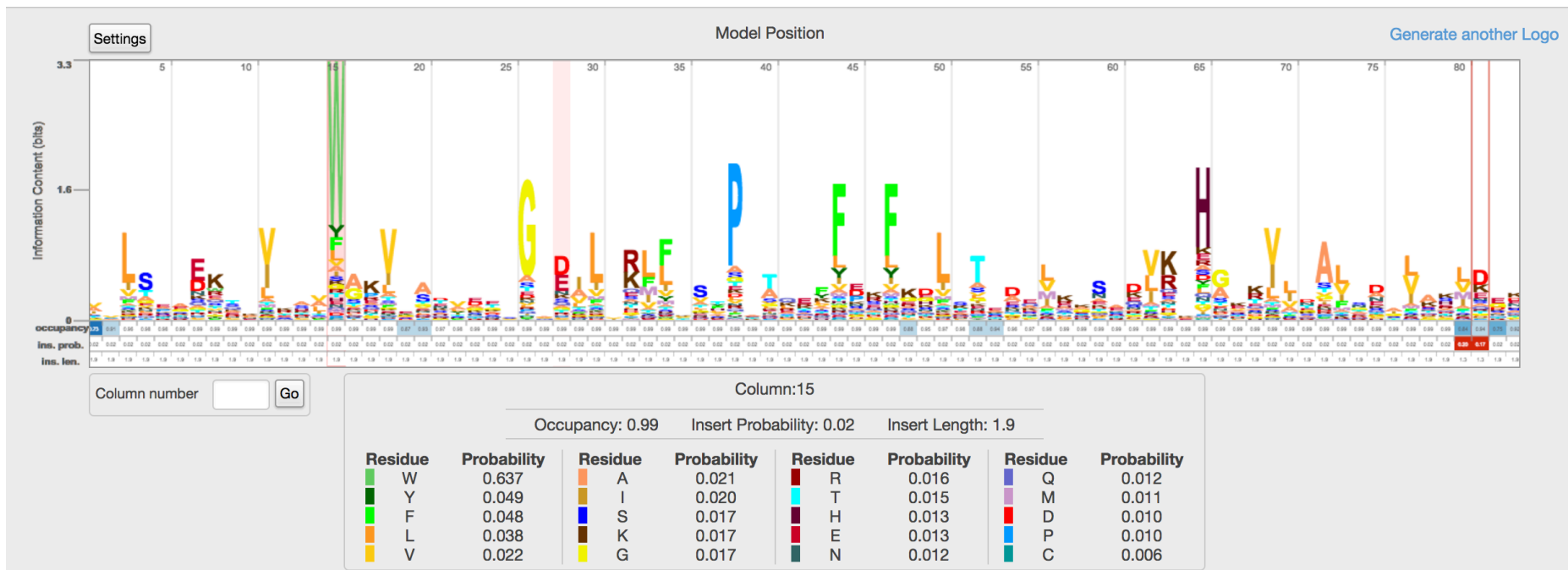
- ☒ Information Content - All ?
- ☐ Information Content - Above Background ?
- ☐ Score ?





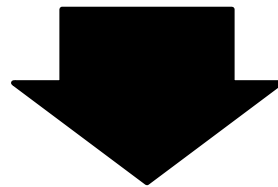
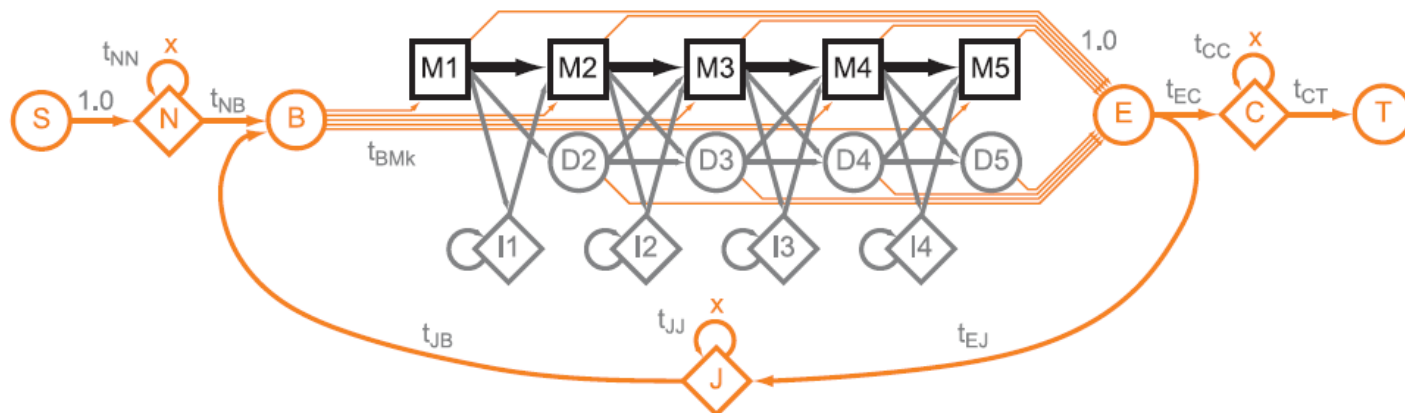
# Skyline Output

Can be computed with the `hmmlogo` command



# Generate sample sequence

Trained HMM



*hmmemit*

Generate (sample) sequences  
from a profile HMM

# Generate Sequences

hmmemit can be used to generate sequence using the Null model

## SINTAX:

**hmmemit** [-N <num>] <hmm\_file>

**hmmemit -N 2 globins4.hmm**

>globins4-sample1

EIPLMDLTEMESIWSGVNAAYKQVGKEEIVMMLQSLPTTVETFEKFHGNVSLDTEYKYRE  
EYTKHAKTLLGAMLAASLSLKQHTENLDHLSKQLAAKVSIGPRPPRLCQRAAVTVLKAKF  
PKNYTKHAMASSKKAMSDQEDLLDGKYK

>globins4-sample2

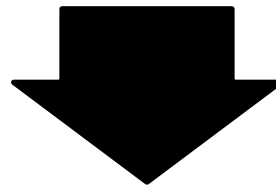
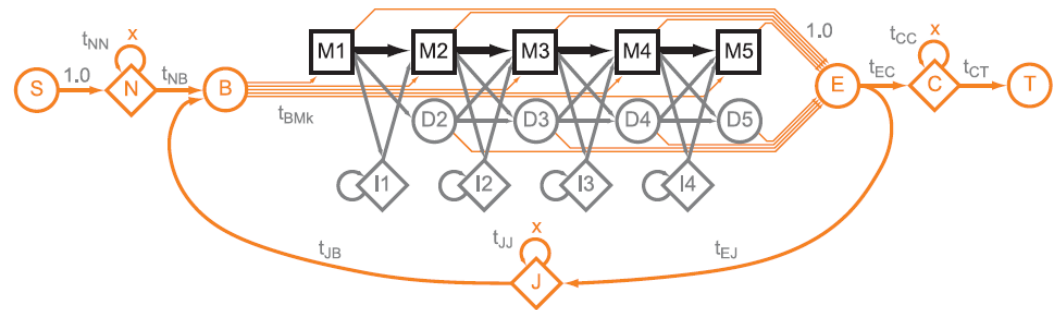
SHIEINPLEAVADLYTTLVIESQYDTPRIQSLHSLEWKKPAACYYRRNFDSDVTTTNM  
MRVSASLRKMTMRVINAFITISATRDNHVQRIIPNAEDHSHKKSNAIDFKAIGVLPEISL  
KMOVPCRHPQDMGNEIHSIEEGLKEGGESADVRY

# Search Matching Sequences

Set of sequences



Trained HMM



*hmmsearch*

List of sequences that match  
the HMM (sorted by E-value)

# Search Sequence

## SINTAX:

**hmmsearch** <hmmfile> <seqdb>

Reads a set of sequences and finds occurrences of the modelled domain. Accepted formats include fasta, embl, genbank, ddbj, uniprot, stockholm, pfam, a2m, and afa.

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b1 (May 2013); http://hmmer.org/
# Copyright (C) 2013 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - -
# query HMM file:                globins4.hmm
# target sequence database:      tutorial/globins45.fa
# output directed to file:       search1.txt
# - - - - -
```

Query: globins4 [M=149]

Scores for complete sequences (score includes all domains):

--- full sequence ---			--- best 1 domain ---			-#dom-		
E-value	score	bias	E-value	score	bias	exp	N	Sequence
-----	-----	-----	-----	-----	-----	----	--	-----
8.7e-67	215.6	2.9	9.7e-67	215.4	2.9	1.0	1	MYG_ESCGI
1.1e-65	211.9	0.1	1.3e-65	211.8	0.1	1.0	1	HBB_MANSP

# Model Line

Domain annotation for each sequence (and alignments):

```
>> MYG_ESCGI
```

#	score	bias	c-Evalue	i-Evalue	hmmfrom	hmm to	alifrom	ali to	envfrom	env to	acc
---	-----	----	-----	-----	-----	-----	-----	-----	-----	-----	---
1 !	215.4	2.9	9.7e-67	9.7e-67	2	149 .]	1	147 [.	1	147 [.	.99

Alignments for each domain:

== domain 1 score: 215.4 bits; conditional E-value: 9.7e-67

```
globins4 2 vLseaektkvkavWakveadvvesGadiLvrlfkstPatqefFekFkdLstedelkksadvkk
          vLs+ae++ v+++Wakveadv+++G+diL+rlfk +P+t+e+F+kFk+L+te+e+k+s+d+kk
MYG_ESCGI 1 VLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHPETLEKFDKFKHLKTEAEMKASEDLKK
          69*****
```

```
globins4 102 dpkyfkl1sevlvdvlaarlpkeftadvqaaleKllalvakllaskYk 149
          ++ky++++s++++v1++r+p++f+ad+qaa++K+l+l++k++a+kYk
MYG_ESCGI 100 PIKYLEFISDAIIHVLHSRHPGDFGADAQAAMNKALELFRKDIAAKYK 147
          *****7 PP
```

IN MODEL LINE: Capital letters represent the most conserved (high information content) positions. Dots (.) in this line indicate insertions in the target sequence with respect to the model.

MIDLINE indicates matches between the query model and target sequence. + indicates positive score (“conservative substitution”, with respect to what the model expects at that position).

BOTTOM LINE represents the posterior probability of each aligned residue. 0: 0-5%, 1: 5-15%, .. 9: 85-95%, \*: 95-100% posterior probability. You can use these posterior probabilities to decide which parts of the alignment are well determined or not.

# Searching Globin Sequence

```
hmmsearch -o search1.txt globins4.hmm tutorial/globins45.fa
```

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.1b1 (May 2013); http://hmmer.org/
# Copyright (C) 2013 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - -
# query HMM file:                globins4.hmm
# target sequence database:      tutorial/globins45.fa
# output directed to file:      search1.txt
# - - - - -
```

Query: globins4 [M=149]

Scores for complete sequences (score includes all domains):

--- full sequence ---			--- best 1 domain ---			-#dom-		
E-value	score	bias	E-value	score	bias	exp	N	Sequence
-----	-----	-----	-----	-----	-----	----	--	-----
8.7e-67	215.6	2.9	9.7e-67	215.4	2.9	1.0	1	MYG_ESCGI
1.1e-65	211.9	0.1	1.3e-65	211.8	0.1	1.0	1	HBB_MANSP

Score is in bits.

Bias is a correction: pay attention when it is on the same order of magnitude of the score (biased compositions/repetitive seq)

# E-values

- if both E-values are significant ( $\ll 1$ ), the **sequence is likely to be homologous to your query**.
- if the **full sequence E-value is significant** but the **single best domain E-value is not**, the target sequence is probably a **multidomain remote homolog**: it contains multiple weakly-scoring domains, even if no single domain is solidly significant on its own; but we need to check if it's just a repetitive sequence.



# The Alignment

Domain annotation for each sequence (and alignments):

```
>> MYG_ESCGI
#      score  bias  c-Evalue  i-Evalue  hmmfrom  hmm to    alifrom  ali to    envfrom  env to    acc
---  -
1 !    215.4    2.9    9.7e-67    9.7e-67         2    149 .]         1    147 [.         1    147 [. 0.99
```

Alignments for each domain:

```
== domain 1 score: 215.4 bits; conditional E-value: 9.7e-67
globins4  2 vLseaektkvkavWakveadvvesGadiLvrlfkstPatqefFekFkdLstedelkksadvkkHgkkvldAlsdalakldekleaklkdLselHakklkv 101
          vLs+ae++ v+++Wakveadv+++G+diL+rlfk +P+t+e+F+kFk+L+te+e+k+s+d+kkHg++vl+Al+ +l+k ++++ea+l+k+L+++Ha+k+k+
MYG_ESCGI 1 VLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGHFPETLEKFDKFKHLKTEAEMKASEDLKKHGNTVLTALGGILKK-KGHHEAELKPLAQSHATKHKI 99
          69*****.99***** PP

globins4 102 dpkyfkllesevlvdvlaarlpkeftadvqaaleKllalvakllaskYk 149
          ++ky++++s+++++vl++r+p++f+ad+qaa++K+l+l++k++a+kYk
MYG_ESCGI 100 PIKYLEFISDAIIHVLSRHPGDFGADAQAAMNKALELFRKDIAAKYK 147
          *****7 PP
```

In a match column, residues are upper case, and a '-' character means a deletion relative to the consensus.

In an insert column, residues are lower case, and a '.' is padding.

Insertions in a profile HMM are unaligned

BOTTOM LINE represents the posterior probability of each aligned residue. 0: 0-5%, 1: 5-15%, .. 9: 85-95%, \*: 95-100% posterior probability.

# Match Scores

Domain annotation for each sequence (and alignments):

```
>> MYG_ESCGI
```

#	score	bias	c-Evalue	i-Evalue	hmmfrom	hmm to	alifrom	ali to	envfrom	env to	acc
1 !	215.4	2.9	9.7e-67	9.7e-67	2	149 .]	1	147 [.	1	147 [.	0.99

!: pass both per-domain and per-sequence E-value thresholds (0.001).

?: pass only one E-value threshold

c-Evalue: conditional E-value: statistical significance of the domain given that we know that the sequence is a true homolog

i-Evalue: independent E-value: statistical significance of the best domain identified in the sequence.

Then the portions of the aligned HMM and the sequence are provided

Envelope is the best aligned sequence portion

Acc: mean per residue alignment a-posteriori probability

# Statistics Summary

Internal pipeline statistics summary:

```
-----  
Query model(s):                1  (149 nodes)  
Target sequences:              45  (6519 residues searched)  
Passed MSV filter:             45  (1); expected 0.9 (0.02)  
Passed bias filter:            45  (1); expected 0.9 (0.02)  
Passed Vit filter:             45  (1); expected 0.0 (0.001)  
Passed Fwd filter:             45  (1); expected 0.0 (1e-05)  
Initial search space (Z):      45  [actual number of targets]  
Domain search space (domZ):    45  [number of targets  
                                   reported over threshold]  
# CPU time: 0.02u 0.01s 00:00:00.03 Elapsed: 00:00:00.03  
# Mc/sec: 32.38  
//
```

MSV: Multi-Segment Viterbi filter: sort of «local» BLAST-like alignments  
(heuristic)

Expected counts must be much lower than real counts.

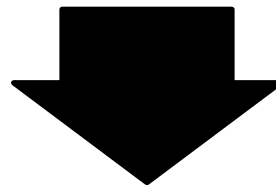
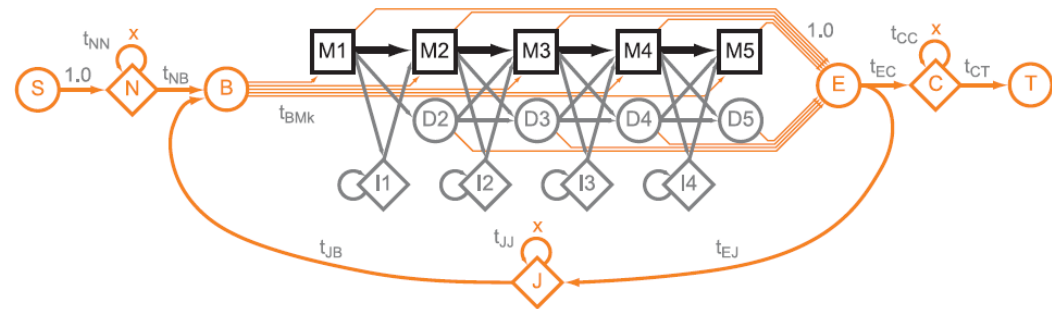
Reported thresholds (in parenthesis) are in terms of P-values

# MSA with HMM

Set of sequences



Trained HMM



*hmmalign*

Alignment of all sequences to the model

# HMMALIGN

## SINTAX:

**hmmalign** [-options] <hmmfile> <seqdb>

Reads a set of sequences and builds a MSA based on the model. Accepted formats include FASTA, EMBL, GenBank, DDBJ, UniProt, Stockholm, and SELEX.

```
# STOCKHOLM 1.0
```

```
MYG_ESCGI      .-VLSDAEWQLVLNIWAKVEADVAGHGQDILIRLFKGH PETLEKFDKFKHLKTEAEMKASEDLKK
#=GR MYG_ESCGI  PP  ..69*****
MYG_HORSE      g--LSDGEWQQVLNVWGKVEADIAGHGQEV LIRLFTGHPETLEKFDKFKHLKTEAEMKASEDLKK
#=GR MYG_HORSE  PP  8..89*****
MYG_PROGU      g--LSDGEWQLVLNVWGKVEGDL SGHGQEV LIRLFKGH PETLEKFDKFKHLKAEDEMRASEELKK
#=GR MYG_PROGU  PP  8..89*****
MYG_SAISC      g--LSDGEWQLVLNIWGKVEADIPSHGQEV LISLFKGH PETLEKFDKFKHLKSEDEM KASEELKK
#=GR MYG_SAISC  PP  8..89*****
```

In a match column, residues are upper case, and a '-' character means a deletion relative to the consensus.

In an insert column, residues are lower case, and a '.' is padding.

Insertions in a profile HMM are unaligned

BOTTOM LINE represents the posterior probability of each aligned residue. 0: 0-5%, 1: 5-15%, .. 9: 85-95%, \*: 95-100% posterior probability.

# Exercise

In the tutorial directory (<https://goo.gl/DsE2im>) two protein MSAs are present: Pkinase.sto (protein kinase) and fn3.sto (fibronectin 3).

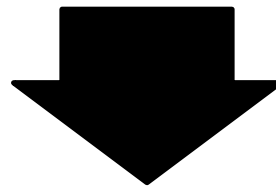
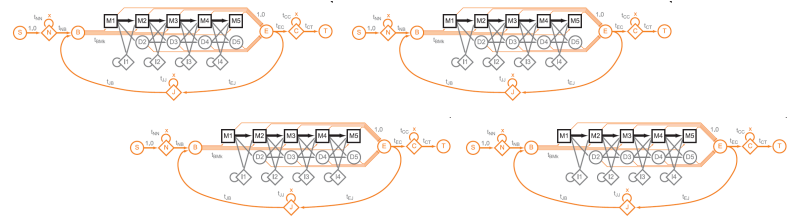
- Build two HMMs for the two alignment
- Check if the sequence 7LESS\_DROME contains protein kinase or fibronectin3 domains. If yes, how many?

# Scan HMM Library

## Protein Sequence

```
>protein_id  
VQLTVETITELAKNSYVAWGLSAAPISQNK  
GKNGLHKFYFKMDNSEDFFEKLQELAGKDE  
TYKGANIRWLGENVFDANSTIVSQDQEHHS  
AEVMDSL SRELHAKVARYDMAYVEYLSMCI  
APGFFANNEPIGAVECVSGIAHKMLKLI AA  
LLSAKY
```

## HMM Library



*hmmsearch*

List of HMMs that best match the sequence

# HMMPRESS & HMMSCAN

## SINTAX:

```
hmmcompress [-options] <hmmfile>
```

```
hmmsearch [-options] <hmmdb> <seqfile>
```

1. Generate a unique file putting together with “cat” all the previously generate hmm models. Use press to generate an hmm library.
2. Scan the new sequence against the hmm library

Query: 7LESS\_DROME [L=2554]

Accession: P13368

Description: RecName: Full=Protein sevenless; EC=2.7.10.1;

Scores for complete sequence (score includes all domains):

--- full sequence ---			--- best 1 domain ---			-#dom-		Model	Description
E-value	score	bias	E-value	score	bias	exp	N		
5.6e-57	178.0	0.4	3.5e-16	47.2	0.9	9.4	9	fn3	Fibronectin
1.1e-43	137.2	0.0	1.7e-43	136.5	0.0	1.3	1	Pkinase	Protein kinas