

If you learn just one thing Bioinformatics
just learn this:

Basic Local Alignment Search
Tool (BLAST)

BLAST programs

blastp: protein

blastn: DNA

query word ($W = 3$)

Query: GSVEDTTGSQSLAALLNKCKTP**PQG**QRLVNQWIKQPLMDKNRIEERLNLVAFVEDAELRQTLQEDL

neighborhood
words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc...	

neighborhood
score threshold
($T = 13$)

Query: 325 SLAALLNKCKTP**PQG**QRLVNQWIKQPLMDKNRIEERLNVEA 365
+LA++L+ TP G R++ +W+ P+ D + ER + A
Sbjct: 290 TLASVLDCTVT**PMG**SRMLKRULHMPVRDTRVLLERQQTIGA 330

High-scoring Segment Pair (HSP)

E-value

$$E \approx 1 - e^{-p(S \geq x)D}$$

Where,

x = a score cutoff

D = database size

p = P-value

Example BLAST output

See the file in the directory

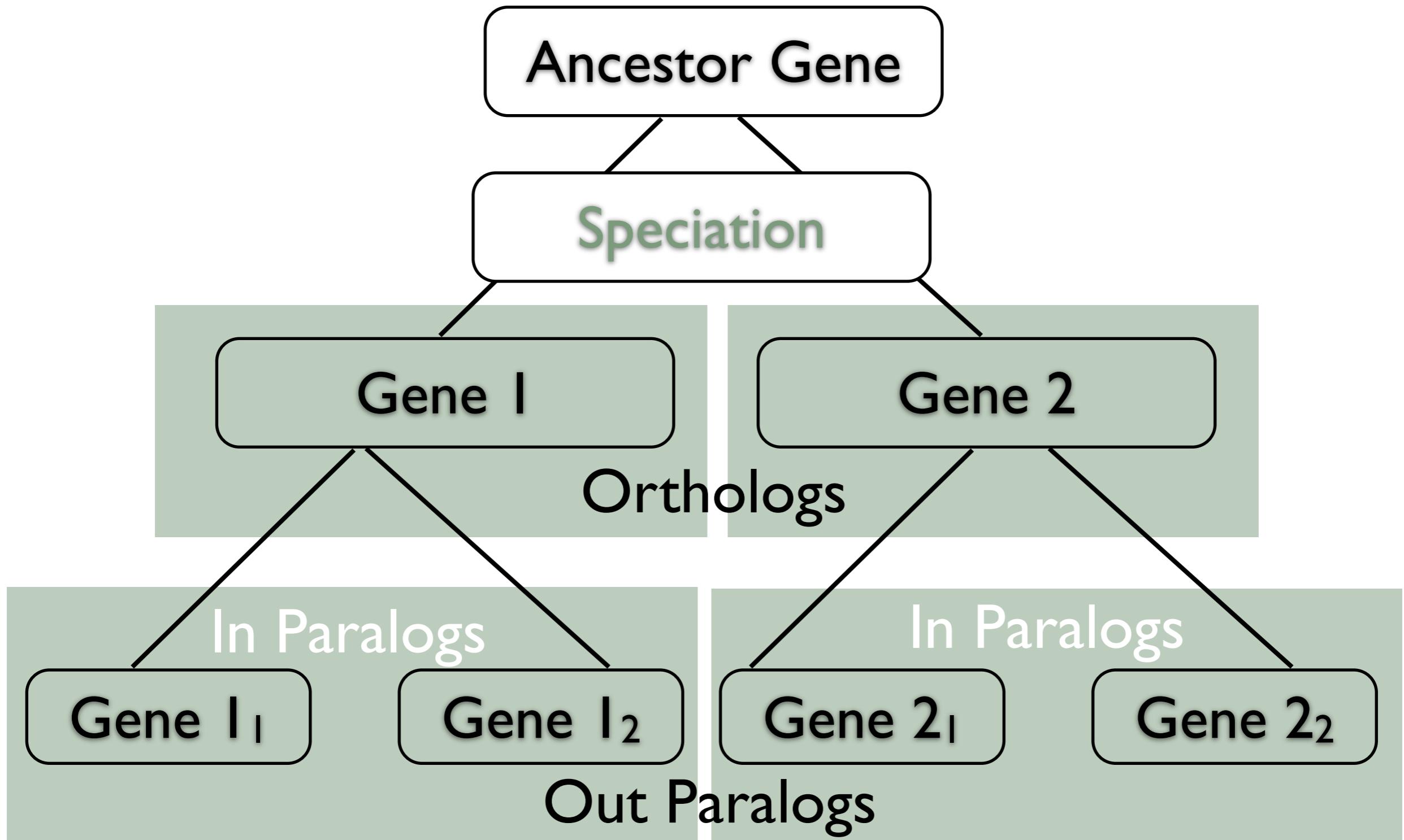
BLOSUM62

Find the score of PQG
matching PQG using
BLOSUM62

Homologs

Genes related by evolution.

Orthologs





Fitch W. (1970). "Distinguishing homologous from analogous proteins".
Syst Zool **19** (2): 99–113.

DISTINGUISHING HOMOLOGOUS FROM ANALOGOUS PROTEINS

WALTER M. FITCH

Abstract

Fitch, W. M. (Dept. Physiological Chem., U. Wisconsin, Madison 53706) 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19:99–113.—This work provides a means by which it is possible to determine whether two groups of related proteins have a common ancestor or are of independent origin. A set of 16 random amino acid sequences were shown to be unrelated by this method. A set of 16 real but presumably unrelated proteins gave a similar result. A set of 24 model proteins which was composed of two independently evolving groups, converging toward the same chemical goal, was correctly shown to be convergently related, with the probability that the result was due to chance being $<10^{-n}$. A set of 24 cytochromes composed of 5 fungi and 19 metazoans was shown to be divergently related, with the probability that the result was due to chance being $< 10^{-n}$. A process was described which leads to the absolute minimum of nucleotide replacements required to account for the divergent descent of a set of genes given a particular topology for the tree depicting their ancestral relations. It was also shown that the convergent processes could realistically lead to amino acid sequences which would produce positive tests for relatedness, not only by a chemical criterion, but by a genetic (nucleotide sequence) criterion as well. Finally, a realistic case is indicated where truly homologous traits, behaving in a perfectly expectable way, may nevertheless lead to a ludicrous phylogeny.

The demonstration that two proteins are related has been attempted using two different criteria. One criterion is to show that their chemical structures are very similar. An early example of this approach was the observation of the relatedness of the oxygen carrying proteins, myoglobin and hemoglobin (Watson and Kendrew, 1961). More recent is the relatedness of two enzymes in carbohydrate metabolism, lysozyme and alpha-lactalbumin (Brew, Vanaman and Hill, 1967). The other criterion is to show that underlying genetic structures of the proteins are more alike than one would expect by chance. This is now possible because our knowledge of the genetic code permits us to determine how many nucleotide positions, at the minimum, must differ in the genes encoding the two presumptively homologous proteins. One then compares the answer obtained to the number of differences one would expect for unrelated proteins. An example of this approach is the observation of the relatedness of plant and bacterial ferredoxins (Matsubara,

Jukes and Cantor, 1969) for which added evidence has been produced (Fitch, 1970a). But regardless of the approach, the impulse, too powerful to resist, is to conclude that a particular pair of proteins had a common genic ancestor if they meet whichever criterion the observer uses.

Now two proteins may appear similar because they descend with divergence from a common ancestral gene (i.e., are homologous in a time-honoured meaning dating back at the least to Darwin's *Origin of Species*) or because they descend with convergence from separate ancestral genes (i.e., are analogous). And, if a common genic ancestor is to be the conclusion, a genetic criterion should be superior to a chemical criterion. This is because analogous gene products, although they have no common ancestor, do serve similar functions and may well be expected to have similar chemical structures and thereby be confused with homologous gene products. This danger can only be increased by using a chemical, as opposed to a genetic, criterion.

Ortholog determination

Fundamental for comparative genomics

Open problem

No clear winner

Briefings in Bioinformatics

[ABOUT THIS JOURNAL](#) [CONTACT THIS JOURNAL](#) [SUBSCRIPTIONS](#)
[CURRENT ISSUE](#) [ARCHIVE](#) [SEARCH](#)
Institution: J Craig Venter Institute [Sign In as Personal Subscriber](#)
[Oxford Journals](#) > [Life Sciences & Mathematics & Physical Sciences](#) > [Briefings in Bioinformatics](#) > Volume 12, Issue 5

Special Issue: Orthology and Applications

Volume 12 Issue 5 September 2011

[Clear](#)
[Get All Checked Abstracts](#)

▲ Editorial

- Christophe Dessimoz
Editorial: Orthology and applications
Brief Bioinform (2011) 12(5): 375-376 doi:10.1093/bib/bbr057
[» Extract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#) [» Permissions](#)

▲ Obituary

- Eugene V. Koonin
Obituary: Walter Fitch and the orthology paradigm
Brief Bioinform (2011) 12(5): 377-378 doi:10.1093/bib/bbr058
[» Extract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#) [» Permissions](#)

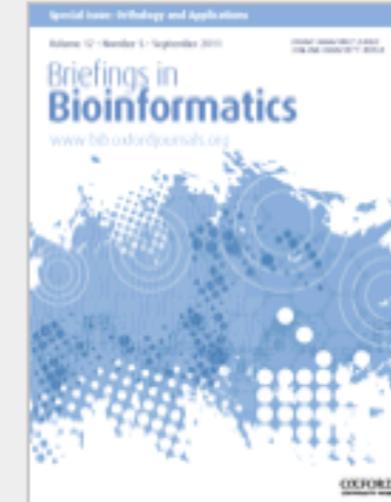
▲ Special Issue Papers

- David M. Kristensen, Yuri I. Wolf, Arcady R. Mushegian, and Eugene V. Koonin
Computational methods for Gene Orthology inference
Brief Bioinform (2011) 12(5): 379-391 doi:10.1093/bib/bbr030
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#) [» Permissions](#)
- Jean-Philippe Doyon, Vincent Ranwez, Vincent Daubin, and Vincent Berry
Models, algorithms and programs for phylogeny reconciliation
Brief Bioinform (2011) 12(5): 392-400 doi:10.1093/bib/bbr045
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#) [» Permissions](#)

[« Previous](#) | [Next Issue »](#)

This Issue

September 2011 12 (5)



- [» Index By Author](#)
- [» Front Matter \(PDF\)](#)
- [» Table of Contents \(PDF\)](#)
- [» Back Matter \(PDF\)](#)

- [» Editorial](#)
- [» Obituary](#)
- [» Special Issue Papers](#)
- [» Letter to the Editor](#)
- [» Non Special Issue Papers](#)
- [» Letter to the Editor](#)

Search this journal:

[GO](#)
[Advanced »](#)

Current Issue

March 2012 13 (2)


[Alert me to new issues](#)

The Journal

- [About the Journal](#)
- [Rights & Permissions](#)
- [Publishers' Books for Review](#)
- [Dispatch date of the next issue](#)
- [We are mobile – find out more](#)
- [This journal is a member of the Committee on Publication Ethics \(COPE\)](#)

Ortholog determination

Sequence similarity based clustering

Tree based

Hybrid approach

Sequence similarity

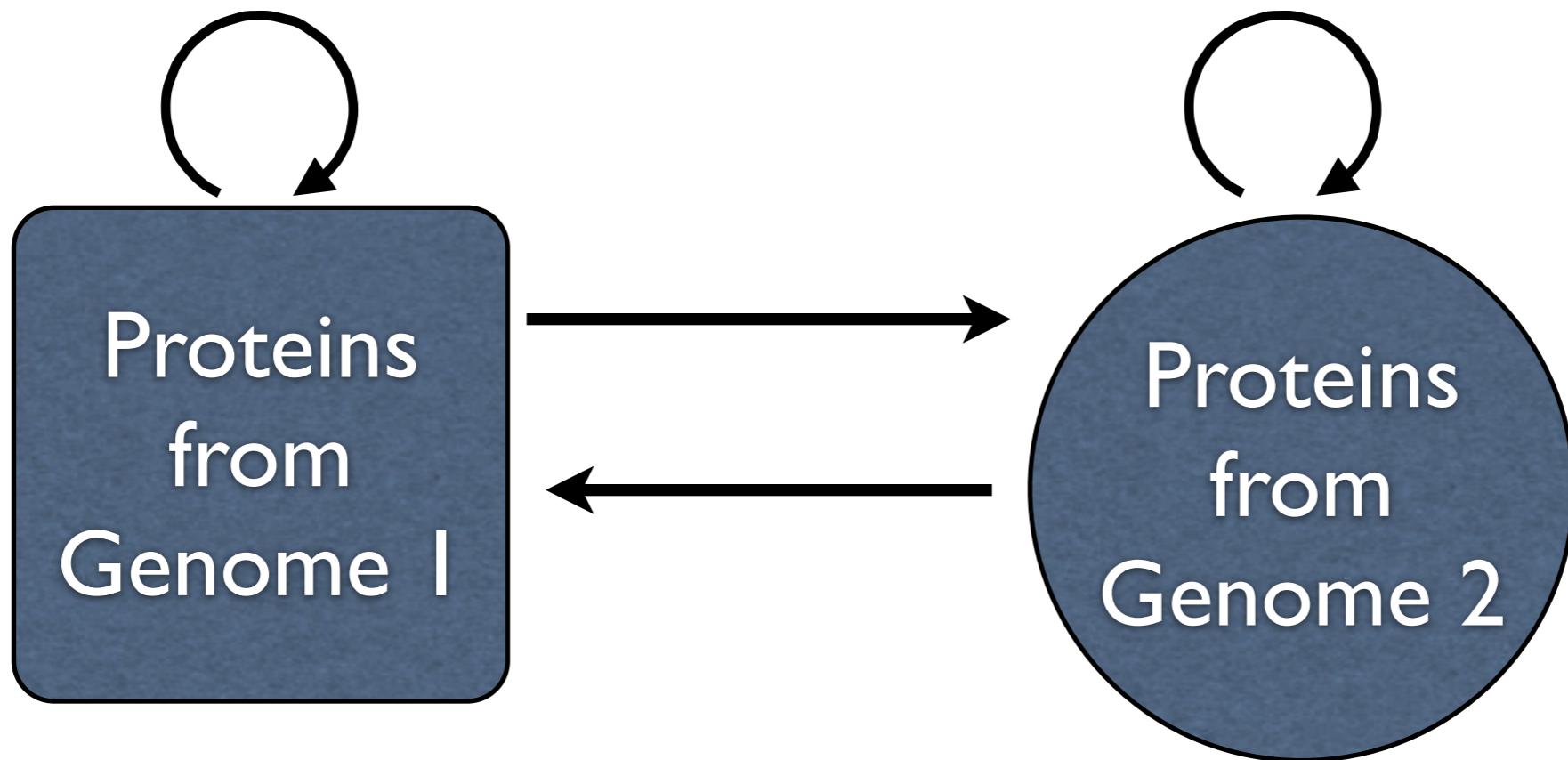
Pioneered by “COG”

Reciprocal Best Hit (usually BLAST)

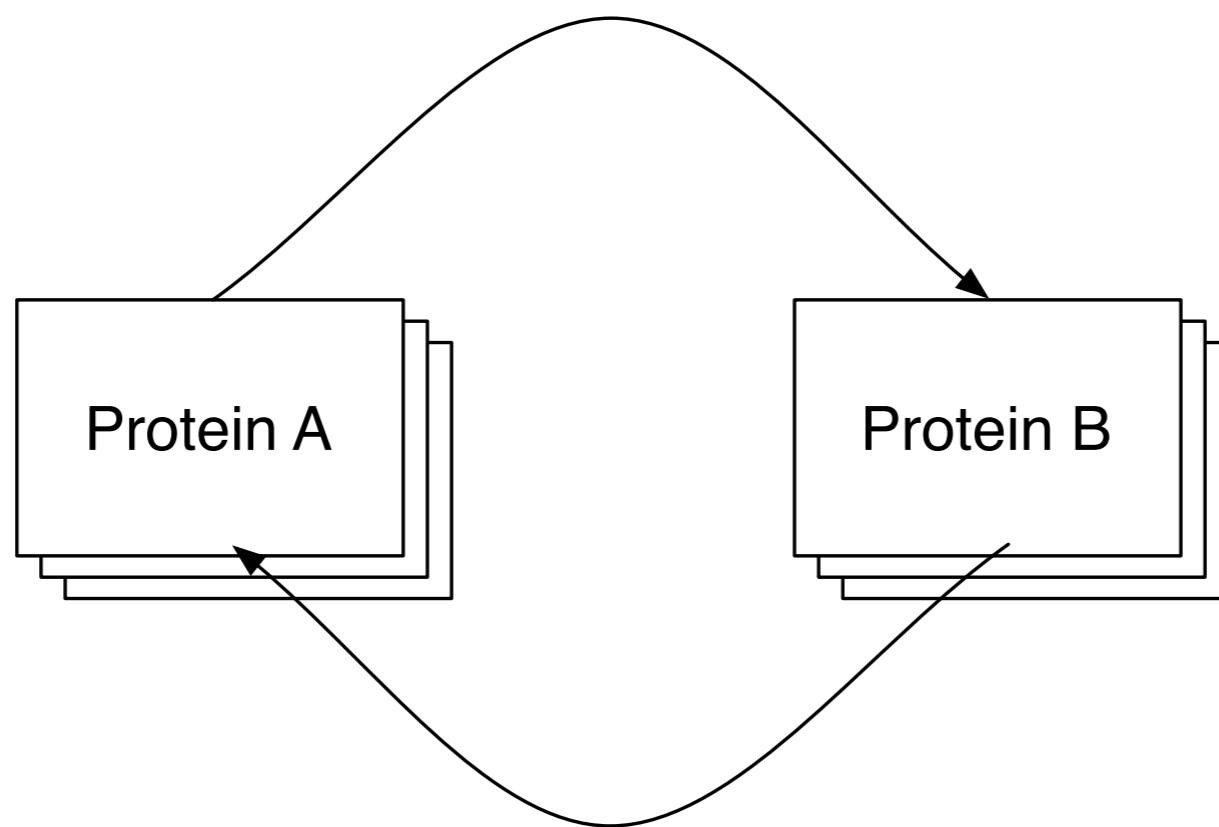
Additional clustering on top of RBH
(OrthoMCL)

Numerous databases: COG, eggNOC,
OrthoMCL, InParanoid...

All vs All BLAST



Reciprocal Best BLAST Hit



Orthologs

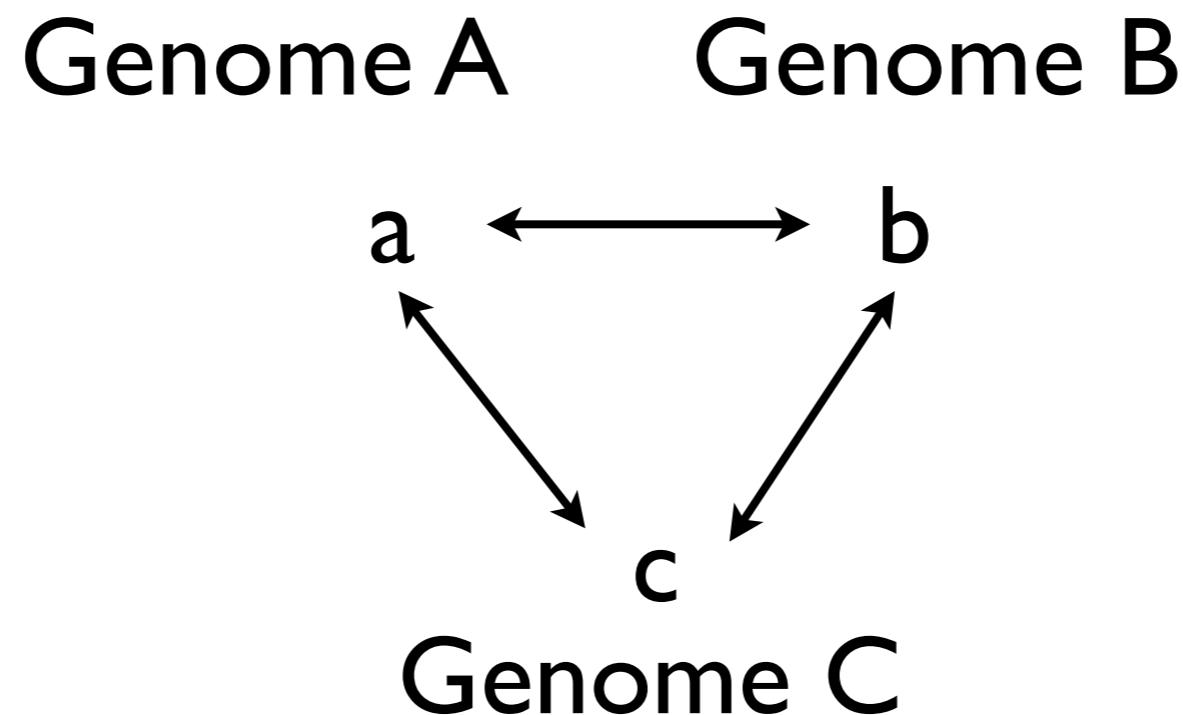
Nothing to do with function!

Homology vs Homoplasy



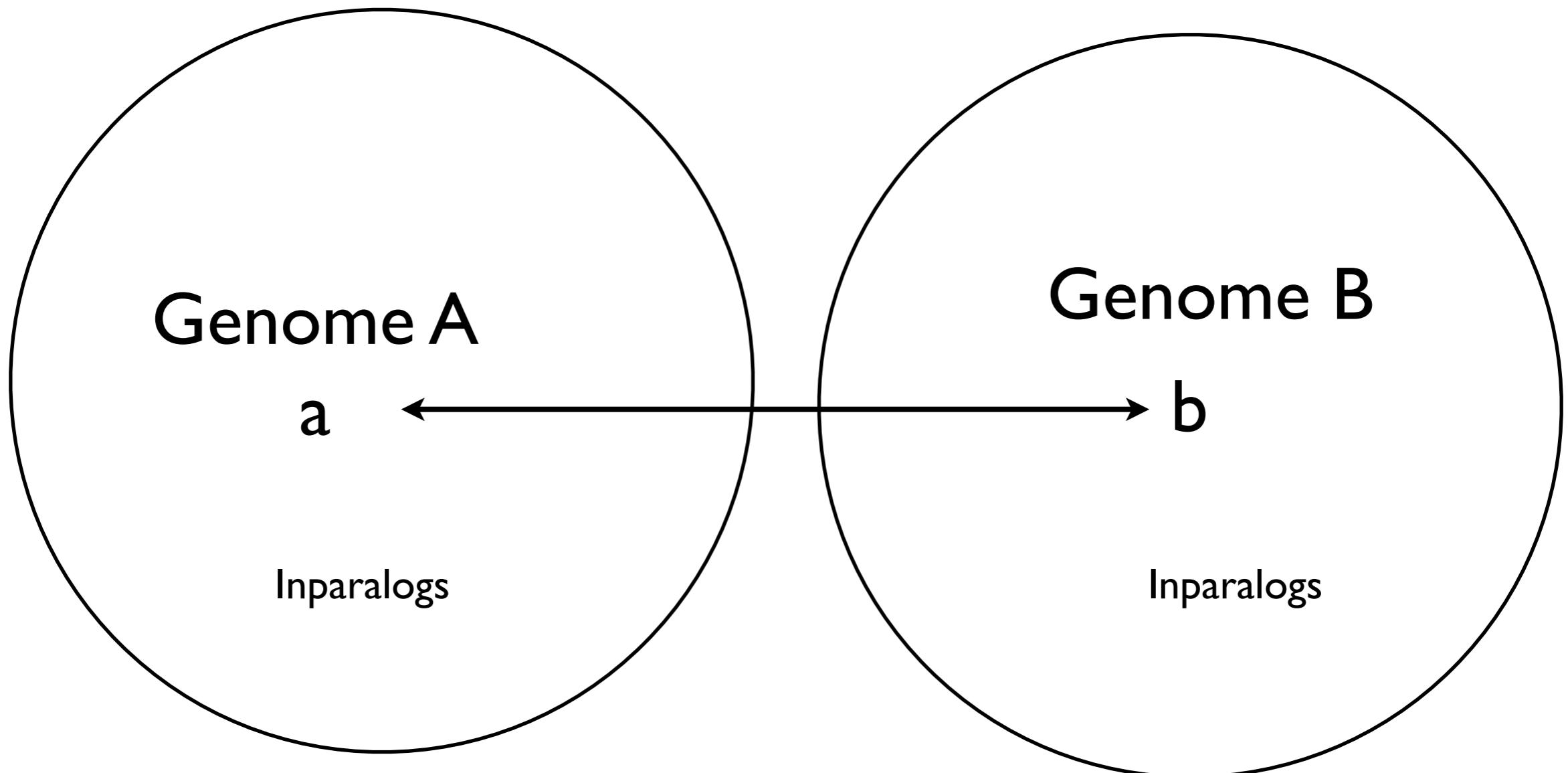
Cluster of orthologous groups (COG)

<http://www.ncbi.nlm.nih.gov/COG/>



InParanoid

<http://inparanoid.sbc.su.se/cgi-bin/index.cgi>



Download BLAST

<ftp://ftp.ncbi.nih.gov/blast/executables/release/2.2.25/>

Creating a BLAST DB from a multifasta file

```
formatdb -i multifasta
```

BLASTP

```
blastall -i input.fas -d dbname  
-o outputfile
```

Position Specific Scoring Matrix (PSSM)

$$P_{ca} = (n_{ca} + b_{ca}) / (N_c + B_c)$$

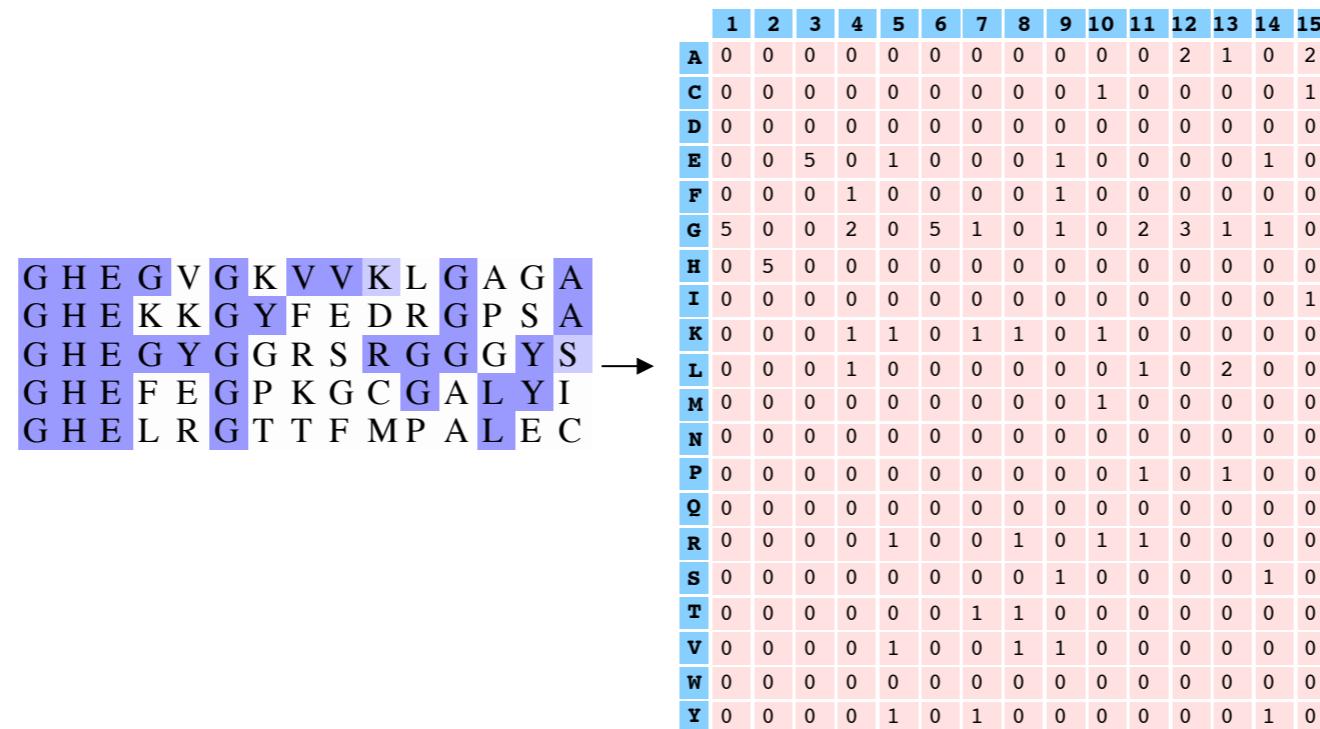
N_{ca} = real count

b_{ca} = pseudo count

N_c = Number of sequence rows in
alignment

B_c = total pseudo count (20 for protein
and 4 for nucleotides)

A *PSSM* is based on the *frequencies* of each residue in a specific position of a multiple alignment.



- Column 1: $f_{A,1} = \frac{0}{5} = 0, f_{G,1} = \frac{5}{5} = 1, \dots$
- Column 2: $f_{A,2} = \frac{0}{5} = 0, f_{H,2} = \frac{5}{5} = 1, \dots$
- ...
- Column 15: $f_{A,15} = \frac{2}{5} = 0.4, f_{C,15} = \frac{1}{5} = 0.2, \dots$

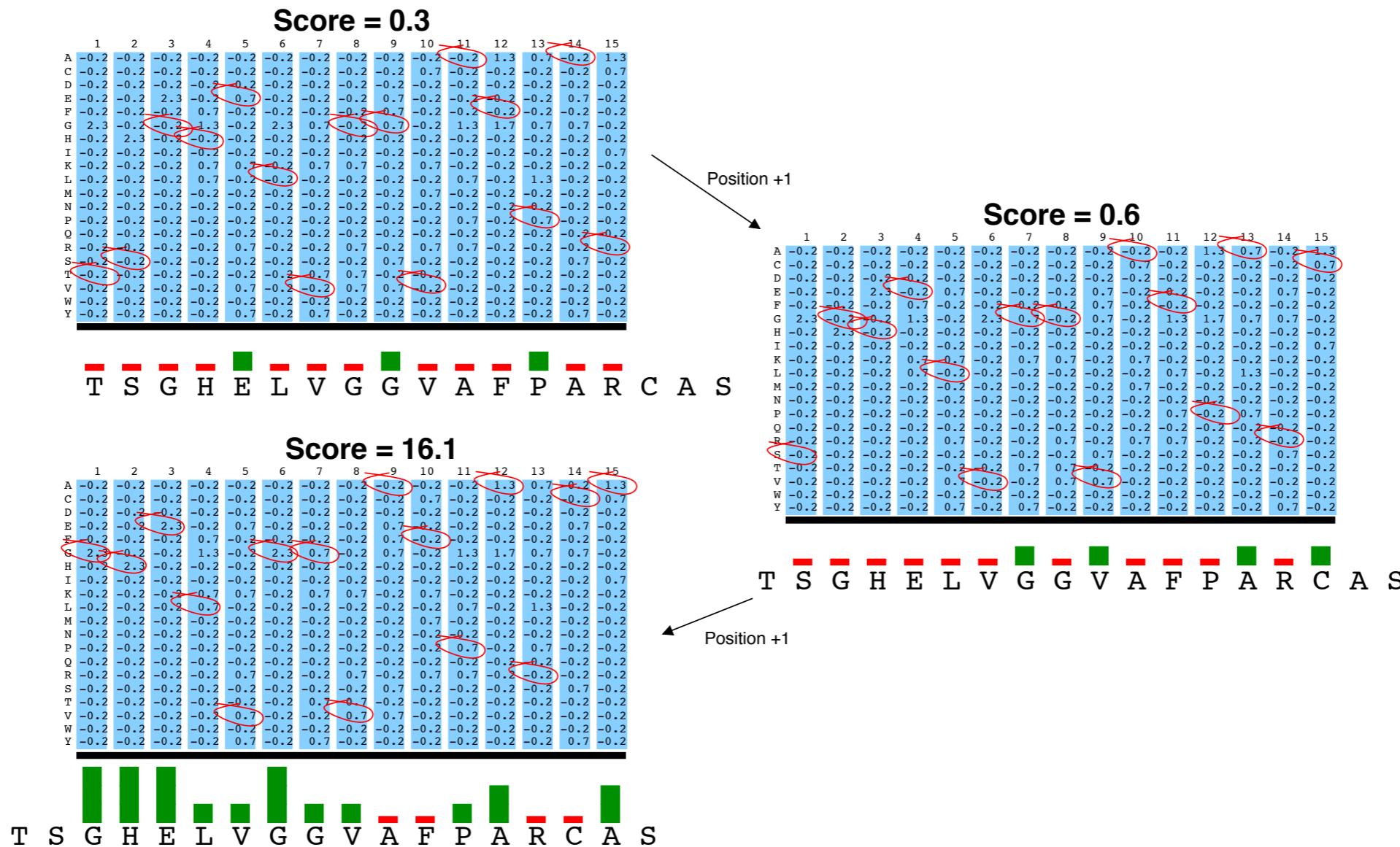
$$Score_{ij} = \log\left(\frac{f'_{ij}}{q_i}\right)$$

$qi = 1/20 = 0.05$ for protein

For A in position 1 = $\log((0+1)/(5+20)) / 0.05 = -0.2$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	1.3	0.7	-0.2	1.3
C	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7
D	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
E	-0.2	-0.2	2.3	-0.2	0.7	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
F	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
G	2.3	-0.2	-0.2	1.3	-0.2	2.3	0.7	-0.2	0.7	-0.2	1.3	1.7	0.7	0.7	-0.2
H	-0.2	2.3	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
I	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7
K	-0.2	-0.2	-0.2	0.7	0.7	-0.2	0.7	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
L	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	1.3	-0.2	-0.2
M	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2
N	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
P	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2
Q	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
R	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	0.7	0.7	0.7	-0.2	-0.2	-0.2	-0.2
S	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	0.7	-0.2
T	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
V	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
W	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2
Y	-0.2	-0.2	-0.2	-0.2	0.7	-0.2	0.7	-0.2	-0.2	-0.2	-0.2	-0.2	-0.2	0.7	-0.2

- At every position, a PSSM score is calculated by summing the scores of all columns;
 - The highest scoring position is reported.



- ▶ modeling positional dependencies
- ▶ recognizing pattern instances with indels
- ▶ modeling variable length patterns
- ▶ detecting boundaries

PSSM search

rpsblast can be used to search a PSSM.

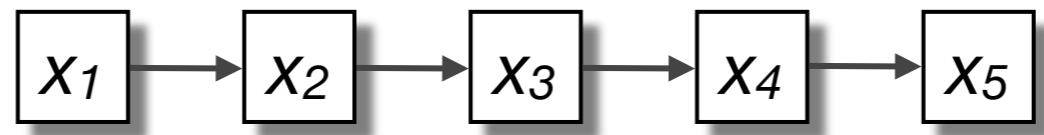
NCBI Conserved Domain Database (CDD) is
a collection of PSSMs.

Markov process

No state information

Memoryless

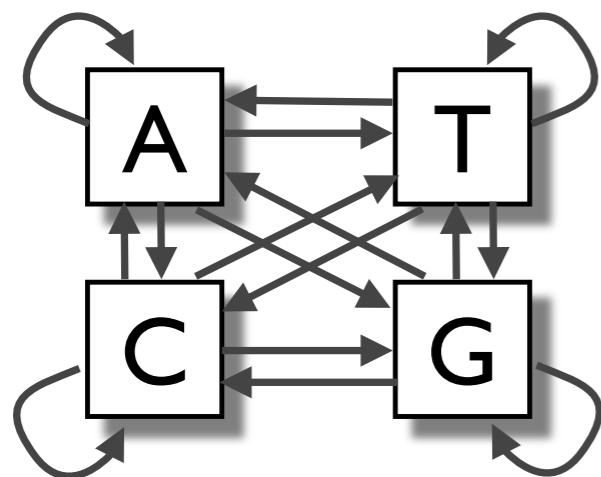
Markov Chains



$$p(x_1, x_2, x_3, \dots) = p(x_1) p(x_2|x_1) p(x_3|x_2) p(x_4|x_3) \dots$$

Markov Chains are memory less:
probability of a state depends only on the
previous state

Markov chains are defined as a state diagram



A Markov chain is defined by:

- a finite set of **states**, $S_1, S_2 \dots S_N$
- a set of **transition probabilities**: $a_{ij} = P(q_{t+1}=S_j|q_t=S_i)$
- and an **initial state probability distribution**, $\pi_i = P(q_0=S_i)$

Markov chains example

Observed sequence: $x = \text{abaaababbaa}$

Model:

*transition
probabilities*

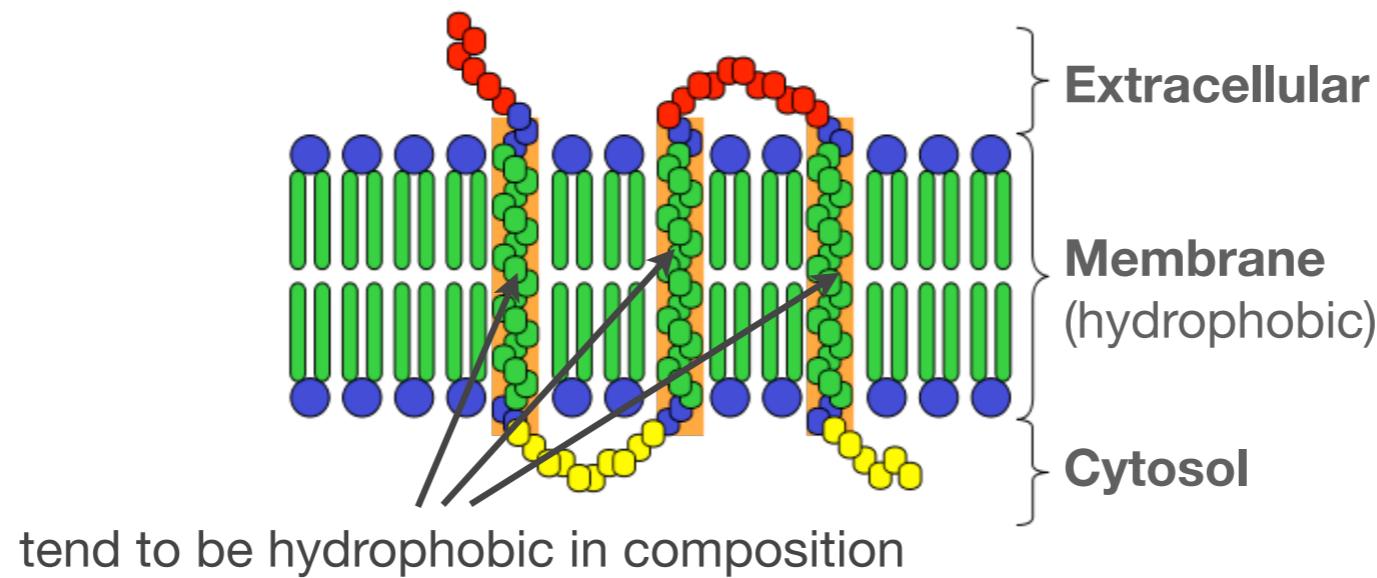
Prev <i>i</i>	Next <i>j</i>	Prob a_{ij}
a	a	0.7
a	b	0.3
b	a	0.5
b	b	0.5

*initial state
probability
distribution*

Start probs	π_i	a 0.5	b 0.5

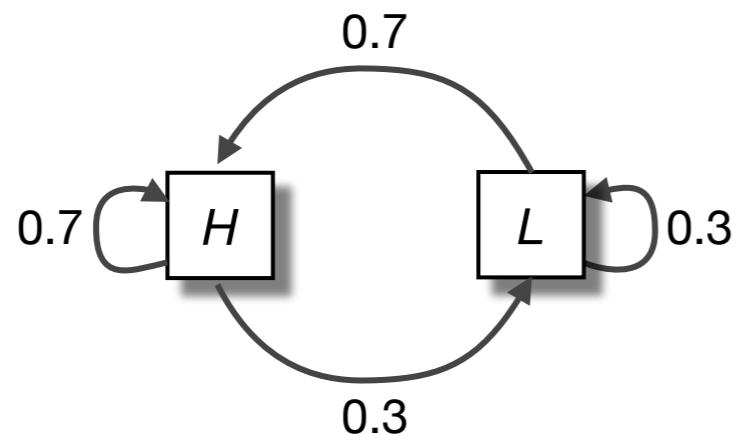
$$P(x) = 0.5 \times 0.3 \times 0.5 \times 0.7 \times 0.7 \times 0.3 \times 0.5 \times 0.3 \times 0.5 \times 0.5 \times 0.7$$

Markov chain example



Question: Is sequence **HHHLHH** a transmembrane protein?

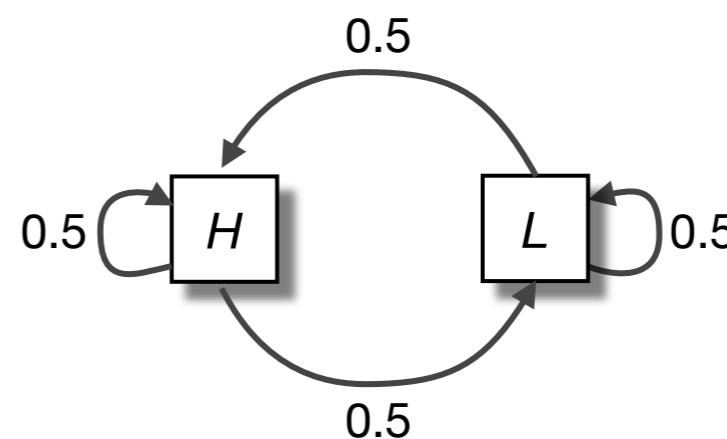
transmembrane model



Transmembrane (TM)

- $\pi(H) = 0.6, \pi(L) = 0.4$

null model



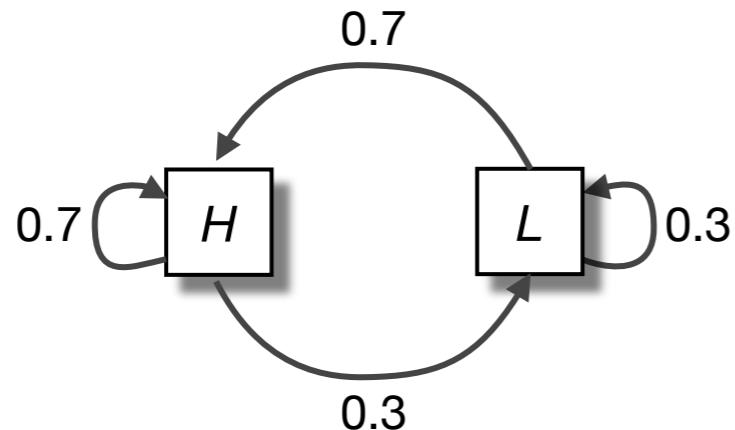
Extracellular/cytosolic (E/C)

- $\pi(H) = 0.5, \pi(L) = 0.5$

$$\frac{P(HHLHH \mid TM)}{P(HHLHH \mid EC)} = \frac{0.6 \times 0.7 \times 0.7 \times 0.3 \times 0.7 \times 0.7}{0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5} = \frac{0.043}{0.016} = 2.69$$

In other words, it is more than twice as likely that **HHLHH** is a transmembrane sequence. The log-odds score is: $\log_2(2.69) = 1.43$

Markov chain Parameter estimation



$\pi(H) = \# \text{ of sequences that begin with } H,$
normalized by the total # of training
sequences

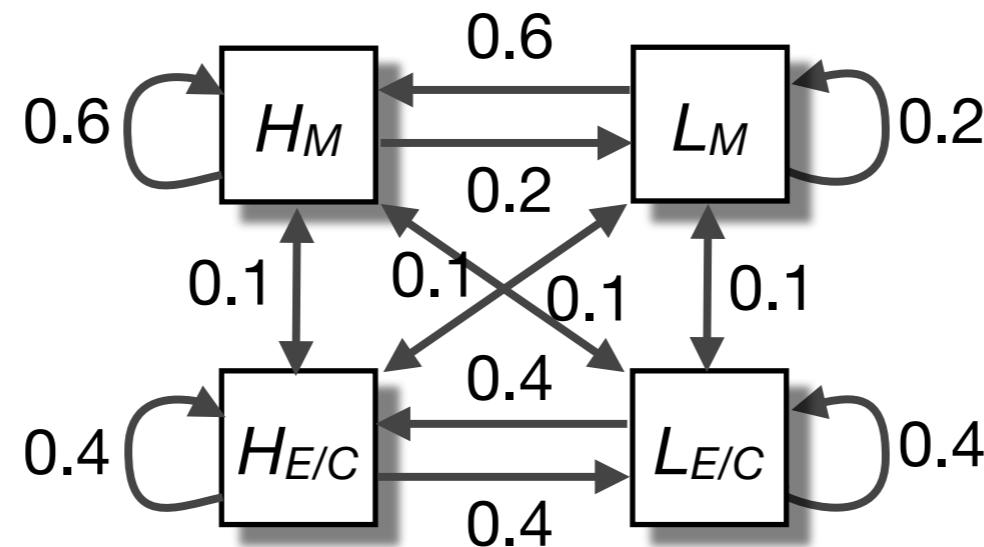
- $\pi(H) = 0.6, \pi(L) = 0.4$

HHHLLHHHHLLLHLHLLHLLLHLHHHL
HHHLHHLHLLLLLHHHHHLLLHHHHHL
HH . . . ($A_{HL} = 12, A_{H^*} = 40$)

$$a_{HL} = \frac{A_{HL}}{\sum_i A_{Hi}} \quad \frac{\#\text{HL pairs}}{\#\text{H}^*\text{ pairs}} \quad \frac{12}{40}$$

HMM:

Given a sequence of H and L find the transmembrane region

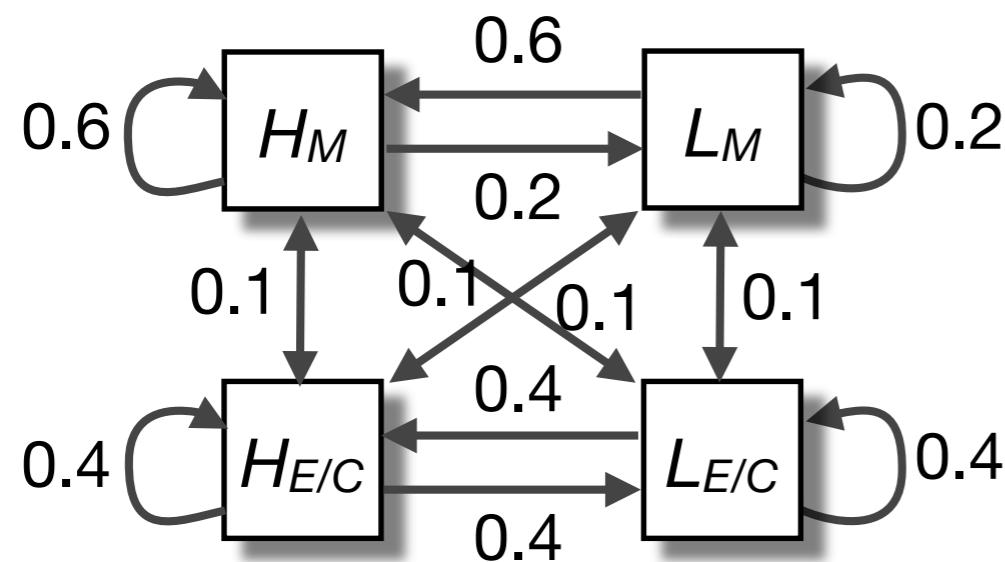


In our new model, there are multiple states that could account for each part of the observed sequence
i.e. we don't know which state emitted a given symbol from knowledge of the sequence and the structure of the model

- ▶ This is the *hidden* part of the problem

For our HMM

- Given HLLH..., we must infer the most probable state sequence
- This HMM state sequence will yield the boundaries between likely TM and E/C regions



HM, LM, LM, HM
 HM, LM, LM, HE/C
 HM, LM, LH/C, HM
 HM, LM, LH/C, HE/C
 HM, LE/C, LM, HM
 HM, LE/C, LM, HE/C
 HM, LE/C, LH/C, HM,
 HM, LE/C, LH/C, HE/C,
 HE/C, LM, LM, HM
 HE/C, LM, LM, HE/C
 HE/C, LM, LH/C, HM
 HE/C, LM, LH/C, HE/C
 HE/C, LE/C, LM, HM
 HE/C, LE/C, LM, HE/C
 HE/C, LE/C, LH/CM, HM
 HE/C, LE/C, LH/CM, HE/C

Markov Chains

- States: $S_1, S_2 \dots S_N$
- Initial probabilities: π_i
- Transition probabilities: a_{ij}

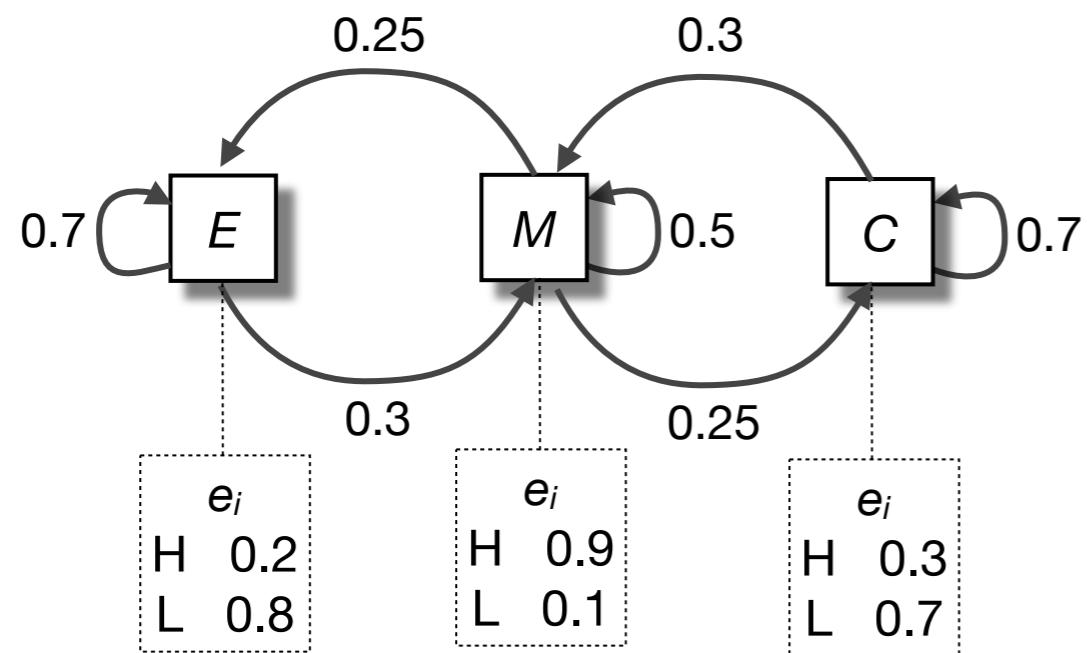
One-to-one correspondence
between states and symbols

Hidden Markov Models

- States: $S_1, S_2 \dots S_N$
- Initial probabilities: π_i
- Transition probabilities: a_{ij}
- **Alphabet** of emitted symbols, Σ
- **Emission probabilities:** $e_i(a)$
probability state i emits symbol a

Symbol may be emitted by more
than one state

Similarly, a state can emit more
than one symbol



Transition prob:

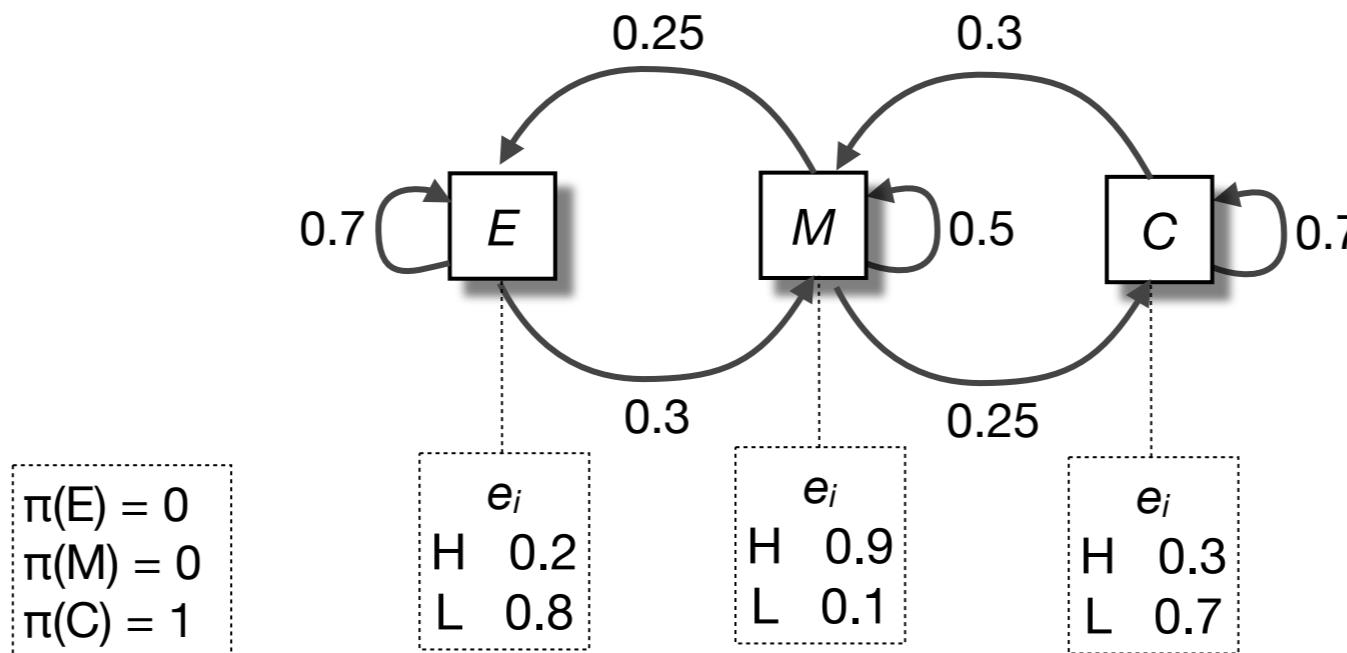
$$a_{ij} = \begin{bmatrix} E & M & C \\ 0.7 & 0.3 & 0 \\ 0.25 & 0.5 & 0.25 \\ 0 & 0.3 & 0.7 \end{bmatrix}$$

Emission prob:

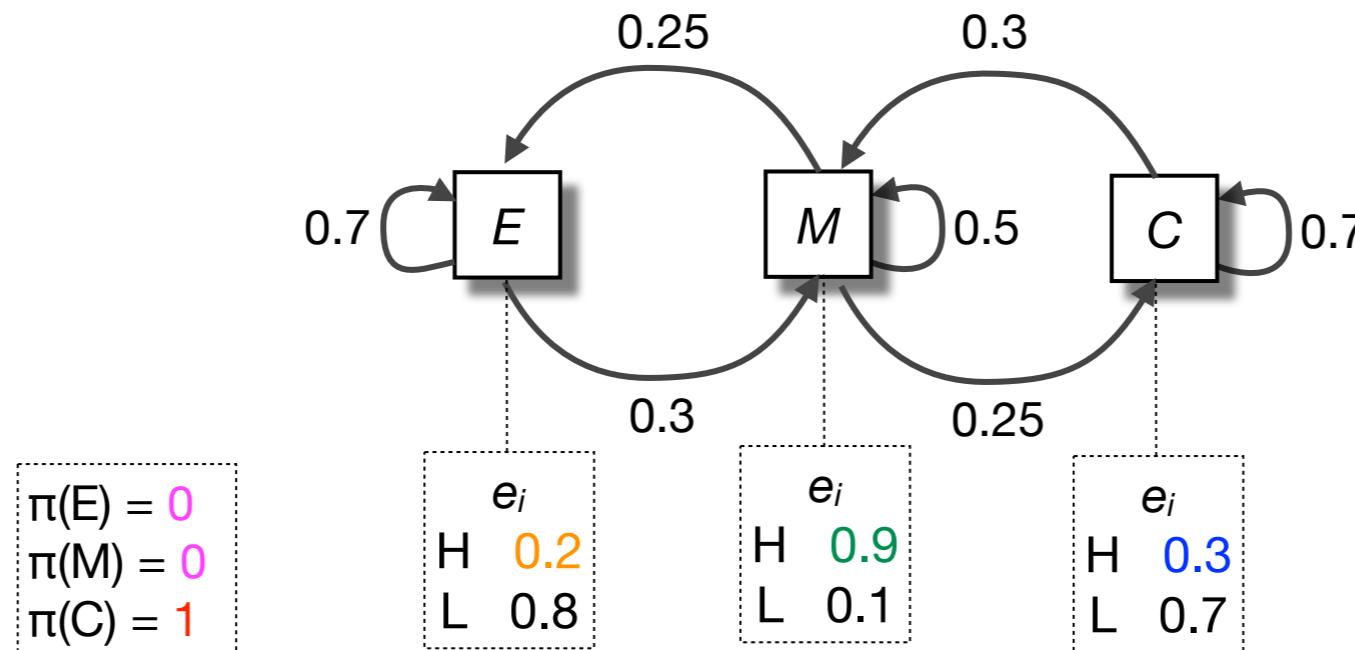
$$a_{ij} = \frac{A_{ij}}{\sum_j A_{ij}}$$

$$e_i(x) = \frac{E_i(x)}{\sum_x E_i(x')}$$

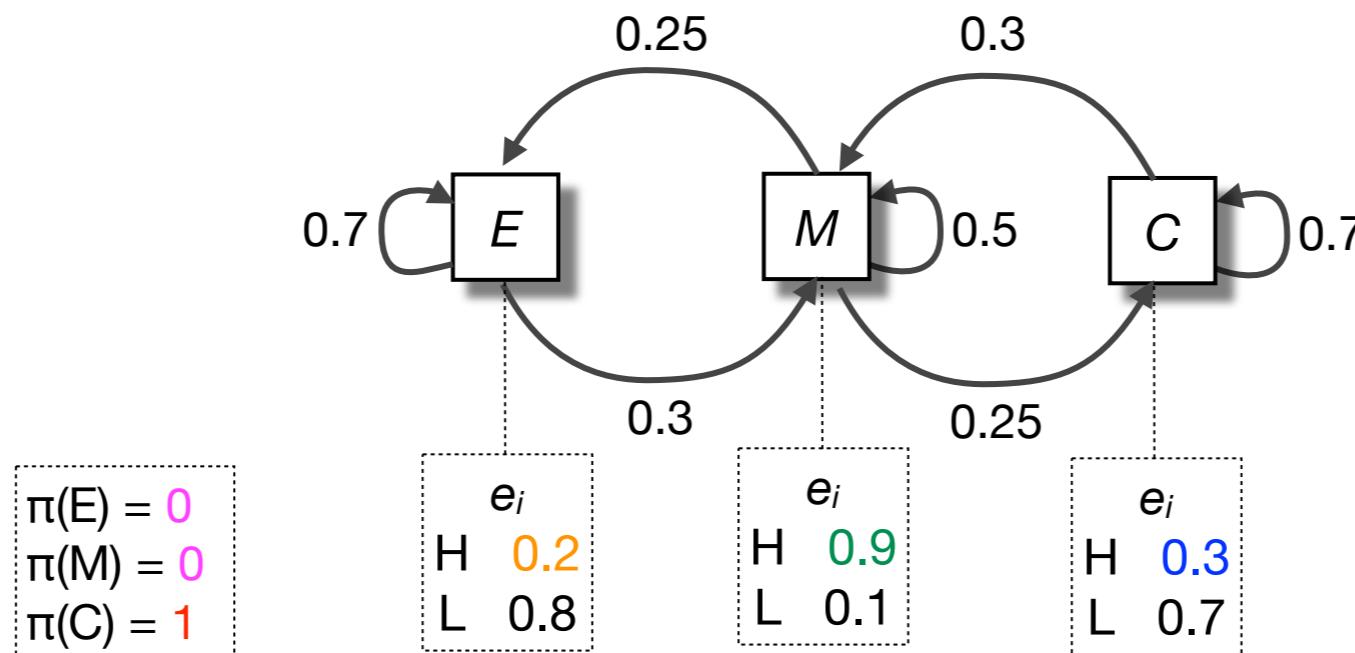
	<i>E</i>	<i>M</i>	<i>C</i>
π_i	0	0	1
$e_i(H)$	0.2	0.9	0.3
$e_i(L)$	0.8	0.1	0.7



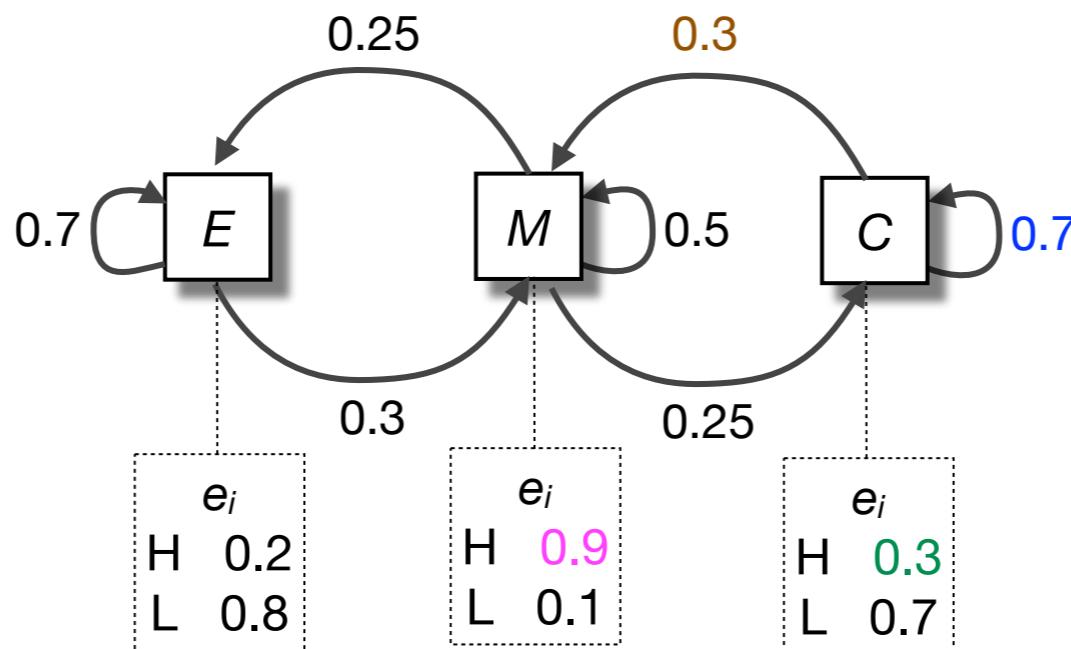
		Query Sequence				
States		H	H	L	L	H
E						
M						
C						
START						



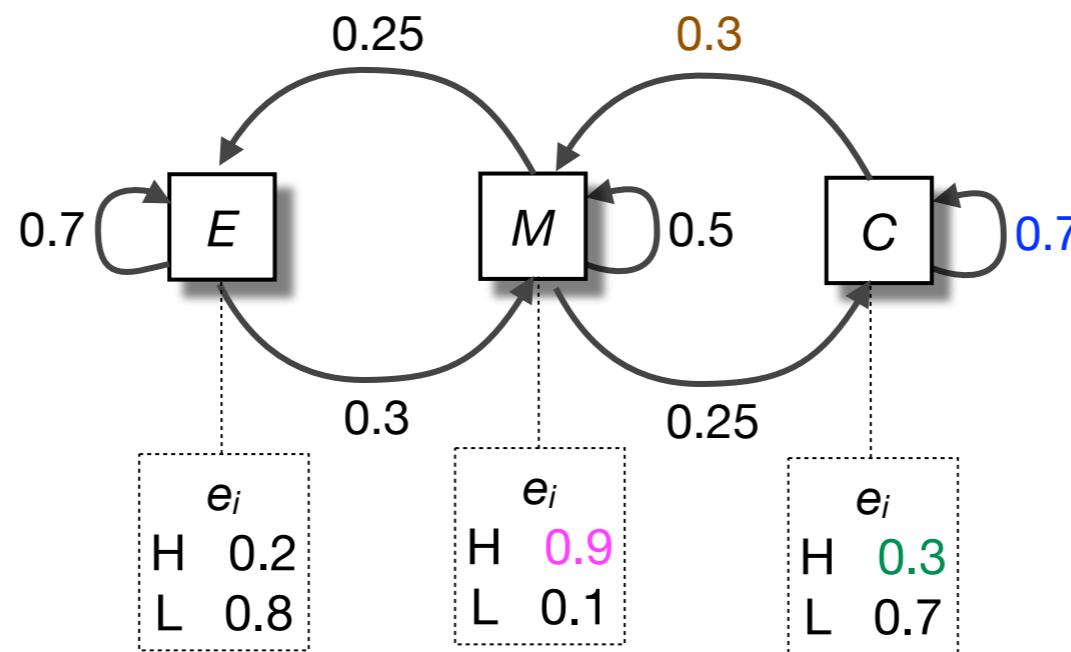
States	Query Sequence				
	H	H	L	L	H
E	0×0.2 =0				
M	0×0.9 =0				
C	1×0.3 =0.3				
START					



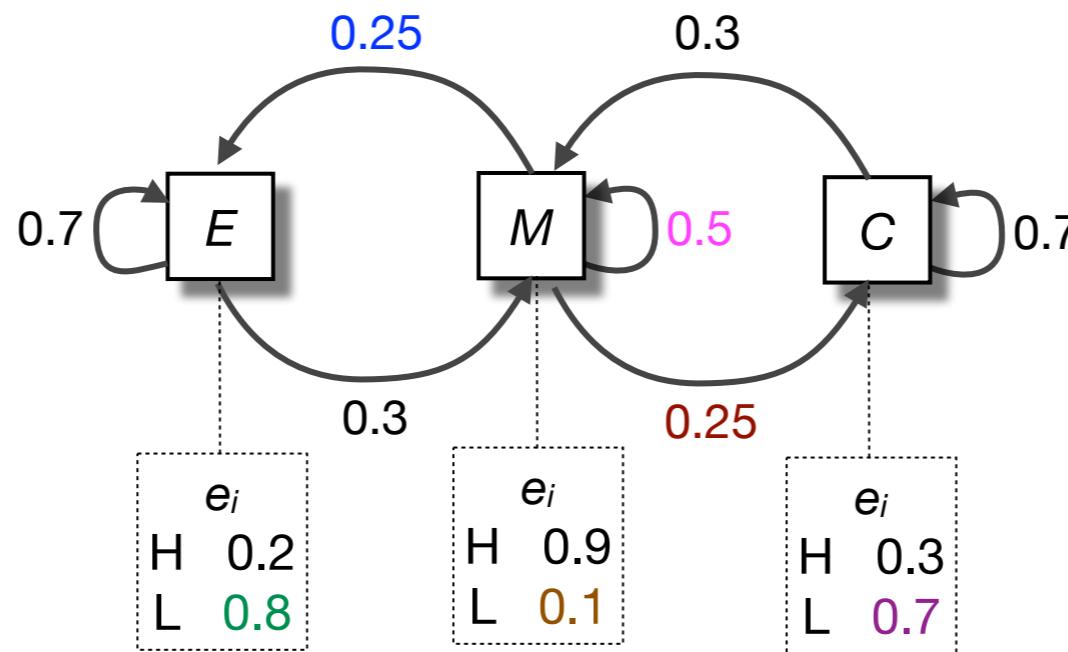
		Query Sequence				
States		H	H	L	L	H
E		0×0.2 = 0				
M		0×0.9 = 0				
C		1×0.3 = 0.3				
START						



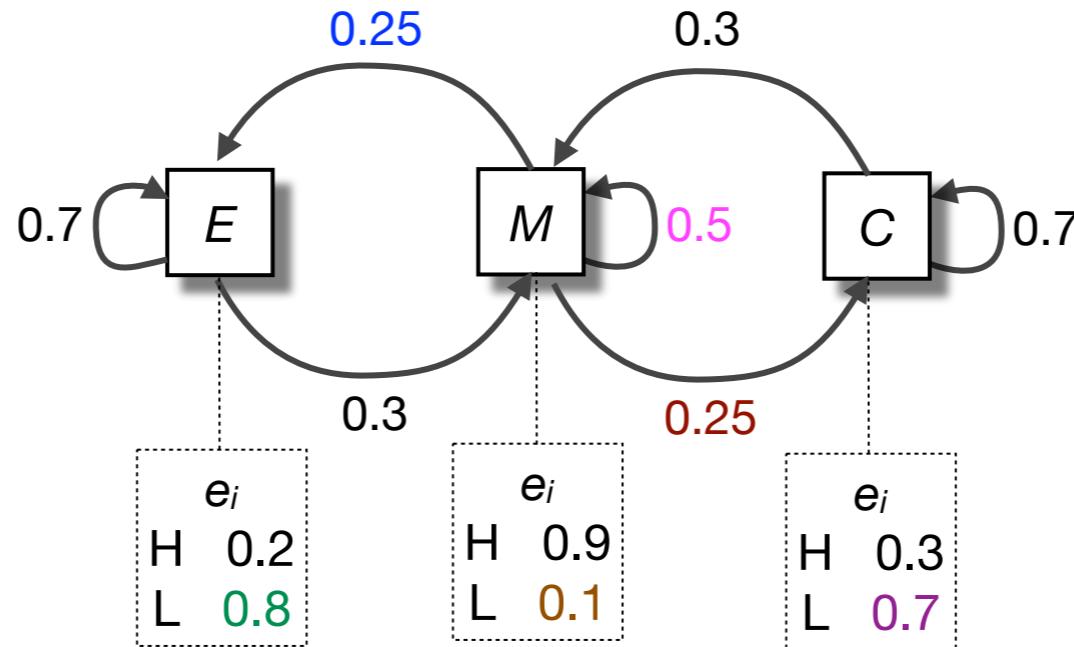
		Query Sequence				
States		H	H	L	L	H
E		$0 \times 0.2 = 0$	-			
M		$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$			
C		$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$			
START						



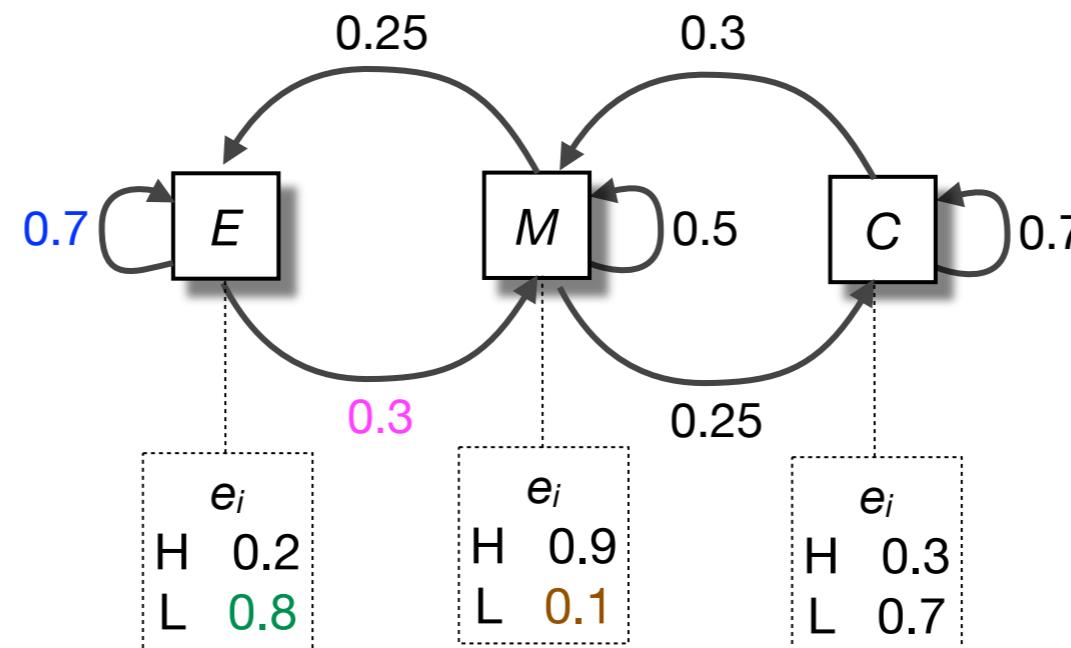
States	Query Sequence				
	H	H	L	L	H
E	$0 \times 0.2 = 0$	-			
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$			
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$			
START					



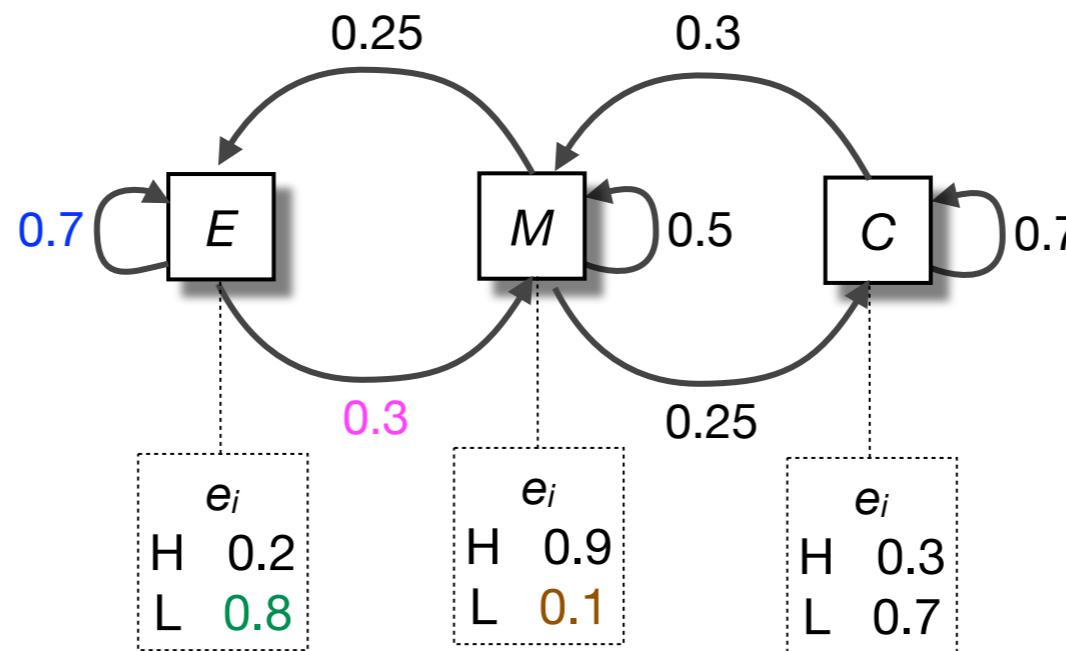
States	Query Sequence				
	H	H	L	L	H
E	$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$		
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.04$		
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$		
START					



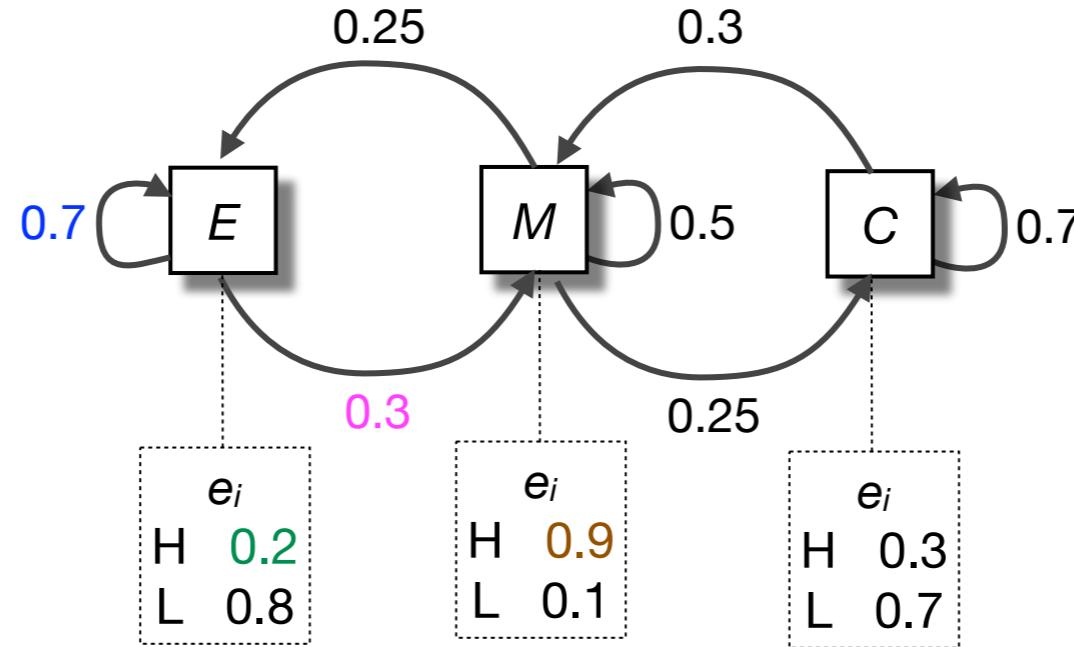
States	Query Sequence				
	H	H	L	L	H
E	$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$		
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.004$		
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$		
START					



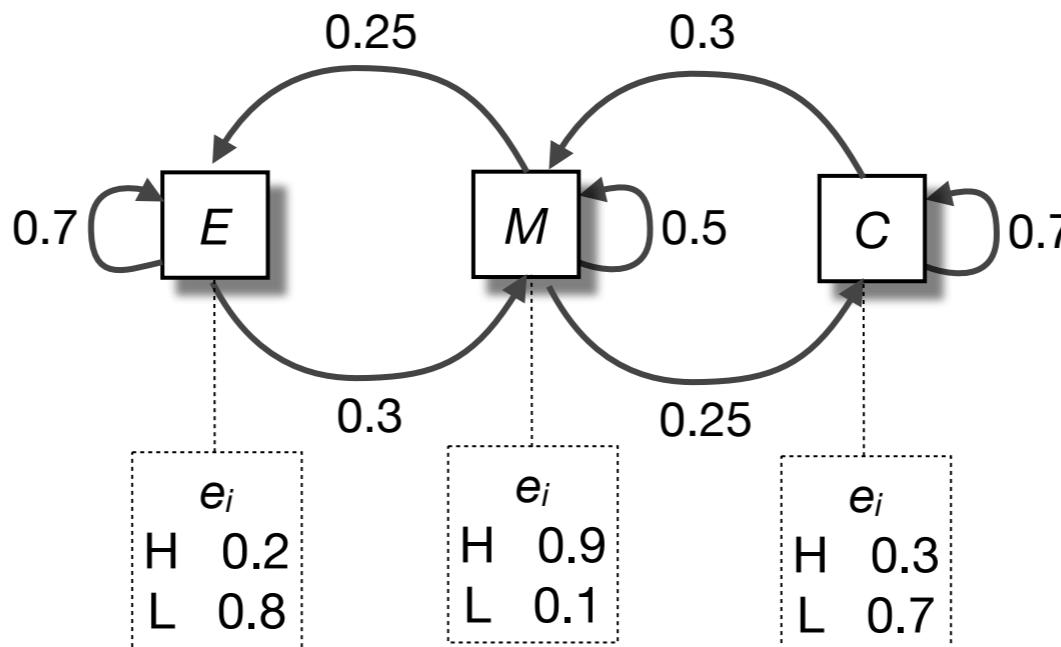
States	Query Sequence				
	H	H	L	L	H
E	0×0.2 =0	-	$0.25 \times 0.8 \times 0.081$ $\quad\quad\quad = 0.016$	$0.7 \times 0.8 \times 0.016$ $\quad\quad\quad = 0.009$	
M	0×0.9 =0	$0.3 \times 0.9 \times 0.3$ $\quad\quad\quad = 0.081$	$0.5 \times 0.1 \times 0.081$ $\quad\quad\quad = 0.04$	$0.3 \times 0.1 \times 0.016$ $\quad\quad\quad = 0.0005$	
C	1×0.3 =0.3	$0.7 \times 0.3 \times 0.3$ =0.063	$0.25 \times 0.7 \times 0.081$ $\quad\quad\quad = 0.014$	-	
START					



States	Query Sequence				
	H	H	L	L	H
E	$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$	$0.7 \times 0.8 \times 0.016 = 0.009$	
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.004$	$0.3 \times 0.1 \times 0.016 = 0.0005$	
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$	-	
START					



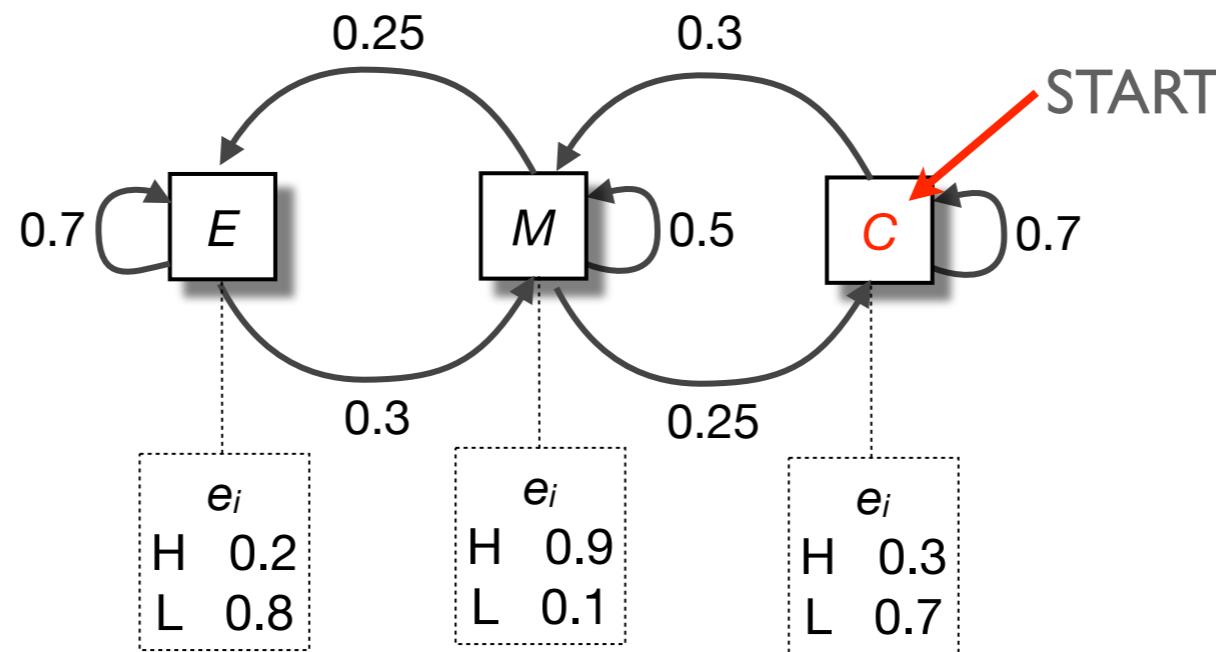
States	Query Sequence				
	H	H	L	L	H
E	0×0.2 =0	-	$0.25 \times 0.8 \times 0.081$ $=0.016$	$0.7 \times 0.8 \times 0.016$ $=0.009$	$0.7 \times 0.2 \times 0.009$ $=0.001$
M	0×0.9 =0	$0.3 \times 0.9 \times 0.3$ $=0.081$	$0.5 \times 0.1 \times 0.081$ $=0.04$	$0.3 \times 0.1 \times 0.016$ $=0.0005$	$0.3 \times 0.9 \times 0.009$ $=0.002$
C	1×0.3 =0.3	$0.7 \times 0.3 \times 0.3$ $=0.063$	$0.25 \times 0.7 \times 0.081$ $=0.014$	-	-
START					



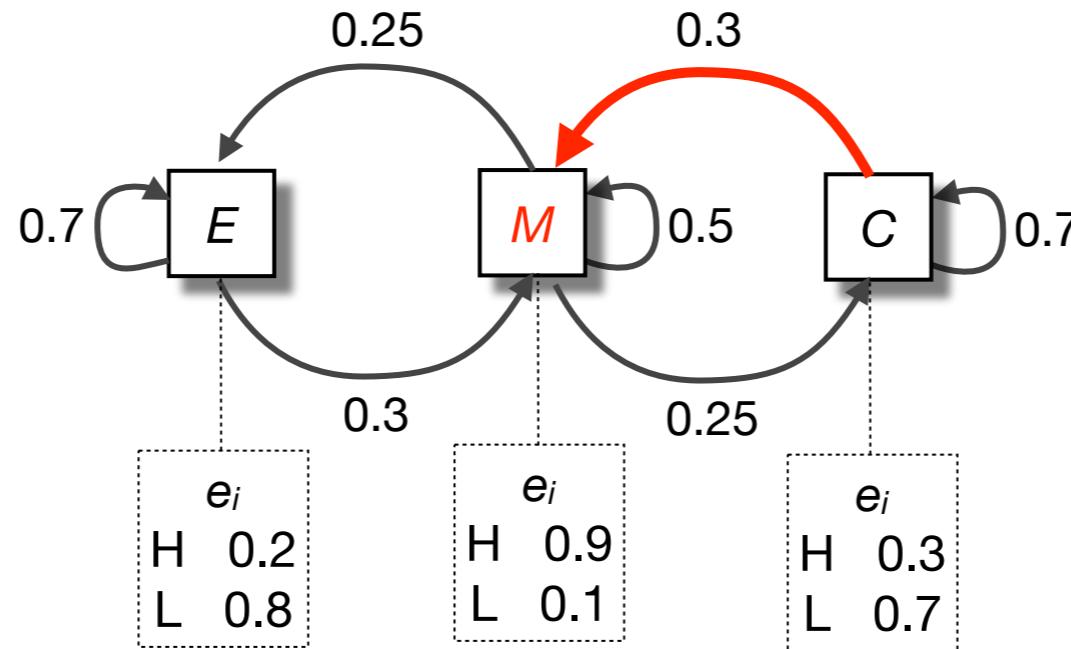
Query Sequence

States	H	H	L	L	H
E	0×0.2 =0	-	$0.25 \times 0.8 \times 0.081$ =0.016	$0.7 \times 0.8 \times 0.016$ =0.009	$0.7 \times 0.2 \times 0.009$ =0.001
M	0×0.9 =0	$0.3 \times 0.9 \times 0.3$ =0.081	$0.5 \times 0.1 \times 0.081$ =0.04	$0.3 \times 0.1 \times 0.016$ =0.0005	$0.3 \times 0.9 \times 0.009$ =0.002
C	1×0.3 =0.3	$0.7 \times 0.3 \times 0.3$ =0.063	$0.25 \times 0.7 \times 0.081$ =0.014	-	-

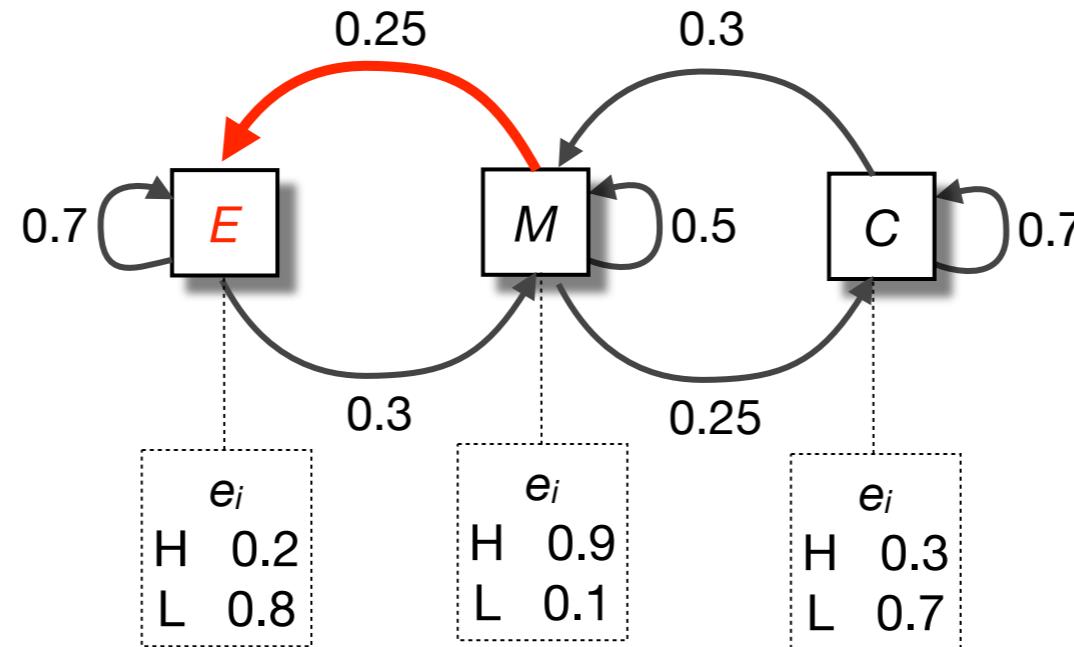
START



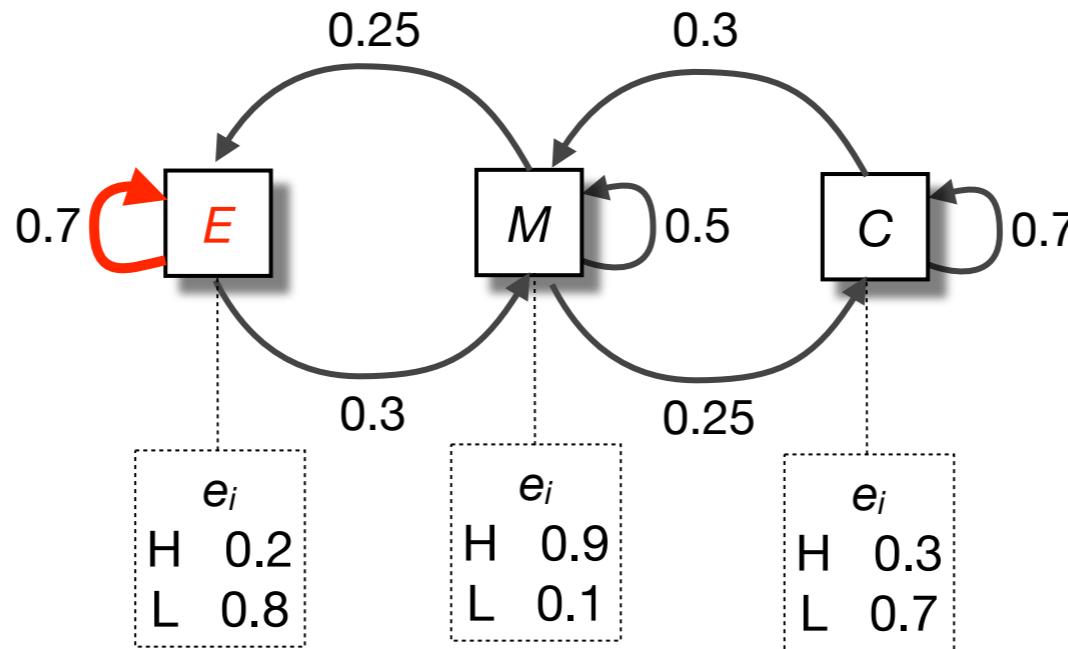
		Query Sequence				
States		H	H	L	L	H
E		$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$	$0.7 \times 0.8 \times 0.016 = 0.009$	$0.7 \times 0.2 \times 0.009 = 0.001$
M		$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.04$	$0.3 \times 0.1 \times 0.016 = 0.0005$	$0.3 \times 0.9 \times 0.009 = 0.002$
C		$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$	-	-
START		C				



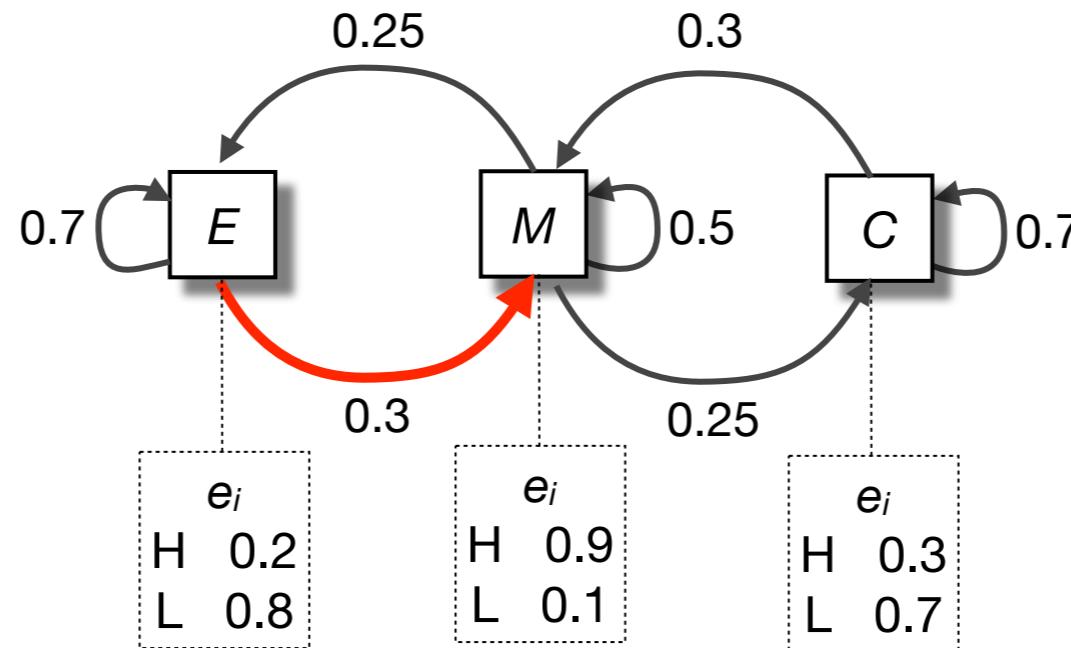
States	Query Sequence				
	H	H	L	L	H
E	$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$	$0.7 \times 0.8 \times 0.016 = 0.009$	$0.7 \times 0.2 \times 0.009 = 0.001$
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.04$	$0.3 \times 0.1 \times 0.016 = 0.0005$	$0.3 \times 0.9 \times 0.009 = 0.002$
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$	-	-
START	C	M			



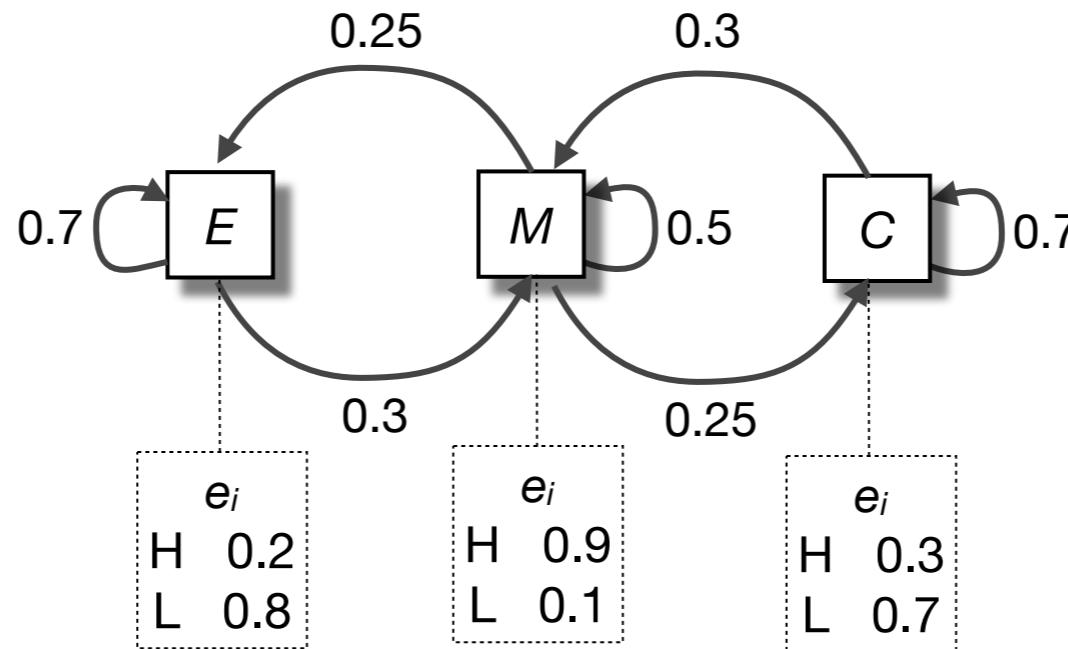
States	Query Sequence				
	H	H	L	L	H
E	$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$	$0.7 \times 0.8 \times 0.016 = 0.009$	$0.7 \times 0.2 \times 0.009 = 0.001$
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.04$	$0.3 \times 0.1 \times 0.016 = 0.0005$	$0.3 \times 0.9 \times 0.009 = 0.002$
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$	-	-
START	C	M	E		



States	Query Sequence				
	H	H	L	L	H
E	$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$	$0.7 \times 0.8 \times 0.016 = 0.009$	$0.7 \times 0.2 \times 0.009 = 0.001$
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.04$	$0.3 \times 0.1 \times 0.016 = 0.0005$	$0.3 \times 0.9 \times 0.009 = 0.002$
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$	-	-
START		C	M	E	E



Query Sequence					
States	H	H	L	L	H
E	$0 \times 0.2 = 0$	-	$0.25 \times 0.8 \times 0.081 = 0.016$	$0.7 \times 0.8 \times 0.016 = 0.009$	$0.7 \times 0.2 \times 0.009 = 0.001$
M	$0 \times 0.9 = 0$	$0.3 \times 0.9 \times 0.3 = 0.081$	$0.5 \times 0.1 \times 0.081 = 0.04$	$0.3 \times 0.1 \times 0.016 = 0.0005$	$0.3 \times 0.9 \times 0.009 = 0.002$
C	$1 \times 0.3 = 0.3$	$0.7 \times 0.3 \times 0.3 = 0.063$	$0.25 \times 0.7 \times 0.081 = 0.014$	-	-
START		C	M	E	M



States	Query Sequence				
	H	H	L	L	H
E	0x0.2 =0	-	0.25x0.8x0.081 =0.016	0.7x0.8x0.016 =0.009	0.7x0.2x0.009 =0.001
M	0x0.9 =0	0.3x0.9x0.3 =0.081	0.5x0.1x0.081 =0.04	0.3x0.1x0.016 =0.0005	0.3x0.9x0.009 =0.002
C	1x0.3 =0.3	0.7x0.3x0.3 =0.063	0.25x0.7x0.081 =0.014	-	-
START	C	M	E	E	M
	Most Probable State Sequence				

Viterbi Algorithm

Training set

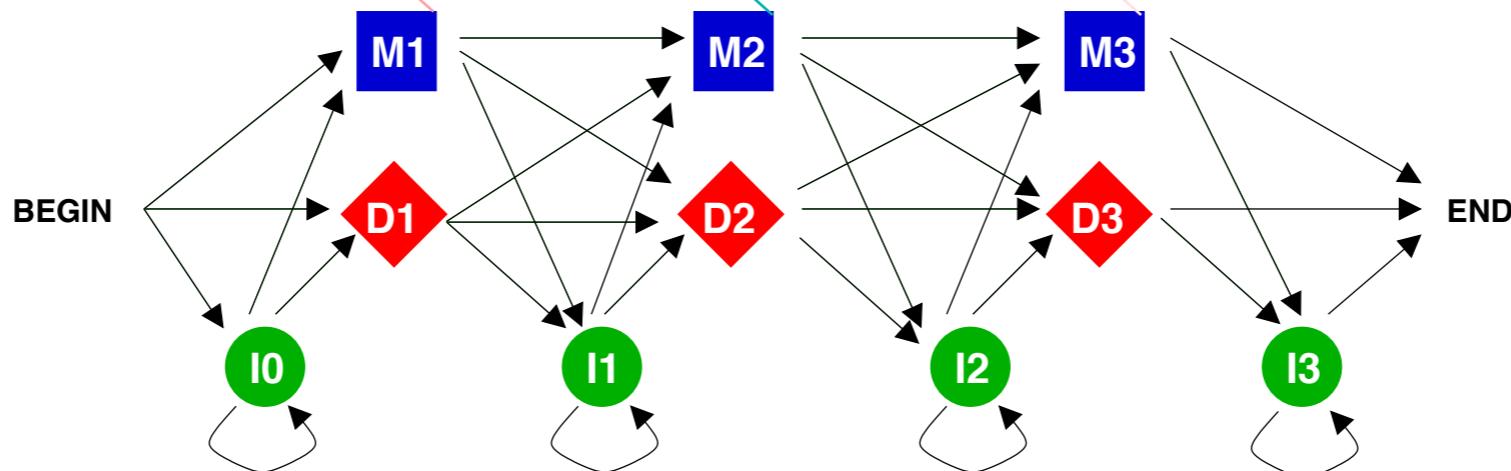
-	A	D	T	C
W	A	E	-	C
-	V	E	-	C
-	A	D	-	C
-	A	E	-	C

HMM model

A 0.74
C 0.01
D 0.03
E 0.03
...

A 0.01
C 0.01
D 0.41
E 0.44
...

A 0.01
C 0.92
D 0.01
E 0.01
...



Hidden Markov Model

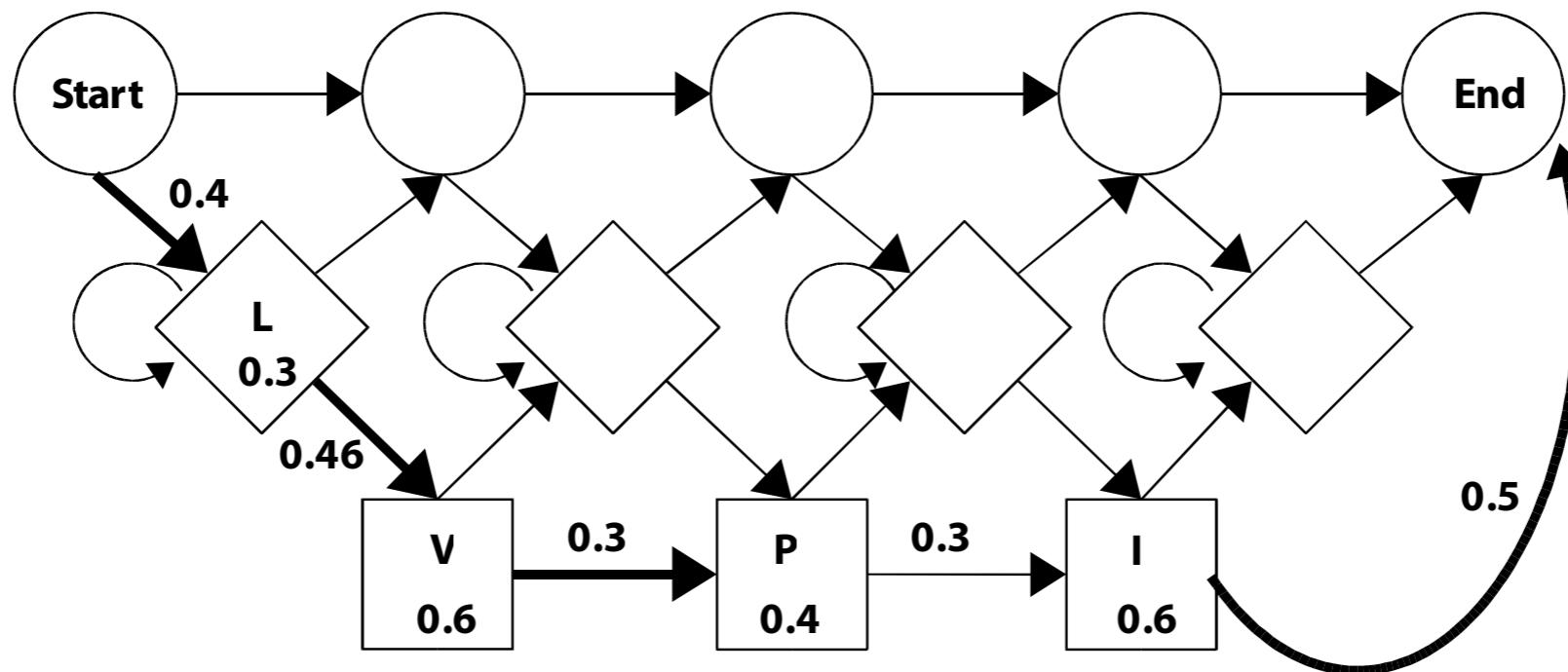


Figure 6: A possible hidden Markov model of protein LVPI. The numbers in the box indicates the emission probabilities and numbers next to arrows indicate transition probabilities. The probability of the protein LVPI is show in bold.

HMMER3

<http://hmmer.janelia.org>

cd ~/Desktop/h<tab>

cd binaries

sudo cp * /usr/bin/

Creating a HMM model of p53

Align:

```
muscle -stable -in infile -out outfile
```

Create HMM:

```
hmmbuild --informat afa p53.hmm  
outfile
```

Search human genome:

```
hmmsearch -o hits.txt p53.hmm  
human.faa
```

HMMER result

```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.0 (March 2010); http://hmmer.org/
# Copyright (C) 2010 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# -----
# query HMM file: PF00870.hmm
# target sequence database: PF00870_full_length_sequences-1.fasta
# output directed to file: result.out
# -----
```

Query: PF00870 [M=612]

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	Description	
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
6e-226	746.2	22.8	7.3e-226	745.9	15.8	1.0	1	P63_MOUSE	(088898)	
7.7e-226	745.8	21.8	9.7e-226	745.5	15.1	1.0	1	P63_RAT	(Q9JJP6)	
1.7e-225	744.7	4.7	3.5e-225	743.6	3.2	1.5	1	P73_HUMAN	(015350)	
1.6e-224	741.5	23.2	2e-224	741.2	16.1	1.0	1	P63_HUMAN	(Q9H3D4)	
2e-223	737.9	20.5	2.2e-223	737.7	14.2	1.0	1	Q3UVI3_MOUSE	(Q3UVI3)	
1.5e-222	735.0	3.4	4.3e-222	733.4	2.3	1.6	1	P73_CERAE	(Q9XSK8)	
2.1e-222	734.5	20.2	2.3e-222	734.3	14.0	1.0	1	Q5CZX0_MOUSE	(Q5CZX0)	
2.1e-221	731.1	34.0	2.4e-221	731.0	23.6	1.0	1	C4Q601_SCHMA	(C4Q601)	

PFAM readymade HMM library

The screenshot shows a web browser window displaying the Pfam 24.0 homepage. The address bar shows the URL <http://pfam.sanger.ac.uk/>. The page header includes the Wellcome Trust Sanger Institute logo and the Pfam logo with a search bar. The main content area is titled "Pfam 24.0 (October 2009, 11912 families)". It provides an overview of the database, mentioning multiple sequence alignments and hidden Markov models (HMMs). Below this, there are sections for "QUICK LINKS" (SEQUENCE SEARCH, VIEW A PFAM FAMILY, VIEW A CLAN, VIEW A SEQUENCE, VIEW A STRUCTURE, KEYWORD SEARCH) and "YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS..." (Analyze your protein sequence for Pfam matches, View Pfam family annotation and alignments, See groups of related families, Look at the domain organisation of a protein sequence, Find the domains on a PDB structure, Query Pfam by keywords). A "JUMP TO" search bar is also present. At the bottom, there is a section for "Recent Pfam blog posts" featuring links to "Job opportunities and staff changes at Xfam" (posted 1 September 2010) and "Naming by numbers" (posted 21 July 2010), along with a "Hide this" link.

Pfam: Home page (Build 20101203074432)

wellcome trust
sanger institute

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam
keyword search Go

Pfam 24.0 (October 2009, 11912 families)

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

QUICK LINKS

- SEQUENCE SEARCH** Analyze your protein sequence for Pfam matches
- VIEW A PFAM FAMILY** View Pfam family annotation and alignments
- VIEW A CLAN** See groups of related families
- VIEW A SEQUENCE** Look at the domain organisation of a protein sequence
- VIEW A STRUCTURE** Find the domains on a PDB structure
- KEYWORD SEARCH** Query Pfam by keywords

JUMP TO [Go](#) [Example](#)

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

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

Recent Pfam blog posts [Hide this](#)

[Job opportunities and staff changes at Xfam](#) (posted 1 September 2010)

We have been very sad to see a few people leave the group recently. Rob Finn has been the dedicated and hard working project leader of Pfam for many years. In fact as a summer student he is credited with preparing most of the families for Pfam 2.0 [1]! We're expecting to see great things [...]

[Naming by numbers](#) (posted 21 July 2010)

A user recently asked us why two highly similar sequences that contain a PAS domain are in different Pfam families within the PAS clan. The PAS domain clan (CL0183) currently contains seven different families: PAS, PAS_2, PAS_3, etc up to PAS_6, as well as the MEKHLA family. We thought we would take the opportunity to [...]