7.91 / 20.490 / 6.874 / HST.506

7.36 / 20.390 / 6.802

C. Burge Lecture #10

March 11, 2014

Markov & Hidden Markov Models of Genomic & Protein Features

Modeling & Discovery of Sequence Motifs

- Motif Discovery with Gibbs Sampling Algorithm
- Information Content of a Motif
- Parameter Estimation for Motif Models (+ others)

Relative Entropy*

Relative entropy, D(p||q) = mean bit-score: $\sum_{k=1}^{n} p_k \log_2 (\frac{p_k}{q_k})$

If $q_k = \frac{1}{4^w}$ *then* mean bit-score = RelEnt = $2w - H_{motif} = I_{motif}$

RelEnt is a measure of **information**, not entropy/uncertainty.
In general RelEnt is different from $H_{before} - H_{after}$ and is a better measure when background is non-random

Example: $q_A = q_T = 3/8$, $q_C = q_G = 1/8$

Suppose: $p_C = 1$. $H(q) - H(p) < 2$

But RelEnt $D(p||q) = \log_2(1/(1/8)) = 3$ bits

Which one better describes frequency of C in background seq?

* Alternate names: “Kullback-Leibler distance”, “information for discrimination”

Position-specific probability matrix (PSPM)

5' Splice Site Motif:

Pos	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	**0.3**	**0.6**	0.1	0.0	0.0	**0.4**	**0.7**	0.1	0.1
C	**0.4**	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.2
G	0.2	0.2	**0.8**	**1.0**	0.0	**0.4**	0.1	**0.8**	0.2
T	0.1	0.1	0.1	0.0	**1.0**	0.1	0.1	0.0	**0.5**

Ex: **TAGGTCAGT**

$S = S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_9$

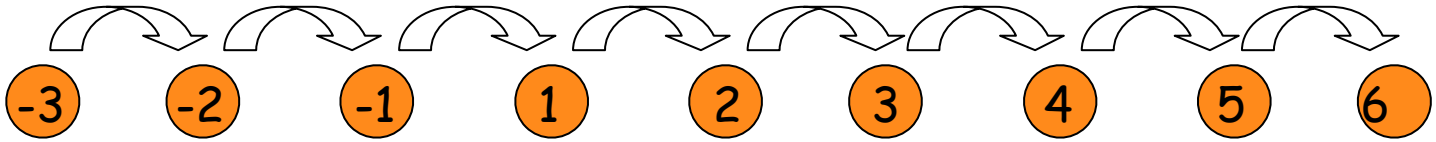
$P(S|+) = P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3) \cdots P_5(S_8)P_6(S_9)$

'Inhomogeneous', assumes independence between positions

What if this is not true?

Inhomogeneous 1st-Order Markov Model

$$P_{-2}(A \mid C) = \frac{N_{CA}^{(-3,-2)}}{N_{C}^{(-3)}}$$



$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

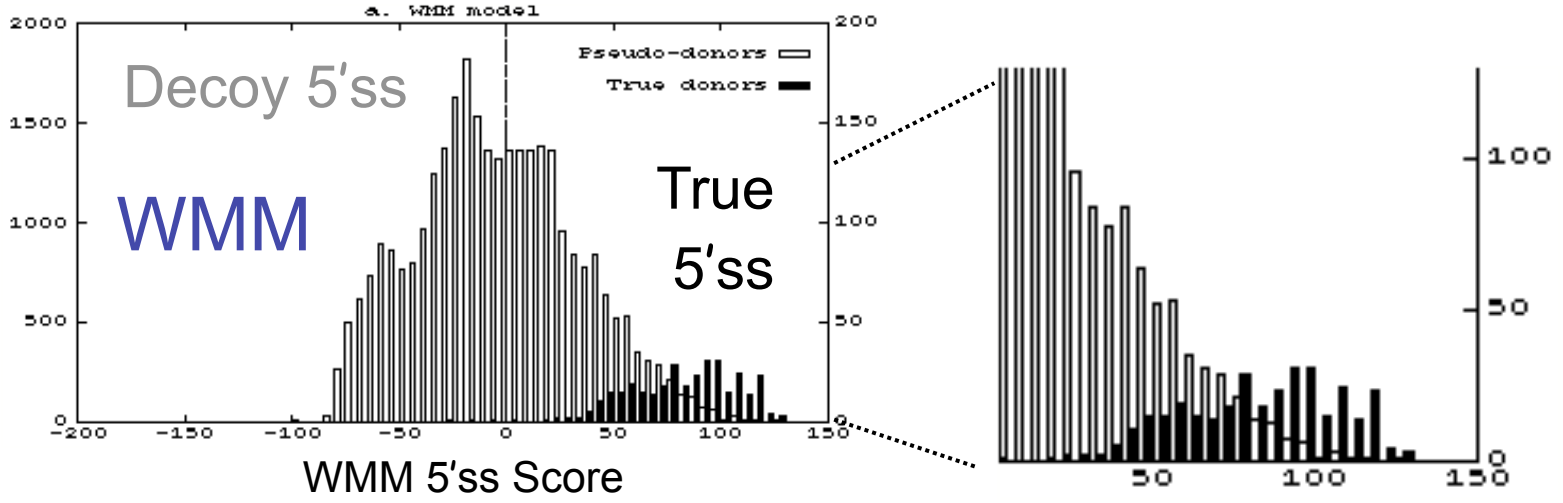
Inhomogeneous

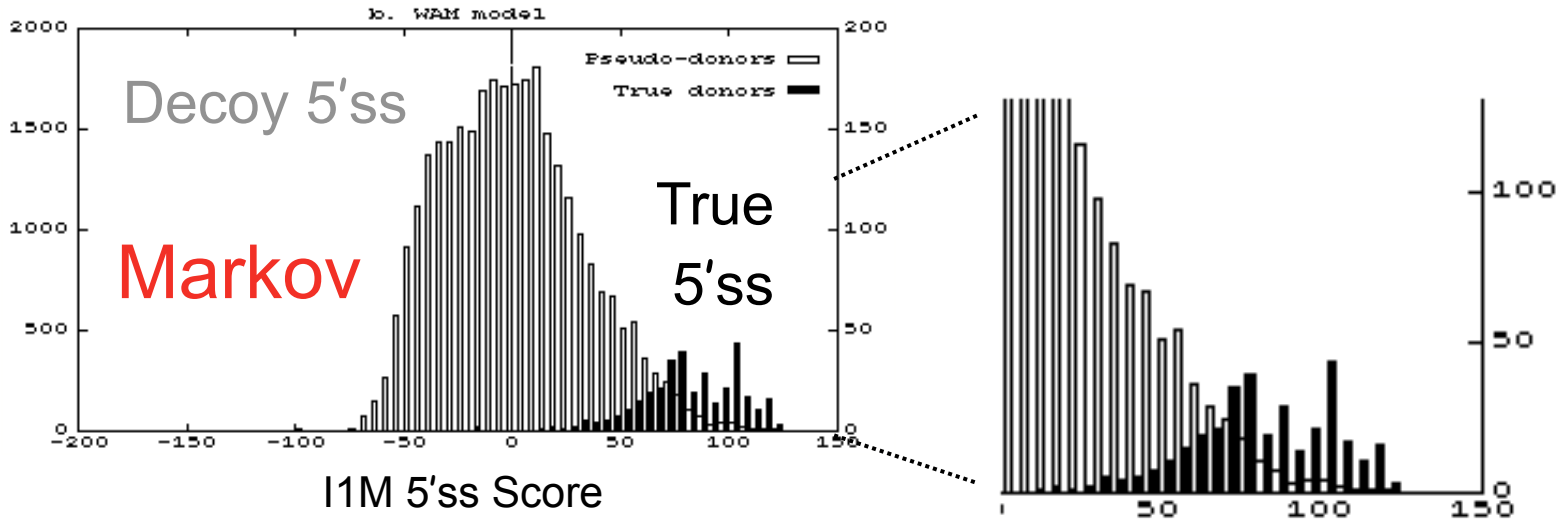
$$R = \frac{P(S \mid +)}{P(S \mid -)} = \frac{P_{-3}(S_1)P_{-2}(S_2 \mid S_1)P_{-1}(S_3 \mid S_2) \cdots P_{6}(S_9 \mid S_8)}{P_{bg}(S_1)P_{bg}(S_2 \mid S_1)P_{bg}(S_3 \mid S_2) \cdots P_{bg}(S_9 \mid S_8)}$$

Homogeneous

$$s = \log_2 R$$

WMM vs 1st-order Markov Models of Human 5′ss

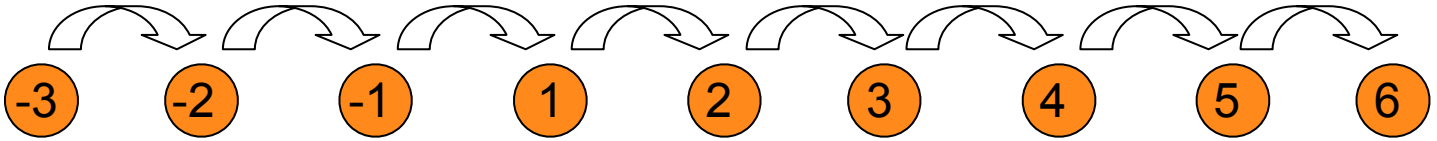




Markov models also improve modeling of transcriptional motifs - Zhou & Liu Bioinformatics 2004

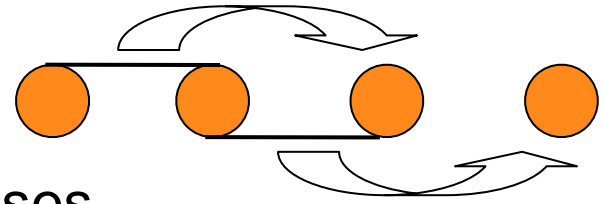
Estimating Parameters for a Markov Model

$$P_{-2}(A \mid C) = \frac{N_{CA}^{(-3,-2)}}{N_{C}^{(-3)}}$$



What about longer-range dependence?

- k-order Markov model
  - next base depends on previous k bases



2nd-order Markov model

Parameters per position for Markov model of order k: $\sim 4^{k+1}$

Dealing With Limited Training Sets

Position:	1	2	3	4	5
A	8				
C	1				
G	1				
T	0				

If the true frequency of T at pos. 1 was 10%, what's the probability we wouldn't see any Ts in a sample of 10 seqs?

$P(N=0) = (10!/0!10!)(0.1)^0(0.9)^{10} = \sim 35\%$

Motivates adding "pseudocounts"

Training Set

ACCTG
AGCTG
ACCCG
ACCTG
ACCCA
GACTG
ACGTA
ACCTG
CCCCG
ACATC

Pseudocounts (Ψcounts)

Nt	Count	Ψcount	Bayescount	ML est.	Bayes est.
A	8	+ 1	9	0.80	0.64
C	1	+ 1	2	0.10	0.14
G	1	+ 1	2	0.10	0.14
T	0	+ 1	1	0.00	0.07
	10		14	1.00	1.00

ML = maximum likelihood (of generating the observed data)

Bayes est. = Bayesian posterior relative to Dirichlet prior

Good treatment of this in appendix of:
Biological Sequence Analysis by Durbin, Eddy, Krogh, Mitchison
See also: Probability and Statistics Primer (under Materials > Resources)

Hidden Markov Models of Genomic & Protein Features

- Hidden Markov Model terminology
- Viterbi algorithm
- Examples
  - CpG Island HMM
  - TMHMM (transmembrane helices)

Background reading for today's lecture:

NBT Primer on HMMs, Z&B Chapter 6, Rabiner tutorial on HMMs

For Thursday's lecture:

NBT Primer on RNA folding, Z&B Ch. 11.9

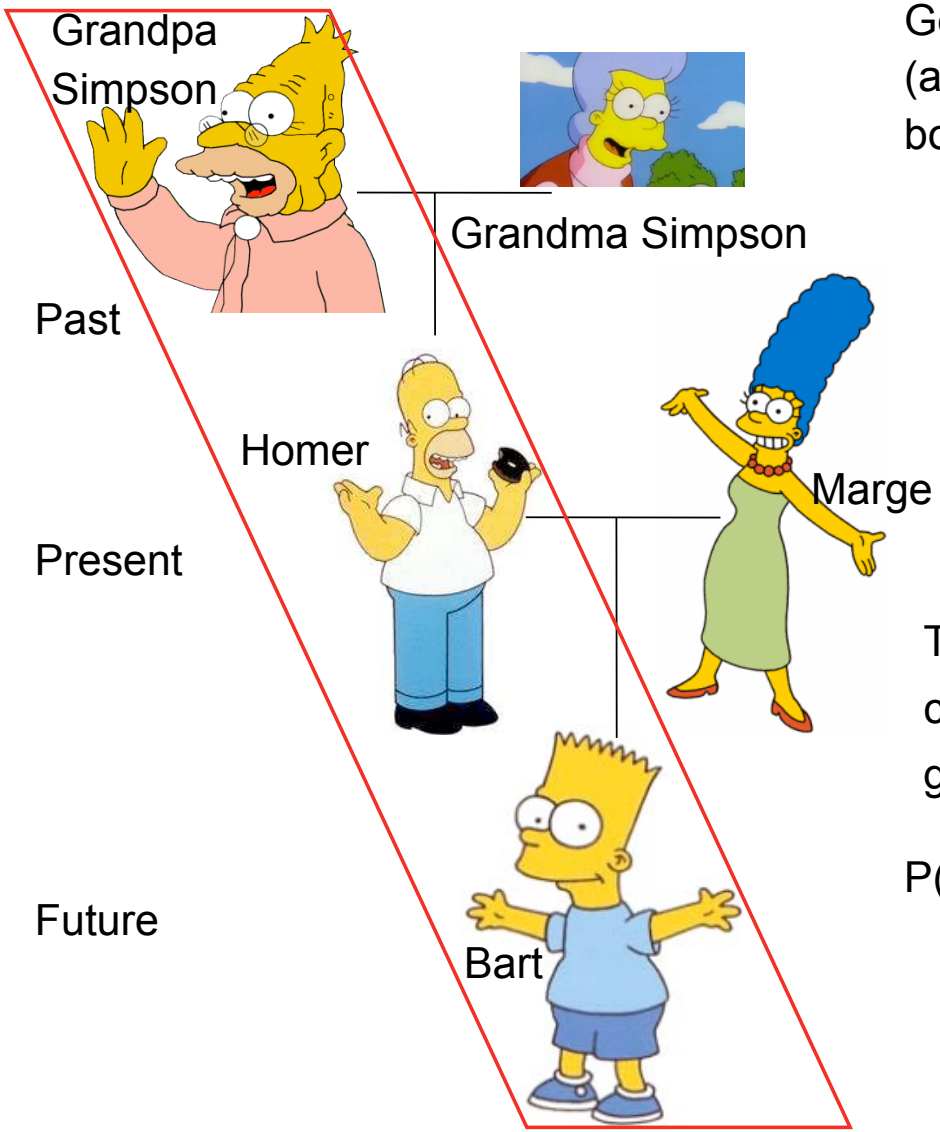
Hidden Markov Models (HMMs)

- Provide a foundation for probabilistic models of linear sequence 'labeling' problems
- Can be designed just by drawing a graph diagram
- The 'Legos' of computational sequence analysis

Developed in Electrical Engineering for applications to voice recognition

Read Rabiner's "Tutorial on hidden Markov models with applications ..."

Markov Model Example

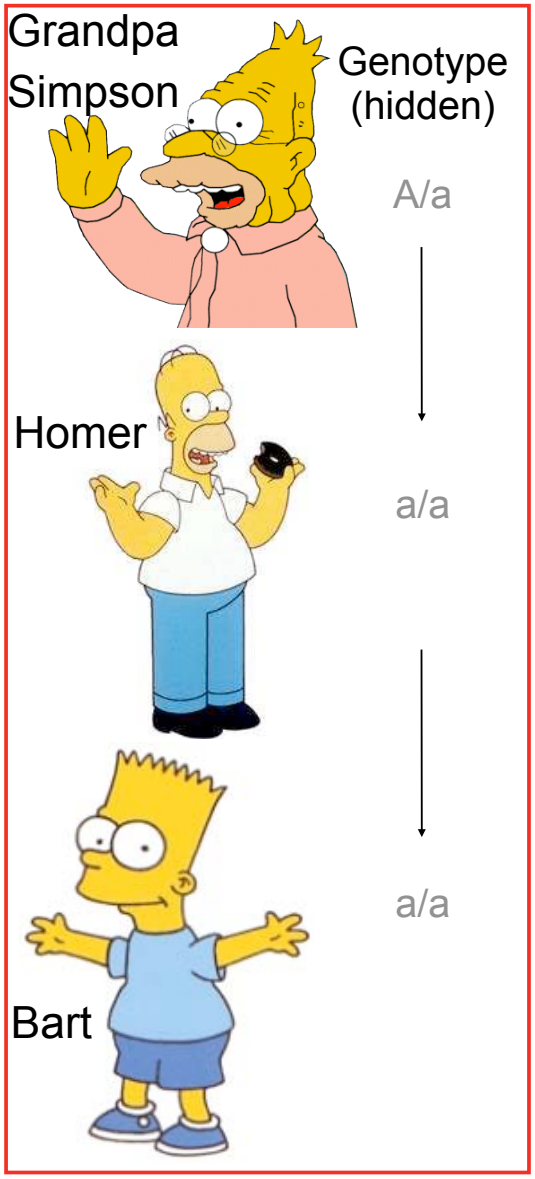


Genotype at the Apolipoprotein locus (alleles A and a) in successive generations of boxed Simpson lineage forms a Markov model

This is because, e.g., Bart's genotype is conditionally independent of Grandpa Simpson's genotype given Homer's genotype:

P(Bart = a/a | Grandpa = A/a & Homer = a/a)
= P(Bart = a/a | Homer = a/a)

Hidden Markov Model Example



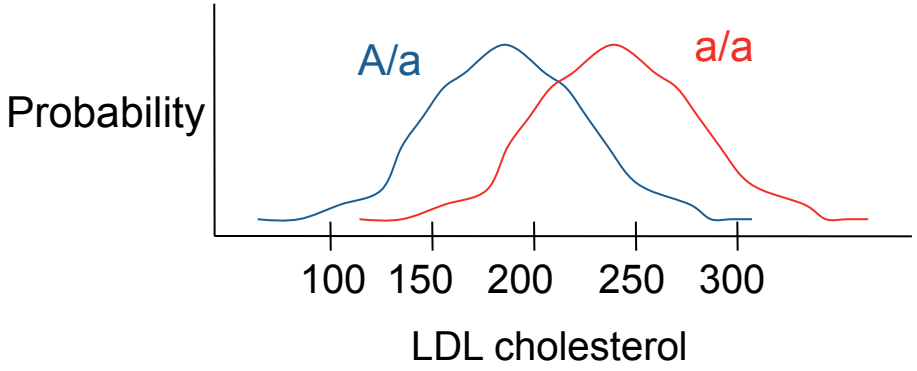
Phenotype -
LDL cholesterol
(observed)

150

250

200

Suppose that we can't observe genotype directly, only some phenotype related to the A locus, and this phenotype depends probabilistically on the genotype. **Then we have a Hidden Markov Model.**



HMMs as Generative Models

An HMM can be used as a generator to give an observation sequence

$$O = O_1 O_2 \cdots O_T \tag{10}$$

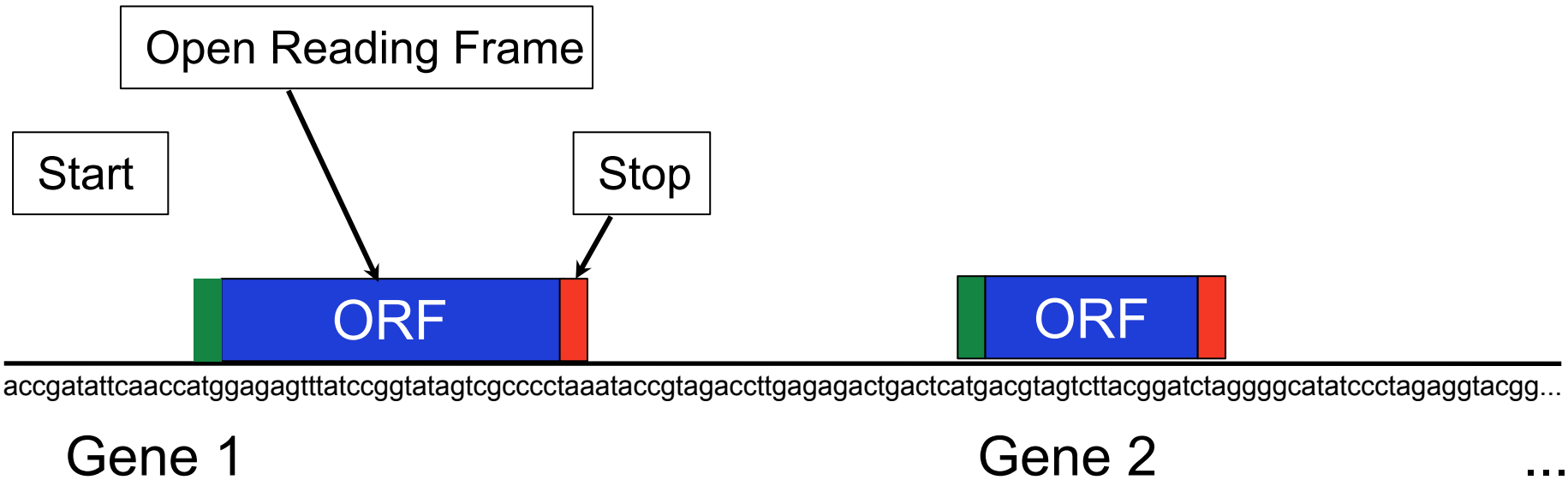
(where each observation O_t is one of the symbols from V, and T is the number of observations in the sequence) as follows:

1) Choose an initial state $q_1 = S_i$ according to the initial state distribution π.
2) Set $t = 1$.
3) Choose $O_t = v_k$ according to the symbol probability distribution in state S_i, i.e., $b_i(k)$.
4) Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i, i.e., a_{ij}.
5) Set $t = t + 1$; return to step 3)if $t < T$; otherwise terminate the procedure.

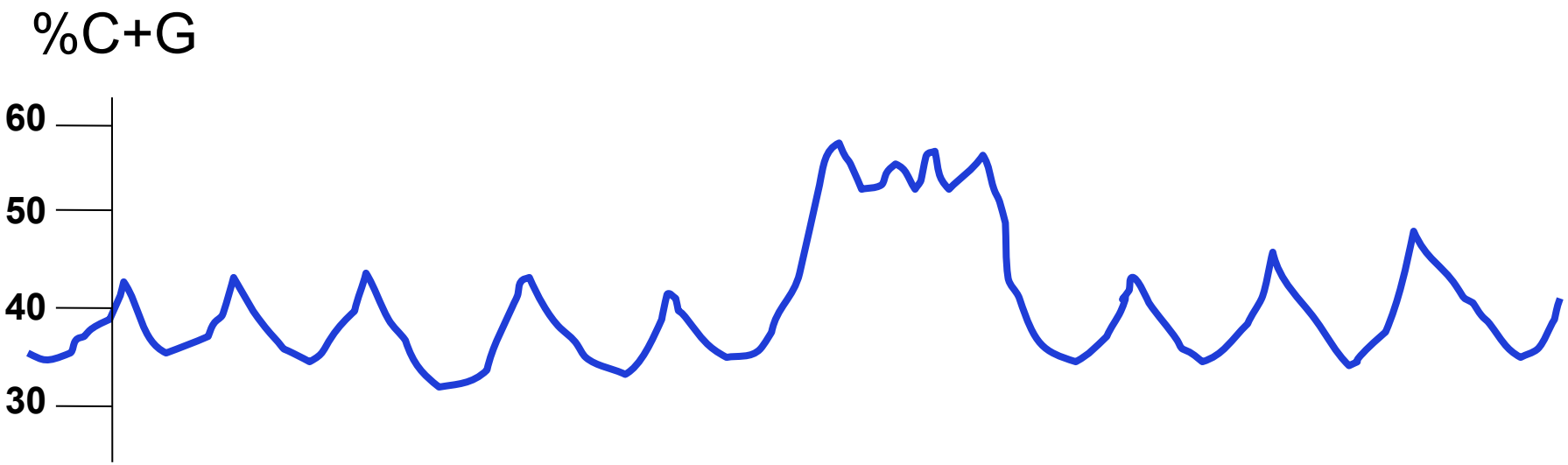
From Rabiner Tutorial

“Sequence Labeling” Problems

Example: Bacterial gene finding

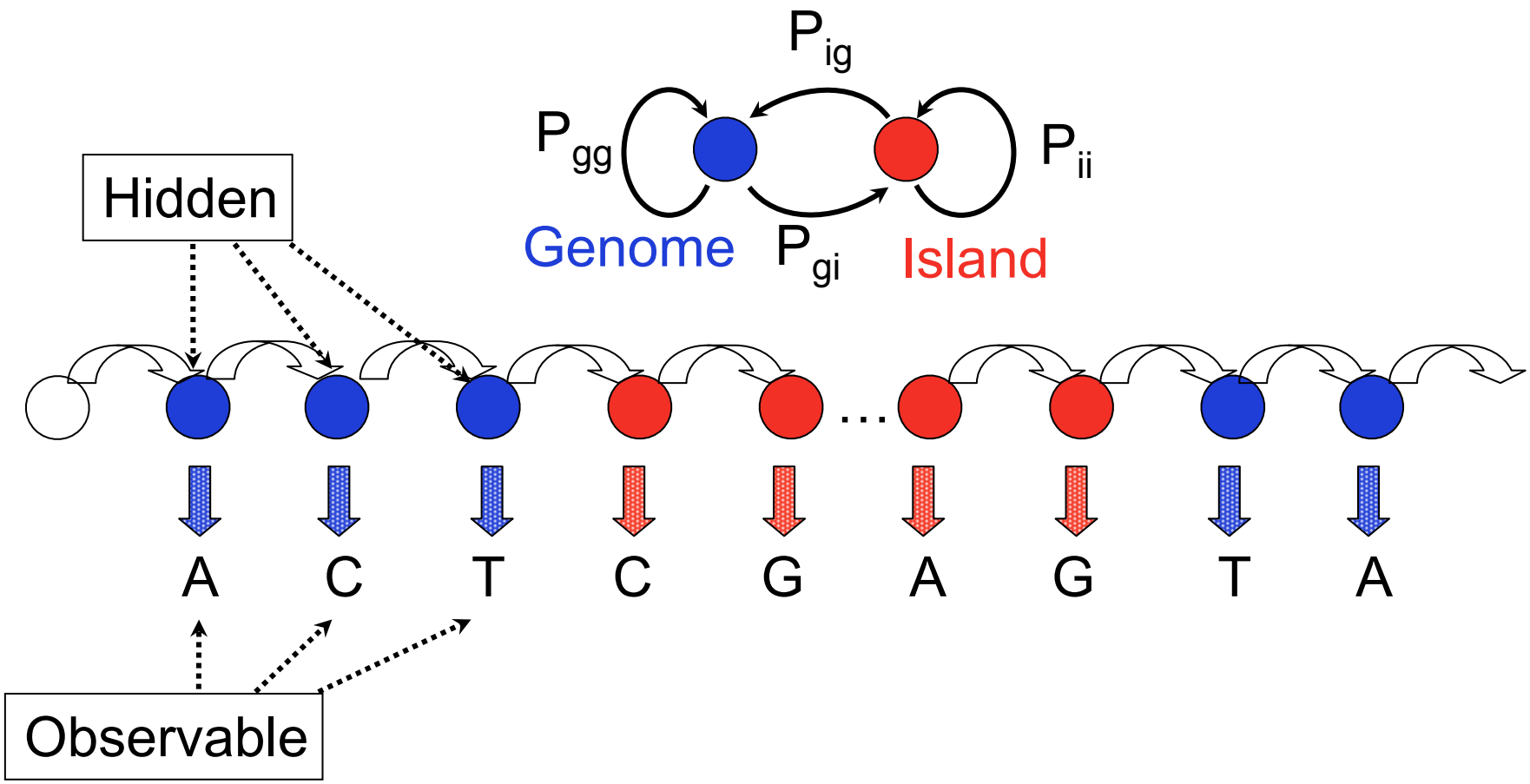


CpG Islands



- Regions of high C+G content and relatively high abundance of CpG dinucleotides (normally rare) which are unmethylated
- Associated with promoters of many human genes (~ 1/2)

CpG Island Hidden Markov Model



CpG Island HMM

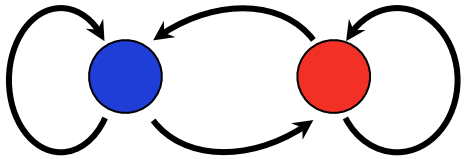
Rabiner notation

"Initiation probabilities" π_j

$P_g = 0.99$, $P_i = 0.01$

$P_{gg} = 0.99999$ $P_{ig} = 0.001$

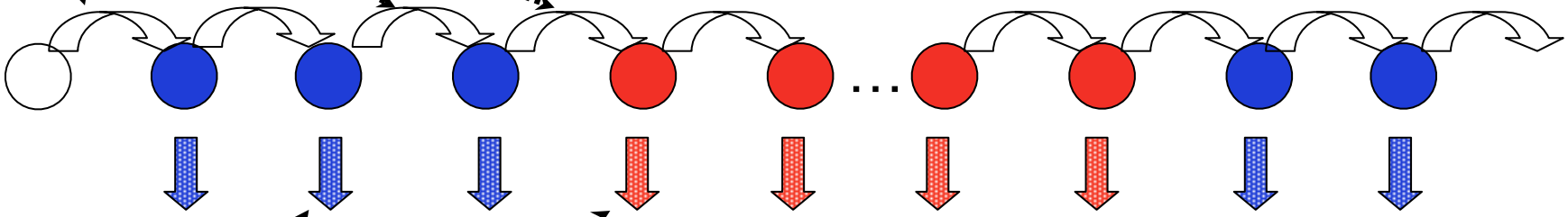
Genome



$P_{ii} = 0.999$

"Transition probabilities" a_{ij}

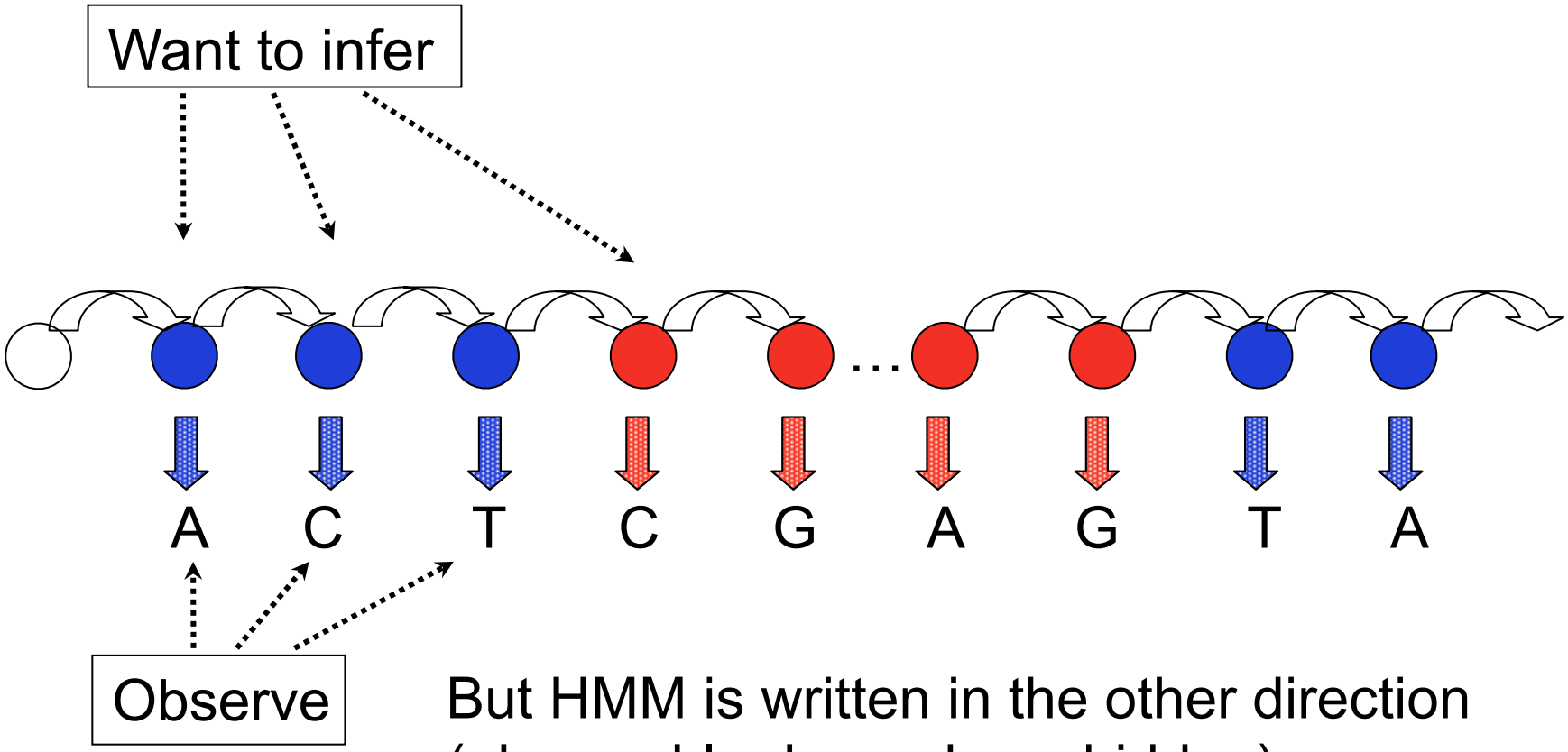
$P_{gi} = 0.00001$ Island



"Emission Probabilities" $b_j(k)$

	C	G	A	T
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

CpG Island HMM III



But HMM is written in the other direction
(observable depends on hidden)

Reversing the Conditioning (Bayes' Rule)

Definition of Conditional Probability:
$P(A|B) = P(A,B) / P(B)$

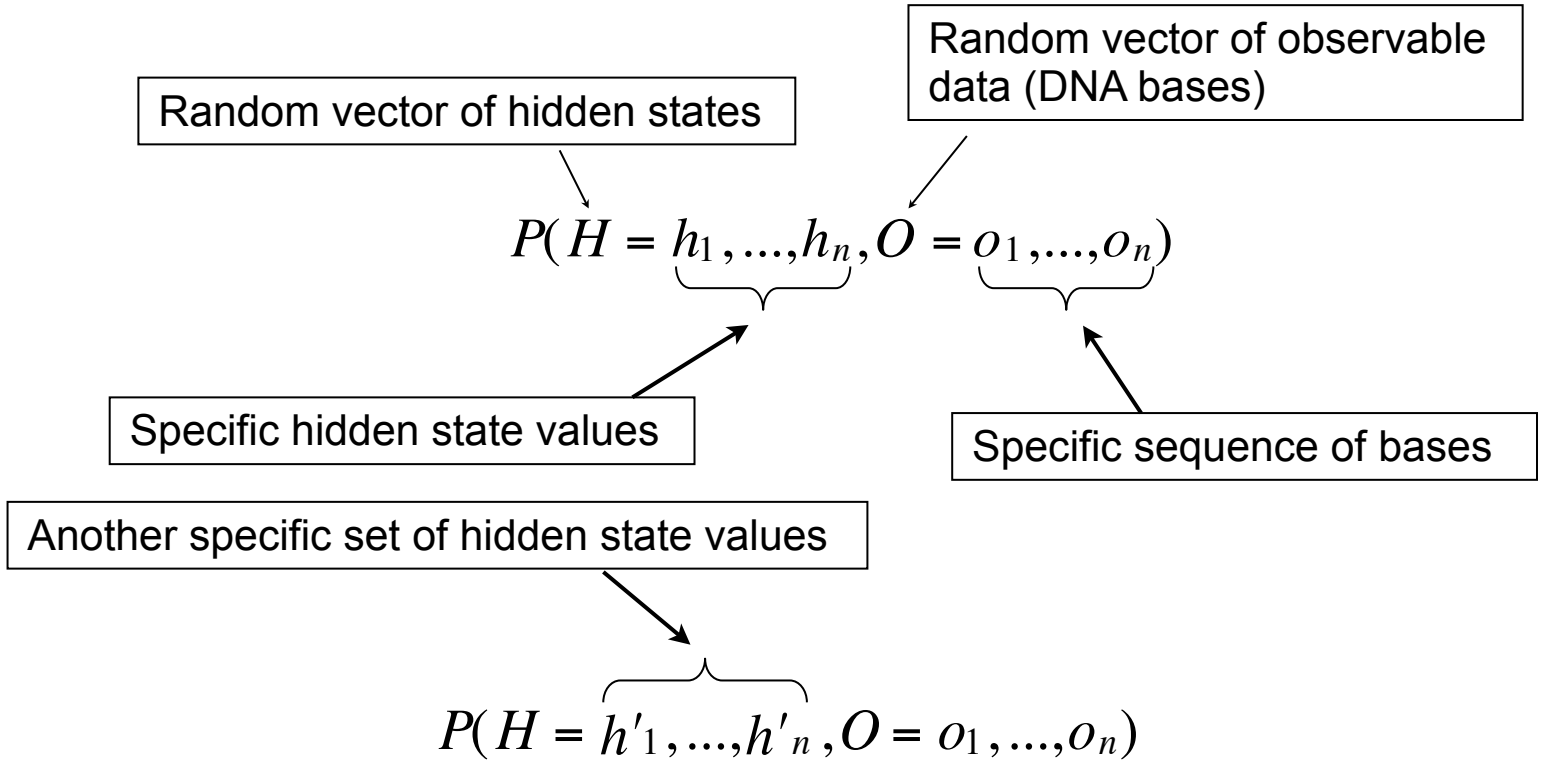
Bayes' Rule (simple form)

$$P(B|A) = P(B)P(A|B) / P(A)$$

Bayes' Rule (more general form)

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_k P(B_k)\, P(A|B_k)}$$

Notation for HMM Calculations



Reversing the Hidden/Observable Conditioning (Bayes' Rule)

$$P(H = h_1, h_2, \ldots, h_n \mid O = o_1, o_2, \ldots, o_n)$$

Conditional Prob:
P(A|B) = P(A,B)/P(B)

$$= \frac{P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)}{P(O = o_1, \ldots, o_n)}$$

$$= \frac{P(H = h_1, \ldots, h_n) P(O = o_1, \ldots, o_n \mid H = h_1, \ldots, h_n)}{P(O = o_1, \ldots, o_n)}$$

$P(O = o_1, \ldots, o_n)$ a bit tricky to calculate, but is independent of $h_1, \ldots, h_n$ so can treat as a constant and simply maximize

$$P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)$$

Inferring the Hidden from the Observable (Viterbi Algorithm)

Want to find sequence of hidden states $H^{opt} = h_1^{opt}, h_2^{opt}, h_3^{opt}, \ldots$

that maximizes joint probability: $P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)$

(optimal "parse" of sequence)

Solution:

Define $R_i^{(h)}$ = probability of optimal parse of the subsequence 1..i ending in state h

Solve recursively, i.e. determine $R_2^{(h)}$ in terms of $R_1^{(h)}$, etc.

Andrew Viterbi, an MIT BS/MEng student in E.E. - founder of Qualcomm

CpG Island HMM

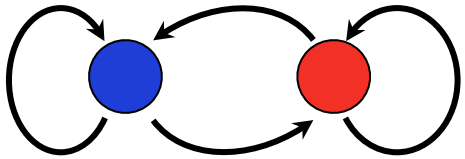
Rabiner notation

"Initiation probabilities" π_j

$P_g = 0.99$, $P_i = 0.01$

$P_{gg} = 0.99999$ $P_{ig} = 0.001$

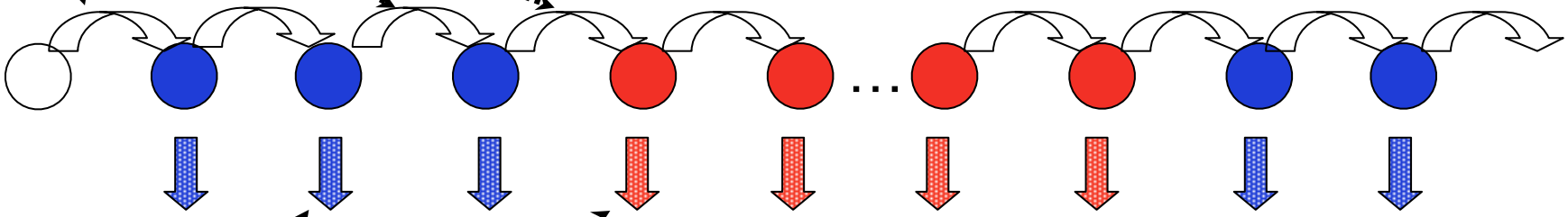
Genome



$P_{ii} = 0.999$

"Transition probabilities" a_{ij}

$P_{gi} = 0.00001$ Island



"Emission Probabilities" $b_j(k)$

	C	G	A	T
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

$\delta_t(i)$ probability of optimal parse of the subsequence 1..t ending in state i

$\psi_t(i)$ the state at t-1 that resulted in the optimal parse of 1..t ending in i

N no. of states

T length of sequence

Viterbi Algorithm

1) Initialization:

$$\delta_1(i) = \pi_i b_i(O_1), \qquad 1 \le i \le N \tag{32a}$$

$$\psi_1(i) = 0. \tag{32b}$$

2) Recursion:

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(O_t), \qquad 2 \le t \le T, \quad 1 \le j \le N \tag{33a}$$

$$\psi_t(j) = \operatorname*{argmax}_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}], \qquad 2 \le t \le T, \quad 1 \le j \le N. \tag{33b}$$

3) Termination: (34a)

$$P^* = \max_{1 \le i \le N} [\delta_T(i)]$$

$$q_T^* = \operatorname*{argmax}_{1 \le i \le N} [\delta_T(i)]. \tag{34b}$$

4) Path (state sequence) backtracking:

$$q_t^* = \psi_{t+1}(q_{t+1}^*), \qquad t = T-1, T-2, \cdots, 1. \tag{35}$$

Rabiner 1989

Viterbi Example

ACG

More Viterbi Examples

What is the optimal parse of the sequence for the CpG island HMM defined previously?

- $(ACGT)_{10000}$
- $A_{1000}C_{80}T_{1000}C_{20}A_{1000}G_{60}T_{1000}$

<u>Powers of 1.5:</u>

N =	20	40	60	80
$(1.5)^N$ =	$3x10^3$	$1x10^7$	$3x10^{10}$	$1x10^{14}$

Run time for k-state HMM on sequence of length L?

$O(k^2L)$

The computational efficiency of the Viterbi algorithm is a major reason for the popularity of HMMs

Midterm Logistics

Midterm 1 is **Tuesday, March 18th during regular class time/room***

Will start promptly at 1:05pm and end at 2:25pm - arrive in time to get settled

***except for 6.874 students who will meet at 12:40 PM.**

Closed book, open notes:

- you may bring **up to two pages** (double-sided) of notes if you wish

No calculators or other electronic aids (you won't need them anyway)

Study lecture notes, readings/tutorials and past exams/Psets 1st, textbook 2nd

Midterm exams from previous years are posted on course web site

Note: there is some variation in topics from year to year

Midterm 1

Exam will cover course topics from Topics 1, 2 and 3 through Hidden Markov Models (but will NOT cover RNA Secondary Structure)

R Feb 06 CB L2 DNA Sequencing, Local Alignment (BLAST) and Statistics
T Feb 11 CB L3 Global Alignment of Protein Sequences
R Feb 13 CB L4 Comparative Genomic Analysis of Gene Regulation
R Feb 20 DG L5 Library complexity and BWT
T Feb 25 DG L6 Genome assembly
R Feb 27 DG L7 ChIP−Seq analysis (DNA−protein interactions)
T Mar 04 DG L8 RNA−seq analysis (expression, isoforms)
R Mar 06 CB L9 Modeling & Discovery of Sequence Motifs
T Mar 11 CB L10 Markov & Hidden Markov Models (+HMM content on 3/13)

Exam may have some overlap with topics from Pset 1+2 but will be biased towards topics NOT covered on PSets

There may be questions on algorithms, but none related to python or programming

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014