

Today

- Midterm
- Project proposals are due Thursday
 - A brief review of citation practices
- Introduction to Hidden Markov Models

Midterm grades

Midterm exam distribution:

Max: 91%, Min: 63%, Mean: 78.5%

Total midterm score = $(\sum_i PS(i) + 0.6 * \text{midterm score})/140$

Midterm score distribution:

Max: 94%, Min: 75%, Mean: 85%

A: > 90%

B: 80% - 90%

B-, C+: < 80%

Midterm

1a) As a preprocessing step for sequence assembly, it is necessary to identify pairs of DNA sequence fragments that overlap. Let T and U be two sequence fragments. We wish to determine whether the end of T overlaps with the beginning of U .

- Semiglobal alignment
- Can use similarity or distance scoring

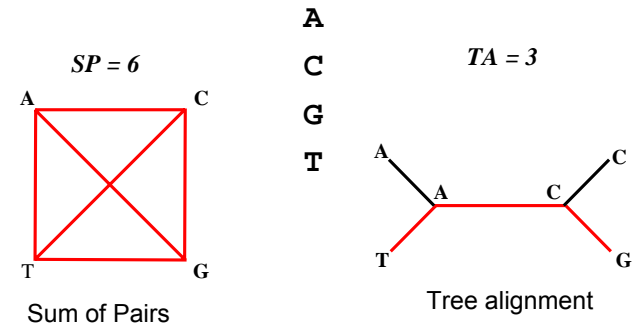
1b) *Given the operations insertion, deletion and substitution, the problem is to determine the minimum number of operations required to transform T into U .*

- Global alignment
- To count operations, you *must* use edit distance:
 - $s[i,j] = 1$, if $i \neq j$,
 - $s[i,i] = 0$,
 - $g = 1$

1c) T and U are genomic DNA sequences containing the albumin gene, in mouse and human, respectively. We wish to find the intron/exon boundaries of the albumin gene in the genomic DNA.

- Local alignment
- Constraints:
 - $s[i,j]$ must be a similarity function
 - $s[i,i] > s[i,j] > 2g$
 - Expected random alignment must be negative
 - At least one, $s[i,i]$ must be positive.

Problem 2 (a and b)



Problem 3

c) Maximum likelihood considers all possible internal labels when scoring a tree: $O(|\Sigma|^k)$. With an alphabet that is twice the size of Σ , the running time will be proportional to $|2\Sigma|^k$, i.e., increased by a factor of 2^k . This is why maximum likelihood is used more for DNA than amino acid sequences

d) Both maximum likelihood and parsimony use character data. Both maximum likelihood and distance based methods assume a model of neutral evolution.

Problem 3

c) Species: bacterium, archaebacterium and two vertebrates. Data is an ultrametric distance matrix

- Use UPGMA or midpoint rooting. Note there is no obvious out group for these species.

d) Data: immunoglobulin sequences for four old world monkeys

- Immunoglobulin sequences mutate rapidly, so use NJ or maximum likelihood. Root with an outgroup, e.g., a new world monkey or an ape sequence.

e) Data: ribosomal sequences in closely related species.

- Both maximum parsimony or maximum likelihood will infer ancestral sequences, but maximum parsimony is a better model for slowly evolving sequences in closely related species.

Problem 4

- a) The distances came from a tree, so you didn't have to check the four point condition.
- c) According to the *tree* distances, are the four taxa changing at the same rate.
The tree distances are additive, but fail the 3pt condition so they are not ultrametric. Therefore they are not changing at the same rate.
- e) Data: ribosomal sequences in closely related species.
➤ Both maximum parsimony or maximum likelihood will infer ancestral sequences, but maximum parsimony is a better model for slowly evolving sequences in closely related species.

Today

- Midterm
- Project proposals are due Thursday
 - A brief review of citation practices
- Introduction to Hidden Markov Models

Why cite?

- Citations reflect *the careful and thorough work you have put in* to locating and exploring your sources.
- Citations *help readers understand the context* of your argument, and locate your work within other conversations on your topic.
- Citations allow you to *acknowledge those authors who made possible particular aspects of your work*. Failure to provide adequate citations constitutes plagiarism.
- Citations, by delineating your intellectual debts, also *draw attention to the originality and legitimacy of your own ideas*.

<http://www.dartmouth.edu/~sources/about/what.html>

When to cite?

- Cite sources for all verbatim quotations of two or more consecutive words.
- Cite sources from which you paraphrase or summarize facts or ideas.
- Cite sources for ideas or information that could be regarded as common knowledge but which you think your reader might still find unfamiliar.

<http://www.dartmouth.edu/~sources/about/what.html>

What to cite?

Primarily: *refereed, archival* materials. (Archival materials are materials that are available in libraries or bookstores, have an ISBN number, etc.)

- Books
- Journal articles
- Refereed conference proceedings

Avoid:

- Websites, news stories, photocopied workshop handouts, personal communications.

The original text:

The main image in *Othello* is that of animals in action, preying upon one another, mischievous, lascivious, cruel or suffering, and through these, the general sense of pain and unpleasantness is much increased and kept constantly before us.

More than half the animal images in the play are lago's, and all these are contemptuous or repellent: a plague of flies, a quarrelsome dog, the recurrent image of bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, wolves, goats and monkeys¹.

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

<http://www.dartmouth.edu/~sources/about/what.html>

Students paper:

The majority of the animal images in the play are lago's, and all of these are contemptuous or repellent. He refers to a plague of flies, a quarrelsome dog, bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, goats and monkeys. Through these images the general sense of pain and unpleasantness is increased and kept constantly before us.

The original text:

The main image in *Othello* is that of animals in action, preying upon one another, mischievous, lascivious, cruel or suffering, and **through these, the general sense of pain and unpleasantness is much increased and kept constantly before us.**

More than half **the animal images in the play are lago's, and all these are contemptuous or repellent: a plague of flies, a quarrelsome dog, the recurrent image of bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, wolves, goats and monkeys¹.**

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

<http://www.dartmouth.edu/~sources/about/what.html>

Not OK:

Students paper:

The majority of **the animal images in the play are lago's, and all of these are contemptuous or repellent.** He refers to **a plague of flies, a quarrelsome dog, bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, goats and monkeys.** **Through these** images the general sense of pain and unpleasantness is increased and kept constantly before us.

The original text:

The **main image in *Othello* is that of animals in action, preying upon one another, mischievous, lascivious, cruel or suffering, and through these, the general sense of pain and unpleasantness is much increased and kept constantly before us.**

More than half the animal images in the play are lago's, and all these are contemptuous or repellent: a plague of flies, a quarrelsome dog, the recurrent image of bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, wolves, goats and monkeys¹.

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

<http://www.dartmouth.edu/~sources/about/what.html>

Not OK:

Students paper:

I believe that the main image in Shakespeare's tragedy, *Othello*, is that of animals. These creatures are constantly in action, preying upon one another, and they are depicted as mischievous, wanton, cruel or suffering. By Shakespeare's ingenious use of these animal images, the general sense of pain and unpleasantness that pervades the entire story is much increased and kept constantly before the reader.

The original text:

The main image in *Othello* is that of animals in action, preying upon one another, mischievous, lascivious, cruel or suffering, and through these, the general sense of pain and unpleasantness is much increased and kept constantly before us.

More than half the animal images in the play are Iago's, and all these are contemptuous or repellent: a plague of flies, a quarrelsome dog, the recurrent image of bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, wolves, goats and monkeys¹.

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

<http://www.dartmouth.edu/~sources/about/what.html>

Not OK:

Students paper:

In *Othello*, Shakespeare makes frequent use of animal imagery. The specific images he uses are generally distasteful and convey to the reader a constant impression of conflict and misery.

The original text:

The main image in *Othello* is that of animals in action, preying upon one another, mischievous, lascivious, cruel or suffering, and through these, the general sense of pain and unpleasantness is much increased and kept constantly before us.

More than half the animal images in the play are Iago's, and all these are contemptuous or repellent: a plague of flies, a quarrelsome dog, the recurrent image of bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, wolves, goats and monkeys¹.

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

<http://www.dartmouth.edu/~sources/about/what.html>

OK:

Students paper:

In the play, *Othello*, the character of Iago is associated with unpleasant animal imagery[1]....

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

The original text:

The main image in *Othello* is that of animals in action, preying upon one another, mischievous, lascivious, cruel or suffering, and through these, the general sense of pain and unpleasantness is much increased and kept constantly before us.

More than half the animal images in the play are Iago's, and all these are contemptuous or repellent: a plague of flies, a quarrelsome dog, the recurrent image of bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, wolves, goats and monkeys¹.

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

<http://www.dartmouth.edu/~sources/about/what.html>

Not OK:

Students paper:

The majority of "the animal images in the play are Iago's, and all of these are contemptuous or repellent". He refers to "a plague of flies, a quarrelsome dog," "bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, goats and monkeys." "Through these" images "the general sense of pain and unpleasantness is increased and kept constantly before us." [1]

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

The original text:

The main image in *Othello* is that of animals in action, preying upon one another, mischievous, lascivious, cruel or suffering, and through these, the general sense of pain and unpleasantness is much increased and kept constantly before us.

More than half the animal images in the play are Iago's, and all these are contemptuous or repellent: a plague of flies, a quarrelsome dog, the recurrent image of bird-snaring, leading asses by the nose, a spider catching a fly, beating an offenceless dog, wild cats, wolves, goats and monkeys¹.

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

<http://www.dartmouth.edu/~sources/about/what.html>

OK:

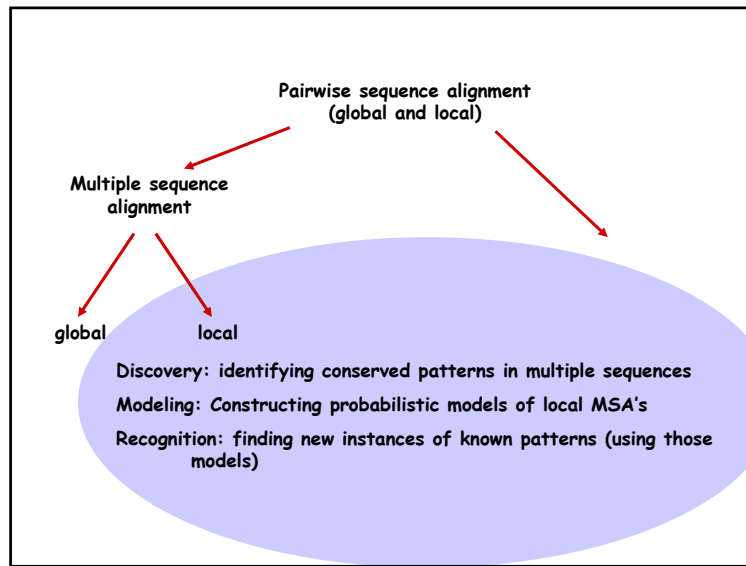
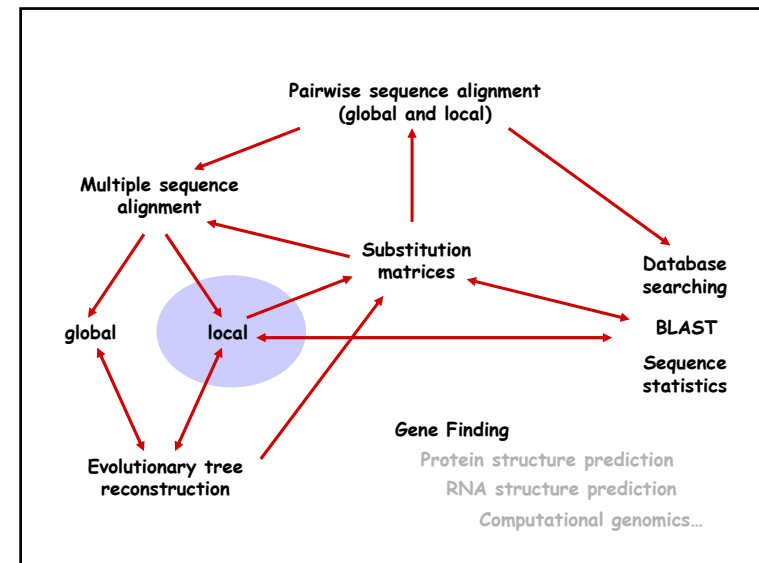
Students paper:

Caroline Spurgeon uses the words "contemptuous" and "repellent" in describing the animal imagery associated with Iago in *Othello* [1]. In my opinion, her choice of words indicates that...

1. Caroline F. E. Spurgeon, *Shakespeare's Imagery* (Cambridge: Cambridge UP, 1935) 335.

Today

- Midterm
- Project proposals are due Thursday
 - A brief review of citation practices
- Introduction to Hidden Markov Models



Local Multiple Alignment

- Position Specific Scoring Matrices (PSSMs)
 - Modeling, Recognition
- Gibbs sampler
 - Discovery
- Hidden Markov Models (HMMs)
 - Discovery, Modeling, Recognition
 - Can represent gaps, positional dependencies

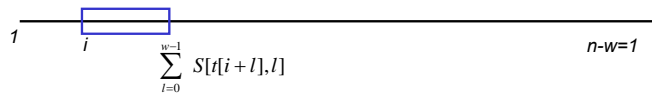
Position Specific Scoring Matrices

Given $A[l, k]$ (k sequences, l positions),

– Frequency of aa i at position j $F[i, j] = \frac{n_i}{k}$

– Propensity of aa i at position j $P[i, j] = \frac{F[i, j]}{p_i}$

– Log odds scoring matrix $S[i, j] = \log_2 P[i, j]$



Problems with PSSMs

Do not capture positional dependencies

WEIRD

WEIRD

WEIQH

WEIRD

WEIQH



D					0.60
E		1.00			
H					0.40
I			1.00		
Q				0.40	
R				0.60	
W	1.00				

Note: We never see QD or RH, only RD and QH.
But, $P(RH) = P(QD) = 0.24$, while $P(QH) = 0.16$

Problems with PSSMs

Hard to recognize pattern instances that contain indels

D	0.8	0.8	0.8	0.8	2.4
E	0.6	2.9	0.6	0.6	1.6
H	2.0	2.0	2.0	2.0	3.0
I	0.8	0.8	3.1	0.8	0.8
Q	1.1	1.1	1.1	2.1	1.1
R	0.8	0.8	0.8	2.8	0.8
W	5.0	2.7	2.7	2.7	1.8

W E T I R D

$$5.0 + 2.9 + 1.2 + 1.4 + 1.5 = 11$$

W E T I R D

$$1.2 + 1.8 + 3.1 + 3.0 + 3.4 = 12.5$$

W E T I R D

$$5.0 + 2.9 + 3.1 + 3.0 + 3.4 = 18.4$$

Problems with PSSMs

Variable length motifs

WETIRD

WE_IRD

WETIQH

WE_IRD

WETIQH

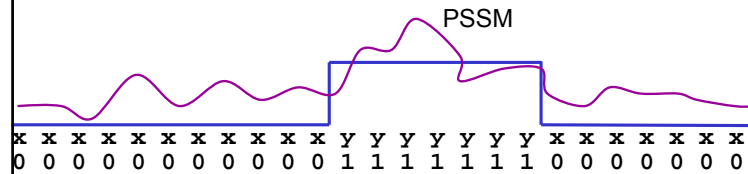
Gaps can be represented by expanding Σ , but what size window should be used to score new instances of the motif?

x x x W E T I R D x x x x x x W E I Q H x x x x x

Problems with PSSMs

Do not handle boundary detection problems well

Goal: label every element in the sequence with a zero (not in pattern) or a one (in pattern)



Examples of boundary detection problems

- Recognition of regulatory motifs
- Recognition of protein domains
- Intron/exon boundaries
- Gene boundaries
- Transmembrane regions
- Secondary structures (α helices, β sheets)

Plan

- Review Markov chains
- Extend to Hidden Markov Models
 - Boundary detection
 - Scoring sequences
- HMM construction
- Biological applications: revisit gaps and dependencies.

Markov chains

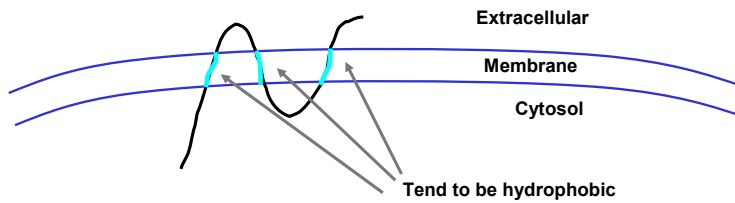
- States: S_1, S_2, \dots, S_N
- Initial distribution of states: $\pi(i) = P(q_0 = S_i)$
- Transition probabilities:
 - $a_{ij} = P(q_t = S_j | q_{t-1} = S_i)$

Questions we can ask:

What is the probability of being in a particular state at a particular time?

What is the probability of seeing a particular sequence of states?

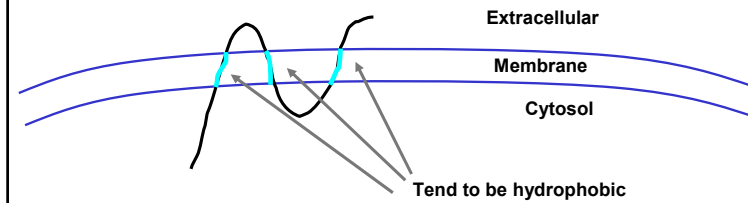
An example: transmembrane regions



Model each amino acid as hydrophobic (H) or hydrophilic (L)
 → A peptide sequence can be represented as a sequence of H's and Ls.

HHHLLHLHHLHL...

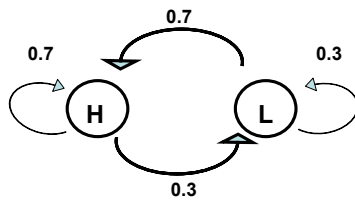
An example: transmembrane regions



A simpler question:
 is a given sequence a transmembrane sequence?

HHHLLHLHHLHL...

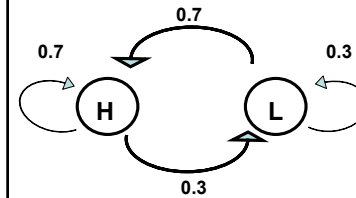
A Markov chain for recognizing transmembrane sequences



- States: S_H, S_L
- $\Sigma = \{H, L\}$
- $\pi(H) = 0.7, \pi(L) = 0.3$

Is a given sequence, say HHLHH,
 a transmembrane sequence?

Transmembrane model:



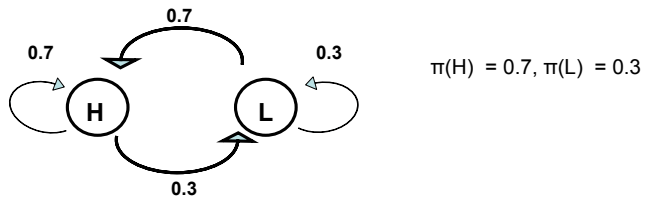
$$\pi(H) = 0.7, \pi(L) = 0.3$$

$$P(\text{HHLHH}) = 0.7 \times 0.7 \times 0.3 \times 0.7 \times 0.7 = 0.072$$

Is it a transmembrane protein?

Problem: need a threshold,
 threshold must be length dependent

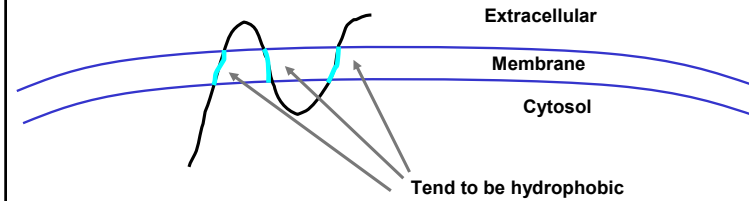
Transmembrane model:



HHHLLHHHLLLLLHLHLLHLLHLLHHHL
 HHHLHHLHLLLLLHHHLLHLLHHHHHL
 HH...

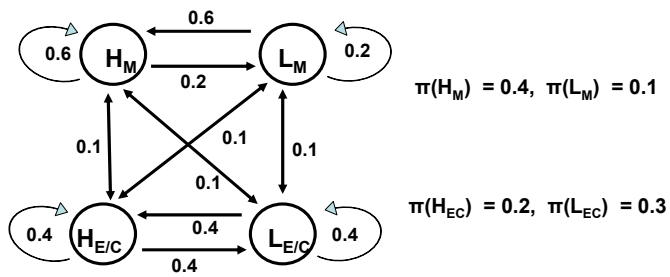
$\pi(H)$ = # of sequences that begin with H,
 normalized by the total # of training sequences

An example: transmembrane regions



Boundary detection problem:
 Given sequence of H's & L's, find all transmembrane regions

A four state transmembrane HMM



$e_H(H_M) = 1.0, e_L(L_M) = 1.0, e_H(H_{EC}) = 1.0, e_L(L_{EC}) = 1.0$

Markov Chains

States: S_1, S_2, \dots, S_N
 Initial state probabilities: $\pi(i)$
 Transition probabilities: a_{ij}

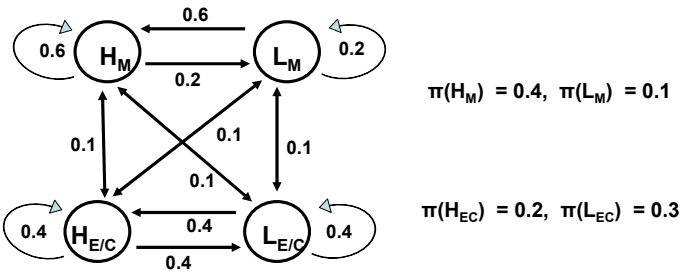
HMMs

States: S_1, S_2, \dots, S_N
 Initial state probabilities: $\pi(i)$
 Transition probabilities: a_{ij}
 Alphabet, Σ
 Emission probabilities: e_i

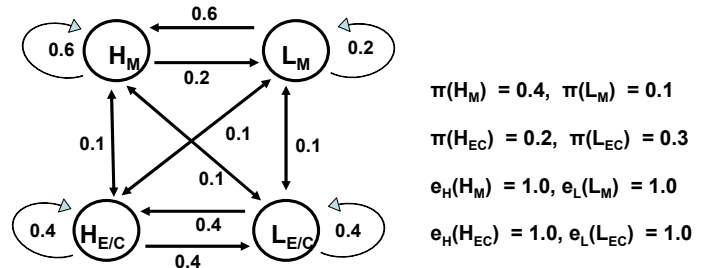
For the TM and E/C Markov chains,

sequence of symbols = sequence of states

Given HLLLHLLL, know $q_0 = S_H$, $q_1 = S_L$, $q_2 = S_L \dots$



In this HMM model, what is $q_0, q_1, q_2 \dots$?



Unlike Markov chains, in HMMs the states are *hidden*.

Given an unlabeled sequence, HHHLLHL...,

we infer the most probable state sequence to obtain the boundaries