

Advanced Algorithms and Models for Computational Biology -- a machine learning approach

Computational Genomics II: HMM variants and Comparative Gene Finding

Eric Xing

Lecture 5, February 1, 2005

Reading: Chap 3, 5 DEKM book
Chap 9, DTW book



Higher-order HMMs

- The Genetic Code**

- 3 nucleotides make 1 amino acid
- Statistical dependencies in triplets

- Question:**

- Recognize protein-coding segments with an HMM

	U	C	A	G
U	UUU phe UUC UUA leu UUG	UCU ser UCC UCA UCG	UAU tyr UAC UAA Stop UAG Stop	UGU cys UGC UGA Stop UGG Stop
C	CUU leu CUC CUA CUG	CCU pro CCC CCA CCG	CAU his CAC CAA gln CAG	CGU arg CGC CGA CGG
A	AUU ile AUC AUA AUG met	ACU thr ACC ACA ACG	AAU asn AAC AAA AAG lys	AGU ser AGC AGA AGG arg
G	GUU val GUC GUA GUG	GCU ala GCC GCA GCG	GAU asp GAC GAA GAG glu	GGU gly GGC GGA GGG



Higher-order HMMs

- Every state of the HMM emits 1 nucleotide

- Transition probabilities:

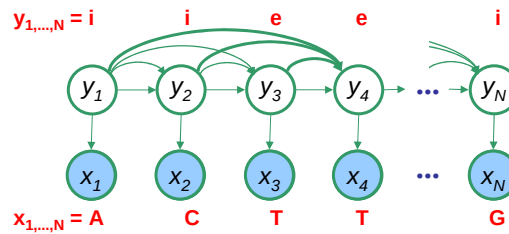
Probability of a state at one position, given those of 3 previous positions (triplets):

$$P(y_i | y_{i-1}, y_{i-2}, y_{i-3})$$

- Emission probabilities:

$$P(x_i | y_i)$$

- Algorithms extend with small modifications



Inference on Higher-order HMMs

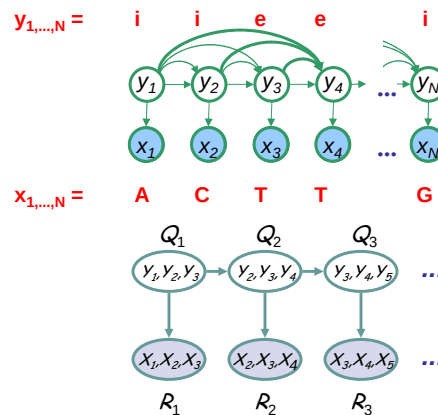
- Building 1st-order HMM on "mega" state

- Use FB algorithm as usual

$$P(Q_2 | R)$$

$$\rightarrow P(y_2, y_3, y_4 | X)$$

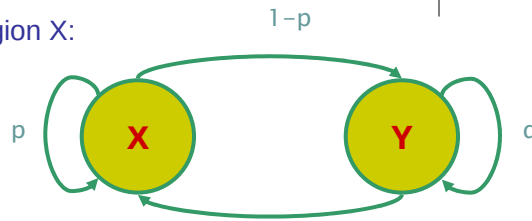
$$\rightarrow P(y_3 | X) = \sum_{y_2, y_4} P(y_2, y_3, y_4 | X)$$



Modeling the Duration of States

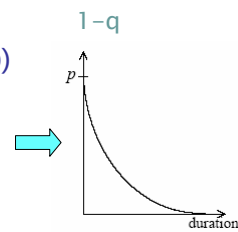
- Length distribution of region X:

$$E[l_X] = 1/(1-p)$$



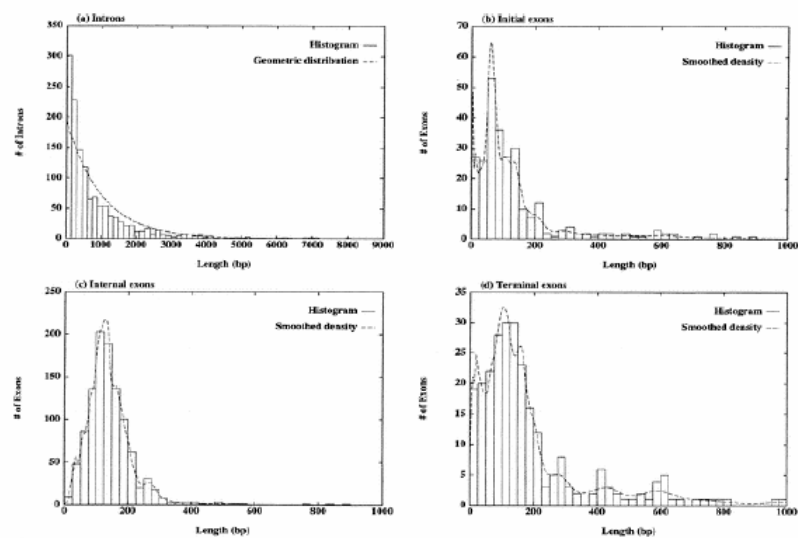
- Geometric distribution, with mean $1/(1-p)$

- (homework: derive this)



- This is a significant disadvantage of HMMs
- Several solutions exist for modeling different length distributions

Observed Duration Time

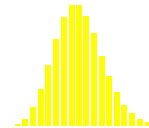


Poisson Point Process



- A counting process that represents the total number of occurrences of discrete events during a temporal/spatial interval
- the number of occurrences in any interval of length τ is Poisson distributed with parameter $\lambda\tau$:

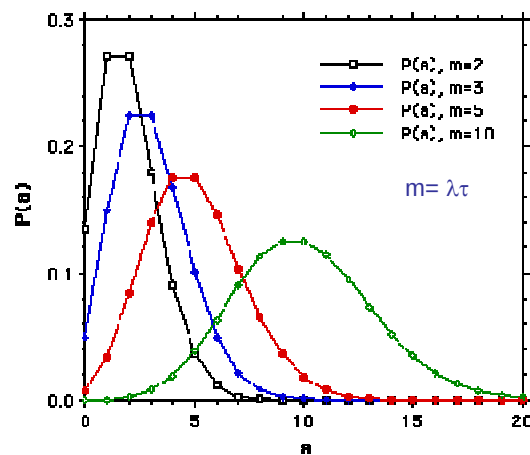
$$p(A(t + \tau) - A(n) = n) = e^{-\lambda\tau} \frac{(\lambda\tau)^n}{n!}$$



- the number of occurrences in disjoint intervals are independent
- the duration of the interval between two consecutive occurrences has the following distribution:

$$p(\tau < s) = 1 - e^{-\lambda s}$$

Poisson point process

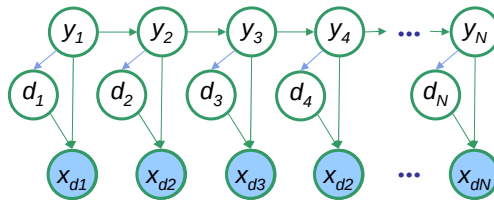


Truncation is needed at both ends!

Generalized HMM

Upon entering a state:

1. Choose duration d , according to probability distribution
2. Generate d letters according to emission probs
3. Take a transition to next state according to transition probs



Disadvantage: Increase in complexity:

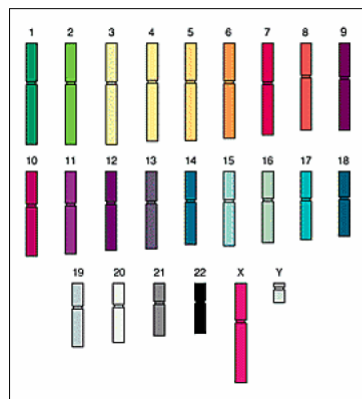
Time: $O(D^2)$

Space: $O(D)$

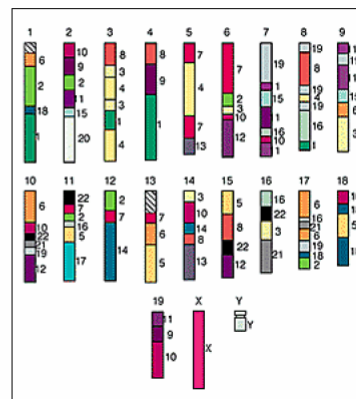
where D = maximum duration of state

Comparative Genomics

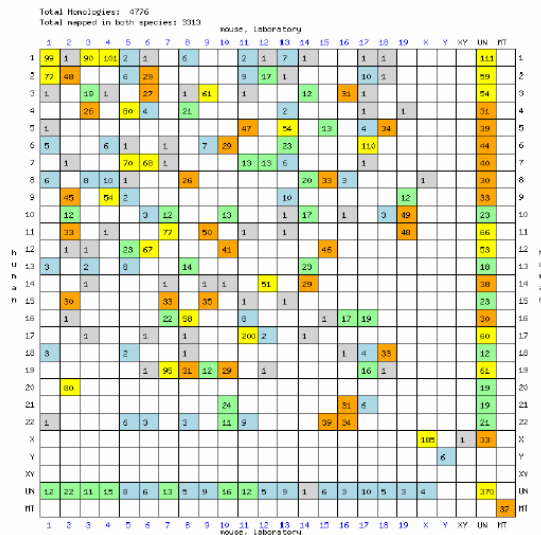
Human



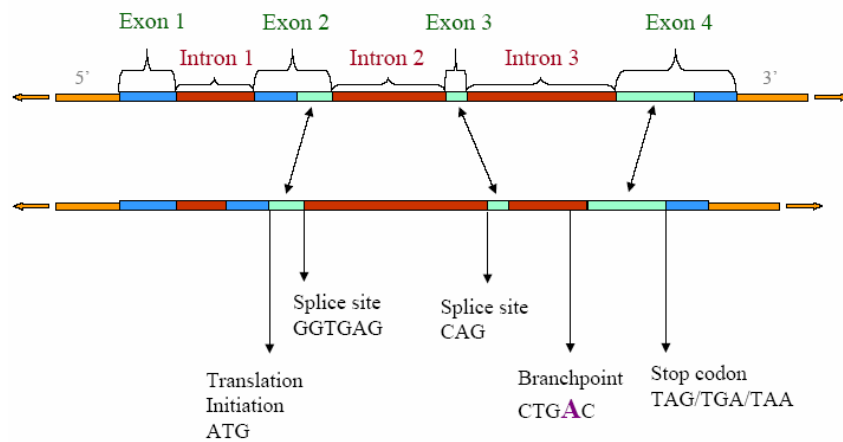
Mouse



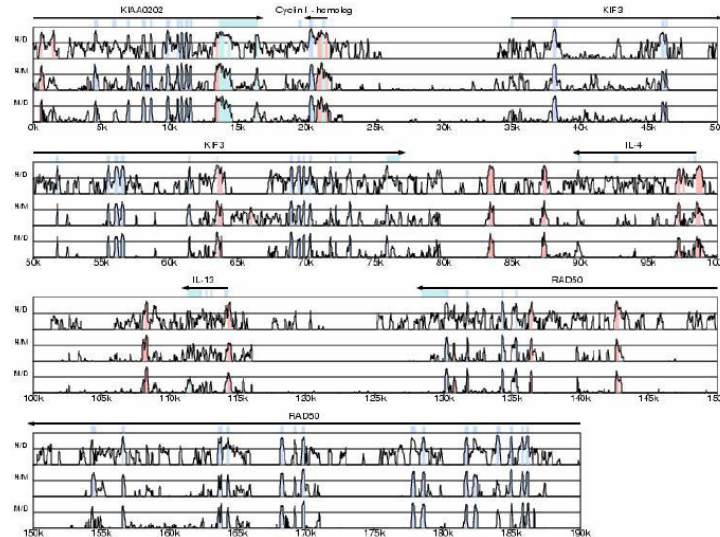
A pairwise comparison between human and mouse genome



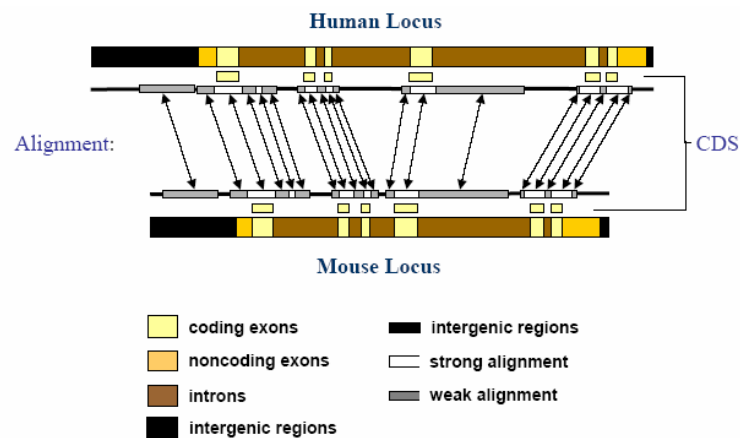
Aligning One Locus



Three Pairwise Alignments



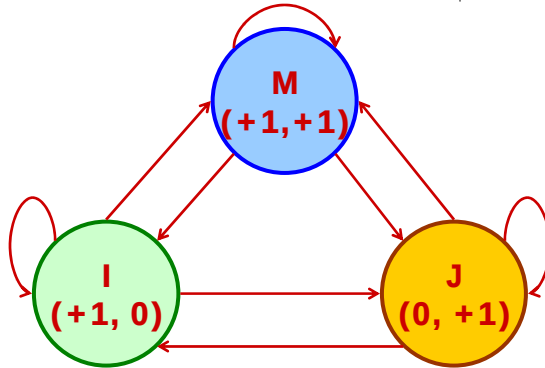
Example: a human/mouse ortholog



Paired HMM



Alignments correspond
1-to-1 with sequences
of states M, I, J

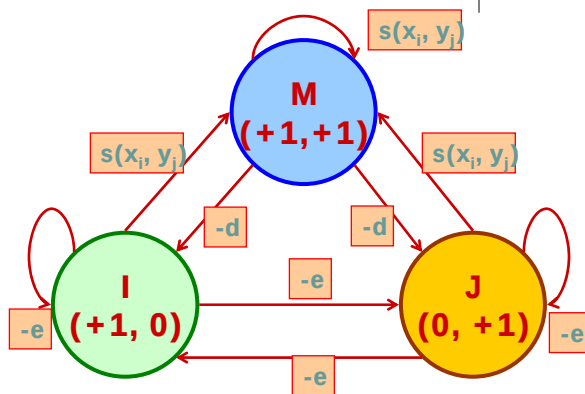


- AGGCTATCACCTGACCTCCAGGCCGA - - TGCCC - - -
TAG - CTATCAC - - GACCGC - GGTGATTGCCCCGACC
I M M J M M M M M M M J J M M M M M M J M M M M M M M I I M M M M M I I I

Let's score the transitions

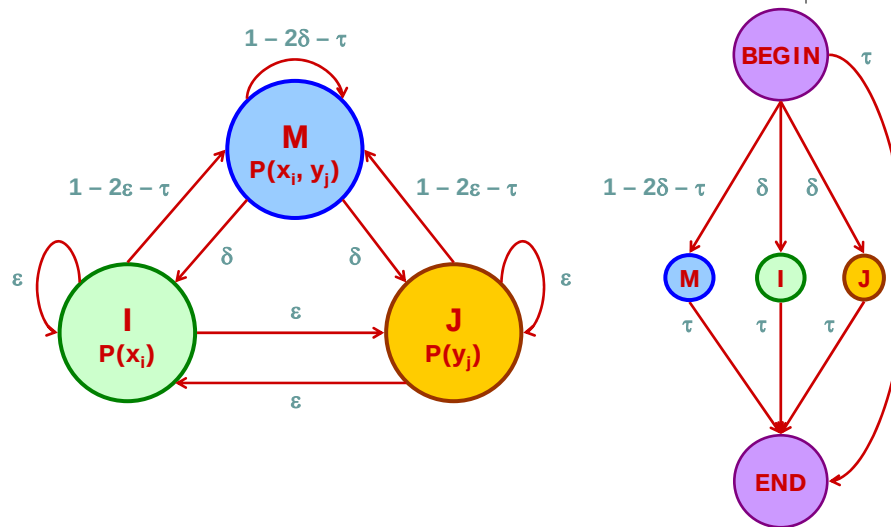


Alignments correspond
1-to-1 with sequences
of states M, I, J

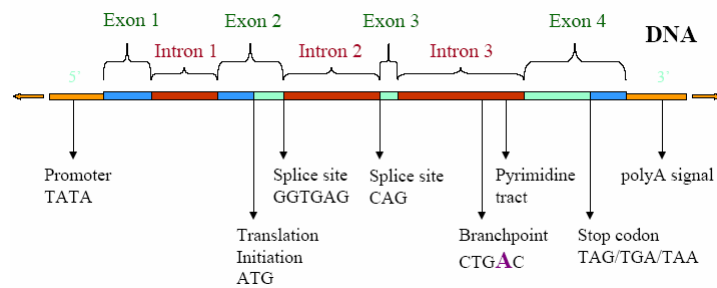


- AGGCTATCACCTGACCTCCAGGCCGA - - TGCCC - - -
TAG - CTATCAC - - GACCGC - GGTGATTGCCCCGACC
I M M J M M M M M M M J J M M M M M M J M M M M M M M I I M M M M M I I I

A Pair HMM for alignments



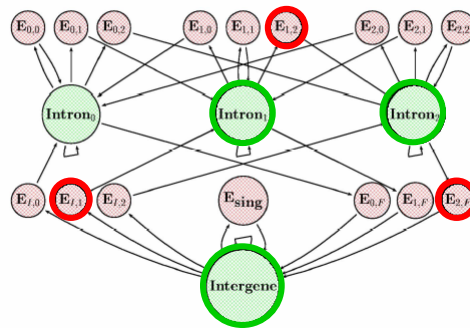
Gene Finding



Generalized HMM Gene finder



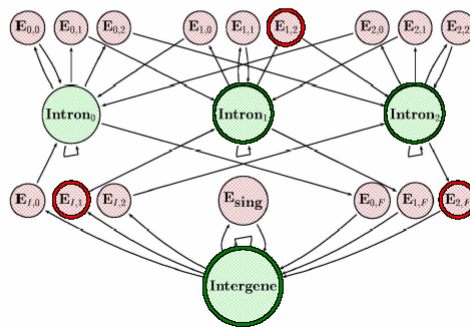
TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA



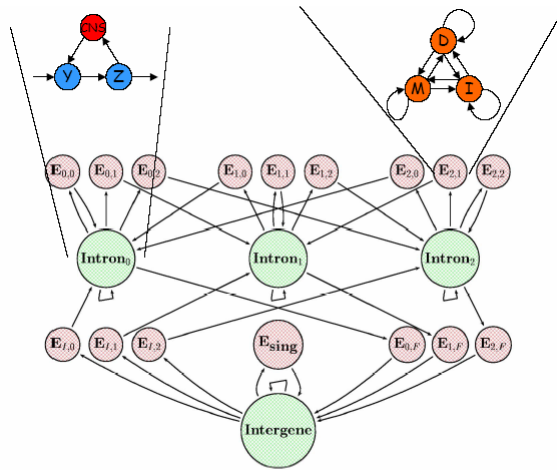
Generalized Pair-HMM gene finder



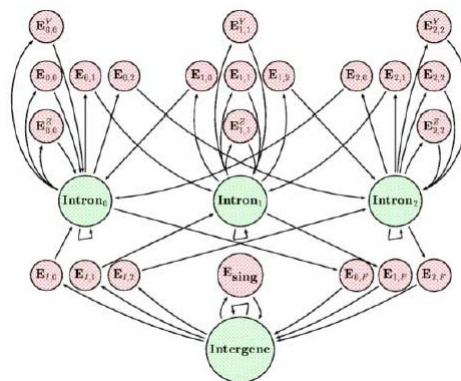
TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA
 CTG ATGTACACTG GTTGGTCCTCAG CTTTGACGGG GTG CATGTAA TGTG



Hierarchical state transition in pHMM



Allowing for inserted exons



Acknowledgments



- **Serafim Batzoglou**: for some of the slides adapted or modified from his lecture slides at Stanford University
- **Lior Pachter**: for some of the slides modified from his lectures at UC Berkeley