HMM variants

--- for comparative gene finding, etc.

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

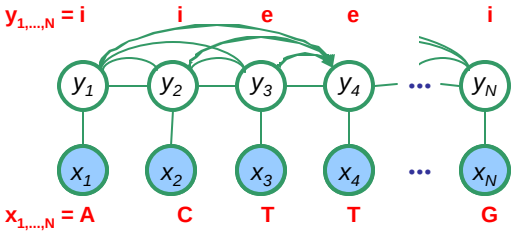
Higher-order HMMs

Every state of the HMM emits 1 nucleotide

Transition probabilities:

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

$$P(\pi_i \mid \pi_{i-1}, \pi_{i-2}, \pi_{i-3})$$



Emission probabilities:

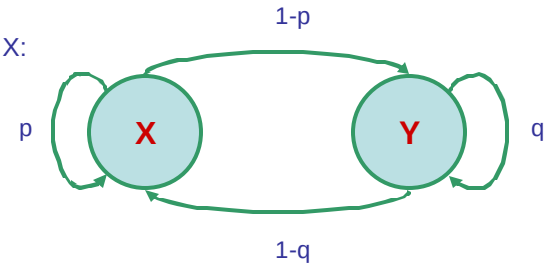
$$P(x_i \mid \pi_i)$$

Algorithms extend with small modifications

Modeling the Duration of States

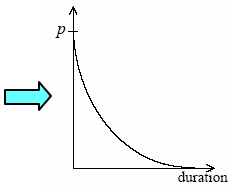
Length distribution of region X:

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



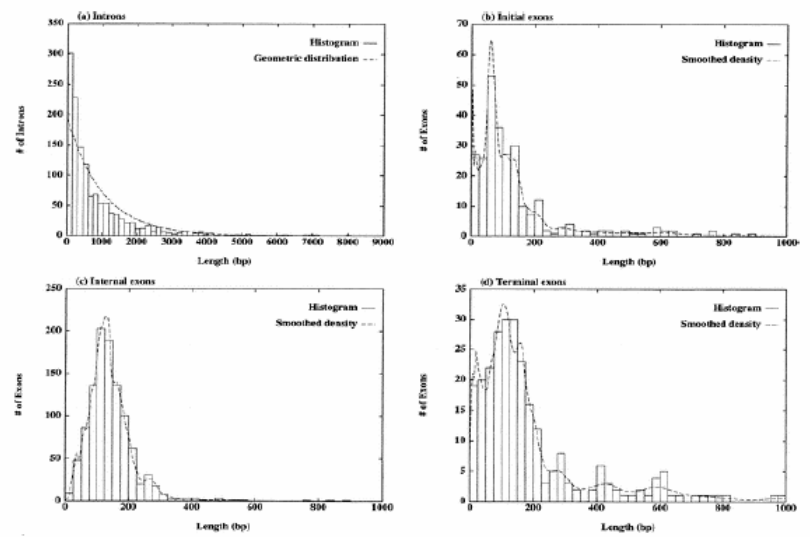
- Geometric distribution, with mean 1/(1-p)

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

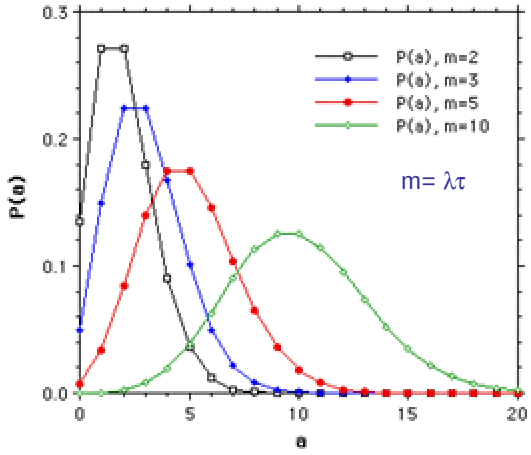
- the number of occurrences in any internal of length τ is Poisson distributed with parameter $\lambda\tau$:

$$p(A(t+t) - A(n) = n) = e^{-lt} \frac{(l\,t)^n}{n!}$$

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

$$p(t < s) = 1 - e^{-ls}$$

Poisson point process

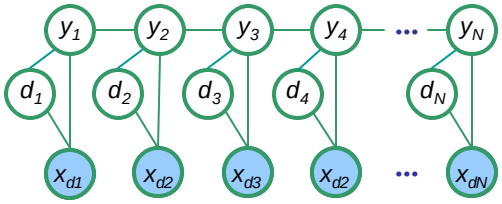


Truncation is needed at both ends!

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

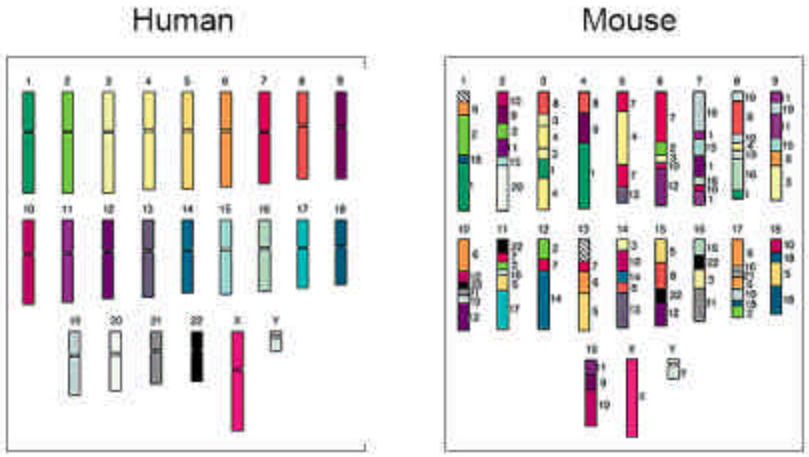


Disadvantage: Increase in complexity:

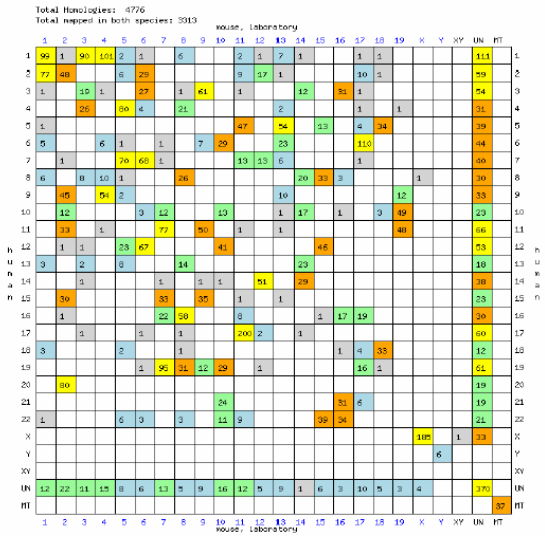
Time: $O(D^2)$
Space: $O(D)$

where D = maximum duration of state

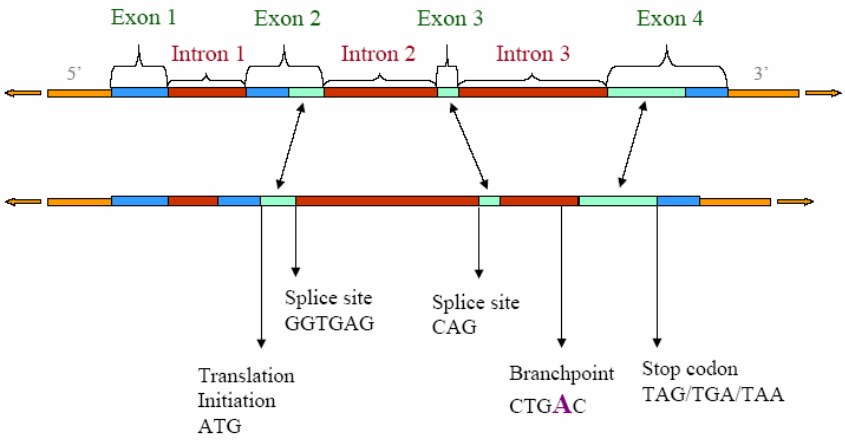
Comparative genomics



A pairwise comparison between human and mouse genome



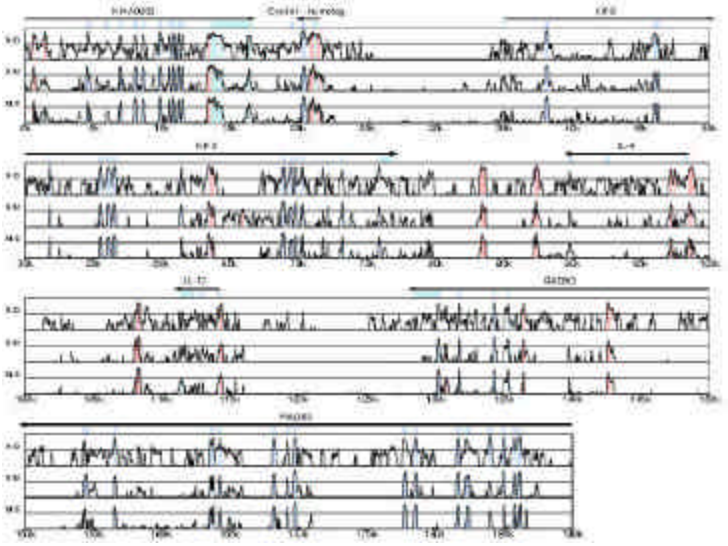
Aligning one locus



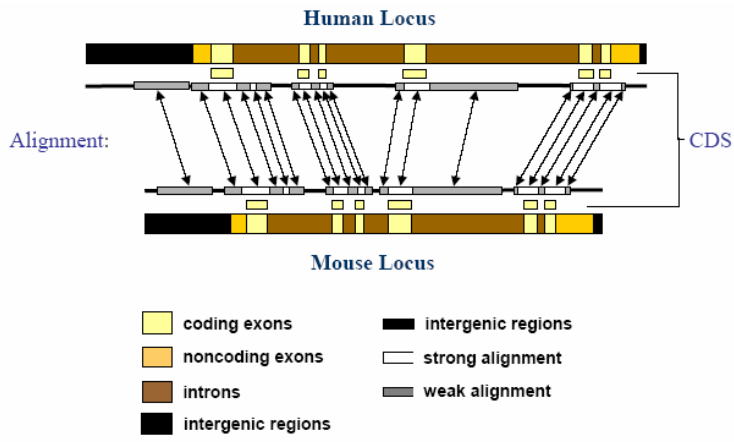
Pairwise alignment - a close-up view

```
 50     .    :     .    :     .    :     .    :     .    :
247 GGTGAGGTCGAGGACCCTGCA  CGGAGCTGTATGGAGGGCA   AGAGC
    |:   ||  ||||:  |||| --:||  ||| |::|   |||---||||
368 GAGTCGGGGGAGGGGGCTGCTGTTGGCTCTGGACAGCTTGCATTGAGAGG

100     .    :     .    :     .    :     .    :     .    :
292 TTC          CTACAGAAAAGTCCCAGCAAGGAGCCACACTTCACTG
    |||----------|| |   |::| |: ||||::|:||:-||  ||:| |
418 TTCTGGCTACGCTCTCCCTTAGGGACTGAGCAGAGGGCT CAGGTCGCGG

150     .    :     .    :     .    :     .    :     .    :
332               ATGTCGAGGGGAAGACATCATTCGGGATGTCAGTG
    ---------------|||||||||||||||||||||:||||||||||||
467 TGGGAGATGAGGCCAATGTCGAGGGGAAGACATCATTTGGGATGTCAGTG

200     .    :     .    :     .    :     .    :     .    :
367 TTCAACCTCAGCAATGCCATCATGGGCAGCGGCATCCTGGGACTCGCCTA
    |||||:||||||||:|||||||||||||:|| ||:|||||:||||||||
517 TTCAATCTCAGCAACGCCATCATGGGCAGTGGAATTCTGGGGCTCGCCTA
```

Three pairwise alignments

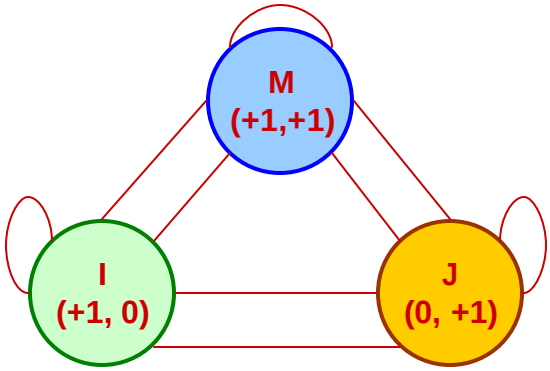


Example: a human/mouse ortholog



Paired HMM

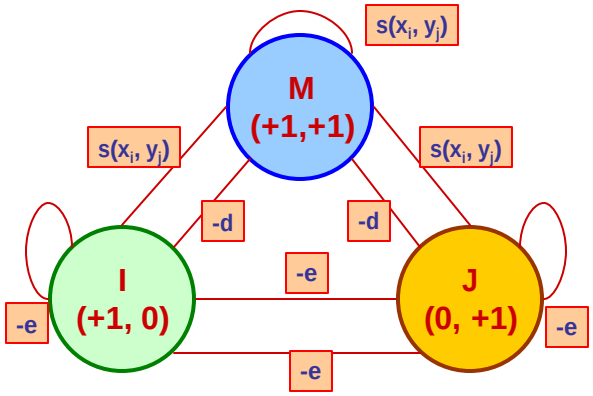
Alignments correspond
1-to-1 with sequences
of states M, I, J




```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
IMMJMMMMMMMJJMMMMMMMJMMMMMMMIIMMMMMIII
```

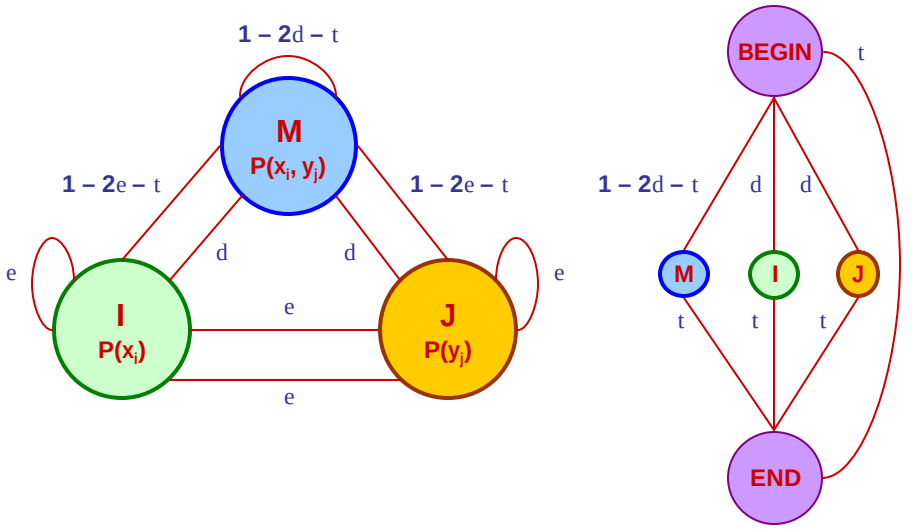
Let's score the transitions

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

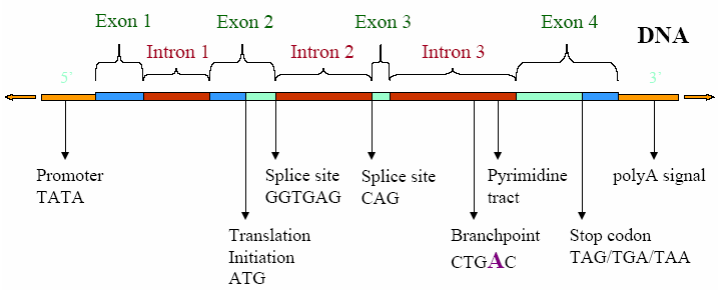



```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
IMMJMMMMMMMJJMMMMMMMJMMMMMMMIIMMMMMIII
```

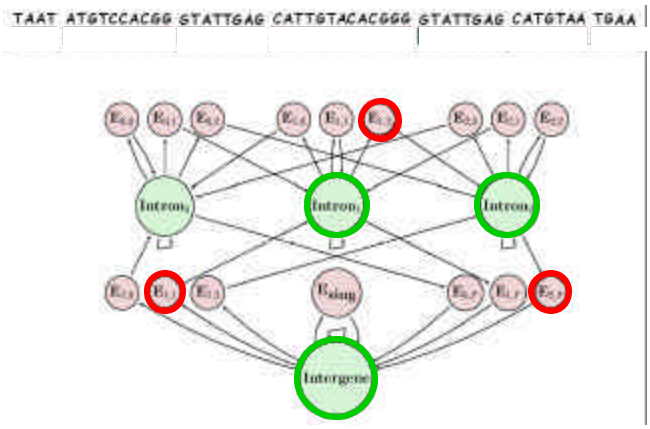
A Pair HMM for alignments



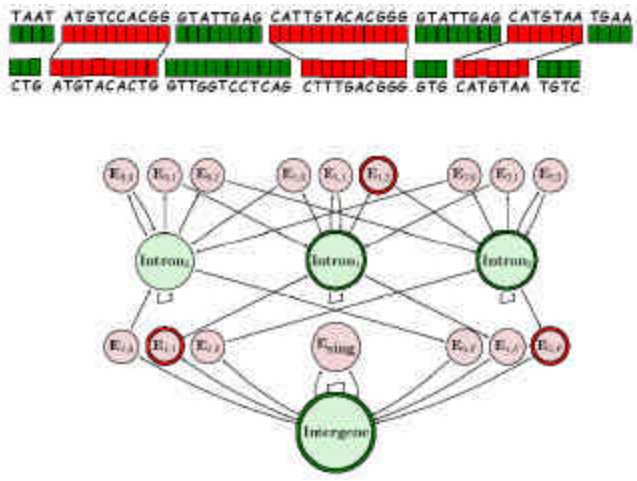
Gene finding



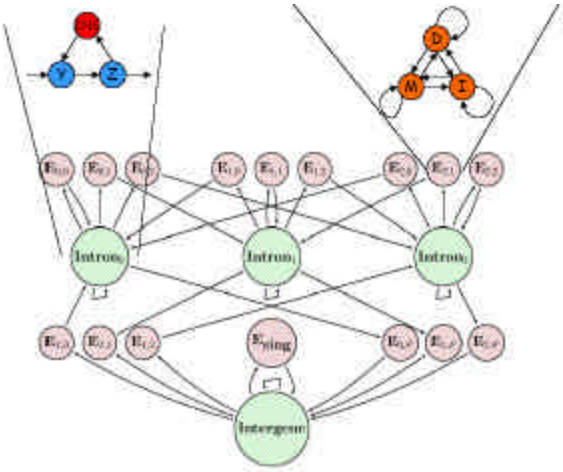
Generalized HMM Gene finder



Generalized Pair-HMM gene finder



Hierarchical state transition in pHMM



Allowing for inserted exons