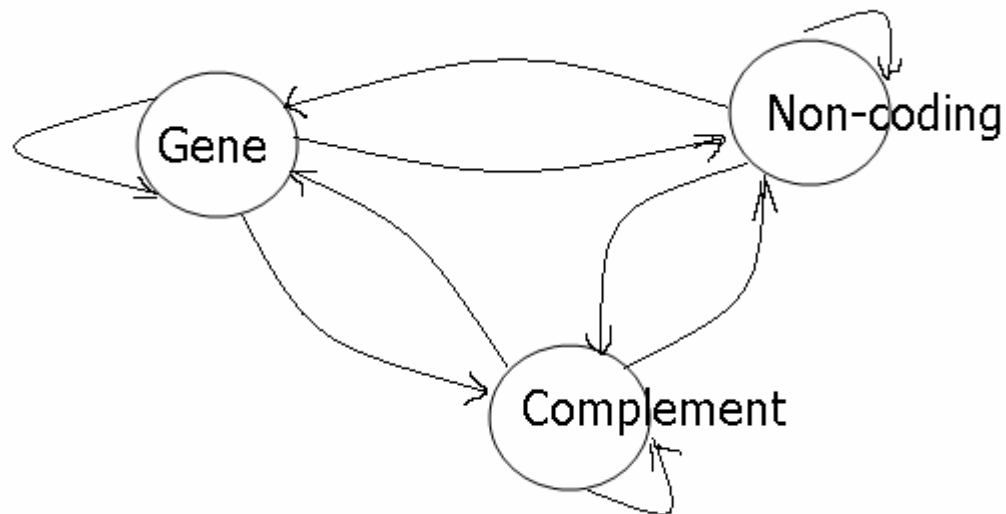
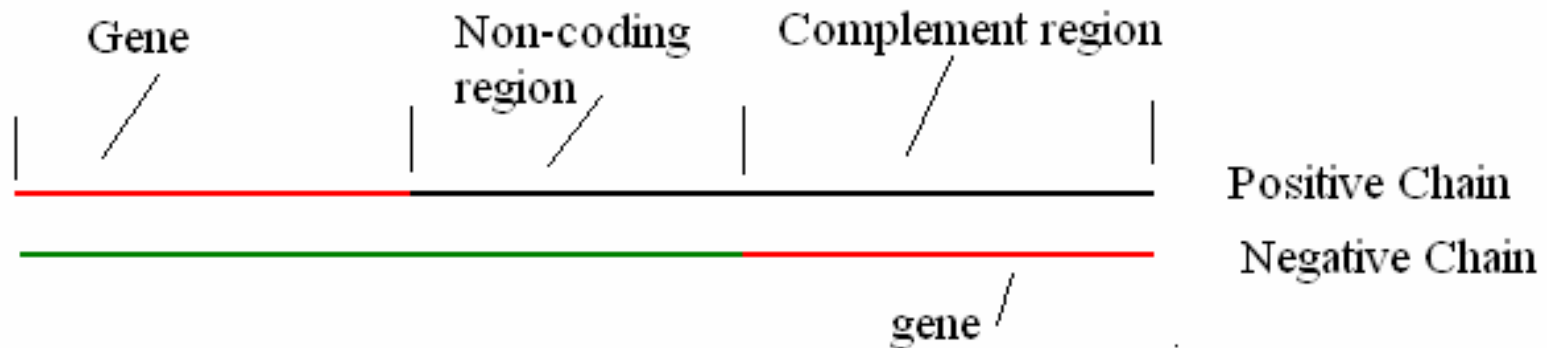




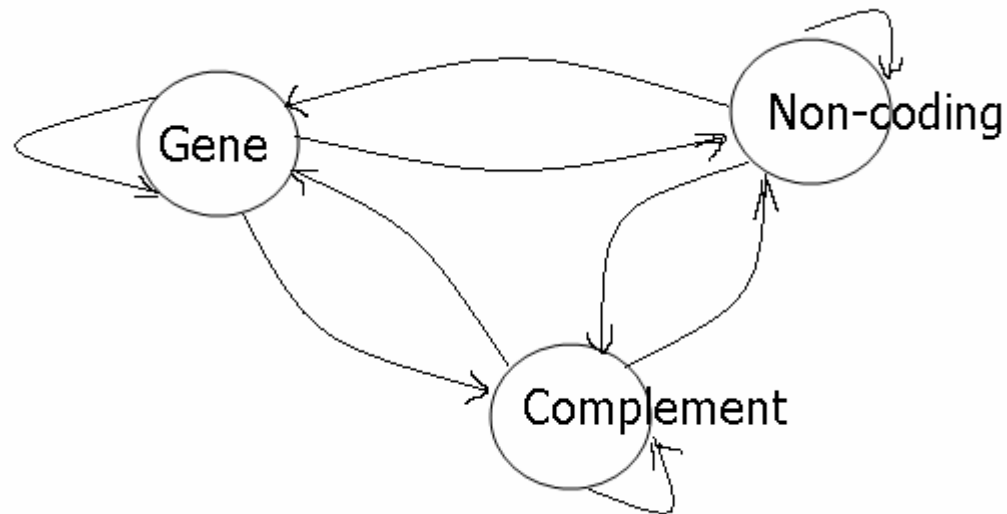
Gene Finding (Cont.)

Charles Yan

Hidden Markov Model (HMM) with Duration

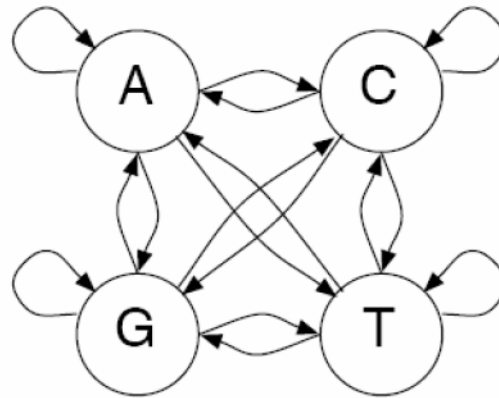


Hidden Markov Model (HMM) with Duration



ATGGTCGATAGGGCCAATGCATACATAGACATAGAATAGGGCCAATG
ggggggggggggggggggggnnnnnnnnnnnnnnnnnnnncccccccccc
 (g, 18) → (n, 5) → (g, 8) → (n, 6) → (c, 10)

Markov Chain Model



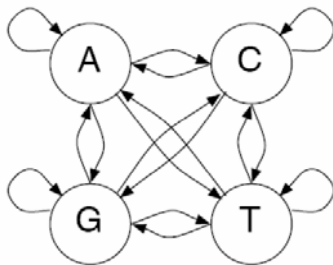
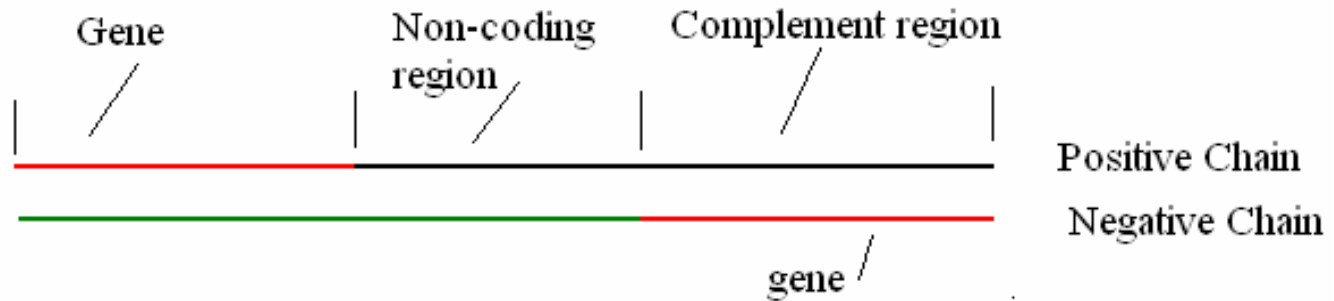
States: A,C,G,T

Emissions: corresponding letter

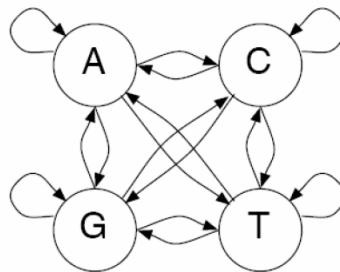
Transitions: $a_{st} = P(x_i = t \mid x_{i-1} = s)$

ATGGTCGATAGGGCCAATGCATACATAGACATAGAATAGGGCCAATG
ATGGTCGATAGGGCCAATGCATACATAGACATAGAATAGGGCCAATG

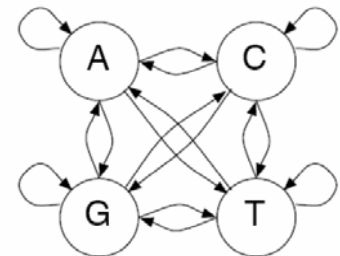
Markov Chain Model



States: A,C,G,T
 Emissions: corresponding letter
 Transitions: $a_{st} = P(x_i = t | x_{i-1} = s)$

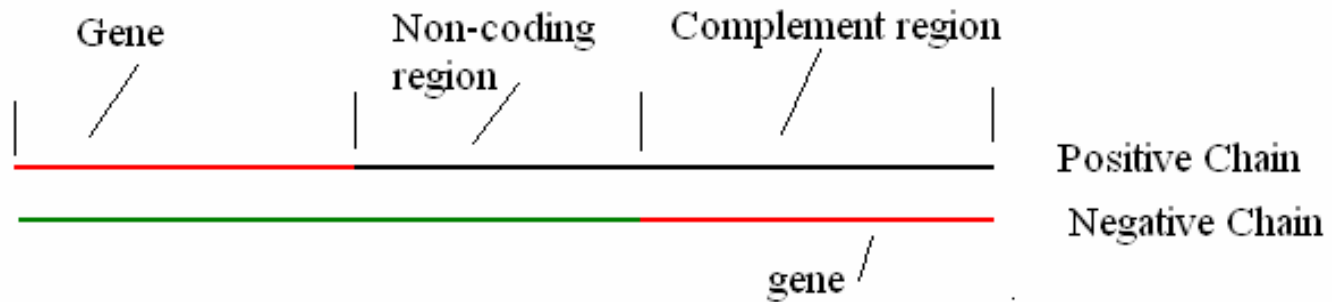


States: A,C,G,T
 Emissions: corresponding letter
 Transitions: $a_{st} = P(x_i = t | x_{i-1} = s)$



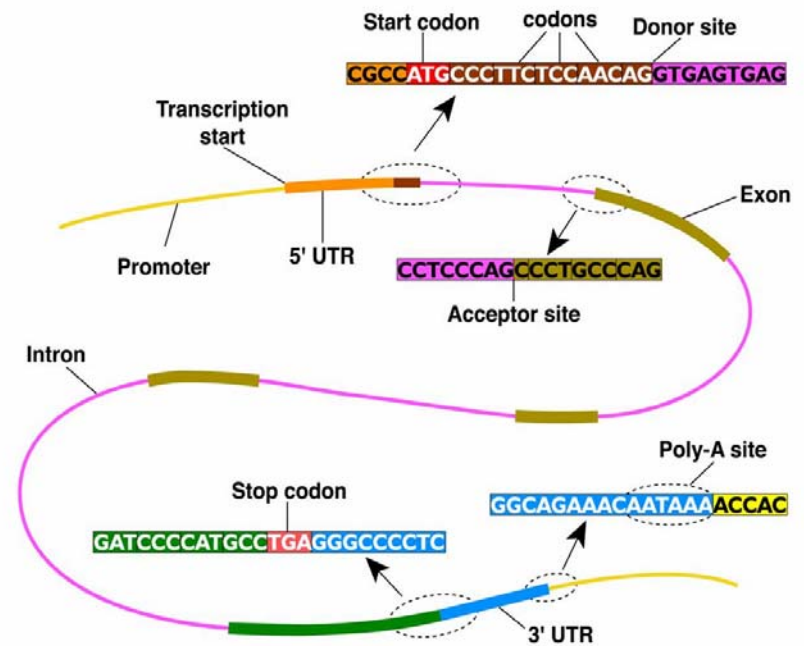
States: A,C,G,T
 Emissions: corresponding letter
 Transitions: $a_{st} = P(x_i = t | x_{i-1} = s)$

Prokaryotes



Eukaryotes

- Transcription (transcription factor binding sites and TATA boxes)
- Splicing (donor and acceptor sites and branch points)
- Polyadenylation [poly(A) site],
- Translation (initiation site, generally ATG with exceptions, and stop codons)



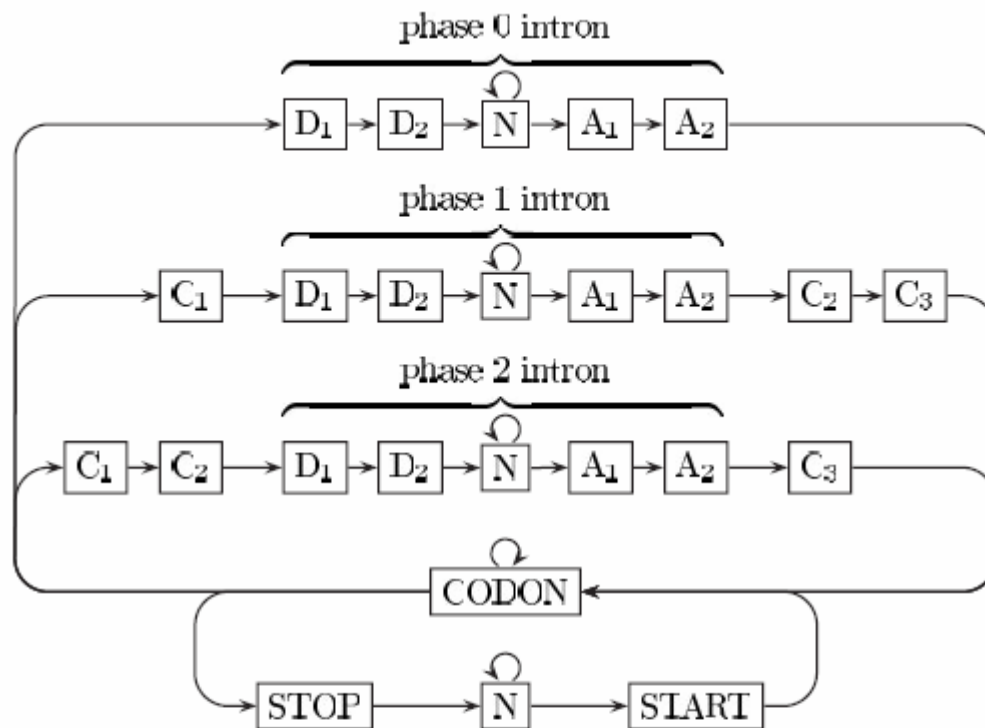
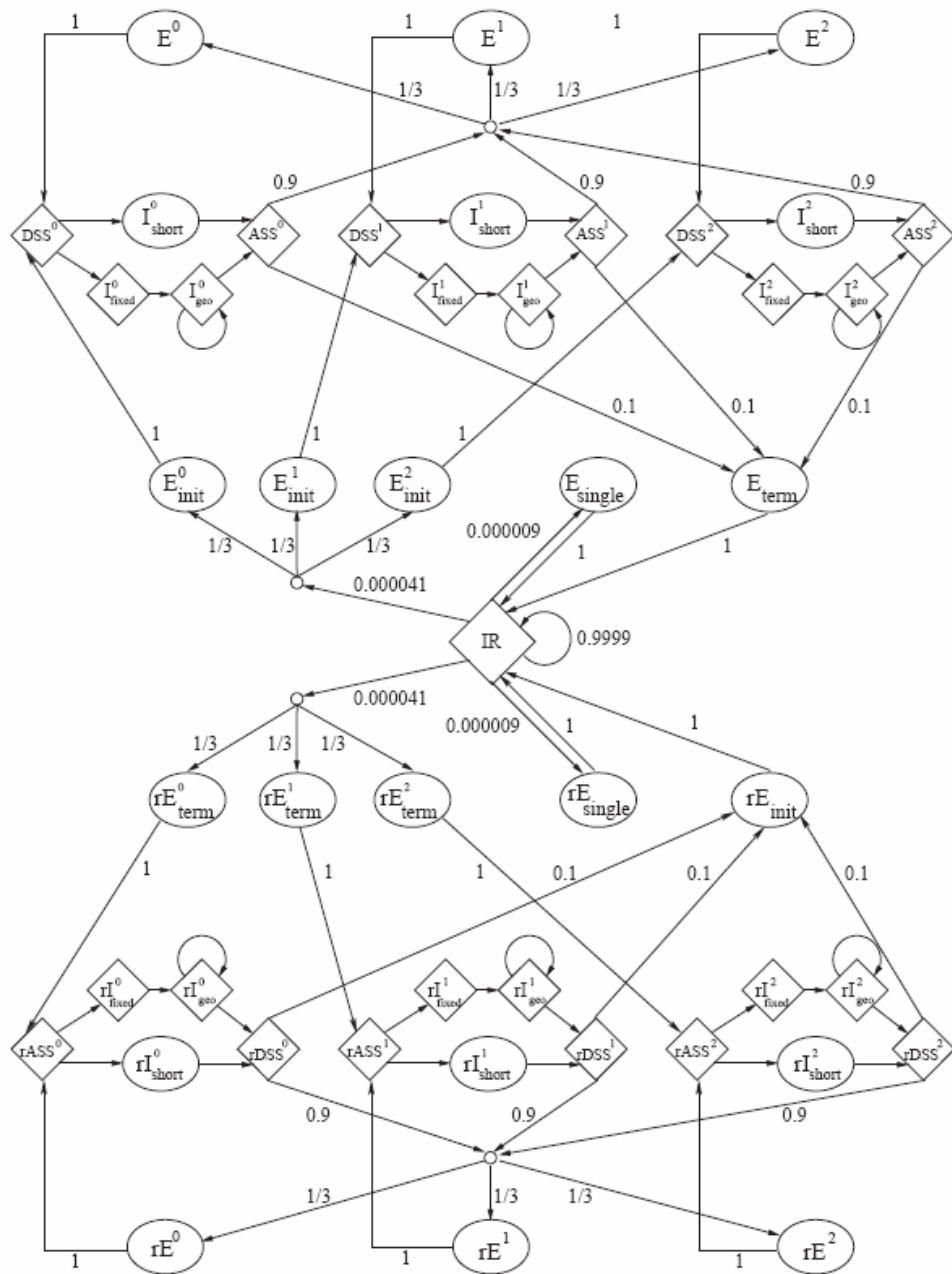
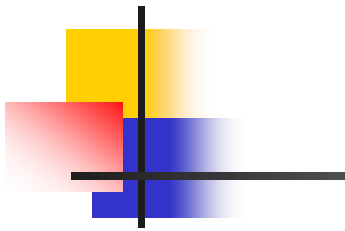


Fig. 2. Architecture of eukaryotic EHMM. Each box of the graph is a state of the model, some states are of the same type. Start, stop, and codon denotes start codon, stop codon, and inner codon, respectively. N denotes intergenic/intron state. C_1 , C_2 , and C_3 denote single-codon positions, in combination they model the codons which are split by introns. D_1 , and D_2 denote first and second position of the splice donor site. A_1 , and A_2 denote first and second position of the splice acceptor site.



forward strand

reverse strand



Pitfalls and Issues

Several issues make the problem of eukaryotic gene finding extremely difficult.

- 1) **Very long genes**: for example, the largest human gene, the dystrophin gene, is composed of 79 exons spanning nearly 2.3 Mb.
- 2) **Very long introns**: again, in the human dystrophin gene, some introns are >100 kb long and >99% of the gene is composed of introns.



Pitfalls and Issues

- 3) **Very conserved introns.** this is particularly a problem when gene prediction is addressed through similarity searches.



Pitfalls and Issues

- 4) **Very short exons**: some exons are only 3 bp long in Arabidopsis genes and probably even 1 bp for the coding part of exons at either end of the coding sequence, meaning that **start or stop codons can be interrupted** by an intron. Such small exons are easily missed by all content sensors, especially if bordered by large introns. The more difficult cases are those where the length of a coding exon is **a multiple of three** (typically 3, 6 or 9 bp long), because missing such exons will not cause a problem in the exon assembly as they do not introduce any change in the frame.



Pitfalls and Issues

- 5) **Overlapping genes**: though very rare in eukaryotic genomes, there are some documented cases in animals as well as in plants
- 6) **Polycistronic gene arrangement**: one gene, and one mRNA, but two or more proteins.



Pitfalls and Issues

- 7) **Frameshifts:** some sequences stored in databases may contain errors (either sequencing errors or simply errors made when editing the sequence) resulting in the introduction of artificial frameshifts (deletion or insertion of one base). Such frameshifts greatly increase the difficulty of the computational gene finding problem by producing erroneous statistics and masking true solutions.



Pitfalls and Issues

- 8) **Introns in non-coding regions**: there are genes for which the genomic region corresponding to the 5` - and/or 3` -UTR in the mature mRNA is interrupted by one or more intron(s).
- 9) **Alternative transcription start**: e.g. three alternative promoters regulate the transcription of the 14 kb full-length dystrophin mRNAs and four `intragenic' promoters control that of smaller isoforms.



Pitfalls and Issues

10) Alternative splicing.

11) Alternative polyadenylation: 20% of human transcripts showing evidence of alternative polyadenylation.



Pitfalls and Issues

12) **Alternative initiation of translation**: finding the right AUG initiator is still a major concern for gene prediction methods. the rule stating that the firrst AUG in the mRNA is the initiator codon can be escaped through three mechanisms: context-dependent **leaky scanning**, **re-initiation** and **direct internal initiation**. **Non-AUG triplet can sometimes act as the functional codon for translation initiation**, as ACG in Arabidopsis or CUG in human sequences