Example 1: Finding CpG islands

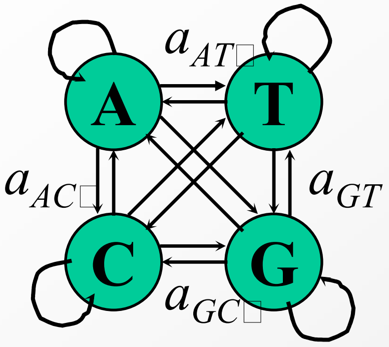
What are CpG islands?

- Regions of regulatory importance in promoters of many genes
  - Defined by their methylation state (epigenetic information)
- Methylation process in the human genome:
  - Very high chance of methyl-C mutating to T in CpG

    ➔ CpG dinucleotides are much rarer
  - BUT it is suppressed around the promoters of many genes

    ➔ CpG dinucleotides are much more frequent than elsewhere
    - Such regions are called **CpG islands**
    - A few hundred to a few thousand bases long
- Problems:
  - Given a short sequence, does it come from a CpG island or not?
  - How to find the CpG islands in a long sequence

Training Markov Chains for CpG islands



- Training Set:
  - set of DNA sequences w/ known CpG islands
- Derive two Markov chain models:
  - **'+' model**: from the CpG islands
  - **'-' model**: from the remainder of sequence
- Transition probabilities for each model:

Probability of C following A

+	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

$$a^{+}_{st} = \frac{c^{+}_{st}}{\sum_{t'} c^{+}_{st'}}$$

$\boldsymbol{c^{+}_{st}}$ is the number of times letter ***t*** followed letter ***s*** inside the CpG islands

$$a^{-}_{st} = \frac{c^{-}_{st}}{\sum_{t'} c^{-}_{st'}}$$

c^{-}_{st} is the number of times letter ***t*** followed letter ***s*** outside the CpG islands

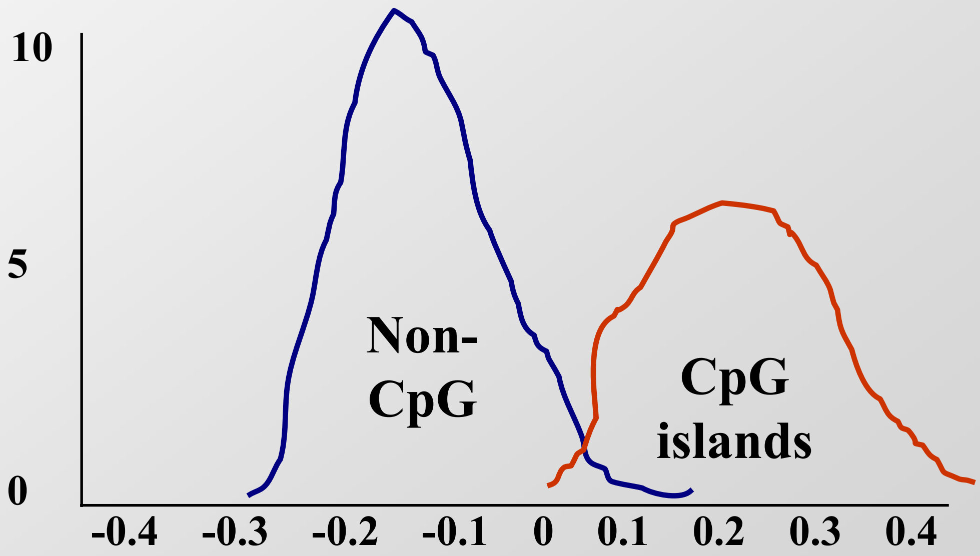
Using Markov Models for CpG classification

Q1: Given a short sequence *x*, does it come from CpG island (**Yes-No** question)

- To use these models for discrimination, calculate the log-odds ratio:

$$\boldsymbol{S(x)} \equiv \log \frac{\boldsymbol{P(x}|\mathrm{model}+\boldsymbol{)}}{\boldsymbol{P(x}|\mathrm{model}-\boldsymbol{)}} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}}$$

Histogram of log odds scores



Using Markov Models for CpG classification

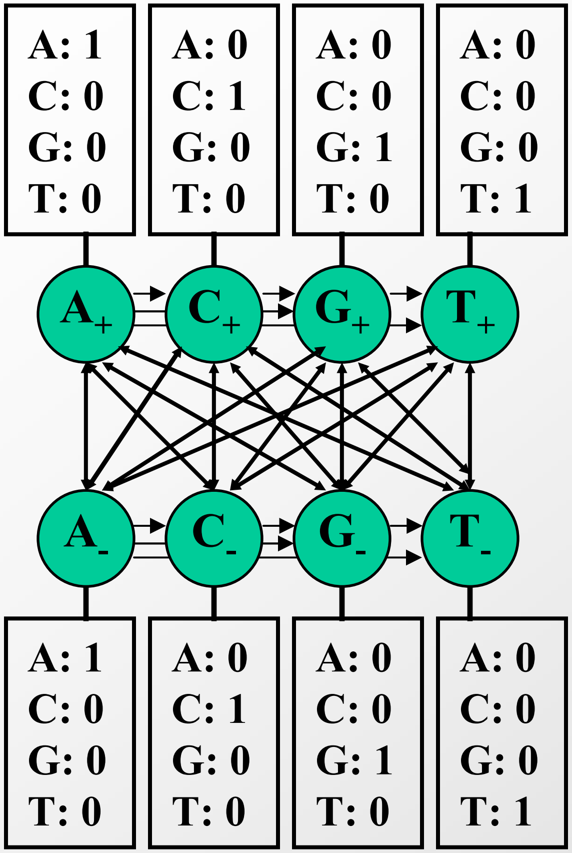
Q2: Given a long sequence *x*, how do we find CpG islands in it

(**Where** question)

- Calculate the log-odds score for a window of, say, 100 nucleotides around every nucleotide, plot it, and predict CpG islands as ones w/ positive values
- Drawbacks: Window size

Use a hidden state: CpG (+) or non-CpG (-)

HMM for CpG islands



A: 1 C: 0 G: 0 T: 0

A: 0 C: 1 G: 0 T: 0

A: 0 C: 0 G: 1 T: 0

A: 0 C: 0 G: 0 T: 1

A: 1 C: 0 G: 0 T: 0

A: 0 C: 1 G: 0 T: 0

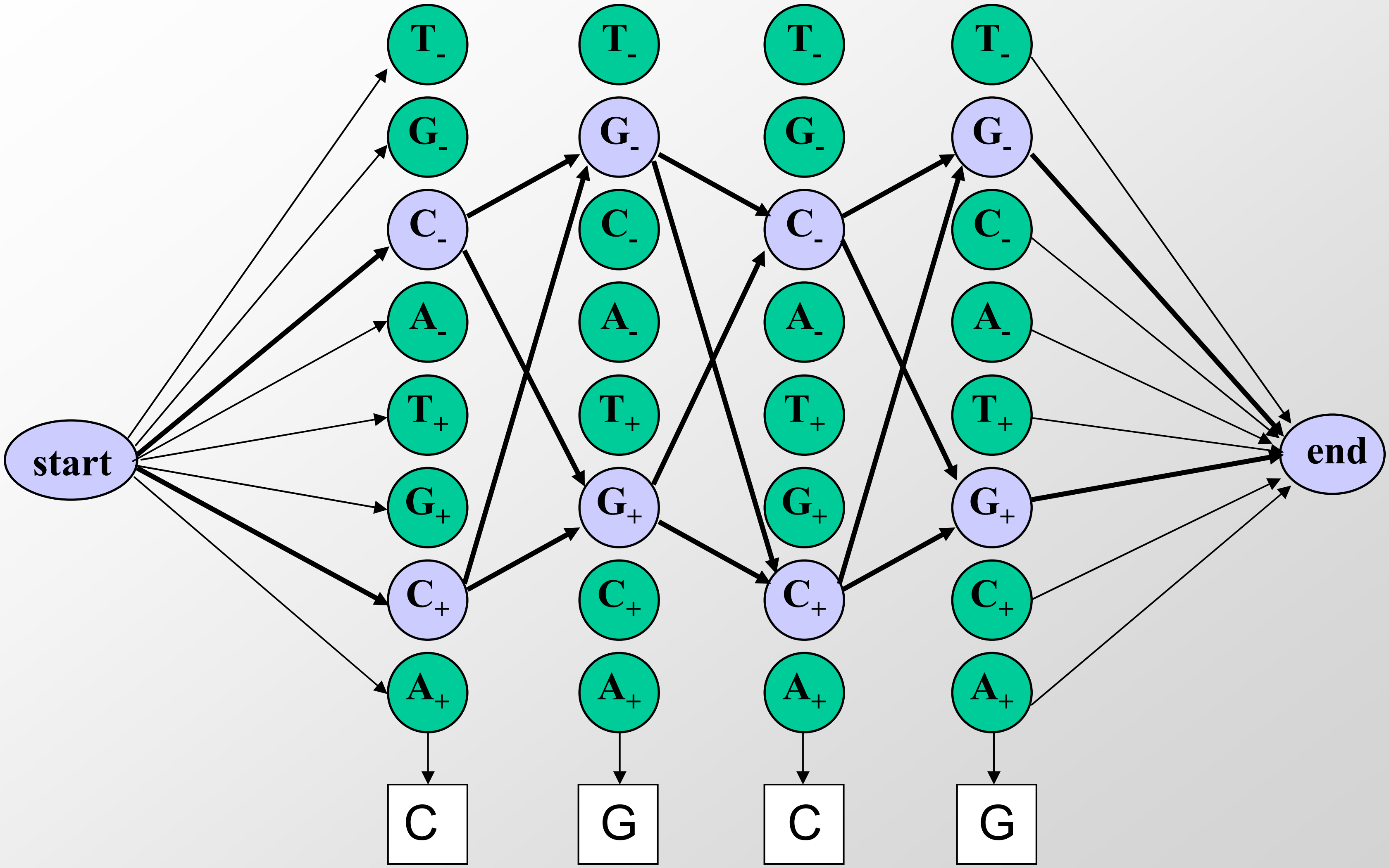
A: 0 C: 0 G: 1 T: 0

A: 0 C: 0 G: 0 T: 1

- Build a single model that combines both Markov chains:
  - **'+' states**: A_+, C_+, G_+, T_+
    - Emit symbols: A, C, G, T in CpG islands
  - **'-' states**: A_-, C_-, G_-, T_-
    - Emit symbols: A, C, G, T in non-islands
- Emission probabilities distinct for the '+' and the '-' states
  - Infer most likely set of states, giving rise to observed emissions

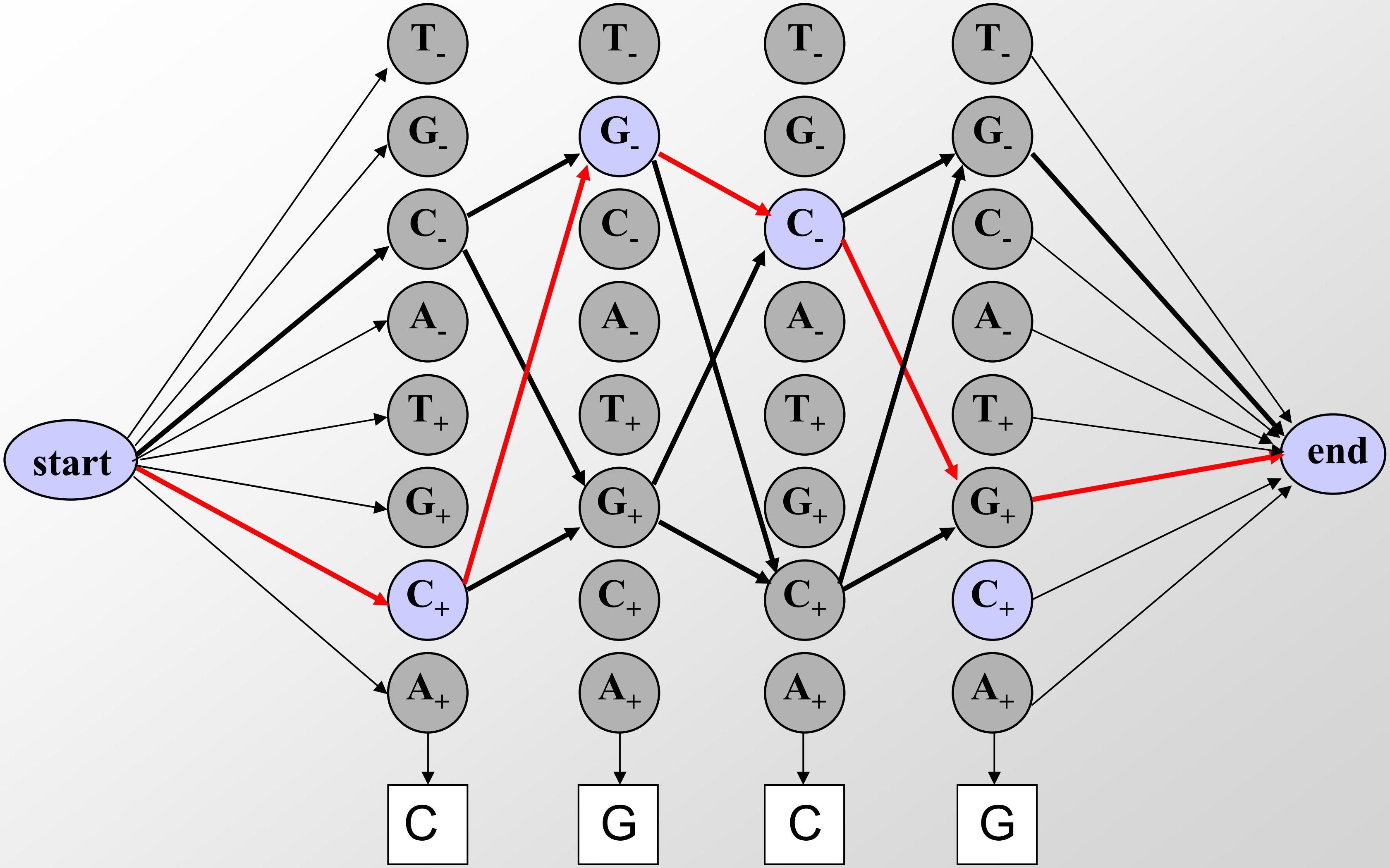
  ➔ 'Paint' the sequence with + and - states

Finding most likely state path



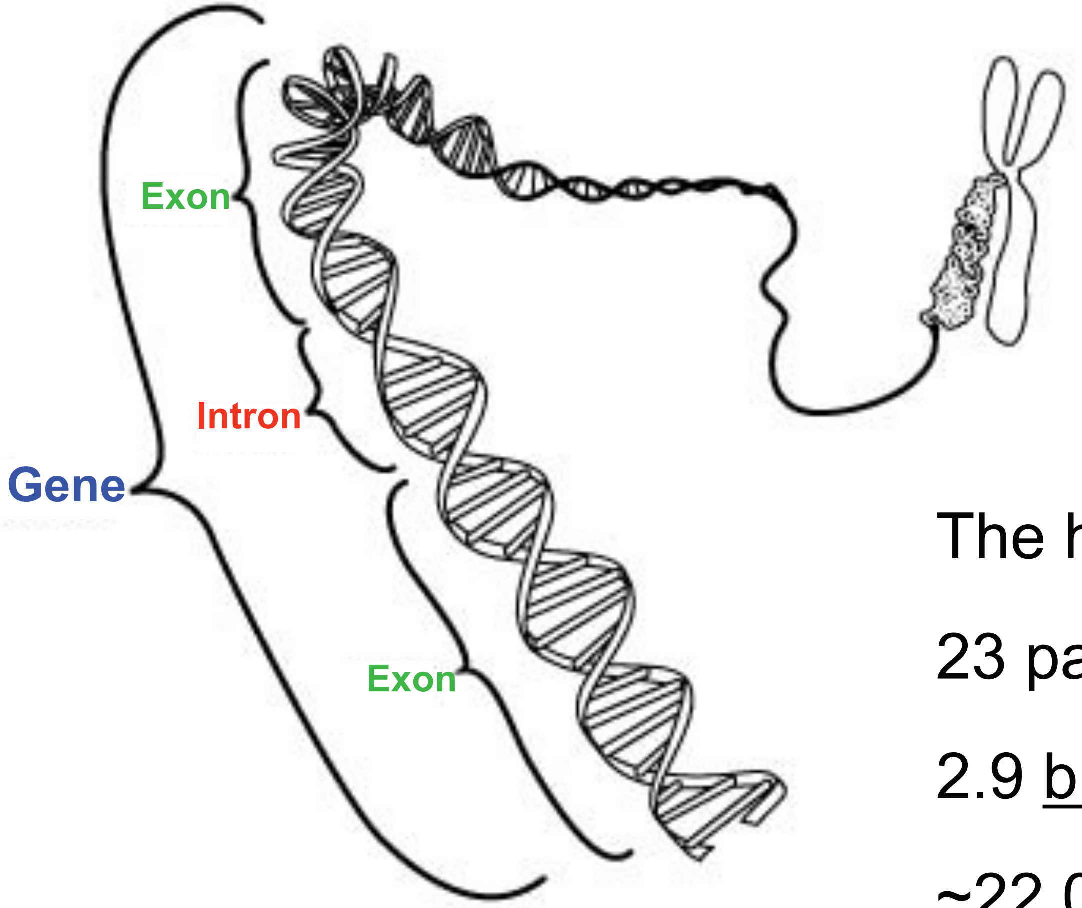
- Given the observed emissions, what was the path?

Probability of given path *p* & observations *x*



- Known observations: CGCG
- Known sequence path: C+, G-, C-, G+

Exons, Introns, and Genes



The human genome:

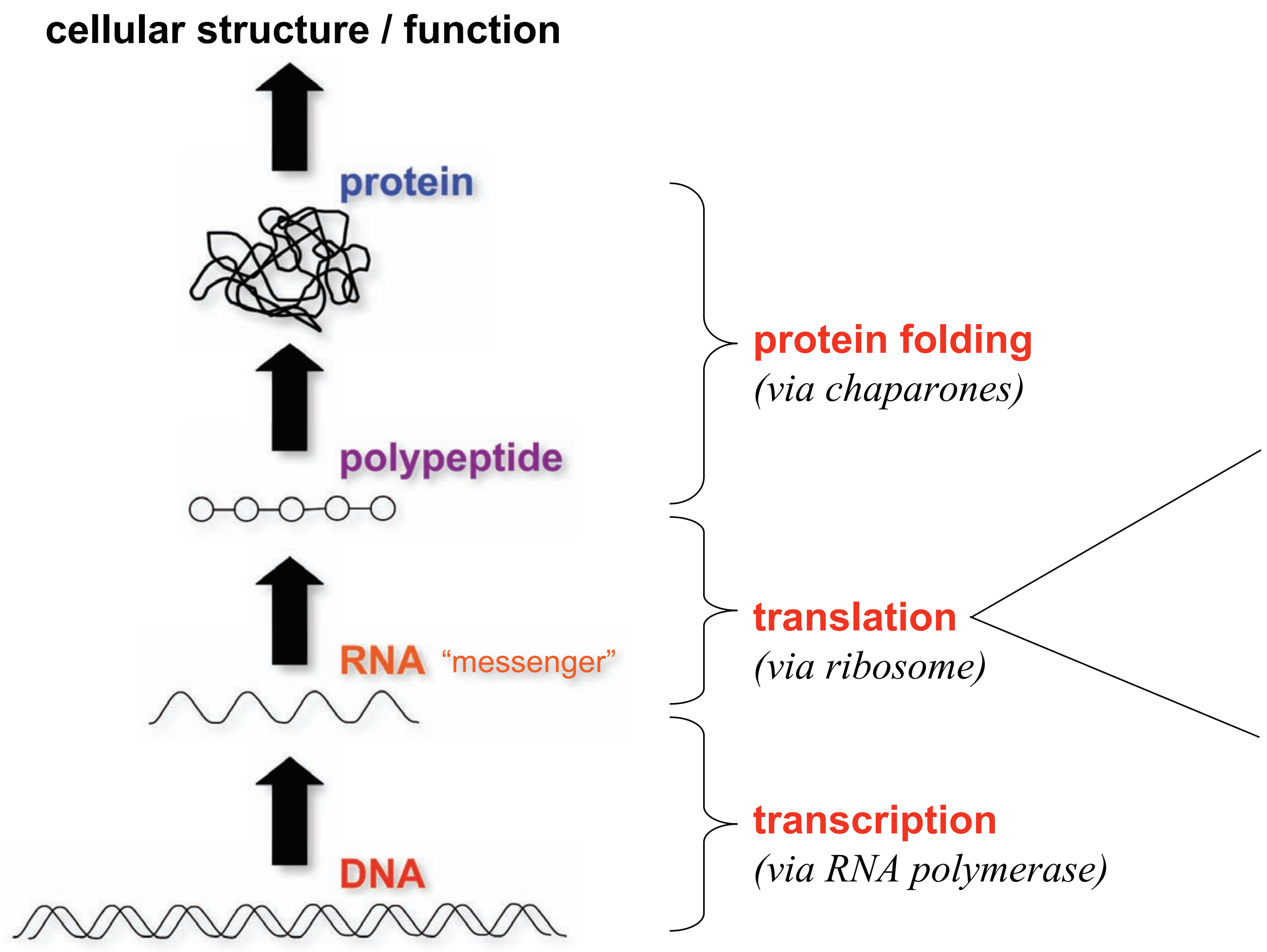
23 pairs of chromosomes

2.9 billion A's, T's, C's, G's

~22,000 genes (?)

~1.4% of genome is coding

The Central Dogma



protein folding
(via chaparones)

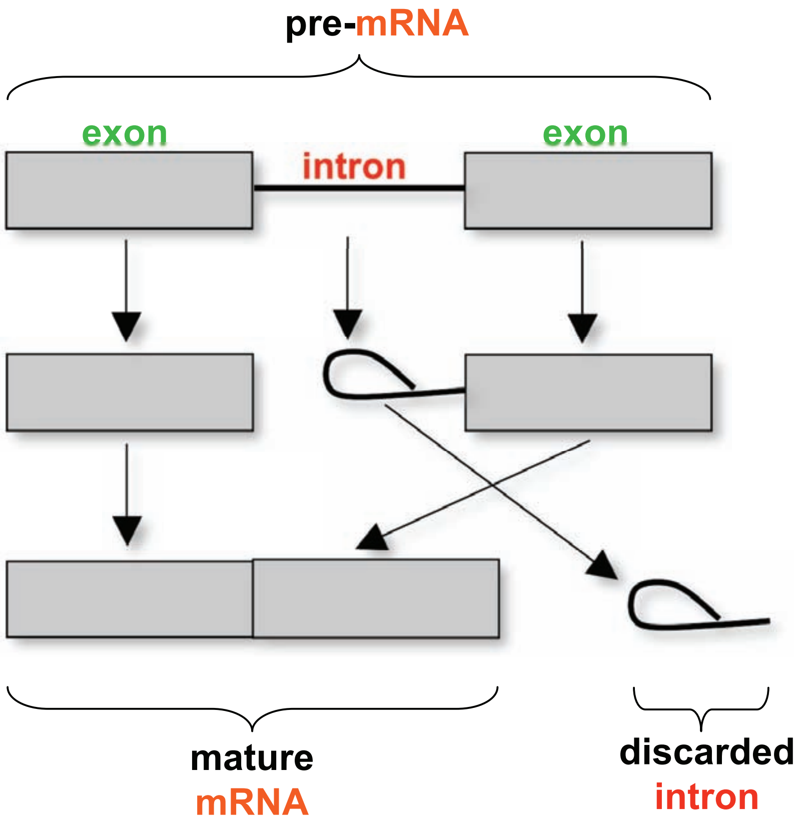
translation
(via ribosome)

transcription
(via RNA polymerase)

RNA		amino acid
AGC	→	S
CGA	→	R
UUR	→	L
GCU	→	A
GUU	→	V
. . .		. . .

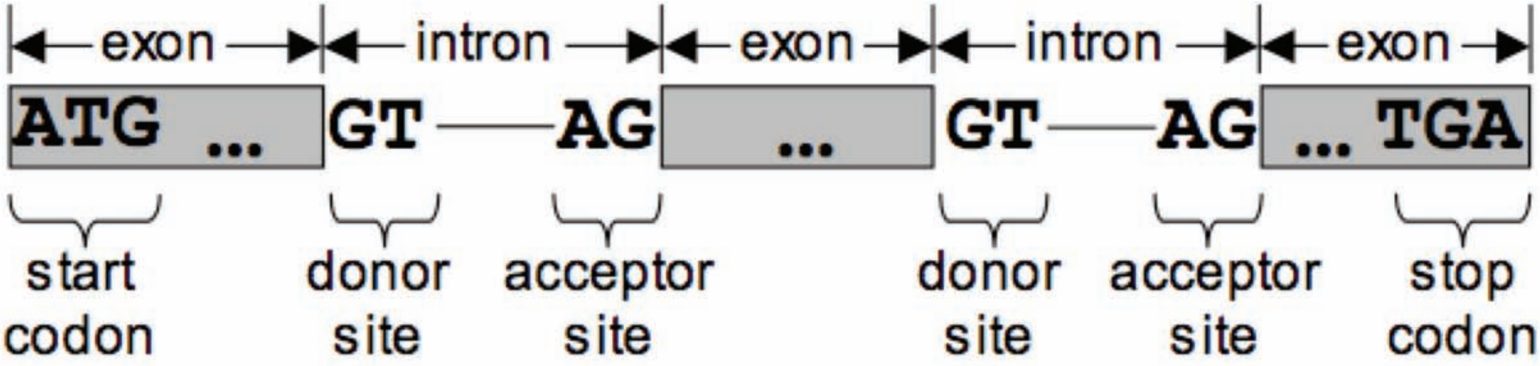
Splicing of Eukaryotic mRNA's

After transcription by the *polymerase*, eukaryotic pre-mRNA's are subject to splicing by the *spliceosome*, which removes introns:



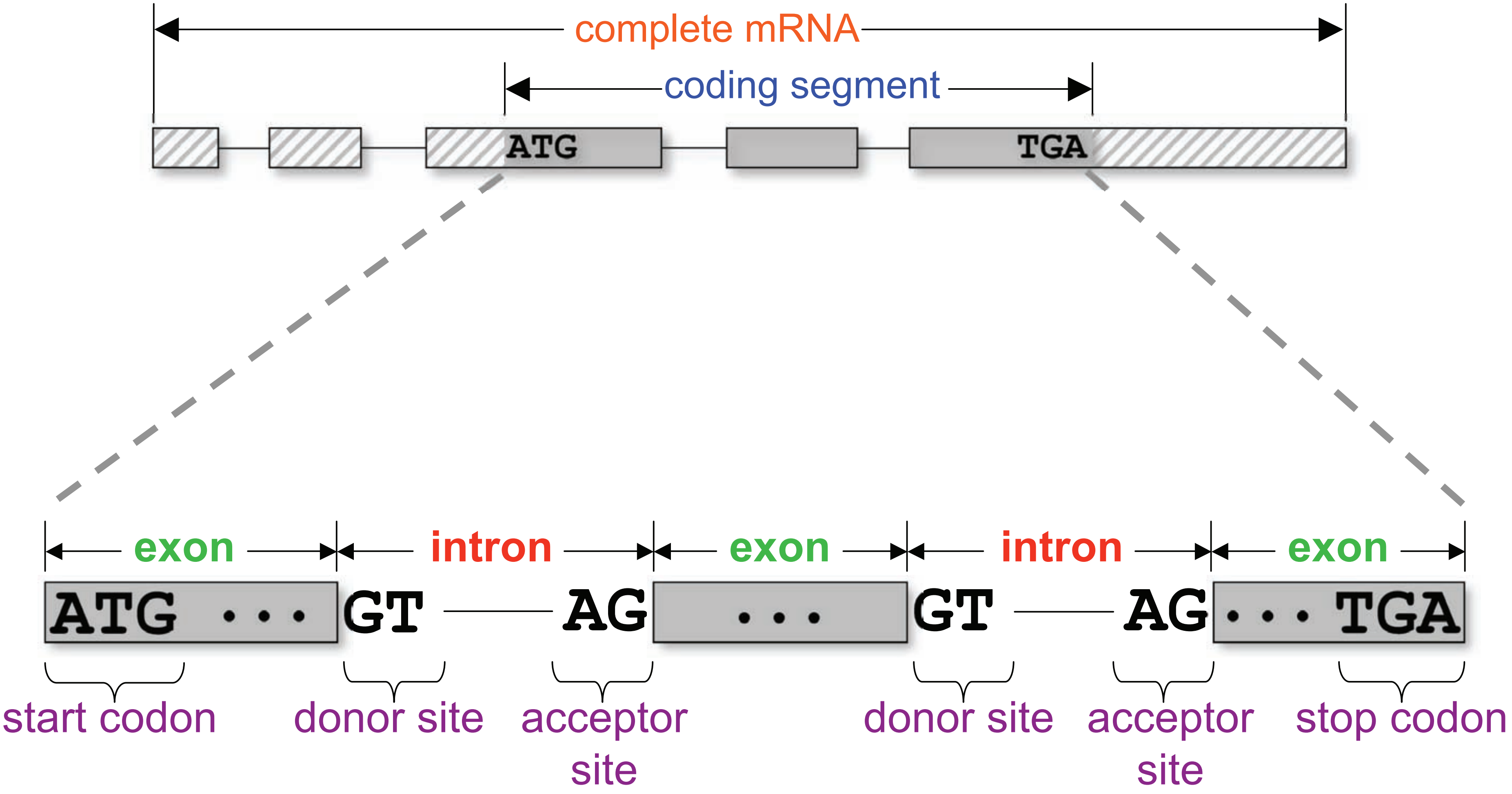
Signals Delimit Gene Features

Coding segments (*CDS*'s) of genes are delimited by four types of signals: *start codons* (ATG in eukaryotes), *stop codons* (usually TAG, TGA, or TAA), *donor sites* (usually GT), and *acceptor sites* (AG):



For initial and final exons, only the coding portion of the exon is generally considered in most of the gene-finding literature; thus, we redefine the word "exon" to include only the coding portions of exons, for convenience.

Eukaryotic Gene Syntax



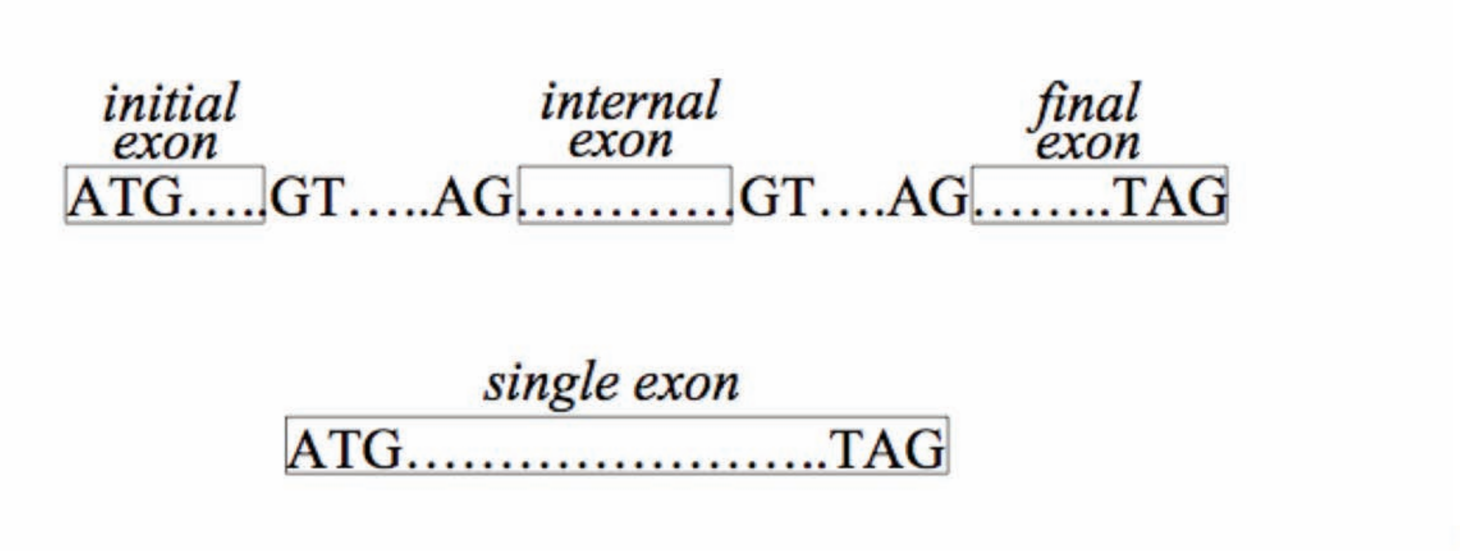
Regions of the gene outside of the CDS are called *UTR*'s (*untranslated regions*), and are mostly ignored by gene finders, though they are important for regulatory functions.

Duke UNIVERSITY

Types of Exons

Three types of exons are defined, for convenience:

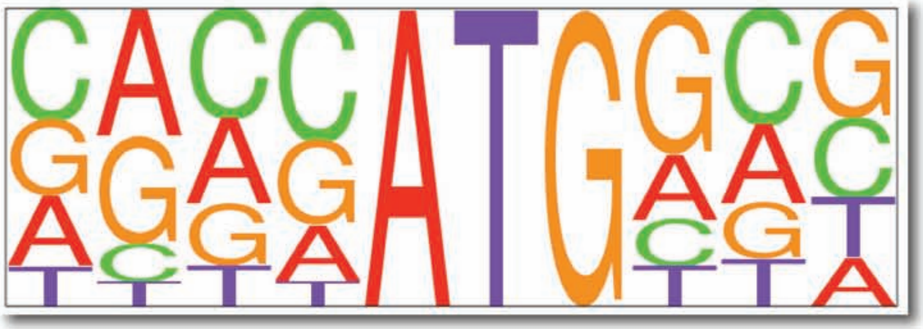
- *initial exons* extend from a start codon to the first donor site;
- *internal exons* extend from one acceptor site to the next donor site;
- *final exons* extend from the last acceptor site to the stop codon;
- *single exons* (which occur only in *intronless genes*) extend from the start codon to the stop codon:



The Stochastic Nature of Signal Motifs

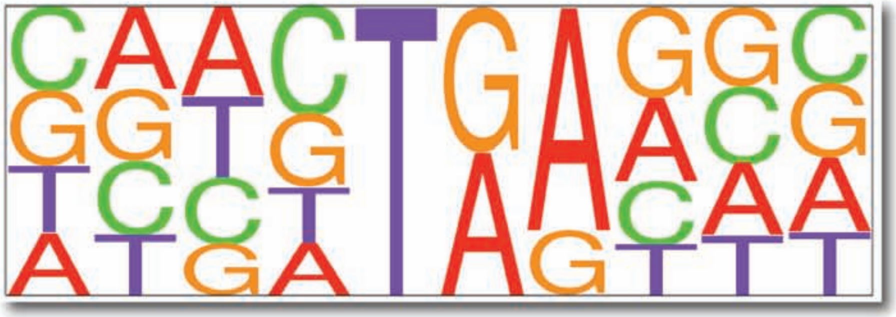
(start codons)

A T G



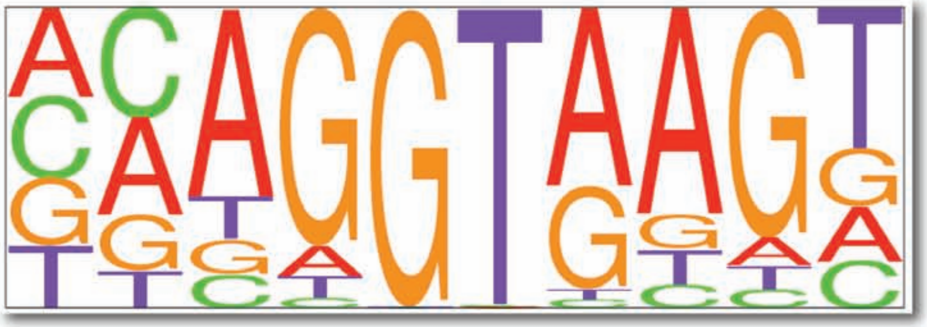
(stop codons)

T G A
T A A
T A G



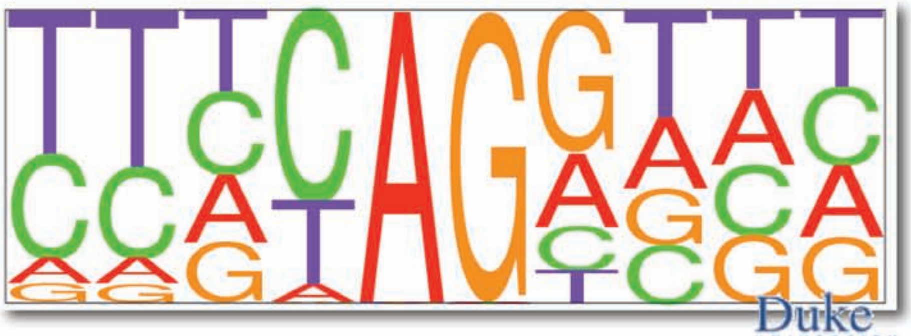
(donor splice sites)

G T



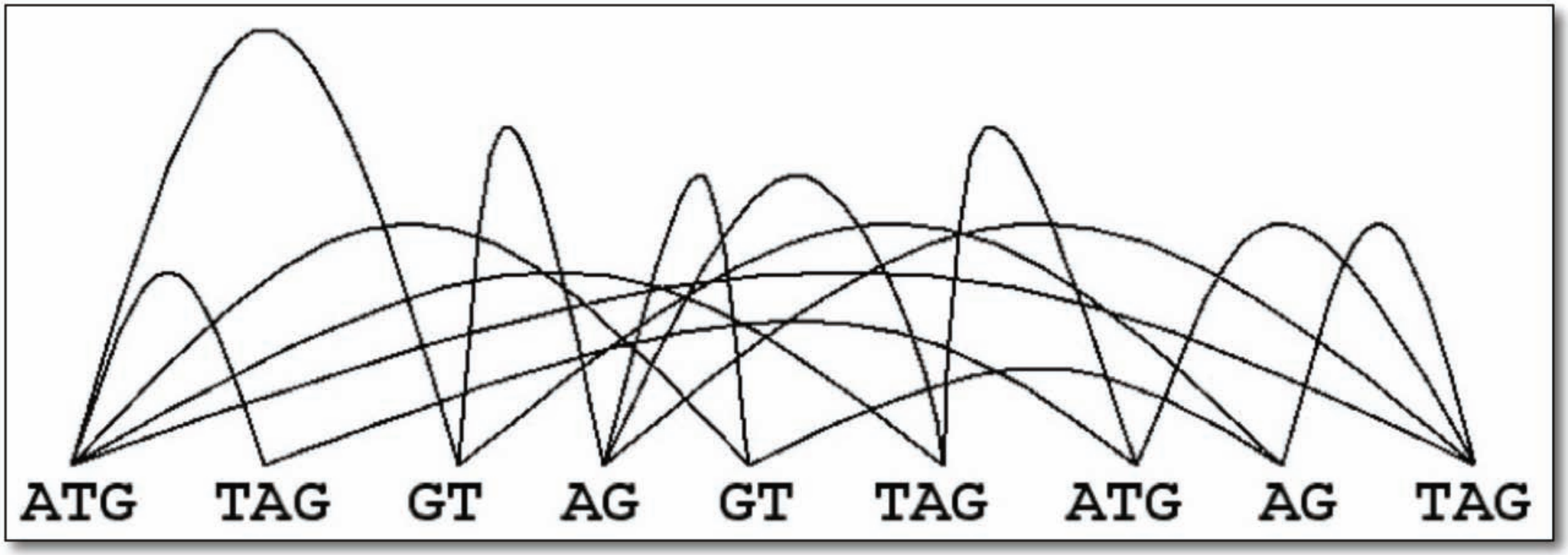
(acceptor splice sites)

A G



Representing Gene Syntax with ORF Graphs

After identifying the most promising (i.e., highest-scoring) signals in an input sequence, we can apply the gene syntax rules to connect these into an *ORF graph*:

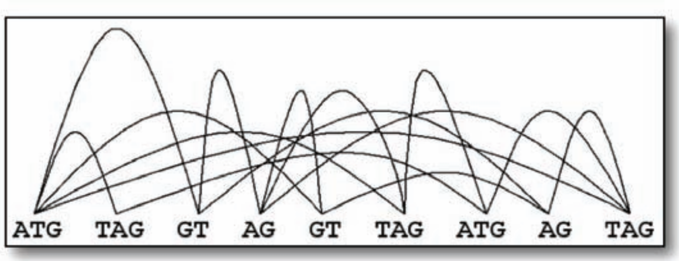


An ORF graph represents all possible *gene parses* (and their scores) for a given set of putative signals. A *path* through the graph represents a single gene parse.

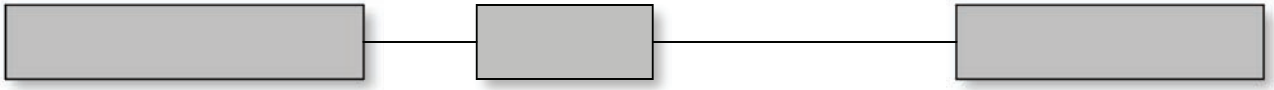
Conceptual Gene-finding Framework

```
TATTCCGATCGATCGATCTCTCTAGCGTCTACG
CTATCATCGCTCTCTATTATCGCGCGATCGTCG
ATCGCGCGAGAGTATGCTACGTCGATCGAATTG
```

identify most promising signals, score signals and content regions between them; induce an ORF graph on the signals



find highest-scoring path through ORF graph; interpret path as a gene parse = gene structure



Recall: “Pure” HMMs

An HMM is a *stochastic machine* M=(Q, α, P_t, P_e) consisting of the following:

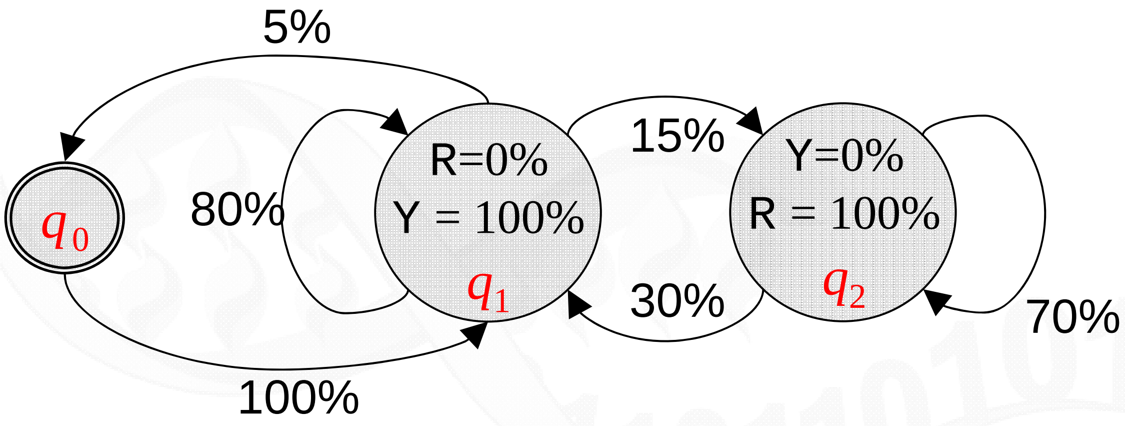
- a finite set of <u>states</u>, $Q=\{q_0, q_1, \ldots, q_m\}$
- a finite <u>alphabet</u> $\alpha=\{s_0, s_1, \ldots, s_n\}$
- a <u>transition</u> distribution $P_t : Q\times Q \rightarrow [0,1]$ i.e., $P_t(q_j \mid q_i)$
- an <u>emission</u> distribution $P_e : Q\times\alpha \rightarrow [0,1]$ i.e., $P_e(s_j \mid q_i)$

An Example

M_1=({q_0,q_1,q_2},{Y,R},P_t,P_e)

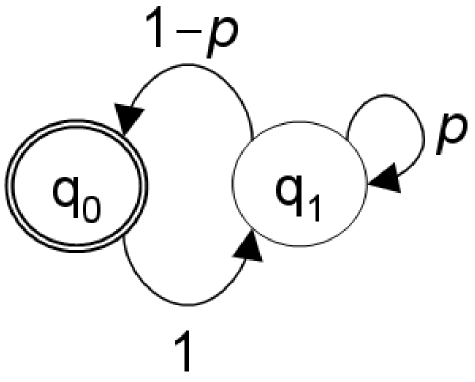
P_t={(q_0,q_1,1), (q_1,q_1,0.8),
(q_1,q_2,0.15), (q_1,q_0,0.05),
(q_2,q_2,0.7), (q_2,q_1,0.3)}

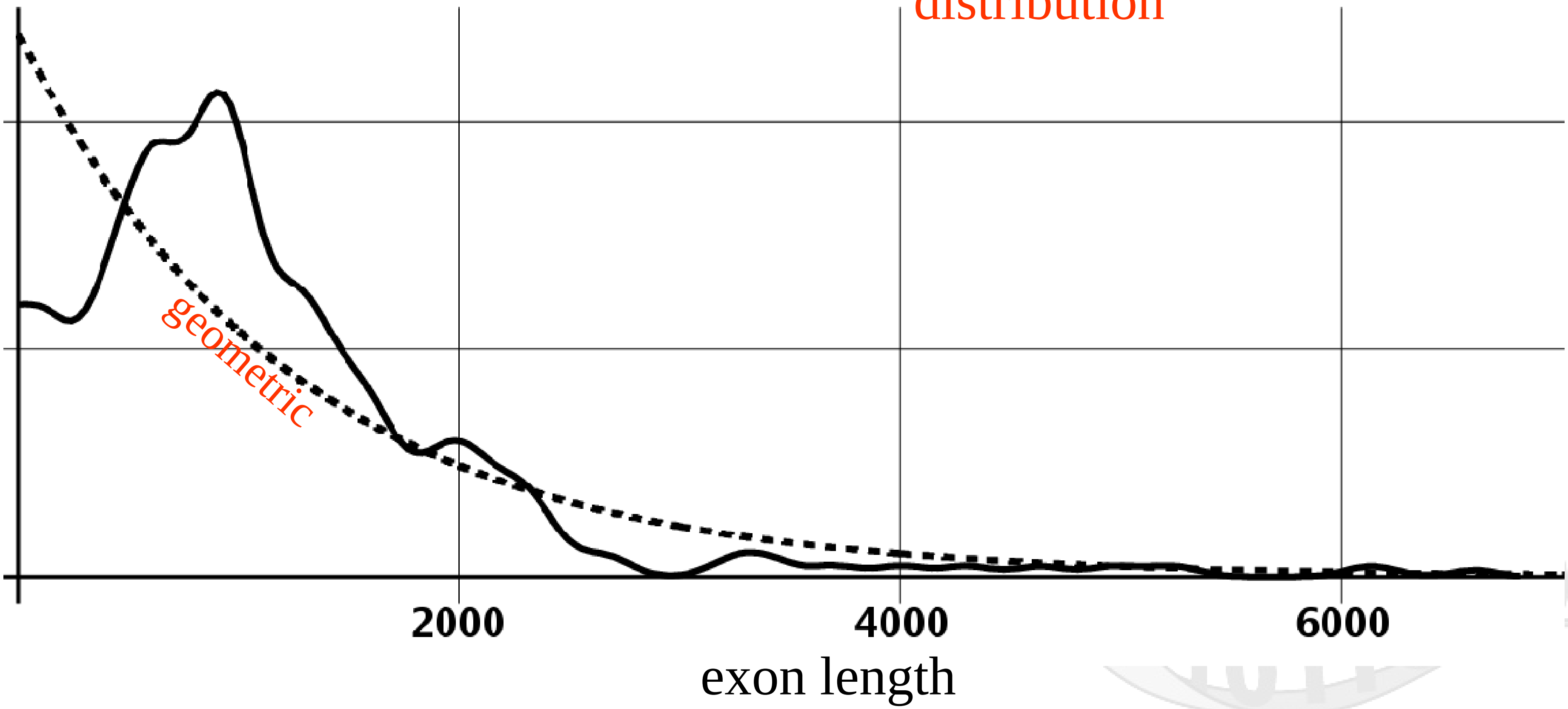
P_e={(q_1,Y,1), (q_1,R,0), (q_2,Y,0), (q_2,R,1)}



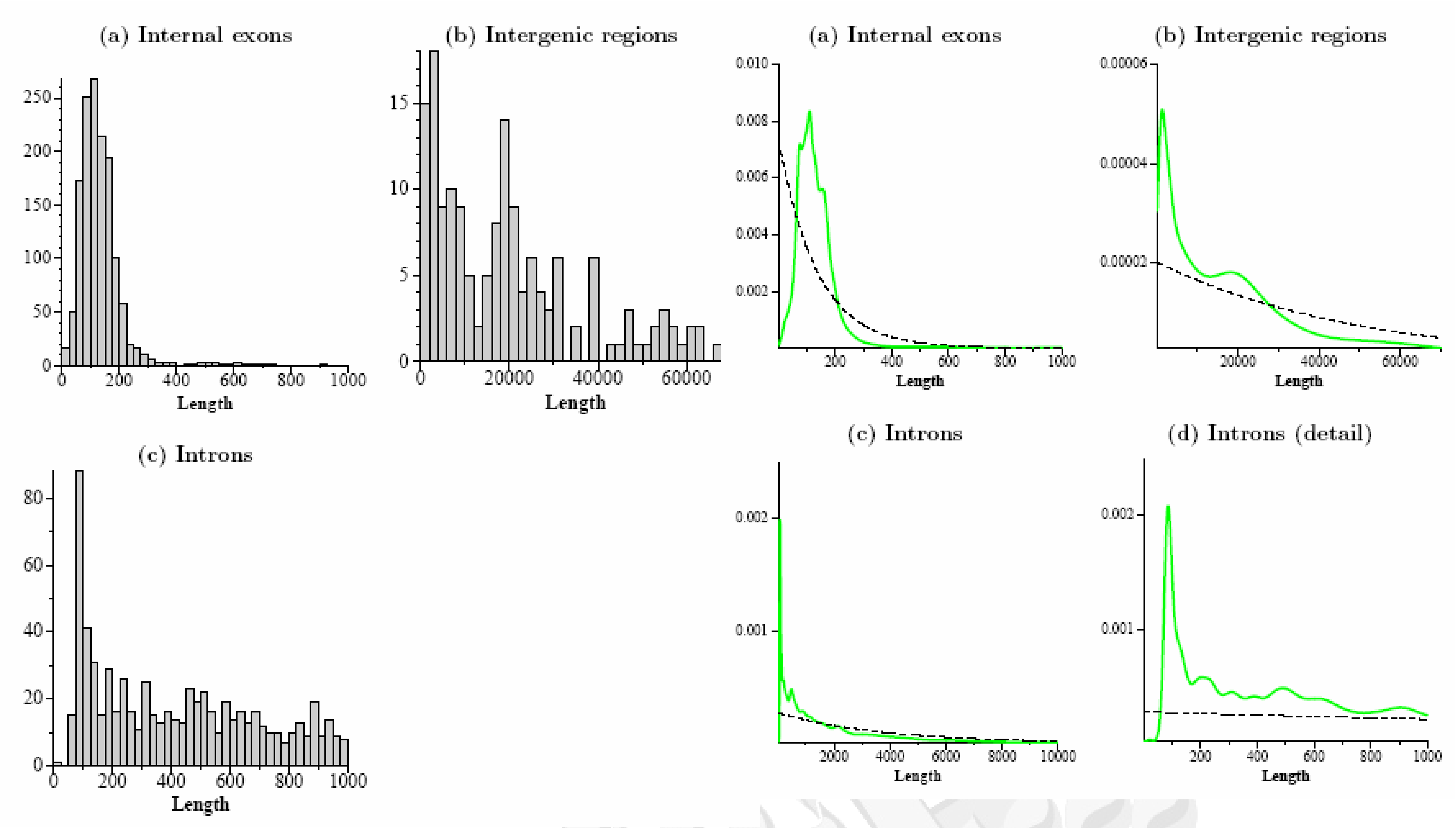
HMMs & Geometric Feature Lengths

$$P(x_0 \ldots x_{d-1} \mid \theta) = \left(\prod_{i=0}^{d-1} P_e(x_i \mid \theta) \right) \underbrace{p^{d-1}(1-p)}_{\text{geometric distribution}}$$



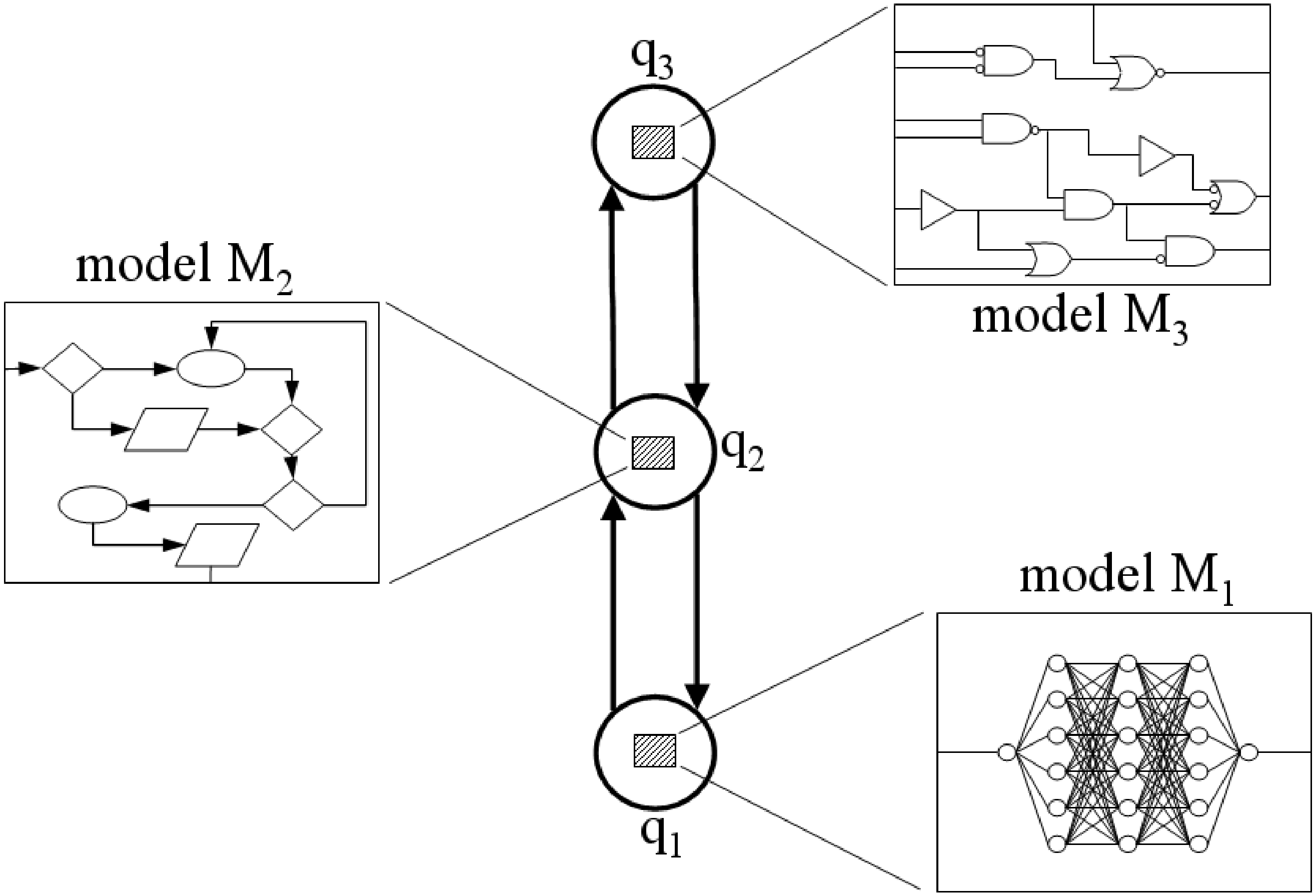


Lengths Distribution in Human



Feature lengths were computed for Human chromosome 22 with RefSeq annotation (as of July 2005).

Generalized Hidden Markov Models



Advantages:

- * Submodel abstraction
- * Architectural simplicity
- * State duration modeling

Disadvantages:

- * Decoding complexity

Generalized HMMs

A GHMM is a stochastic machine M=(Q, α, P_t, P_e, P_d) consisting of the following:

- a finite set of states, $Q=\{q_0, q_1, \ldots, q_m\}$
- a finite alphabet $\alpha=\{s_0, s_1, \ldots, s_n\}$
- a transition distribution $P_t : Q \times Q \rightarrow [0,1]$ i.e., $P_t(q_j \mid q_i)$
- an emission distribution $P_e : Q \times \alpha^* \times \mathrm{N} \rightarrow [0,1]$ i.e., $P_e(s_j \mid q_i, d_j)$
- a duration distribution $P_e : Q \times \mathrm{N} \rightarrow [0,1]$ i.e., $P_d(d_j \mid q_i)$

Key Differences

- each state now emits an entire subsequence rather than just one symbol
- feature lengths are now explicitly modeled, rather than implicitly geometric
- emission probabilities can now be modeled by any arbitrary probabilistic model
- there tend to be far fewer states => simplicity & ease of modification

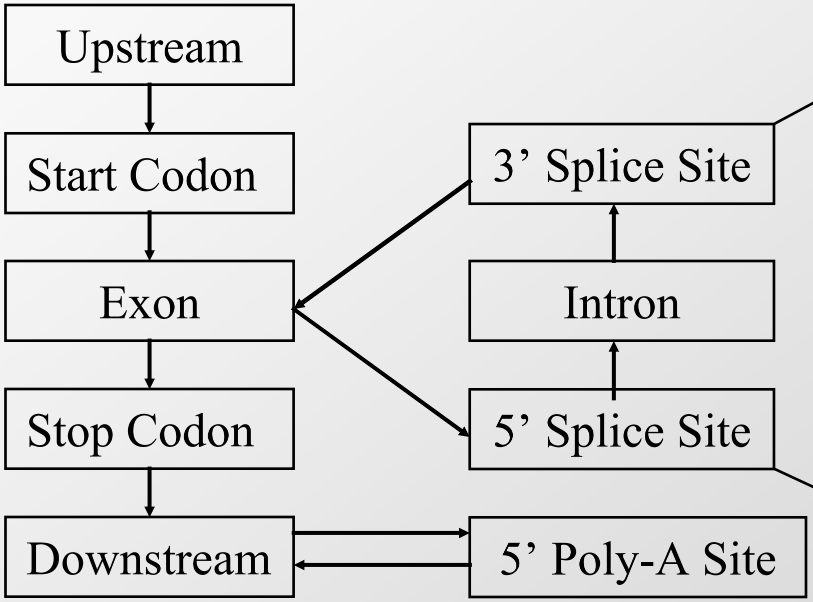
Ref: Kulp D, Haussler D, Reese M, Eeckman F (1996) A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB '96*.

HMM-based Gene Finding

- GENSCAN (Burge 1997)
- FGENESH (Solovyev 1997)
- HMMgene (Krogh 1997)
- GENIE (Kulp 1996)
- GENMARK (Borodovsky & McIninch 1993)
- VEIL (Henderson, Salzberg, & Fasman 1997)

VEIL: Viterbi Exon-Intron Locator

- Contains 9 hidden states or features
- Each state is a complex internal Markovian model of the feature
- Features:
  - Exons, introns, intergenic regions, splice sites, etc.



Upstream

Start Codon

Exon

Stop Codon

Downstream

3’ Splice Site

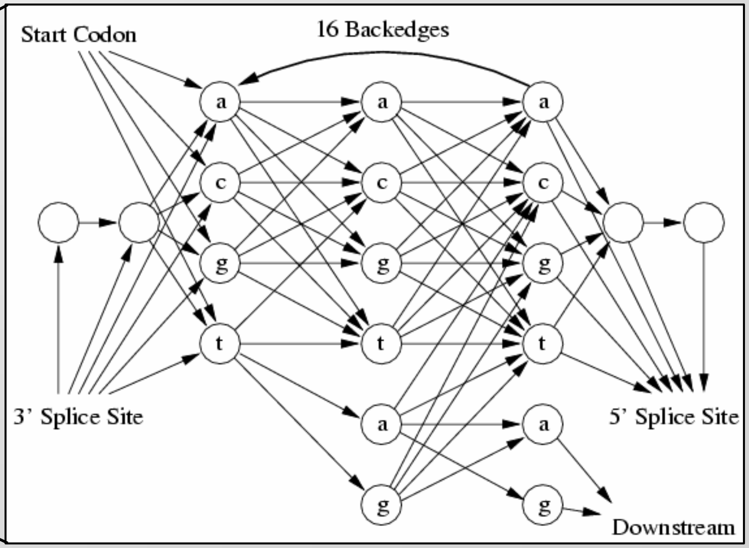
Intron

5’ Splice Site

5’ Poly-A Site

VEIL Architecture

Exon HMM Model

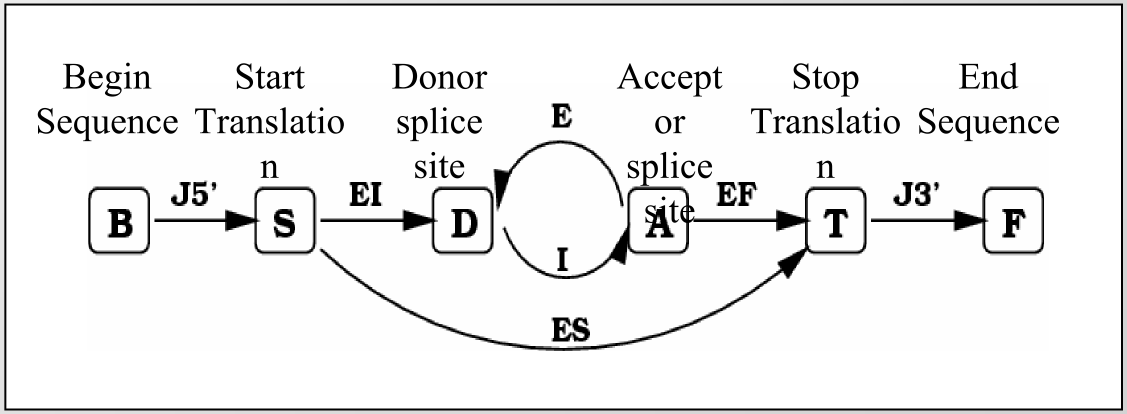


- Enter: start codon or intron (3’ Splice Site)
- Exit: 5’ Splice site or three stop codons (taa, tag, tga)

Genie

- Uses a generalized HMM (GHMM)
- Edges in model are complete HMMs
- States can be any arbitrary program
- States are actually neural networks specially designed for signal finding

- J5’ – 5’ UTR
- EI – Initial Exon
- E – Exon, Internal Exon
- I – Intron
- EF – Final Exon
- ES – Single Exon
- J3’ – 3’UTR



Genscan Overview

- Developed by Chris Burge (Burge 1997), in the research group of Samuel Karlin, Dept of Mathematics, Stanford Univ.
- Characteristics:
  - Designed to predict complete gene structures
    - Introns and exons, Promoter sites, Polyadenylation signals
  - Incorporates:
    - Descriptions of transcriptional, translational and splicing signal
    - Length distributions (Explicit State Duration HMMs)
    - Compositional features of exons, introns, intergenic, C+G regions
  - Larger predictive scope
    - Deal w/ partial and complete genes
    - Multiple genes separated by intergenic DNA in a seq
    - Consistent sets of genes on either/both DNA strands
- Based on a general probabilistic model of genomic sequences composition and gene structure

Genscan Architecture

- It is based on Generalized HMM (GHMM)
- Model both strands at once
  - Other models: Predict on one strand first, then on the other strand
  - Avoids prediction of overlapping genes on the two strands (rare)
- Each state may output a string of symbols (according to some probability distribution).
- Explicit intron/exon length modeling
- Special sensors for Cap-site and TATA-box
- Advanced splice site sensors

Image removed due to copyright restrictions.

Fig. 3, Burge and Karlin 1997

GenScan States

- N - intergenic region
- P - promoter
- F - 5' untranslated region
- E_{sngl} – single exon (intronless) (translation start -> stop codon)
- E_{init} – initial exon (translation start -> donor splice site)
- E_k – phase k internal exon (acceptor splice site -> donor splice site)
- E_{term} – terminal exon (acceptor splice site -> stop codon)
- I_k – phase k intron: 0 – between codons; 1 – after the first base of a codon; 2 – after the second base of a codon

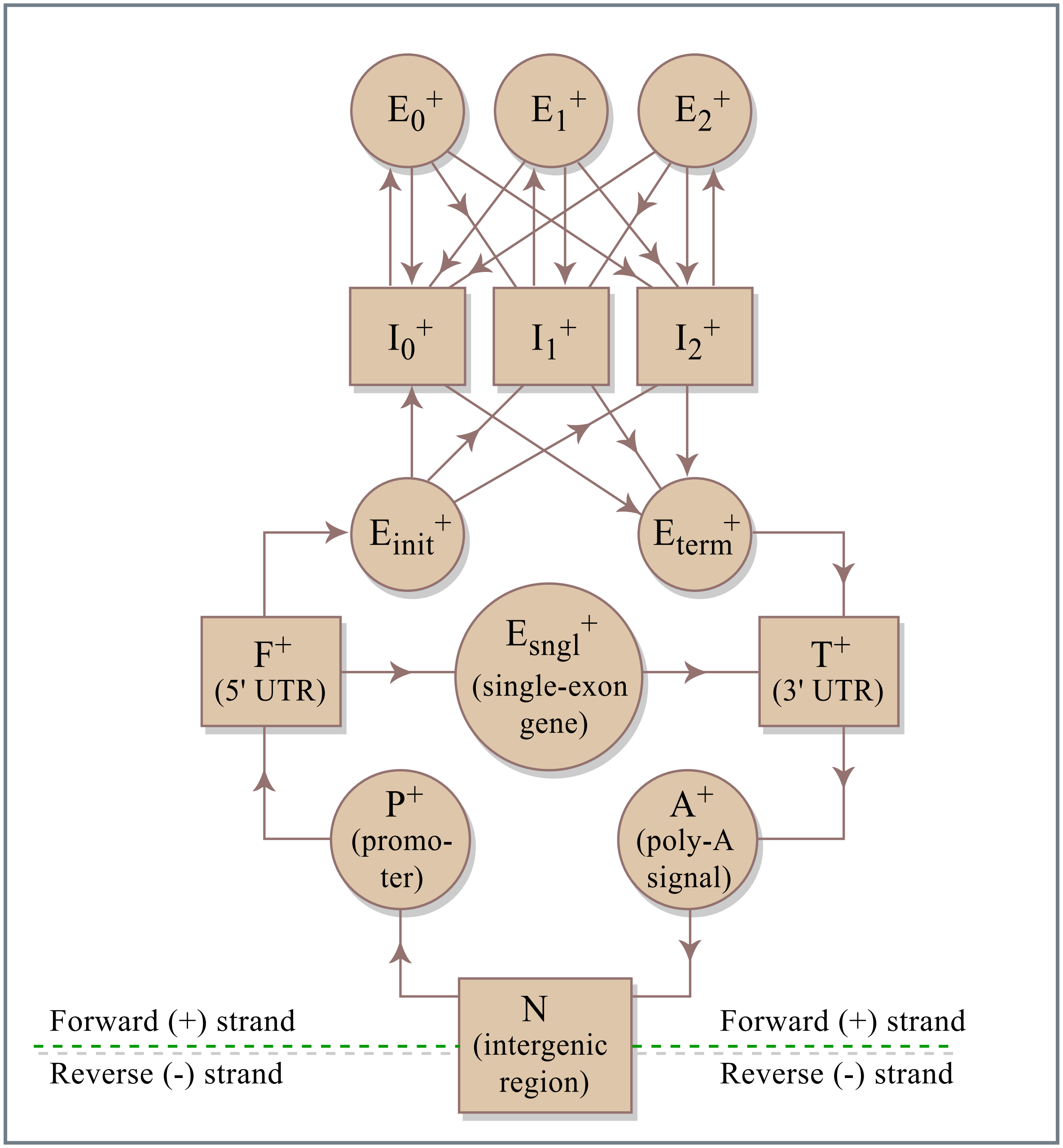


Figure by MIT OCW.

Acknowledgement

- Slides are due to Manolis Kellis and William Majoros.