Eukaryotic Gene Finding

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2018

Anthony Gitter

gitter@biostat.wisc.edu

Goals for Lecture

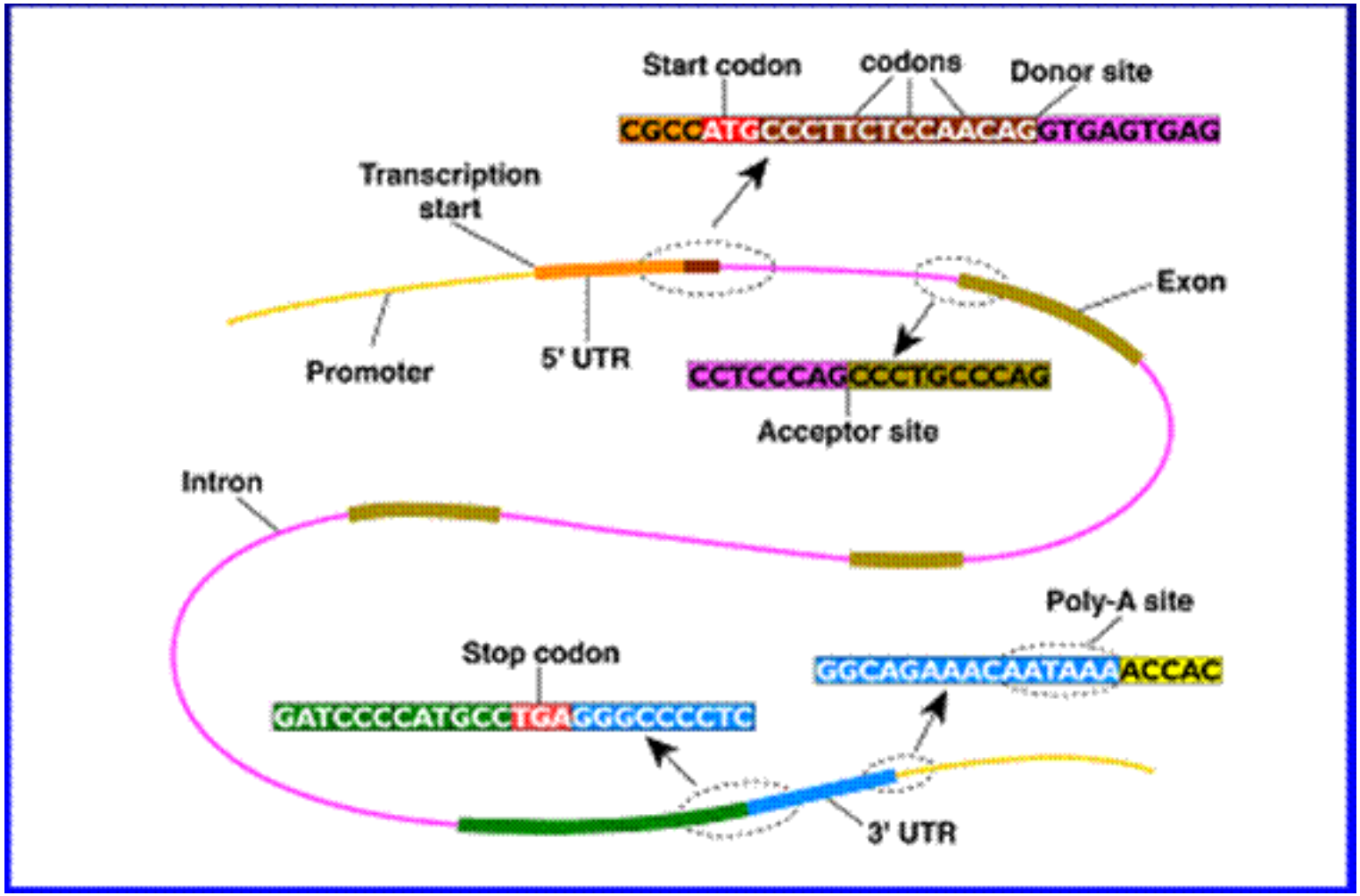
Key concepts

- Incorporating sequence signals into gene finding with HMMs
- Modeling durations with generalized HMMs
- Modeling conversation with pair HMMs
- Modern gene finding and genome annotation

Sources of Evidence for Gene Finding

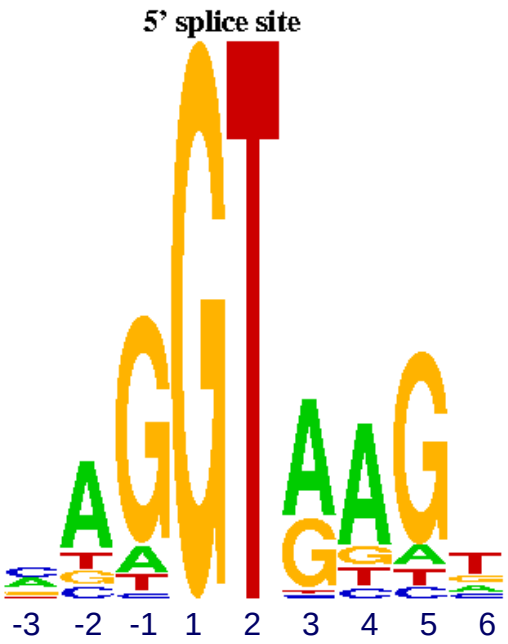
- **Signals**: the sequence *signals* (e.g. splice junctions) involved in gene expression
- **Content**: statistical properties that distinguish protein-coding DNA from non-coding DNA
- **Conservation**: signal and content properties that are conserved across related sequences (e.g. orthologous regions of the mouse and human genome)

Eukaryotic Gene Structure



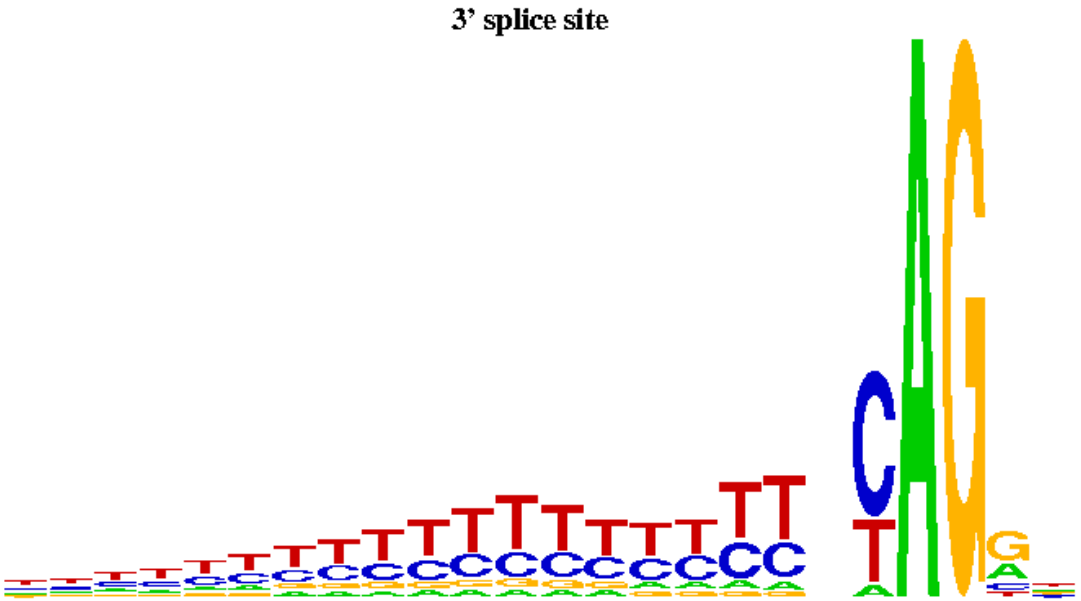
Splice Signals Example

donor sites



exon

acceptor sites



exon

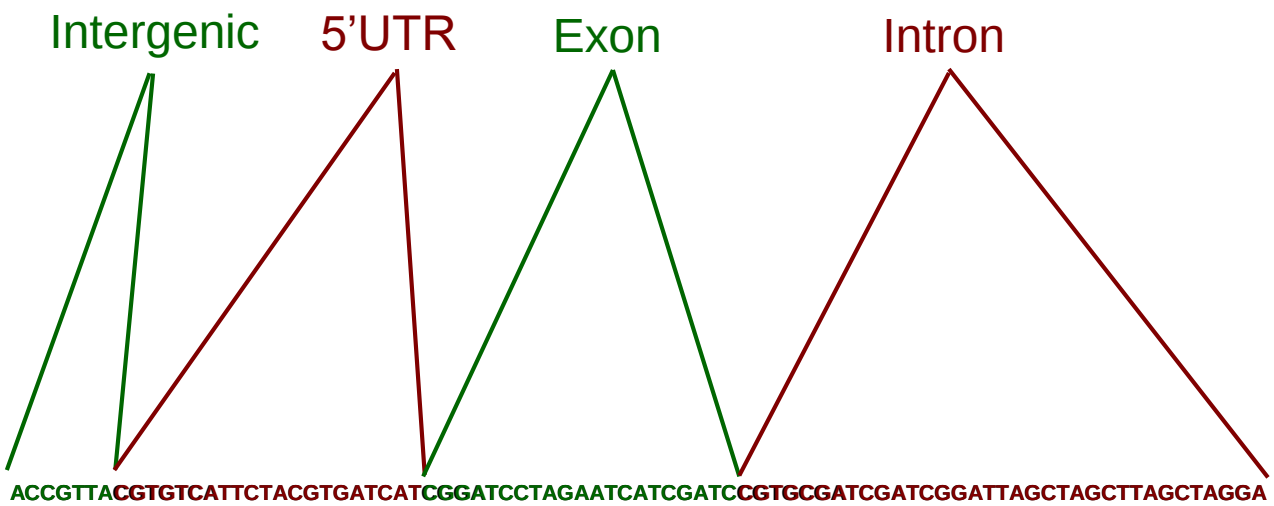
Figures from Yi Xing

- There are significant dependencies among non-adjacent positions in donor splice signals
- Informative for inferring hidden state of HMM

Parsing a DNA Sequence

- The HMM Viterbi path represents a parse of a given sequence, predicts exons, acceptor sites, introns, etc.

Hidden state

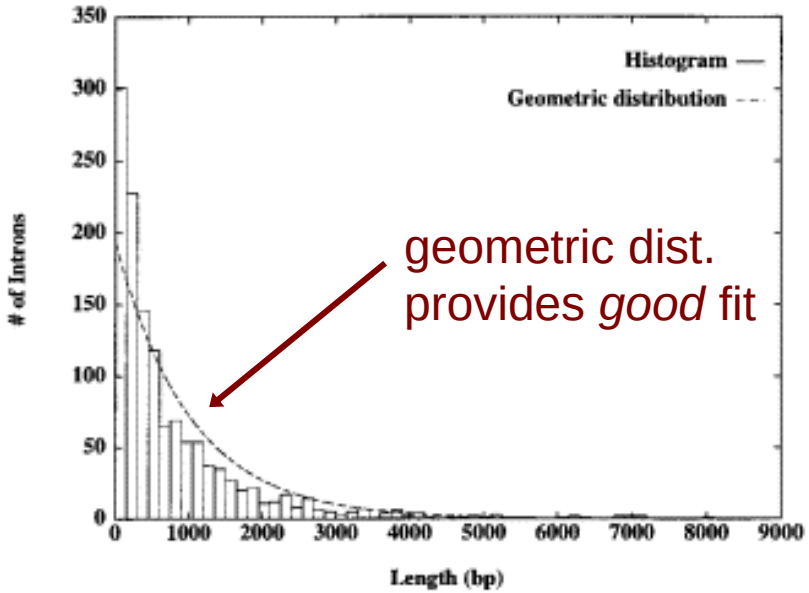


Observed sequence

- How can we properly model the transitions from one state to another?

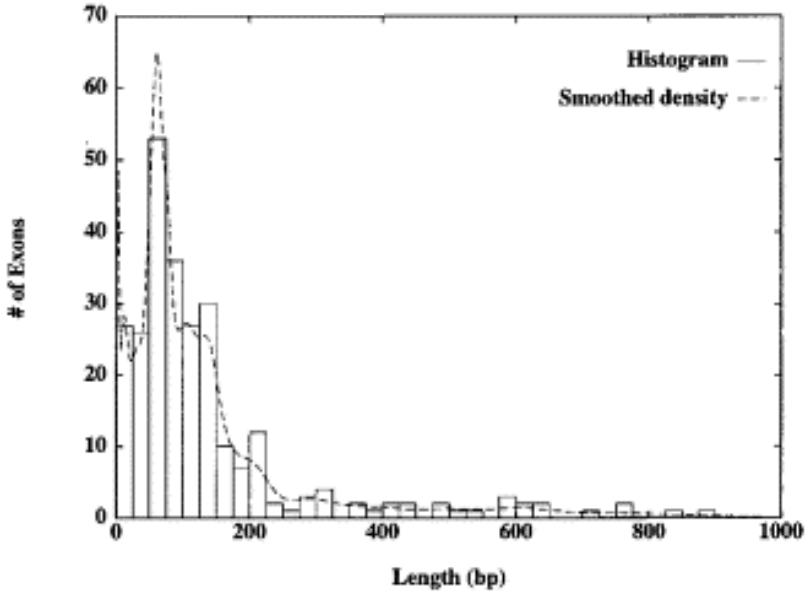
Length Distributions of Introns/Exons

Introns

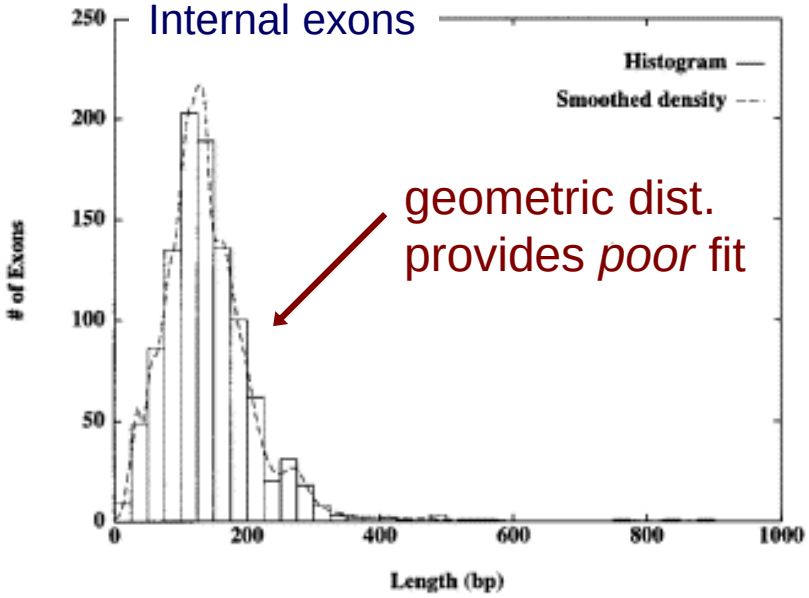


geometric dist. provides *good* fit

Initial exons



Internal exons



geometric dist. provides *poor* fit

Terminal exons

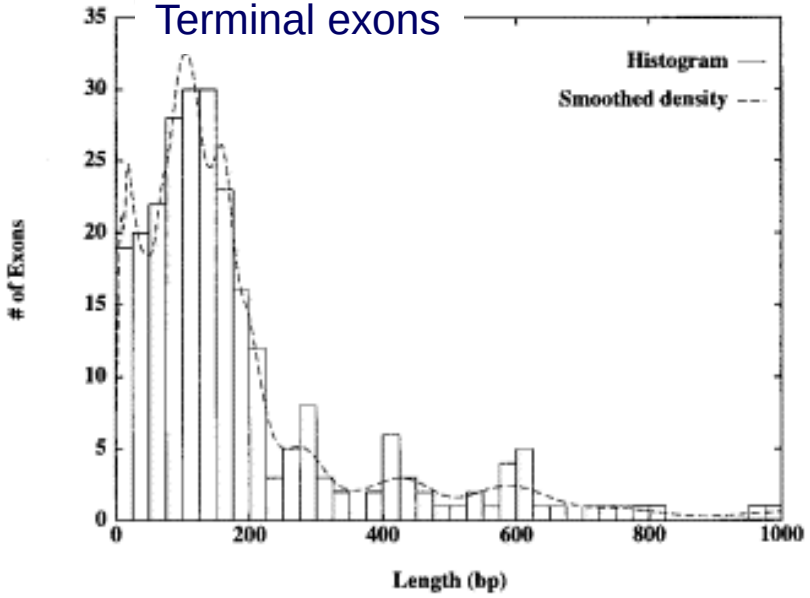


Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

Duration Modeling in HMMs

- Semi-Markov models are well-motivated for some sequence elements (e.g. exons)
  - **Semi-Markov**: explicitly model length duration of hidden states
  - Also called generalized hidden Markov model

The GENSCAN HMM for Eukaryotic Gene Finding [Burge & Karlin '97]

Each shape represents a functional unit of a gene or genomic region

Pairs of intron/exon units represent the different ways an intron can interrupt a coding sequence (after 1st base in codon, after 2nd base or after 3rd base)

Complementary submodel (not shown) detects genes on opposite DNA strand

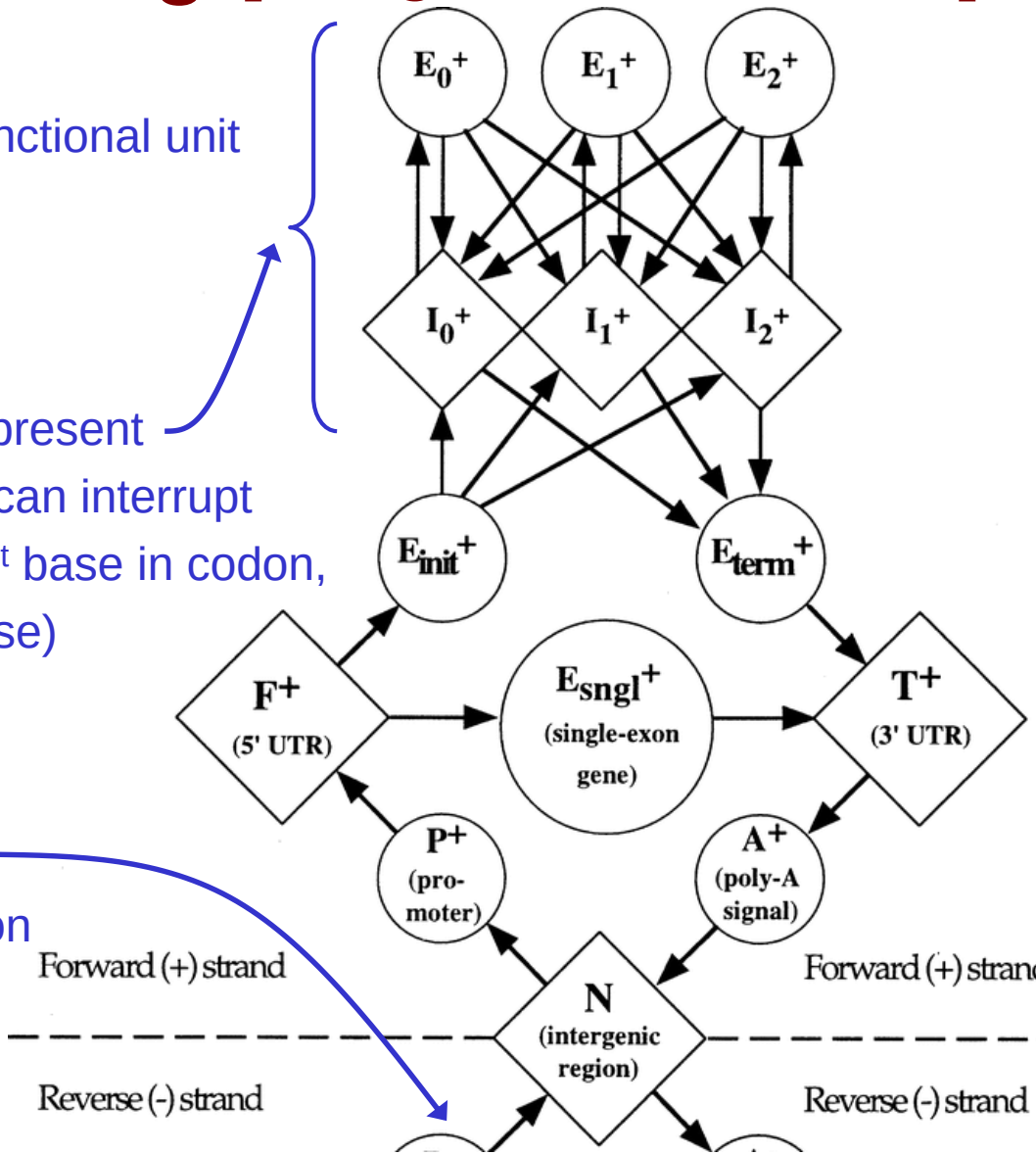
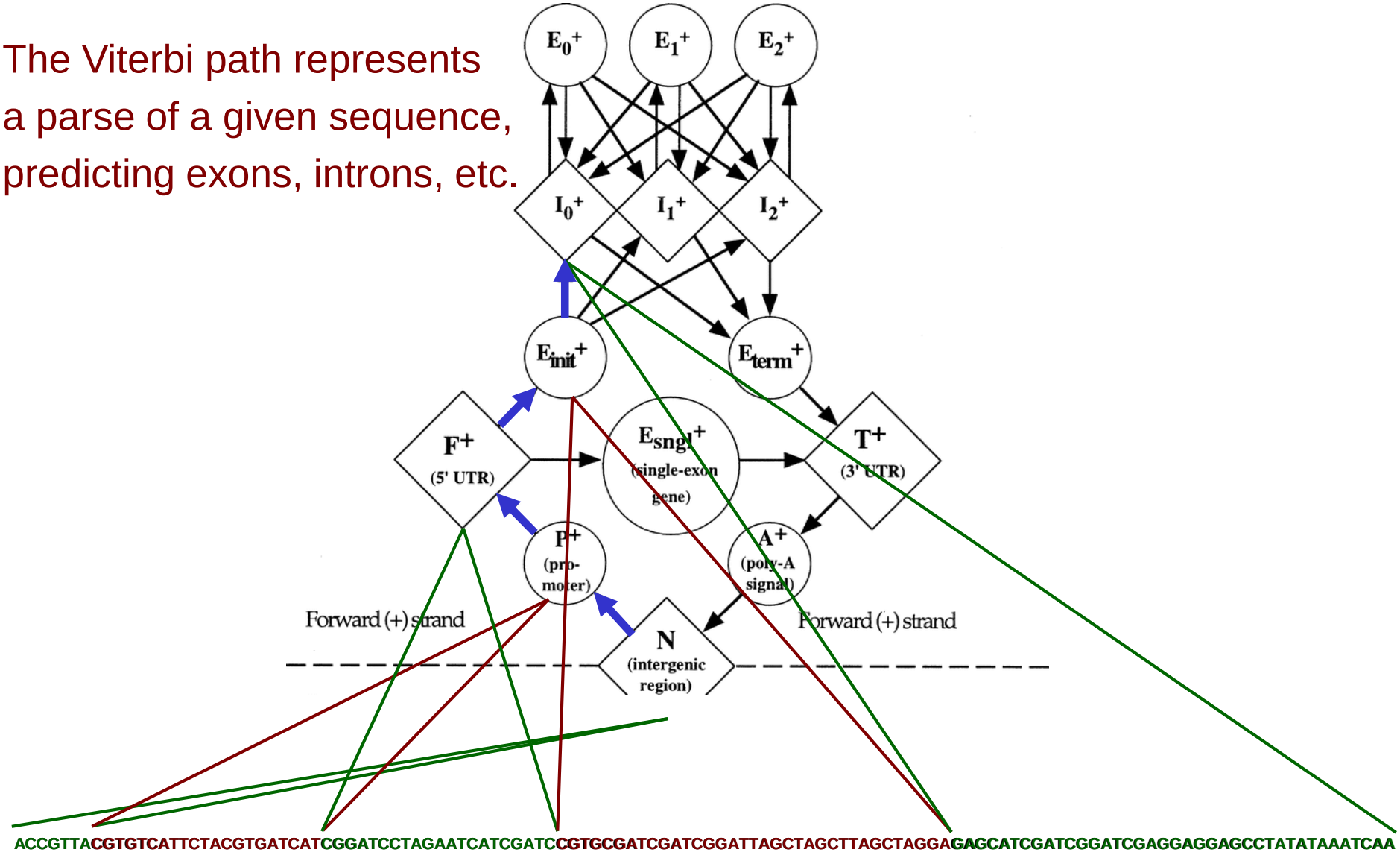


Figure from Burge & Karlin, *Journal of Molecular Biology*, 1997

Parsing a DNA Sequence

The Viterbi path represents a parse of a given sequence, predicting exons, introns, etc.

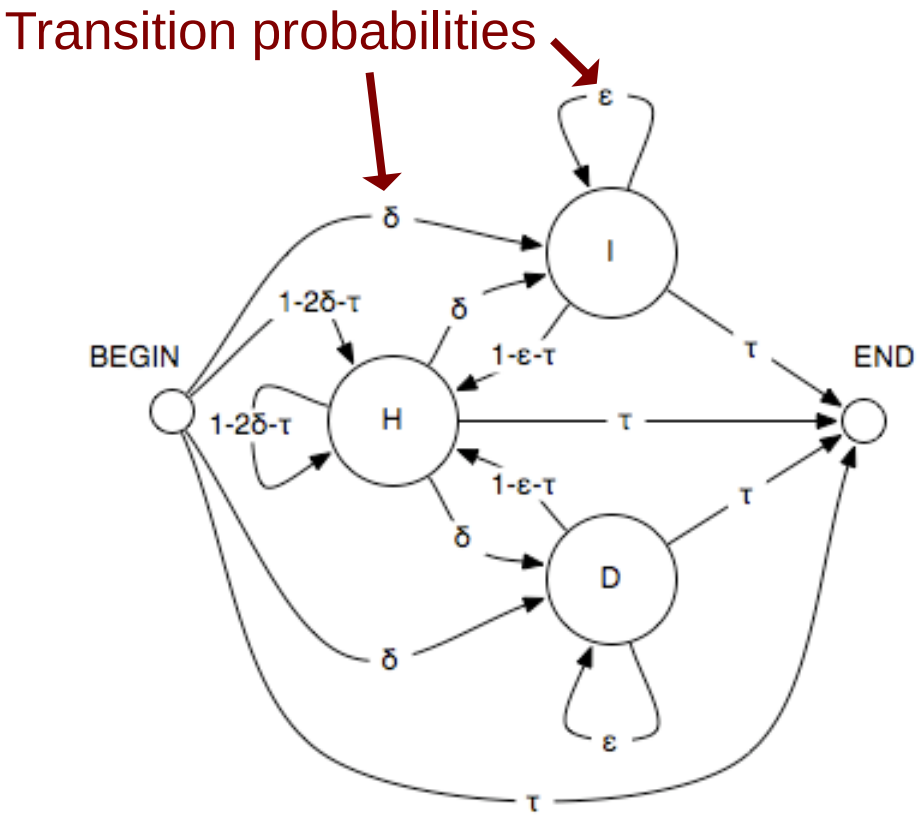


Comparative Algorithms

- Genes are among the most conserved elements in the genome
  - use conservation to help infer locations of genes
- Some signals associated with genes are short and occur frequently in the genome
  - use conservation to eliminate false candidate sites from consideration

Pair Hidden Markov Models

- Each non-silent state emits one or a pair of characters

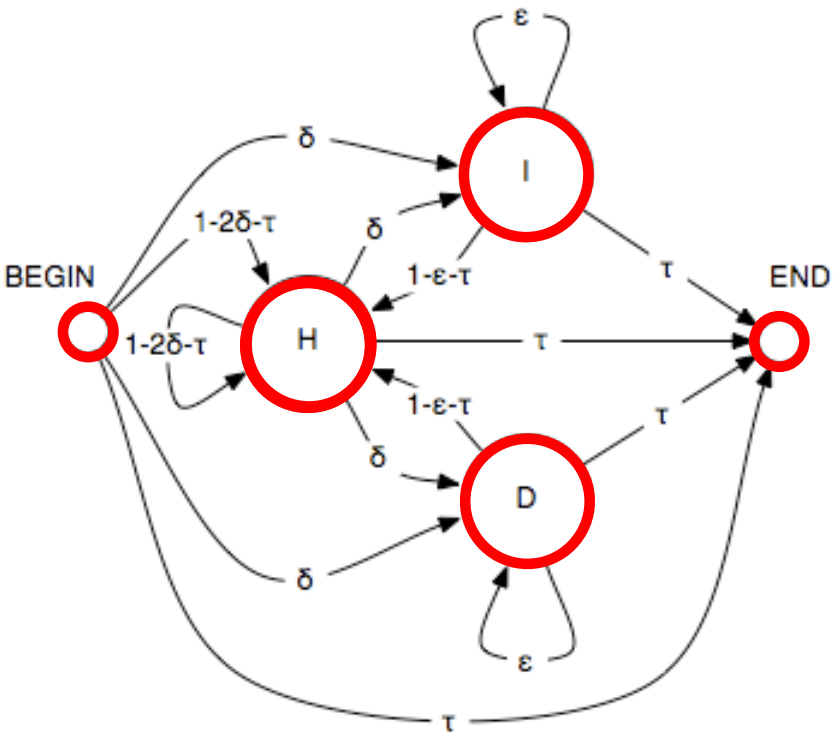


H: homology (match) state

I: insert state

D: delete state

Pair HMM Paths are Alignments



sequence 1: **AAGCGC**
sequence 2: **ATGTC**

hidden: **B H H I I H D H E**
observed: **A A G C G C**
A T G T C

Generalized Pair HMMs

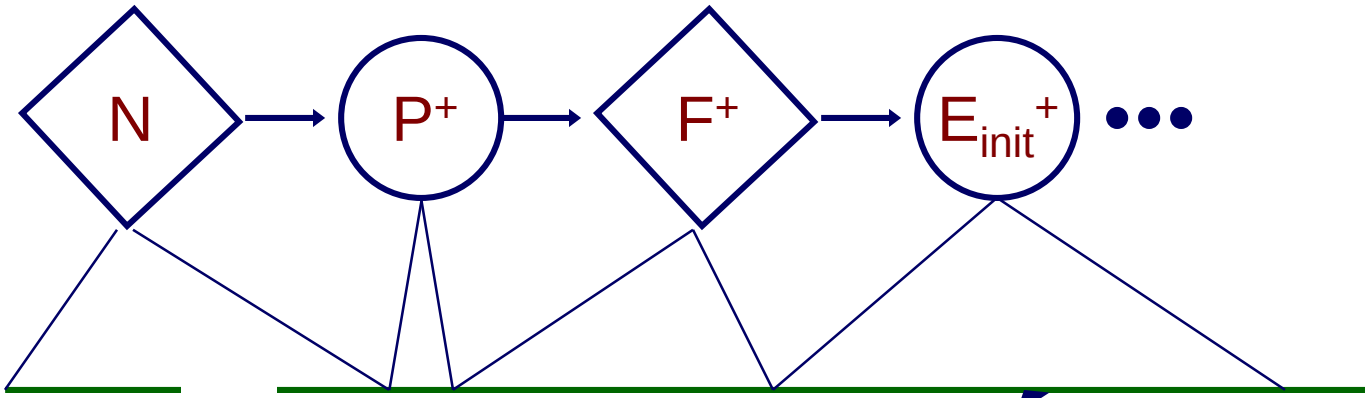
- Represent a parse π, as a sequence of states and a sequence of associated lengths for each input sequence

sequence of hidden states

$\vec{q} = \{q_1, q_2, \ldots, q_n\}$

$\vec{d} = \{d_1, d_2, \ldots, d_n\}$

$\vec{e} = \{e_1, e_2, \ldots, e_n\}$



pair of duration times generated by hidden state

may be gaps in the sequences

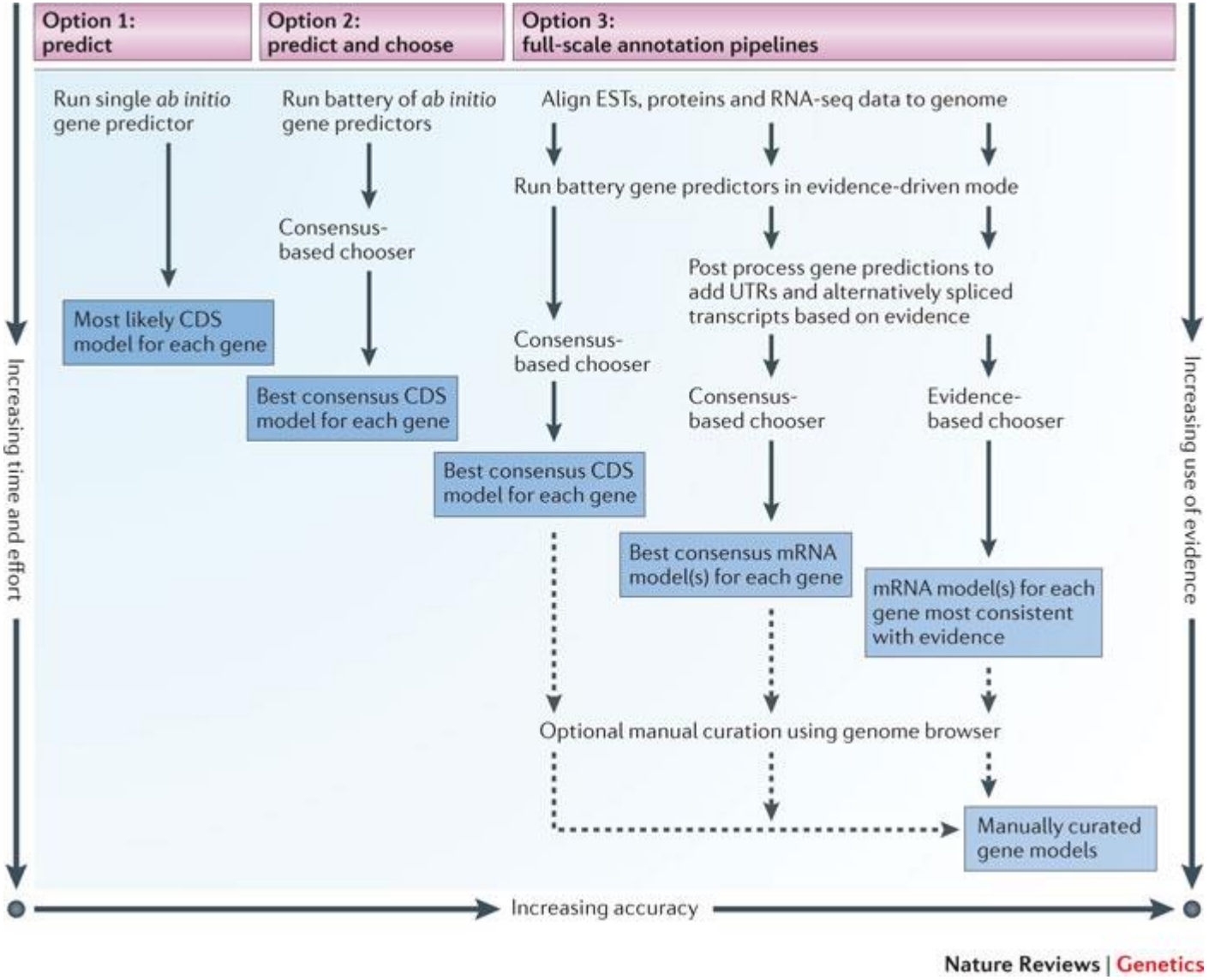
pair of sequences generated by hidden state

SLAM: Pachter et al. *RECOMB* 2001

Modern Genome Annotation

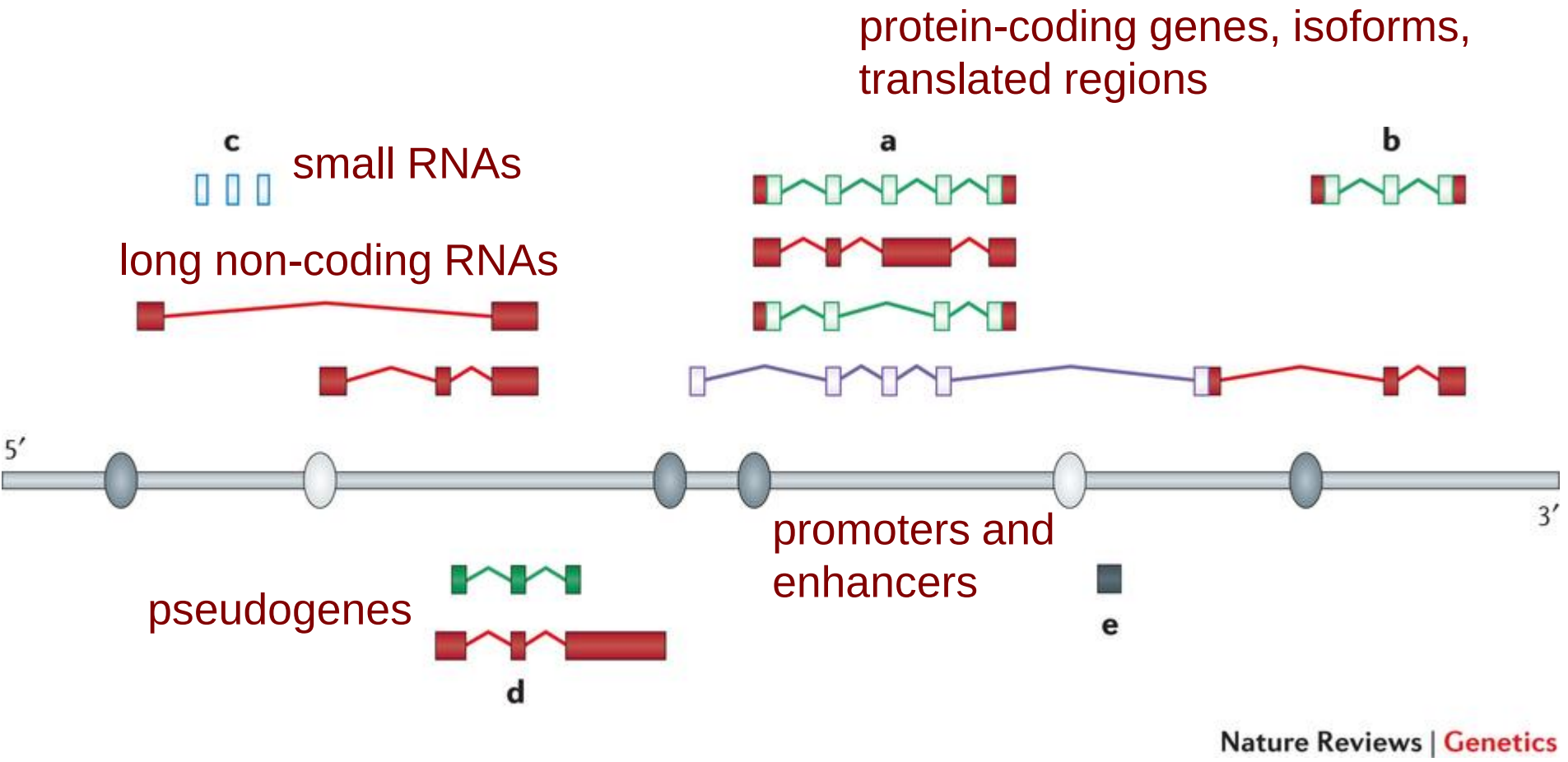
- RNA-Seq, mass spectrometry, and other technologies provide powerful information for genome annotation

Modern Genome Annotation



Yandell et al. *Nature Reviews Genetics* 2012

Modern Genome Annotation



Mudge and Harrow *Nature Reviews Genetics* 2016