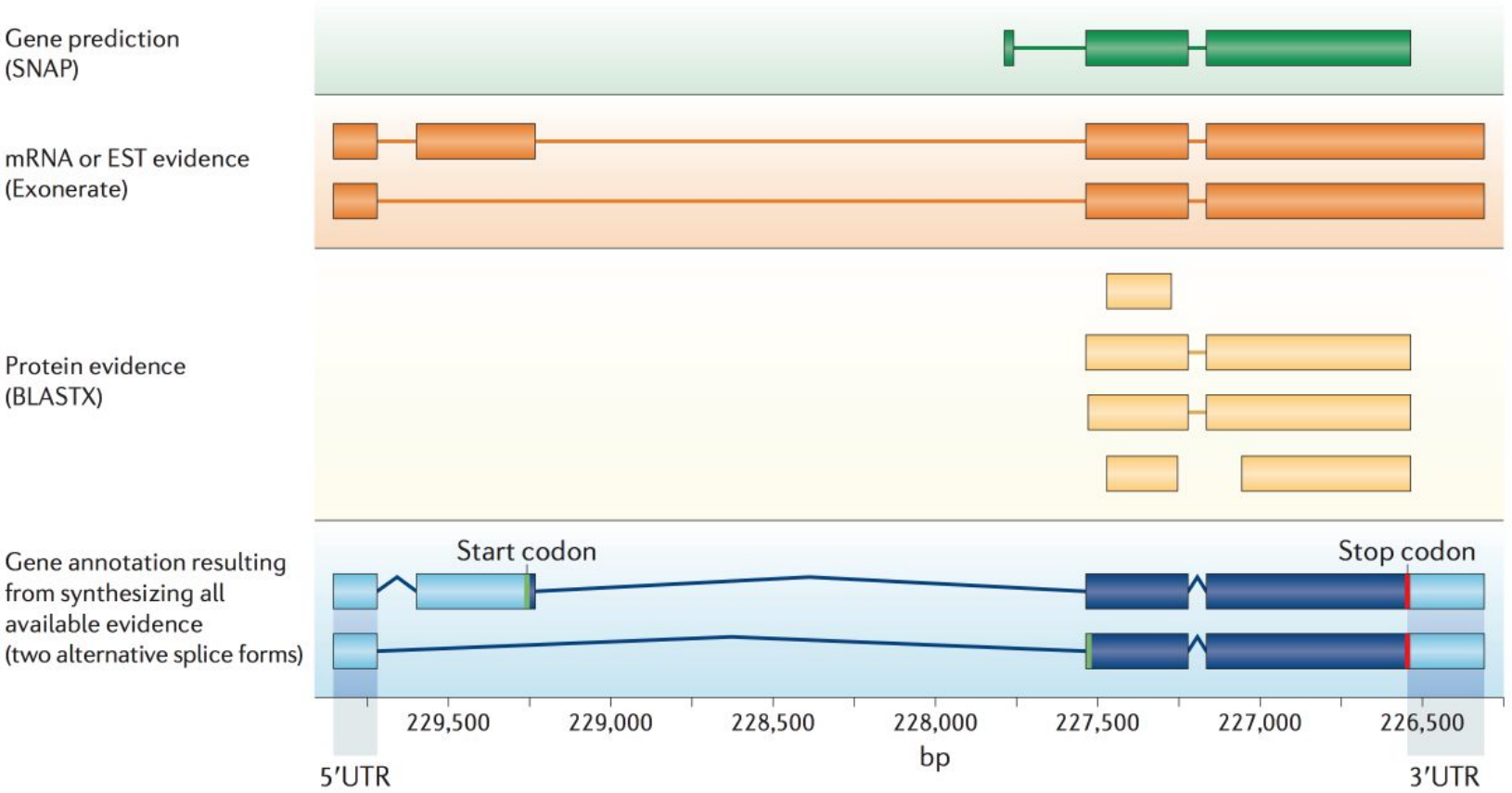
Lectures 5: Genome assembly & annotation

- Genome assembly
  - de Bruijn graphs
- Genome annotation
  - Hidden Markov Models

Topics

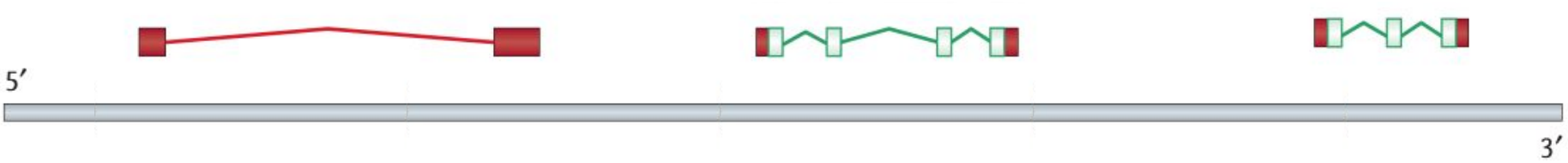
topic	V. High	High	Med	Low
Genome assembly, alignment, & annotation	6	2	6	0
Sequence alignment & pattern finding	8	1	4	1
Comparative genomics; Phylogenomics	9	0	4	1
Genetic variation & quantitative genetics	6	4	3	1
Regulatory genomics	4	3	5	2
Functional genomics	6	2	5	1
Single-cell genomics	3	4	5	2
Molecular dynamics; Protein structure prediction	5	2	6	1
Modeling cellular pathways; Digital evolution	6	4	4	0
Biological networks	9	1	4	0
Cancer genomics	5	2	3	4
Personal genomics	3	3	3	5
Genome engineering	5	3	5	1
DataSci Primers	7	3	4	0
ML Primers	8	3	3	0

Genome annotation



Yandell & Ence (2012)

Genome annotation – Gene finding

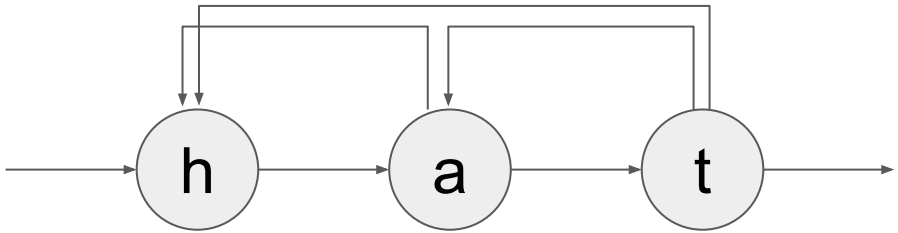


Problem: Given the entire genome, label the nucleotides as exons, introns, UTRs, or intergenic.

Requirements:

- Combine splice-site consensus, codon bias, exon/intron length preferences, and open reading frame analysis into one scoring system.
- Provide results that can be interpreted probabilistically. (How confident are we that the best scoring answer is correct?)
- Should be extensible, capable of modeling additional genomic features like translational initiation consensus, alternative splicing, and a polyadenylation signal.

Markov models



Current state depends only on previous state and transition probability.

- $\Pr(\text{‘at’}) = \Pr(\text{‘a’}) \cdot \Pr(\text{‘t’} \mid \text{‘a’})$
- $\Pr(x_1 \ldots x_n) = \Pr(x_1) \prod \Pr(x_i \mid x_{i-1})$

Hidden Markov Models (HMMs)

HMM for probabilistic sequence classification:

- HMMs are a way of relating a sequence of observations to a sequence of hidden classes or hidden states that explain the observations.
- An HMM is a full probabilistic model:
  - the model parameters and the overall sequence 'scores' are all probabilities.
  - Therefore, we can use Bayesian probability theory to manipulate these numbers in standard, powerful ways, including optimizing parameters and interpreting the significance of scores.

Brent (2007)

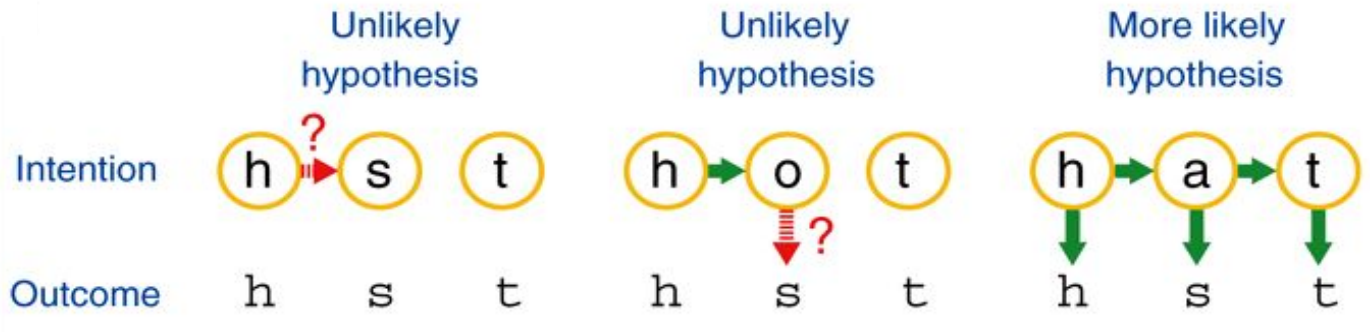
Hidden Markov Models (HMMs)

HMM for probabilistic sequence classification:

- HMMs are a way of relating a sequence of observations to a sequence of hidden classes or hidden states that explain the observations.
- The process of discovering the sequence of hidden states, given the sequence of observations, is known as **decoding** or inference. The **Viterbi algorithm** is commonly used for decoding.
- The **parameters** of an HMM are:
    - the transition probability matrix and
    - the observation likelihood (emission probability) matrix
    - Both can be trained with the Baum-Welch or forward-backward algorithm.

Brent (2007)

Hidden Markov Models (HMMs)

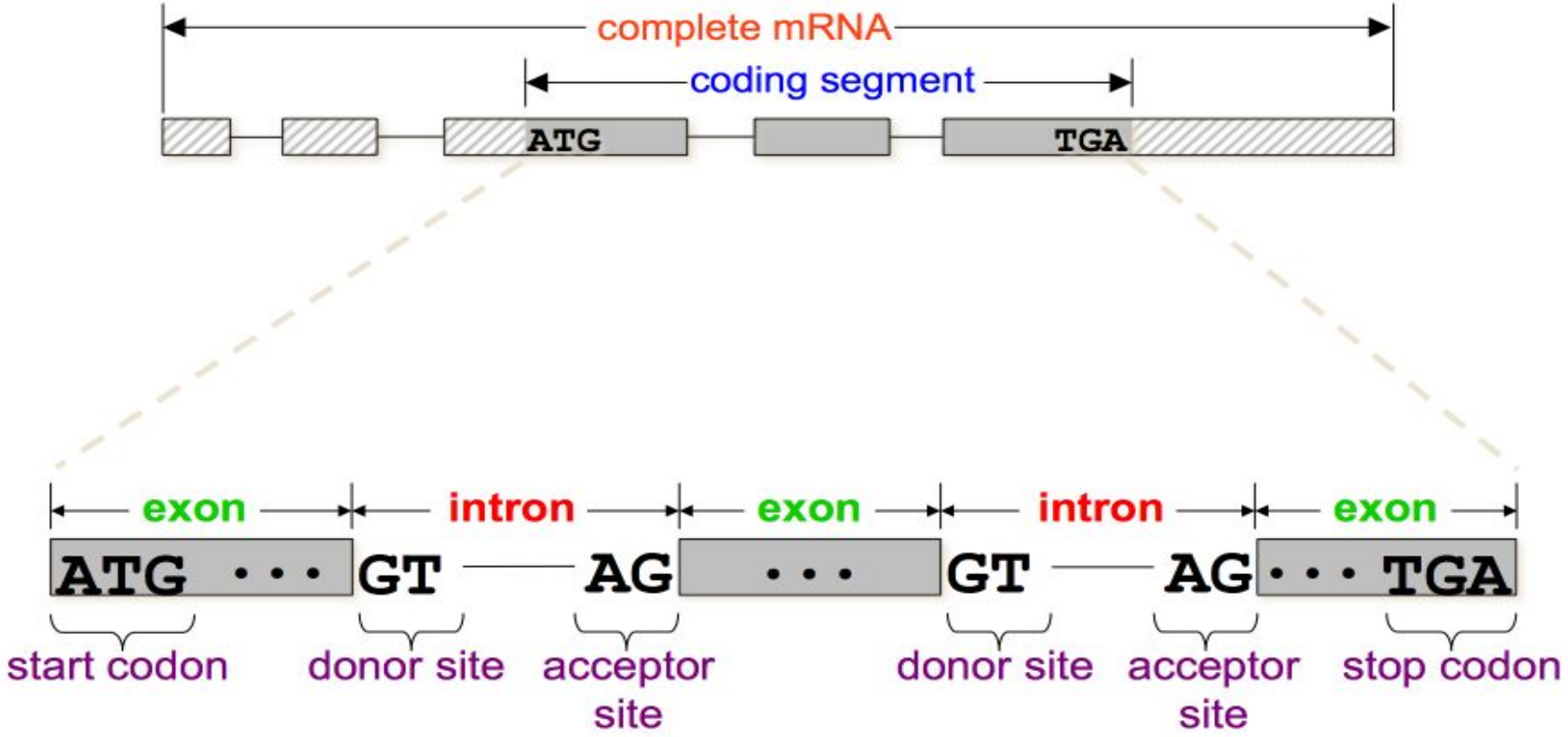


Transition probabilities: model letter sequences in correctly spelled words

Emission probabilities: model the probability of each possible typographical error.

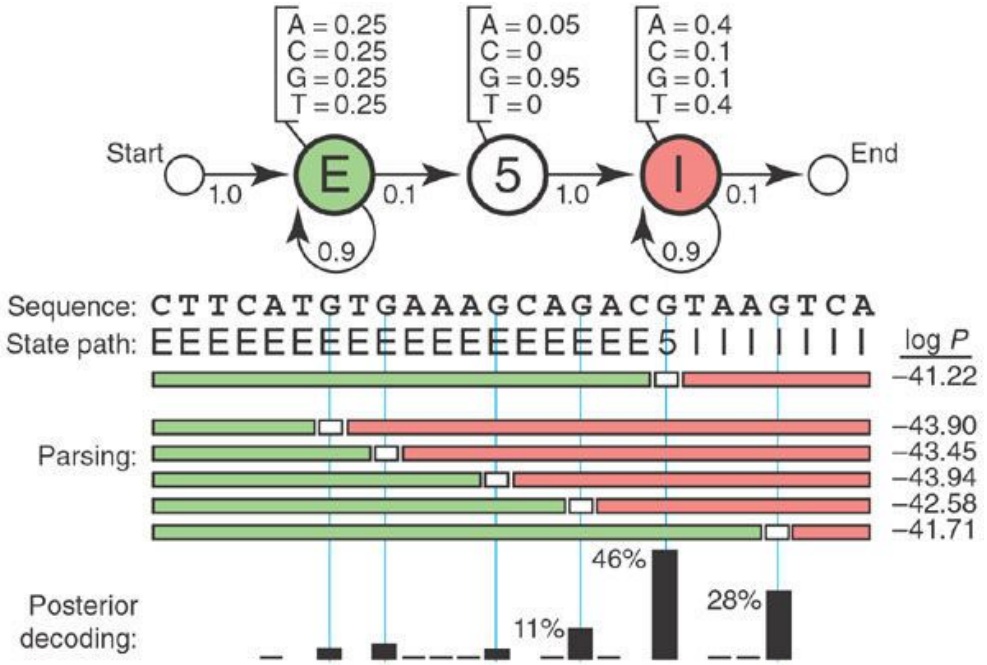
Brent (2007)

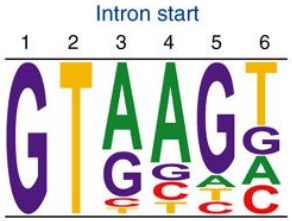
A simple HMM for modeling eukaryotic genes



A simple HMM for modeling eukaryotic genes

A toy HMM for 5′ splice site recognition





Sequence logo representing the weight matrix for the first six bases of an intron.

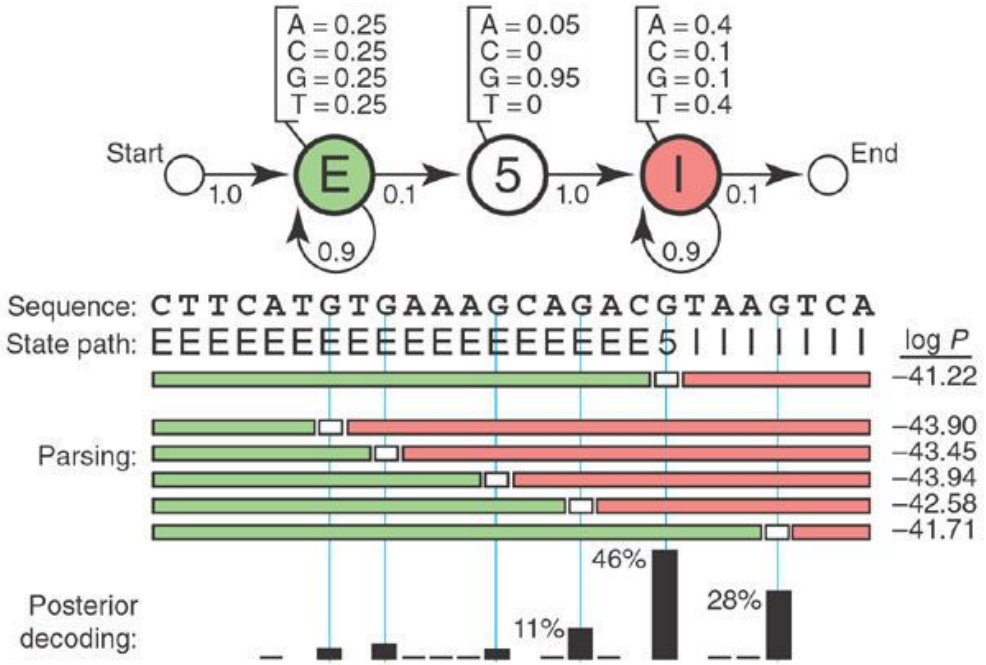
Designing the HMM:

1. Alphabet with K symbols.
2. No. of states in the model M.
3. Emission probabilities $e_i(x)$ for each state i, that sum to one over the K symbols x, $\sum_x e_i(x) = 1$.
4. Transition probabilities for each state i going to any other state j (including itself) that sum to one over the M states j, $\sum_j t_i(j) = 1$.

Eddy (2004)

A simple HMM for modeling eukaryotic genes

A toy HMM for 5′ splice site recognition



S: observed sequence

π: state path, a Markov chain that's hidden.

θ: parameters of the model

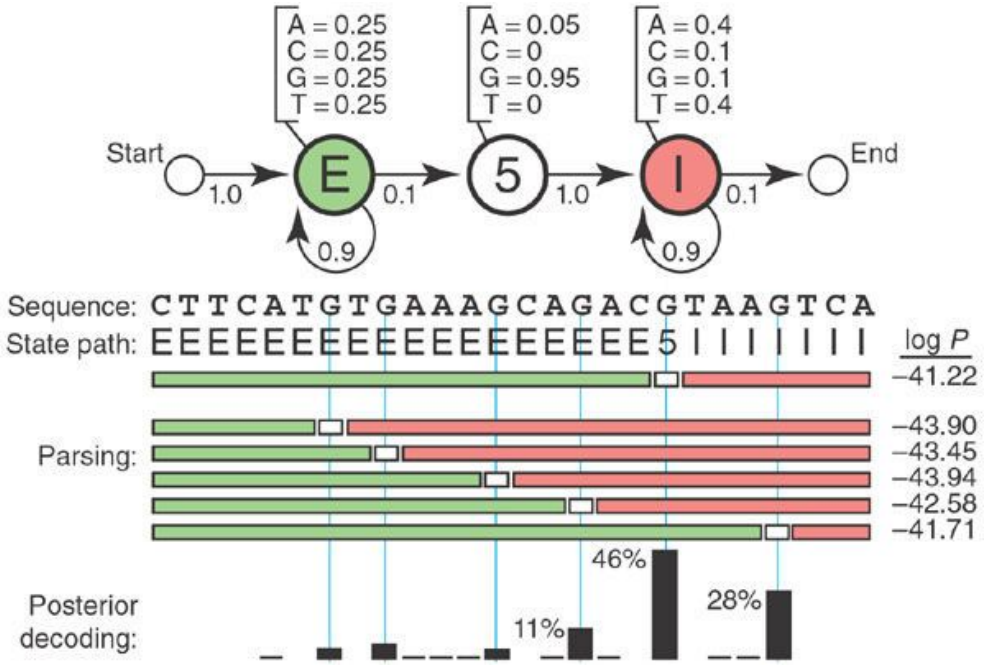
Pr(S, π | HMM, θ): product of all emission & transition probabilities. Here, 26 emissions & 27 transitions.

- How many possible paths?
- Which path to pick? The Viterbi algorithm (dynamic programming) is guaranteed to find the most probable path given seq & HMM.

Eddy (2004)

A simple HMM for modeling eukaryotic genes

A toy HMM for 5′ splice site recognition



S: observed sequence

π: state path, a Markov chain that's hidden.

θ: parameters of the model

Pr(S, π | HMM, θ): product of all emission & transition probabilities. Here, 26 emissions & 27 transitions.

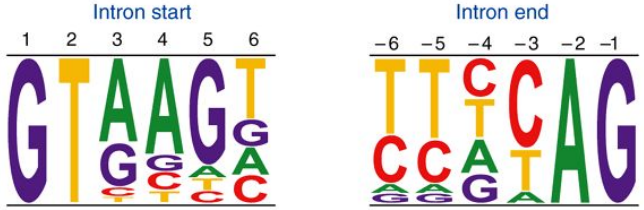
- How confident are we that the 5th G is the right choice?
  - Posterior decoding using two dynamic programming algorithms – Forward and Backward – that sum over possible paths instead of choosing the best.

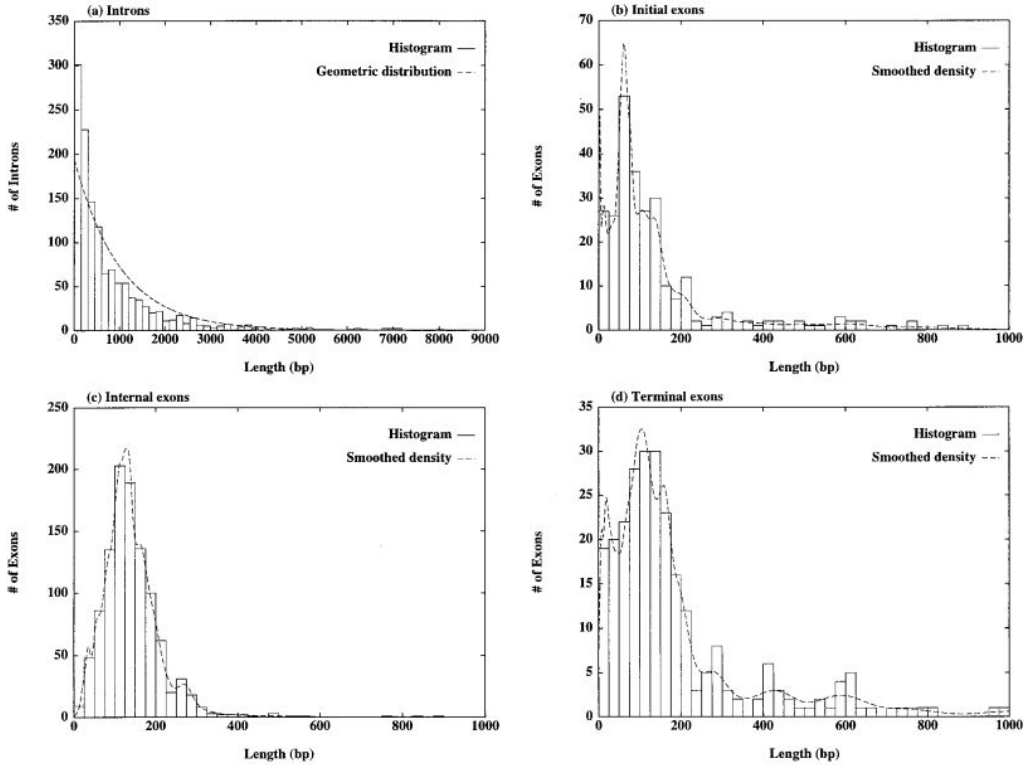
Eddy (2004)

More realistic HMMs for modeling eukaryotic genes

Adding more details:

- Six-nucleotide consensus GTRAGT at the 5′ splice site.
- Similarly for the 3′ splice site.
- Add a 3′ exon state.
- Length constraints (Generalized hidden Markov models - GHMMs).



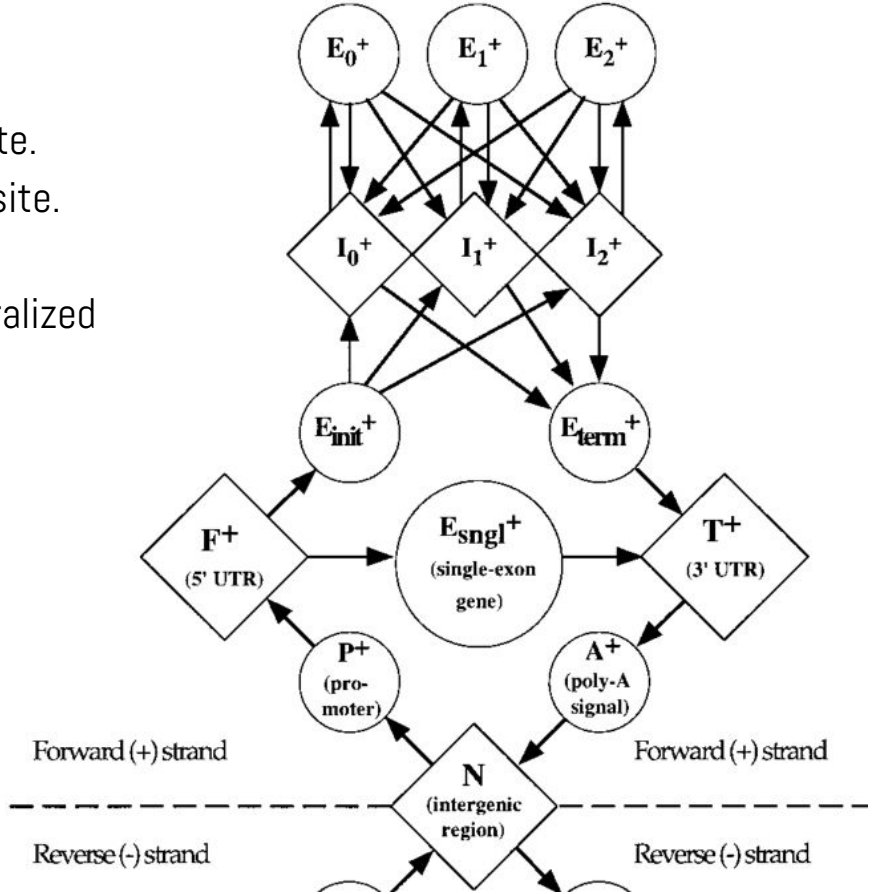


Eddy (2004)

More realistic HMMs for modeling eukaryotic genes

Adding more details:

- Six-nucleotide consensus GTRAGT at the 5′ splice site.
- Similarly for the 3′ splice site.
- Add a 3′ exon state.
- Length constraints (Generalized hidden Markov models - GHMMs).



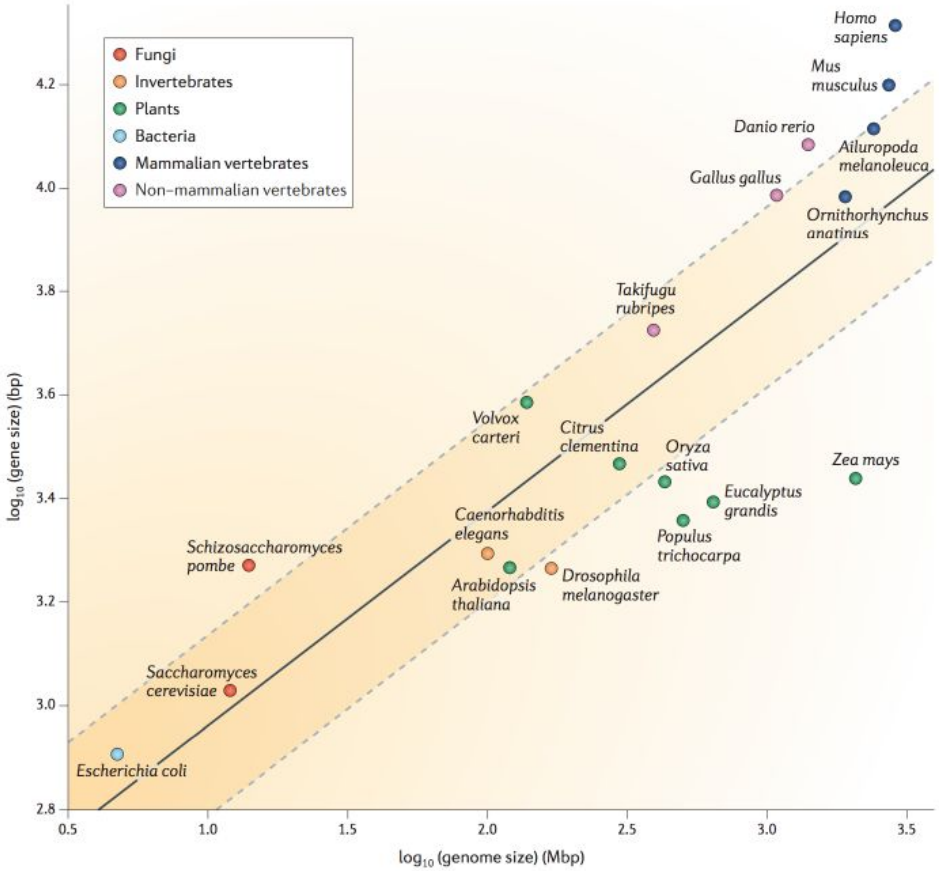
Burge & Karlin (1997)

Some modern ab-initio gene prediction tools

Software	Description
Ab initio *and evidence-drivable gene predictors*	
Augustus	Accepts expressed sequence tag (EST)-based and protein-based evidence hints. Highly accurate
mGene	Support vector machine (SVM)-based discriminative gene predictor. Directly predicts 5′ and 3′ untranslated regions (UTRs) and poly(A) sites
SNAP	Accepts EST and protein-based evidence hints. Easily trained
FGENESH	Training files are constructed by SoftBerry and supplied to users
Geneid	First published in 1992 and revised in 2000. Accepts external hints from EST and protein-based evidence
Genemark	A self-training gene finder
Twinscan	Extension of the popular Genscan algorithm that can use homology between two genomes to guide gene prediction
GAZE	Highly configurable gene predictor
GenomeScan	Extension of the popular Genscan algorithm that can use BLASTX searches to guide gene prediction
Conrad	Discriminative gene predictor that uses conditional random fields (CRFs)
Contrast	Discriminative gene predictor that uses both SVMs and CRFs
CRAIG	Discriminative gene predictor that uses CRFs
Gnomon	Hidden Markov model (HMM) tool based on Genscan that uses EST and protein alignments to guide gene prediction
GeneSeqer	A tool for identifying potential exon–intron structure in precursor mRNAs (pre-mRNAs) by splice site prediction and spliced alignment

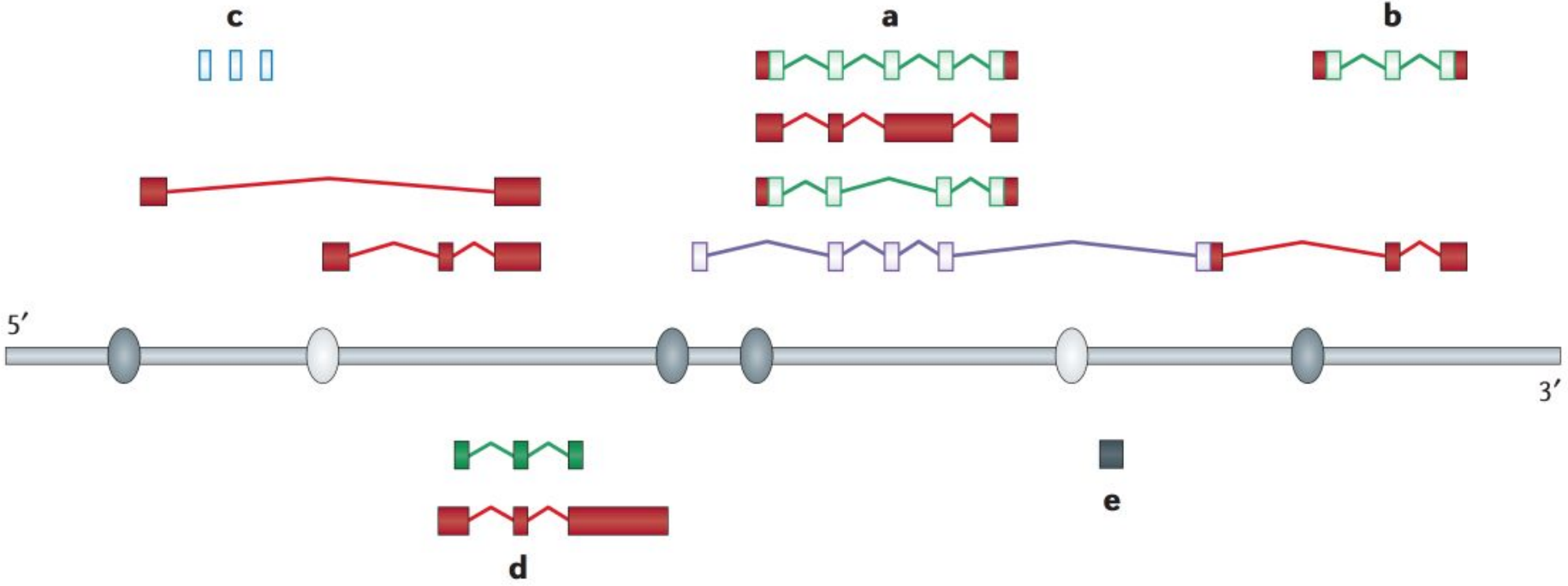
Yandell & Ence (2012)

Genomic complexity



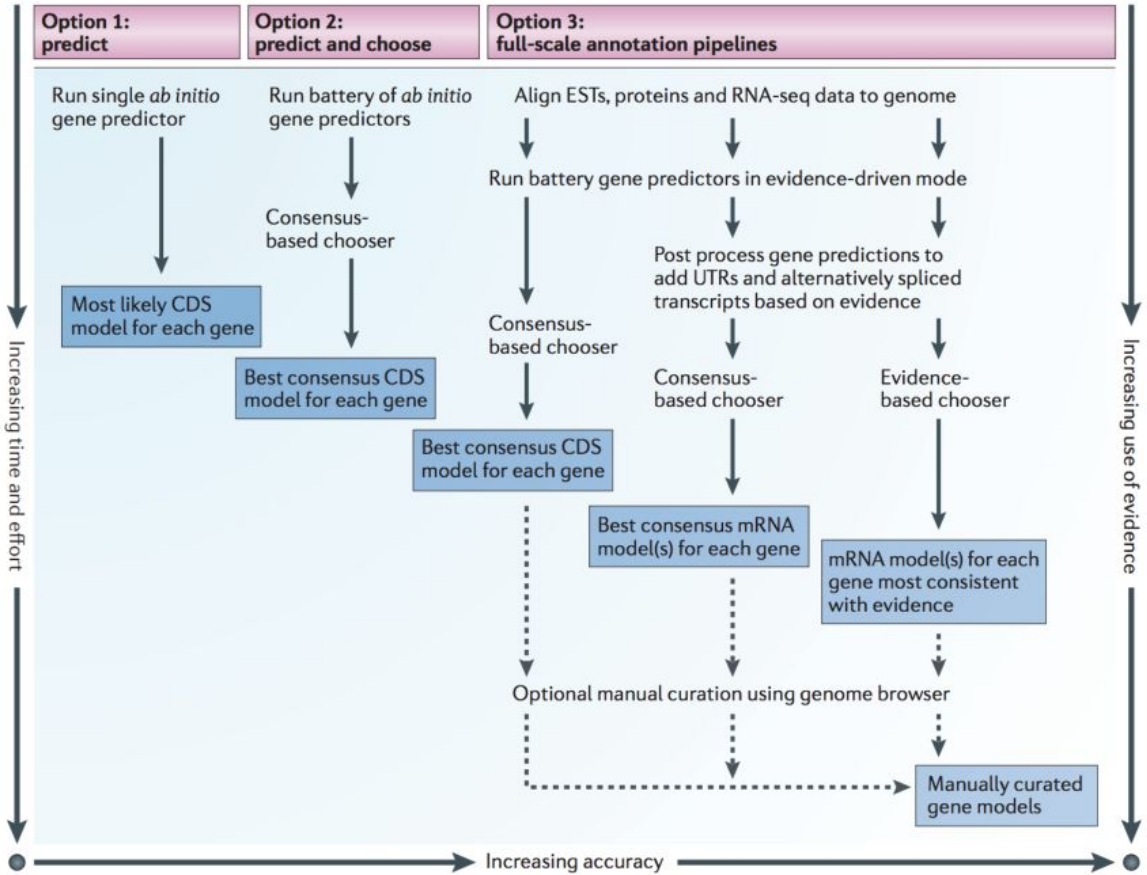
Yandell & Ence (2012)

Genomic complexity



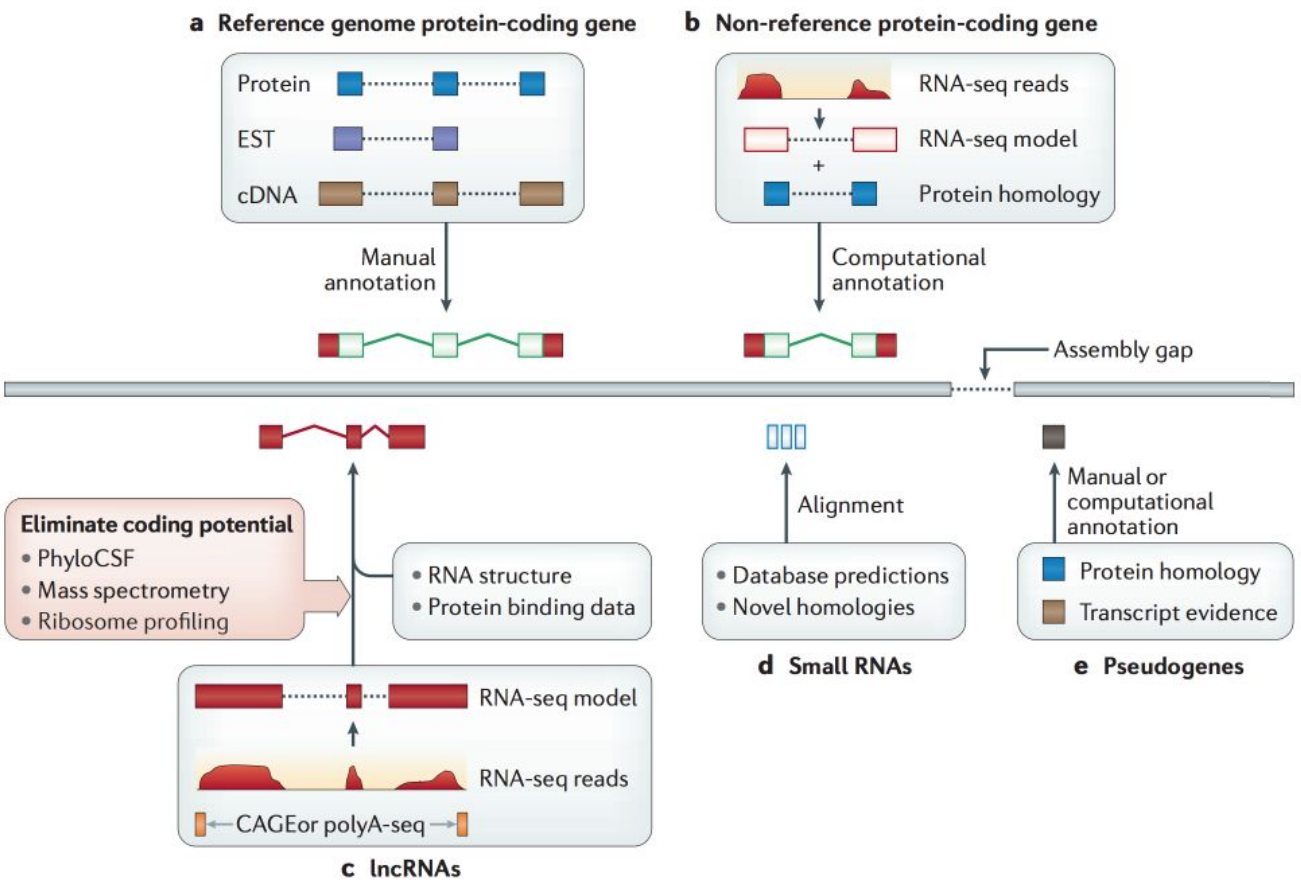
Mudge & Harrow (2016)

Approaches to genome annotation



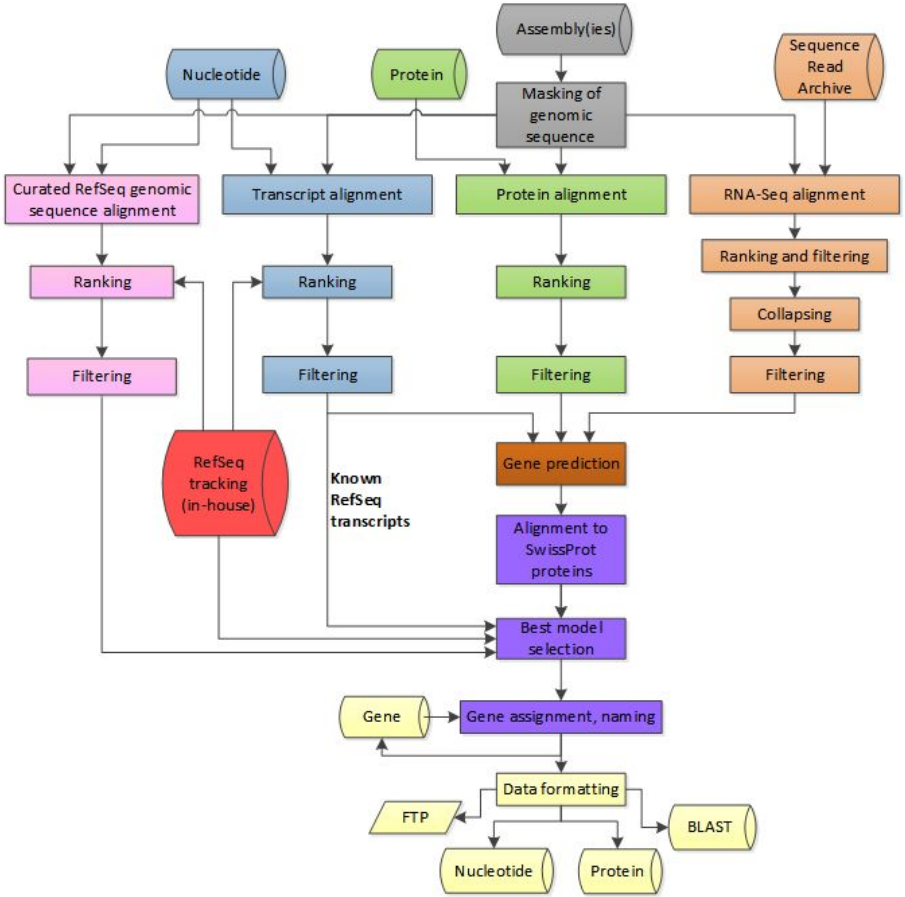
Yandell & Ence (2012)

Annotation workflows for different gene types



Mudge & Harrow (2016)

NCBI eukaryotic genome annotation pipeline



Genomic sequence preparation (grey)

Alignments of transcripts (blue)

Alignment of proteins (green)

Alignment of short reads (orange)

Alignment of curated genomic sequences (pink)

Gene prediction based on all avail. Alignments (brown)

Internal tracking database of RefSeq seqs (red)

Selection of best models & protein naming (purple)

Formatting of ann sets for public resources (yellow)

Paper discussion

- Ask questions & offer answers/thoughts.
  - Let's collectively engage.
  - Good for you & the presenters.
- Also be ready with all the questions you had from your reading.