Gene prediction

Annotation of Genomic Sequence

- Given the sequence of an organism's genome, we would like to be able to identify:
  - Genes
  - Exon boundaries & splice sites
  - Beginning and end of translation

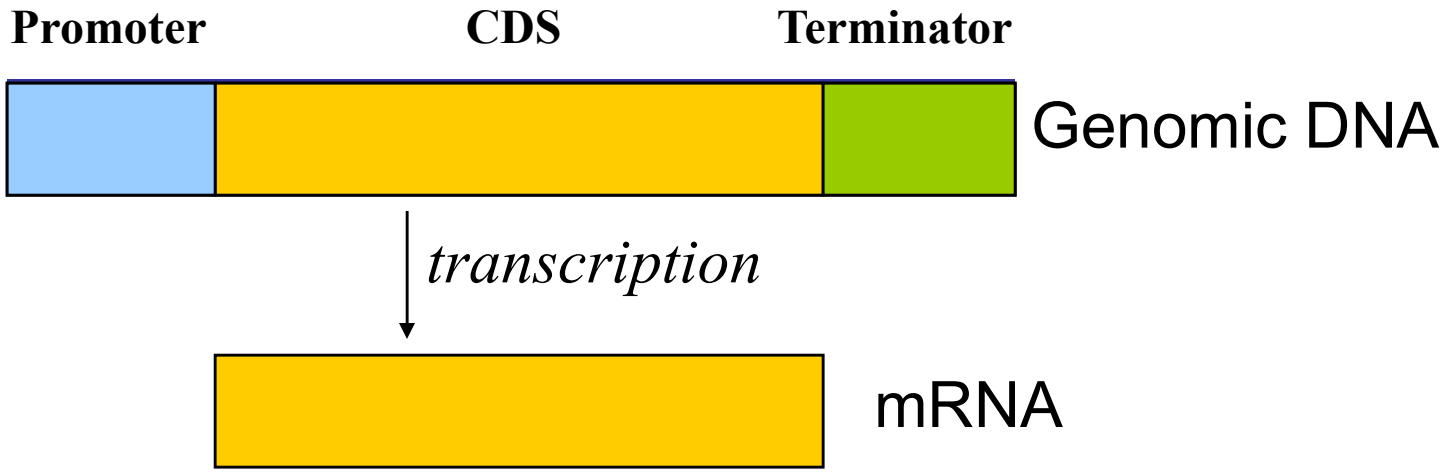
  } primary goals

  - Alternative splicings
  - Regulatory elements (e.g. promoters)

  } secondary goals

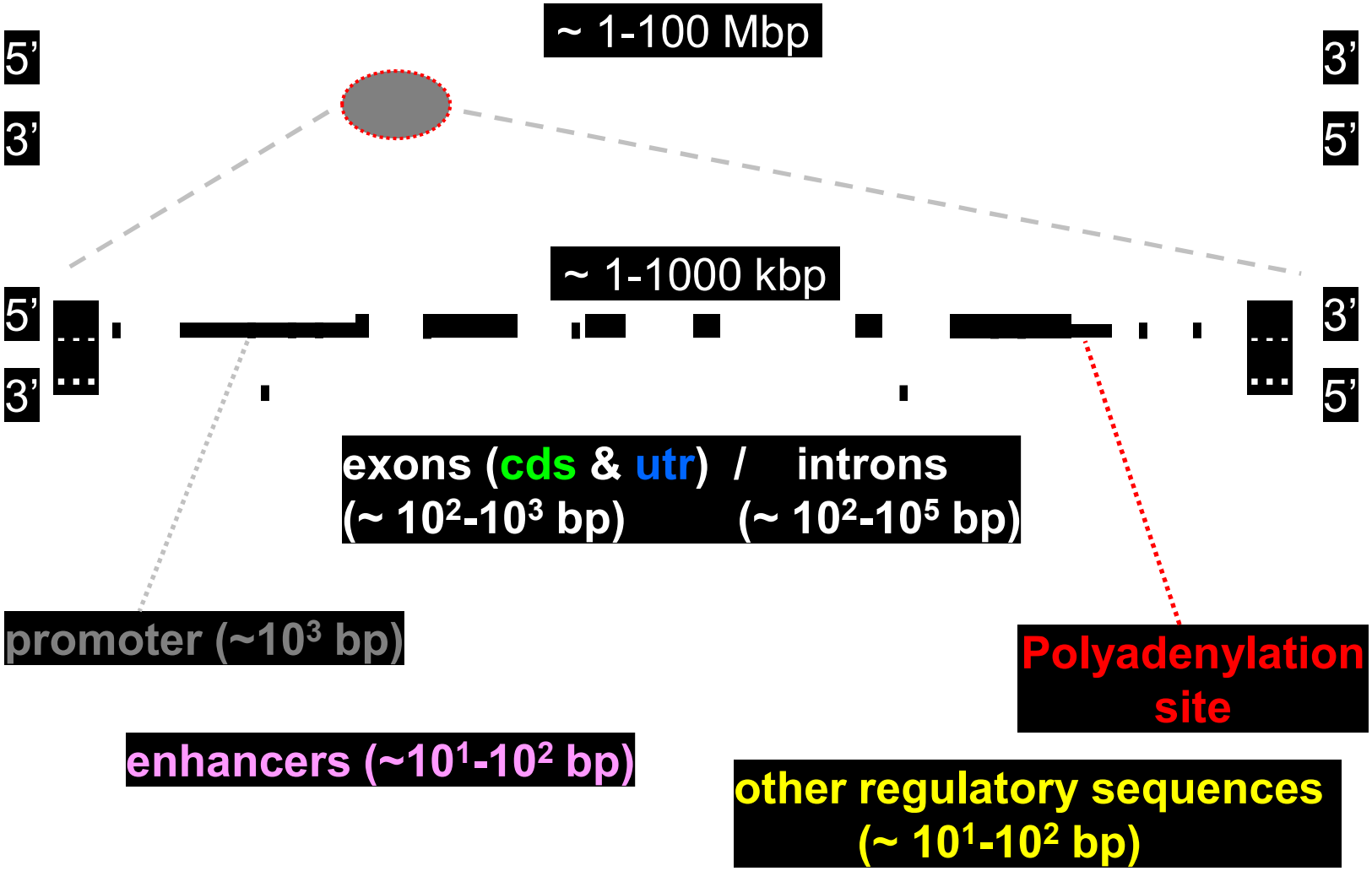
- The only certain way to do this is experimentally, but it is time consuming and expensive. Computational methods can achieve reasonable accuracy quickly, and help direct experimental approaches.

Prokaryotic Gene Structure



- Most bacterial ***promoters*** contain the Shine-Delgarno signal, at about -10 that has the consensus sequence: 5'-TATAAT-3'.
- The ***terminator***: a signal at the end of the coding sequence that terminates the transcription of RNA
- The ***coding sequence*** is composed of nucleotide triplets. Each triplet codes for an amino acid. The AAs are the building blocks of proteins.

Pieces of a (Eukaryotic) Gene (on the genome)



What is it about genes that we can measure (and model)?

- Most of our knowledge is biased towards protein-coding characteristics
  - ORF (Open Reading Frame): a sequence defined by in-frame AUG and stop codon, which in turn defines a putative amino acid sequence.
  - Codon Usage: most frequently measured by CAI (Codon Adaptation Index)

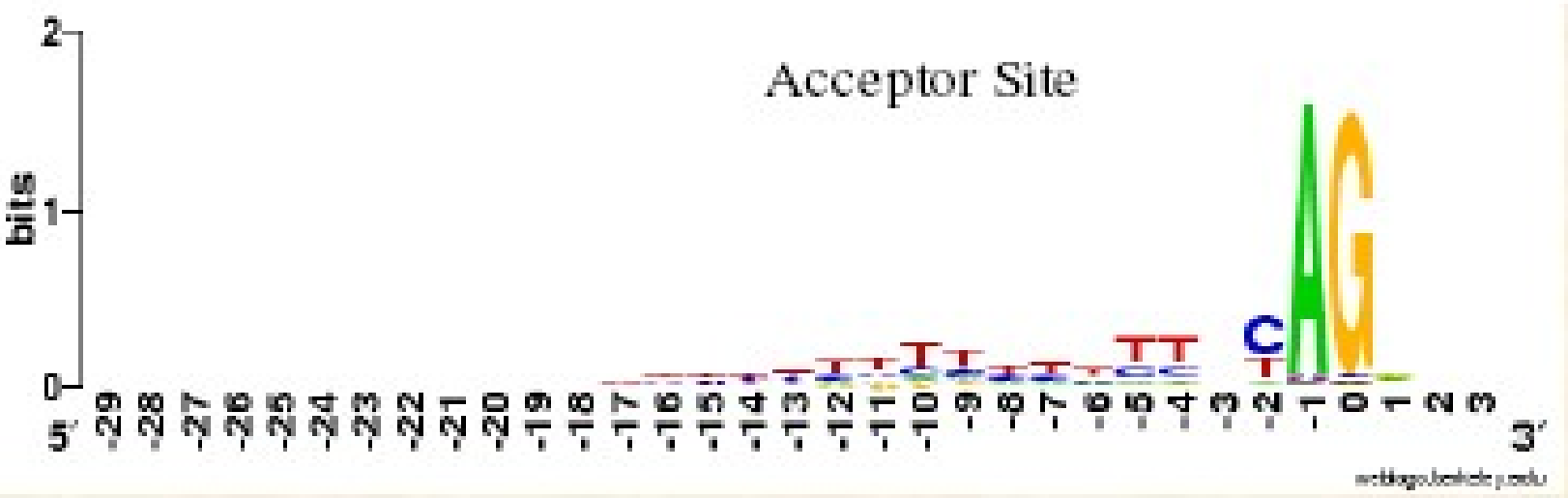
- Other phenomena
  - Nucleotide frequencies and correlations:
    - value and structure
  - Functional sites:
    - splice sites, promoters, UTRs, polyadenylation sites

Codon Adaptation Index (CAI)

$$CAI = \prod_{i=codons} \left[\frac{f_{codon_i}}{f_{(codon_i)_{\max}}} \right]$$

- Parameters are empirically determined by examining a "large" set of example genes
- This is not perfect
  - Genes sometimes have unusual codons for a reason
  - The predictive power is dependent on length of sequence

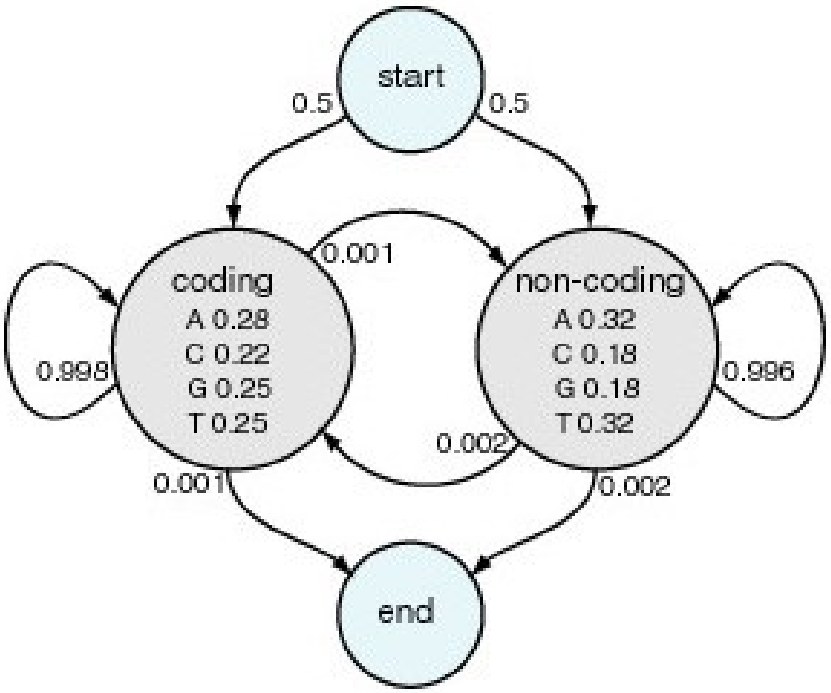
Splice signals (mice): GT , AG



General Things to Remember about (Protein-coding) Gene Prediction Software

- It is, in general, organism-specific
- It works best on genes that are *reasonably* similar to something seen previously
- It finds protein coding regions far better than non-coding regions
- In the absence of external (direct) information, alternative forms will not be identified
- It is imperfect! (It's biology, after all…)

Simple HMM : Prokaryotes



$$\Phi = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5 & 0.998 & 0.002 & 0 \\ 0.5 & 0.001 & 0.996 & 0 \\ 0 & 0.001 & 0.002 & 0 \end{bmatrix}$$

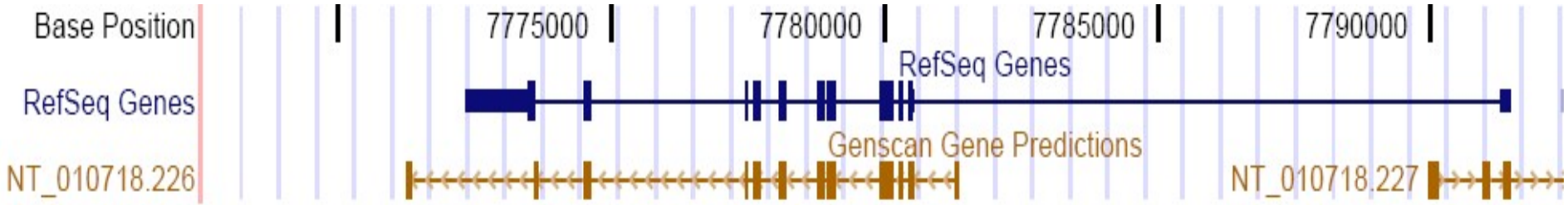
$$H = \begin{bmatrix} 0.28 & 0.32 \\ 0.22 & 0.18 \\ 0.25 & 0.18 \\ 0.25 & 0.32 \end{bmatrix}$$

$x_m(i)$ = probability of being in state m at position i;

$H(m,y_i)$ = probability of emitting character y_i in state m;

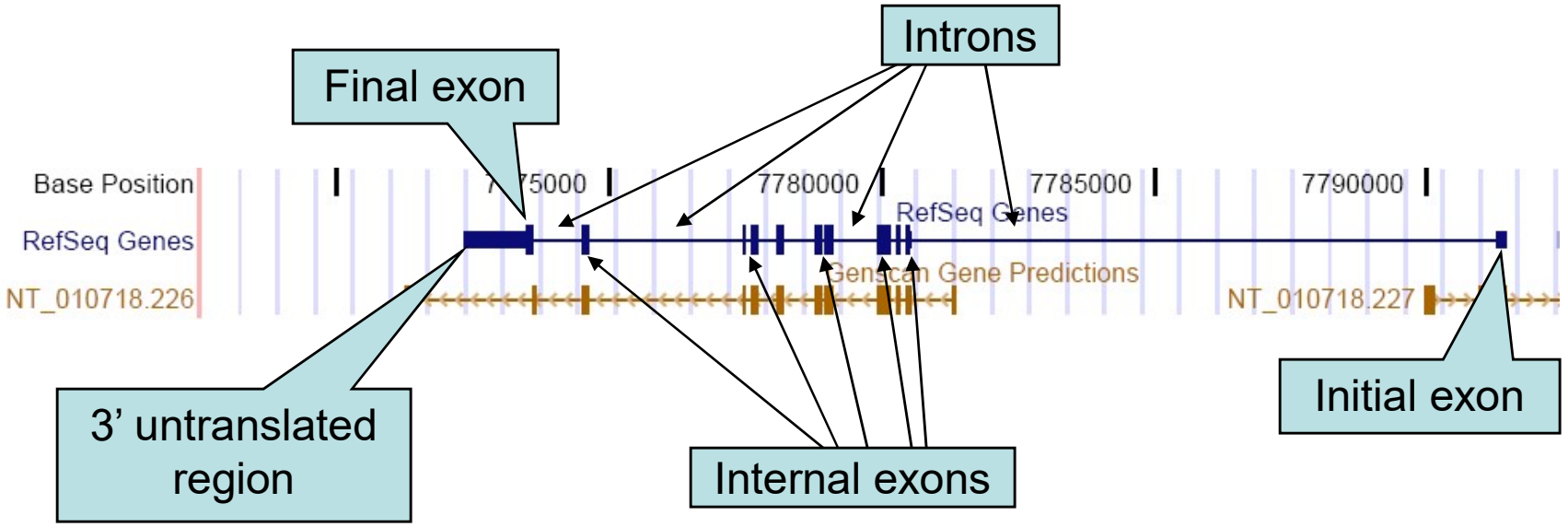
Φ_{mk} = probability of transition from state k to m.

A eukaryotic gene



- This is the human p53 tumor suppressor gene on chromosome 17.
- Genscan is one of the most popular gene prediction algorithms.

A eukaryotic gene



This particular gene lies on the reverse strand.

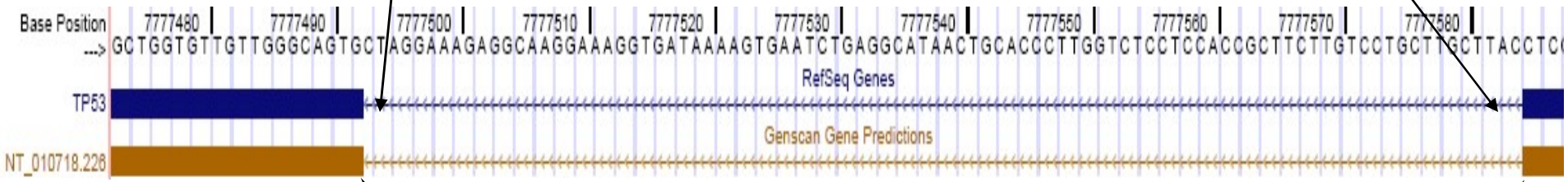
An Intron

revcomp(CT)=AG

GT: signals start of intron

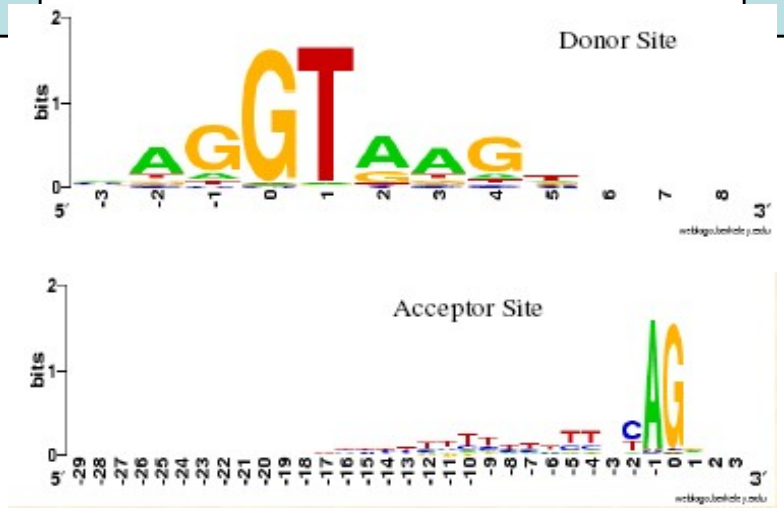
AG: signals end of intron

revcomp(AC)=GT



3’ splice site

5’ splice site



Signals vs Contents

- In gene finding, a small pattern within the genomic DNA is referred to as a signal, whereas a region of genomic DNA is a content.
- Examples of signals: splice sites, starts and ends of transcription or translation, branch points, transcription factor binding sites
- Examples of contents: exons, introns, UTRs, promoter regions

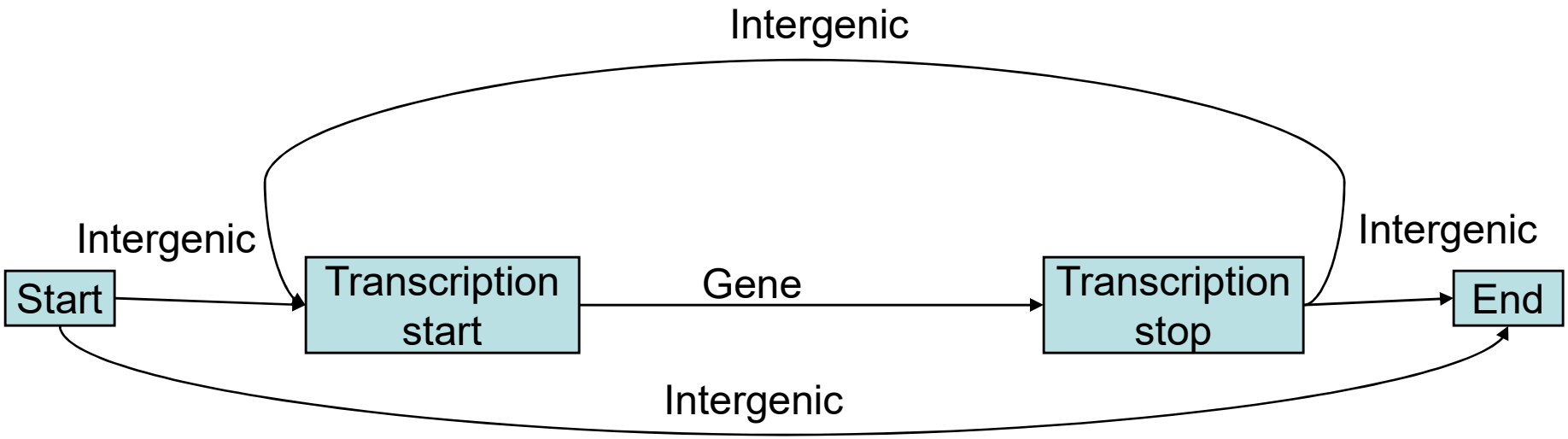
Prior knowledge

- We want to build a probabilistic model of a gene that incorporates our prior knowledge.
- E.g., the translated region must have a length that is a multiple of 3.

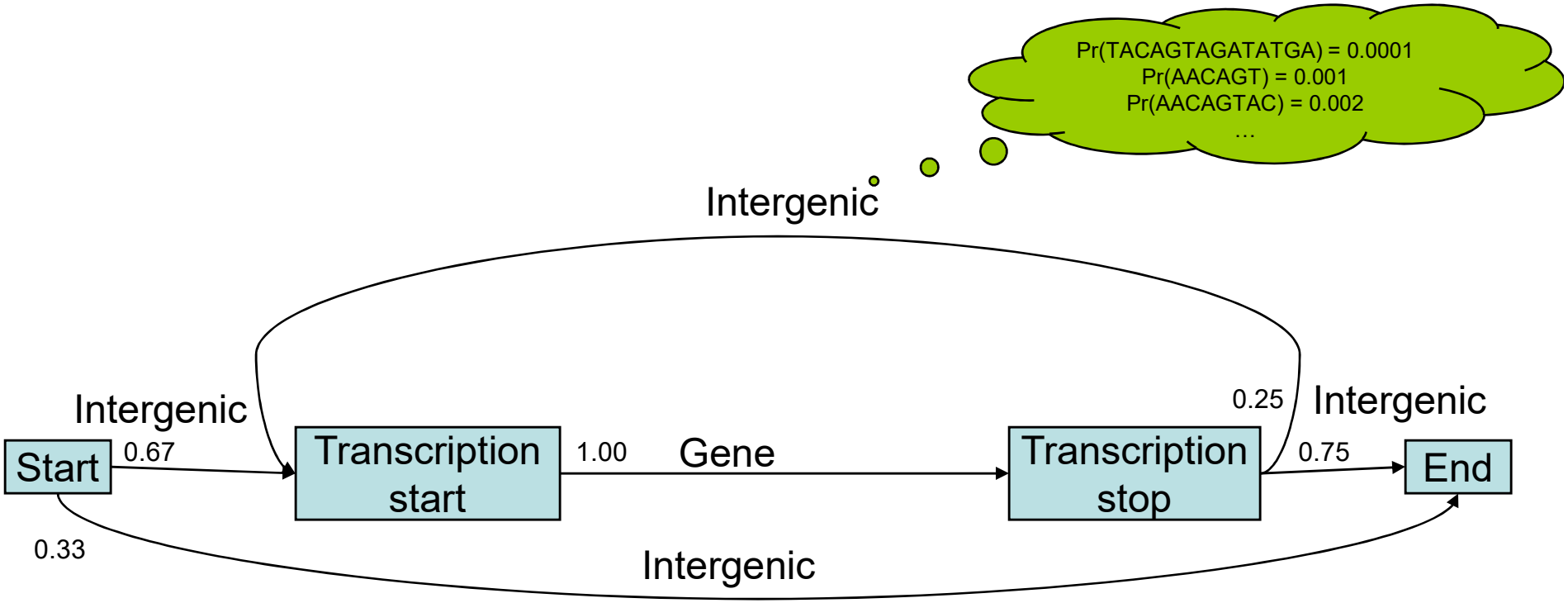
Prior knowledge

- The translated region must have a length that is a multiple of 3.
- Some codons are more common than others.
- Exons are usually shorter than introns.
- The translated region begins with a start signal and ends with a stop codon.
- 5′ splice sites (exon to intron) are usually *GT*;
- 3′ splice sites (intron to exon) are usually *AG*.
- The distribution of nucleotides and dinucleotides is usually different in introns and exons.

A simple gene model



A probabilistic gene model



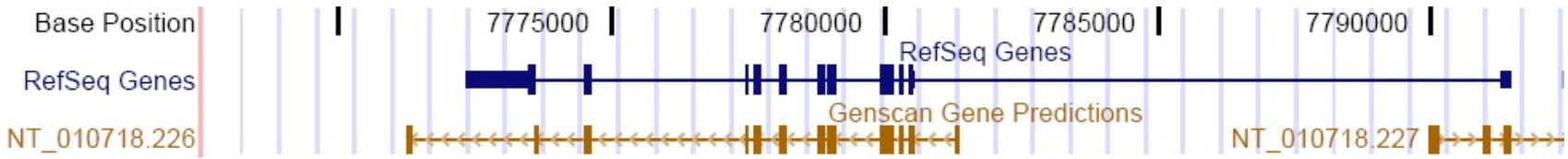
Every box stores transition probabilities for outgoing arrows.
Every arrow stores emission probabilities for emitted nucleotides.

Parse

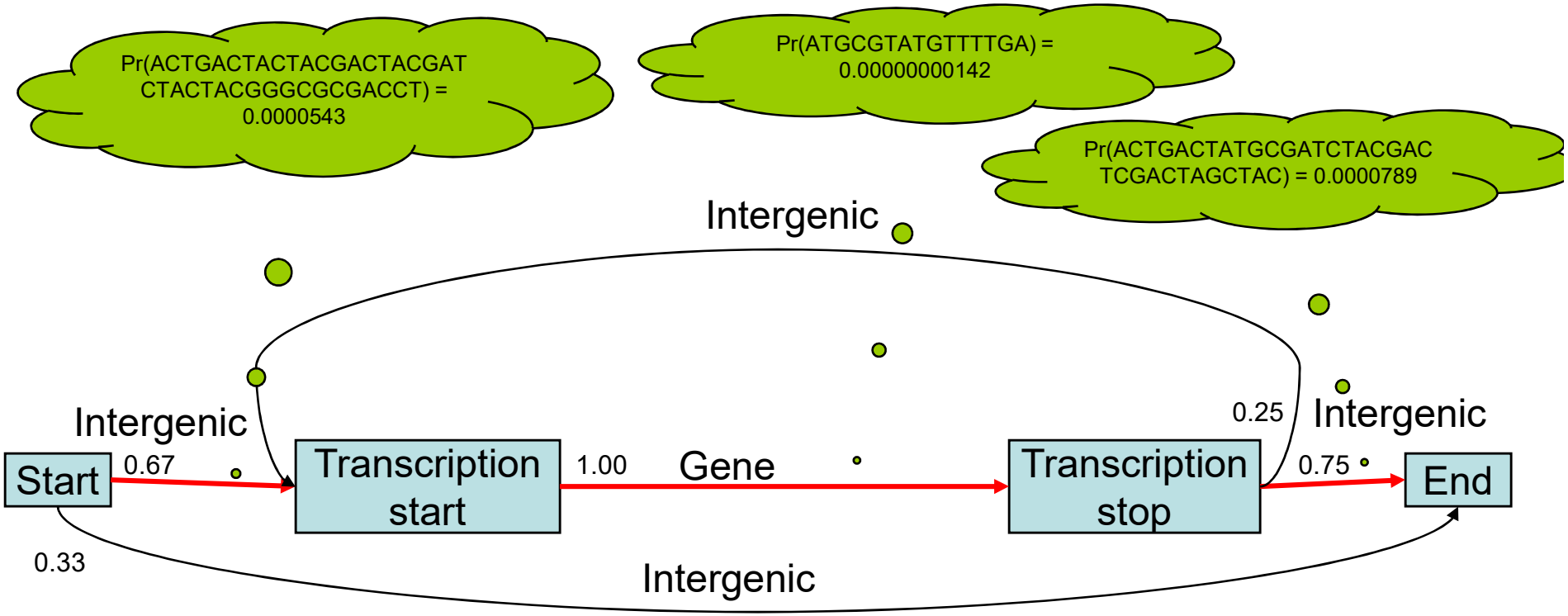
```
S = ACTGACTACTACGACTACGATCTACTACGGGCGCGACCTATGCG
P = IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGG

    TATGTTTTGAACTGACTATGCGATCTACGACTCGACTAGCTAC
    GGGGGGGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

- For a given sequence, a parse is an assignment of gene structure to that sequence.
- In a parse, every base is labeled, corresponding to the content it (is predicted to) belongs to.
- In our simple model, the parse contains only "I" (intergenic) and "G" (gene).
- A more complete model would contain, e.g., "-" for intergenic, "E" for exon and "I" for intron.



The probability of a parse



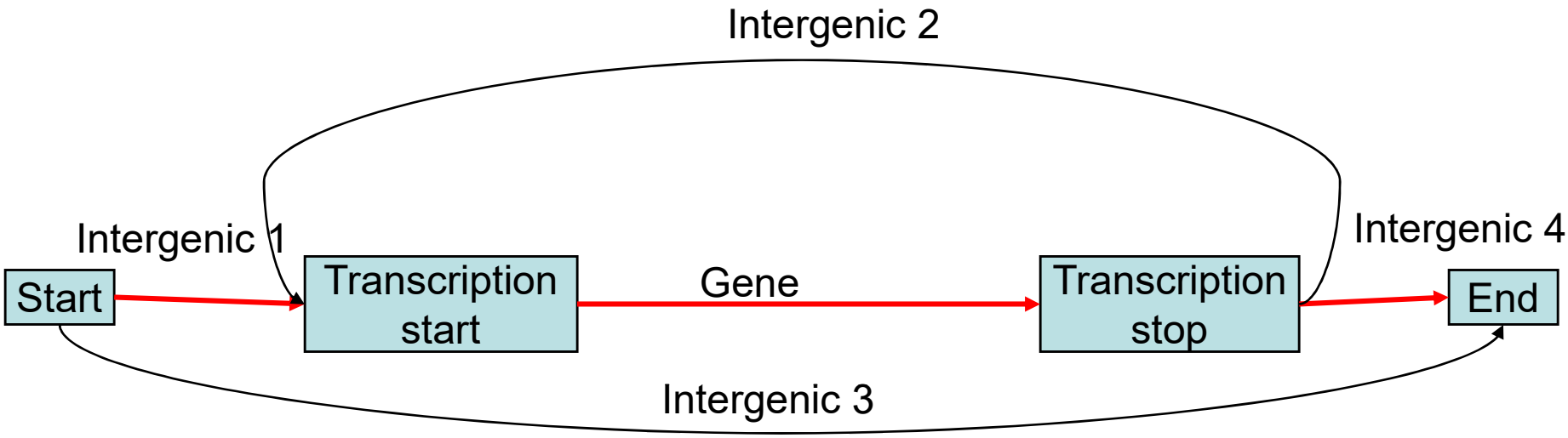

```
S = ACTGACTACTACGACTACGATCTACTACGGGCGCGACCTATGCGTATGTTTTGAACTGACTATGCGATCTACGACTCGACTAGCTAC
P = IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIGGGGGGGGGGGGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Pr(parse P| sequence S, model M)

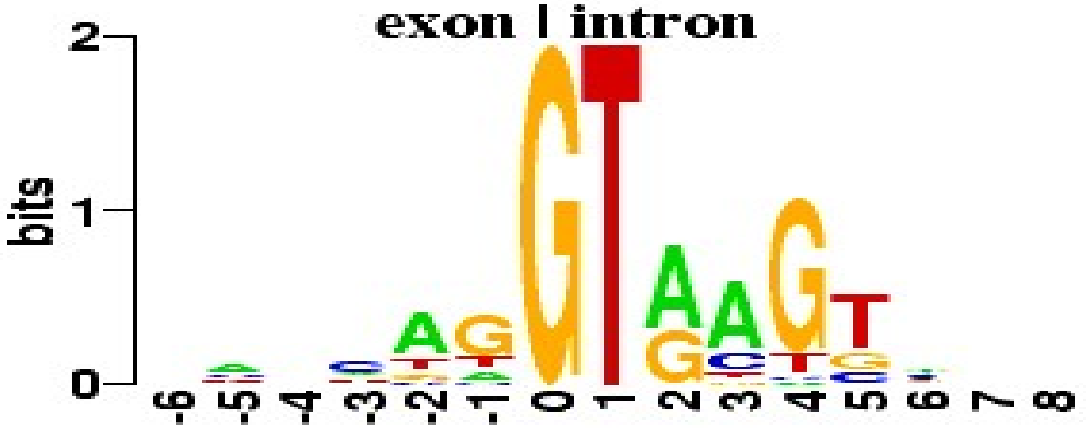
= $0.67 \times 0.0000543 \times 1.00 \times 0.00000000142 \times 0.75 \times 0.0000789$

= 3.057×10^{-18}

Improved model topology



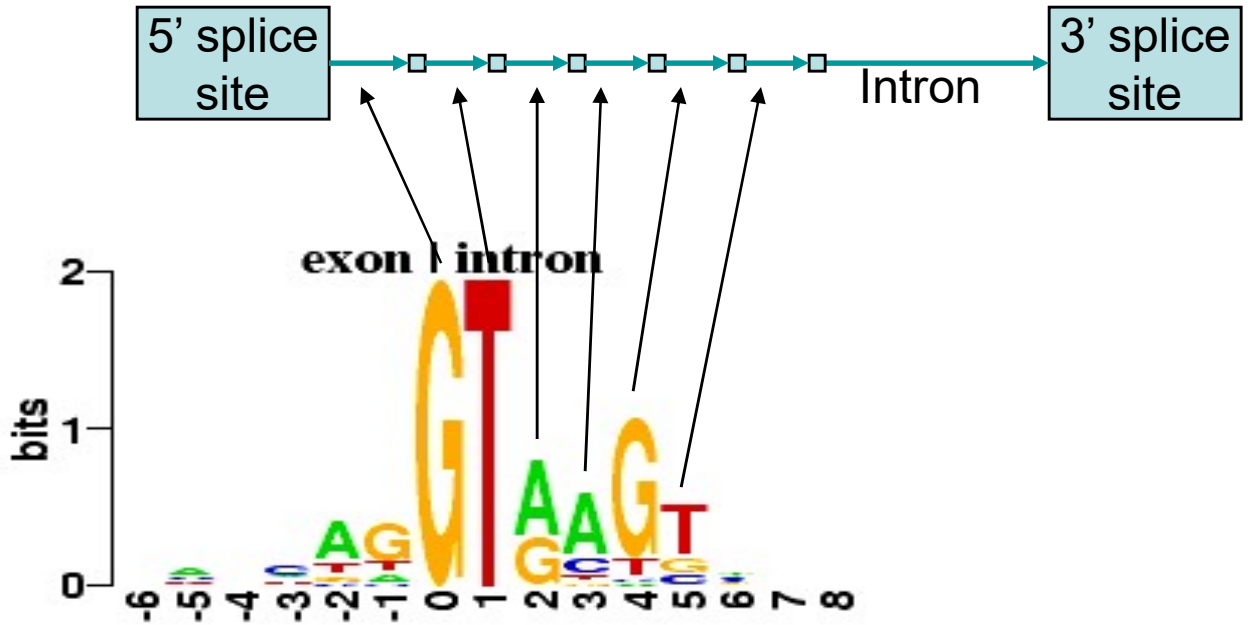
Real splice sites



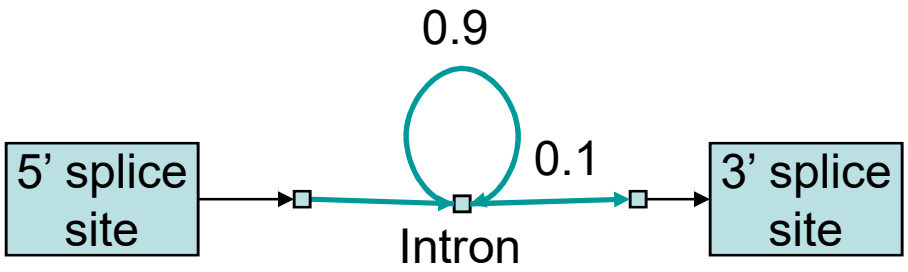
- Real splice sites show some conservation at positions beyond the first two.
- We can add additional arrows to model these states.

weblogo.berkeley.edu

Modeling the 5′ splice site

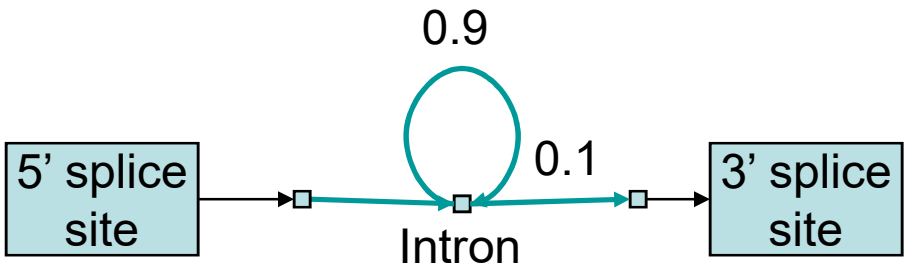


A small problem



- Say that each blue arrow emits one letter.
- What is the probability that the intron will be exactly 2 letters long?
- 3 letters long?
- 4 letters long?

A small problem



- Say that each blue arrow emits one letter.
- What is the probability that the intron will be exactly 2 letters long? 10%
- 3 letters long? 9%
- 4 letters long? 8.1%