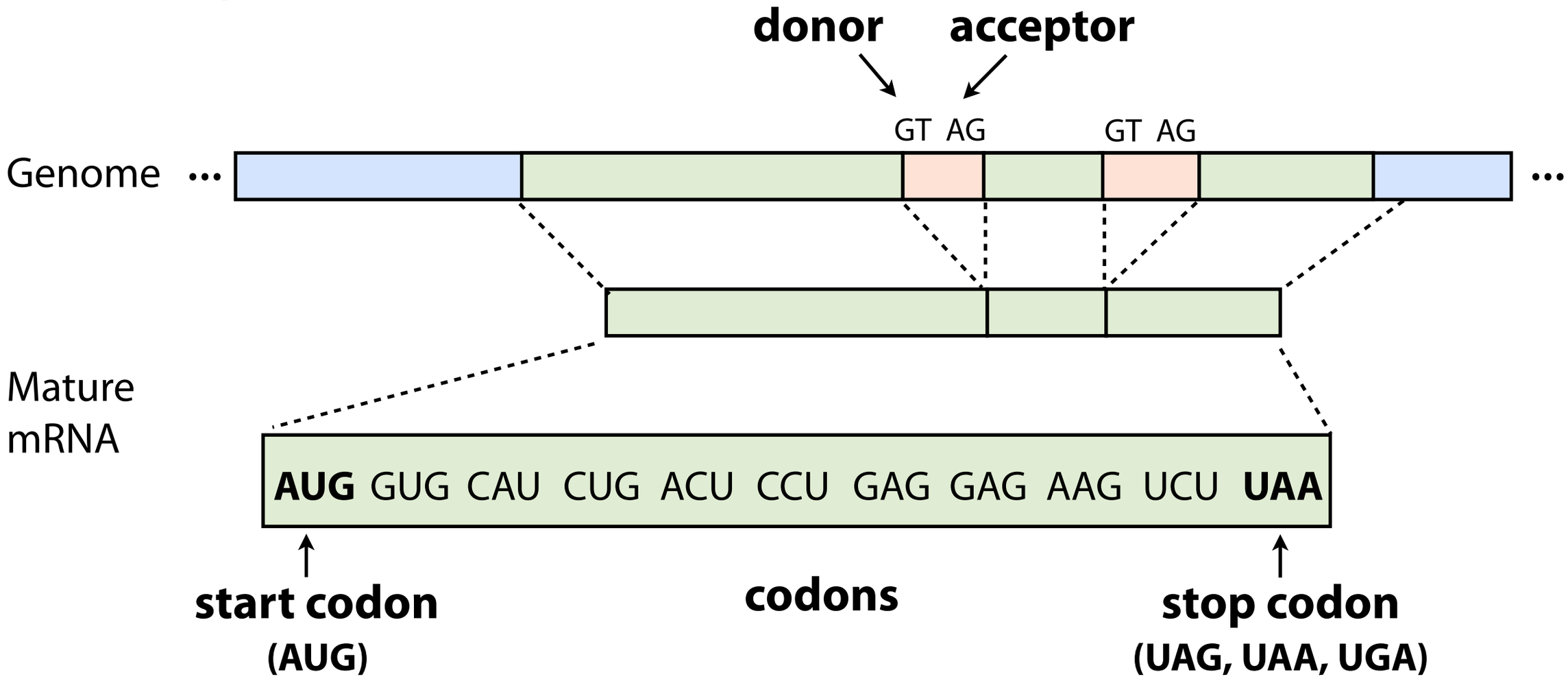
# Example: HMMs for gene finding, part 2

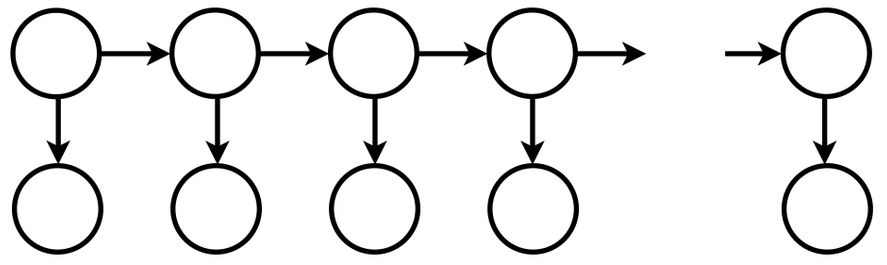
Ben Langmead

# Transcription

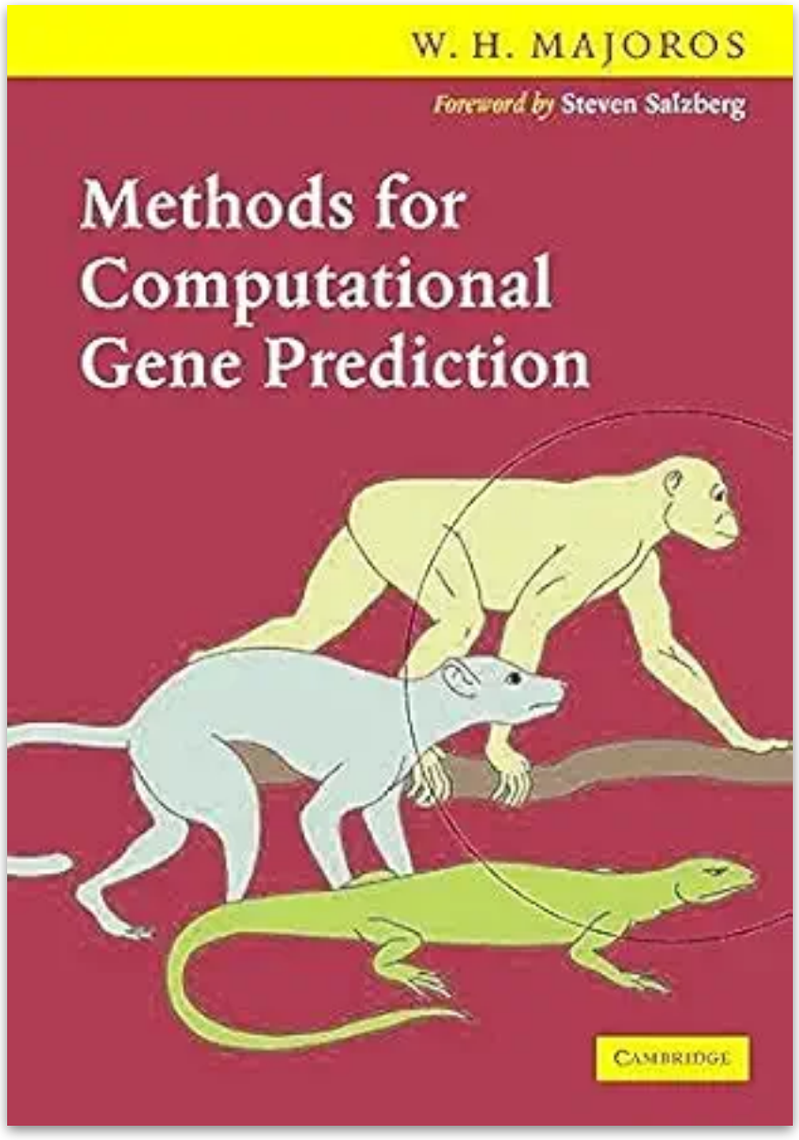



Can HMM help us find genes and their constituent parts?

What will be the states? Emissions?



# Today's discussion follows…




Majoros, W.H., 2007. Methods for computational gene prediction. Cambridge University Press.

# We can do this the easy way or the hard way

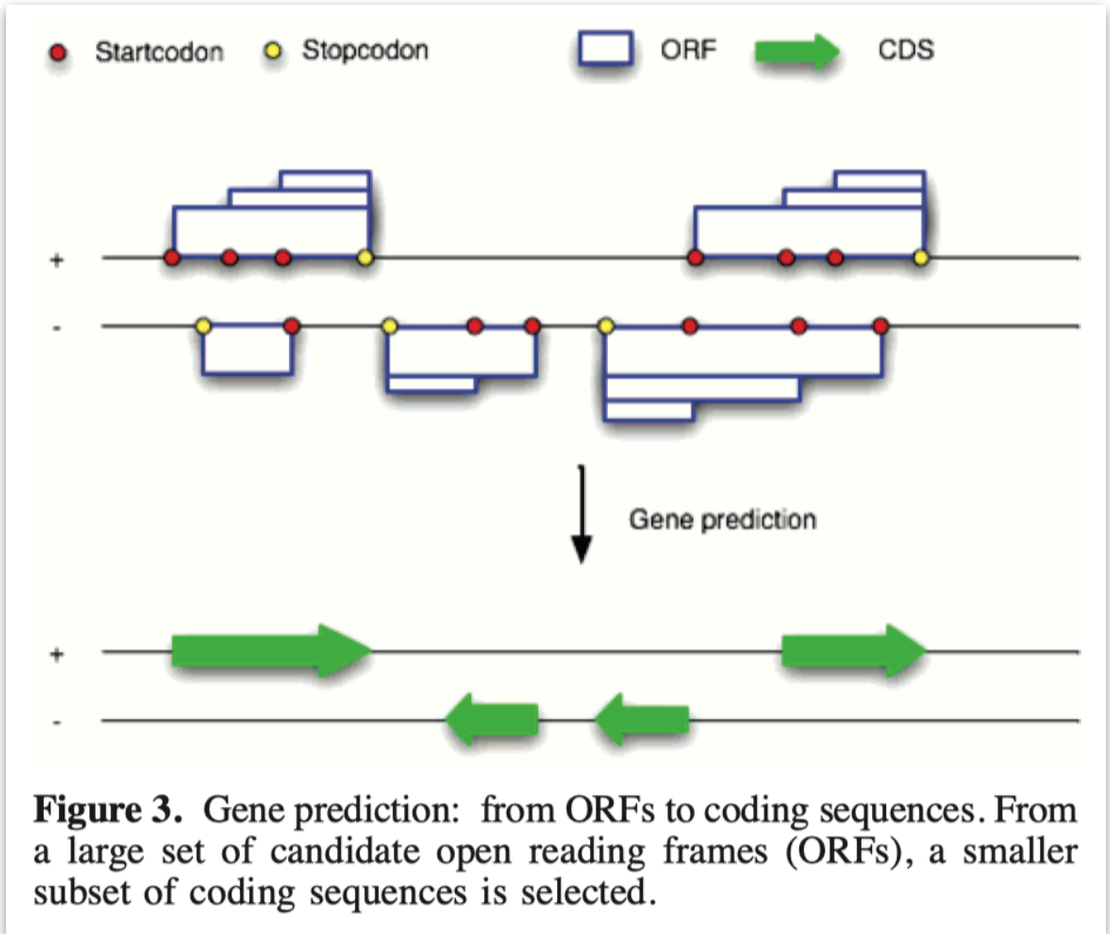



**Figure 3.** Gene prediction: from ORFs to coding sequences. From a large set of candidate open reading frames (ORFs), a smaller subset of coding sequences is selected.

Gene finding is much easier in some parts of the tree of life, e.g. for prokaryotes

In prokaryotes, genes are *unspliced* and *packed into the genome* very tightly

Overbeek, R., Bartels, D., Vonstein, V. and Meyer, F., 2007. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chemical reviews*, *107*(8), pp.3431-3447.

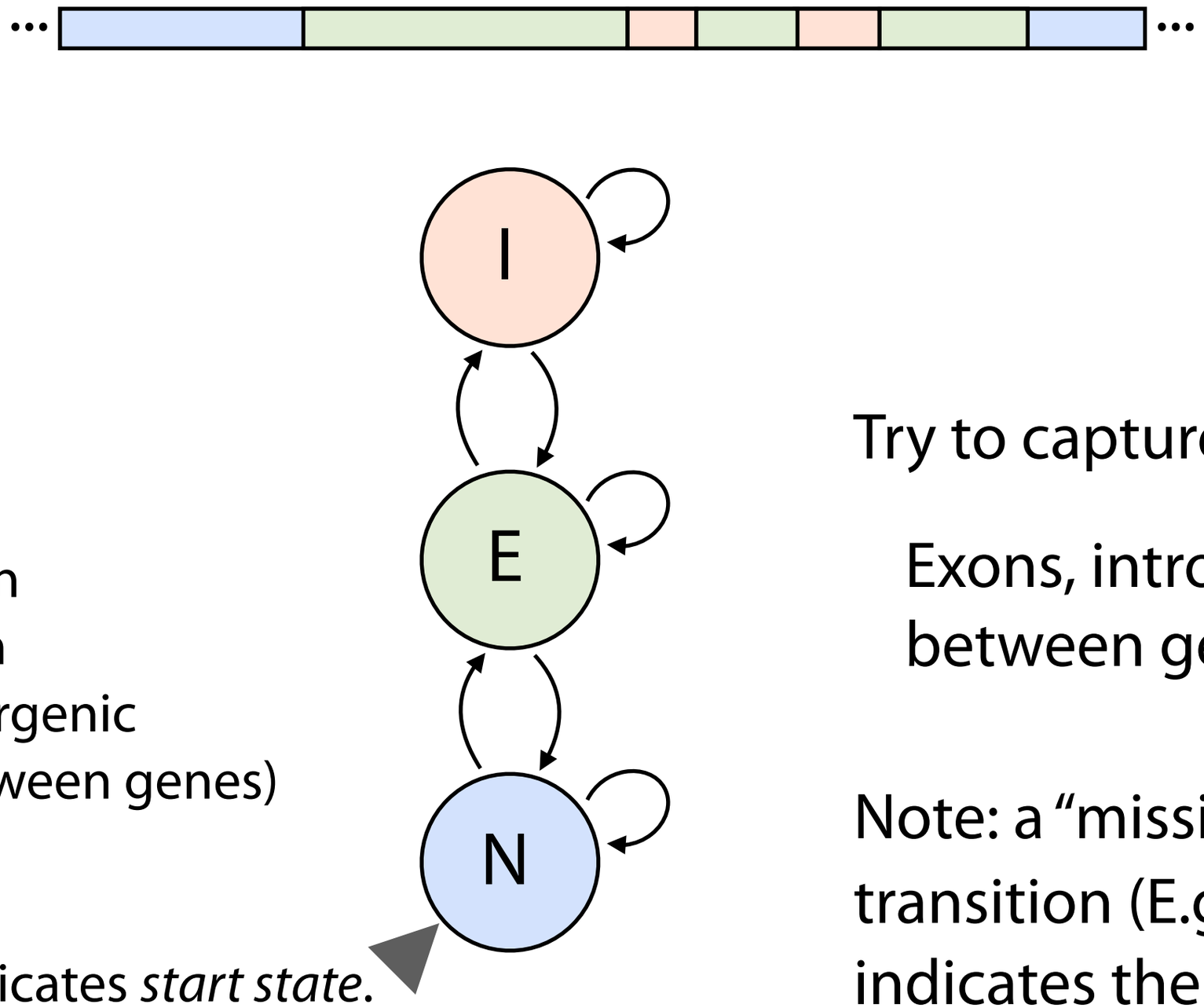
Gene finding is much harder in eukaryotes: animals, humans, etc

# Gene finding

Try to capture:

Exons, introns, spaces between genes

# Gene finding




I = intron
E = exon
N = intergenic
(between genes)

Indicates *start state*.
State's initial probability is 1

Try to capture:

Exons, introns, spaces between genes

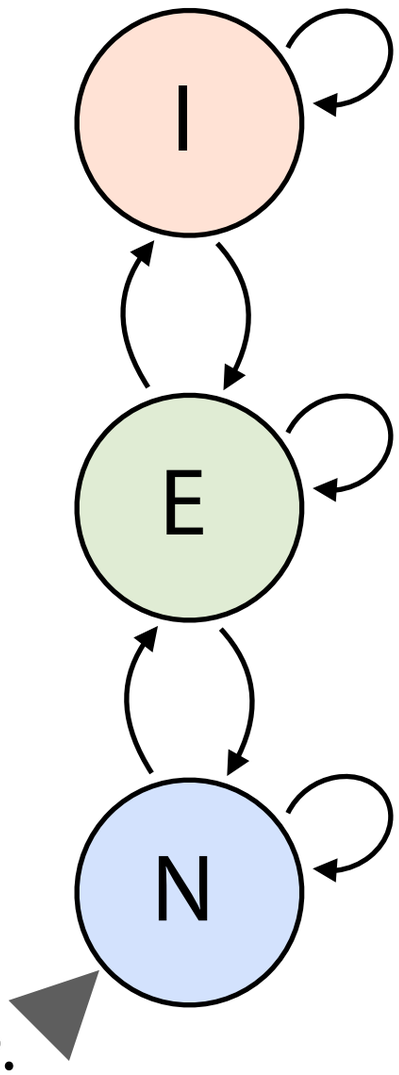
Note: a "missing" transition (E.g. N → I) indicates the transition probability is 0

# Gene finding

Emissions are:

Nucleotides

I = intron
E = exon
N = intergenic
(between genes)




Indicates *start state*.
State's initial probability is 1

Try to capture:

Exons, introns, spaces between genes

Note: a "missing" transition (E.g. N → I) indicates the transition probability is 0

# Gene finding

What if we wanted to model the three codon positions separately?

**Second letter**

| First letter | U | C | A | G | Third letter |
|---|---|---|---|---|---|
| **U** | UUU, UUC } Phe<br>UUA, UUG } Leu | UCU, UCC, UCA, UCG } Ser | UAU, UAC } Tyr<br>**UAA Stop**<br>**UAG Stop** | UGU, UGC } Cys<br>**UGA Stop**<br>UGG Trp | U C A G |
| **C** | CUU, CUC, CUA, CUG } Leu | CCU, CCC, CCA, CCG } Pro | CAU, CAC } His<br>CAA, CAG } Gln | CGU, CGC, CGA, CGG } Arg | U C A G |
| **A** | AUU, AUC, AUA } Ile<br>**AUG Met** | ACU, ACC, ACA, ACG } Thr | AAU, AAC } Asn<br>AAA, AAG } Lys | AGU, AGC } Ser<br>AGA, AGG } Arg | U C A G |
| **G** | GUU, GUC, GUA, GUG } Val | GCU, GCC, GCA, GCG } Ala | GAU, GAC } Asp<br>GAA, GAG } Glu | GGU, GGC, GGA, GGG } Gly | U C A G |

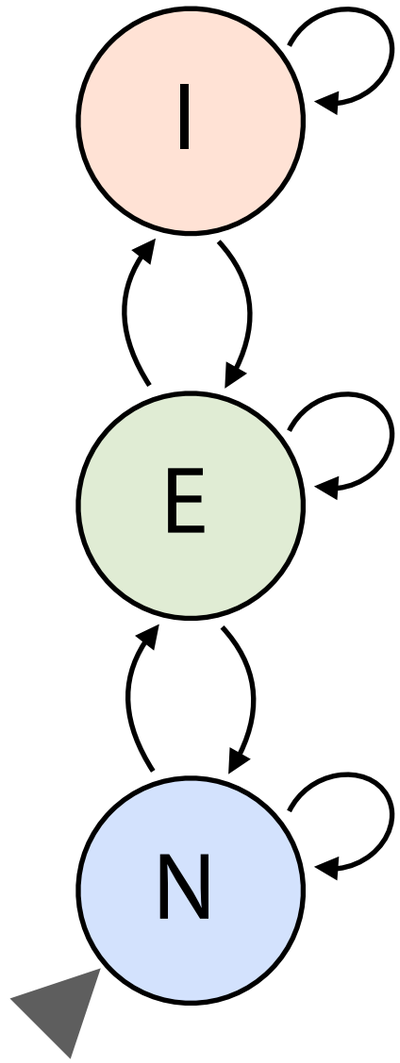
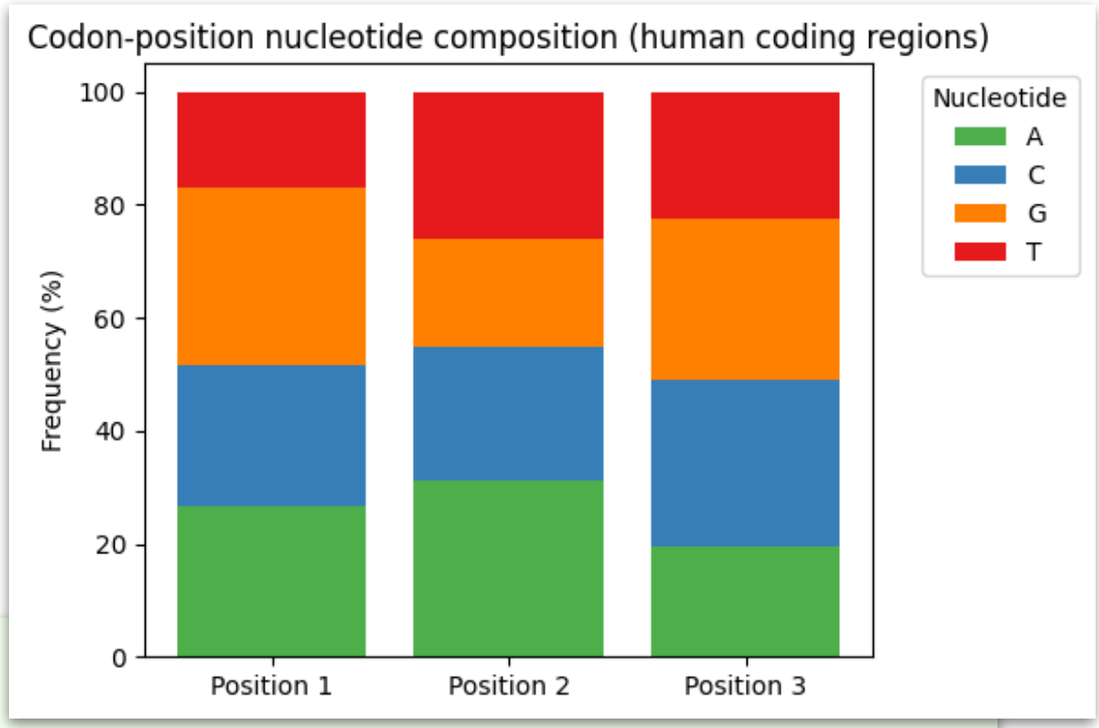



Image: https://openstax.org/books/biology/pages/15-1-the-genetic-code

# Gene finding

What if we wanted to model the three codon positions separately?



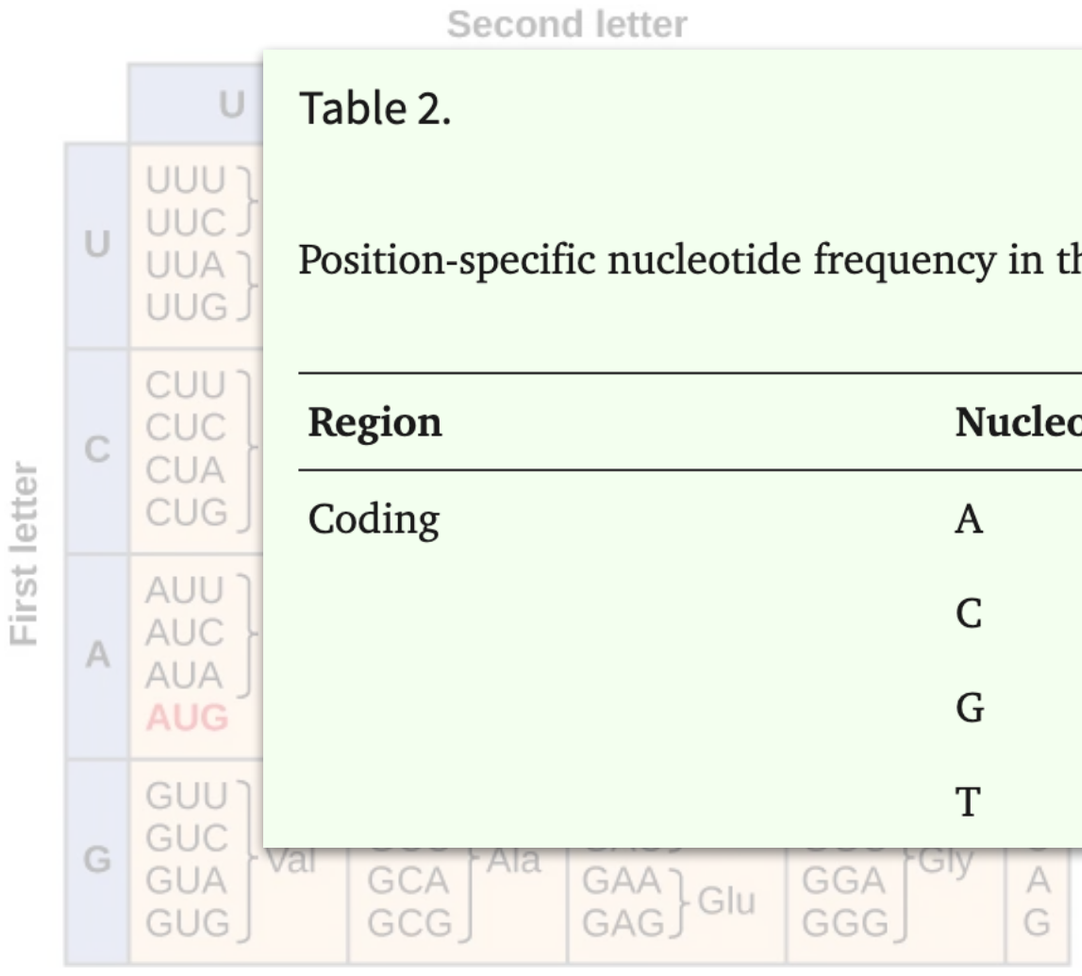




Image: https://openstax.org/books/biology/pages/15-1-the-genetic-code

Table 2.

Position-specific nucleotide frequency in the human genome

| Region | Nucleotide | Position 1 (%) | Position 2 (%) | Position 3 (%) |
|---|---|---|---|---|
| Coding | A | 26.7 | 31.1 | 19.4 |
| | C | 24.9 | 23.7 | 29.7 |
| | G | 31.4 | 19.1 | 28.6 |
| | T | 17.0 | 26.1 | 22.3 |

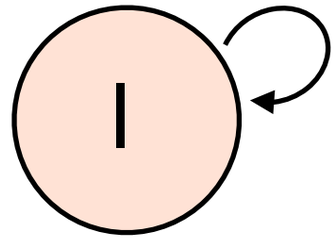
Howe, E.D. and Song, J.S., 2013. Categorical spectral analysis of periodicity in human and viral genomes. *Nucleic acids research*, *41*(3), pp.1395-1405.

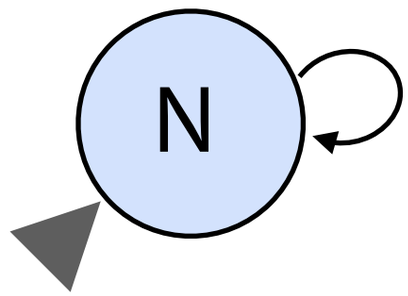
# Gene finding

What if we wanted to model the three codon positions separately?

| First letter | Second letter: U | Second letter: C | Second letter: A | Second letter: G | Third letter |
|---|---|---|---|---|---|
| U | UUU, UUC: Phe; UUA, UUG: Leu | UCU, UCC, UCA, UCG: Ser | UAU, UAC: Tyr; **UAA, UAG: Stop** | UGU, UGC: Cys; **UGA: Stop**; UGG: Trp | U C A G |
| C | CUU, CUC, CUA, CUG: Leu | CCU, CCC, CCA, CCG: Pro | CAU, CAC: His; CAA, CAG: Gln | CGU, CGC, CGA, CGG: Arg | U C A G |
| A | AUU, AUC, AUA: Ile; **AUG: Met** | ACU, ACC, ACA, ACG: Thr | AAU, AAC: Asn; AAA, AAG: Lys | AGU, AGC: Ser; AGA, AGG: Arg | U C A G |
| G | GUU, GUC, GUA, GUG: Val | GCU, GCC, GCA, GCG: Ala | GAU, GAC: Asp; GAA, GAG: Glu | GGU, GGC, GGA, GGG: Gly | U C A G |

Image: https://openstax.org/books/biology/pages/15-1-the-genetic-code



# Gene finding

What if we wanted to model the three codon positions separately?

**Second letter**

| First letter | U | C | A | G | Third letter |
|---|---|---|---|---|---|
| U | UUU, UUC } Phe<br>UUA, UUG } Leu | UCU, UCC, UCA, UCG } Ser | UAU, UAC } Tyr<br>**UAA Stop**<br>**UAG Stop** | UGU, UGC } Cys<br>**UGA Stop**<br>UGG Trp | U C A G |
| C | CUU, CUC, CUA, CUG } Leu | CCU, CCC, CCA, CCG } Pro | CAU, CAC } His<br>CAA, CAG } Gln | CGU, CGC, CGA, CGG } Arg | U C A G |
| A | AUU, AUC, AUA } Ile<br>**AUG Met** | ACU, ACC, ACA, ACG } Thr | AAU, AAC } Asn<br>AAA, AAG } Lys | AGU, AGC } Ser<br>AGA, AGG } Arg | U C A G |
| G | GUU, GUC, GUA, GUG } Val | GCU, GCC, GCA, GCG } Ala | GAU, GAC } Asp<br>GAA, GAG } Glu | GGU, GGC, GGA, GGG } Gly | U C A G |

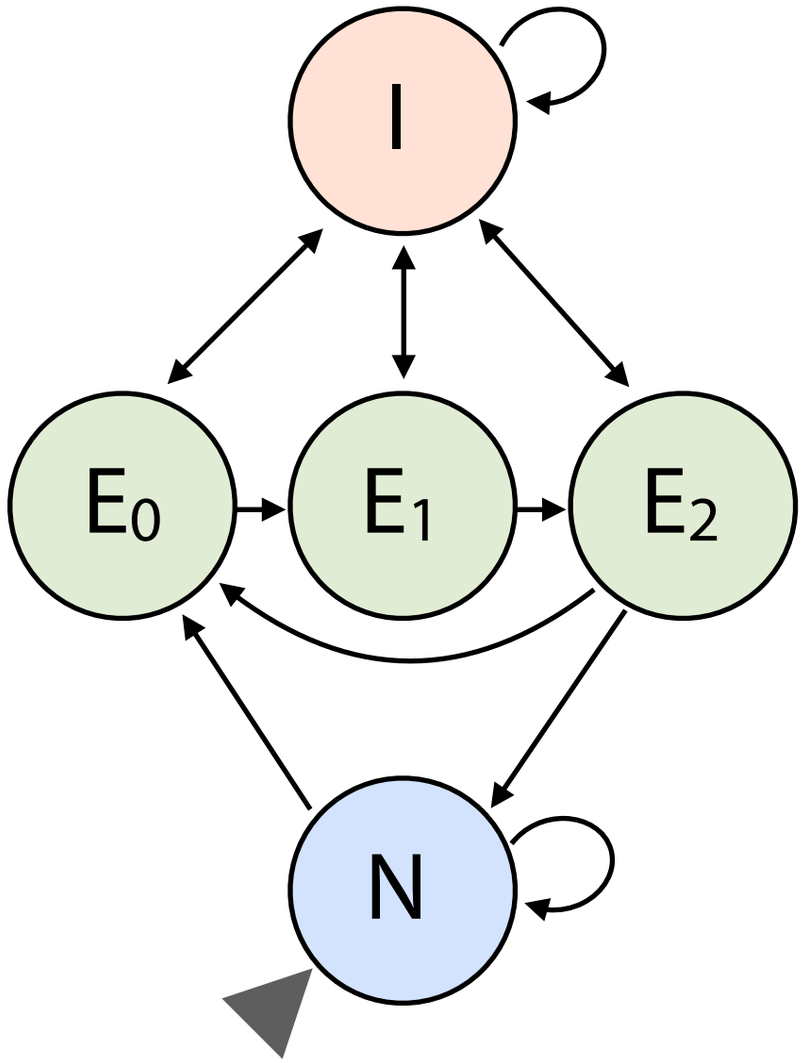
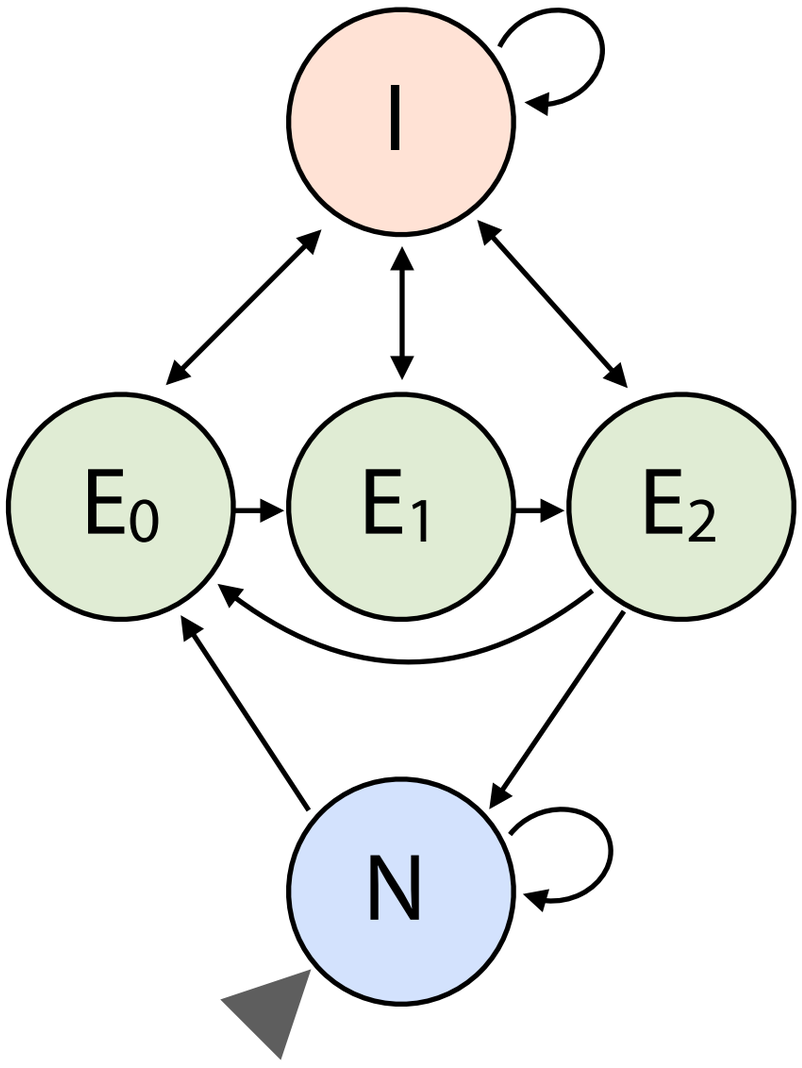


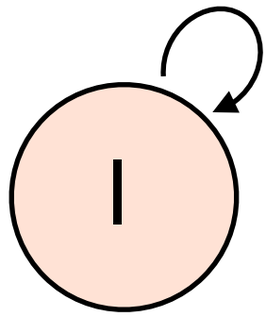

Image: https://openstax.org/books/biology/pages/15-1-the-genetic-code

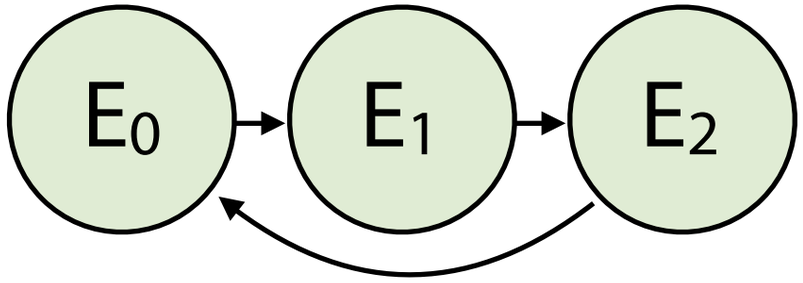
# Gene finding



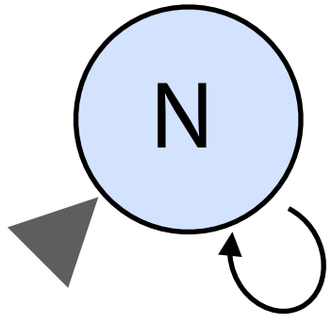

Can we additionally model **start** & **stop** codons and **donors** & **acceptors**?

# Gene finding

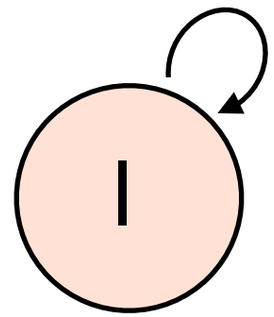


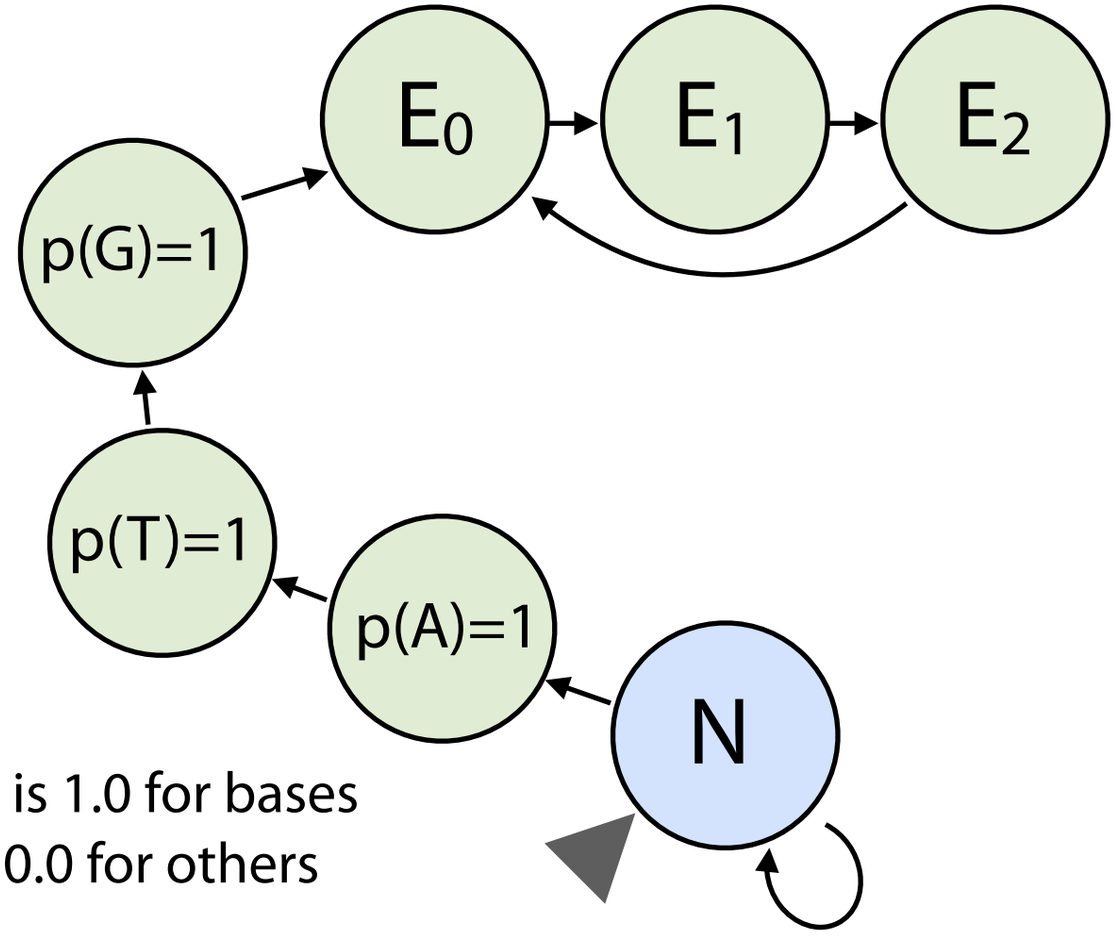




Start codon



# Gene finding



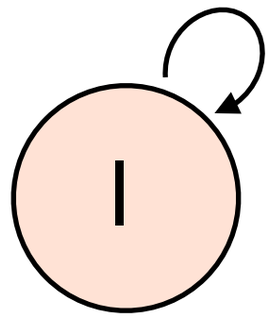

## Start codon

Emission probability is 1.0 for bases shown (AUG/ATG), 0.0 for others

## Stop codons

UAG/TAG,
UAA/TAA,
UGA/TGA

# Gene finding

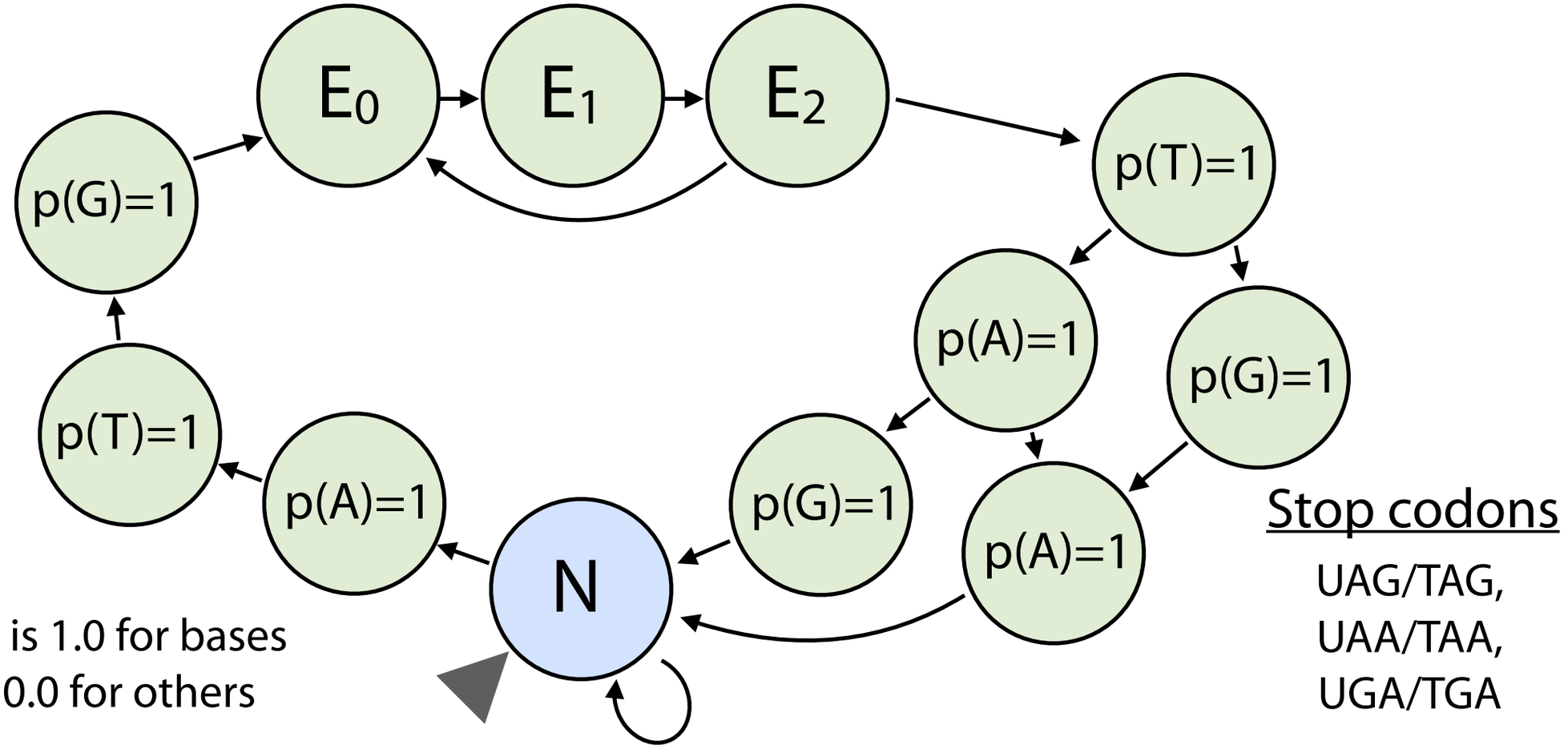



<u>Donor</u>

GU/GT

<u>Acceptor</u>

AG



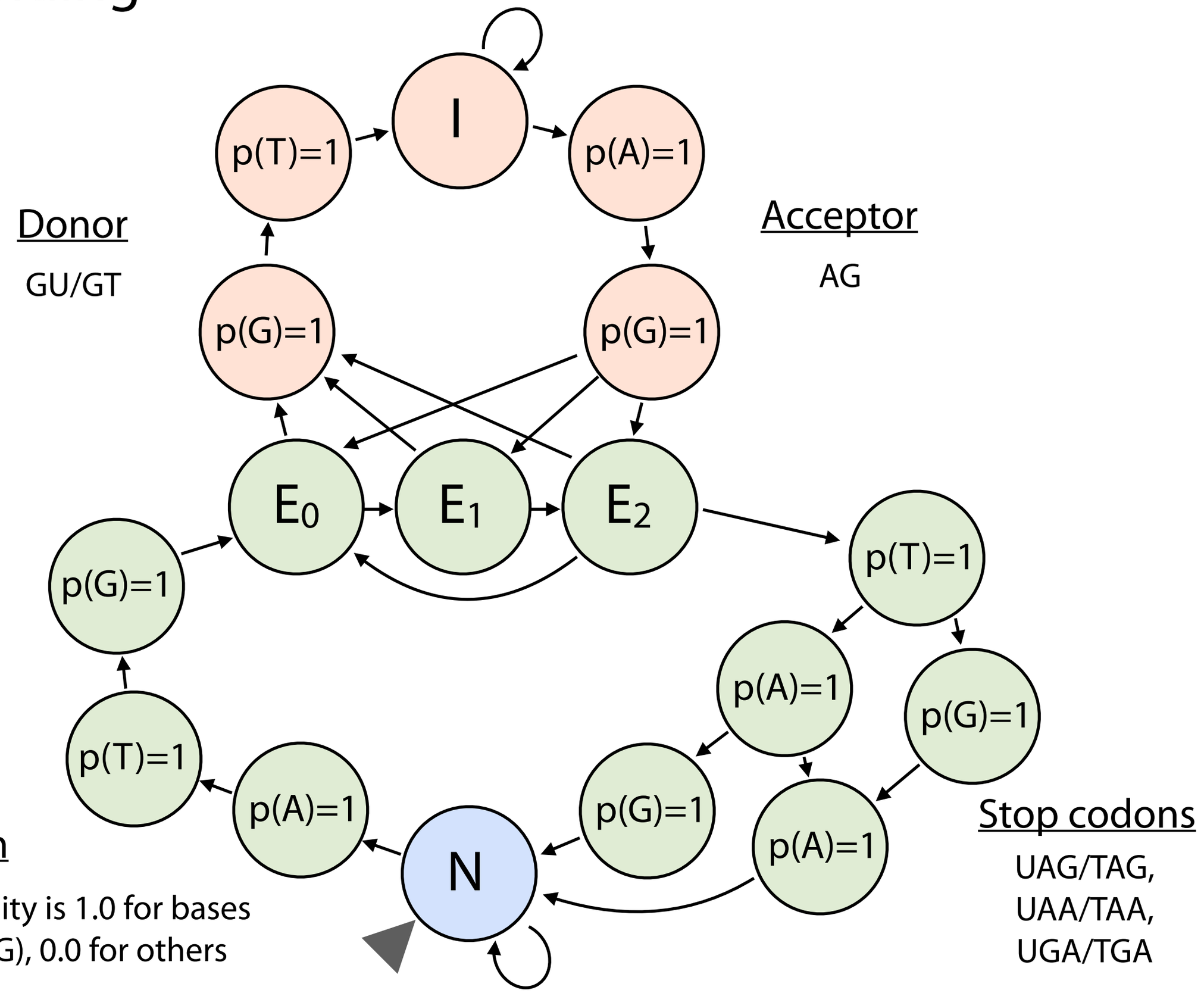

<u>Start codon</u>

Emission probability is 1.0 for bases shown (AUG/ATG), 0.0 for others

<u>Stop codons</u>

UAG/TAG,
UAA/TAA,
UGA/TGA

# Gene finding




Donor

GU/GT

Acceptor

AG

Start codon

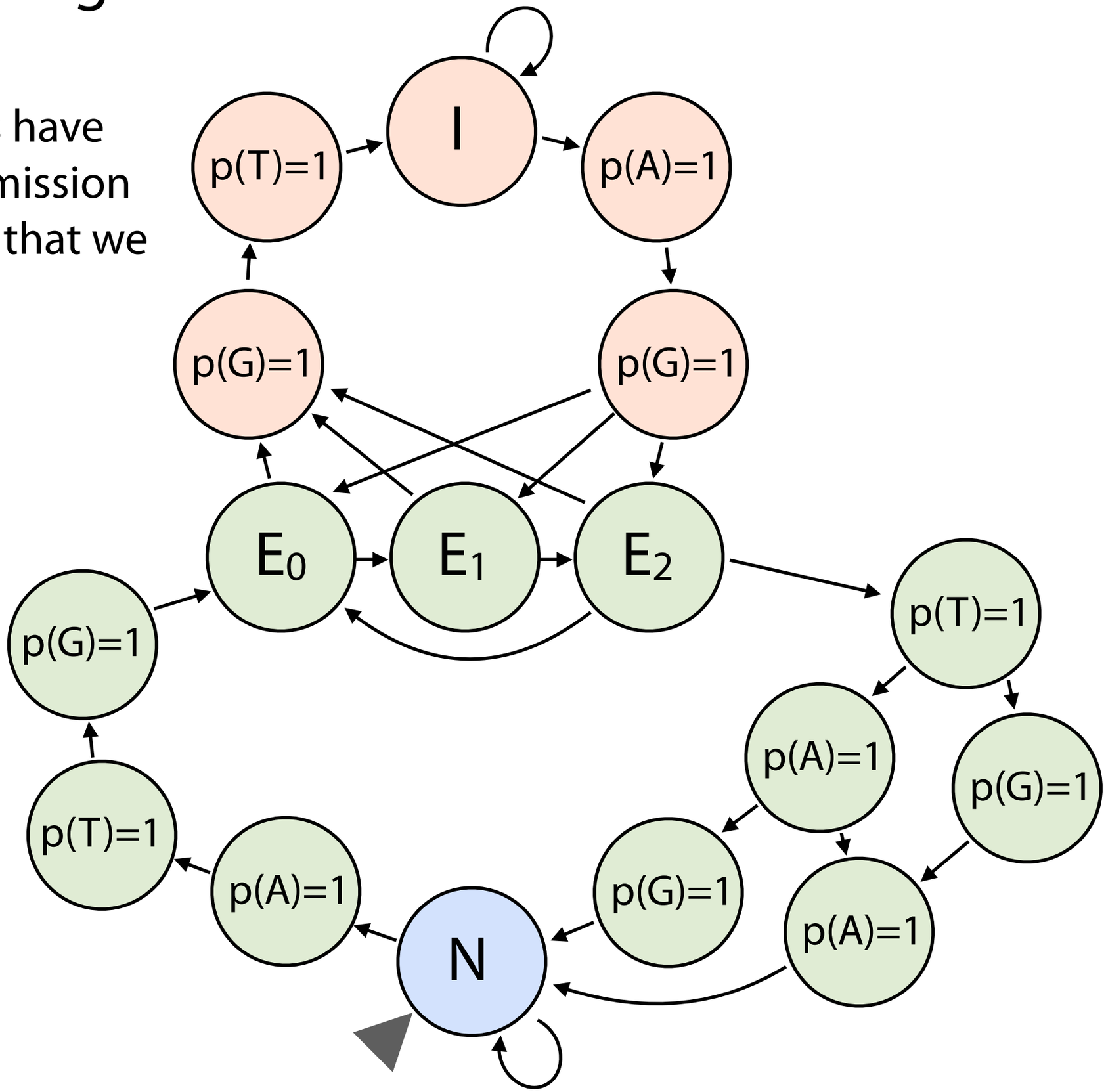
Emission probability is 1.0 for bases shown (AUG/ATG), 0.0 for others

Stop codons

UAG/TAG,
UAA/TAA,
UGA/TGA

# Gene finding

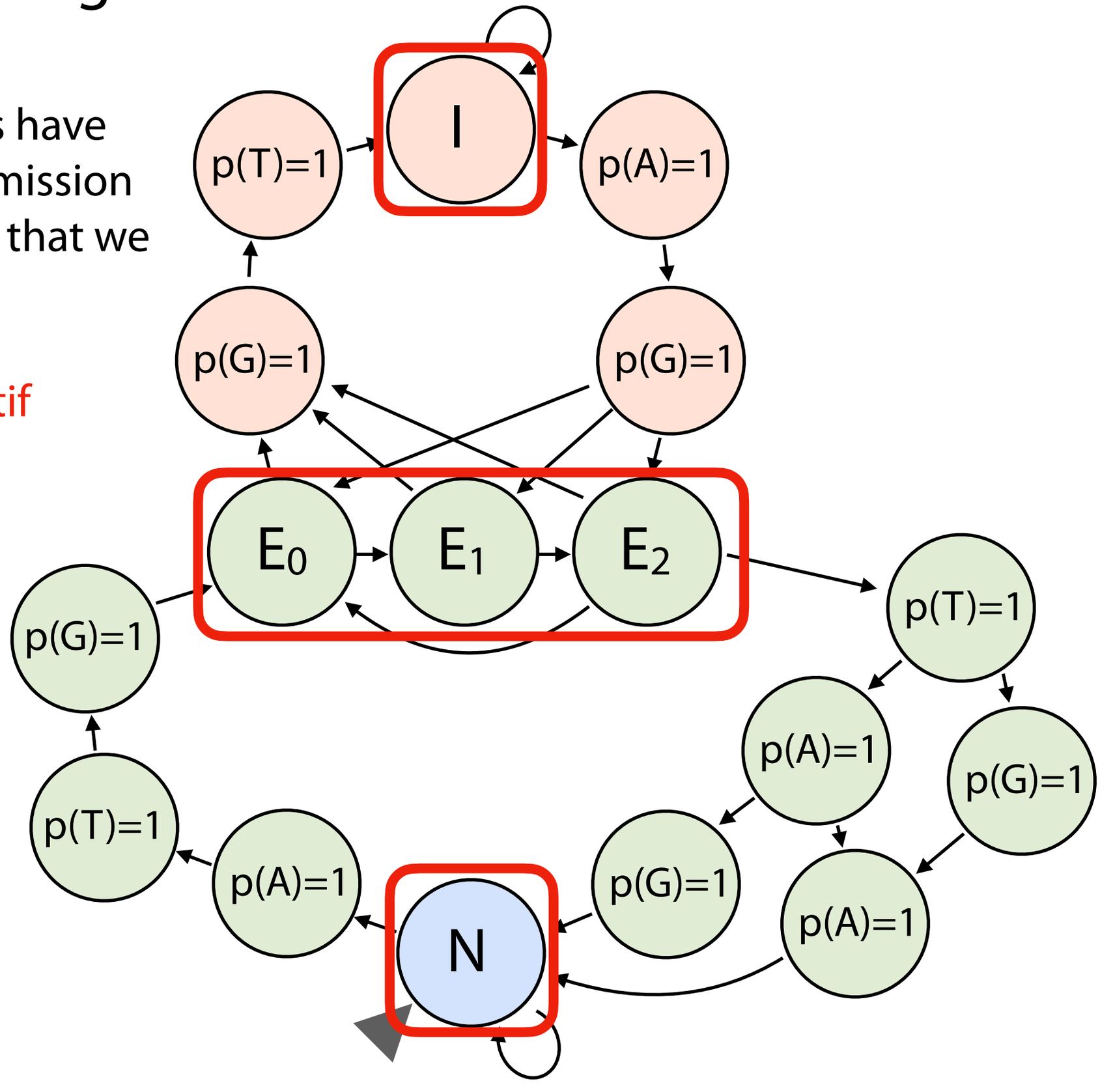
Which nodes have non-trivial emission probabilities that we must learn?



# Gene finding

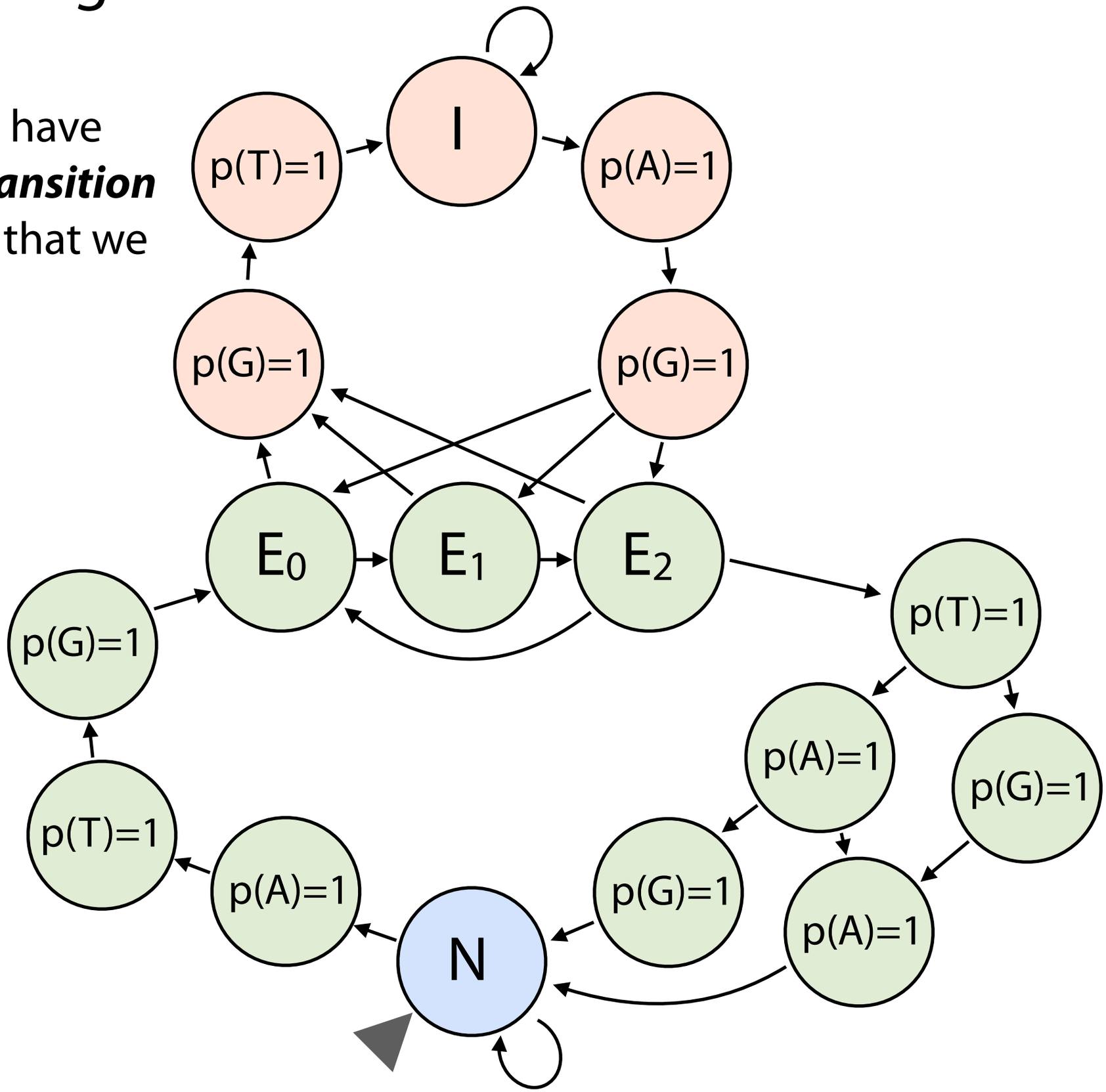
Which nodes have non-trivial emission probabilities that we must learn?

The non-motif nodes



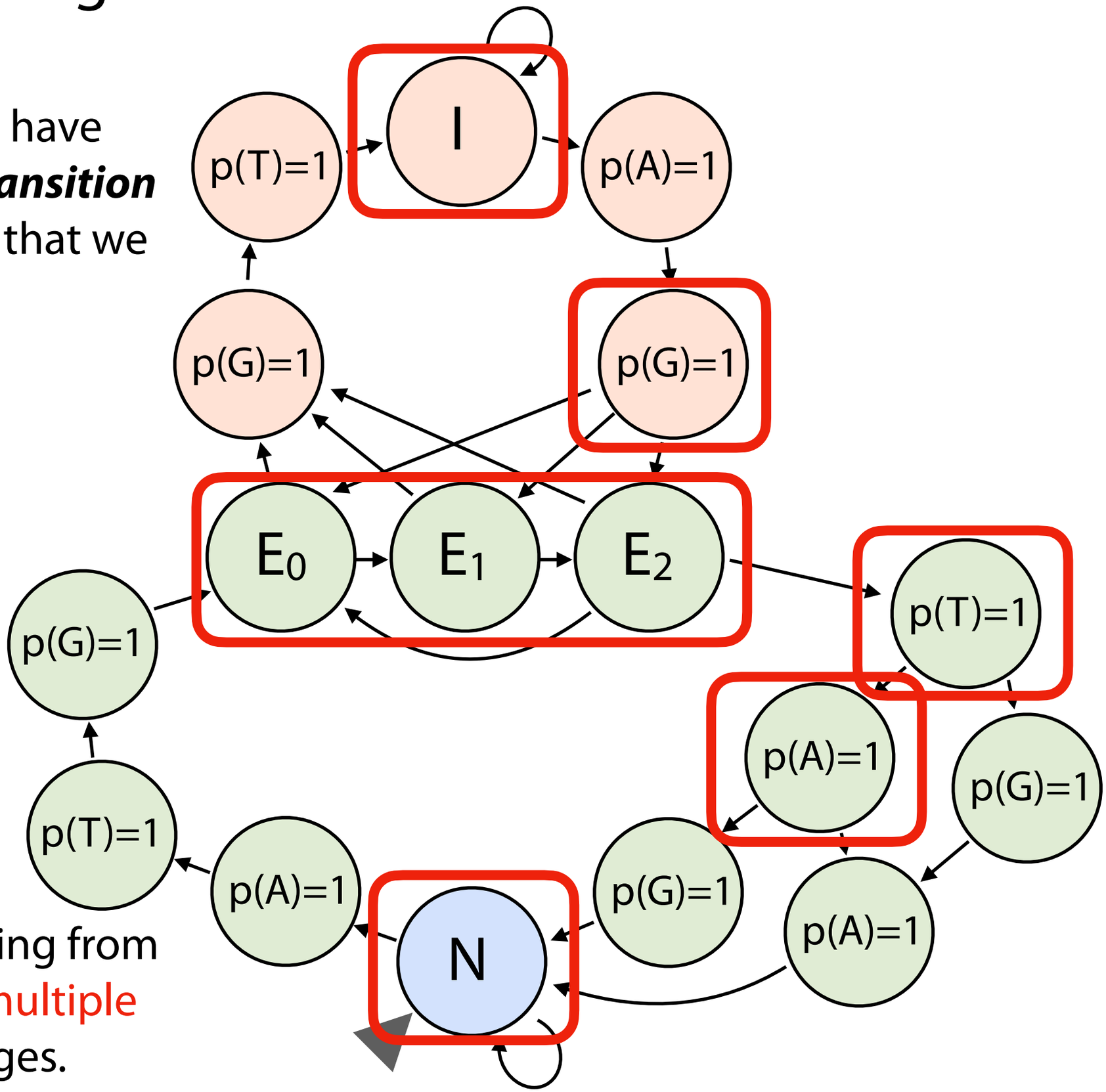
# Gene finding

Which edges have non-trivial ***transition*** probabilities that we must learn?



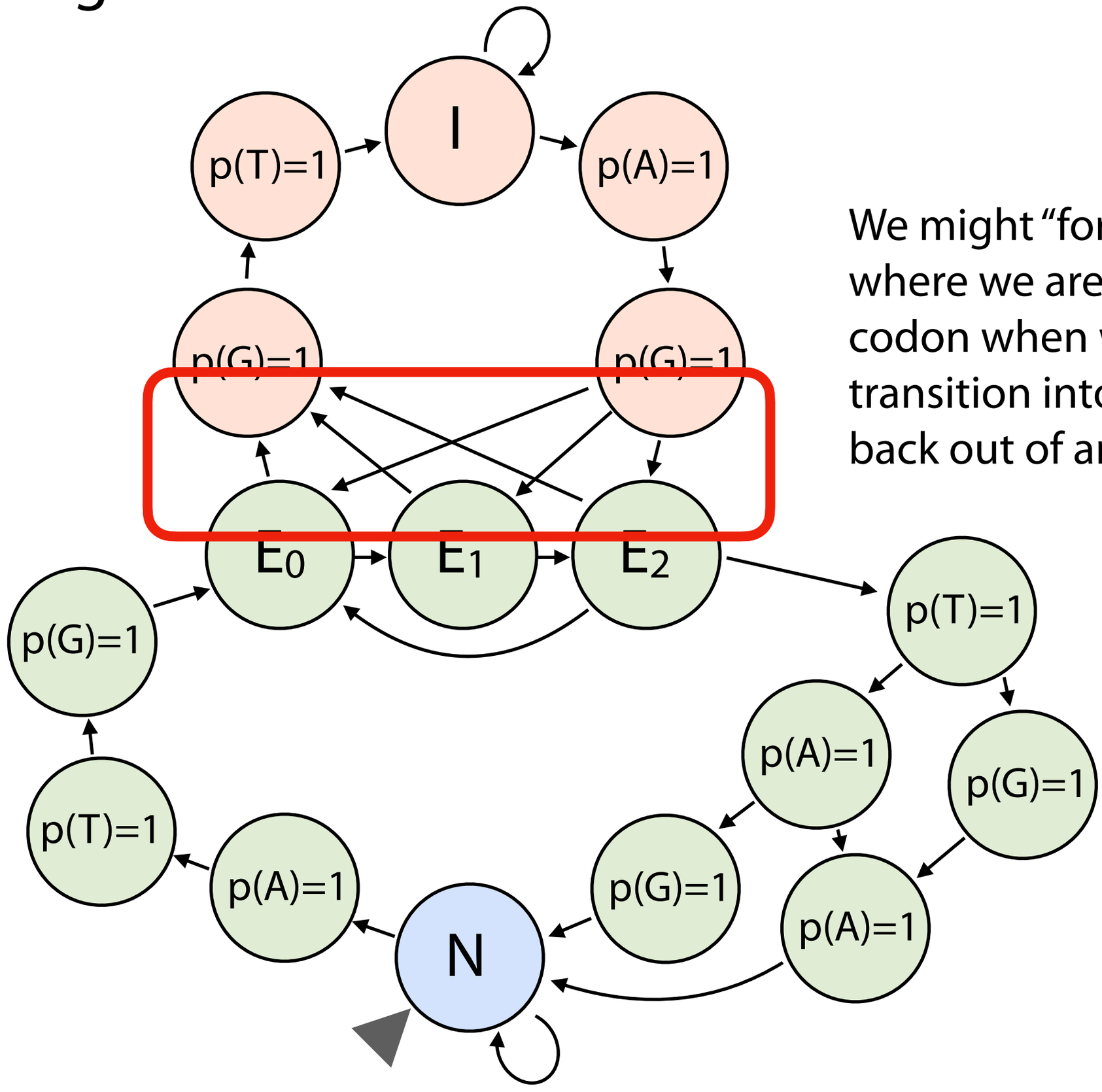
# Gene finding

Which edges have non-trivial ***transition*** probabilities that we must learn?



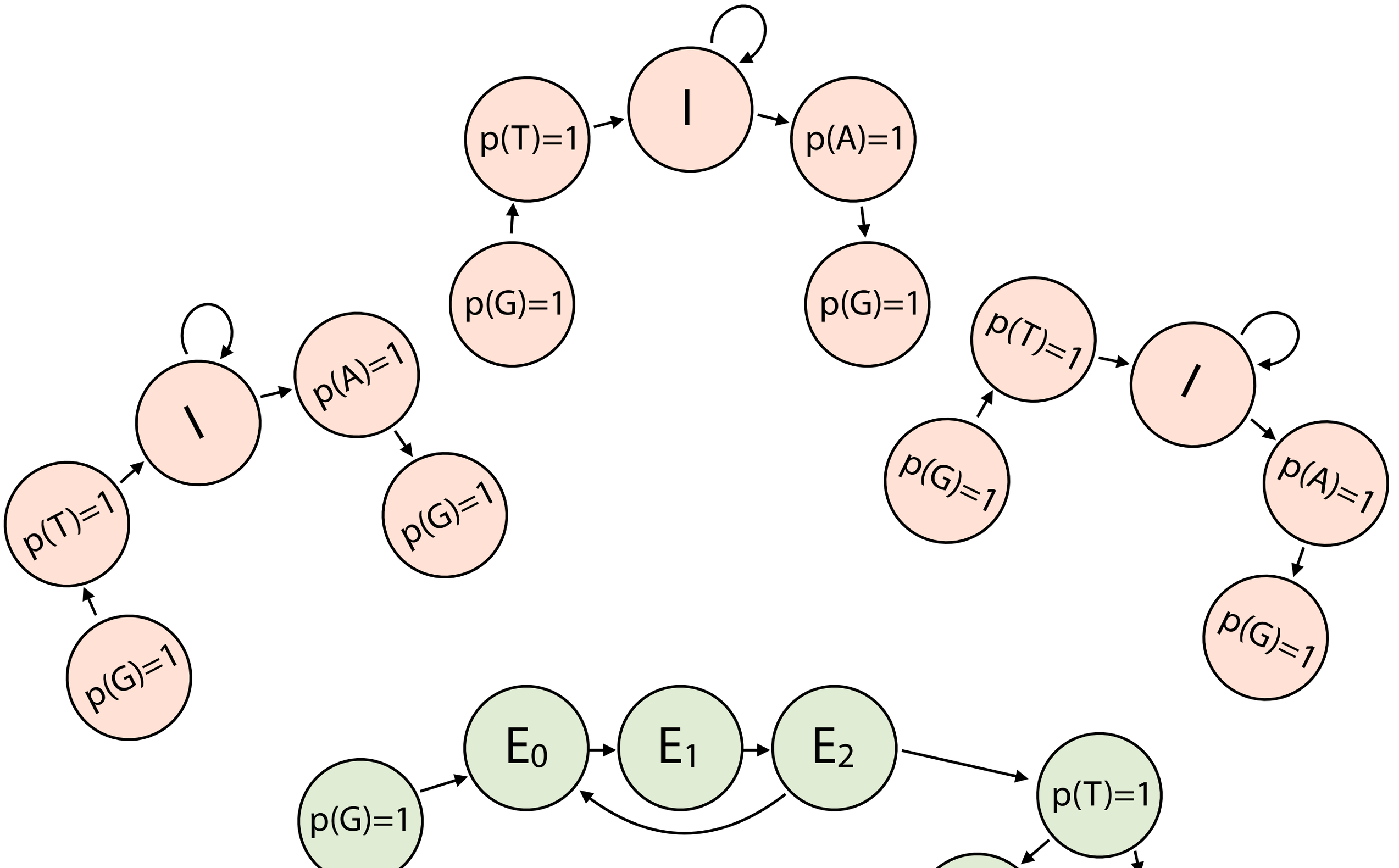

Edges outgoing from nodes with multiple outgoing edges.

# Gene finding



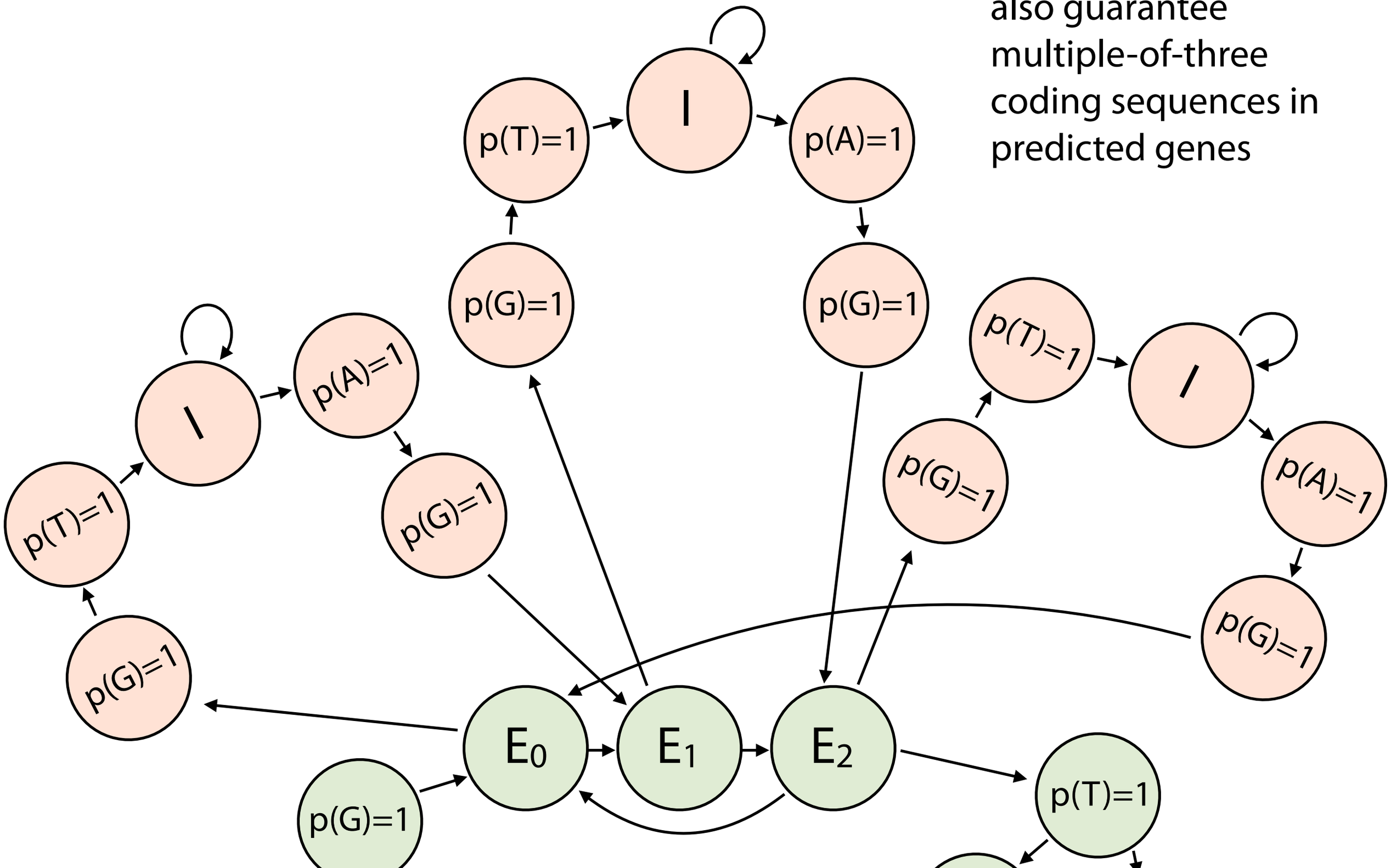

We might "forget" where we are in the codon when we transition into then back out of an intron

# Gene finding



# Gene finding

This change should also guarantee multiple-of-three coding sequences in predicted genes



# Gene finding

In short, we can capture exons and their **codons**

with distinct emissions for distinct codon positions

with special consideration for start & stop codons

without forgetting codon position during introns

requiring multiple-of-three coding sequences

As well as **introns**

with special consideration for donors & acceptors

with an emission profile distinct from exons

All of which emerges from an **intergenic** background

with its own distinct emission profile

# Gene finding

We can sequence all the mRNAs in a cell and use read alignment to find where they came from — those are genes!

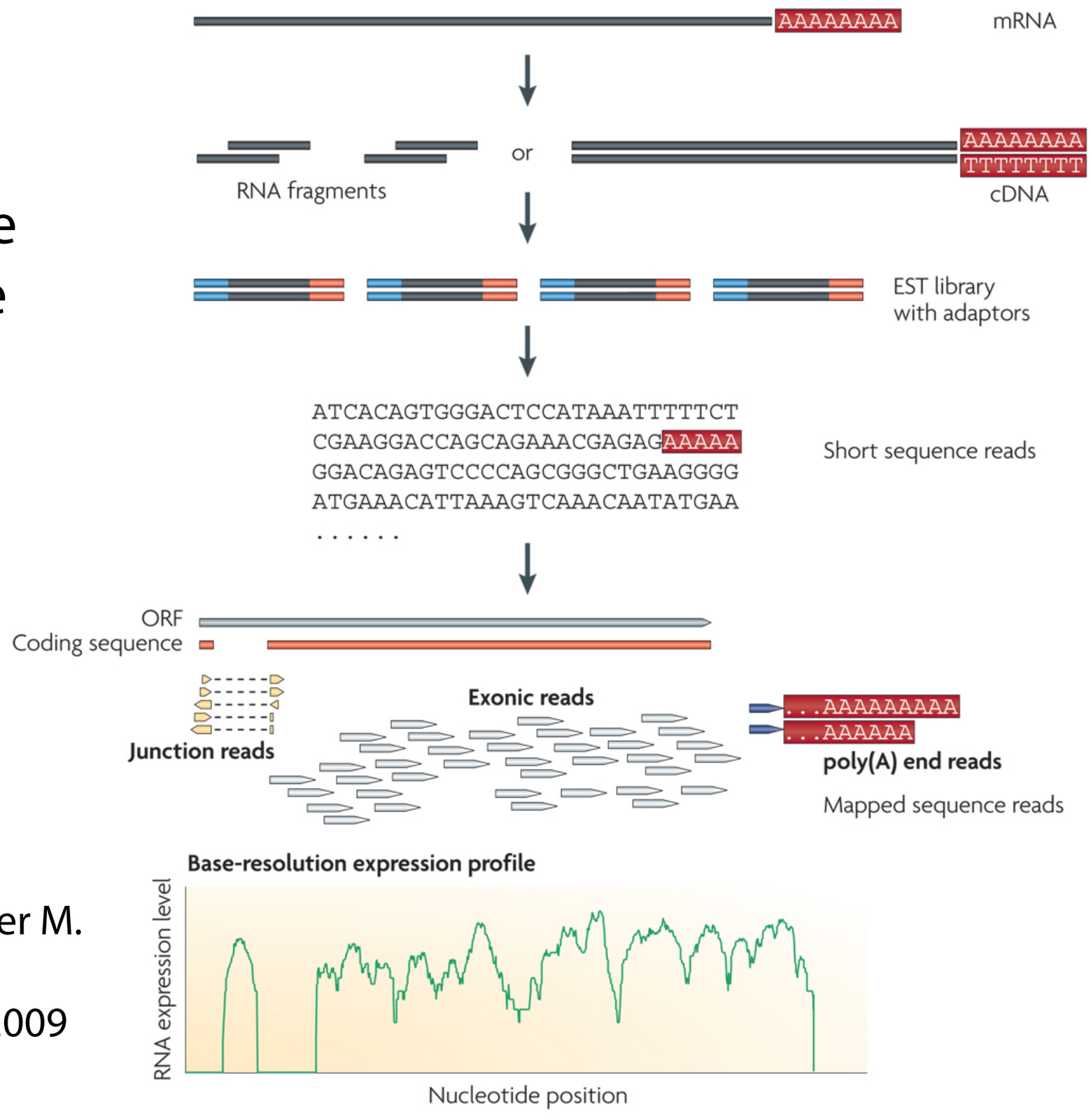



Image: Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63.

# Gene finding

Pertea *et al. Genome Biology* (2018) 19:208
https://doi.org/10.1186/s13059-018-1590-2

Genome Biology

DATABASE Open Access

## CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise

Mihaela Pertea[1,3], Alaina Shumate[1,2], Geo Pertea[1], Ales Varabyou[1,3], Florian P. Breitwieser[1], Yu-Chi Chang[3], Anil K. Madugundu[4,5,6,8], Akhilesh Pandey[4,7,8] and Steven L. Salzberg[1,2,3,9*]

**Abstract**

We assembled the sequences from deep RNA sequencing experiments by the Genotype-Tissue Expression (GTEx) project, to create a new catalog of human genes and transcripts, called CHESS. The new database contains 42,611 genes, of which 20,352 are potentially protein-coding and 22,259 are noncoding, and a total of 323,258 transcripts. These include 224 novel protein-coding genes and 116,156 novel transcripts. We detected over 30 million additional transcripts at more than 650,000 genomic loci, nearly all of which are likely nonfunctional, revealing a heretofore unappreciated amount of transcriptional noise in human cells. The CHESS database is available at http://ccb.jhu.edu/chess.

**Keywords:** Human gene count, GTEx, RNA sequencing, Transcriptome, Transcriptome assembly

### Background

Scientists have been attempting to estimate the number of human genes for more than 50 years, dating back to 1964 [1]. In the decade preceding the initial publication of the human genome, multiple estimates were made based on sequencing of short messenger RNA frag- analysis suggested 20,500 [9], and a proteomics-based study in 2014 estimated 19,000 [10].

One striking feature of most early attempts to catalog all human genes was their lack of precision. Most estimates have only one to two significant digits, indicating major uncertainty about the exact number. As we re-