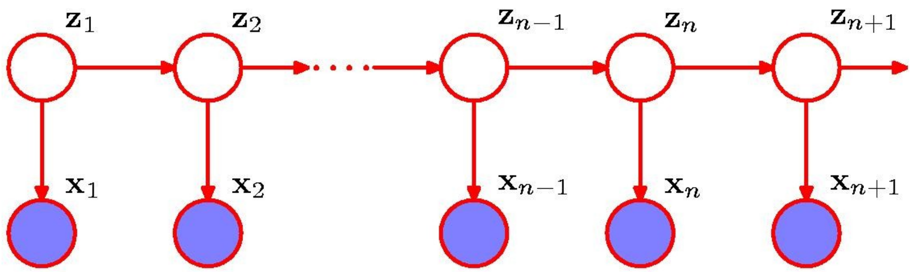
Hidden Markov Models

Selecting the initial model parameters

Using HMMs for (simpel) gene finding



Christian Nørgaard Storm Pedersen

cstorm@birc.au.dk

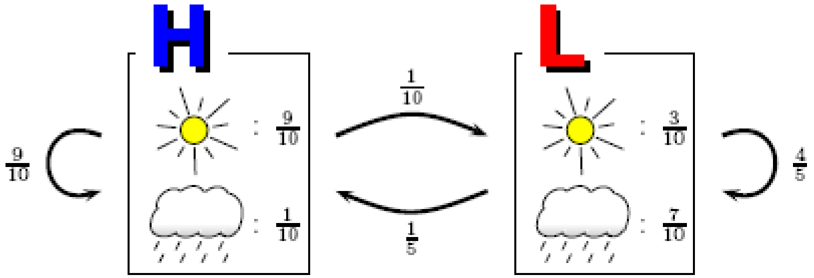
Last time

Training, or how to select model parameters (transition and emission probabilities) to reflect either a set of corresponding (**X**,**Z**)'s (Training by Counting), or just a set of **X**'s (Viterbi Training, and EM Training).

HMMs as a generative model

A HMM ***generates a sequence of observables*** by moving from hidden state to hidden state according to the transition probabilities and ***emitting an observable*** (from a discrete set of observables, i.e. a finite alphabet) from each hidden state visited ***according to the emission probabilities*** of the state ...

Model *M*:



A run follows a sequence of states:

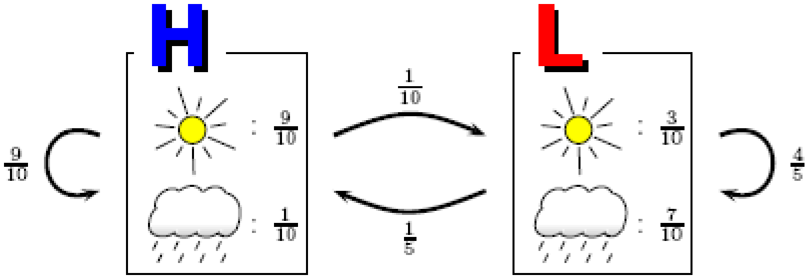
H H L L H

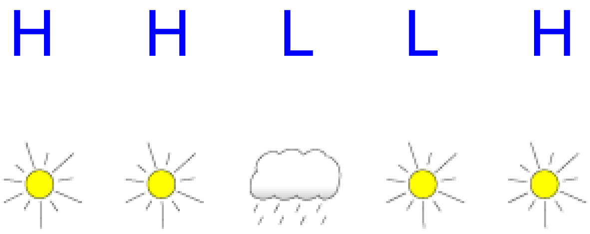
And emits a sequence of symbols:

For a HMM that generates finite strings (e.g. a HMM with an end-state), the language L = {**X** | *p*(**X**) > 0} is regular ...

Selecting initial model parameters

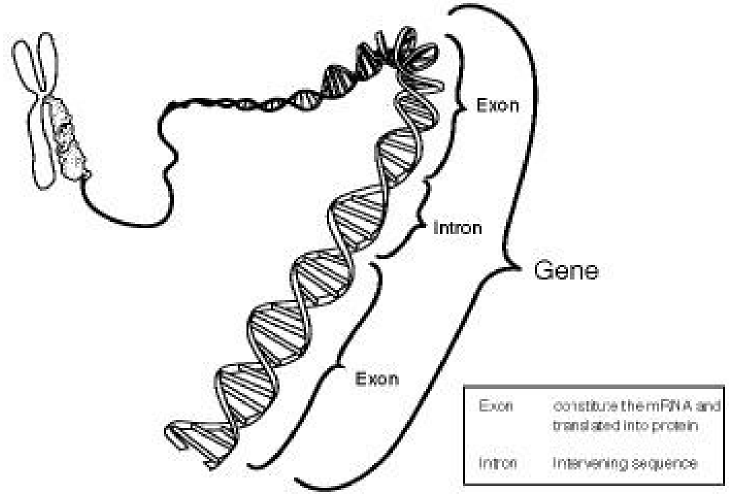
The initial selection of transition and emission probabilities, i.e. A, π, Φ, models (how we see) the underlying structure of the observations, i.e. the syntax of possible sequences of observations, recall that the language $L = \{x \mid P(x \mid \theta) > 0\}$ is regular.

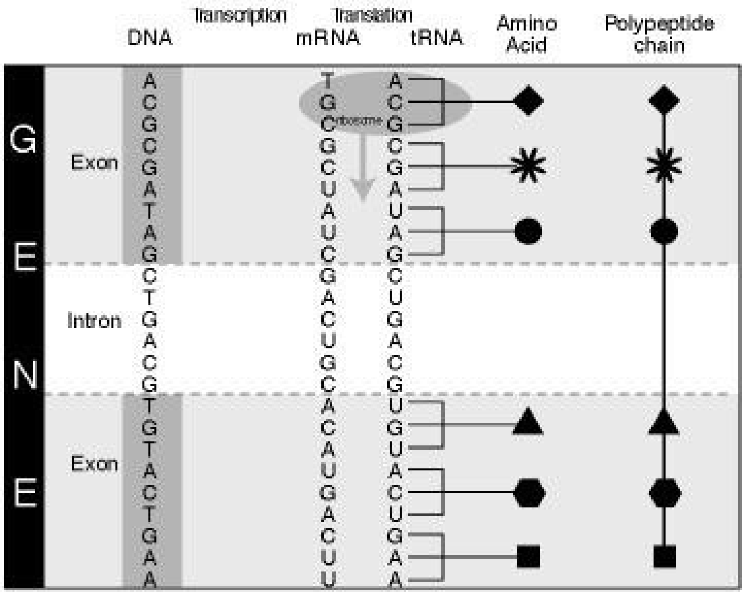




The initial selection of parameters is essential just to decide which parameters are 0 (or 1), i.e. to decide which transitions of emission should never (or always) be possible ...

Example – Gene finding





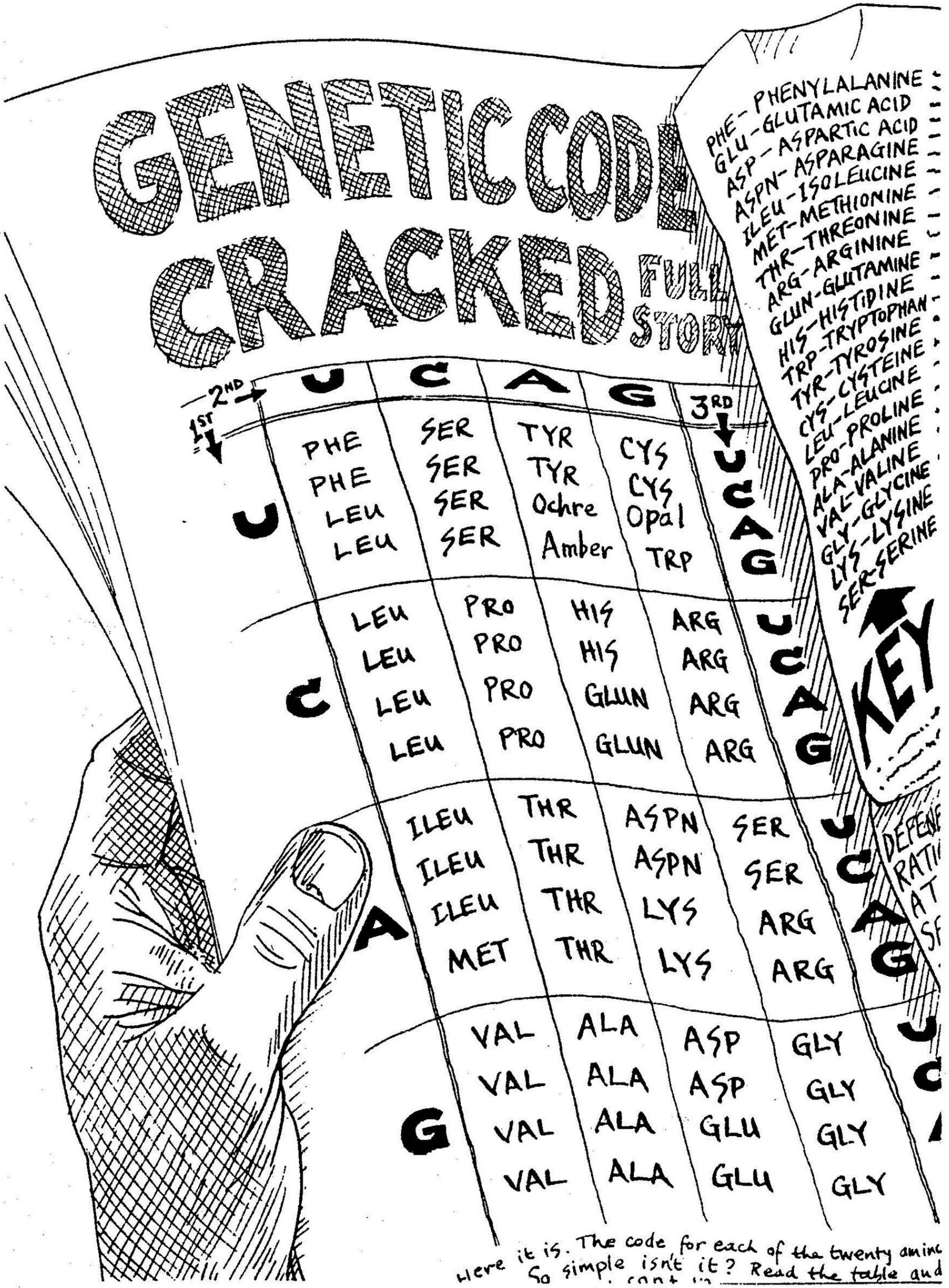
Each protein is encoded in a stretch of DNA. A gene ...

Which is expressed when the protein is needed ...

Important problem

Locating genes on the genome and determining how they get expressed ...

Recognizing the patterns that indicates a gene ...



GENETIC CODE CRACKED FULL STORY

1ST ↓ / 2ND →	U	C	A	G	3RD ↓
U	PHE PHE LEU LEU	SER SER SER SER	TYR TYR Ochre Amber	CYS CYS Opal TRP	U C A G
C	LEU LEU LEU LEU	PRO PRO PRO PRO	HIS HIS GLUN GLUN	ARG ARG ARG ARG	U C A G
A	ILEU ILEU ILEU MET	THR THR THR THR	ASPN ASPN LYS LYS	SER SER ARG ARG	U C A G
G	VAL VAL VAL VAL	ALA ALA ALA ALA	ASP ASP GLU GLU	GLY GLY GLY GLY	U C

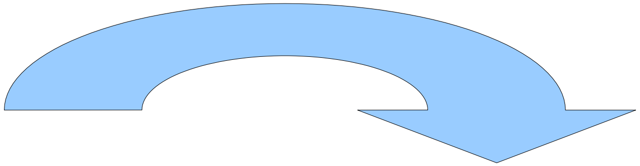
PHE - PHENYLALANINE
GLU - GLUTAMIC ACID
ASP - ASPARTIC ACID
ASPN - ASPARAGINE
ILEU - ISOLEUCINE
MET - METHIONINE
THR - THREONINE
ARG - ARGININE
GLUN - GLUTAMINE
HIS - HISTIDINE
TRP - TRYPTOPHAN
TYR - TYROSINE
CYS - CYSTEINE
LEU - LEUCINE
PRO - PROLINE
ALA - ALANINE
VAL - VALINE
GLY - GLYCINE
LYS - LYSINE
SER - SERINE

KEY

Here it is. The code for each of the twenty aminc
So simple isn't it? Read the table and

```
>NC_002737.1 Streptococcus pyogenes M1 GAS
TTGTTGATATTCTGTTTTTTCTTTTTTAGTTTTCCACATGAAAAATAGTTGAAAACAATA
GCGGTGTCCCCTTAAAATGGCTTTTCCACAGGTTGTGGAGAACCCAAATTAACAGTGTTA
ATTTATTTTCCACAGGTTGTGGAAAAACTAACTATTATCCATCGTTCTGTGGAAAACTAG
AATAGTTTATGGTAGAATAGTTCTAGAATTATCCACAAGAAGGAACCTAGTATGACTGAA
AATGAACAAATTTTTTGGAACAGGGTCTTGGAATTAGCTCAGAGTCAATTAAAACAGGCA
ACTTATGAATTTTTTGTTCATGATGCCCGTCTATTAAAGGTCGATAAGCATATTGCAACT
ATTTACTTAGATCAAATGAAAGAGCTCTTTTGGGAAAAAAATCTTAAAGATGTTATTCTT
ACTGCTGGTTTTGAAGTTTATAACGCTCAAATTTCTGTTGACTATGTTTTCGAAGAAGAC
CTAATGATTGAGCAAAATCAGACCAAAATCAACCAAAAACCTAAGCAGCAAGCCTTAAAT
TCTTTGCCTACTGTTACTTCAGATTTAAACTCGAAATATAGTTTTGAAAACTTTATTCAA
GGAGATGAAAATCGTTGGGCTGTTGCTGCTTCAATAGCAGTAGCTAATACTCCTGGAACT
ACCTATAATCCTTTGTTTATTTGGGGTGGCCCTGGGCTTGGAAAAACCCATTTATTAAAT
GCTATTGGTAATTCTGTACTATTAGAAAATCCAAATGCTCGAATTAAATATATCACAGCT
GAAAACTTTATTAATGAGTTTGTTATCCATATTCGCCTTGATACCATGGATGAATTGAAA
GAAAAATTTCGTAATTTAGATTTACTCCTTATTGATGATATCCAATCTTTAGCTAAAAAA
ACGCTCTCTGGAACACAAGAAGAGTTCTTTAATACTTTTAATGCACTTCATAATAATAAC
AAACAAATTGTCCTAACAAGCGACCGTACACCAGATCATCTCAATGATTTAGAAGATCGA
TTAGTTACTCGTTTTAAATGGGGATTAACAGTCAATATCACACCTCCTGATTTTGAAACA
CGAGTGGCTATTTTGACAAATAAAATTCAAGAATATAACTTTATTTTTCCTCAAGATACC
ATTGAGTATTTGGCTGGTCAATTTGATTCTAATGTCAGAGATTTAGAAGGTGCCTTAAAA
GATATTAGTCTGGTTGCTAATTTCAAACAAATTGACACGATTACTGTTGACATTGCTGCC
GAAGCTATTCGCGCCAGAAAGCAAGATGGACCTAAAATGACAGTTATTCCCATCGAAGAA
ATTCAAGCGCAAGTTGGAAAATTTTACGGTGTTACCGTCAAAGAAATTAAAGCTACTAAA
CGAACACAAAATATTGTTTTAGCAAGACAAGTAGCTATGTTTTTAGCACGTGAAATGACA
GATAACAGTCTTCCTAAAATTGGAAAAGAATTTGGTGGCAGAGACCATTCAACAGTACTC
CATGCCTATAATAAAATCAAAAACATGATCAGCCAGGACGAAAGCCTTAGGATCGAAATT
GAAACCATAAAAAACAAAATTAAATAACATGTGGAAAAGAATATCTTTTATGAAATAGTT
ATCCACAAGTTGTGAACATCCATTTAGTCTTGGATTCTCTCGTTTATTTAGAGTTATCCA
CTATATACACAAGACCTACTACTACTACTTATTATTATACTTATTAAATAAAGGAGTTCT
```

Viterbi decoding



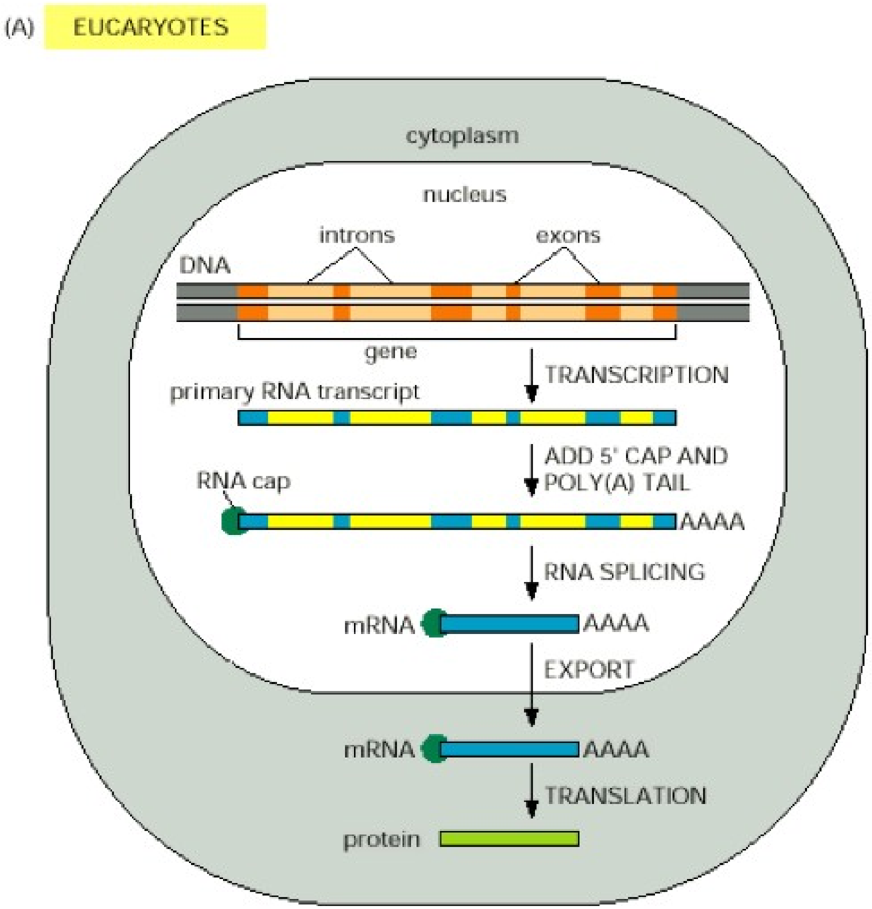
```
>NC_002737.1 Streptococcus pyogenes M1 GAS
TTGTTGATATTCTGTTTTTTCTTTTTTAGTTTTCCACATGAAAAATAGTTGAAAACAATA
GCGGTGTCCCCTTAAAATGGCTTTTCCACAGGTTGTGGAGAACCCAAATTAACAGTGTTA
ATTTATTTTCCACAGGTTGTGGAAAAACTAACTATTATCCATCGTTCTGTGGAAAACTAG
AATAGTTTATGGTAGAATAGTTCTAGAATTATCCACAAGAAGGAACCTAGTATGACTGAA
AATGAACAAATTTTTTGGAACAGGGTCTTGGAATTAGCTCAGAGTCAATTAAAACAGGCA
ACTTATGAATTTTTTGTTCATGATGCCCGTCTATTAAAGGTCGATAAGCATATTGCAACT
ATTTACTTAGATCAAATGAAAGAGCTCTTTTGGGAAAAAAATCTTAAAGATGTTATTCTT
ACTGCTGGTTTTGAAGTTTATAACGCTCAAATTTCTGTTGACTATGTTTTCGAAGAAGAC
CTAATGATTGAGCAAAATCAGACCAAAATCAACCAAAAACCTAAGCAGCAAGCCTTAAAT
TCTTTGCCTACTGTTACTTCAGATTTAAACTCGAAATATAGTTTTGAAAACTTTATTCAA
GGAGATGAAAATCGTTGGGCTGTTGCTGCTTCAATAGCAGTAGCTAATACTCCTGGAACT
ACCTATAATCCTTTGTTTATTTGGGGTGGCCCTGGGCTTGGAAAAACCCATTTATTAAAT
GCTATTGGTAATTCTGTACTATTAGAAAATCCAAATGCTCGAATTAAATATATCACAGCT
GAAAACTTTATTAATGAGTTTGTTATCCATATTCGCCTTGATACCATGGATGAATTGAAA
GAAAAATTTCGTAATTTAGATTTACTCCTTATTGATGATATCCAATCTTTAGCTAAAAAA
ACGCTCTCTGGAACACAAGAAGAGTTCTTTAATACTTTTAATGCACTTCATAATAATAAC
AAACAAATTGTCCTAACAAGCGACCGTACACCAGATCATCTCAATGATTTAGAAGATCGA
TTAGTTACTCGTTTTAAATGGGGATTAACAGTCAATATCACACCTCCTGATTTTGAAACA
CGAGTGGCTATTTTGACAAATAAAATTCAAGAATATAACTTTATTTTTCCTCAAGATACC
ATTGAGTATTTGGCTGGTCAATTTGATTCTAATGTCAGAGATTTAGAAGGTGCCTTAAAA
GATATTAGTCTGGTTGCTAATTTCAAACAAATTGACACGATTACTGTTGACATTGCTGCC
GAAGCTATTCGCGCCAGAAAGCAAGATGGACCTAAAATGACAGTTATTCCCATCGAAGAA
ATTCAAGCGCAAGTTGGAAAATTTTACGGTGTTACCGTCAAAGAAATTAAAGCTACTAAA
CGAACACAAAATATTGTTTTAGCAAGACAAGTAGCTATGTTTTTAGCACGTGAAATGACA
GATAACAGTCTTCCTAAAATTGGAAAAGAATTTGGTGGCAGAGACCATTCAACAGTACTC
CATGCCTATAATAAAATCAAAAACATGATCAGCCAGGACGAAAGCCTTAGGATCGAAATT
GAAACCATAAAAAACAAAATTAAATAACATGTGGAAAAGAATATCTTTTATGAAATAGTT
ATCCACAAGTTGTGAACATCCATTTAGTCTTGGATTCTCTCGTTTATTTAGAGTTATCCA
CTATATACACAAGACCTACTACTACTACTTATTATTATACTTATTAAATAAAGGAGTTCT
```

```
>NC_002737.1 gene annotation Streptococcus pyogenes M1 GAS
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
CCCCCCCCCCCCCCCCCCCCCCCCCCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

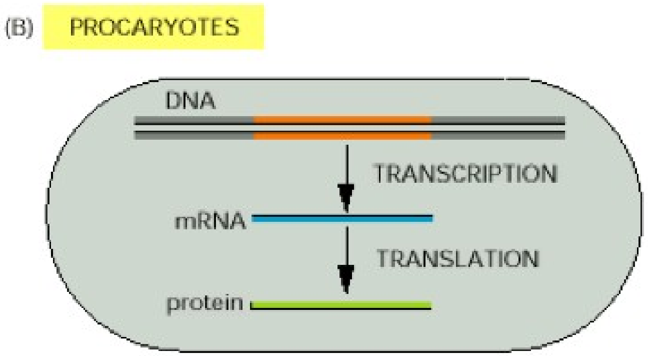
Design a HMM that models the syntax of genes

Gene structure

Depends on the organism (eucaryote or procaryote)

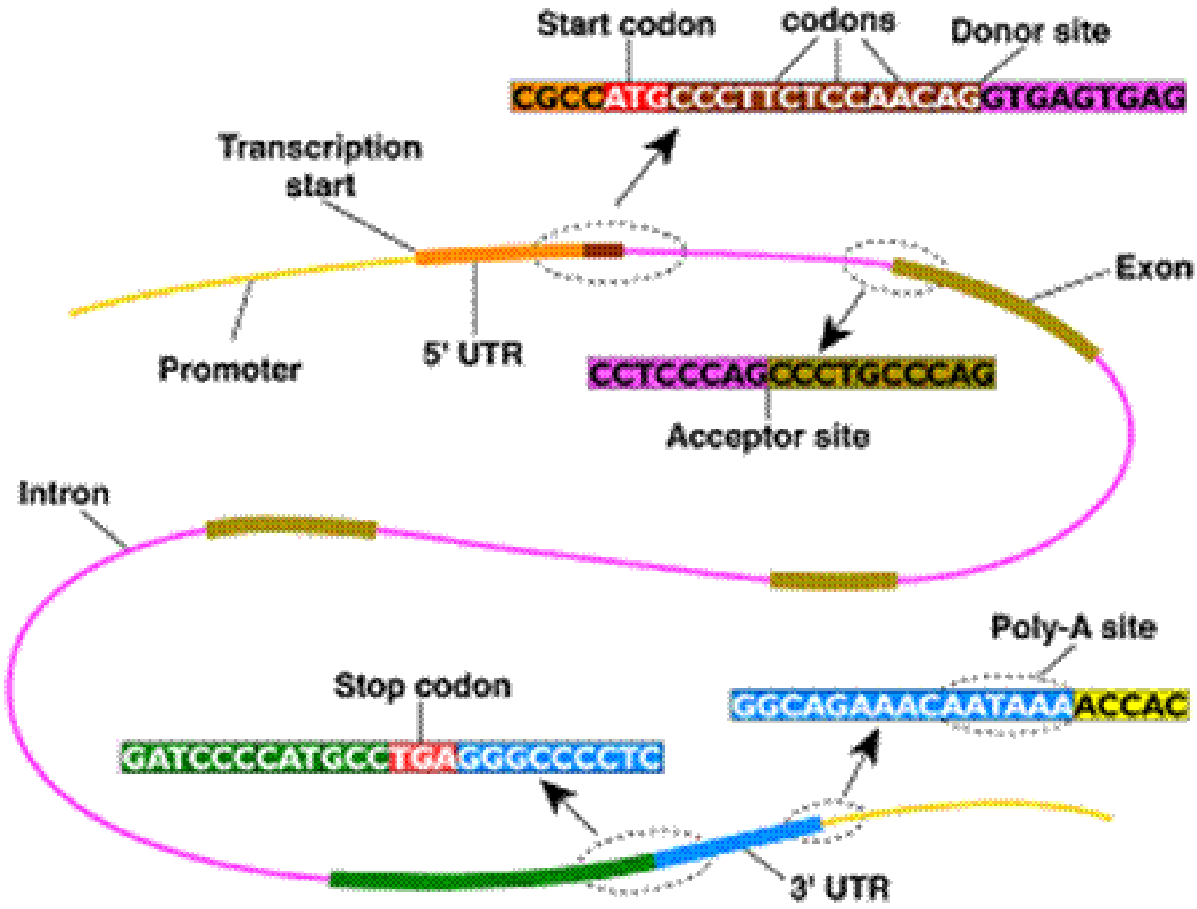


Large genomes. Intron/exon structure and low coding density



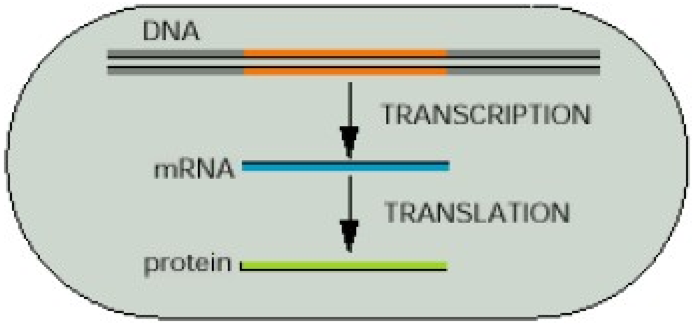
Smaller genomes and high coding density.

Gene structure in eukaryotes



Eukaryotic gene structure in more details

Gene structure in procaryotes

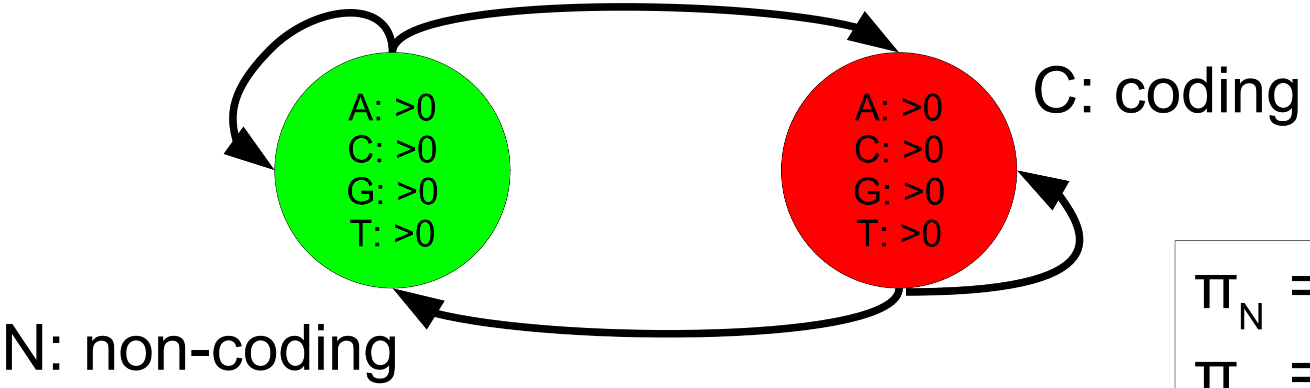


Biological facts

- The gene is a substring of the DNA sequence of A,C,G,T's

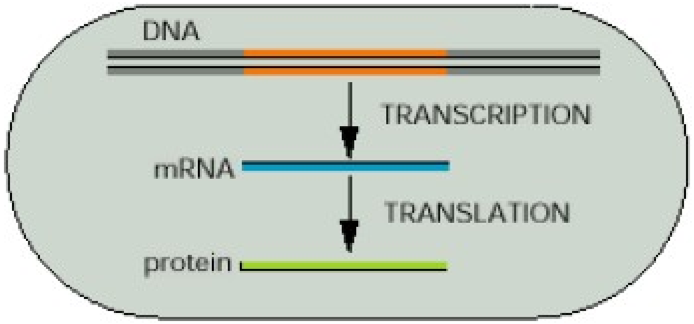
Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: `acgatgcgctaatatgtccgatgacgtgagcataagcgacatgcag`



$\pi_N = 1$
$\pi_C = 0$

Gene structure in procaryotes

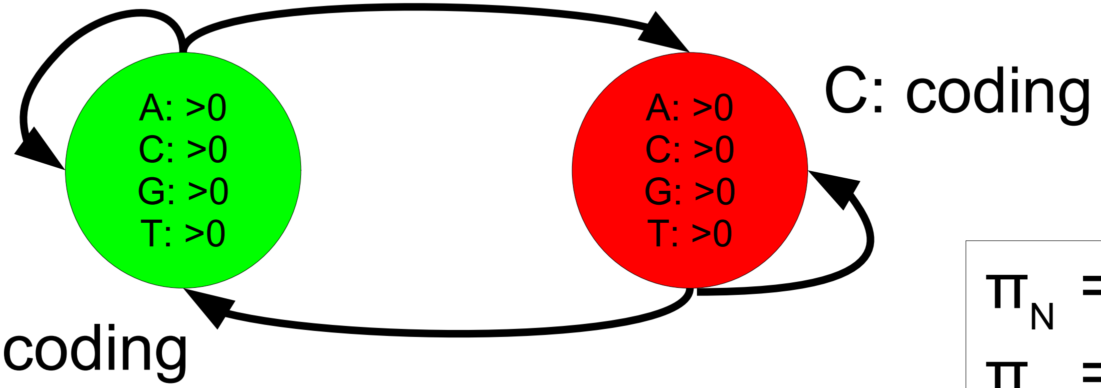


Biological facts

- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

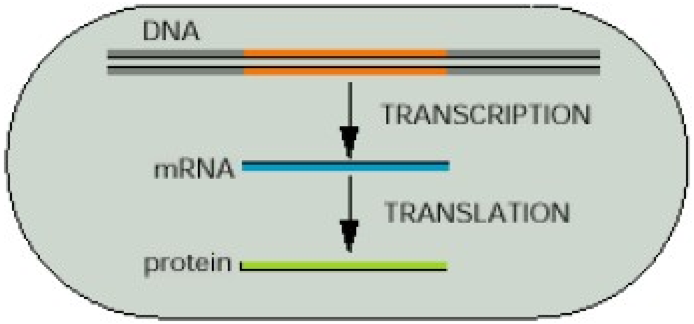
X: acgatgcgctaatatgtccgatgacgtgagcataagcgacatgcag



$\pi_N = 1$

$\pi_C = 0$

Gene structure in procaryotes



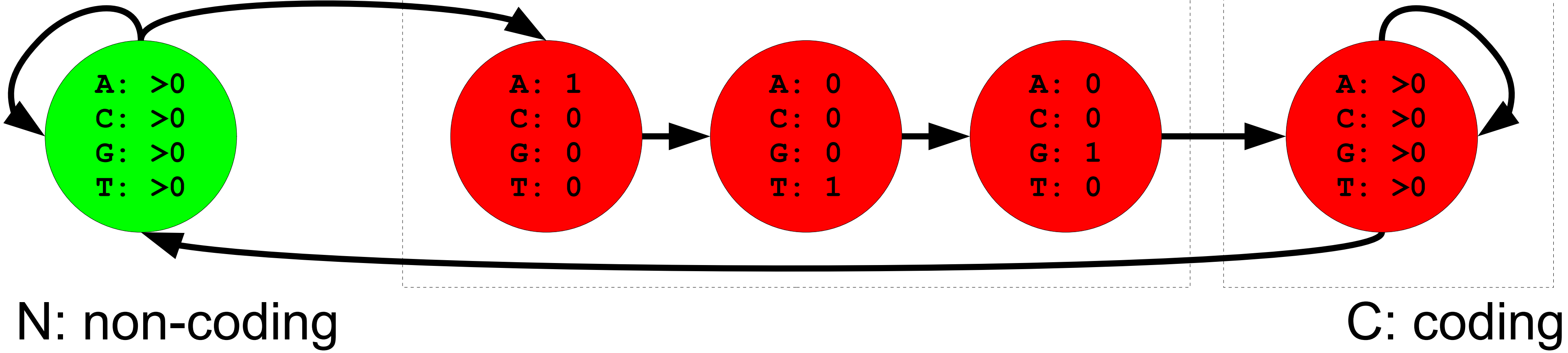
Biological facts

- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**

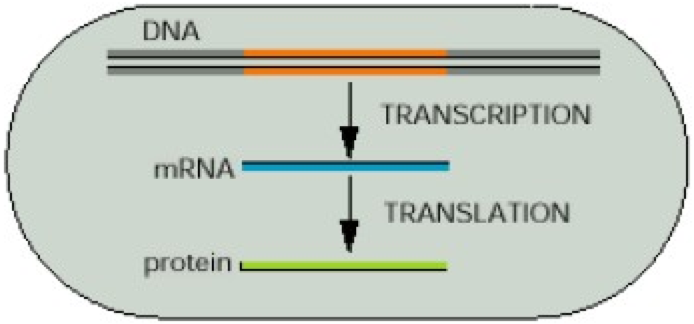
Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: `acgatgcgctaatatgtccgatgacgtgagcataagcgacat`

$\pi_N = 1$
$\pi_C = 0$



Gene structure in procaryotes



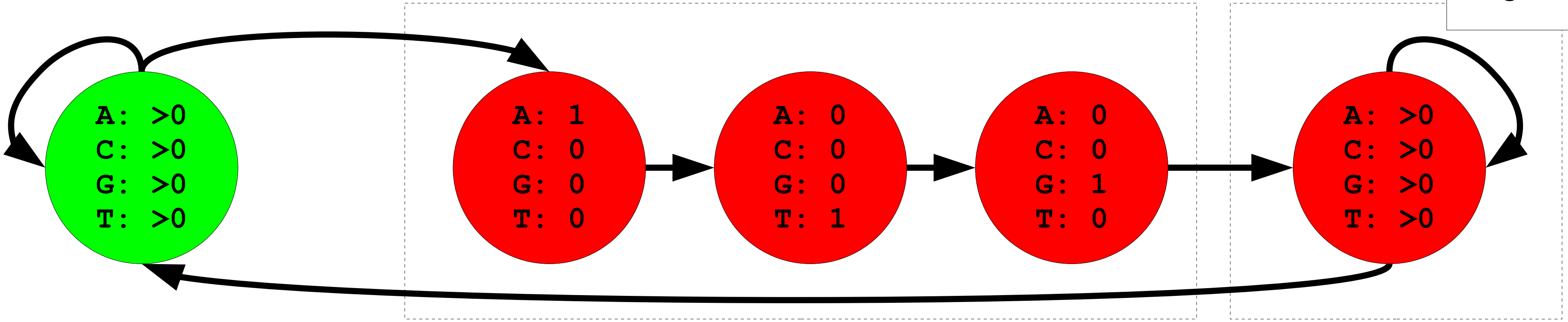
Biological facts

- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**
- The gene ends with a stop-codon **taa**, **tag** or **tga**

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacat

$\pi_N = 1$
$\pi_C = 0$

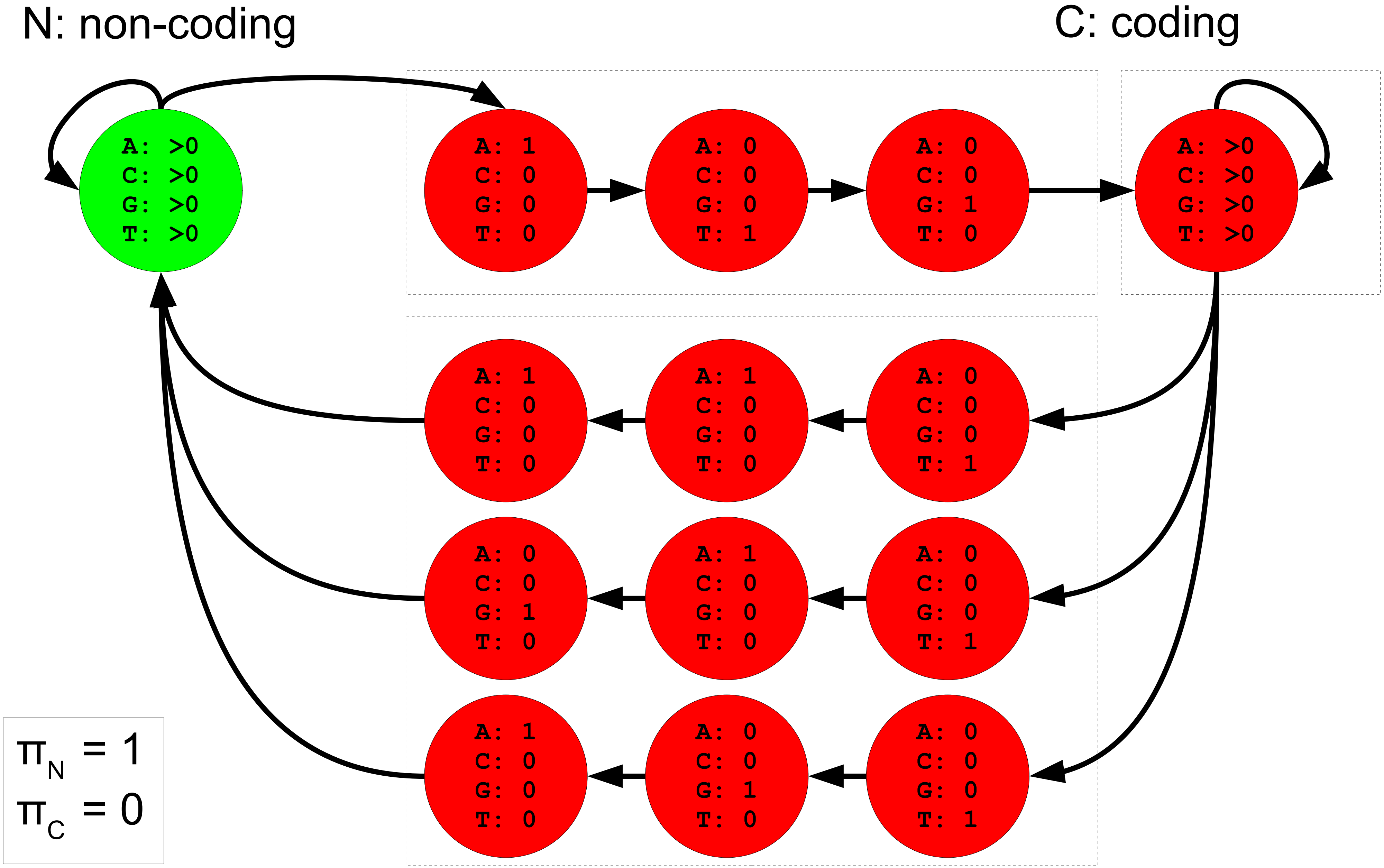


N: non-coding

C: coding

Gene struct

- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**
- The gene ends with a stop-codon **taa**, **tag** or **tga**

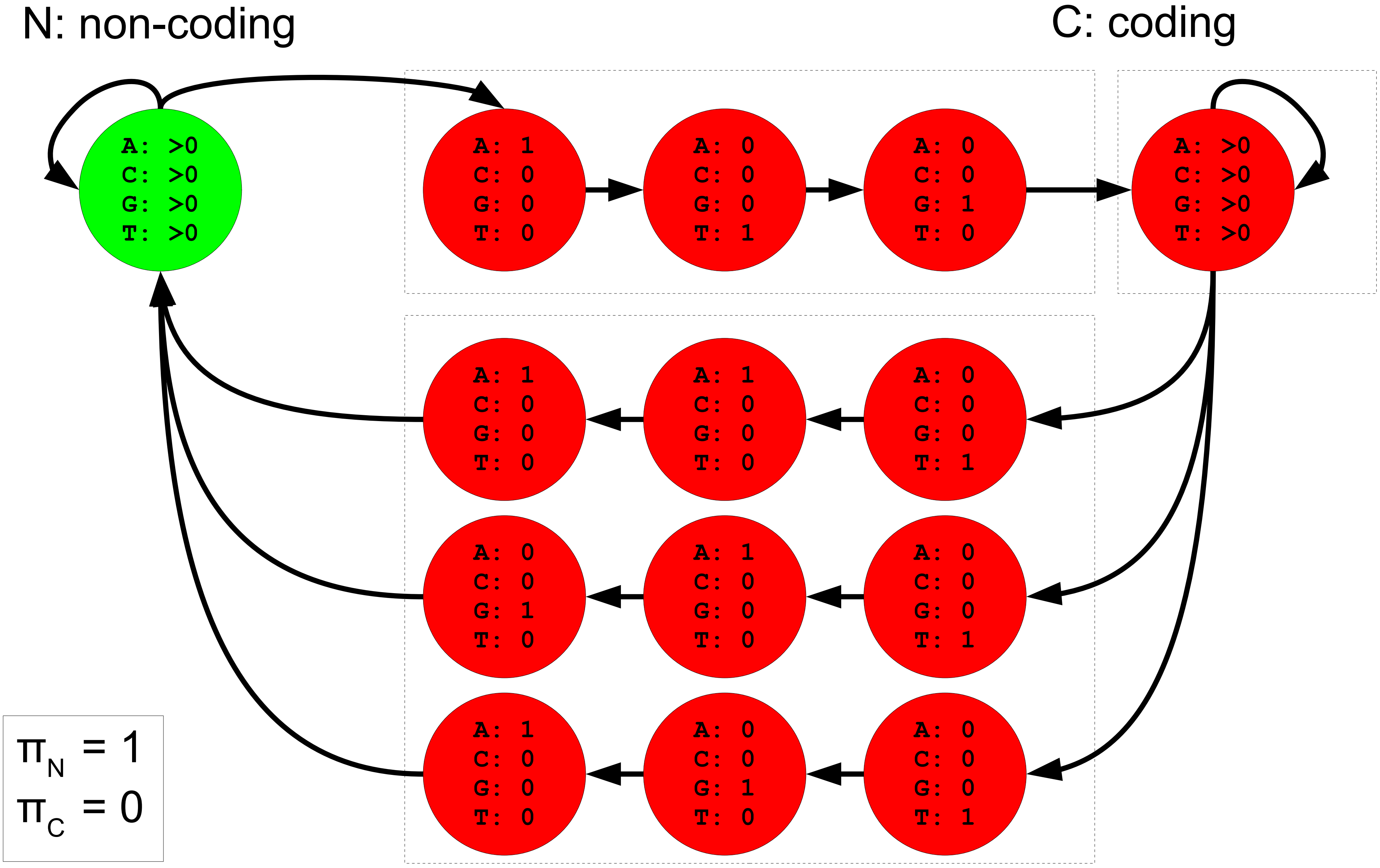


$\pi_N = 1$

$\pi_C = 0$

Gene struct

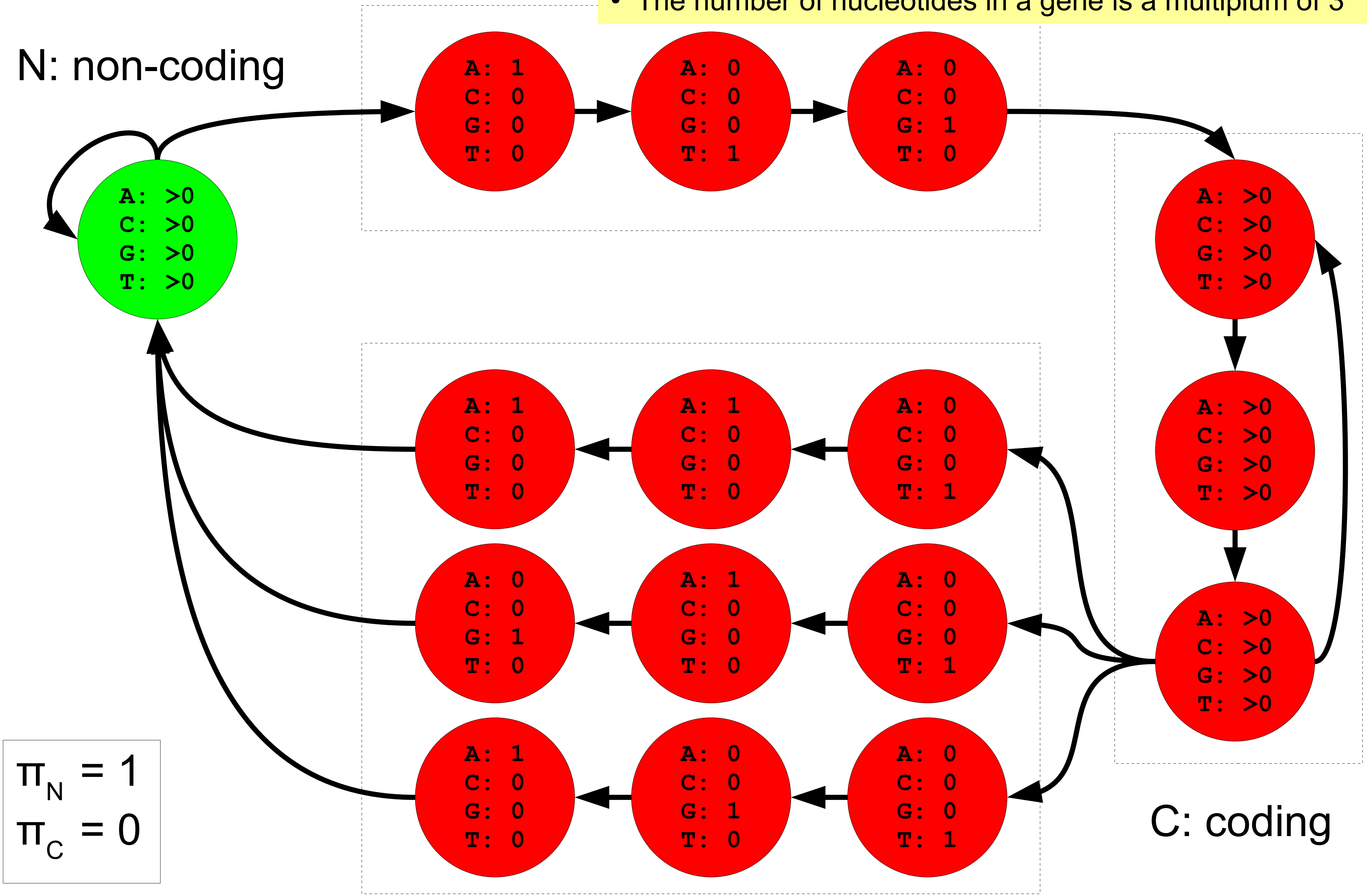
- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**
- The gene ends with a stop-codon **taa**, **tag** or **tga**
- The number of nucleotides in a gene is a multiplum of 3



$\pi_N = 1$
$\pi_C = 0$

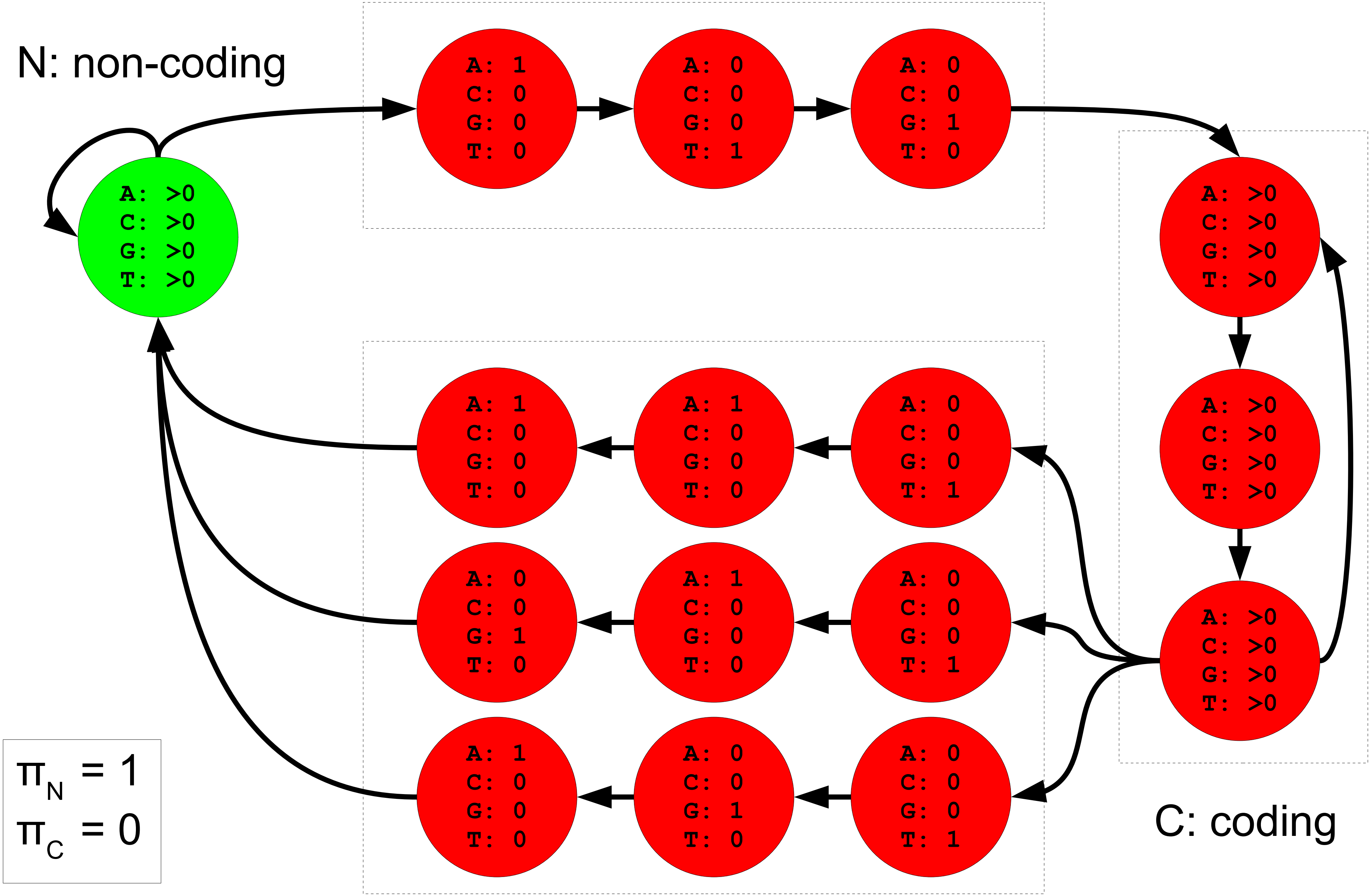
Gene struct

- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**
- The gene ends with a stop-codon **taa**, **tag** or **tga**
- The number of nucleotides in a gene is a multiplum of 3



$\pi_N = 1$
$\pi_C = 0$

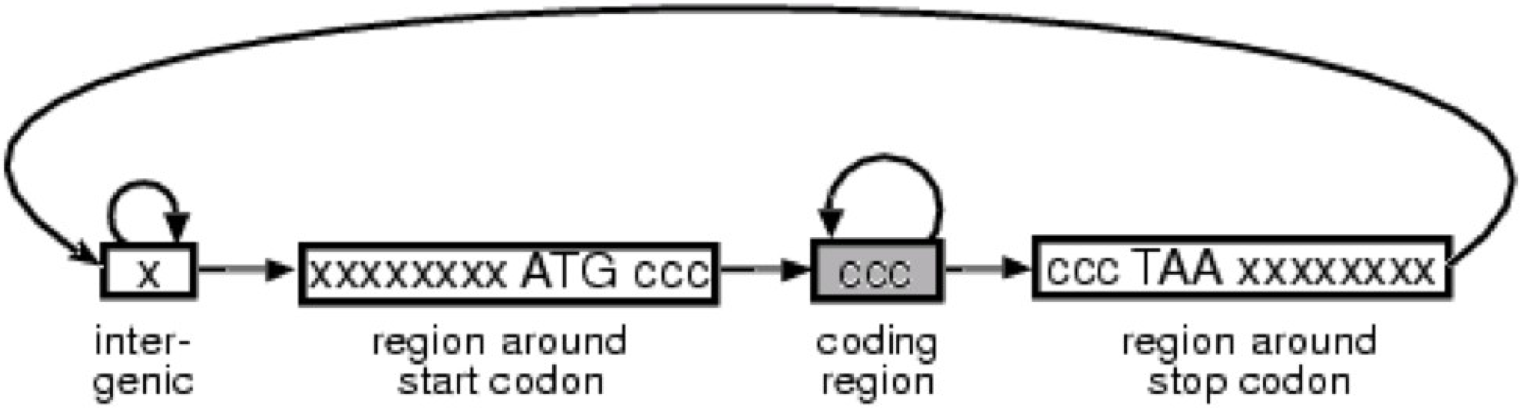
Gene structure in procaryotes



$\pi_N = 1$

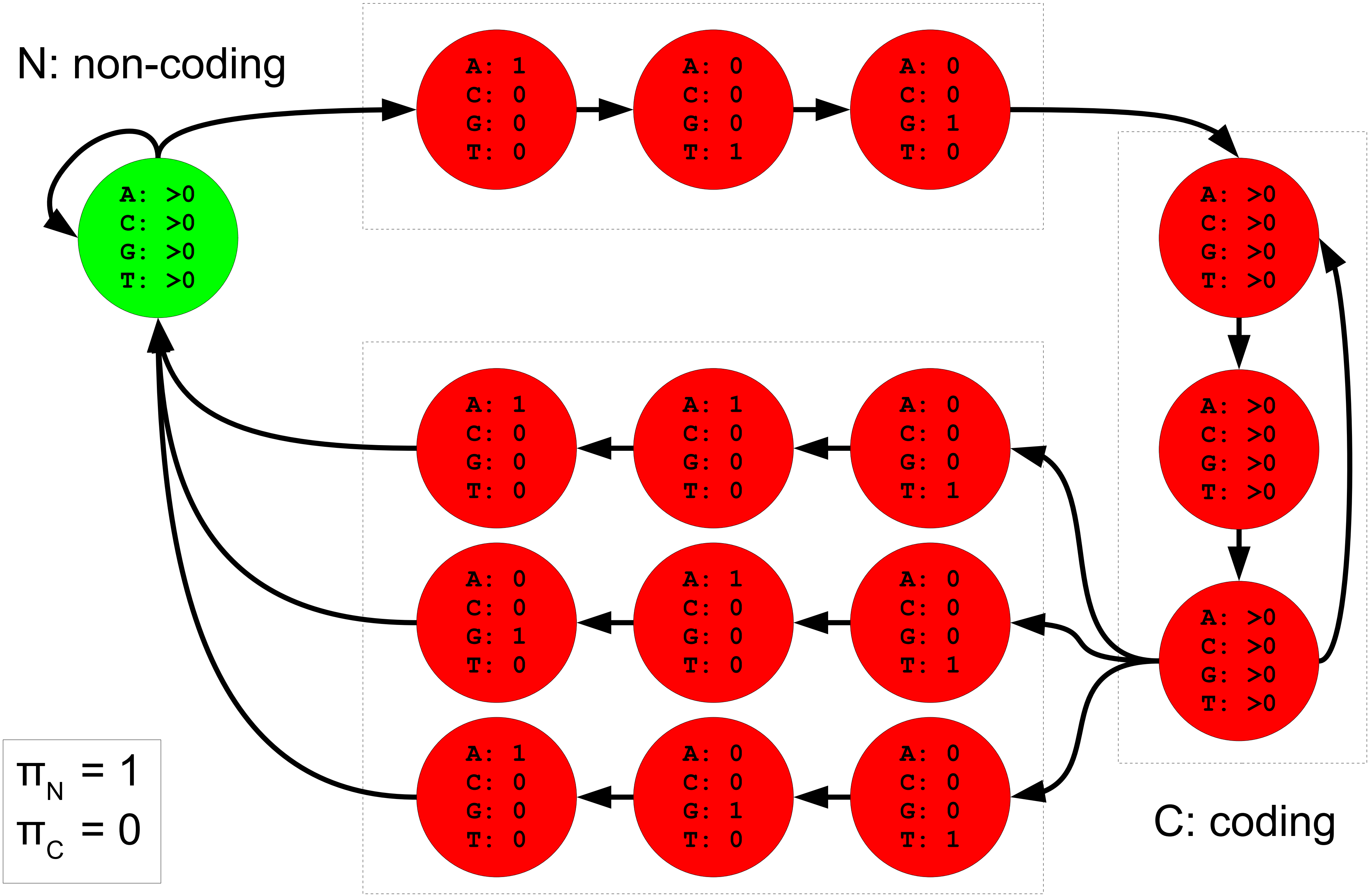
$\pi_C = 0$

Gene structure in procaryotes



From “An Introduction to HMMs for Biological Sequences”, A. Krogh, 1998

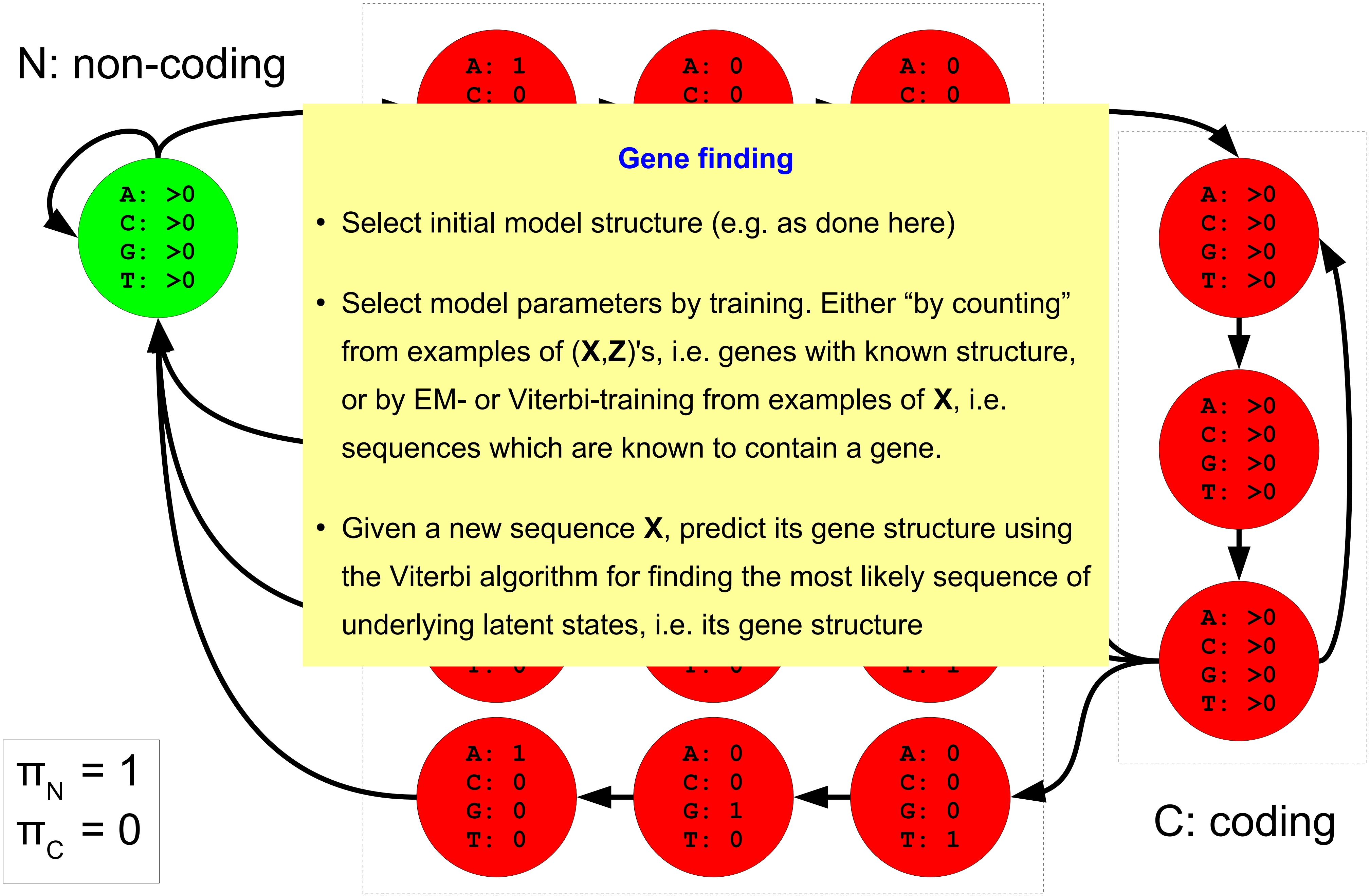
Example – Gene finding



$\pi_N = 1$

$\pi_C = 0$

Example – Gene finding



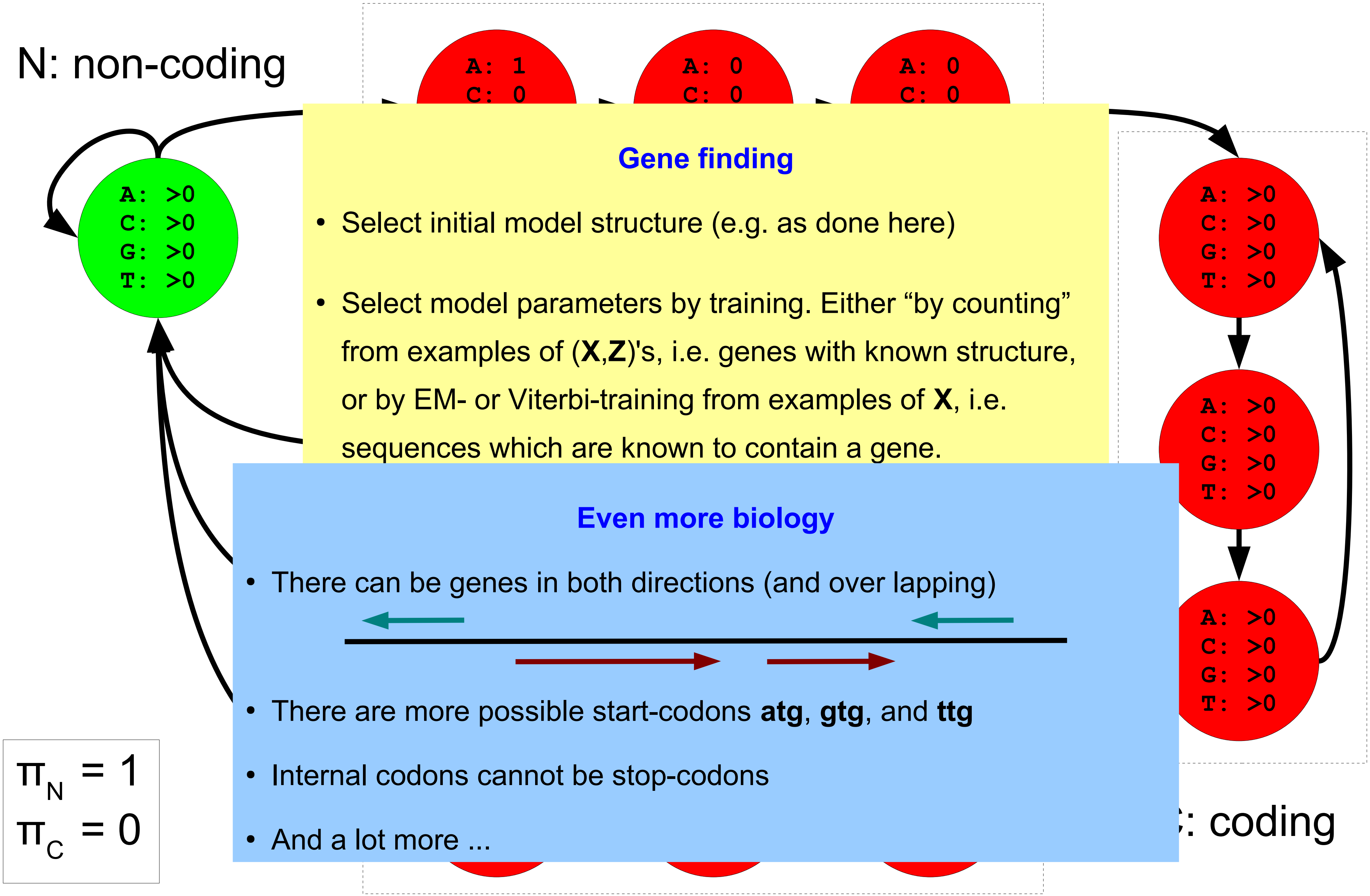
$\pi_N = 1$

$\pi_C = 0$

Gene finding

- Select initial model structure (e.g. as done here)
- Select model parameters by training. Either “by counting” from examples of (**X**,**Z**)'s, i.e. genes with known structure, or by EM- or Viterbi-training from examples of **X**, i.e. sequences which are known to contain a gene.
- Given a new sequence **X**, predict its gene structure using the Viterbi algorithm for finding the most likely sequence of underlying latent states, i.e. its gene structure

Example – Gene finding



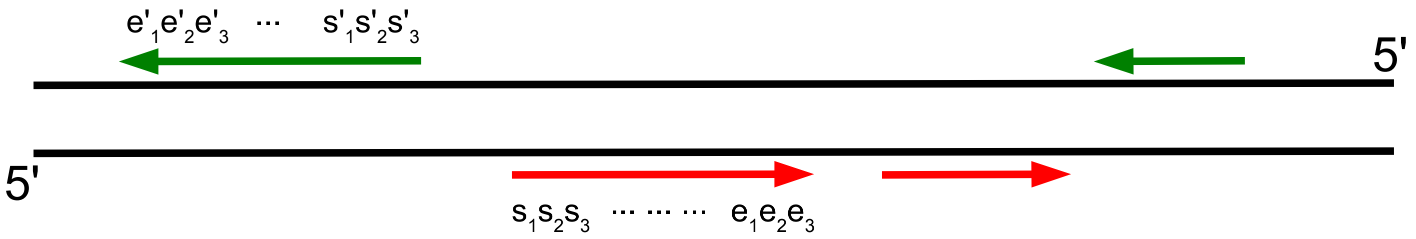
Gene finding

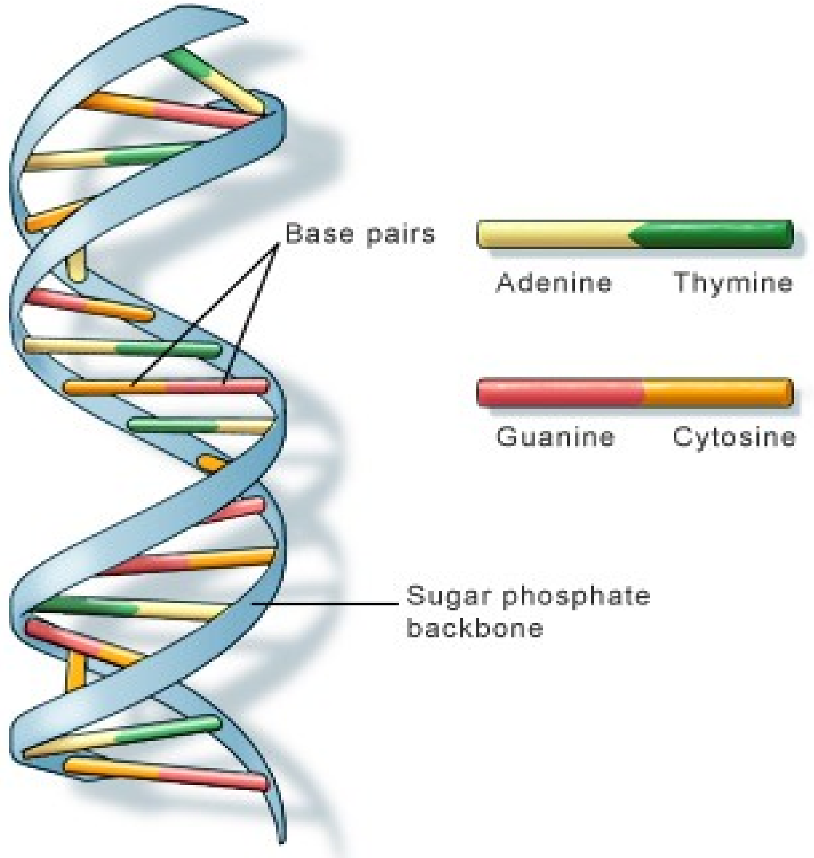
- Select initial model structure (e.g. as done here)
- Select model parameters by training. Either “by counting” from examples of (**X**,**Z**)'s, i.e. genes with known structure, or by EM- or Viterbi-training from examples of **X**, i.e. sequences which are known to contain a gene.

Even more biology

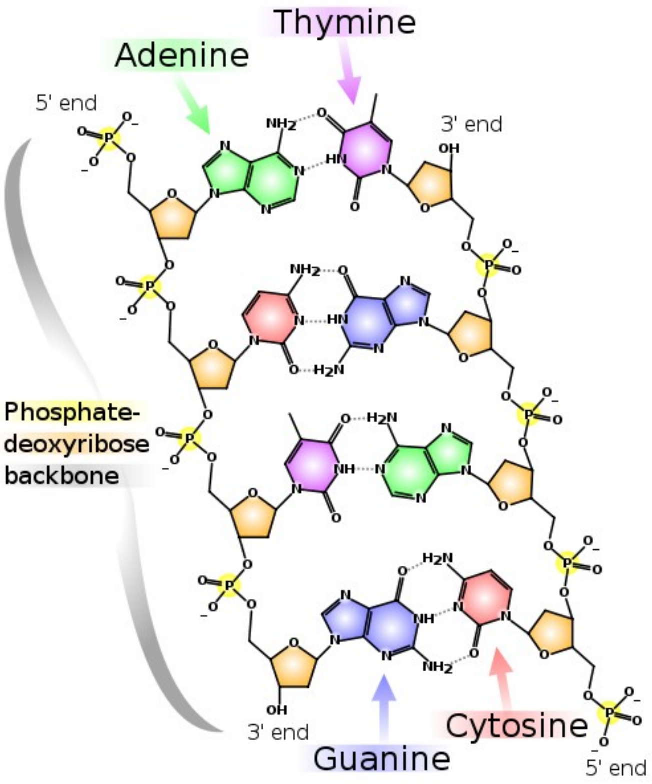
- There can be genes in both directions (and over lapping)
- There are more possible start-codons **atg**, **gtg**, and **ttg**
- Internal codons cannot be stop-codons
- And a lot more ...

DNA





U.S. National Library of Medicine



DNA

AGT
GAT
AAT

GTA

$e'_1e'_2e'_3$ ··· $s'_1s'_2s'_3$

5'

5'

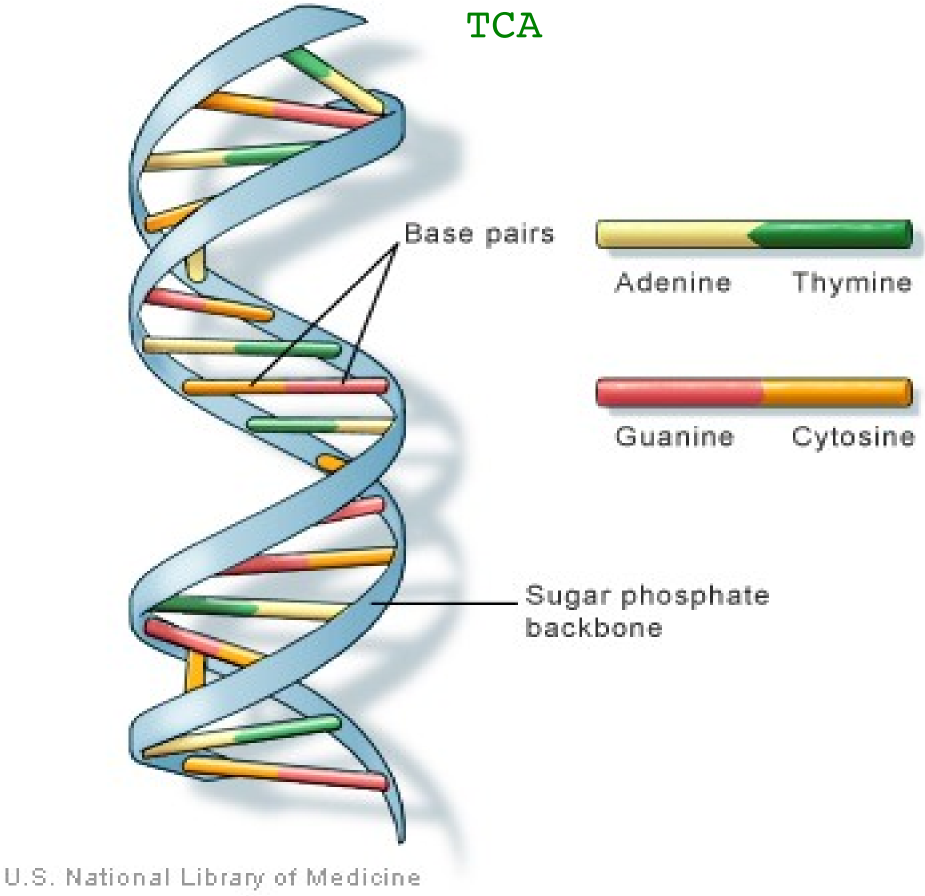
$s_1s_2s_3$ ··· ··· ··· $e_1e_2e_3$

TTA
CTA
TCA

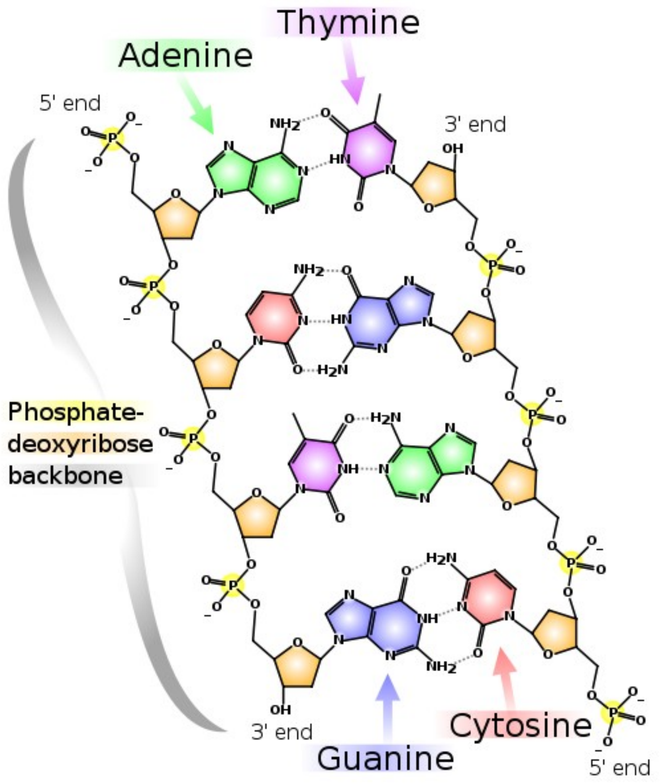
CAT

ATG

TAA
TAG
TGA

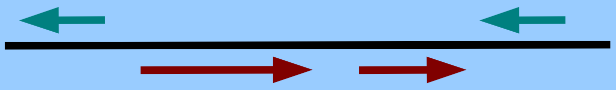


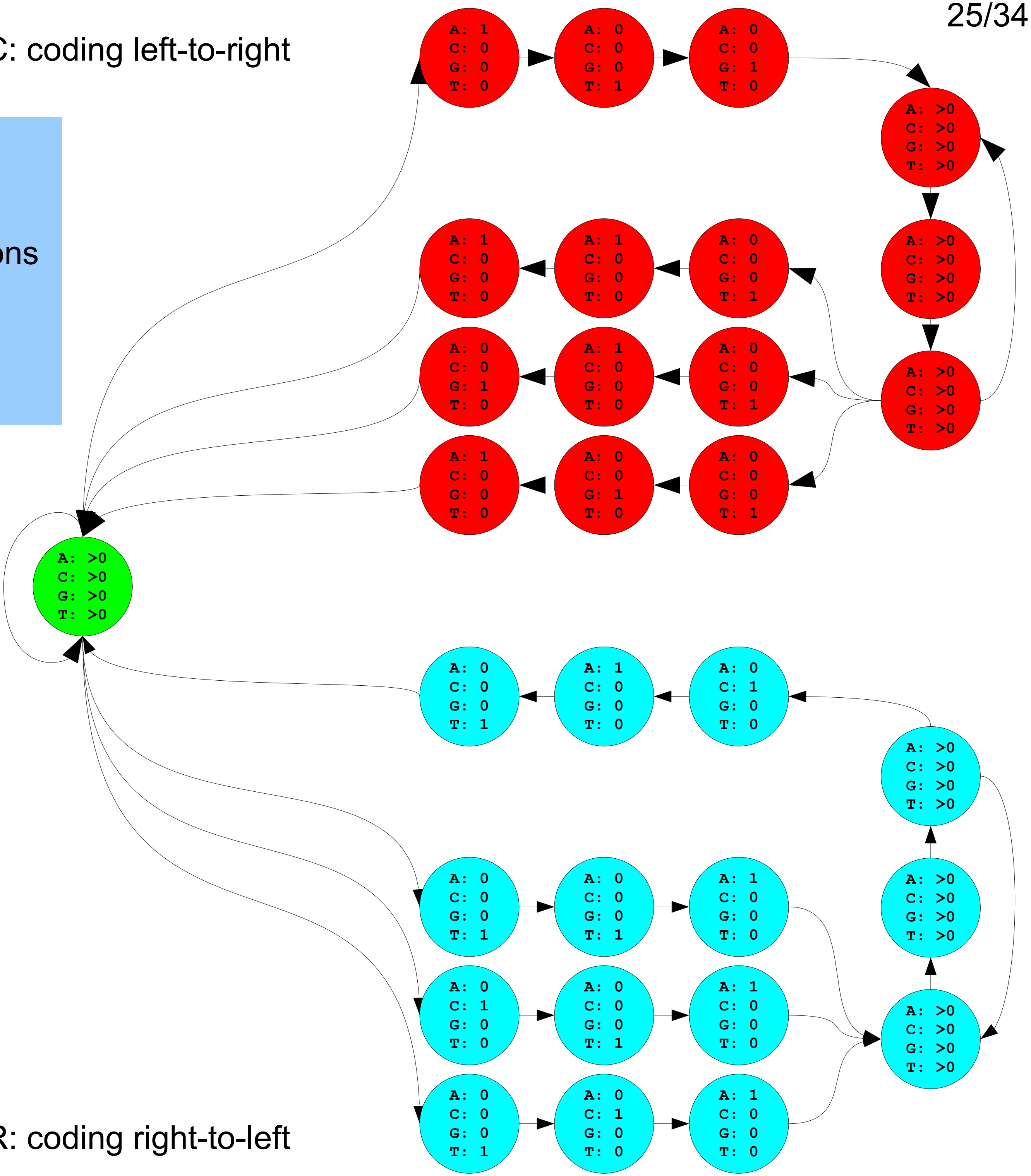
U.S. National Library of Medicine



Even more biology

There can be genes in both directions





C: coding left-to-right

N: Non-coding

R: coding right-to-left

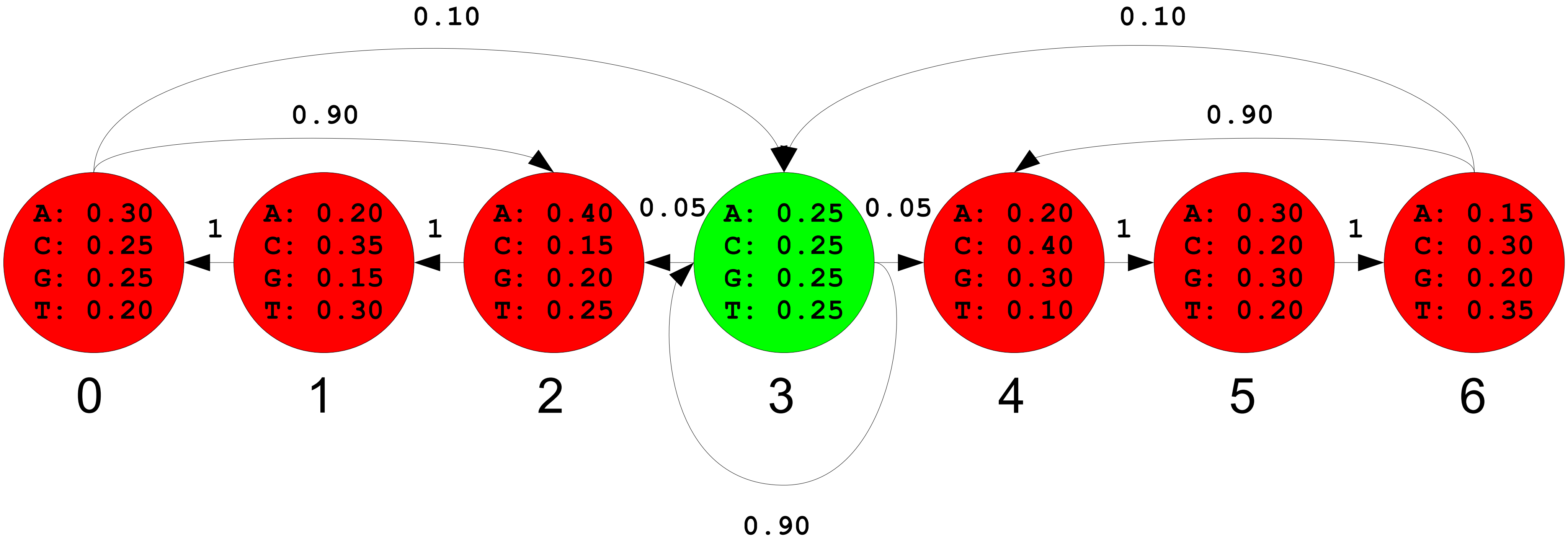
$\pi_N = 1$

$\pi_C = 0$

Example – 7-state HMM

Observable: {A, C, G, T}, States: {0,1, 2, 3, 4, 5, 6}

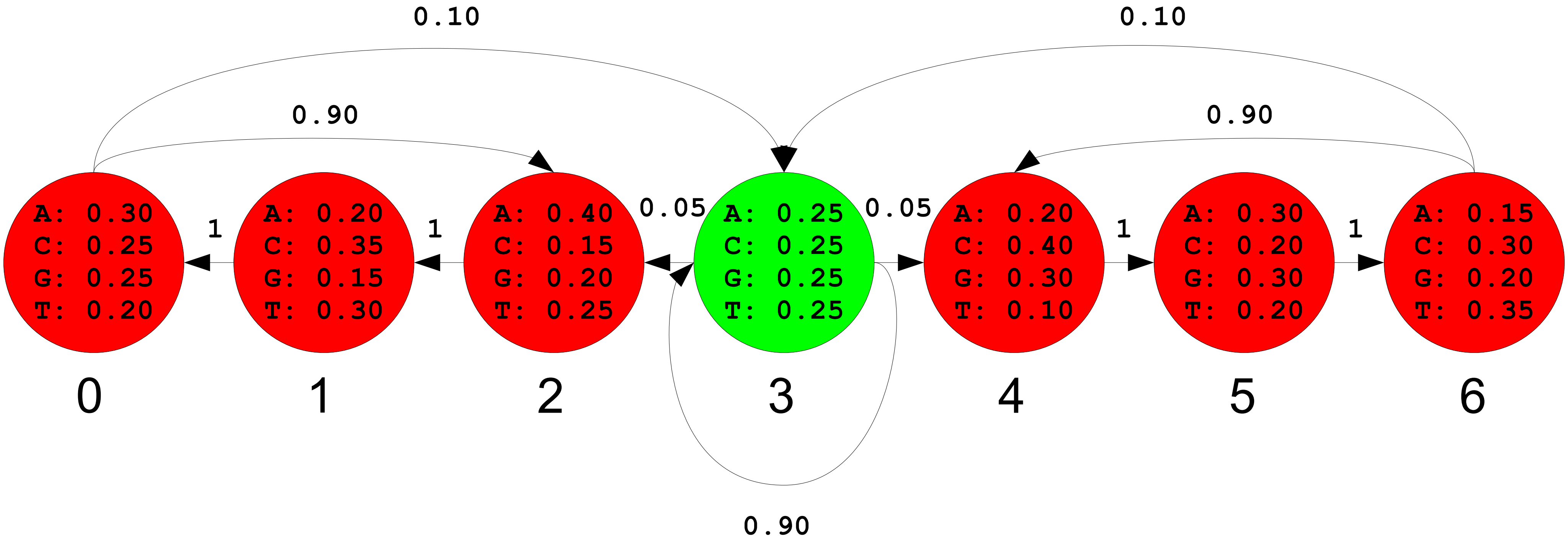
$$A \quad \begin{matrix} 0.00 & 0.00 & 0.90 & 0.10 & 0.00 & 0.00 & 0.00 \\ 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.05 & 0.90 & 0.05 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.10 & 0.90 & 0.00 & 0.00 \end{matrix} \qquad \pi \quad \begin{matrix} 0.00 \\ 0.00 \\ 0.00 \\ 1.00 \\ 0.00 \\ 0.00 \\ 0.00 \end{matrix} \qquad \varphi \quad \begin{matrix} 0.30 & 0.25 & 0.25 & 0.20 \\ 0.20 & 0.35 & 0.15 & 0.30 \\ 0.40 & 0.15 & 0.20 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.20 & 0.40 & 0.30 & 0.10 \\ 0.30 & 0.20 & 0.30 & 0.20 \\ 0.15 & 0.30 & 0.20 & 0.35 \end{matrix}$$



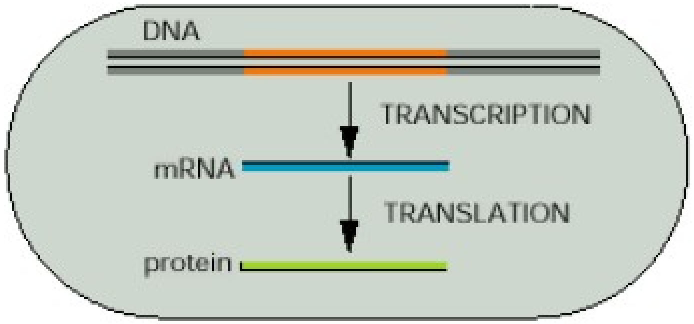
This model is also applicable for gene finding.

It does not model start- and stop-codons explicitly, but models that genes in both directions are a sequence of triplets.

$$A \quad \begin{matrix} 0.00 & 0.00 & 0.90 & 0.10 & 0.00 & 0.00 & 0.00 \\ 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.05 & 0.90 & 0.05 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \\ 0.00 & 0.00 & 0.00 & 0.10 & 0.90 & 0.00 & 0.00 \end{matrix} \qquad \pi \quad \begin{matrix} 0.00 \\ 0.00 \\ 0.00 \\ 1.00 \\ 0.00 \\ 0.00 \\ 0.00 \end{matrix} \qquad \varphi \quad \begin{matrix} 0.30 & 0.25 & 0.25 & 0.20 \\ 0.20 & 0.35 & 0.15 & 0.30 \\ 0.40 & 0.15 & 0.20 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.20 & 0.40 & 0.30 & 0.10 \\ 0.30 & 0.20 & 0.30 & 0.20 \\ 0.15 & 0.30 & 0.20 & 0.35 \end{matrix}$$



Problem: From annotation to Z



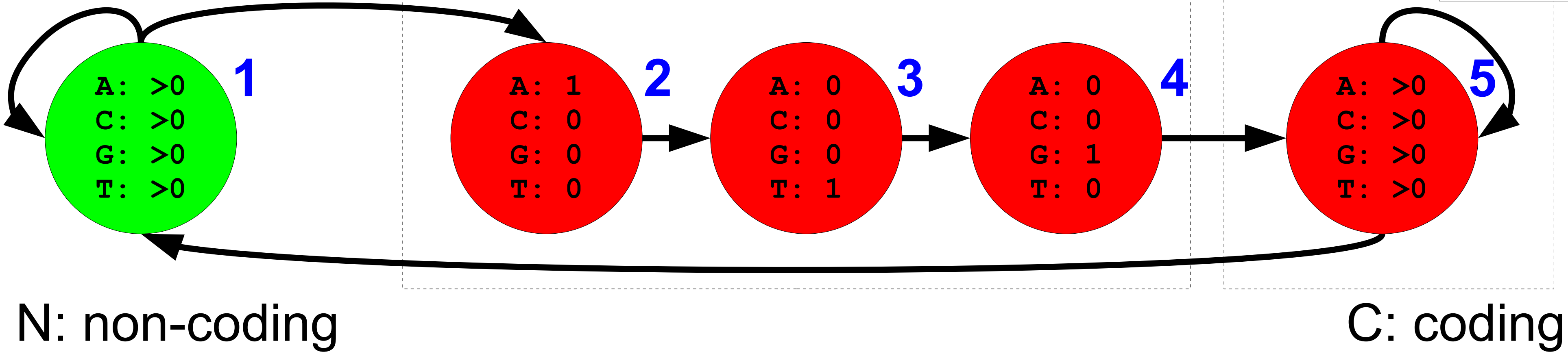
Biological facts

- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacat

$\pi_N = 1$
$\pi_C = 0$



Problem: From annotation to Z

Problem: The string **Z=**NNNCCC.... is not a prober sequence of states in the illustrated HMM, but it can easily be converted into one (because there in this case is a 1-1 matching between a sequence of Ns and Cs and a sequence of states).

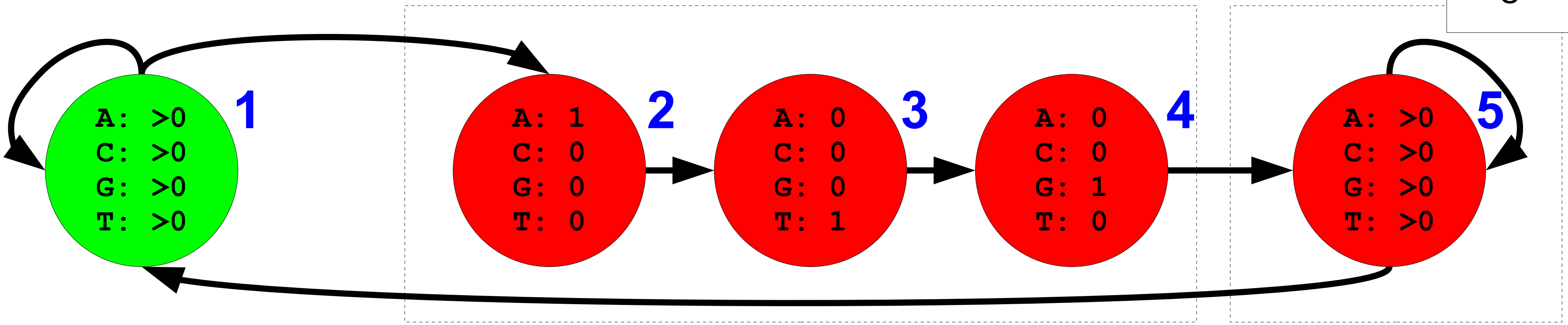
ence of A,C,G,T's

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacat

$\pi_N = 1$

$\pi_C = 0$



N: non-coding

C: coding

Problem: From annotation to Z

Problem: The string **Z=**NNNCCC.... is not a prober sequence of states in the illustrated HMM, but it can easily be converted into one (because there in this case is a 1-1 matching between a sequence of Ns and Cs and a sequence of states).

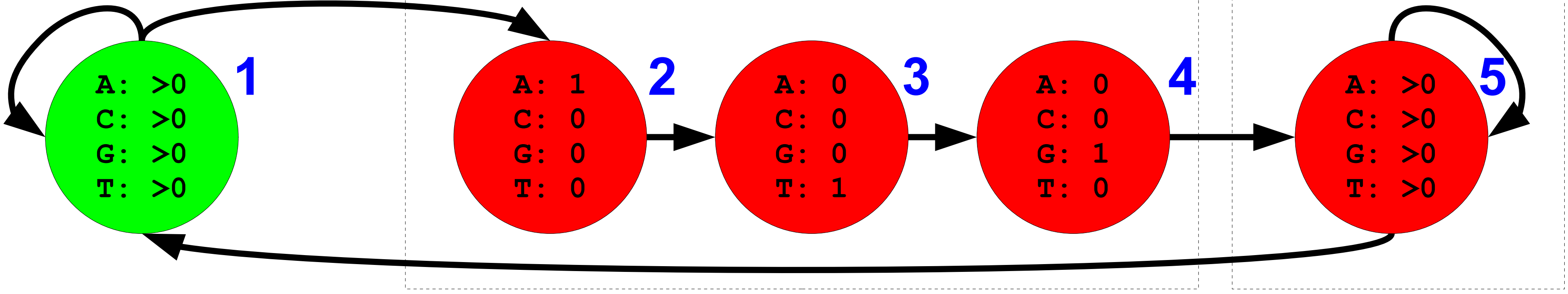
ence of A,C,G,T's

```
    1112345555551111111123455555555555511111111111
Z:  NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X:  acgatgcgctaatatgtccgatgacgtgagcataagcgacat
```

$\pi_N = 1$

$\pi_C = 0$

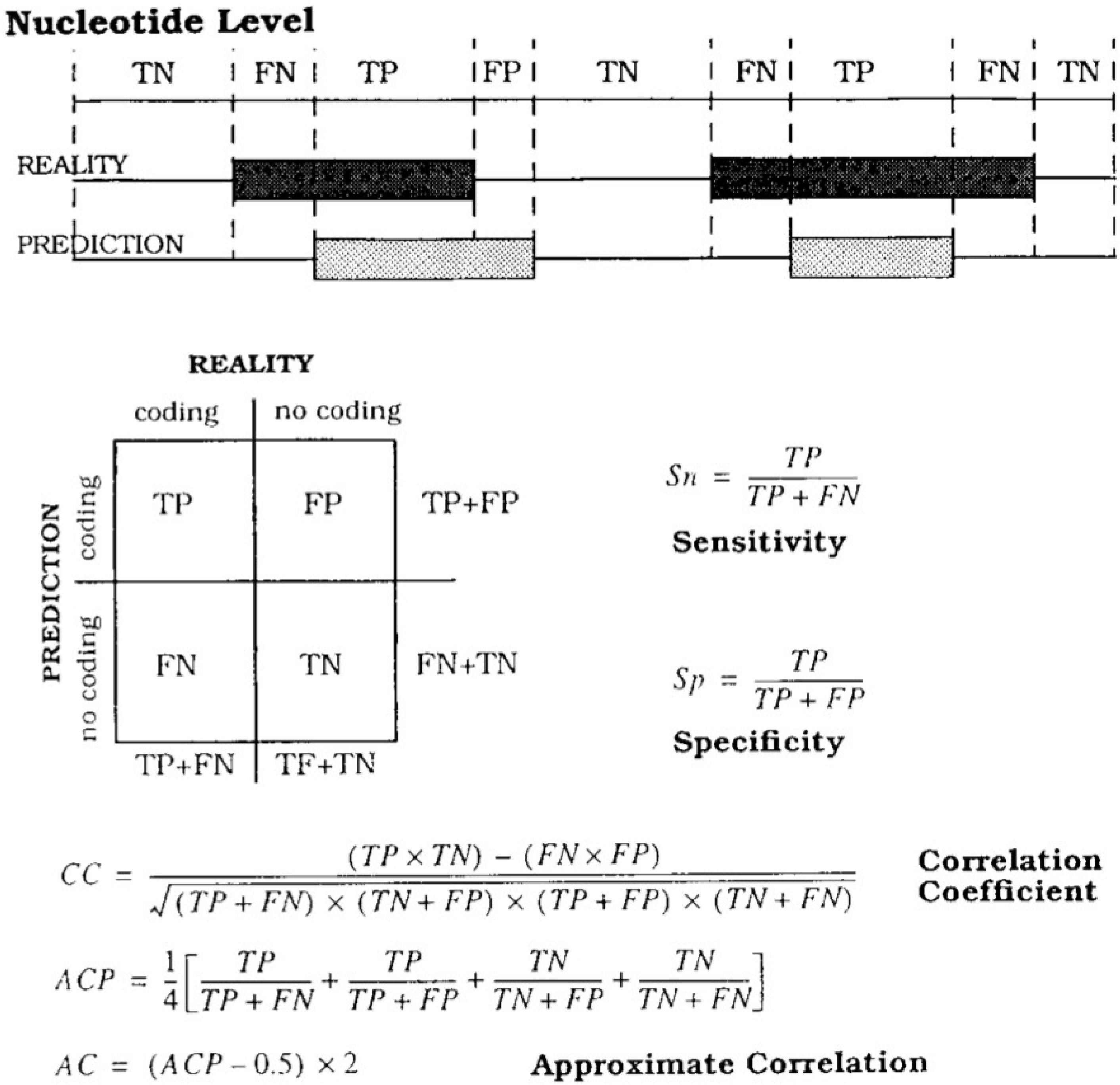


N: non-coding

C: coding

Evaluating performance

Evaluation of Gene Structure Prediction Programs (Burset and Guigo, 1996)



Nucleotide Level

		REALITY		
		coding	no coding	
PREDICTION	coding	TP	FP	TP+FP
	no coding	FN	TN	FN+TN
		TP+FN	TF+TN	

$$Sn = \frac{TP}{TP + FN}$$

Sensitivity

$$Sp = \frac{TP}{TP + FP}$$

Specificity

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

Correlation Coefficient

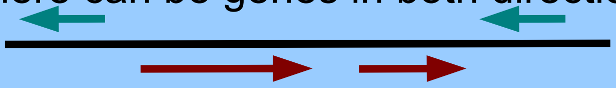
$$ACP = \frac{1}{4}\left[\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right]$$

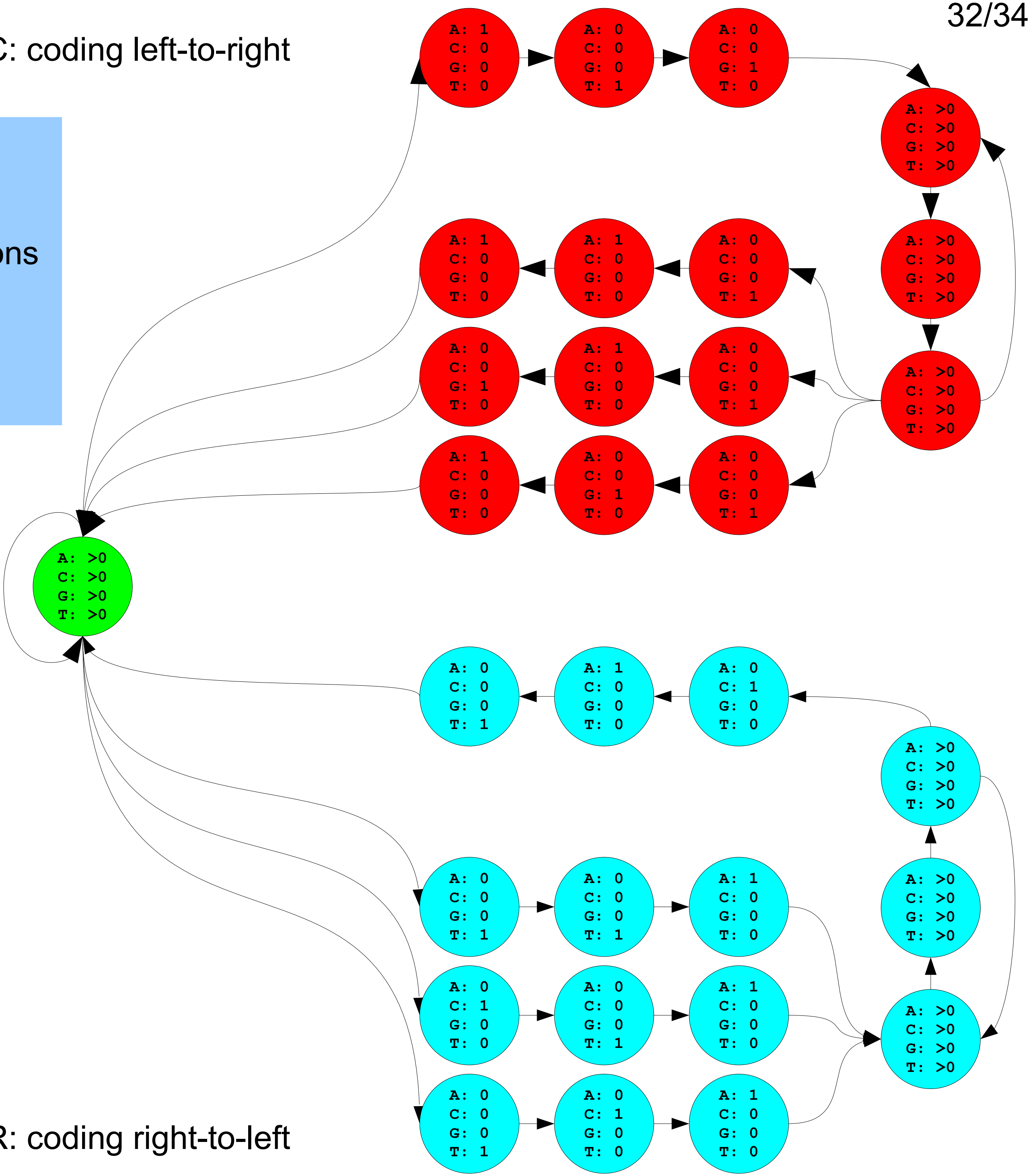
$$AC = (ACP - 0.5) \times 2$$

Approximate Correlation

Even more biology

There can be genes in both directions





C: coding left-to-right

N: Non-coding

R: coding right-to-left

$\pi_N = 1$

$\pi_C = 0$

Analysis of some genomes

Even more biology

- There can be genes in both directions
- There are more possible start-codons **atg**, **gtg**, and **ttg**
- Internal codons cannot be stop-codons
- And a lot more ...

```
Length of genome1: 1852441 (1852441)
Length of genome2: 2211485 (2211485)
Length of genome3: 2499279 (2499279)
Length of genome4: 1796846 (1796846)
Length of genome5: 2685015 (2685015)
Length of genome6: 2127839 (2127839)
Length of genome7: 2742531 (2742531)
Length of genome8: 2046115 (2046115)
Length of genome9: 2388435 (2388435)
Length of genome10: 1570485 (1570485)
Length of genome11: 2096309 (2096309)
```

```
Start-codon in normal genes:
ATG [8423, 'NCCC']
ATC [3, 'NCCC']
ATA [1, 'RCCC']
GTG [713, 'NCCC']
ATT [3, 'NCCC']
CTG [2, 'NCCC']
GTT [1, 'NCCC']
CTC [1, 'NCCC']
TTA [1, 'NCCC']
TTG [1020, 'NCCC']

Stop-codon in normal genes:
TAG [1949, 'CCCN']
TGA [1531, 'CCCN']
TAA [6686, 'CCCN']

Reversed stop-codon in reversed genes:
TTA (reverse-complement: TAA) [6596, 'NRRR']
CTA (reverse-complement: TAG) [2014, 'NRRR']
TCA (reverse-complement: TGA) [1148, 'NRRR']

Reversed start-codon in reversed genes:
TAT (reverse-complement: ATA) [2, 'RRRN']
ATG (reverse-complement: CAT) [1, 'RRRN']
GAT (reverse-complement: ATC) [1, 'RRRN']
CAT (reverse-complement: ATG) [8077, 'RRRN']
AAT (reverse-complement: ATT) [4, 'RRRN']
TAC (reverse-complement: GTA) [1, 'RRRN']
CAC (reverse-complement: GTG) [715, 'RRRN']
CAA (reverse-complement: TTG) [953, 'RRRN']
CAG (reverse-complement: CTG) [4, 'RRRN']
```

Summary

- We have considered the problem of selecting initial model parameters, and exemplified this problem by designing different HMMs for gene finding.

- Next time: See various extensions to the ‘standard’ HMM, which might be useful in some applications of HMMs, e.g. for gene finding.