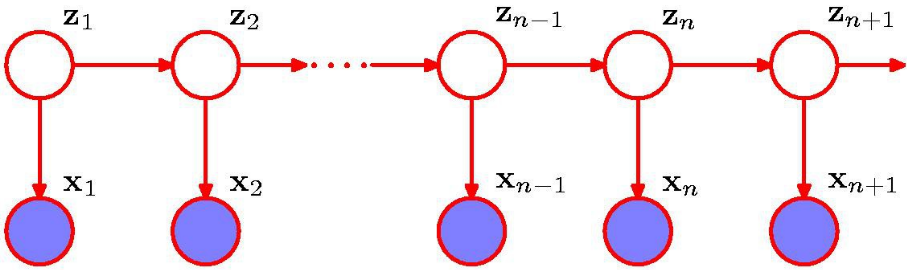
Hidden Markov Models

Selecting the initial model parameters

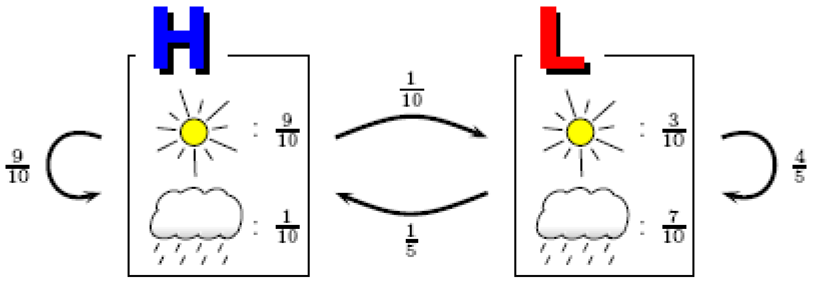
Using HMMs for gene finding



HMMs as a generative model

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

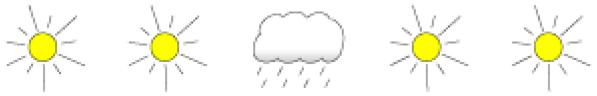
Model *M*:



A run follows a sequence of states:

H H L L H

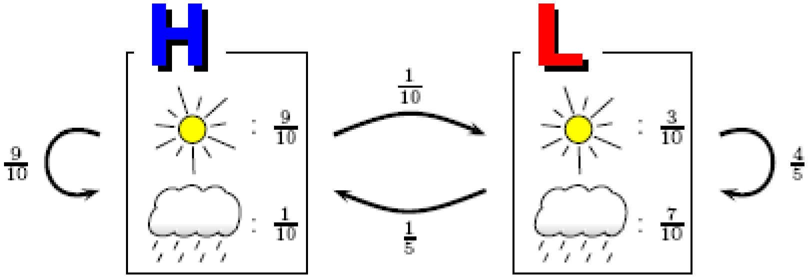
And emits a sequence of symbols:

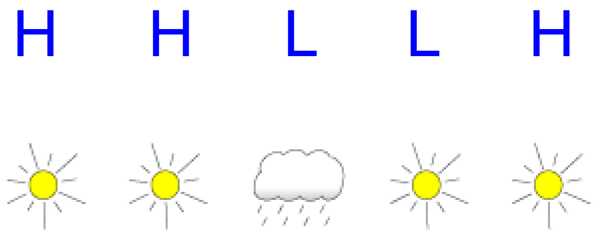


For a HMM that generates finite strings (e.g. a HMM with an end-state), the language L = {**X** | *p*(**X**) > 0} is regular ...

Selecting initial model parameters

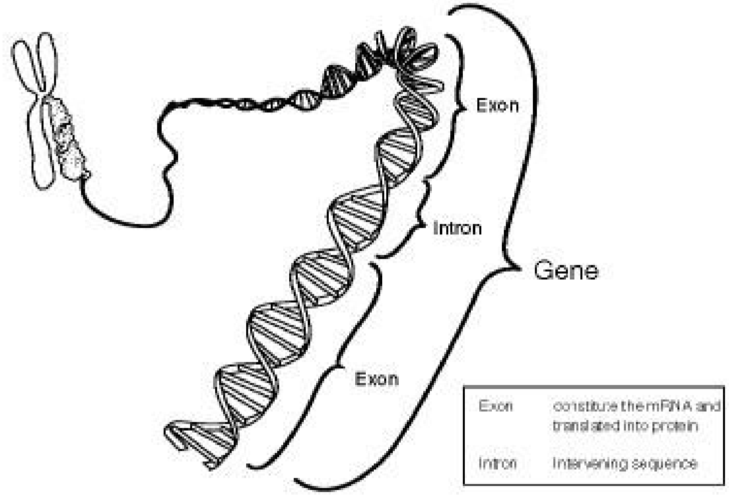
The initial selection of transition and emission probabilities, i.e. A, π, Φ, should model (how we see) the underlying structure of the observations, i.e. the syntax of possible sequences of observations, recall that the language L = {x | P(x | θ) > 0} is regular.

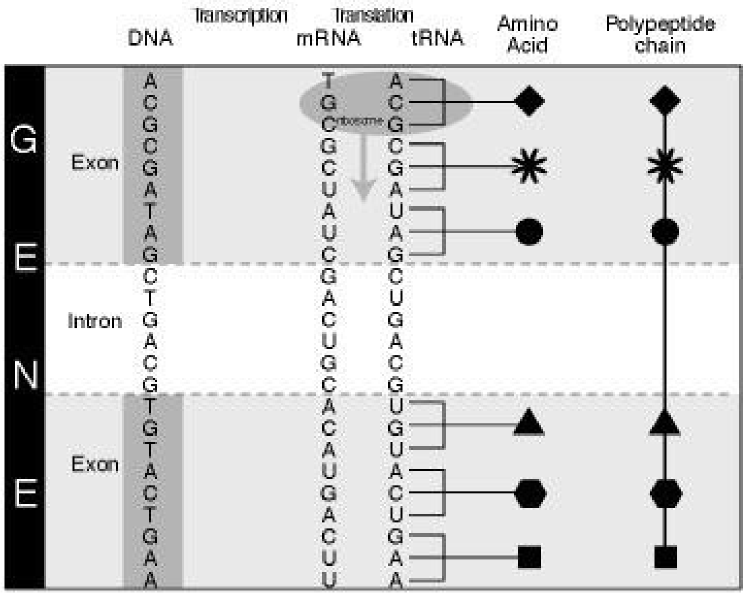




The initial selection of parameters is essential just to decide which parameters are 0 (or 1), i.e. to decide which transitions of emission should never (or always) be possible ...

Example – Gene finding


Exon
Intron
Gene
Exon
Exon constitute the mRNA and translated into protein
Intron Intervening sequence


DNA
Transcription
mRNA
Translation
tRNA
Amino Acid
Polypeptide chain
G
E
N
E
Exon
Intron
Exon
A C G C G A T A G C T G A C G T G T A C T G A A
T G C G C U A U C G A C U G C A C A U G A C U U
A C G C G A U A G C U G A C G U G U A C U G A A
ribosome

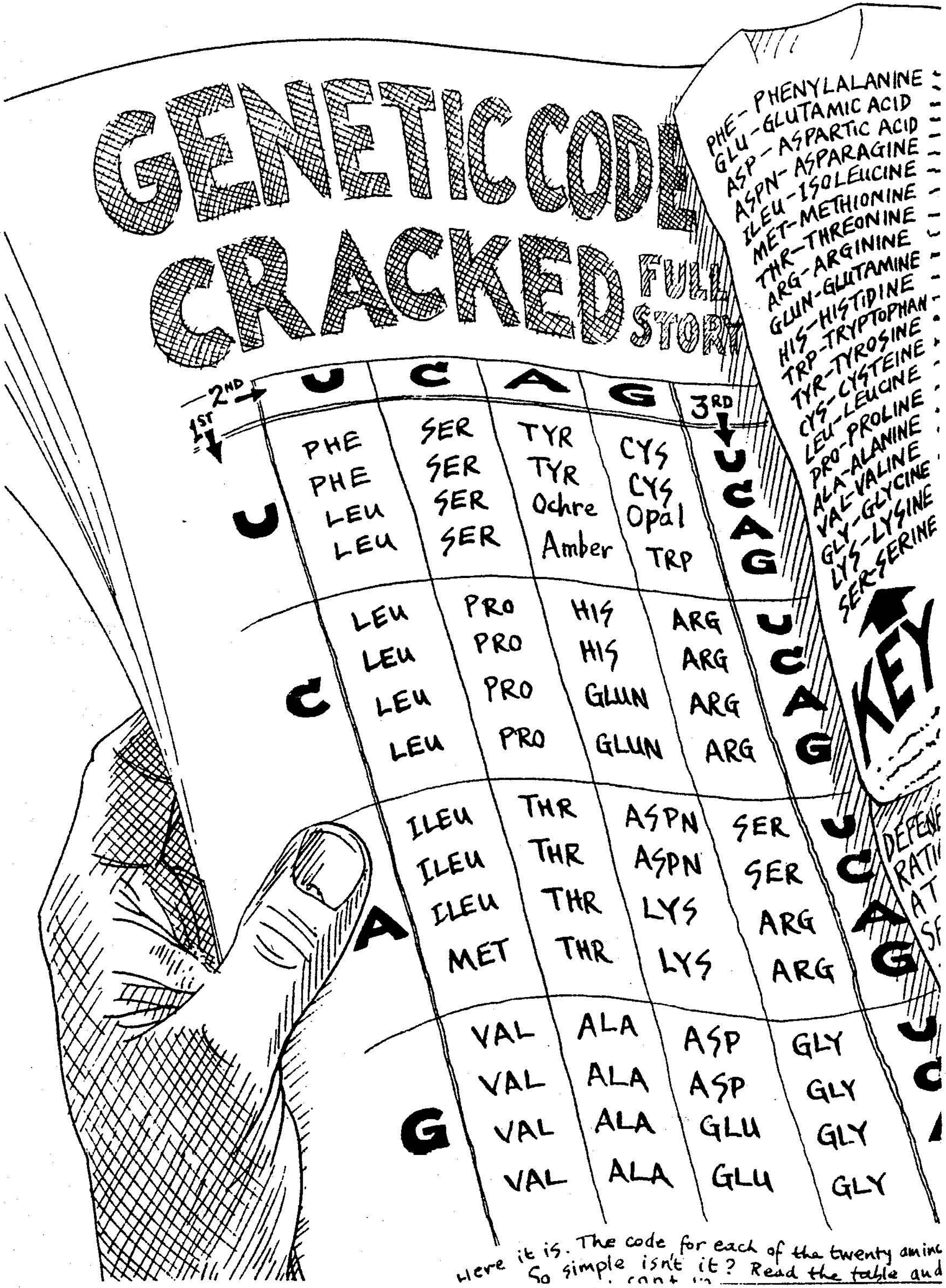
Each protein is encoded in a stretch of DNA. A gene ...

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 ...



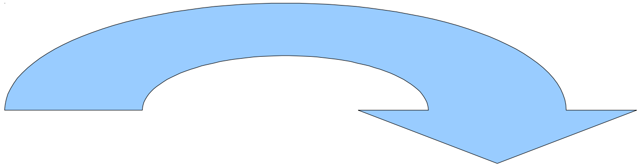
1st ↓ \ 2nd →	U	C	A	G	3rd ↓
U	PHE	SER	TYR	CYS	U
	PHE	SER	TYR	CYS	C
	LEU	SER	Ochre	Opal	A
	LEU	SER	Amber	TRP	G
C	LEU	PRO	HIS	ARG	U
	LEU	PRO	HIS	ARG	C
	LEU	PRO	GLUN	ARG	A
	LEU	PRO	GLUN	ARG	G
A	ILEU	THR	ASPN	SER	U
	ILEU	THR	ASPN	SER	C
	ILEU	THR	LYS	ARG	A
	MET	THR	LYS	ARG	G
G	VAL	ALA	ASP	GLY	U
	VAL	ALA	ASP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

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

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

```
>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



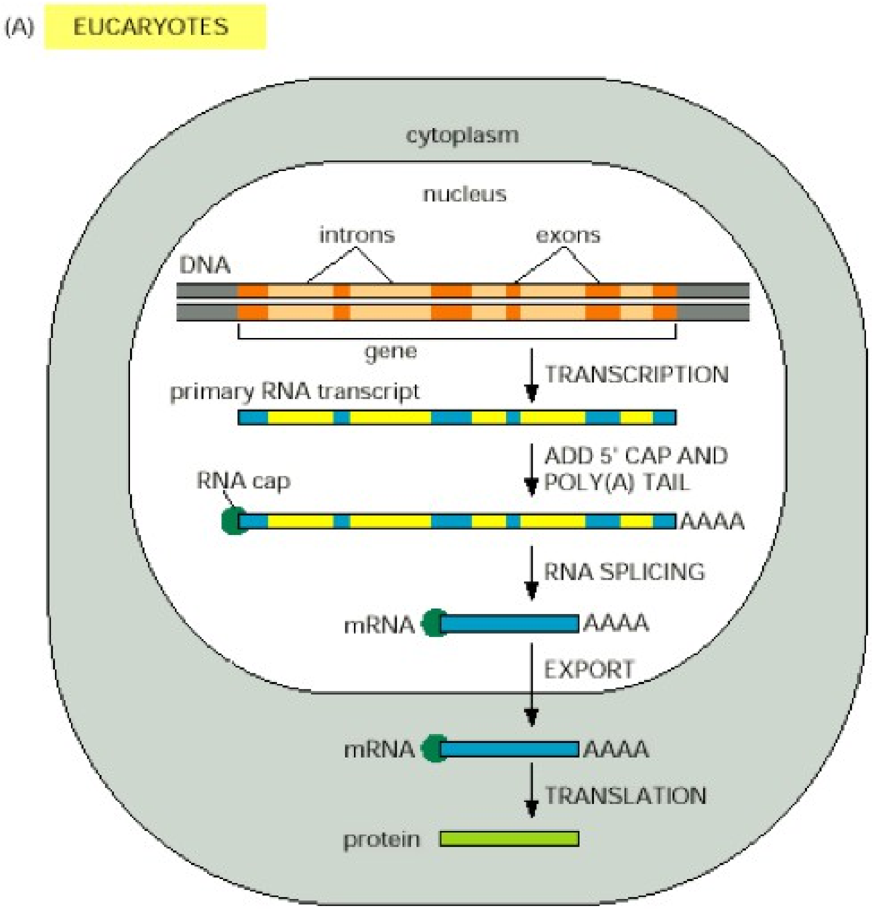
```
>NC_002737.1 Streptococcus pyogenes M1 GAS
TTGTTGATATTCTGTTTTTTCTTTTTTAGTTTTCCACATGAAAAATAGTTGAAAACAATA
GCGGTGTCCCCTTAAAATGGCTTTTCCACAGGTTGTGGAGAACCCAAATTAACAGTGTTA
ATTTATTTTCCACAGGTTGTGGAAAAACTAACTATTATCCATCGTTCTGTGGAAAACTAG
AATAGTTTATGGTAGAATAGTTCTAGAATTATCCACAAGAAGGAACCTAGTATGACTGAA
AATGAACAAATTTTTTGGAACAGGGTCTTGGAATTAGCTCAGAGTCAATTAAAACAGGCA
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
```

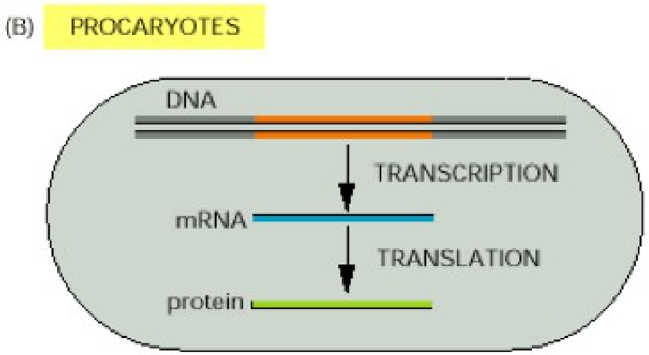
Design a HMM that models the syntax of genes

Gene structure

Depends on the organism (eucaryote or procaryote)

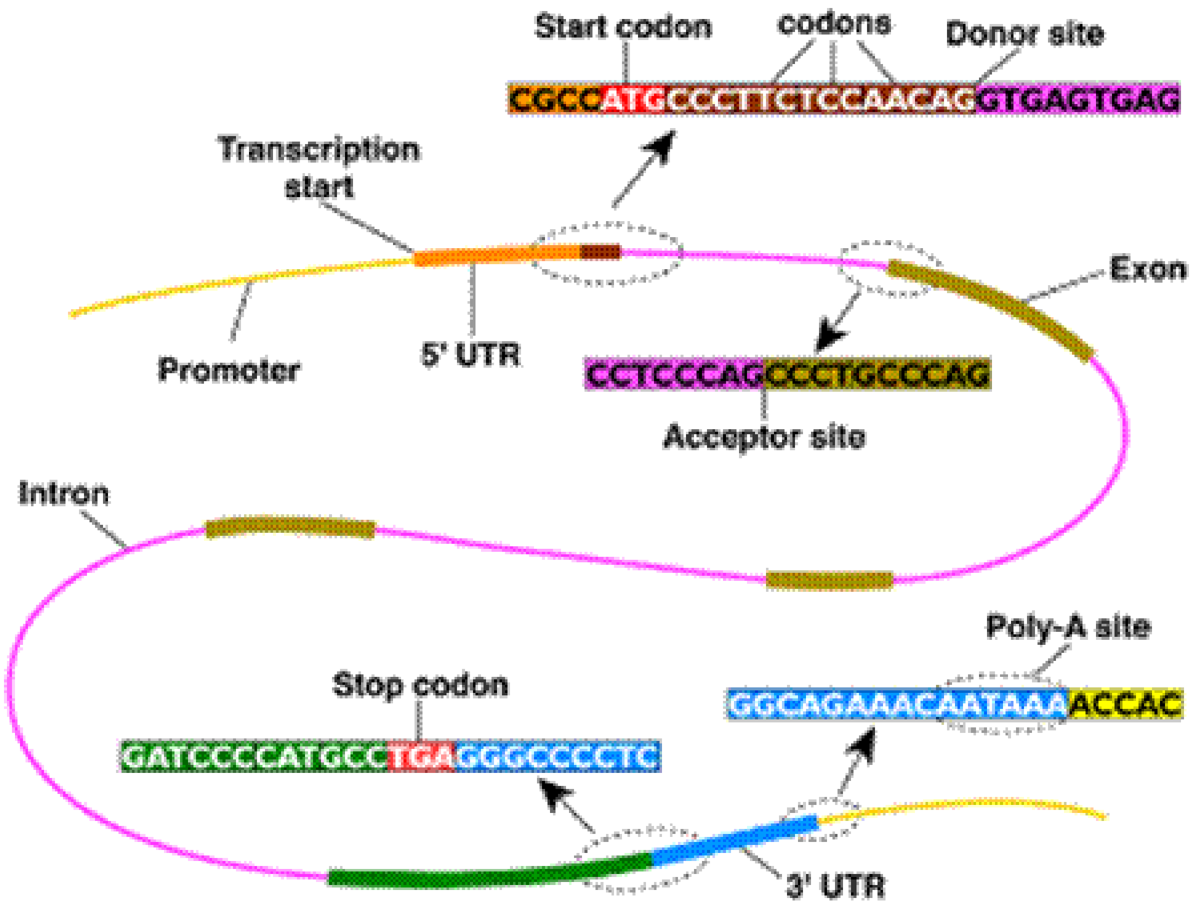


Large genomes. Intron/exon structure and low coding density



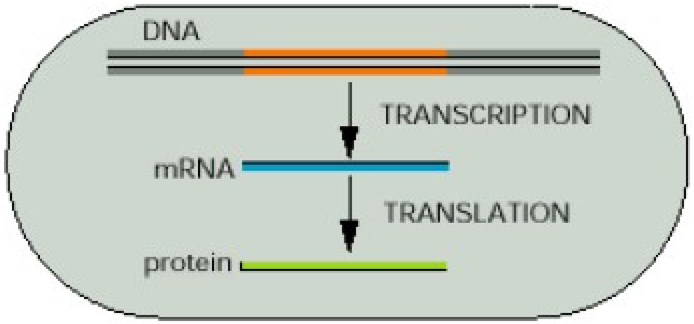
Smaller genomes and high coding density.

Gene structure in eukaryotes



Eukaryotic gene structure in more details

Gene structure in procaryotes

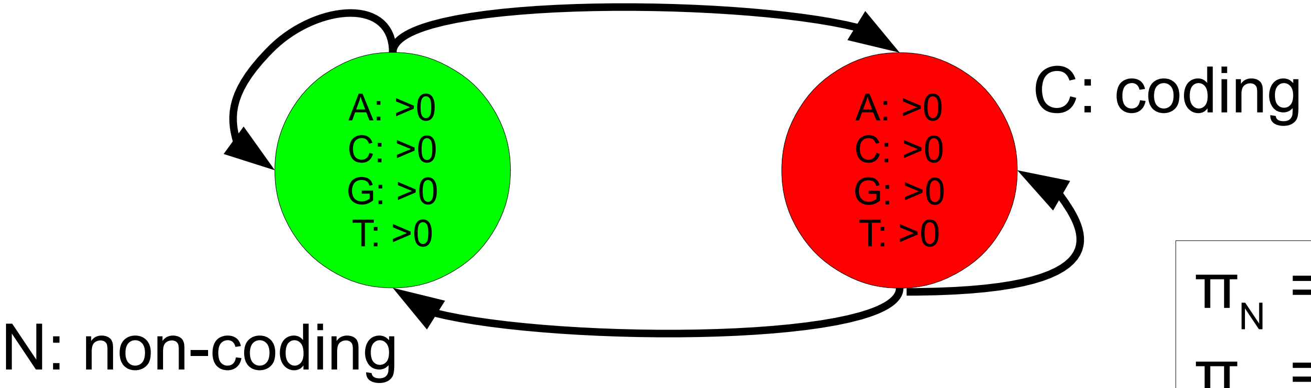


Biological facts

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

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

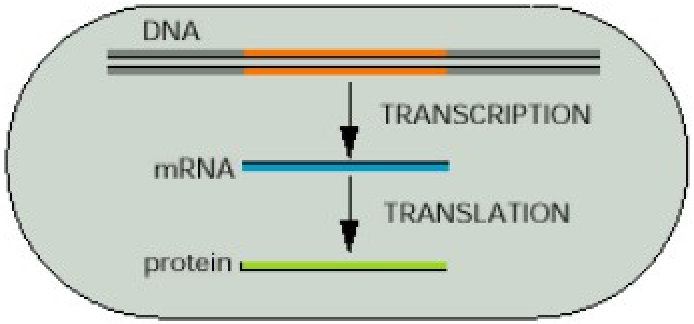
X: acgatgcgctaatatgtccgatgacgtgagcataagcgacatgcag



$\pi_N = 1$

$\pi_C = 0$

Gene structure in procaryotes

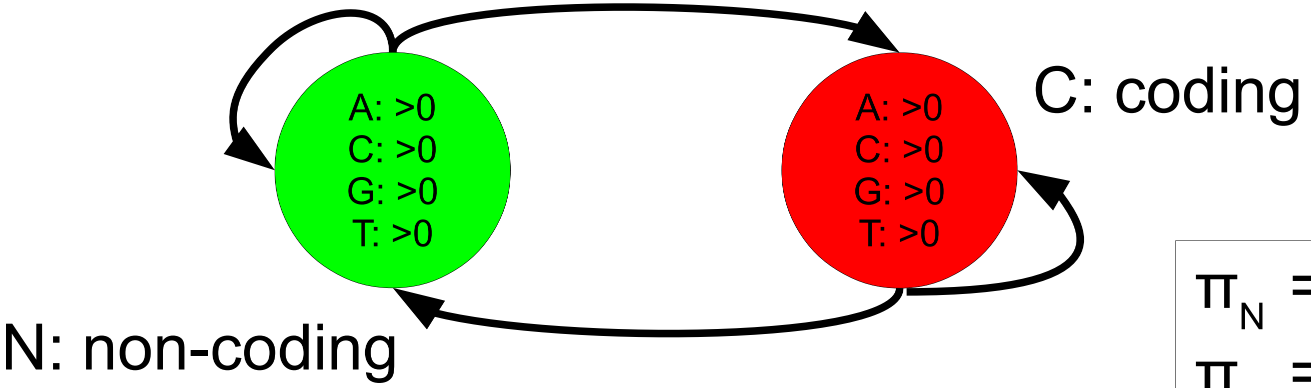


Biological facts

- 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

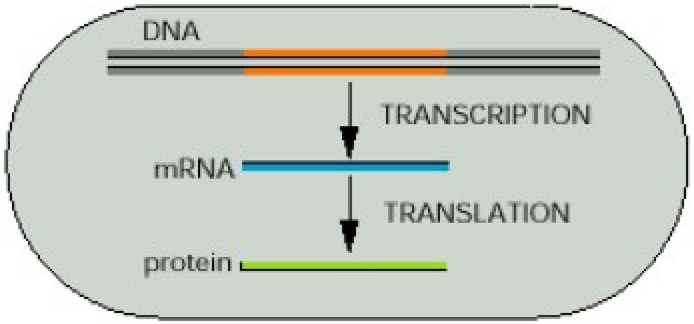
X: acgatgcgctaatatgtccgatgacgtgagcataagcgacatgcag



$\pi_N = 1$

$\pi_C = 0$

Gene structure in procaryotes



Biological facts

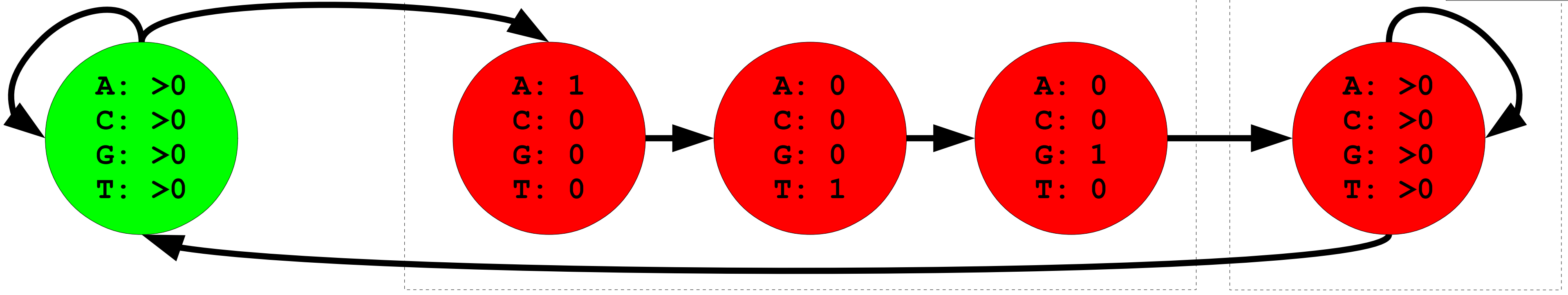
- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacat

$\pi_N = 1$

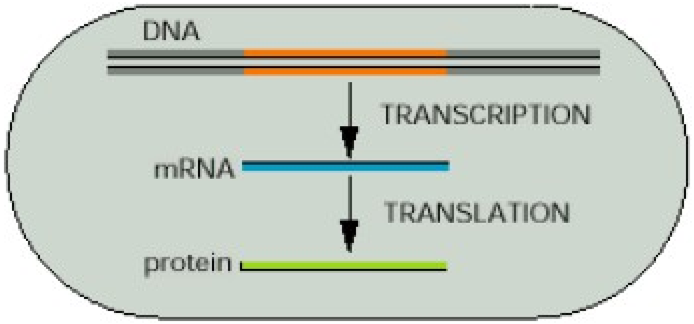
$\pi_C = 0$



N: non-coding

C: coding

Gene structure in procaryotes



Biological facts

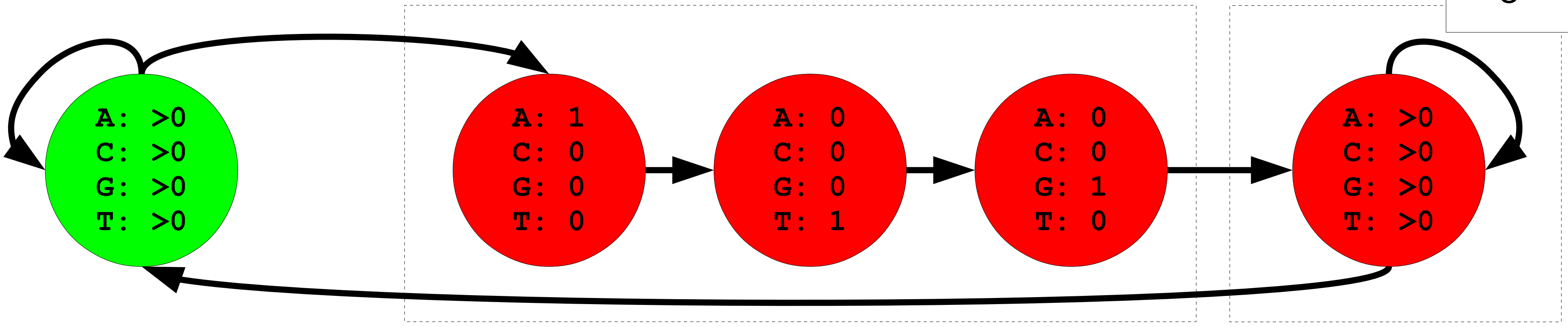
- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**
- The gene ends with a stop-codon **taa**, **tag** or **tga**

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacat

$\pi_N = 1$

$\pi_C = 0$



N: non-coding

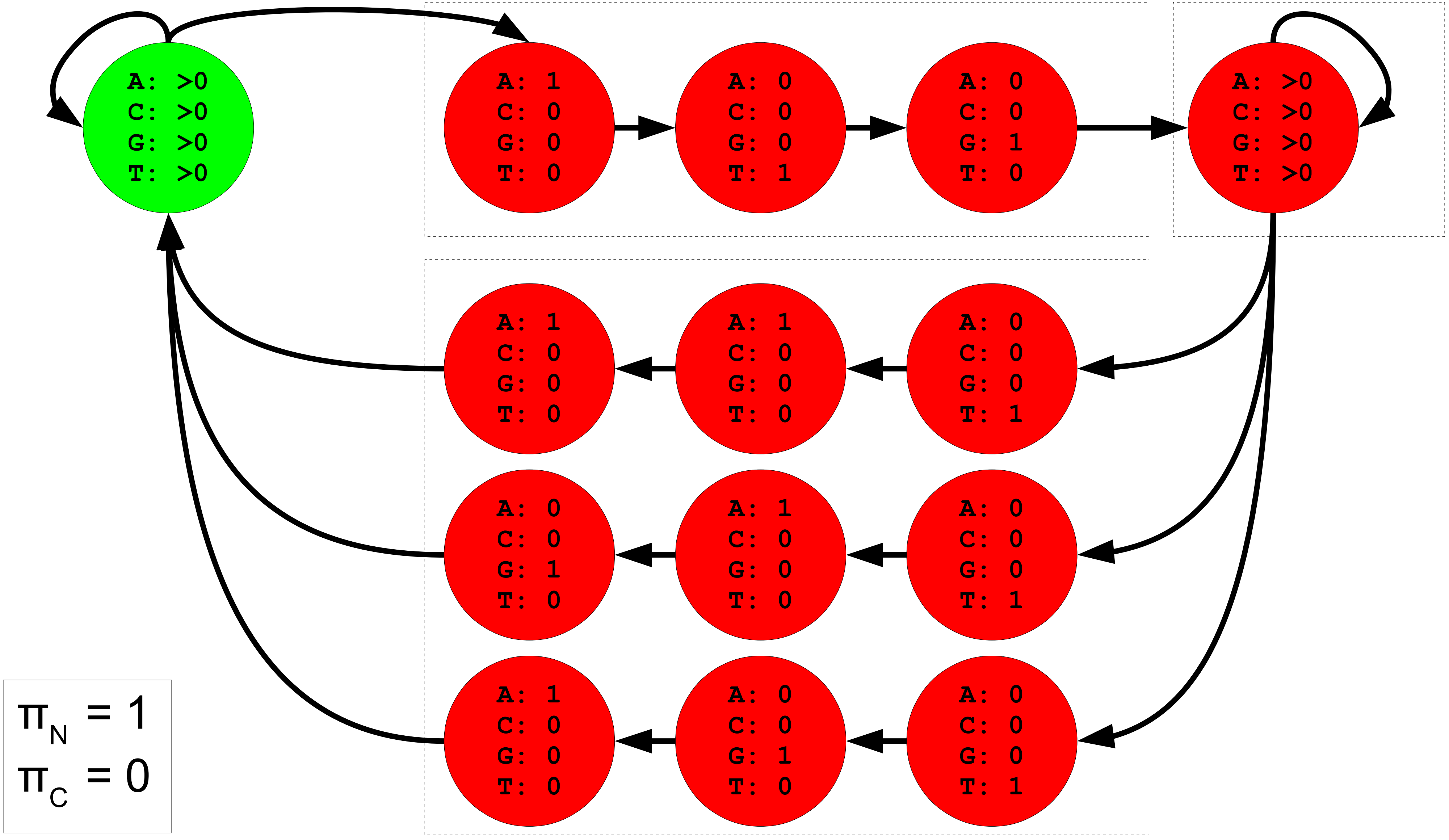
C: coding

Gene structu

- 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**

N: non-coding

C: coding



A: >0
C: >0
G: >0
T: >0

A: 1
C: 0
G: 0
T: 0

A: 0
C: 0
G: 0
T: 1

A: 0
C: 0
G: 1
T: 0

A: >0
C: >0
G: >0
T: >0

A: 1
C: 0
G: 0
T: 0

A: 1
C: 0
G: 0
T: 0

A: 0
C: 0
G: 0
T: 1

A: 0
C: 0
G: 1
T: 0

A: 1
C: 0
G: 0
T: 0

A: 0
C: 0
G: 0
T: 1

A: 1
C: 0
G: 0
T: 0

A: 0
C: 0
G: 1
T: 0

A: 0
C: 0
G: 0
T: 1

$\pi_N = 1$

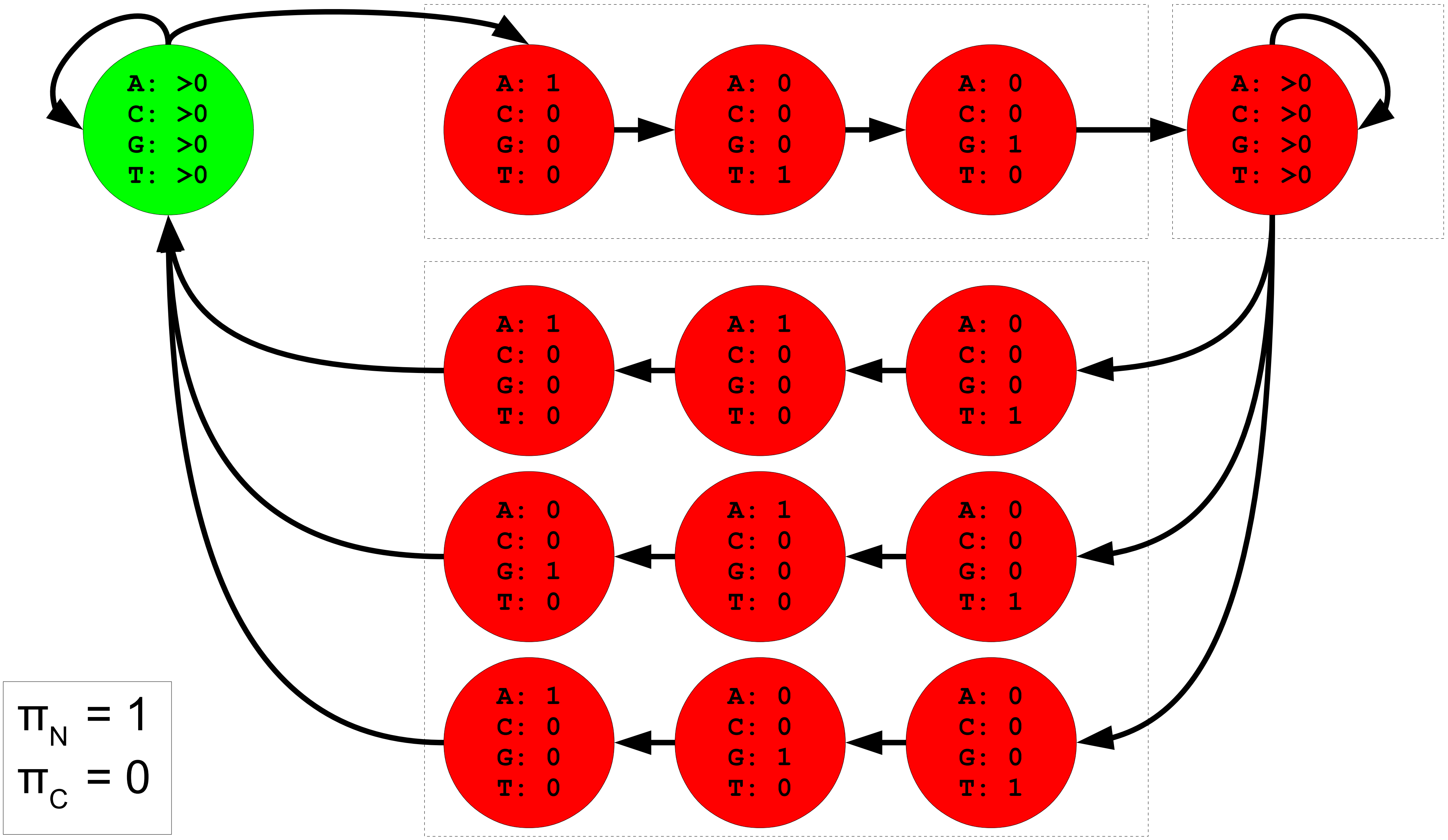
$\pi_C = 0$

Gene structu

- 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

N: non-coding

C: coding

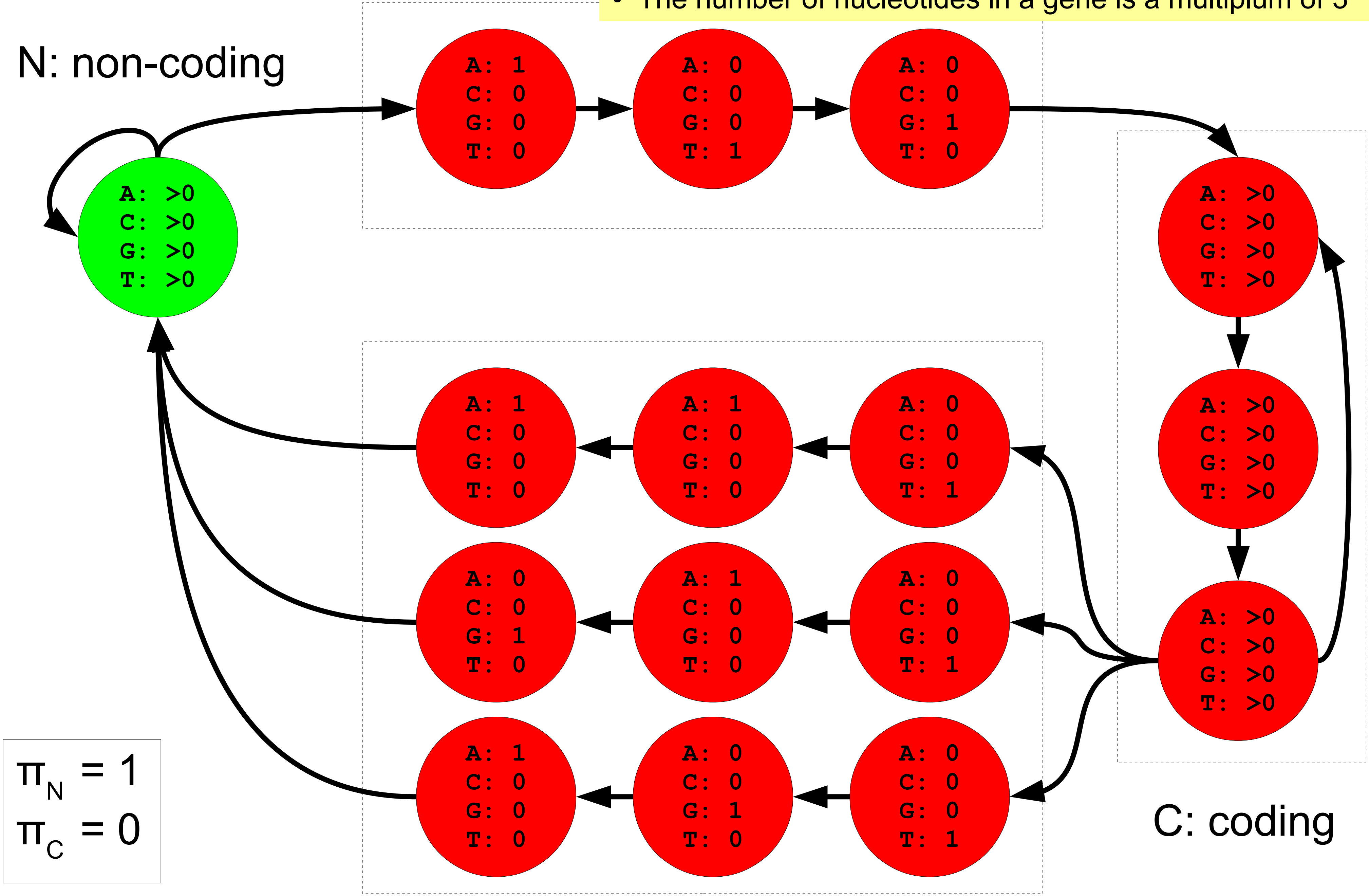


$\pi_N = 1$

$\pi_C = 0$

Gene structu

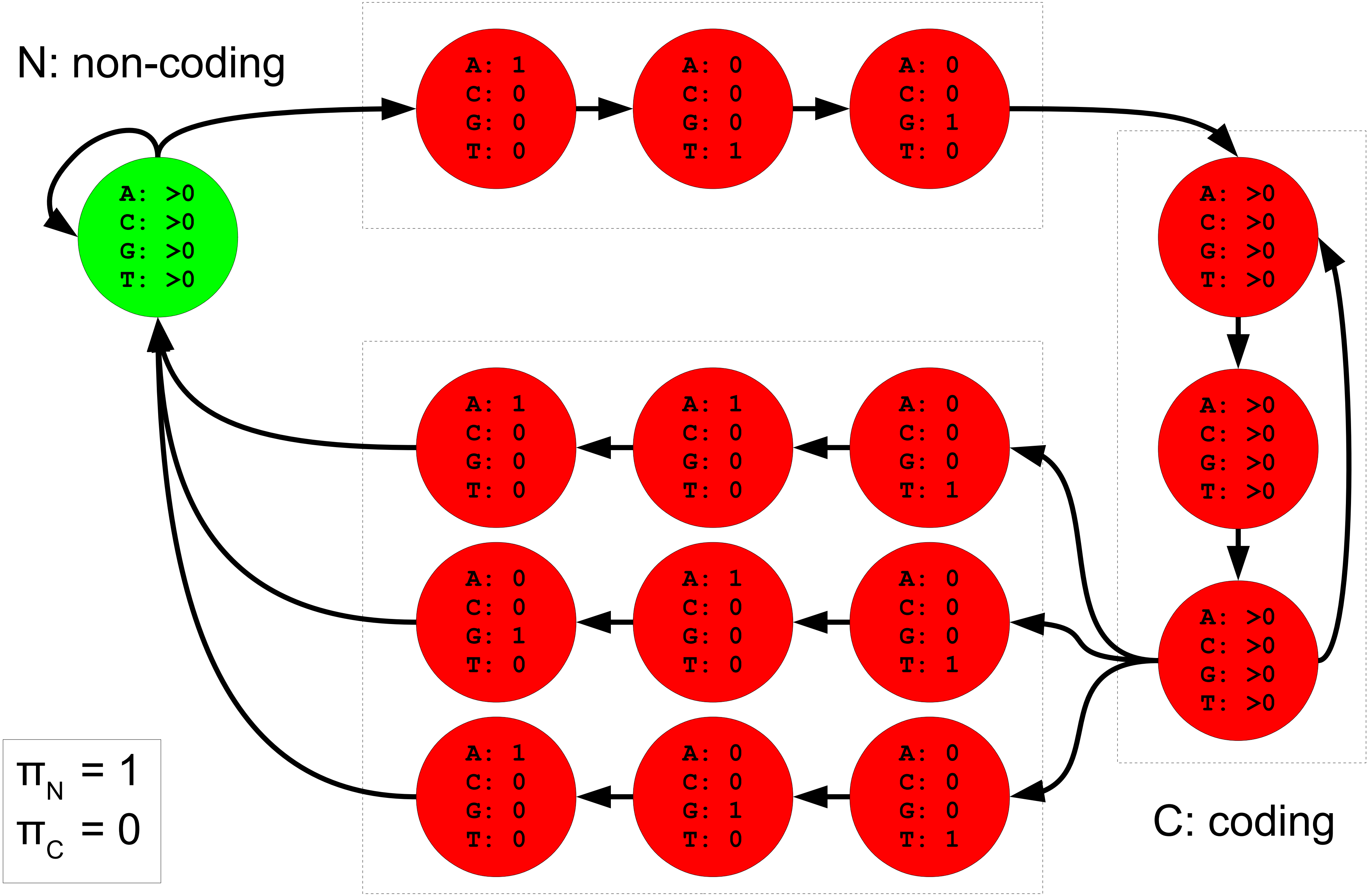
- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**
- The gene ends with a stop-codon **taa**, **tag** or **tga**
- The number of nucleotides in a gene is a multiplum of 3



$\pi_N = 1$

$\pi_C = 0$

Gene structure in procaryotes



N: non-coding

A: >0
C: >0
G: >0
T: >0

A: 1
C: 0
G: 0
T: 0

A: 0
C: 0
G: 0
T: 1

A: 0
C: 0
G: 1
T: 0

A: >0
C: >0
G: >0
T: >0

A: >0
C: >0
G: >0
T: >0

A: >0
C: >0
G: >0
T: >0

A: 1
C: 0
G: 0
T: 0

A: 1
C: 0
G: 0
T: 0

A: 0
C: 0
G: 0
T: 1

A: 0
C: 0
G: 1
T: 0

A: 1
C: 0
G: 0
T: 0

A: 0
C: 0
G: 0
T: 1

A: 1
C: 0
G: 0
T: 0

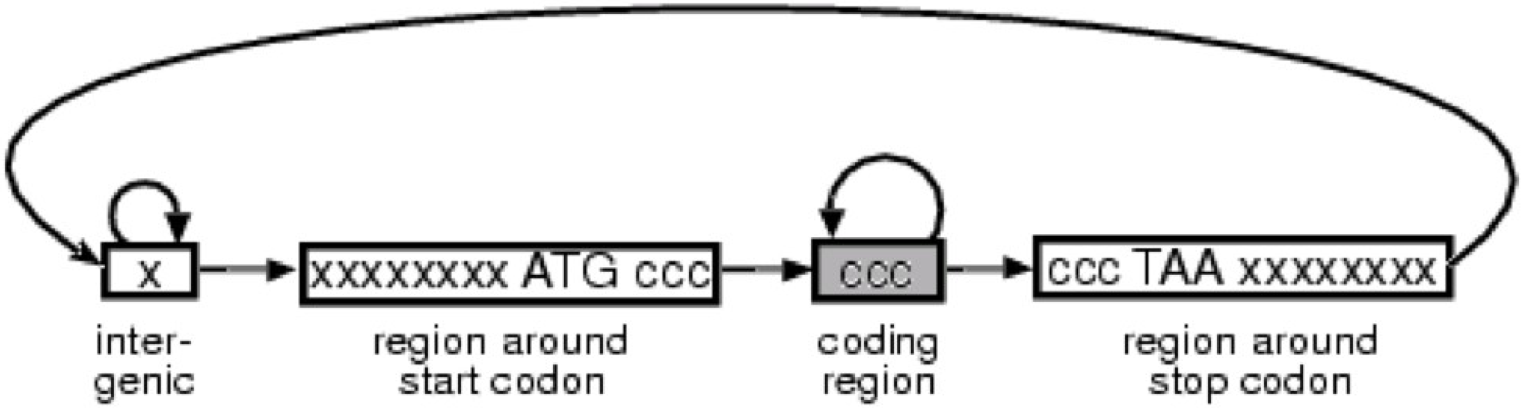
A: 0
C: 0
G: 1
T: 0

A: 0
C: 0
G: 0
T: 1

C: coding

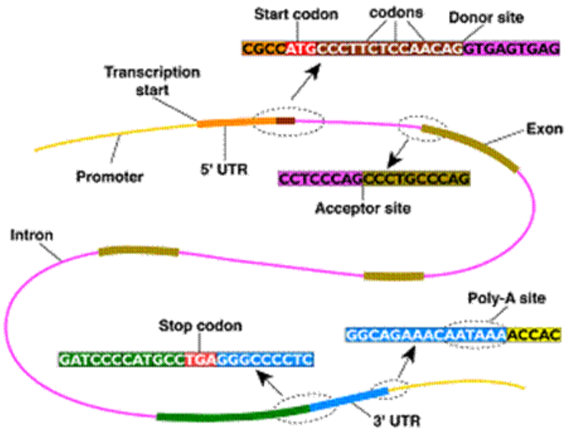
$\pi_N = 1$
$\pi_C = 0$

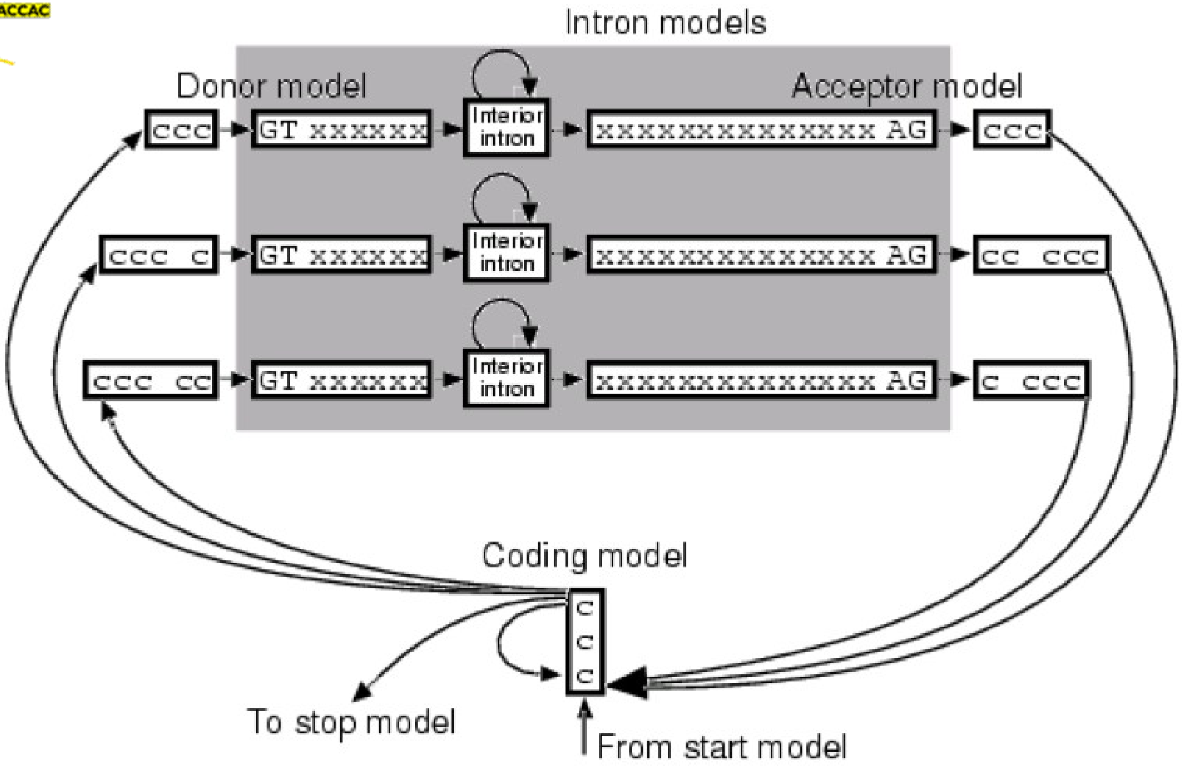
Gene structure in procaryotes



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

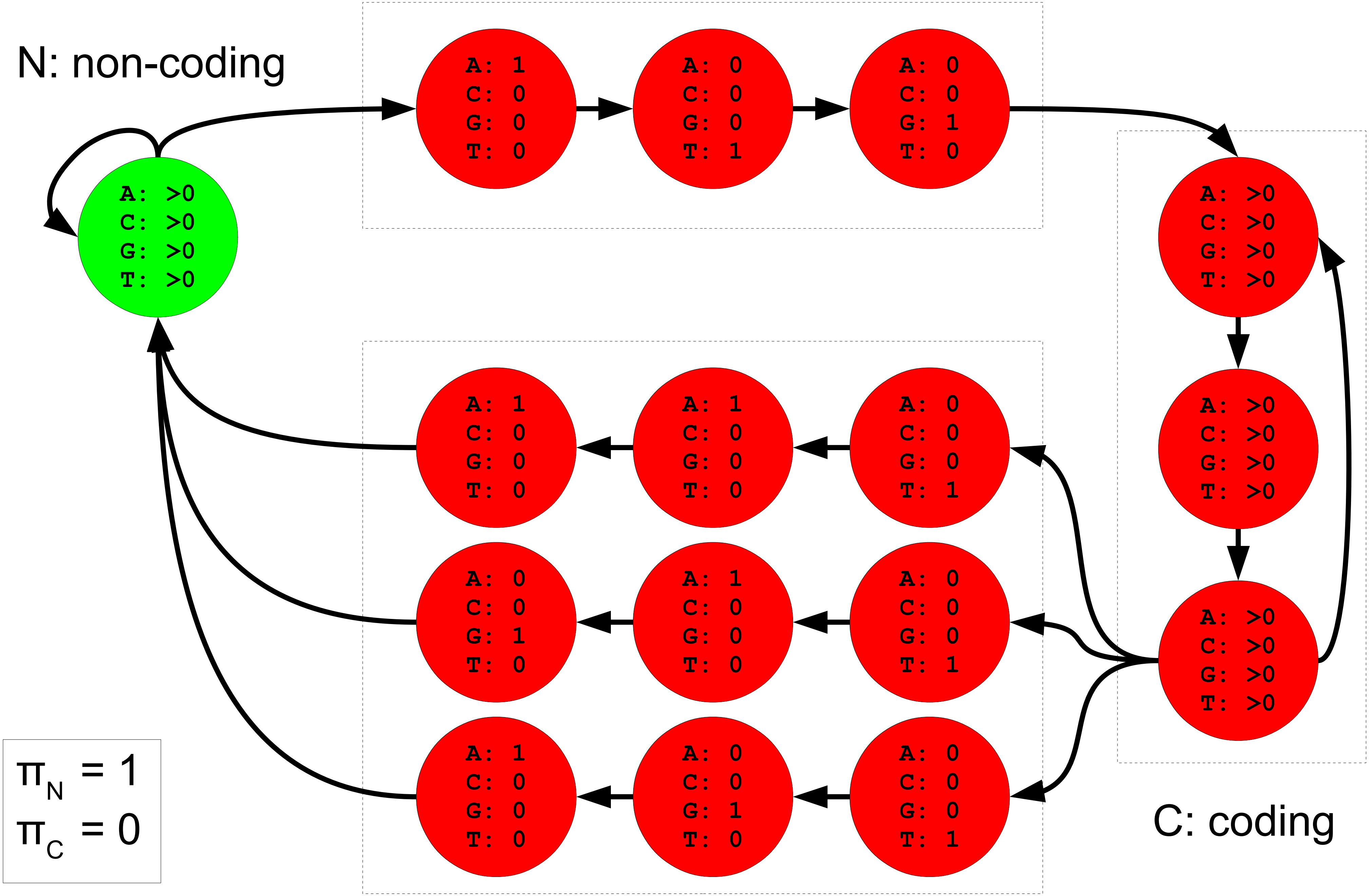
Gene structure in eukaryotes





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

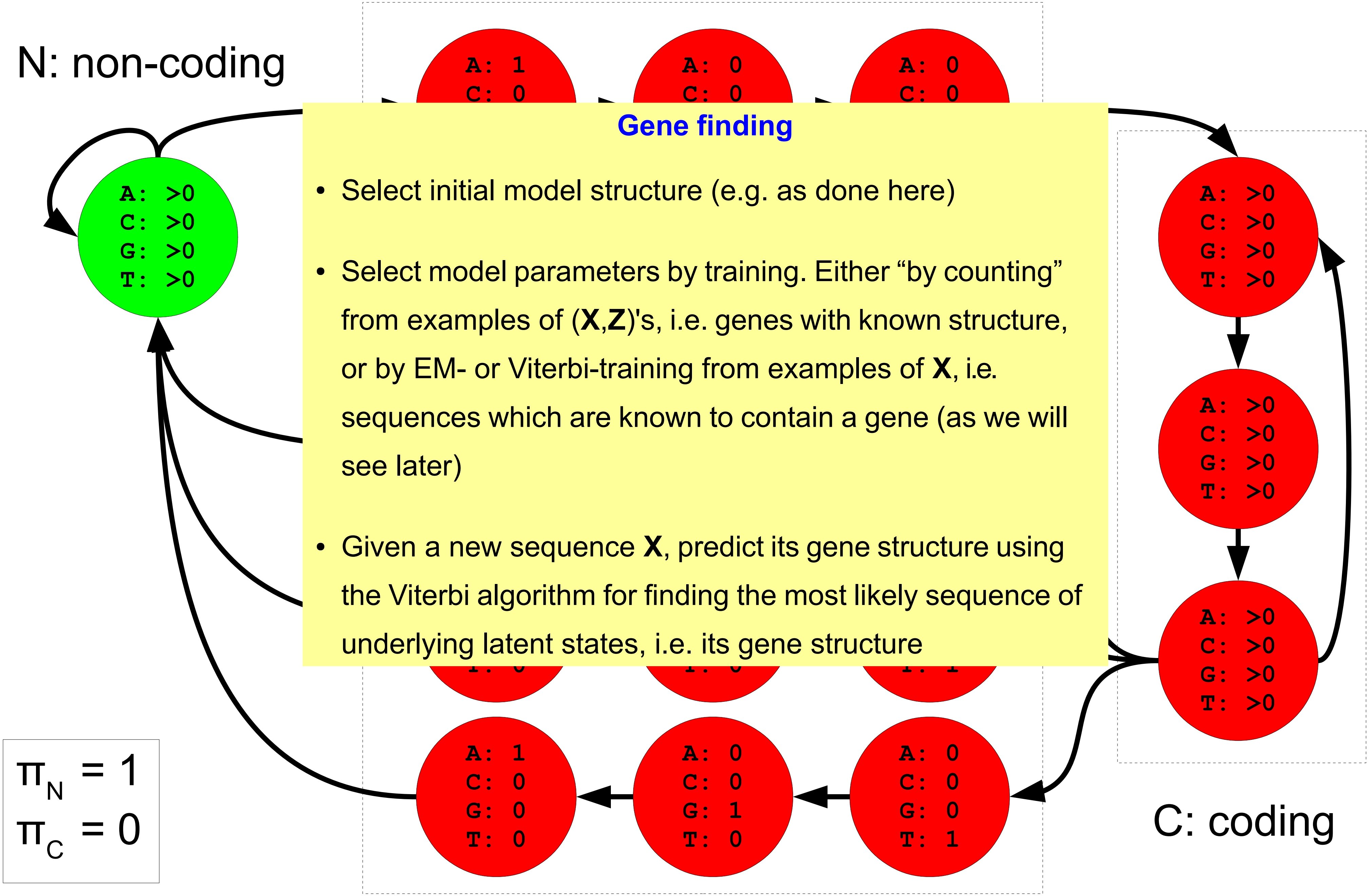
Gene structure in procaryotes



$\pi_N = 1$

$\pi_C = 0$

Gene structure in procaryotes



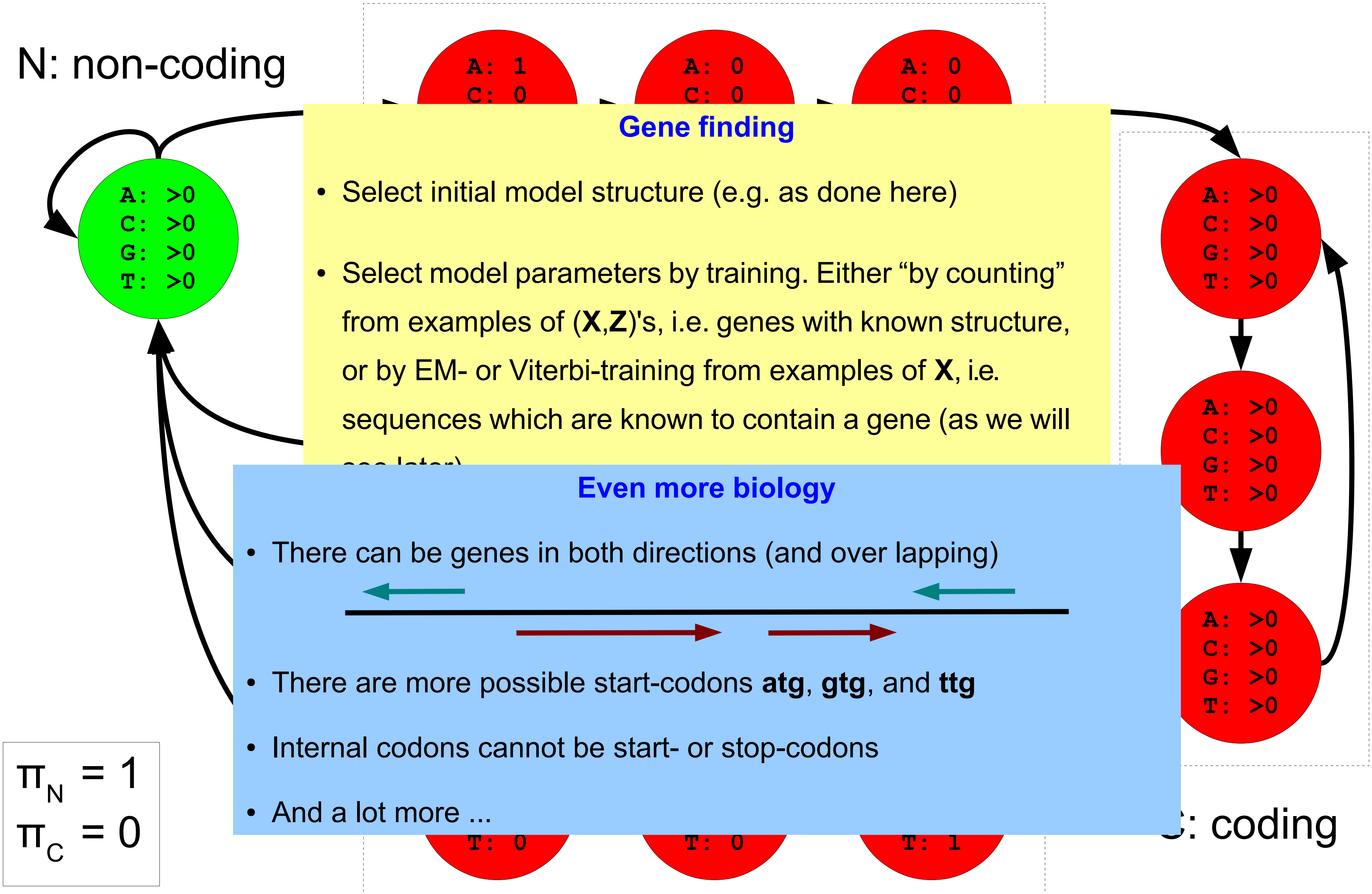
$\pi_N = 1$

$\pi_C = 0$

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 (as we will see later)
- 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



$\pi_N = 1$

$\pi_C = 0$

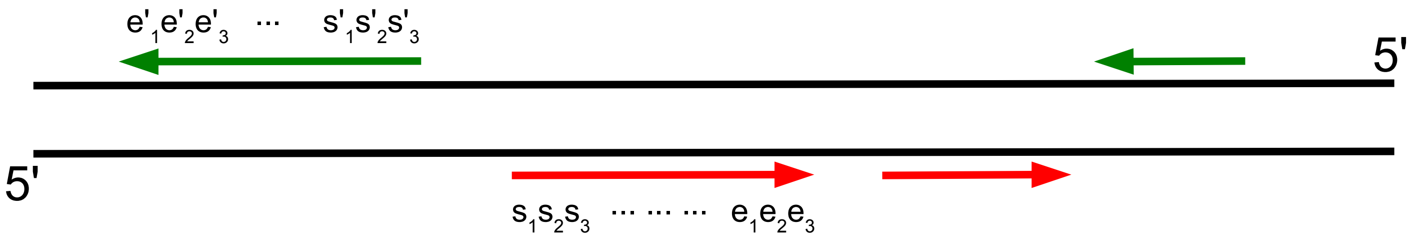
Gene finding

- 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 (as we will see later)

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 start- or stop-codons
- And a lot more ...

DNA

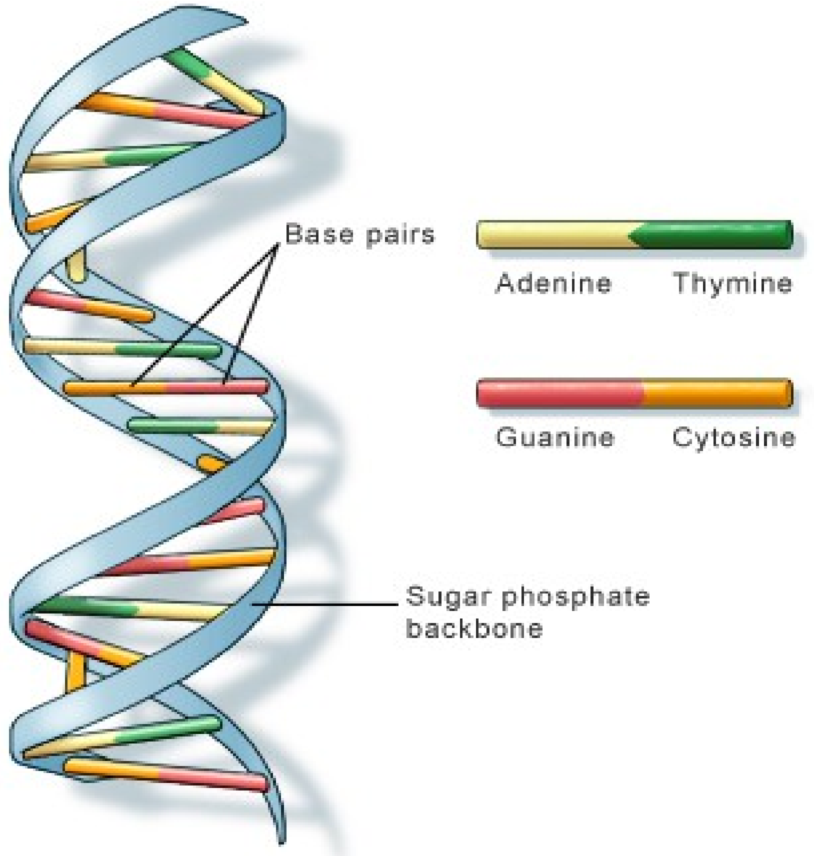


$e'_1e'_2e'_3$ ··· $s'_1s'_2s'_3$

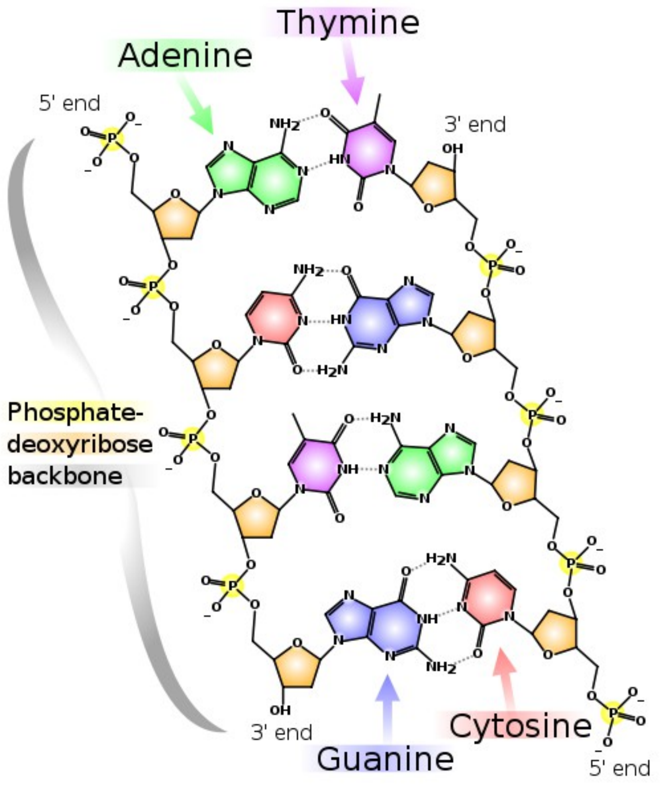
5'

5'

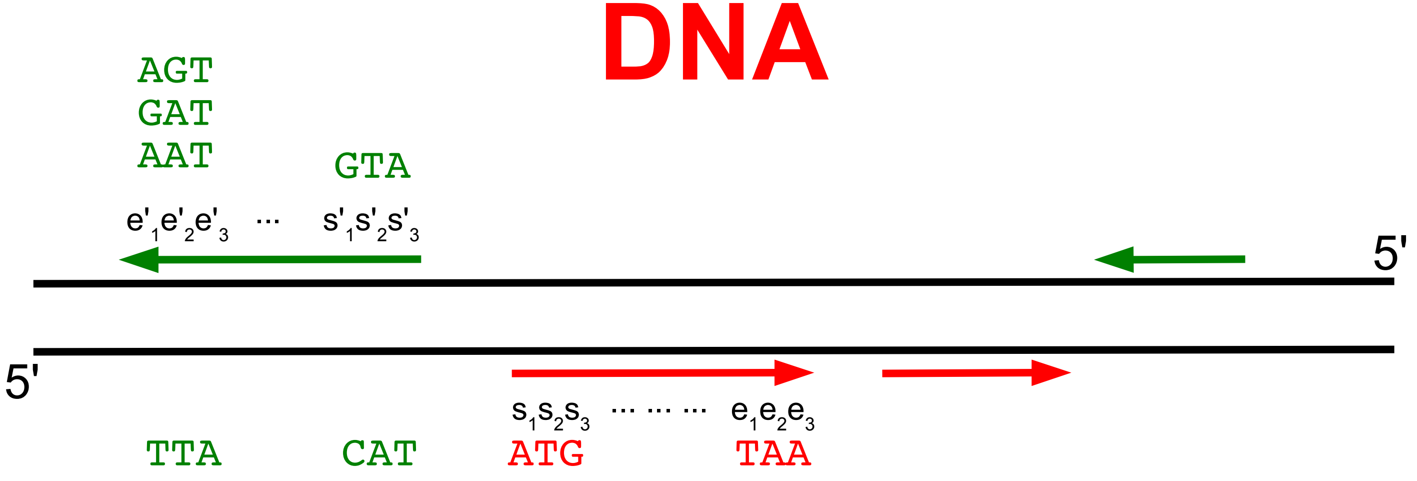
$s_1s_2s_3$ ··· ··· ··· $e_1e_2e_3$



U.S. National Library of Medicine



DNA


AGT
GAT
AAT

GTA

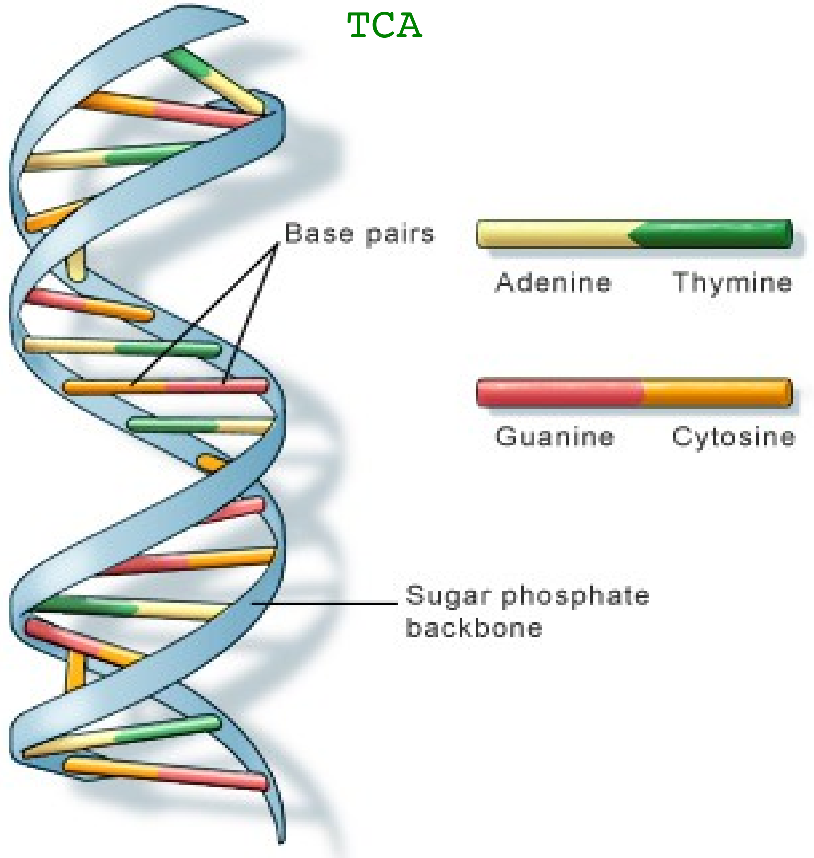
$e'_1e'_2e'_3$ ··· $s'_1s'_2s'_3$

5'

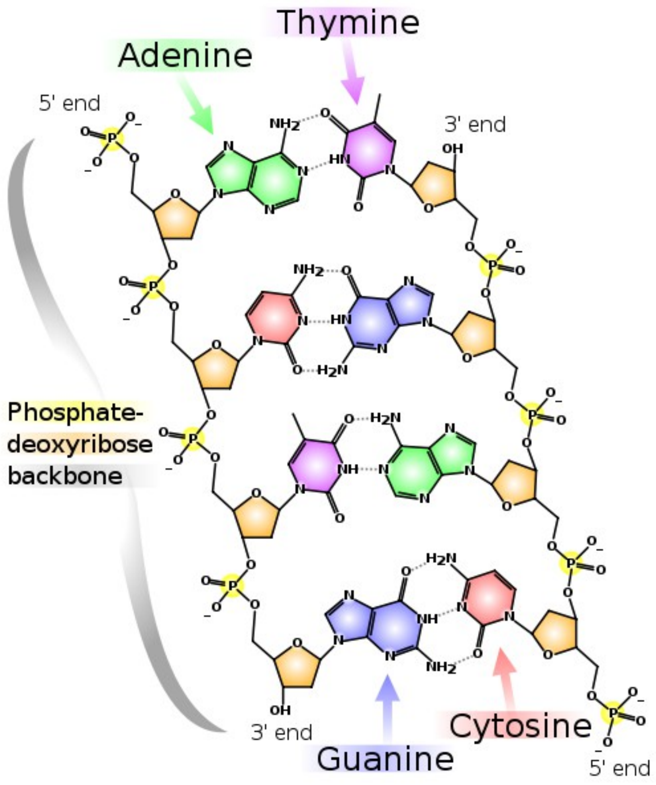
5'

$s_1s_2s_3$ ··· ··· ··· $e_1e_2e_3$

TTA CAT ATG TAA
CTA TAG
TCA TGA

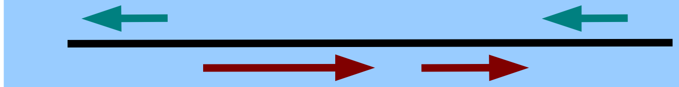


U.S. National Library of Medicine



Even more biology

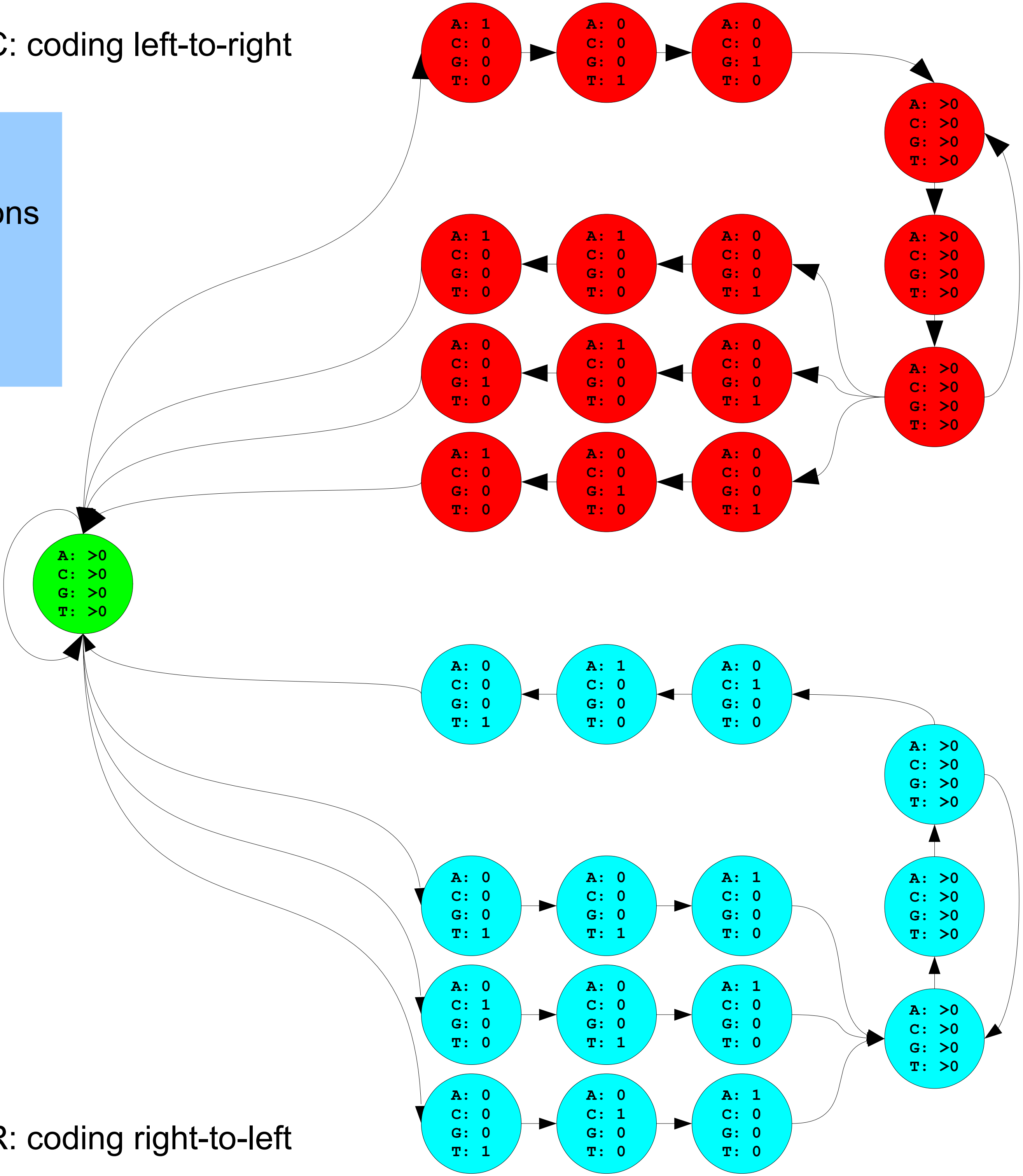
There can be genes in both directions



C: coding left-to-right

N: Non-coding

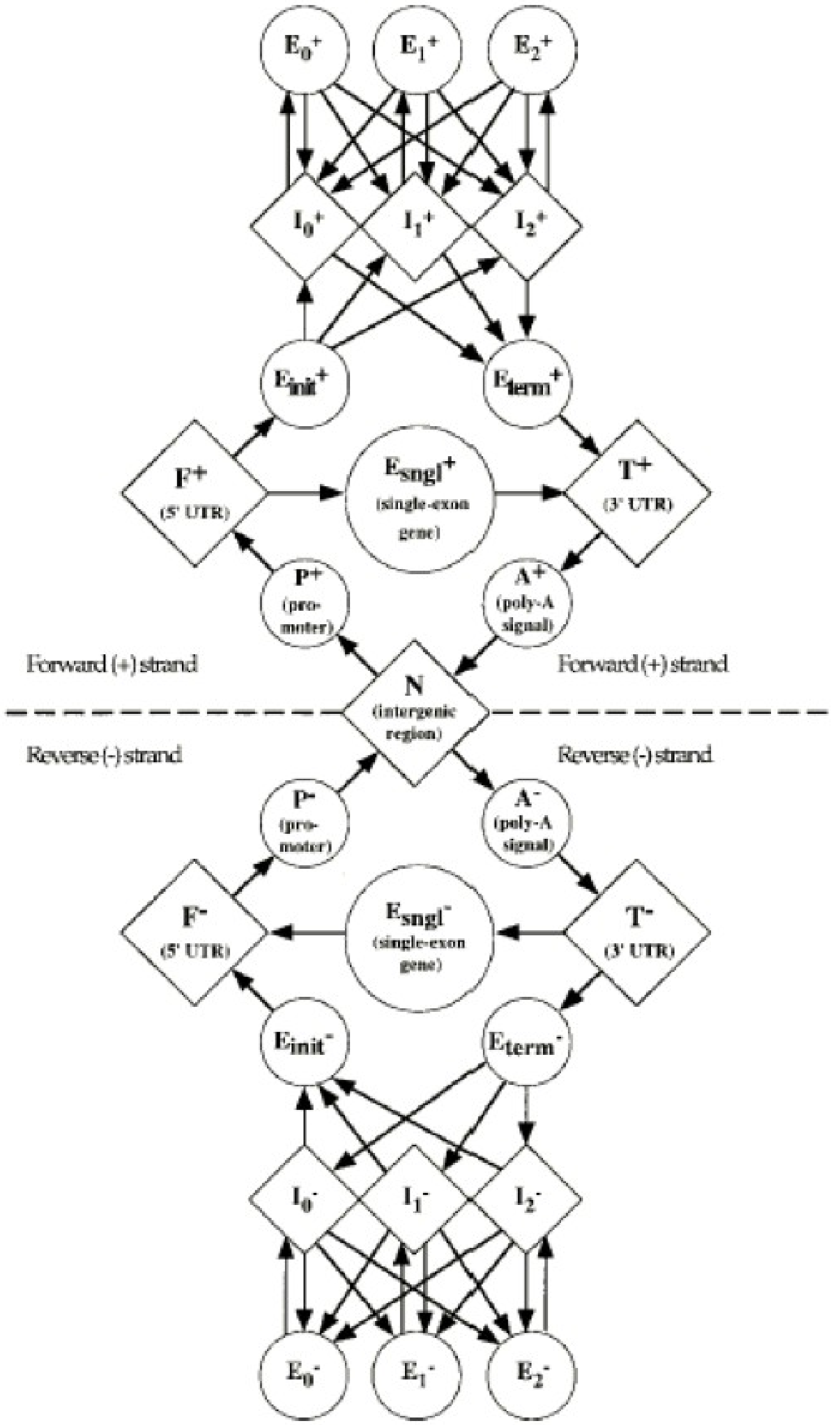
R: coding right-to-left



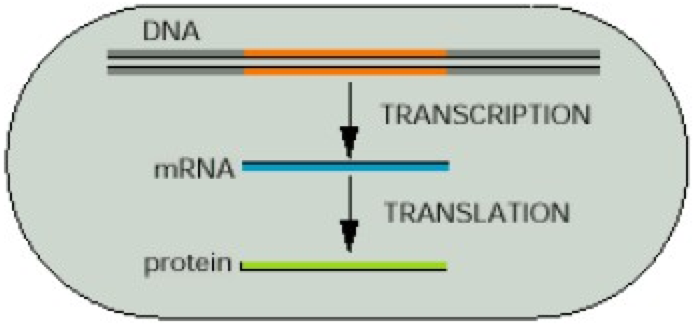
$\pi_N = 1$

$\pi_C = 0$

Eukaryotic gene structure in both directions (GenScan)



Problem: From annotation to Z



Biological facts

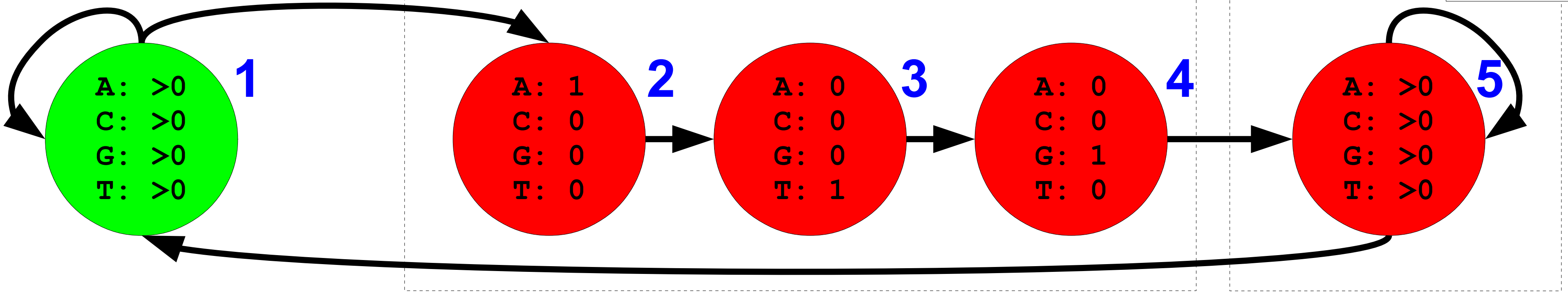
- The gene is a substring of the DNA sequence of A,C,G,T's
- The gene starts with a start-codon **atg**

Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacatc

$\pi_N = 1$

$\pi_C = 0$



N: non-coding

C: coding

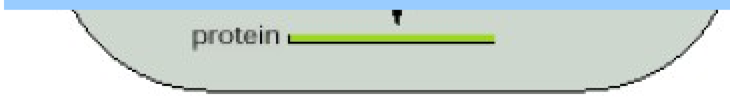
Problem: From annotation to Z

Problem: The string **Z=**NNNCCC.... is not a prober sequence of states in the illustrated HMM, but is can easily be converted into one.

Biological facts

s a substring of the DNA sequence of A,C,G,T's

tarts with a start-codon **atg**

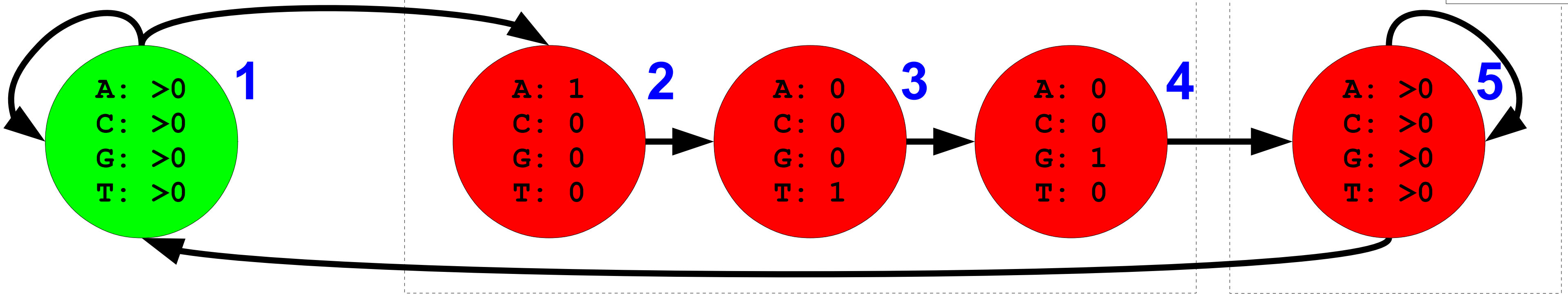


Z: NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacat

$\pi_N = 1$

$\pi_C = 0$



N: non-coding

C: coding

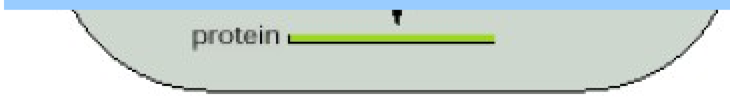
Problem: From annotation to Z

Problem: The string **Z=**NNNCCC.... is not a prober sequence of states in the illustrated HMM, but is can easily be converted into one.

Biological facts

s a substring of the DNA sequence of A,C,G,T's

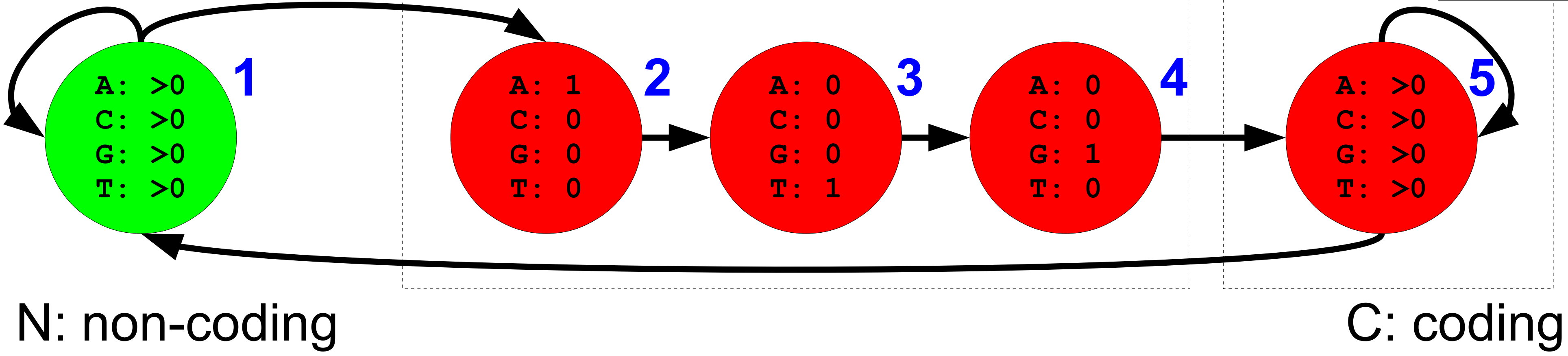
tarts with a start-codon **atg**



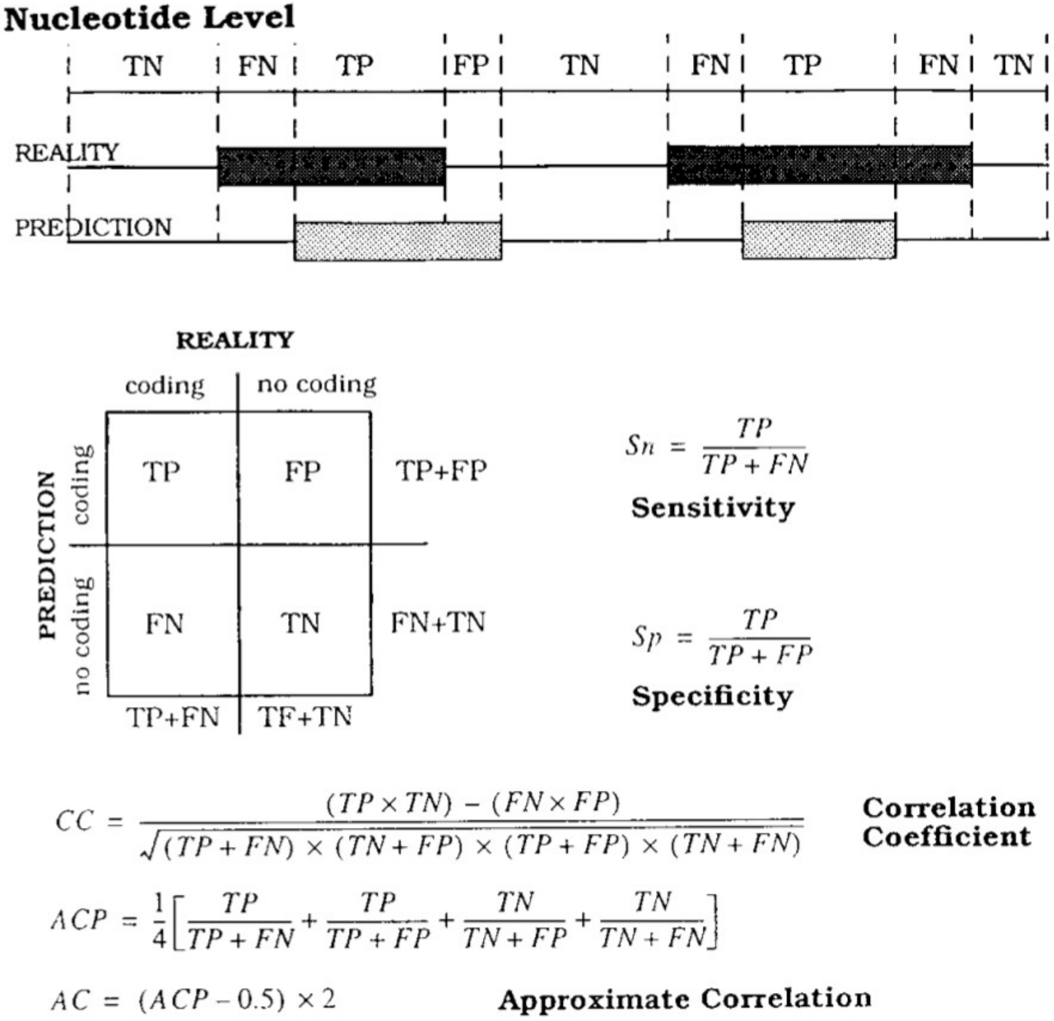
```
    1112345555551111111123455555555555551111111111
Z:  NNNCCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCNNNNNNNNNNN

X:  acgatgcgctaatatgtccgatgacgtgagcataagcgacat
```

$\pi_N = 1$

$\pi_C = 0$



Evaluating performance



Nucleotide Level

REALITY

PREDICTION	coding	no coding	
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

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