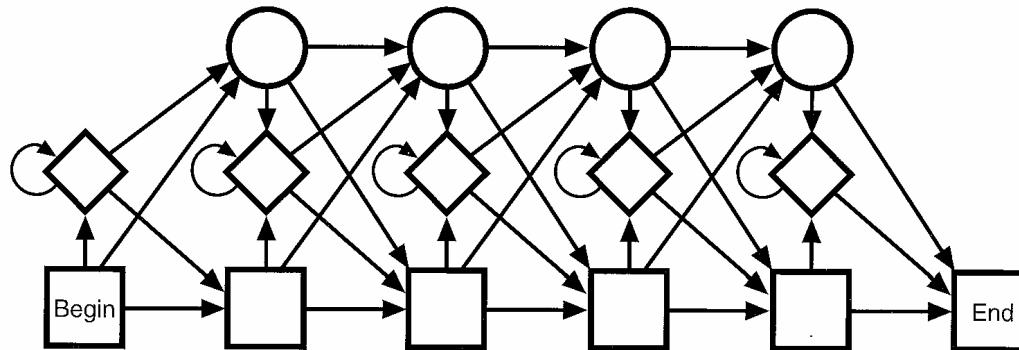


Genome Annotation (protein coding genes)



Genome analysis; Nov 14 2007

Motivation

Problem: Once we know the sequence, we wish to *understand* it.

First step is to understand find the genes.

Finding the genes is the first step in

- Determining the proteome
- Study gene regulation
- Map genotypes to phenotypes
- ...

Protein coding genes

Protein coding genes are the best understood.

Several approaches exist for searching for them.

We will cover only a small fraction of the approaches here.

Protein coding genes

Search for functional sites:

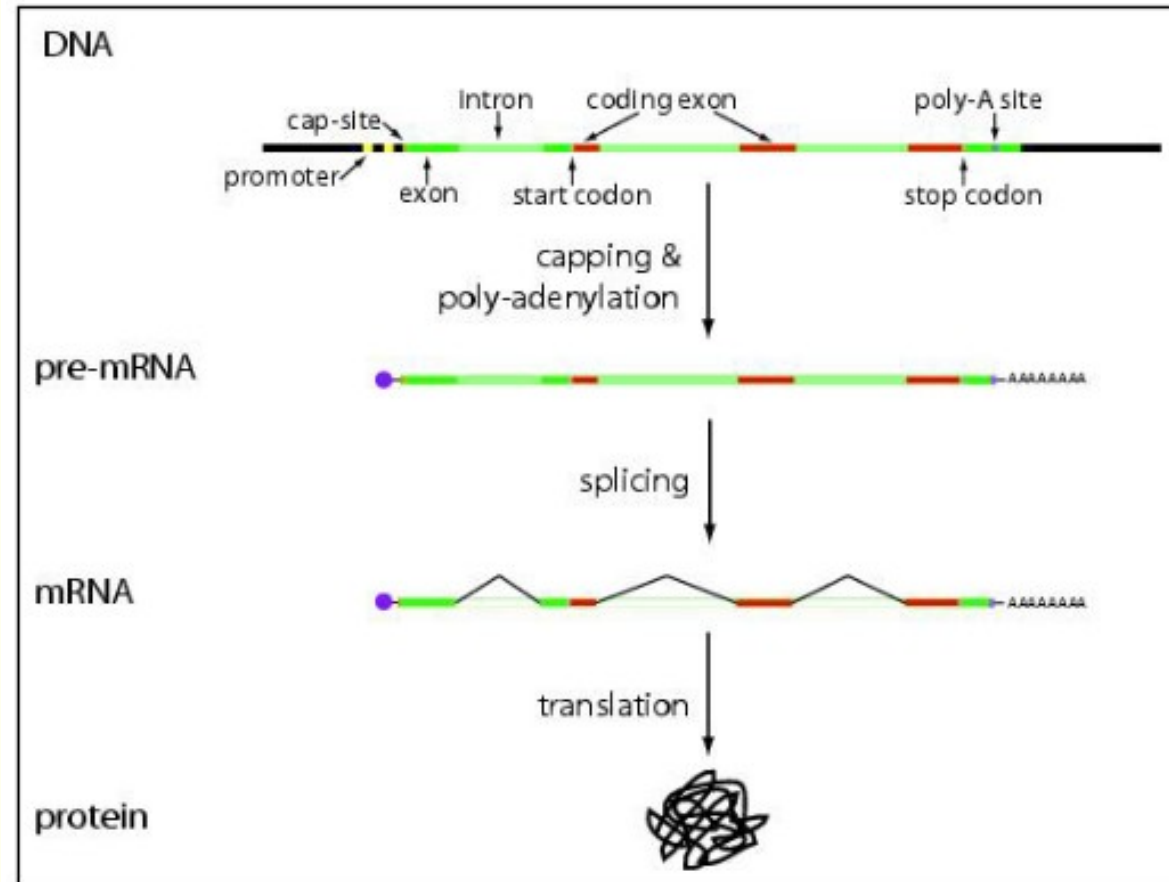
- e.g. promoters, cap-sites, poly-A sites, splice sites, start and stop codons...

Analysis of sequence composition:

- e.g. nucleotide usage and triple distribution...

Homology search:

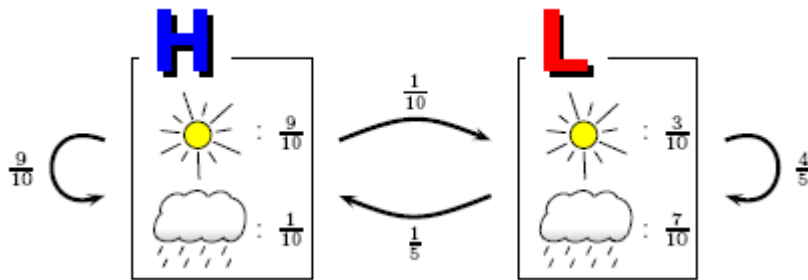
- e.g. search against databases of known proteins or known genes...



Hidden Markov models

Hidden Markov model: Moves from state to state in the same way as a Markov model, but **emits a symbol** (from a finite alphabet of the model) from each state visited (except silent states) **according to a probability distribution of the state** (emission probabilities) ...

Model M :



A **run** follows a sequence of states:

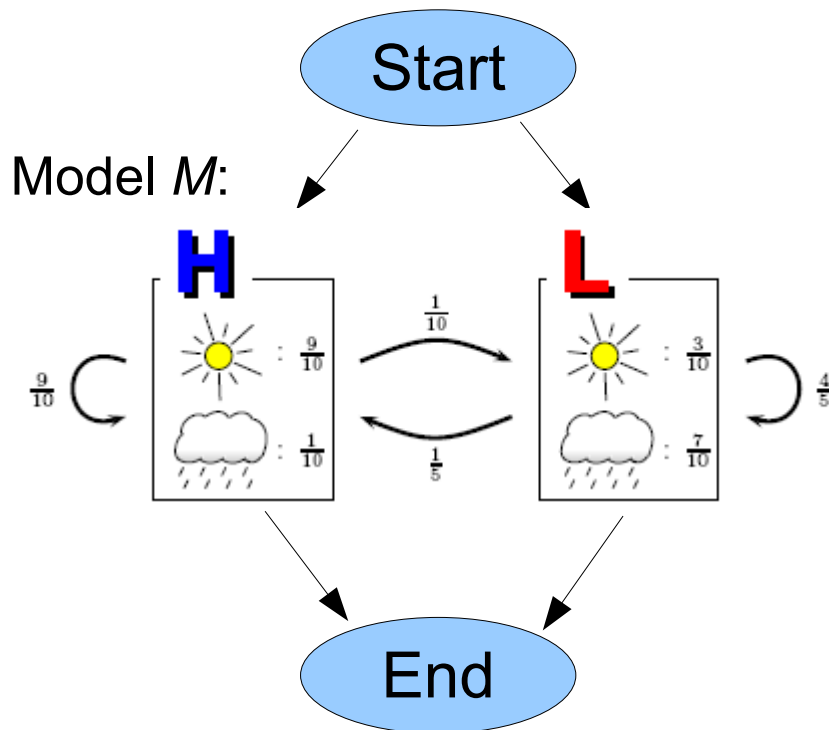
H H L L H

And **emits** a sequence of symbols:



Hidden Markov models

Hidden Markov model: Moves from state to state in the same way as a Markov model, but **emits a symbol** (from a finite alphabet of the model) from each state visited (except silent states) **according to a probability distribution of the state** (emission probabilities) ...



A **run** follows a sequence of states:

H H L L H

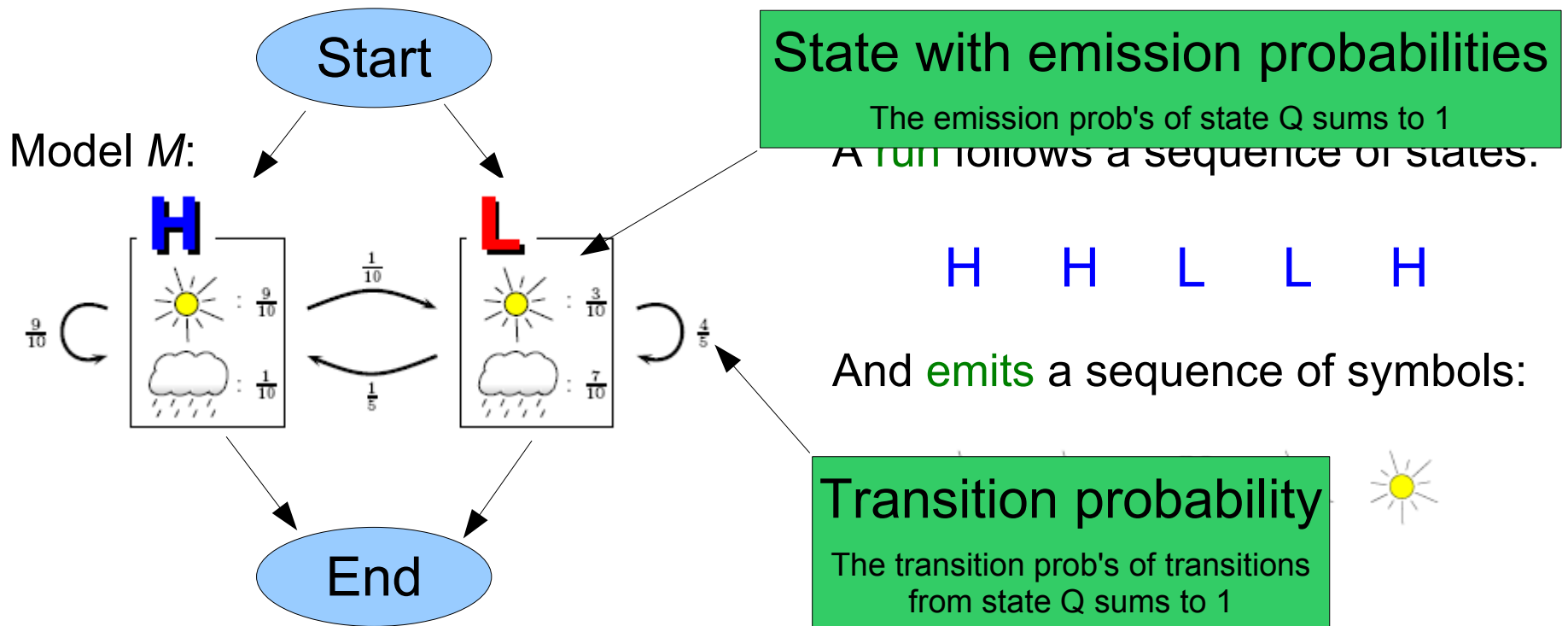
And **emits** a sequence of symbols:



Special **Start**- and **End**-states can be added to generate finite output

Hidden Markov models

Hidden Markov model: Moves from state to state in the same way as a Markov model, but **emits a symbol** (from a finite alphabet of the model) from each state visited (except silent states) **according to a probability distribution of the state** (emission probabilities) ...



Special **Start**- and **End**-states can be added to generate finite output

History of HMMs

History of HMMs

Hidden Markov Models were introduced in statistical papers by Leonard E. Baum and others in the late 1960s. One of the first applications of HMMs was speech recognition in the mid-1970s.

In the late 1980s, HMMs were applied to the analysis of biological sequences. Since then, many applications in bioinformatics...

Applications of HMMs in bioinformatics

prediction of protein-coding regions in genome sequences
modeling families of related DNA or protein sequences
prediction of secondary structure elements in proteins
... and many others ...

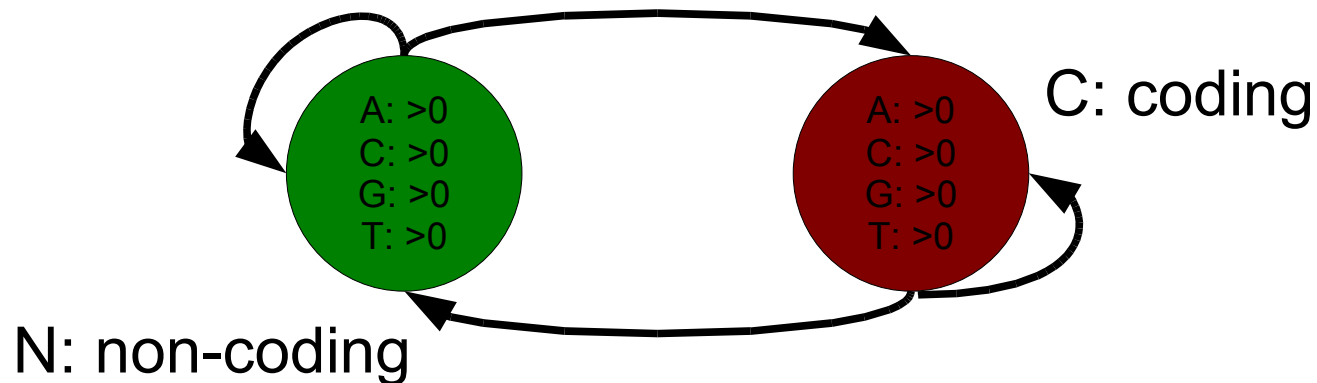
HMM based gene finding

Biological facts

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

Z: NNNCCCCCCCCNNNNNNNNCCCCCCCCCCCCCCCCNNNNNNNNNNNN

X: acgatgcgctaatatgtccgatgacgtgagcataagcgacatgcag



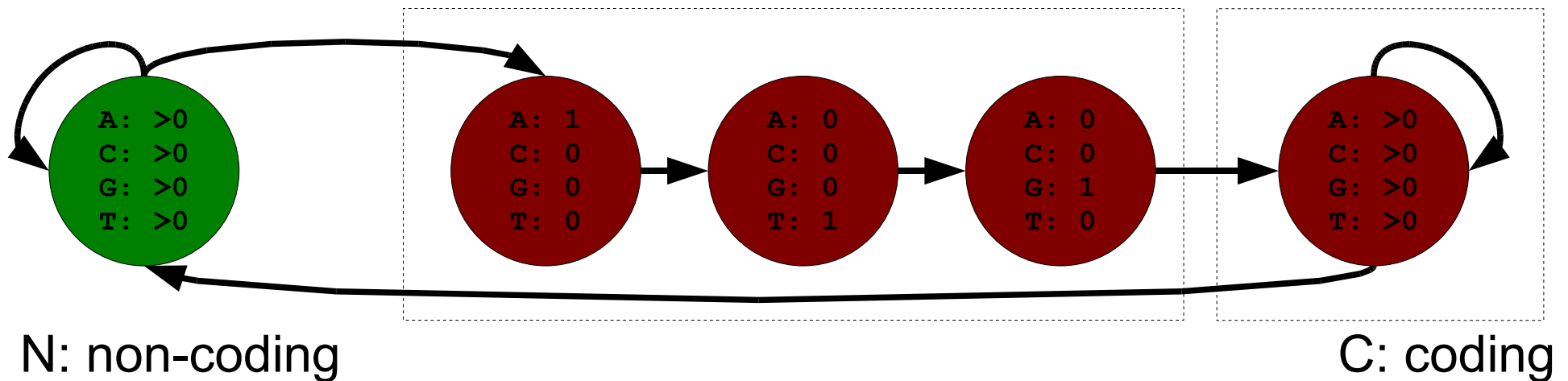
HMM based gene finding

Biological facts

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

Z: NNNCCCCCCCCNNNNNNNNCCCCCCCCCCCCNNNNNNNNNN

X: acgatgcgctaataatgtccgatgacgtgagcataagcgacatgcag

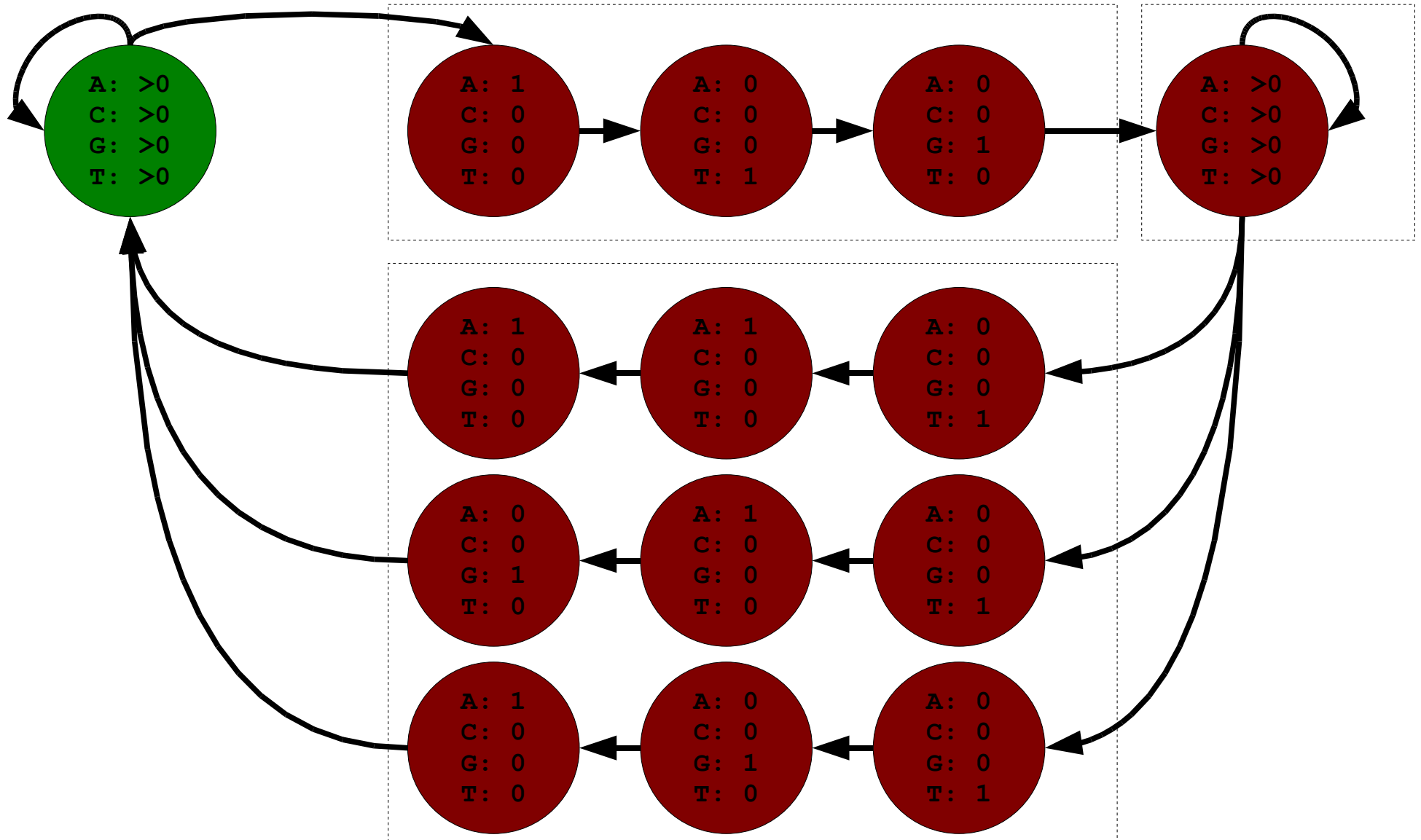


HMM bas

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

N: non-coding

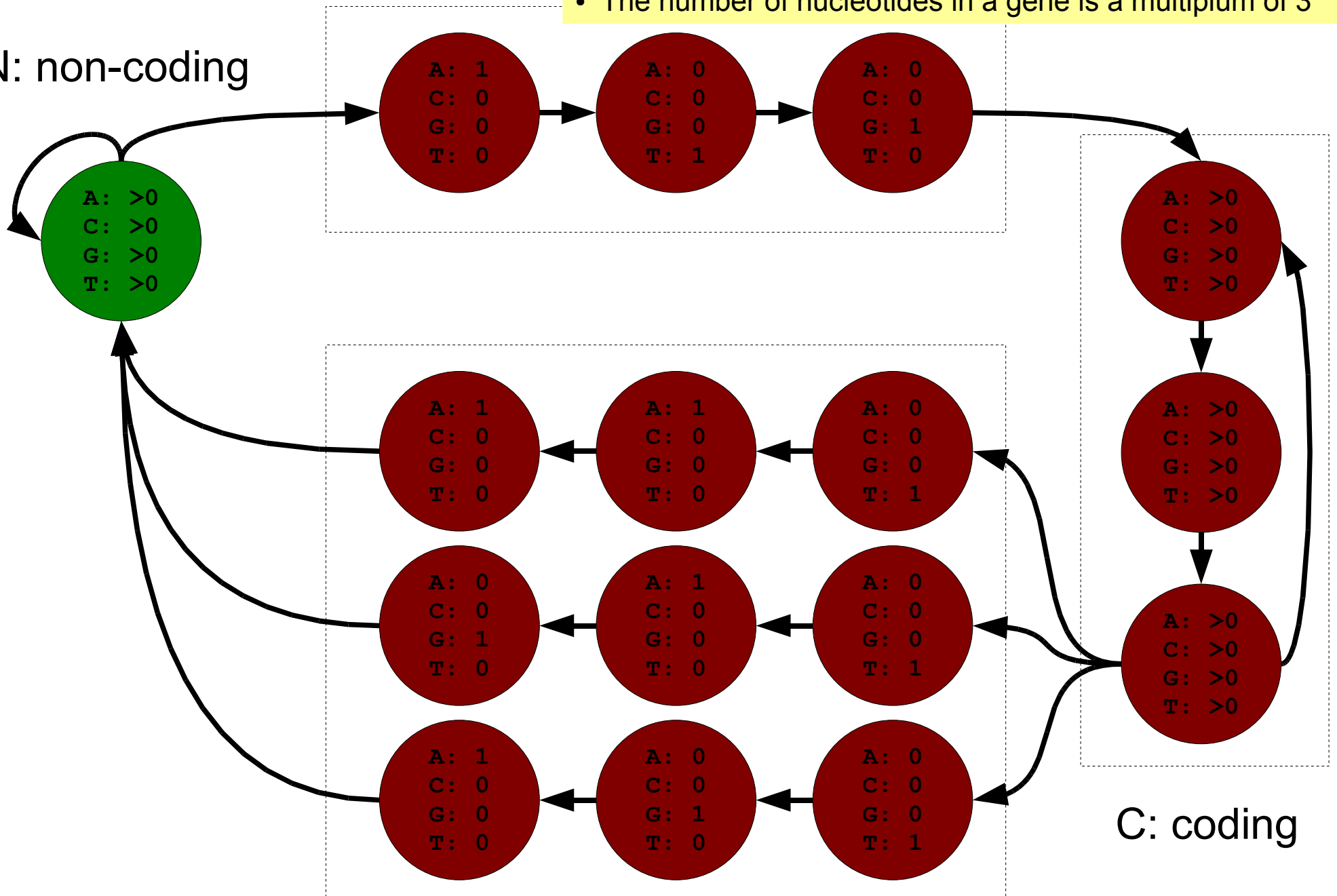
C: coding



HMM bas

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

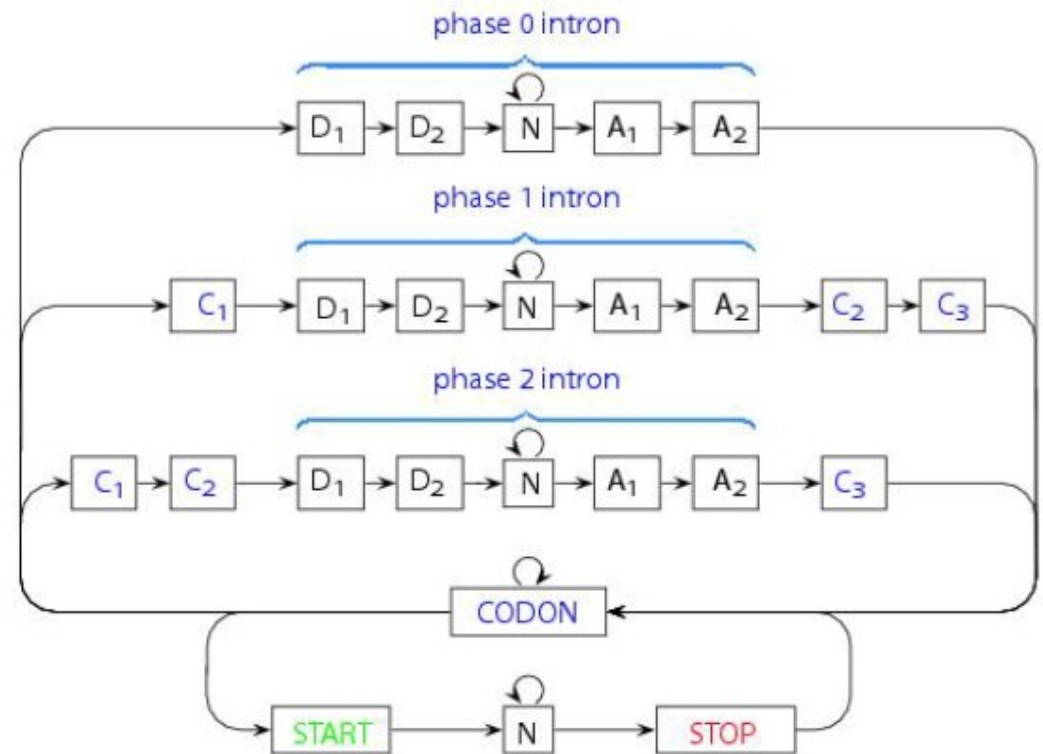


C: coding

HMM based gene finding

Hidden states correspond to functional sites and regions.

| | |
|--------|-----------------------|
| N: | intergenic or intron |
| START: | start codon |
| CODON: | inner codon |
| STOP: | stop codon |
| C: | single codon position |
| D: | splice donor |
| A: | splice acceptor |



HMM modeling eukaryotic gene structure.

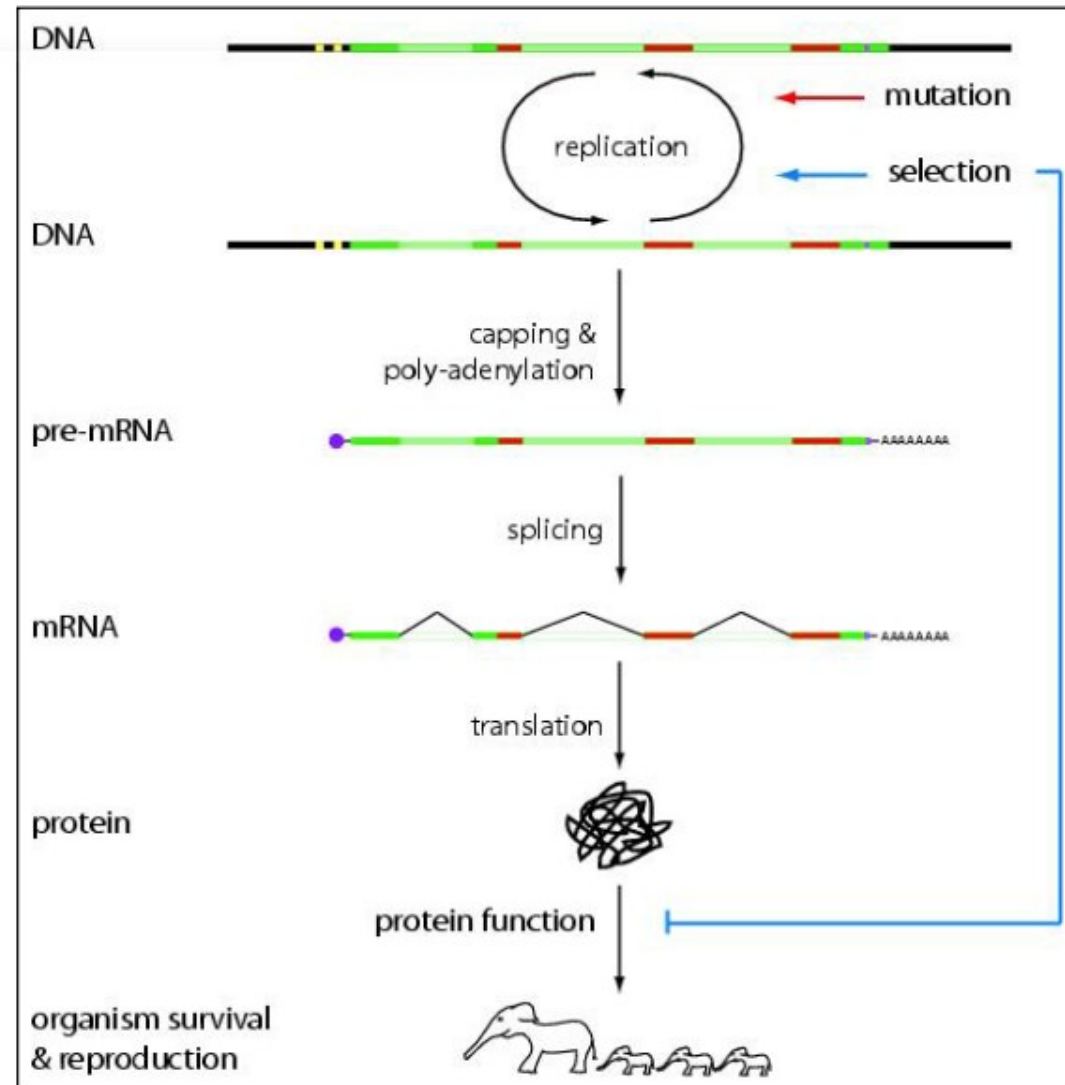
Possible observation sequence:

X=TGAAC ATG CAT TTA ACG GGG CTC CAT ... ACG TGA GCCGAGATCATG

Comparative gene finding

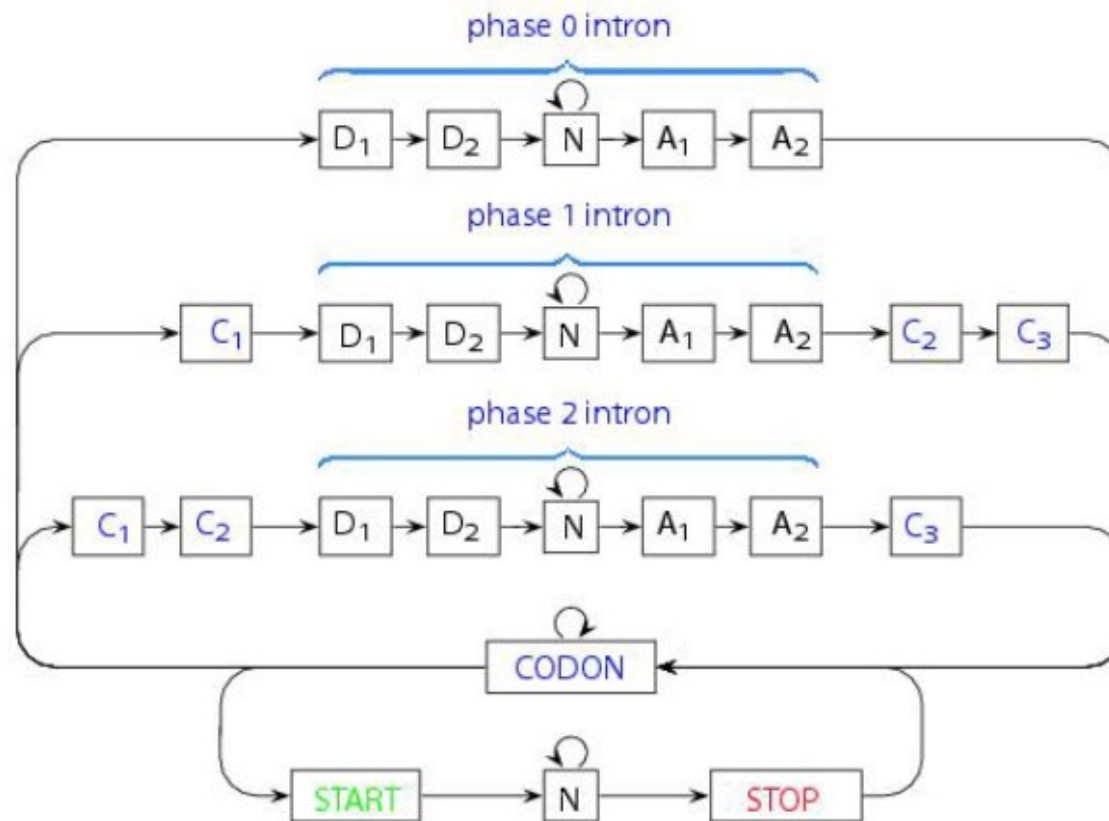
The pattern of substitutions and indels will carry information on the function of the region:

- Coding regions will evolve relatively slow
- Coding regions will only have indels of length three
- Third codon positions will evolve relatively fast
- Introns and intergenic regions will evolve relatively fast
- Functional sites will evolve slowly with very few indels



Comparative HMM gene finding

EHMM architecture

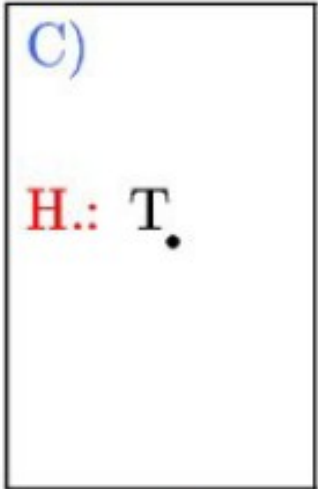
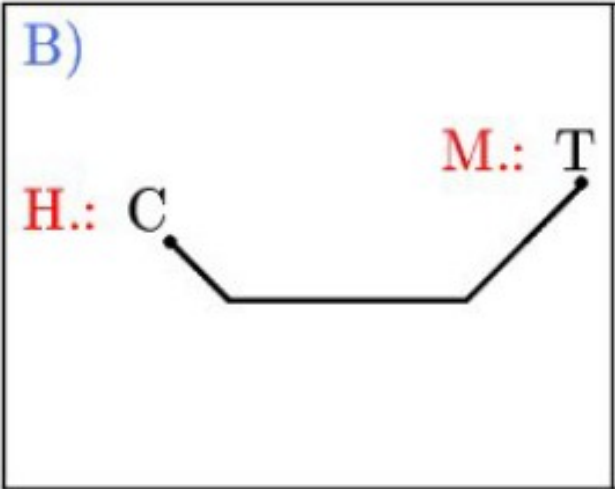
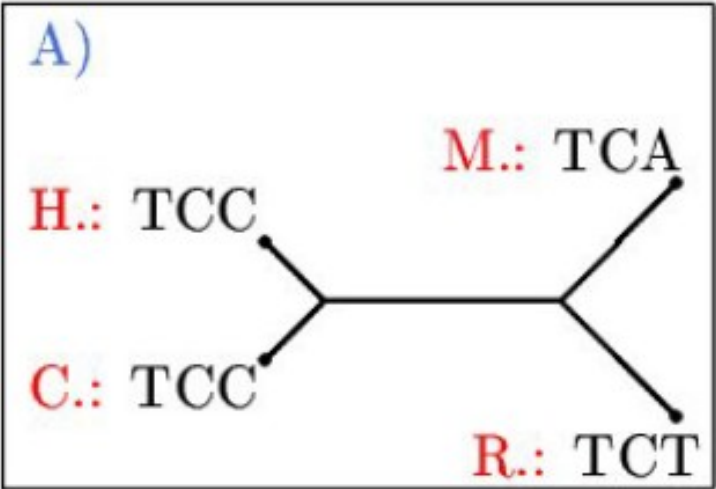


Example of generated alignment with color annotation:

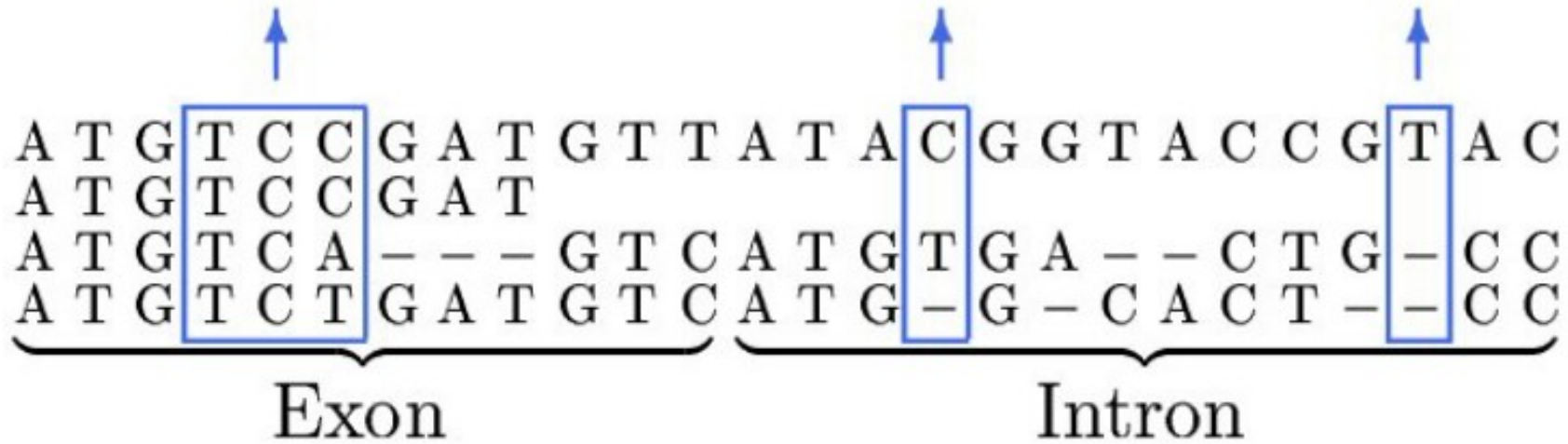
```

TGGAC ATG CAT TTA ACG GGG CTC CAT ... ACA TGA GCAAT
x = TCGAC ATG CAC TTA ACA GGG CTC CAC ... ACA TGA ACAGT
CGGGC ATG CAT TGA ACA GGA CTT CAC ... ACA TGA GCAAT
  
```

Comparative HMM gene finding



Human:
Chimp.:
Mouse:
Rat:



Quality of gene finders

| | Nucleotide No genes | SN | SP |
|------------------------|------------------------|------|------|
| Twinscan | 7 | 0.90 | 0.95 |
| GenomeScan | 43 | 0.88 | 0.83 |
| GlimmerHMM | 9 | 0.89 | 0.79 |
| Augustus | 0 | 0.81 | 0.78 |
| GeneZilla | 0 | 0.70 | 0.67 |
| SNAP (<i>H.sap</i>) | 9 | 0.72 | 0.71 |
| SNAP (<i>A.thal</i>) | 7 | 0.47 | 0.22 |

Using existing knowledge of genes

Just finding sequences need not be the whole story.

Filtering candidates through databases of known genes, for example, can improve the quality of gene finders (but only for finding “known” genes).

“Profiles”

An aligned sequence family or “region of interest”

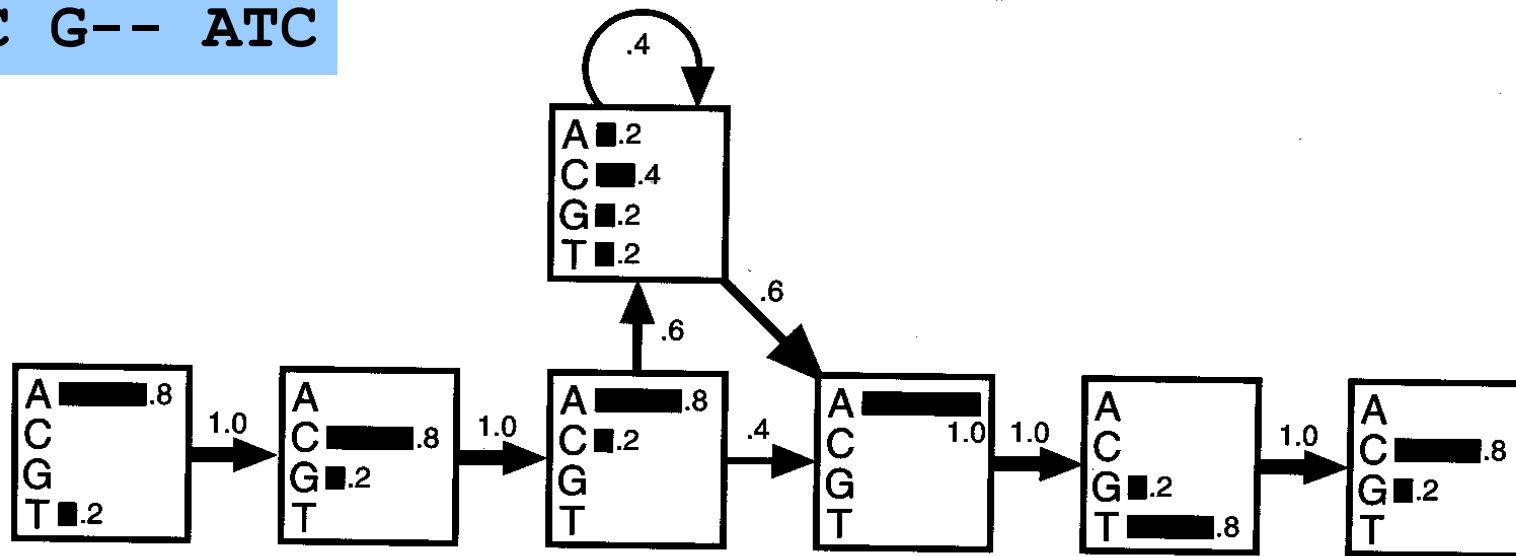
```
... ACA --- ATG ...  
... TCA ACT ATC ...  
... ACA C-- AGC ...  
... AGA --- ATC ...  
... ACC G-- ATC ...
```

A “classic” **profile** summarizing the sequence family

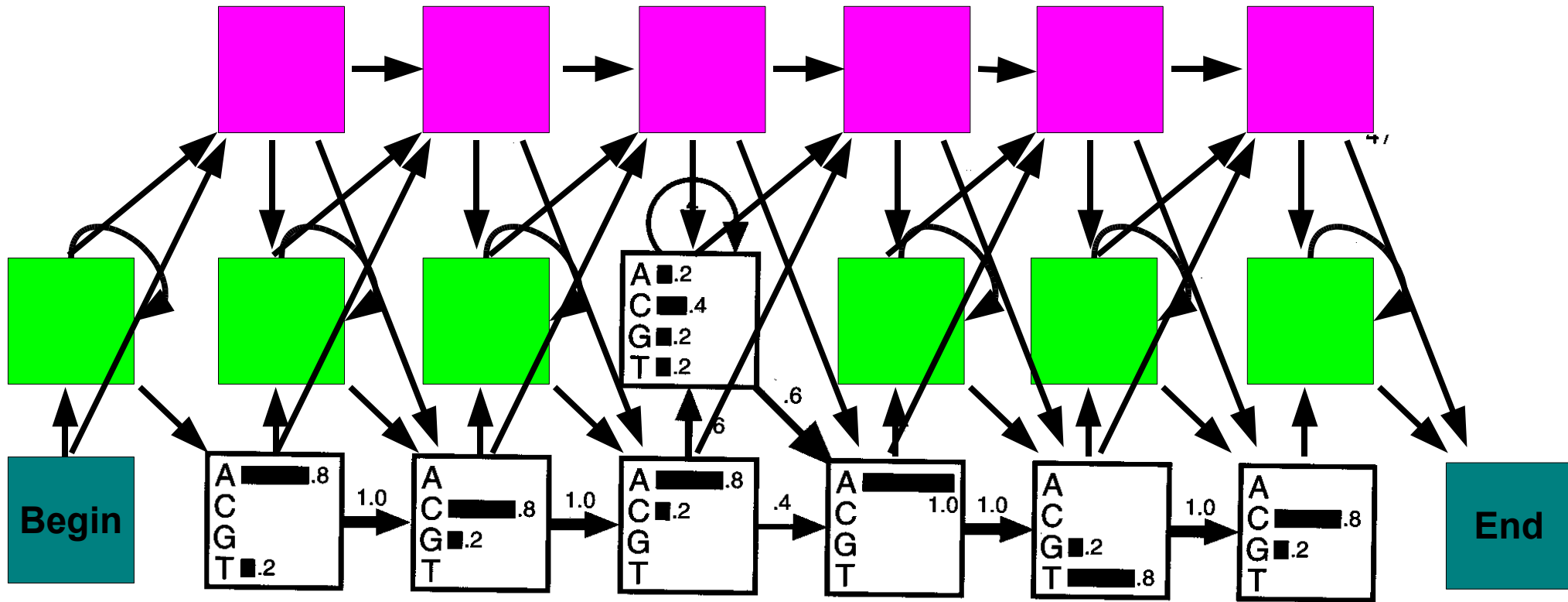
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| A: | 0.8 | . | 0.8 | 0.2 | . | . | 1.0 | . | . |
| C: | . | 0.8 | 0.2 | 0.2 | 0.2 | . | . | . | 0.8 |
| G: | . | 0.2 | . | 0.2 | . | . | . | 0.2 | 0.2 |
| T: | 0.2 | . | . | . | . | 0.2 | . | 0.8 | . |
| -: | . | . | . | 0.4 | 0.8 | 0.8 | . | . | . |
| | A | C | A | - | - | - | A | T | C |

Profile HMMs

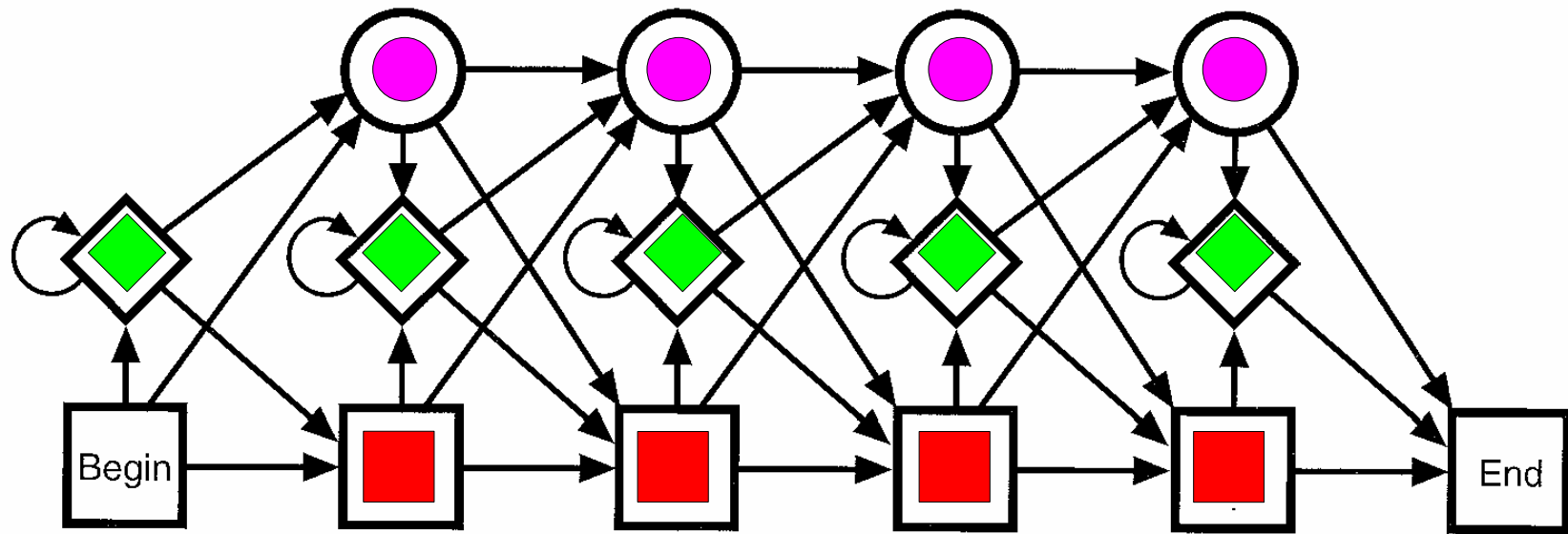
ACA --- ATG
 TCA ACT ATC
 ACA C-- AGC
 AGA --- ATC
 ACC G-- ATC



Profile HMMs



Profile HMMs

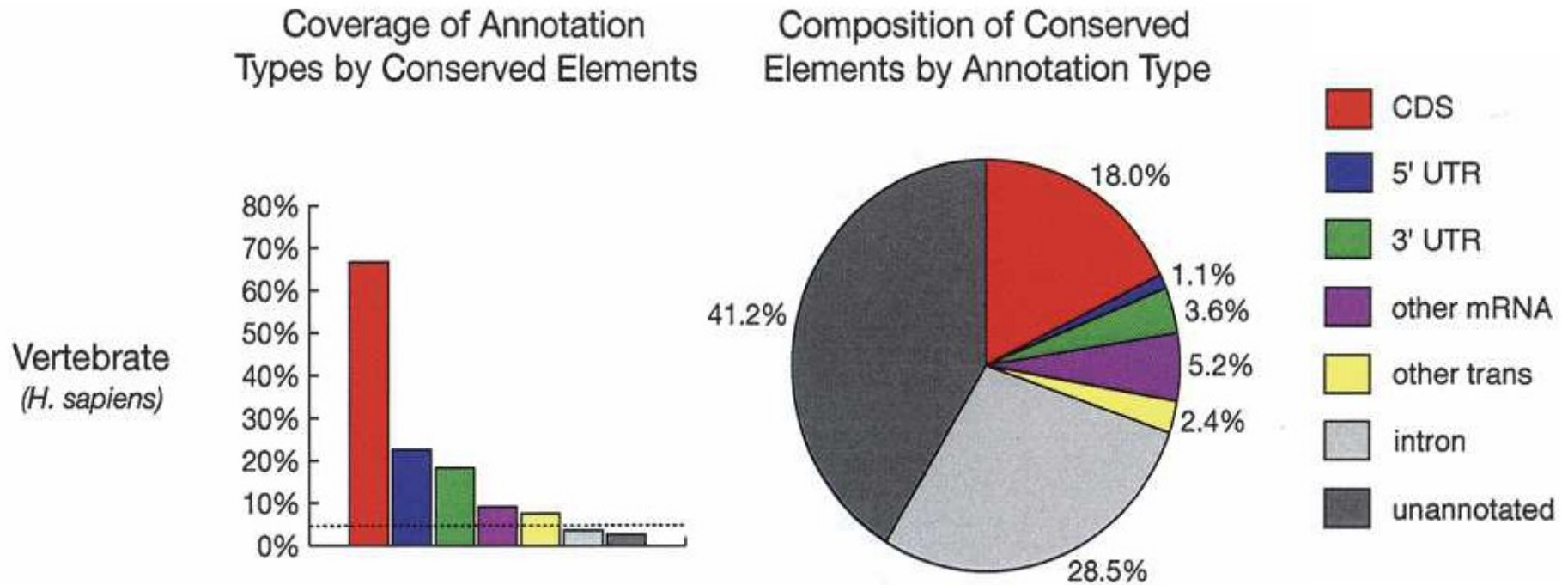


Consists of **Match**-, **Insert**-, and **Delete**-states. A run generates a sequence (DNA or protein). The hidden path of states explains how the generated sequence relates to the sequence family ...

Beyond protein coding genes

There seem to be more conserved sequence than coding sequence in genomes.

Beyond protein coding genes



Beyond protein coding genes

There seem to be more conserved sequence than coding sequence in genomes.

If we believe **conservation** ← **purifying selection** ← **function** then there is much non-coding but functionally important sequence to be found and figured out.

Beyond protein coding genes

There seem to be more conserved sequence than coding sequence in genomes.

If we believe **conservation** ← **purifying selection** ← **function** then there is much non-coding but functionally important sequence to be found and figured out.

...regulatory sequences, splicing machinery, RNA genes...

RNA genes

Stochastic Context-Free Grammars (SCFGs) – a formalism similar to HMMs – can be used for recognizing RNA structure.

RNA genes

Stochastic Context-Free Grammars (SCFGs) – a formalism similar to HMMs – can be used for recognizing RNA structure.

Not quite as computationally efficient, but can analyse genome sequences in sliding windows.

Simple example of CFG

$S \rightarrow aSu \mid uSa \mid gSc \mid cSc \mid \dots$
| SS
| $aS \mid uS \mid cS \mid gS$
| $""$

$S \rightarrow a S u$

$\rightarrow ac S gu$

$\rightarrow acg S cgu$

$\rightarrow acgg S ccgu$

$\rightarrow acgga S ccgu \rightarrow acggag S ccgu$

$\rightarrow acggagu S ccgu \rightarrow acggagug S ccgu$

$\rightarrow acggagugc S ccgu \rightarrow acggagugc "" ccgu$

ACGG AG
UGCC U
CG

Simple example of CFG

$$\begin{array}{l}
 S \rightarrow aSu \mid uSa \mid gSc \mid cSc \mid \dots \\
 \quad \mid SS \\
 \quad \mid aS \mid uS \mid cS \mid gS \\
 \quad \mid ""
 \end{array}$$

Adding probabilities to the rules gives us *stochastic* CFGs.

→ These can be used for annotations (similar to HMMs).

→ acg S ccgu

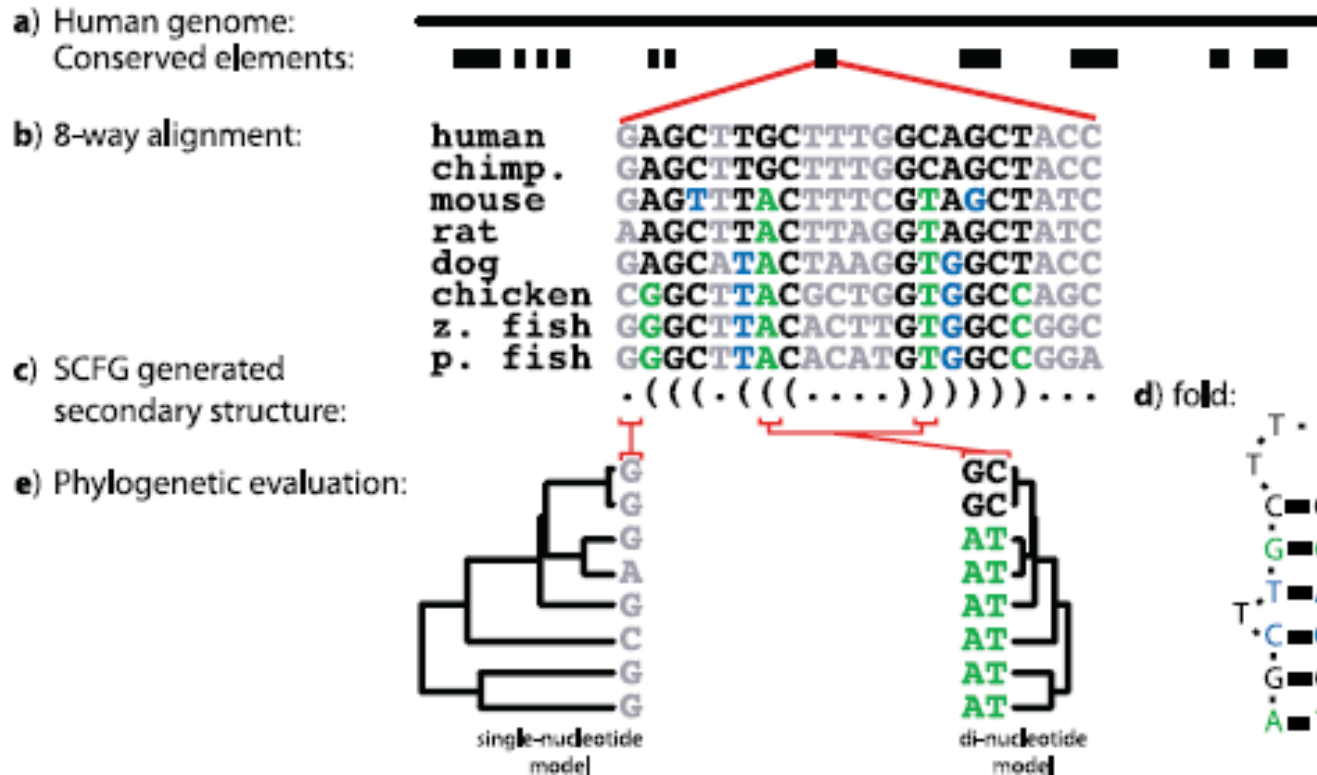
→ acgg S ccgu

→ acgga S ccgu → acggag S ccgu

→ acggagu S ccgu → acggagug S ccgu

→ acggagugc S ccgu → acggagugc "" ccgu

Comparative RNA gene finding



Summary

Using stochastic methods (HMMs or SCFGs) we can model our biological knowledge to extract sequence and evolutionary signal to automatically annotate genomes.

- We have seen hidden Markov models for annotating protein coding genes
- and stochastic context free grammars to annotate for RNA genes