Genome 540 Class 20

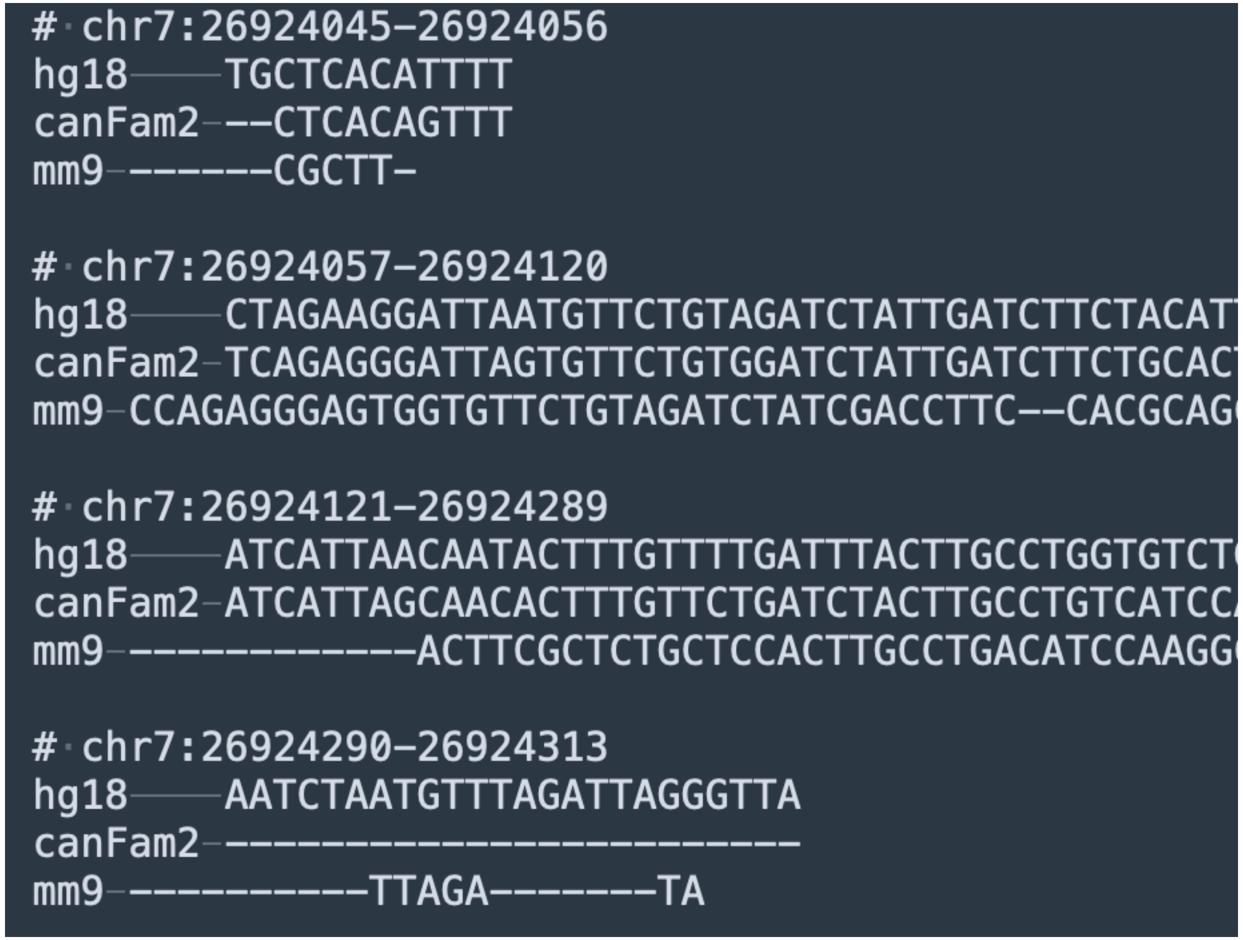
Chengxiang Qiu

HW8 questions?

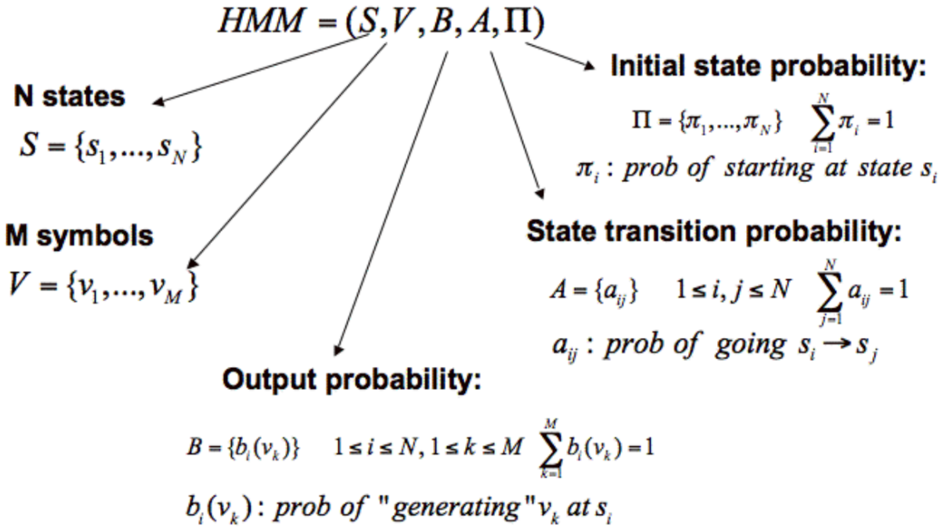
HW9: Evolutionarily conserved segments Due Sunday March-13 11:59pm

- ENCODE region 010 (chromosome 7)
- Multiple alignment of human, dog, and mouse
- 2 states: neutral (fast-evolving), conserved (slow-evolving)
- Emitted symbols are multiple alignment columns (e.g. ‘AAT’)
- Viterbi parse (no iteration)

Input data



```
# chr7:26924045-26924056
hg18    TGCTCACATTTT
canFam2 --CTCACAGTTT
mm9 ------CGCTT-

# chr7:26924057-26924120
hg18    CTAGAAGGATTAATGTTCTGTAGATCTATTGATCTTCTACAT
canFam2 TCAGAGGGATTAGTGTTCTGTGGATCTATTGATCTTCTGCAC
mm9 CCAGAGGGAGTGGTGTTCTGTAGATCTATCGACCTTC--CACGCAG

# chr7:26924121-26924289
hg18    ATCATTAACAATACTTTGTTTTGATTTACTTGCCTGGTGTCT
canFam2 ATCATTAGCAACACTTTGTTCTGATCTACTTGCCTGTCATCC
mm9 ------------ACTTCGCTCTGCTCCACTTGCCTGACATCCAAGG

# chr7:26924290-26924313
hg18    AATCTAATGTTTAGATTAGGGTTA
canFam2 ------------------------
mm9 ----------TTAGA-------TA
```



$HMM = (S, V, B, A, \Pi)$

N states

$S = \{s_1, \ldots, s_N\}$

M symbols

$V = \{v_1, \ldots, v_M\}$

Initial state probability:

$\Pi = \{\pi_1, \ldots, \pi_N\} \quad \sum_{i=1}^{N} \pi_i = 1$

π_i : *prob of starting at state* s_i

State transition probability:

$A = \{a_{ij}\} \quad 1 \le i, j \le N \quad \sum_{j=1}^{N} a_{ij} = 1$

a_{ij} : *prob of going* $s_i \rightarrow s_j$

Output probability:

$B = \{b_i(v_k)\} \quad 1 \le i \le N, 1 \le k \le M \quad \sum_{k=1}^{M} b_i(v_k) = 1$

$b_i(v_k)$: *prob of "generating"* v_k *at* s_i

N = 2 states

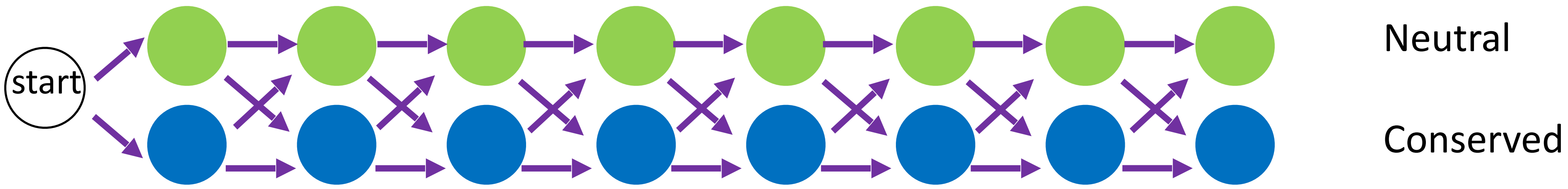
M = 100 symbols

Finding the most likely series of hidden states (Viterbi Path)

- Step 1: given an observed alignment, determine the most probable series of states
  - ○ This depends on the specified probabilities:
    - ▪ Initiation
    - ▪ Transition
    - ▪ Emission
  - ○ Process nodes in a sliding window

Observation: Alignment

```
A  T  T  C  A  G  C  A
C  T  -  C  A  G  C  A
C  G  -  C  A  G  C  A
```



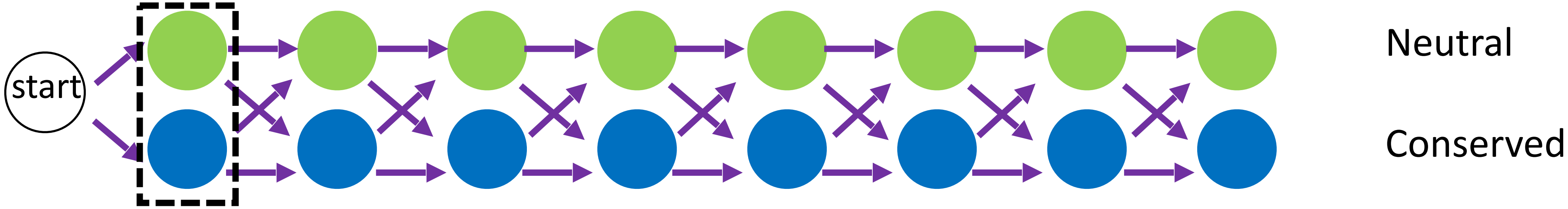
Finding the most likely series of hidden states (Viterbi Path)

- Step 1: given an observed alignment, determine the most probable series of states
  - This depends on the specified probabilities:
    - Initiation
    - Transition
    - Emission
  - Process nodes in a sliding window

Initial Prob * Prob Emitting (ACC)

Observation: Alignment

```
A   T   T   C   A   G   C   A
C   T   -   C   A   G   C   A
C   G   -   C   A   G   C   A
```

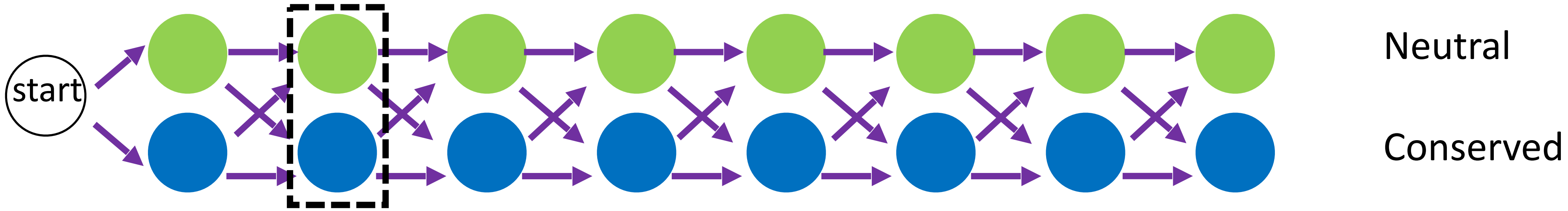


Finding the most likely series of hidden states (Viterbi Path)

- Step 1: given an observed alignment, determine the most probable series of states
  - This depends on the specified probabilities:
    - Initiation
    - Transition
    - Emission
  - Process nodes in a sliding window

Previous Prob * Transition Prob * Prob Emitting (TTG)

Observation: Alignment	A	T	T	C	A	G	C	A
	C	T	-	C	A	G	C	A
	C	G	-	C	A	G	C	A



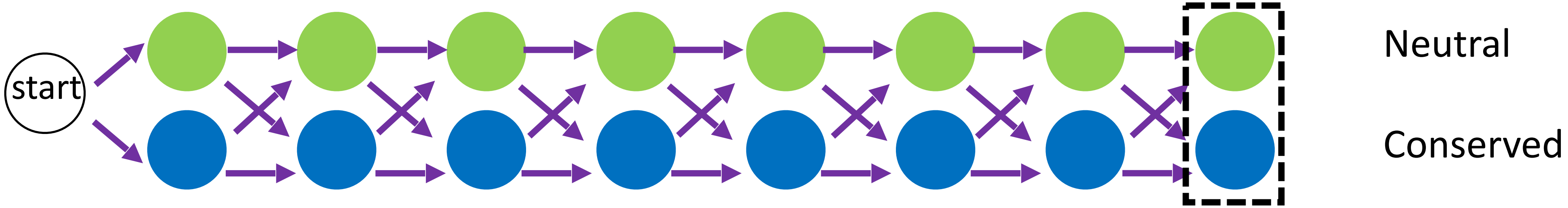
Finding the most likely series of hidden states (Viterbi Path)

- Step 1: given an observed alignment, determine the most probable series of states
  - This depends on the specified probabilities:
    - Initiation
    - Transition
    - Emission
  - Process nodes in a sliding window

Previous Prob * Transition Prob * Prob Emitting (AAA)

Observation: Alignment

```
A   T   T   C   A   G   C   A
C   T   -   C   A   G   C   A
C   G   -   C   A   G   C   A
```

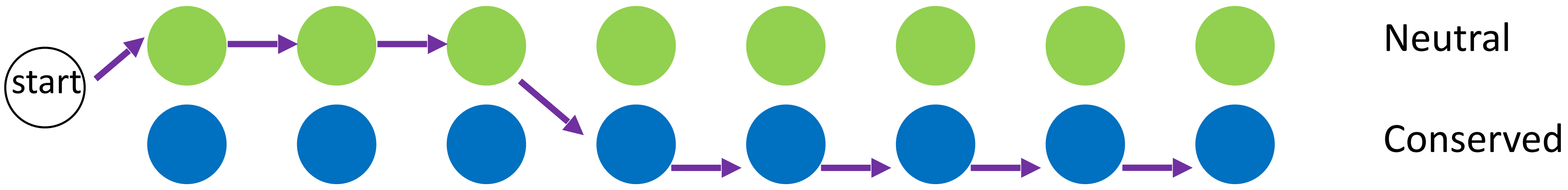


Finding the most likely series of hidden states (Viterbi Path)

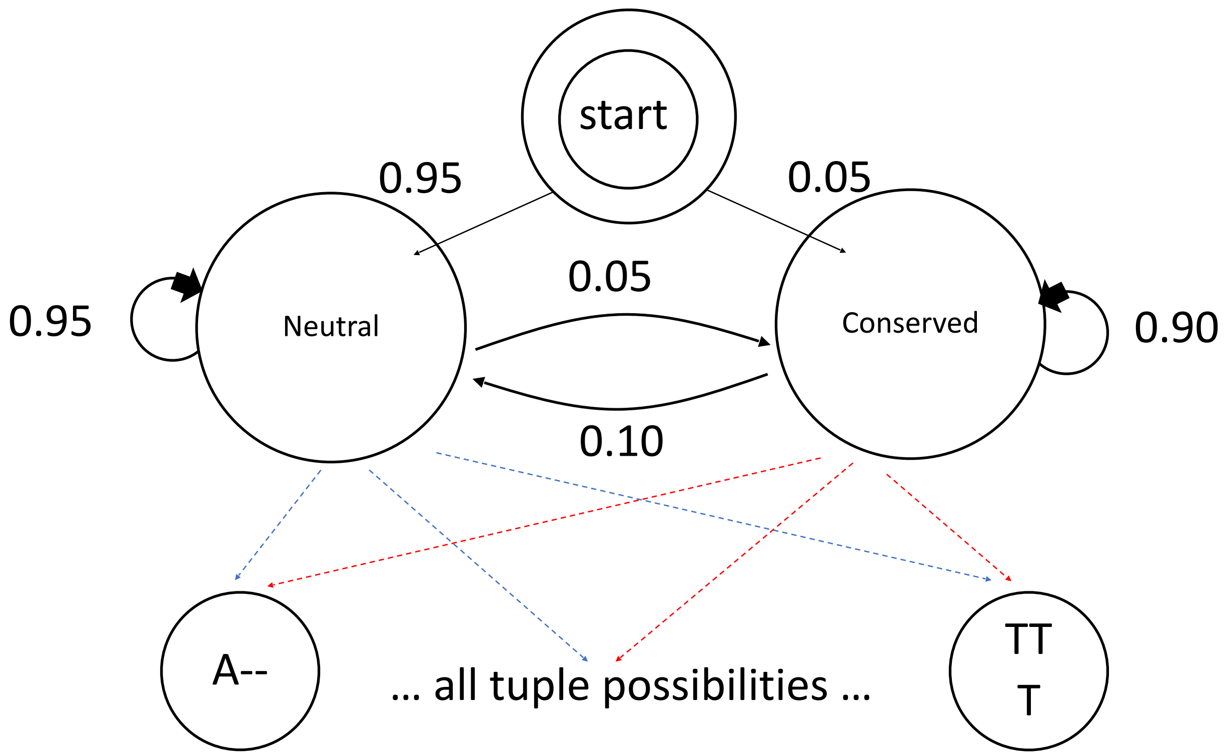
- Step 1: given an observed alignment, determine the most probable series of states
  - This depends on the specified probabilities:
    - Initiation
    - Transition
    - Emission
  - Process nodes in a sliding window

Observation: Alignment

```
A   T   T   C   A   G   C   A
C   T   -   C   A   G   C   A
C   G   -   C   A   G   C   A
```

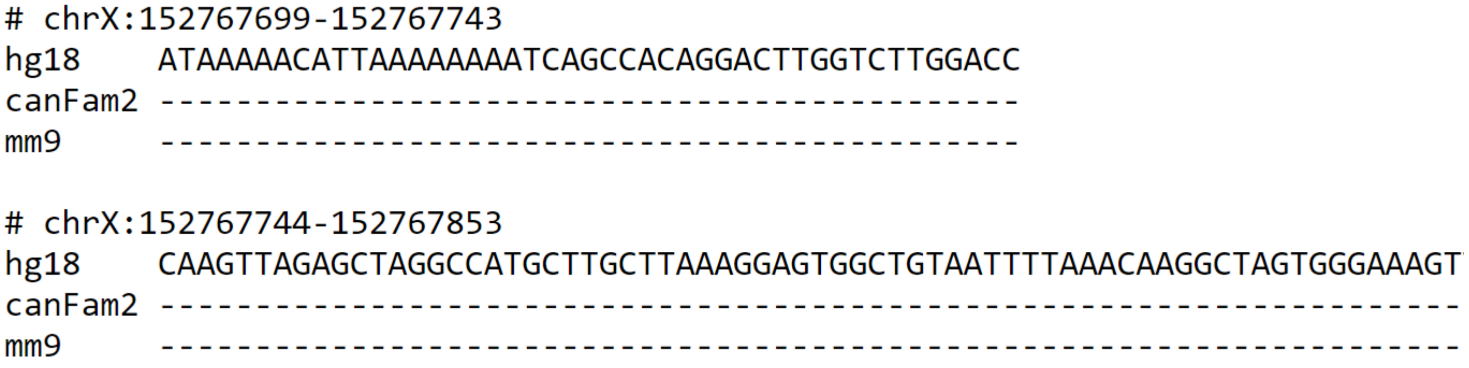


HMM Diagram


start
0.95
0.05
0.95
Neutral
0.05
Conserved
0.90
0.10
A--
... all tuple possibilities ...
TT
T

Input

- Original maf format
  - Sequences broken into alignment blocks based on the species included
  - Official file format specs
- Homework file format
  - Only 3 species
  - Gaps in human sequence were removed and ambiguous bases replaced with ‘A’ for simplicity



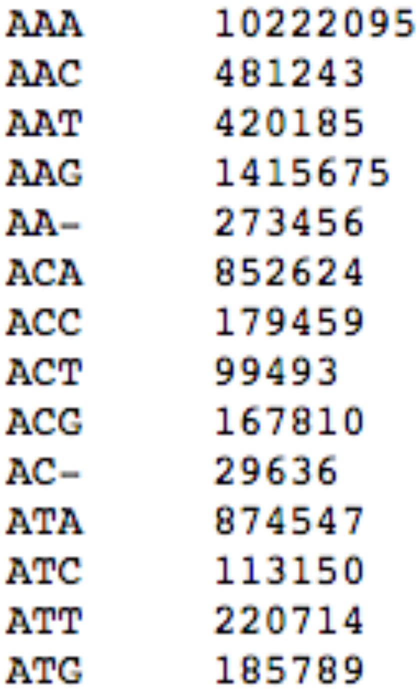
```
# chrX:152767699-152767743
hg18    ATAAAAACATTAAAAAAAATCAGCCACAGGACTTGGTCTTGGACC
canFam2 ---------------------------------------------
mm9     ---------------------------------------------

# chrX:152767744-152767853
hg18    CAAGTTAGAGCTAGGCCATGCTTGCTTAAAGGAGTGGCTGTAATTTTAAACAAGGCTAGTGGGAAAGT
canFam2 --------------------------------------------------------------------
mm9     --------------------------------------------------------------------
```

Setting parameters

- Emission probabilities
  - Neutral state: observed frequencies in neutral data set
  - Conserved state: observed frequencies in functional data set
- Transition probabilities
  - Given in the assignment; more likely to go from conserved to neutral
- Initiation probabilities
  - Given in the assignment; more likely to start in the neutral state

Calculating Emission Probabilities

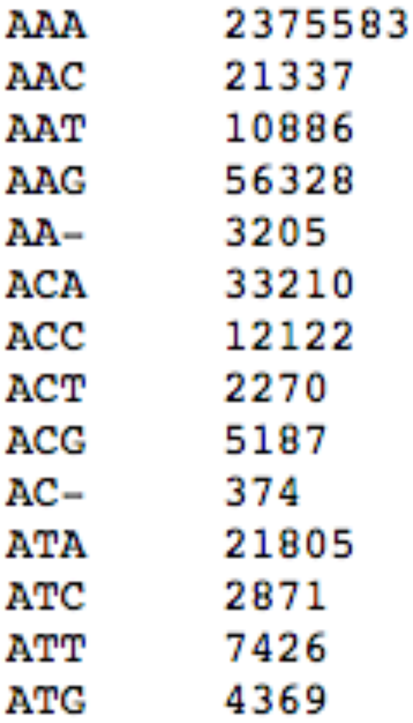
Neutral State: Ancient Repeat Sequences



AAA	10222095
AAC	481243
AAT	420185
AAG	1415675
AA-	273456
ACA	852624
ACC	179459
ACT	99493
ACG	167810
AC-	29636
ATA	874547
ATC	113150
ATT	220714
ATG	185789

etc …

Conserved State: Putative Functional Sites



AAA	2375583
AAC	21337
AAT	10886
AAG	56328
AA-	3205
ACA	33210
ACC	12122
ACT	2270
ACG	5187
AC-	374
ATA	21805
ATC	2871
ATT	7426
ATG	4369

etc …

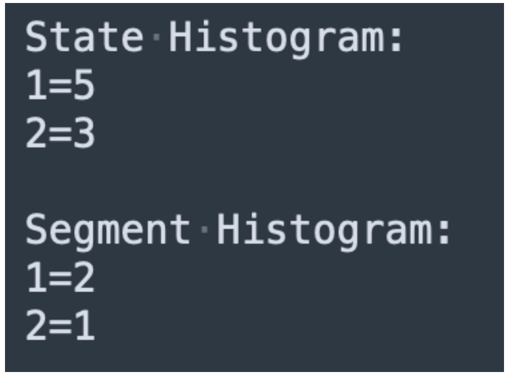
1st base: human

2nd base: dog

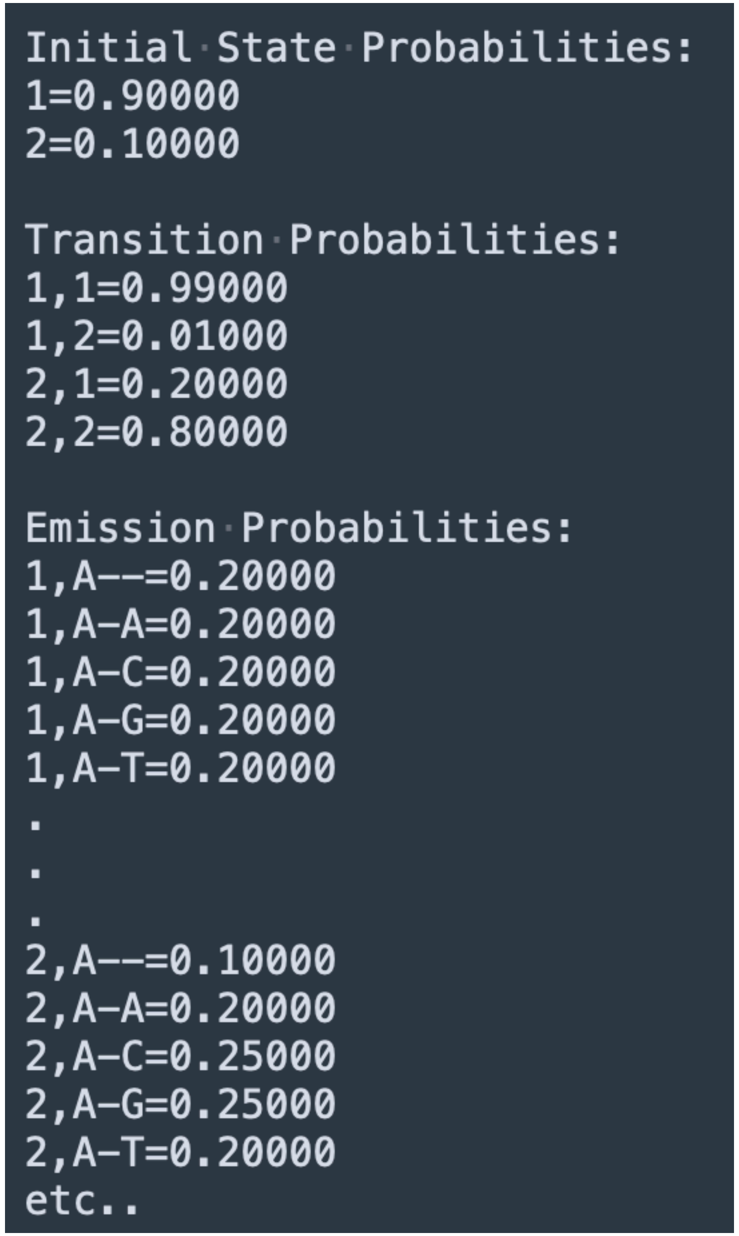
3rd base: mouse

Output

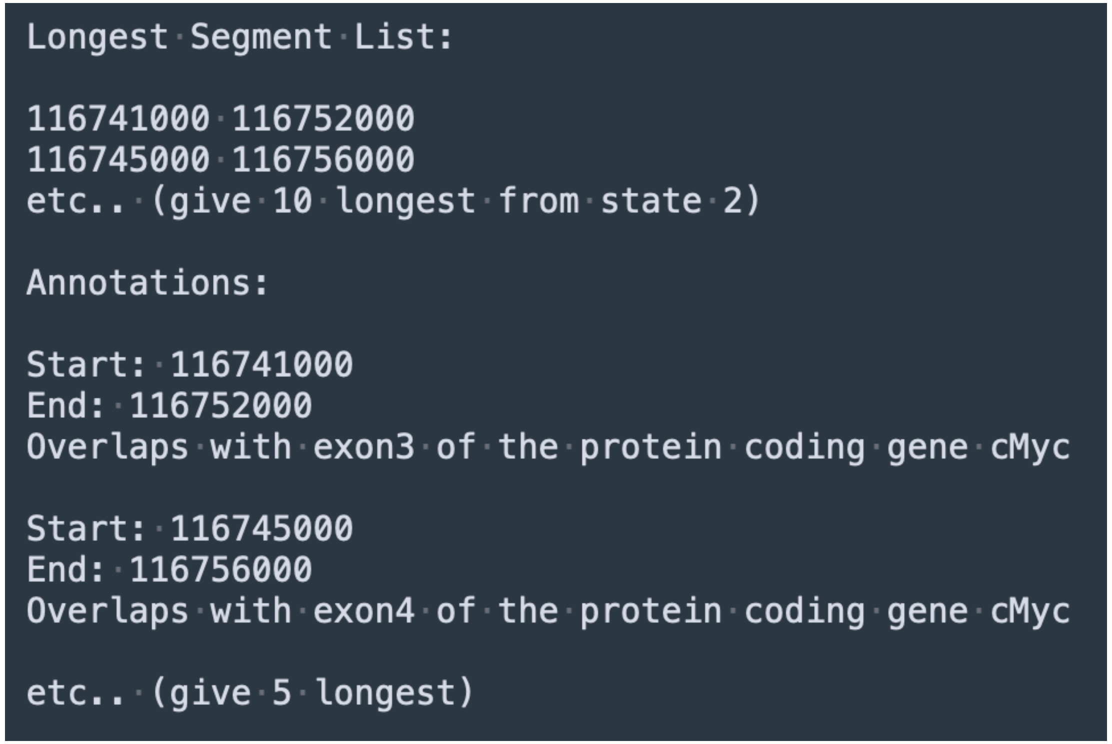
- State and segment histograms
- Parameter values
    - Initiation/transition probabilities you were given in the assignment
    - Emission probabilities you calculated from neutral and conserved data sets
- Coordinates of 10 longest conserved segments (report positions relative to the start of the chromosome)
- Brief annotations for the 5 longest conserved segments (look at UCSC genome browser, and make sure using the correct genome version, e.g. hg18)



```
State Histogram:
1=5
2=3

Segment Histogram:
1=2
2=1
```



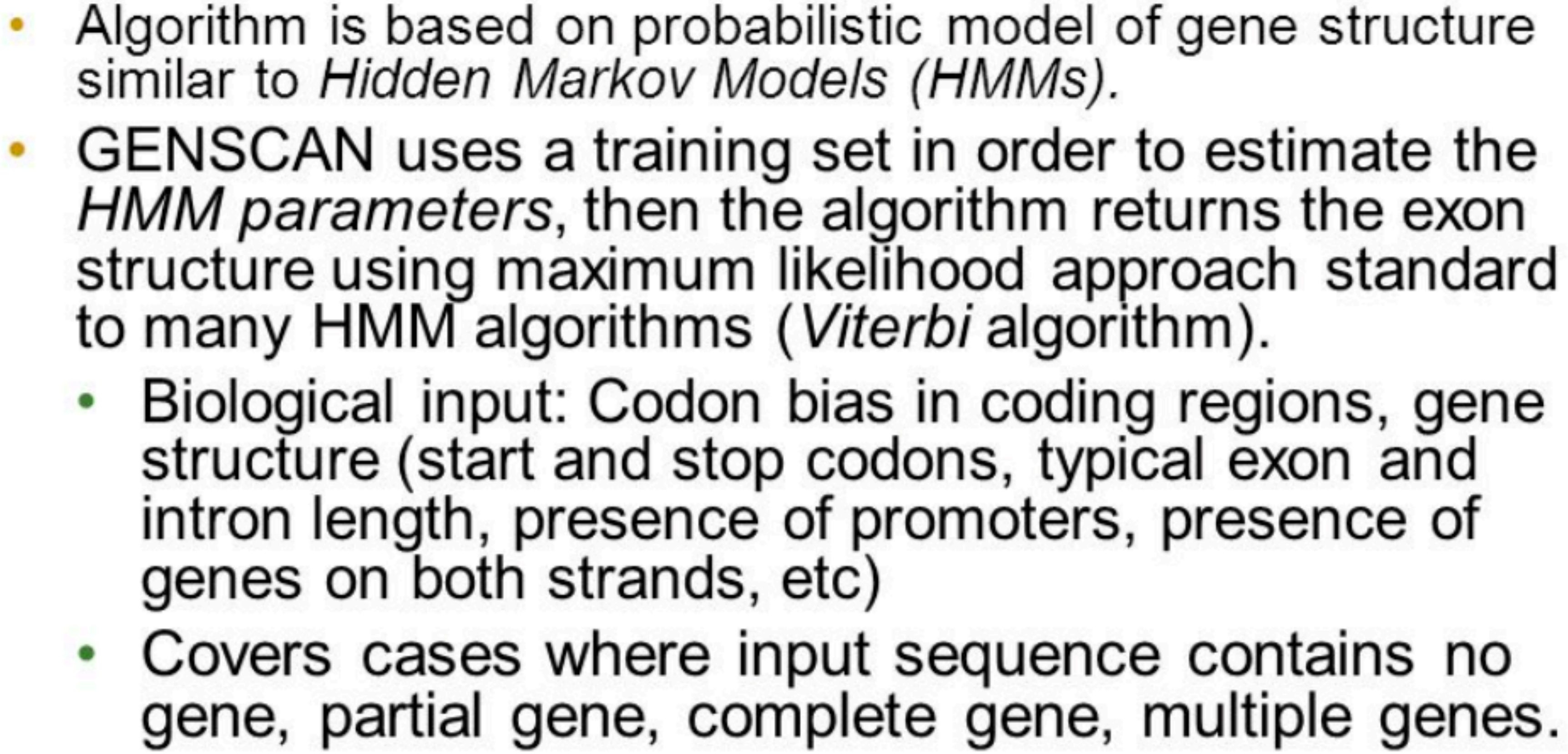
```
Initial State Probabilities:
1=0.90000
2=0.10000

Transition Probabilities:
1,1=0.99000
1,2=0.01000
2,1=0.20000
2,2=0.80000

Emission Probabilities:
1,A--=0.20000
1,A-A=0.20000
1,A-C=0.20000
1,A-G=0.20000
1,A-T=0.20000
.
.
.
2,A--=0.10000
2,A-A=0.20000
2,A-C=0.25000
2,A-G=0.25000
2,A-T=0.20000
etc..
```



```
Longest Segment List:

116741000 116752000
116745000 116756000
etc.. (give 10 longest from state 2)

Annotations:

Start: 116741000
End: 116752000
Overlaps with exon3 of the protein coding gene cMyc

Start: 116745000
End: 116756000
Overlaps with exon4 of the protein coding gene cMyc

etc.. (give 5 longest)
```

Questions?

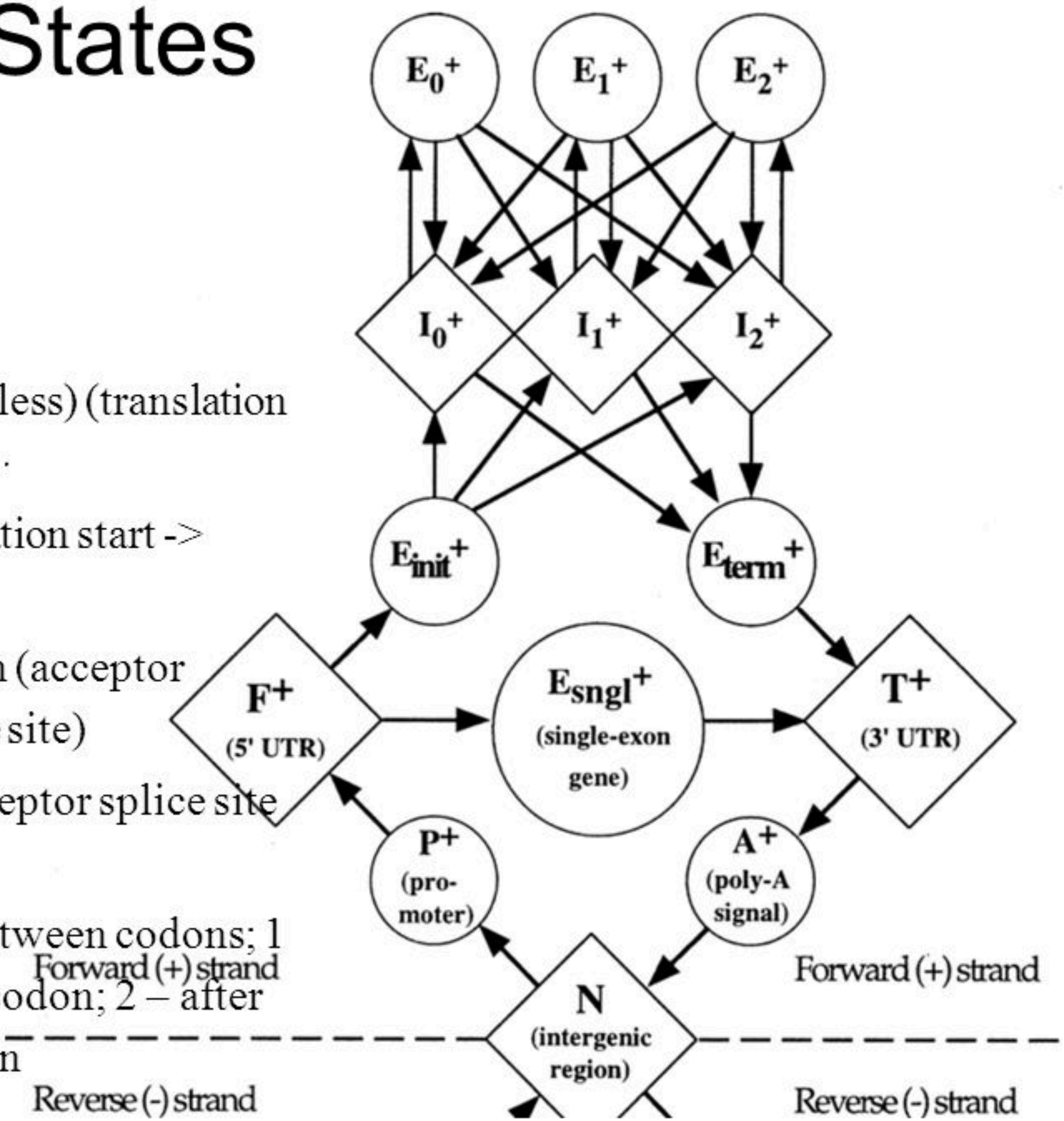
Gene detection: GENSCAN



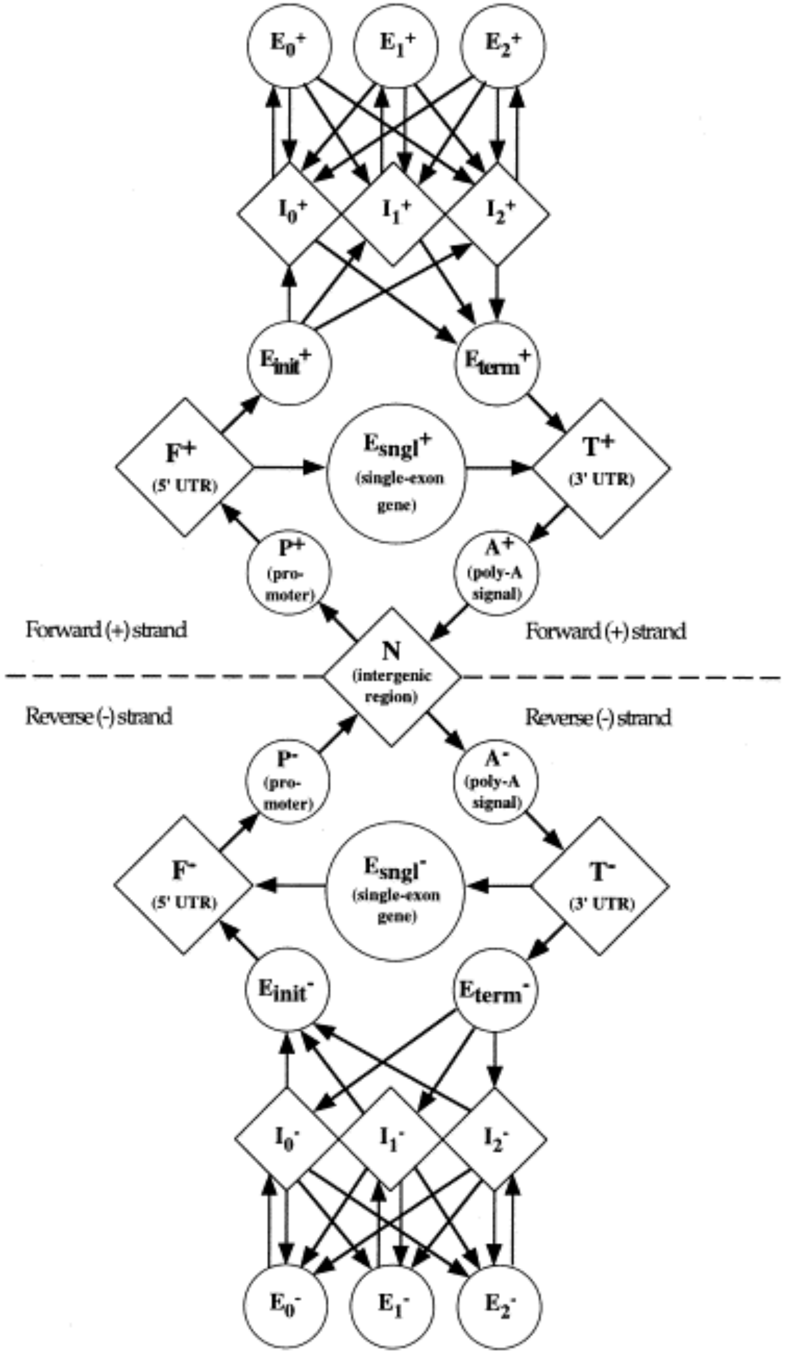
- Algorithm is based on probabilistic model of gene structure similar to *Hidden Markov Models (HMMs).*
- GENSCAN uses a training set in order to estimate the *HMM parameters*, then the algorithm returns the exon structure using maximum likelihood approach standard to many HMM algorithms (*Viterbi* algorithm).
  - Biological input: Codon bias in coding regions, gene structure (start and stop codons, typical exon and intron length, presence of promoters, presence of genes on both strands, etc)
  - Covers cases where input sequence contains no gene, partial gene, complete gene, multiple genes.

GenScan States

- N - intergenic region
- P - promoter
- F - 5’ untranslated region
- E_{sngl} – single exon (intronless) (translation start -> stop codon)
- E_{init} – initial exon (translation start -> donor splice site)
- E_k – phase k internal exon (acceptor splice site -> donor splice site)
- E_{term} – terminal exon (acceptor splice site -> stop codon)
- I_k – phase k intron: 0 – between codons; 1 – after the first base of a codon; 2 – after the second base of a codon



GENSCAN HMM Architecture


E0+
E1+
E2+
I0+
I1+
I2+
Einit+
Eterm+
F+ (5' UTR)
Esngl+ (single-exon gene)
T+ (3' UTR)
P+ (pro-moter)
A+ (poly-A signal)
Forward (+) strand
Forward (+) strand
N (intergenic region)
Reverse (-) strand
Reverse (-) strand
P- (pro-moter)
A- (poly-A signal)
F- (5' UTR)
Esngl- (single-exon gene)
T- (3' UTR)
Einit-
Eterm-
I0-
I1-
I2-
E0-
E1-
E2-