Computational Biology Lectures: Hidden Markov Models

Dylan Taylor & Sara Carioscia

Agara Bio Community Lab

Part 1: December 6, 2020

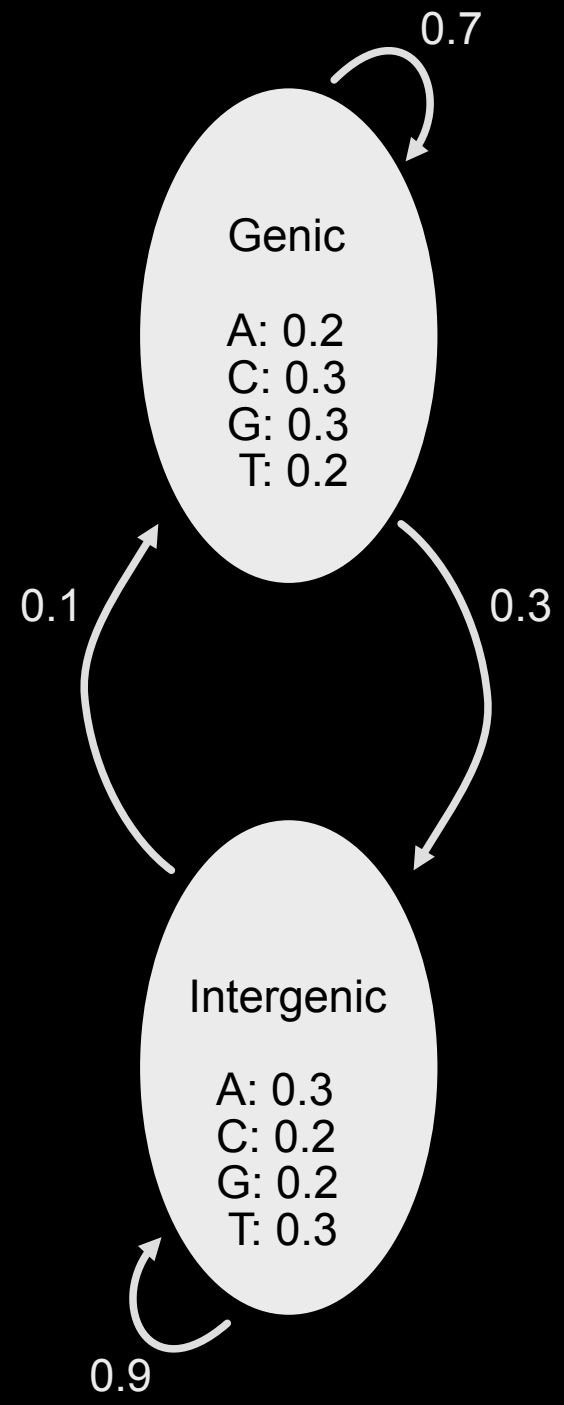
Part 2: February 7, 2020

Why Use an HMM?

- Gene finding
- Sequence alignment
- Inference of parental haplotypes
- Predict protein structure

Genic vs. Intergenic

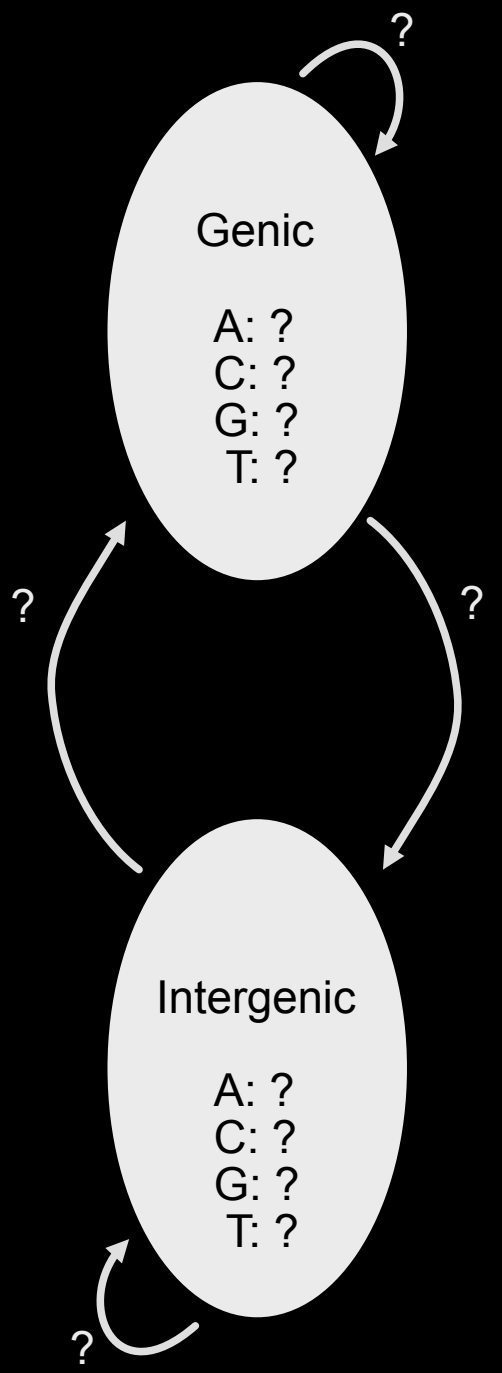
- Any given nucleotide can be in one of two **states**, Genic or Intergenic
- Within each of these states, there is some probability of an A, C, G, or T
- As we move from one nucleotide to the next, there is some probability that we stay in the same state and some probability that we switch to the other



Inferring Probabilities

- Before we move on to decoding the HMM, let's talk about how we actually get these probabilities
- If we know the structure of the HMM and we have some training data with "correct" labels, how can we infer these probabilities?

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCAGGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGGIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIIGIIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

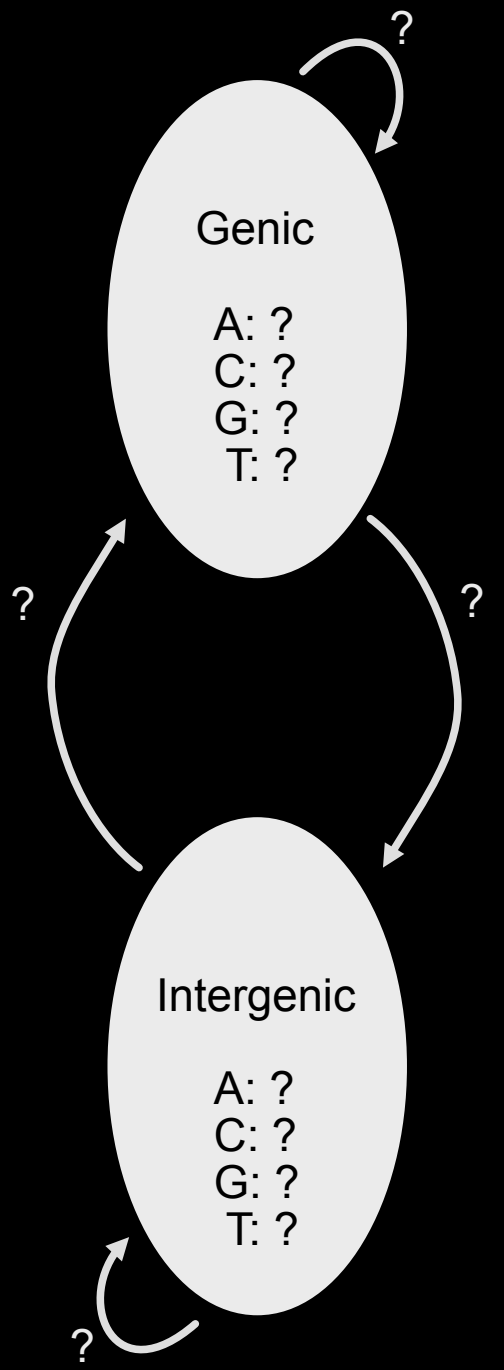


Inferring Probabilities

- Before we move on to decoding the HMM, let's talk about how we actually get these probabilities
- If we know the structure of the HMM and we have some training data with "correct" labels, how can we infer these probabilities?

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCAGGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGGIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIIGIIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Count!



Inferring Probabilities

Training Data

Genic Emission

A: 34 = 0.241
C: 40 = 0.284
G: 38 = 0.270
T: 29 = 0.205

Intergenic Emission

A: 110 = 0.307
C: 73 = 0.203
G: 75 = 0.209
T: 101 = 0.281

Transitions

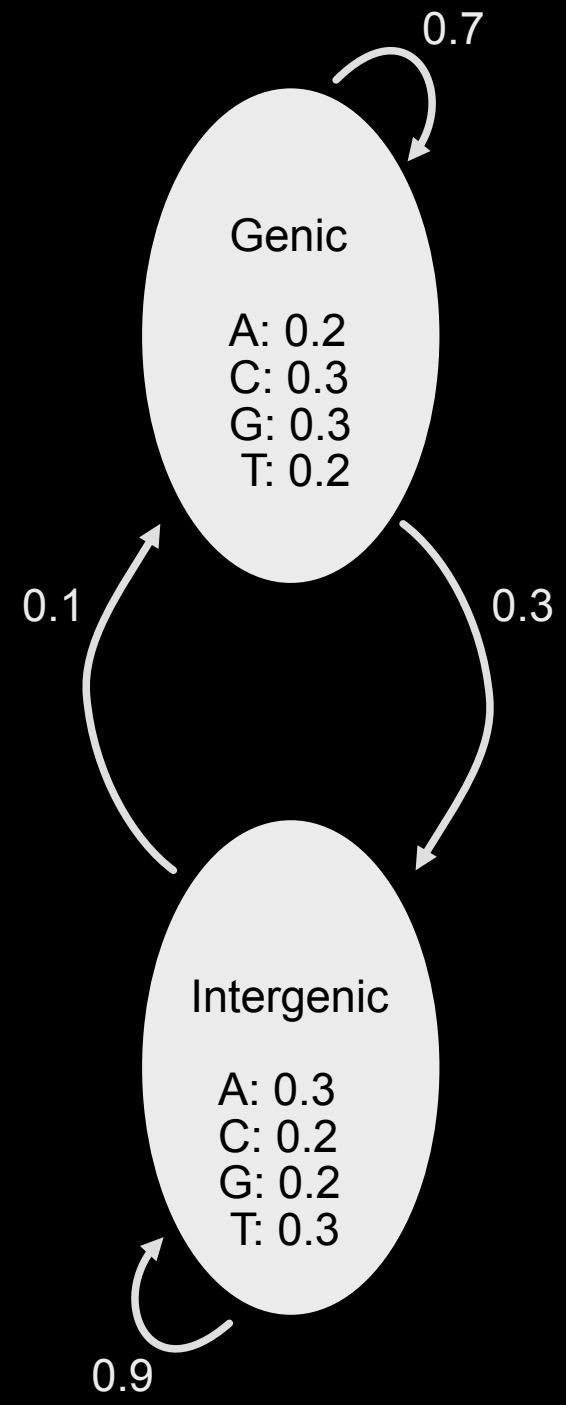
G → G: 97 = 0.703
G → I: 41 = 0.297
I → I: 311 = 0.884
I → G: 41 = 0.116

Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x$$
$$p(C|genic)\ x\ p(genic\ \rightarrow inter)\ x\ p(T|inter)\ \ldots$$

$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ x\ 0.3\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3 =$$

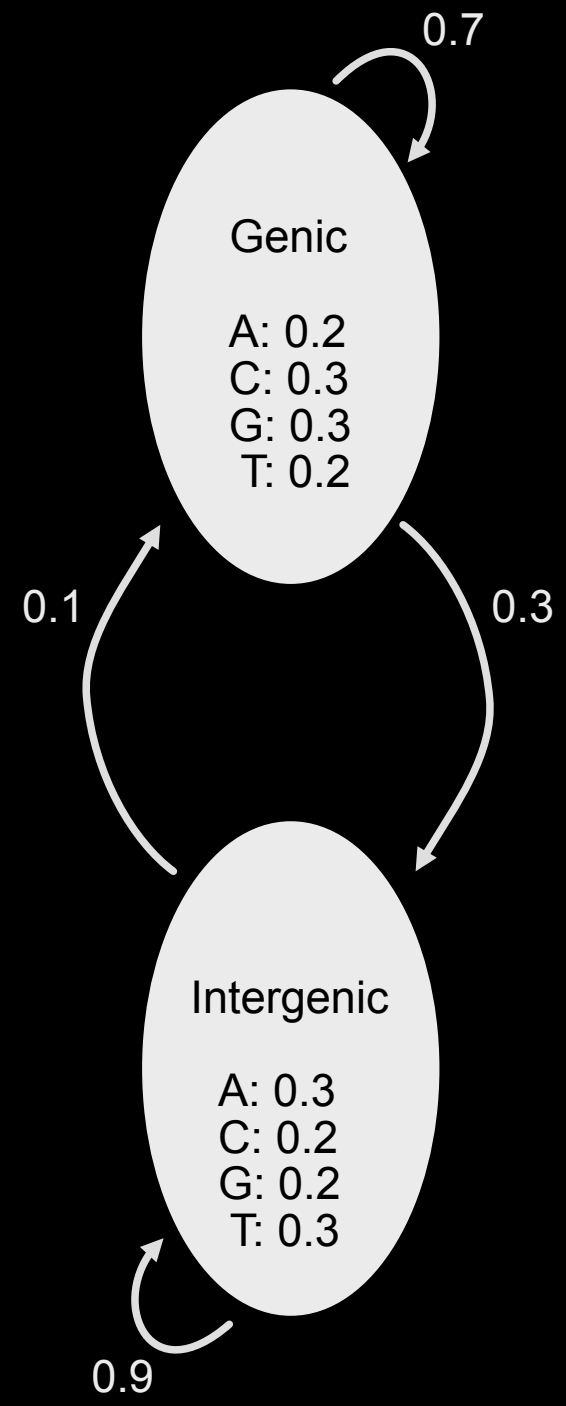


Calculating Probability of a State Path

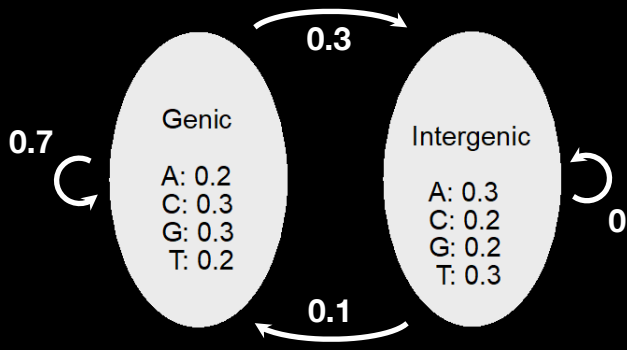
- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x$$
$$p(C|genic)\ x\ p(genic\ \rightarrow inter)\ x\ p(T|inter)\ \ldots$$

$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ x\ 0.3\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3 = 5.58\ x\ 10^{-5}$$



How likely is this sequence, given our model of how the DNA works?

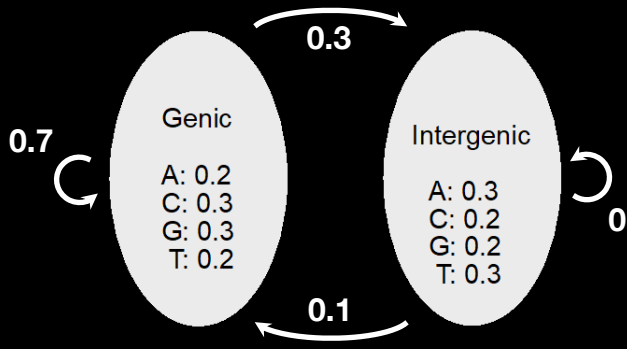


Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic									
Intergenic									

$$P(G, \text{Genic}) = \text{P}(\text{Genic}|\text{start}) \times P(G|\text{Genic})$$

$$P(G, \text{Genic}) = 0.5 \times 0.3 = 0.15$$

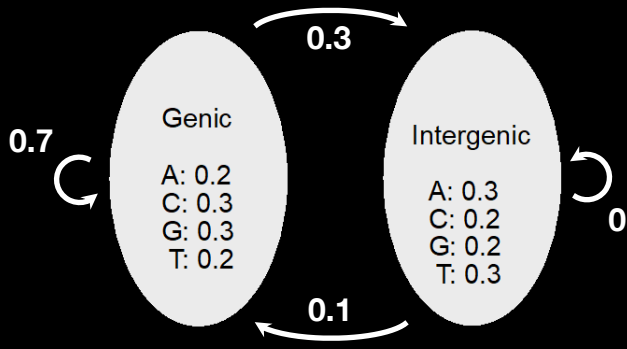


Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic									

$$P(G, \text{Intergenic}) = \text{P}(\text{Intergenic}|\text{start}) \times P(G|\text{Intergenic})$$

$$P(G, \text{Intergenic}) = 0.5 \times 0.2 = 0.1$$



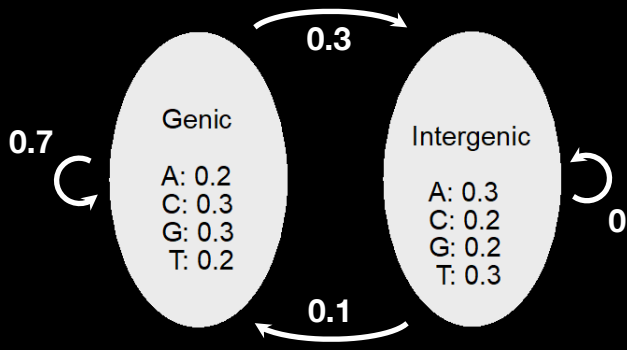
Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

When we were doing the Viterbi algorithm, we calculated the probability of emitting a “C” in the genic state given that the previous state was genic OR given that the previous state was intergenic

We then chose the maximum of the two, and kept track of the path

This time, we’re going to sum the probabilities



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

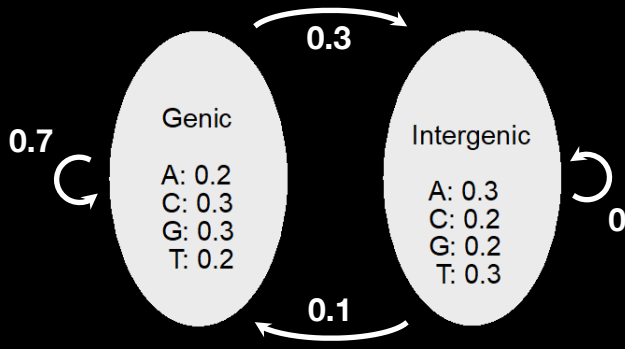
$$P(GC, \text{Genic} \rightarrow \text{Genic}) = P(G, \text{Genic}) \times P(\text{Genic} \rightarrow \text{Genic}) \times P(C|\text{Genic})$$

$$P(GC, \text{Genic} \rightarrow \text{Genic}) = 0.15 \times 0.7 \times 0.3 = 0.0315$$

$$P(GC, \text{Intergenic} \rightarrow \text{Genic}) = P(G, \text{Intergenic}) \times P(\text{Intergenic} \rightarrow \text{Genic}) \times P(C|\text{Genic})$$

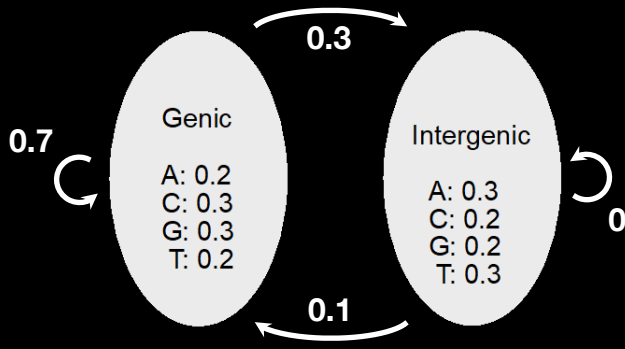
$$P(GC, \text{Intergenic} \rightarrow \text{Genic}) = 0.1 \times 0.1 \times 0.3 = 0.003$$

$$P(GC, \text{End on Genic}) = 0.0315 + 0.003 = 0.0345$$



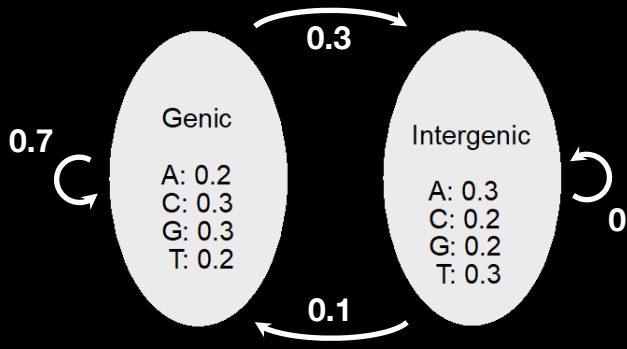
Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345							
Intergenic	0.1								



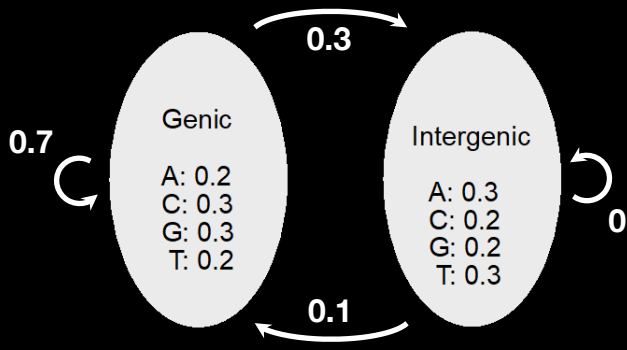
Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345							
Intergenic	0.1								



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345							
Intergenic	0.1	0.027							



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345	0.008055	…	…	…	…	…	1.0×10^{-6}
Intergenic	0.1	0.027	0.00693	…	…	…	…	…	2.3×10^{-6}

$$P(GCCGCGTAG, \text{End on Genic}) = 1.012\times10^{-6}$$

$$P(GCCGCGTAG, \text{End on Intergenic}) = 2.296\times10^{-6}$$

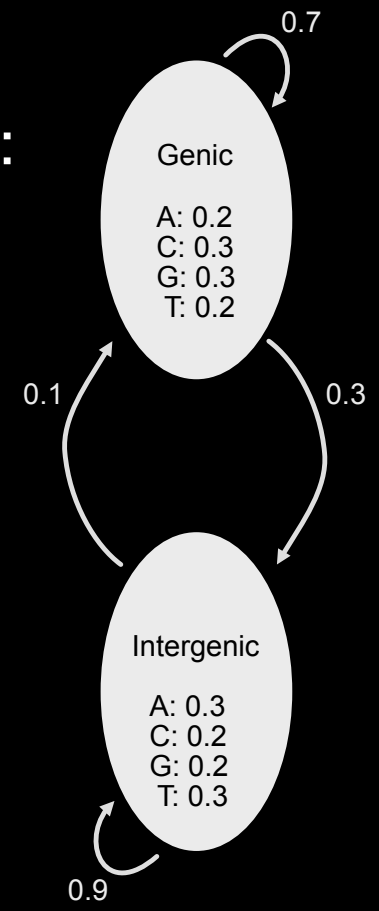
$$\boldsymbol{P}(GCCGCGTAG) = 1.012\times10^{-6} + 2.296\times10^{-6} = 3.308\times10^{-6}$$

Log Odds

The probability that our model generated the sequence: “GCCGCGTAG” is:

$$\boldsymbol{P}(GCCGCGTAG|\text{Full Model}) = 3.308\times10^{-6}$$

The probability that an *intergenic-only* model generated the sequence is:



Log Odds

The probability that our model generated the sequence: “GCCGCGTAG” is:

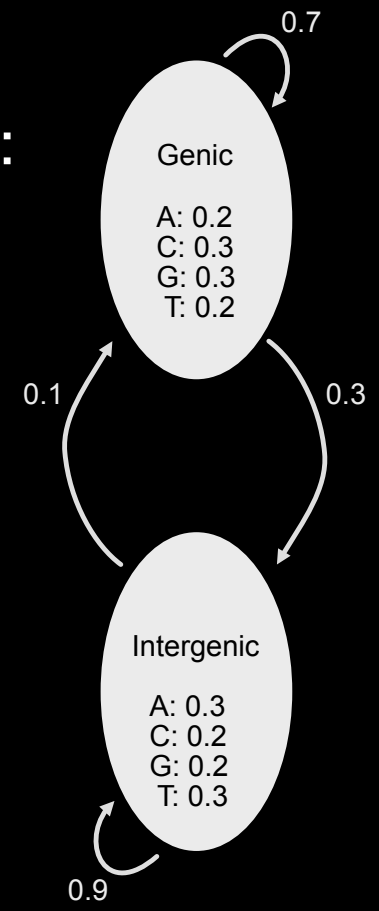
$$\boldsymbol{P}(GCCGCGTAG|\text{Full Model}) = 3.308\times10^{-6}$$

The probability that an *intergenic-only* model generated the sequence is:

$$0.2\times0.2\times0.2\times0.2\times0.2\times0.2\times0.3\times0.3\times0.2 = 1.152\times10^{-6}$$

So the log-odds are as follows:

$$\log(3.308\times10^{-6} \div 1.152\times10^{-6}) = 1.055$$



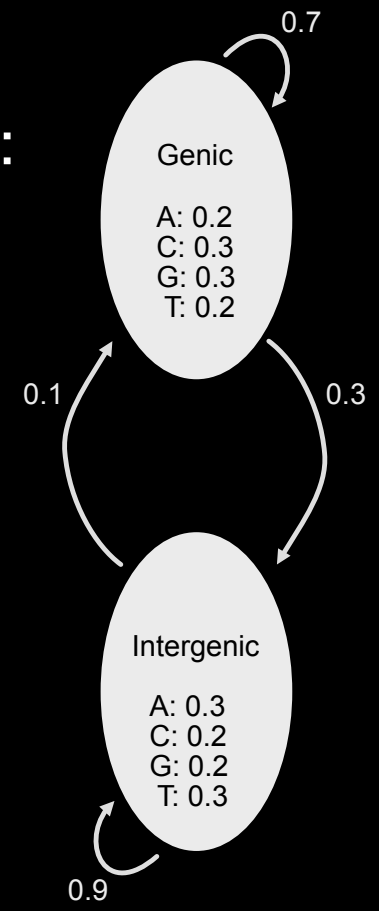
Log Odds

The probability that our model generated the sequence: “GCCGCGTAG” is:

$$\boldsymbol{P}(GCCGCGTAG|\text{Full Model}) = 3.308\times10^{-6}$$

The probability that an *intergenic-only* model generated the sequence is:

$$0.2\times0.2\times0.2\times0.2\times0.2\times0.2\times0.3\times0.3\times0.2 = 1.152\times10^{-6}$$



Log Odds

The probability that our model generated the sequence: “GCCGCGTAG” is:

$$\boldsymbol{P}(GCCGCGTAG|\text{Full Model}) = 3.308\times10^{-6}$$

The probability that an *intergenic-only* model generated the sequence is:

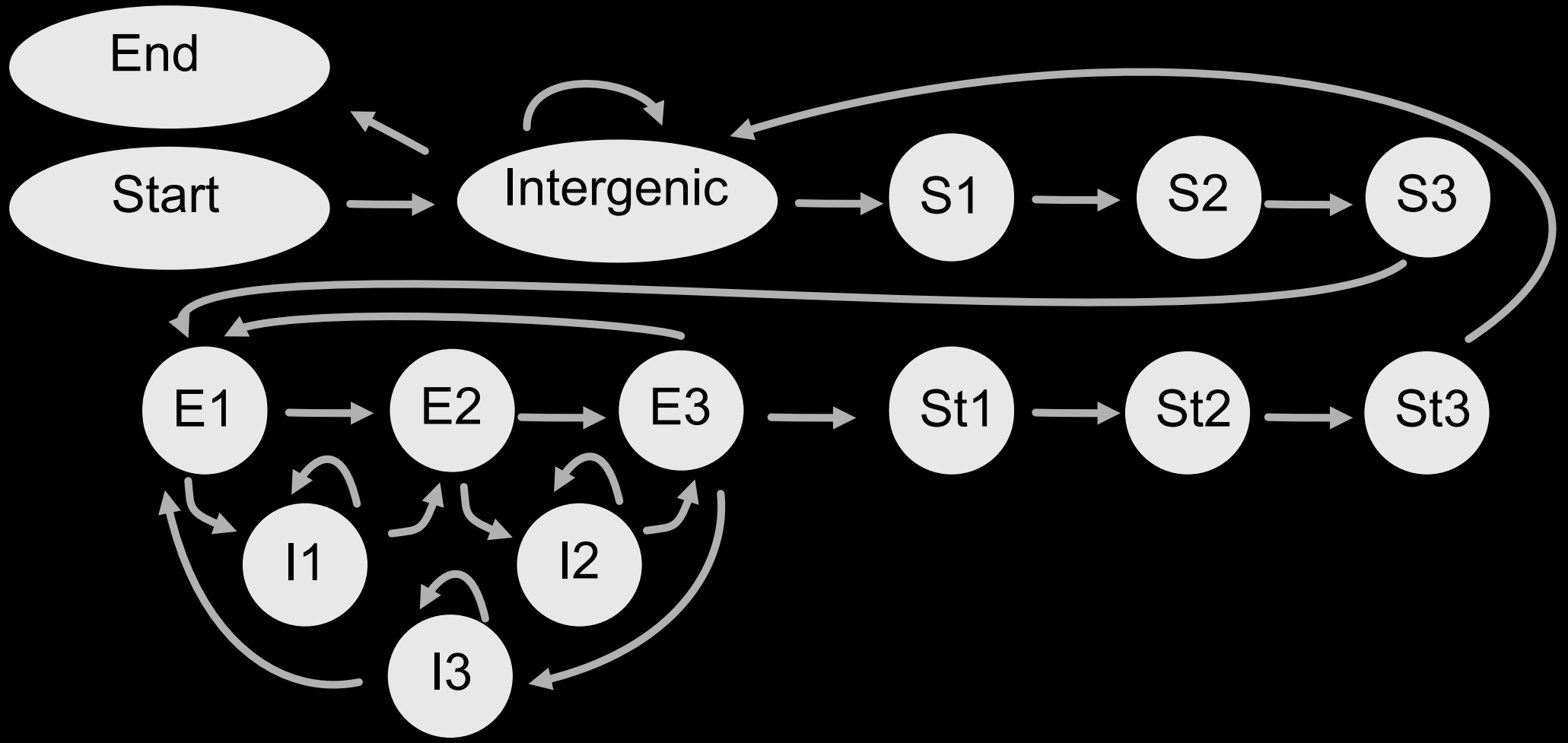
$$0.2\times0.2\times0.2\times0.2\times0.2\times0.2\times0.3\times0.3\times0.2 = 1.152\times10^{-6}$$

So the log-odds are as follows:

$$\log(3.308\times10^{-6} \div 1.152\times10^{-6}) = 1.055$$

This is useful because it centers the values around zero
If positive, numerator was more likely
If negative, denominator was more likely

Genic vs. Intergenic



Today

- Walk through each step and write `pseudo code`: in words what you'd like to do
- We'll give you the real python code for each step + go over what the code does
- Using `jupyter notebooks`

- https://github.com/scarioscia/hmm_workshop and click

- `HMM_Building_Key` is the answer key – you'll copy the python from there
- `HMM_Building` is where you'll do you work (save as pdf, HTML, etc.)

Today

DO NOT REFRESH

`HMM_Building`

- Walk through each step and write `pseudo code`: in words what you'd like to do
- We'll give you the real python code for each step + go over what the code does
- Using `jupyter notebooks`
- https://github.com/scarioscia/hmm_workshop and click launch binder
- `HMM_Building_Key` is the answer key – you'll copy the python from there
- `HMM_Building` is where you'll do you work

Thank you!

AGARA BIO

@tlrdln22
@saracarioscia

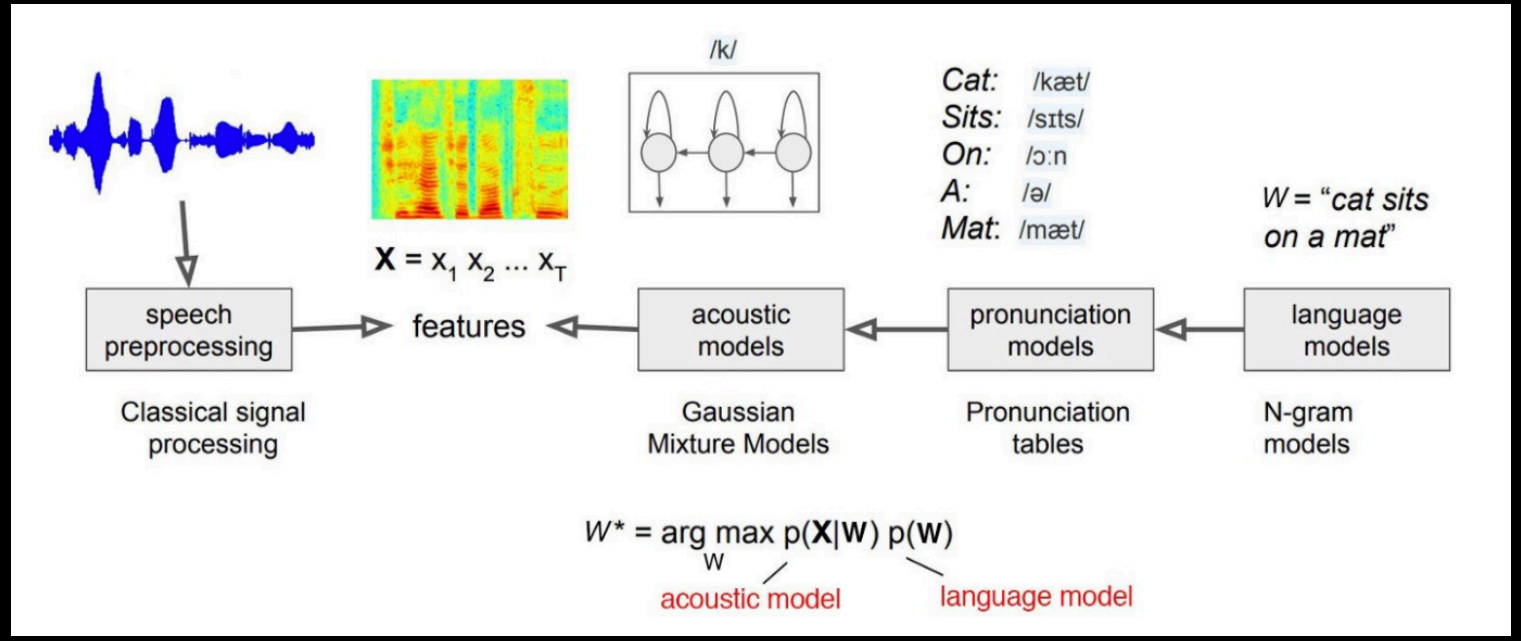
McCoy Lab

Backup

Why Use an HMM?

- Gene finding
- Sequence alignment
- Inference of parental haplotypes
- Predict protein structure

- Speech recognition
- Characterize human gait
- Detecting cyber attacks



Probability

Let's say we have a sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

What is $p(s)$, the probability of s?

Probability

Let's say we have a sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

What is $p(s)$, the probability of s?

If each observation is independent:

$$p(s) = p(s_1) \times p(s_2) \times \cdots \times p(s_n)$$

Probability

Let's say we have a sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

What is $p(s)$, the probability of s?

If each observation is independent:

$$p(s) = p(s_1) \times p(s_2) \times \cdots \times p(s_n)$$

Only need to model one parameter per possible observation

Probability

Let's say we have a sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

What is $p(s)$, the probability of s?

If each observation is independent:

$$p(s) = p(s_1) \times p(s_2) \times \cdots \times p(s_n)$$

Only need to model one parameter per possible observation

But what if the observations are not independent?

Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

If the observations are not independent, how do we find the probability of s?

Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

If the observations are not independent, how do we find the probability of s?

$$p(s) = p(s_1, s_2, \ldots, s_n)$$
$$p(s) = p(s_1) \times p(s_2, s_3, \ldots, s_n)$$
$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3, \ldots, s_n)$$
$$\vdots$$
$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3|s_2, s_1) \times \cdots \times p(s_n|s_{n-1}, s_{n-2}, \ldots, s_1)$$

Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

If the observations are not independent, how do we find the probability of s?

$$p(s) = p(s_1, s_2, \ldots, s_n)$$
$$p(s) = p(s_1) \times p(s_2, s_3, \ldots, s_n)$$
$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3, \ldots, s_n)$$
$$\vdots$$
$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3|s_2, s_1) \times \cdots \times p(s_n|s_{n-1}, s_{n-2}, \ldots, s_1)$$

The probability of each observation is influenced by all other observations

Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

If the observations are not independent, how do we find the probability of s?

$$p(s) = p(s_1, s_2, \ldots, s_n)$$
$$p(s) = p(s_1) \times p(s_2, s_3, \ldots, s_n)$$
$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3, \ldots, s_n)$$
$$\vdots$$
$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3|s_2, s_1) \times \cdots \times p(s_n|s_{n-1}, s_{n-2}, \ldots, s_1)$$

The probability of each observation is influenced by all other observations

Need to model one parameter for every single possible sequence

(Simplified) Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

What's the simplest way we can model the dependence between observations – **without** modelling the full conditional probability?

(Simplified) Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

What's the simplest way we can model the dependence between observations – **without** modelling the full conditional probability?

Markov assumption: each observation is only on the observation directly before

$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3|s_2) \times \cdots \times p(s_n|s_{n-1})$$

(Simplified) Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \ldots, s_n\}$

What's the simplest way we can model the dependence between observations – **without** modelling the full conditional probability?

Markov assumption: each observation is only on the observation directly before

$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3|s_2) \times \cdots \times p(s_n|s_{n-1})$$

Need to model (number of possible observations)^2 parameters

(Simplified) Conditional Probability

Sequence of observations: $s = \{s_1, s_2, \dots, s_n\}$

Independent Probability

$$p(s) = p(s_1) \times p(s_2) \times \cdots \times p(s_n)$$

Full Conditional Probability

$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3|s_2, s_1) \times \cdots \times p(s_n|s_{n-1}, s_{n-2}, \dots, s_1)$$

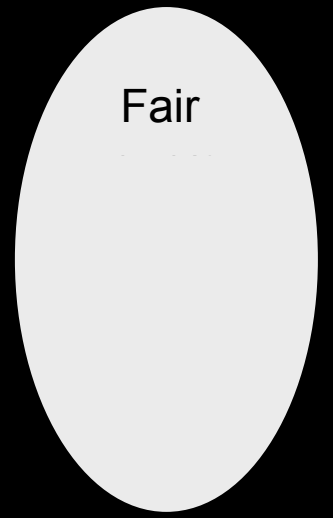
Markov Probability

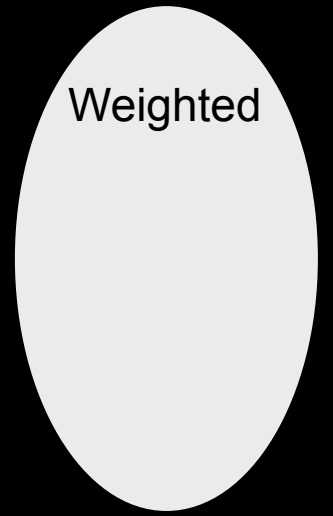
$$p(s) = p(s_1) \times p(s_2|s_1) \times p(s_3|s_2) \times \cdots \times p(s_n|s_{n-1})$$

The weighted die

Let’s imagine we run a casino, and we like to cheat our customers. 🙂

- At any point in time, we can either be using a fair die or a weighted die → **states**

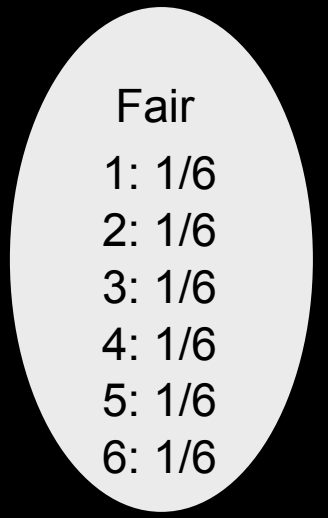


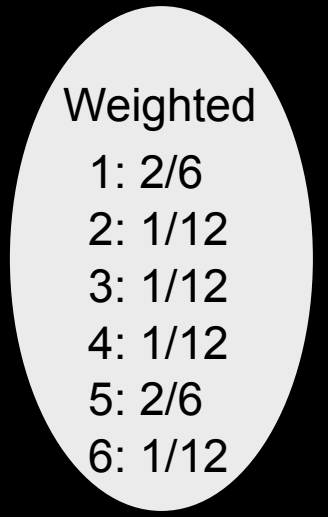


The weighted die

Let’s imagine we run a casino, and we like to cheat our customers. 🙂

- At any point in time, we can either be using a fair die or a weighted die → **states**
- Each state has a probability of revealing the numbers 1 through 6 → **emission probabilities**

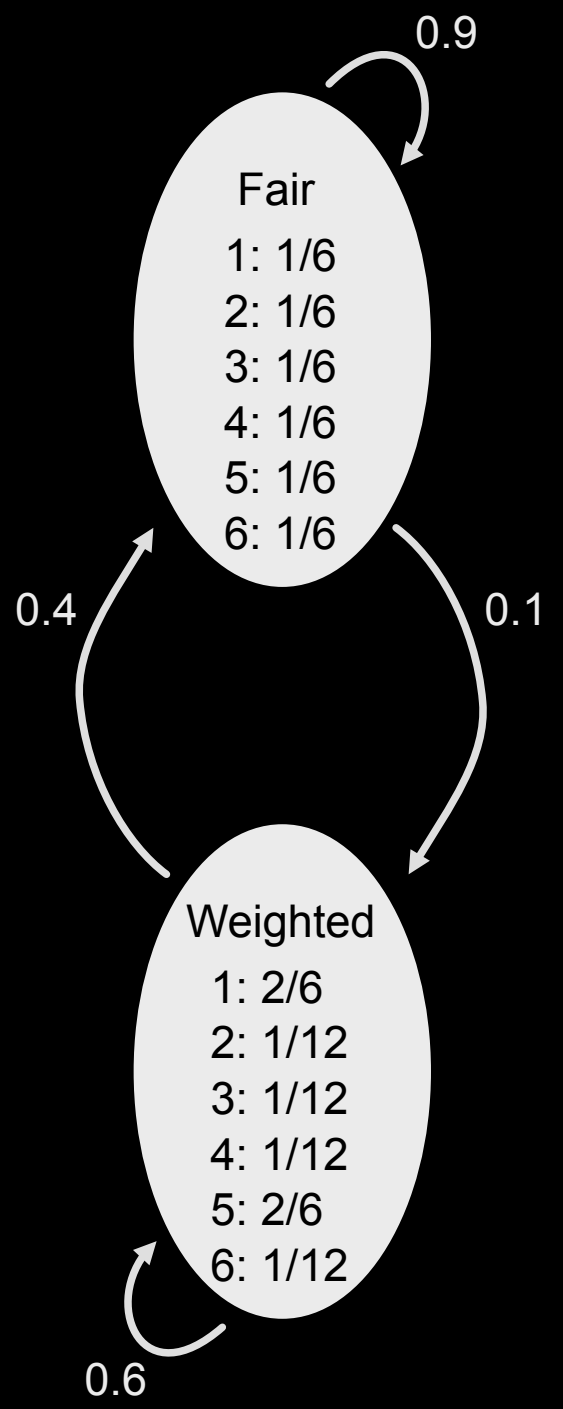




The weighted die

Let’s imagine we run a casino, and we like to cheat our customers. 🙂

- At any point in time, we can either be using a fair die or a weighted die → **states**
- Each state has a probability of revealing the numbers 1 through 6 → **emission probabilities**
- With each roll, the casino either keeps using the same die or switches dice → **transition probabilities**

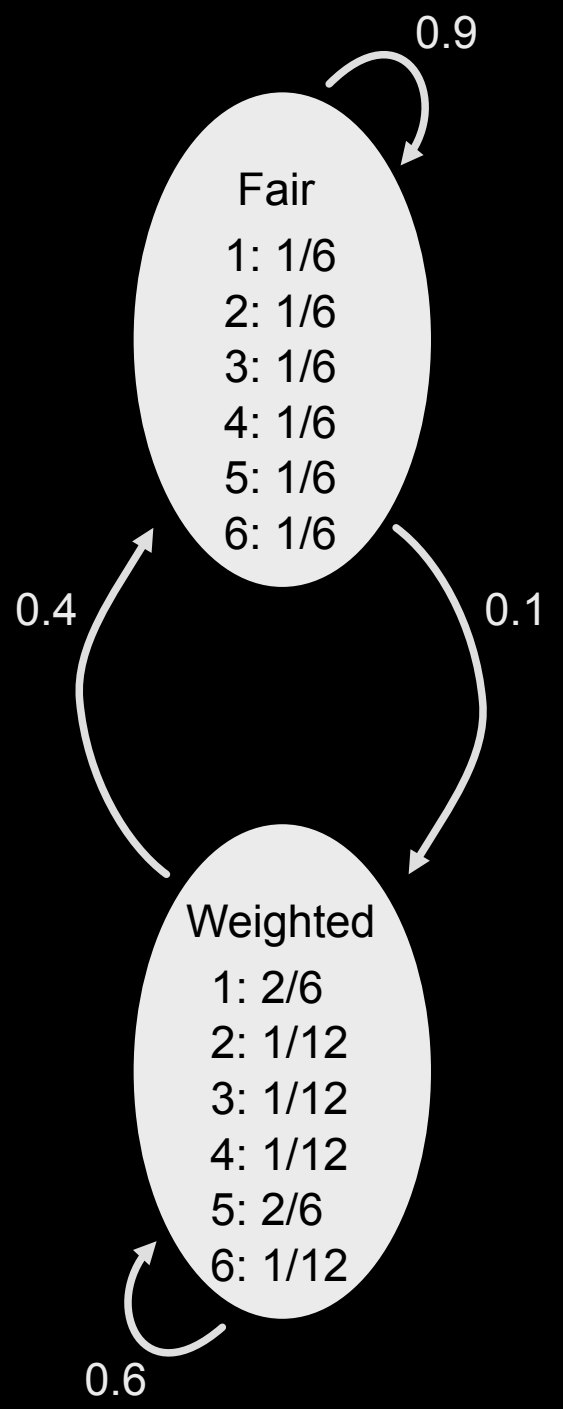


The weighted die

Let's imagine we run a casino, and we like to cheat our customers. 🙂

- At any point in time, we can either be using a fair die or a weighted die → **states**
- Each state has a probability of revealing the numbers 1 through 6 → **emission probabilities**
- With each roll, the casino either keeps using the same die or switches dice → **transition probabilities**

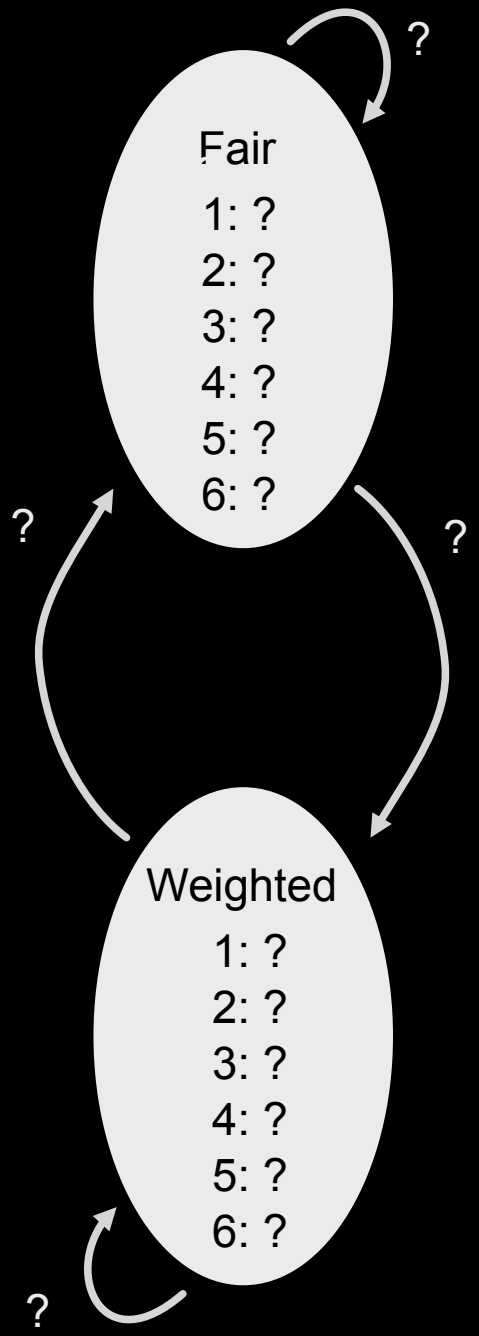
Together, these form a **Hidden Markov Model**



Why is it Hidden?

Now, let's imagine we're a gambler, and we want to figure out when the casino is using the fair die and when they're using the weighted die.

What information do we actually have?



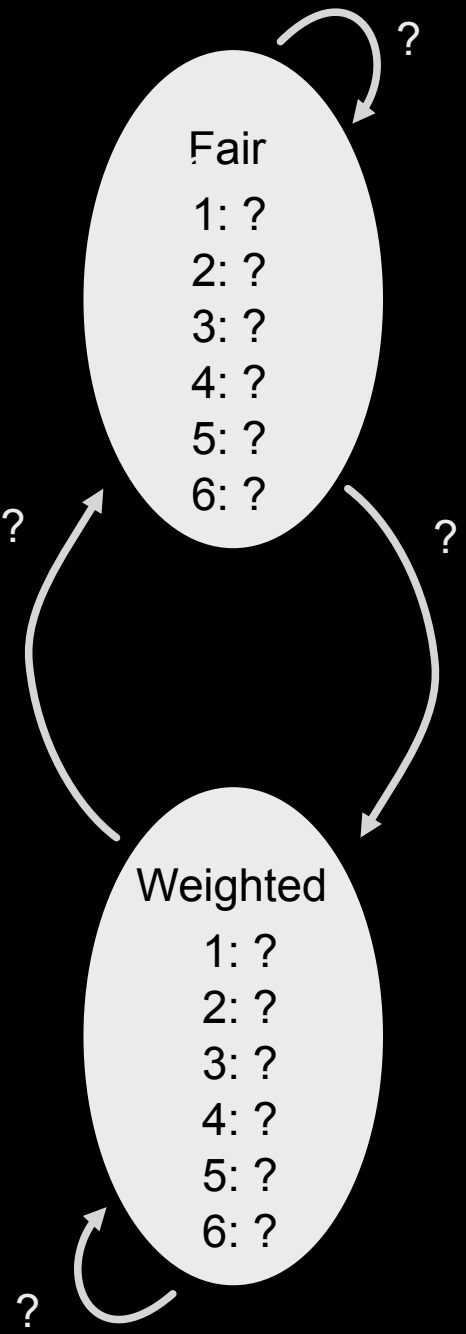
Why is it Hidden?

Now, let's imagine we're a gambler, and we want to figure out when the casino is using the fair die and when they're using the weighted die.

What information do we actually have?

The sequence of rolls:

`1236534563126635254261616552414142536123654 3`



Why is it Hidden?

Now, let's imagine we're a gambler, and we want to figure out when the casino is using the fair die and when they're using the weighted die.

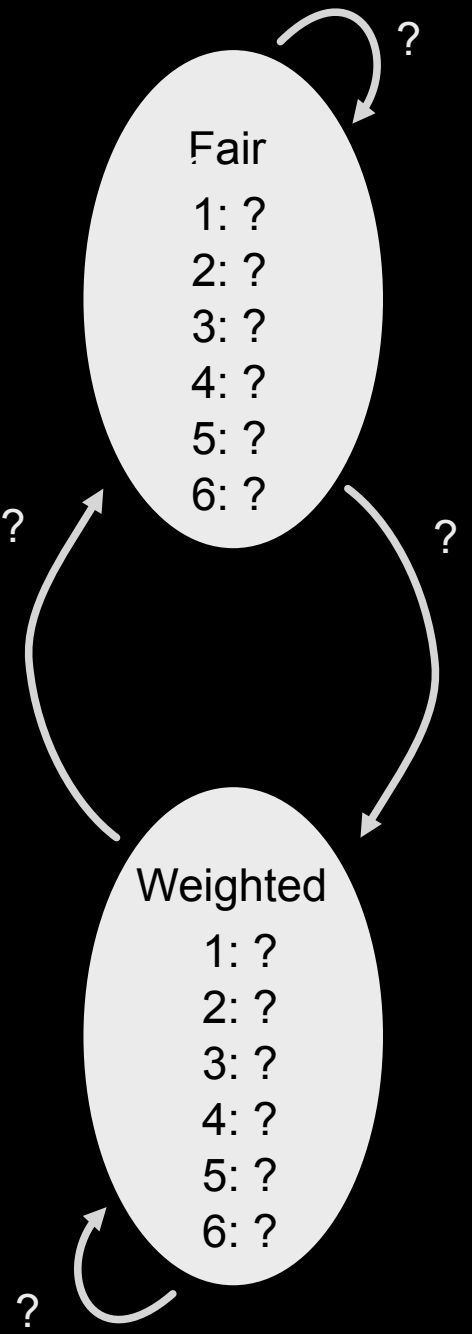
What information do we actually have?

The sequence of rolls:

```
1236534563126635254261655241414253612365 43
```



```
FWWWFFFFFFFFWWFFFFWWFWFFFFFFFFWWWWWWWFFFFWFF
```



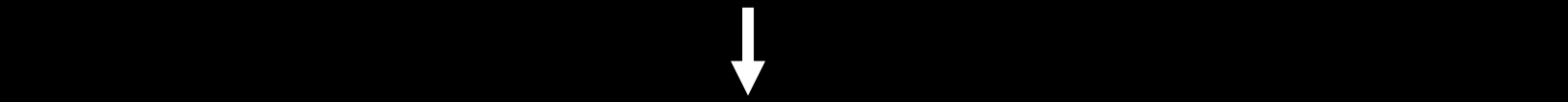
Why is it Hidden?

Now, let's imagine we're a gambler, and we want to figure out when the casino is using the fair die and when they're using the weighted die.

What information do we actually have?

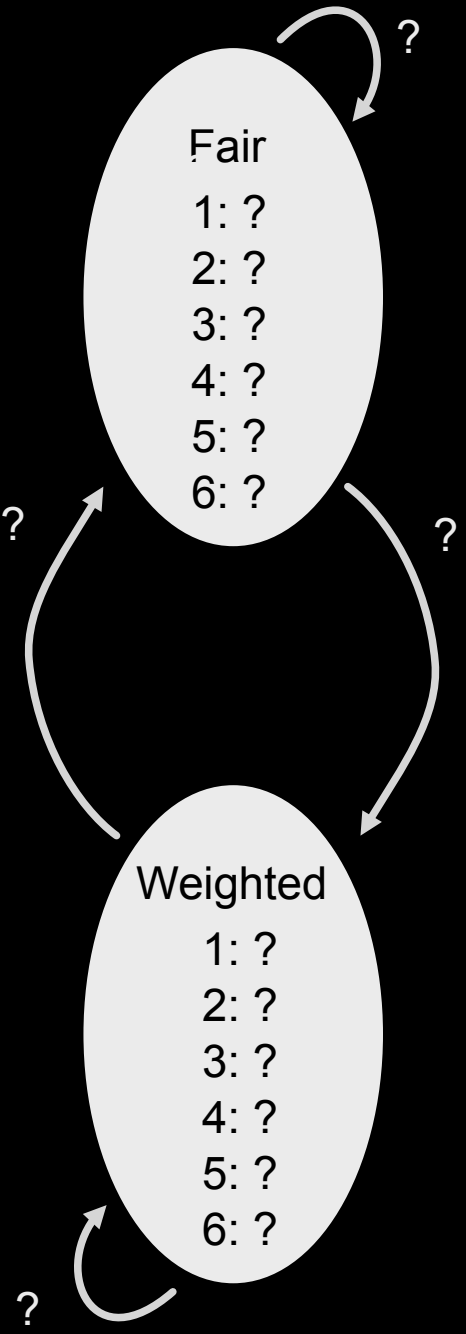
The sequence of rolls:

`12365345631266352542616165524141425361236543`



`FWWWFFFFFFFFWWFFFFWWFWFFFFFFFFWWWWWWWFFFFWFF`

The **states** are hidden



Gene Finding

Let’s say we have a DNA sequence:

AGGCAGTGACGCTAGGGCAGAGAACCTAATTTTGAAAGCTTCGCCGGGCGAACGTTTGGG...

Is there a gene within this sequence?

If so, where is it?

Genic vs. Intergenic

Whether a given nucleotide is in a gene is going to depend on the context:
If a nucleotide is near others that are in a gene, it’s more likely that that nucleotide is within a gene

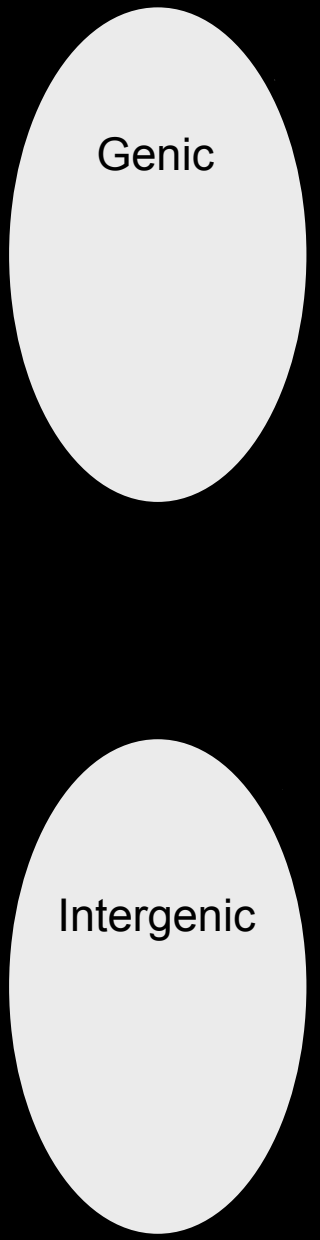
Genic vs. Intergenic

Whether a given nucleotide is in a gene is going to depend on the context:
If a nucleotide is near others that are in a gene, it's more likely that that nucleotide is within a gene

- We assume the only thing that determines each position (position x) is the previous position (position $x - 1$)

Genic vs. Intergenic

- Any given nucleotide can be in one of two **states**, Genic or Intergenic



Genic vs. Intergenic

- Any given nucleotide can be in one of two **states**, Genic or Intergenic
- Within each of these states, there is some probability of an A, C, G, or T

Genic

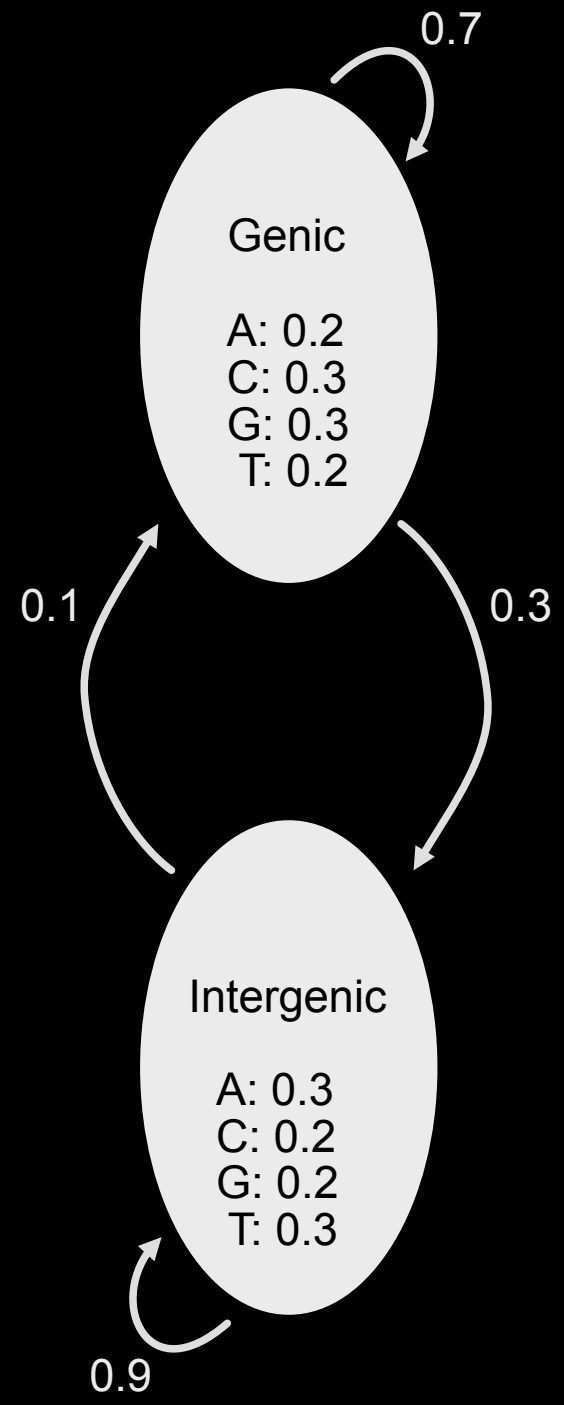
A: 0.2
C: 0.3
G: 0.3
T: 0.2

Intergenic

A: 0.3
C: 0.2
G: 0.2
T: 0.3

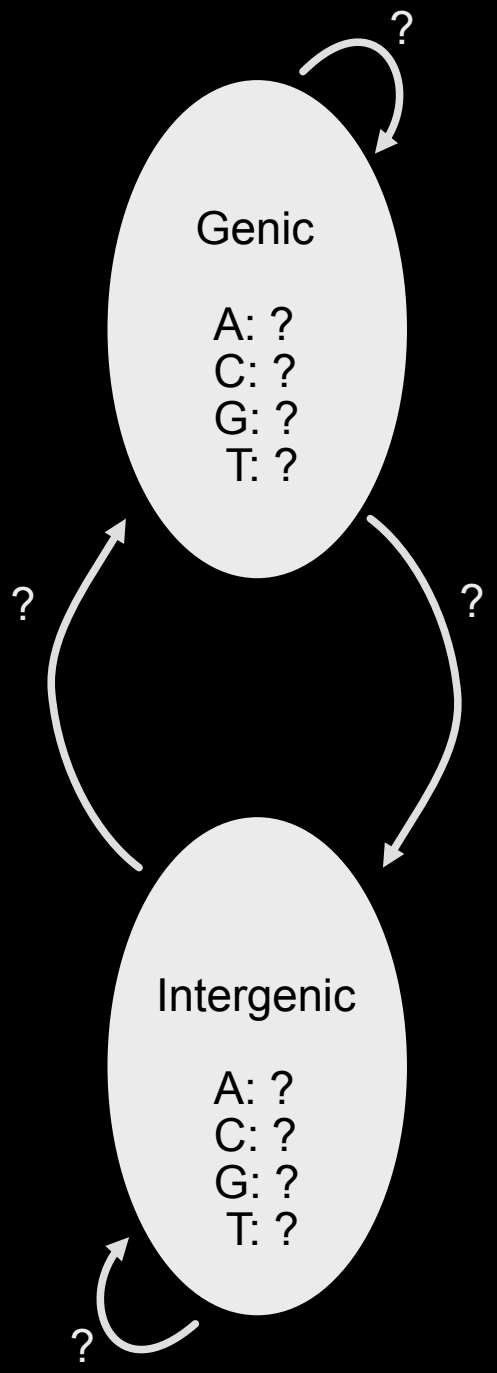
Genic vs. Intergenic

- Any given nucleotide can be in one of two **states**, Genic or Intergenic
- Within each of these states, there is some probability of an A, C, G, or T
- As we move from one nucleotide to the next, there is some probability that we stay in the same state and some probability that we switch to the other



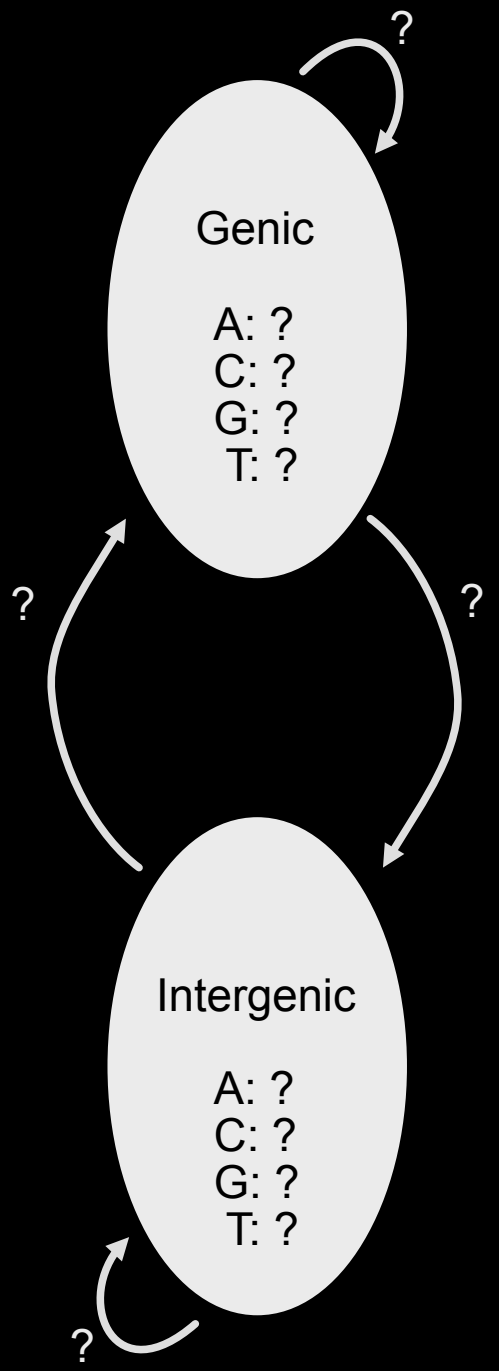
Inferring Probabilities

- Before we move on to decoding the HMM, let's talk about how we actually get these probabilities



Inferring Probabilities

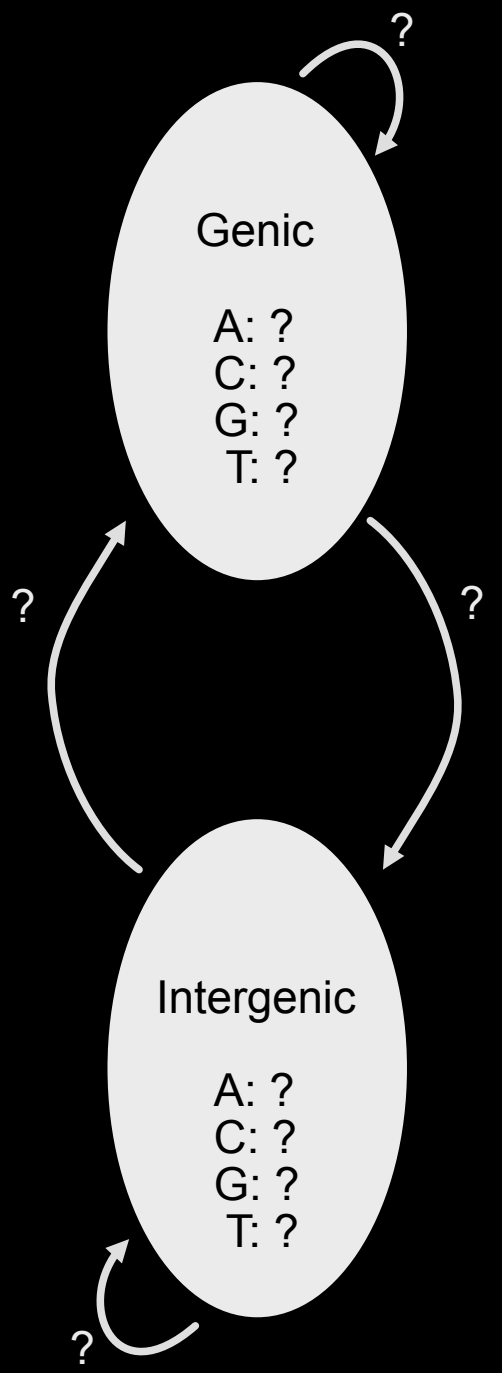
- Before we move on to decoding the HMM, let's talk about how we actually get these probabilities
- If we know the structure of the HMM and we have some training data with "correct" labels, how can we infer these probabilities?



Inferring Probabilities

- Before we move on to decoding the HMM, let's talk about how we actually get these probabilities
- If we know the structure of the HMM and we have some training data with "correct" labels, how can we infer these probabilities?

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCAGGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGGIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIIGIIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

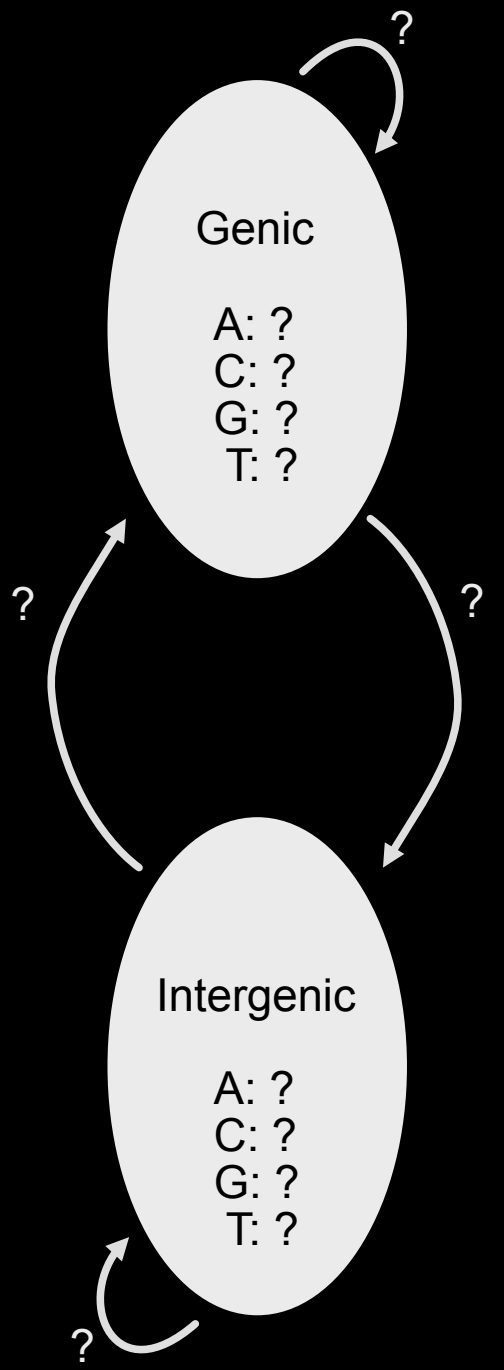


Inferring Probabilities

- Before we move on to decoding the HMM, let's talk about how we actually get these probabilities
- If we know the structure of the HMM and we have some training data with "correct" labels, how can we infer these probabilities?

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCAGGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGGIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIIGIIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII
```

Count!



Inferring Probabilities

Training Data

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCAGGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGGIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIIGIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII

CTTGGACGGCACGGGGATGAAGGCTTAAATTGGTAATCCGTGTAAAATAT
GIIIIIIIIIIIIIIIIGGGGGGIIIIIIIIIGIIIIIGIIIIIIIIIIGG

ATAAACGCCTCAGCATGCAACCGATTCTAATGTCGTAGCCGCAATTAGTT
IIIIIIIIIGGIIIIIIIGGGGGIIIIIIGGGIIIIIGGGIIIIIIIIII

ATATGCGTCAACAGTGCATCCAGCGCTATTCGCGCGTTTCCCGTAAGCAC
IIIGIIIIIIIIGIIIIIIIIIIIIIIIIIGGIIIIGGGGGGGGGGGGGG

GACGCTATGTGGACCATTCGAATTAAGCTAATACTAGATCGTGCCCGGTA
IIGGGGGGIGGGGGGGGIIIIIIGIIGIIIIIIGGIIIIIIIIIIGIGGI
```

Inferring Probabilities

Training Data

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCAGGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGGIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIIGIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII

CTTGGACGGCACGGGGATGAAGGCTTAAATTGGTAATCCGTGTAAAATAT
GIIIIIIIIIIIIIIIIGGGGGGIIIIIIIIIGIIIIIGIIIIIIIIIIGG

ATAAACGCCTCAGCATGCAACCGATTCTAATGTCGTAGCCGCAATTAGTT
IIIIIIIIIGGIIIIIIIGGGGGIIIIIIGGGIIIIIGGGIIIIIIIIII

ATATGCGTCAACAGTGCATCCAGCGCTATTCGCGCCTTTCCCGTAAGCAC
IIIGIIIIIIIIGIIIIIIIIIIIIIIIIIIGGIIIIGGGGGGGGGGGGGG

GACGCTATGTGGACCATTCGAATTAAGCTAATACTAGATCGTGCCCGGTA
IIGGGGGGIGGGGGGGGIIIIIIGIIGIIIIIIGGIIIIIIIIIIGIGGI
```

Genic Emission

A: 34 = 0.241

C: 40 = 0.284

G: 38 = 0.270

T: 29 = 0.205

Inferring Probabilities

Training Data

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCACGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGCIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIIGIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII

CTTGGACGGCACGGGGATGAAGGCTTAAATTGGTAATCCGTGTAAAATAT
GIIIIIIIIIIIIIIIIGGGGGGIIIIIIIIIGIIIIIGIIIIIIIIIIGG

ATAAACGCCTCAGCATGCAACCGATTCTAATGTCGTAGCCGCAATTAGTT
IIIIIIIIIGGIIIIIIIGGGGGIIIIIIGGGIIIIIGGGIIIIIIIIII

ATATGCGTCAACAGTGCATCCAGCGCTATTCGCGCGTTTCCCGTAAGCAC
IIIGIIIIIIIIGIIIIIIIIIIIIIIIIIGGIIIIGGGGGGGGGGGGGG

GACGCTATGTGGACCATTCGAATTAAGCTAATACTAGATCGTGCCCGGTA
IIGGGGGGIGGGGGGGGIIIIIIGIIGIIIIIIGGIIIIIIIIIIGIGGI
```

Genic Emission

A: 34 = 0.241
C: 40 = 0.284
G: 38 = 0.270
T: 29 = 0.205

Intergenic Emission

A: 110 = 0.307
C: 73 = 0.203
G: 75 = 0.209
T: 101 = 0.281

Inferring Probabilities

Training Data

Genic Emission

A: 34 = 0.241
C: 40 = 0.284
G: 38 = 0.270
T: 29 = 0.205

Intergenic Emission

A: 110 = 0.307
C: 73 = 0.203
G: 75 = 0.209
T: 101 = 0.281

Transitions

G → G: 97 = 0.703
G → I: 41 = 0.297
I → I: 311 = 0.884
I → G: 41 = 0.116

Inferring Probabilities

Training Data

```
TTGCGTTGAAATATTAAATAGGAACGAAACGGATCCCTGGCACCATCAAA
IIIIIIIIIIIIIIIIIIGGGGIIIIIGGGIIGGGGGGIIIIIIIIIII

TGCCCCAGATTCCAGAGTCGCGTGGAACTCTGGAAAAATACCTGAAGTCT
GGIIGGGGGGIIIIGGGGGGGGGGGGIIGIIIIIIGGGGGGGGGGGGGGI

GCTTAGTAGGATTACACCAGGGAGTCTCCACGGTTATTCCACTATCTATT
IIIIGIIIGIIIIIIIGGGGIIIIIIIIIGGGGIIGIIIIIIIIIIIIII

CGTTTGACAAAGTAACGCTTCGGAATAGACCTATGCTTGTGGGCACTCAG
IIIIIIIIIIIIGGIIGIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII

ACGGCAACAATCCGTATAATATATTTCGCATAGGCGTTGCTTCTGGGTAG
GGIIGIIIIIIIIGIIIIIIIIIIIIIGIIIIIIIIIIIGGGGIIIGIIG

ATTGAGTCGCCAGGGTCTTGCAAAAATTACGTTCACAAGAACTAATATCA
IIIIIIIIIGIIIIIIIIGGGGIIIIIIIIIIIIIIIIIIIIIIIIIIII

CTTGGACGGCACGGGGATGAAGGCTTAAATTGGTAATCCGTGTAAAATAT
GIIIIIIIIIIIIIIIIGGGGGGIIIIIIIIIGIIIIIGIIIIIIIIIIGG

ATAAACGCCTCAGCATGCAACCGATTCTAATGTCGTAGCCGCAATTAGTT
IIIIIIIIIGGIIIIIIIGGGGGIIIIIIGGGIIIIIGGGIIIIIIIIII

ATATGCGTCAACAGTGCATCCAGCGCTATTCGCGCGTTTCCCGTAAGCAC
IIIGIIIIIIIIGIIIIIIIIIIIIIIIIIGGIIIIGGGGGGGGGGGGGG

GACGCTATGTGGACCATTCGAATTAAGCTAATACTAGATCGTGCCCGGTA
IIGGGGGGIGGGGGGGGIIIIIIGIIGIIIIIIGGIIIIIIIIIIGIGGI
```

Genic Emission

A: 34 = 0.241
C: 40 = 0.284
G: 38 = 0.270
T: 29 = 0.205

Intergenic Emission

A: 110 = 0.307
C: 73 = 0.203
G: 75 = 0.209
T: 101 = 0.281

Transitions

G → G: 97 = 0.703
G → I: 41 = 0.297
I → I: 311 = 0.884
I → G: 41 = 0.116

Decoding the HMM

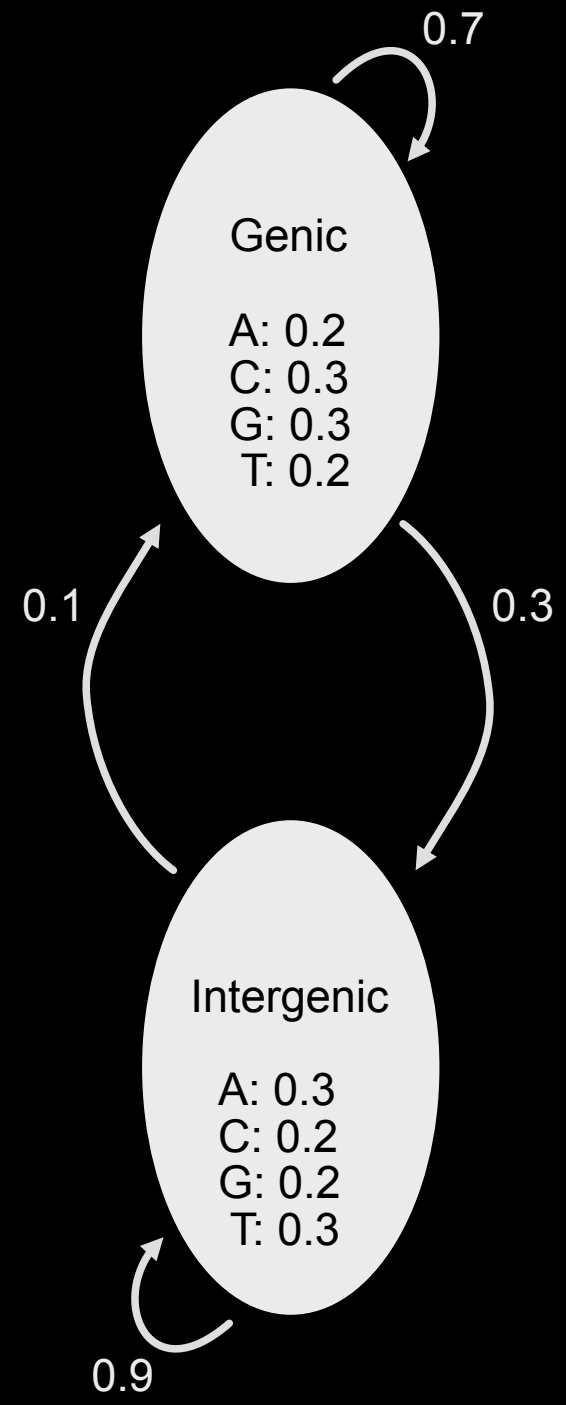
Now that we've "learned" the emission and transition probabilities in our HMM, how can we use these probabilities to answer our question: given a sequence of observations (nucleotides), which state (genic vs. intergenic) is each observation in?

```
TTTATAAGTATTGCAGCCCGCCCGATCGTTCAGACGTCCATATTCCCCTT
```

```
IGIIIIIIIIIIIIGGGGGGGGGGGGIIIIIIIIIIIIIIIGGGGGG
```

Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

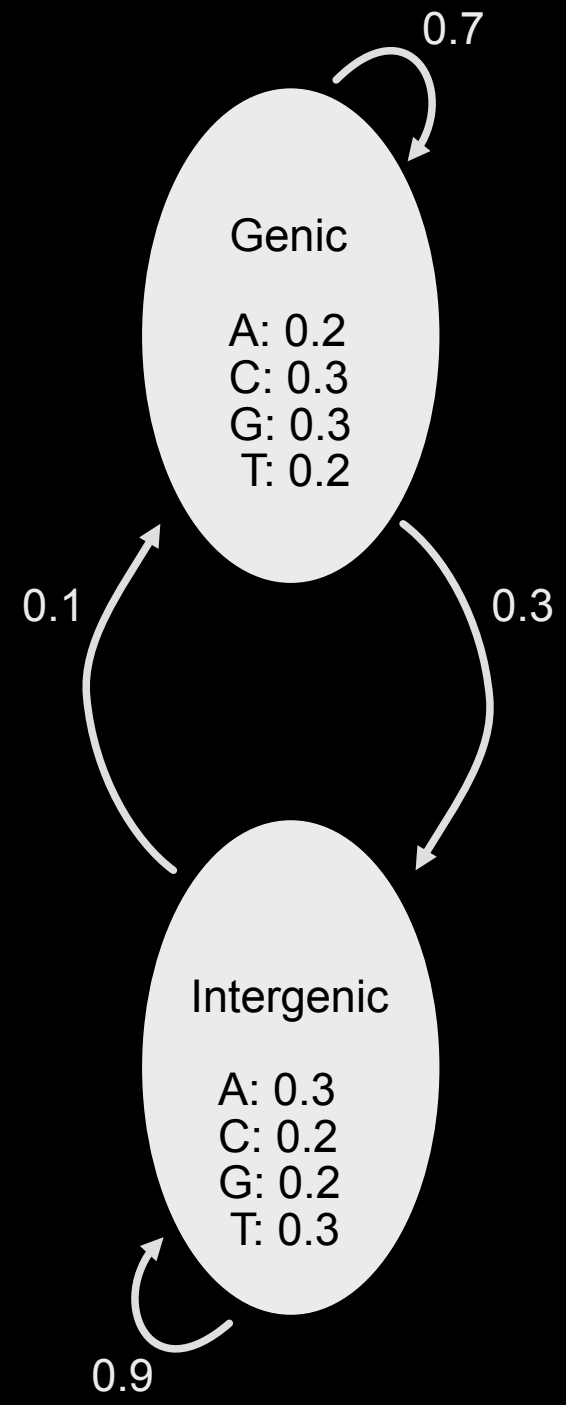


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$p(genic|start)$ ...

0.5 ...

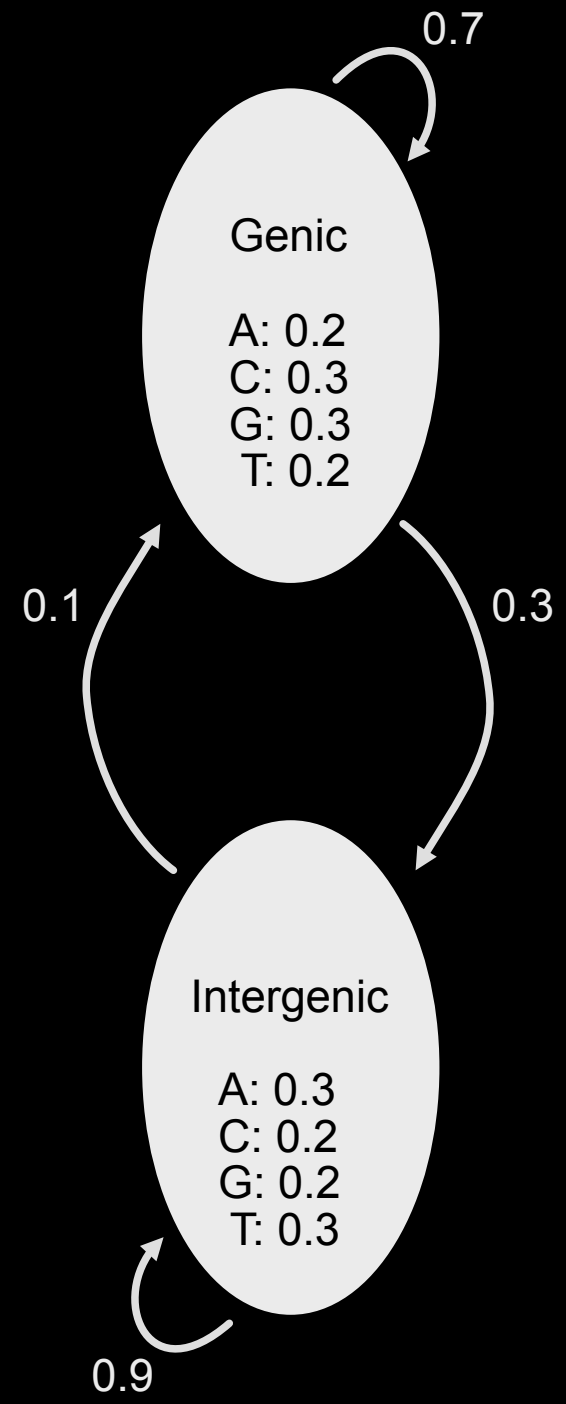


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ \ldots$$

$$0.5\ x\ 0.3\ \ldots$$

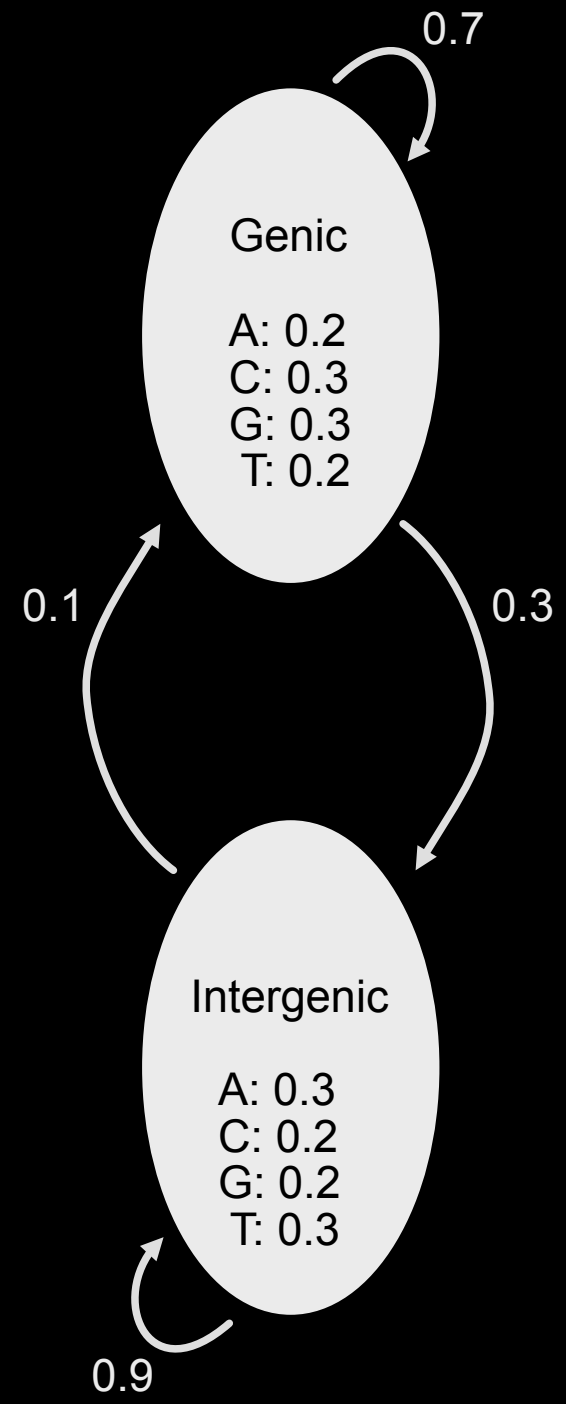


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ ...$$

$$0.5\ x\ 0.3\ x\ 0.7\ ...$$

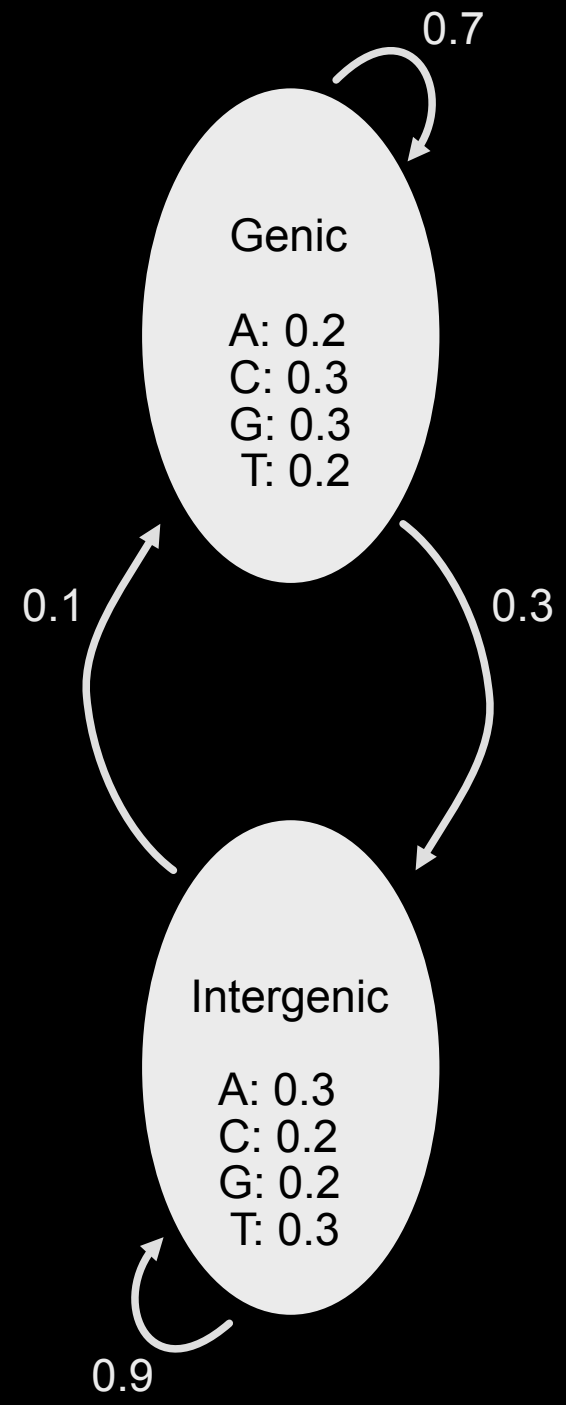


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x\ p(C|genic)\ \ldots$$

$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ \ldots$$

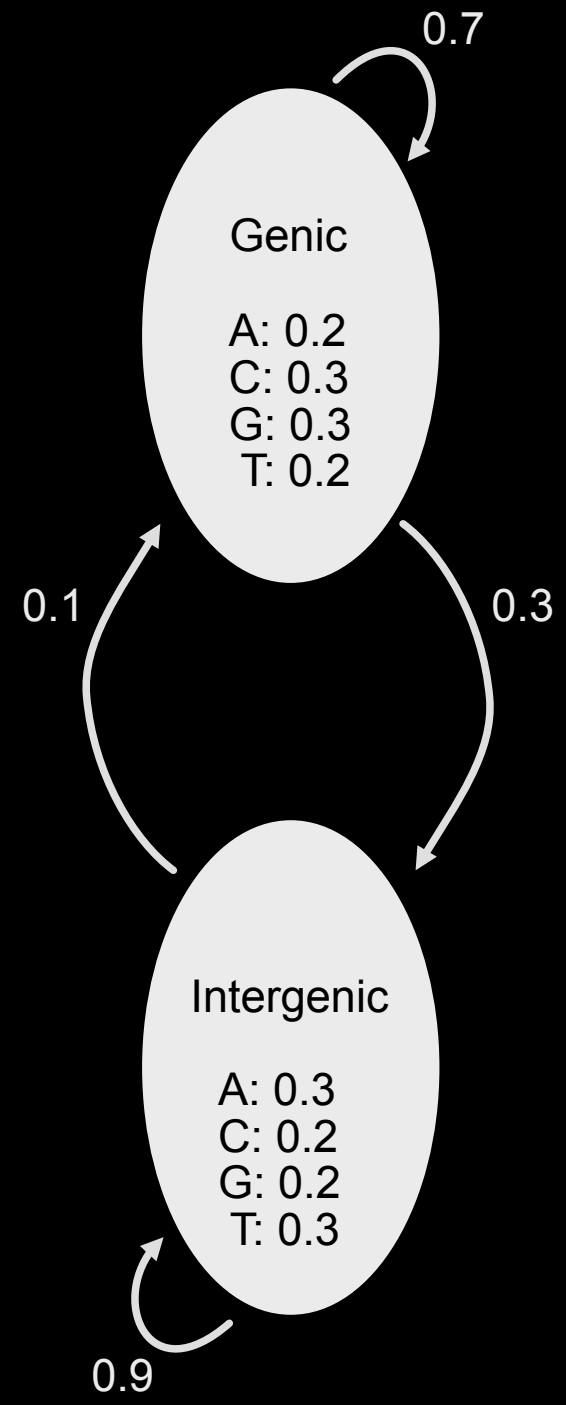


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x\ p(C|genic)\ x\ p(genic\ \rightarrow inter)\ \ldots$$

$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ x\ 0.3\ \ldots$$

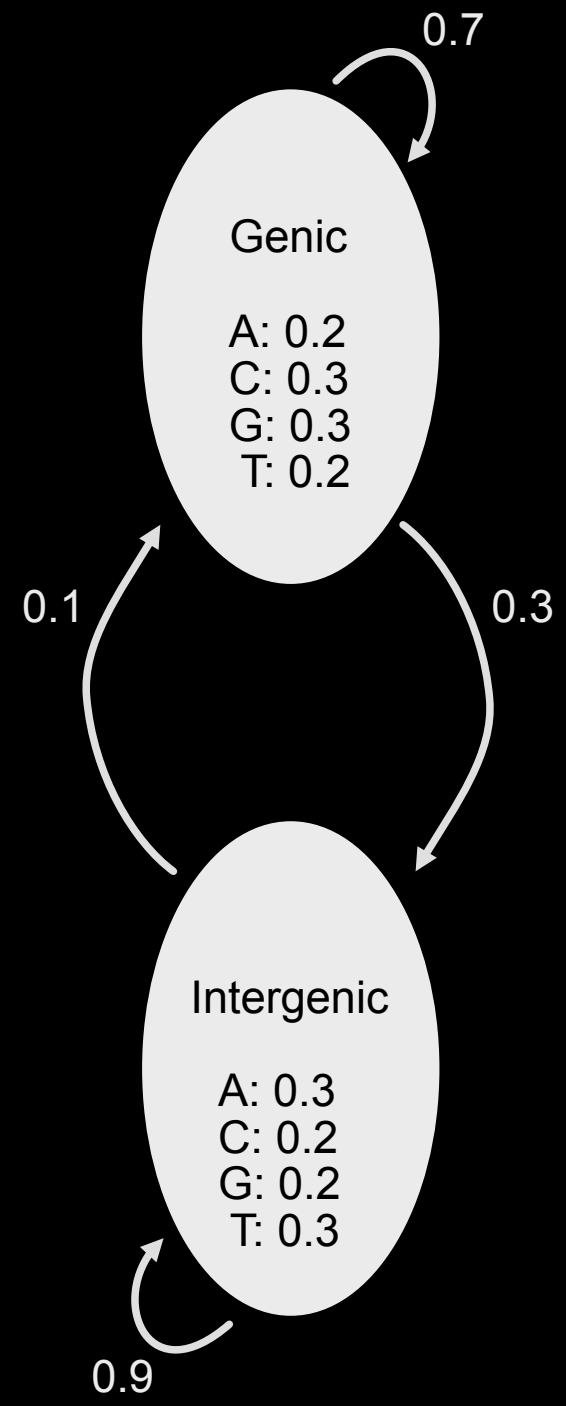


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x$$
$$p(C|genic)\ x\ p(genic\ \rightarrow inter)\ x\ p(T|inter)\ \ldots$$

$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ x\ 0.3\ x\ 0.3\ \ldots$$

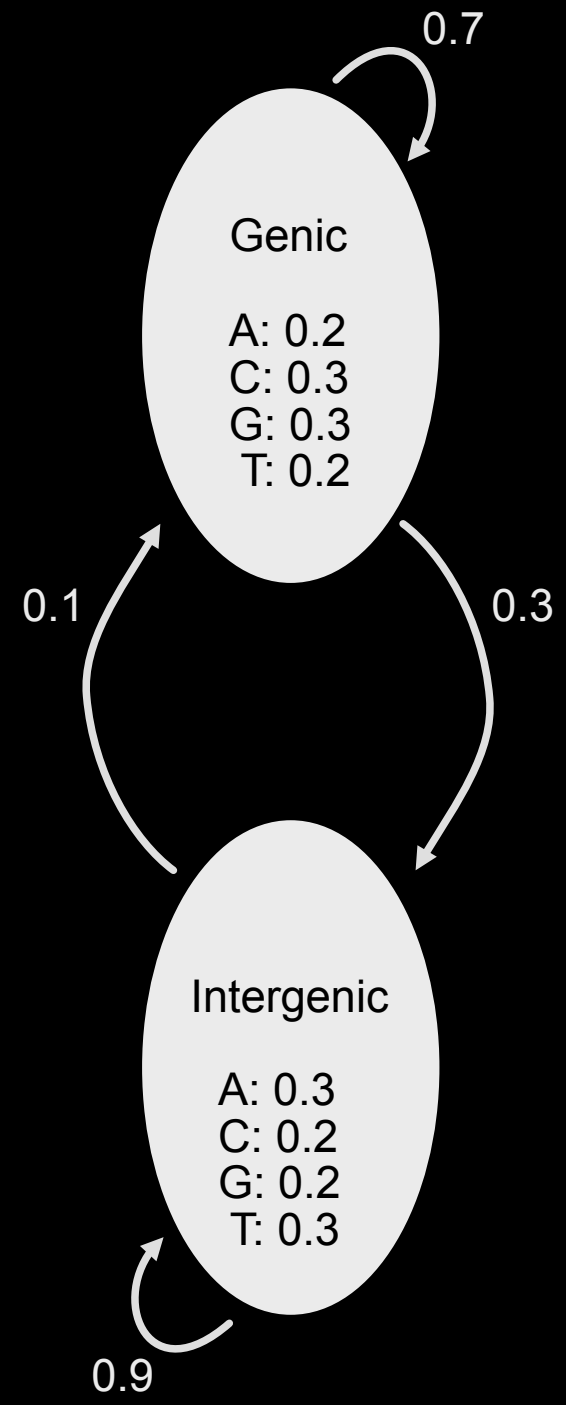


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x$$
$$p(C|genic)\ x\ p(genic\ \rightarrow inter)\ x\ p(T|inter)\ \ldots$$

$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ x\ 0.3\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3 =$$

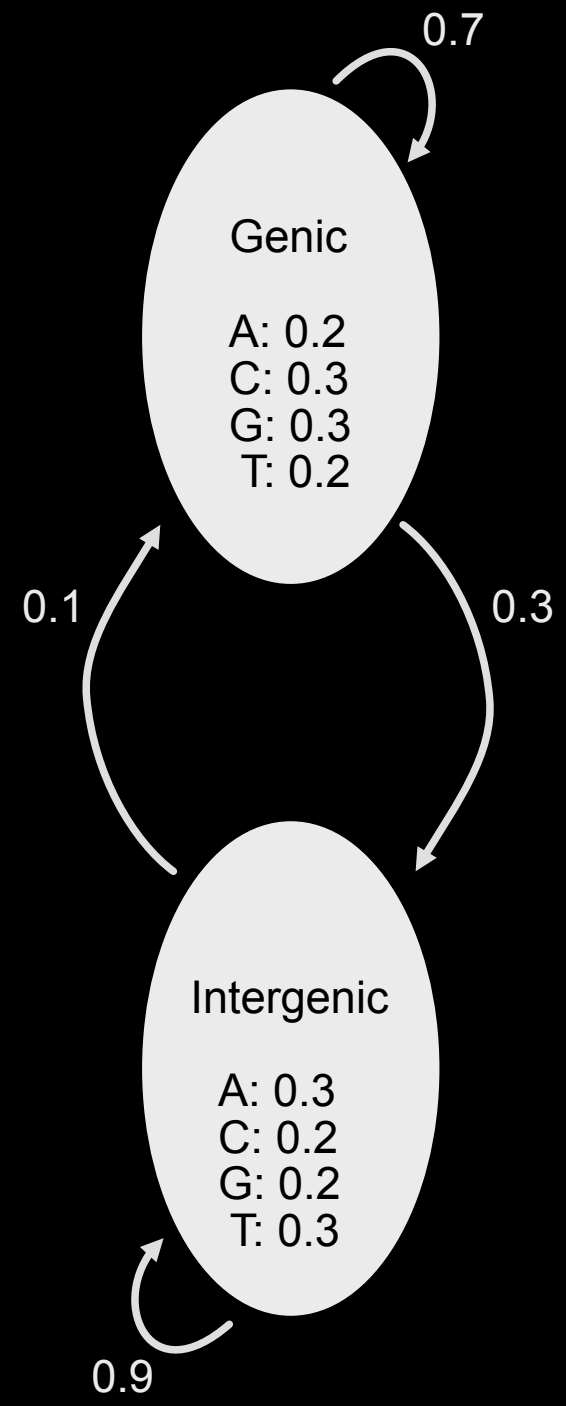


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x$$
$$p(C|genic)\ x\ p(genic\ \rightarrow inter)\ x\ p(T|inter)\ \ldots$$

$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ x\ 0.3\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3 = 5.58\ x\ 10^{-5}$$

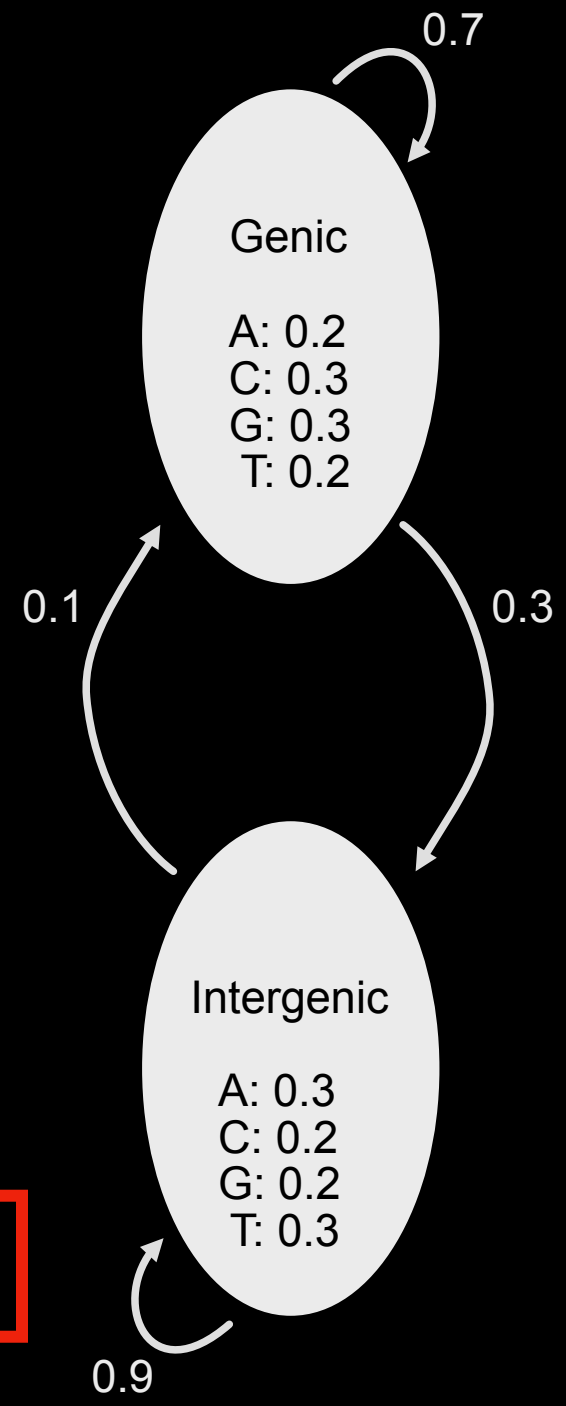


Calculating Probability of a State Path

- Observations: `GCTAAT`
- Proposed sequence of states: `GGIIII`
- What is the probability of our observations being generated by the state path `GGIIII`? → P(`GGIIII` | `GCTAAT`, Model)

$$p(genic|start)\ x\ p(G|genic)\ x\ p(genic\ \rightarrow genic)\ x$$
$$p(C|genic)\ x\ p(genic\ \rightarrow inter)\ x\ p(T|inter)\ \ldots$$

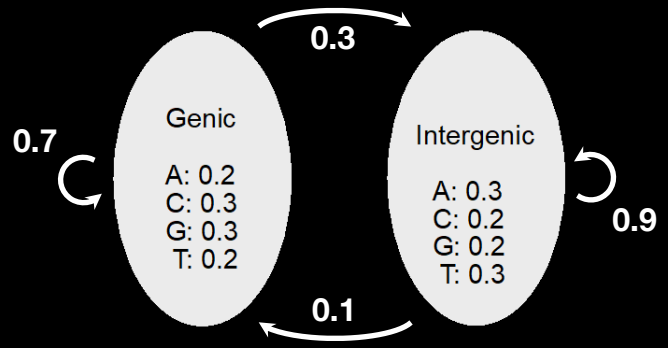
$$0.5\ x\ 0.3\ x\ 0.7\ x\ 0.3\ x\ 0.3\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3\ x\ 0.9\ x\ 0.3 = 5.58\ x\ 10^{-5}$$



What is the *most likely* state sequence given the observations?

What is the *most likely* state sequence given the observations?

Viterbi gives us the most probable path through the two hidden states (genic and intergenic) that could generate our observed sequence of nucleotides.

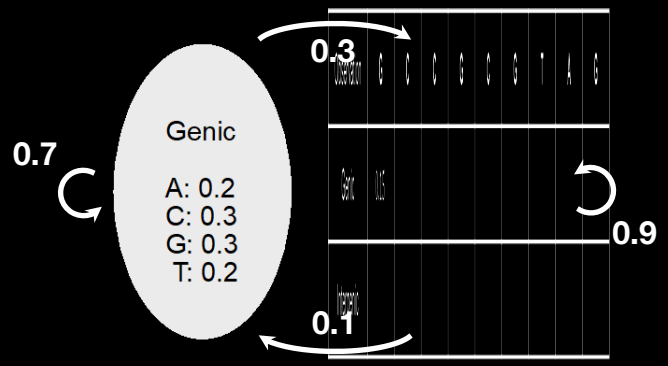


Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic									
Intergenic									

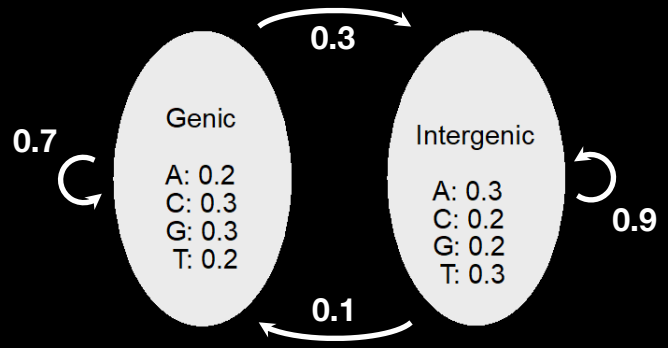
The probability that the first nucleotide is from a genic region is the chance that the sequence starts in a genic region, times the probability that the genic region emits a “G”.

The probability that the first nucleotide is from a genic region is $0.5 \times 0.3 = 0.15$



Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic									

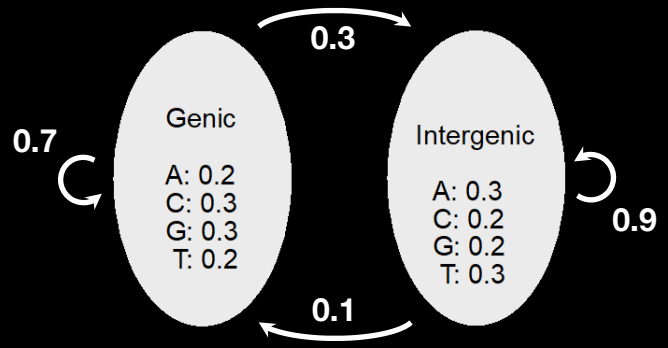


Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic									

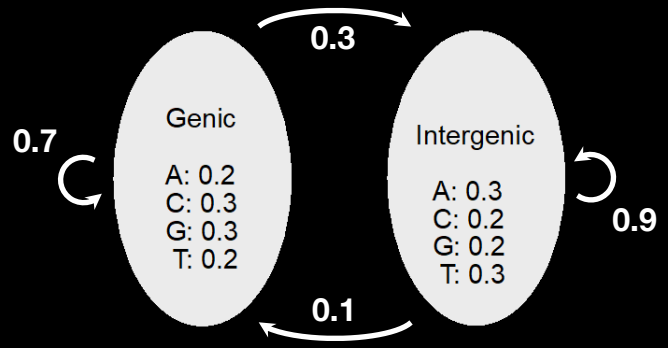
The probability that the first nucleotide is from an intergenic region is the chance that the sequence starts in an intergenic region, times the probability that the intergenic region emits a “G”.

The probability that the first nucleotide is from an intergenic region is:
$0.5 \times 0.2 = 0.1$



Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

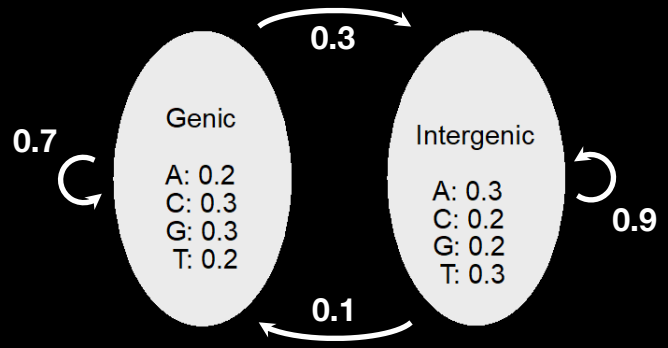


Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

The probability that the second nucleotide came from either a genic or intergenic region depends on the “state” of the first nucleotide.

We need to consider the probability of the previous state times the transition probability times the emission probability of a “C” from a genic region.



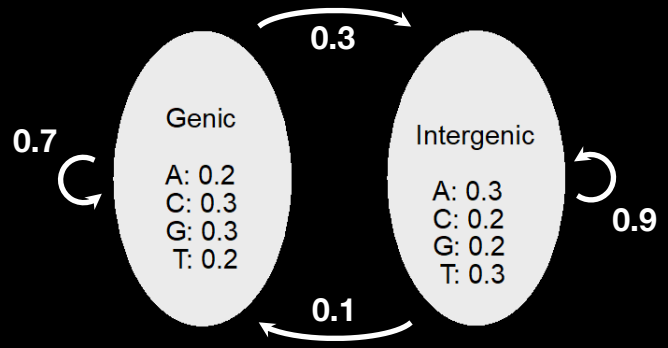
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

If the previous nucleotide was from a genic region, the transition probability is 0.7 (i.e., that we stayed in a genic region).

The emission probability of a “C” from a genic region is 0.3.

Therefore, the probability that this nucleotide is in a genic region AND that the previous nucleotide is in a genic region is $0.15 \times 0.7 \times 0.3 = 0.0315$



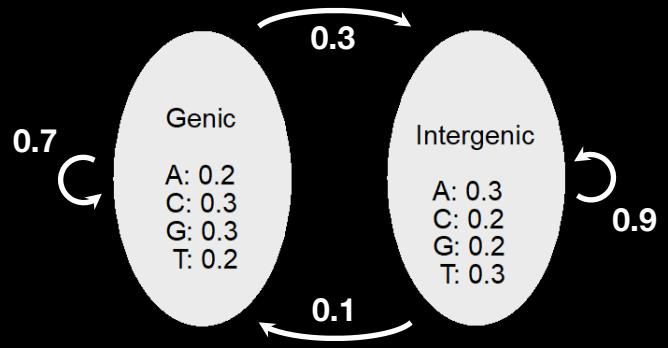
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

If the previous nucleotide was instead in an intergenic region, the transition probability is 0.1 (i.e. that we switched from intergenic to genic)

The emission probability of a “C” from a genic state is 0.3.

Therefore, the probability that this nucleotide is in a genic region AND that the previous nucleotide is in an intergenic region is $0.1 \times 0.1 \times 0.3 = 0.003$.



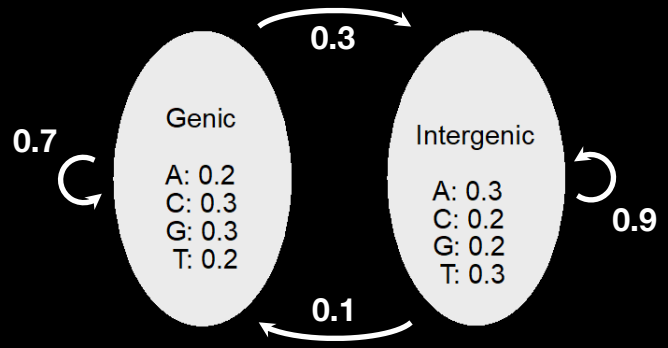
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

Remember, we're trying to find the MOST probable path through the HMM

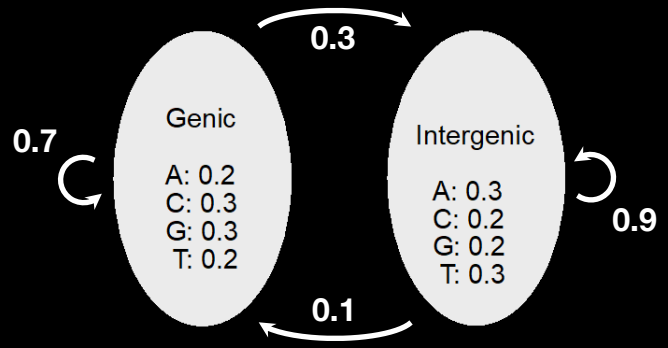
$P(GG) = 0.0315$ and $P(IG) = 0.003$

Because $P(GG)$ is higher, we can ignore all paths that start intergenic and switched to genic. We note that we're taking the **G → G path.**



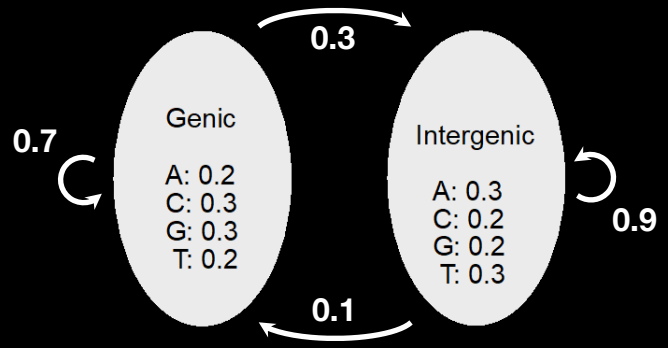
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15 ←	0.0315							
Intergenic	0.1								



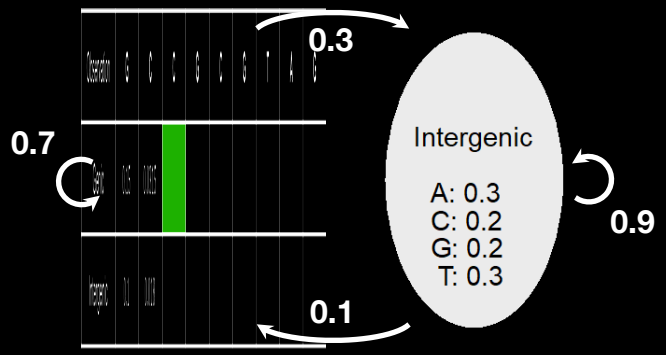
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0315							
Intergenic	0.1								



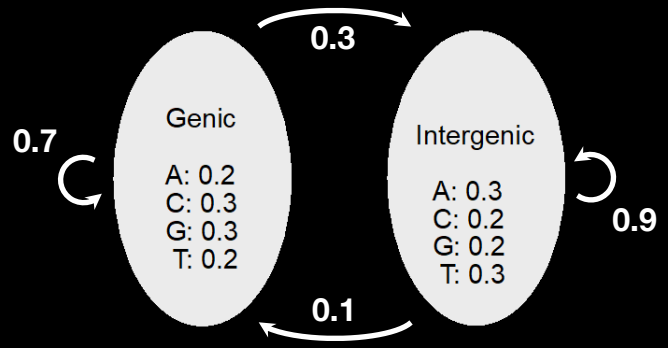
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15 ←	0.0315							
Intergenic	0.1 ←	0.018							



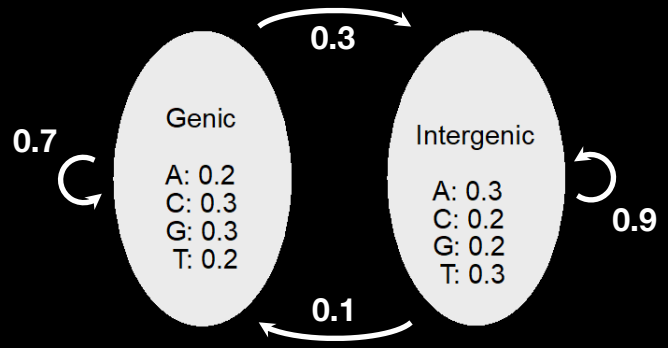
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15 ←	0.0315							
Intergenic	0.1 ←	0.018							



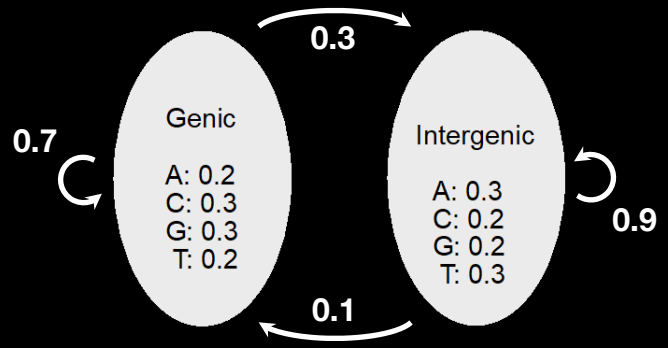
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15 ←	0.0315 ←	0.006615						
Intergenic	0.1 ←	0.018							



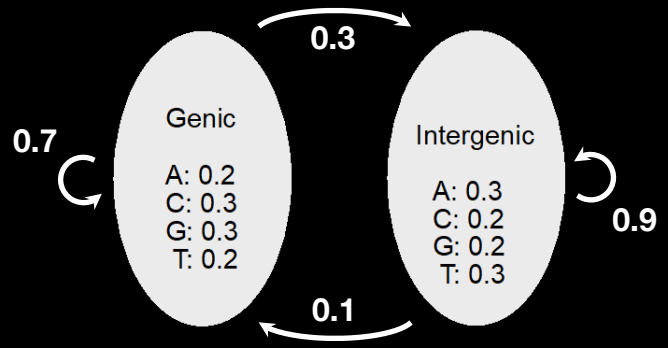
Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15 ←	0.0315 ←	0.006615						
Intergenic	0.1 ←	0.018 ←	0.00324						



Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15 ←	0.0315 ←	0.006615 ←	… ←	… ←	… ←	… ←	… ←	2.5×10^{-7}
Intergenic	0.1 ←	0.018 ←	0.00324 ←	… ←	… ←	…	… ←	… ←	2.7×10^{-7}



Viterbi Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0315	0.006615	…	…	…	…	…	2.5×10^{-7}
Intergenic	0.1	0.018	0.00324	…	…	…	…	…	2.7×10^{-7}

Most Probable Path: $GGGGGGIII$

$P(GGGGGGIII) = 2.7 \times 10^{-7}$

Probability of a Sequence of Observations

- The Viterbi algorithm allows us to find the most probable path through an HMM given some sequence of observations.
- But nature is by no means ideal… so it's also useful to be able to calculate the *overall* probability that our model generated some sequence of observations.
- This will help us determine how confident we are in our findings

Log Odds

Let's say our DNA sequence is GCCGTA. We can calculate the odds that our model generated it: p(GCCGTA | Model). We compare this to another model… say, a null model where there are no genes and it's all intergenic space.

p(GCCGTA | Model) = 2.54×10^{-4}

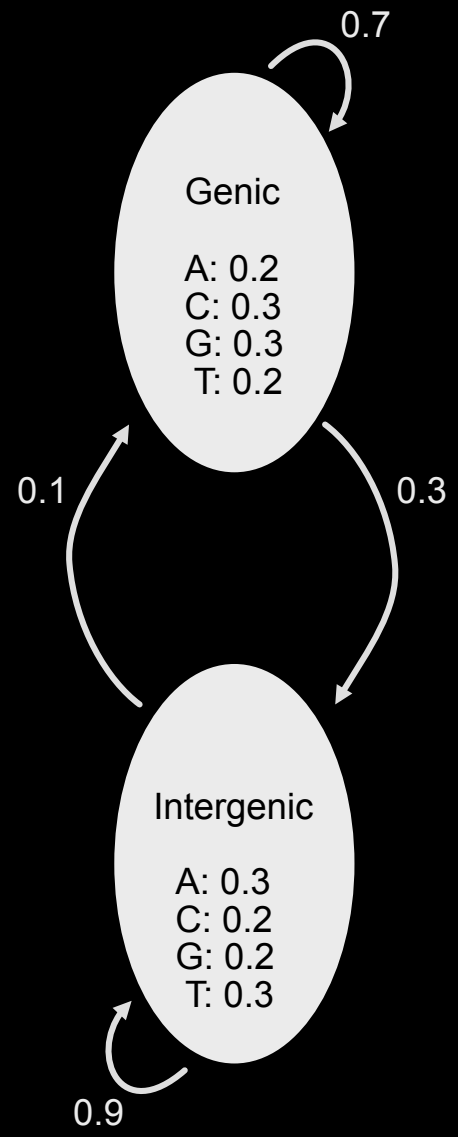
p(GCCGTA | Null) = 1.44×10^{-4}

Log(our model / null) = $\log(2.54 \times 10^{-4} / 1.44 \times 10^{-4})$ = 0.2465

This is useful because it centers the values around zero

If positive, numerator was more likely

If negative, denominator was more likely



Probability of a Sequence of Observations

- We understand now why it's useful to be able to get the full probability that a sequence of observations was generated by our model. But how do we do that?
- Let's say we have the sequence: "CAT"
- There are eight state paths that could have generated this sequence of observations: *GGG, GGI, GIG, GII, IGG, IGI, IIG, III*
- **For short sequences, we can easily calculate the probability of each of these state paths generating the sequence of observations and then sum those probabilities to get the total probability that the model generated that sequence of observations**
- **This falls apart for longer sequences, as the number of state paths to check is S^n where S is the number of states, and n is the number of observations**

How likely is this sequence, given our model of how the DNA works?

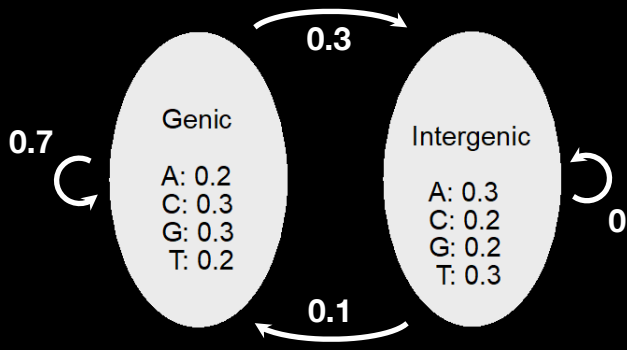
The Forward Algorithm

- Can we use dynamic programming to reduce the number of calculations we have to do to calculate the probability that our model generated a sequence of observations?

  YES!

- The forward algorithm is incredibly similar to the Viterbi algorithm

- On each iteration, instead of choosing the maximum probability path, we just sum the probabilities of all possible paths

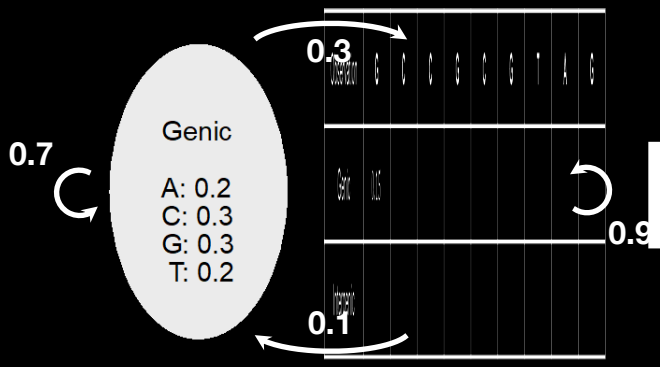


Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic									
Intergenic									

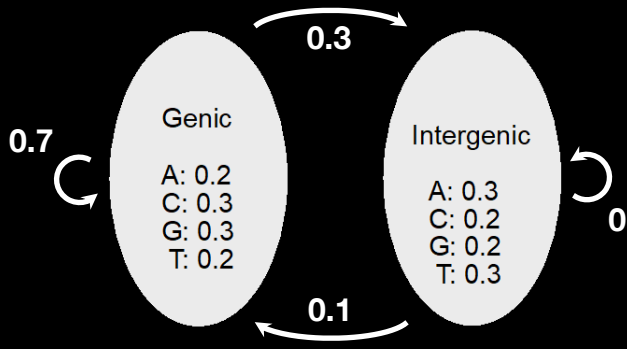
$$P(G, \text{Genic}) = \text{P}(\text{Genic}|\text{start}) \times P(G|\text{Genic})$$

$$P(G, \text{Genic}) = 0.5 \times 0.3 = 0.15$$



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic									

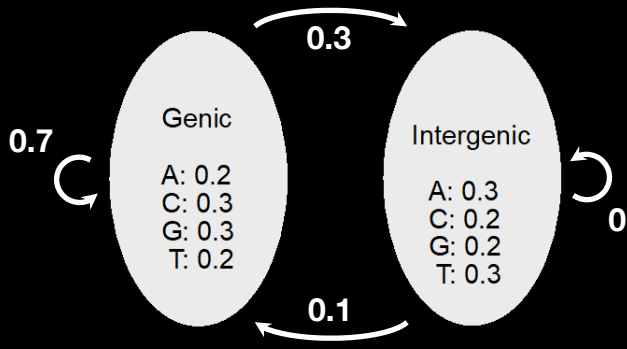


Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic									

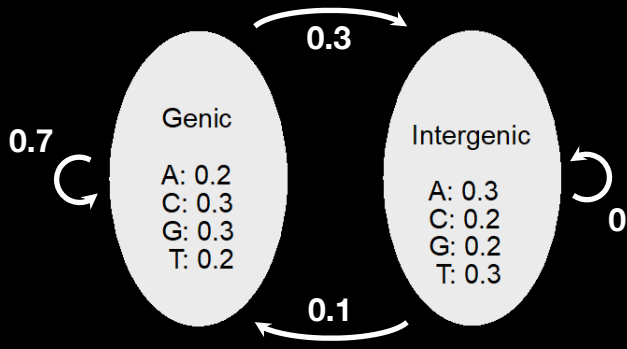
$$P(G, \text{Intergenic}) = \text{P}(\text{Intergenic}|\text{start}) \times P(G|\text{Intergenic})$$

$$P(G, \text{Intergenic}) = 0.5 \times 0.2 = 0.1$$



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								



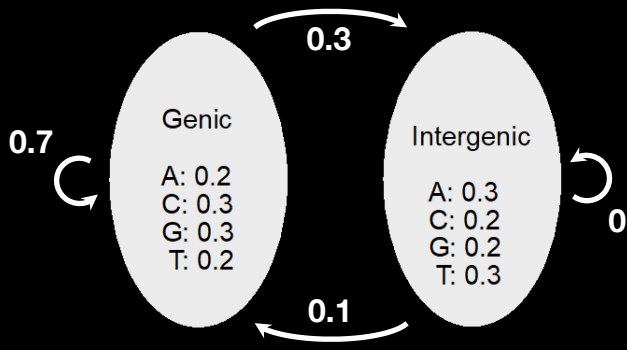
Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

When we were doing the Viterbi algorithm, we calculated the probability of emitting a “C” in the genic state given that the previous state was genic OR given that the previous state was intergenic

We then chose the maximum of the two, and kept track of the path

This time, we’re going to sum the probabilities



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15								
Intergenic	0.1								

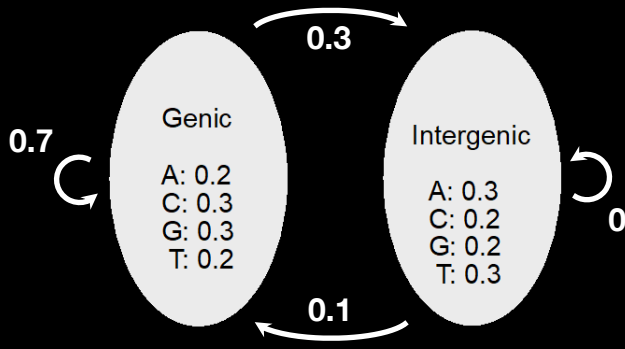
$$P(GC, \text{Genic} \rightarrow \text{Genic}) = P(G, \text{Genic}) \times P(\text{Genic} \rightarrow \text{Genic}) \times P(C|\text{Genic})$$

$$P(GC, \text{Genic} \rightarrow \text{Genic}) = 0.15 \times 0.7 \times 0.3 = 0.0315$$

$$P(GC, \text{Intergenic} \rightarrow \text{Genic}) = P(G, \text{Intergenic}) \times P(\text{Intergenic} \rightarrow \text{Genic}) \times P(C|\text{Genic})$$

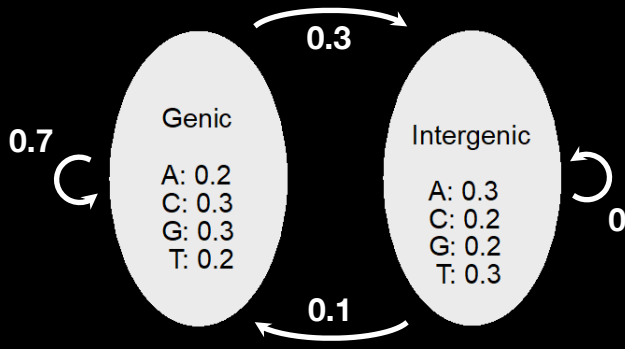
$$P(GC, \text{Intergenic} \rightarrow \text{Genic}) = 0.1 \times 0.1 \times 0.3 = 0.003$$

$$P(GC, \text{End on Genic}) = 0.0315 + 0.003 = 0.0345$$



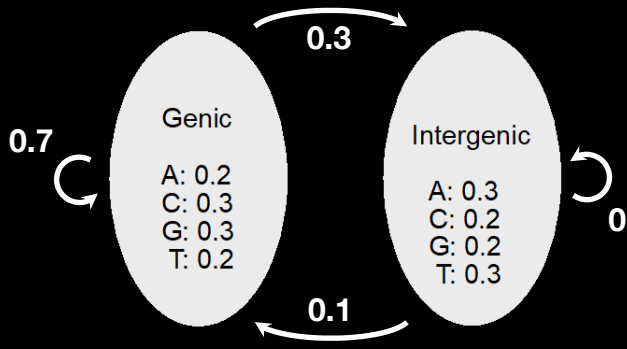
Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345							
Intergenic	0.1								



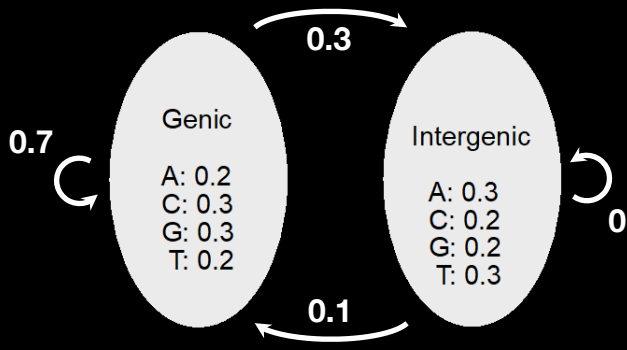
Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345							
Intergenic	0.1								



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345							
Intergenic	0.1	0.027							



Forward Algorithm

Observation	G	C	C	G	C	G	T	A	G
Genic	0.15	0.0345	0.008055	…	…	…	…	…	1.0×10^{-6}
Intergenic	0.1	0.027	0.00693	…	…	…	…	…	2.3×10^{-6}

$$P(GCCGCGTAG, \text{End on Genic}) = 1.012\times10^{-6}$$

$$P(GCCGCGTAG, \text{End on Intergenic}) = 2.296\times10^{-6}$$

$$\boldsymbol{P}(GCCGCGTAG) = 1.012\times10^{-6} + 2.296\times10^{-6} = 3.308\times10^{-6}$$

Back to Log Odds

The probability that our model generated the sequence: “GCCGCGTAG” is:

$$\boldsymbol{P}(GCCGCGTAG|\text{Full Model}) = 3.308\times10^{-6}$$

The probability that an *intergenic-only* model generated the sequence is:

$$0.2\times0.2\times0.2\times0.2\times0.2\times0.2\times0.3\times0.3\times0.2 = 1.152\times10^{-6}$$

So the log-odds are as follows:

$$\log(3.308\times10^{-6} \div 1.152\times10^{-6}) = 1.055$$

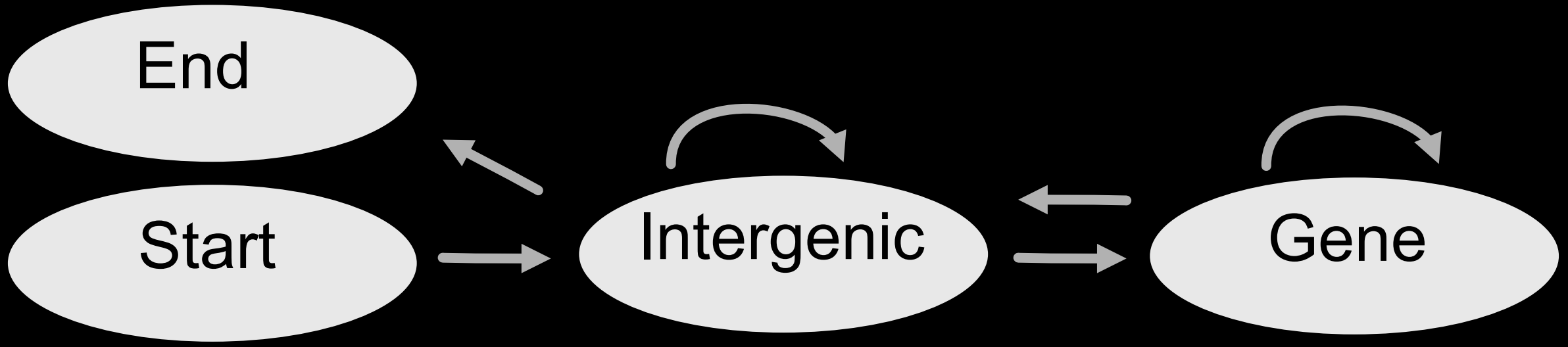
Applying Algs to Our Casino

- **Evaluation**: How likely is this sequence of observations, given our mode? **Forward Algorithm**
- **Decoding (1)**: What is the most likely sequence of states given the observations? **Viterbi Algorithm**

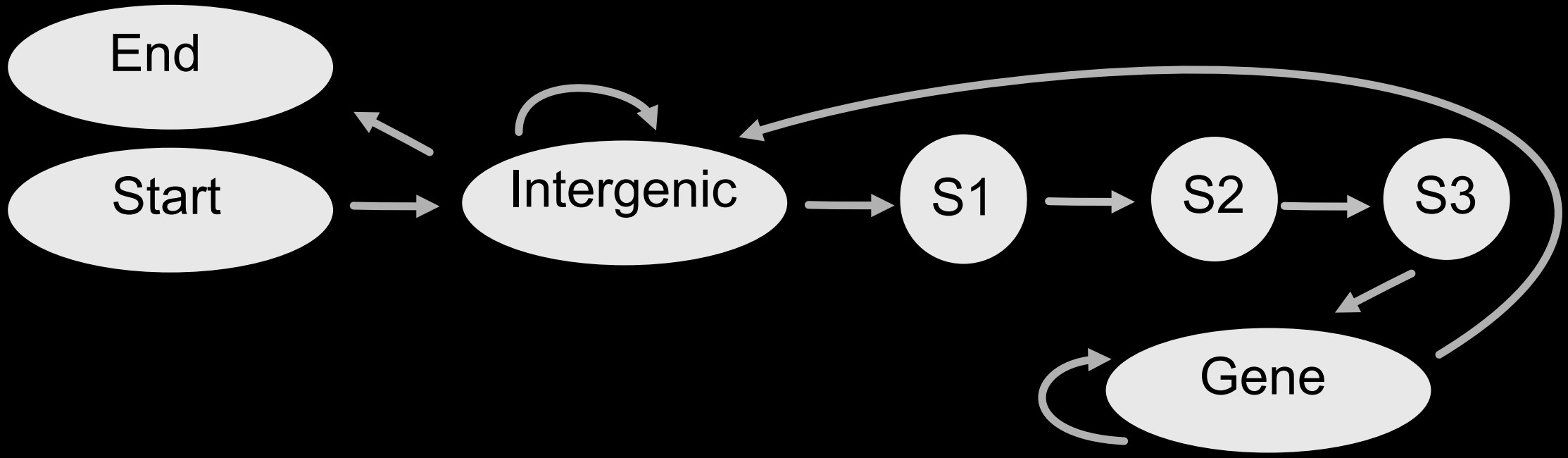
Applying Algs to Our Casino

- **Evaluation**: How likely is this sequence of observations, given our mode? **Forward Algorithm**
- **Decoding (1)**: What is the most likely sequence of states given the observations? **Viterbi Algorithm**
- **Decoding (2)**: What is the probability that the n^{th} observation was generated by a given state? **Forward-Backward Algorithm**
- **Learning**: What are the parameters of the model? **Baum-Welch Algorithm** (Expectation Maximization)

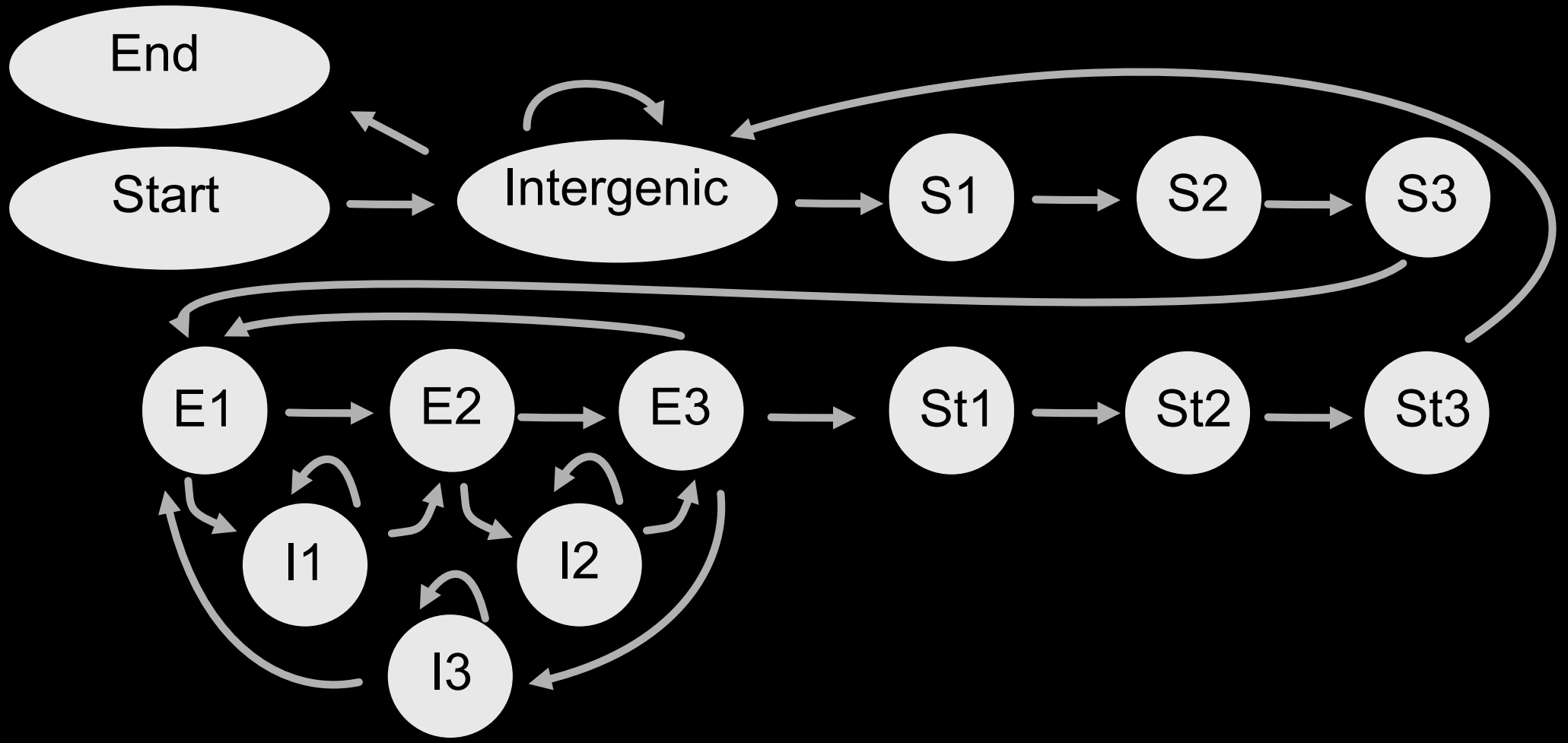
Genic vs. Intergenic



Genic vs. Intergenic



Genic vs. Intergenic



Markov when he hid the model idk I never studied system modelling

Thank you!

AGARA BIO

@tlrdln22
@saracarioscia

McCoy Lab