Biology 644

Old Title: Bioinformatics for Molecular Biologists

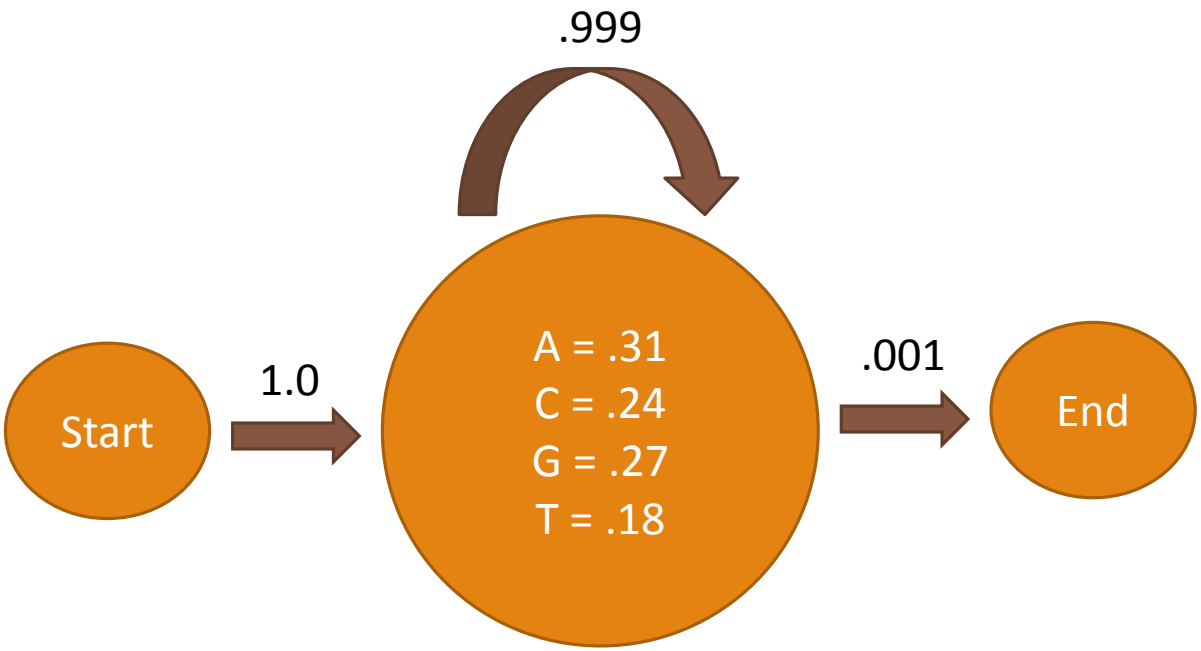
Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Markov Models

- A stochastic (probabilistic) model that assumes the Markov property
  - Markov property is satisfied when the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state; that is, given the present, the future does not depend on the past
  - Markov chain (discrete-time Markov chain or DTMC) undergoes transitions from one state to another in a state space.
    - It is a random process usually characterized as memory-less: the next state depends only on the current state and not on the sequence of events that preceded it.
  - A Hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states.
    - First used in speech and handwriting recognition
    - In biology, frequently used to predict gene tracks, splice sites, chromatin states, CpG islands, protein folding conformations, binding sites, CNV, etc.

Nth Order Markov Model

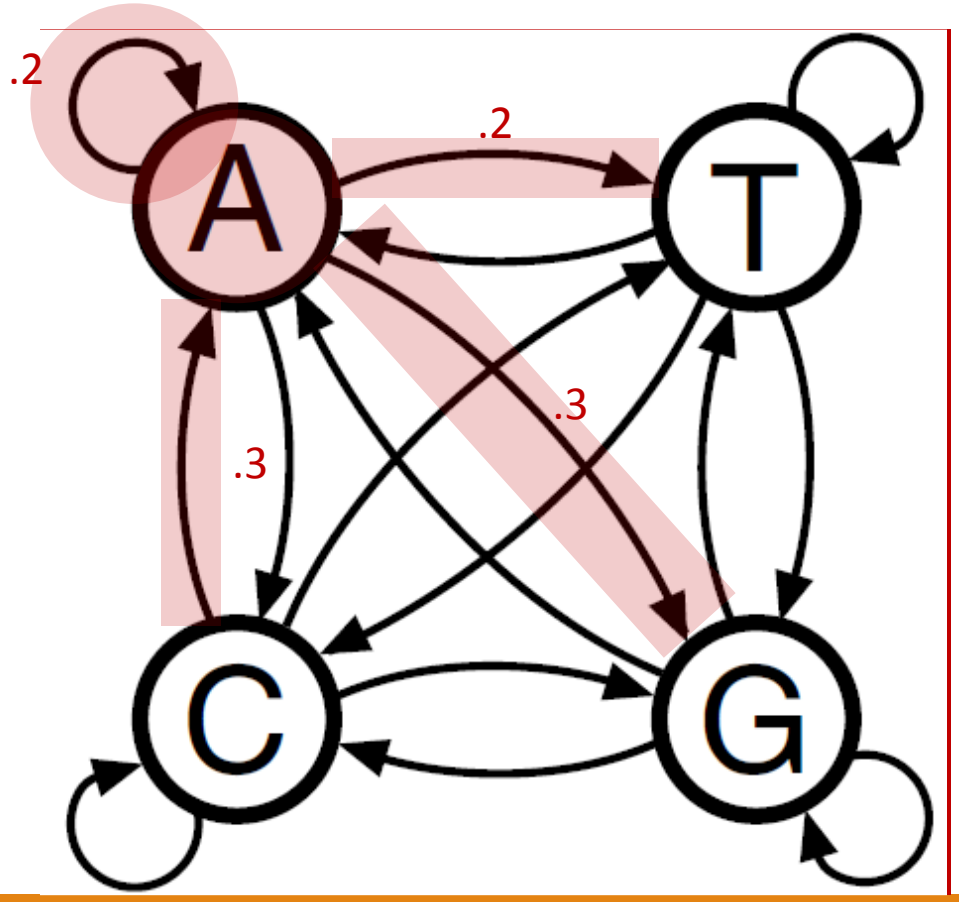
- **N refers to the amount of "sequence memory" in the model**
  - **How many positions in the sequence you have to look back to predict the next symbol**
- **Mononucleotide Content is a 0th Order Markov Model**
  - **Best visualized by a State Transition Diagram**



Random Walk: ATGCGAGATCAAGC….

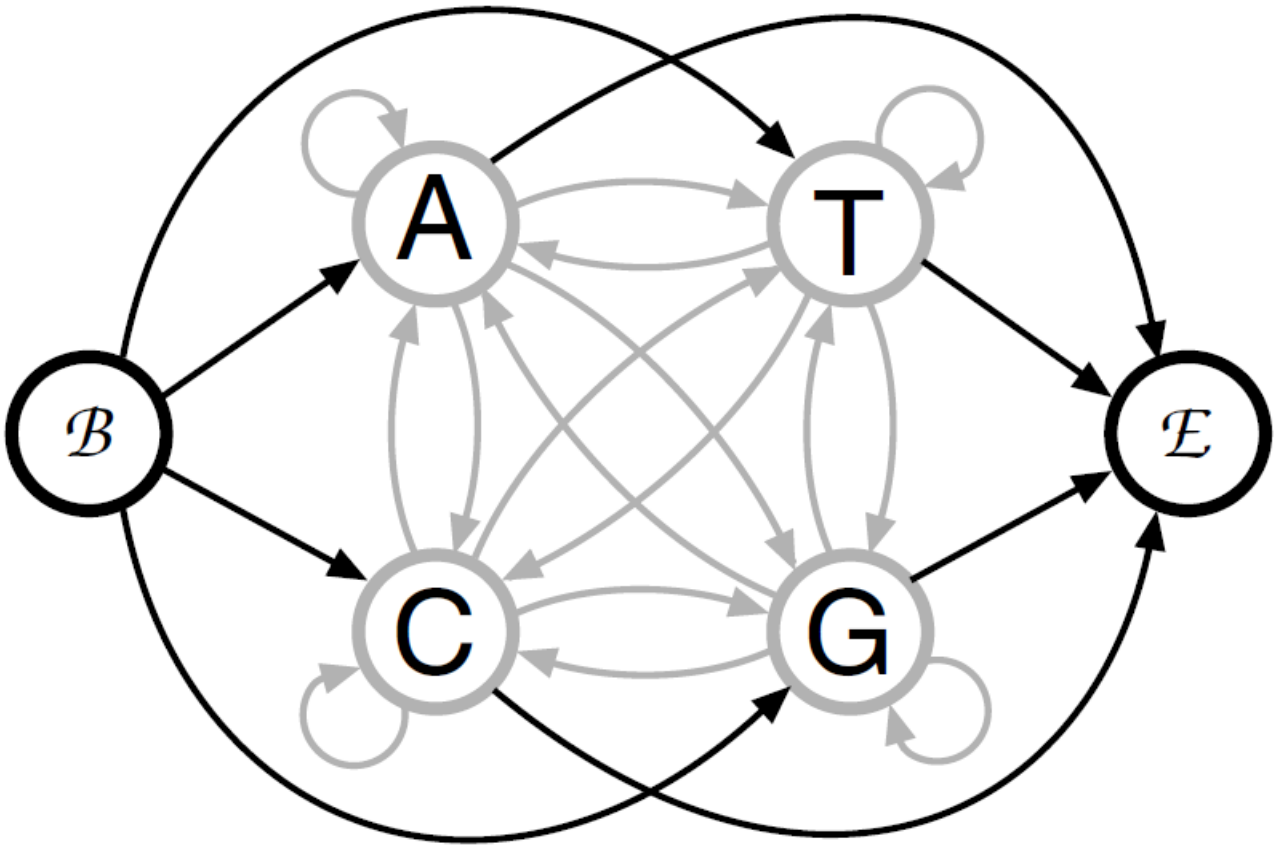
1st Order Markov Model

- Looks back 1 position in the sequence CGATCGATCGATAC….. to determine the probability of the next nucleotide
- Best visualized by a State Transition Diagram



1st Order Markov Model with Begin and End States

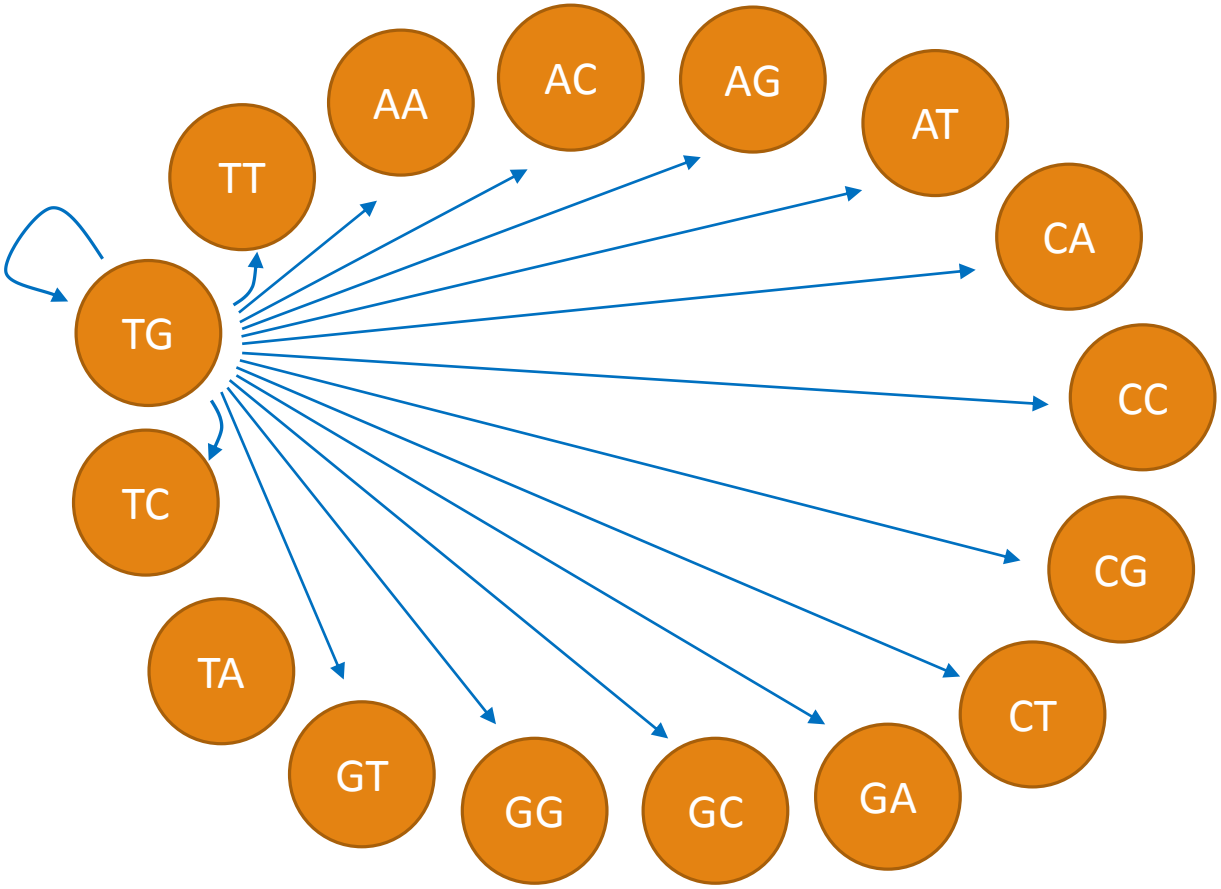
- **Add 2 additional States, Begin and End, that model the starting and ending behavior for a sequence**



2nd Order Markov Model

- Looks back 2 positions in the sequence CGATCGATCGATAC..... to determine the probability of the next nucleotide

- For 1st Order Model that were 4 states in the State Transition Diagram (not including the Begin and End States). How many will there be the 2nd Order Model? How many transitions will there be the 2nd Order Model?



Nth Order Markov Model

- In general, there are L^N states in a State Transition Diagram of an Nth Order Markov Model L is the size of the alphabet
  - A 5th Order Markov Model to model Genomic Nucleotide Biases will have 4^5 states
- In general, there are $(L^N)^2$ transitions (parameters to train) in an Nth Order Markov Model where L is the size of the alphabet
  - A 5th Order Markov Model to model Genomic Nucleotide Biases will have $(4^5)^2$ transitions to train
- Higher order Markov Models are able to capture longer range dependencies between the positions
  - If there aren't dependencies longer than N positions apart, then an (N+1)th Order Markov Model won't be any more accurate than an Nth Order Markov Model
- Higher order Markov Models become more difficult to train due to the exponential increase in the number of parameters to train
- State Transition Diagrams also become too large to draw out for higher order models

Nth Order Markov Model

- Generally, in order to computationally model Nth Order Markov Models we count (N+1)-mers and N-mers in order to build Markov Chains of Conditional Probabilities

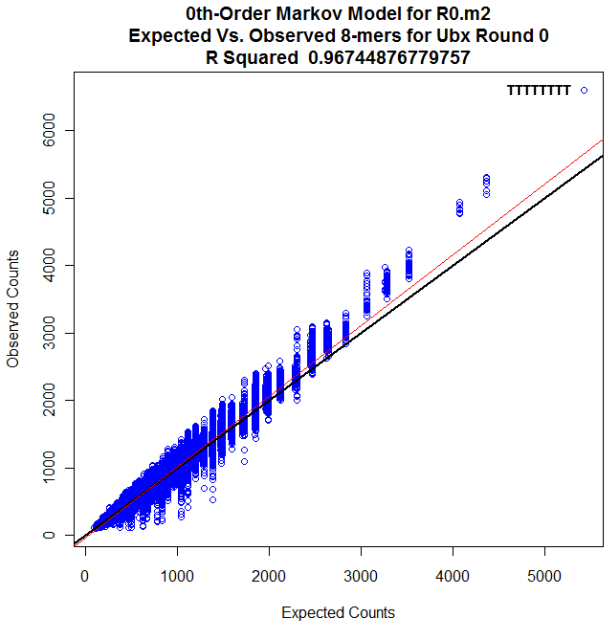
- 1st Order Markov Model Example
  - Looks back 1 position in the sequence CGATCGATCGATAC..... to determine the probability of the next nucleotide
  - Count 2-mers and 1-mers and calculate frequencies
  - $P(CGATCGATCGATAC) = P(CG) \cdot P(GA|G) \cdot P(AT|A) \cdot P(TC|T) \cdot P(CG|C) \cdot P(GA|G) \cdot \ldots$
  - $P(CGATCGATCGATAC) = P(CG) \cdot \frac{P(GA)}{P(G)} \cdot \frac{P(AT)}{P(A)} \cdot \frac{P(TC)}{P(T)} \cdot \frac{P(CG)}{P(C)} \cdot \frac{P(GA)}{P(G)} \cdot \ldots$

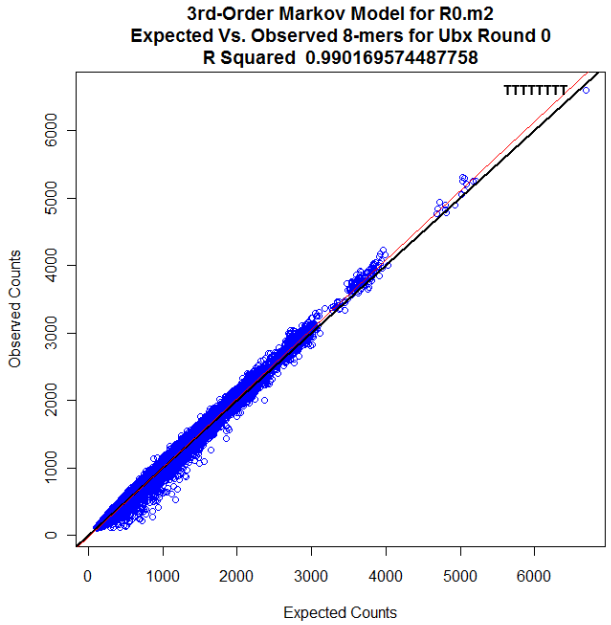
Nth Order Markov Model

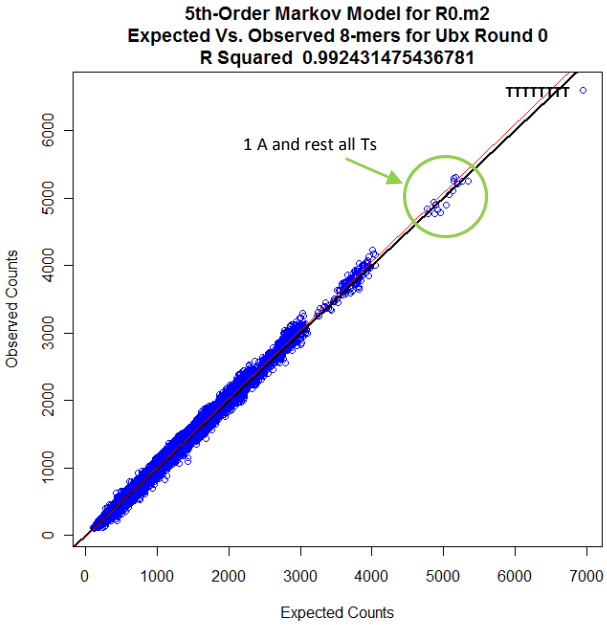
- Generally, in order to computationally model Nth Order Markov Models we count (N+1)-mers and N-mers in order to build Markov Chains of Conditional Probabilities

- 2nd Order Markov Model Example
  - Looks back 2 positions in the sequence CGATCGATCGATAC..... to determine the probability of the next nucleotide
  - Count 3-mers and 2-mers and calculate frequencies
  - $P(CGATCGATCGATAC) = P(CGA) \cdot P(GAT|GA) \cdot P(ATC|AT) \cdot P(TCG|TC) \cdot P(CGA|CG) \cdot P(GAT|GA) \cdot \ldots$
  - $P(CGATCGATCGATAC) = P(CGA) \cdot \frac{P(GAT)}{P(GA)} \cdot \frac{P(ATC)}{P(AT)} \cdot \frac{P(TCG)}{P(TC)} \cdot \frac{P(CGA)}{P(CG)} \cdot \frac{P(GAT)}{P(GA)} \cdot \ldots$

Using Nth Order Markov Models to Model Biases in Random Synthesized DNA







- The 3rd and 5th Order Markov Models are much better at recapitulating the observed 8-mer Counts in the "Random" Pool.
- The T -> A -> G -> C biases are due to melting rates in the PCR reactions and synthesis biases

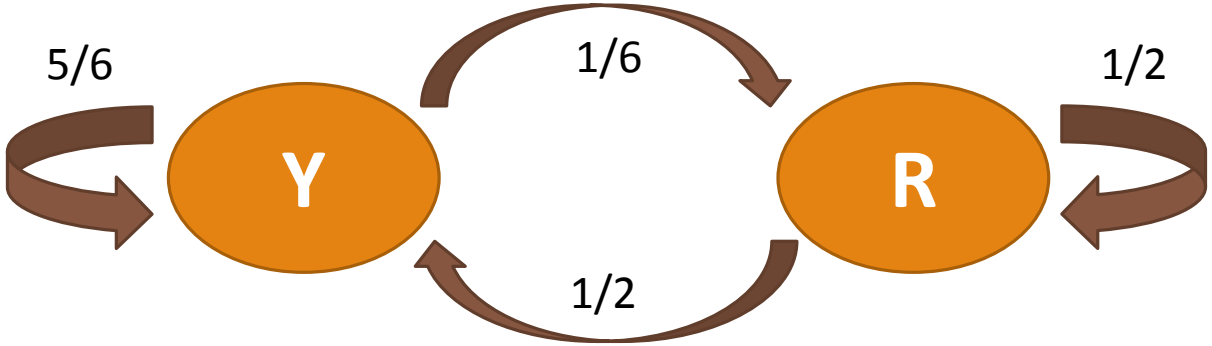
Probability Transition Matrix

- The transition probabilities of Markov Models can be conveniently organized into a matrix

To

From

	Y	R
Y	5/6	1/6
R	1/2	1/2



Random Walk: RYYYRYYYYRRYYYYY….

Probability Transition Matrix

- The transition probabilities of Markov Models can be conveniently organized into a matrix

To

From

	A	C	G	T
A	1/6	5/6	0	0
C	1/8	1/2	1/4	1/8
G	0	1/3	1/2	1/6
T	0	1/6	1/2	1/3

Random Walk: ACGGCGTGGCCGGCG….
What kind of DNA sequence might this transition matrix model?

Probability Transition Matrix

- The transition probabilities of Markov Models can be conveniently organized into a matrix

To

From	Y	R
Y	5/6	1/6
R	1/2	1/2

- Let $\boldsymbol{\pi}_0$ be the vector of initial probabilities, then $\boldsymbol{\pi}_n^T \boldsymbol{P} = \boldsymbol{\pi}_{n+1}^T$ represents the probability distributions at 1 time step:

$$\boldsymbol{\pi}_1^T = \boldsymbol{\pi}_0^T \boldsymbol{P} = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{5}{6} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{2}{3} \cdot \frac{5}{6} + \frac{1}{3} \cdot \frac{1}{2} & \frac{2}{3} \cdot \frac{1}{6} + \frac{1}{3} \cdot \frac{1}{2} \end{pmatrix} = \begin{pmatrix} \frac{13}{18} & \frac{5}{18} \end{pmatrix}$$

Probability Transition Matrix

- The transition probabilities of Markov Models can be conveniently organized into a matrix

To

From

	Y	R
Y	5/6	1/6
R	1/2	1/2

- And in general we have that $\boldsymbol{\pi}_n^T \boldsymbol{P} = \boldsymbol{\pi}_{n+1}^T$

- So from an initial probability vector $\boldsymbol{\pi}_0$, we can calculate the probability distributions at time n in the model by taking n powers of the probability transitions matrix P:

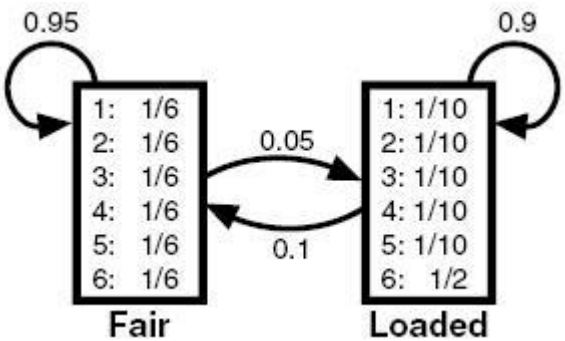
$$\boldsymbol{\pi}_0^T \boldsymbol{P}^n = \boldsymbol{\pi}_n^T$$

Stationary Distribution

- A probability distribution π satisfying $\pi^T = \pi^T P$ is stationary because the transition matrix does not change the probabilities of the states of the process
- Such a distribution exists and is unique if the Markov Chain is Ergodic or irreducible:
  - if it is possible to eventually get from every state to every other state with positive probability
- Describes the long term behavior of the process
- For any initial distribution π_0, it follows that $\pi_0 P^n$ tends to π^T as $n \to \infty$

Hidden Markov Models

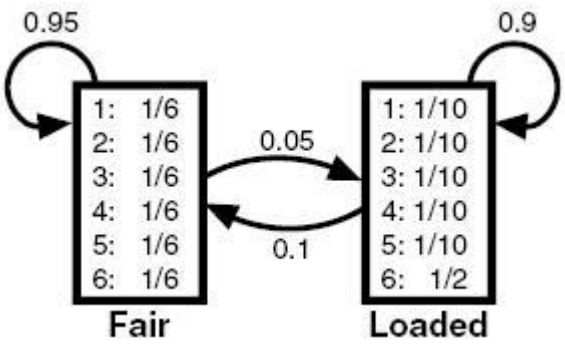
- When at least some of the data labels are missing (hidden) in the training data, then we must infer (label) the missing hidden states
  - Requires correctly inferring the model topology of the system
  - Requires a lot of training data to train the additional parameters
  - Successful training of the parameters is highly dependent on the initial conditions
  - Famous example: Occasionally Dishonest Casino



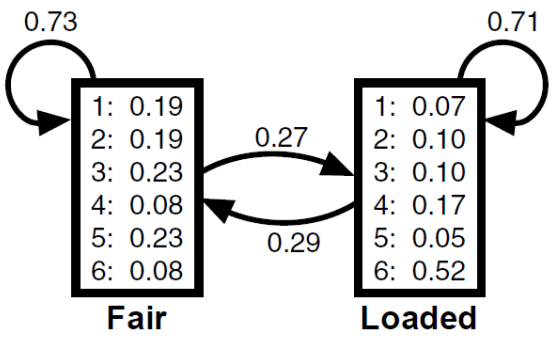

```
Rolls    3151162464466442453113216311641521336251445436316566265666 66
Die      FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLL

Rolls    6511664531326512456366646316366631623264552362666666251516 31
Die      LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFFFFF
Viterbi  LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF

Rolls    2225554416665665635643243641315134651463534111264146262533 56
Die      FFFFFFFFLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL

Rolls    3661636664662325344136616611632525624622552652522664353533 36
Die      LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi  LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls    2331216253644144323351632436336655624666626326661235524524 2
Die      FFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFFF
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLFFFFFFFFFFF
```

Hidden Markov Models

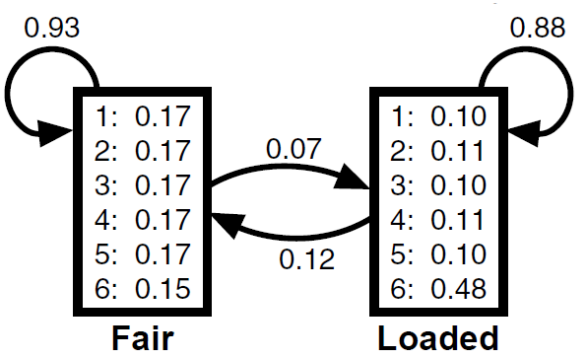
- When at least some of the data labels are missing (hidden) in the training data, then we must infer (label) the missing hidden states
  - Requires correctly inferring the topology of the system
  - Requires a lot of training data to train the additional parameters
  - Successful training of the parameters is highly dependent on the initial conditions
  - Famous example: Occasionally Dishonest Casino



Used to generate rolls

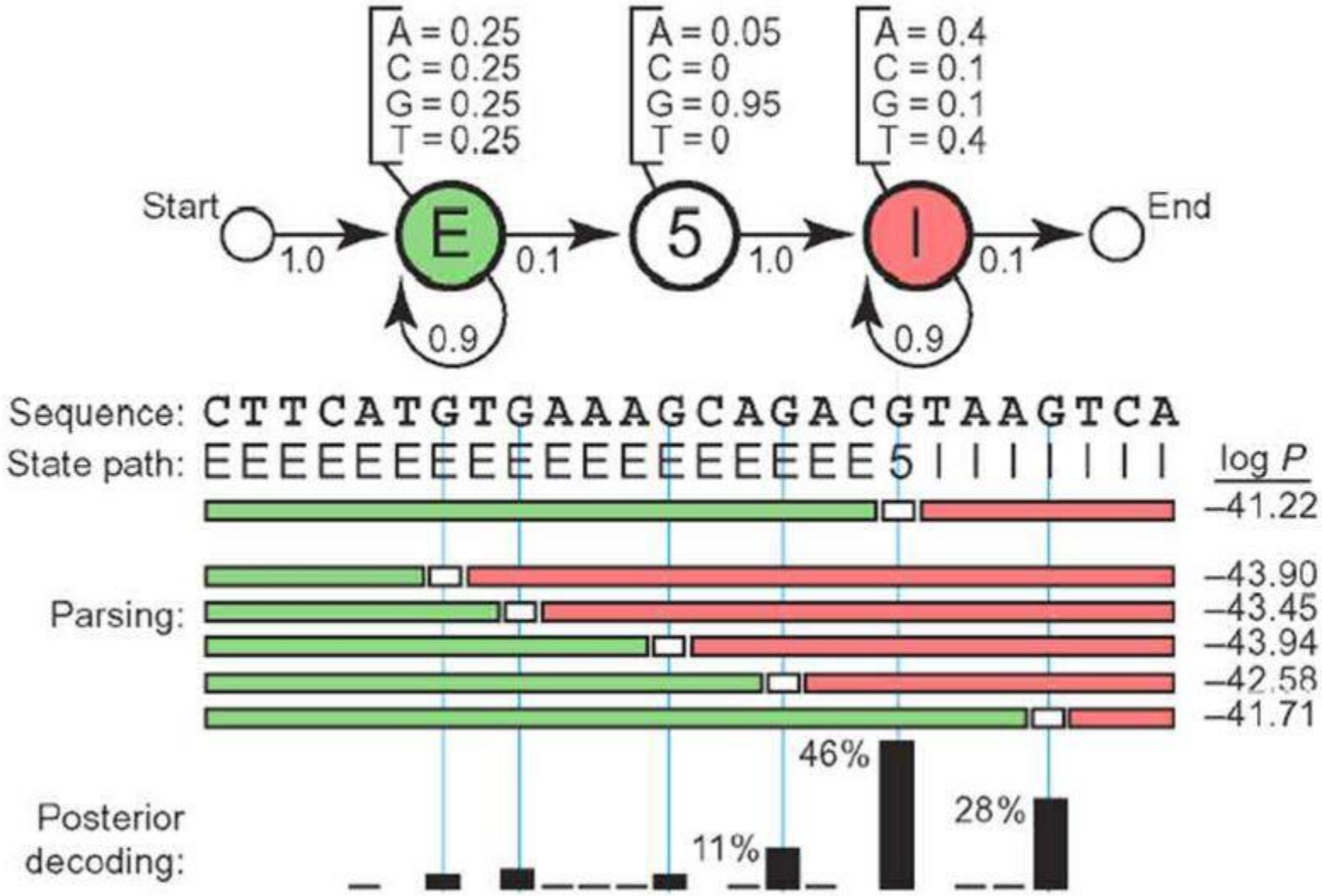


Training with 300 rolls



Training with 30,000 rolls

Using HMMs in 5' splice site recognition



Using HMMs to identify the CpG islands