
# Hidden Markov Models

## BIOL 7711

## Computational Bioscience

Biochemistry and Molecular Genetics
Computational Bioscience Program
Consortium for Comparative Genomics
University of Colorado School of Medicine

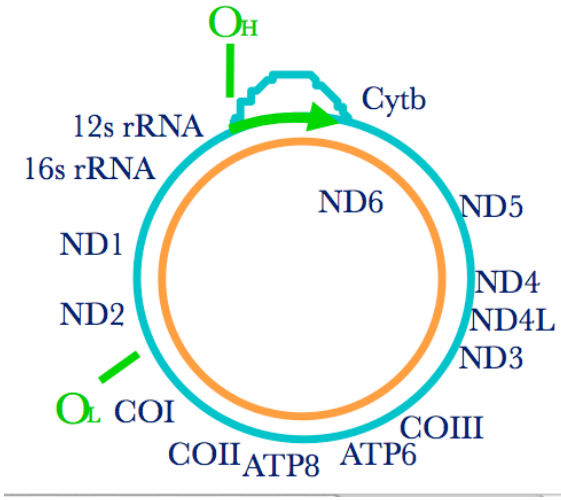
David.Pollock@uchsc.edu
www.EvolutionaryGenomics.com

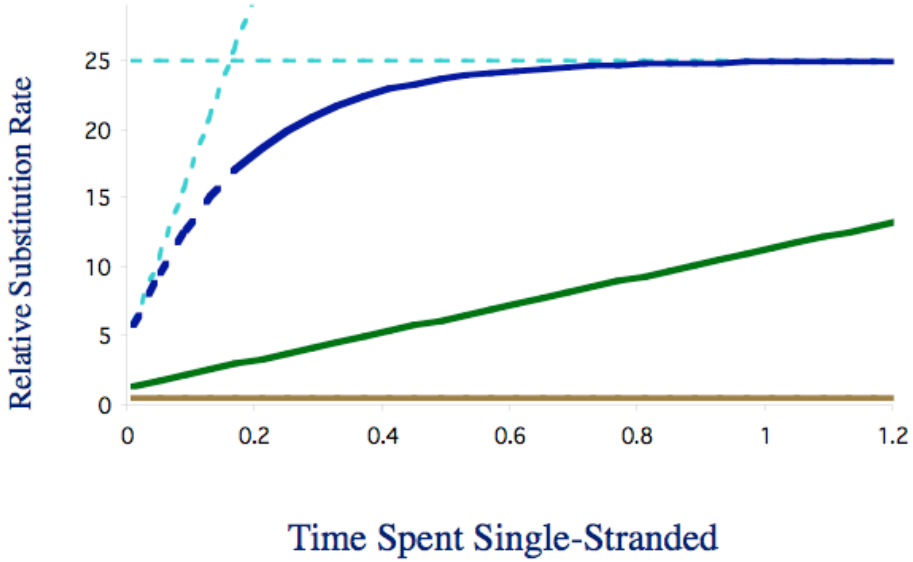
Evolution
Genomics
Biodiversity

# Why a Hidden Markov Model?

- Data elements are often linked by a string of connectivity, a linear sequence
  - Secondary structure prediction (Goldman, Thorne, Jones)
  - CpG islands

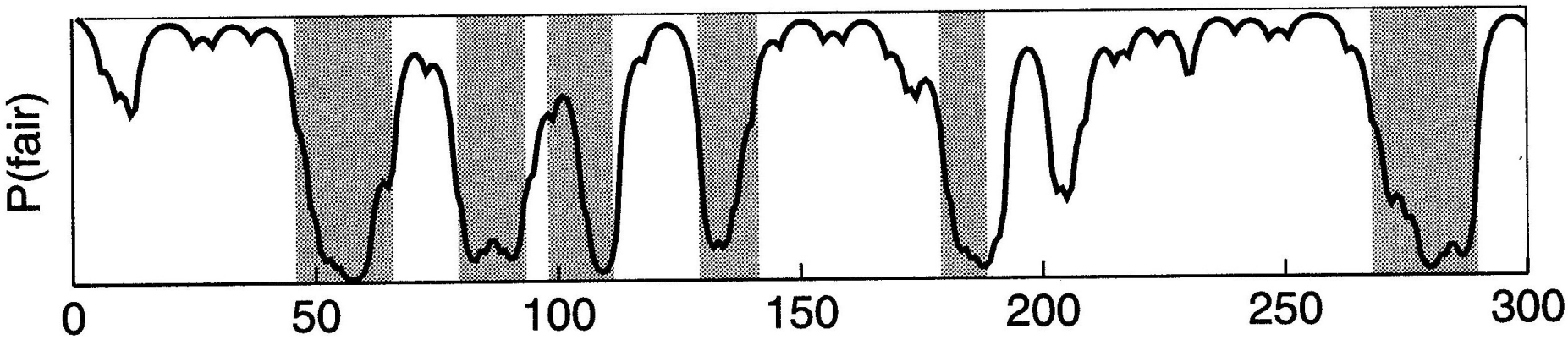
  - Models of exons, introns, regulatory regions, genes
  - Mutation rates along genome



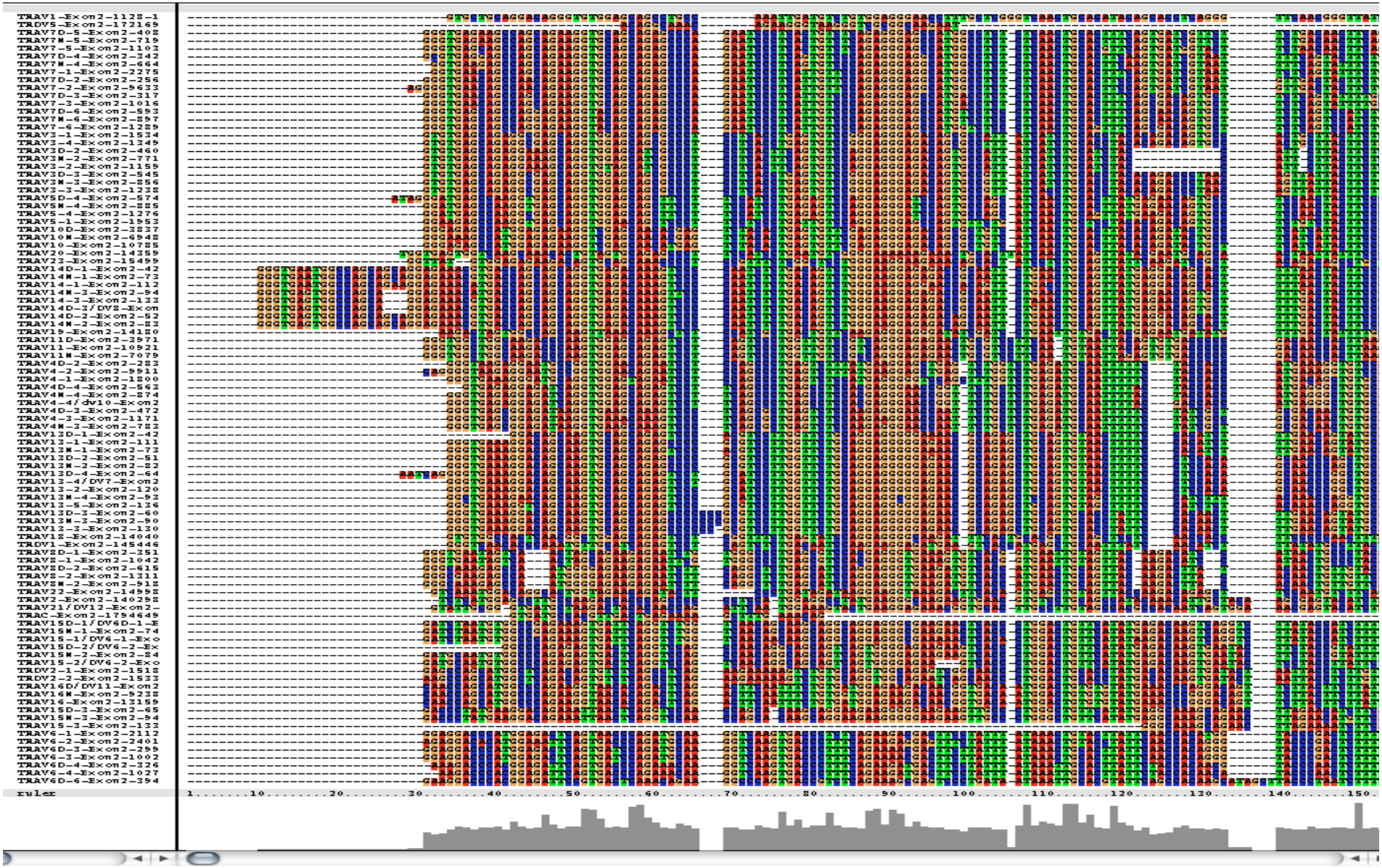
# Occasionally Dishonest Casino

# Posterior Probability of Dice



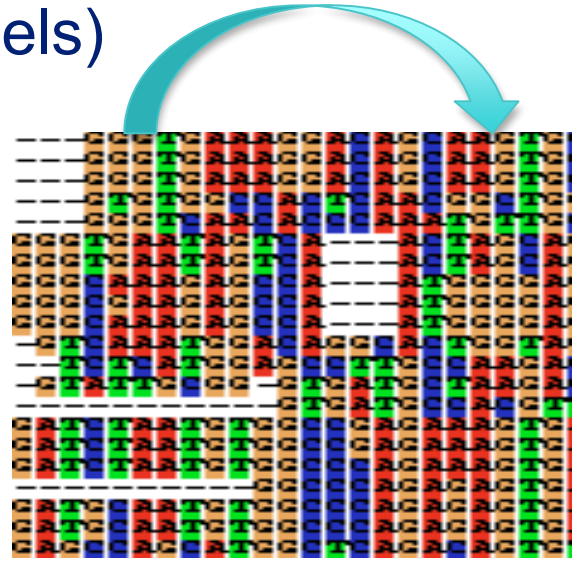
# Sequence Alignment Profiles

## Mouse TCR Vα



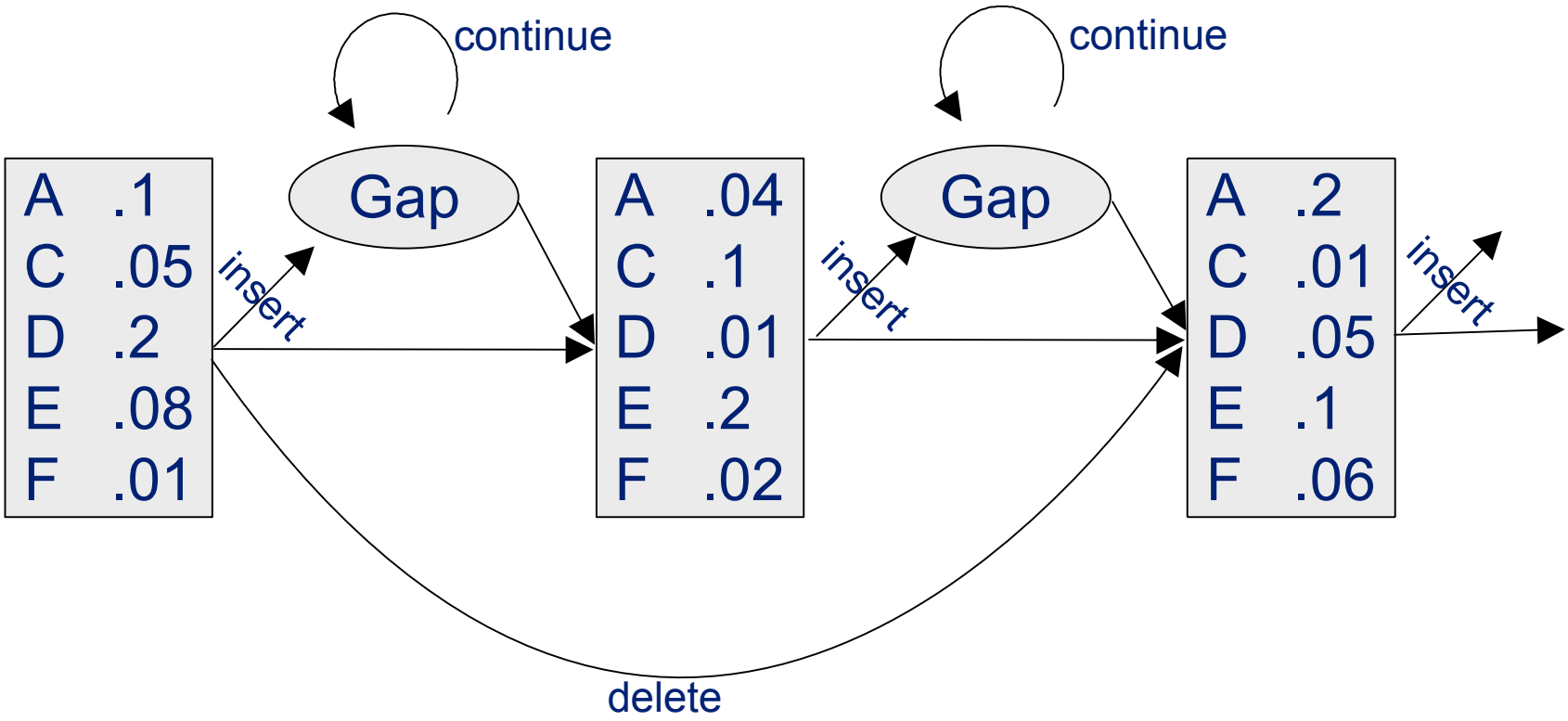
# Hidden Markov Models: Bugs and Features

- Memoryless
- Sum of states is conserved (rowsums =1)
- Complications?
  - Insertion and deletion of states (indels)
  - Long-distance interactions

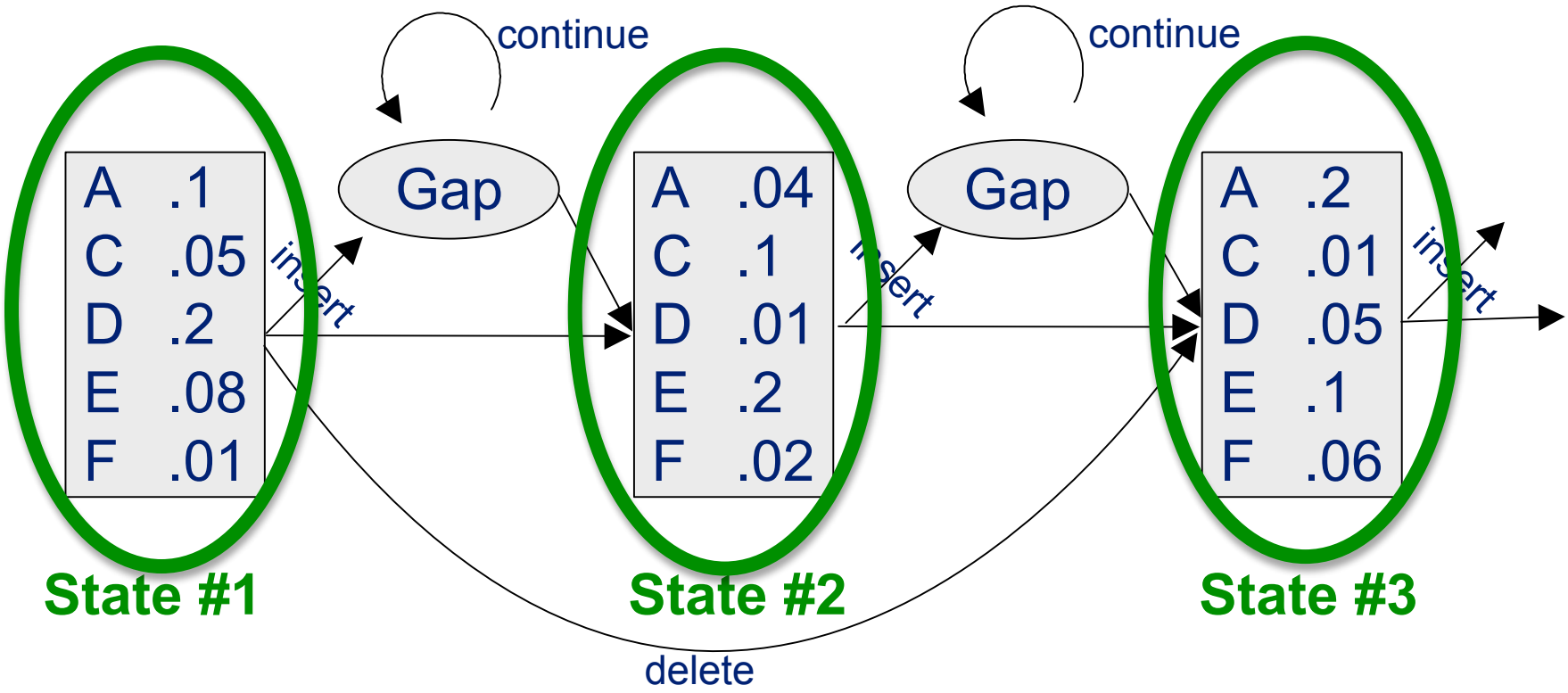


- Benefits
  - Flexible probabilistic framework
    - E.g., compared to regular expressions

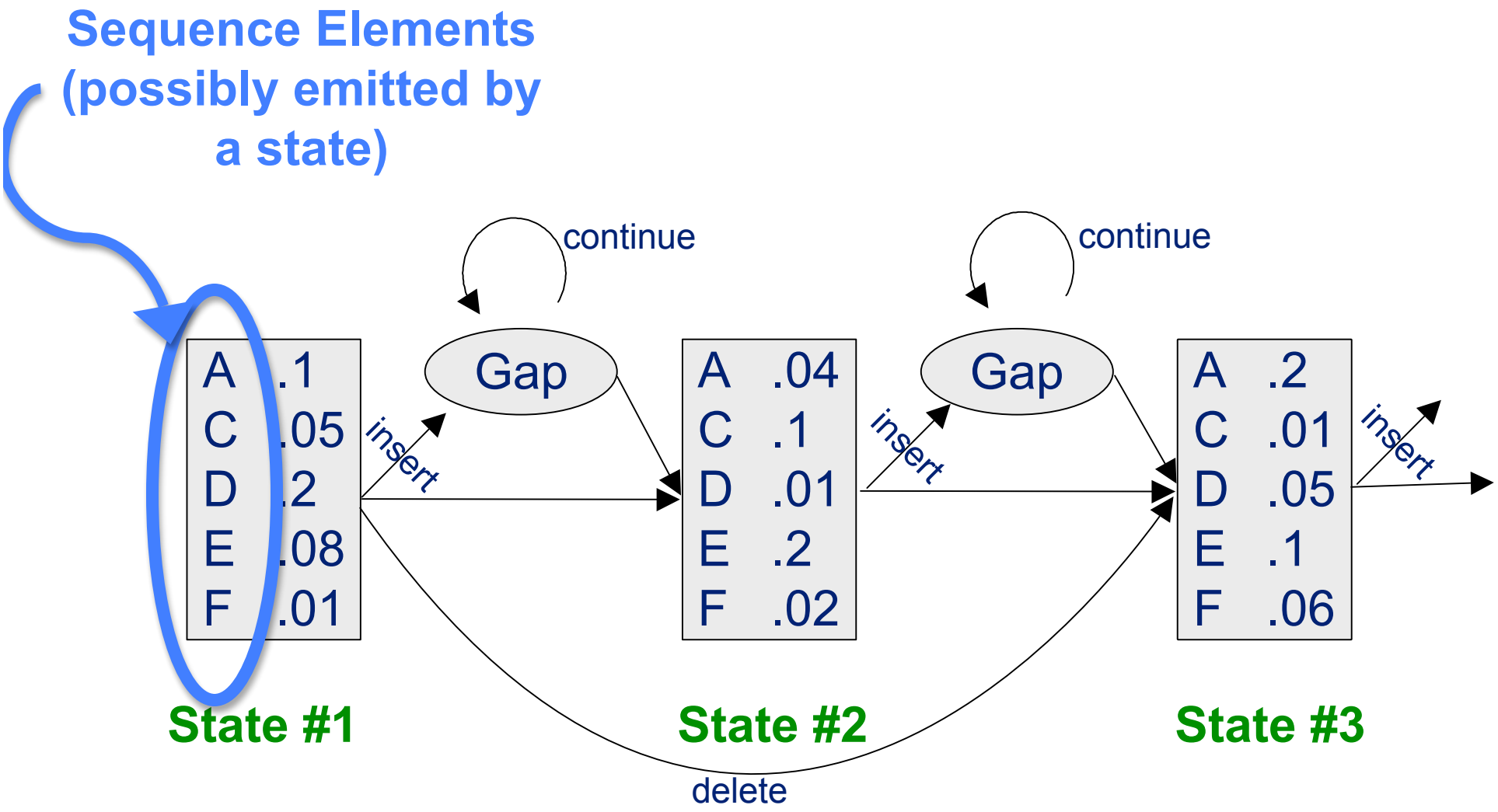
# Profiles: an Example



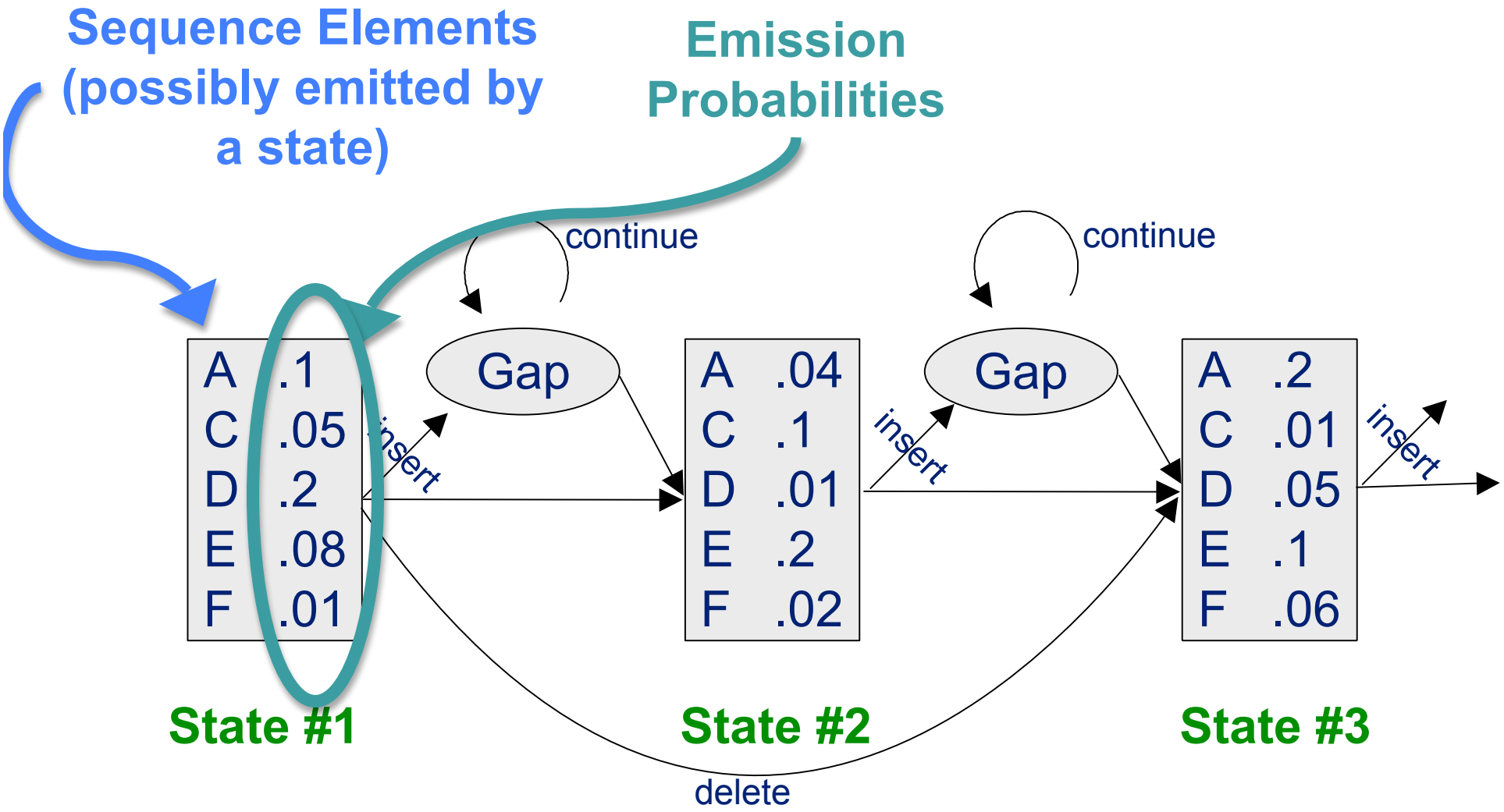
# Profiles, an Example: States



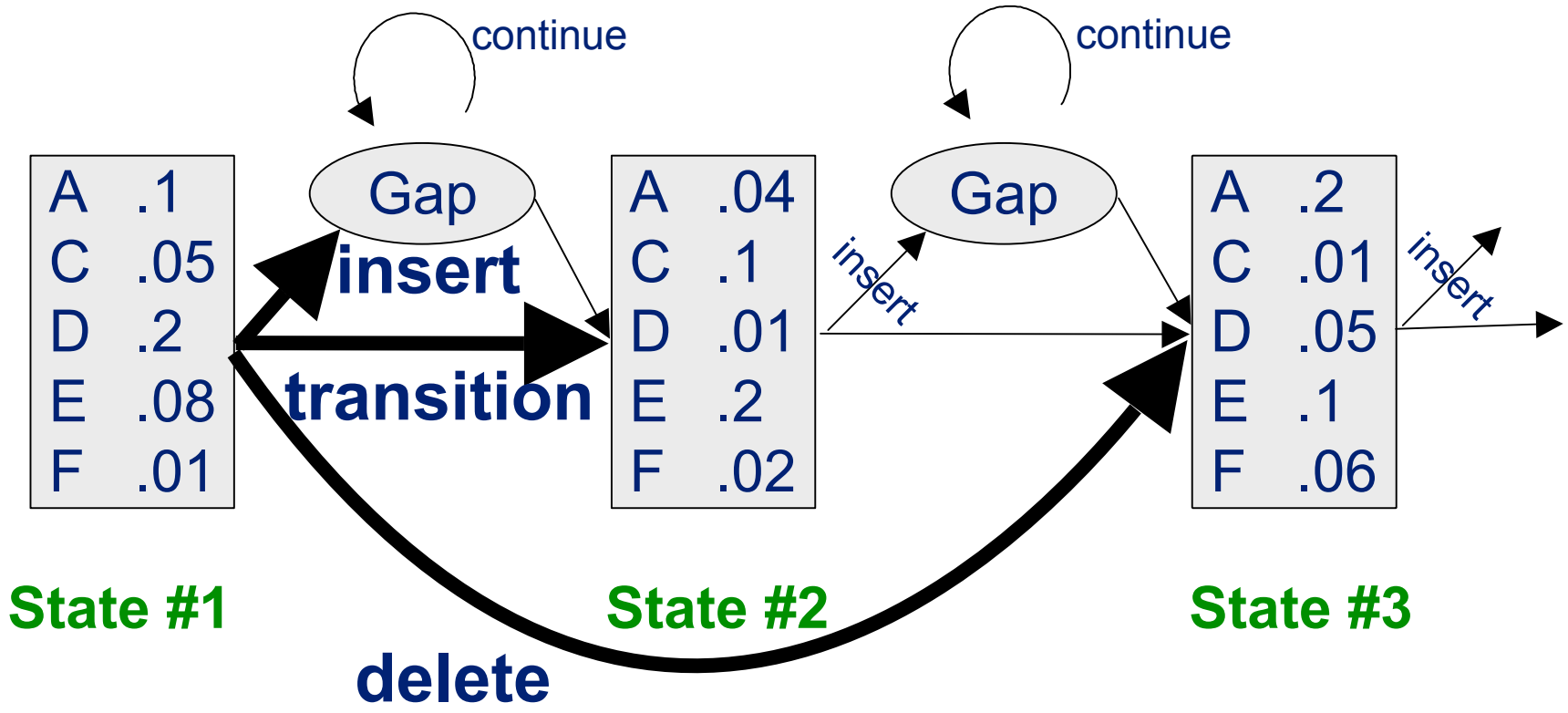
# Profiles, an Example: Emission



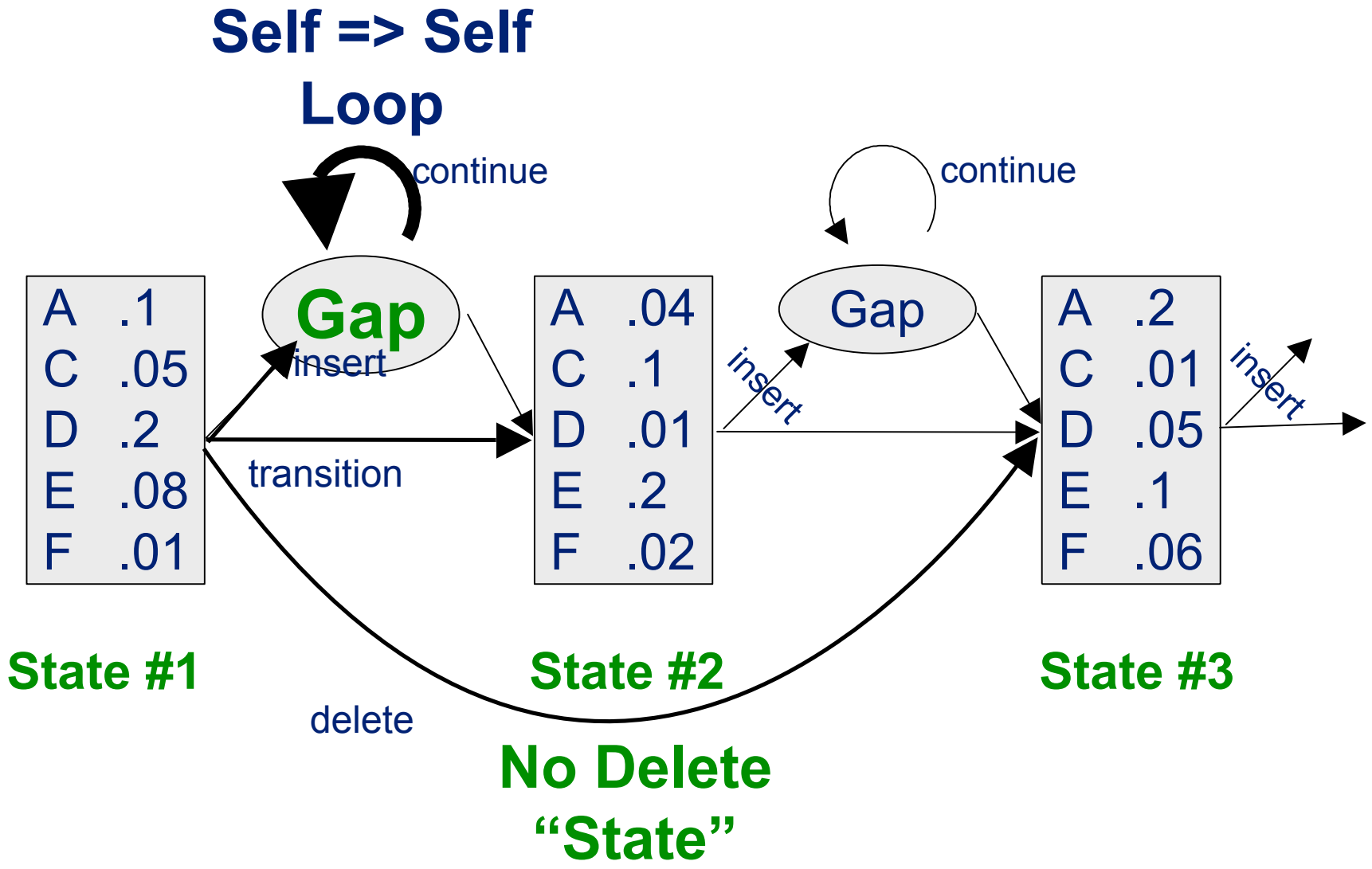
# Profiles, an Example: Emission

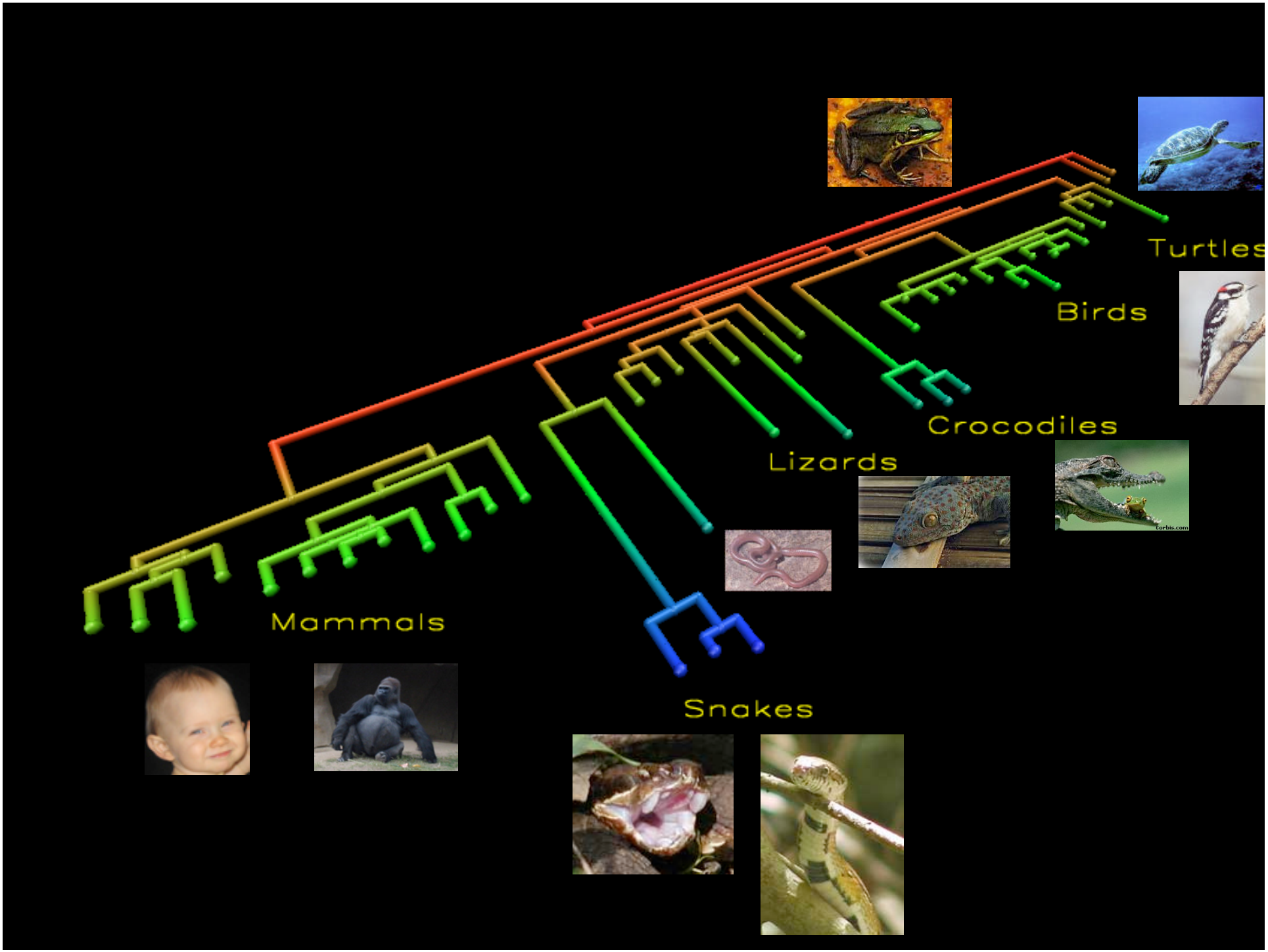


# Profiles, an Example: Arcs



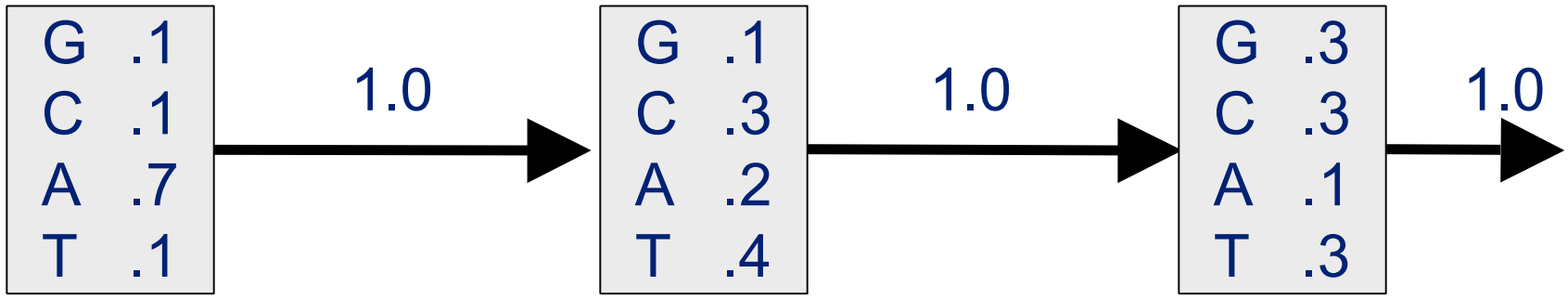
# Profiles, an Example: Special States




Turtles
Birds
Crocodiles
Lizards
Mammals
Snakes

# A Simpler not very Hidden MM

## Nucleotides, no Indels, Unambiguous Path



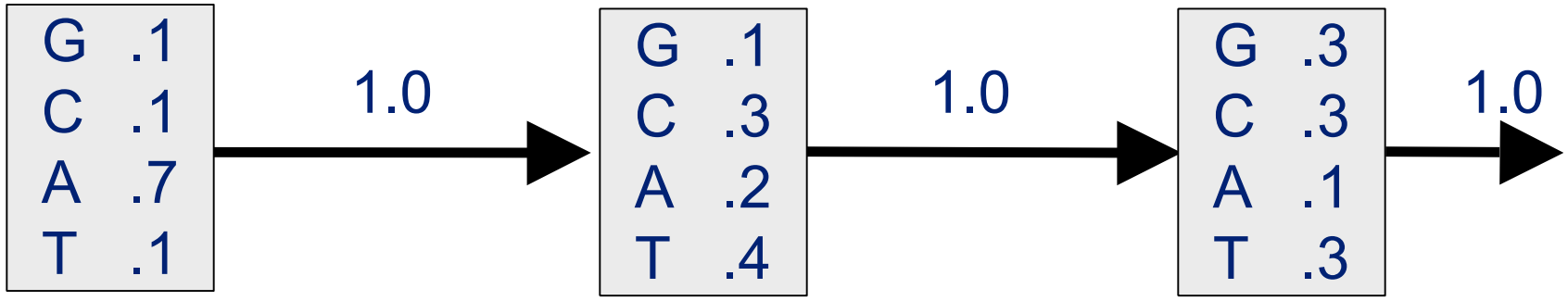

**A** **T** **T**

**0.7** **0.4** **0.3**

$$P(D \mid M) = 0.7 * 1.0 * 0.4 * 1.0 * 0.3 * 1.0$$

# A Simpler not very Hidden MM

## Nucleotides, no Indels, Unambiguous Path




**A** **T** **T**

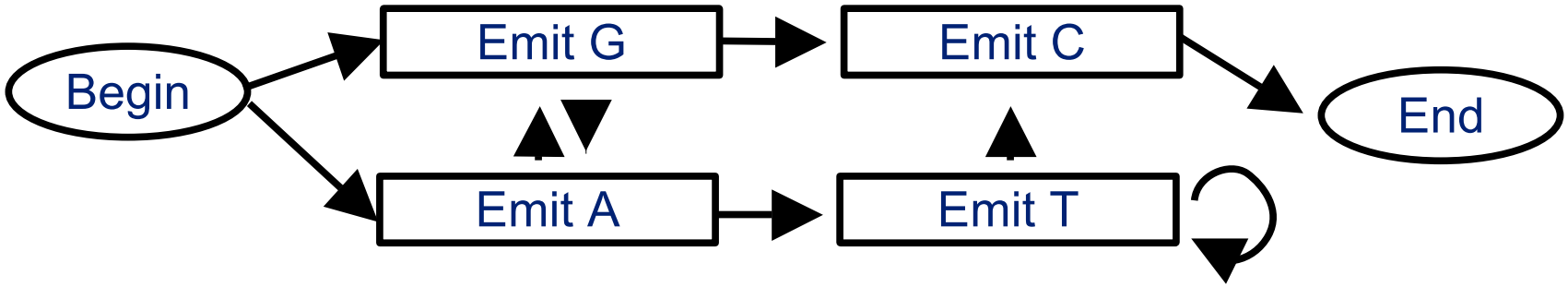
**0.7** **0.4** **0.3**

$$\ln P(D \mid M) = \sum_{states} \ln P(E_D \mid state) + \sum_{arcs} \ln P(x -> y)$$

# A Toy not-Hidden MM

## Nucleotides, no Indels, Unambiguous but Variable Path

## All arcs out are equal



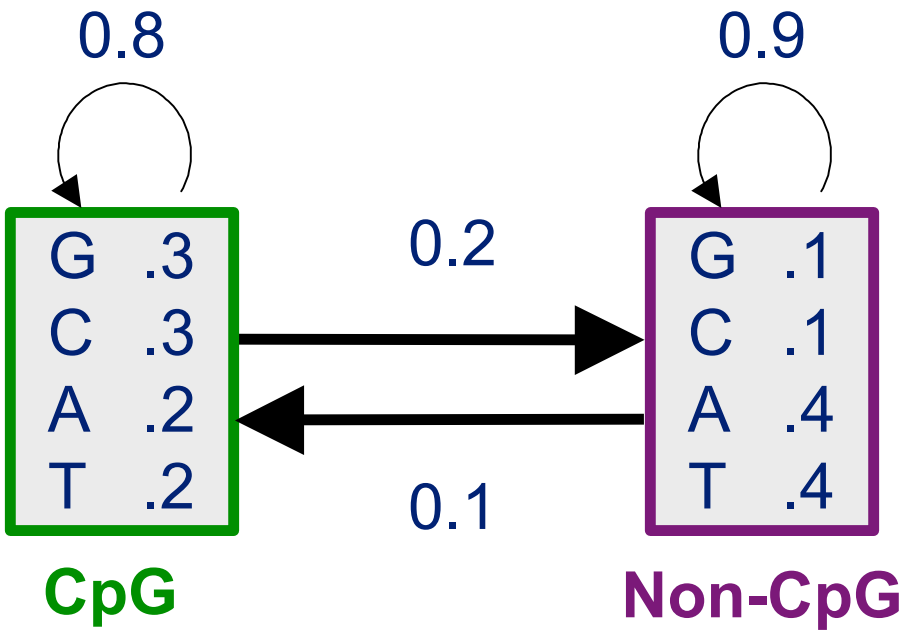

Example sequences: GATC ATC GC GAGAGC AGATTTC

$$P(AGATTTC \mid M) = (0.5 * 1.0)^{l=7}$$

**Arc** **Emission**

# A Simple HMM

## CpG Islands; Methylation Suppressed in Promoter Regions; States are Really Hidden Now

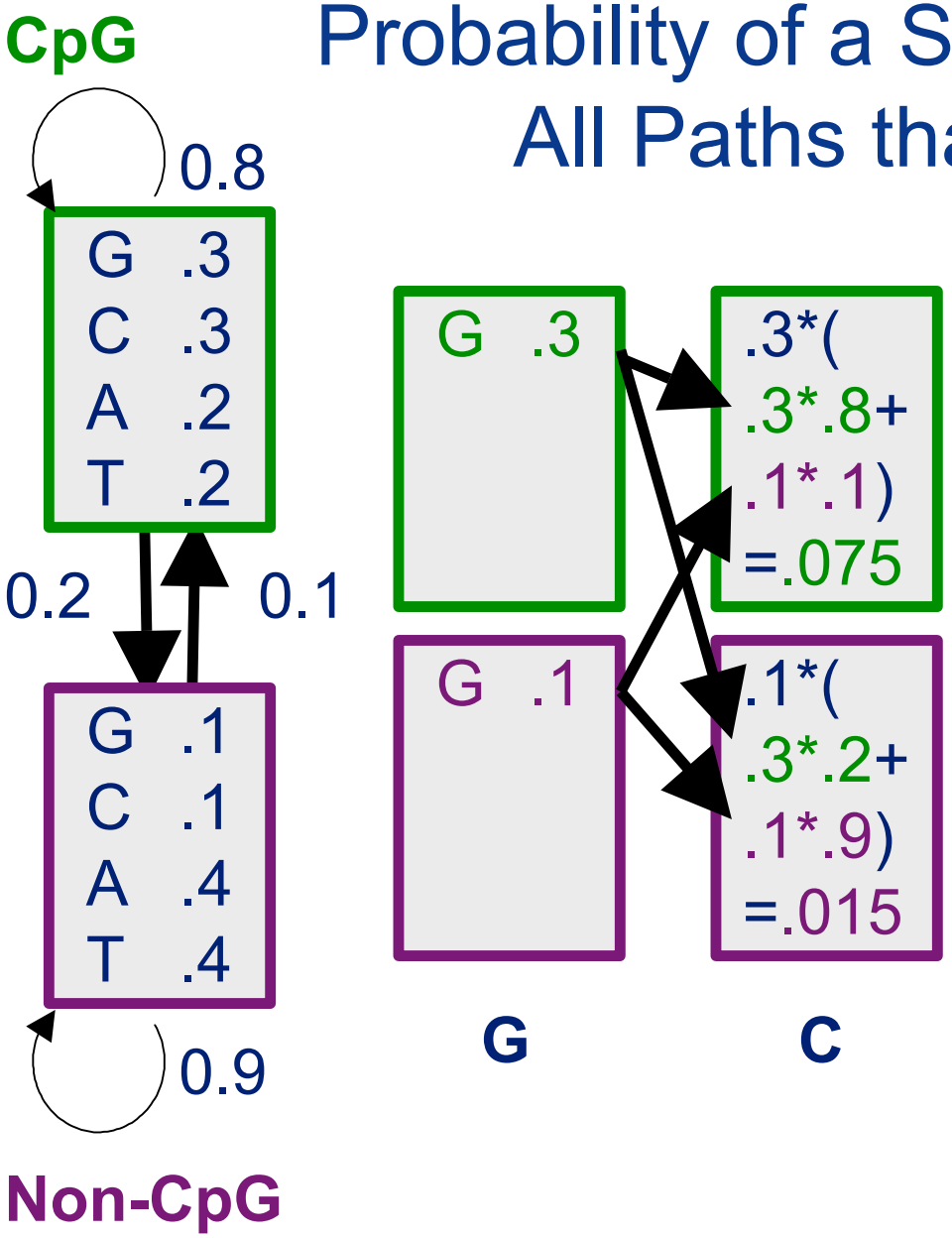



$$P(state_y^i \mid D <= i) = \sum_x P(state_x^{i-1}) * P(x -> y) * P(E_D \mid state_y^i)$$

Fractional likelihood

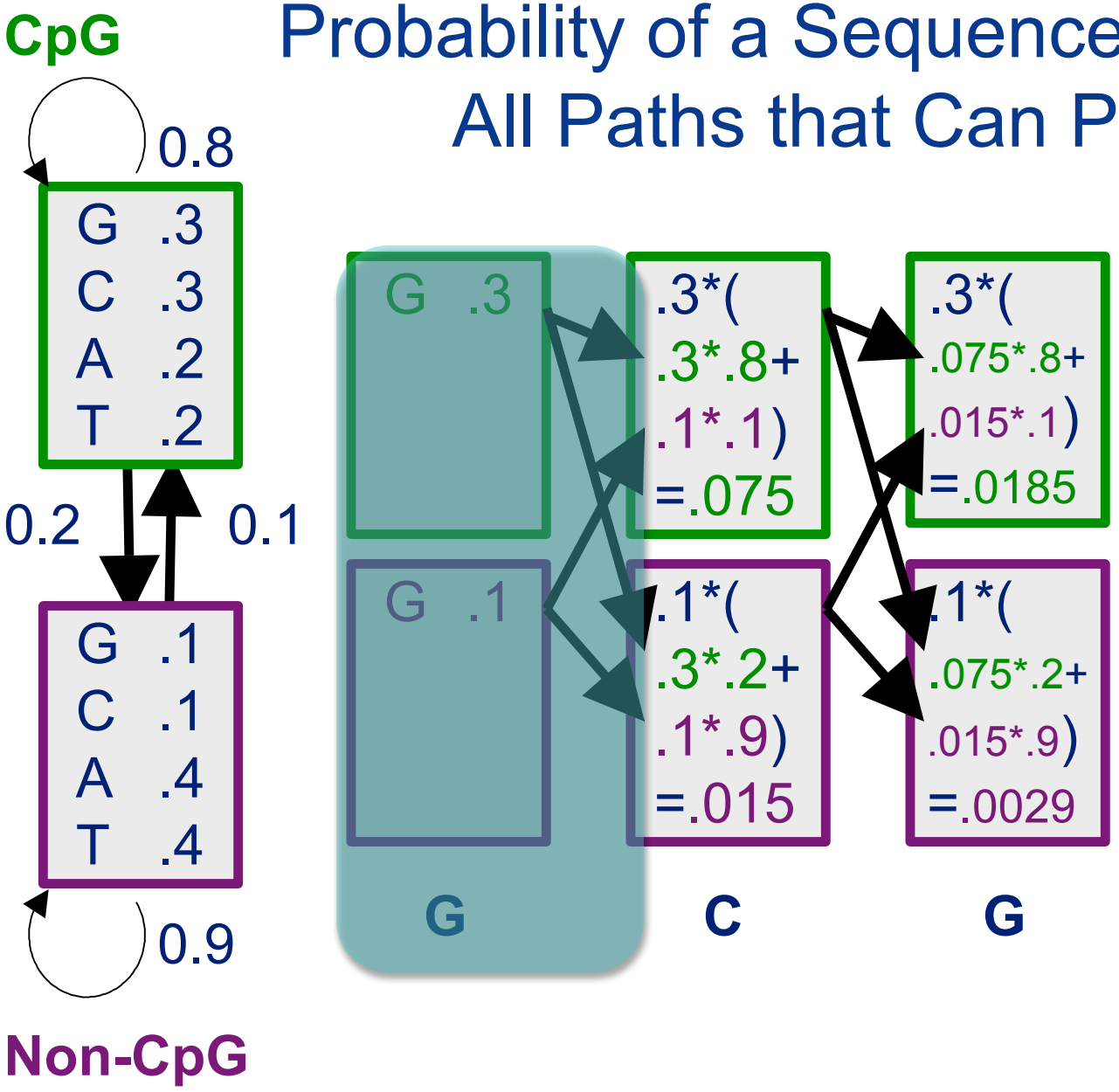
# The Forward Algorithm

## Probability of a Sequence is the Sum of All Paths that Can Produce It



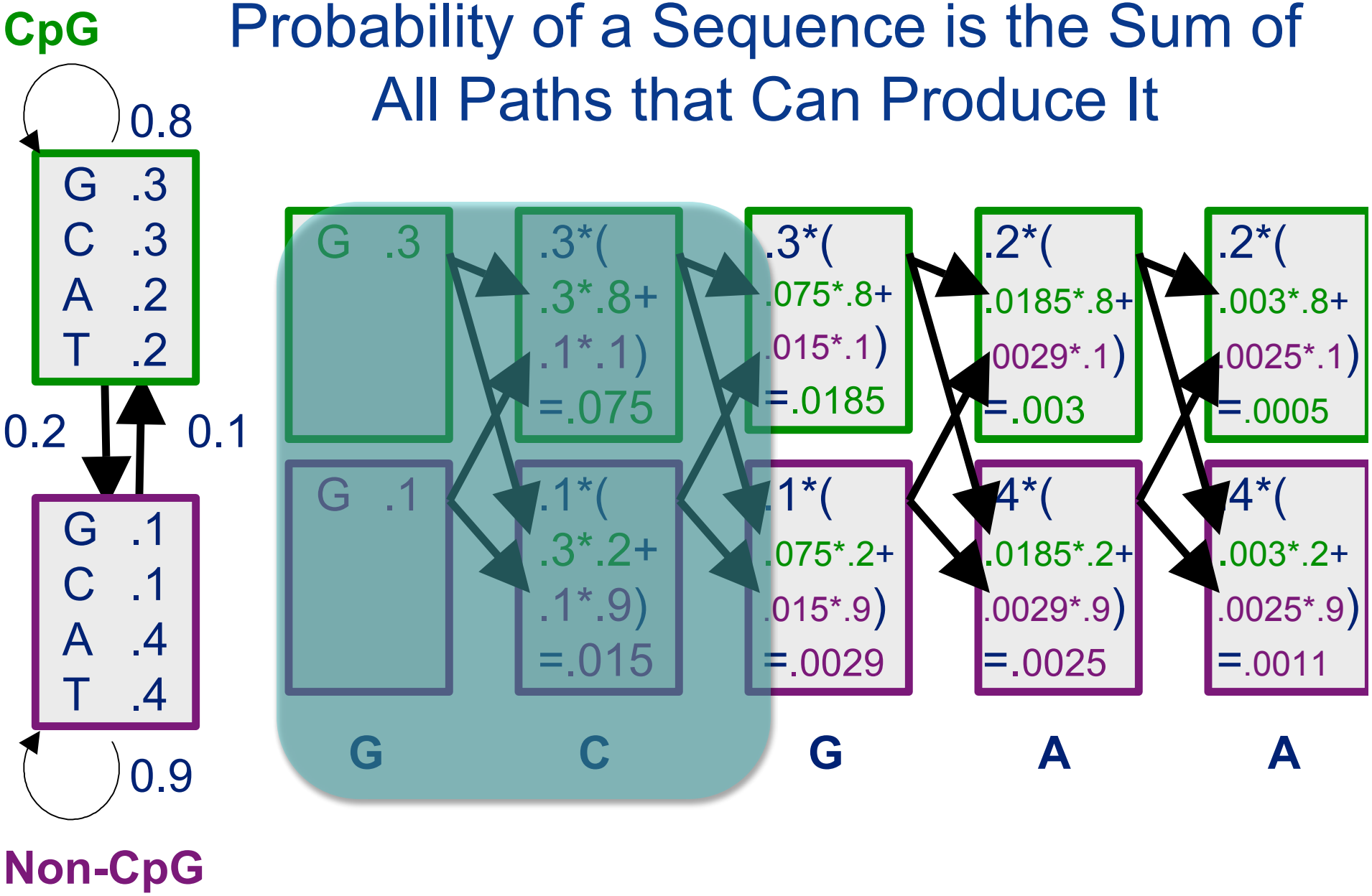
# The Forward Algorithm

## Probability of a Sequence is the Sum of All Paths that Can Produce It



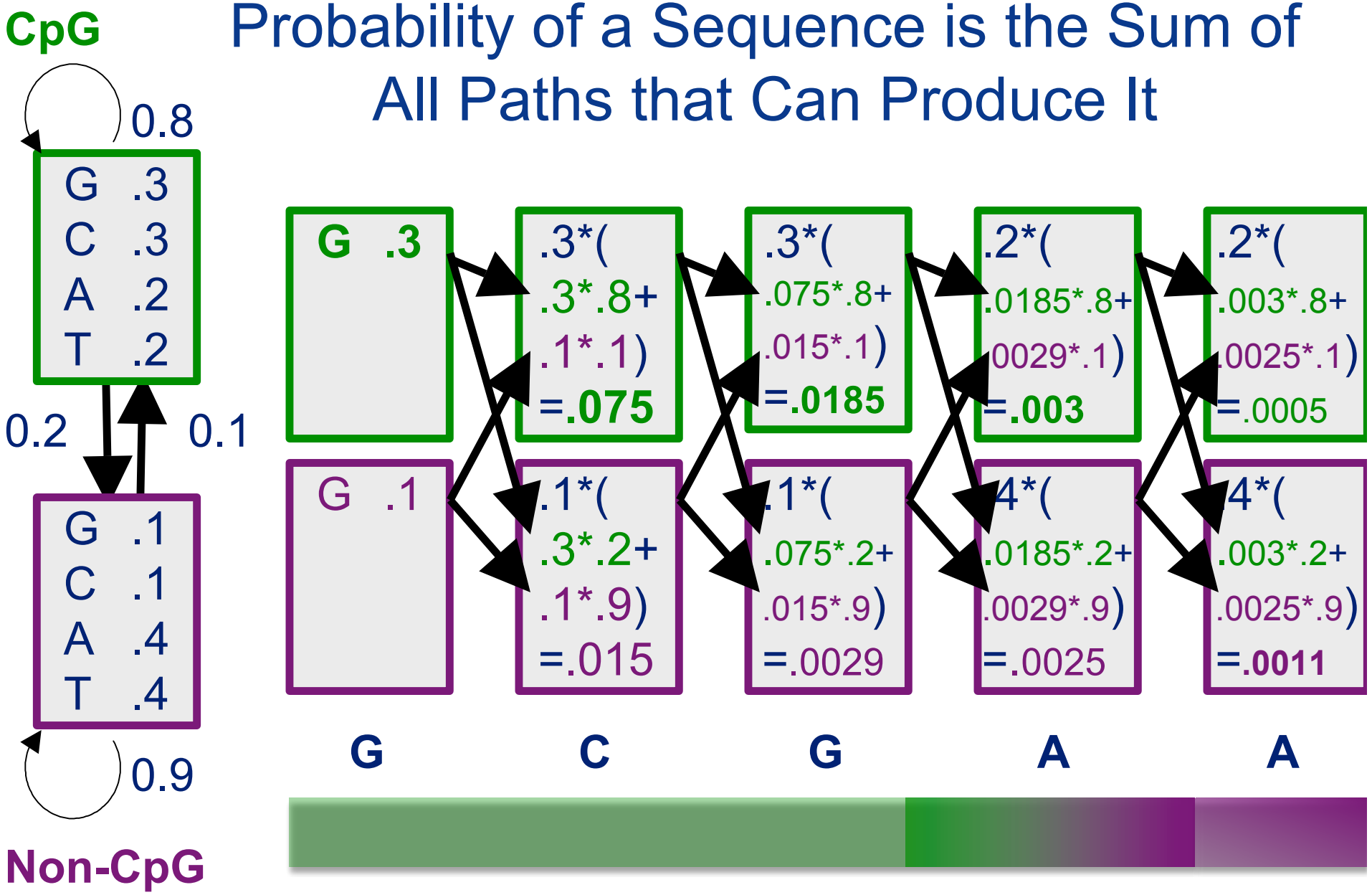
# The Forward Algorithm

## Probability of a Sequence is the Sum of All Paths that Can Produce It



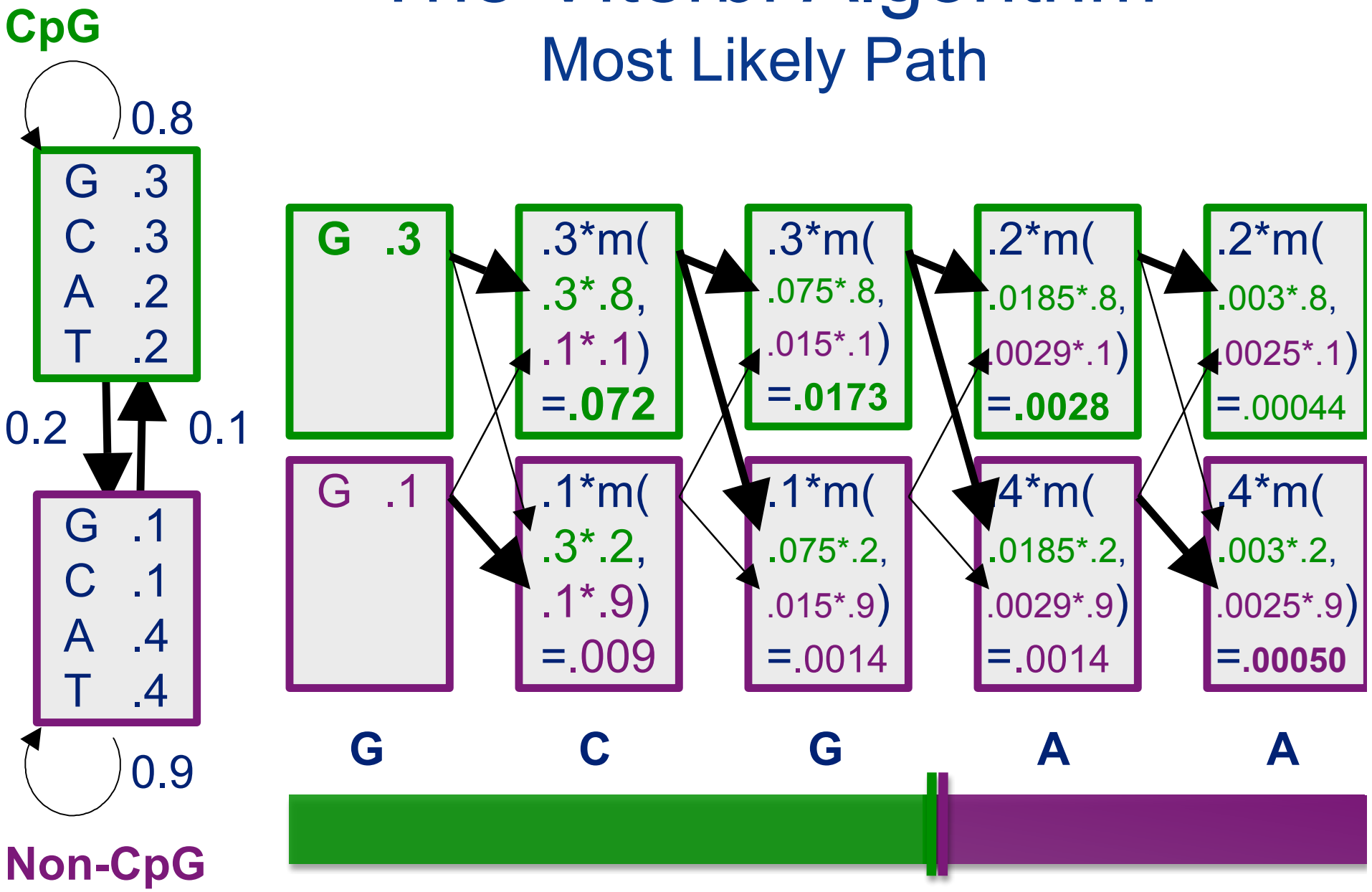
# The Forward Algorithm

## Probability of a Sequence is the Sum of All Paths that Can Produce It



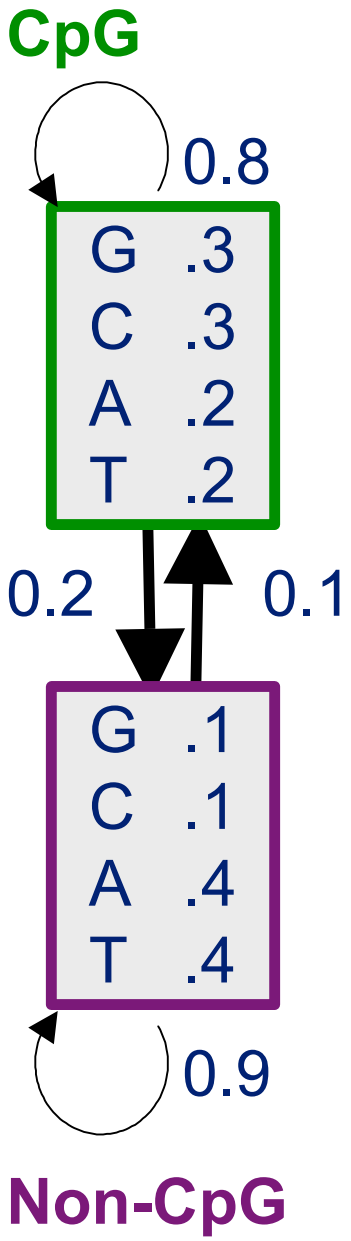
# The Viterbi Algorithm

## Most Likely Path

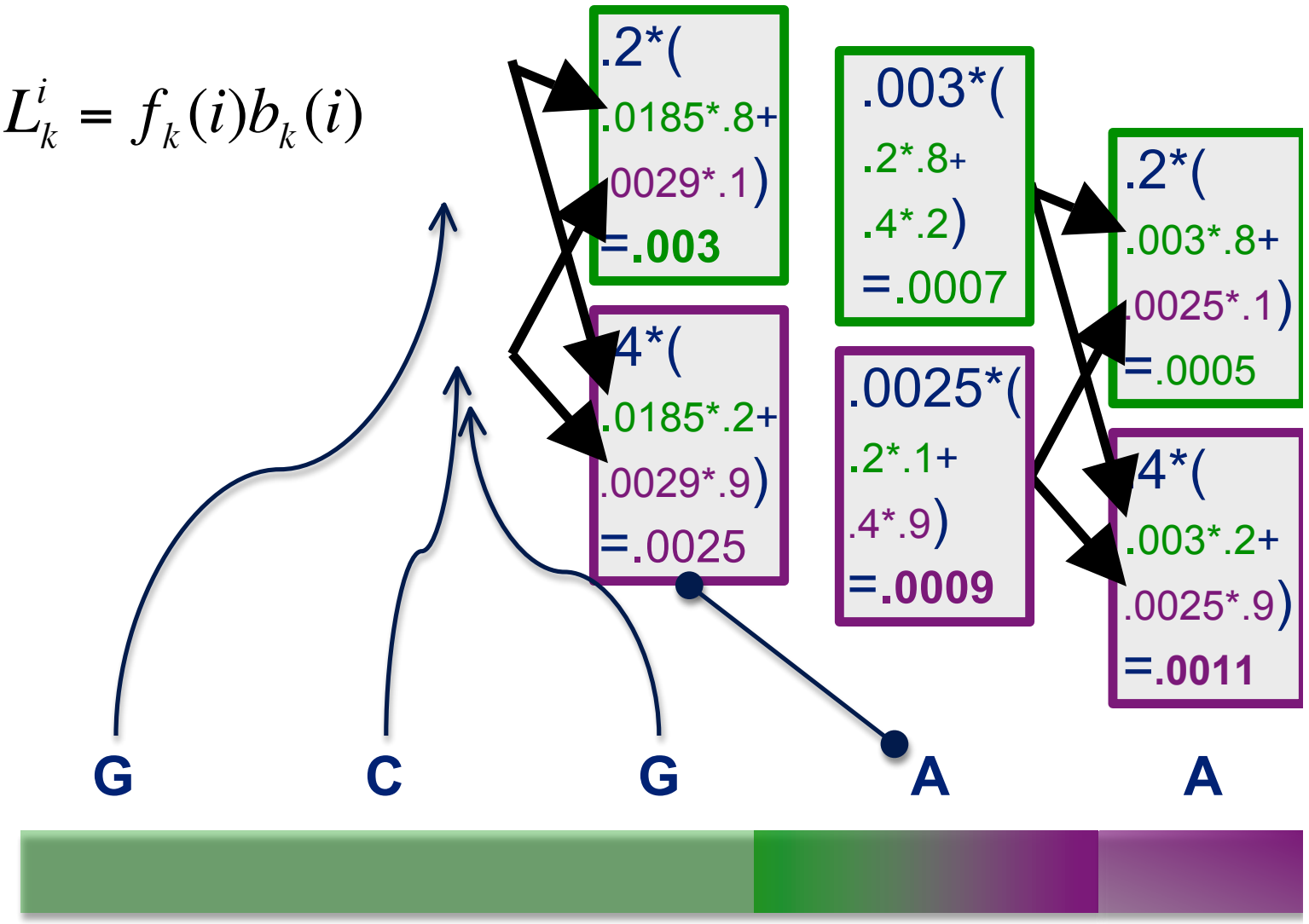


# Forwards and Backwards

## Probability of a State at a Position




$$L_k^i = f_k(i)b_k(i)$$



# Forwards and Backwards

## Probability of a State at a Position

$$P(CpG \mid i = 4, D)$$

$$= \frac{P(CpG)}{\left[P(CpG) + P(non - CpG)\right]}$$

$$= \frac{0.0007}{0.0007 + 0.0009} = 0.432$$

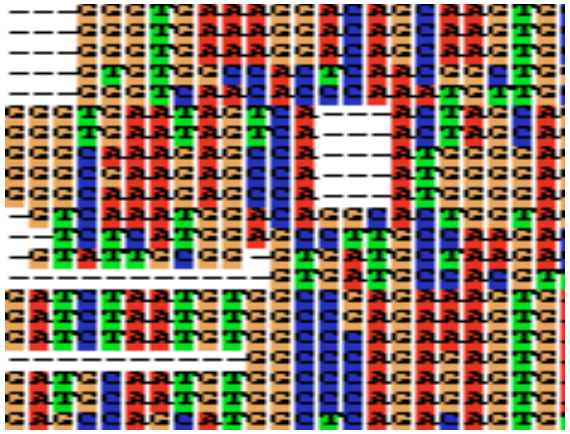
.003*(
.2*.8+
.4*.2)
=.0007

.0025*(
.2*.1+
.4*.9)
=**.0009**

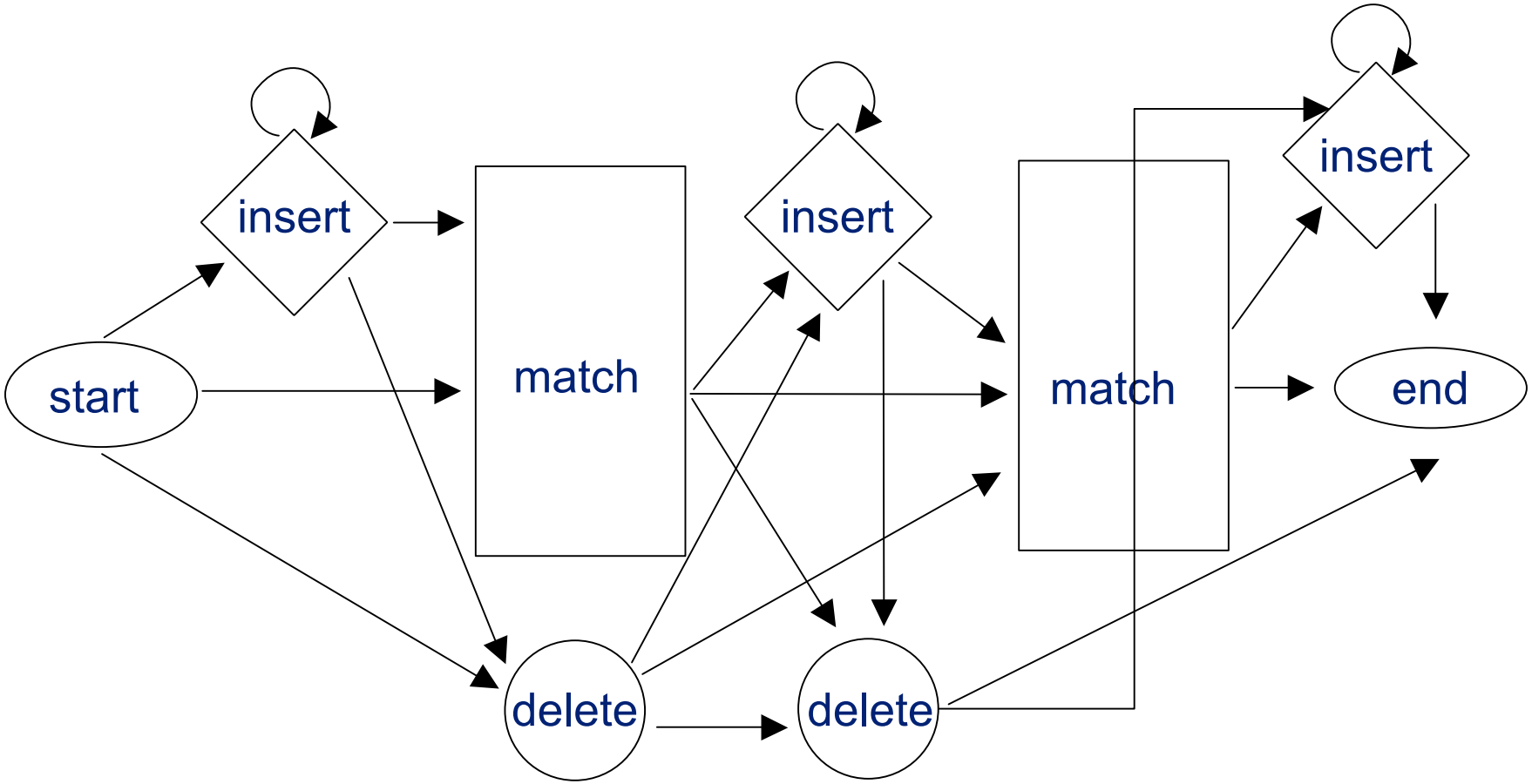
# Homology HMM

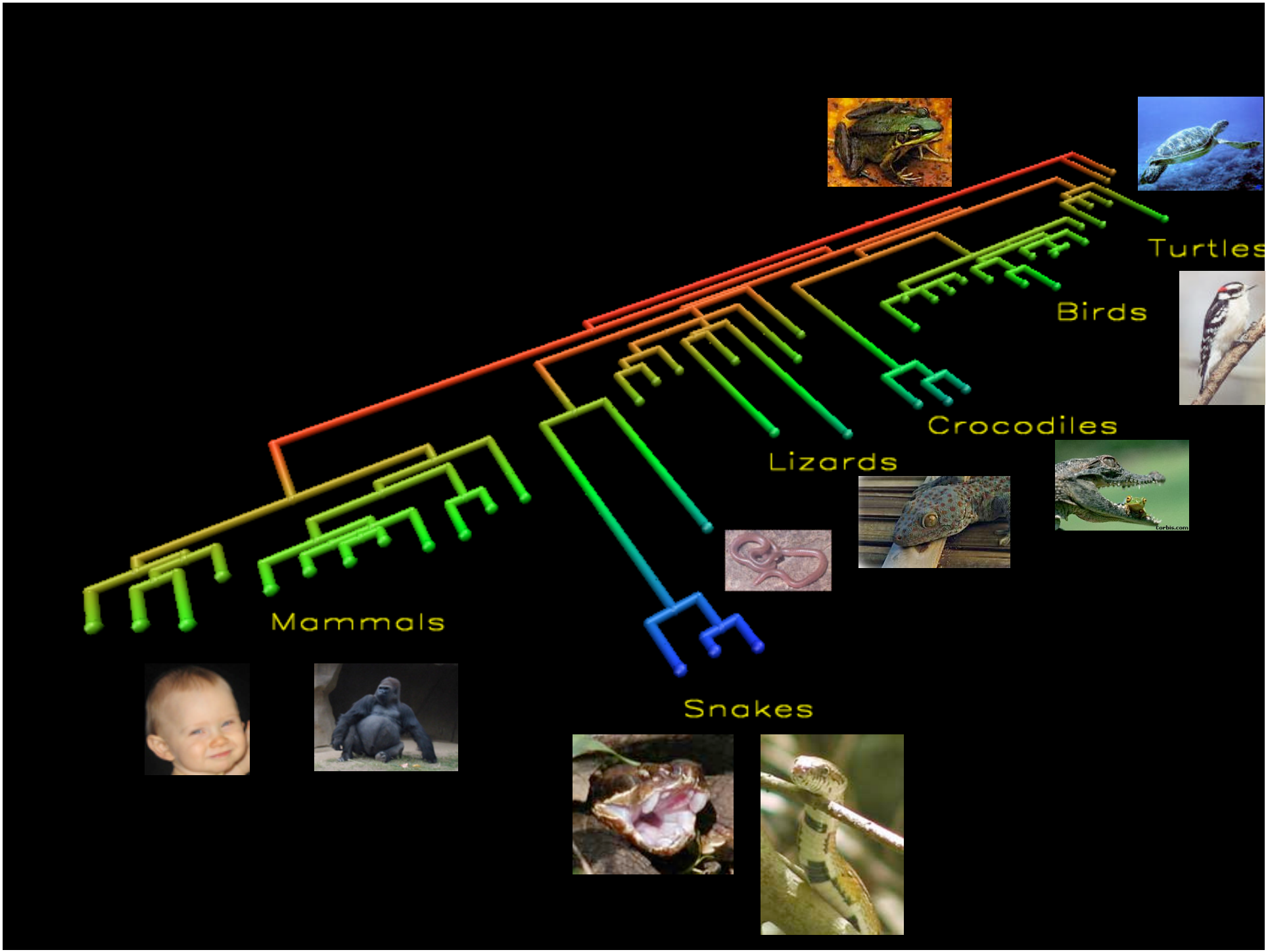
- Gene recognition, identify distant homologs

- Common Ancestral Sequence
  - Match, site-specific emission probabilities
  - Insertion (relative to ancestor), global emission probs
  - Delete, emit nothing
  - Global transition probabilities



# Homology HMM




Turtles
Birds
Crocodiles
Lizards
Mammals
Snakes

# Multiple Sequence Alignment HMM

- Defines predicted homology of positions (sites)
  - Recognize region within longer sequence
  - Model domains or whole proteins
  - Structural alignment
  - Compare alternative models
- Can modify model for sub-families
- Ideally, use phylogenetic tree
  - Often not much back and forth
  - Indels a problem

# Model Comparison

- Based on $P(D \mid \theta, M)$
  - For ML, take $P_{\max}(D \mid \theta, M)$
    - Usually $-\ln P_{\max}(D \mid \theta, M)$ to avoid numeric error
  - For heuristics, "score" is $-\log_2 P(D \mid \theta_{fixed}, M)$
  - For Bayesian, calculate

$$P_{\max}(\theta, M \mid D) = \frac{P(D \mid \theta, M) * P(\theta) * P(M)}{\sum P(D \mid \theta, M) * P(\theta) * P(M)}$$

# Parameters, $\theta$

- Types of parameters
  - Amino acid distributions for positions
  - Global AA distributions for insert states
  - Order of match states
  - Transition probabilities
  - Tree topology and branch lengths
  - Hidden states (integrate or augment)
- Wander parameter space (search)
  - Maximize, or move according to posterior probability (Bayes)

# Expectation Maximization (EM)

- Classic algorithm to fit probabilistic model parameters with unobservable states
  - Or missing data
- Two Stages, iterate
  - Maximize
    - If know hidden variables (states), maximize model parameters with respect to that knowledge
  - Expectation
    - If know model parameters, find expected values of the hidden variables (states)
- Works well even with e.g., Bayesian to find near-equilibrium space

# Homology HMM EM

- Start with heuristic (e.g., ClustalW)
- Maximize
  - Match states are residues aligned in most sequences
  - Amino acid frequencies observed in columns
- Expectation
  - Realign all the sequences given model
- Repeat until convergence
- Problems: Local, not global optimization
  - Use procedures to check how it worked

# Model Comparison

- Determining significance depends on comparing two models
  - Usually null model, $H_0$, and test model, $H_1$
  - Models are nested if $H_0$ is a subset of $H_1$
  - If not nested
    - Akaike Iinformation Criterion (AIC) [similar to empirical Bayes] or
    - Bayes Factor (BF) [but be careful]
- Generating a null distribution of statistic
  - Z-factor, bootstrapping, $\chi^2_\nu$, parametric bootstrapping, posterior predictive

# Z Test Method

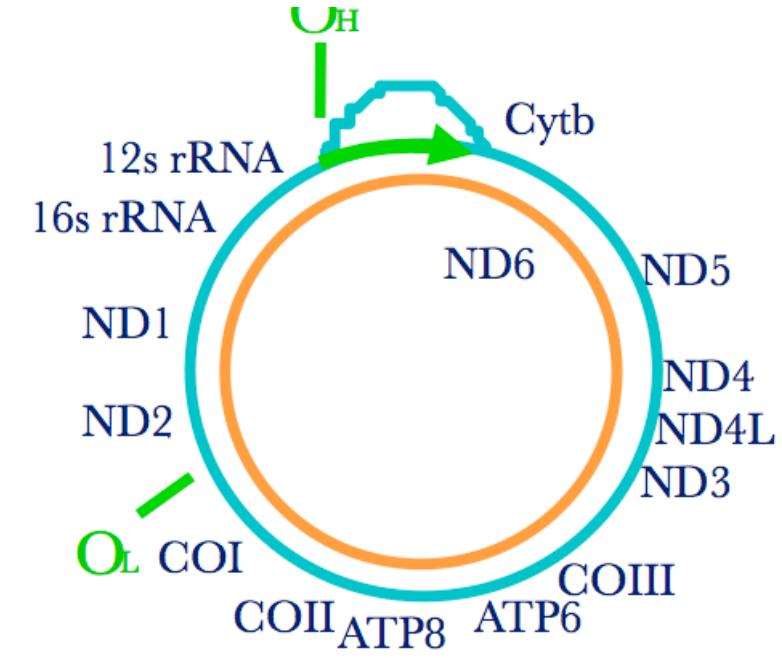
- Database of known negative controls
  - E.g., non-homologous (NH) sequences
  - Assume NH scores $\sim N(\mu,\sigma)$
    - i.e., you are modeling known NH sequence scores as a normal distribution
  - Set appropriate significance level for multiple comparisons (more below)
- Problems
  - Is homology certain?
  - Is it the appropriate null model?
    - Normal distribution often not a good approximation
  - Parameter control hard: e.g., length distribution

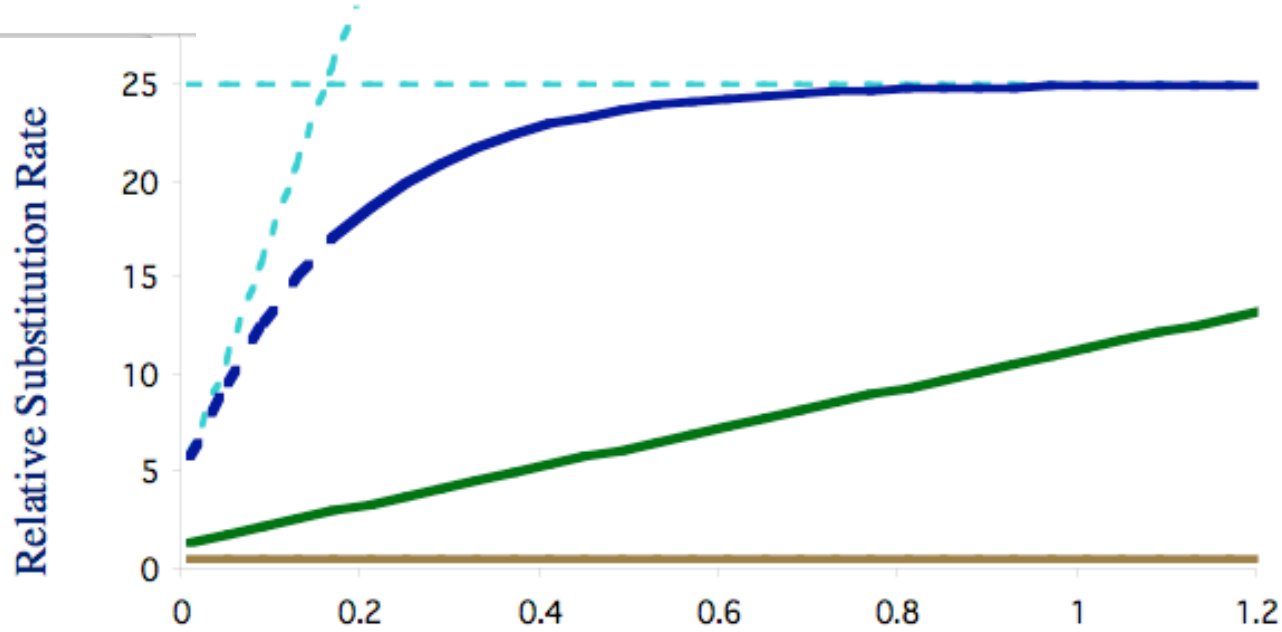
# Bootstrapping and Parametric Models

- Random sequence sampled from the same set of emission probability distributions
  - Same length is easy
  - Bootstrapping is re-sampling columns
  - Parametric models use estimated frequencies, may include variance, tree, etc.
    - More flexible, can have more complex null
    - Allows you to consider carefully what the null means, and what null is appropriate to use!
    - Pseudocounts of global frequencies if data limit
- Insertions relatively hard to model
  - What frequencies for insert states? Global?

# Homology HMM Resources

- UCSC (Haussler)
  - SAM: align, secondary structure predictions, HMM parameters, etc.
- WUSTL/Janelia (Eddy)
  - Pfam: database of pre-computed HMM alignments for various proteins
  - HMMer: program for building HMMs


OH
Cytb
12s rRNA
16s rRNA
ND6
ND5
ND1
ND4
ND4L
ND2
ND3
OL
COI
COIII
COII
ATP8
ATP6


Relative Substitution Rate
25
20
15
10
5
0
0
0.2
0.4
0.6
0.8
1
1.2
Time Spent Single-Stranded

# Increasing Asymmetry with Increasing Single Strandedness

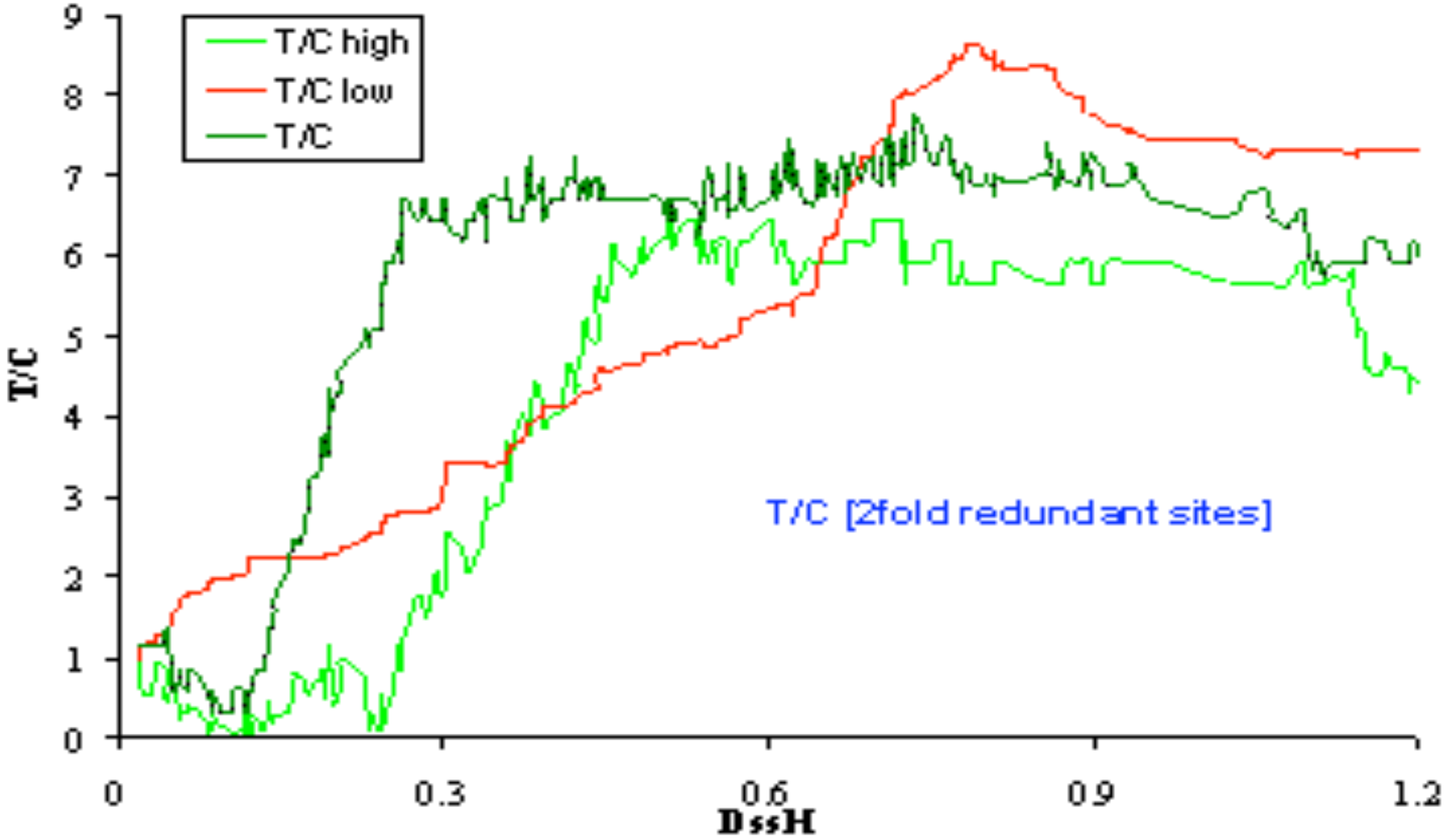
$$\begin{array}{c} \\ A \\ C \\ T \\ G \end{array}\begin{array}{c} \begin{array}{cccc} A & C & T & G \end{array} \\ \begin{bmatrix} - & \lambda_{AC}\pi_C & \lambda_{AT}\pi_T & \lambda_{AG}\pi_G \\ \lambda_{CA}\pi_A & - & \lambda_{CT}\pi_T & \lambda_{CG}\pi_G \\ \lambda_{TA}\pi_A & \lambda_{TC}\pi_C & - & \lambda_{TG}\pi_G \\ \lambda_{GA}\pi_A & \lambda_{GC}\pi_C & \lambda_{GT}\pi_T & - \end{bmatrix} \end{array}$$

$$\begin{array}{c} \\ A \\ C \\ T \\ G \end{array}\begin{array}{c} \begin{array}{cccc} A & C & T & G \end{array} \\ \begin{bmatrix} - & a & b & c \\ d & - & e & f \\ b & c & - & a \\ e & f & d & - \end{bmatrix} \end{array}$$

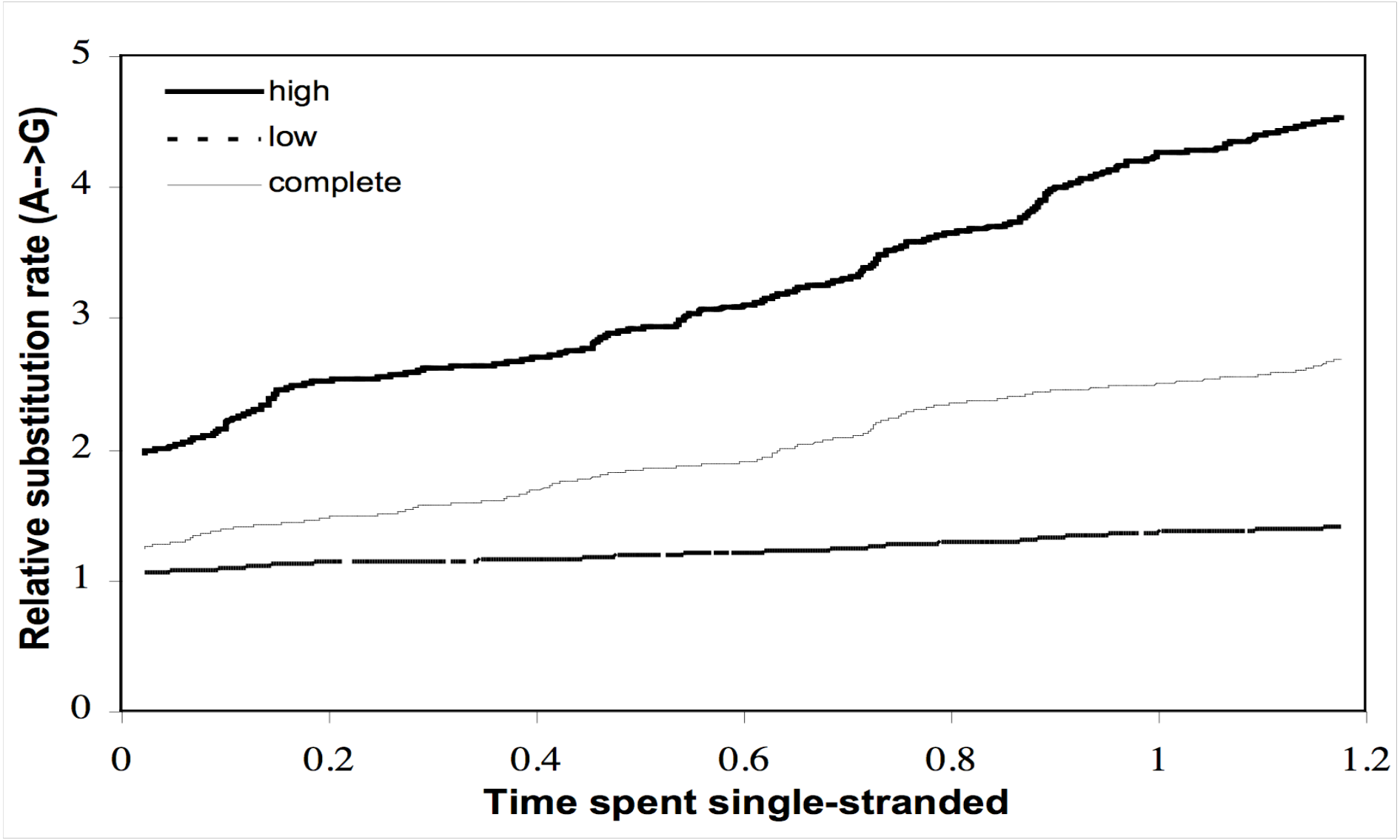
**e.g., P ( A=> G) = c + τ**

**τ = ( $D_{ssH}$ * Slope ) + Intercept**

# 2x Redundant Sites



# 4x Redundant Sites



# Beyond HMMs

- Neural nets
- Dynamic Bayesian nets
- Factorial HMMs
- Boltzmann Trees
- Kalman filters
- Hidden Markov random fields

# COI Functional Regions

***$O_2$ + protons+ electrons = $H_2O$ + secondary proton pumping (=ATP)***