Hidden Markov Models and Their Applications in Bioinformatics

Instructor: Connie Li

Given at the University of Calgary

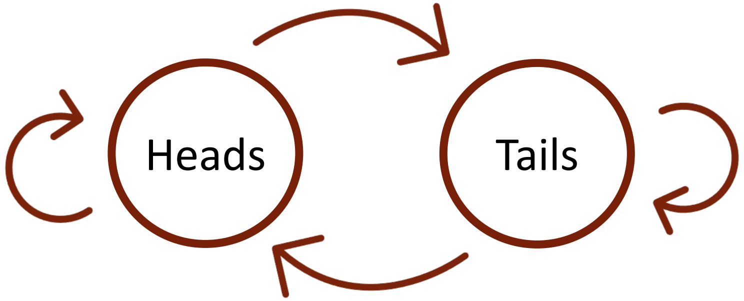
November 14, 2024

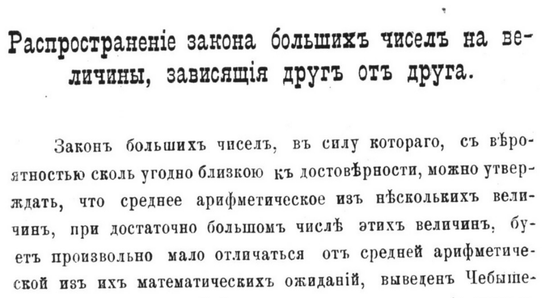
Outline: From theory to application

1. Introducing hidden Markov models
2. Formal definitions
3. Solving and inference
4. HMMs in Bioinformatics
5. Wrap up

Markov models

- A family of models for stochastic processes
- Key 1: Represent a system as a set of states and transitions between the states
  - The simplest Markov model is the Markov chain
- Key 2: The Markov property says the future state only needs depends on the present state
  - “The future only depends on today, ignore the past





Распространеніе закона большихъ чиселъ на величины, зависящія другъ отъ друга.

Законъ большихъ чиселъ, въ силу котораго, съ вѣроятностью сколь угодно близкою къ достовѣрности, можно утверждать, что среднее арифметическое изъ нѣсколькихъ величинъ, при достаточно большомъ числѣ этихъ величинъ, буетъ произвольно мало отличаться отъ средней арифметической изъ ихъ математическихъ ожиданій, выведенъ Чебыше-

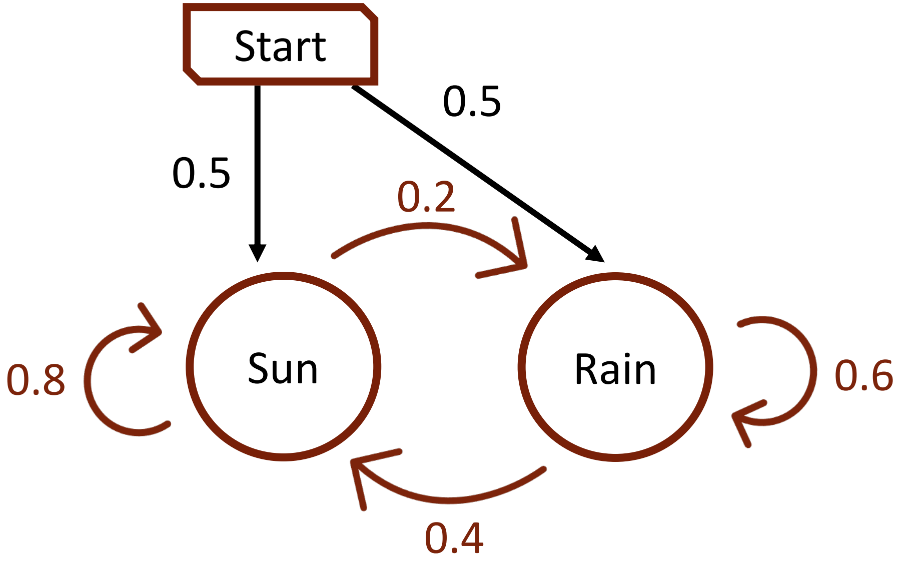
“Extension of the law of large numbers to quantities that depend on each other”

Andrey Markov*

https://en.wikipedia.org/wiki/Andrey_Markov
https://www.jehps.net/Novembre2006/Markov3pdf.pdf
*Not Andrei Markov, the former Canadiens hockey player

A Markov chain for the weather

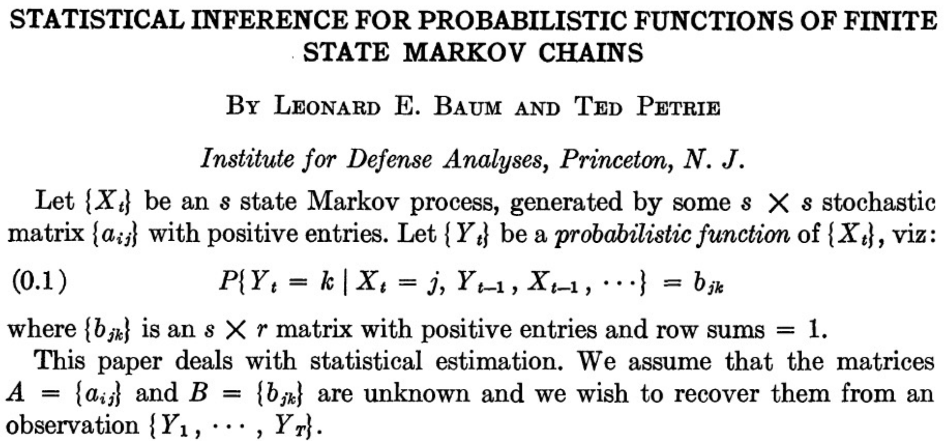
- Key 1: Represent a system as a set of states and transitions between the states
  - We can model the weather from day to day as transitions between sunny and rainy days
- Key 2: The Markov property says the future state only depends on the present state
  - The weather tomorrow depends only on whether it's sunny or rainy today
- The transition probabilities describe the probability of moving from state to state
- The initial state probabilities describe the probability of starting in each state



Hidden Markov models

- An extension of Markov chains by Leonard E. Baum and colleagues in the 1960s

- We can't see the states ("hidden")
- We can see emissions or observations from the states



STATISTICAL INFERENCE FOR PROBABILISTIC FUNCTIONS OF FINITE STATE MARKOV CHAINS

BY LEONARD E. BAUM AND TED PETRIE

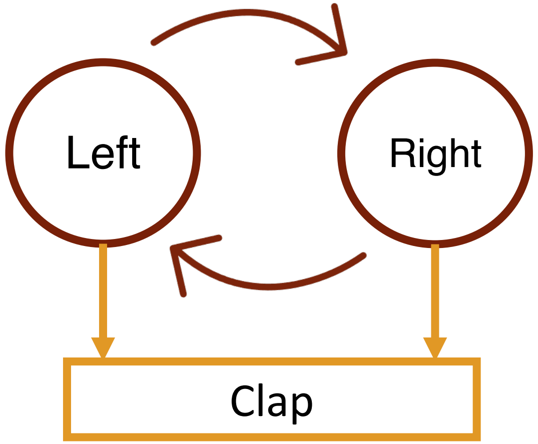
Institute for Defense Analyses, Princeton, N. J.

Let $\{X_t\}$ be an s state Markov process, generated by some $s \times s$ stochastic matrix $\{a_{ij}\}$ with positive entries. Let $\{Y_t\}$ be a *probabilistic function* of $\{X_t\}$, viz:

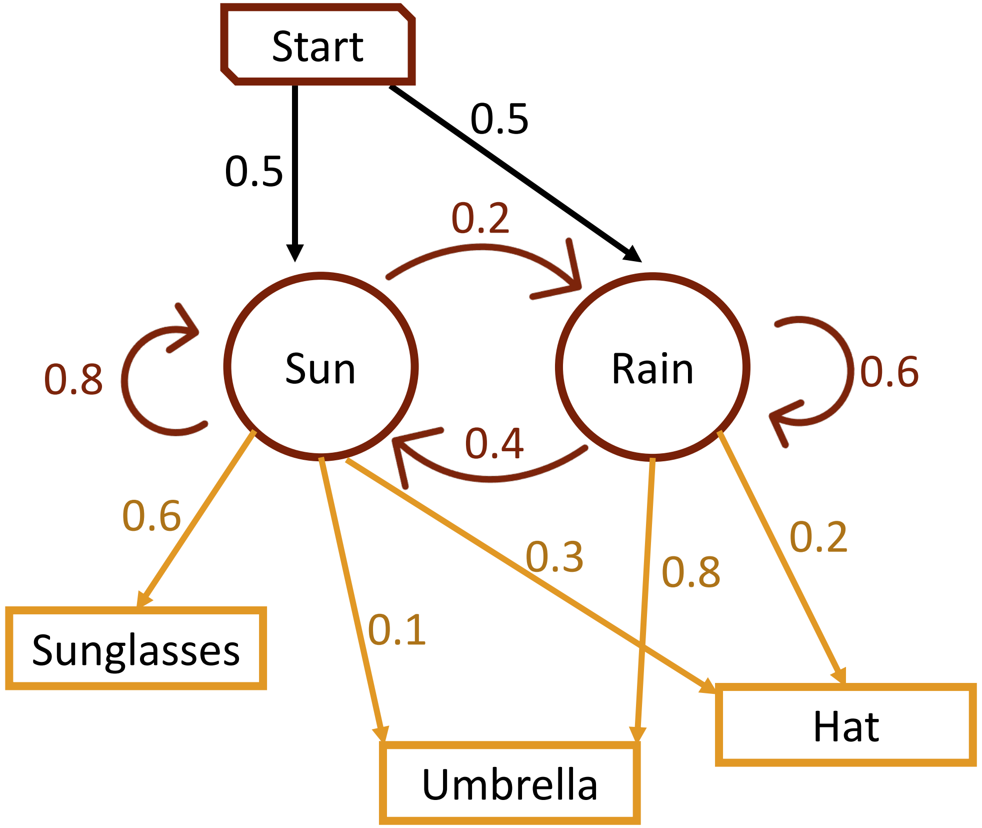
$$P\{Y_t = k \mid X_t = j, Y_{t-1}, X_{t-1}, \cdots\} = b_{jk} \qquad (0.1)$$

where $\{b_{jk}\}$ is an $s \times r$ matrix with positive entries and row sums = 1.

This paper deals with statistical estimation. We assume that the matrices $A = \{a_{ij}\}$ and $B = \{b_{jk}\}$ are unknown and we wish to recover them from an observation $\{Y_1, \cdots, Y_T\}$.



How can you know the weather without seeing it?



Initial probabilities

	Sun	Rain
Start	0.5	0.5

Transition probabilities

	Sun	Rain
Sun	0.8	0.2
Rain	0.4	0.6

Emission probabilities

	Sunglasses	Umbrella	Hat
Sun	0.6	0.1	0.3
Rain	0	0.8	0.2

Sunglasses Sunglasses Umbrella Sunglasses Sunglasses Sunglasses Sunglasses

Formal definitions and notation

- A hidden Markov model consists of random variables q_i and y_i where

$\mathrm{Q} = \{q_1\ q_2 \dots q_{\boldsymbol{N}}\}$ — Q is a set of N hidden states q_i

$\boldsymbol{A} = a_{1,1}\ a_{1,2}\ a_{n,1}, a_{N,N}$ **s.t.** $\sum_{j=1}^{N} a_{i,j} = 1\ \forall i$ — with transition probabilities $a_{i,j}$ describing the probability of moving from state i to state j

$\pi = \pi_1, \pi_2, \dots, \pi_N$ **s.t.** $\sum_{j=1}^{N} \pi_j = 1$ — and initial probabilities π_i describing the probability of starting in state i.

$\mathrm{Y} = \{y_1\ y_2 \dots y_M\}$ — Y is a set of M possible emissions (or observations) y_i

$\boldsymbol{B} = b_i(y_t)$ — each with an emission probability $b_i(y_t)$ of being generated from state q_i at time t

- We call a sequence of emission observations $Y = y_1, y_2, \dots y_T$ and a path of hidden states $Q = q_1, q_2, \dots q_T$, both with length T

Key assumptions for the HMM

- Markov property:

$$P(q_t = i|\, q_1, q_2, \ldots, q_{t-1}) = P(q_t = i|\, q_{t-1})$$

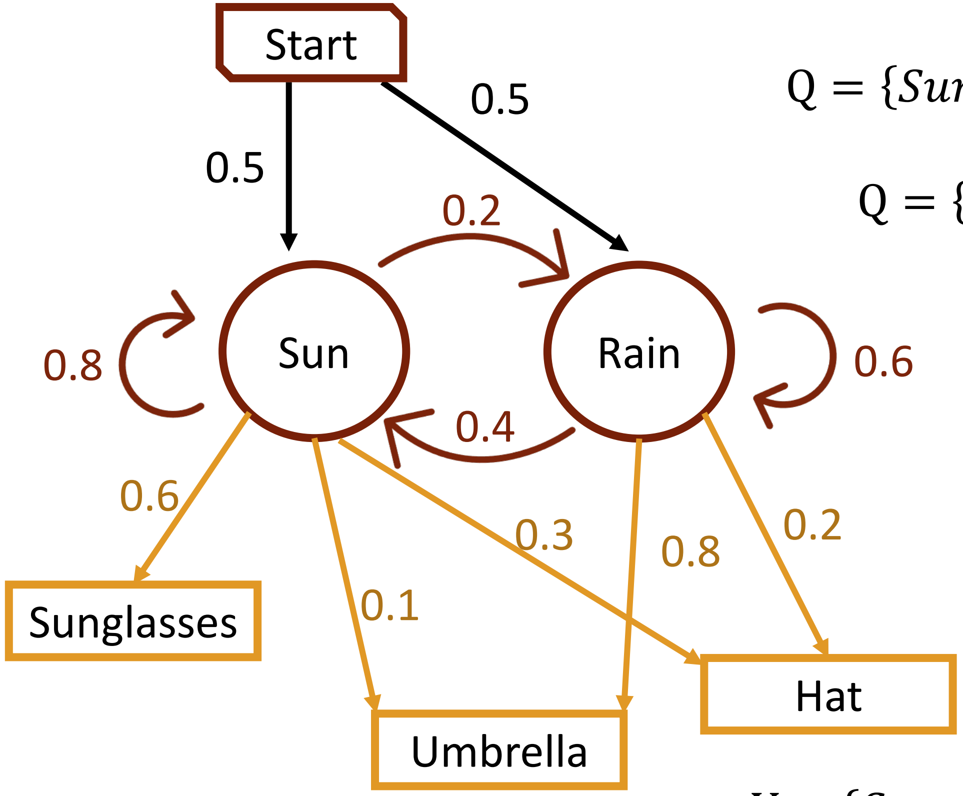
“the future only depends on the present, not the past”

- Output independence:

$$P(y_i|\, q_1, \ldots q_i \ldots q_T, y_1 \ldots y_i, \ldots y_T) = P(y_i|\, q_i)$$

“the probability observing the emission y_i depends only on the state that produced it q_i, not any other state or emissions”

Adding labels to our weather HMM



$Q = \{Sun, Rain\}$

or

$Q = \{q_s, q_r\}$

$Y = \{Sunglasses, Umbrella, Hat\}$

Initial probabilities

π_i	**Sun**	**Rain**
Start	0.5	0.5

Transition probabilities

$a_{i,j}$	**Sun**	**Rain**
Sun	0.8	0.2
Rain	0.4	0.6

Emission probabilities

$b_i(y_t)$	**Sunglasses**	**Umbrella**	**Hat**
Sun	0.6	0.1	0.3
Rain	0	0.8	0.2

- **What can we learn from our HMM?**

What is the probability of…

- Observing a particular sequence of emissions?

  sunglasses, hat, umbrella

  - First, think of one possible hidden state path

    sun, sun, rain

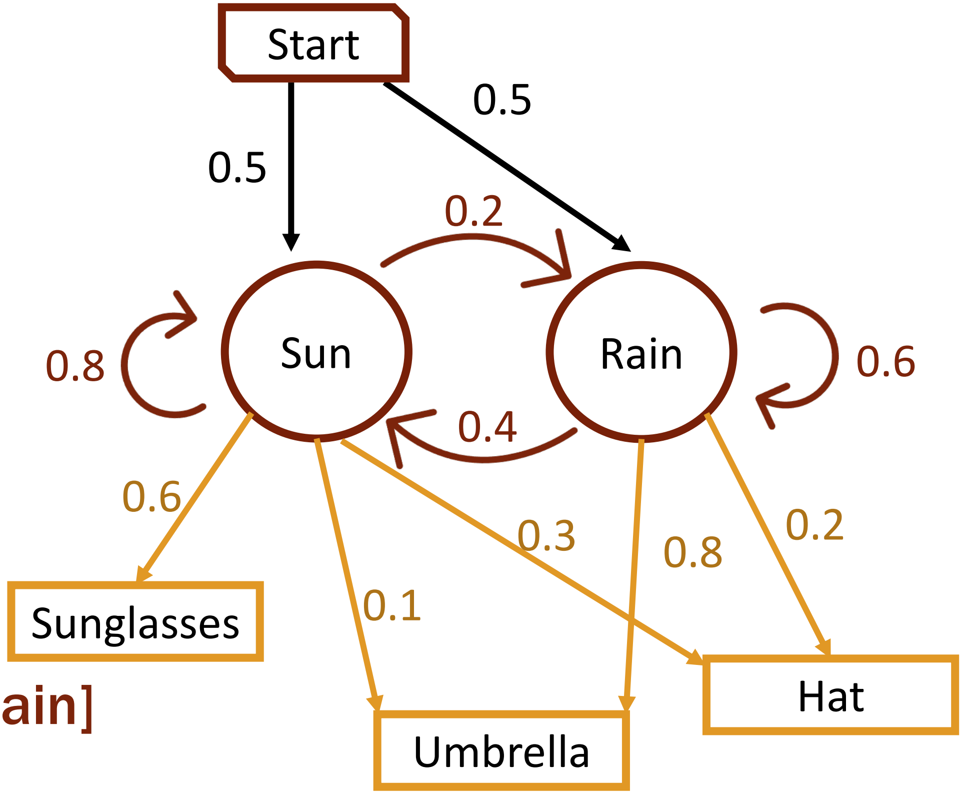
    $$P(SHU|sun\ sun\ rain)$$
    $$= \pi_{sun} b_{sun}(S) * a_{sun,sun} b_{sun}(H) * a_{sun,rain} b_{rain}(U)$$
    $$= (0.5 * 0.6) * (0.8 * 0.3) * (0.2 * 0.8)$$
    $$= 0.01152$$

  - How many possible hidden state paths are there?

    [sun, sun, rain] [sun, sun, sun] [sun, rain, sun] [sun, rain, rain]

    $$P(SHU) = 0.01152 + 0.00576 + 0.00048 + 0.00576$$
    $$= 0.02352$$

  - What about the paths starting with rain?
- Notice: You can also identify the most likely path of hidden states



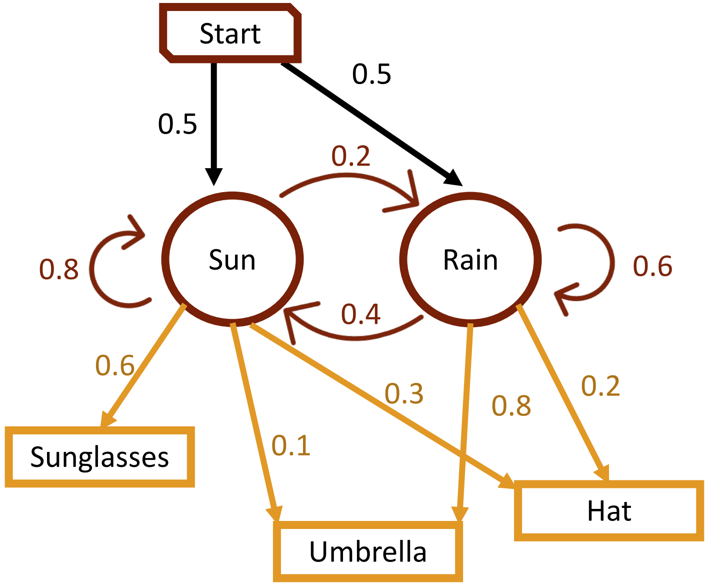
Four typical HMM inference problems

- Given the parameters of an HMM and an observed sequence of T emissions,
  - what is the probability of that observed sequence?
    (i. Scoring)
  - what is the most likely hidden states path?
    (ii. Decoding)
  - what is the distribution of hidden states at time k?
    (iii. Filtering when $k = \mathrm{T}$; iv. Smoothing when $k < \mathrm{T}$)
- These problems become increasingly complex with increasing N, M, and T
  - How complex?
- Usually not possible to compute by traversing all the paths

Some insights about state sequences

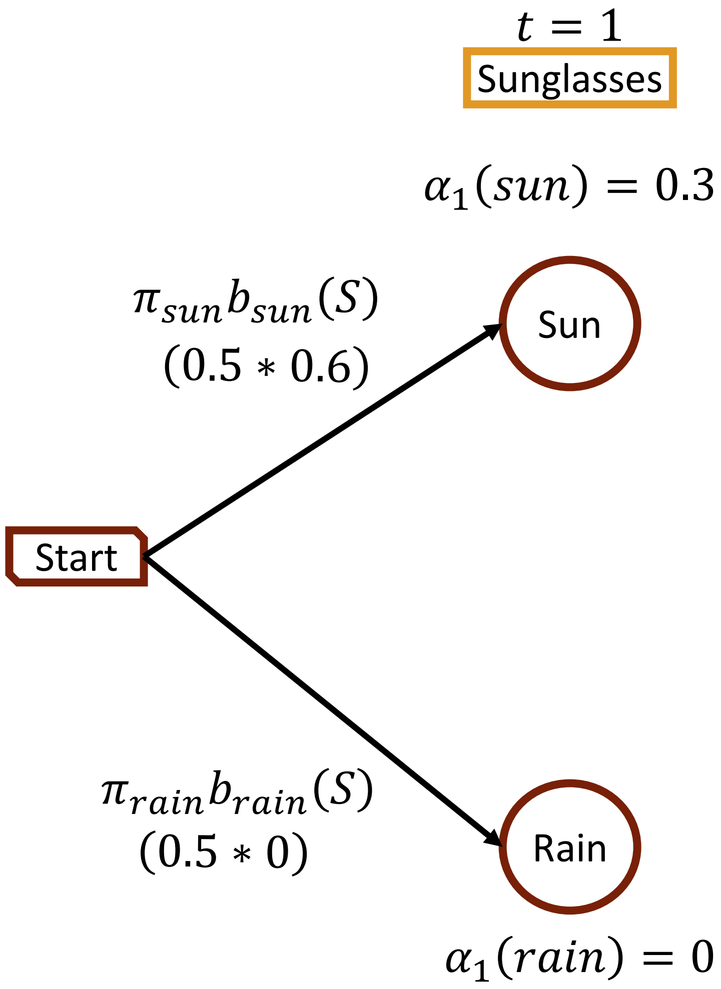
- Given emission sequence $Y =$ sunglasses, hat, umbrella,
  All possible paths are:

$t = 1$ $y_1 = sunglasses$	$t = 2$ $y_2 = hat$	$t = 3$ $y_1 = umbrella$
sun	sun	rain
sun	sun	sun
sun	rain	sun
sun	rain	rain
rain	sun	rain
rain	sun	sun
rain	rain	sun
rain	rain	rain



- Note: at $t = 1$, there are two unique paths
  at $t = 2$, there are four unique paths
  at $t = 3$, there are eight unique paths
- At $t = T$, there are N^T unique paths, but we can save time by re-using information

A trellis representation of our process



$t = 1$

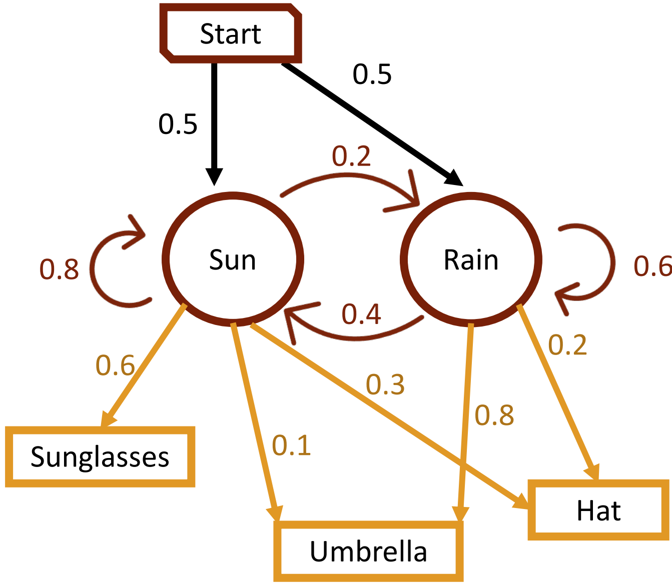
Sunglasses

$\alpha_1(sun) = 0.3$

$\pi_{sun} b_{sun}(S)$
$(0.5 * 0.6)$

$\pi_{rain} b_{rain}(S)$
$(0.5 * 0)$

$\alpha_1(rain) = 0$



A trellis representation of our process

$t = 1$

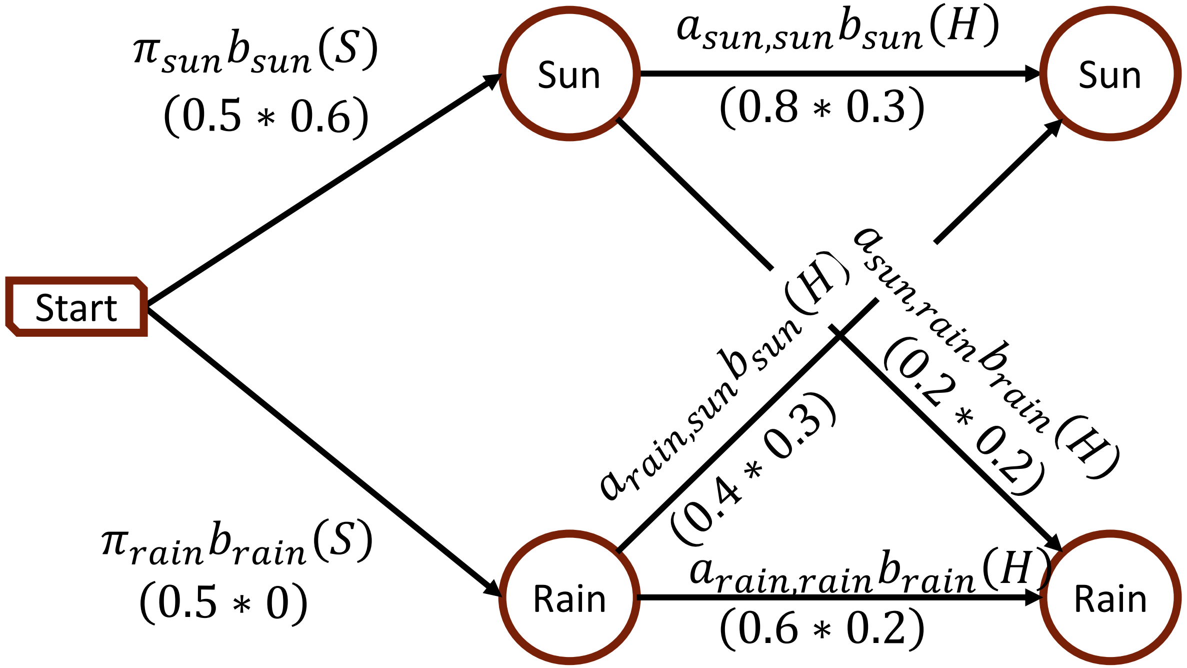
Sunglasses

$t = 2$

Hat

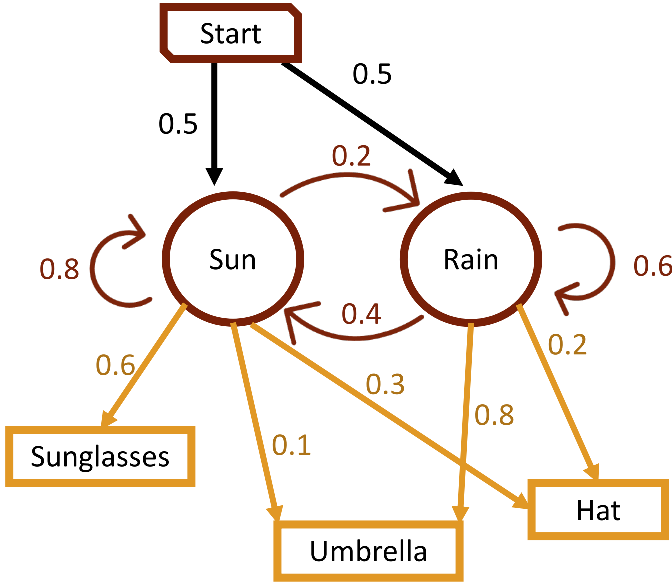
$\alpha_1(sun) = 0.3$

$$\alpha_2(sun) = \alpha_1(sun) * a_{sun,sun} b_{sun}(H) + \alpha_1(rain)\ a_{rain,sun} b_{sun}(H)$$
$$= 0.3 * 0.24 + 0 * 0.12 = 0.072$$

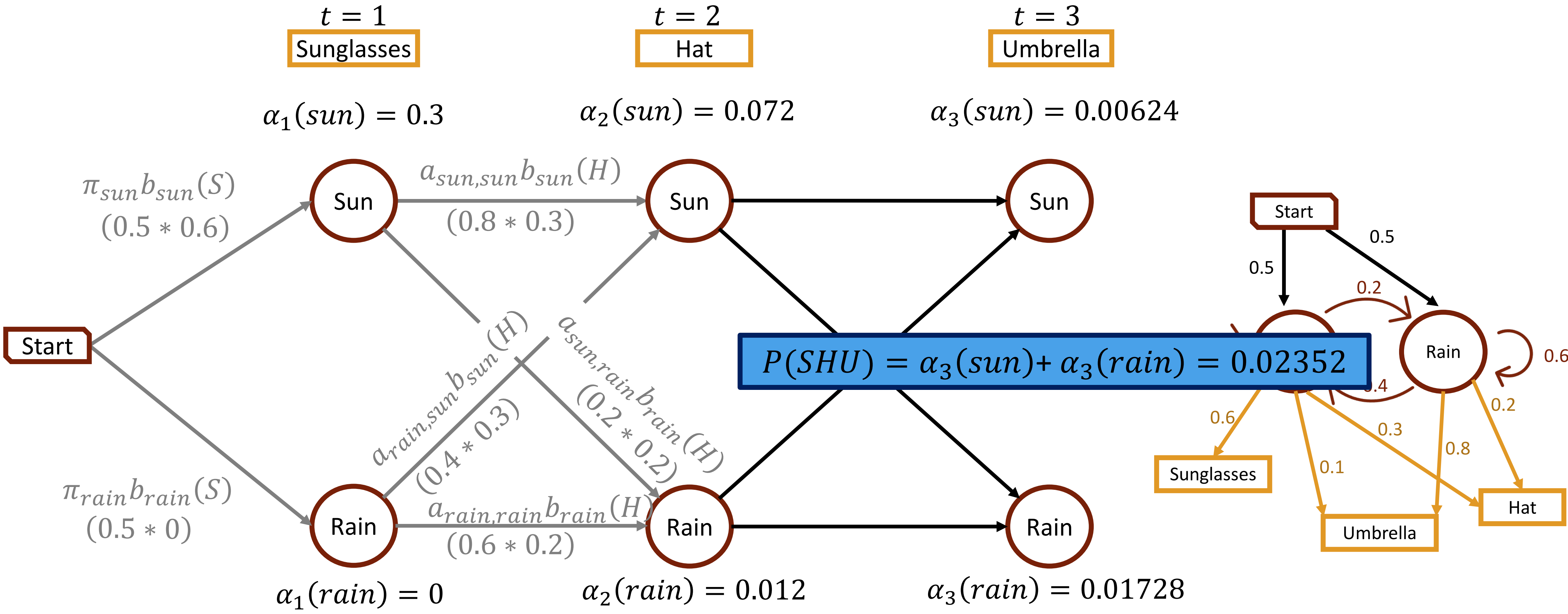


$\alpha_1(rain) = 0$

$$\alpha_2(rain) = \alpha_1(sun) * a_{sun,rain} b_{rain}(H) + \alpha_1(rain)\ a_{rain,rain} b_{rain}(H)$$
$$= 0.3 * 0.04 + 0 * 0.12 = 0.012$$



A trellis representation of our process



$$P(SHU) = \alpha_3(sun) + \alpha_3(rain) = 0.02352$$

Dynamic programming for HMMs

- Our trellis example motivates the Forward Algorithm, which computes the forward probabilities $\alpha_t(j)$
  - Wrote the probability of being in state j at time t in terms of the previous step $t-1$:

$$\alpha_t(j) = \sum_{i}^{N} \underbrace{\alpha_{t-1}(i)}_{\text{The previous forward probability}} \underbrace{a_{ij}}_{\text{The transition probability from } q_i \text{ to } q_j} \underbrace{b_j(y_t)}_{\text{The emission probability of observation } y_t \text{ given state } j}$$

- Dynamic programming uses recursive relationships to simplify problems
  - Recursion involves breaking up a problem into simpler versions of itself
  - We require a base case, recursive step, and a termination step

The Forward Algorithm

Scoring problem: Given an HMM $\Theta = (A, B)$ and an observation sequence Y, what is the likelihood $\mathrm{P}(Y|\Theta)$?

1. Base case:

$$\alpha_1(j) = \pi_j b_j(y_1);\ \ 1 \le j \le N$$

2. Recursion:

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\, b_j(y_t);\ \ \ 1 \le j \le N, 1 \le t \le T$$

3. Termination:

$$P(Y|\Theta) = \sum_{i=1}^{N} \alpha_T(i)$$

Notes on the Forward Algorithm

- The Forward Algorithm is much more efficient than traversing all paths N^T
  - $\mathbf{O}(N^2 T)$ time
- Formal derivation relies on liberal application of chain rule and Markov property (see additional materials on git)
- We saw how to use it to score a sequence of emissions
- The Forward Algorithm is also useful in filtering
  - What is the most likely hidden state at the end of a sequence of emissions?

- A small adjustment allows us to identify the most likely underlying hidden state sequence

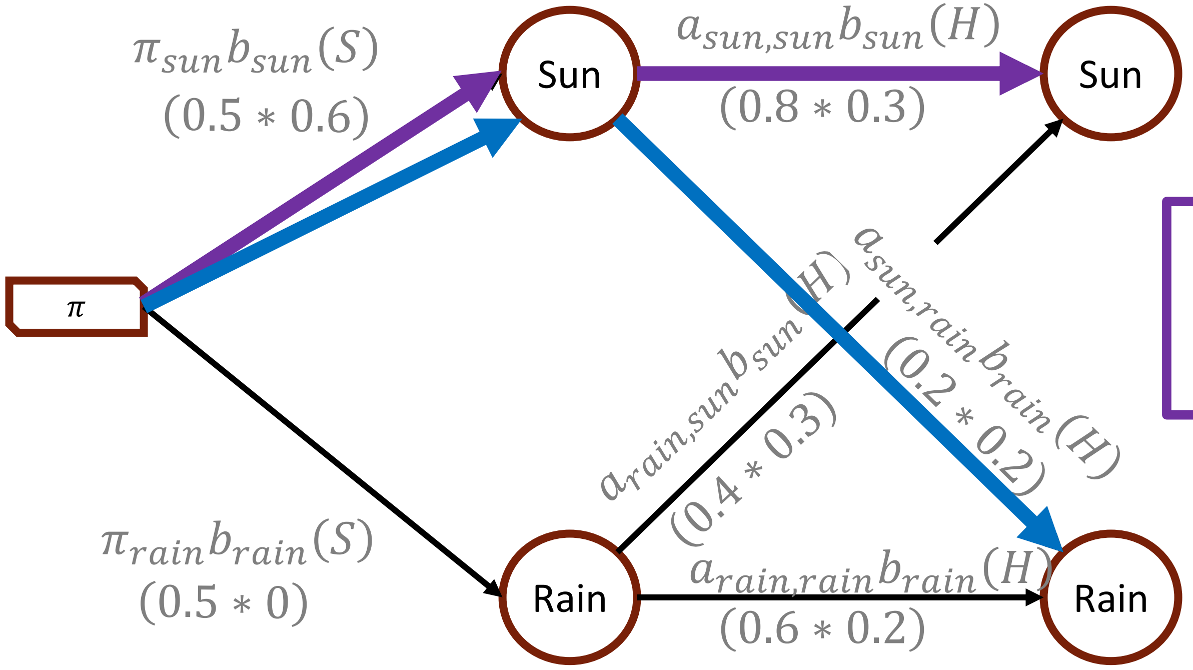
The trellis again, but we don't take the sum

$t = 1$ Sunglasses

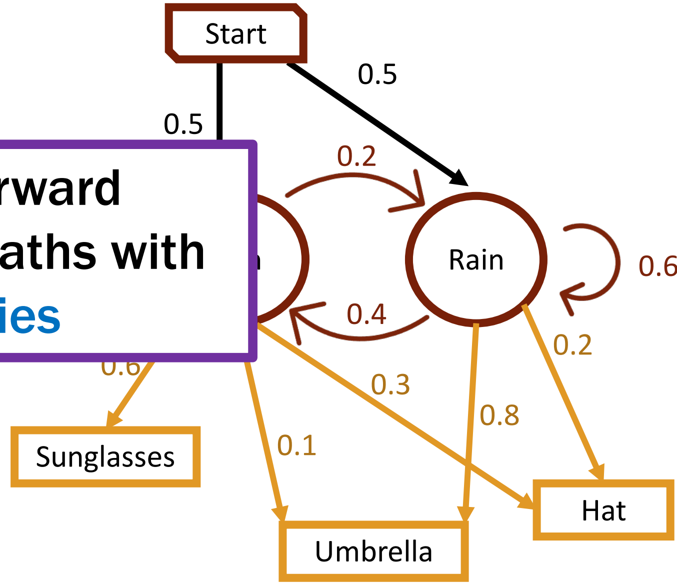
$t = 2$ Hat

$$\alpha_2(sun) = \alpha_1(sun) * a_{sun,sun} b_{sun}(H) + \alpha_1(rain)\ a_{rain,sun} b_{sun}(H)$$
$$= 0.3 * 0.24 + 0 * 0.12$$
$$= 0.072 + 0$$

$\alpha_1(sun) = 0.3$



Instead of adding the forward probabilities, we track the paths with the highest probabilities



$\alpha_1(rain) = 0$

$$\alpha_2(rain) = \alpha_1(sun) * a_{sun,rain} b_{rain}(H) + \alpha_1(rain)\ a_{rain,rain} b_{rain}(H)$$
$$= 0.3 * 0.04 + 0 * 0.12$$
$$= 0.012 + 0$$

The Viterbi Algorithm

Decoding problem: **Given an HMM $\Theta = (A, B)$ and an observation sequence Y, what is the most likely sequence of hidden states?**

Forward algorithm

1. **Base case:**

$$v_1(j) = \pi_j b_j(y_1);\ \ 1 \leq j \leq N$$
$$bt_1(j) = 0;\ \ 1 \leq j \leq N$$

$$\alpha_1(j) = \pi_j b_j(y_1);\ \ 1 \leq j \leq N$$

2. **Recursion:**

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(y_t)\,;\quad 1 \leq j \leq N, 1 \leq t \leq T$$

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\, b_j(y_t);$$

$$bt_t(j) = \operatorname*{argmax}_{i=1}^{N} v_{t-1}(i) a_{ij} b_j(y_t)\,;\quad 1 \leq j \leq N, 1 \leq t \leq T$$

2. **Termination:**

$$P^* = \max_{i=1}^{N} v_T(i)$$

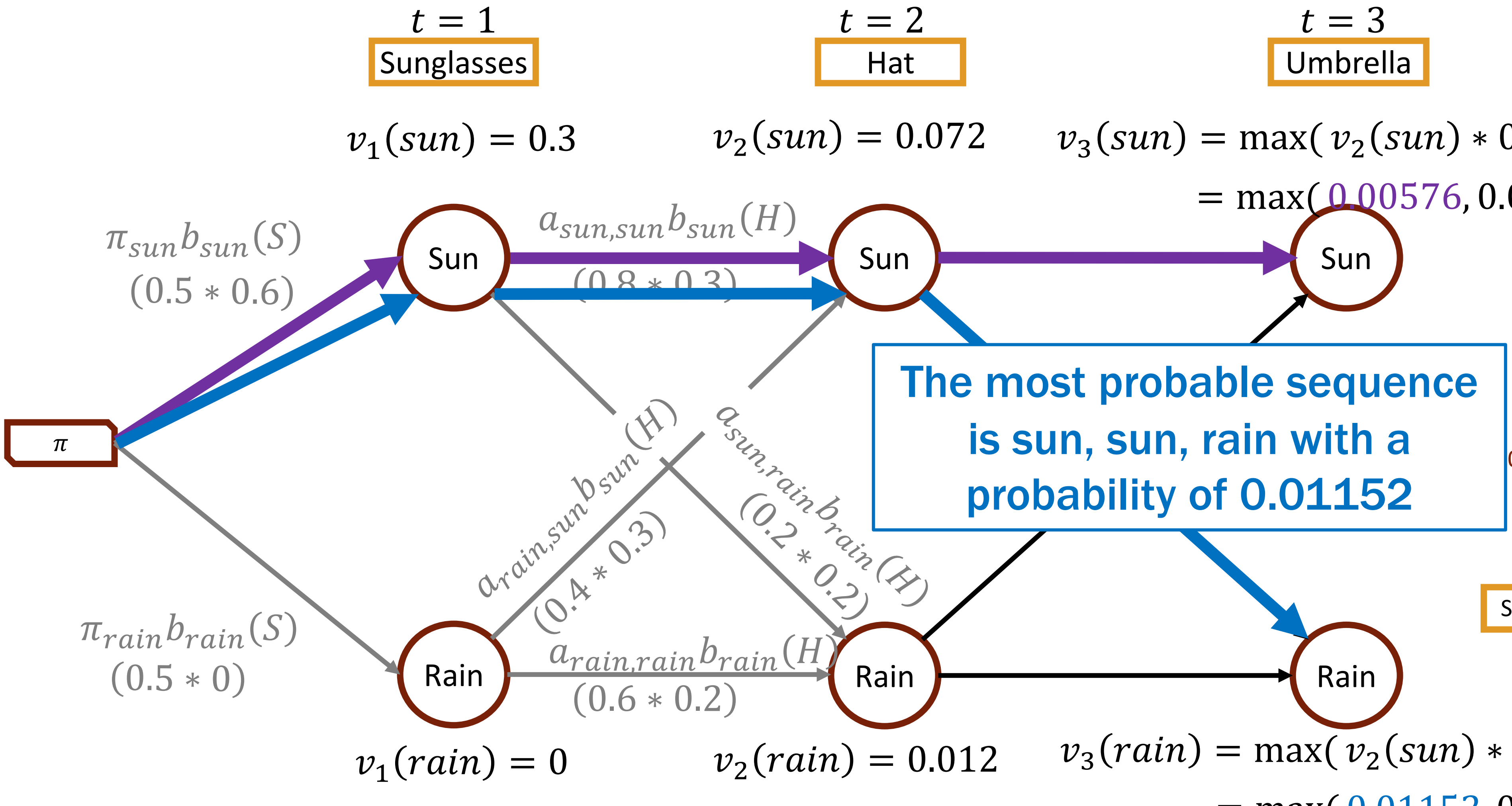
Probability of the best path

$$P(Y|\Theta) = \sum_{i=1}^{N} \alpha_T(i)$$

$$q_T^* = \operatorname*{argmax}_{i=1}^{N} v_T(i)$$

Backtrace to the best path

Tracing the last Viterbi step through the trellis



$v_1(sun) = 0.3$

$v_2(sun) = 0.072$

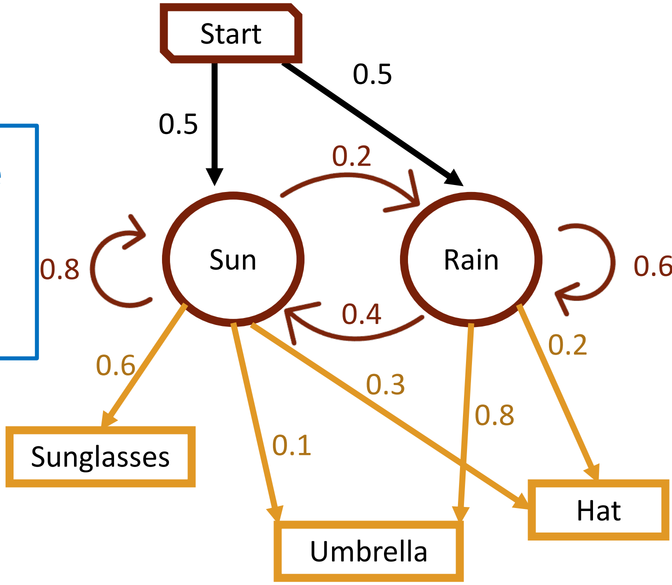
$$v_3(sun) = \max(\, v_2(sun) * 0.8 * 0.1, v_2(rain) * 0.4 * 0.1)$$
$$= \max(\, 0.00576, 0.00048)$$

The most probable sequence is sun, sun, rain with a probability of 0.01152

$v_1(rain) = 0$

$v_2(rain) = 0.012$

$$v_3(rain) = \max(\, v_2(sun) * 0.2 * 0.8\,, v_2(rain) * 0.6 * 0.8)$$
$$= \max(\, 0.01152, 0.00576)$$



Four typical HMM inference problems and one more

- Given the parameters of an HMM and an observed sequence of emissions,
  - what is the probability of that observed sequence?

    i. Scoring → Forward Algorithm, Backward Algorithm
  - what is the most likely sequence of hidden states?

    ii. Decoding → Viterbi Algorithm
  - what is the distribution of hidden states at time k?

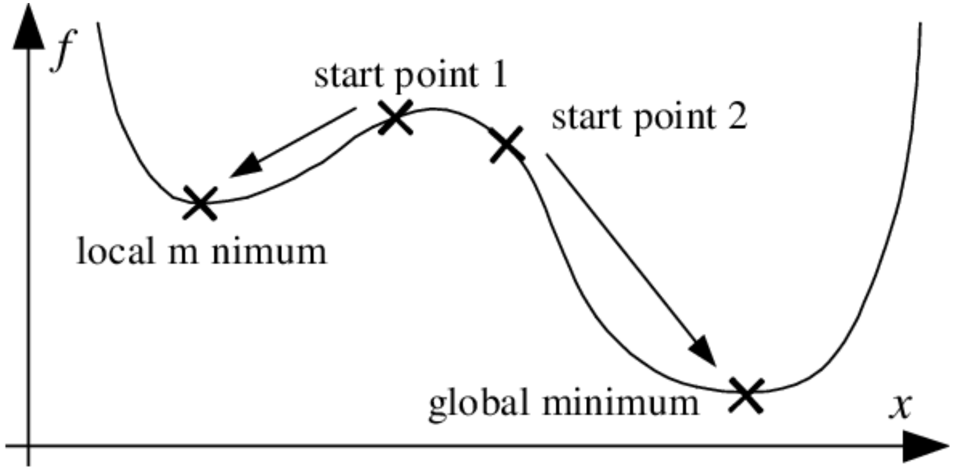
    iii. Filtering when $t = \mathrm{T}$ → Forward Algorithm

    iv. Smoothing when $t < \mathrm{T}$ → Forward-Backward Algorithm
- How do we parameterize an HMM to begin with?
  - Given all hidden states and emission sequences, how do we compute π, A and B?

    v. Parameterization → Baum-Welch Algorithm

Notes on the Baum-Welch Algorithm

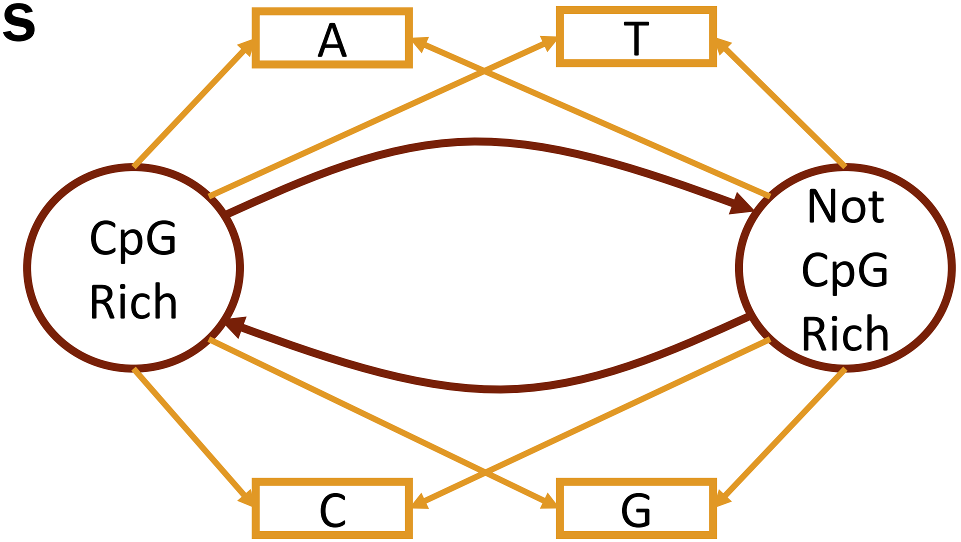
- The Baum-Welch is an expectation-maximization algorithm
  - An iterative method that converges on a local optimum
- It combines concepts from the forward and backward algorithms
- The outputs are estimates for initial probabilities π, transition probabilities A and emission probabilities B to characterize the HMM $\Theta = (A, B, \pi)$



https://www.researchgate.net/publication/228563694_A_genetic_algorithm_for_calculating_minimum_distance_between_convex_and_concave_bodies

An HMM for CpG islands

- CpG sites are dinucleotide pairings of 5’- cytosine - phosphate - guanine - 3’
- CpG islands have a high frequency of CpG sites
    - Associated with gene promoters
    - Almost all housekeeping genes have promoters in CpG islands
- Finding CpG islands can help us identify genes



- What are reasonable guesses for (A, B, π)?

Intuition on the Baum-Welch Algorithm

- Given a DNA sequence (emission sequence), we can guess at the underlying hidden states

AGTGGATGCTGACGCGCGCGCGCGCGC

NNNNNNNNNNNNNNNNNRRRRRRRRRR

GCGCGCGCGCATGATTAAGCGTACCTC

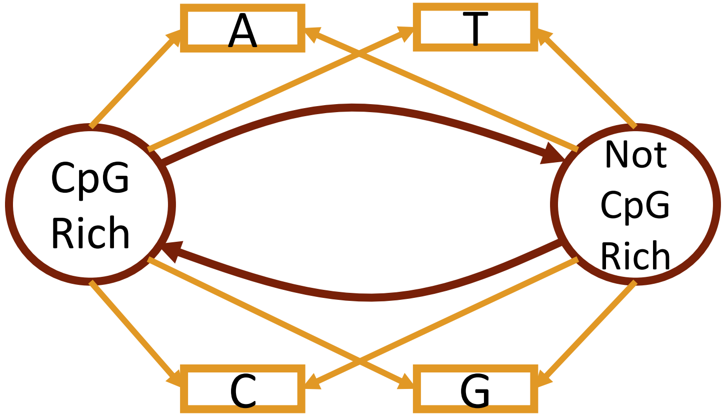
RRRRRRRRRRRRRRRRRNNNNNNNNNN

ATCTCACCAATACATAGAGAGAGTACA

NNNNNNNNNNRRRRRNNNNNNNNNNNN

TATCGCGCGCGCGCGCGCGCGCGCGCG

NNNRRRRRRRRRRRRRRRRRRRRRRRR

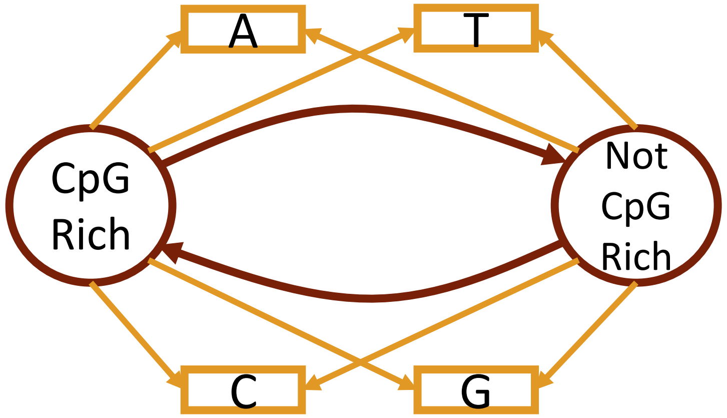


- We can estimate $\left(\hat{A}, \hat{B}, \hat{\pi}\right)$ based on this guess
- This is our initialization of the algorithm

Intuition on the Baum-Welch Algorithm

1. Using estimated $(\hat{A}, \hat{B}, \hat{\pi})$, we look for the most likely hidden state path

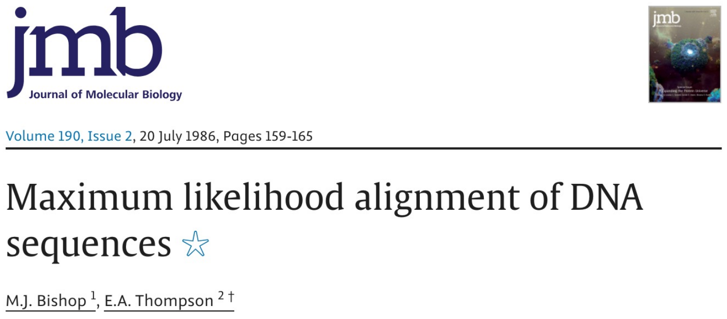
A G T G G A T G C T G A C G C G C G C G C G C G C G C
N N N N N N N N N N N N N N N R R R R R R R R R R R R
G C G C G C G C G C A T G A T T A A G C G T A C C T C
R R R R R R R R R R R R R R R N N N N N N N N N N N N
A T C T C A C C A A T A C A T A G A G A G A G T A C A
N N N N N N N N N N N N N R R N N N N N N N N N N N N
T A T C G C G C G C G C G C G C G C G C G C G C G C G
N N N R R R R R R R R R R R R R R R R R R R R R R R R



2. Based on the new estimated hidden sequence, re-estimate $(\hat{A}, \hat{B}, \hat{\pi})$
3. Repeat 1 - 2 until convergence

Why do we care about HMMs?

- Earliest applications in speech recognition (1970s)
- They have since found widespread applications in myriad fields
  - Think of any system that involves change between states
  - And where we observe data emitted from those states
- First bioinformatics application in DNA sequence alignment (1986)
- And more:
  - Gene prediction
  - Pairwise and multiple sequence alignment
  - Base-calling
  - Protein structure prediction
  - Chromatin domains
  - Copy number calling



Volume 190, Issue 2, 20 July 1986, Pages 159-165

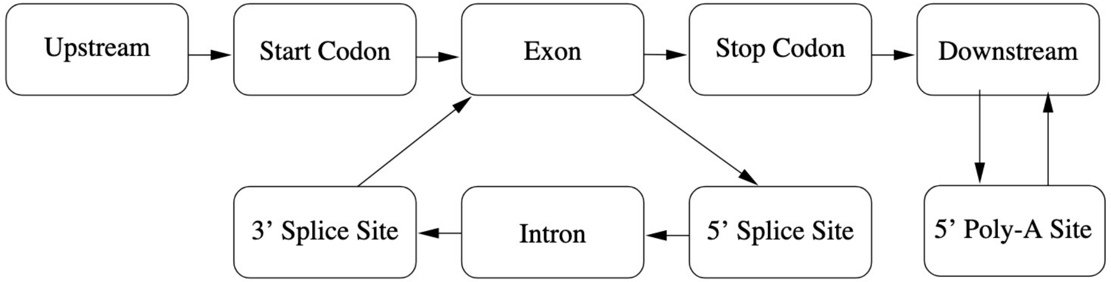
Maximum likelihood alignment of DNA sequences ☆

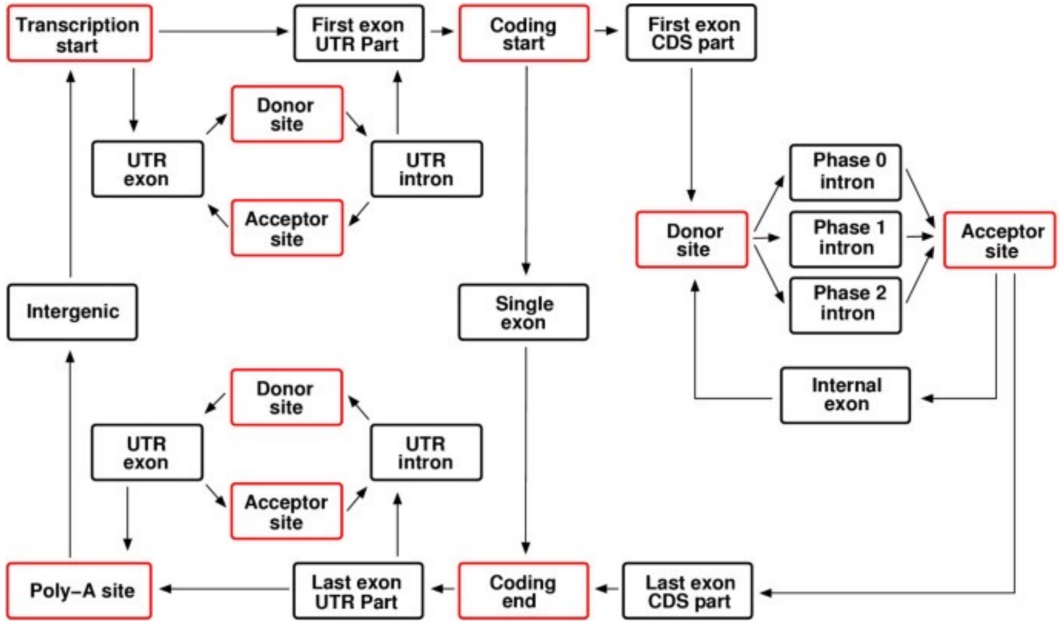
M.J. Bishop [1], E.A. Thompson [2 †]

More history? Check out https://ethw.org/First-Hand:The_Hidden_Markov_Model
https://www.jehps.net/Novembre2006/Markov3pdf.pdf

Gene prediction

- We saw a simple HMM for CpG islands
  - Encode underlying characteristics of a DNA sequence
- We can extend the idea to genome annotation
- What characterizes a gene?
  - What is a gene?
  - Where is the start? The end?
  - What differentiates exon from intron?
- After learning the parameters, we can compute the most likely hidden state sequence and identify genic regions

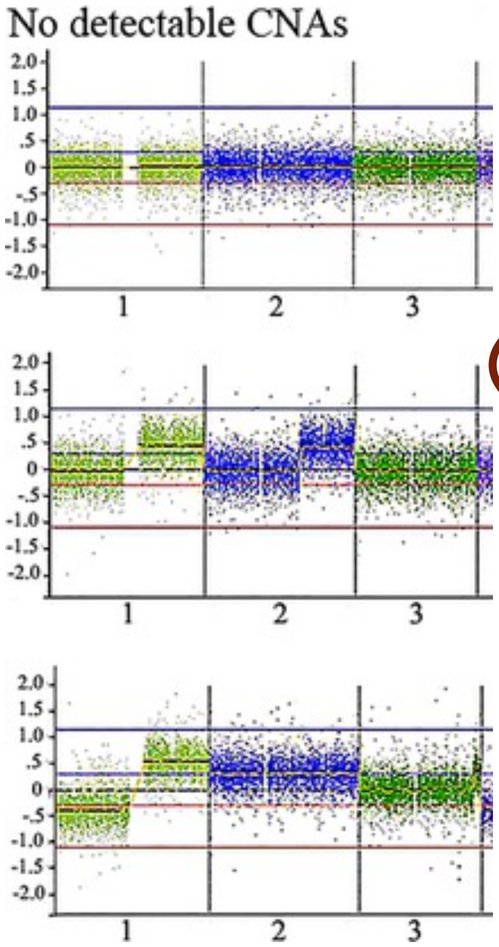


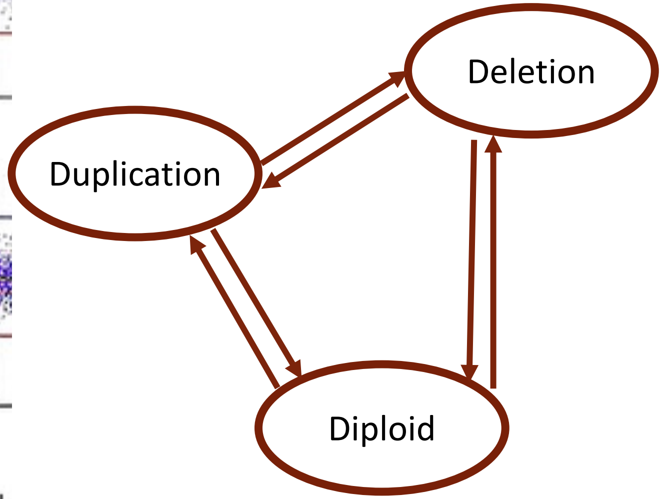


VEIL: http://www.ece.iit.edu/~biitcomm/research/resource/Gene%20Identification/veil.pdf
Agene: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-263

Detecting copy number changes

- Normal human genome is diploid
- An individual may:
  - inherit variation in the copy number of a gene (copy number variant; CNV)
  - acquire copy number changes in lifetime (copy number alteration; CNA)
  - CNAs frequently observed and implicated in cancer
- We are interested in calling copy number changes from DNA microarray and sequencing data





https://www.frontiersin.org/journals/oncology/articles/10.3389/fonc.2021.741526/full

Sequence alignment (especially proteins)

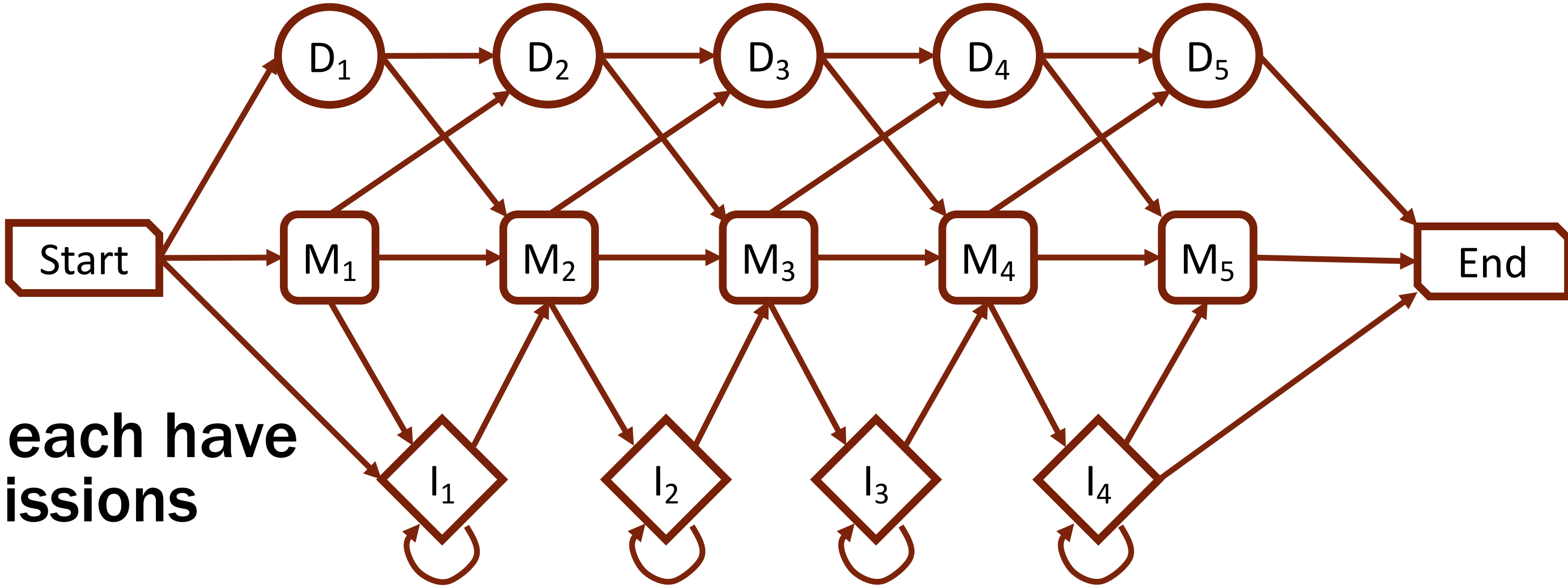
- Given a protein, identify its protein family
    - A group of proteins that share an evolutionary origin
    - They share similarities in function, structure, and sequence
- We can construct an HMM describing sequence patterns in a protein family
    - Profile HMM
    - Family protein sequences can be aligned or not aligned (learn by Baum-Welch EM)
- Given a new protein sequence, we can check its family membership and the most likely alignment
- Pfam is a database of profile HMMs
    - 21,979 families as of 2024

Sequence alignment (especially proteins)

- Example of a Profile HMM

Aligned sequences

Seq1	H N Y - H S
Seq2	H H Y - H G
Seq3	N H Y - - S
Seq4	T N Y g F S
Seq5	N G Y - H G



- Match and Insertion states each have 20 possible amino acid emissions
- Deletion states are silent
- We can use the aligned sequences to parameterize the profile HMM

Take homes

- The Hidden Markov model is a powerful tool for modeling systems we cannot directly observe
- We use dynamic programming to efficiently solve inference problems
  - We saw the Forward and Viterbi algorithms
- We use expectation maximization to parameterize, remembering that we are not guaranteed the absolute best answer
  - We have some intuition for the Baun-Welch algorithm
- HMMs are common in bioinformatics
  - Genome annotation (including CpG island and gene identification)
  - Copy number calling
  - And especially sequence alignment

Additional resources

- https://github.com/conniehli/HMM_Materials
  - Formal derivation of the Forward algorithm + R implementation
  - Introduction to the Backward algorithm + exercises
  - Self study exercises on the Forward-Backward and Baum-Welch algorithms
  - Exercises expanding on our protein family profile HMM
- Interested in HMMs in sequence alignment?
  - *Biological sequence analysis,* Cambridge University Press Durbin, Eddy, Krogh, & Mitchison (1998)
  - *Pfam: The protein families database in 2021*, Nucleic Acids Research, Mistry et al (2021)
- Curious about gene finding?
  - *Finding genes in DNA with a Hidden Markov Model*, Henderson, Salzberg & Fasman, Journal of Computational Biology (1997)
  - *Using database matches with HMMGene for automated gene detection in* Drosophila, Krogh, Genome Research (2000)