

Probabilistic Models for Biological Sequences

Laboratory of Bioinformatics I
Module 2

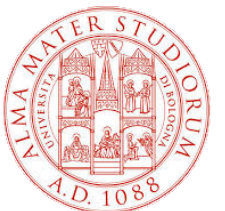
Emidio Capriotti

<http://biofold.org/>



Biomolecules
Folding and
Disease

Department of Pharmacy and
Biotechnology (FaBiT)
University of Bologna



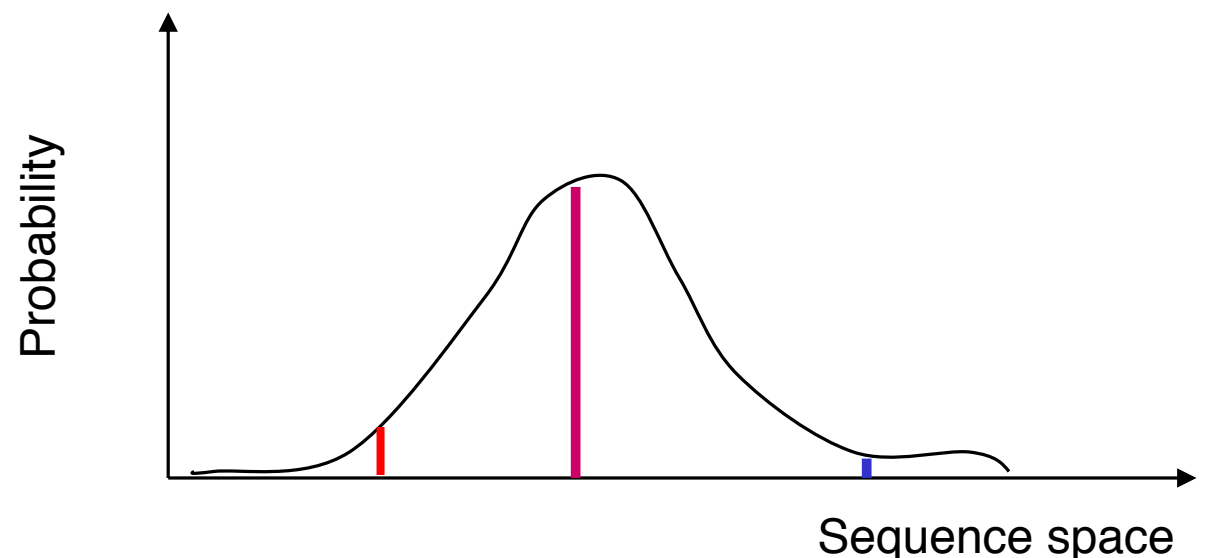
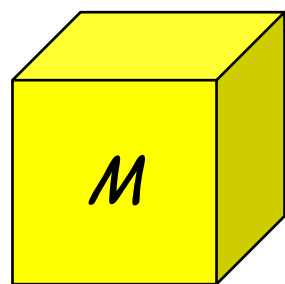
Models for Sequence

Generative definition:

- Objects producing **different outcomes (sequences)** with different probabilities
- The **probability distribution** over the sequences space determines the model specificity

Generates s_i with probability $P(s_i \mid M)$

e.g.: M is the representation of the family of globins



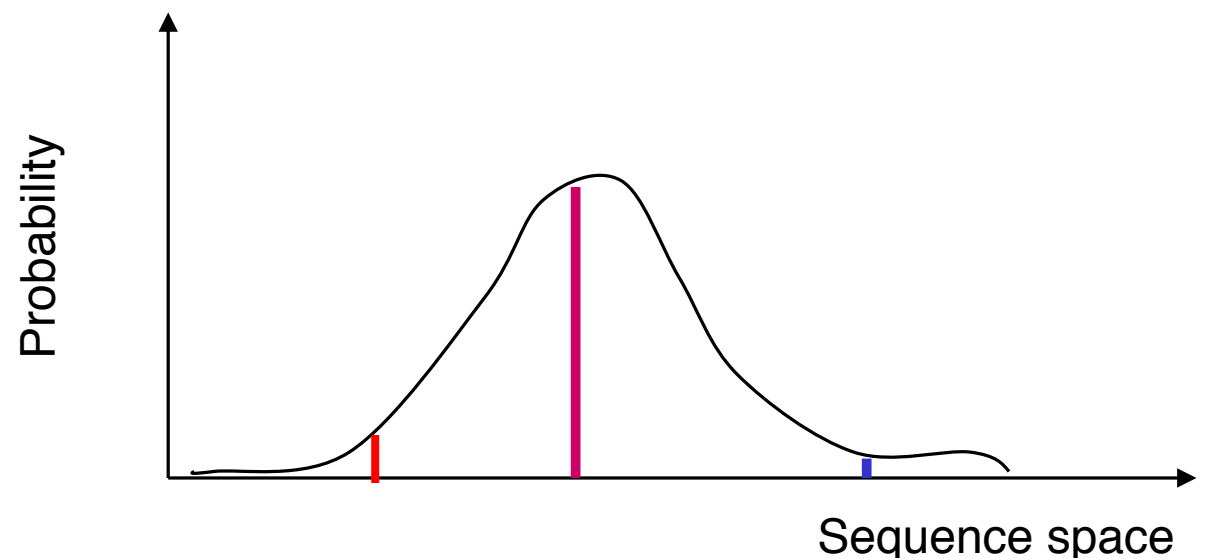
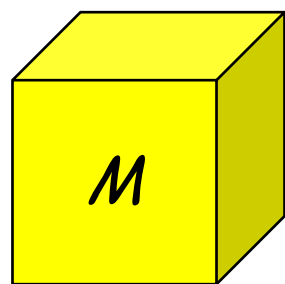
Associative Definition

The generative definition is useful as operative definition

- Objects that, **given an outcome (sequence), compute a probability value**

Calculates the associated probability $P(s_i \mid M)$ to s_i .

e.g.: M is the representation of the family of globins

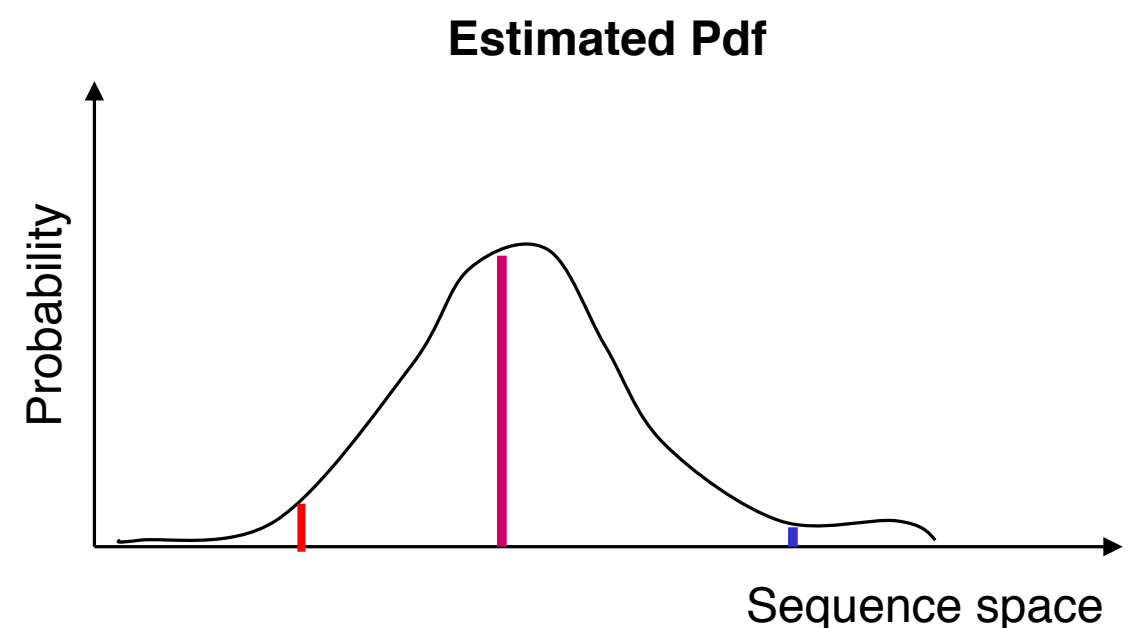
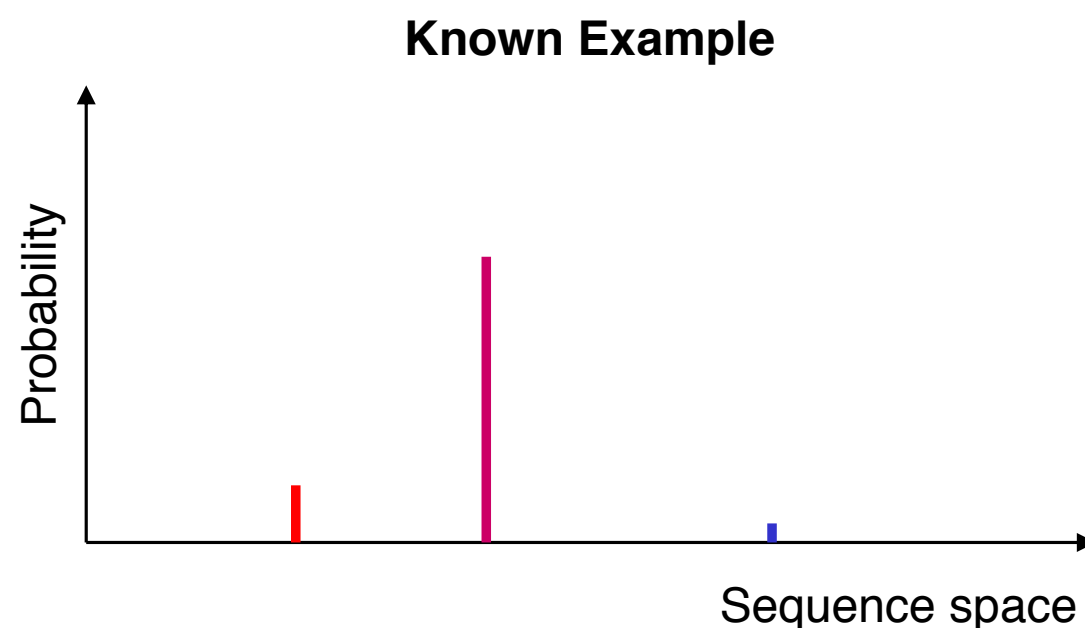


Which Model?

The most useful probabilistic models are Trainable systems

The **probability density function** over the sequence space **can be estimated** from known examples by a learning algorithm


Define a generic representation of the sequences of globins starting from a set of known globins



Similarity Measure

Given a class of proteins (e.g. Globins), a probabilistic model trained on this family can be **adopted to compute a probability value for new sequences**

Seq1	0.98
Seq2	0.21
Seq3	0.12
Seq4	0.89
Seq5	0.47
Seq6	0.78



This value measures the **similarity between the new sequence and the family** described by the model

Which Probability?

A model M associates to a sequence s_i the probability $P(s_i \mid M)$

This probability answers the question:

Which is the probability for a model M (e.g. describing the Globins) to generate the sequence s_i ?

The question we want to answer is:

Given a sequence s_i , does it belong to the class described by the model M ? (e.g. is it a Globin?)

We need to compute $P(M \mid s_i)$

Bayes Theorem

$$P(X, Y) = P(X | Y) P(Y) = P(Y | X) P(X) \quad \text{Joint probability}$$

$$P(Y | X) = \frac{P(X | Y) P(Y)}{P(X)}$$

$$P(M | s_i) = \frac{P(s_i | M) P(M)}{P(s_i)}$$

$P(M)$ and $P(s_i)$
A priori probabilities

$P(M)$ is the probability of the model (i.e. of the class described by the model)
BEFORE we know the sequence:

Can be estimated as the **abundance of the class**

$P(s_i)$ is the probability of the sequence in the sequence space.

Cannot be reliably estimated!!

Comparing Models

We can overcome the problem comparing the probability of generating s_i from different models

$$\frac{P(M_1 | s_i)}{P(M_2 | s_i)} = \frac{P(s_i | M_1) P(M_1)}{P(s_i)} \frac{P(s_i)}{P(s_i | M_2) P(M_2)} = \frac{P(s_i | M_1) P(M_1)}{P(s_i | M_2) P(M_2)}$$

$$\frac{P(M_1)}{P(M_2)}$$

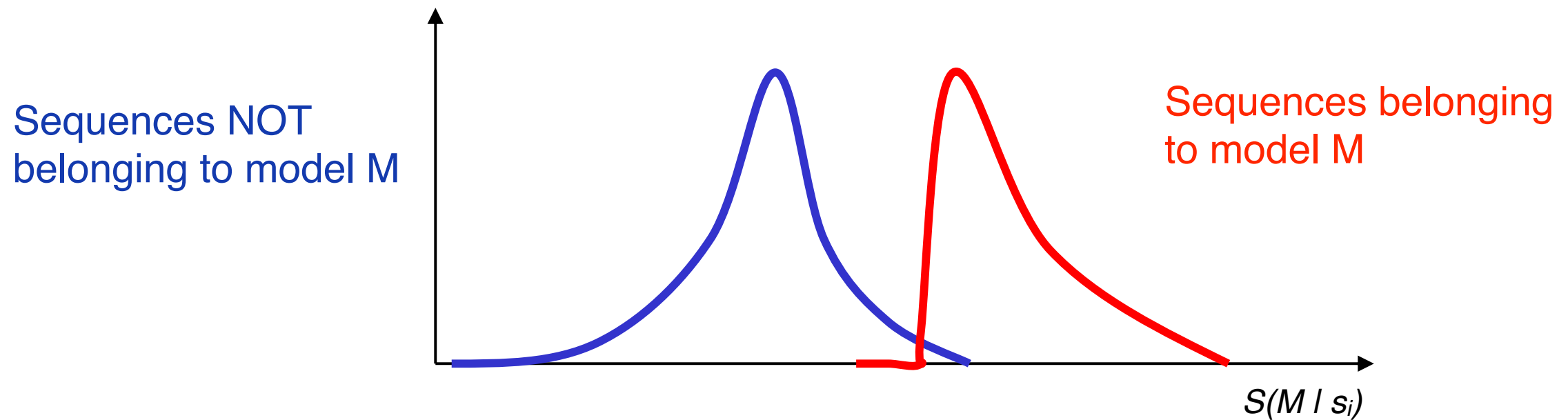
Ratio between the abundances of the classes

Null Model

Alternatively, we can score a sequence for a model M comparing it to a Null Model:

a model that generates ALL the possible sequences with probabilities depending ONLY on letter (e.g. residue) statistical abundance

$$S(M \mid s_i) = \log \frac{P(s_i \mid M)}{P(s_i \mid N)}$$



In this case we need a threshold and a statistic for evaluating the significance (E-value, P-value)

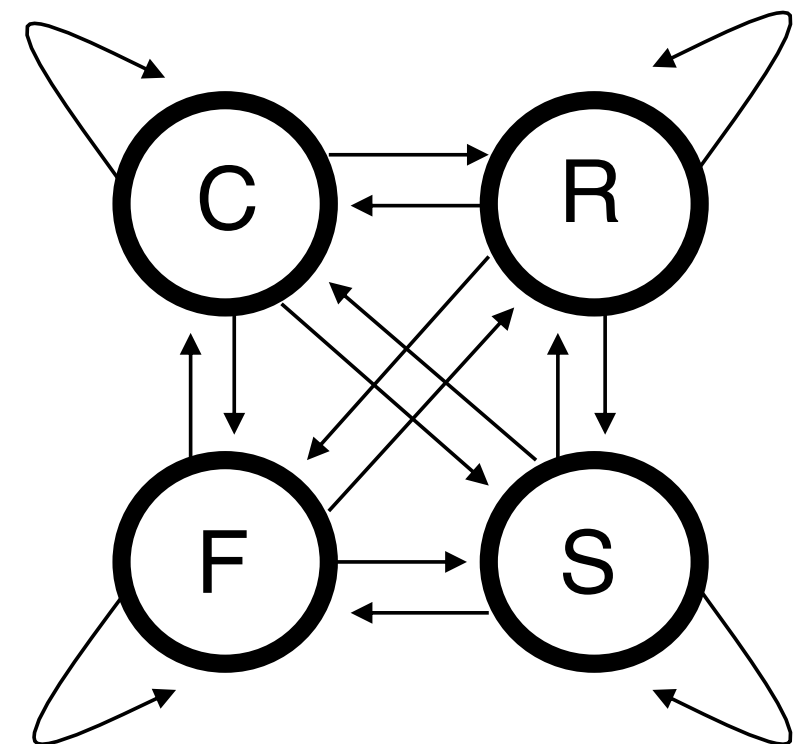
A Simple Model

Time series of the weather conditions

as a first hypothesis the weather condition in a day probabilistically depends ONLY on the weather conditions in the day before.

Define the conditional probabilities

$P(C|C)$, $P(C|R)$, $P(R|C)$,



The probability for the 5-days registration
CRRCS

$$P(CRRCS) = P(C) \cdot P(R|C) \cdot P(R|R) \cdot P(C|R) \cdot P(S|R)$$

C: Clouds

R: Rain

F: Fog

S: Sun

Markov Model

Stochastic generator of sequences in which the probability of state in position i depends ONLY on the state in position $i-1$

Given a set of states (== alphabet)

$$C = \{C_1; C_2; C_3; \dots C_N\}$$

a Markov model is described with $N \times (N+2)$ parameters

$$\{a_{r,t}, a_{BEGIN,t}, a_{r,END} \text{ with } r, t \in C\}$$

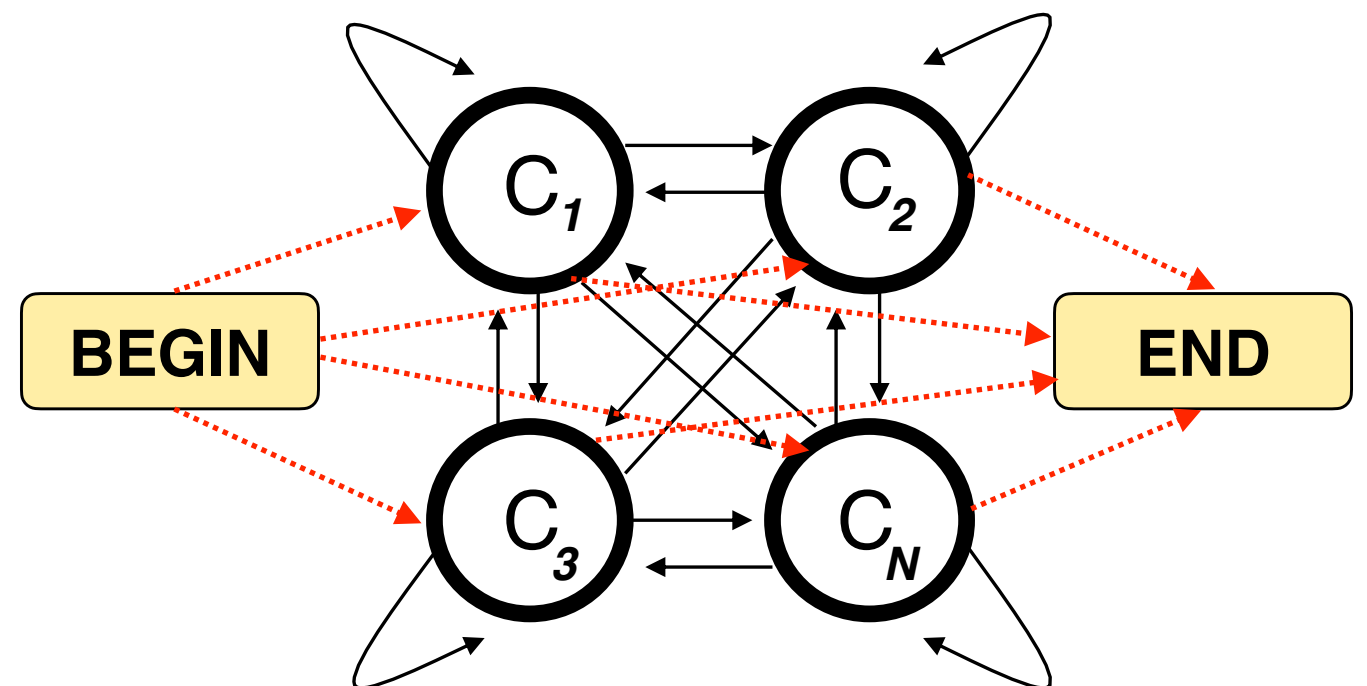
$$a_{r,q} = P(s_i = q \mid s_{i-1} = r)$$

$$a_{BEGIN,q} = P(s_1 = q)$$

$$a_{r,END} = P(s_T = END \mid s_{T-1} = r)$$

$$\sum_t a_{r,t} + a_{r,END} = 1 \quad \forall r$$

$$\sum_t a_{BEGIN,t} = 1$$



All transitions going out from a state sum up to 1

Sequence Probability

Given the sequence:

$$S = s_1 s_2 s_3 s_4 s_6 \dots s_T \quad \text{with} \quad s_i \in C = \{c_1; c_2; c_3; \dots c_N\}$$

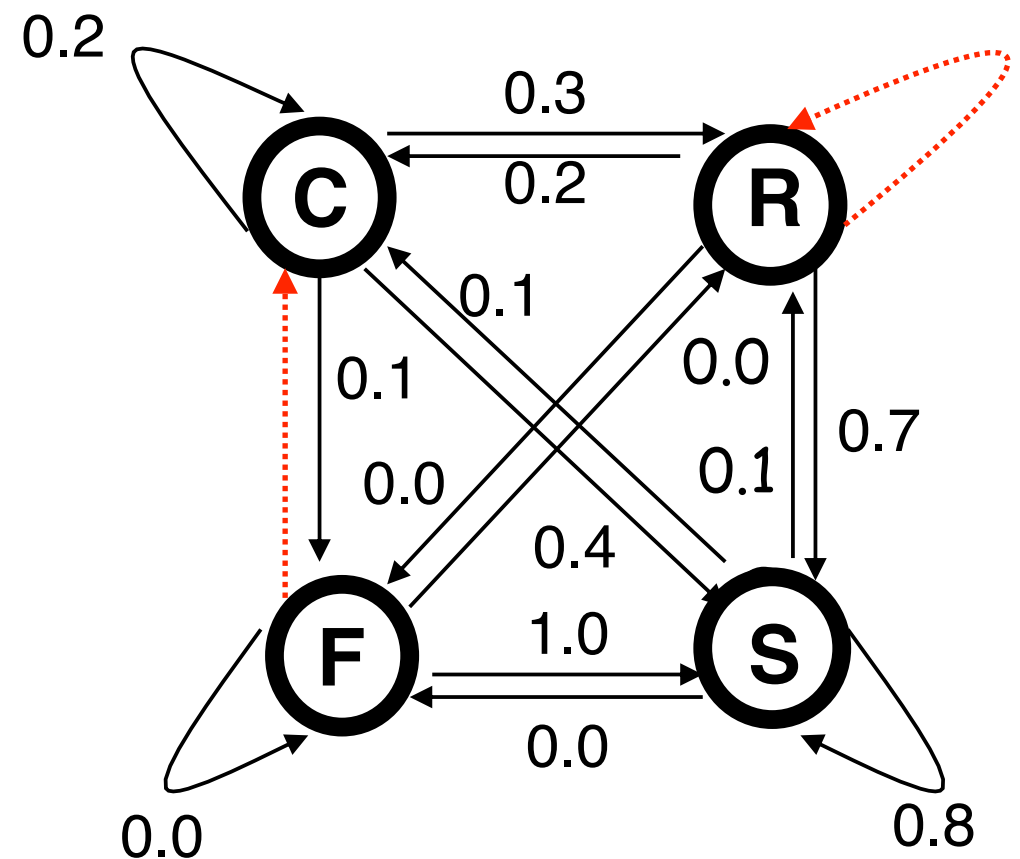
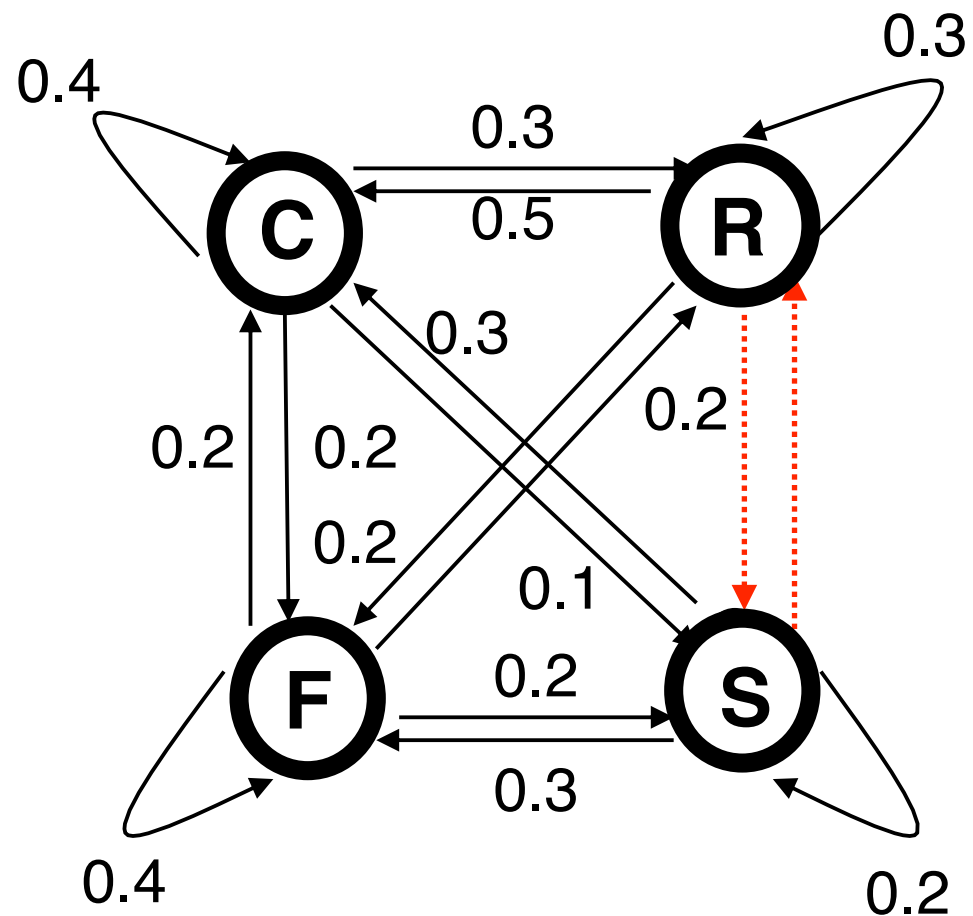
$$P(s | M) = P(s_1) \prod_{i=2}^T P(s_i | s_{i-1}) =$$

$$a_{BEGIN, s_1} \times \prod_{i=2}^T a_{s_{i-1}, s_i} \times a_{s_T, END}$$

$$P(ALKALI) = a_{BEGIN, A} \times a_{A, L} \times a_{L, K} \times a_{K, A} \times a_{A, L} \times a_{L, I} \times a_{I, END}$$

Probability Constrains

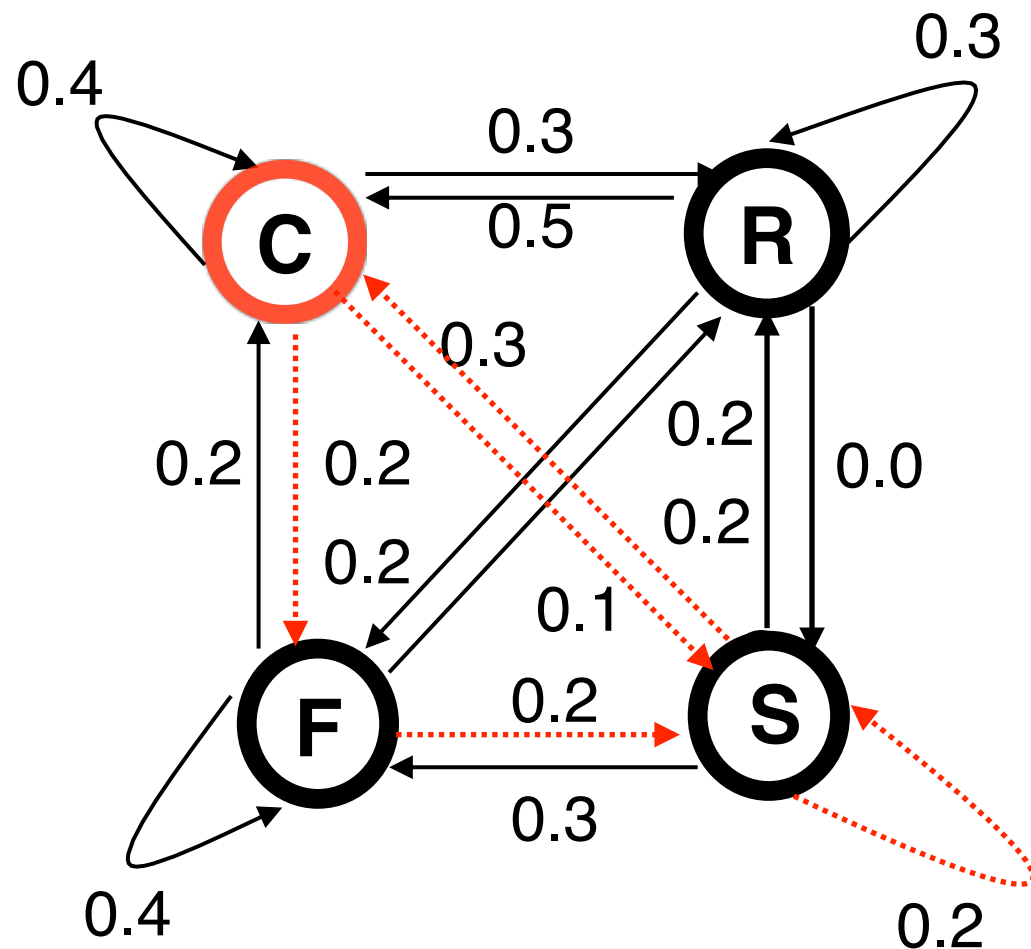
What are the missing probabilities given the constraints?



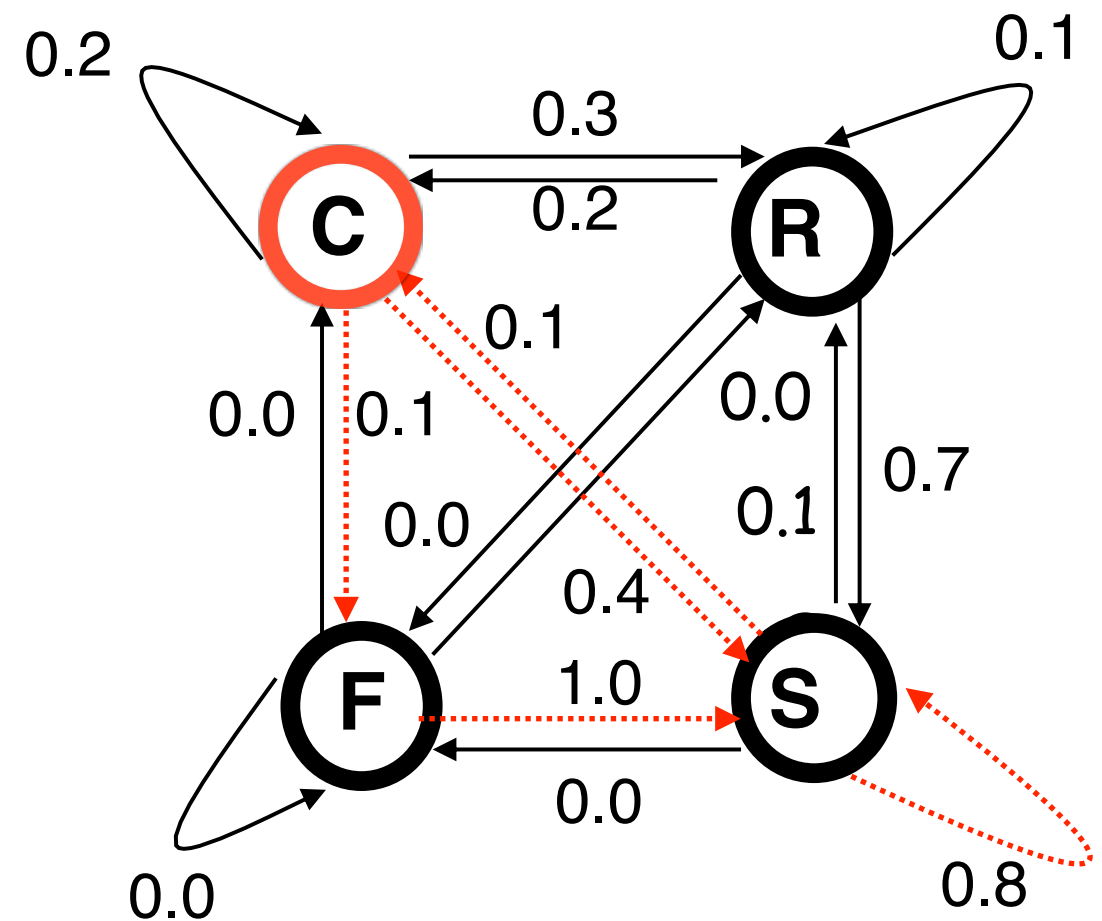
What is the better model to describe the **weather in winter**?

Probability Calculation

Consider the sequence “CSSSCFS” and calculate its probability with both models
when $P(X \mid \text{BEGIN}) = 0.25$



$$P(\text{CSSSCFS} \mid \text{Winter}) = 0.25 \times 0.1 \times 0.2 \times 0.2 \times 0.3 \times 0.2 \times 0.2$$

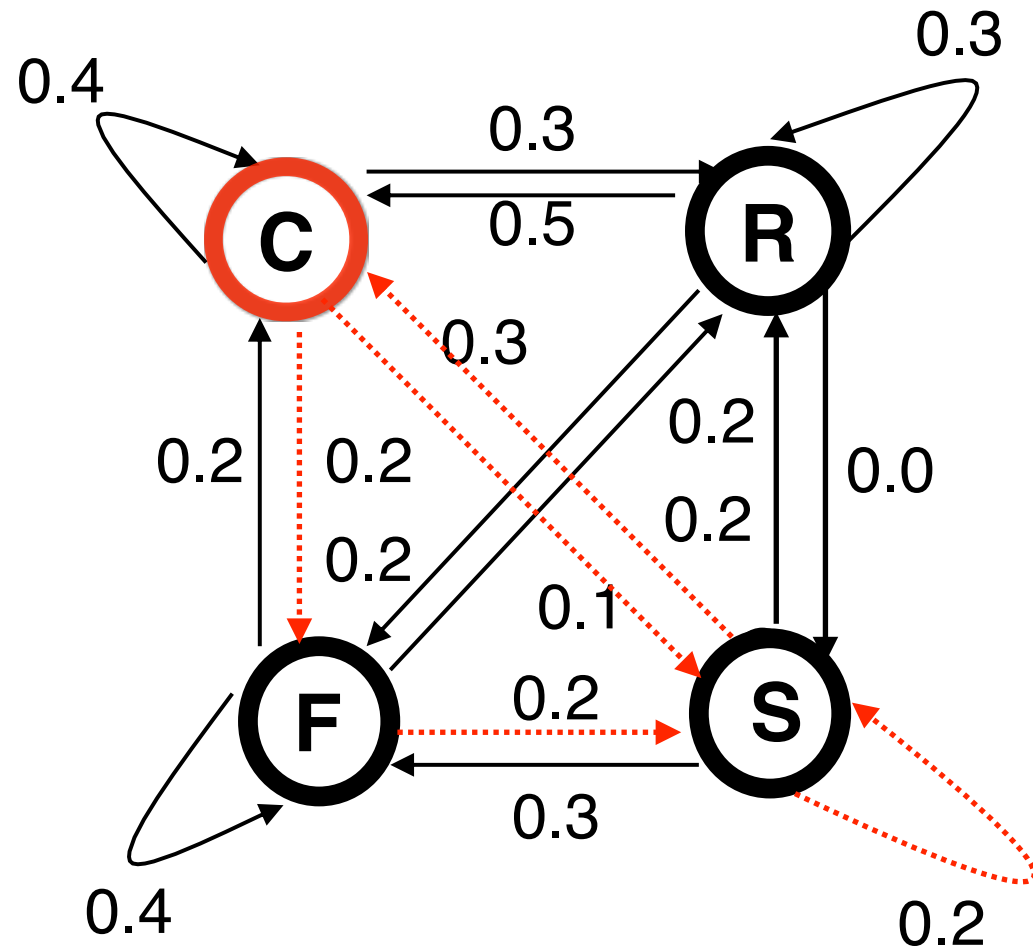


$$P(\text{CSSSCFS} \mid \text{Summer}) = 0.25 \times 0.4 \times 0.8 \times 0.8 \times 0.1 \times 0.1 \times 1.0$$

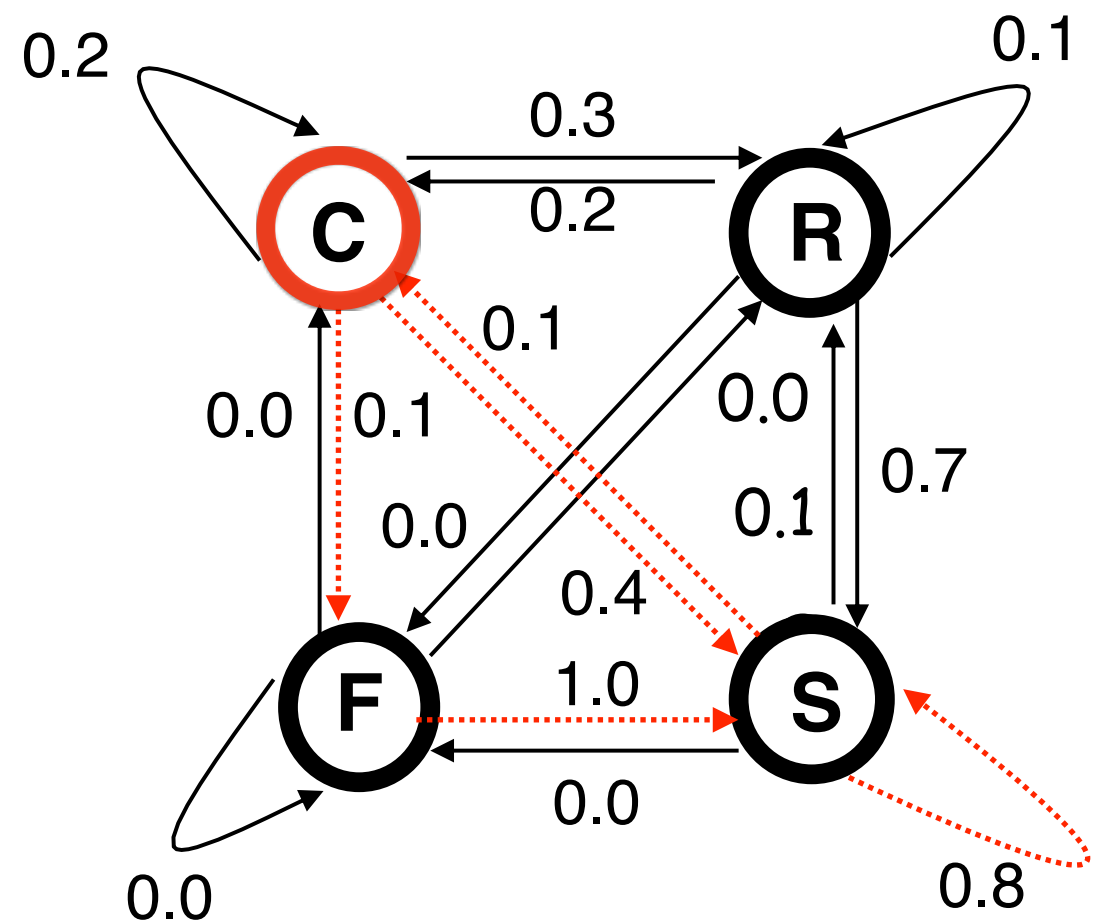
To which **season** the weather sequence is **more likely to belong**?

Models Comparison

$P(\text{Seq} \mid \text{Winter}) = 1.2 \times 10^{-5}$

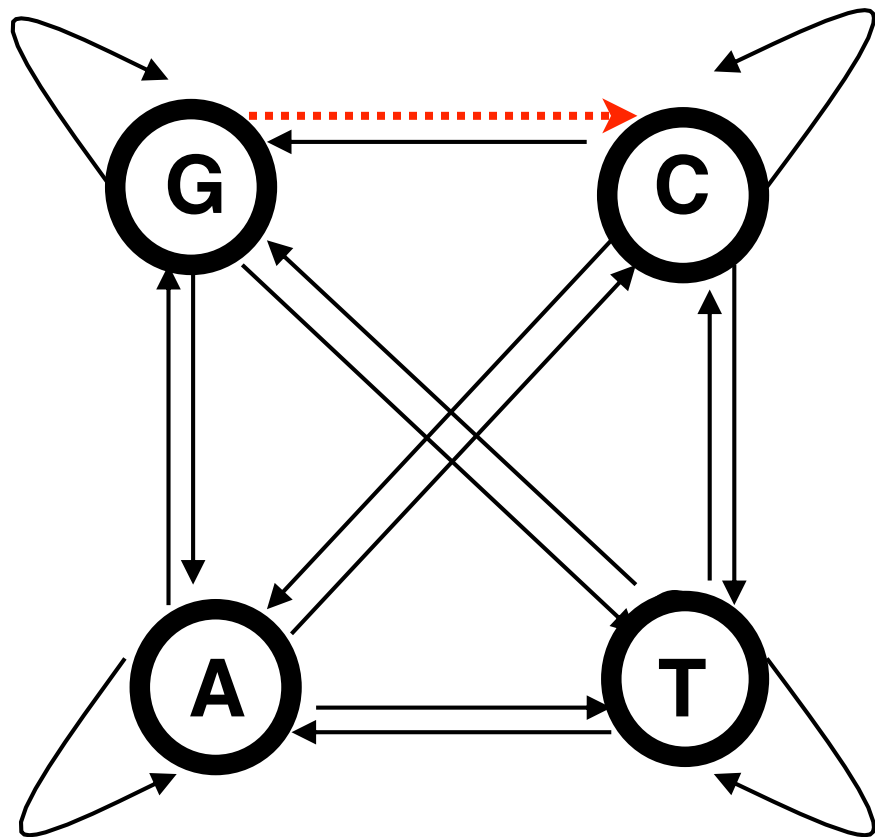


$P(\text{Seq} \mid \text{Summer}) = 6.4 \times 10^{-4}$

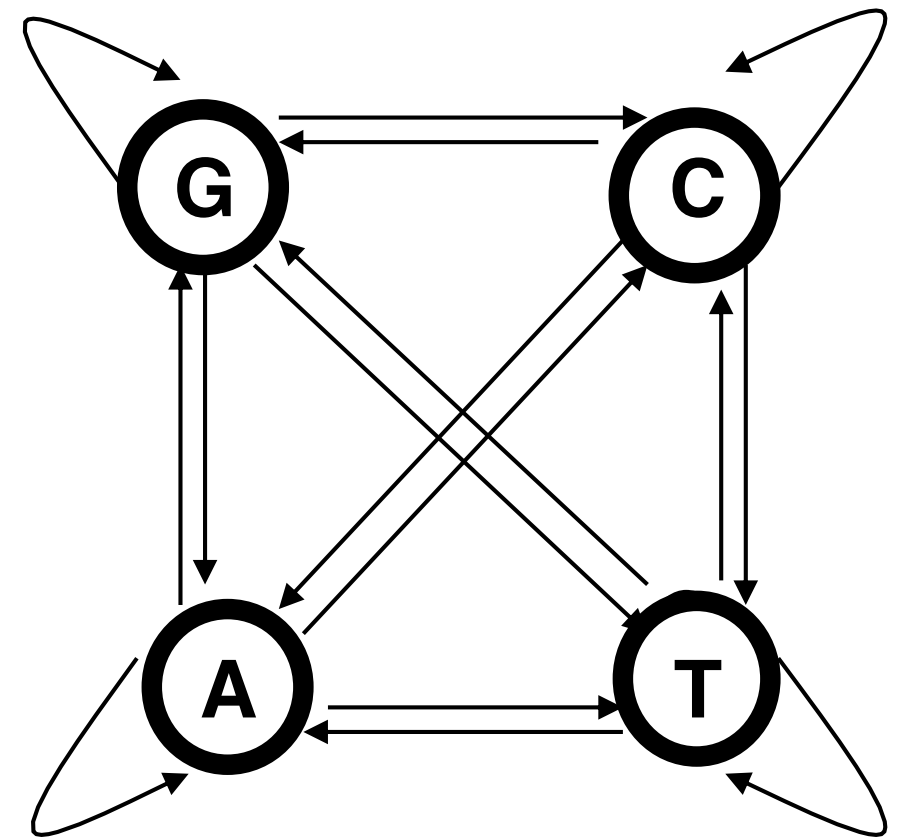


$$\frac{P(\text{Summer} \mid \text{Seq})}{P(\text{Winter} \mid \text{Seq})} = \frac{P(\text{Seq} \mid \text{Summer})}{P(\text{Seq} \mid \text{Winter})} \times \frac{P(\text{Summer})}{P(\text{Winter})} \quad \text{with} \quad \frac{P(\text{Summer})}{P(\text{Winter})} \approx 1$$

Modeling CpG Islands



CpG Islands



Non-CpG Islands

In the Markov Model of CpG Islands a_{GC} is higher than in Markov Model Non-CpG Islands

$$P(\text{CpG} \mid s) = \frac{P(s \mid \text{CpG}) \times P(\text{CpG})}{P(s \mid \text{CpG}) \times P(\text{CpG}) + P(s \mid \text{not CpG}) \times P(\text{notCpG})}$$

Demonstration

We assume that only two models (CpG and notCpG) are possible.

$$P(s) = P(s, \text{CpG}) + P(s, \text{notCpG})$$

Given the Bayes Theorem

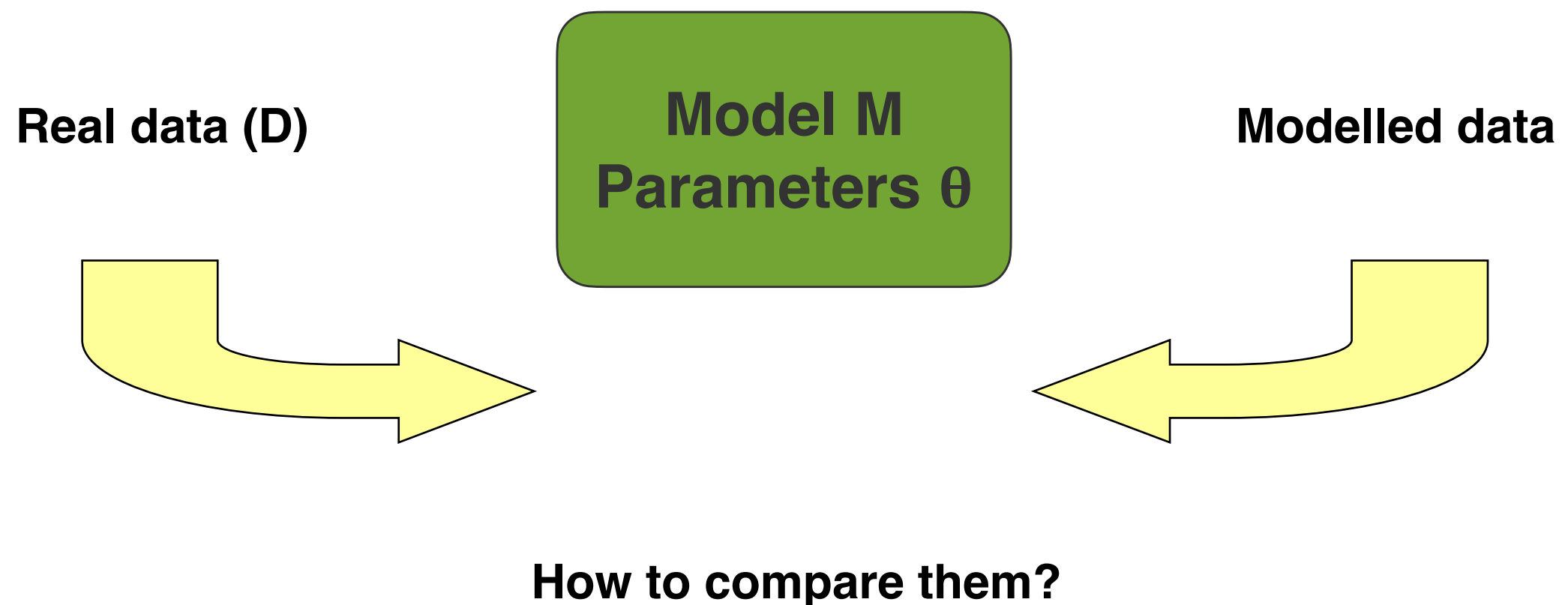
$$P(s) = \frac{P(s \mid \text{CpG}) \times P(\text{CpG})}{P(\text{CpG} \mid s)} = P(s \mid \text{CpG}) \times P(\text{CpG}) + P(s \mid \text{notCpG}) \times P(\text{notCpG})$$

Thus,

$$P(\text{CpG} \mid s) = \frac{P(s \mid \text{CpG}) \times P(\text{CpG})}{P(s \mid \text{CpG}) \times P(\text{CpG}) + P(s \mid \text{notCpG}) \times P(\text{notCpG})}$$

Training of the Method

Generally speaking, a parametric model M aims to reproduce a set of known data



Maximum Likelihood

Let θ_M be the set of parameters of model M.

During the training phase, θ_M parameters are estimated from the set of known data D

Maximum Likelihood Estimation (ML)

$$\theta^{ML} = \operatorname{argmax}_{\theta} P(D \mid M, \theta)$$

Training Proof

Given a sequence s contained in D : $s = s^1 s^2 s^3 s^4 s^6 \dots s^T$

$$P(s \mid M) = a_{BEGIN, s^1} \cdot \prod_{i=2}^{T-1} a_{s^i s^{i+1}} \cdot a_{s^T END}$$

We can count the number of transitions between any two states j and k : n_{jk}

$$P(s \mid M) = \prod_{j=0}^{N+1} \prod_{k=0}^{N+1} a_{jk}^{n_{jk}}$$

Where states 0 and $N+1$ are BEGIN and END

On top of this, keep in mind that **normalization constraints must be satisfied** for each state

$$\forall j: \sum_{k'=0}^N a_{jk'} = 1$$

So the likelihood has to be maximized on the variety defined by the normalization constraints. **How we do that?**

Maximum Likelihood

Let θ_M be the set of parameters of model M.

During the training phase, θ_M parameters are estimated from the set of known data D

Maximum Likelihood Estimation (ML)

$$\theta^{ML} = \operatorname{argmax}_{\theta} P(D \mid M, \theta)$$

It can be proved that: $a_{ik} = \frac{n_{ik}}{\sum_j n_{ij}}$ Frequency of occurrence as counted in the data set D

Maximum A Posteriori Estimation

$$\theta^{MAP} = \operatorname{argmax}_{\theta} P(\theta \mid M, D) = \operatorname{argmax}_{\theta} [P(D \mid M, \theta) \times P(\theta)]$$

Example with Dice

We have 99 regular dice (***R***) and 1 loaded die (***L***).

	P(1)	P(2)	P(3)	P(4)	P(5)	P(6)
<i>R</i>	1/6	1/6	1/6	1/6	1/6	1/6
<i>L</i>	1/10	1/10	1/10	1/10	1/10	1/2

Given a sequence of numbers:

4156266656321636543662152611536264162364261664616263

What is the sequence of dice that generated it?

RRRRRLRLRRRRRRRLRRRRRRRRRRRLRLRRRRRRRRRLRRRLRRRLRR

Hypothesis

We chose a different die for each roll

Two stochastic processes give origin to the sequence of observations.

1) Choosing the die (R o L).

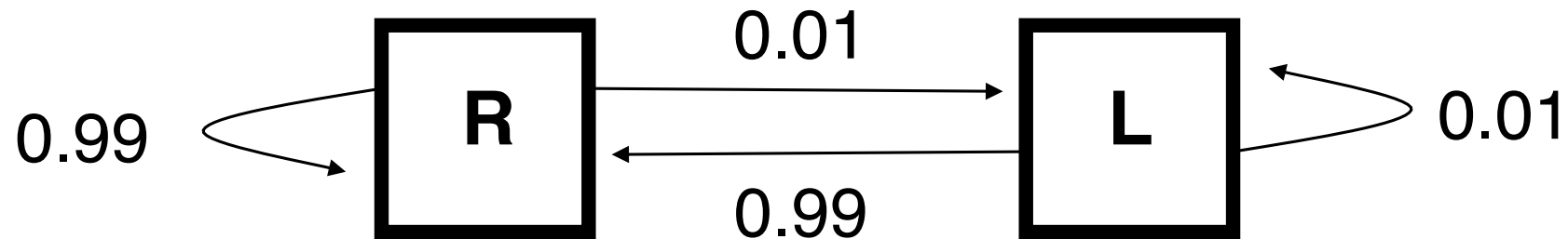
2) Rolling the die

The sequence of dice is hidden

The first process is assumed to be Markovian (in this case a 0-order MM)

The outcome of the second process depends only on the state reached in the first process (that is the chosen die)

Casinò



Each state (***R*** and ***L***) generates a character of the alphabet $\mathbf{C} = \{1, 2, 3, 4, 5, 6\}$

The emission probabilities depend only on the state.

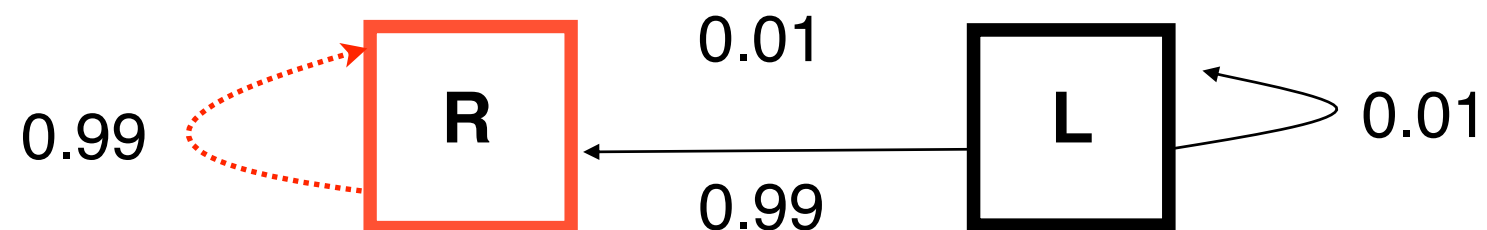
The transition probabilities describe a Markov model that generates a state path: the hidden sequence (π)

The observations sequence (s) is generated by two concomitant stochastic processes

One Step

415626665632163654366215261

RRRRRLRLRRRRRRRLRRRRRRRRRRR



Choose the State : R Probability= 0.99

Chose the Symbol: 1 Probability= 1/6 (given R)

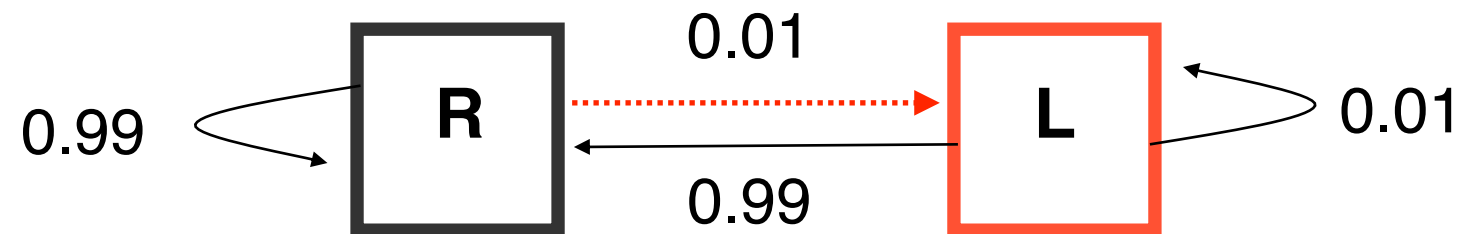
415626665632163654366215261**1**

RRRRRLRLRRRRRRRLRRRRRRRRRRR**R**

Alternative Step

415626665632163654366215261

RRRRRLRLRRRRRRRLRRRRRRRRRRR



Choose the State : **L** Probability= 0.01

Chose the Symbol: 5 Probability= 1/10 (given **L**)

415626665632163654366215261**5**

RRRRRLRLRRRRRRRLRRRRRRRRRRR**L**

Some applications

1) DEMOGRAPHY

Observable: Number of births and deaths in a year in a village.

Hidden variable: Economic conditions (as a first approximation we can consider the success in business as a random variable, and by consequence, the wealth as a Markov variable)

—> can we deduce the economic conditions of a village during a century by means of the register of births and deaths?

2) THE METEOROPATHIC TEACHER

Observable: Average of the marks that a meteoropathic teacher gives to their students during a day.

Hidden variable: Weather conditions

—> can we deduce the weather conditions during a years by means of the class register?

In Bioinformatics

1) SECONDARY STRUCTURE

Observable: protein sequence

Hidden variable: secondary structure

---> can we deduce (predict) the secondary structure of a protein given its amino acid sequence?

2) ALIGNMENT

Observable: protein sequence

Hidden variable: position of each residue along the alignment of a protein family

---> can we align a protein to a family, starting from its amino acid sequence?

Exercise

Given the observation sequence:

CCCFRCRRRCCSSSSFSFRRFFSSFCRRRS

and the model on the left

Write a script to set the parameters of the Markov Model that maximize the probability of the sequence

