Probabilistic Models for Biological Sequences

Laboratory of Bioinformatics I
Module 2

27 March, 2020

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Department of Pharmacy and Biotechnology (FaBiT)
University of Bologna

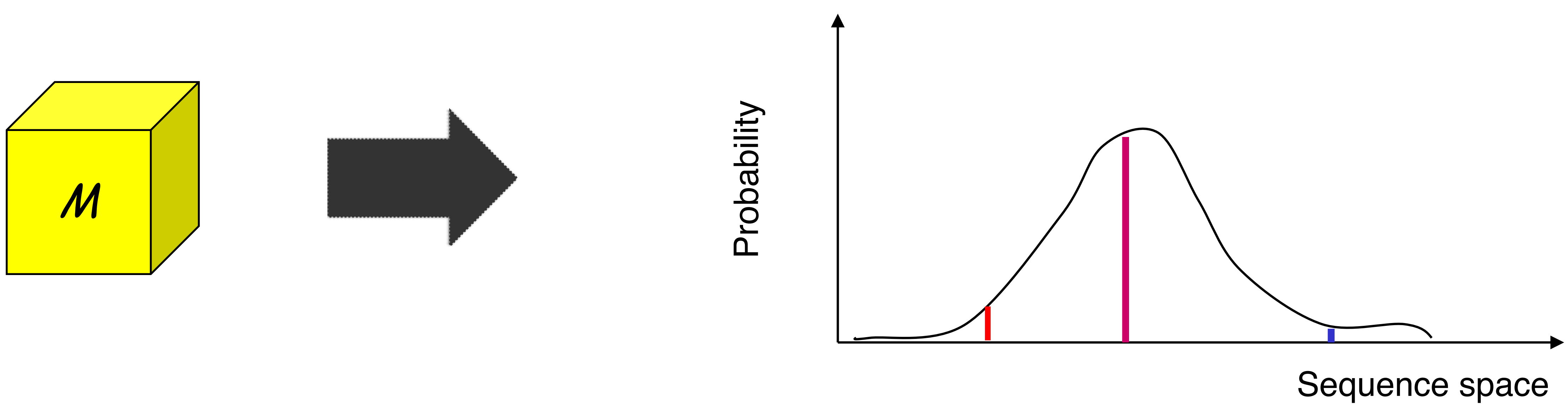
Models for Sequence

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

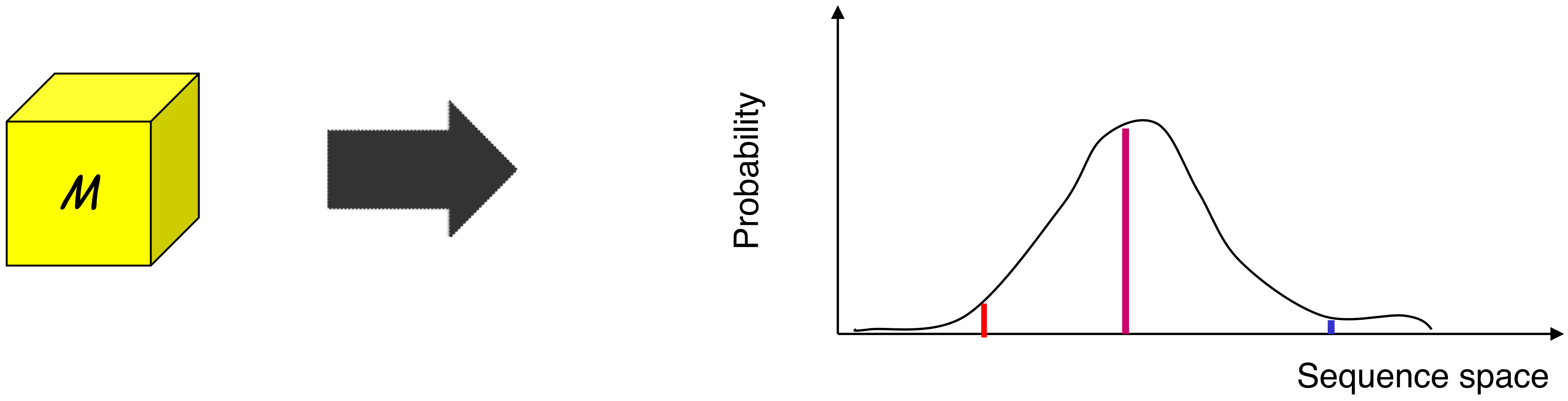


Associative Definition

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 l M)* to s_i .
e.g.: *M* is the representation of the family of globins

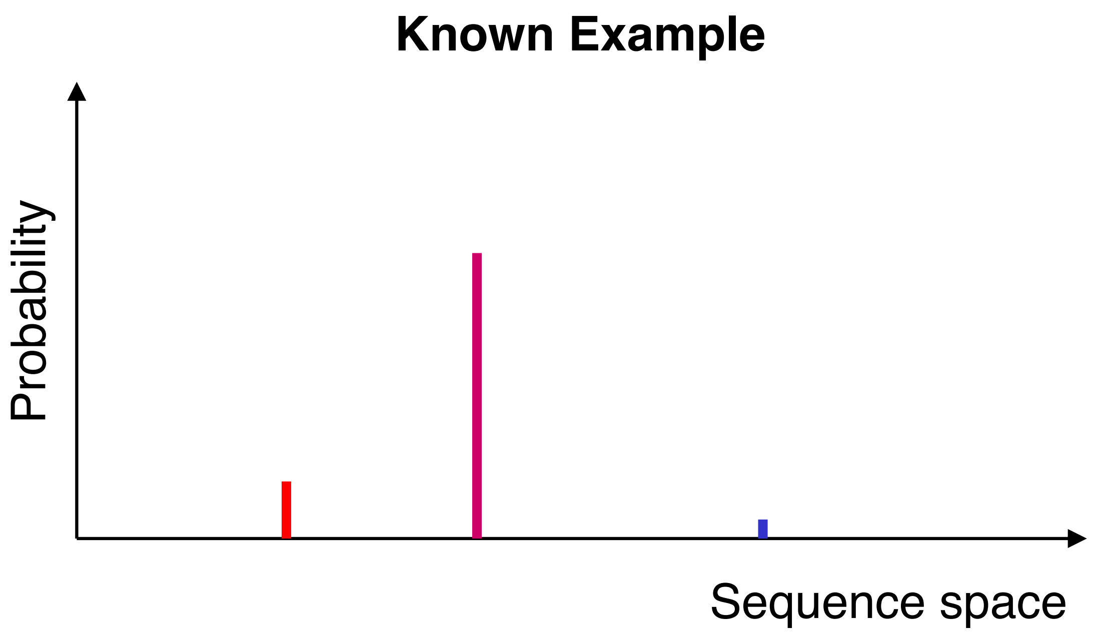


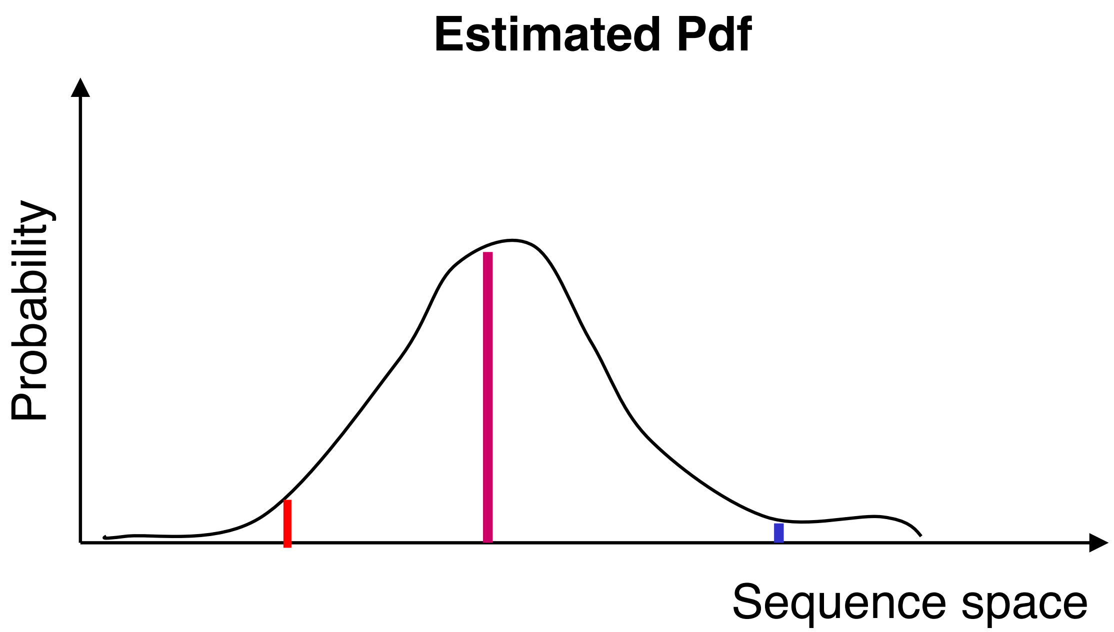
Which Model?

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

$P(X,Y) = P(X \mid Y)\, P(Y) = P(Y \mid X)\, P(X)$ Joint probability

$$P(Y \mid X) = \frac{P(X \mid Y)\, P(Y)}{P(X)}$$

$$P(M \mid s_i) = \frac{P(s_i \mid M)\, P(M)}{P(s_i)}$$

$P(M)$ 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

$$\frac{P(M_1 \mid s_i)}{P(M_2 \mid s_i)} = \frac{P(s_i \mid M_1)\, P(M_1)}{P(s_i)} \quad \frac{P(s_i)}{P(s_i \mid M_2)\, P(M_2)} = \frac{P(s_i \mid M_1)\, P(M_1)}{P(s_i \mid M_2)\, P(M_2)}$$

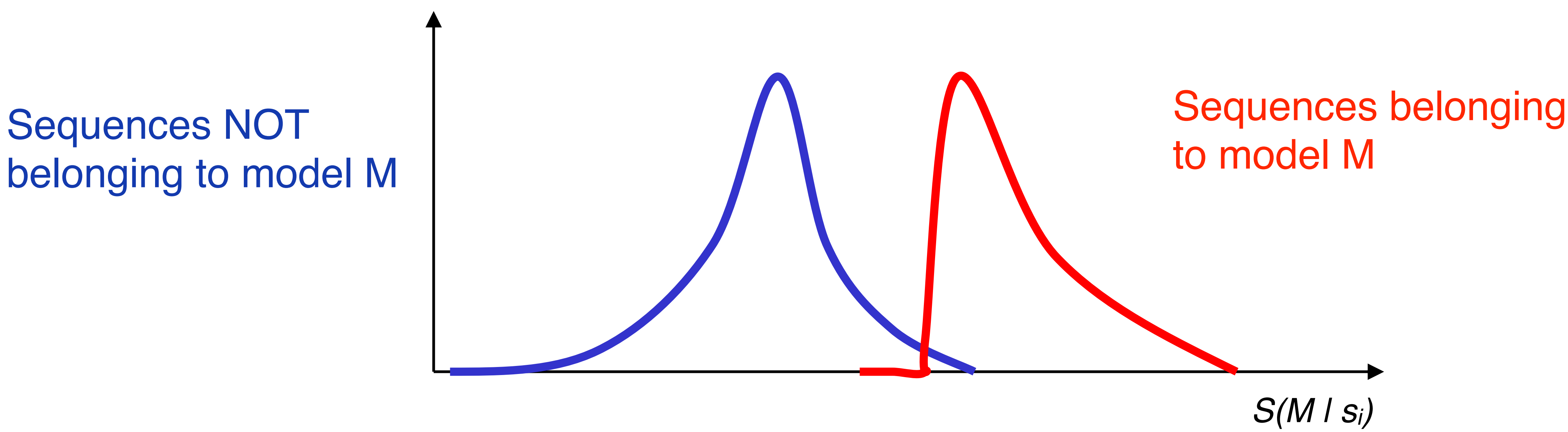
$$\frac{P(M_1)}{P(M_2)}$$ Ratio between the abundances of the classes

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

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

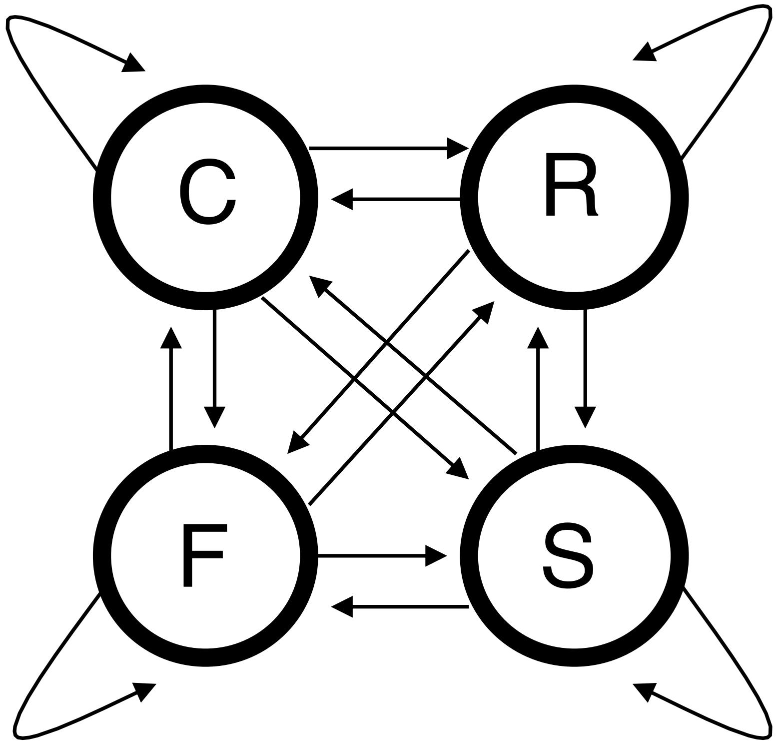


In this case we need a threshold and a statistic for evaluating the significance (E-value, P-value)

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

$P(C|C), P(C|R),\ldots. P(R|C)\ldots..$

The probability for the 5-days registration CRRCS

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

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)

$C = \{c_1;\ c_2;\ c_3;\ \ldots\ldots\ldots c_N\}$

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

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

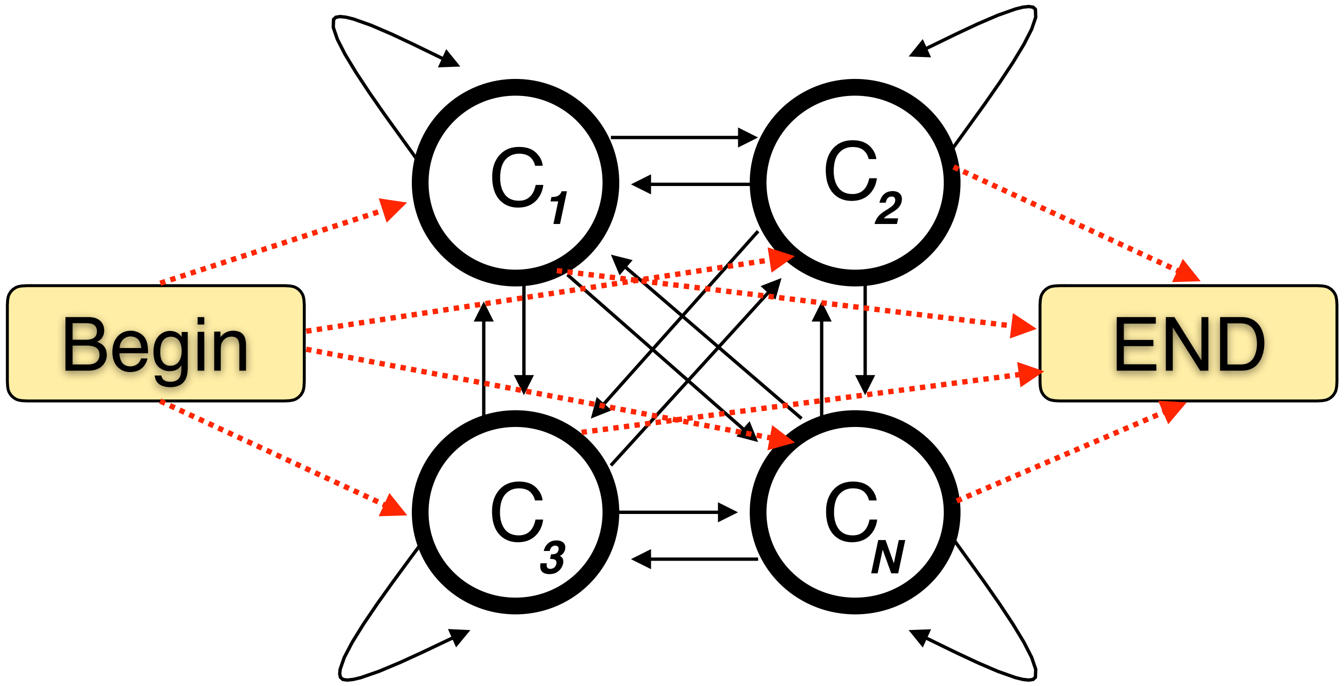
$a_{r,q} = P(\ s_i = q \mid s_{i-1} = r\ )$

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

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

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

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



Sequence Probability

Given the sequence:

$s = s_1\, s_2\, s_3\, s_4\, s_6 \ldots\ldots\ldots\ldots\ldots s_T$ with $s_i \in \mathrm{C} = \{c_1; c_2; c_3; \ldots\ldots\ldots c_N\}$

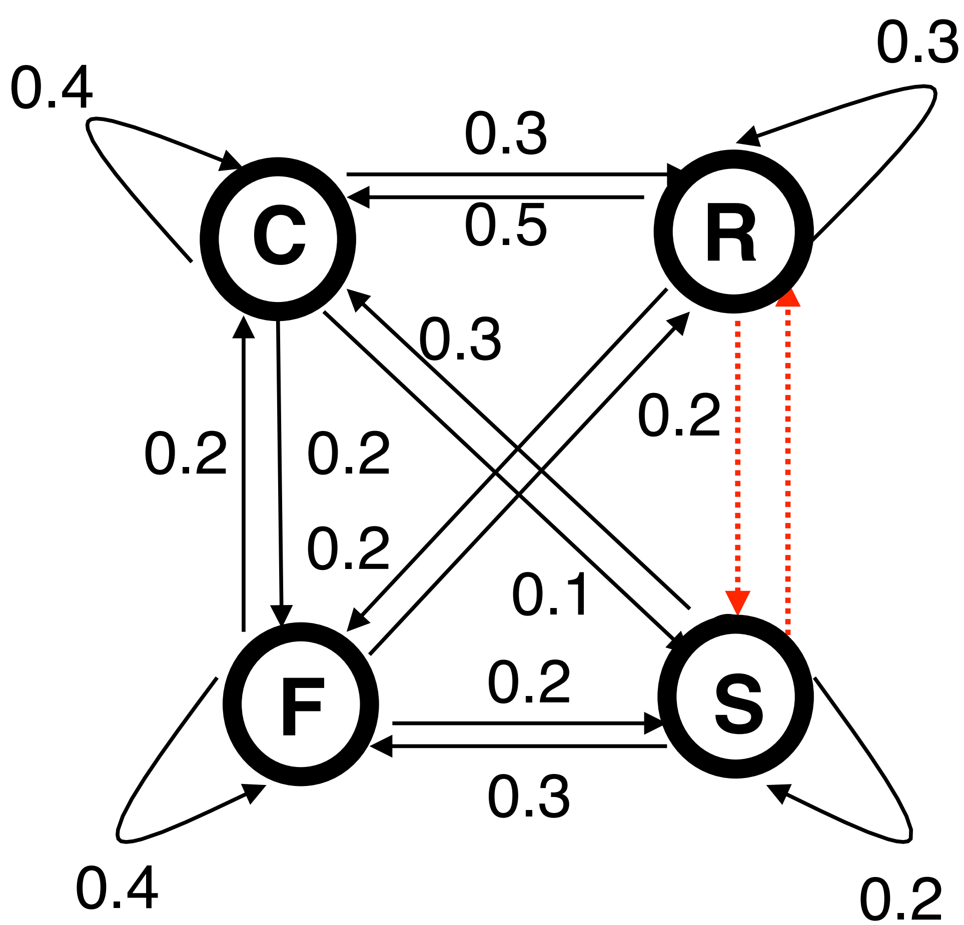
$$P(s \mid M) = P(s_1) \prod_{i=2}^{T} P(s_i \mid s_{i-1}) =$$

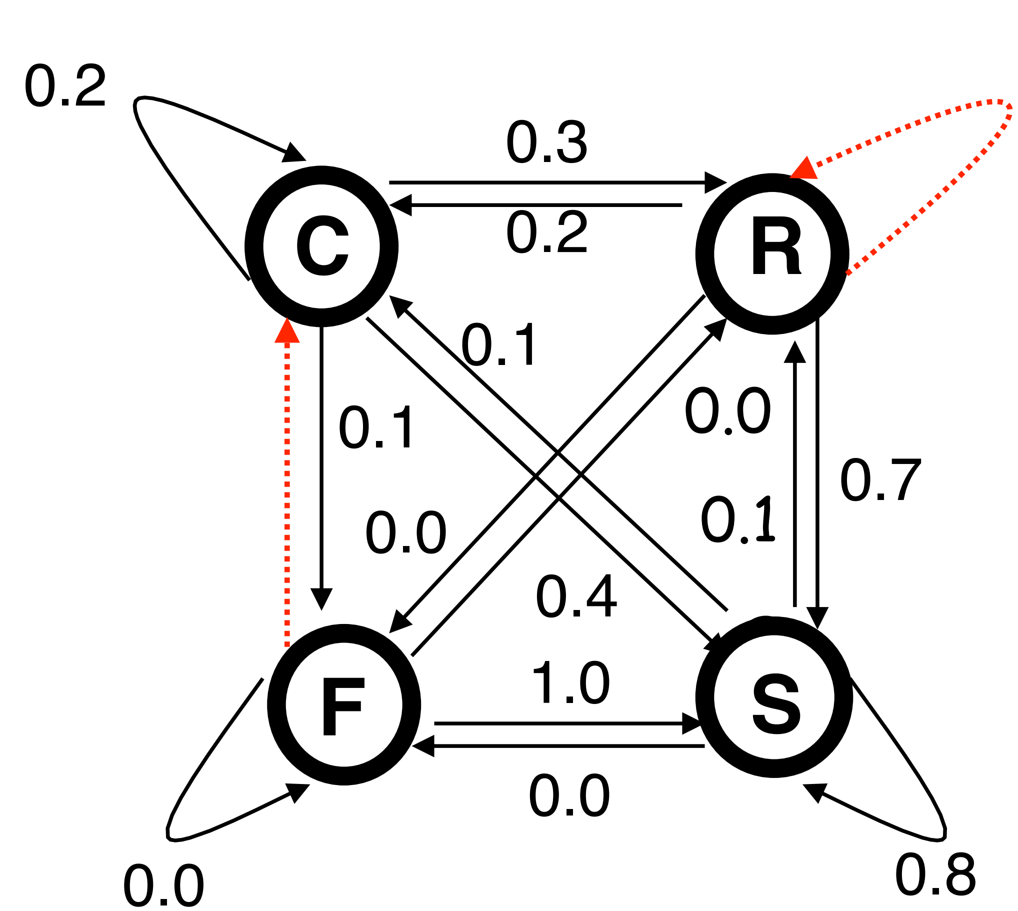
$$a_{BEGIN, s_1} \times \prod_{\mathrm{i}=2}^{T} a_{s_{i-1}, s_i} \times a_{s_T, END}$$

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

Probability Constrains

What are the missing probabilities given the constrains?

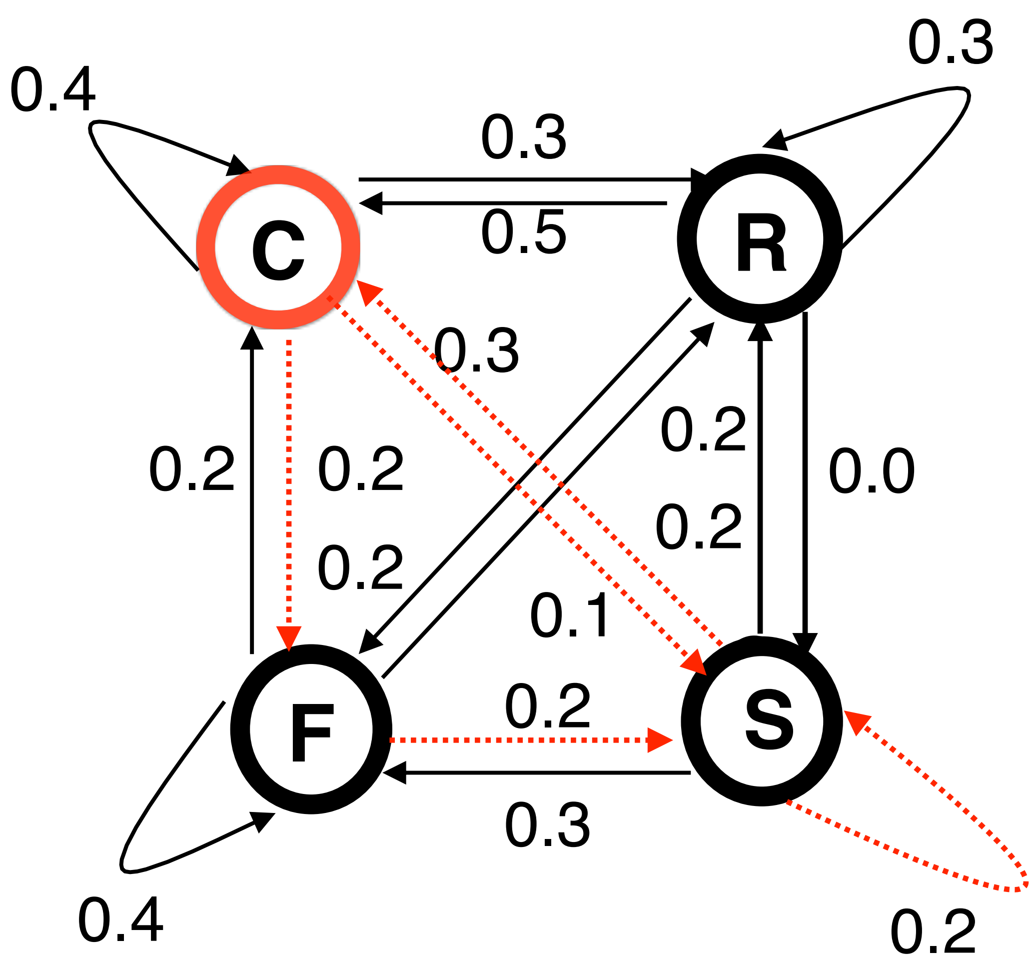




What is the better model to describe the weather in winter?

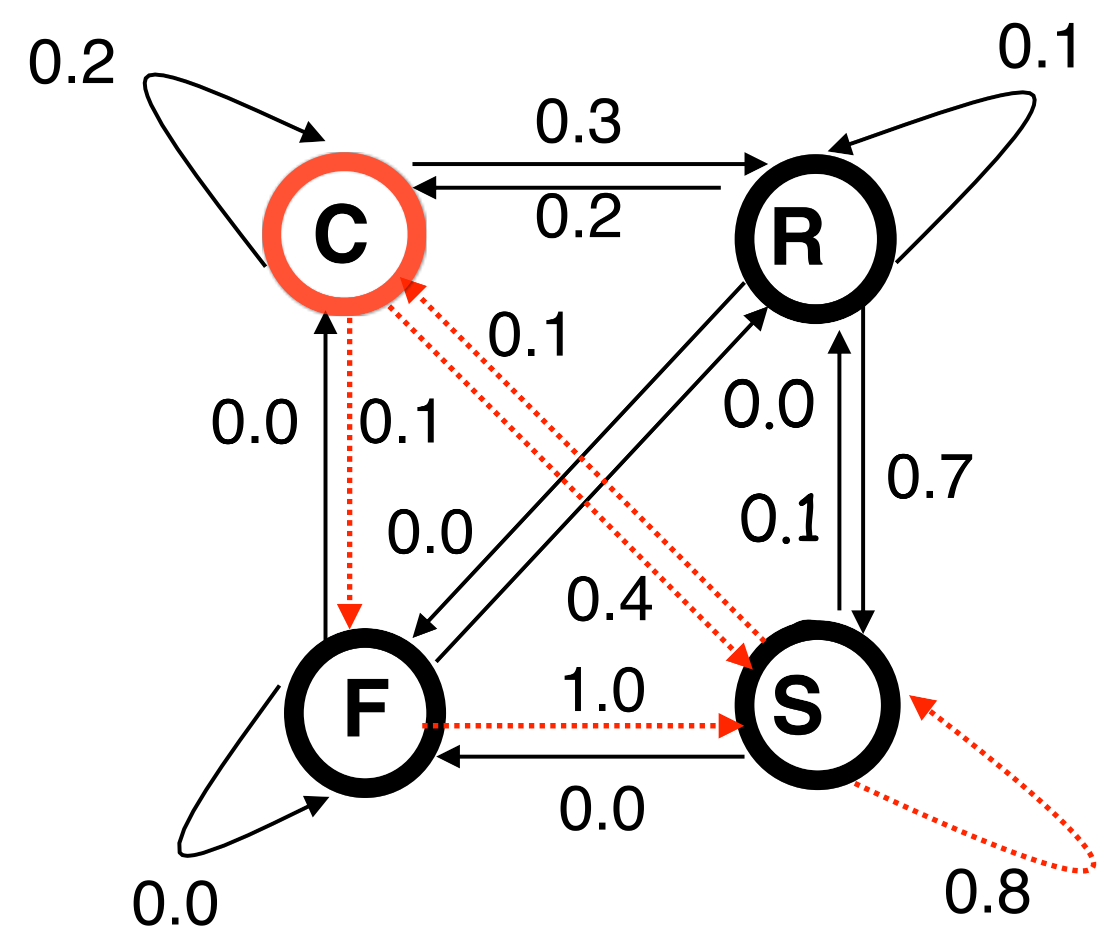
Probability Calculation

Consider the sequence “CSSSCFS” and calculate its probability with both models when *P(X | BEGIN)* = 0.25



P(CSSSCFS | Winter) =

0.25×0.1×0.2×0.2×0.3×0.2×0.2



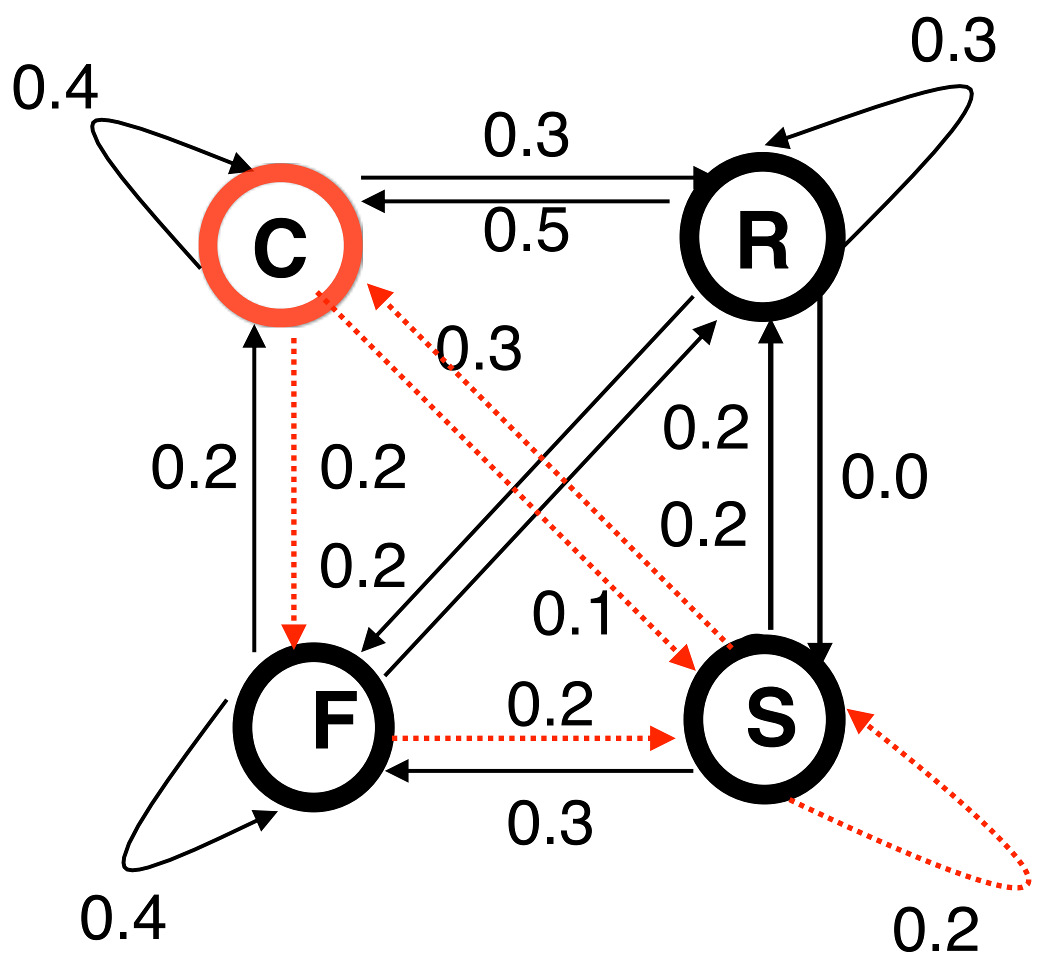
P(CSSSCFS | Summer) =

0.25×0.4×0.8×0.8×0.1×0.1×1.0

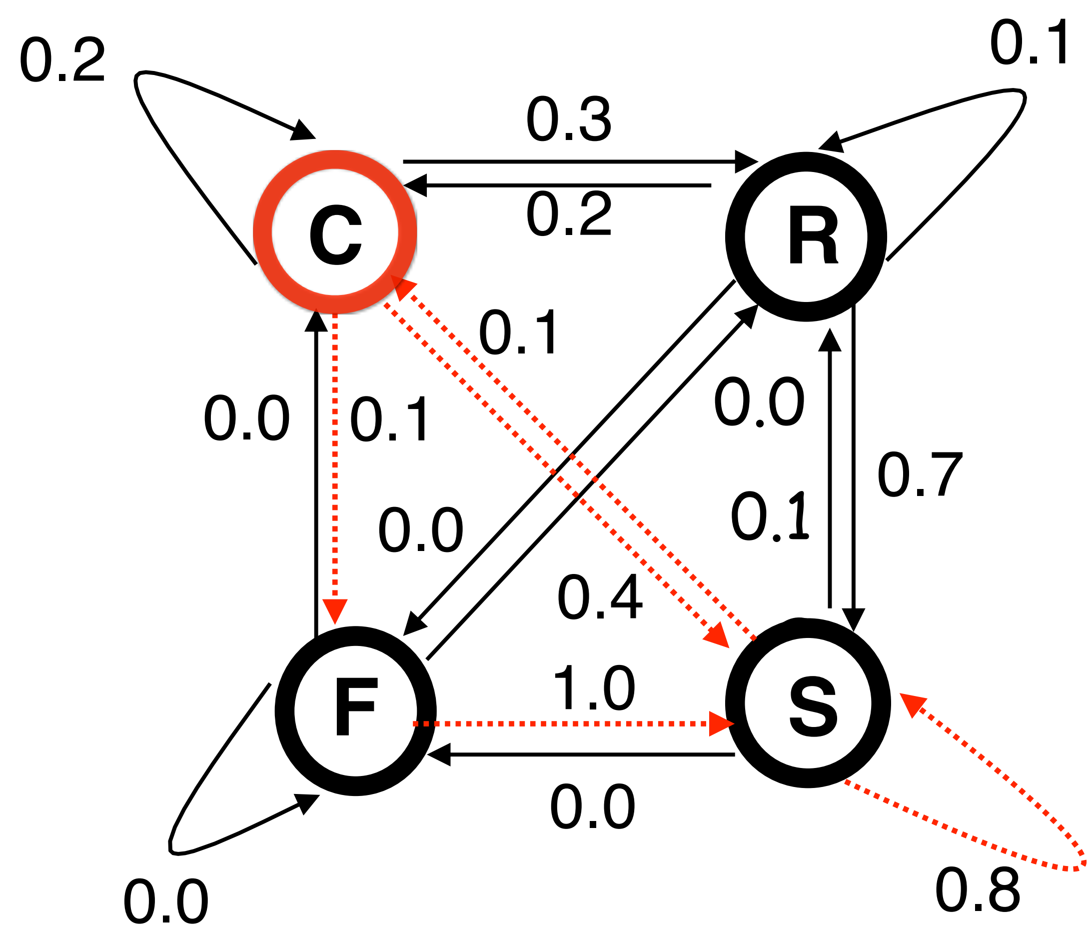
To which season the weather sequence is more likely to belong?

Probability Calculation

P (Seq | Winter) =1.2 x 10^{-5}



P (Seq | Summer) = 6.4 x 10^{-4}



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