Session 4 (Theory)
Introduction to Hidden Markov Models

ESCI upf. School of International Studies

BelongBecome

Date: 29/01/2024, 15:00-17:00

Teacher: **Fernando Cruz** (CNAG)

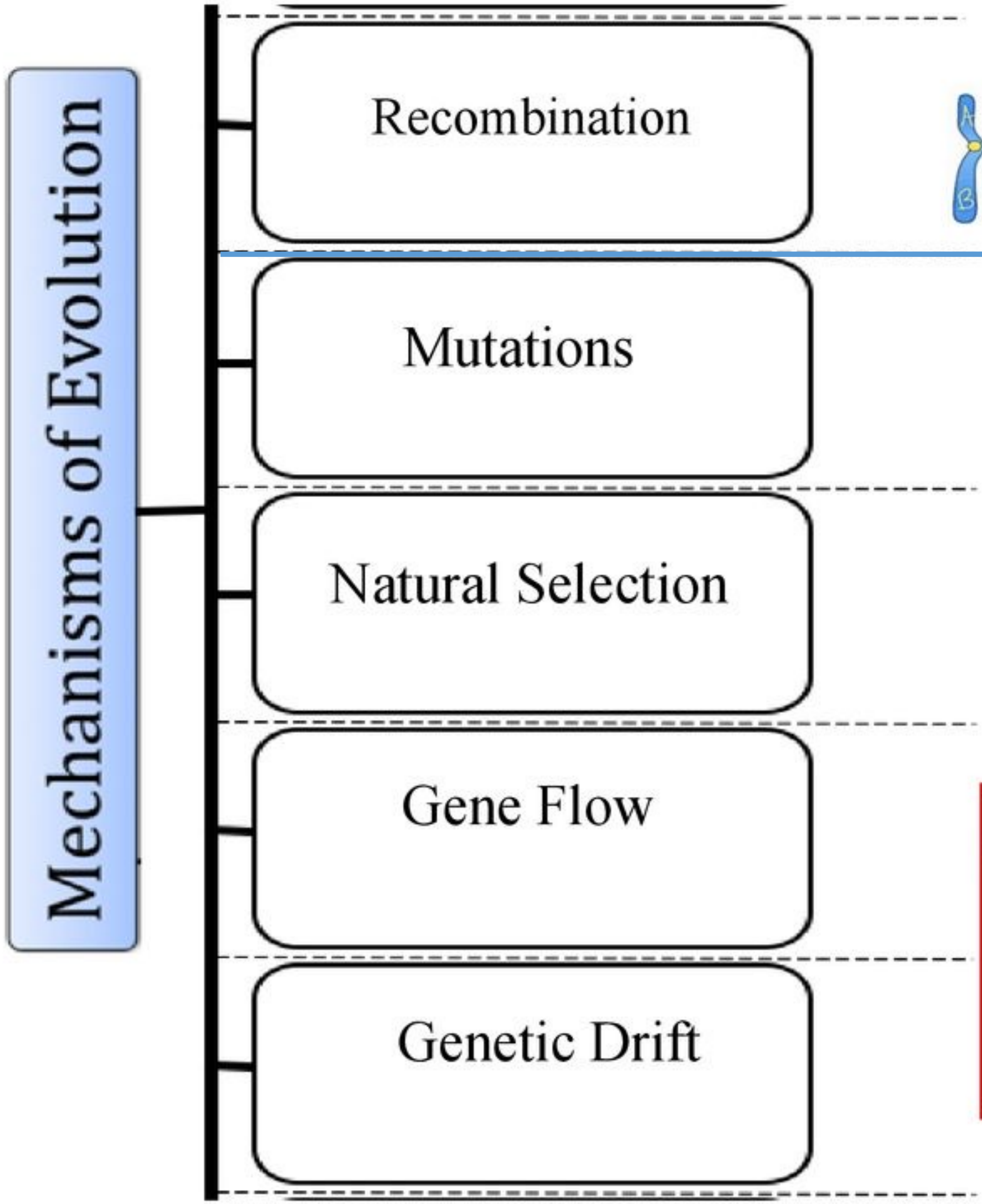
fernando.cruz@prof.esci.upf.edu

Bachelor's Degree in Bioinformatics

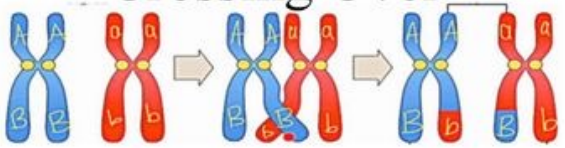
Course 2021-2022

52115 - Algorithms for sequence analysis in Bioinformatics (**ASAB**)

Evolutionary Forces Act at Population Level

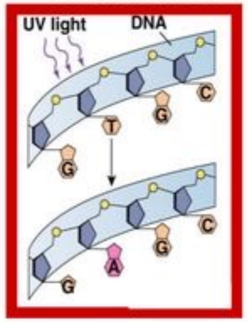


Crossing Over



Exchange of genetic material between chromosomes (change combinations)

Mutation



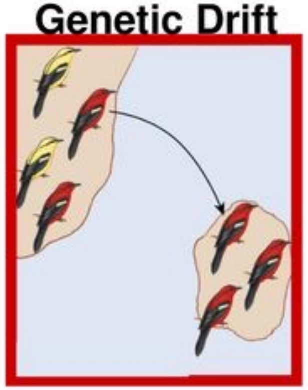
Substitution of bases generates variation

Selection

Variants selected based on their “fitness”

Gene Flow

Genetic Drift



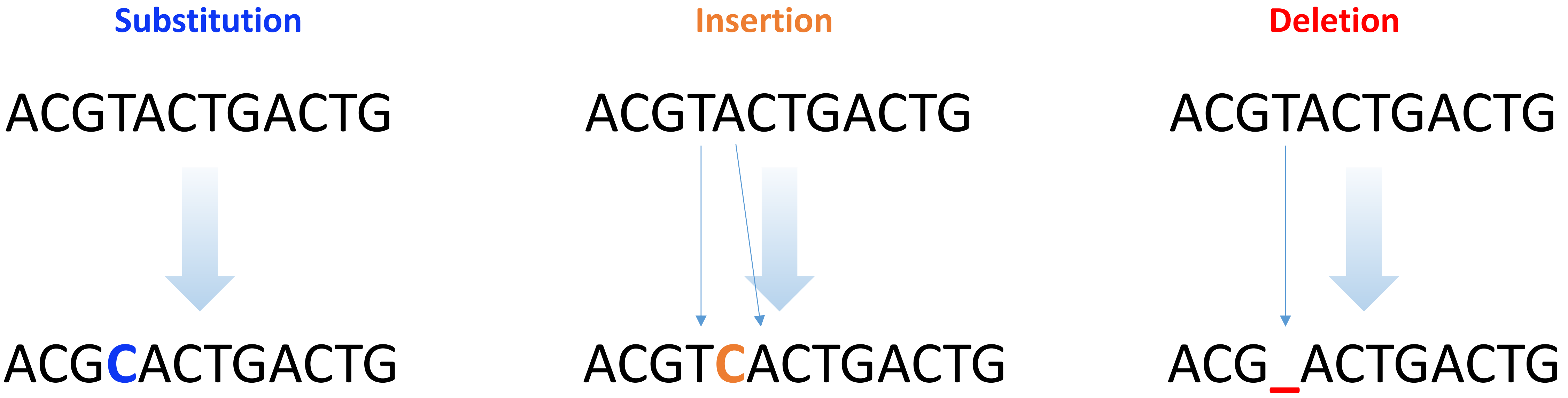
Frequency of variants also varies due to random processes (gamete generation and migration)

Point Mutations

Small scale mutations affecting to a single nucleotide.

Most frequent mutations

If they consist on a replacement they are known as *substitutions*.



Substitution

ACGTACTGACTG

ACGCACTGACTG

Insertion

ACGTACTGACTG

ACGTCACTGACTG

Deletion

ACGTACTGACTG

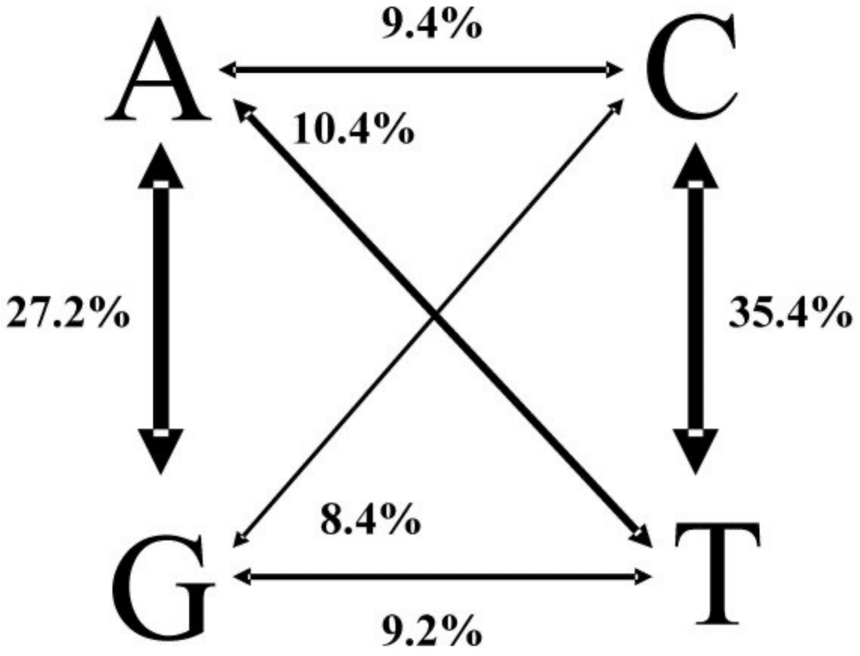
ACG_ACTGACTG

REMIND

All of **these factors** will determine:

- **Substitution rates**
- **Bioinformatic models** and values to build **subtitution matrices**

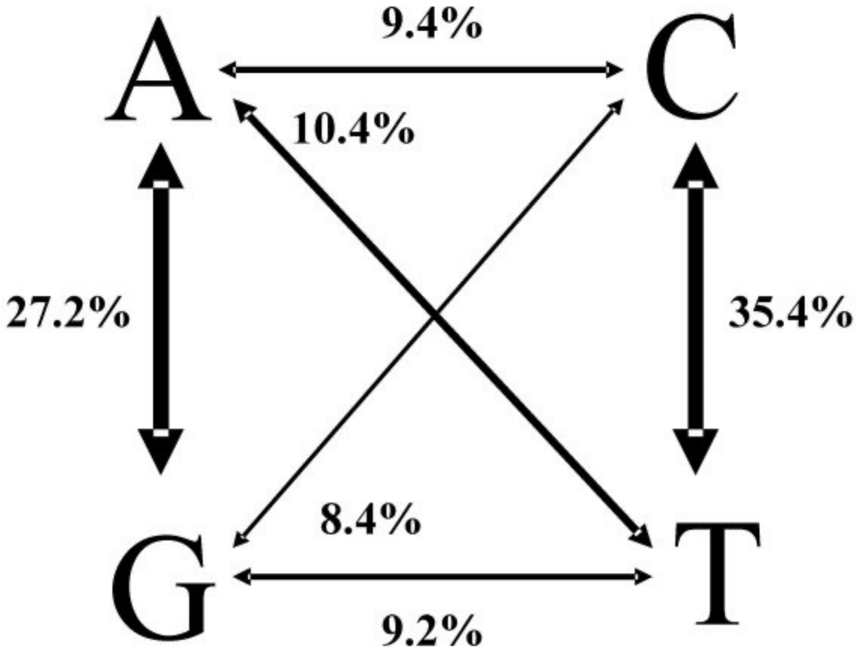
Nucleotide Substitution Matrix



	A	C	T	G
A	-	0.094	0.104	0.272
C	0.094	-	0.354	0.084
T	0.104	0.354	-	0.092
G	0.272	0.084	0.092	-

Identification and analysis of Single Nucleotide Polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector (Wondji et al. 2007)
https://doi.org/10.1186/1471-2164-8-5

Nucleotide Substitution Matrix



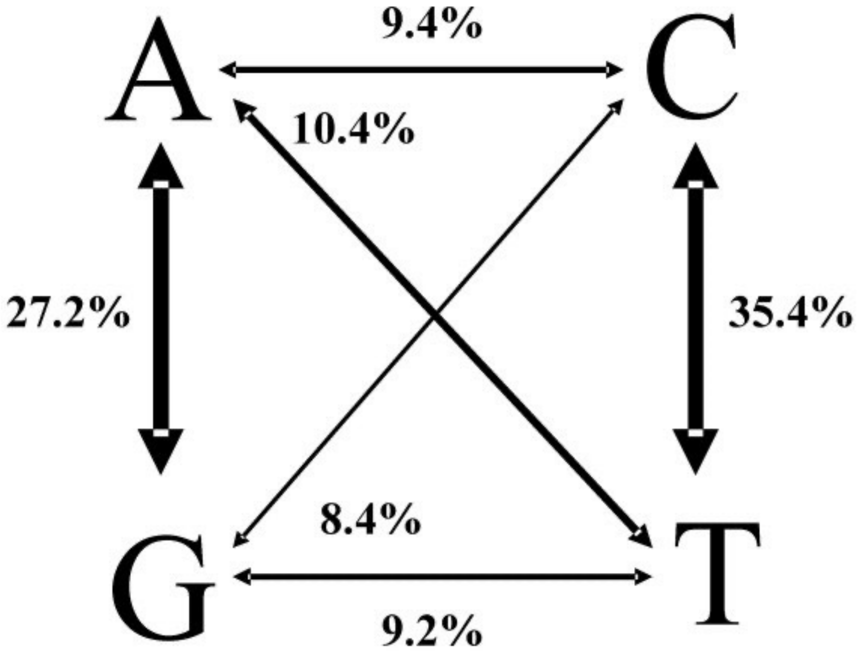
Identification and analysis of Single Nucleotide Polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector (Wondji et al. 2007)
https://doi.org/10.1186/1471-2164-8-5

	A	C	T	G
A	?	0.094	0.104	0.272
C	0.094	?	0.354	0.084
T	0.104	0.354	?	0.092
G	0.272	0.084	0.092	?

Let's assume that the probability of each row totals 1:

P(A->A) = 1 - (0.094+0.104+0.272) = **0.53**

Nucleotide Substitution Matrix



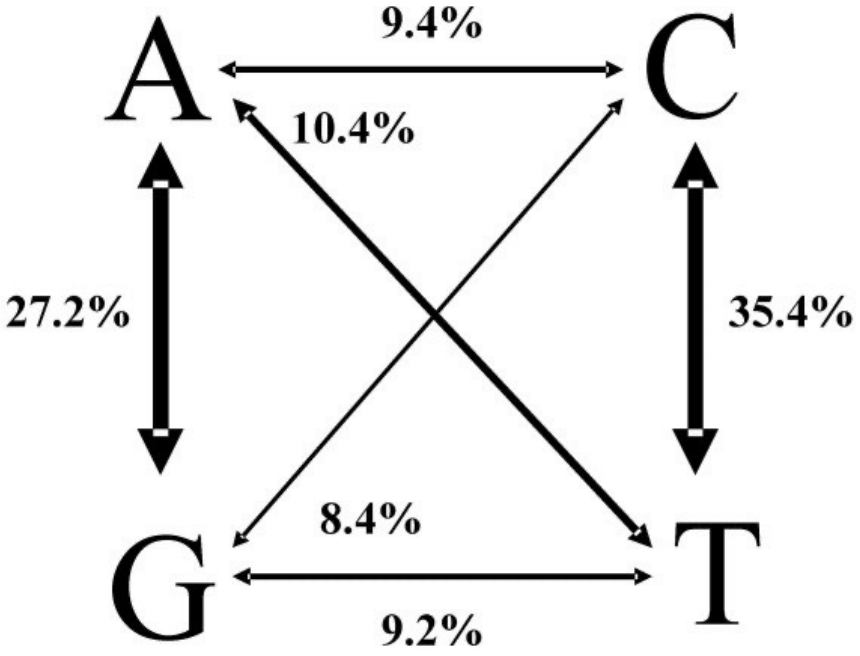
	A	C	T	G
A	0.53	0.094	0.104	0.272
C	0.094	-	0.354	0.084
T	0.104	0.354	-	0.092
G	0.272	0.084	0.092	-

Identification and analysis of Single Nucleotide Polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector (Wondji et al. 2007)
https://doi.org/10.1186/1471-2164-8-5

Let's assume that the probability of each row totals 1:

P(A->A) = 1 - (0.094+0.104+0.272) = **0.53**

Nucleotide Substitution Matrix

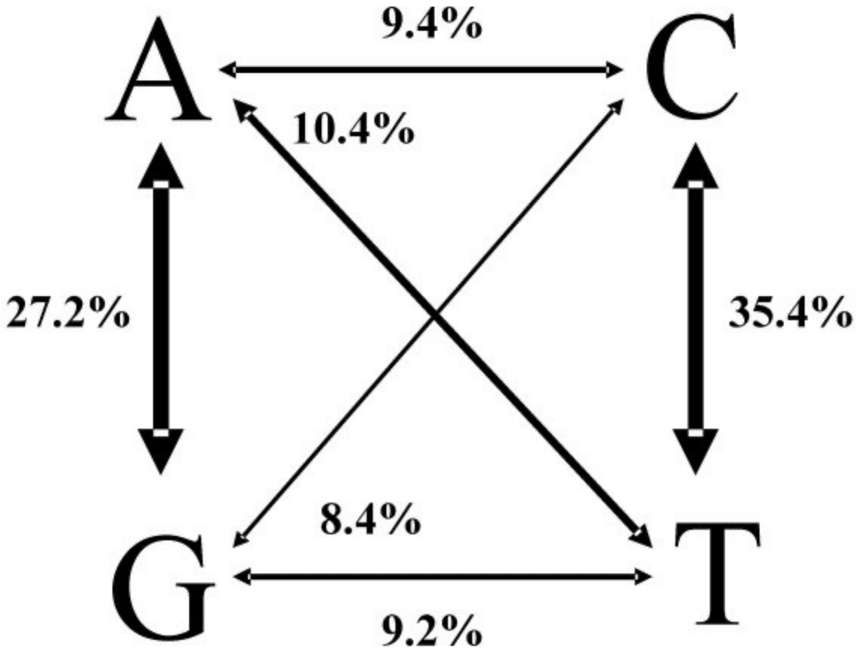


Identification and analysis of Single Nucleotide Polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector (Wondji et al. 2007)
https://doi.org/10.1186/1471-2164-8-5

	A	C	T	G
A	**0.53**	0.09	0.10	0.27
C	0.09	**0.47**	0.35	0.08
T	0.10	0.35	**0.45**	0.09
G	0.27	0.08	0.09	**0.55**

Often symmetrical

Nucleotide Substitution Matrix



Identification and analysis of Single Nucleotide Polymorphisms (SNPs) in the mosquito *Anopheles funestus*, malaria vector (Wondji et al. 2007)
https://doi.org/10.1186/1471-2164-8-5

	A	C	T	G
A	**0.53**	-	-	-
C	0.09	**0.47**	-	-
T	0.10	0.35	**0.45**	-
G	0.27	0.08	0.09	**0.55**

Often symmetrical

Particular patterns repeated in the genome

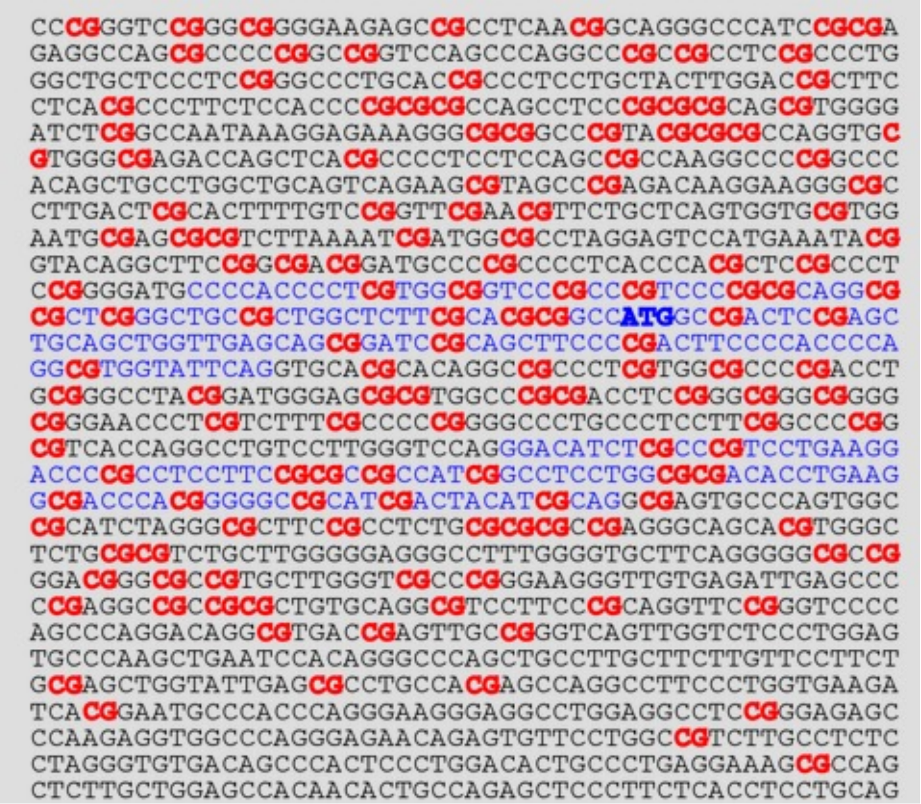
- GC content
- Commonly repeated motifs
- Genes
- CpG islands…

What are *CpG* islands ?

- **Short regions of DNA** in which the **frequency** of the **CG sequence is higher** than in other regions.

- They **often** appear next to **promoters**

- Regulatory relevance **affecting to gene expression**



```
CCCGGGTCCGGGCGGGGAAGAGCCGCCTCAACGGCAGGGCCCATCCGCGA
GAGGCCAGCGCCCCCGGCCGGTCCAGCCCAGGCCCGCCGCCTCCGCCCTG
GGCTGCTCCCTCCGGGCCCTGCACCGCCCTCCTGCTACTTGGACCGCTTC
CTCACGCCCTTCTCCACCCCGCGCGCCAGCCTCCCGCGCGCAGCGTGGGG
ATCTCGGCCAATAAAGGAGAAAGGGCGCGGCCCGTACGCGCGCCAGGTGC
GTGGGCGAGACCAGCTCACGCCCCTCCTCCAGCCGCCAAGGCCCCGGCCC
ACAGCTGCCTGGCTGCAGTCAGAAGCGTAGCCCGAGACAAGGAAGGGCGC
CTTGACTCGCACTTTTGTCCGGTTCGAACGTTCTGCTCAGTGGTGCGTGG
AATGCGAGCGCGTCTTAAAATCGATGGCGCCTAGGAGTCCATGAAATACG
GTACAGGCTTCCGGCGACGGATGCCCCGCCCCTCACCCACGCTCCGCCCT
CCGGGGATGCCCCACCCCTCGTGGCGGTCCCGCCCCGTCCCCGCGCAGGCG
CGCTCGGGCTGCCGCTGGCTCTTCGCACGCGGCCATGGCCGACTCCGAGC
TGCAGCTGGTTGAGCAGCGGATCCGCAGCTTCCCCGACTTCCCCACCCCA
GGCGTGGTATTCAGGTGCACGCACAGGCCGCCCTCGTGGCGCCCCGACCT
GCGGGCCTACGGATGGGAGCGCGTGGCCCGCGACCTCCGGGCGGGCGGGG
CGGGAACCCTCGTCTTTCGCCCCCGGGGCCCTGCCCTCCTTCGGCCCCGG
CGTCACCAGGCCTGTCCTTGGGTCCAGGGACATCTCGCCCCGTCCTGAAGG
ACCCCGCCTCCTTCCGCGCCGCCATCGGCCTCCTGGCGCGACACCTGAAG
GCGACCCACGGGGGCCGCATCGACTACATCGCAGGCGAGTGCCCAGTGGC
CGCATCTAGGGCGCTTCCGCCTCTGCGCGCGCCGAGGGCAGCACGTGGGC
TCTGCGCGTCTGCTTGGGGGAGGGCCTTTGGGGTGCTTCAGGGGGCGCCG
GGACGGGCGCCGTGCTTGGGTCGCCCGGGAAGGGTTGTGAGATTGAGCCC
CCGAGGCCGCCGCGCTGTGCAGGCGTCCTTCCCGCAGGTTCCGGGTCCCC
AGCCCAGGACAGGCGTGACCGAGTTGCCGGGTCAGTTGGTCTCCCTGGAG
TGCCCAAGCTGAATCCACAGGGCCCAGCTGCCTTGCTTCTTGTTCCTTCT
GCGAGCTGGTATTGAGCGCCTGCCACGAGCCAGGCCTTCCCTGGTGAAGA
TCACGGAATGCCCACCCAGGGAAGGGAGGCCTGGAGGCCTCCGGGAGAGC
CCAAGAGGTGGCCCAGGGAGAACAGAGTGTTCCTGGCCGTCTTGCCTCTC
CTAGGGTGTGACAGCCCACTCCCTGGACACTGCCCTGAGGAAAGCGCCAG
CTCTTGCTGGAGCCACAACACTGCCAGAGCTCCCTTCTCACCTCCTGCAG
```

Upstream region of a human gene

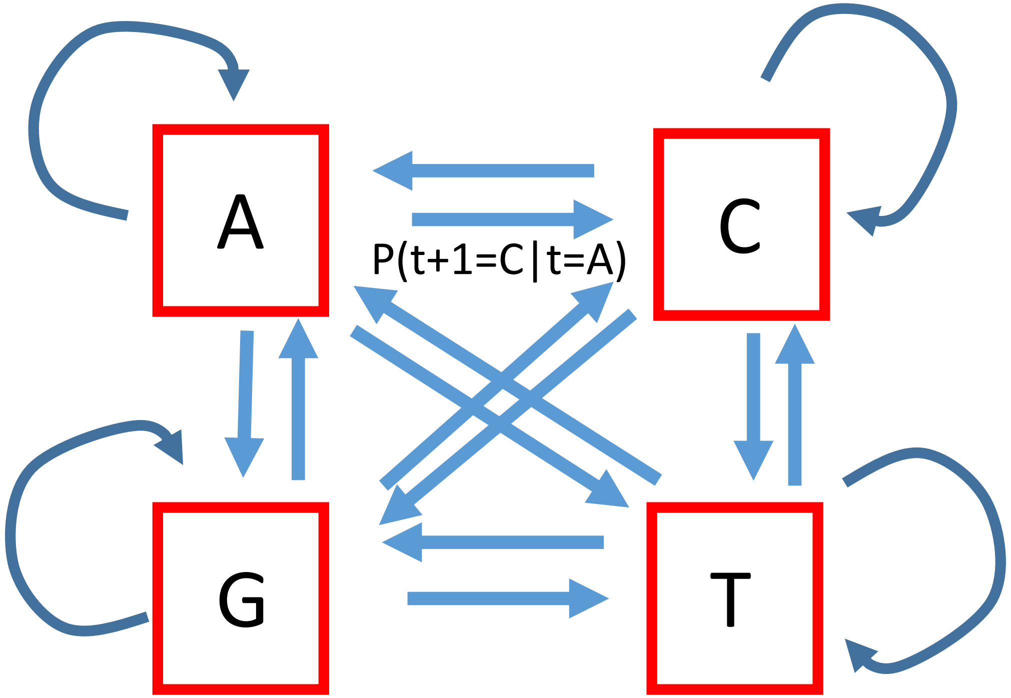
CpG islands in vertebrate genomes. Gardiner-Garden M. and Frommer M.
J Mol Biol . 1987 Jul 20;196(2):261-82. doi: 10.1016/0022-2836(87)90689-9.

How to identify *CpG islands*?

Markov chain
(MC)

t0		t1		t2		t3		t4		t5		t6
A	→	A	→	C	→	A	→	G	→	G	→	T

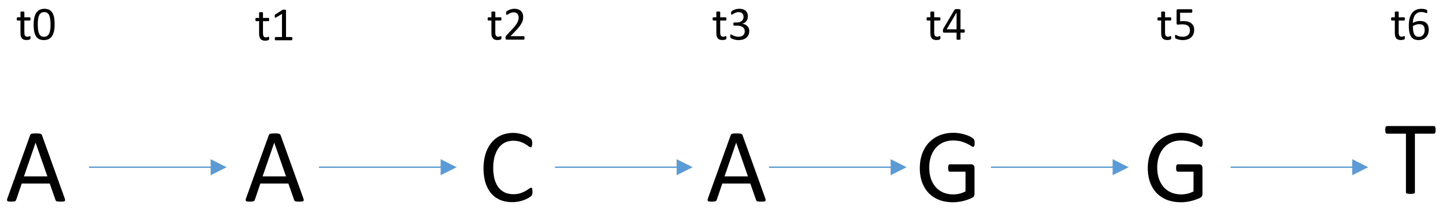
Four possible categories



Probability transition matrix

How to identify *CpG islands*?

Markov chain
(MC)



$$P(AACAGGT) = P(AACAGG)P(T|AACAGG)$$
$$P(AACAGGT) = P(AACAG)P(G|AACAG)P(T|AACAGG)$$

From P(X;Y) = P(Y)P(X|Y)

$$P(AACAGGT) = P(A)P(A|A)P(C|A)P(A|C)P(G|A)P(G|G)P(T|G)$$

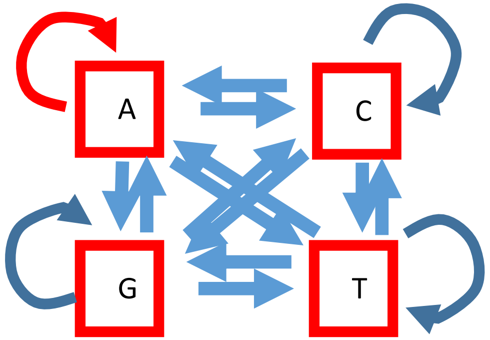
However, in a MC, the **probability at a position depends ONLY on** the **previous state**

How to identify *CpG islands*?

	t0		t1		t2		t3		t4		t5		t6
Markov chain	A	→	A	→	C	→	A	→	G	→	G	→	T

$$P(AACAGGT) = \boxed{P(A)}\boxed{P(A|A)}P(C|A)P(A|C)P(G|A)P(G|G)P(T|G)$$

Prior probability



How to identify *CpG islands*?

1) Compare competing models

AAAGGACCGCCG

Model A

Model B

Sequence comes **from CpG** island

Sequence from **outside a CpG** island

$$\text{L}ikelihoodRatio = \frac{P(Sequence|ModelA)}{P(Sequence|ModelB)}$$

$$\text{LogL}ikelihoodRatio = log\left(\frac{P(Sequence|ModelA)}{P(Sequence|ModelB)}\right)\begin{Bmatrix} L > 0 \ \rightarrow Support\ ModelA \\ L < 0 \rightarrow Support\ Model\ B \end{Bmatrix}$$

How to identify *CpG islands*?

1) Compare competing models

AAAGGACCGCCG

Model A

Model B

Sequence comes **from CpG** island

Sequence from **outside a CpG** island

$$\text{Log}LikelihoodRatio = log\left(\frac{P(AAAGGACCATCA|ModelA)}{P(AAAGGACCATCA|ModelB)}\right)\begin{Bmatrix} L > 0 \ \rightarrow Support\ ModelA \\ L < 0 \rightarrow Support\ Model\ B \end{Bmatrix}$$

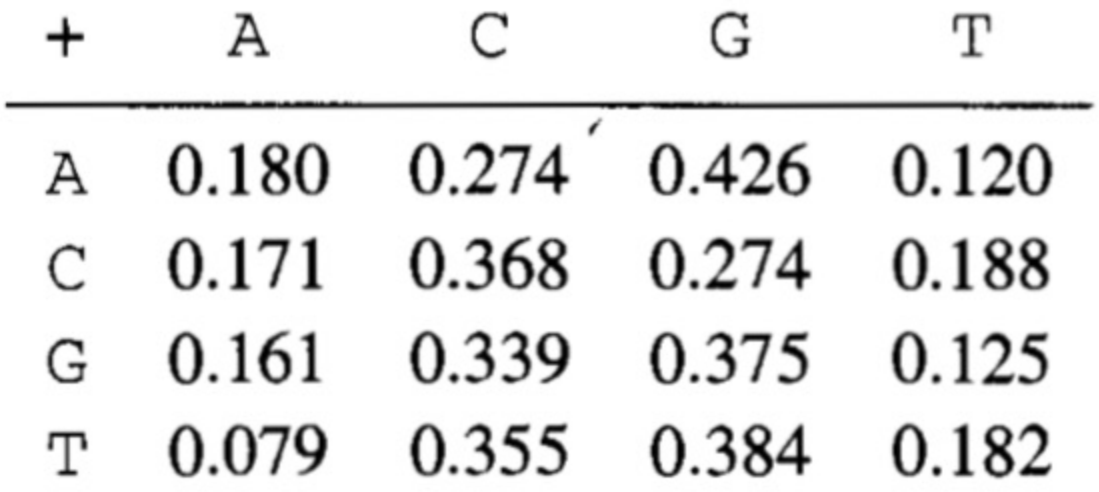
How to identify *CpG islands*?

1) Compare competing models

- Transition probablilty matrices
- Each row sums 1

AAAGGACCGCCG

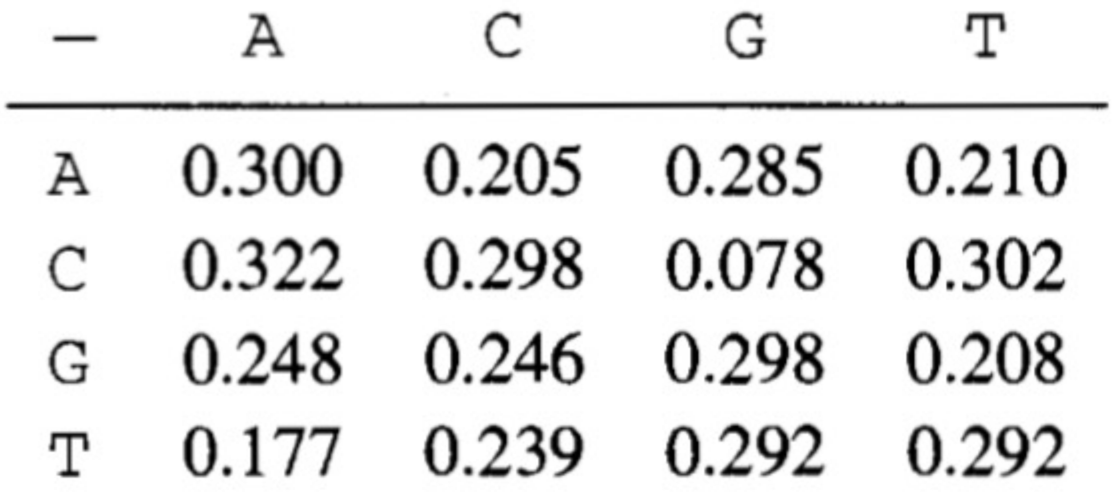
Model A



+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

$P(Sequence|ModelA)$

Model B



−	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

$P(Sequence|ModelB)$

How to identify *CpG islands*?

1) Compare competing models

AAAGGACCGCCG

$\beta =$

	A	C	G	T
A	$\log\left(\frac{P(A \rightarrow A\|ModelA)}{P(A \rightarrow A\|ModelB)}\right)$			
C	$\log\left(\frac{P(C \rightarrow A\|ModelA)}{P(C \rightarrow A\|ModelB)}\right)$			
G				
T				

LogL$ikelihoodRatio = log\left(\frac{P(Sequence|ModelA)}{P(Sequence|ModelB)}\right)$

How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island?

Based on what we have already discussed, how would you find a sequence that corresponds to a ***CpG island***?

How to identify ***CpG islands***?

2) Which part of the sequence corresponds to a CpG island and which not?

ATGATTTCAAAAGGACCGCCGGCCGCG

$$\text{LogL}ikelihoodRatio = log\left(\frac{P(Sequence|ModelA)}{P(Sequence|ModelB)}\right)$$

How to identify ***CpG islands***?

2) Which part of the sequence corresponds to a CpG island and which not?

$$\text{ATGATTTCAA}\boxed{\text{AAGGACCGC}}\text{CGGCCGCG}$$

$$\text{Log}Likelihood Ratio = log\left(\frac{P(Sequence|ModelA)}{P(Sequence|ModelB)}\right)$$

How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?

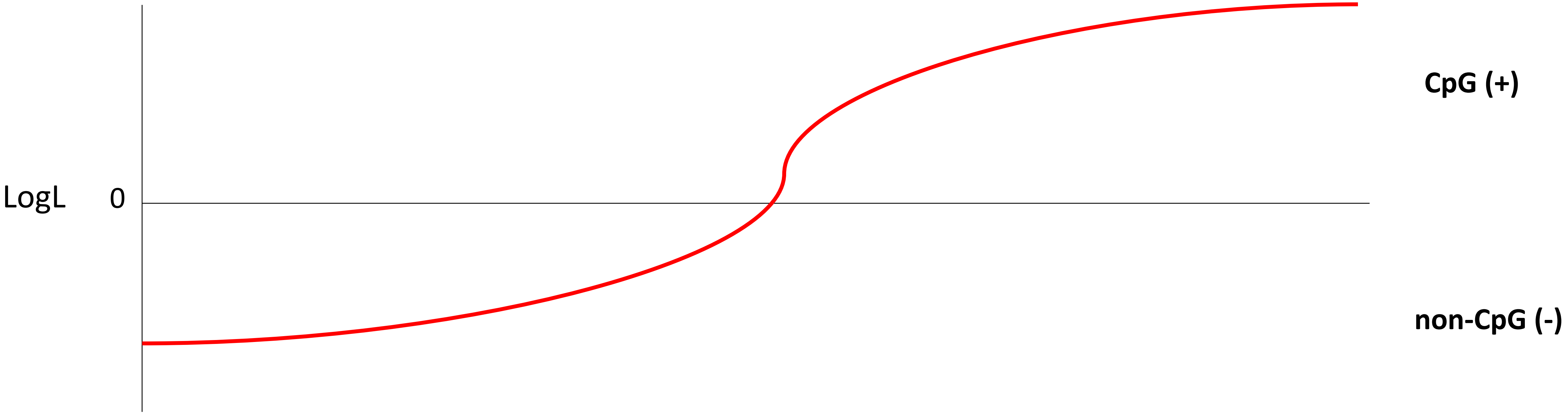
$$\text{ATGATTTCAAAAGGACCGC}\boxed{\text{CGGCCGCG}}$$

$$\text{LogLikelihoodRatio} = log\left(\frac{P(Sequence|ModelA)}{P(Sequence|ModelB)}\right)$$

How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?

ATGATTTCAAAAGGACCGCCGGCCGCG



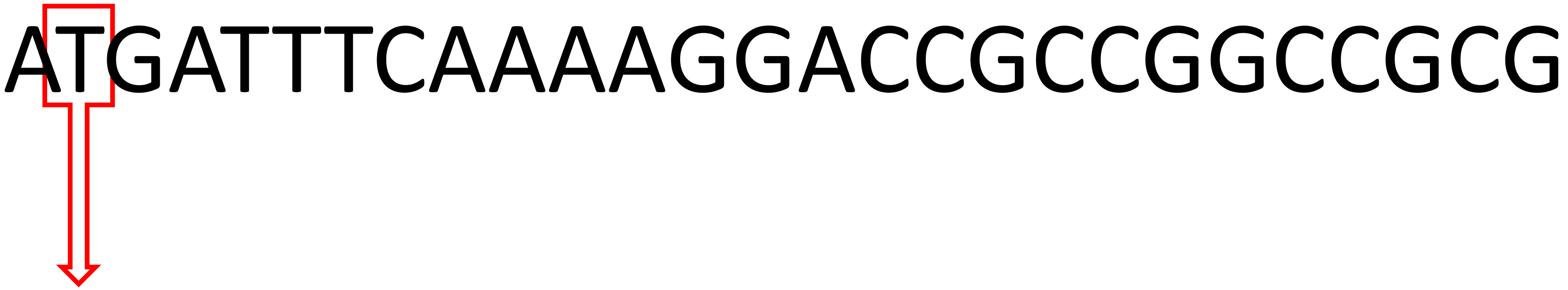
How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?

Problem of this strategy: why this window size? Why not bigger? Or smaller? Where to set the border of the island?

How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?



ATGATTTCAAAAGGACCGCCGGCCGCG

$P(T, GpG|A, GpG)$

$P(T, GpG|A, NoGpG)$

$P(T, NoGpG|A, GpG)$

$P(T, NoGpG|A, NoCpG)$

How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?

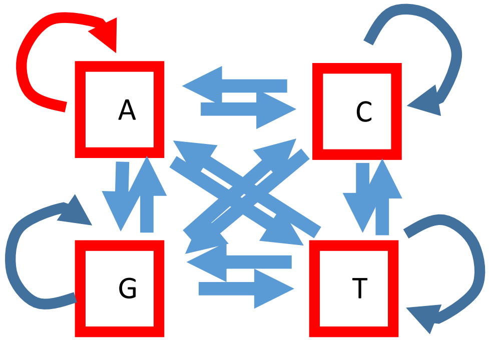
ATGATTTCAAAAGGACCGCCGGCCGCG

$P(T, GpG|A, GpG)$

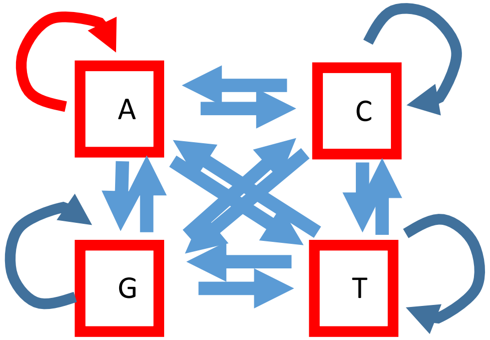
$P(T, GpG|A, NoGpG)$

$P(T, NoGpG|A, GpG)$

$P(T, NoGpG|A, NoCpG)$



CpG (+)



non-CpG (-)

How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?

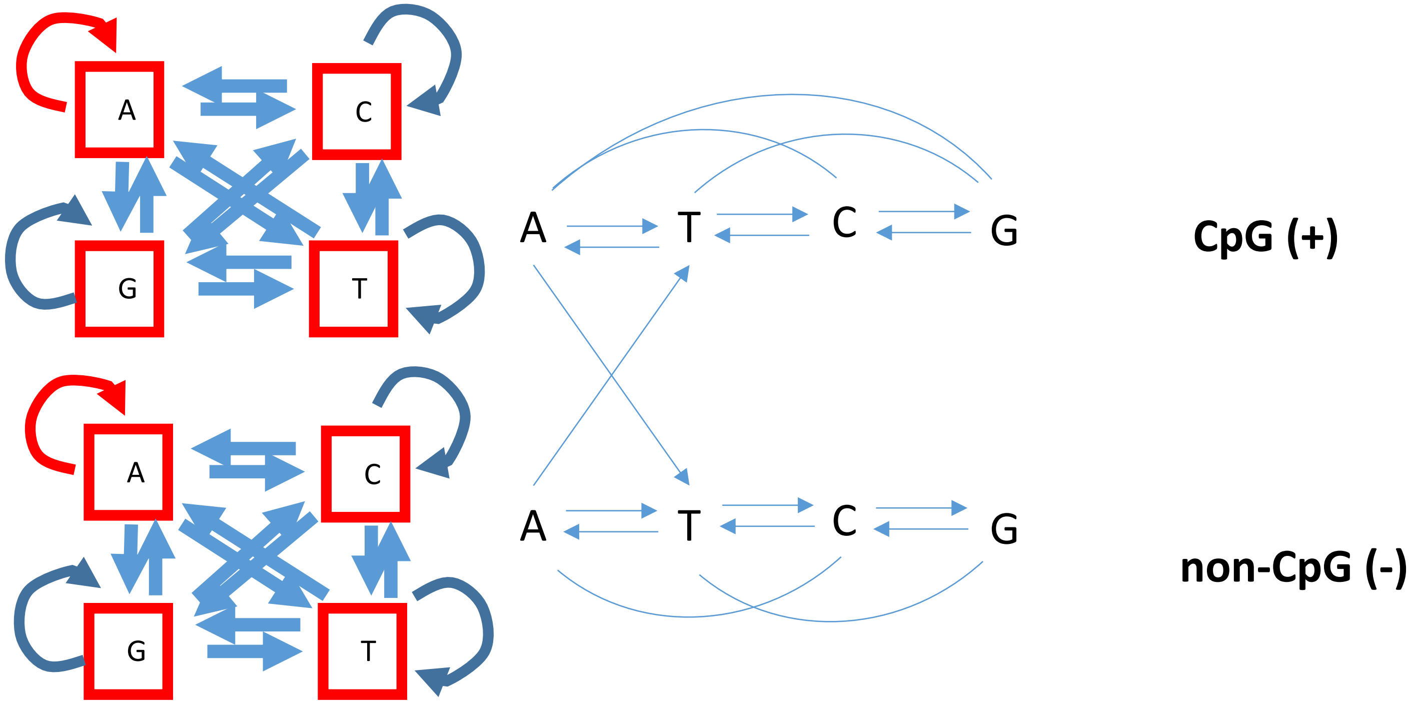
ATGATTTCAAAAGGACCGCCGGCCGCG

$P(T, GpG|A, GpG)$

$P(T, GpG|A, NoGpG)$

$P(T, NoGpG|A, GpG)$

$P(T, NoGpG|A, NoCpG)$



How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?

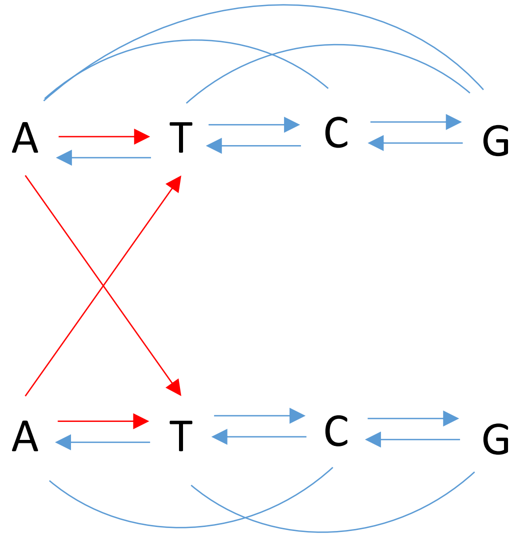
ATGATTTCAAAAGGACCGCCGGCCGCG

$P(T, GpG|A, GpG)$

$P(T, GpG|A, NoGpG)$

$P(T, NoGpG|A, GpG)$

$P(T, NoGpG|A, NoCpG)$

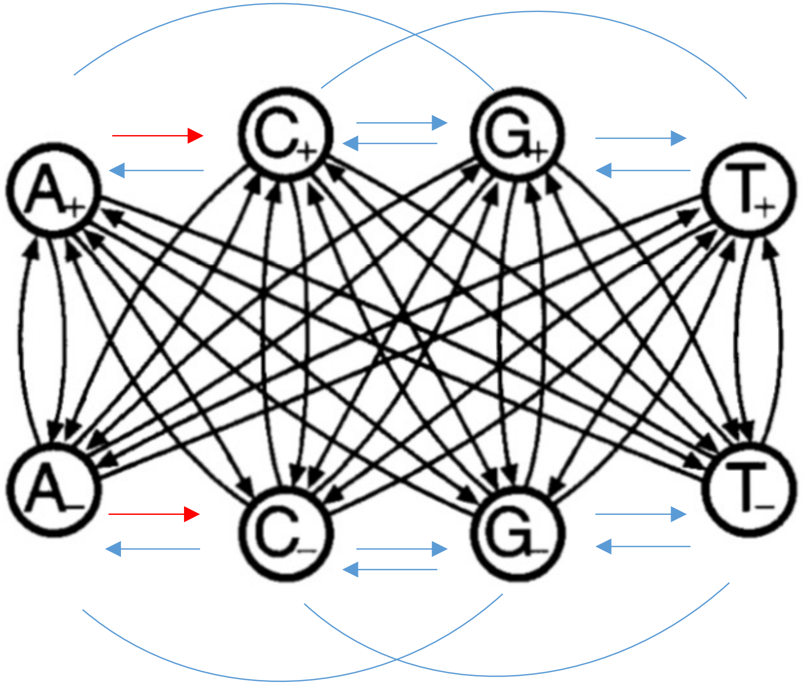


CpG (+)

non-CpG (-)

How to identify *CpG islands*?

Hidden Markov Model in which each nucleotide has **two** possible **states (i.e. CpG vs. non-CpG)**



How to identify *CpG islands*?

2) Which part of the sequence corresponds to a CpG island and which not?

Hidden because the state (CpG/nonCpG) is unknown!

ATGATTTCAAAAGGACCGCCGGCCGCG Observed categories

NNNNNGGGGGGGNNNNNNNGGGG Hidden states

Being in a CpG (G) or out of a CpG (N) is hidden (in fact, this is what we want to estimate!)

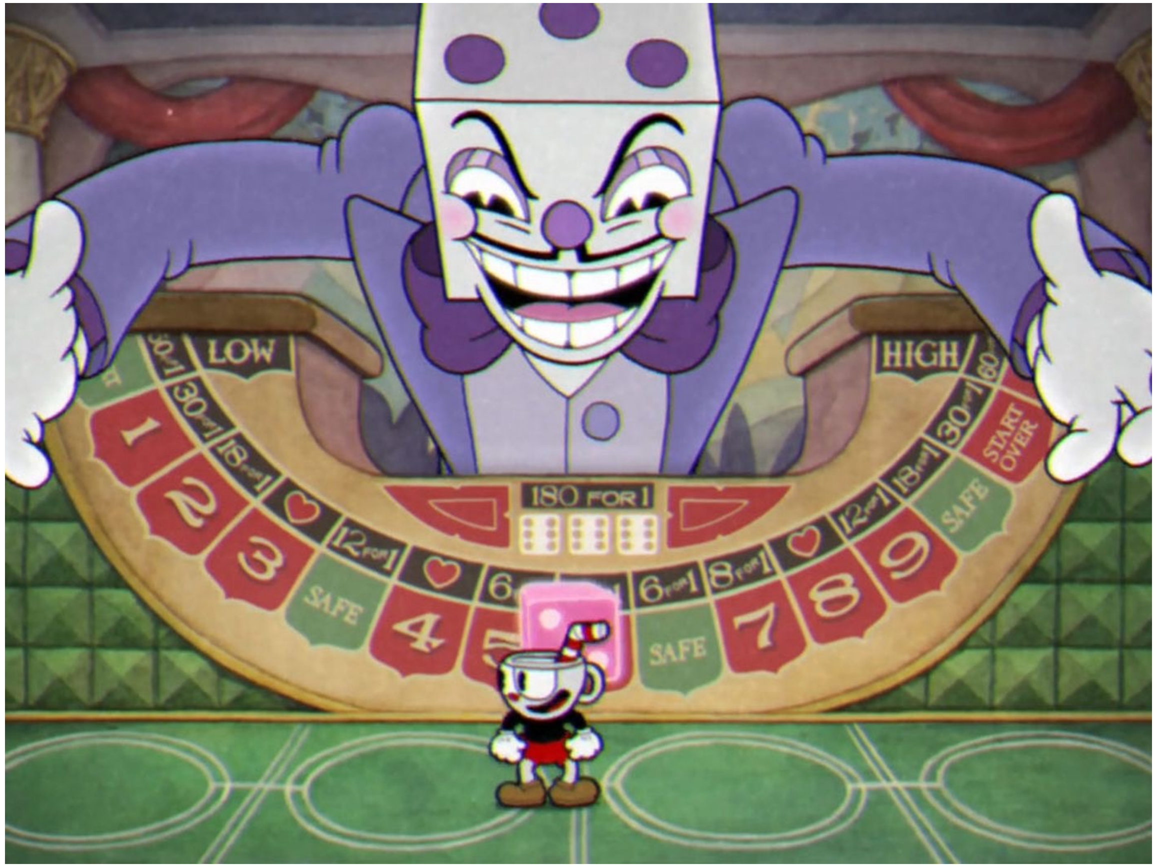
SUMMARY

Markov Chains:

- Can be **applied to biological research** (detect CpGs, genes, etc.)
- The **model** underlying our observations **is unknown (hidden)**
- **The Log Likelihood Ratio Test** to **find the model with higher likelihood**
- Each model has a different **Transition Probability Matrix**

HMM

The occasionally dishonest casino problem



With some probability, the casino uses a dice that is **LOADED** so the number six (the bench wins) occurs more often ($P(6)=0.5$ than expected at random.

HOW CAN WE KNOW WHEN IT IS LOADED OR FAIR?

HMM

The occasionally dishonest casino problem

FAIR State

$P(\boldsymbol{F}|\boldsymbol{F}) = 0.95 \quad P(L|F) = ?$

LOADED State

$P(\boldsymbol{L}|\boldsymbol{L}) = 0.9 \quad P(F|L) = ?$

Categories (Roll Numbers)

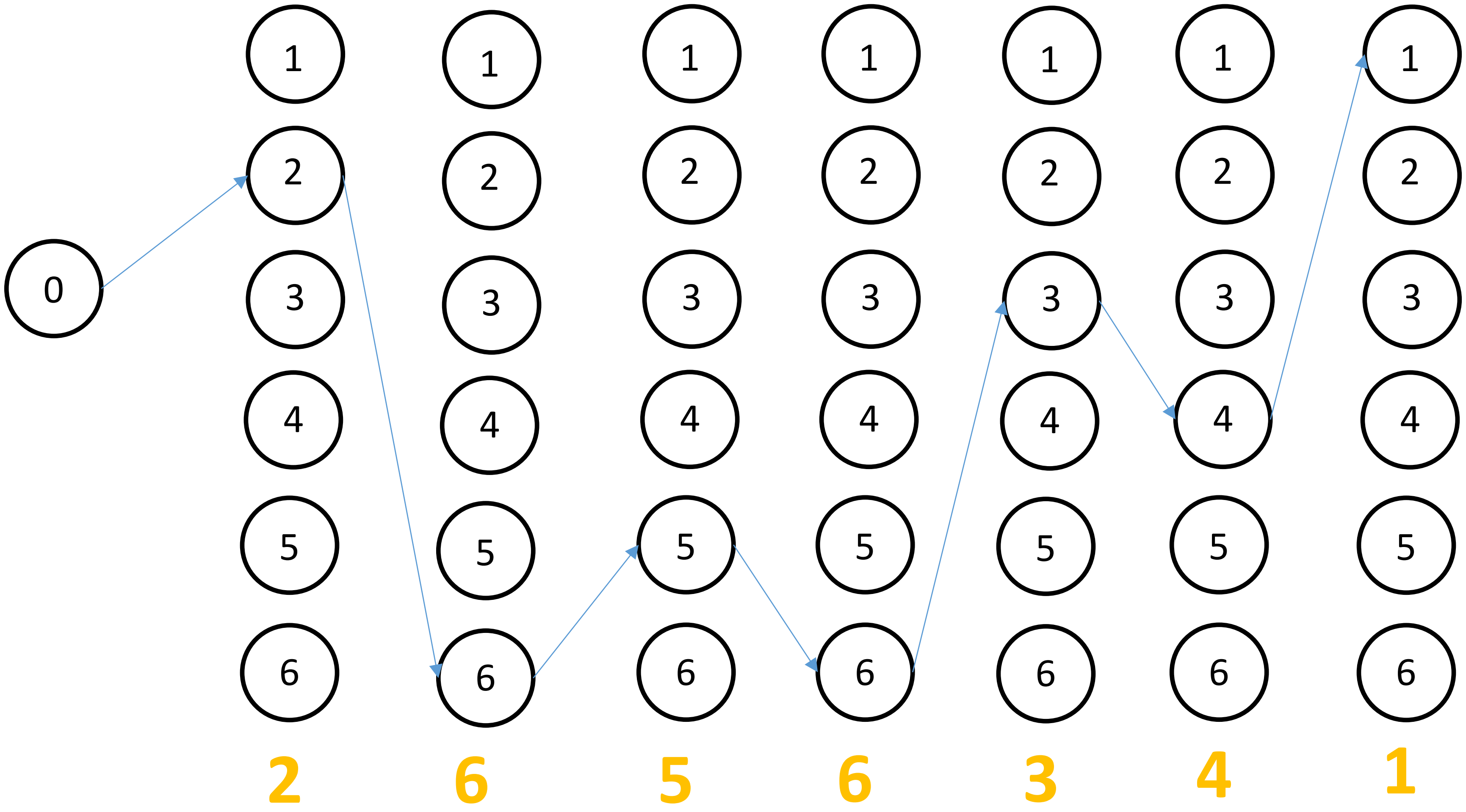
$$P(1|F) = P(2|F) = P(3|F) = P(4|F) = P(5|F) = P(6|F) = \frac{1}{6}$$

$$\begin{cases} P(1|L) = P(2|L) = P(3|L) = P(4|L) = P(5|L) = \frac{1}{10} \\ P(6|L) = \frac{1}{2} \end{cases}$$

Categories (Observed)

HMM

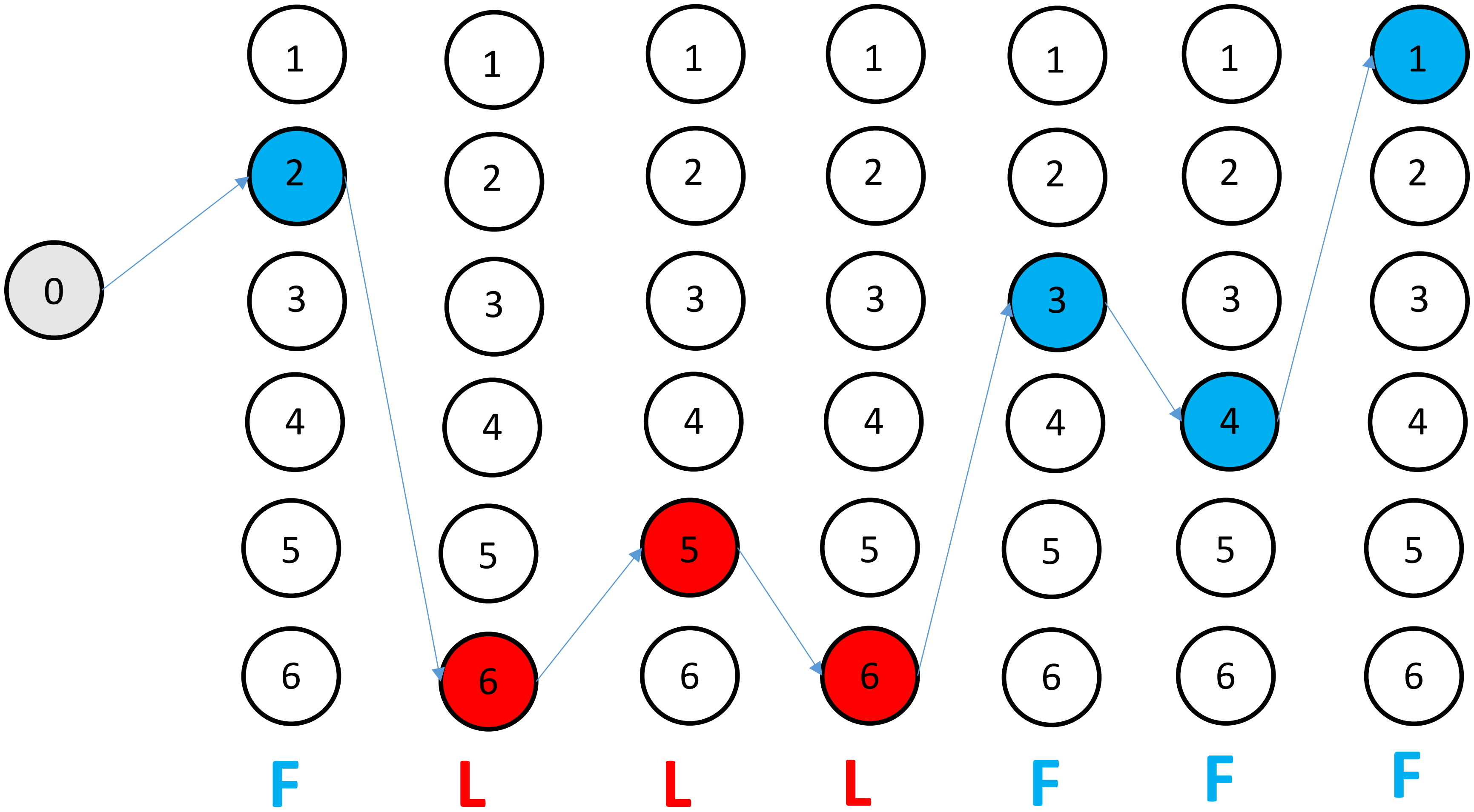
The occasionally dishonest casino problem. A run



HMM

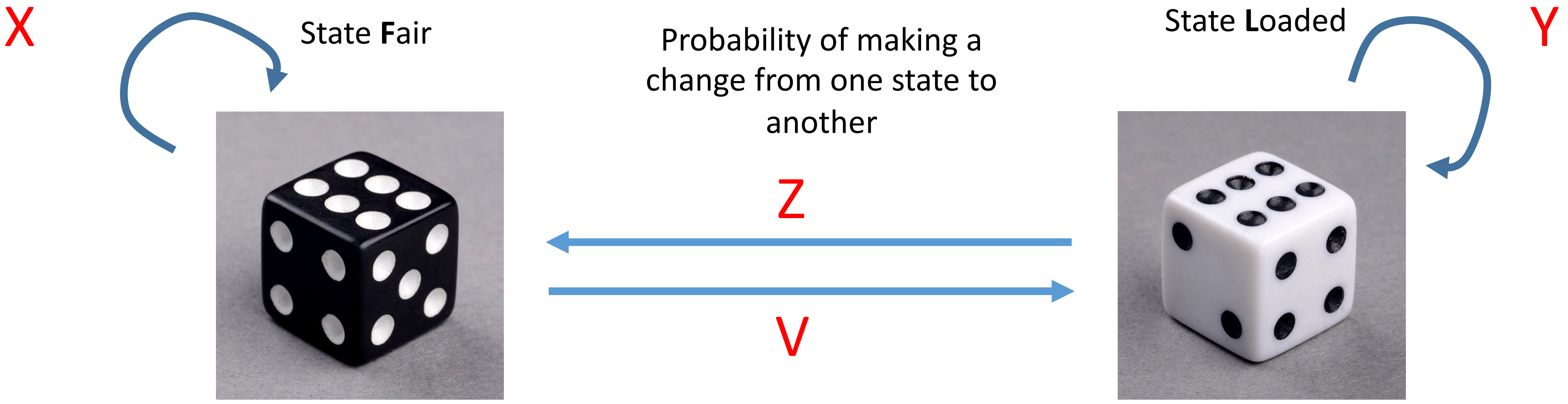
The occasionally dishonest casino problem. A run

States (that were used to generate the data, but in principle unknown)



HMM

The occasionally dishonest casino problem



X

State **F**air

Probability of making a change from one state to another

Z

V

State **L**oaded

Y

Emission probabilities of **F**air dice

1	2	3	4	5	6

Fair (F) dice

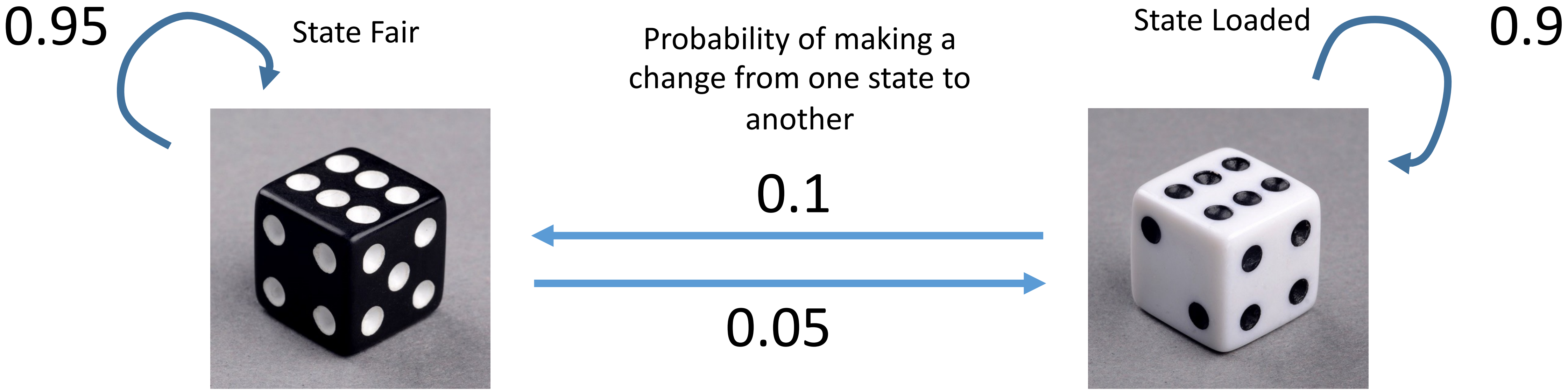
Emission probabilities of **L**oaded dice

1	2	3	4	5	6

Loaded (L) dice

HMM

The occasionally dishonest casino problem. Some notation



0.95

State Fair

Probability of making a change from one state to another

0.1

0.05

State Loaded

0.9

Emission probabilities of **F**air dice

1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

Fair (F) dice

Emission probabilities of **L**oaded dice

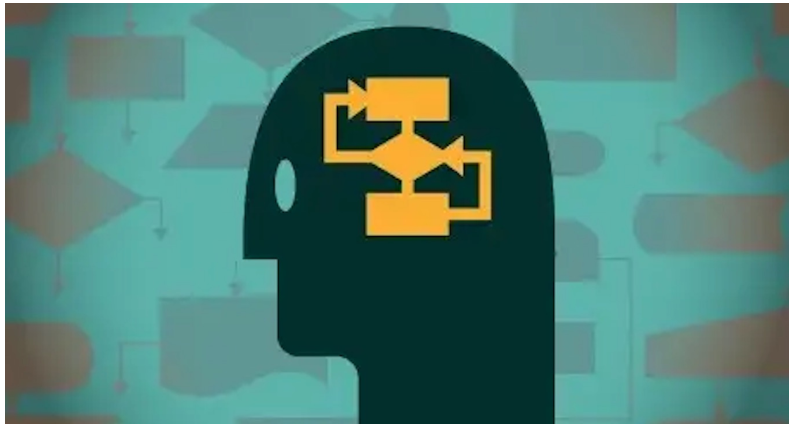
1	2	3	4	5	6
1/10	1/10	1/10	1/10	1/10	1/2

Loaded (L) dice

HMM

The *occasionally dishonest casino* problem.

How data are generated? Propose the *pseudo-algorithm*



HMM

The occasionally dishonest casino problem. Some notation

State at position i $= \pi_i$

Sequence of states $= \pi$

Category at position i $= x_i$

Sequence of categories $= x$

$$P(x,\pi) = a_{0\pi} \prod_{i}^{N} e_{\pi_i}(x_i) a_{\pi_i \pi_i+1}$$

Initial state

Probability of changing from state k to state r $= a_{kr} = P(\pi_i = r | \pi_{i-1} = k)$

I am using **state *r***

Given that I was in the previous position using **state *k***

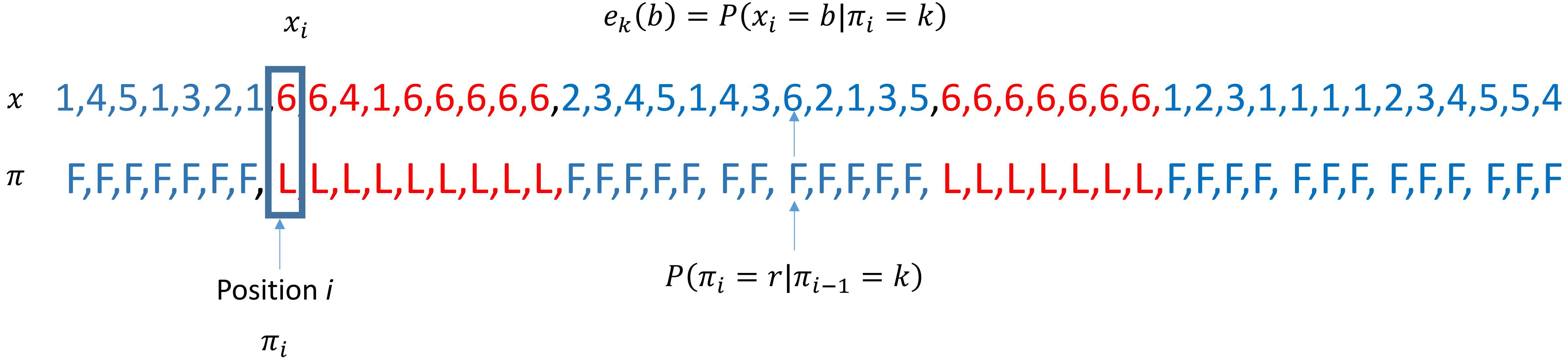
Probability of ***emission*** *of category b at position i given that I am using state k* $= e_k(b) = P(x_i = b | \pi_i = k)$

HMM

The occasionally dishonest casino problem

b = [1,6]

$e_k(b) = P(x_i = b|\pi_i = k)$



x_i

x 1,4,5,1,3,2,1,6,6,4,1,6,6,6,6,6,2,3,4,5,1,4,3,6,2,1,3,5,6,6,6,6,6,6,6,1,2,3,1,1,1,1,2,3,4,5,5,4

π F,F,F,F,F,F,F, L, L,L,L,L,L,L,L,L,F,F,F,F,F, F,F, F,F,F,F,F, L,L,L,L,L,L,L,F,F,F,F, F,F,F, F,F,F, F,F,F

Position *i*

π_i

$P(\pi_i = r|\pi_{i-1} = k)$

HMM

The occasionally dishonest casino problem

1,4,5,1,3,2,1,6,6,4,1,6,6,6,6,6,2,3,4,5,1,4,3,6,2,1,3,5,6,6,6,6,6,6,6,1,2,3,1,1,1,1,2,3,4,5,5,4

When is the casino using the Loaded dice?

HMM

The occasionally dishonest casino problem

1,4,5,1,3,2,1,6,6,4,1,6,6,6,6,6,2,3,4,5,1,4,3,6,2,1,3,5,6,6,6,6,6,6,6,1,2,3,1,1,1,1,2,3,4,5,5,4

Intuitively, when we have an excess of 6 one after the other we can imagine that the casino is using the *loaded* dice, because then the probability of getting 6 is 0.5

HMM

The occasionally dishonest casino problem

1,4,5,1,3,2,1,6,6,4,1,6,6,6,6,6,2,3,4,5,1,4,3,6,2,1,3,5 6,6,6,6,6,6,6 1,2,3,1,1,1,1,2,3,4,5,5,4

Intuitively, when we have an excess of 6 one after the other we can imagine that the casino is using the *Loaded* dice, because then the probability of getting 6 is 0.5

HMM

The occasionally dishonest casino problem

Imagine we observe the sequence x

$$x = \langle x_1, x_2, x_3 \rangle = \langle 6,1,6 \rangle$$

Assume that the prior probability of starting at one state or at the other is the same (0.5)

What would be the probability $P(x,\pi)$ if the state sequence was

$$\pi = \langle \pi_1, \pi_2, \pi_3 \rangle = \langle F, F, F \rangle$$

$$P(x,\pi) = a_{0\pi} \prod_{i=1}^{N} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

$$P(x,\pi) = \underset{1}{0.5 * \frac{1}{6}} * 0.95 * \frac{1}{6} * 0.95 * \frac{1}{6} = 0.00208912$$

Prior of starting at F ($a_{0\pi}$)

Emission prob of 6 at state F (e_{π_1}(x_1=6))

HMM

The occasionally dishonest casino problem

Imagine we observe the sequence x

$$x = \langle x_1, x_2, x_3 \rangle = \langle 6,1,6 \rangle$$

Assume that the prior probability of starting at one state or at the other is the same (0.5)

What would be the probability $P(x,\pi)$ if the state sequence was

$$\pi = \langle \pi_1, \pi_2, \pi_3 \rangle = \langle F, F, F \rangle$$

$$P(x,\pi) = a_{0\pi} \prod_{i=1}^{N} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

$$P(x,\pi) = 0.5 * \frac{1}{6} * 0.95 * \frac{1}{6} * 0.95 * \frac{1}{6} = 0.00208912$$

2

Prior of starting at F

Prob of staying at state F ($a_{\pi 1 \pi 2}$)

Emission prob of 6 at state F (e_{π_2})

HMM

The occasionally dishonest casino problem

Imagine we observe the sequence x

$$x = \langle x_1, x_2, x_3 \rangle = \langle 6,1,6 \rangle$$

Assume that the prior probability of starting at one state or at the other is the same (0.5)

What would be the probability $P(x,\pi)$ if the state sequence was

$$\pi = \langle \pi_1, \pi_2, \pi_3 \rangle = \langle F, F, F \rangle$$

$$P(x,\pi) = a_{0\pi} \prod_{i=1}^{N} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

$$P(x,\pi) = 0.5 * \frac{1}{6} * 0.95 * \frac{1}{6} * 0.95 * \frac{1}{6} = 0.00208912$$

3

Prior of starting at F

Prob of staying at state F ($a_{\pi 2 \pi 3}$)

Emission prob of 6 at state F (e_{π_3})

HMM

dishonest casino problem

$$P(x,\pi) = a_{0\pi} \prod_{i}^{N} e_{\pi_i}(x_i) a_{\pi_i \pi_i+1}$$

$$P(x,\pi) = 0.5 * \frac{1}{6} * 0.95 * \frac{1}{6} * \underset{3}{0.95 * \frac{1}{6}} = 0.00208912$$

Prior of starting at F

Prob of staying at state F ($a_{\pi 2\pi 3}$)

Emission prob of 6 at state F (e_{π_3})

HMM

lishonest casino problem

$$P(x,\pi) = a_{0\pi} \prod_{i}^{N} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

2. Finish MC: Assumes $a_{\pi N \pi N+1}$=1 We are 100% sure it finishes, regardless of the previous state and observation

1. Start MC: Assume that the prior probability of starting at one state or at the other is the same (0.5)

$$P(x,\pi) = \mathbf{0.5} * \frac{1}{6} * 0.95 * \frac{1}{6} * 0.95 * \frac{1}{6} * 1 = 0.00208912$$

3

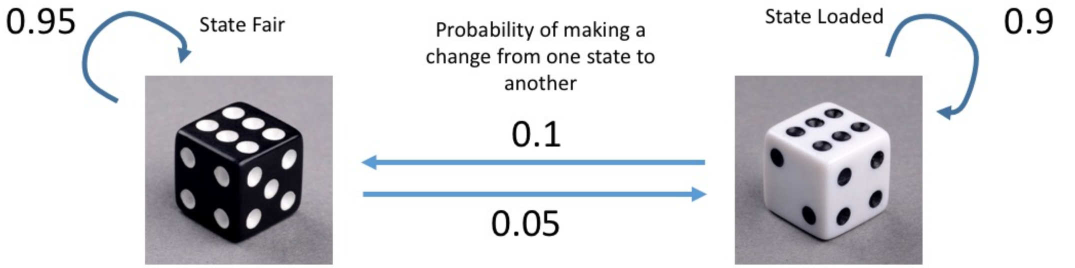
Prior of starting at F

Prob of staying at state F ($a_{\pi 2 \pi 3}$)

Emission prob of 6 at state F (e_{π_3})

HMM

The occasionally dishonest casino



Emission probabilities of Fair dice

1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

Emission probabilities of Loaded dice

1	2	3	4	5	6
1/10	1/10	1/10	1/10	1/10	1/2

6	1	6	$P(x,\pi)$
F	F	F	
F	F	L	
F	L	F	
F	L	L	…
L	F	F	
L	F	L	
L	L	F	
L	L	L	

Pick the hidden combination that maximizes the likelihood

Problem: The number of combinations grows exponentially

HMM

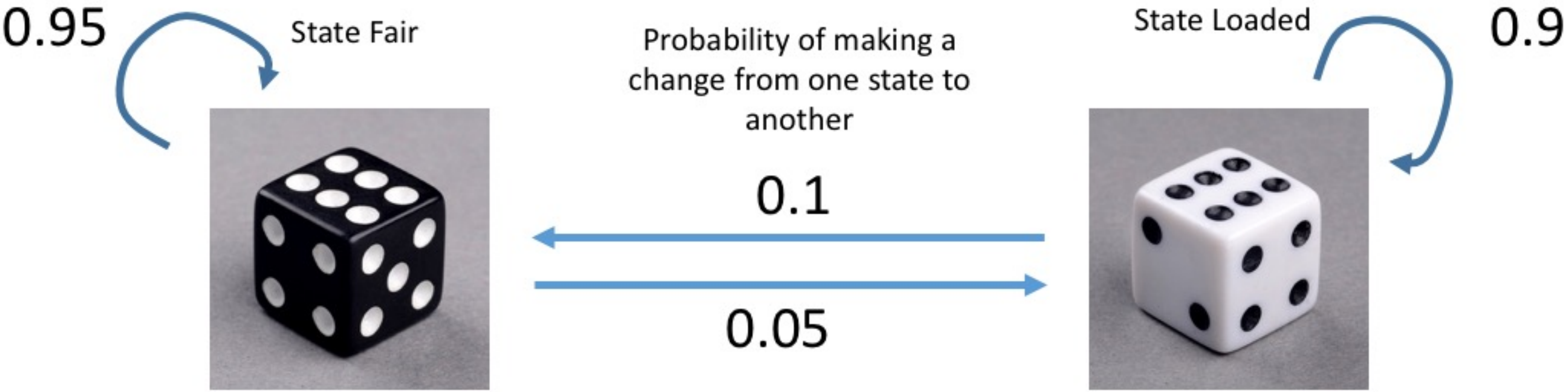
Long decimal numbers require *high precision*

6	1	6	$P(x,\pi)$
F	F	F	0.00208912
F	F	L	
F	L	F	6.94444E-06
F	L	L	…
L	F	F	
L	F	L	
L	L	F	
L	L	L	

Log2(6.94e-06)=-17.13

Bit units are **log2** this avoids *overflow errors*

We could also use log10 or natural log



Probability of making a change from one state to another

Emission probabilities of **F**air dice

1	2	3	4	5	6
1/6	1/6	1/6	1/6	1/6	1/6

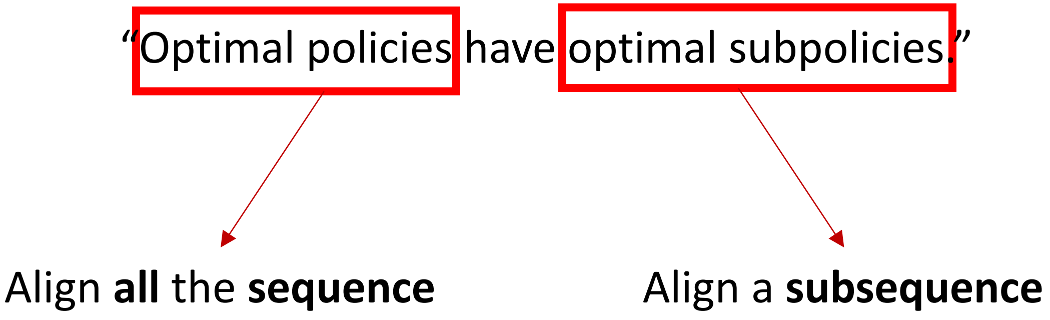
Emission probabilities of **L**oaded dice

1	2	3	4	5	6
1/10	1/10	1/10	1/10	1/10	1/2

Dynamic programming

Bellman’s Principle of Optimality

“An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.”



“Optimal policies have optimal subpolicies.”

Align **all** the **sequence**

Align a **subsequence**

Dynamic Programming, a Computational Tool.

Dynamic programming

- ***Sequential decision process***. Rather than exploring all possible solutions to the problem, decisions may be made in some specified sequence. There is a finite number of sequence operations to be done (*horizon*).

- ***Nested*** set of optimization ***operations***. *Each decision depends on the previous one*. Each decision leads to multiple next-states rather than a single one.

- ***Quantify the costs*** for each of the individual decisions.

- ***Reconstruction process***: determine the initial optimal decision, the optimal second decision that should be made in the next-state that results from the first decision, and so forth.

HMM

The occasionally dishonest casino problem

Markov chain

$$\begin{tikzcd} t0 & t1 & t2 & t3 \\ 0 \arrow[r] & 6 \arrow[r] & 1 \arrow[r] & 6 \end{tikzcd}$$

$$e_k(b) = P(x_i = b|\pi_i = k)$$

Probability of observing element b if I am in state k

HMM

The occasionally dishonest casino problem

	t0	t1	t2	t3
Markov chain	0 →	6 →	1 →	6

$$\max_{l}(p_l(j, x-1)p_{kl}) \quad e_k(b) = P(x_i = b|\pi_i = k)$$

Probability of the *most probable path* ending at position x-1 in state l with element j

Probability of *observing element b* if I am in state k

(emission)

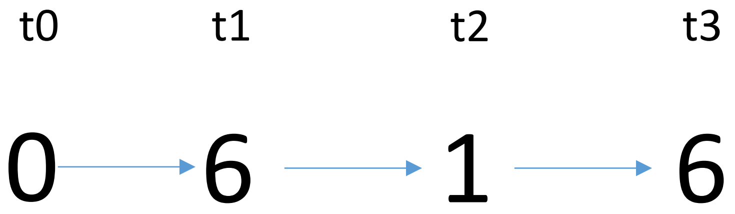
Probability *switching* from state l to k

(transition)

HMM

The occasionally dishonest casino problem

Markov chain



$$p(x,i) = \max_{l}\left(p_l(j, x-1)p_{kl}\right) e_k(b)$$

HMM

The occasionally dishonest casino problem

t0 t1 t2 t3

Markov chain 0 → 6 → 1 → 6

$$P(1|\pi_i = F; \pi_{i-1} = L) = P(\pi_{i-1} = L)P(\pi_i = F|\pi_{i-1} = L)P(1|\pi_i = F) = \log(0.25) + \log(0.1) + \log\left(\frac{1}{6}\right) =$$

	e	6	1	6
t	0	1	2	3
F	0.5	log(0.5*1/6)=log(**0.0833333**)		
L	0.5	log(0.5*1/2)=log(0.25)		

HMM

The occasionally dishonest casino problem

Markov chain

t0	t1	t2	t3

$$0 \rightarrow 6 \rightarrow 1 \rightarrow 6$$

$P(1|\pi_i = F; \pi_{i-1} = L) = P(\pi_{i-1} = L)P(\pi_i = F|\pi_{i-1} = L)P(1|\pi_i = F) = \log(0.25) + \log(0.1) + \log\left(\frac{1}{6}\right)$=

Switch (L->F)

	e	6	1	6
t	0	1	2	3
F	0.5	log(0.5*1/6)=log(**0.0833333**)	-5.480639	
L	0.5	log(0.5*1/2)=log(0.25)		

HMM

The occasionally dishonest casino problem

t0 t1 t2 t3

Markov chain $0 \rightarrow 6 \rightarrow 1 \rightarrow 6$

$P(1|\pi_i = F; \pi_{i-1} = L) = P(\pi_{i-1} = L)P(\pi_i = F|\pi_{i-1} = L)P(1|\pi_i = F) = \log(0.25) + \log(0.1) + \log\left(\frac{1}{6}\right)=$ *Switch (L->F)*

$P(1|\pi_i = L; \pi_{i-1} = L) = P(\pi_{i-1} = L)P(\pi_i = F|\pi_{i-1} = L)P(1|\pi_i = F) = \log(0.25) + \log(0.9) + \log\left(\frac{1}{10}\right)=$ *Stay (L->L)*

	e	6	1	6
t	0	1	2	3
F	0.5	log(0.5*1/6)=log(**0.0833333**)	-5.480639	
L	0.5	log(0.5*1/2)=log(0.25)	-3.740173	

HMM

The occasionally dishonest casino problem

$$\begin{tikzcd}[column sep=large]
t0 & t1 & t2 & t3 \\
0 \arrow[r] & 6 \arrow[r] & 1 \arrow[r] & 6
\end{tikzcd}$$

Markov chain

$$P(1|\pi_i = F; \pi_{i-1} = L) = P(\pi_{i-1} = L)P(\pi_i = F|\pi_{i-1} = L)P(1|\pi_i = F) = \log(0.25) + \log(0.1) + \log\left(\frac{1}{6}\right)=$$

Switch (L->F)

$$P(1|\pi_i = L; \pi_{i-1} = L) = P(\pi_{i-1} = L)P(\pi_i = F|\pi_{i-1} = L)P(1|\pi_i = F) = \log(0.25) + \log(0.9) + \log\left(\frac{1}{10}\right)=$$

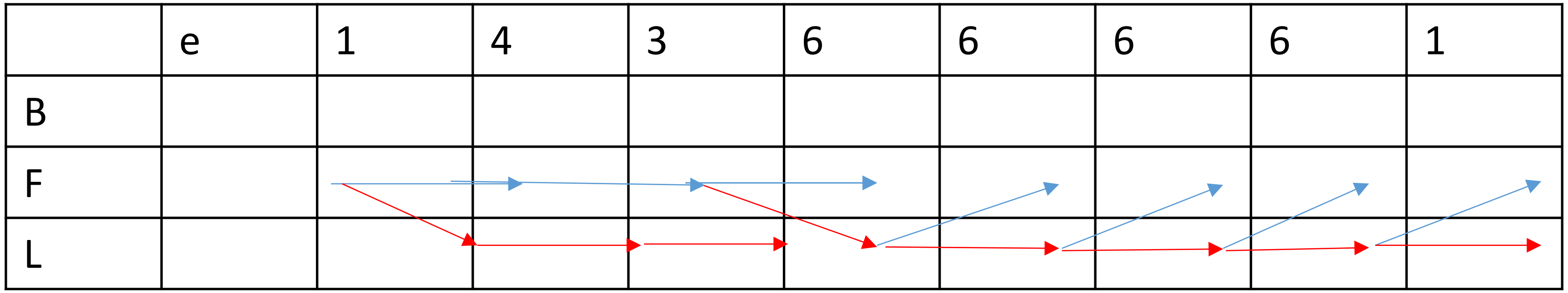
Stay (L->L)

	e	6	1	6
t	0	1	2	3
F	0.5	log(0.5*1/6)=log(**0.0833333**)	-5.480639	?
L	0.5	log(0.5*1/2)=log(0.25)	-3.740173	

HMM

The occasionally dishonest casino problem

Move Forward



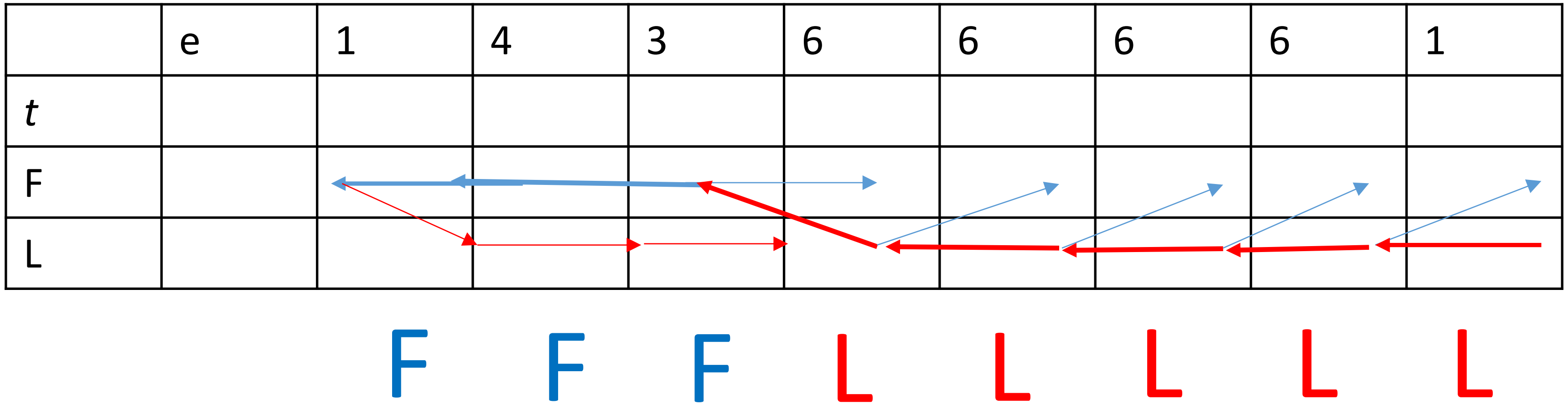
	e	1	4	3	6	6	6	6	1
B									
F									
L									

F F F L L L L L

HMM

The occasionally dishonest casino problem

Move Backwards



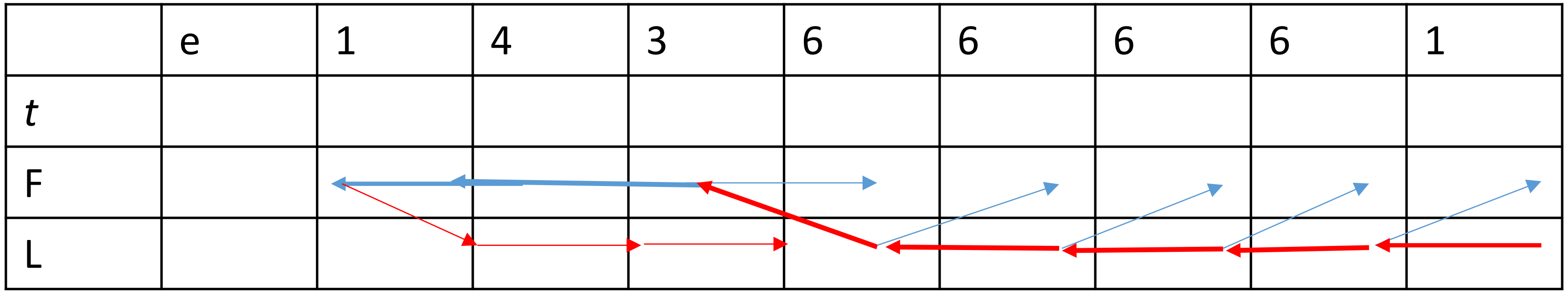
	e	1	4	3	6	6	6	6	1
t									
F									
L									

F F F L L L L L

HMM

The occasionally dishonest casino problem

Find the ***optimal path*** $\pi^{*} = argmax_{\pi} P(x, \pi)$



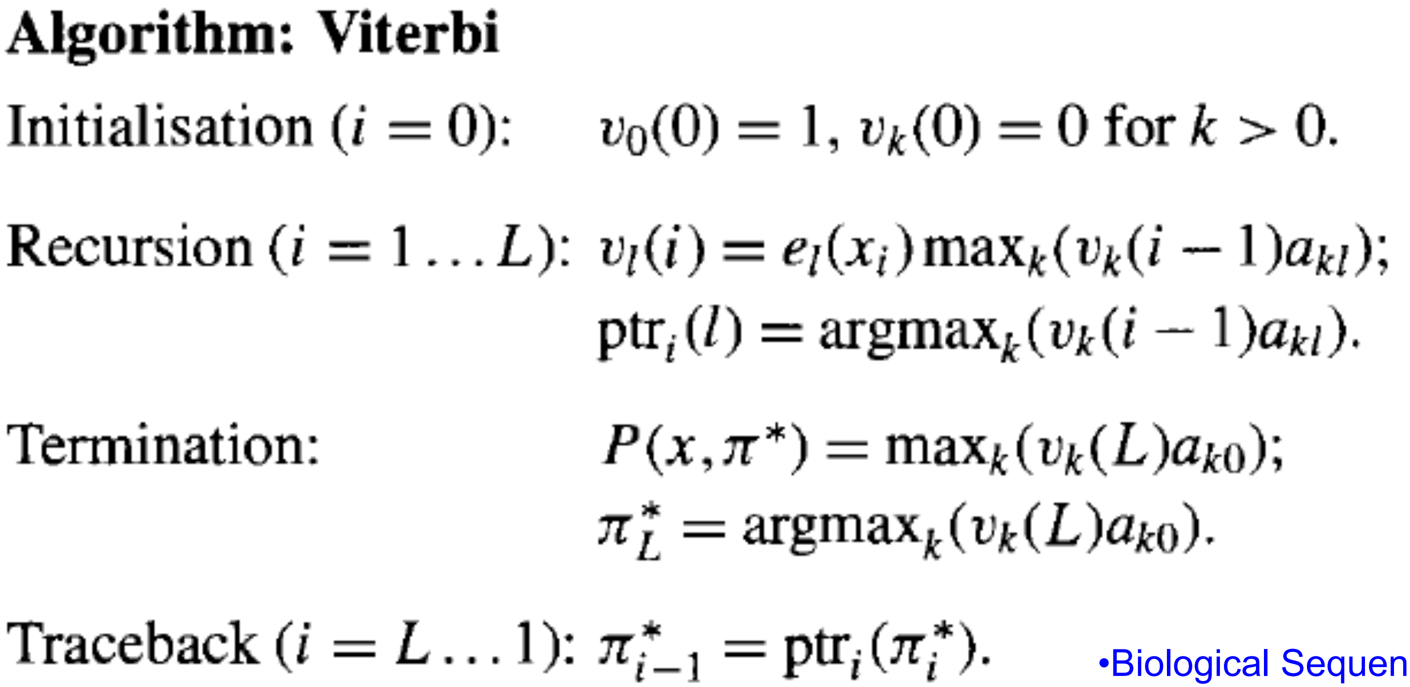
	e	1	4	3	6	6	6	6	1
t									
F									
L									

F F F L L L L L

HMM

Basically what the Viterbi algorithm does

Find the ***optimal path*** $\pi^* = argmax_\pi P(x,\pi)$



Algorithm: Viterbi

Initialisation ($i = 0$): $v_0(0) = 1$, $v_k(0) = 0$ for $k > 0$.

Recursion ($i = 1 \dots L$): $v_l(i) = e_l(x_i)\max_k(v_k(i-1)a_{kl})$;

$\mathrm{ptr}_i(l) = \mathrm{argmax}_k(v_k(i-1)a_{kl})$.

Termination: $P(x,\pi^*) = \max_k(v_k(L)a_{k0})$;

$\pi_L^* = \mathrm{argmax}_k(v_k(L)a_{k0})$.

Traceback ($i = L \dots 1$): $\pi_{i-1}^* = \mathrm{ptr}_i(\pi_i^*)$.

•Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Richard Durbin, Sean R. Eddy,Anders Krogh,Graeme Mitchison, 1998 (last edition 2013). Cambridge University Press.

HMM: Generate HMM model (states, emissions, probability of change from state)

- We want to identify regions of High GC content from regions of Low GC content.
- We know that if we are in a High GC content nucleotide, the probability of moving to a low GC content is 0.6. If we are low, then the probability of changing to high is 0.3
- The nucleotide composition in High GC content is A:0.1, T:0.1, C:0.4, G:0.4
- The nucleotide composition in Low GC content is A:0.4, T:0.4, C:0.1, G:0.1