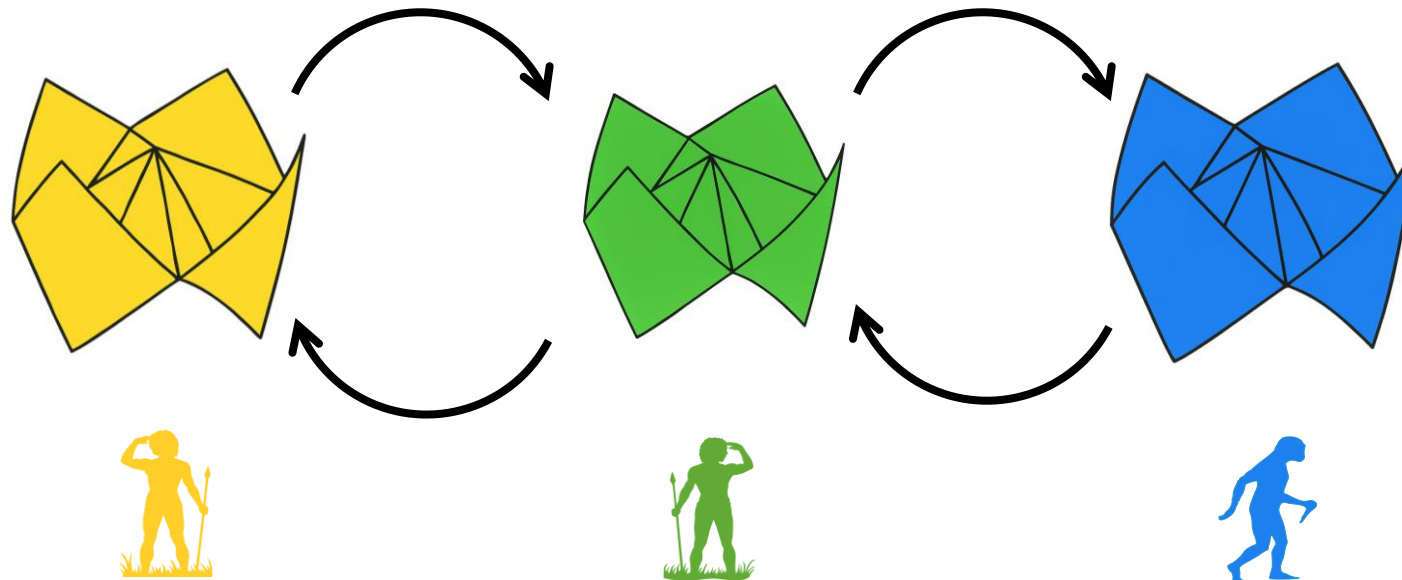


Hidden Markov Models

Fabrizio Mafessoni, Genomica Computazionale, Corso di Laurea Specialistica in Genomica Funzionale, Università di Trieste

Come identificare aplotipi, imputare genomi, trovare nuove forme di CRISPR/Cas, stimare la dimensione di una popolazione dai genomi, caratterizzare domini proteici, identificare zone del genoma metilate o altamente espresso, identificare le zone target di fattori di trascrizione, etc etc.....

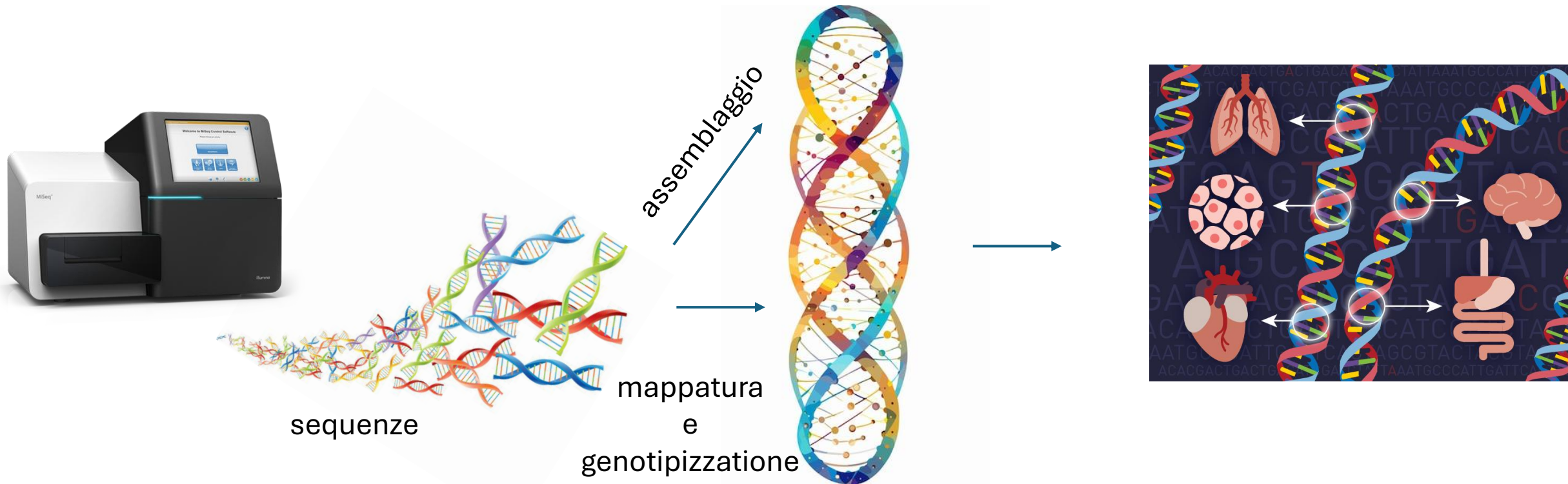


Scopo ultimo del corso: saper partire da sequenze di DNA fino ad assemblare interi genomi e caratterizzarli funzionalmente

Dal sequenziatore..

..al genoma..

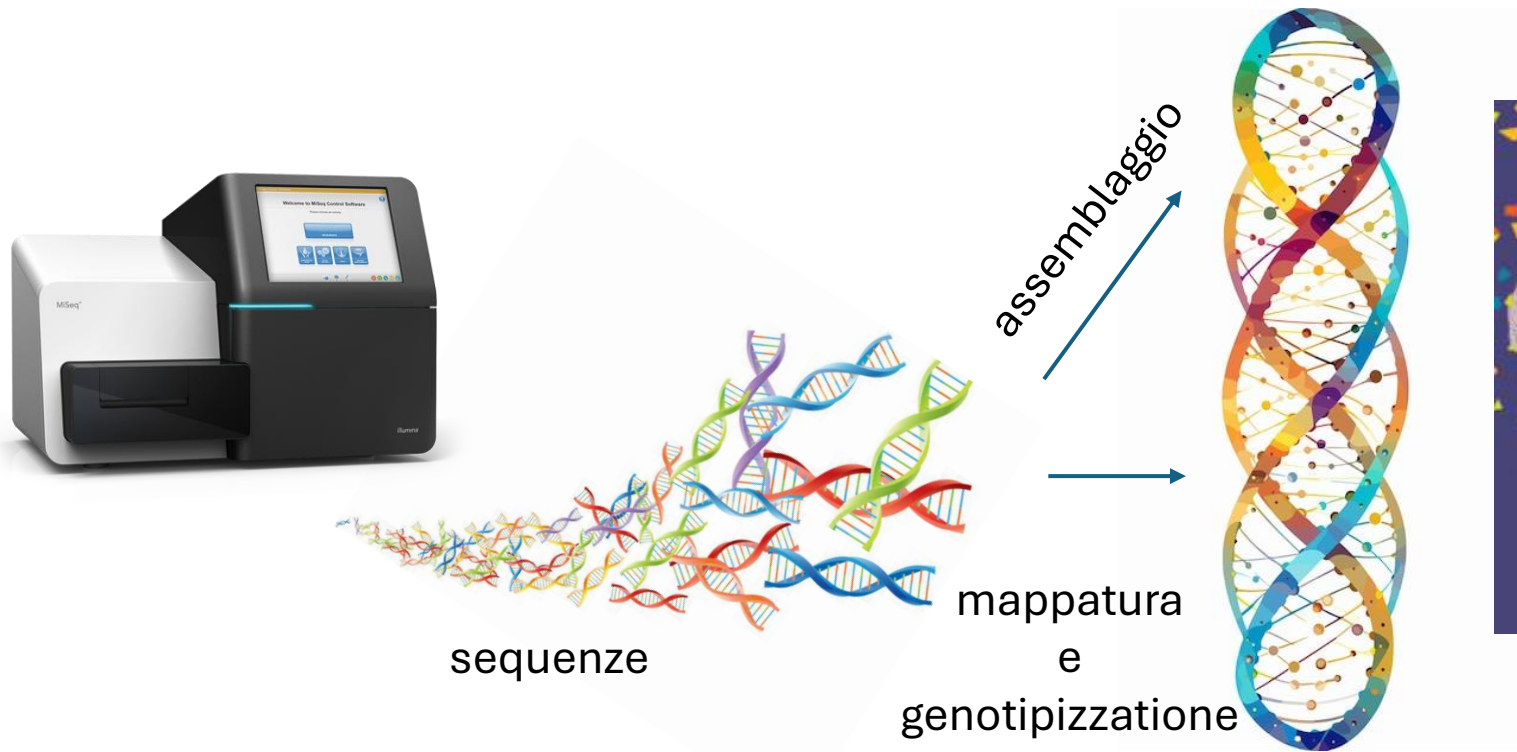
..alla funzione



Complimenti, potete costruire genomi e genotipi di piu' individui partire a partire da sequenze di DNA!

Dal sequenziatore..

..al genoma..



Formati di file bioinformatici

.fastq

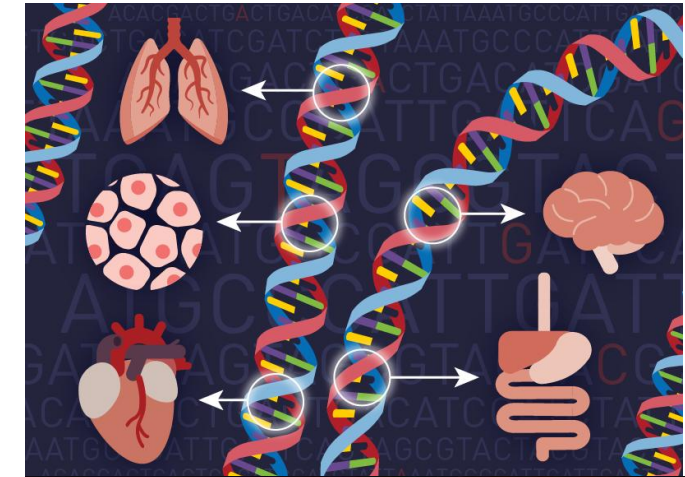
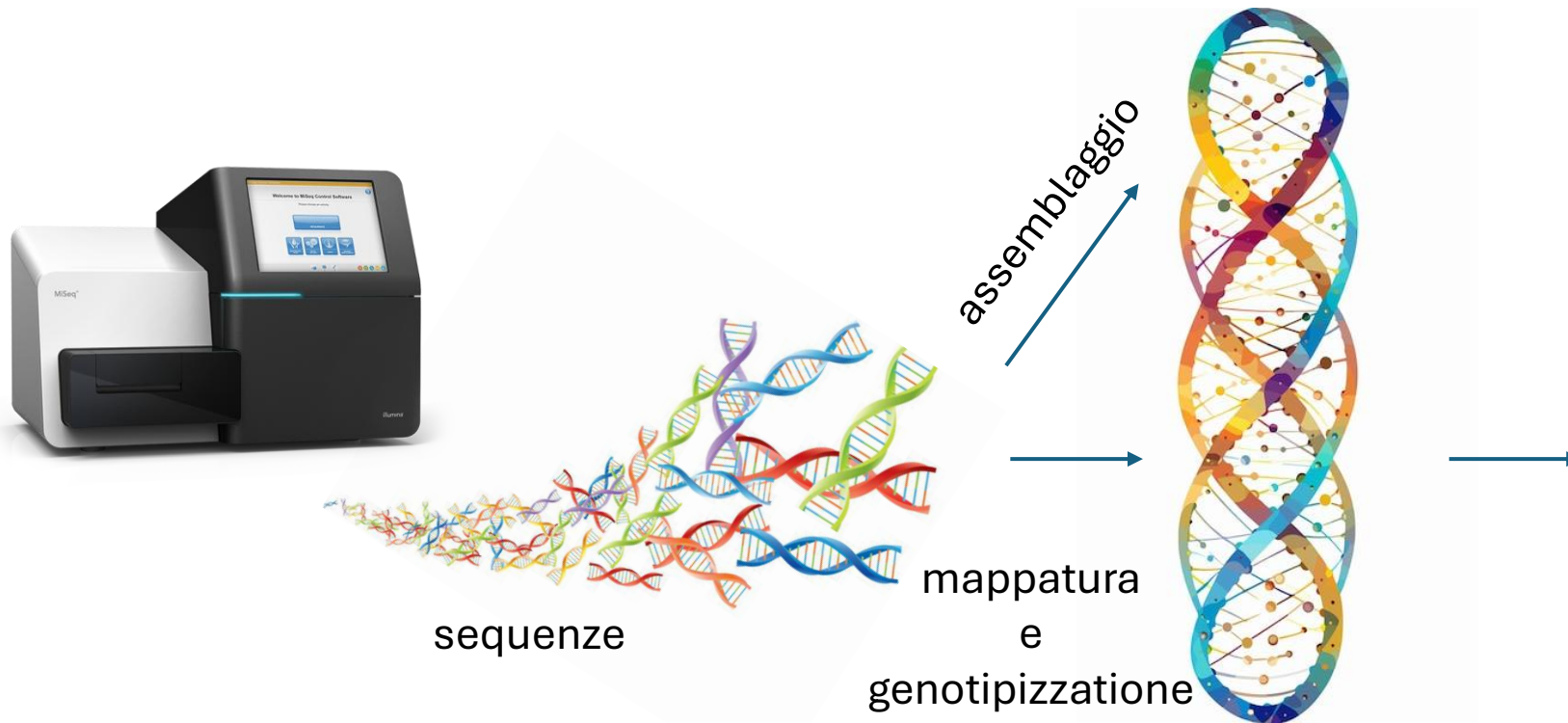
.fasta, .bam, .vcf

Ultima parte: come caratterizzare “funzionalmente” varie regioni del genoma

Dal sequenziatore..

..al genoma..

..alla funzione



SO EXCITED!!!



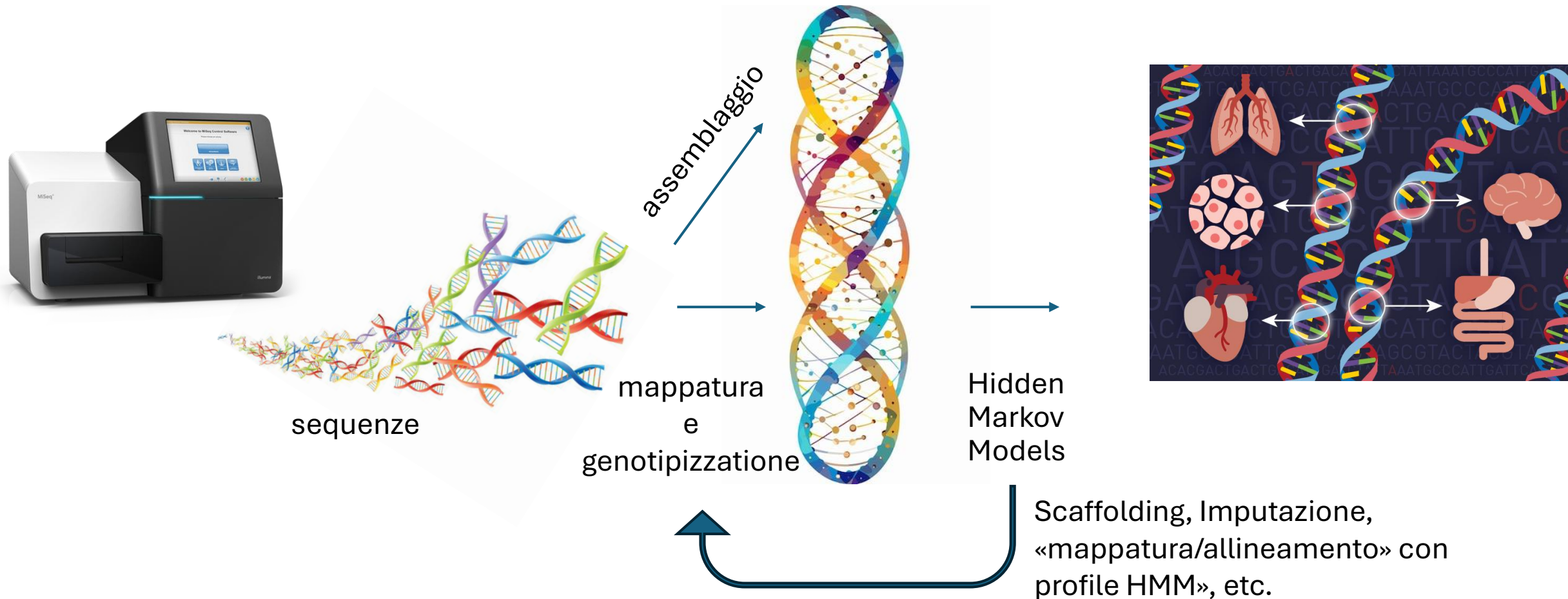
WE CAN'T WAIT!!!!!

Gli Hidden Markov Models sono utilizzatissimi in moltissimi ambiti, inclusi quelli già visti durante il corso!

Dal sequenziatore..

..al genoma..

..alla funzione



Un genoma è una struttura «lineare» con
regioni con caratteristiche diverse



Un genoma è una struttura «lineare» con regioni con caratteristiche diverse

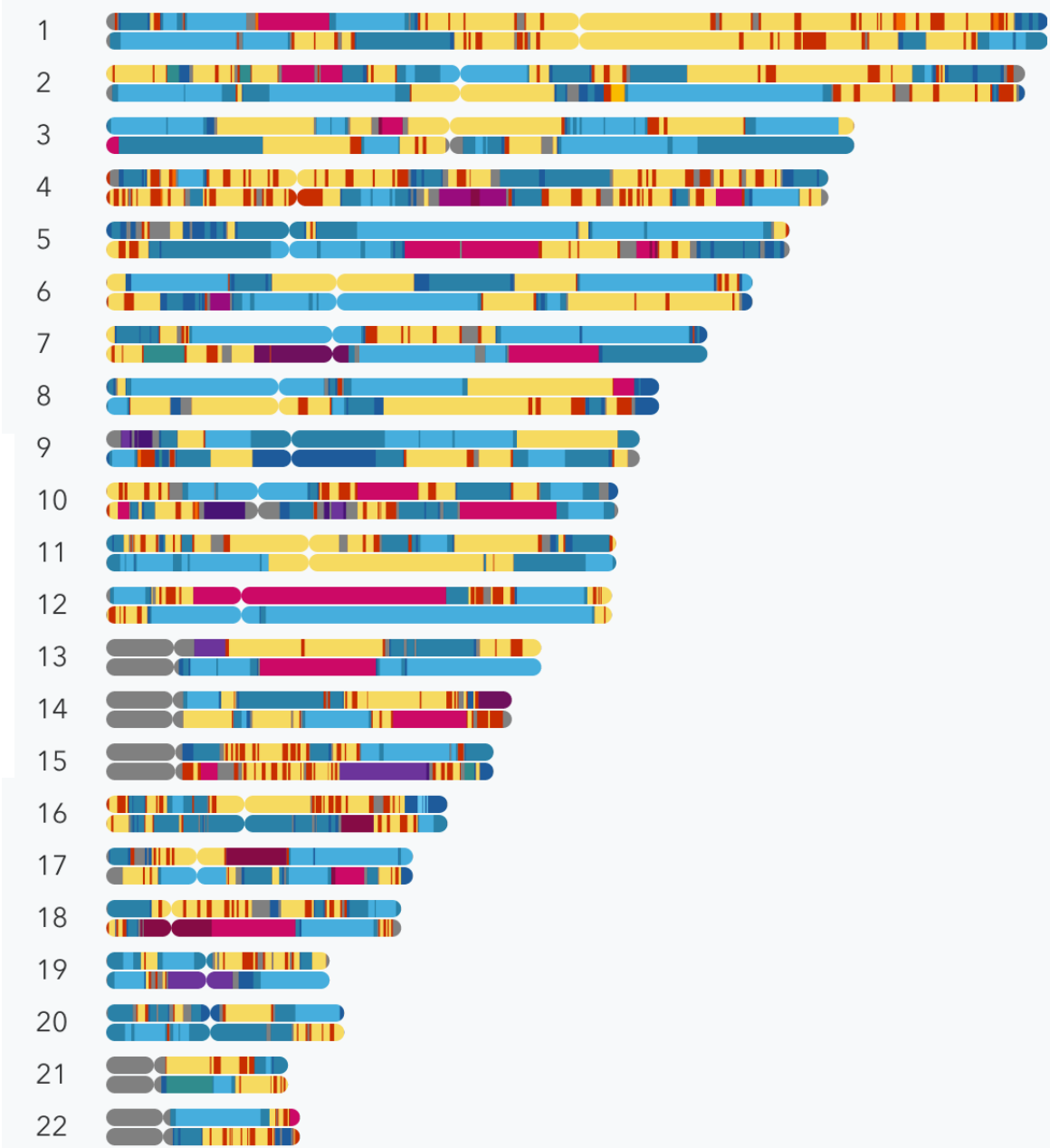


Caratteristiche evolutive

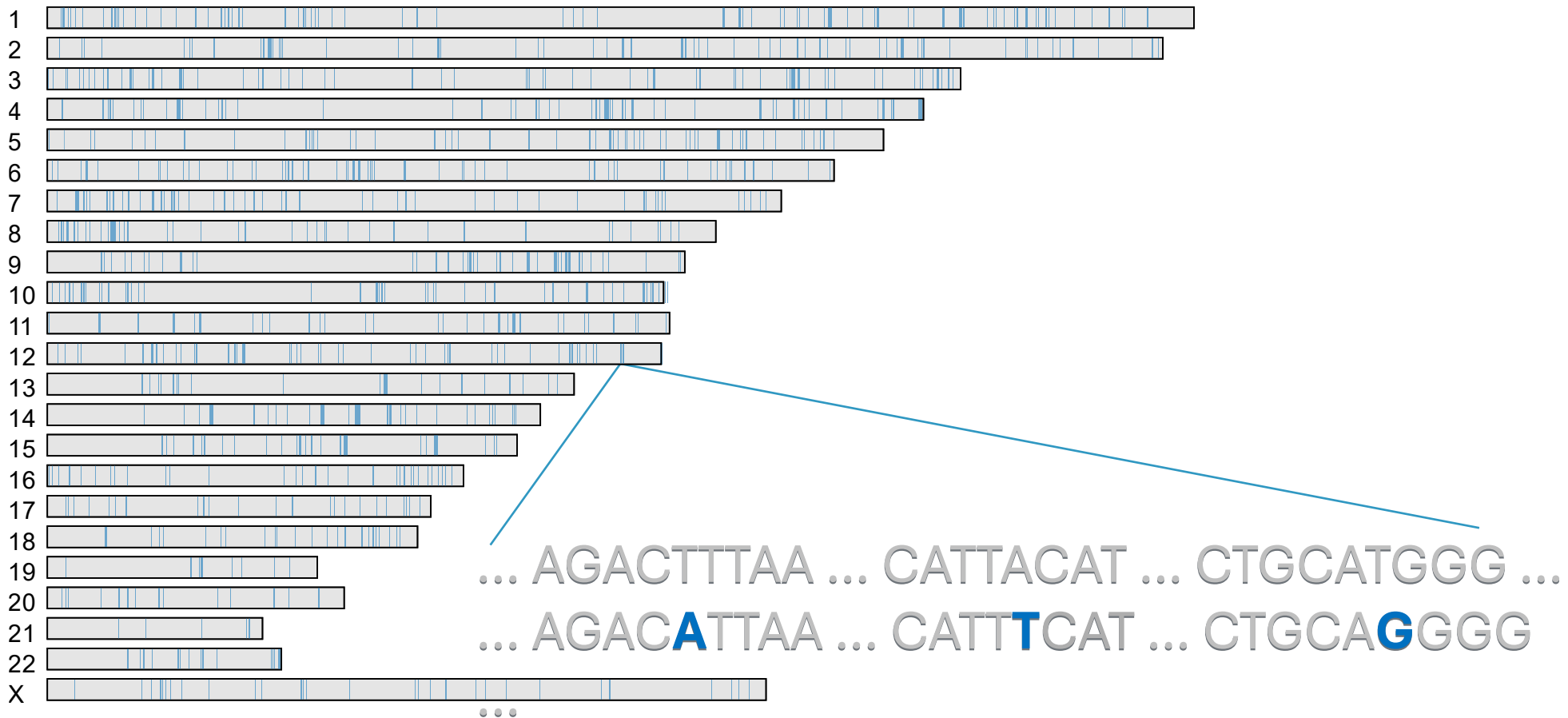
(ancestry, aplotipi, popolazione effettiva, introgressione,...)



Chromosome-painting



European	48.3%
Southern European	43.3%
Iberian	25.3%
Ashkenazi	0.5%
Northwestern European	< 0.1%
East Asian & Native American	38.7%
Native American	30.7%
East Asian	0.1%
Southeast Asian	< 0.1%
Sub-Saharan African	7.1%
West African	5.3%
Central & South African	0.6%
East African	0.3%



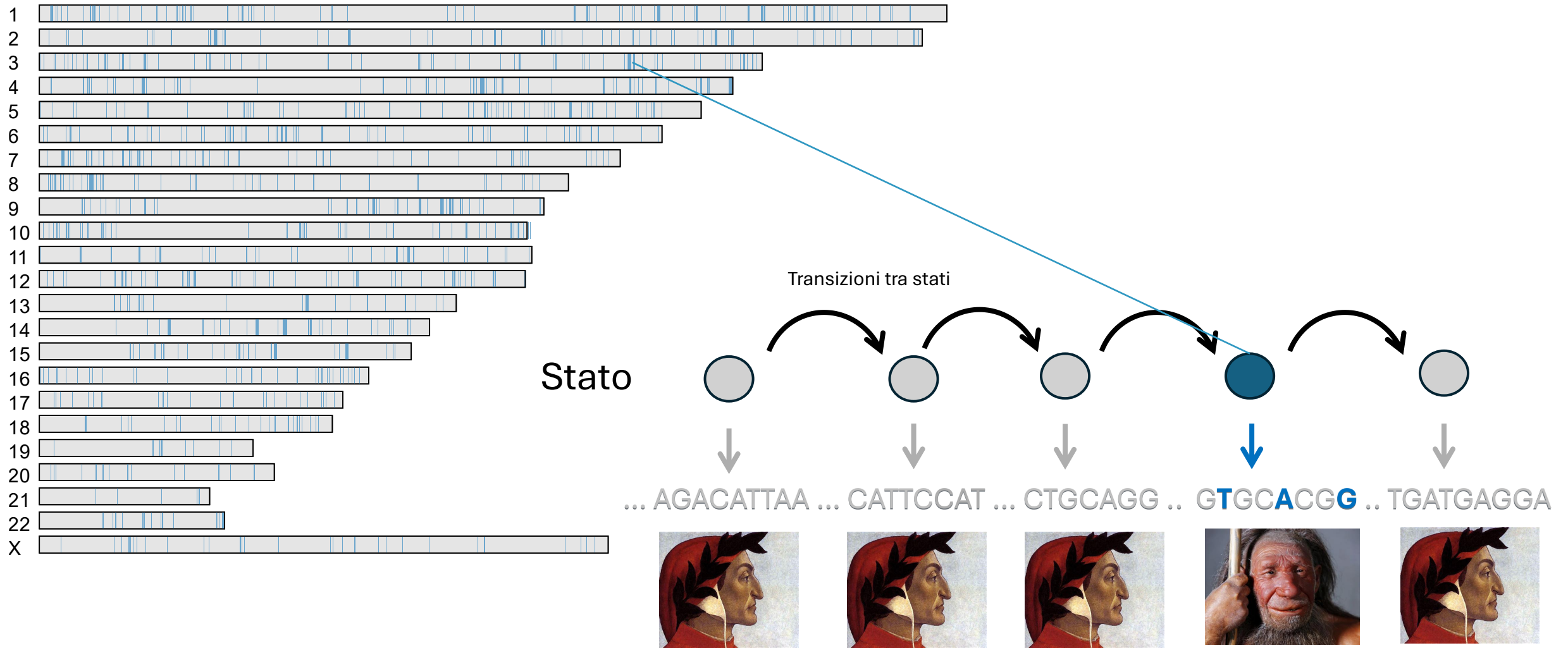
... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG

...

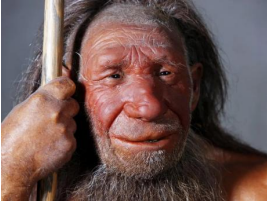
... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG

...





Un genoma è una struttura «lineare» con regioni con caratteristiche diverse



Caratteristiche evolutive

(ancestry, aplotipi, popolazione effettiva, introgressione,...)

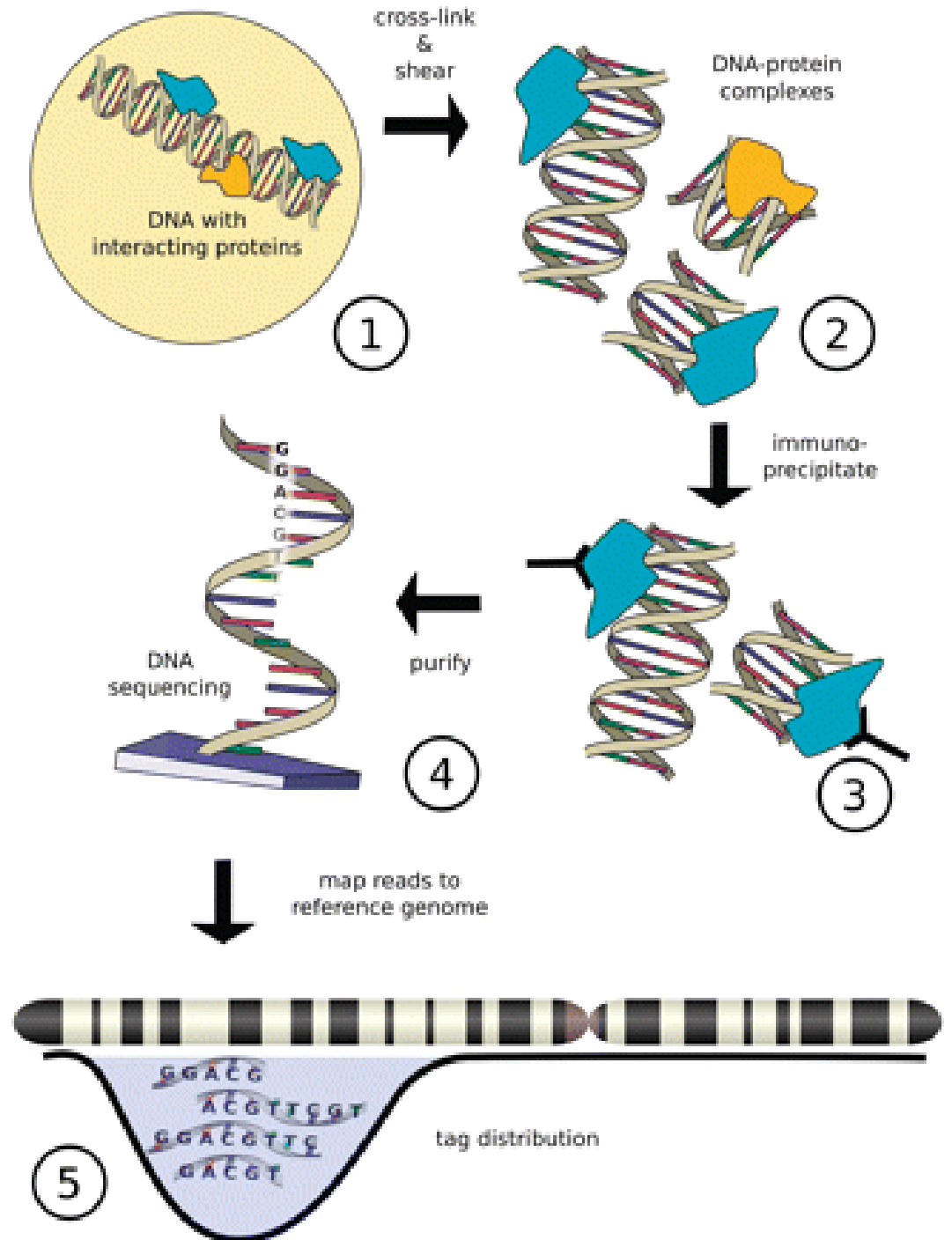


Caratteristiche della cromatina

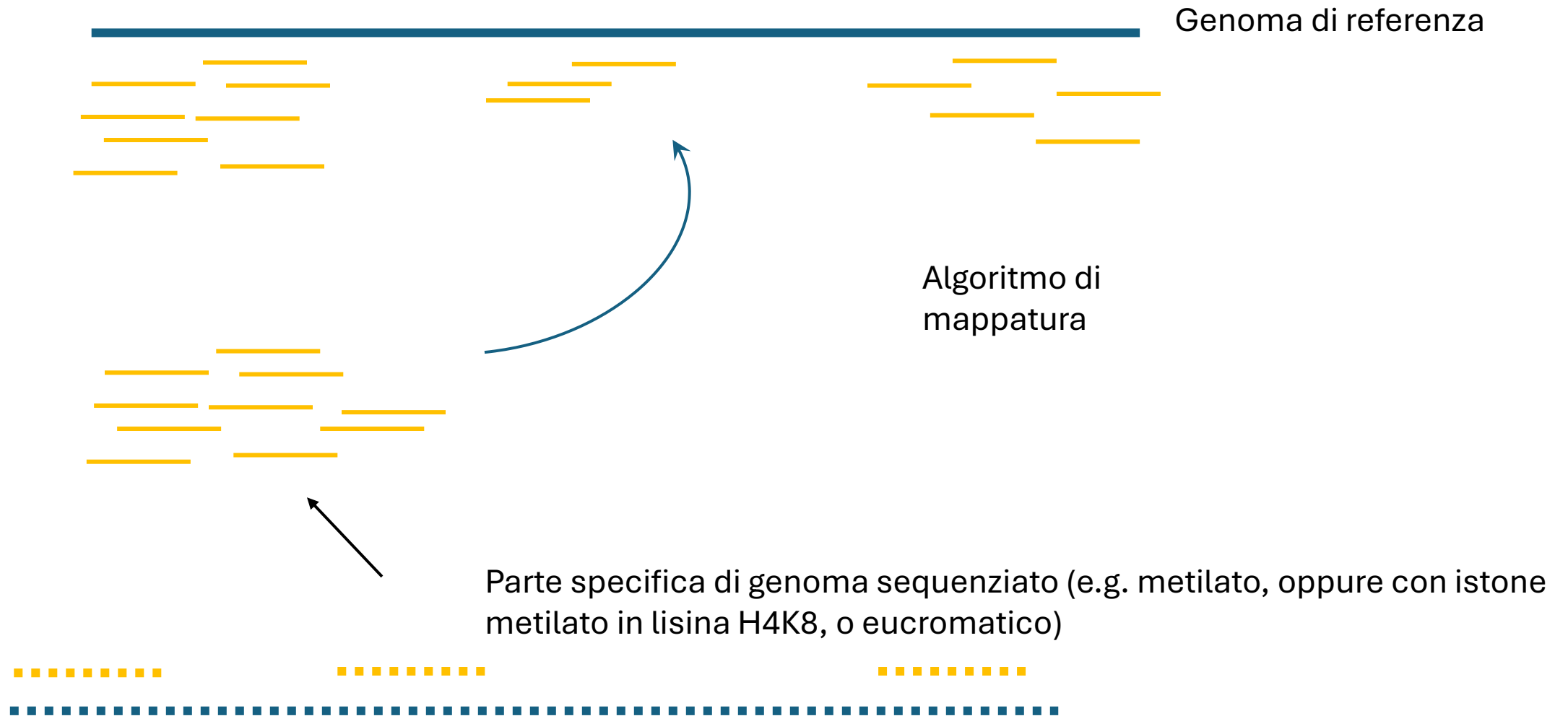
(posizionamento dei nucleosomi, metilazione, Chip-seq, segmentazione multi-marker..)



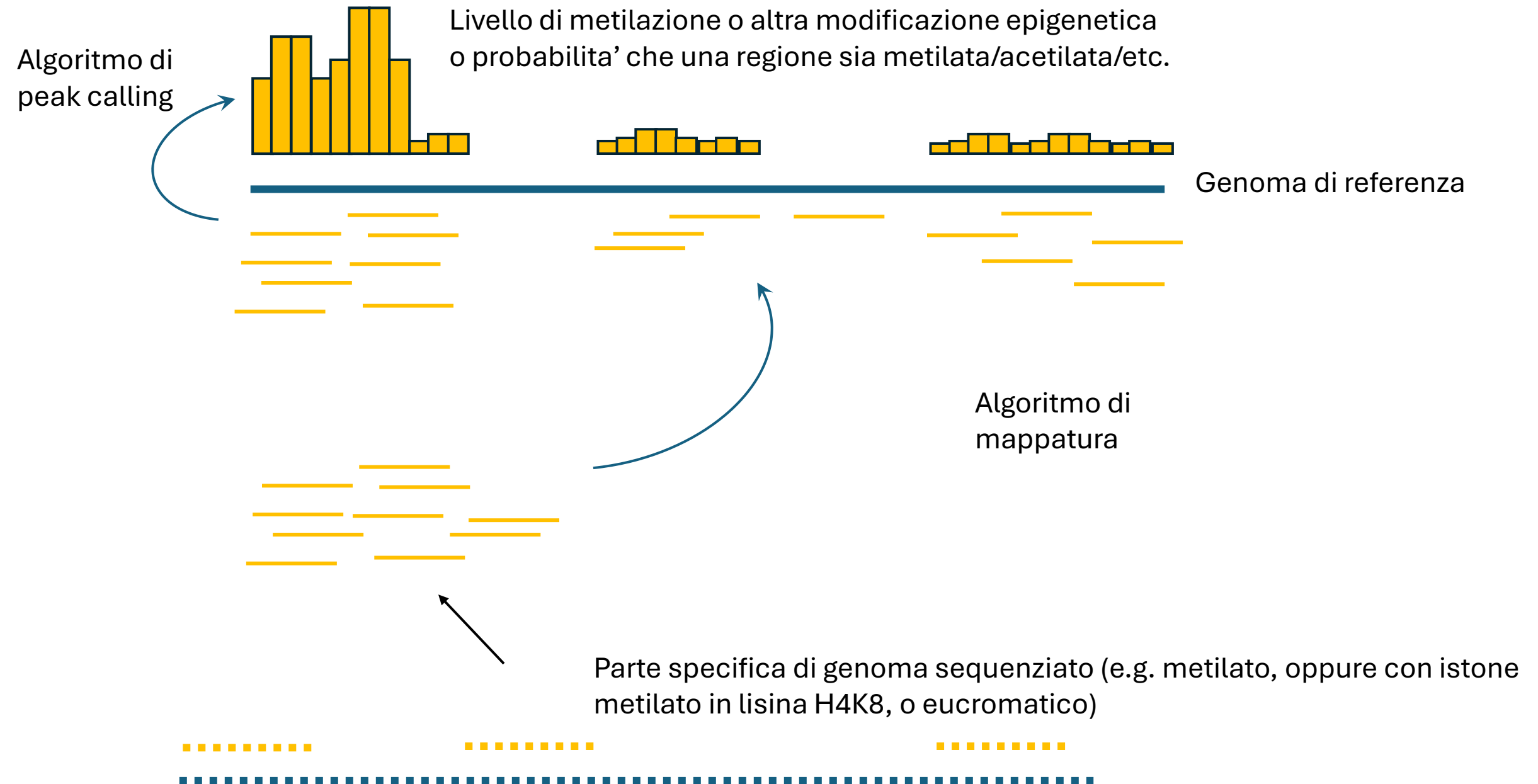
Immunoprecipitazione della cromatina (chip-seq)



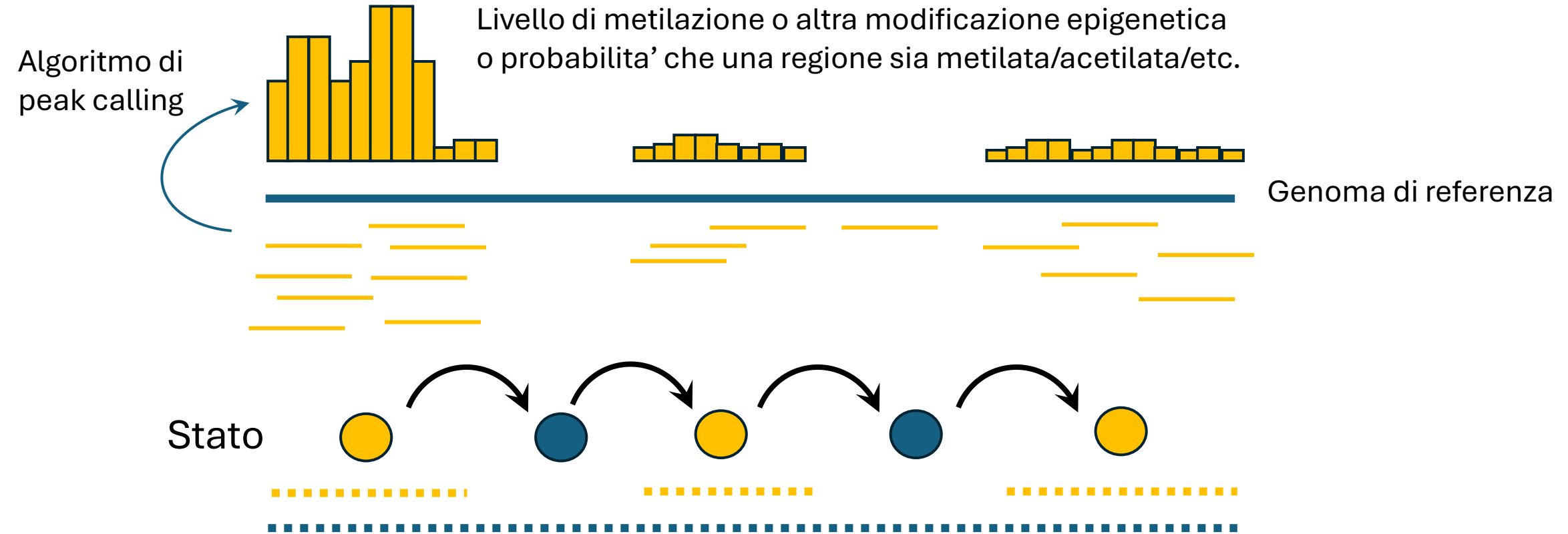
Peak calling per epigenomica



Peak calling per epigenomica



Peak calling per epigenomica



Esempio:

- Regione non-metilata 
- Regione metilata 

Un genoma è una struttura «lineare» con regioni con caratteristiche diverse



Caratteristiche evolutive

(ancestry, aplotipi, popolazione effettiva, introgressione,...)

Caratteristiche di dei patterns di espressione genica

(identificazione di promoters,
ATACseq, DNase I
hypersensitivity)



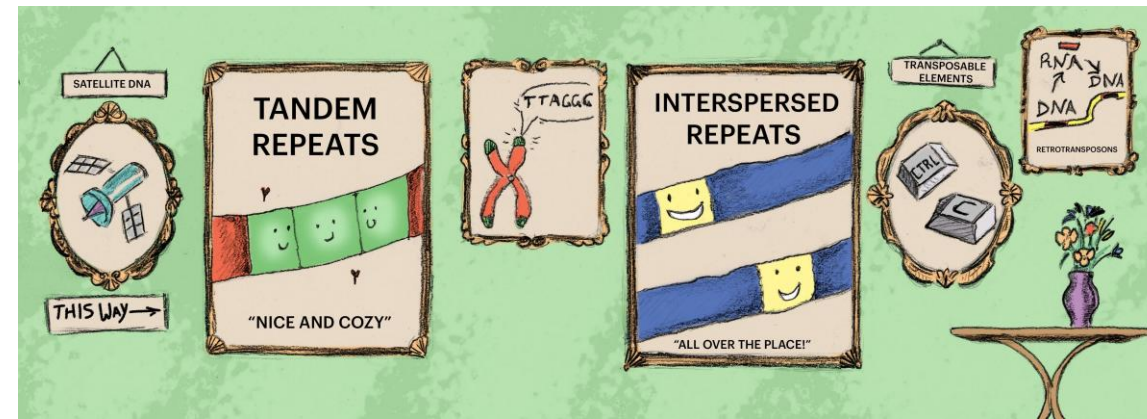
Caratteristiche della cromatina

(posizionamento dei nucleosomi,
metilazione, Chip-seq,
segmentazione multi-marker..)

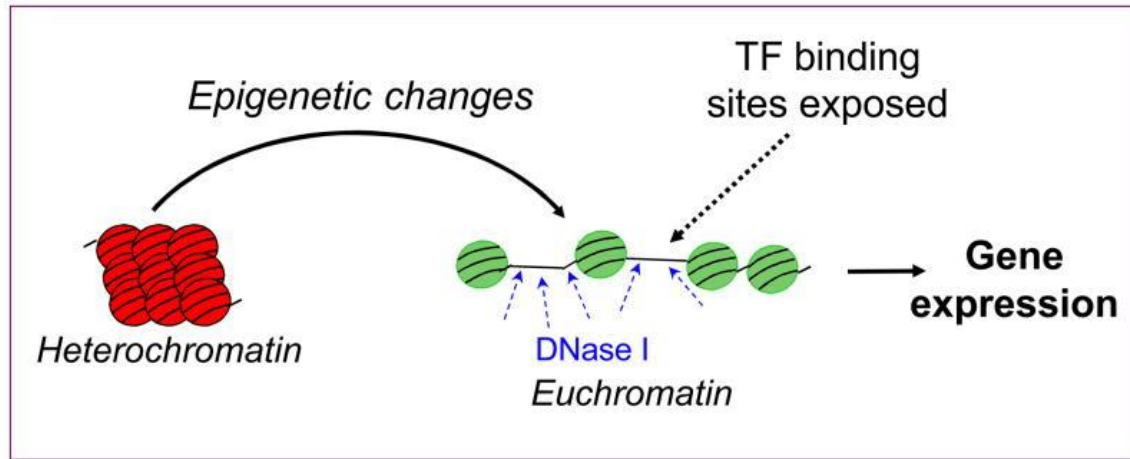


Caratteristiche strutturali

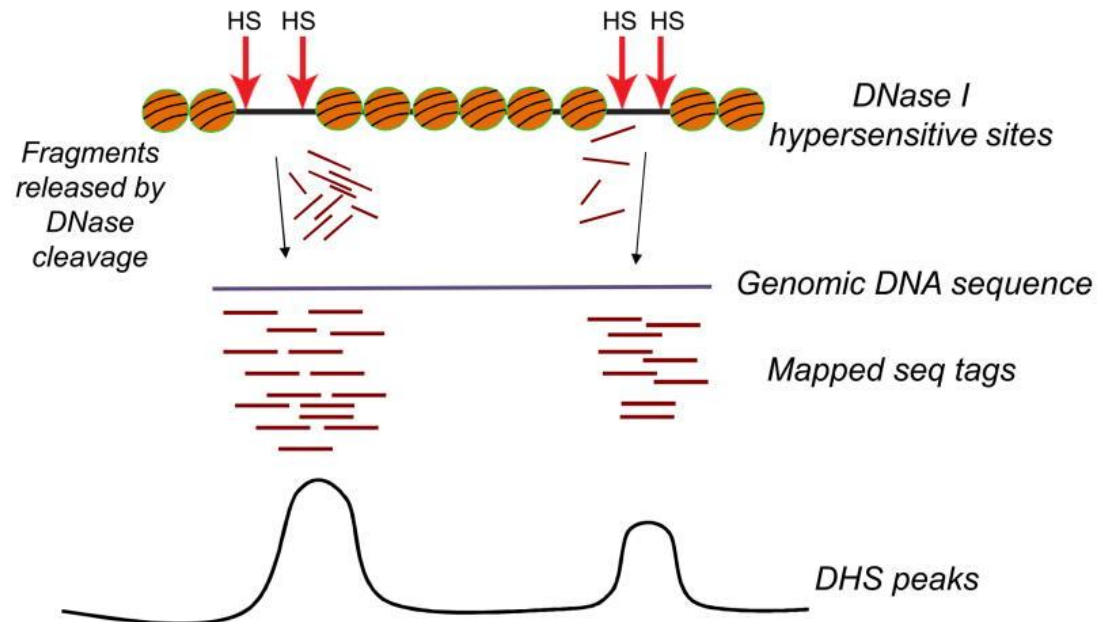
(ripetizioni, isole CpG,..)



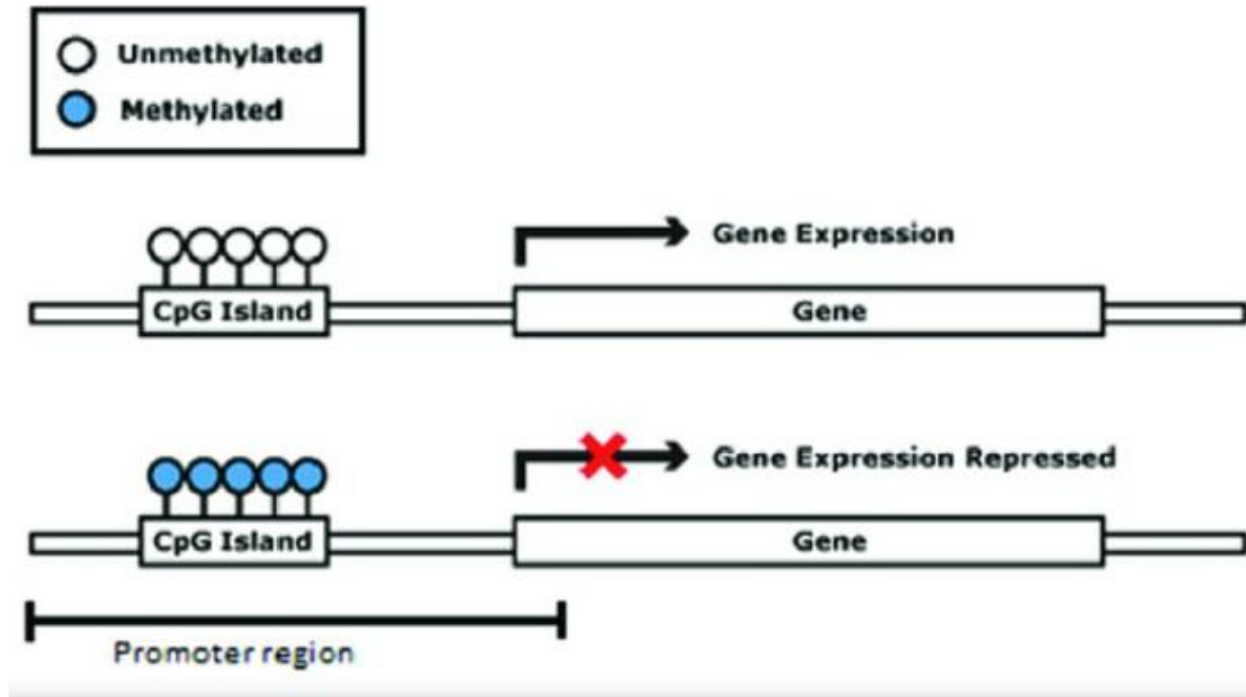
Ipersensitività alla DNAsi I



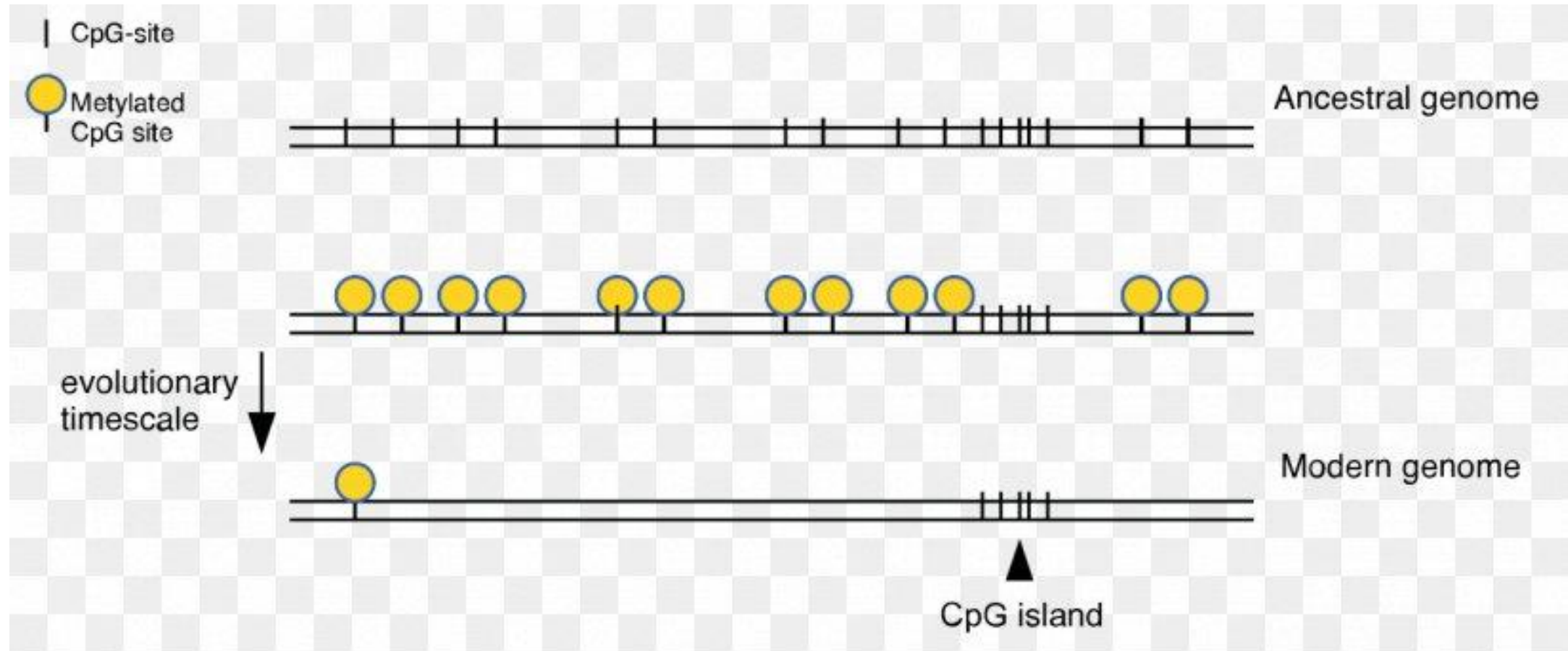
DNase-Seq analysis



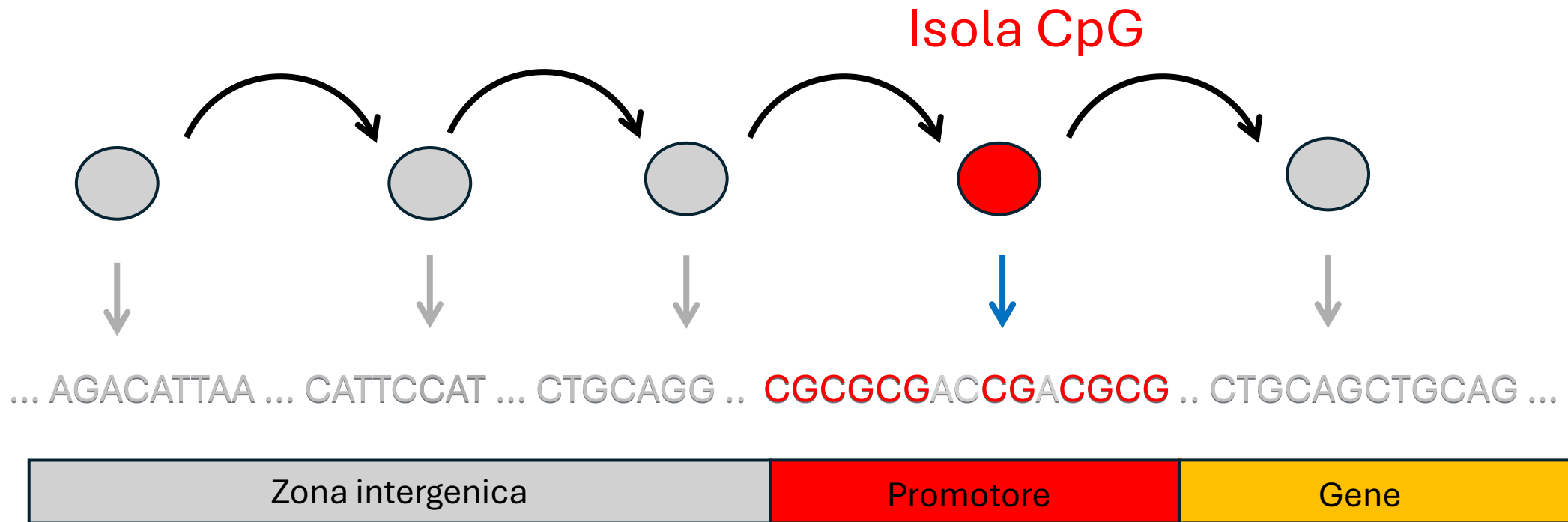
Le isole CpG caratterizzano l'80% dei promotori di mammifero



A causa dell'alto tasso di mutazione delle CpG metilate, i dinucleotidi CG sono rari (1%) fuori dalle isole



Struttura delle isole CpG modellizzate come un Hidden Markov Model



Gli ingredienti di un Hidden Markov Model

stati e probabilità di transizione (tra stati)

2 Stati

Esempio:

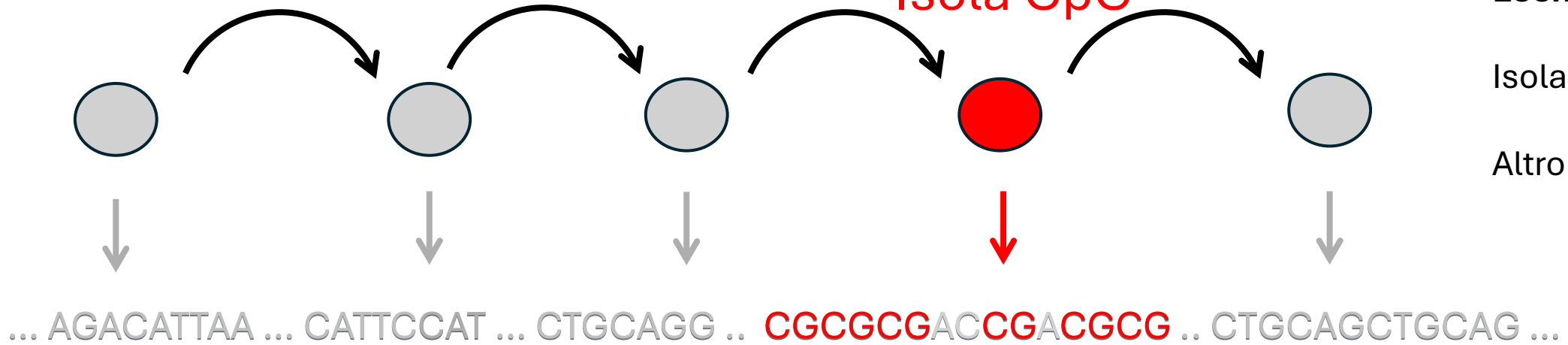
Isola CpG



Altro



Isola CpG



Zona intergenica

Promotore

Gli ingredienti di un Hidden Markov Model

... AGACATTAA ... CATTCCAT ... CTGCAGG .. **CGCGCGACCGACGCG** .. CTGCAGCTGCAG ...

Osservazioni: i nucleotidi A,C,G,T

Gli ingredienti di un Hidden Markov Model

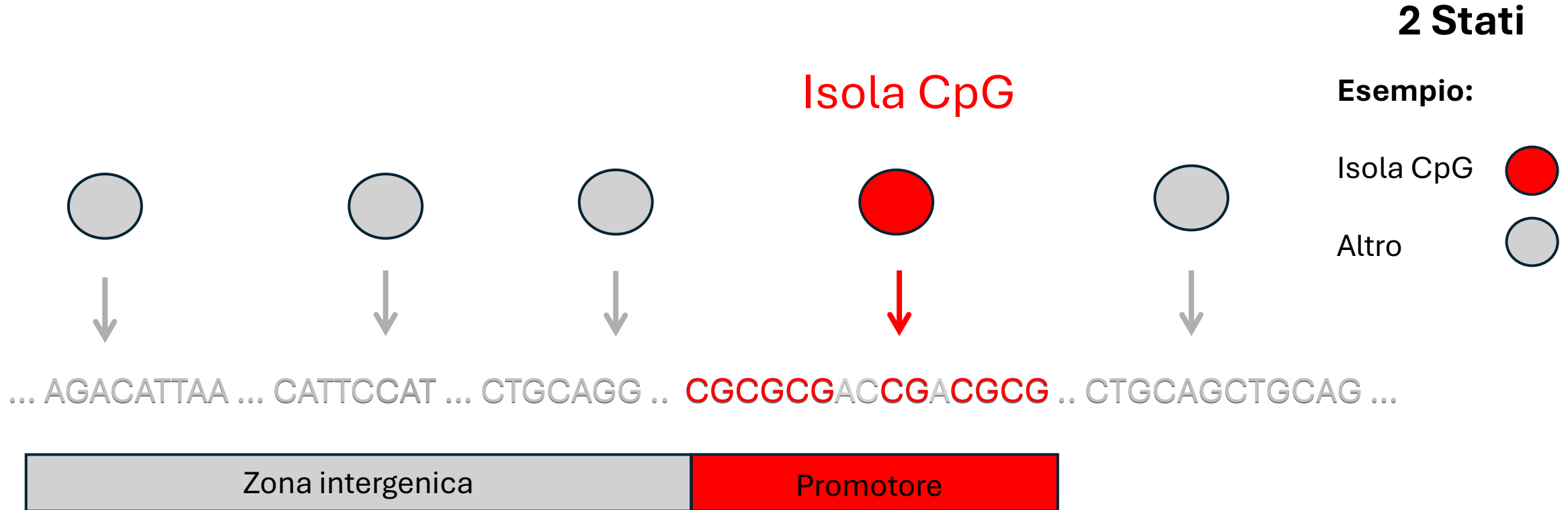
... AGACATTAA ... CATTCCAT ... CTGCAGG .. **CGCGCGACCGACGCG** .. CTGCAGCTGCAG ...



Osservazioni: i nucleotidi A,C,G,T

Gli ingredienti di un Hidden Markov Model

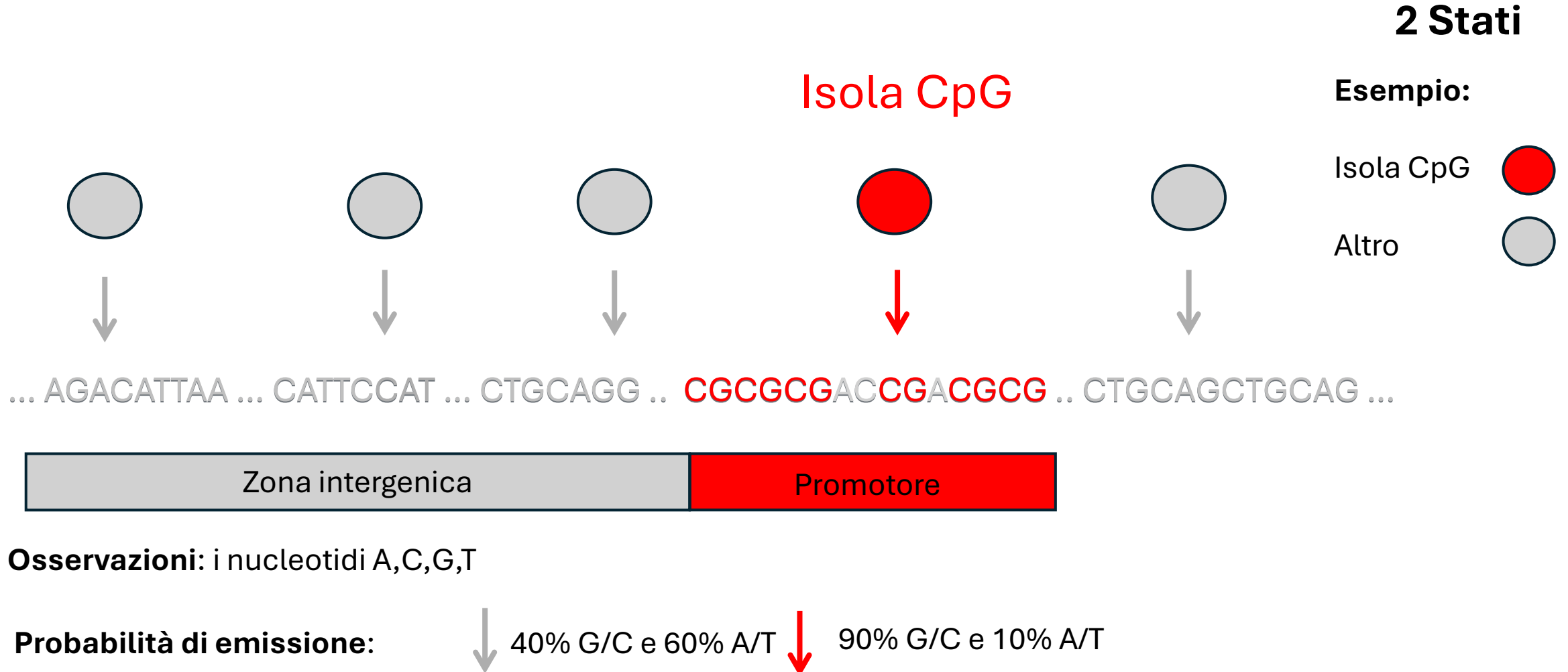
Primo ingrediente: due **stati** (isola CpG e non)



Osservazioni: i nucleotidi A,C,G,T

Gli ingredienti di un Hidden Markov Model

Secondo ingrediente: i due **stati** (nascosti) generano le osservazioni attraverso delle **probabilità di emissione**



Gli ingredienti di un **Hidden** Markov Model

Secondo ingrediente: i due **stati** (**nascosti**) generano le osservazioni attraverso delle **probabilità di emissione**

NOI OSSERVIAMO I DATI, MA NON DIRETTAMENTE GLI STATI!

2 Stati

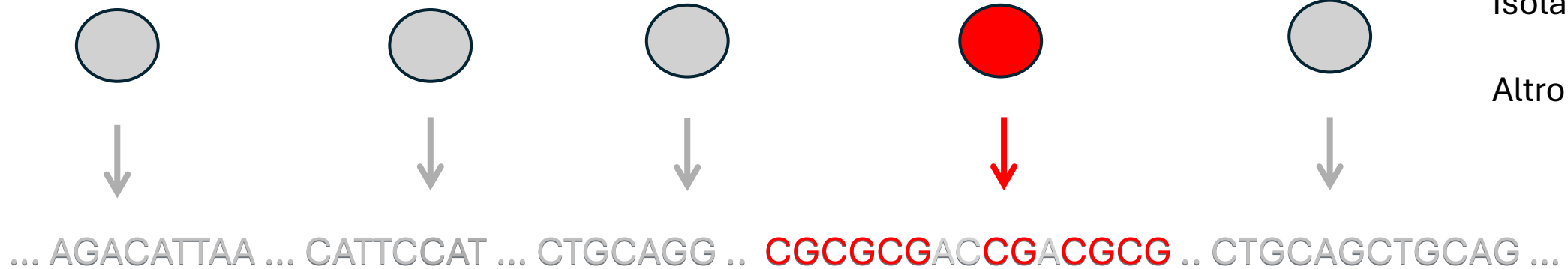
Isola CpG

Esempio:

Isola CpG



Altro



Osservazioni: i nucleotidi A,C,G,T

Probabilità di emissione:



Gli ingredienti di un **Hidden** Markov Model

Secondo ingrediente: i due **stati** (**nascosti**) generano le osservazioni attraverso delle **probabilità di emissione**

**NOI OSSERVIAMO I DATI,
MA NON DIRETTAMENTE
GLI STATI!**

2 Stati

Esempio:

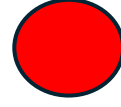
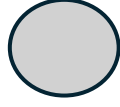
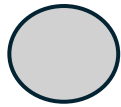
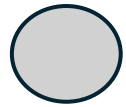
Isola CpG



Altro



Isola CpG



... AGACATTAA ... CATTCCAT ... CTGCAGG .. **CGCGCGACCGACGCG** .. CTGCAGCTGCAG ...

Zona intergenica

Promotore

Osservazioni: i nucleotidi A,C,G,T

Probabilità di emissione:



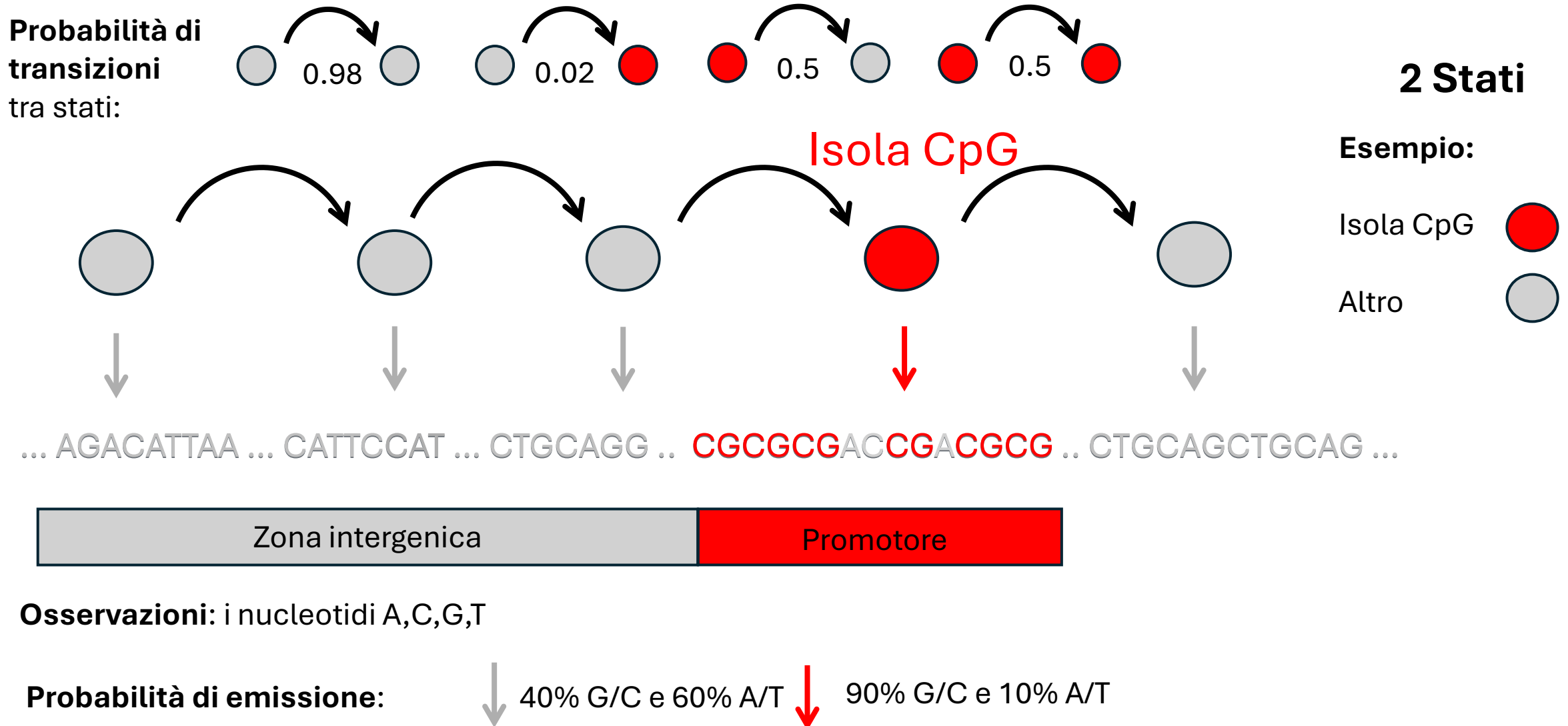
40% G/C e 60% A/T



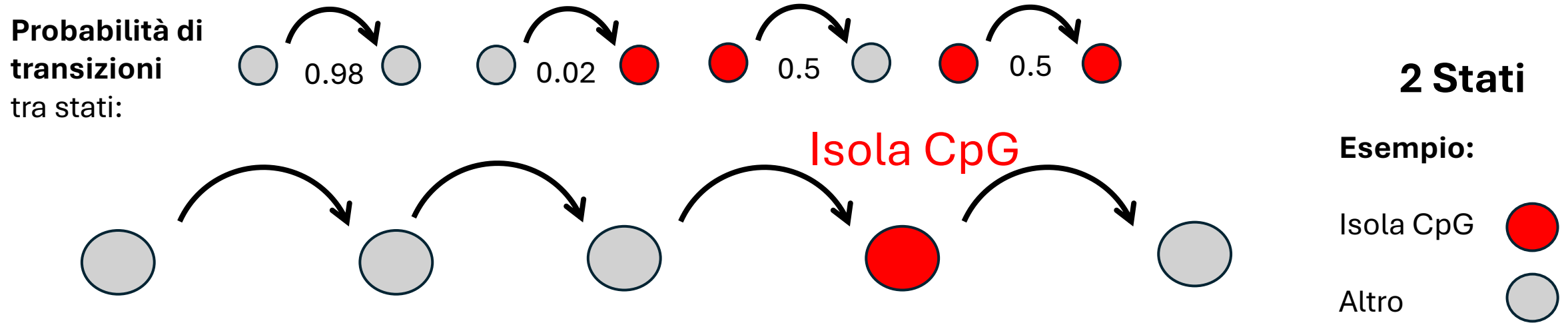
90% G/C e 10% A/T

Gli ingredienti di un Hidden Markov Model

Terzo ingrediente: gli **stati** si alternano seguendo delle **probabilità di transizione**



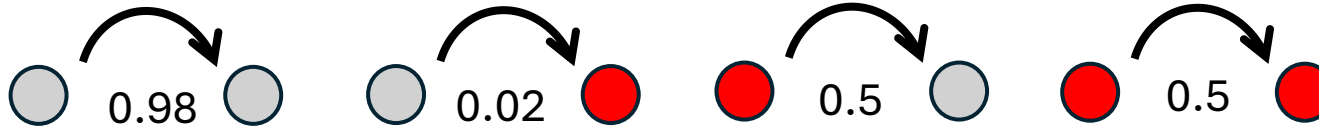
Gli ingredienti di un Hidden **Markov** Model



Non vi dimenticate di me! Come mai “**Markov**”?

Gli ingredienti di un Hidden **Markov** Model

Probabilità di
transizioni
tra stati:



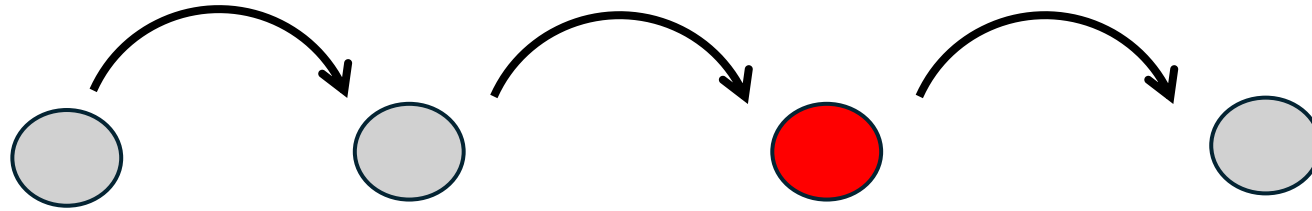
2 Stati

Esempio:

Isola CpG



Altro



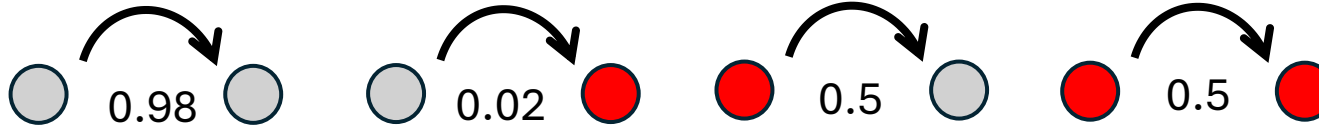
Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.



Gli ingredienti di un Hidden **Markov** Model

Probabilità di
transizioni
tra stati:



2 Stati

Esempio:

Isola CpG



Altro

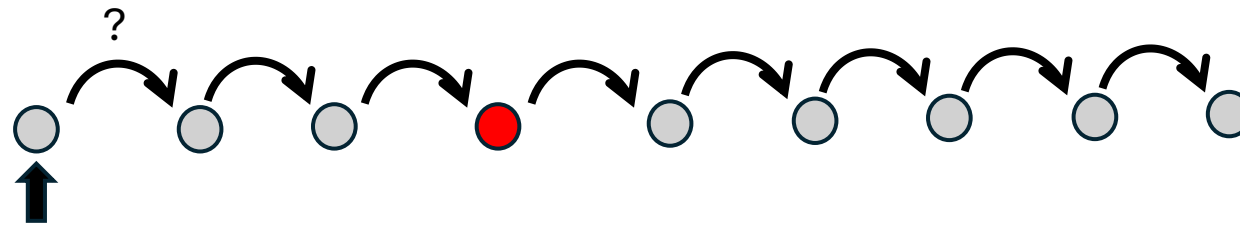


Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.

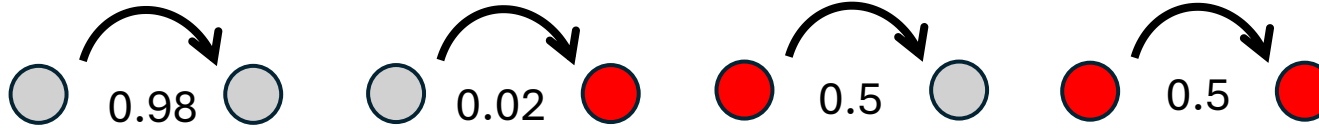
Ad esempio:

La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 



Gli ingredienti di un Hidden **Markov** Model

Probabilità di
transizioni
tra stati:



2 Stati

Esempio:

Isola CpG



Altro

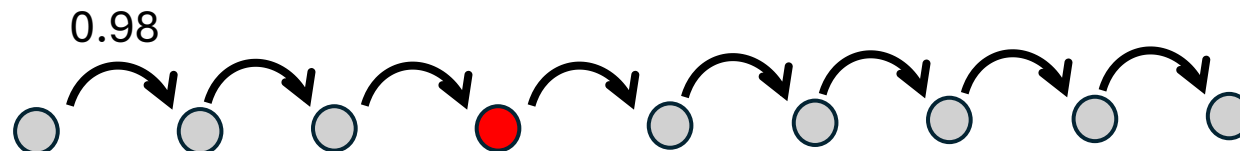


Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.

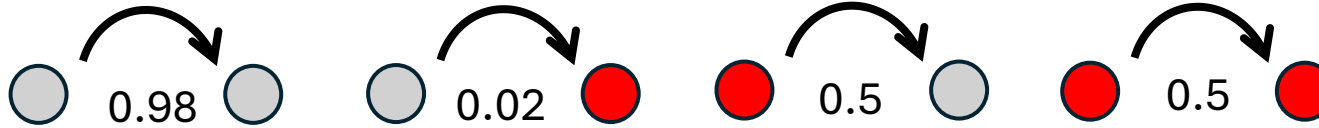
Ad esempio:

La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 



Gli ingredienti di un Hidden **Markov** Model

Probabilità di
transizioni
tra stati:



2 Stati

Esempio:

Isola CpG



Altro

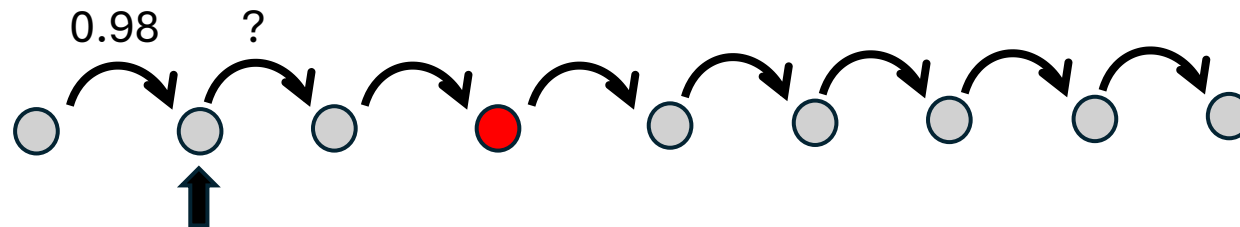


Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.

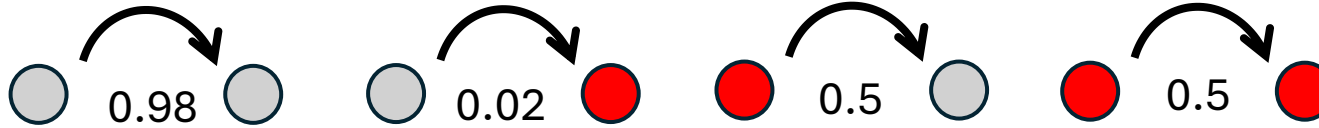
Ad esempio:

La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 



Gli ingredienti di un Hidden **Markov** Model

Probabilità di
transizioni
tra stati:



2 Stati

Esempio:

Isola CpG



Altro

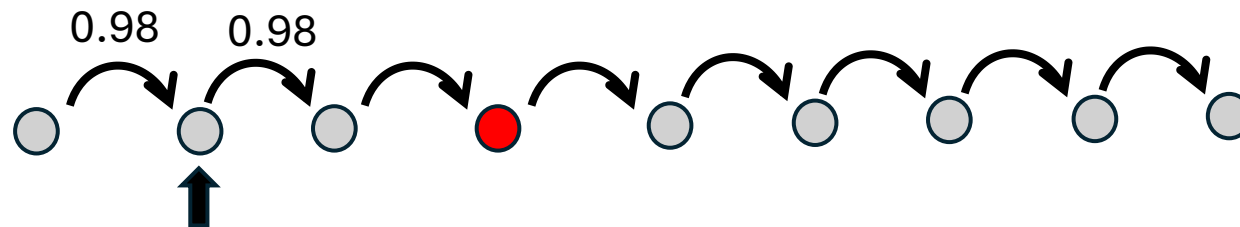


Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.

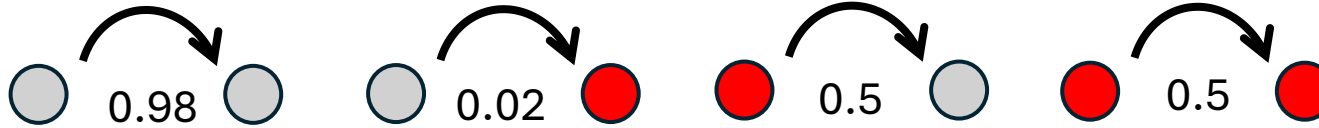
Ad esempio:

La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 



Gli ingredienti di un Hidden **Markov** Model

Probabilità di
transizioni
tra stati:



2 Stati

Esempio:

Isola CpG



Altro

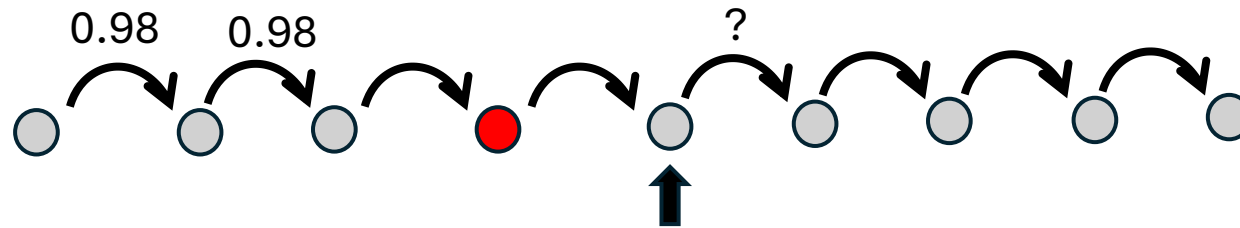


Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.

Ad esempio:

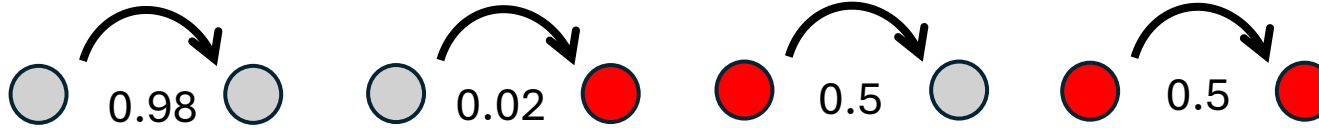
La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 



Gli ingredienti di un Hidden **Markov** Model

Terzo ingrediente: gli **stati** si alternano seguendo delle **probabilità di transizione**

Probabilità di
transizioni
tra stati:



2 Stati

Esempio:

Isola CpG



Altro

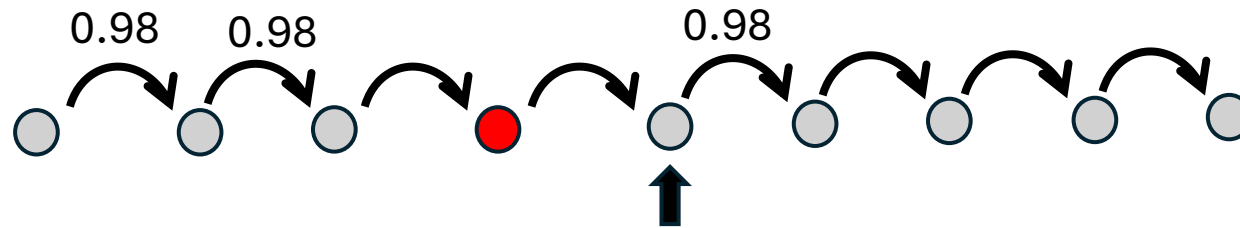


Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.

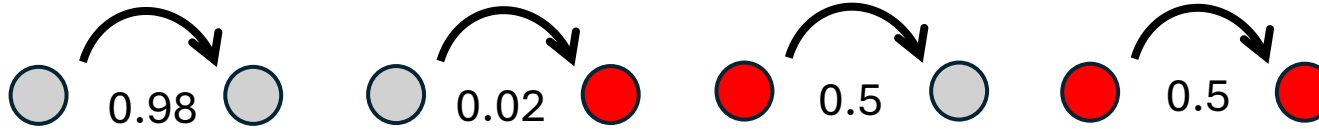
Ad esempio:

La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 



Gli ingredienti di un Hidden **Markov** Model

Probabilità di
transizioni
tra stati:



Esempio:

Isola CpG

Altro



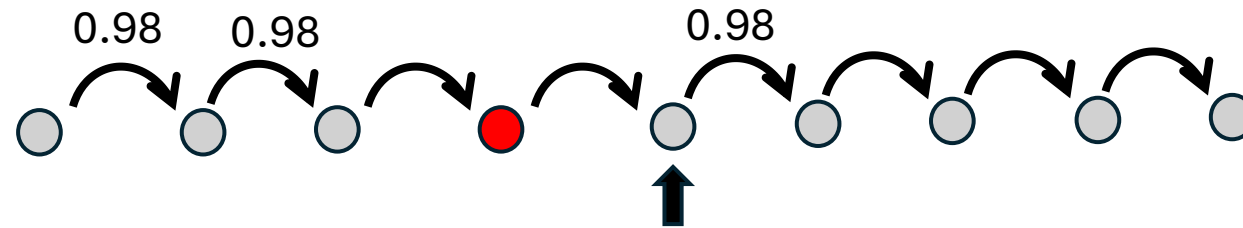
La catena non ha memoria!

Proprietà di una catena Markoviana:

La probabilità di cambiare stato, dipende solamente dallo stato attuale.

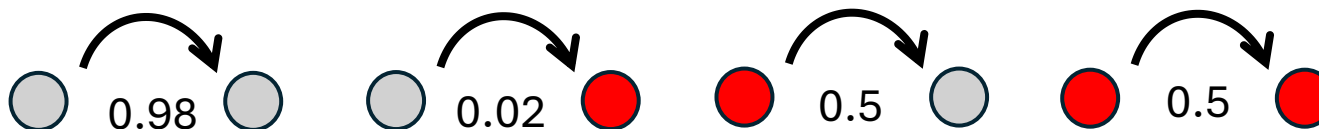
Ad esempio:

La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 



Rappresentazione di un Hidden Markov Model

Probabilità di
transizioni
tra stati:



Esempio:

Isola CpG

Altro



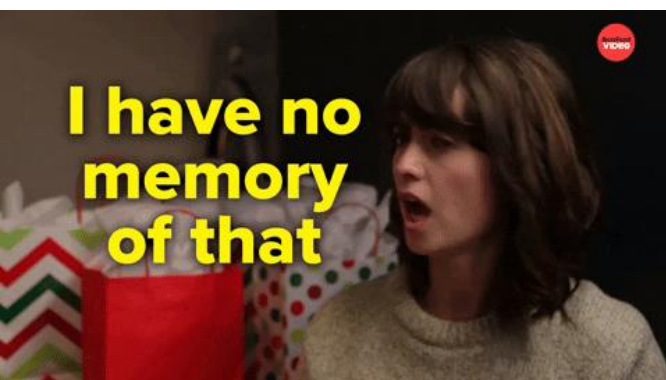
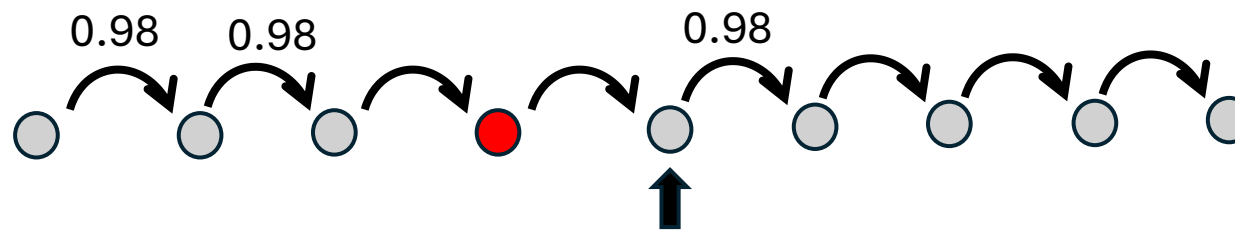
La catena non ha memoria!

Proprietà di una catena Markoviana:

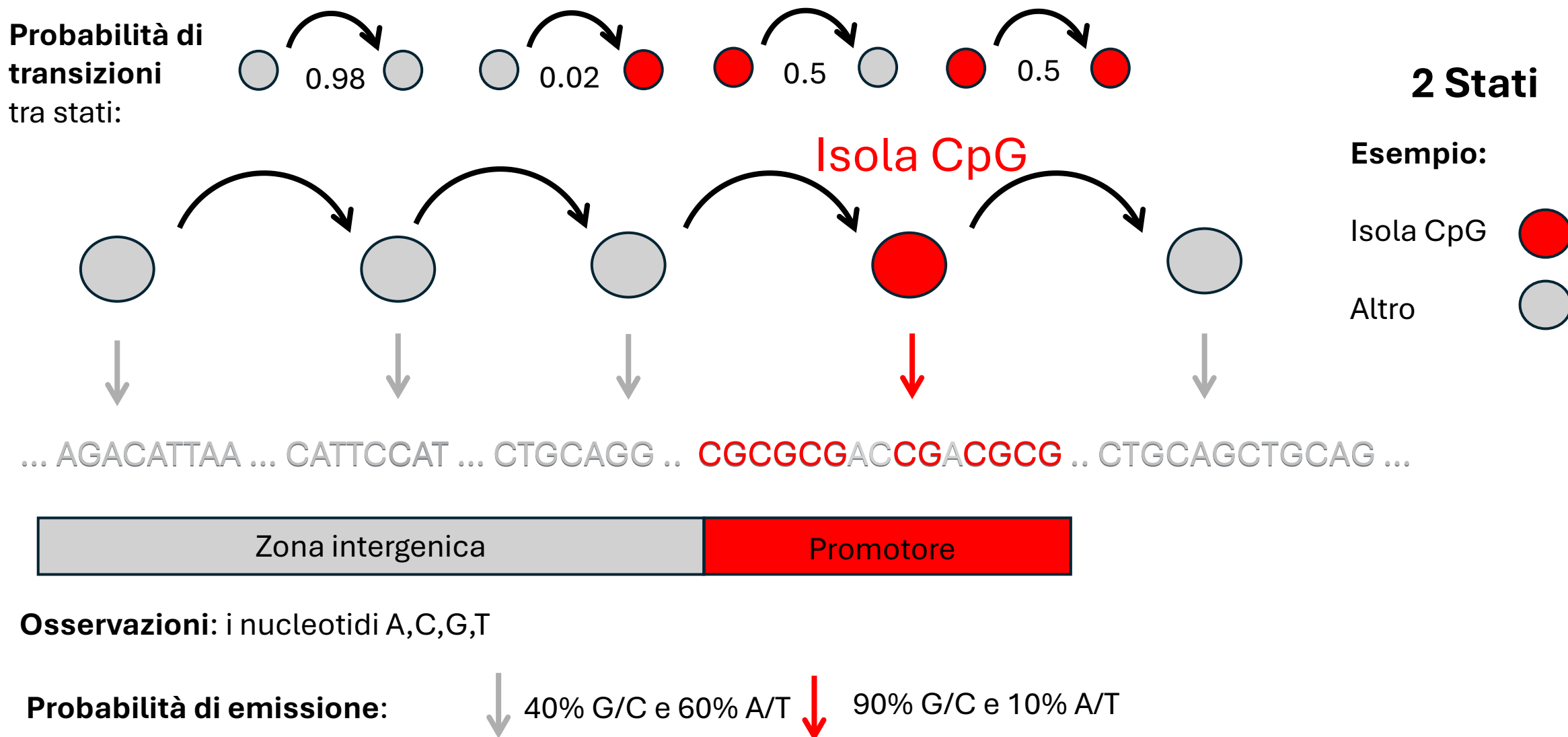
La probabilità di cambiare stato, dipende solamente dallo stato attuale.

Ad esempio:

La probabilità di andare da  a  o  è sempre la stessa per qualsiasi 

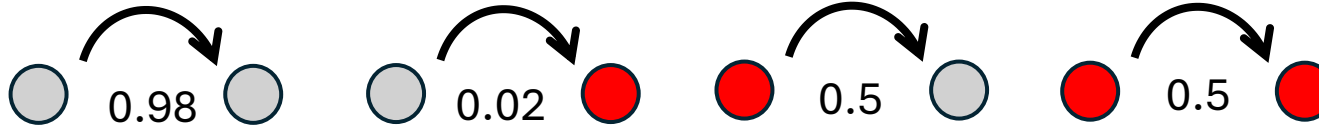


Rappresentazione di un Hidden Markov Model



Rappresentazione di un Hidden Markov Model

Probabilità di
transizioni
tra stati:



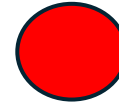
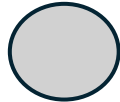
2 Stati

Esempio:

Isola CpG

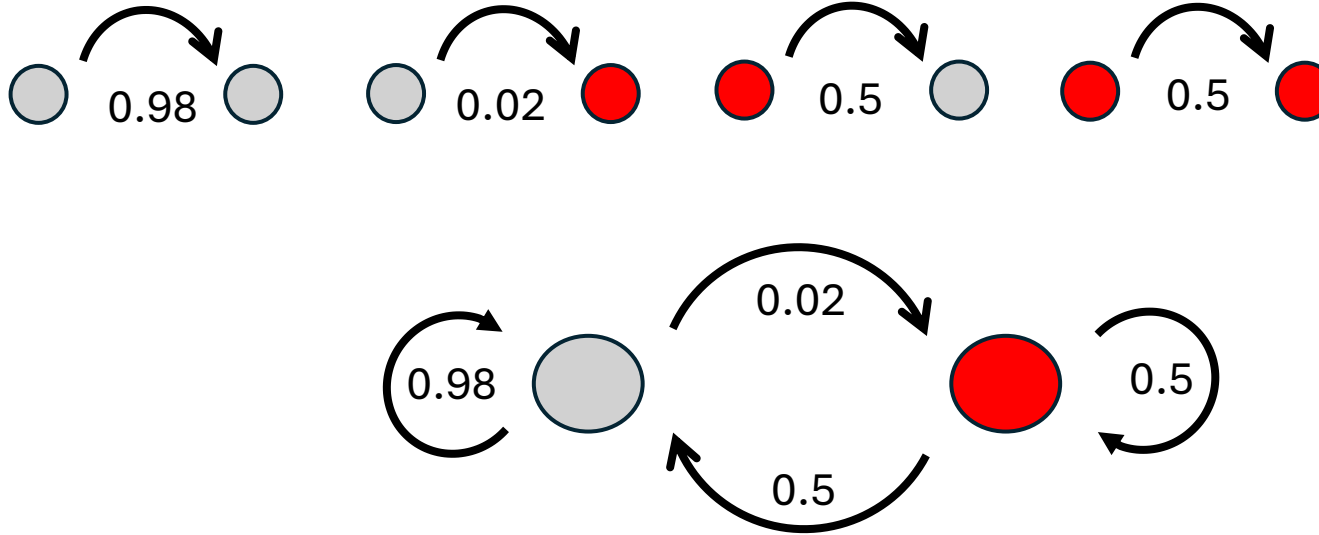


Altro



Rappresentazione di un Hidden Markov Model

Probabilità di
transizioni
tra stati:



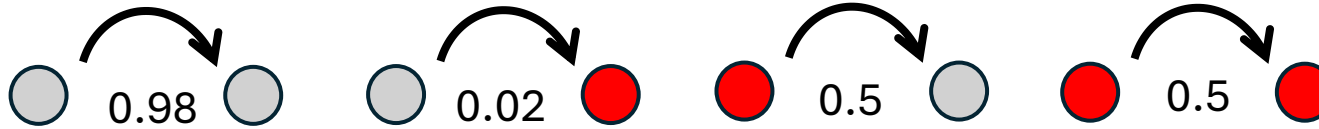
2 Stati

Esempio:

Isola CpG 
Altro 

Rappresentazione di un Hidden Markov Model

Probabilità di transizioni
tra stati:



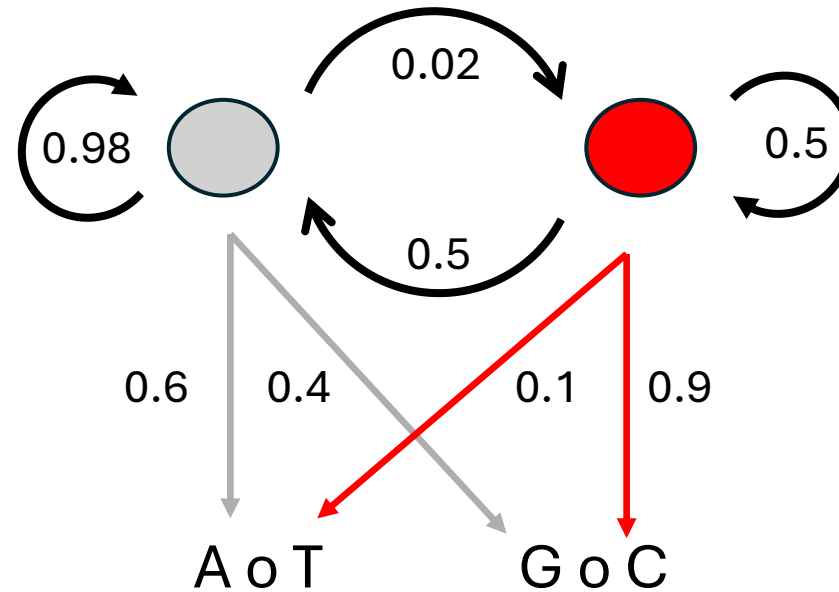
2 Stati

Esempio:

Isola CpG



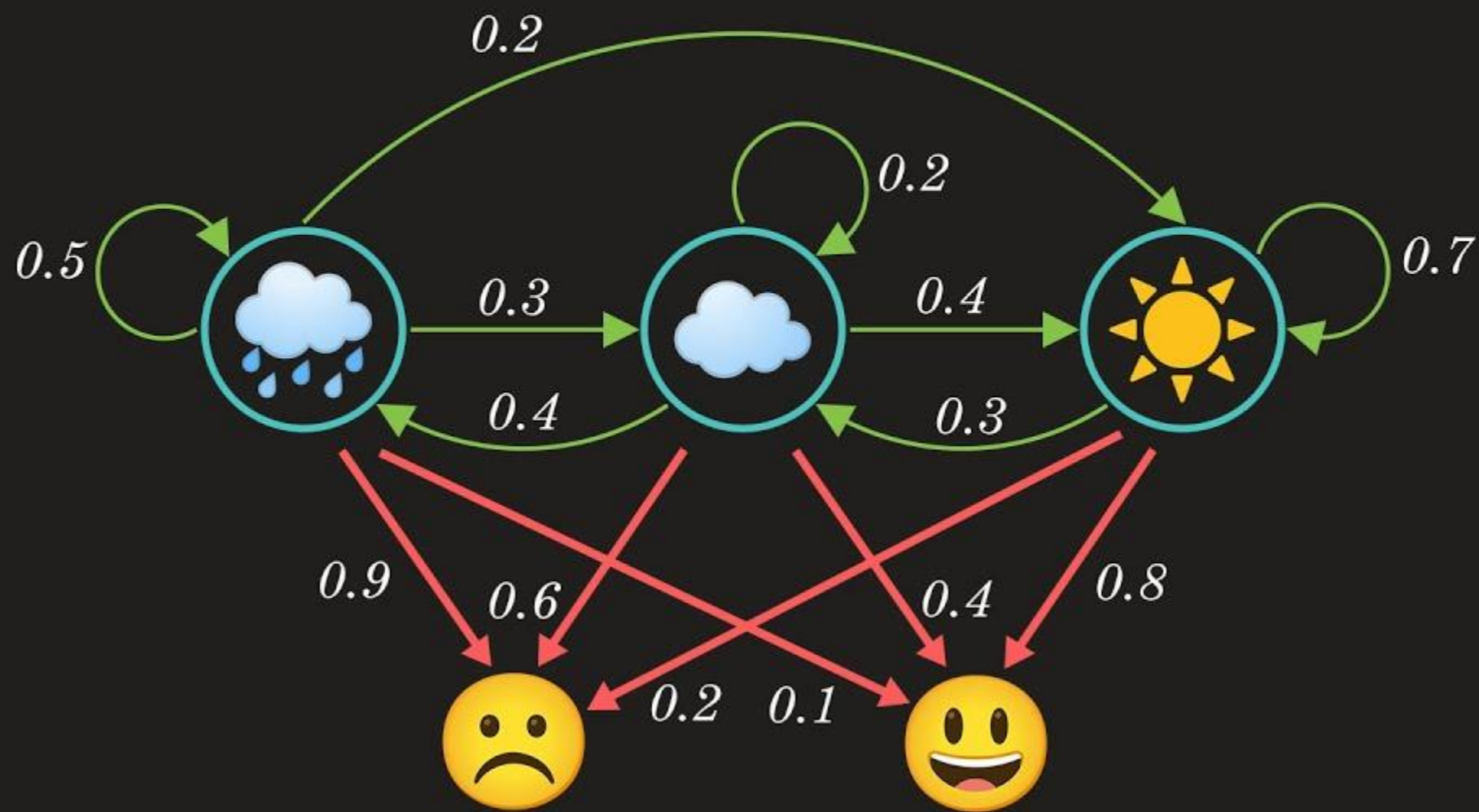
Altro



Probabilità di emissione:



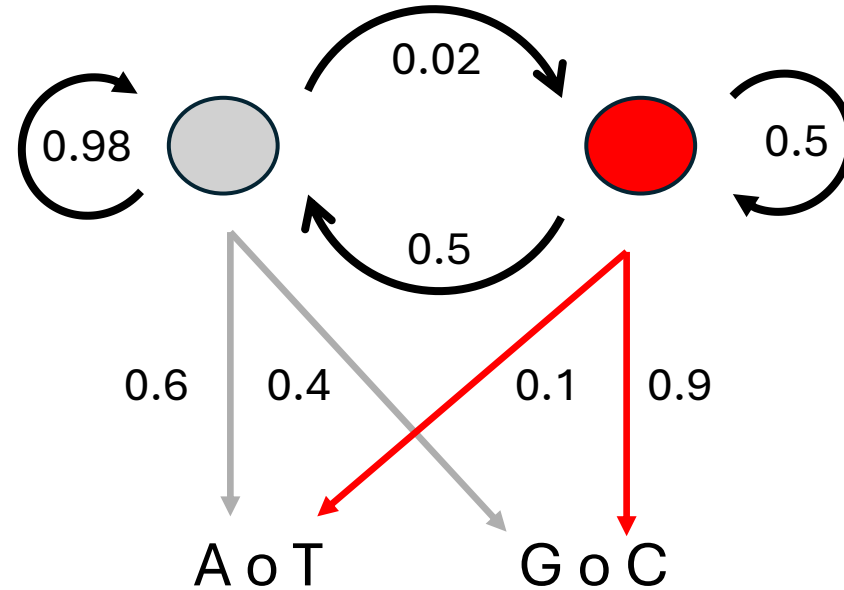
Hidden Markov Model



Aggiungiamo un po' di formalismo

non CpG

Isola CpG



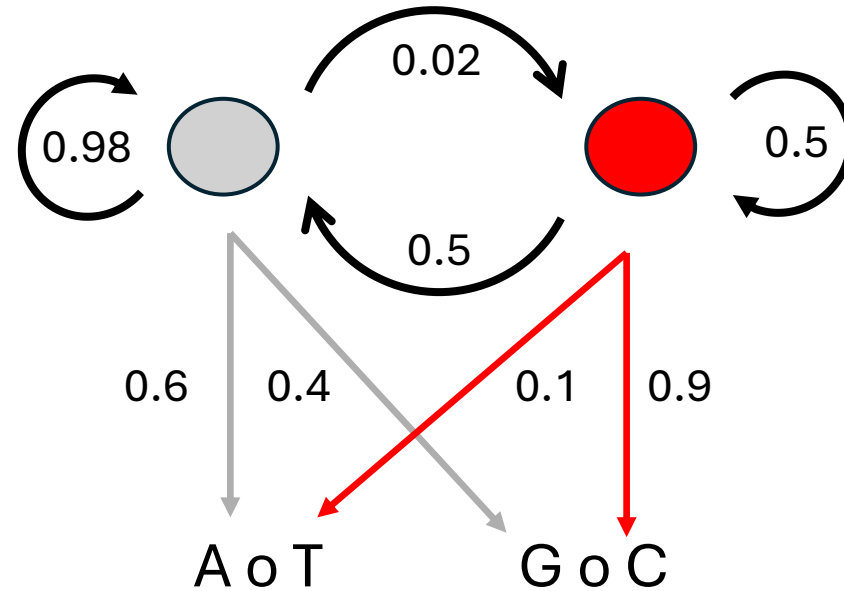
**Probabilità di
transizioni
tra stati:**



Aggiungiamo un po' di formalismo

non CpG

Isola CpG



Probabilità di
transizioni
tra stati:



+



= 1



+

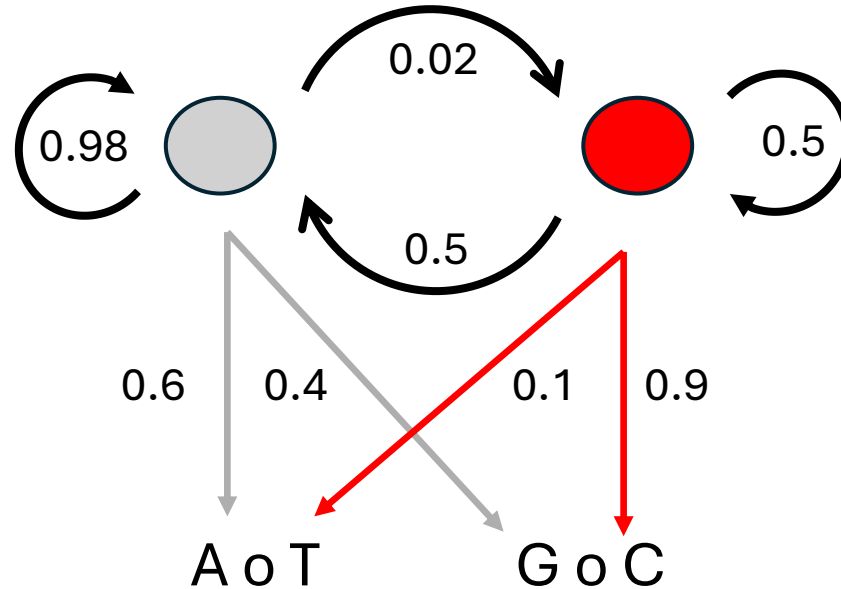


= 1

Aggiungiamo un po' di formalismo

non CpG

Isola CpG



Le **Probabilità di transizione** tra stati sono le probabilità **condizionate** di vedere uno stato in position ***i+1*** dato uno stato in posizione ***i***

Condizionate al fatto che in *i* sia: 



+



= 1

Condizionate al fatto che in *i* sia: 



+



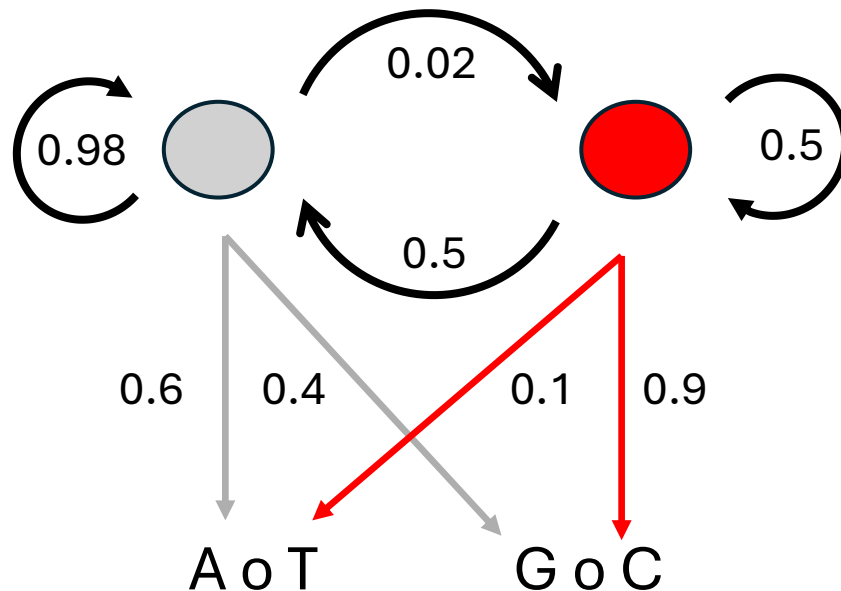
= 1

Probabilità di transizioni tra stati:

Aggiungiamo un po' di formalismo

non CpG

Isola CpG



Le **Probabilità di transizione** tra stati sono le probabilità **condizionate** di vedere uno stato in position ***i+1*** dato uno stato in posizione ***i***

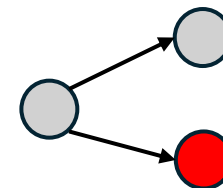
Condizionate al fatto che in *i* sia:



+



= 1



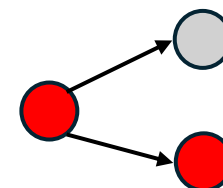
Condizionate al fatto che in *i* sia:



+



= 1

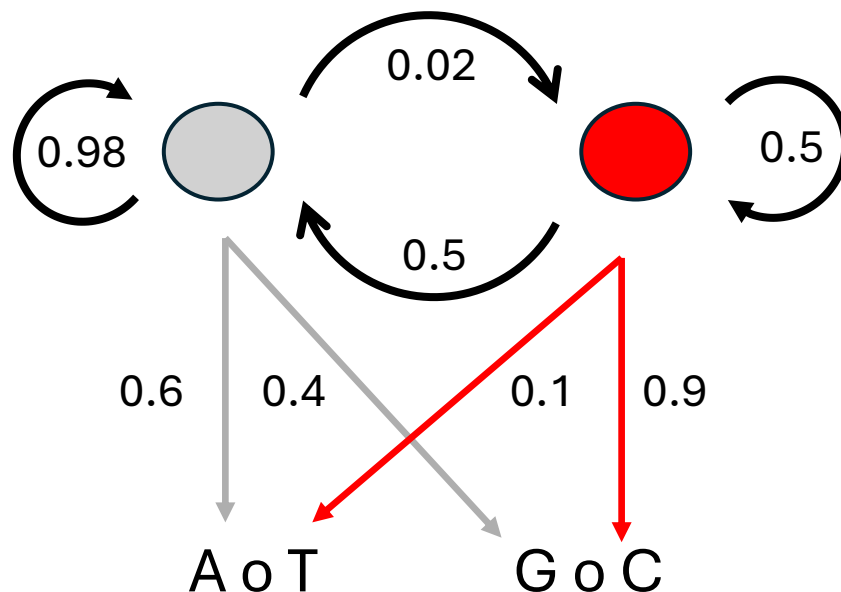


Probabilità di transizioni tra stati:

Aggiungiamo un po' di formalismo

non CpG

Isola CpG



Le **Probabilità di transizione** tra stati sono le probabilità **condizionate** di vedere uno stato in position ***i+1*** dato uno stato in posizione ***i***

Probabilità di transizioni tra stati:



$\Pr(\text{non CpG in } i+1 \mid \text{non CpG in } i)$



$\Pr(\text{non CpG in } i+1 \mid \text{non CpG in } i)$



$\Pr(\text{CpG in } i+1 \mid \text{non CpG in } i)$

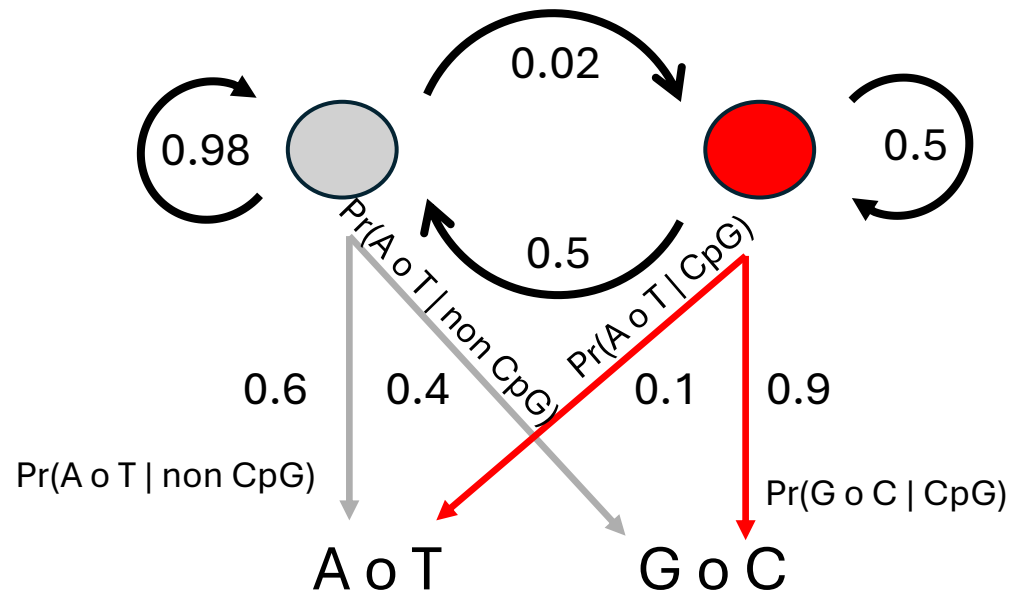


$\Pr(\text{CpG in } i+1 \mid \text{non CpG in } i)$

Aggiungiamo un po' di formalismo

non CpG

Isola CpG



Le **Probabilità di emissione** sono le probabilità **condizionate** di vedere una osservazione dato uno stato in *i* (e solo in *i*, a causa di Markov!)

Condizionate al fatto che in *i* lo stato sia: 

$\Pr(A \circ T | \text{grey circle})$

$\Pr(G \circ C | \text{grey circle})$

Probabilità di emissione:

Condizionate al fatto che in *i* lo stato sia: 

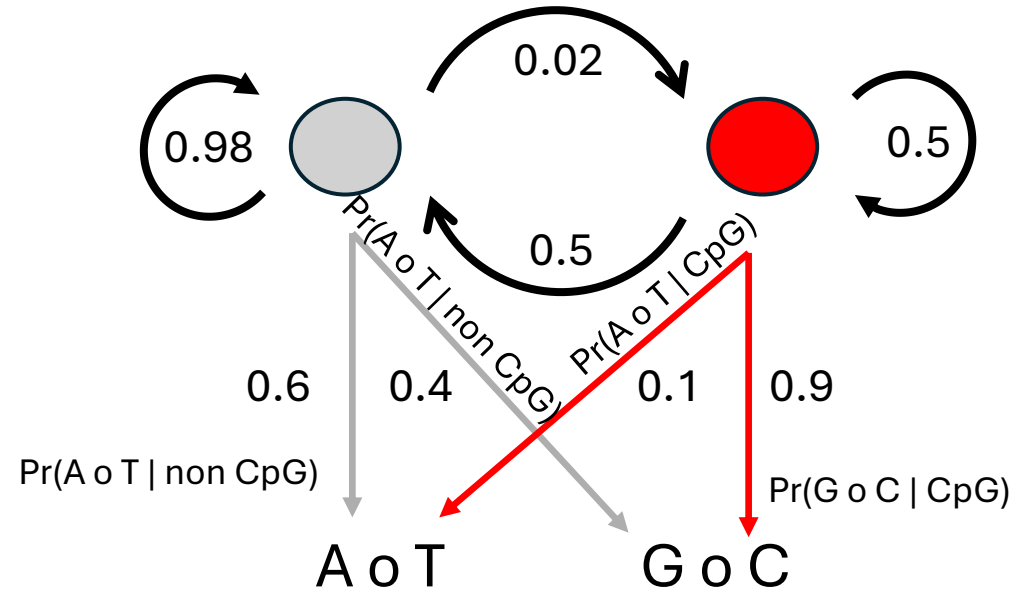
$\Pr(A \circ T | \text{red circle})$

$\Pr(G \circ C | \text{red circle})$

Aggiungiamo un po' di formalismo

non CpG

Isola CpG



Le **Probabilità di emissione** sono le probabilità **condizionate** di vedere una osservazione dato uno stato in *i* (e solo in *i*, a causa di Markov!)

Probabilità di emissione:

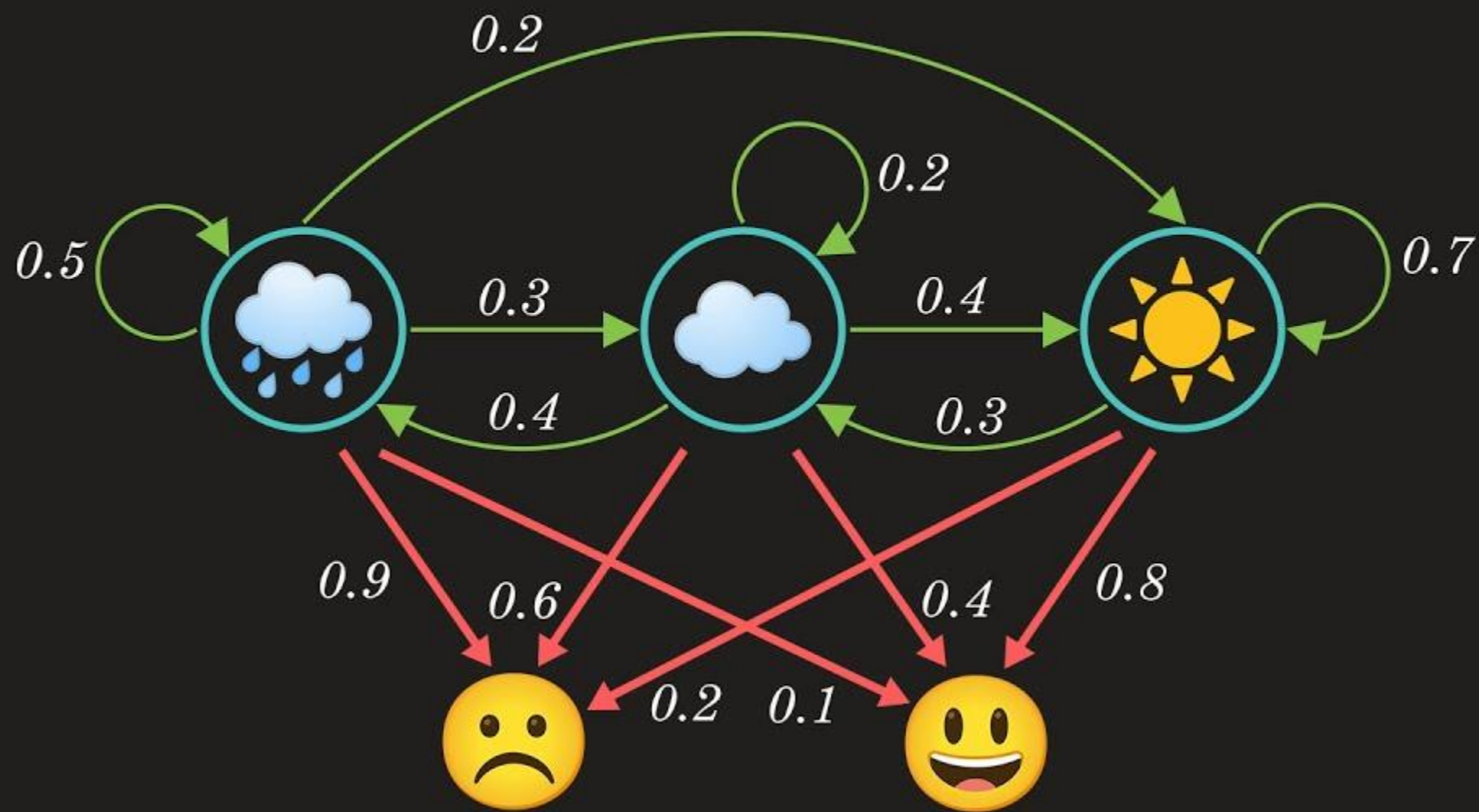
Condizionate al fatto che in *i* lo stato sia: 

$$\Pr(\text{A o T} | \text{grey circle}) + \Pr(\text{G o C} | \text{grey circle}) = 1$$

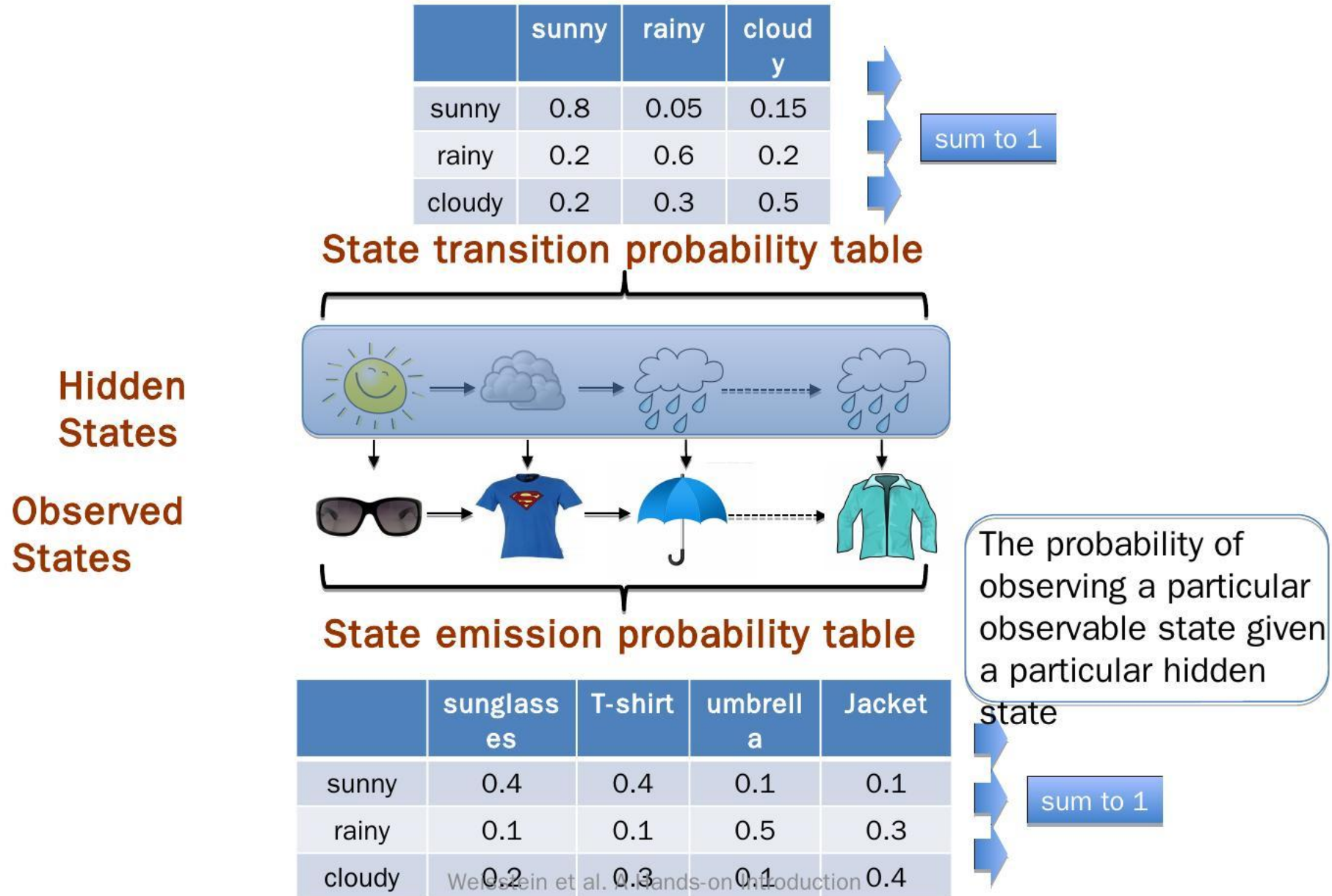
Condizionate al fatto che in *i* lo stato sia: 

$$\Pr(\text{A o T} | \text{red circle}) + \Pr(\text{G o C} | \text{red circle}) = 1$$

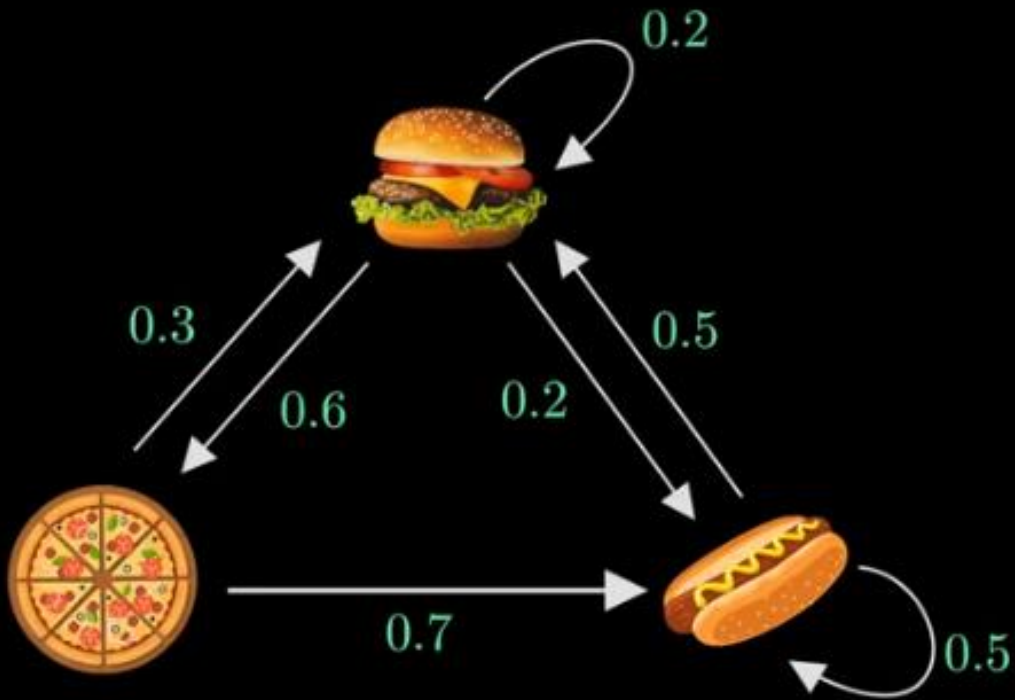
Hidden Markov Model



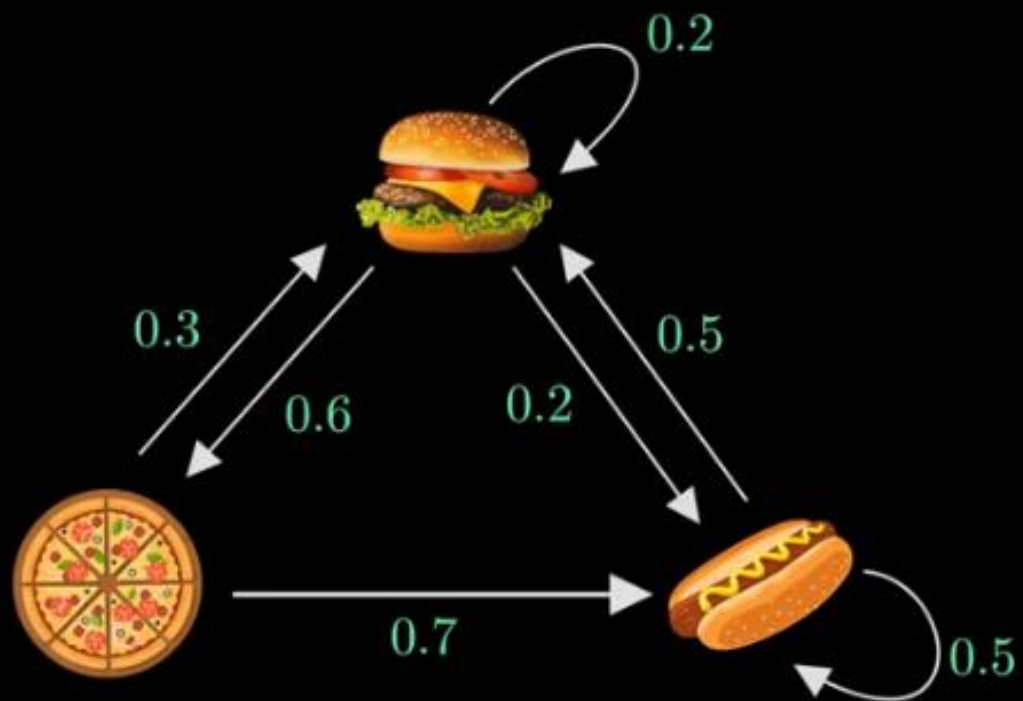
The Hidden Markov Model



Altri modelli di dubbia interpretabilità



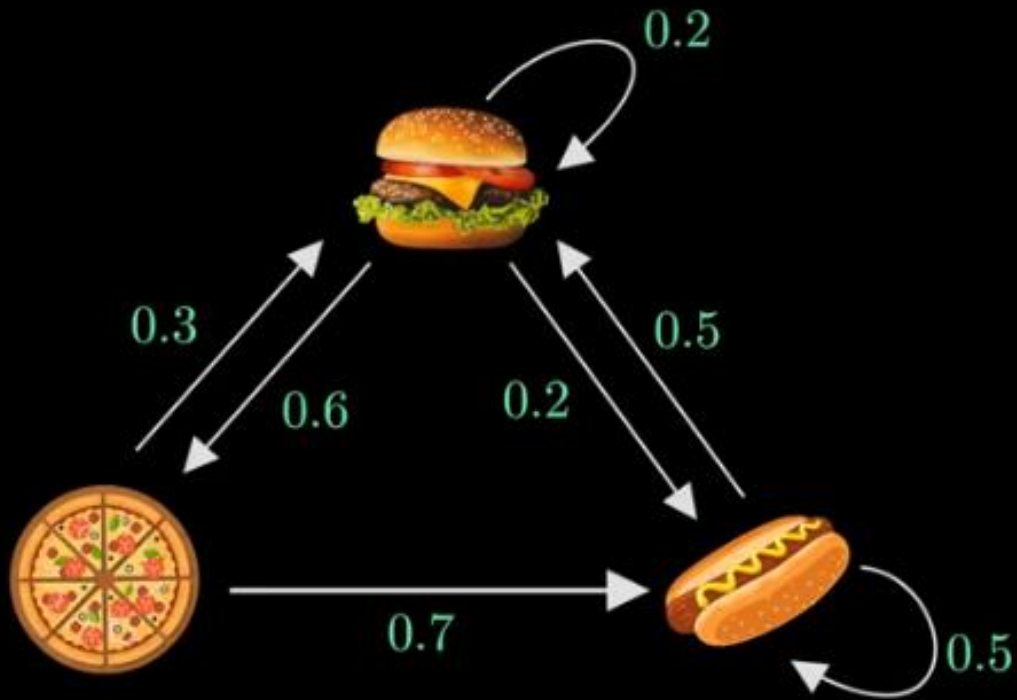
E di dubbio gusto



Random Walk



E di dubbio gusto



After ∞ steps...

$P(\text{Burger})$

0.35191

$P(\text{Pizza})$

0.21245

$P(\text{Hotdog})$

0.43564

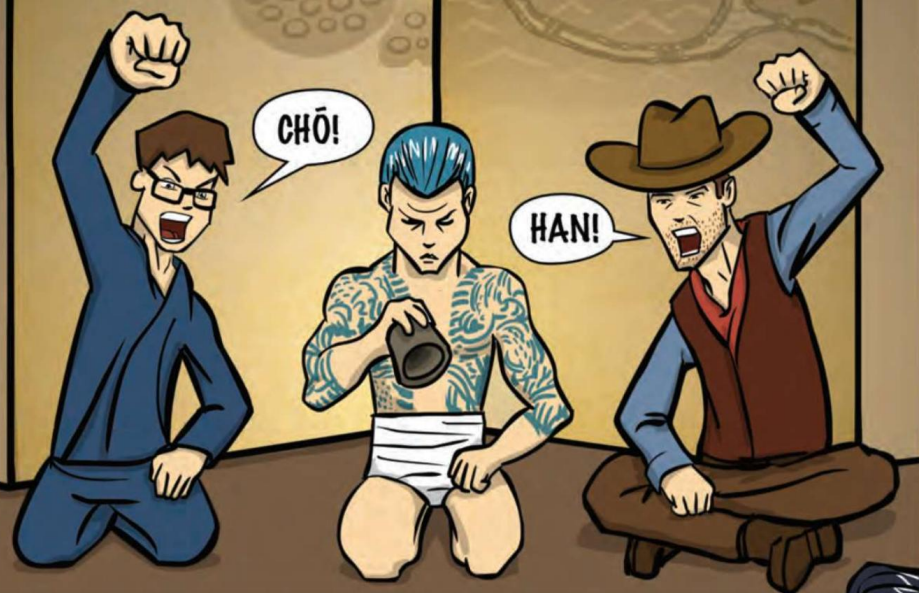
Random Walk

Le probabilità di transizione determinano le frequenze degli stati



WHY HAVE BIOLOGISTS STILL NOT DEVELOPED AN HIV VACCINE?

Hidden Markov Models



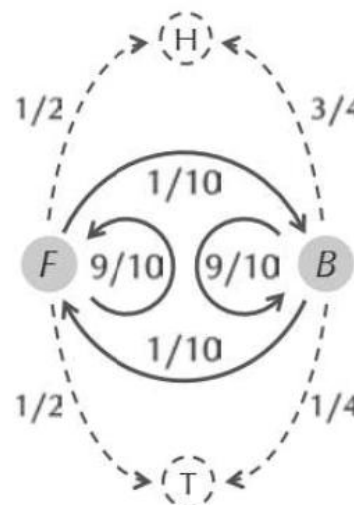
Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Emission

	H	T
F	$1/2$	$1/2$
B	$3/4$	$1/4$

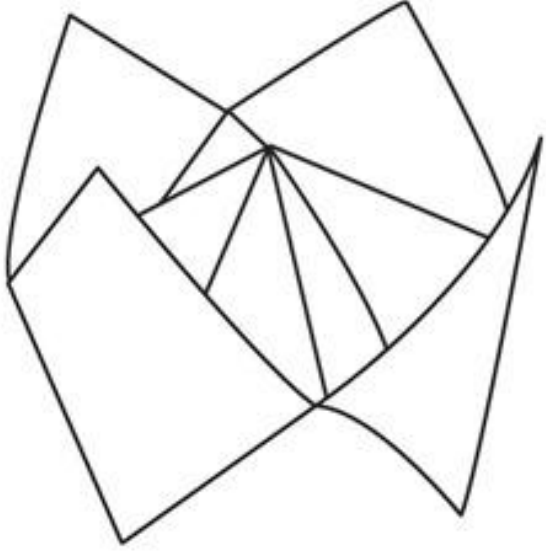


Transmission

	F	B
F	$9/10$	$1/10$
B	$1/10$	$9/10$



Momento Art Attack!



Create le vostre catene di Markov!



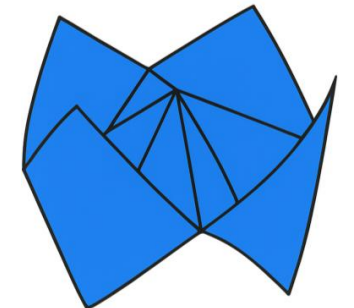
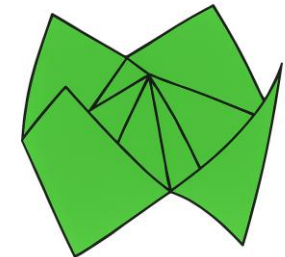
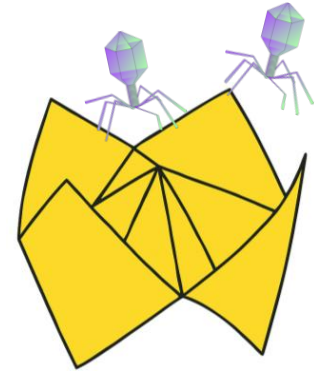
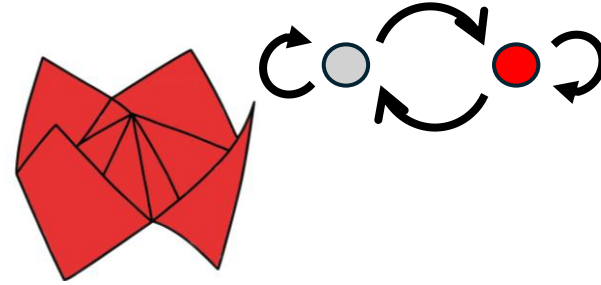
Create le vostre catene di Markov!

- Dividetevi in 3 gruppi.
- Ad ogni gruppo saranno date 2 ranocchiette, ognuna rappresentante uno **stato**.
- Una persona per gruppo sarà lo Yakuza dealer (senza togliervi la maglietta) e alternerà (ogni tanto) le ranocchiette (rappresentando le **transizioni**).
- Una persona del gruppo a turno sceglierà una finestrella della ranocchietta, rappresentando una **emissione**.
- Effettuate circa 20 emissioni e scrivetele
- **Per lo Yakuza dealer:**
 - nascondete le «transizioni» tra stati dietro la schiena o sotto ad un banco per non farle vedere al resto del gruppo (gli stati sono Hidden, nascosti!)
 - Non fate barare il gruppo sbirciando tutte le finestrelle (sarete poi comunque assegnati ad un altro gruppo per risolvere le catene!)



Le quattro ranocchiette rappresentano:

- CpG islands e non (ranocchiette rosse)
- Contenuto GC sbilanciato, ad esempio per integrazione di un virus come Herpesvirus 6, Epstein-Barr o vari fagi nel genoma) (ranocchiette gialle)
- Zone ripetute nel genoma (AT-rich) verso non-ripetute (ranocchiette verdi)
- Aplotipi Neandertal e non per gli studenti da casa (ranocchietta blu). Per voi la sequenza è già emessa e dovete solo cercare di trovare il modo di inferire gli stati:



HHHNNHHHHNHHHHNNHHHHHHNHH

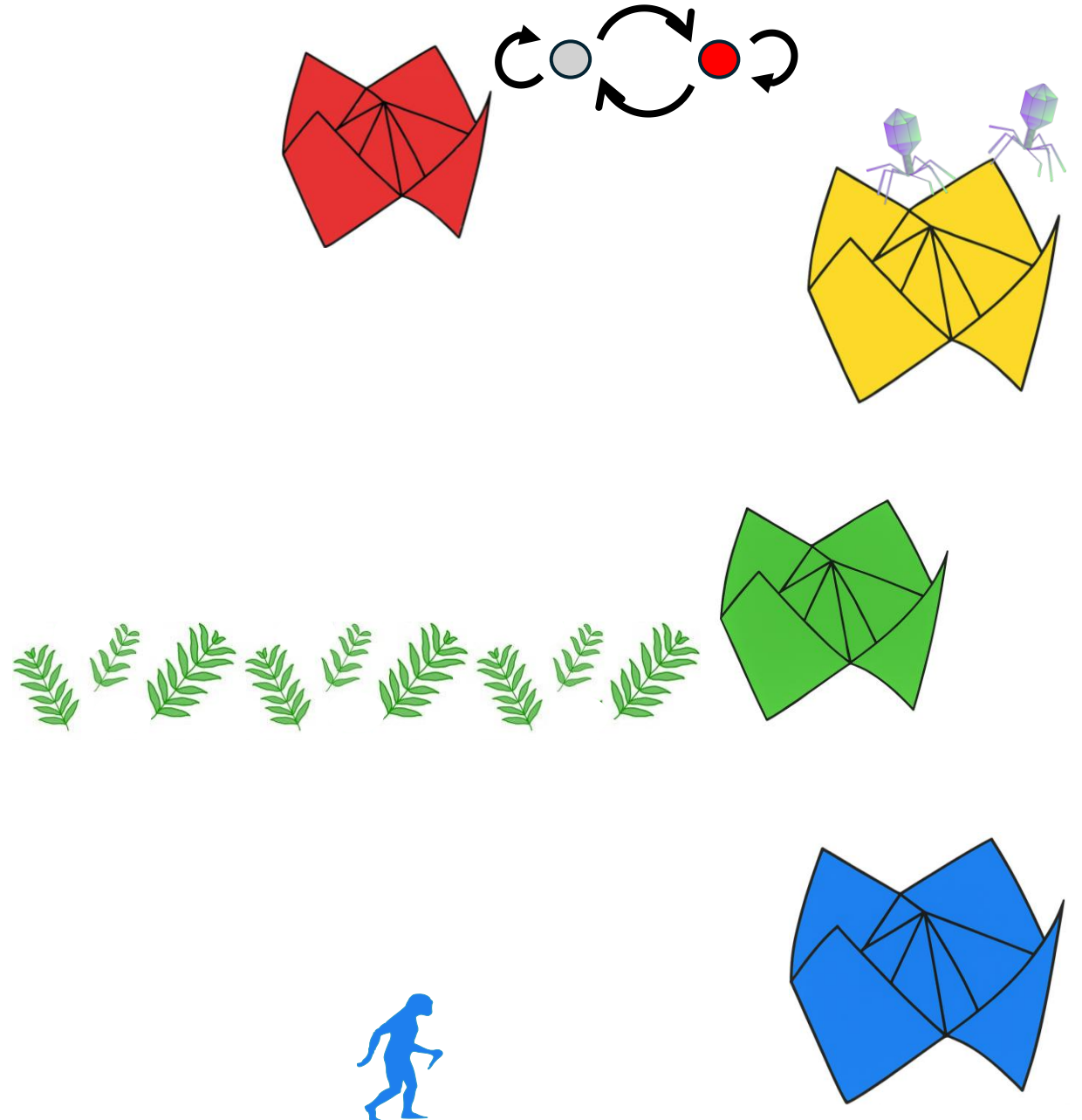


Provate ora ad:

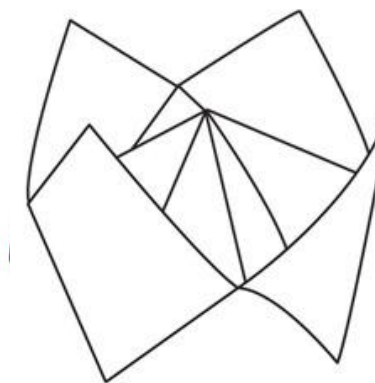
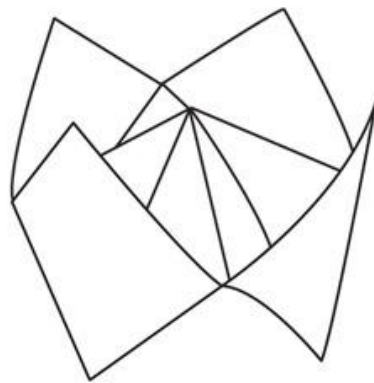
1) inferire la successione degli stati che hanno generato le vostre osservazioni

2) stimare probabilità di emissione e transizione

Come fareste?

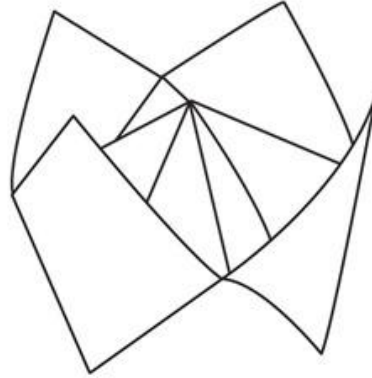


Stato 1

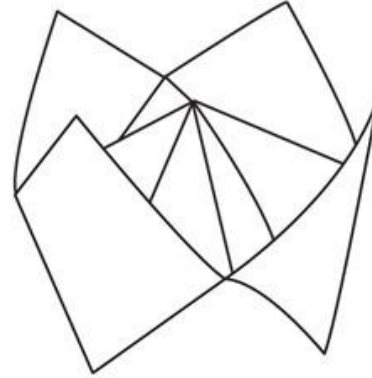


Stato 2

Stato 1

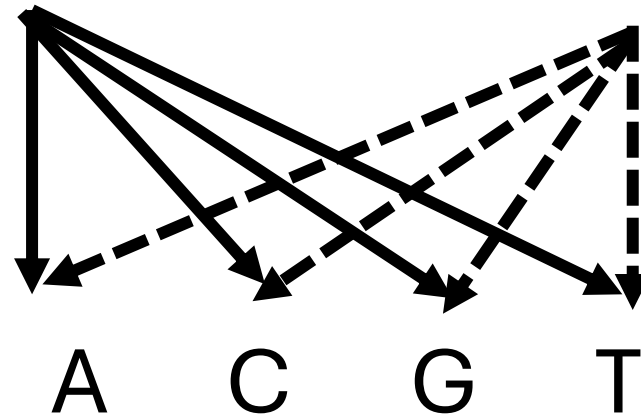


Stato 2



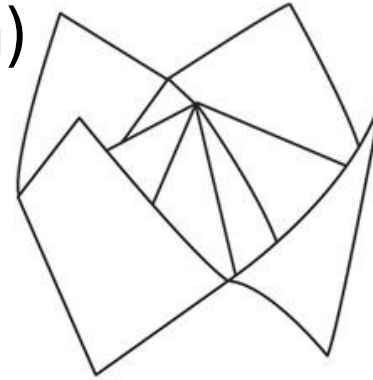
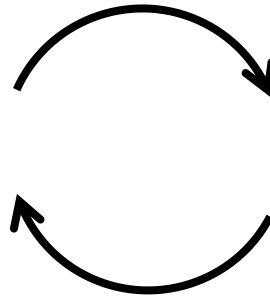
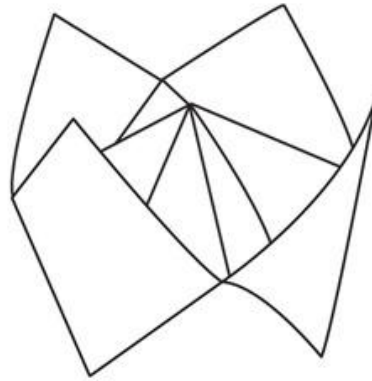
Probabilità di
emissione
differenti

Osservazioni



Probabilità di
transizione (scelte
dallo Yakuza)

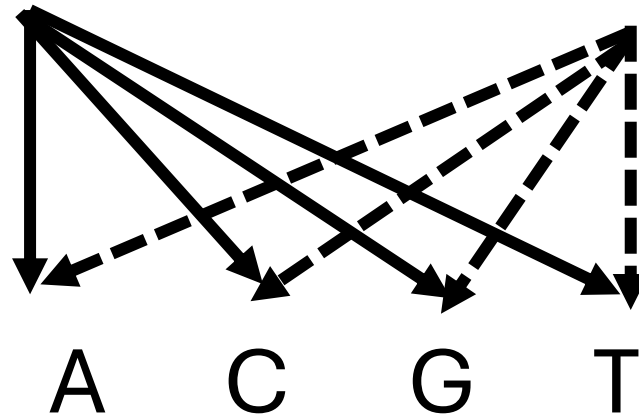
Stato 1



Stato 2

Probabilità di
emissione
differenti

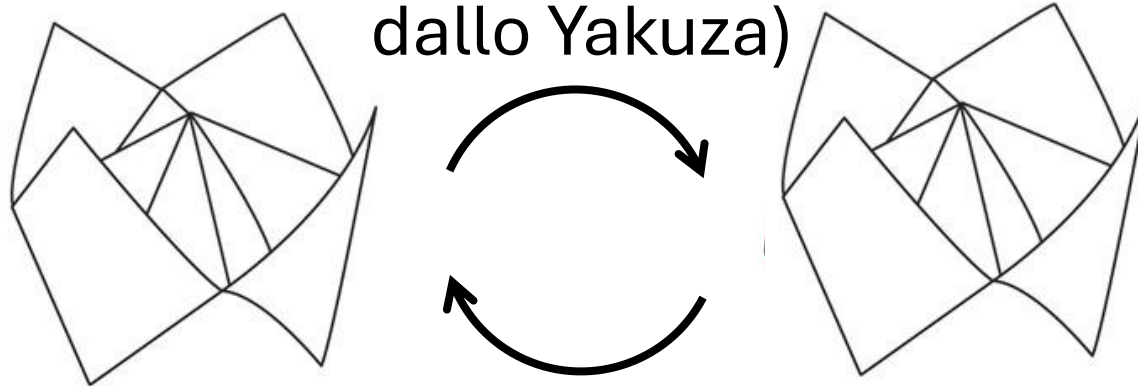
Osservazioni





Stato 1

Probabilità di
transizione (scelte
dallo Yakuza)

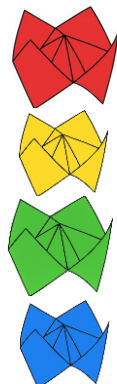


Stato 2



Probabilità di
emissione
differenti

Osservazioni



CG

C o G

AT

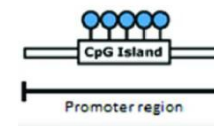
N

altri dinucleotidi

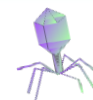
A o T

altri dinucleotidi

N



Isole CpG



Integrazione virale



DNA ripetuto

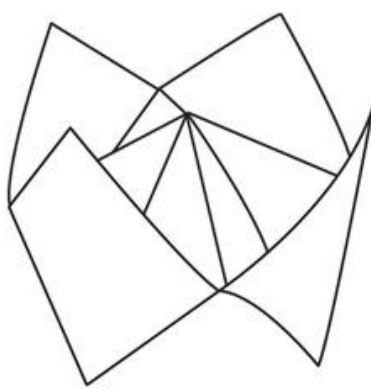


Introgresione Neandertal

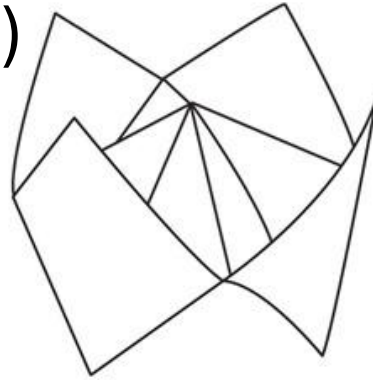
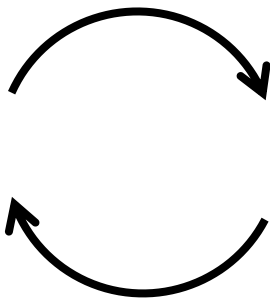
Fase 1: pensare a come semplificare ragionevolmente il problema



Stato 1



Probabilità di transizione (scelte dallo Yakuza)

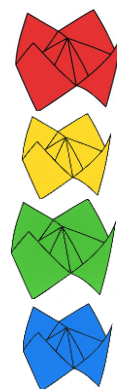


Stato 2



Probabilità di emissione differenti

Osservazioni



CG

C o G

AT

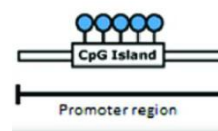
N

altri dinucleotidi

A o T

altri dinucleotidi

N



Isole CpG



Integrazione virale



DNA ripetuto



Integressione Neandertal

Fase 1: pensare a come semplificare ragionevolmente il problema

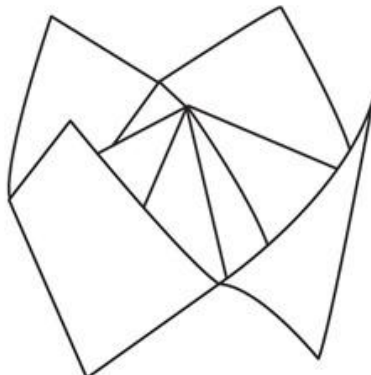
Questo vale per ogni modello, non solo per le HMM. Ma che vuol dire «ragionevolmente»??



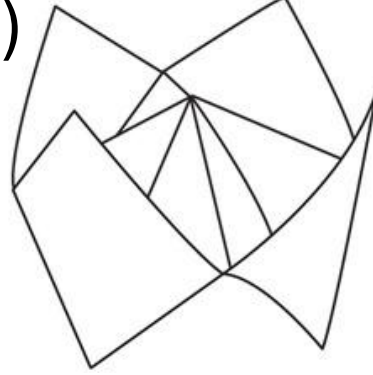
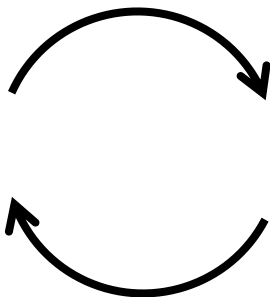
Eh, dipende dai casi!!
Semplifica piu' che puoi, ma non troppo che perda di significato



Stato 1



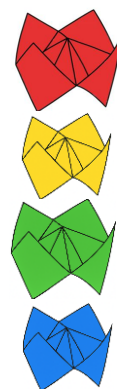
Probabilita' di transizione (scelte dallo Yakuza)



Stato 2

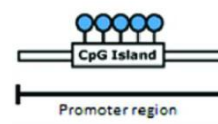
Probabilita' di emissione differenti

Osservazioni



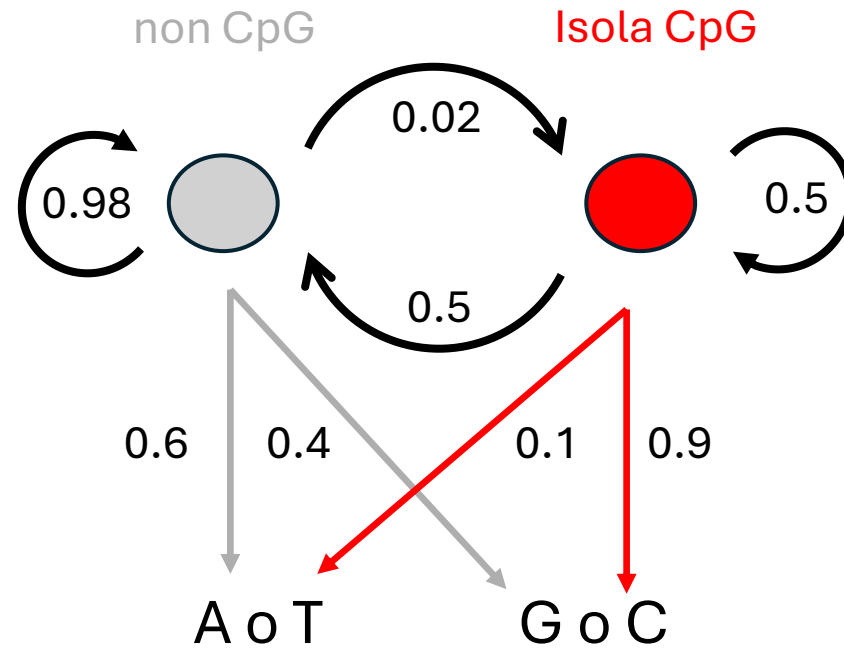
CG
C o G
AT
N

altri dinucleotidi
A o T
altri dinucleotidi
N



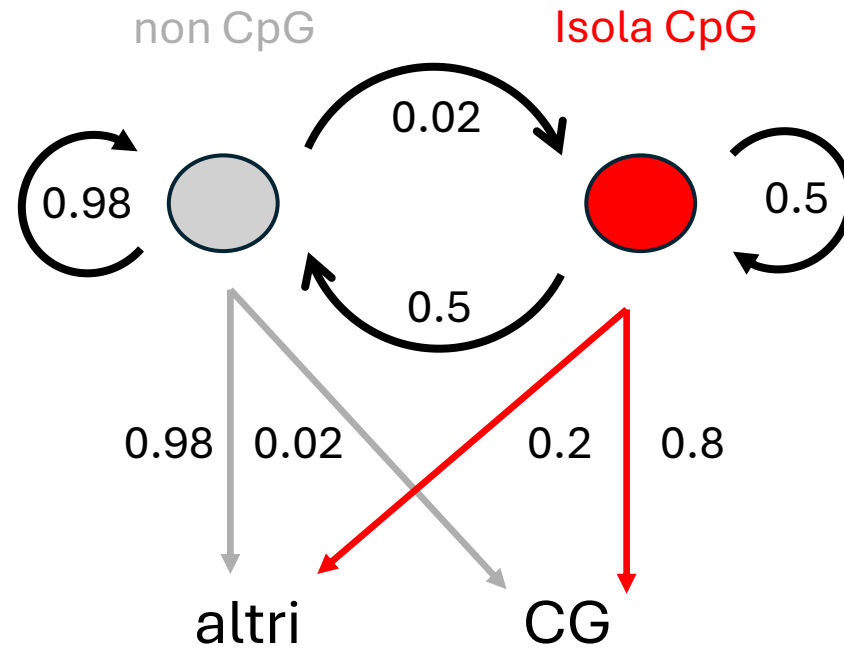
- Isole CpG
- Integrazione virale
- DNA ripetuto
- Introgresione Neandertal

Modello di Isole CpG a singolo nucleotide



Modello di Isole CpG a dinucleotidi

(un dinucleotide alla volta)



Cons:

- Non modella i singoli nucleotidi C e G che sono più abbondanti nelle isole CpG
- Le CG potrebbero non essere «in frame»

Pros:

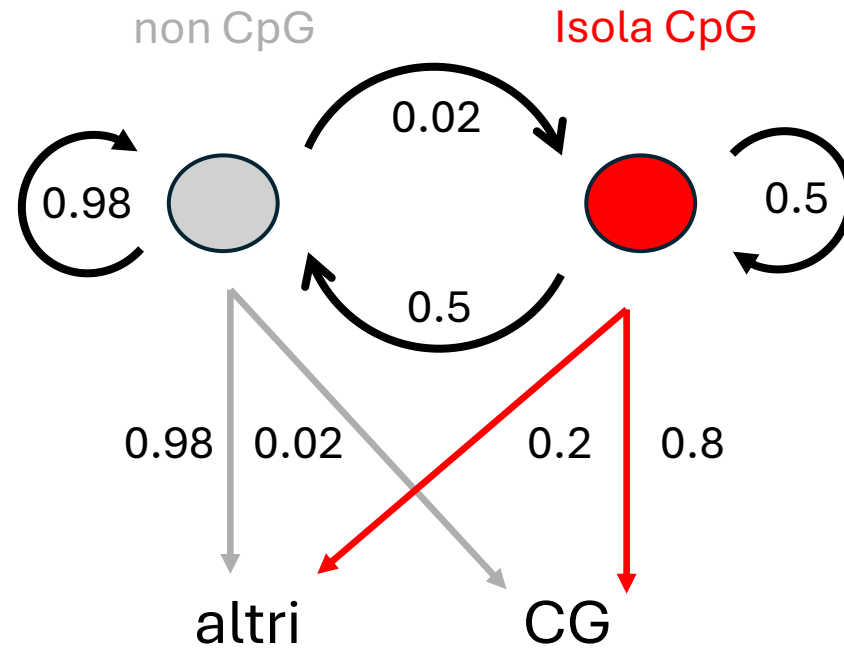
- Le proporzioni di GC sono molto diverse tra non CpG e isole CpG, rendendo più facile la discriminazione tra stati

Useremo questo per semplicità



Modello di Isole CpG a dinucleotidi

(un dinucleotide alla volta)



Cons:

- Non modella i singoli nucleotidi C e G che sono più abbondanti nelle isole CpG
- **Le CG potrebbero non essere «in frame»**

Pros:

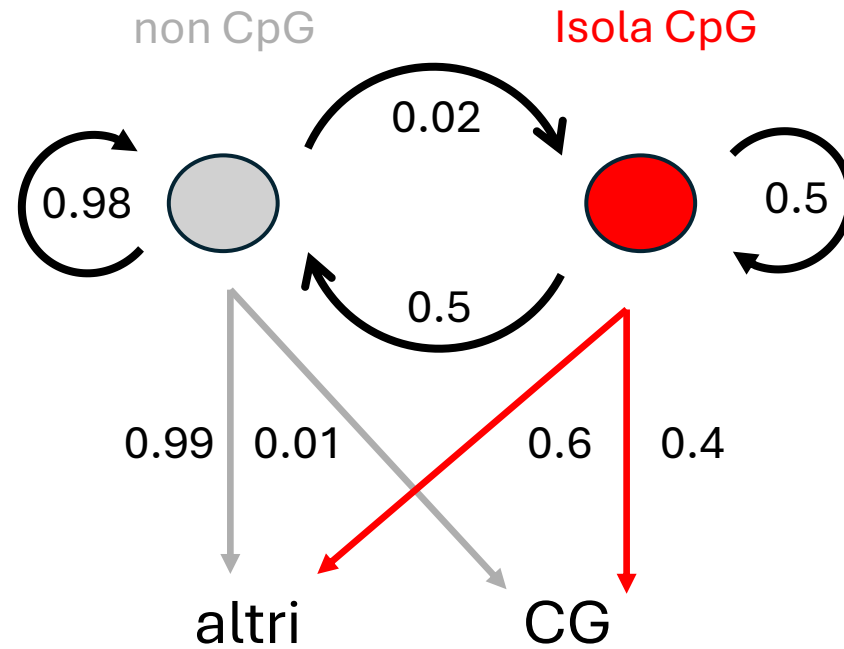
- Le proporzioni di GC sono molto diverse tra non CpG e isole CpG, rendendo più facile la discriminazione tra stati

Useremo questo per semplicità



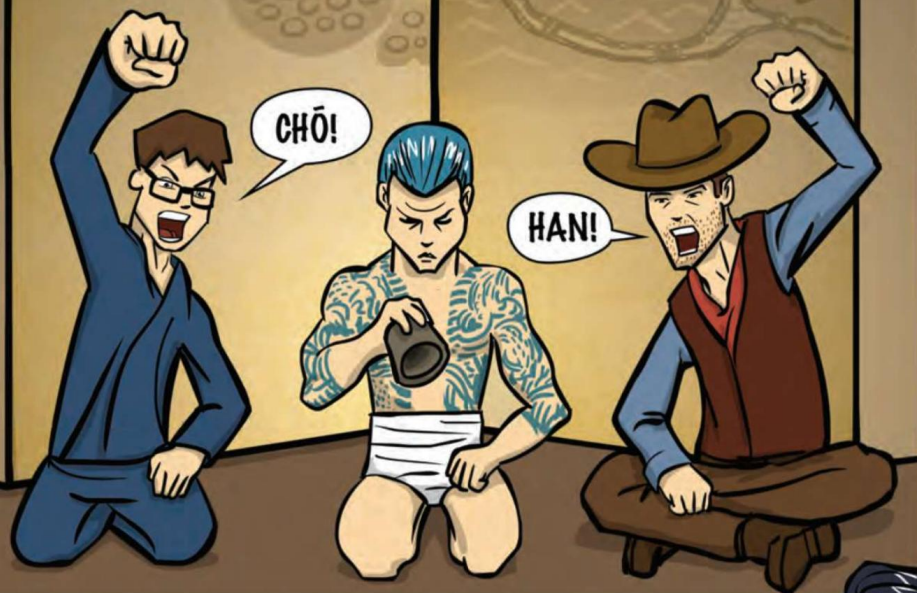
Modello di Isole CpG a dinucleotidi

(un singolo nucleotide alla volta)



WHY HAVE BIOLOGISTS STILL NOT DEVELOPED AN HIV VACCINE?

Hidden Markov Models

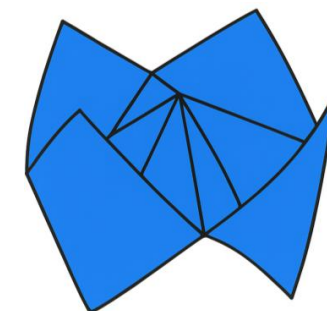
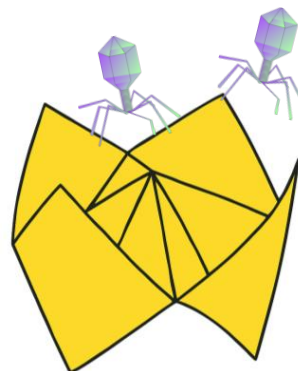
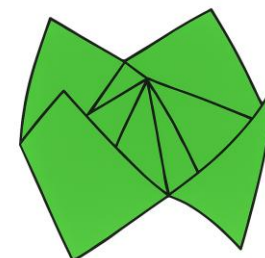
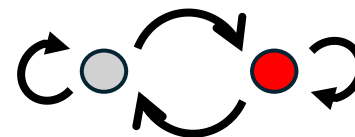
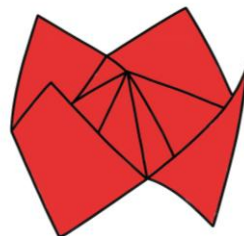


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Come possiamo risolvere il problema?

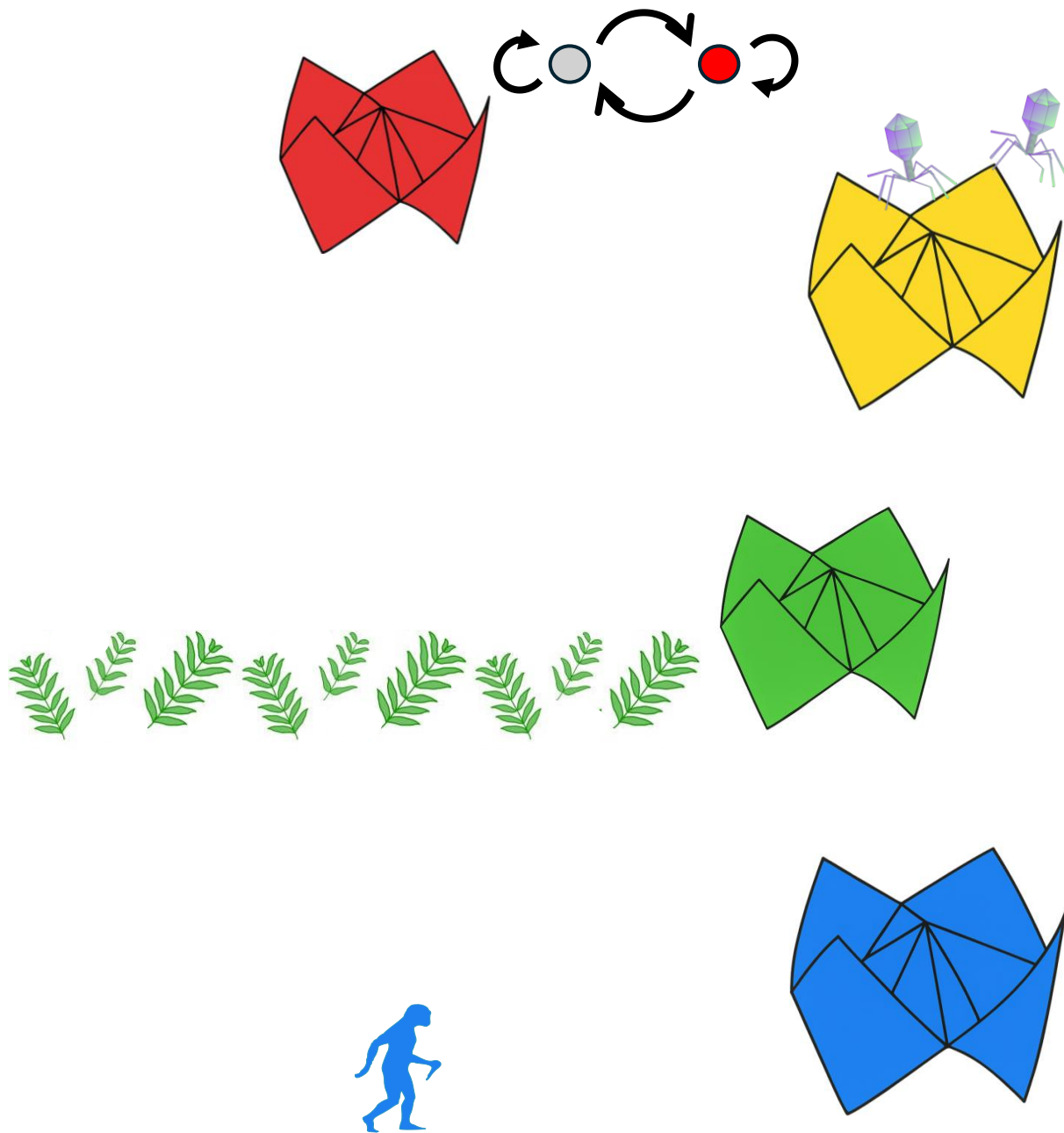


Provate ora ad:

1) inferire la
successione degli
stati che hanno
generato le vostre
osservazioni

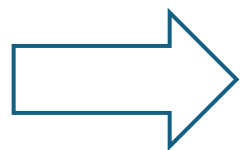
2) stimare probabilità
di emissione e
transizione

Come fareste?

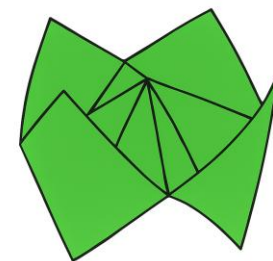
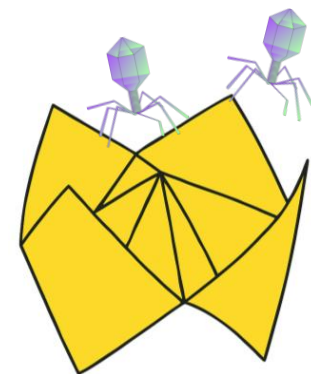
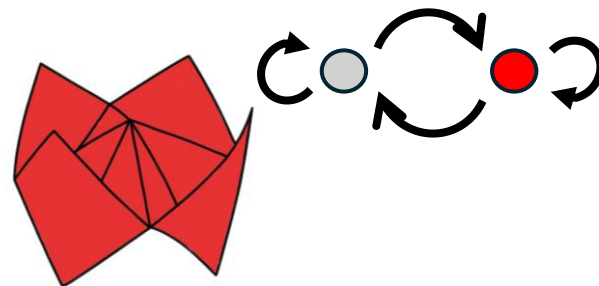


Provate ora ad:

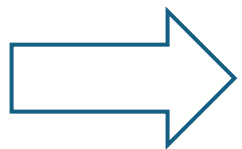
1) inferire la successione degli stati che hanno generato le vostre osservazioni



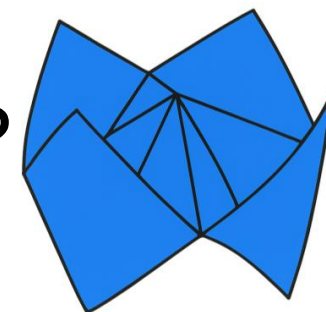
Per ogni emissione, quale ranocchietta stavano usando gli Yakuza Linda, Pedro e Ilenia?



2) stimare probabilità di emissione e transizione

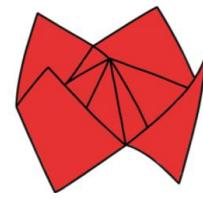


Con che frequenza le hanno cambiate? E quante GC (o AT) aveva ogni ranocchietta?



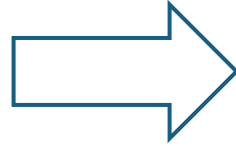
Come fareste?

Provate ora ad:



Esempio Isole CpG

1) inferire la
successione degli
stati che hanno
generato le vostre
osservazioni



AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Problema 1: trovare le nostre regioni CpG

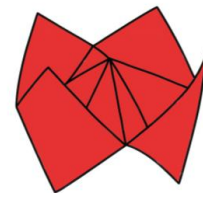
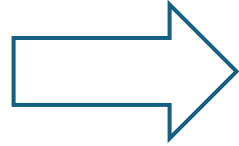


2) stimare probabilità
di emissione e
transizione

Come fareste?

Provate ora ad:

1) inferire la
successione degli
stati che hanno
generato le vostre
osservazioni



Esempio Isole CpG

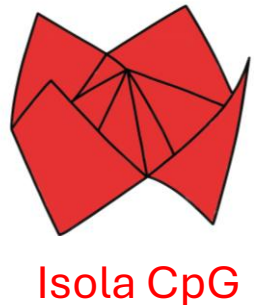
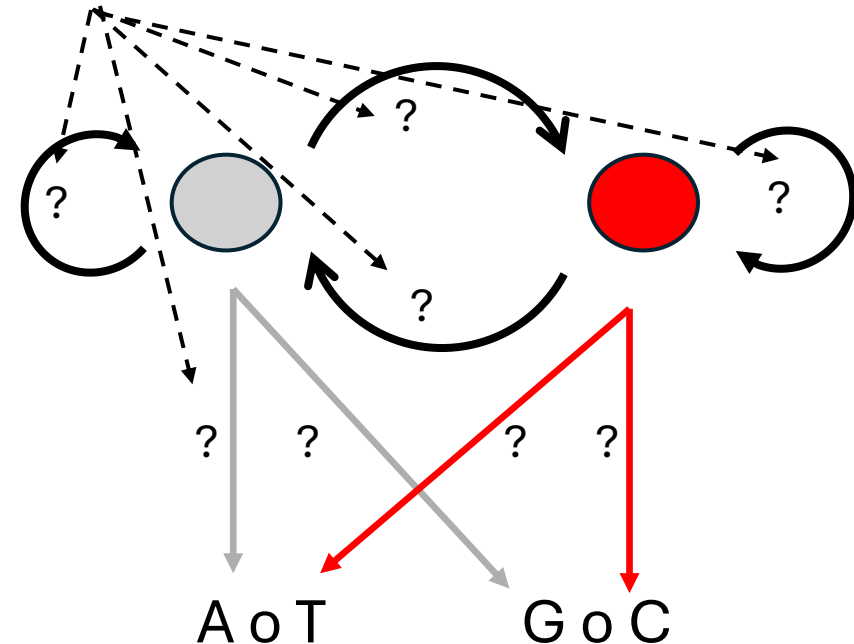
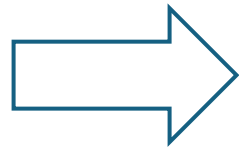
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Problema 1: trovare le nostre regioni CpG

Problema 2: stimare le probabilit  di transizione ed
emission (i parametri del modello)

2) stimare probabilit 
di emissione e
transizione



Come fareste?

Idee che avete avuto!

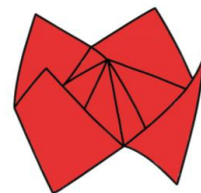
Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!



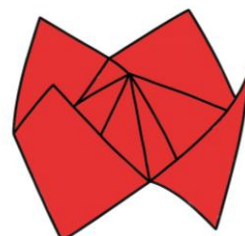
Esempio Isole CpG

AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**

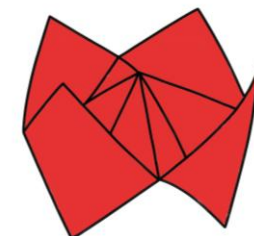
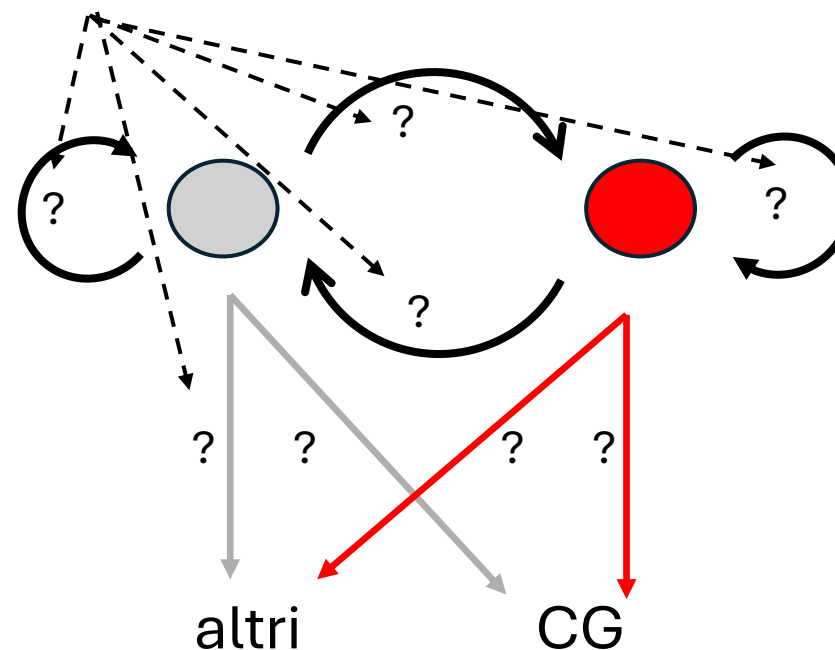


Problema 1: trovare le nostre regioni CpG

Problema 2: stimare le probabilita' di transizione ed emission (i parametri del modello)



non CpG



Isola CpG

Gruppo



Hidden Markov Models



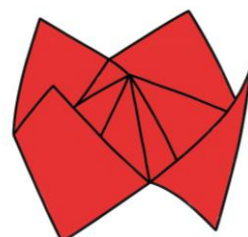
Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

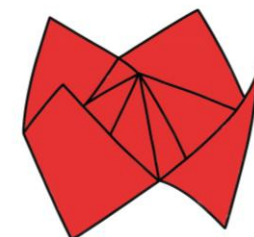
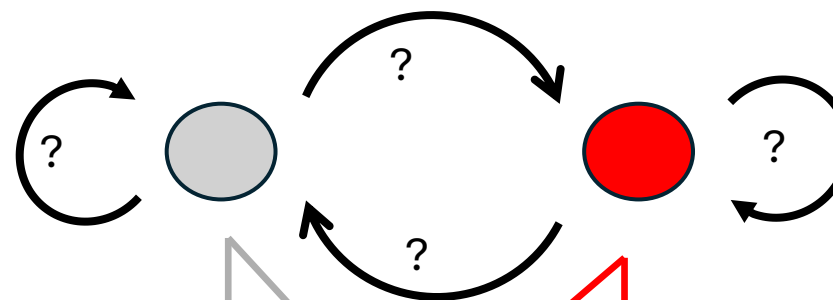
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 2: stimare le probabilita' di transizione ed emissione (i parametri del modello)

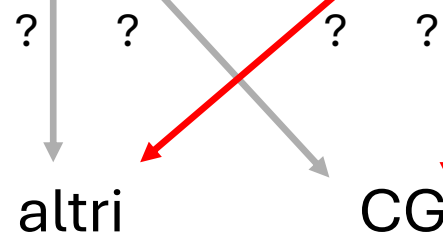
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



non CpG



Isola CpG



Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

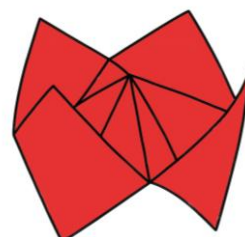
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 2: stimare le probabilita' di transizione ed emissione (i parametri del modello)

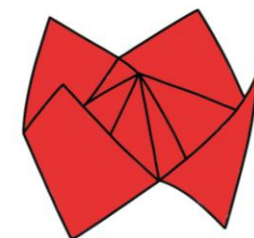
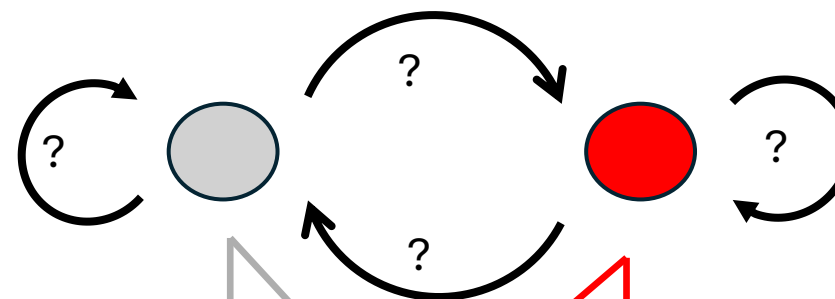
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



4



non CpG



Isola CpG

altri CG

Gruppo



Hidden Markov Models



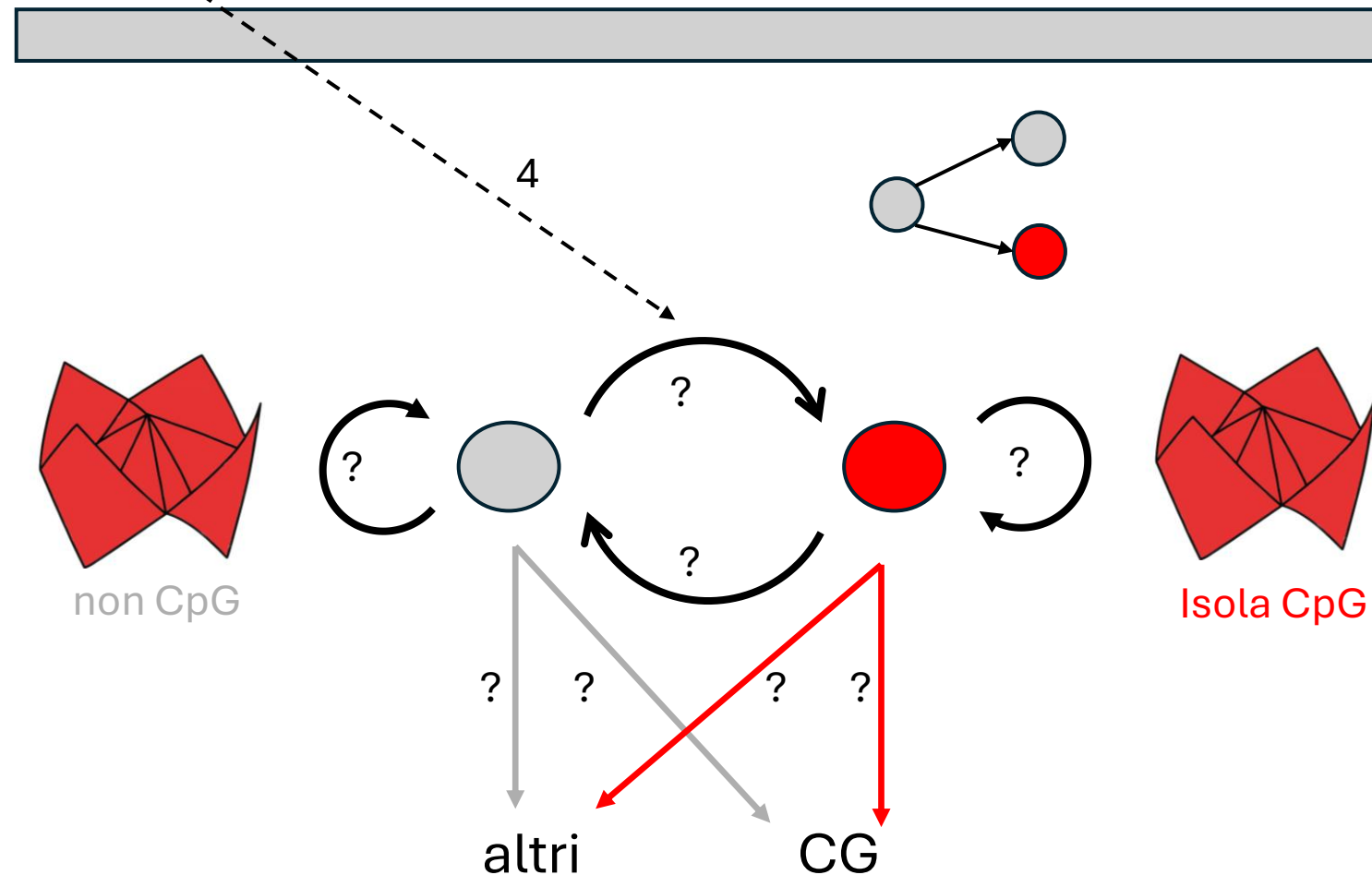
Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 2: stimare le probabilita' di transizione ed emissione (i parametri del modello)

AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Gruppo



Hidden Markov Models

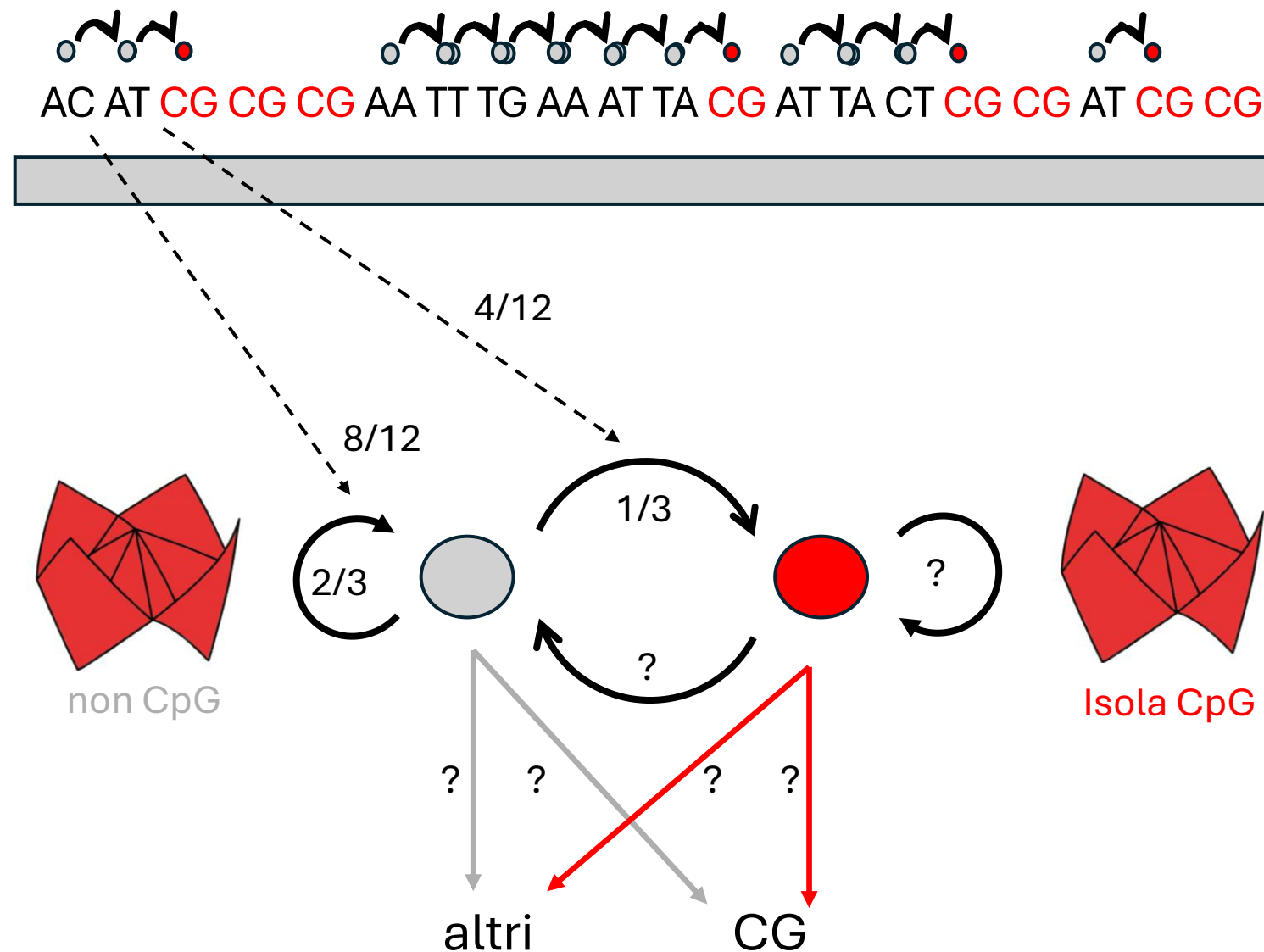


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 2: stimare le probabilita' di transizione ed emissione (i parametri del modello)



Gruppo



Hidden Markov Models

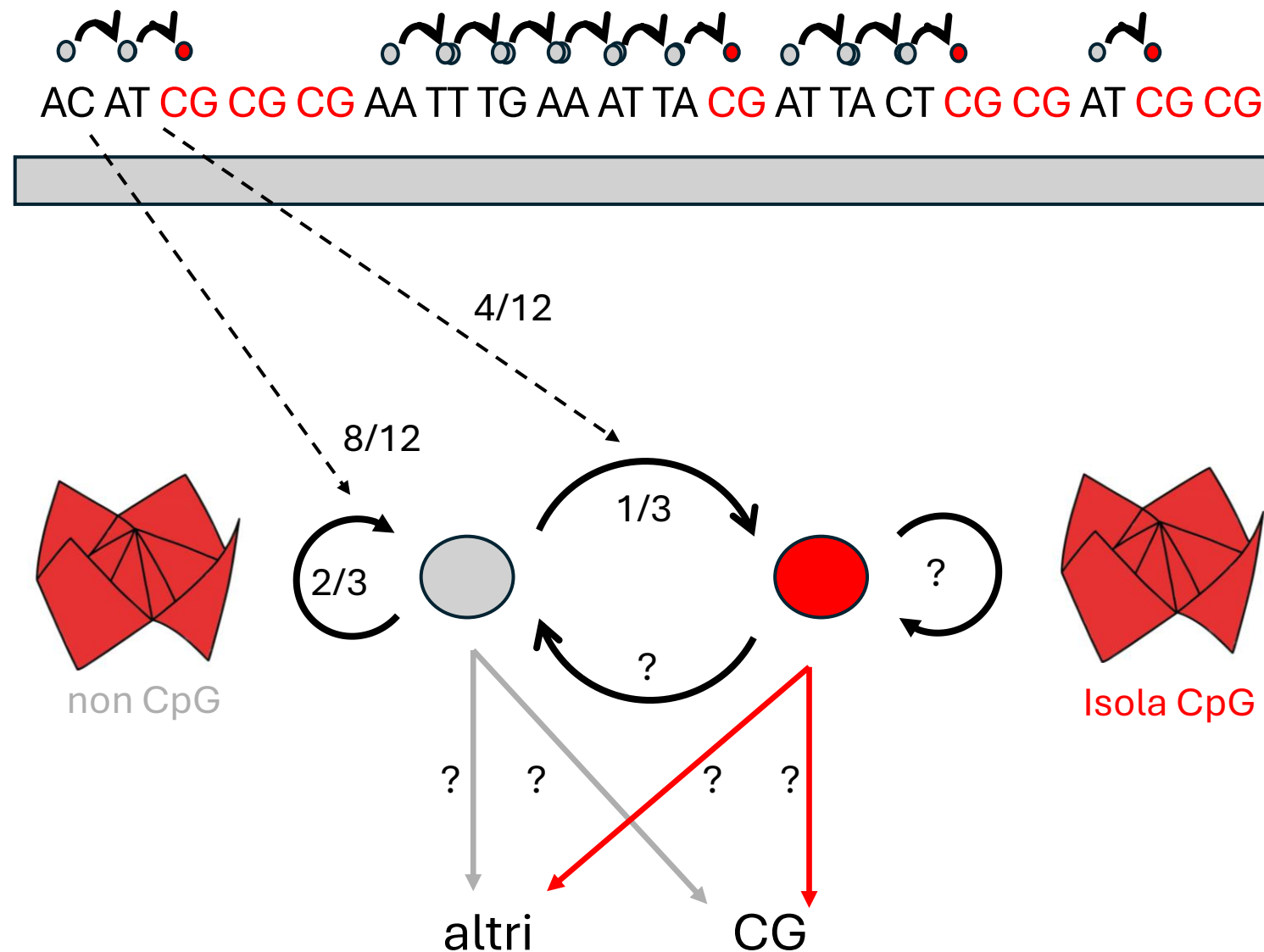


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 2: Quale è il problema di questo approccio?



Gruppo



Hidden Markov Models



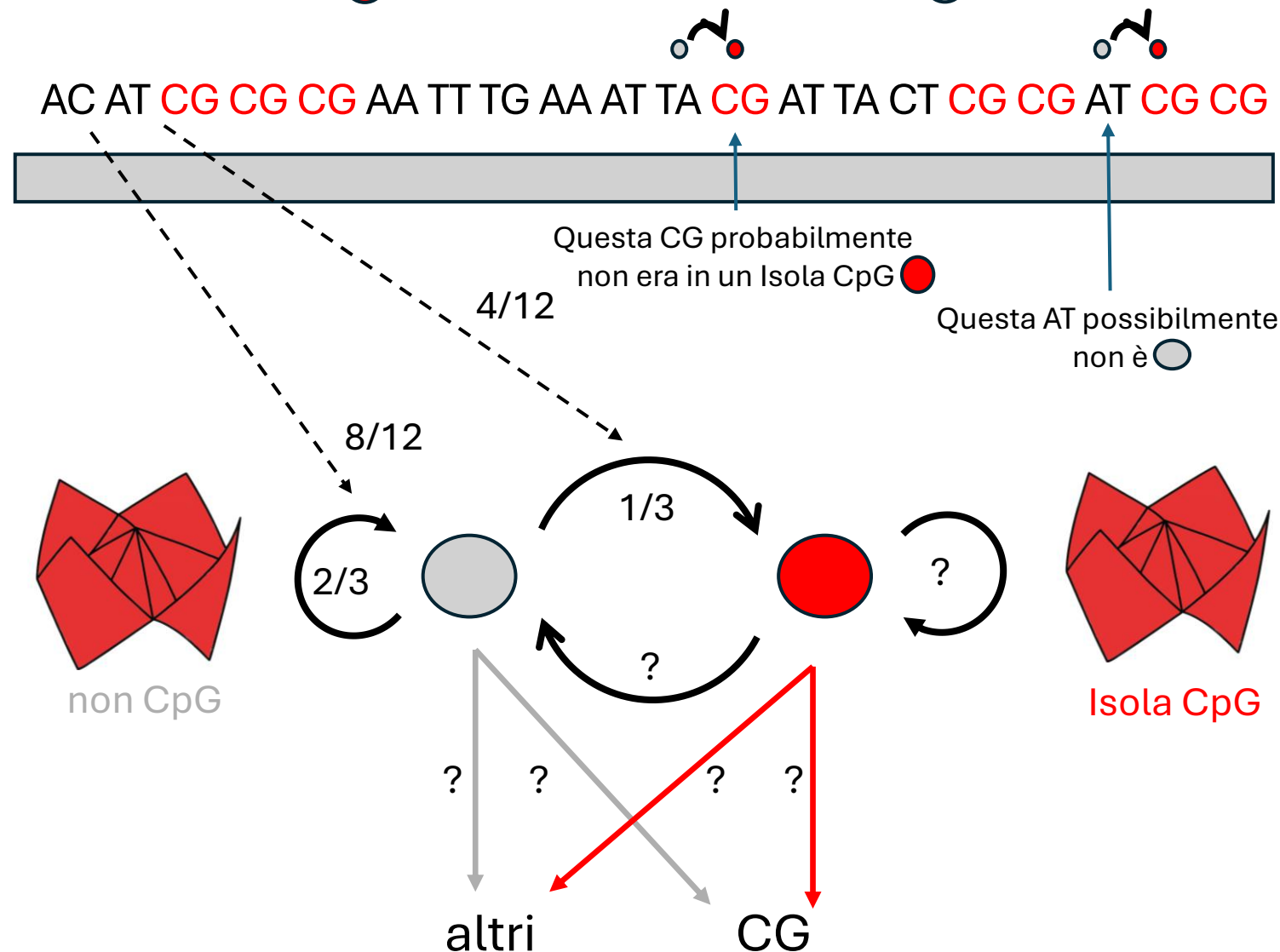
Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

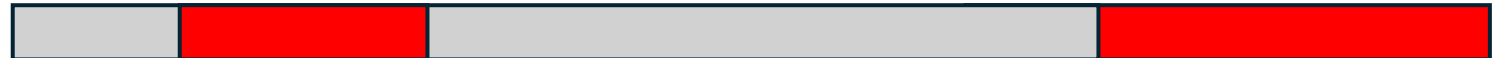
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 2: Quale è il problema?

Noi non conosciamo davvero gli stati!! Potremmo avere dei CG dentro allo stato ● ma anche dentro allo stato ● !!



How were we supposed to know that the casino was run by yakuza?!?!..



Problema 2: stimare le probabilita' di transizione ed emission (i parametri del modello)

Gruppo



Hidden Markov Models

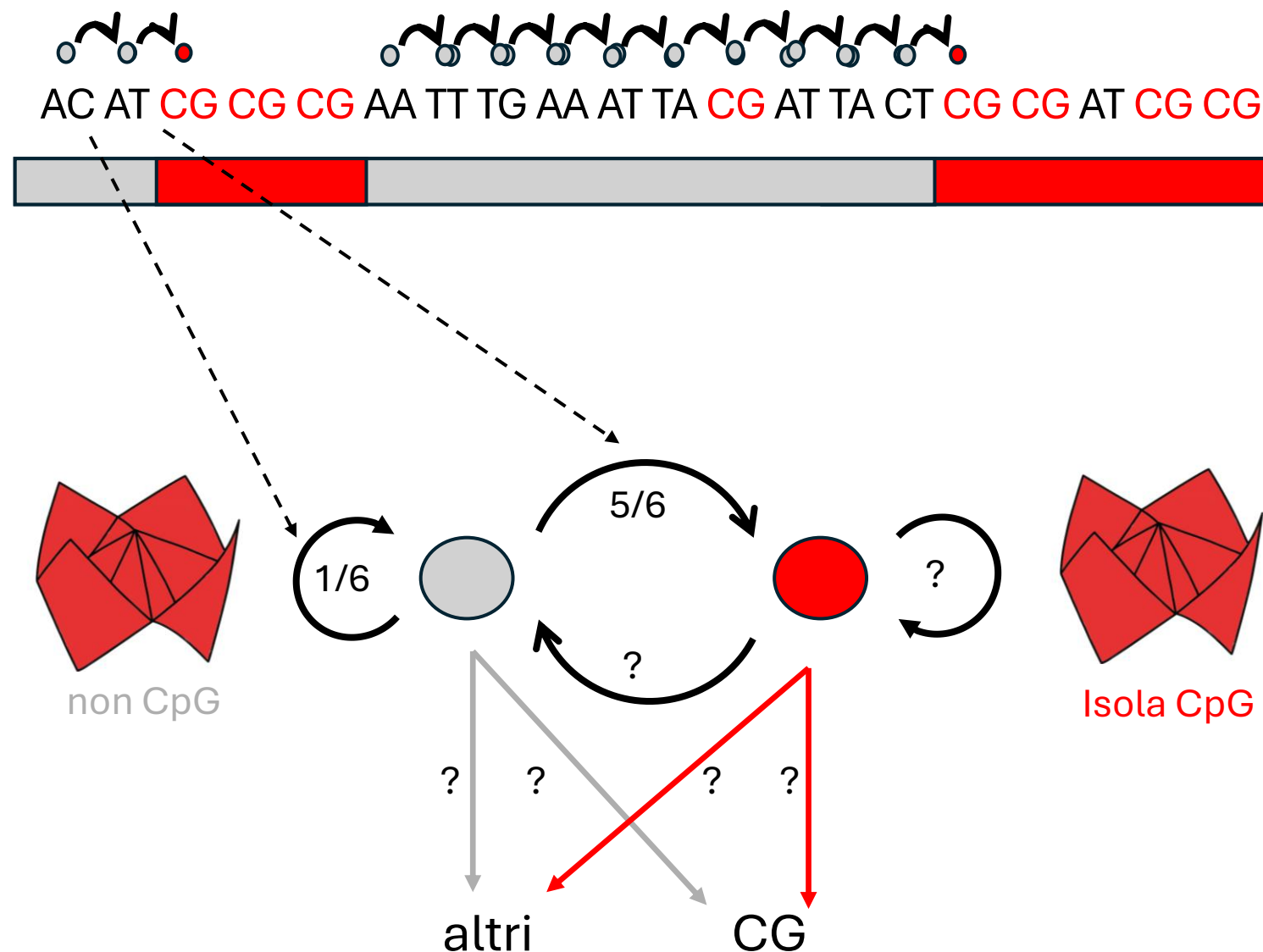


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni



Gruppo



Hidden Markov Models

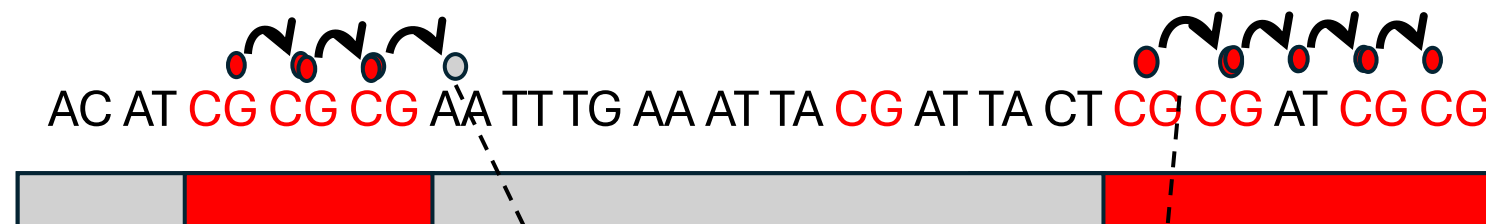


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

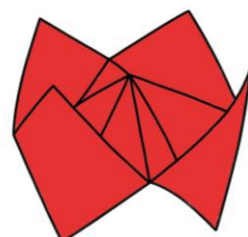
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

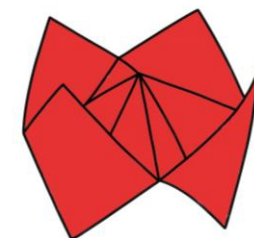
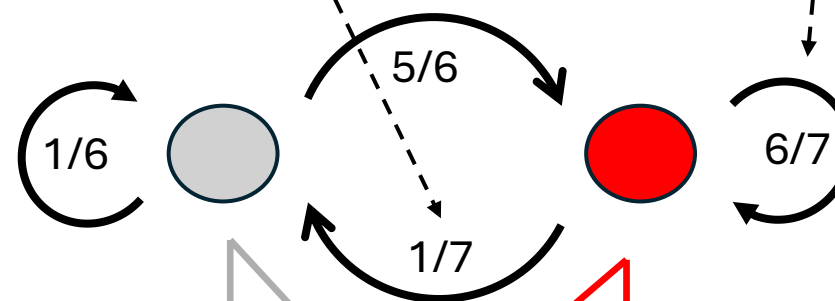
Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni



Le probabilità di emissione sono date dalla probabilità di osservare CG nei vari stati



non CpG



Isola CpG

? ? ? ?
altri CG

Gruppo



Hidden Markov Models

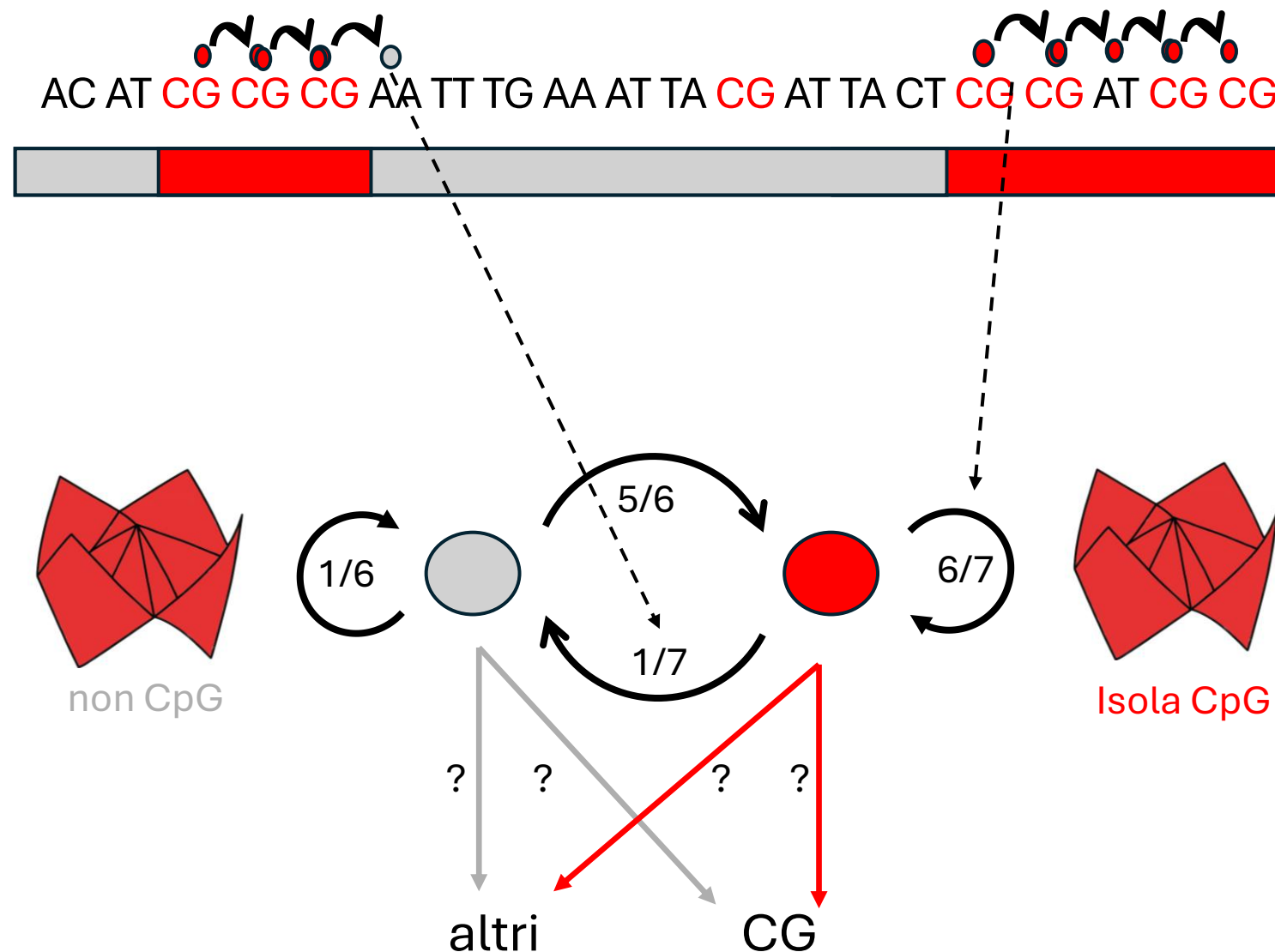


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni



Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

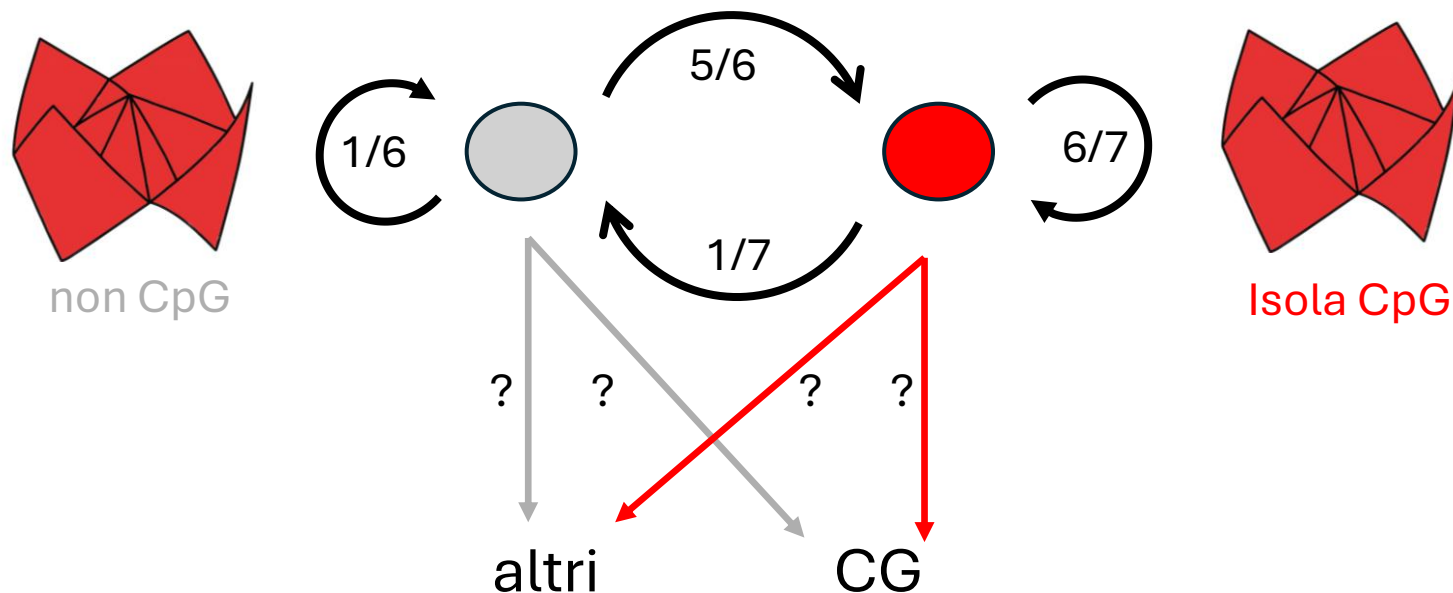
How were we supposed to know that the casino was run by yakuza?!?!..

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni

AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Le probabilità di emissione sono date dalla probabilità di osservare CG nei vari stati



Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

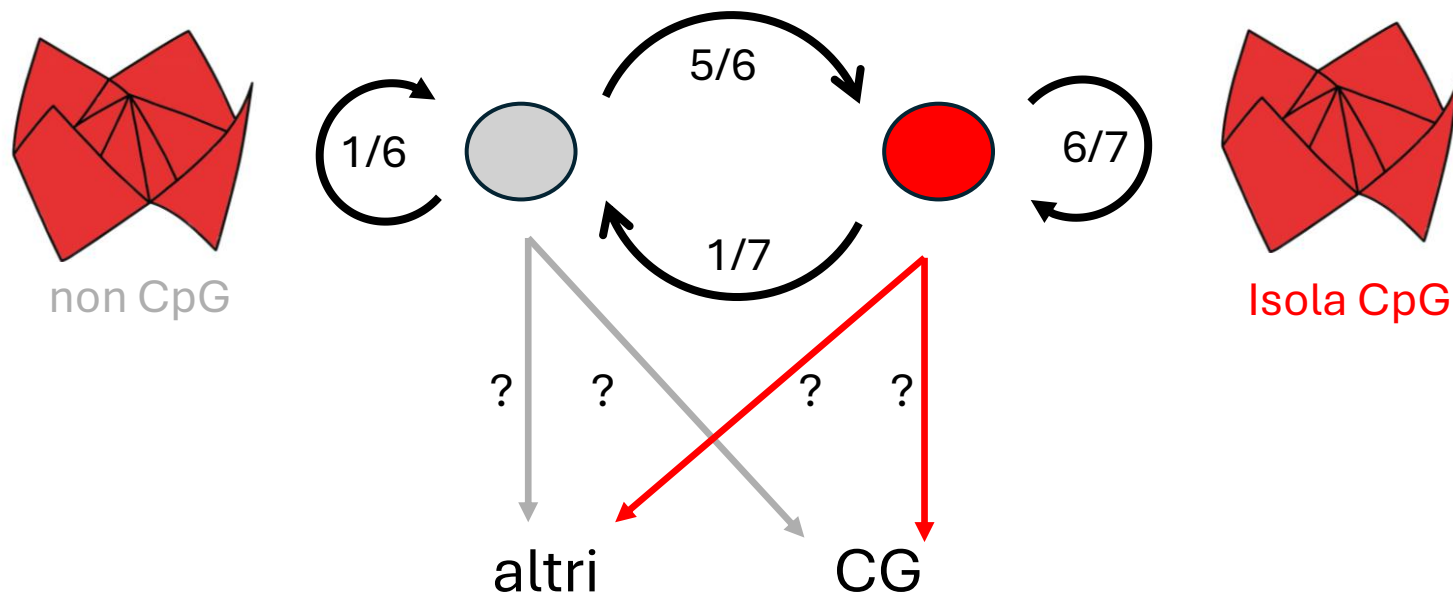
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni

AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Le probabilità di emissione sono date dalla probabilità di osservare CG nei vari stati



Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

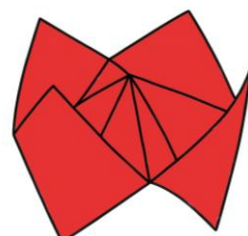
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni

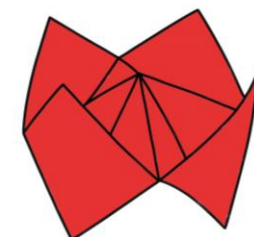
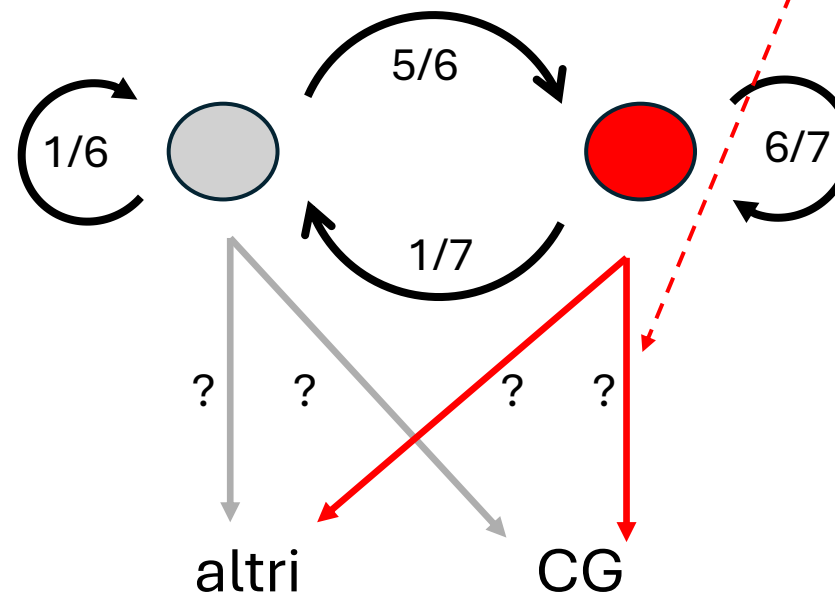
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Le probabilità di emissione sono date dalla probabilità di osservare CG nei vari stati



non CpG



Isola CpG

Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

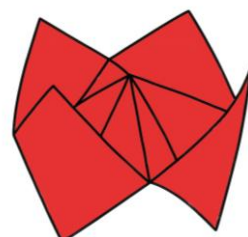
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni

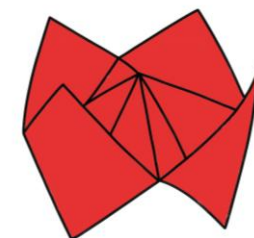
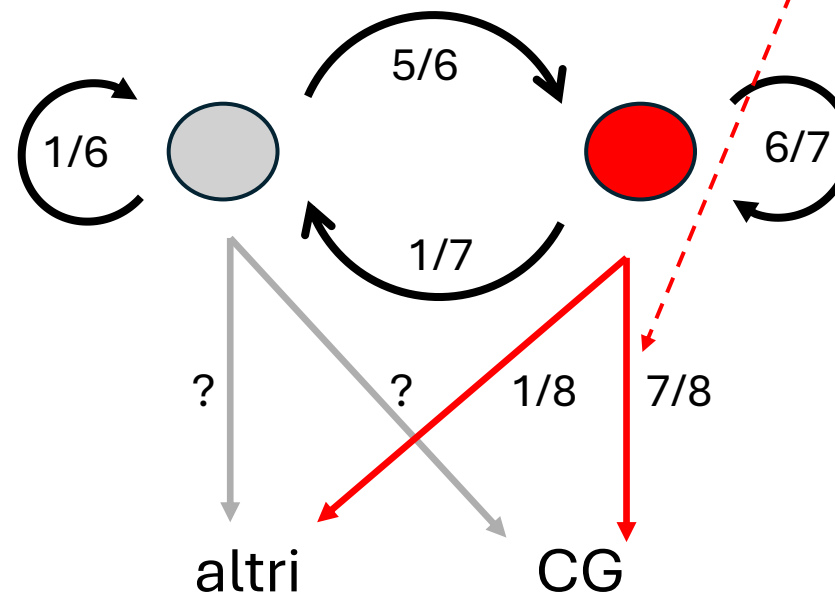
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Le probabilità di emissione sono date dalla probabilità di osservare CG nei vari stati



non CpG



Isola CpG

Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

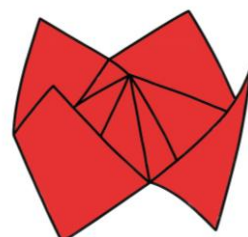
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni

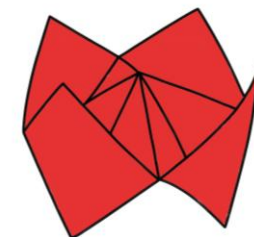
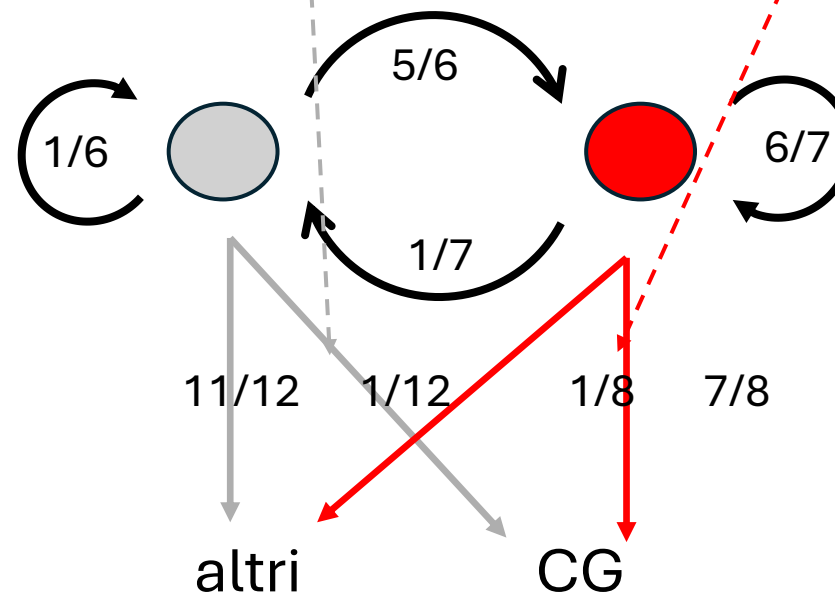
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Le probabilità di emissione sono date dalla probabilità di osservare CG nei vari stati



non CpG



Isola CpG

Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

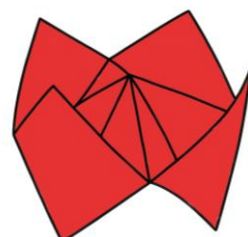
How were we supposed to know that the casino was run by yakuza?!?!

Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni

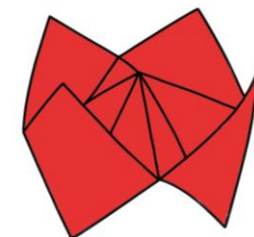
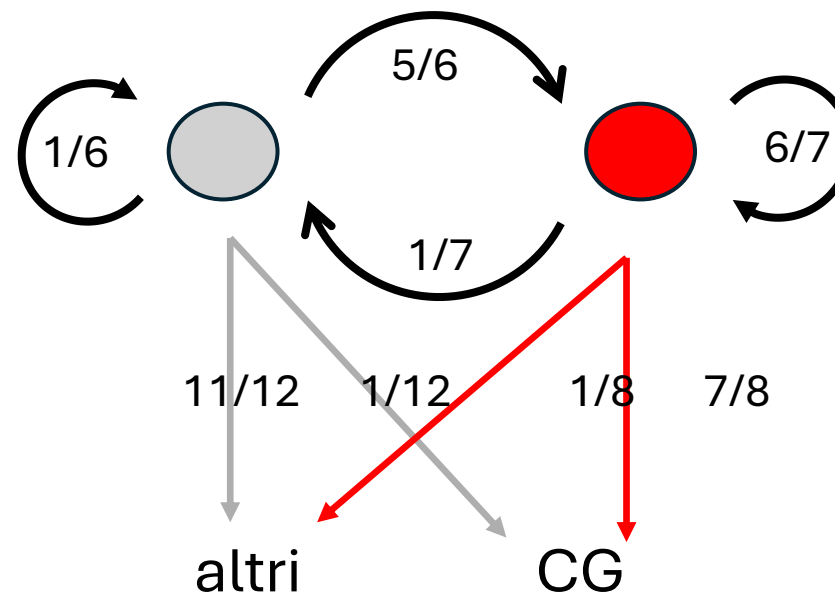
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



Ottimo, sapendo gli stati possiamo stimare le probabilità. Ma quale è il problema di questo approccio?



non CpG

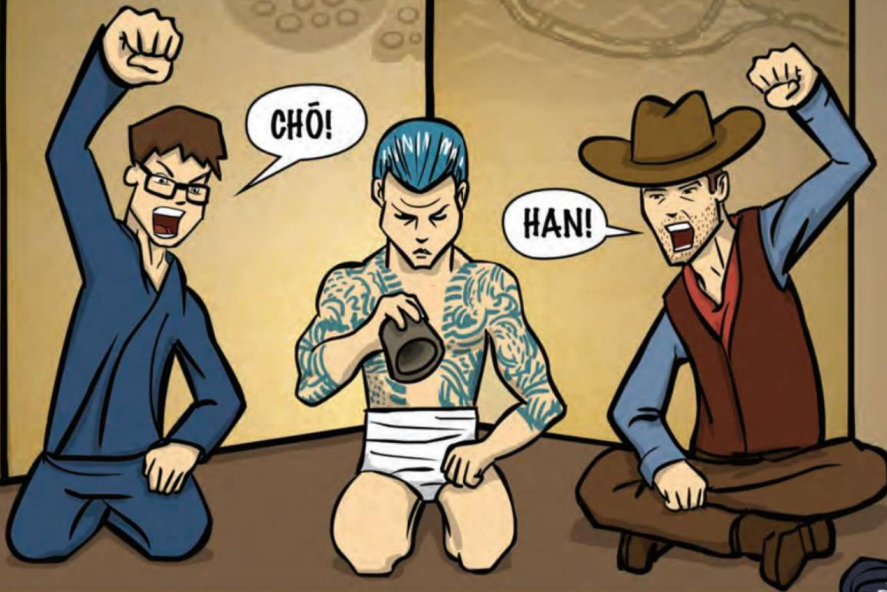


Isola CpG

Gruppo



Hidden Markov Models



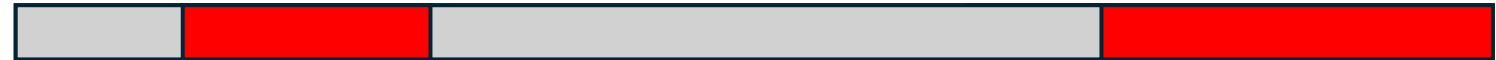
Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

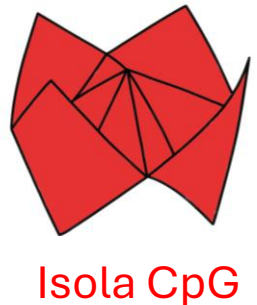
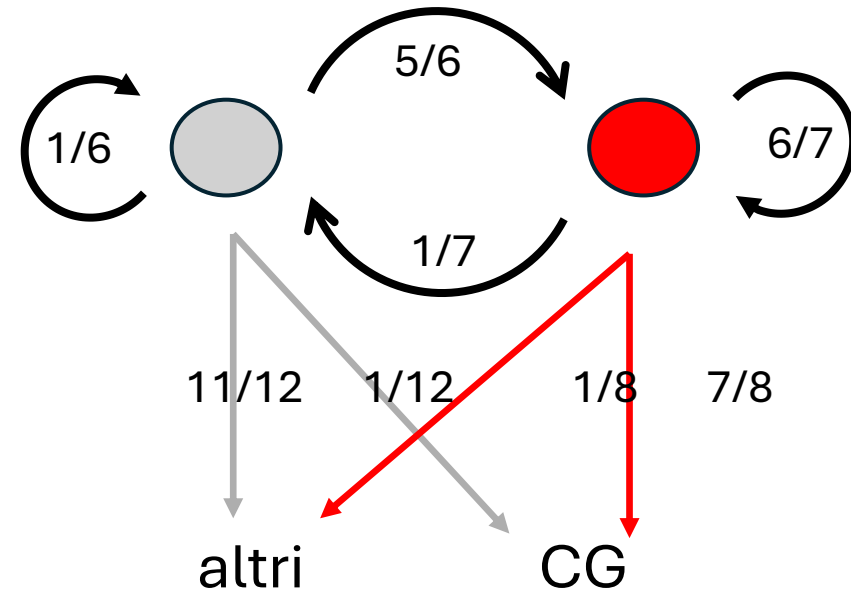
Partiamo da problema 1: Proviamo a partire da delle regioni (individuate ad occhio) e poi calcoliamo emissioni e transizioni

AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



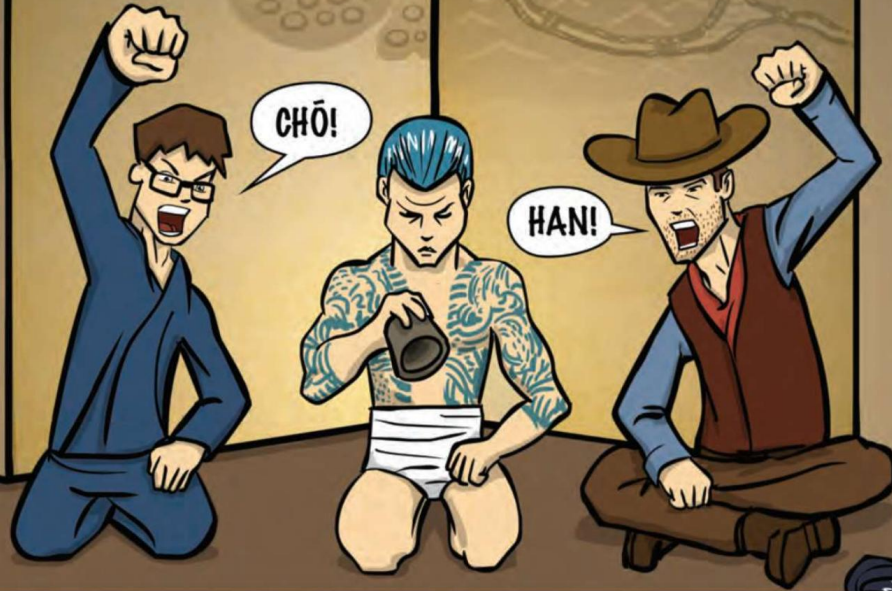
Quale è il problema di questo approccio?

Abbiamo bisogno di un modo per definire gli stati che non sia solo ad occhio!



WHY HAVE BIOLOGISTS STILL NOT DEVELOPED AN HIV VACCINE?

Hidden Markov Models



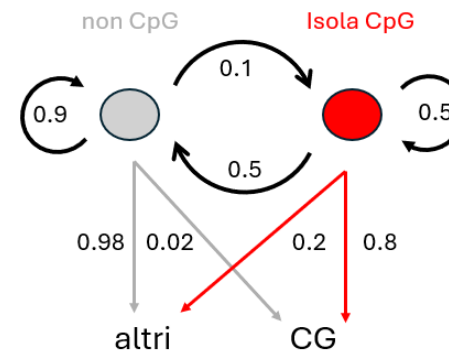
Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

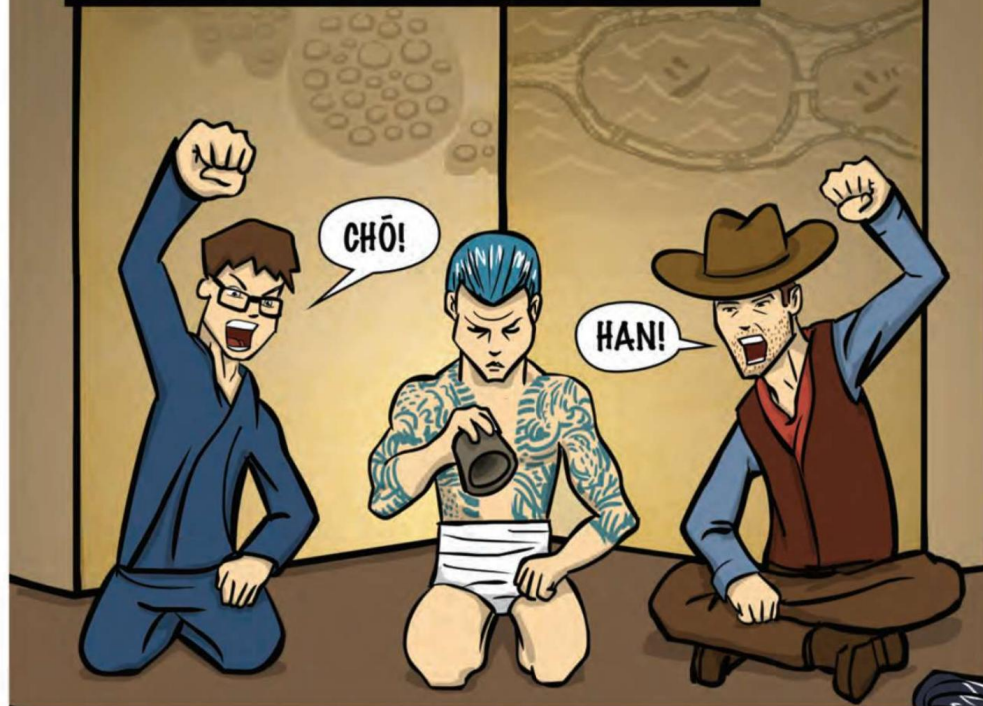
Da dove partire? Quale risolvere prima? Fingiamo di conoscere le probabilità del modello.

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11



Probabilità di un percorso π

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

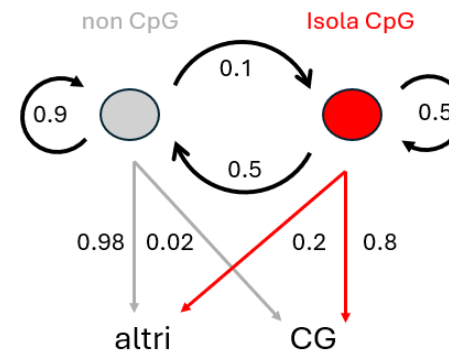


Osservazione x_i

Posizione i

Stato

AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
1	2	3	4	5	6	7	8	9	10	11



Probabilità di un percorso π

Hidden Markov Models

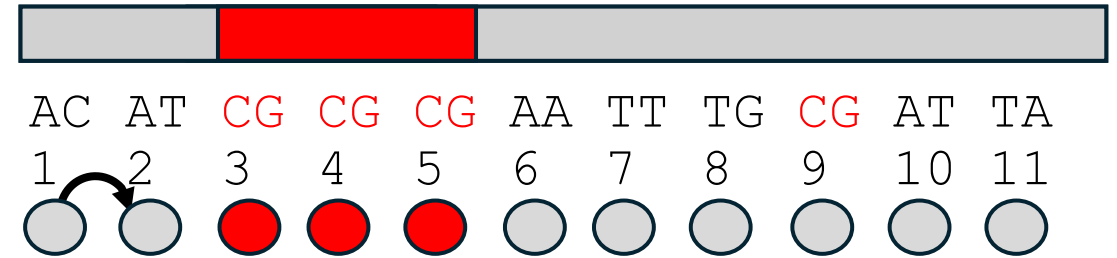


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

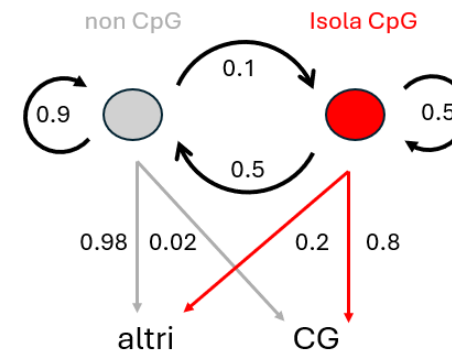
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

?



Probabilità di un percorso π

Hidden Markov Models

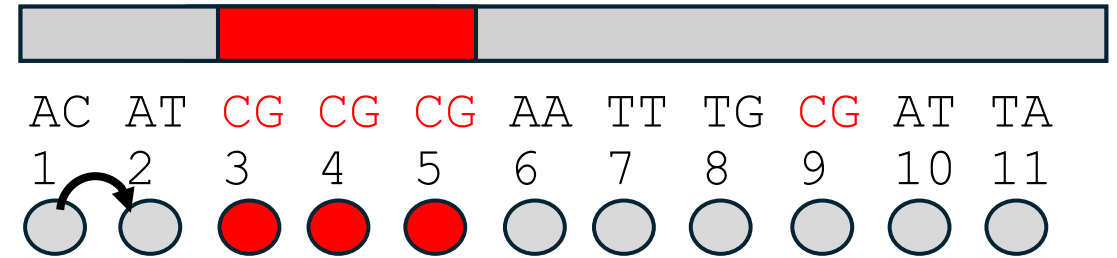


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

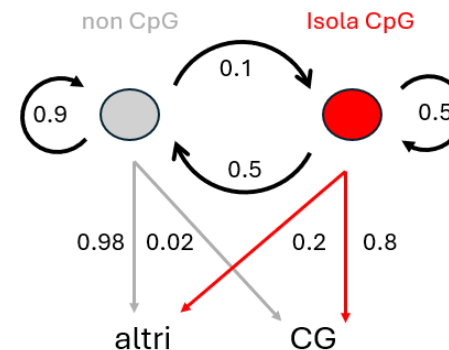
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9



Probabilità di un percorso π

Hidden Markov Models

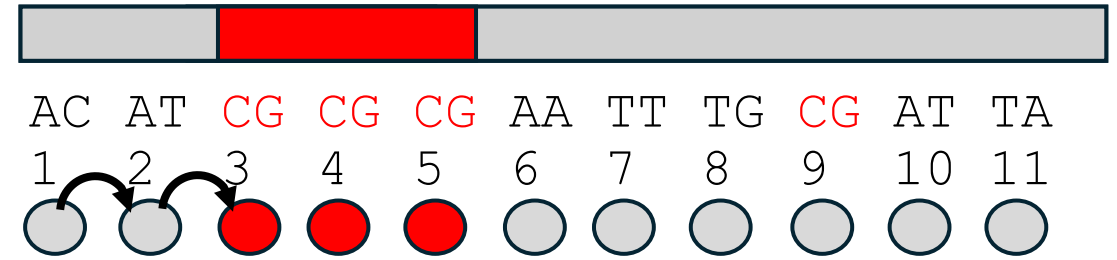


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

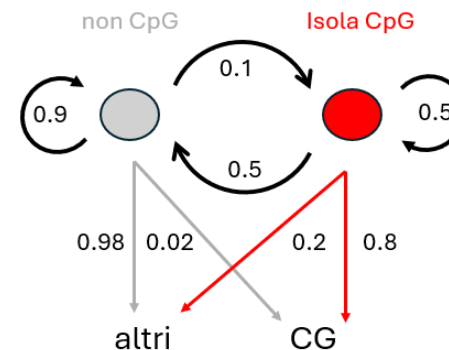
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 ?



Probabilità di un percorso π

Hidden Markov Models

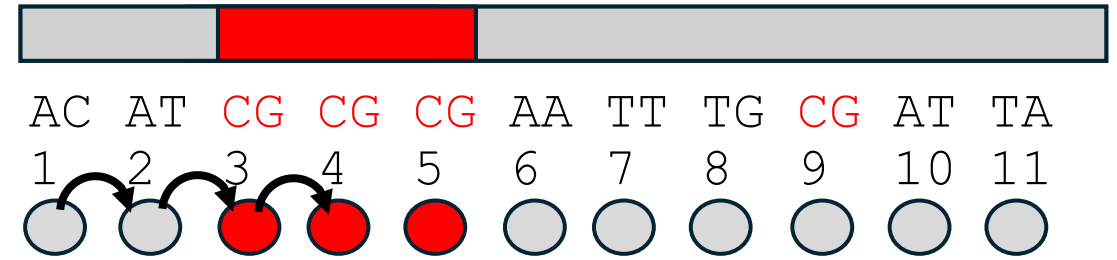


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

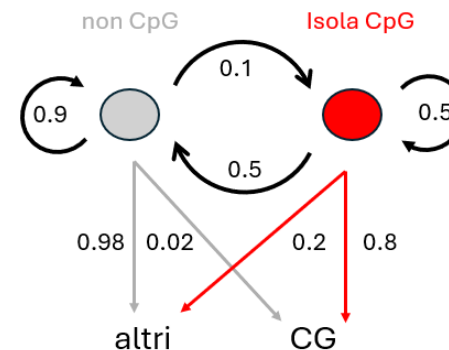
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 0.1 ?



Probabilità di un percorso π

Hidden Markov Models

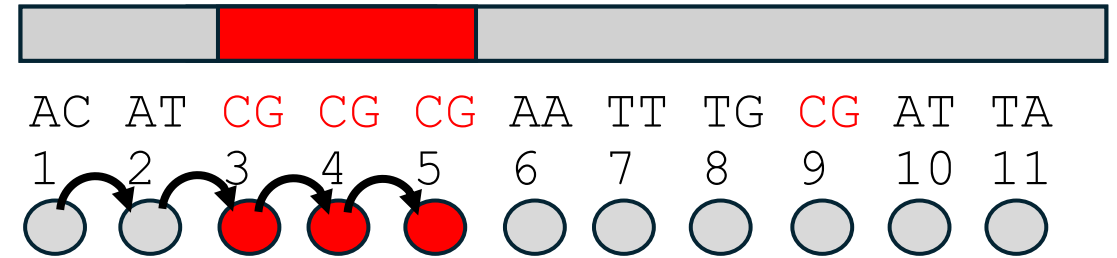


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

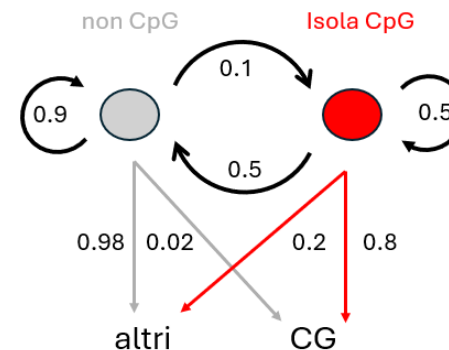
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 0.1 0.5 ?



Probabilità di un percorso π

Hidden Markov Models

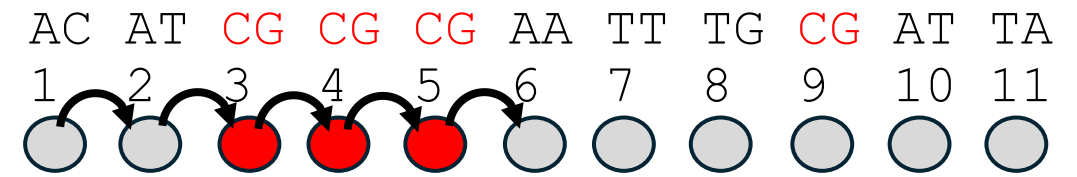


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

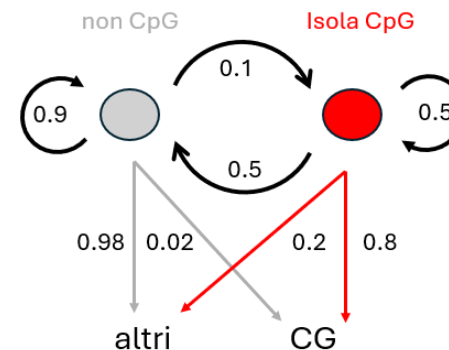
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 0.1 0.5 0.5 ?



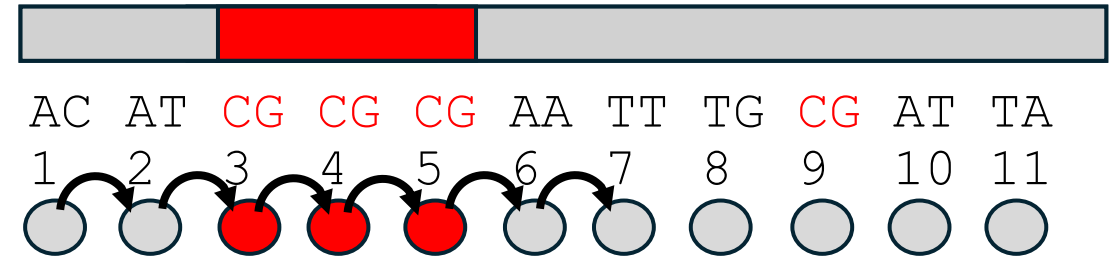
Probabilità di un percorso π

Hidden Markov Models



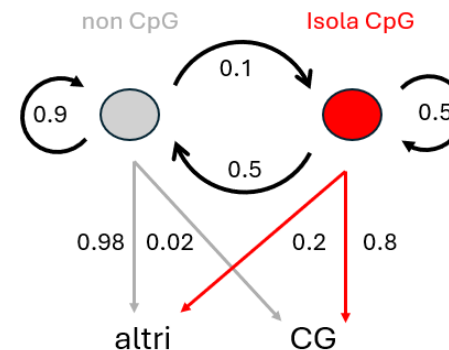
Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i
 Posizione i
 Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 0.1 0.5 0.5 0.5 ?



Probabilità di un percorso π

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

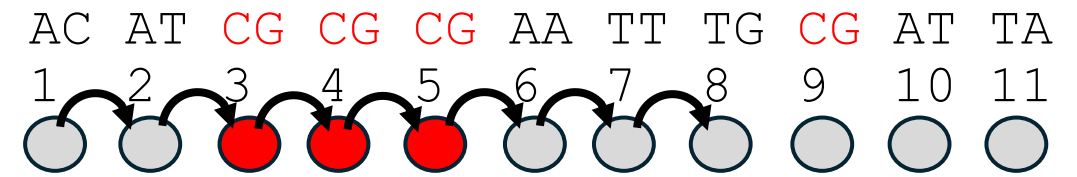
How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

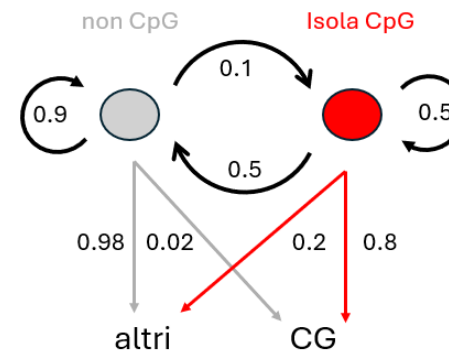
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 0.1 0.5 0.5 0.5 0.9 ?



Probabilità di un percorso π

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

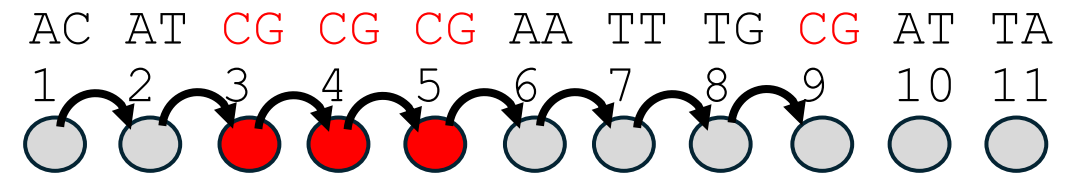
How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

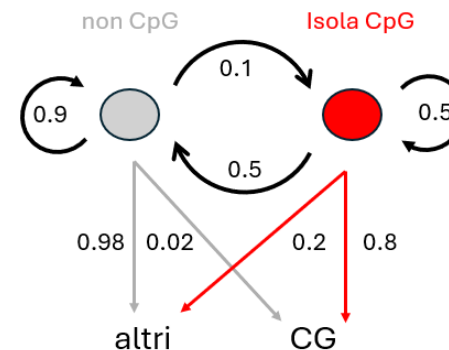
Posizione i

Stato



$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 0.1 0.5 0.5 0.5 0.9 0.9 ?



Probabilità di un percorso π

Hidden Markov Models

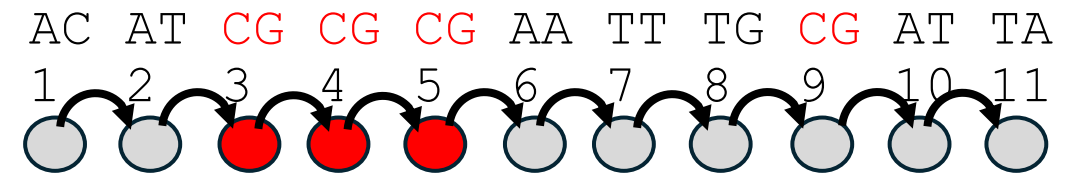


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i

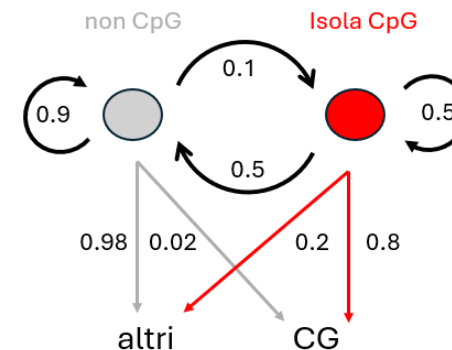
Posizione i

Stato



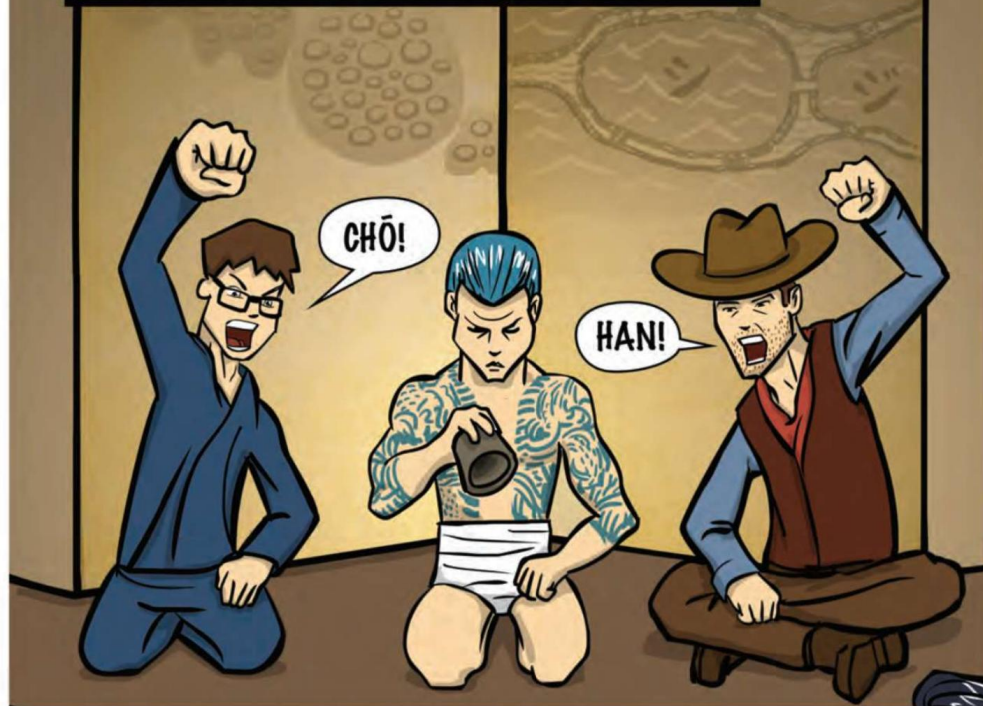
$\Pr(\pi_{i-1} \rightarrow \pi_i)$

0.9 0.1 0.5 0.5 0.5 0.9 0.9 0.9 0.9 0.9



Probabilità di un percorso π

Hidden Markov Models

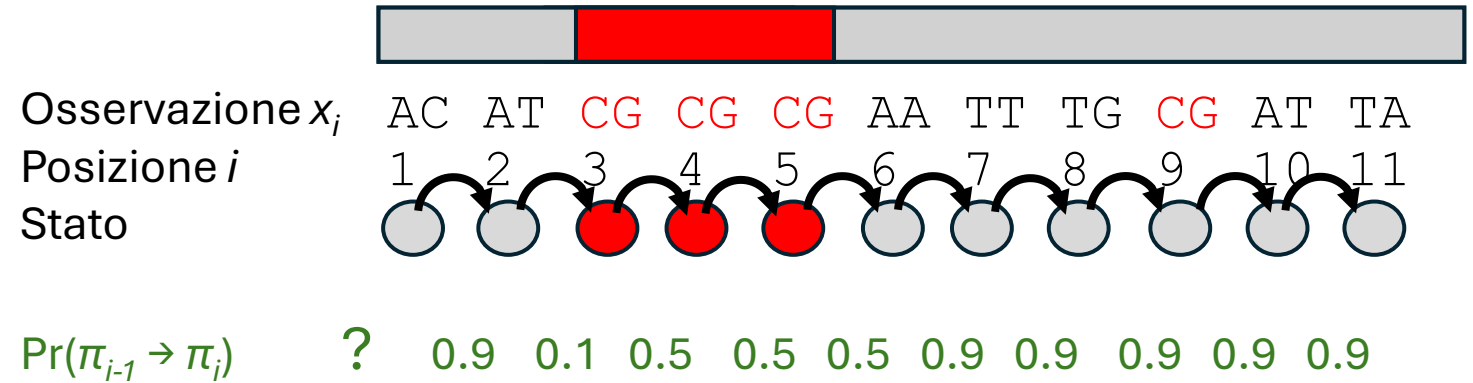


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

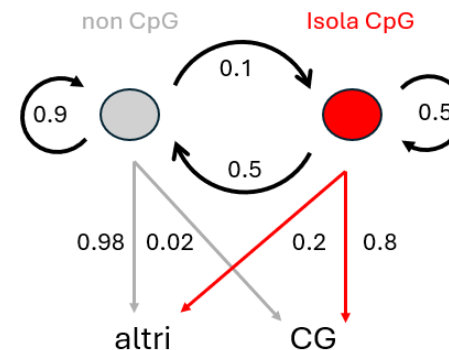
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Per il primo dinucleotide non sappiamo nulla

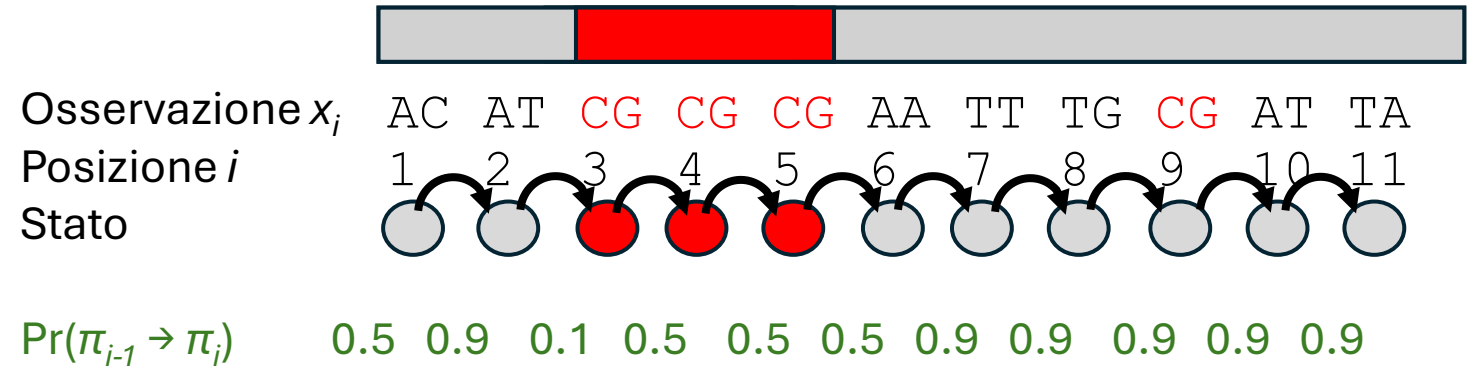


Probabilità di un percorso π

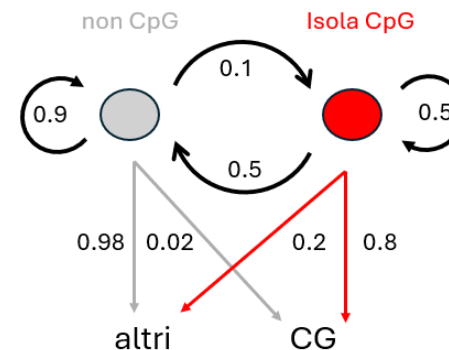
Hidden Markov Models



Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Per il primo dinucleotide non sappiamo nulla, quindi assumiamo che siamo al 50% in  o 

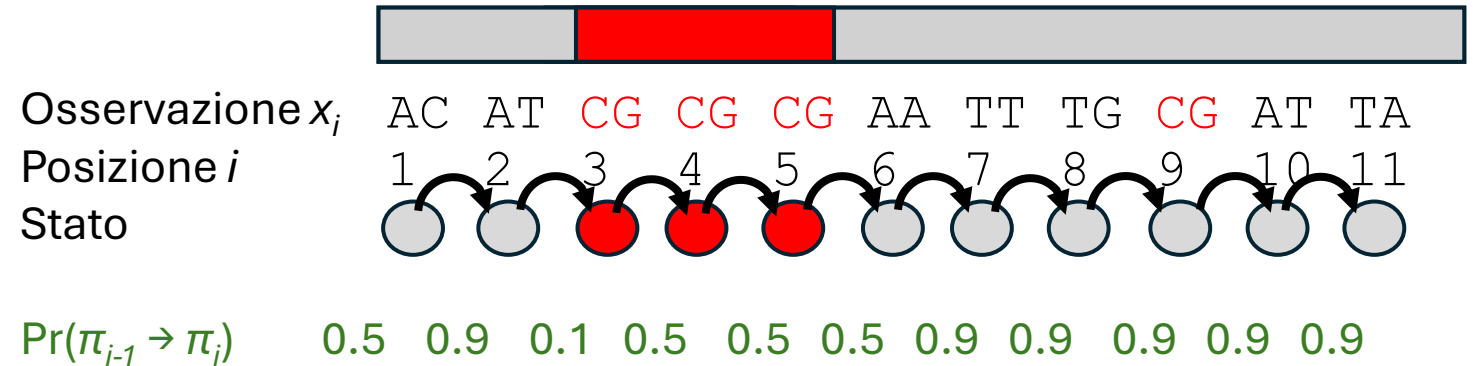


Probabilità di un percorso π

Hidden Markov Models

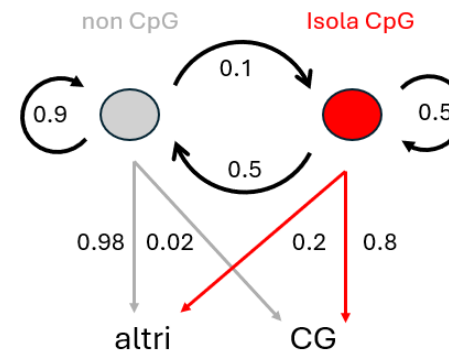


Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



La probabilità del nostro percorso (ancora non considerando le osservazioni) è:

$$Pr(\pi) = 0.5 * 0.9 * 0.1 * 0.5 * 0.5 * 0.5 * 0.9 \dots =$$

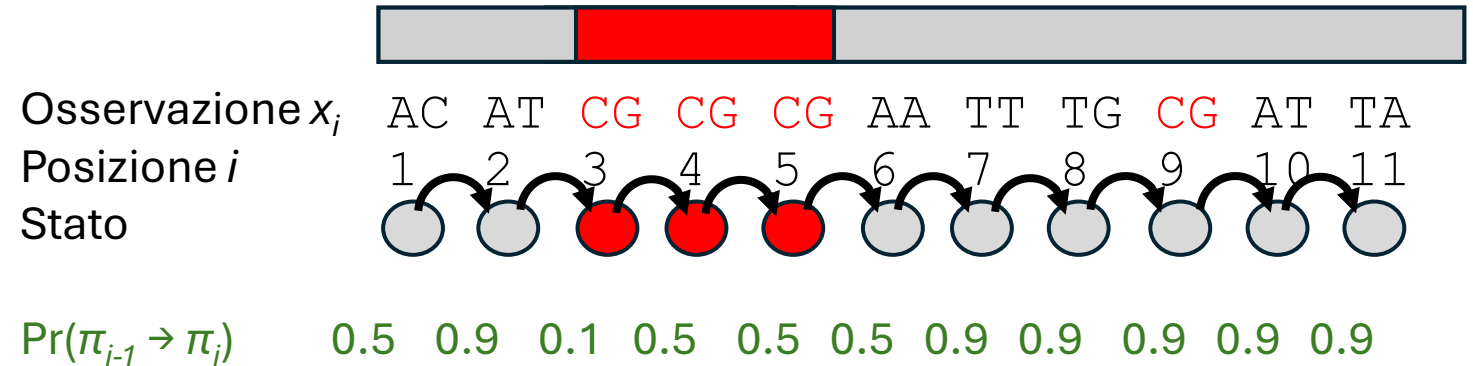


Probabilità di un percorso π

Hidden Markov Models



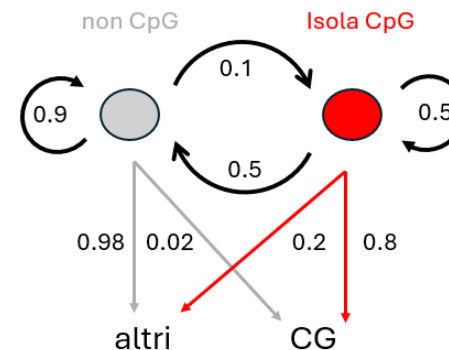
Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



La probabilità del nostro percorso (ancora non considerando le osservazioni) è:

$$\Pr(\pi) = 0.5 * 0.9 * 0.1 * 0.5 * 0.5 * 0.5 * 0.9 * \dots =$$

$$= \prod_i \Pr(\pi_{i-1} \rightarrow \pi_i) \simeq \prod_i \text{Probabilità di transizione}$$



Probabilità delle osservazioni

Hidden Markov Models

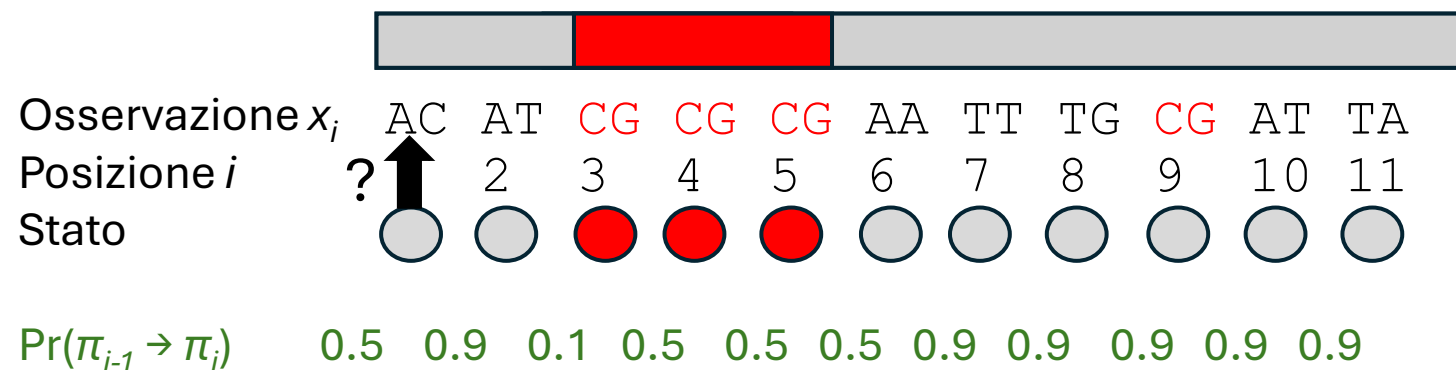


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

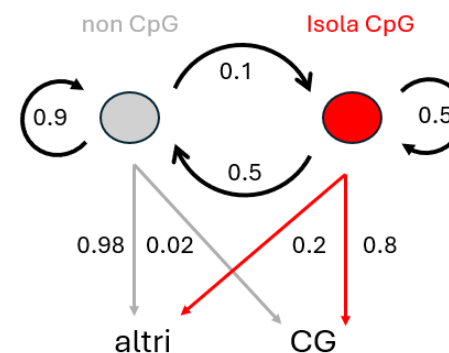
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models

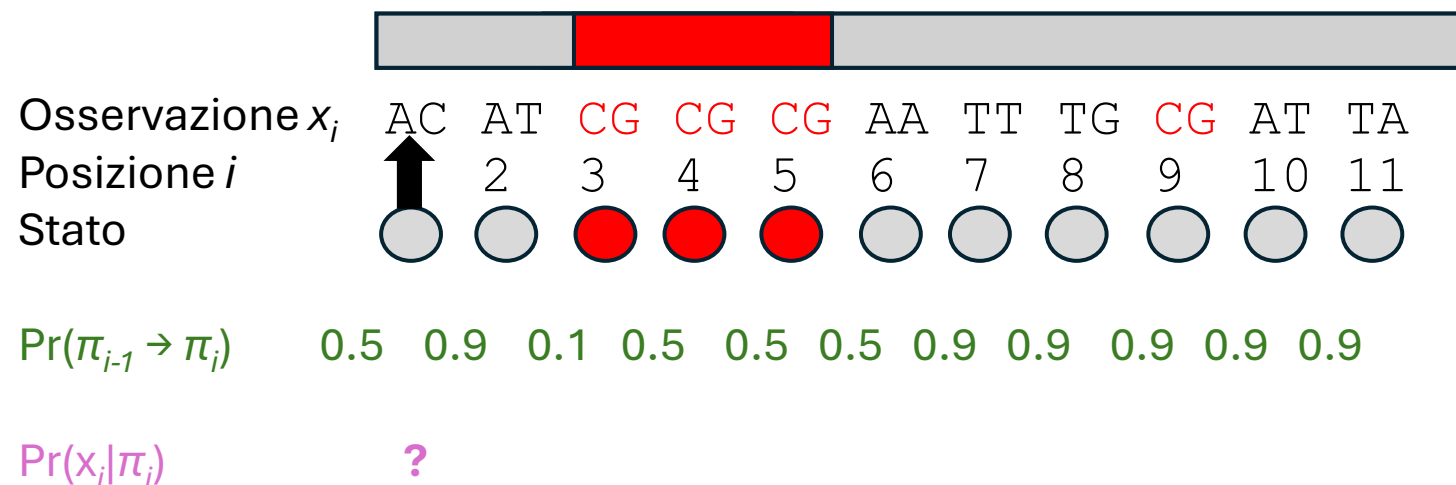


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

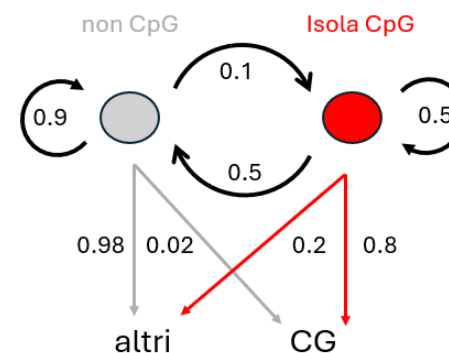
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models

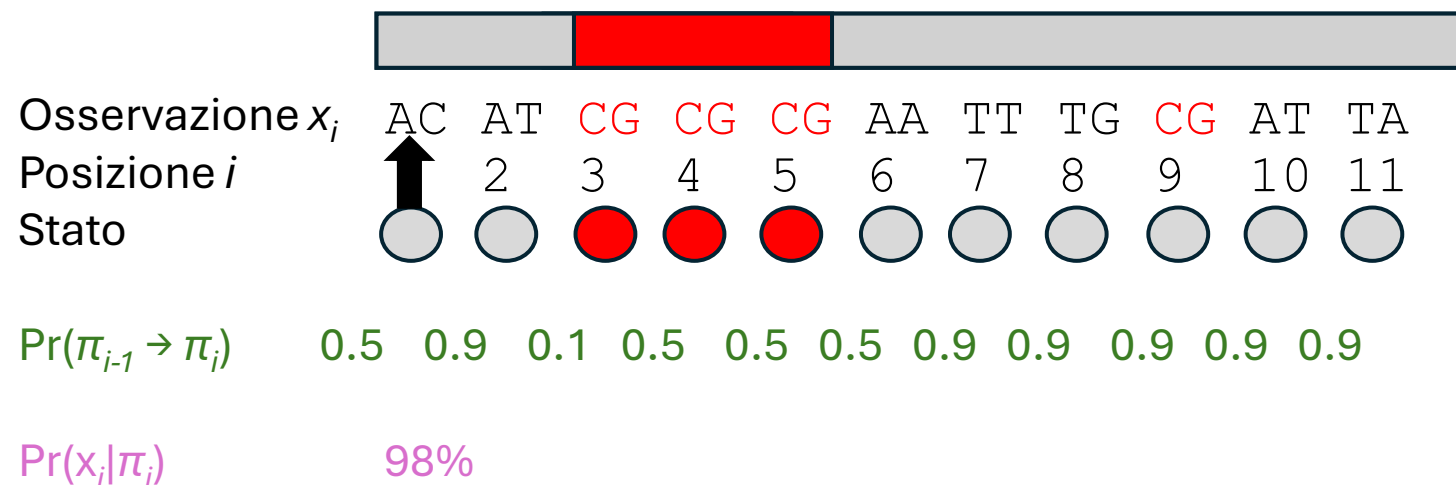


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

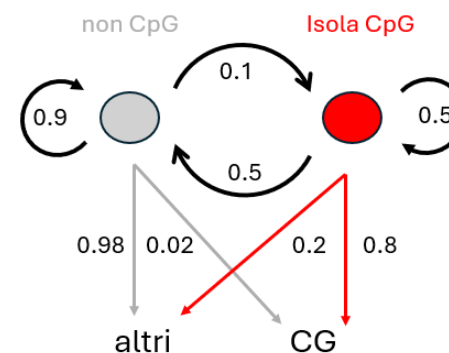
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models

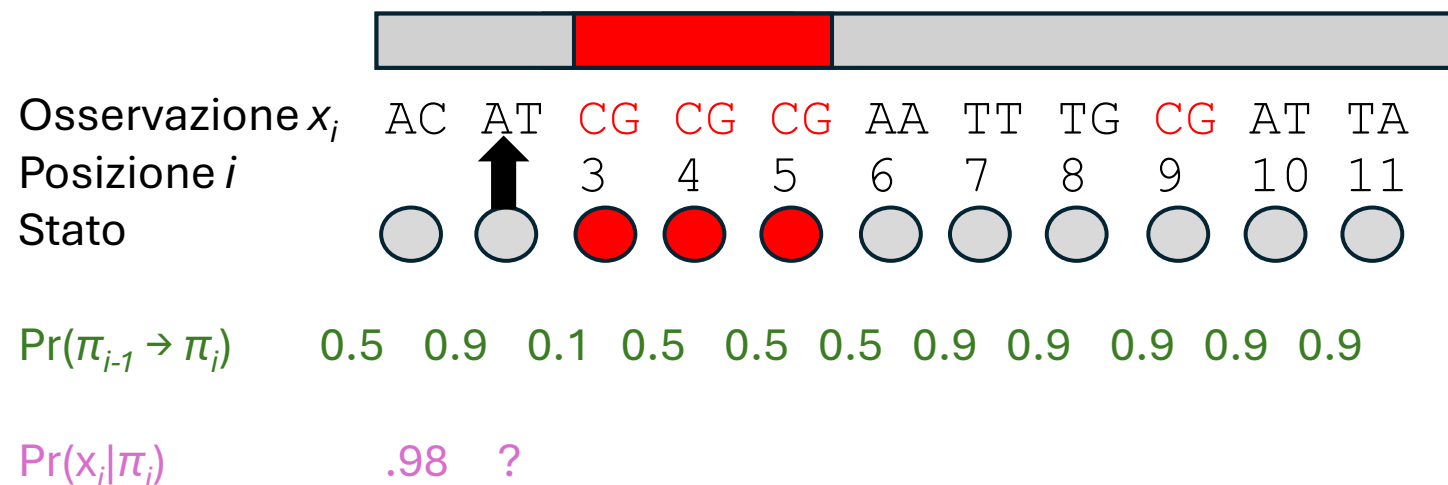


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

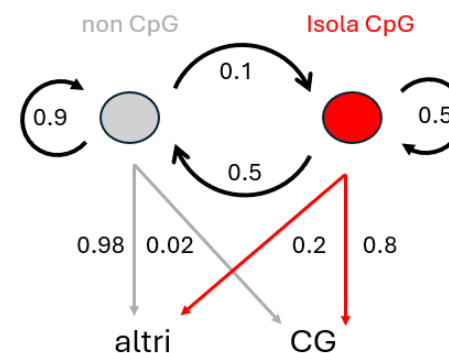
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models

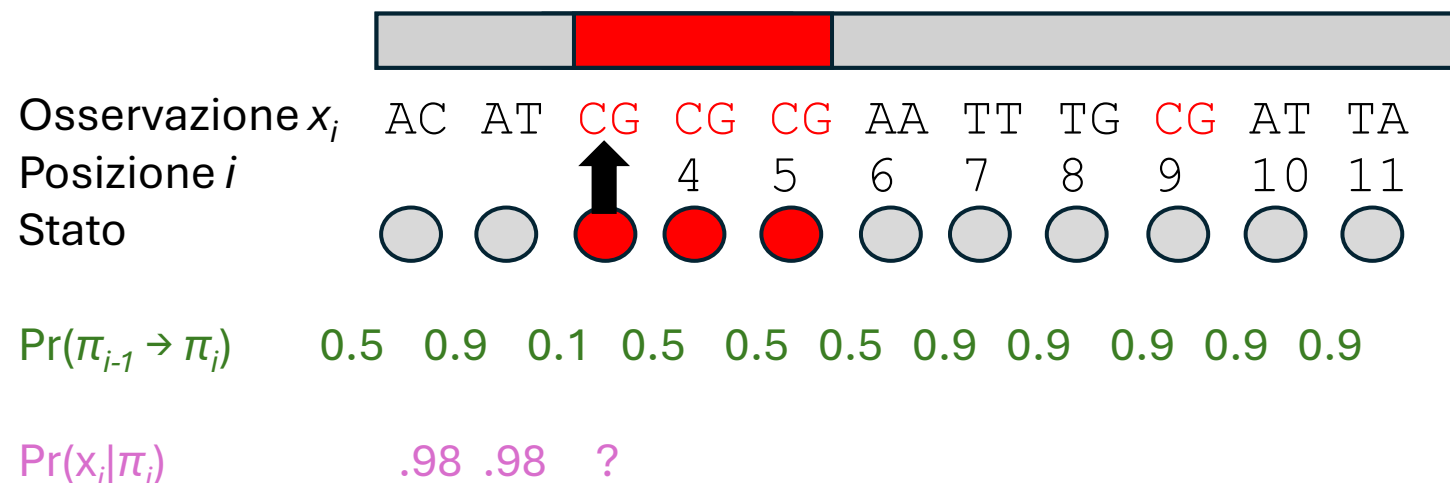


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

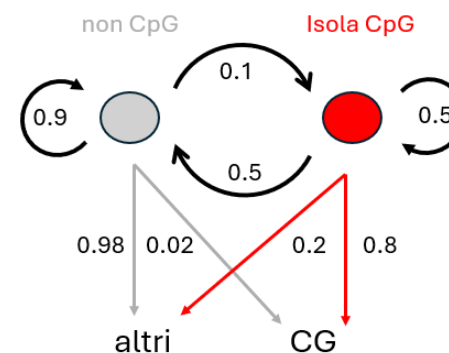
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

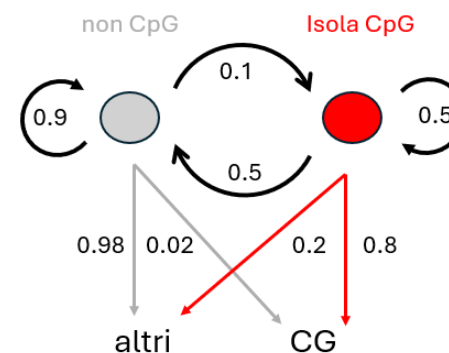
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i					5	6	7	8	9	10	11
Stato											
$Pr(\pi_{i-1} \rightarrow \pi_i)$	0.5	0.9	0.1	0.5	0.5	0.5	0.9	0.9	0.9	0.9	0.9
$Pr(x_i \pi_i)$		.98	.98	.8	?						

Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

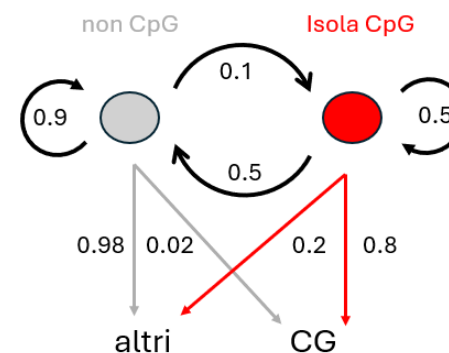
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i						6	7	8	9	10	11
Stato											
$Pr(\pi_{i-1} \rightarrow \pi_i)$	0.5	0.9	0.1	0.5	0.5	0.5	0.9	0.9	0.9	0.9	0.9
$Pr(x_i \pi_i)$		.98	.98	.8	.8	?					

Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

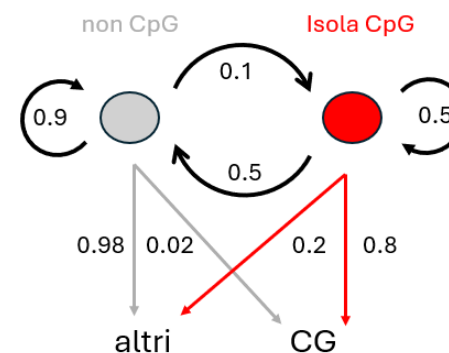
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i						6	7	8	9	10	11
Stato											
$Pr(\pi_{i-1} \rightarrow \pi_i)$	0.5	0.9	0.1	0.5	0.5	0.5	0.9	0.9	0.9	0.9	0.9
$Pr(x_i \pi_i)$		.98	.98	.8	.8	.8	?				

Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

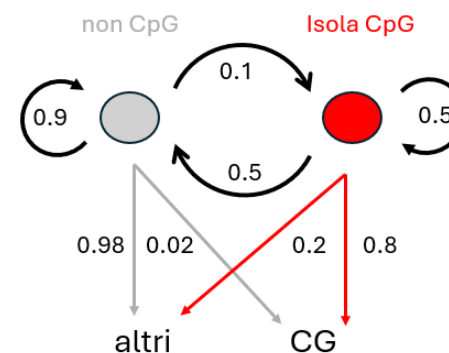
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i							7	8	9	10	11
Stato											
$Pr(\pi_{i-1} \rightarrow \pi_i)$	0.5	0.9	0.1	0.5	0.5	0.5	0.9	0.9	0.9	0.9	0.9
$Pr(x_i \pi_i)$		.98	.98	.8	.8	.8	?				

Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models

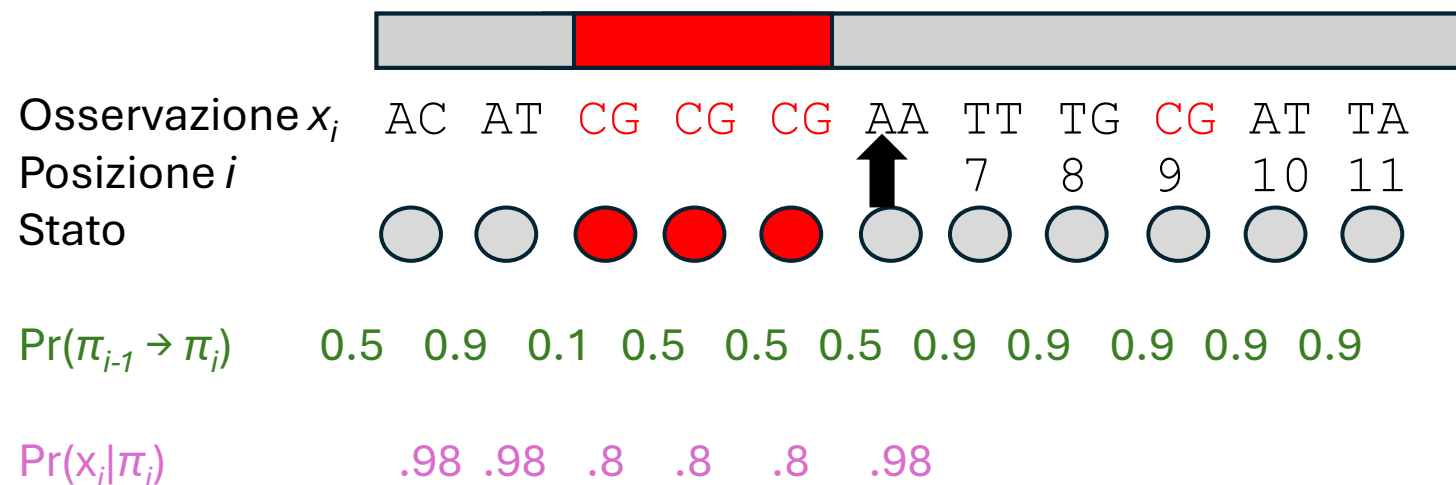


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

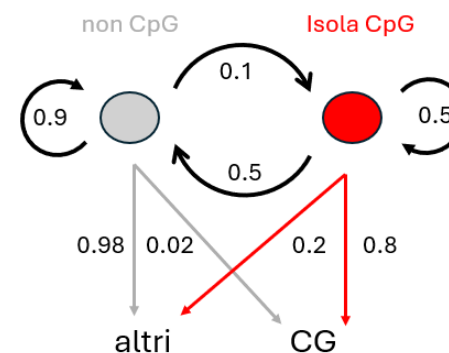
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

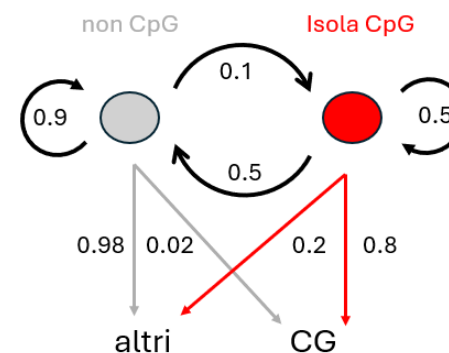
We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA	
Posizione i	1	2	3	4	5	6	7	8	9	10	11	
Stato												
$\Pr(\pi_{i-1} \rightarrow \pi_i)$	0.5	0.9	0.1	0.5	0.5	0.5	0.9	0.9	0.9	0.9	0.9	
$\Pr(x_i \pi_i)$		.98	.98	.8	.8	.8	.98	.98	.98	.02	.98	.98

Ora consideriamo le osservazioni. Per ogni posizione i , quale è la probabilità di osservare i nostri dati?



Probabilità delle osservazioni

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

$$\Pr(\pi_{i-1} \rightarrow \pi_i) \quad 0.5 \quad 0.9 \quad 0.1 \quad 0.5 \quad 0.5 \quad 0.5 \quad 0.9 \quad 0.9 \quad 0.9 \quad 0.9 \quad 0.9$$

$$\Pr(x_i | \pi_i) \quad .98 \quad .98 \quad .8 \quad .8 \quad .8 \quad .98 \quad .98 \quad .98 \quad .02 \quad .98 \quad .98$$

Quale è la probabilità di osservare la nostra sequenza di DNA x **dato** il nostro percorso/sequenza di stati π ?

$$\Pr(x|\pi) = 0.98 * 0.98 * 0.8 * 0.8 * 0.8 * \dots =$$

Probabilità delle osservazioni

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato	○	○	●	●	●	○	○	○	○	○	○

$$\Pr(\pi_{i-1} \rightarrow \pi_i) \quad 0.5 \quad 0.9 \quad 0.1 \quad 0.5 \quad 0.5 \quad 0.5 \quad 0.9 \quad 0.9 \quad 0.9 \quad 0.9 \quad 0.9$$

$$\Pr(x_i | \pi_i) \quad .98 \quad .98 \quad .8 \quad .8 \quad .8 \quad .98 \quad .98 \quad .98 \quad .02 \quad .98 \quad .98$$

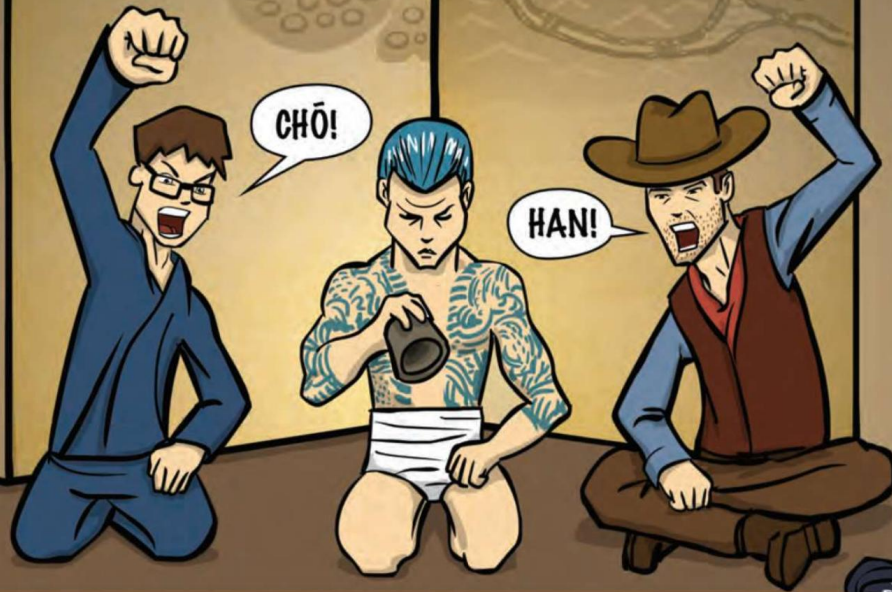
Quale è la probabilità di osservare la nostra sequenza di DNA **x** dato il nostro percorso/sequenza di stati π ?

$$\Pr(x|\pi) = 0.98 * 0.98 * 0.8 * 0.8 * 0.8 * \dots =$$

$$\Pr(x|\pi) = \prod_i \Pr(x_i | \pi_i) = \prod_i \text{Probabilità di emissione in } i \text{ per stato } \pi_i$$

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

$Pr(\pi_{i-1} \rightarrow \pi_i)$ 0.5 0.9 0.1 0.5 0.5 0.5 0.9 0.9 0.9 0.9 0.9

$Pr(x_i|\pi_i)$.98 .98 .8 .8 .8 .98 .98 .98 .02 .98 .98

Quale è la probabilità di osservare la nostra sequenza di DNA x e il nostro percorso/sequenza di stati π ?

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato	○	○	●	●	●	○	○	○	○	○	○

$Pr(\pi_{i-1} \rightarrow \pi_i)$ 0.5 0.9 0.1 0.5 0.5 0.5 0.9 0.9 0.9 0.9 0.9

$Pr(x_i|\pi_i)$.98 .98 .8 .8 .8 .98 .98 .98 .02 .98 .98

Quale è la probabilità di osservare la nostra sequenza di DNA x e il nostro percorso/sequenza di stati π ?

$$Pr(\pi) = \prod_i Pr(\pi_{i-1} \rightarrow \pi_i)$$

Probabilità di transizione

$$Pr(x|\pi) = \prod_i Pr(x_i|\pi_i)$$

Probabilità di emissione

Probabilità di osservazioni e percorso

Hidden Markov Models



Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

$\Pr(\pi_{i-1} \rightarrow \pi_i)$ 0.5 0.9 0.1 0.5 0.5 0.5 0.9 0.9 0.9 0.9 0.9

$\Pr(x_i | \pi_i)$.98 .98 .8 .8 .8 .98 .98 .98 .02 .98 .98

Quale è la probabilità di osservare la nostra sequenza di DNA x e il nostro percorso/sequenza di stati π ?

$$\Pr(\pi) = \prod_i \Pr(\pi_{i-1} \rightarrow \pi_i)$$

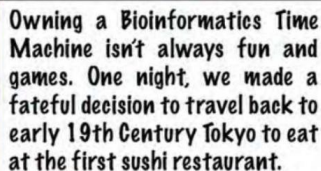
$$\Pr(x|\pi) = \prod_i \Pr(x_i | \pi_i)$$

Probabilità di transizione

Probabilità di emissione

$$\Pr(x, \pi) = \prod_i \Pr(x_i | \pi_i) \Pr(\pi_i \rightarrow \pi_{i+1})$$

Hidden Markov Models



We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!...

--	--	--

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

$\Pr(\pi_{i-1} \rightarrow \pi_i)$	0.5	0.9	0.1	0.5	0.5	0.5	0.9	0.9	0.9	0.9	0.9
------------------------------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

$\Pr(x_j \pi_j)$	.98	.98	.8	.8	.8	.98	.98	.98	.02	.98	.98
------------------	-----	-----	----	----	----	-----	-----	-----	-----	-----	-----

Quale è la probabilità di osservare la nostra sequenza di DNA x e il nostro percorso/sequenza di stati π ?

$$Pr(\pi) = \prod_i \Pr(\pi_{i-1} \rightarrow \pi_i)$$

$$Pr(x|\pi) = \prod_j Pr(x_j | \pi_j)$$

Probabilità di transizione

Probabilità di emissione

$$Pr(x, \pi) = 0.5 * 0.98 * 0.9 * 0.98 * 0.1 * 0.8 \dots$$

Posizione i	1	2	3
---------------	---	---	---

Probabilità di osservazioni e percorso

Hidden Markov Models



Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

$\Pr(\pi_{i-1} \rightarrow \pi_i)$ 0.5 0.9 0.1 0.5 0.5 0.5 0.9 0.9 0.9 0.9 0.9

$\Pr(x_i | \pi_i)$.98 .98 .8 .8 .8 .98 .98 .98 .02 .98 .98

Quale è la probabilità di osservare la nostra sequenza di DNA x e il nostro percorso/sequenza di stati π ?

$$\Pr(\pi) = \prod_i \Pr(\pi_{i-1} \rightarrow \pi_i)$$

Probabilità di transizione

$$\Pr(x|\pi) = \prod_i \Pr(x_i | \pi_i)$$

Probabilità di emissione

$$\Pr(x, \pi) = 0.5 * 0.9 * 0.1 * 0.5 * 0.5 * 0.5 * 0.9 * \dots * 0.98 * 0.98 * 0.8 * 0.8 * 0.8 \dots$$

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

$Pr(\pi_{i-1} \rightarrow \pi_i)$ 0.5 0.9 0.1 0.5 0.5 0.5 0.9 0.9 0.9 0.9 0.9

$Pr(x_i|\pi_i)$.98 .98 .8 .8 .8 .98 .98 .98 .02 .98 .98

Quale è la probabilità di osservare la nostra sequenza di DNA x e il nostro percorso/sequenza di stati π ?

$Pr(\pi)=0.5*0.9*0.1*0.5*0.5*0.5*0.9...$ $Pr(x|\pi)=0.98*0.98*0.8*0.8*0.8...$

Probabilità di transizione

Probabilità di emissione



$Pr(x,\pi)=0.5*0.9*0.1*0.5*0.5*0.5*0.9...*0.98*0.98*0.8*0.8*0.8...$

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Fingiamo di conoscere le probabilità di transizione e emissione.
Calcoliamo le probabilità di osservare un percorso π .



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

$$\Pr(\pi_{i-1} \rightarrow \pi_i) \quad 0.5 \quad 0.9 \quad 0.1 \quad 0.5 \quad 0.5 \quad 0.5 \quad 0.9 \quad 0.9 \quad 0.9 \quad 0.9 \quad 0.9$$

$$\Pr(x_i | \pi_i) \quad .98 \quad .98 \quad .8 \quad .8 \quad .8 \quad .98 \quad .98 \quad .98 \quad .02 \quad .98 \quad .98$$

Quale è la probabilità di osservare la nostra sequenza di DNA x e il nostro percorso/sequenza di stati π ?

$$\Pr(\pi) = 0.5 * 0.9 * 0.1 * 0.5 * 0.5 * 0.5 * 0.9 \dots \quad \Pr(x | \pi) = 0.98 * 0.98 * 0.8 * 0.8 * 0.8 \dots$$

Probabilità di transizione

Probabilità di emissione



$$\Pr(x, \pi) = 0.5 * 0.9 * 0.1 * 0.5 * 0.5 * 0.5 * 0.9 \dots * 0.98 * 0.98 * 0.8 * 0.8 * 0.8 \dots$$

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)



$$Pr(\pi) = \prod_i Pr(\pi_{i-1} \rightarrow \pi_i)$$

Probabilità di transizione

$$Pr(x|\pi) = \prod_i Pr(x_i | \pi_i)$$

Probabilità di emissione

$$Pr(x, \pi) = \prod_i Pr(x_i | \pi_i) Pr(\pi_{i-1} \rightarrow \pi_i)$$

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)



Questa probabilità viene anche detta la **likelihood** di un percorso!

$$Pr(x, \pi) = \prod_i Pr(x_i | \pi_i) Pr(\pi_{i-1} \rightarrow \pi_i)$$

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)

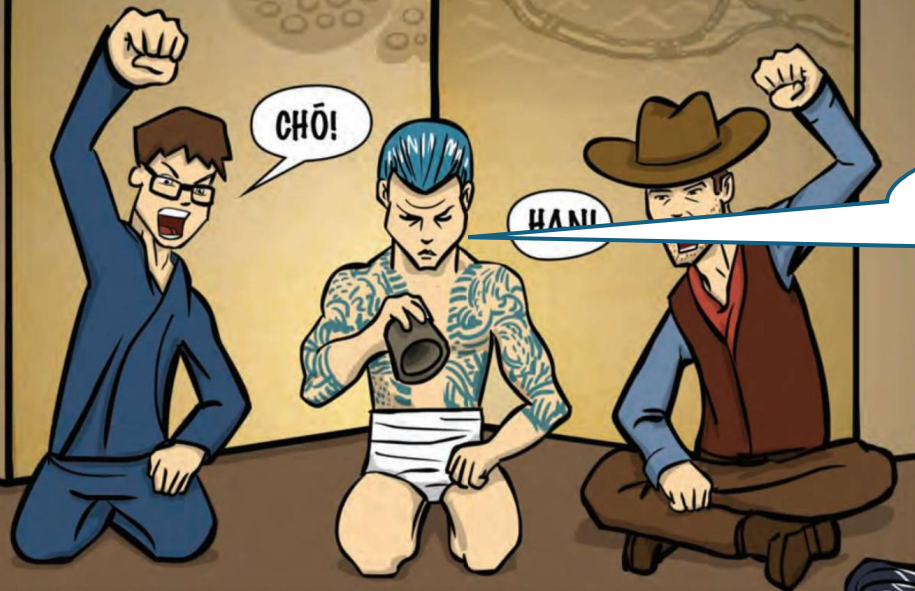


Questa probabilità viene anche detta la **likelihood** di un percorso!

$$Pr(x, \pi) = \prod_i Pr(\text{emissione } x_i | \pi_i) Pr(\text{transizione } \pi_{i-1} \rightarrow \pi_i)$$

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)

Ma come possiamo usarla?

Questa probabilità viene anche detta la **likelihood** di un percorso!

$$Pr(x, \pi) = \prod_i Pr(\text{emissione } x_i | \pi_i) Pr(\text{transizione } \pi_{i-1} \rightarrow \pi_i)$$

Hidden Markov Models



How were we supposed to know that the casino was run by yakuza?!?!..

(dato un modello, cioè delle probabilità di transizione ed emissione)



Stato

AC AT **CG** **CG** **CG** AA TT TG **CG** AT TA

1 2 3 4 5 6 7 8 9 10 11



Probabilità di osservazioni e percorso

Hidden Markov Models



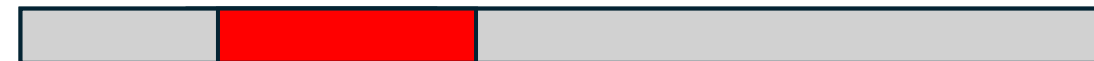
Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

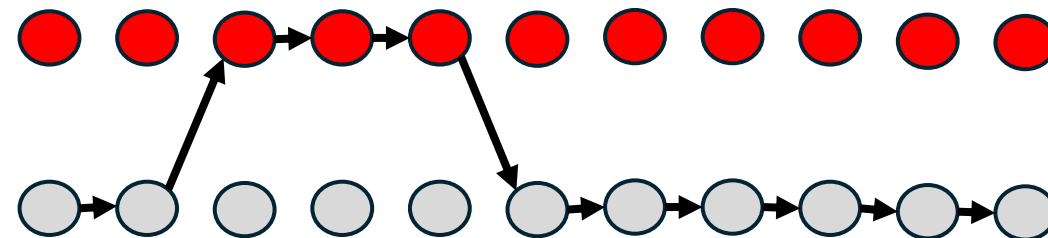
Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)



Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato	○	○	●	●	●	○	○	○	○	○	○

Rappresentazione di Viterbi



Hidden Markov Models



HAN!



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!...

(dato un modello, cioè delle probabilità di transizione ed emissione)



AC

AT

CG

CG

CG

AA

II

TG

CG

AT

TA

Posizione i

1

2

3

4

5

6

7

8

9

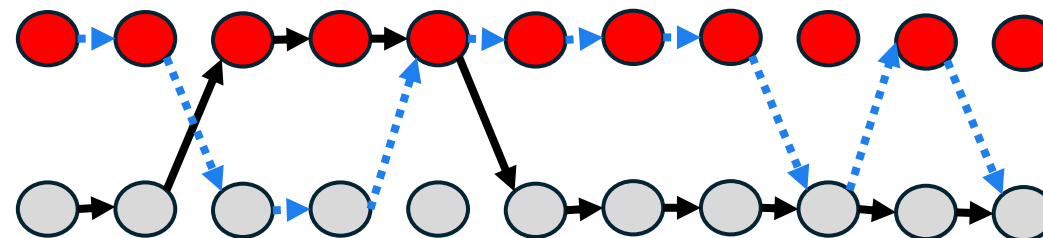
1

1

Stato



Rappresentazione di Viterbi



Questo è solo uno dei possibili **2ⁿ** percorsi!

Probabilità di osservazioni e percorso

Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

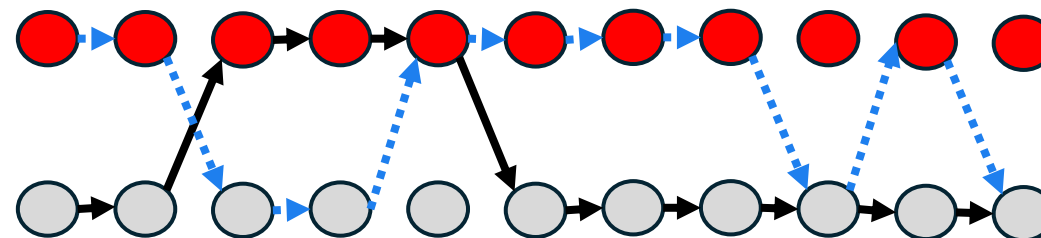
How were we supposed to know that the casino was run by yakuza?!?!

Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

Rappresentazione di Viterbi



Questo è solo uno dei possibili 2^n percorsi!

..e noi per ognuno di questi possiamo calcolarne la probabilità (likelihood)!

$$Pr(x, \pi) = \prod_i Pr(\text{emissione } x_i | \pi_i) Pr(\text{transizione } \pi_{i-1} \rightarrow \pi_i)$$

Probabilità di osservazioni e percorso

Hidden Markov Models



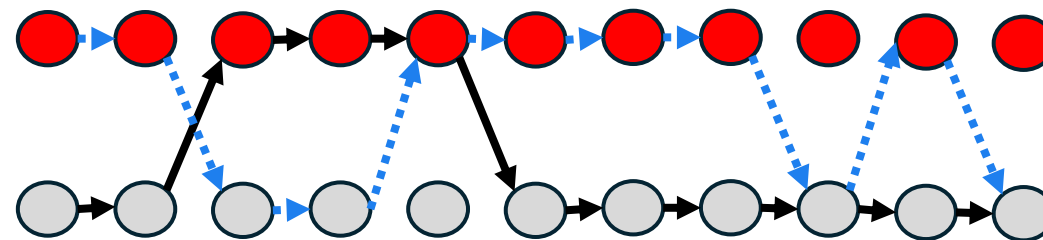
Idea, potremmo calcolare la probabilità per tutti i possibili percorsi e trovare quello giusto (likelihood piu' alta)!!!

Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

Rappresentazione di Viterbi



Questo è solo uno dei possibili 2^n percorsi!

..e noi per ognuno di questi possiamo calcolarne la probabilità (likelihood)!

$$Pr(x, \pi) = \prod_i Pr(\text{emissione } x_i | \pi_i) Pr(\text{transizione } \pi_{i-1} \rightarrow \pi_i)$$

Probabilità di osservazioni e percorso

Hidden Markov Models



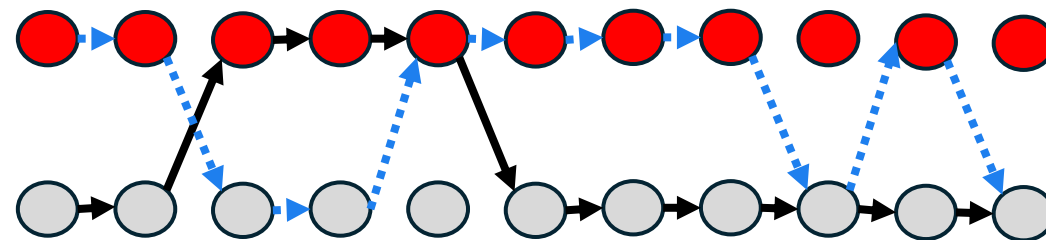
Idea, potremmo calcolare la probabilità per tutti i possibili percorsi e trovare quello giusto (likelihood piu' alta)!!!

Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)

Osservazione x_i	AC	AT	CG	CG	CG	AA	TT	TG	CG	AT	TA
Posizione i	1	2	3	4	5	6	7	8	9	10	11
Stato											

Rappresentazione di Viterbi



Questo è solo uno dei possibili 2^n percorsi!

Sono **troppi** (2^n , dove n sono il numero delle posizioni, che spesso è grandissimo)!!

Senza contare che non conosciamo ancora le vere probabilità di transizione ed emissione, quindi ci potremmo passare tutta la vita!

Probabilità di osservazioni e percorso

Hidden Markov Models

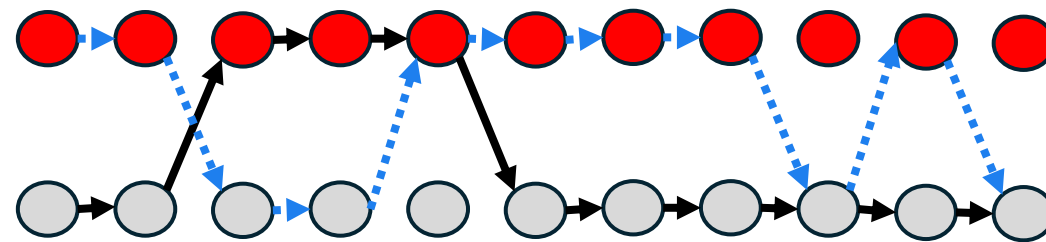


Ora possiamo calcolare la probabilità che i dati x siano generate da percorso π !

(dato un modello, cioè delle probabilità di transizione ed emissione)

Come possiamo trovare il percorso più probabile senza dover esplorare tutti i 2^n percorsi?

Rappresentazione di Viterbi



Questo è solo uno dei possibili 2^n percorsi!

Idea, potremmo calcolare la probabilità per tutti i possibili percorsi e trovare quello giusto (likelihood più alta)!!!

Sono **troppi** (2^n , dove n sono il numero delle posizioni, che spesso è grandissimo)!!

Senza contare che non conosciamo ancora le vere probabilità di transizione ed emissione, quindi ci potremmo passare tutta la vita!

Algoritmo di Viterbi

Hidden Markov Models

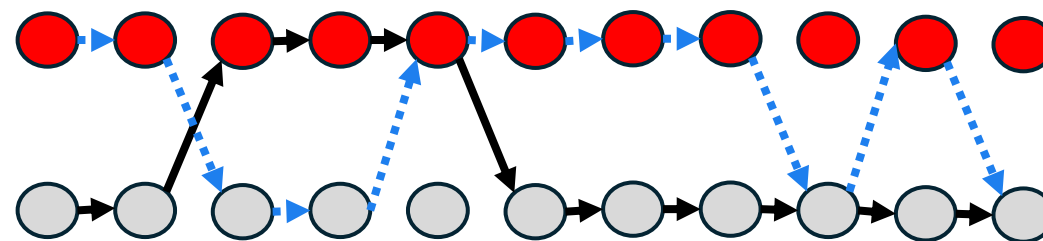


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

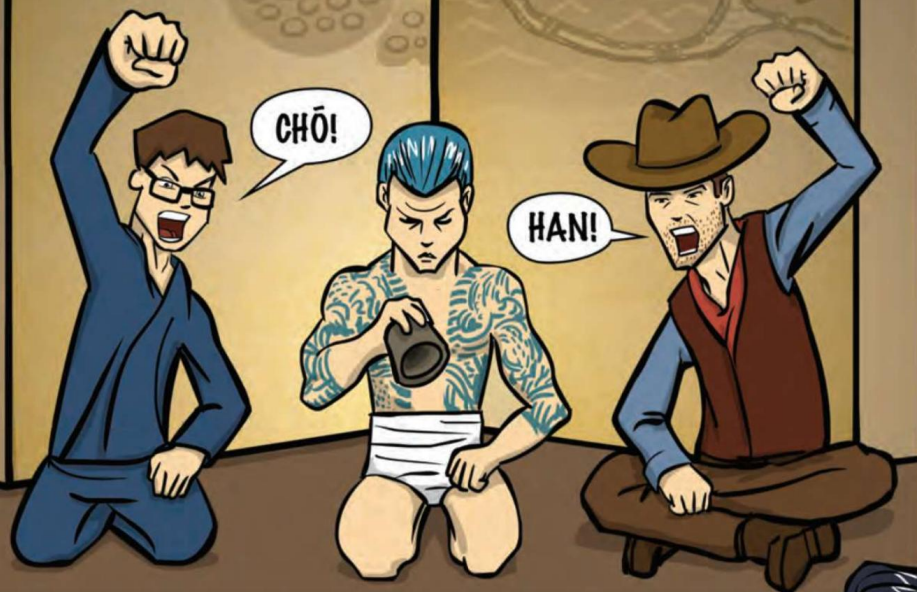
How were we supposed to know that the casino was run by yakuza?!?!..

Rappresentazione di Viterbi

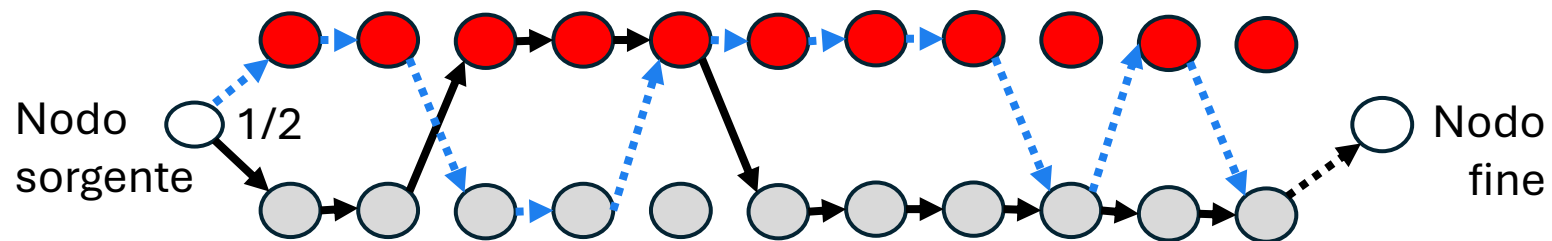


Algoritmo di Viterbi

Hidden Markov Models

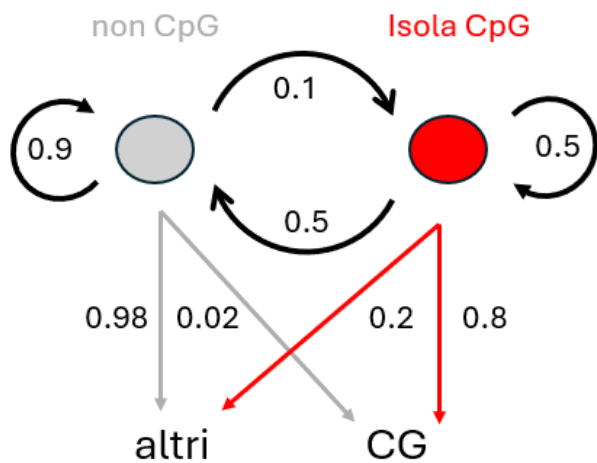


Rappresentazione di Viterbi



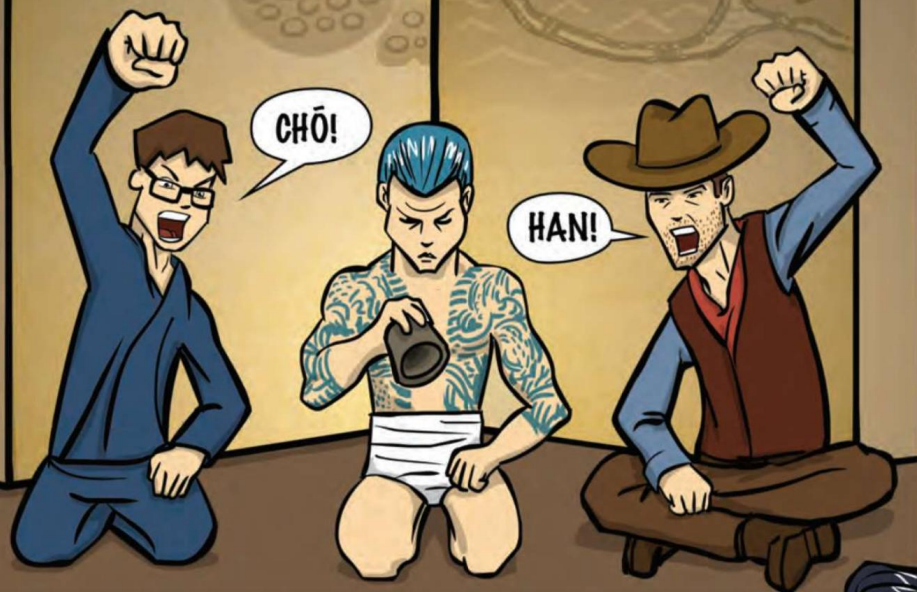
La nostra likelihood/probabilità di un percorso π

$$Pr(x, \pi) = \prod_i \text{Pr}(\text{emissione } x_i | \pi_i) \text{Pr}(\text{transizione } \pi_{i-1} \rightarrow \pi_i)$$

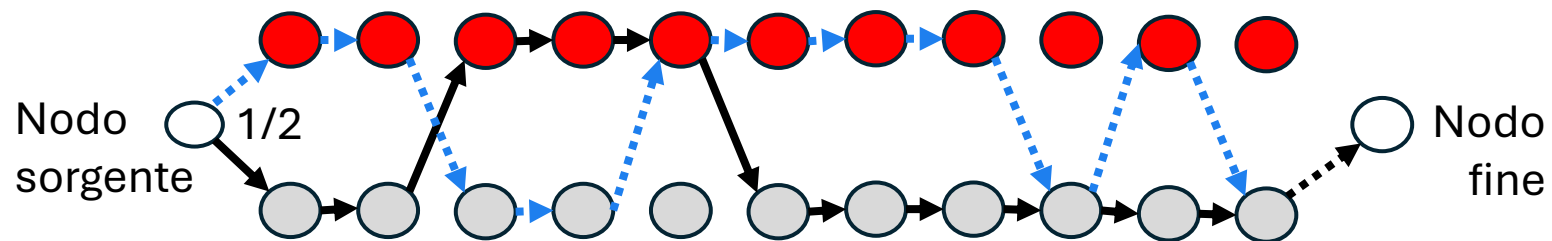


Algoritmo di Viterbi

Hidden Markov Models

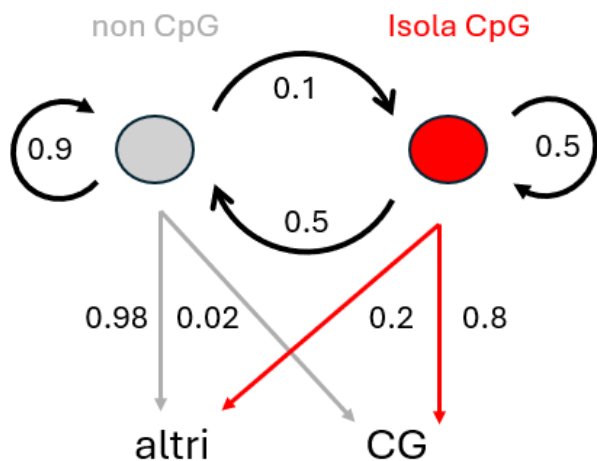


Rappresentazione di Viterbi



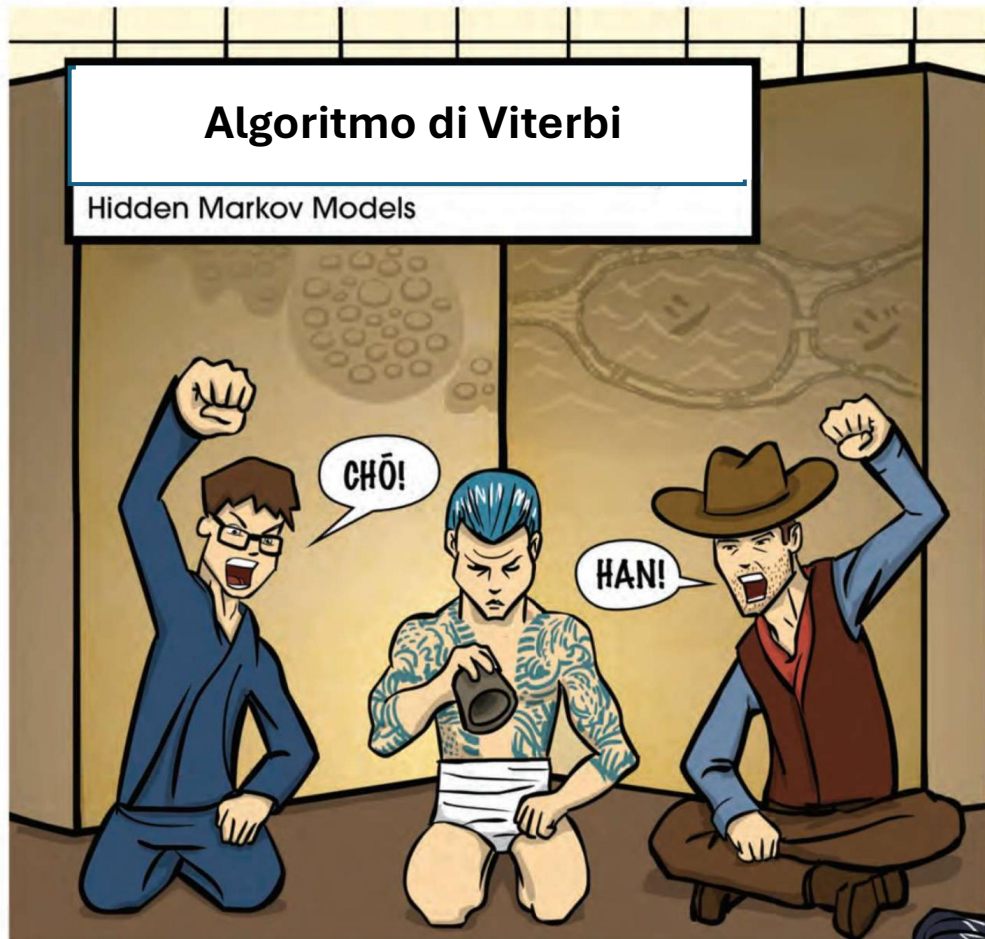
Idea intuitiva: posso provare a percorrere tutto il percorso considerando tutti gli stati e poi selezionare alla fine la traccia w piu' probabile.

$$Pr(x, \pi)_k = \underbrace{\prod_{i \leq k} \Pr(\text{emissione } x_i | \pi_i) \Pr(\text{transizione } \pi_{i-1} \rightarrow \pi_i)}_{\text{«peso» } w \text{ della traccia fino a posizione } k} = \prod_{i \leq k} w_i$$

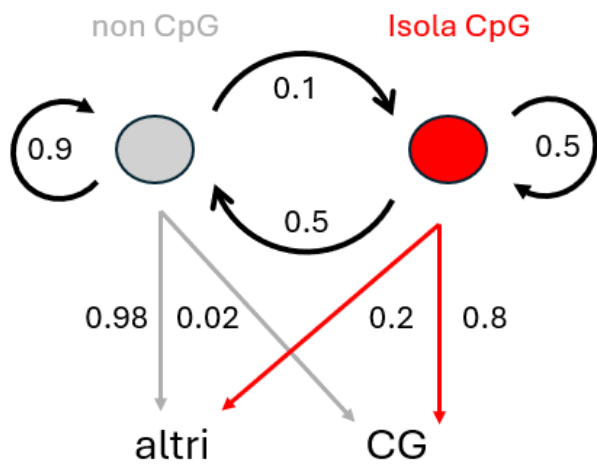
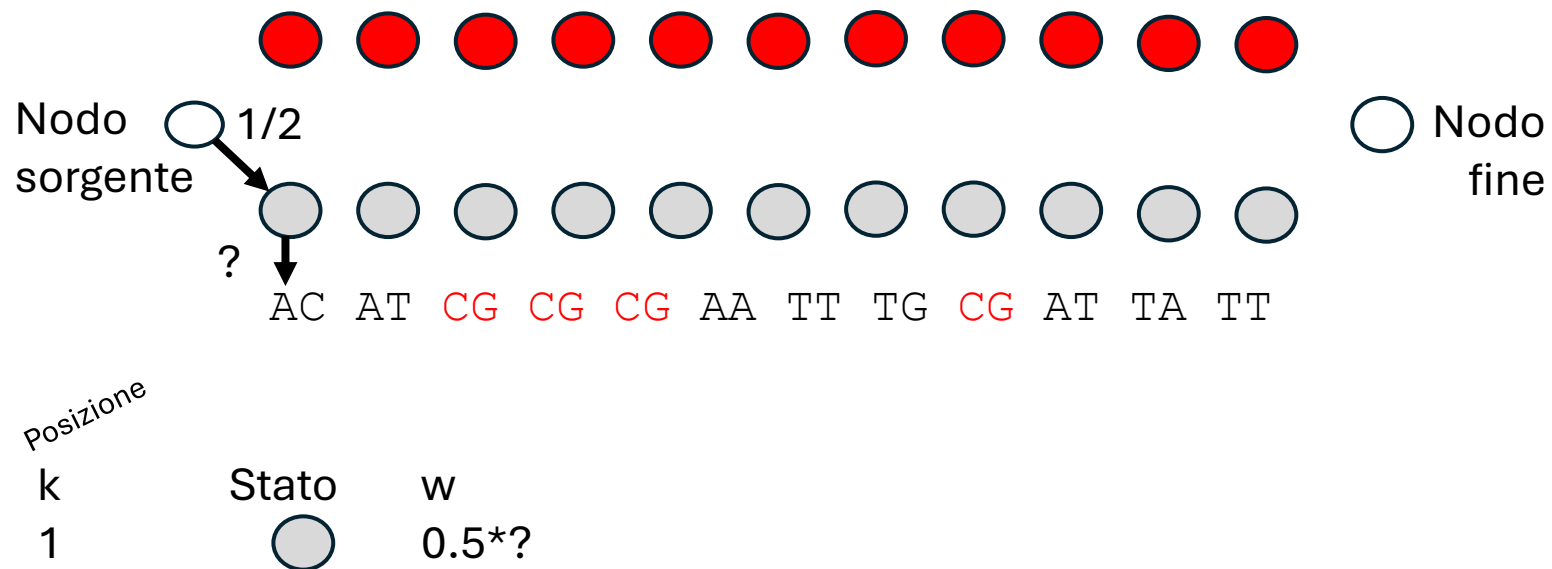


Algoritmo di Viterbi

Hidden Markov Models

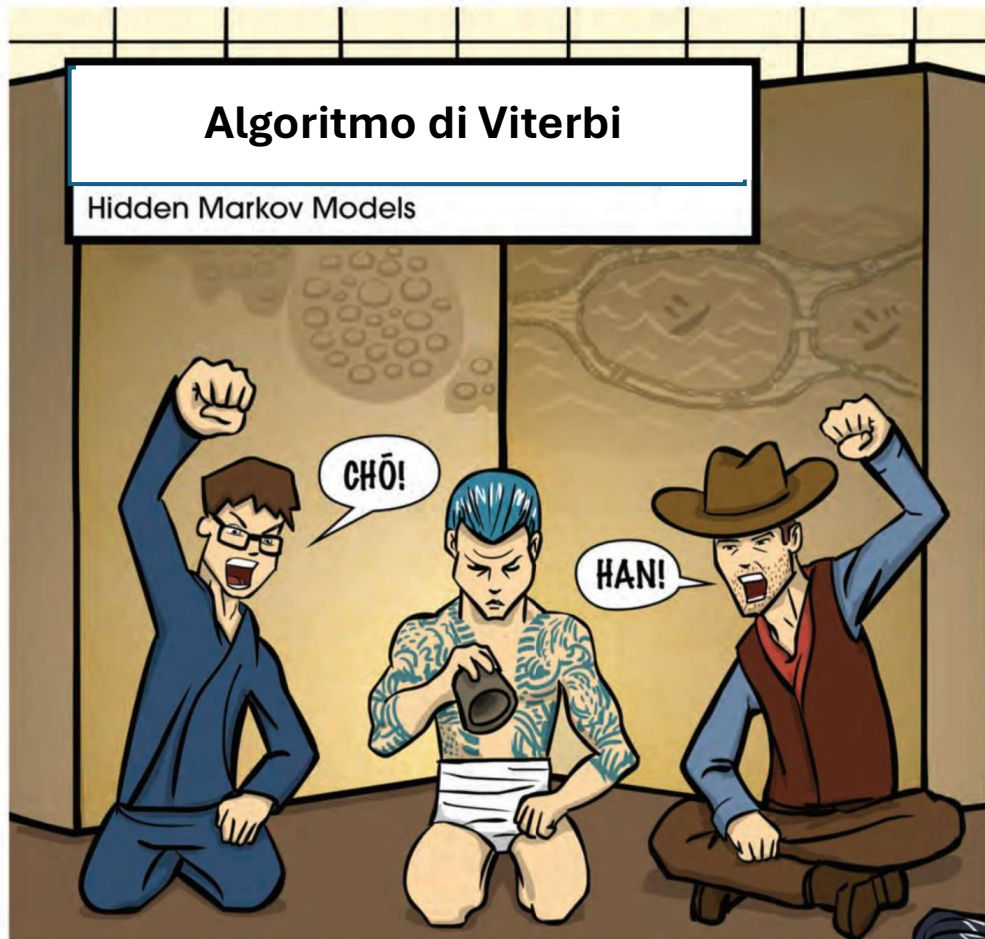


Rappresentazione di Viterbi

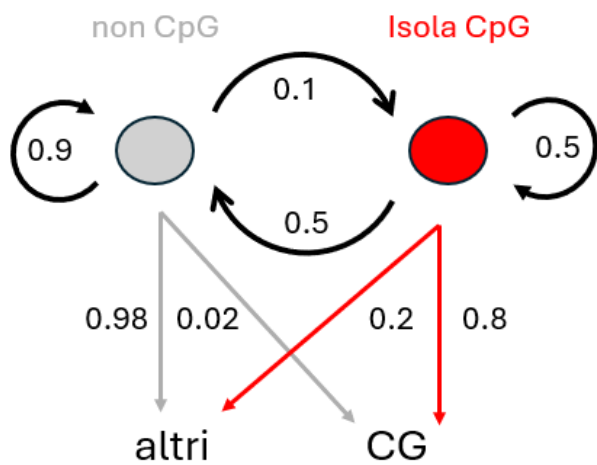
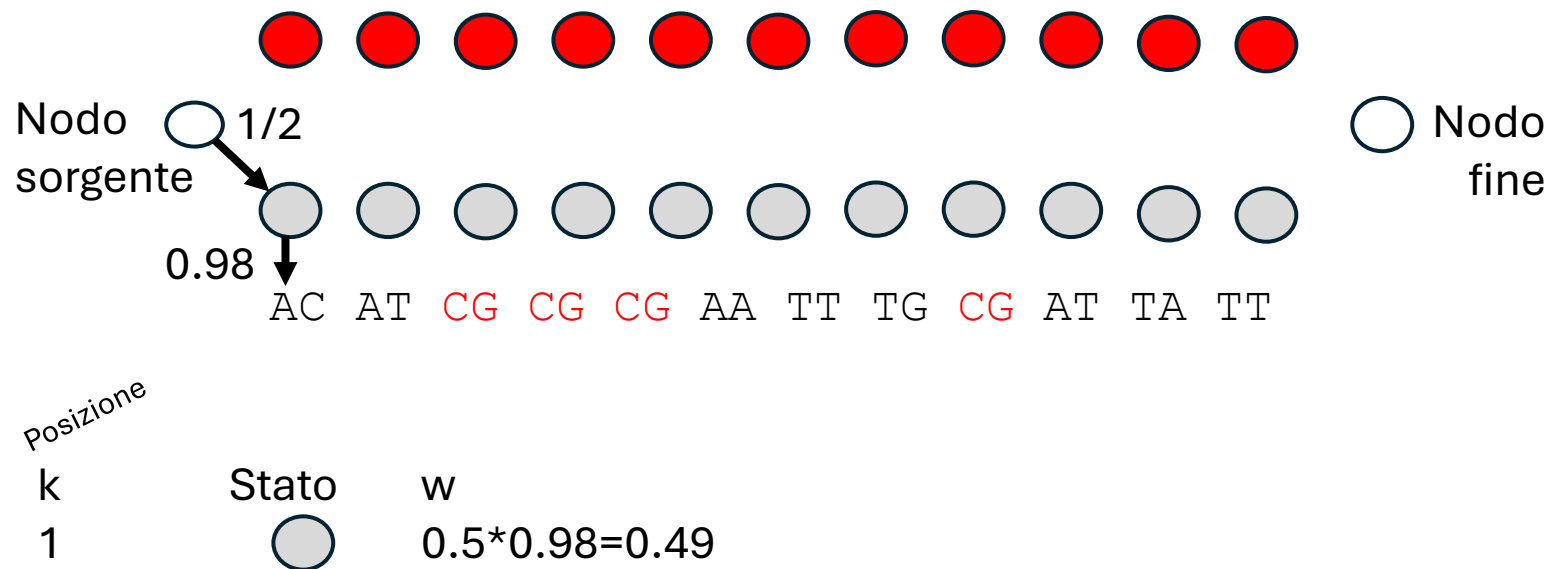


Algoritmo di Viterbi

Hidden Markov Models

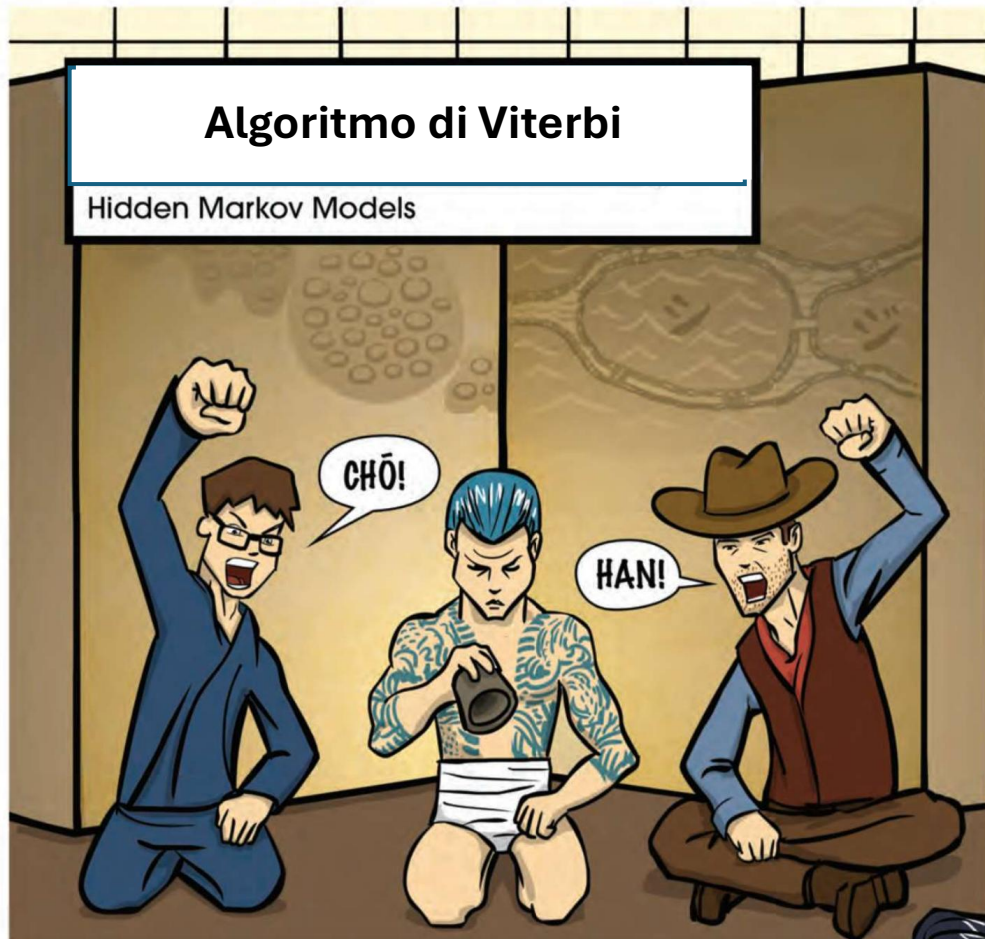


Rappresentazione di Viterbi

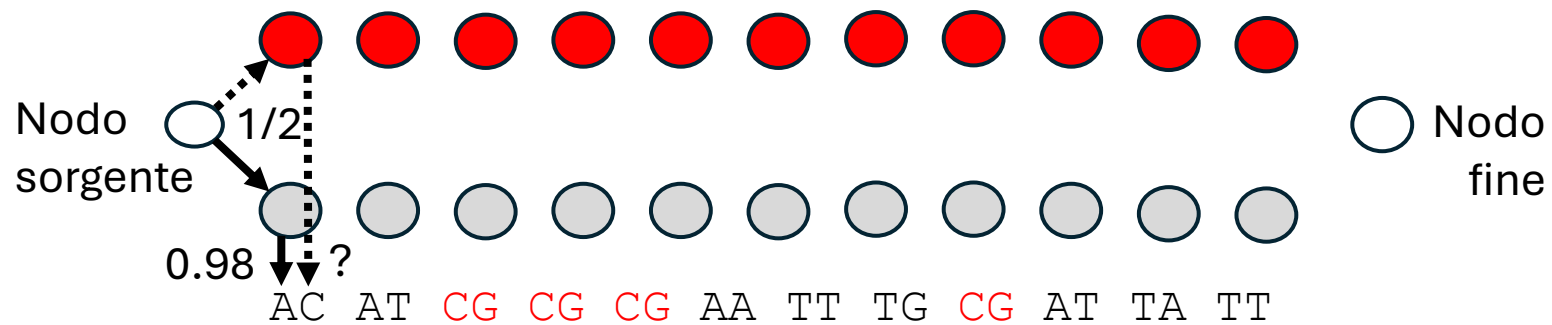


Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi



Posizione

k

1

1

Stato



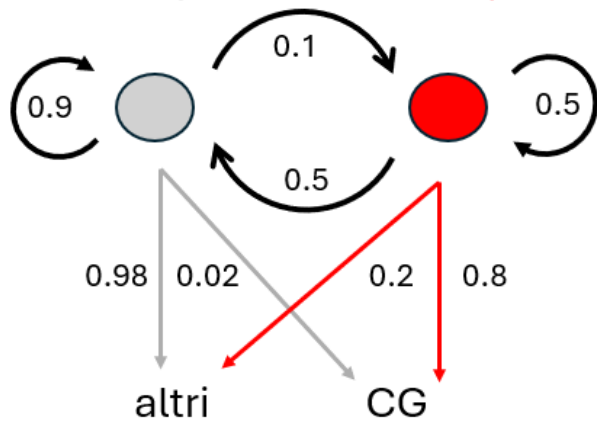
w

$$0.5 * 0.98 = 0.49$$

$$0.5 * ?$$

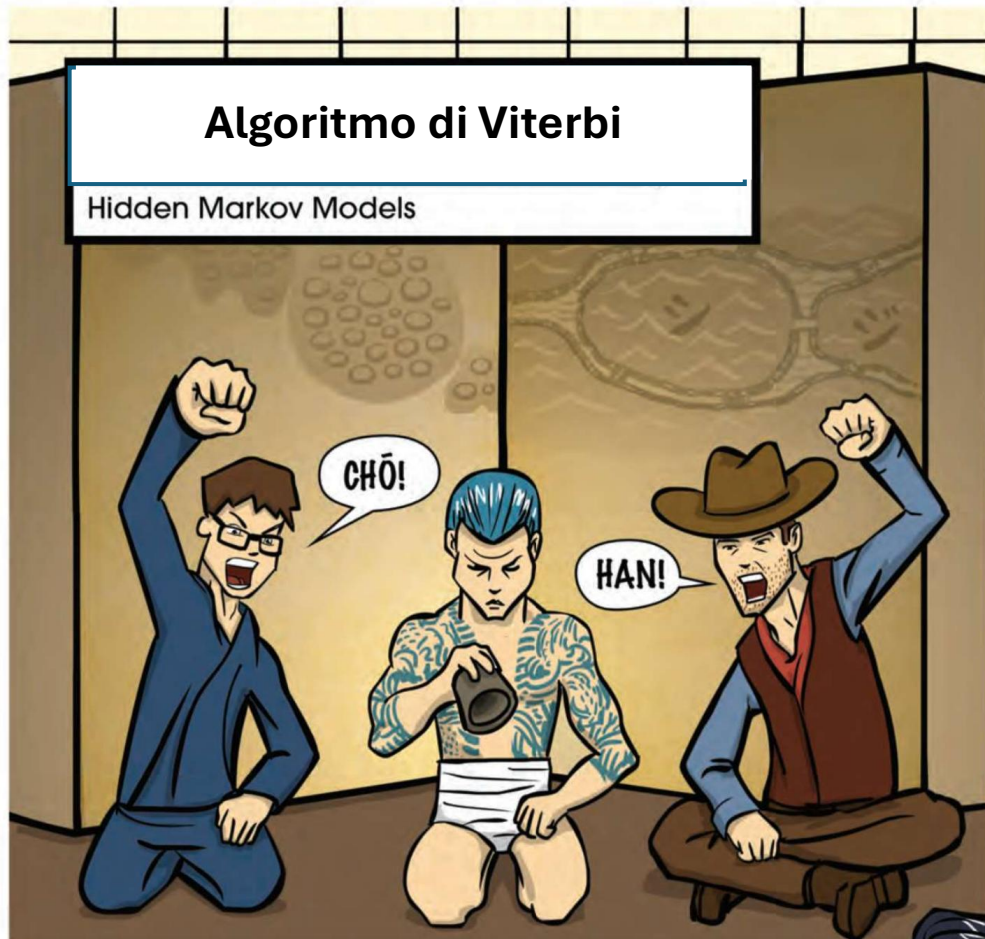
non CpG

Isola CpG

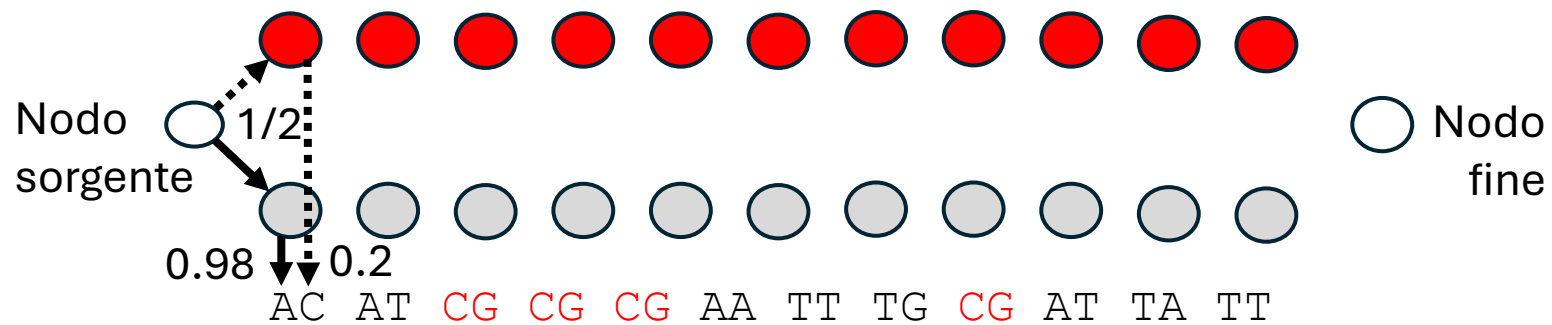


Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi



Posizione

k

Stato

w

1



$$0.5 \cdot 0.98 = 0.49$$

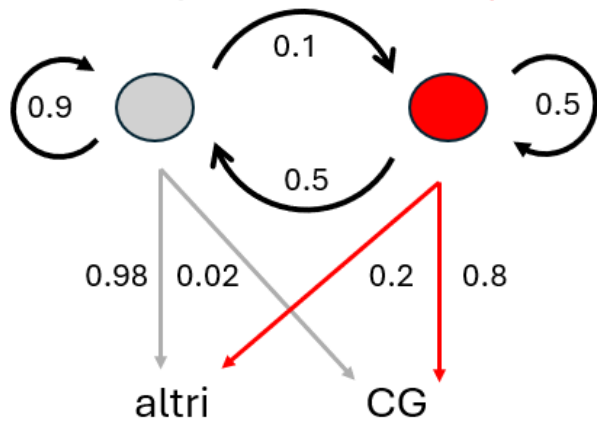
1



$$0.5 \cdot 0.2 = 0.1$$

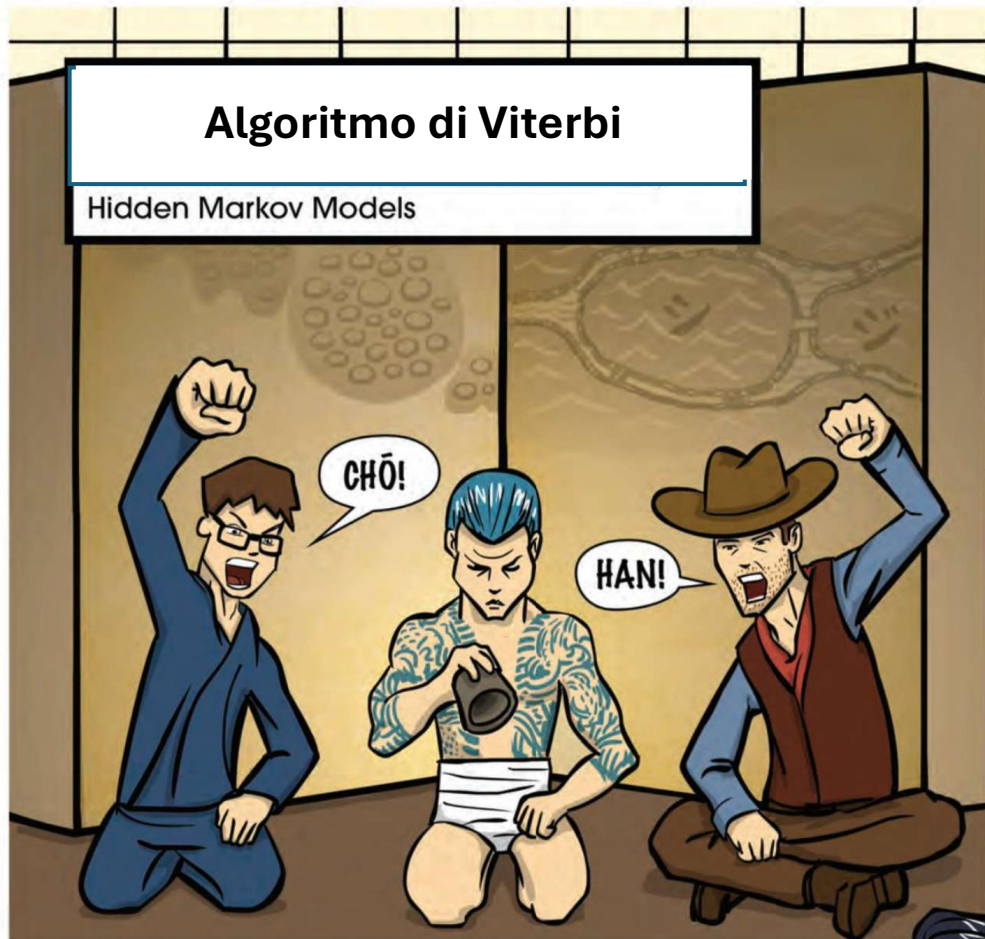
non CpG

Isola CpG

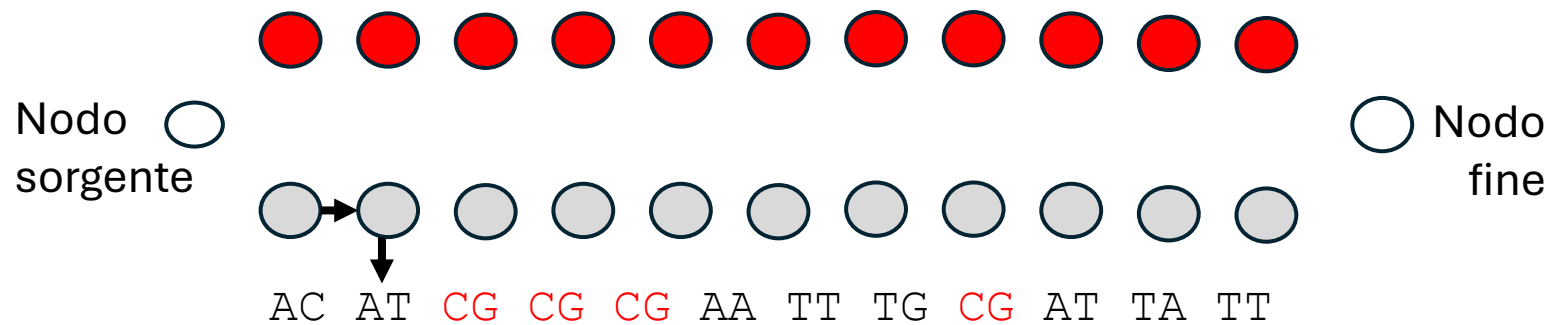


Algoritmo di Viterbi

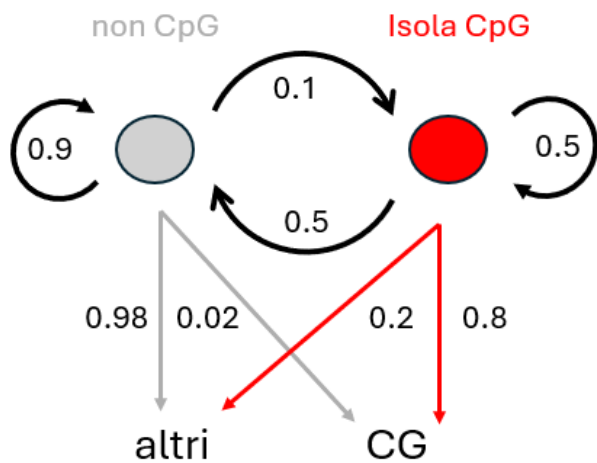
Hidden Markov Models



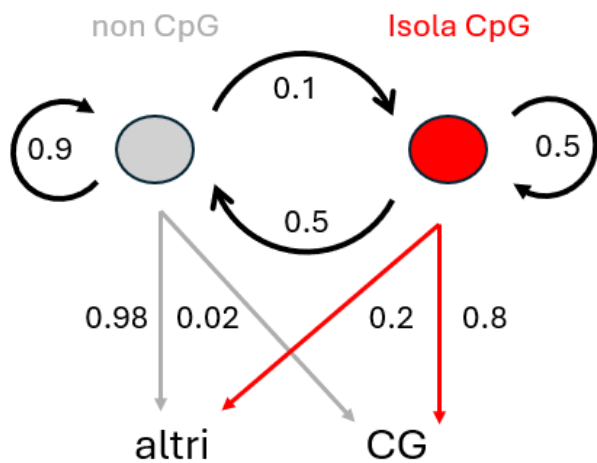
Rappresentazione di Viterbi



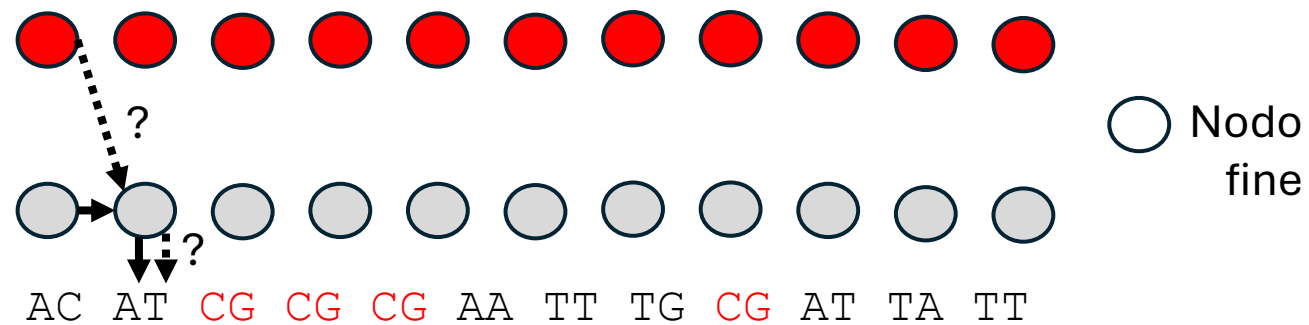
k	Stato	w	Stato precedente
1	○	0.49	-
1	●	0.1	-
2	○	$0.49 \times 0.9 \times 0.98 = 0.882$	○



Hidden Markov Models



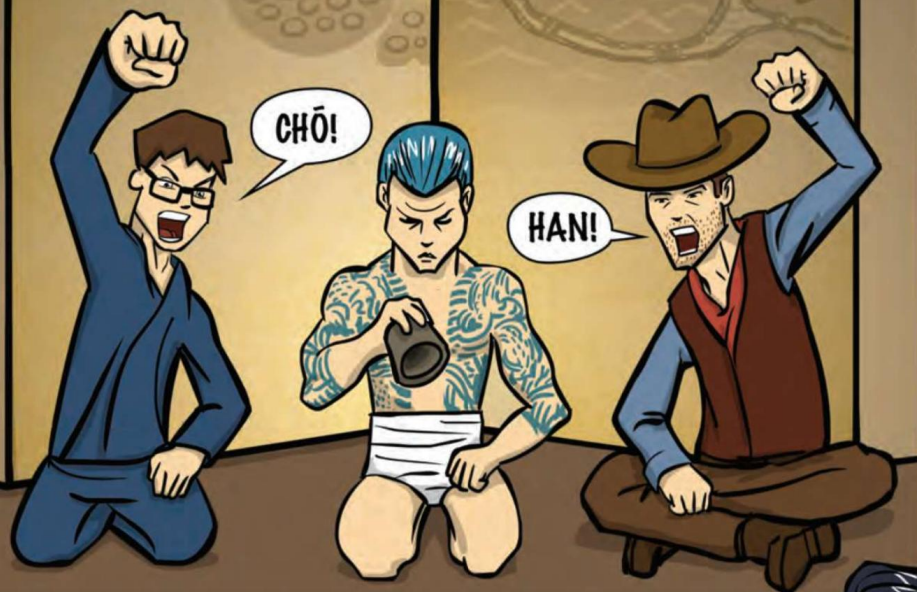
Nodo
sorgente



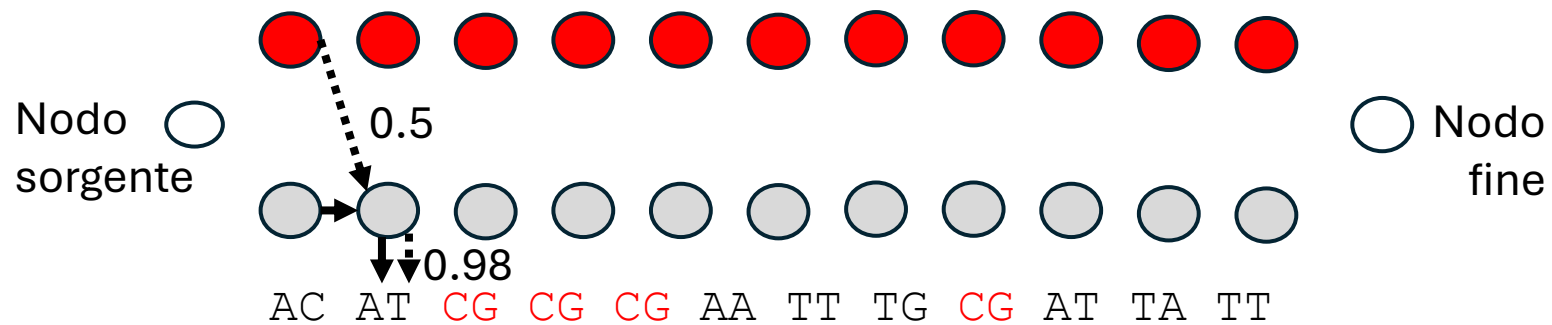
k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \cdot 0.9 \cdot 0.98 = 0.882$	
2		?	

Algoritmo di Viterbi

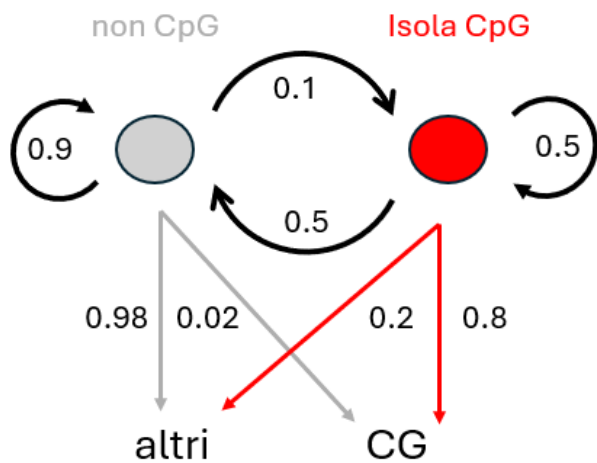
Hidden Markov Models



Rappresentazione di Viterbi

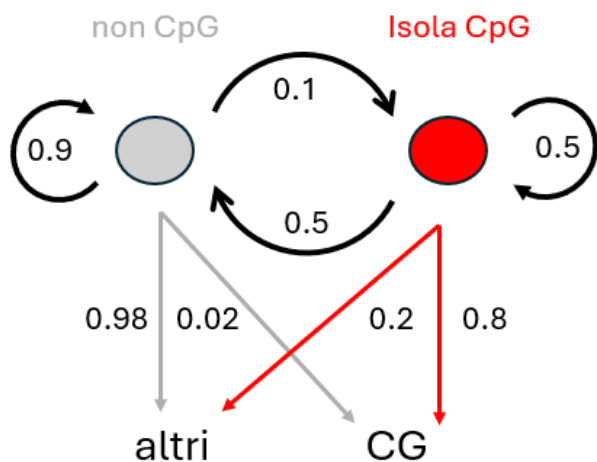
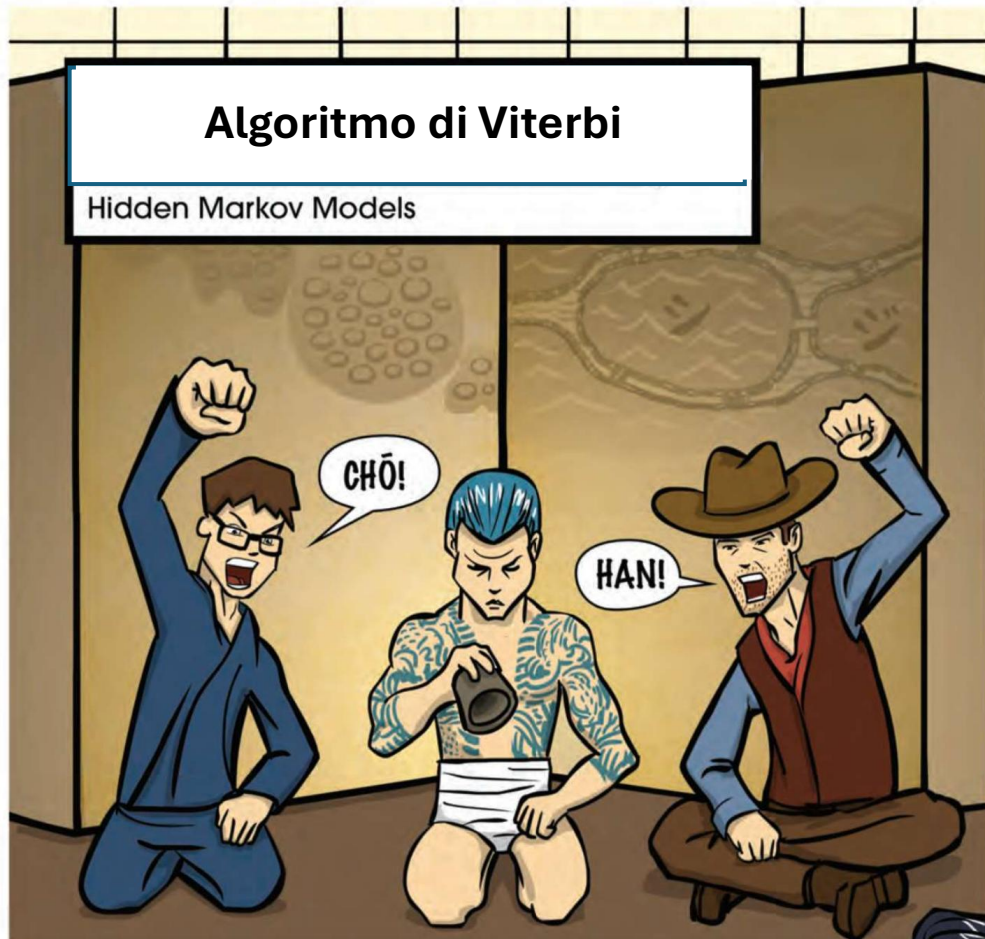


k	Stato	w	Stato precedente
1	○	0.49	-
1	●	0.1	-
2	○	$0.49 \cdot 0.9 \cdot 0.98 = 0.882$	○
2	○	$0.1 \cdot 0.5 \cdot 0.98 = 0.049$	●

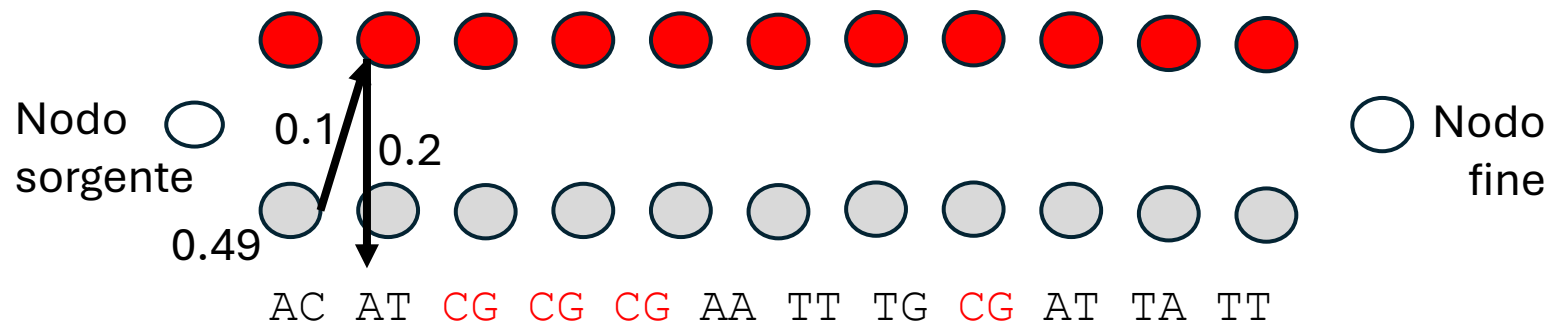


Algoritmo di Viterbi

Hidden Markov Models



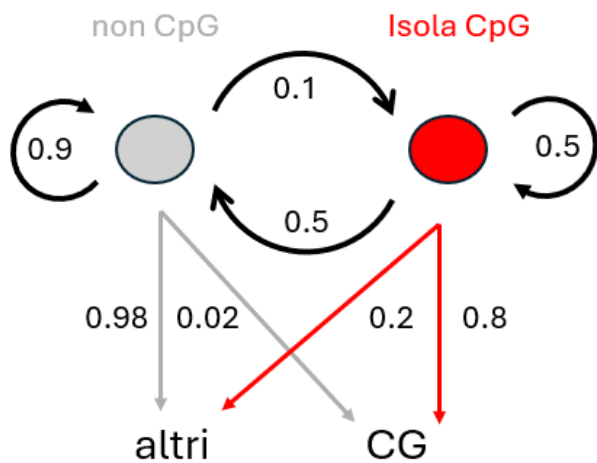
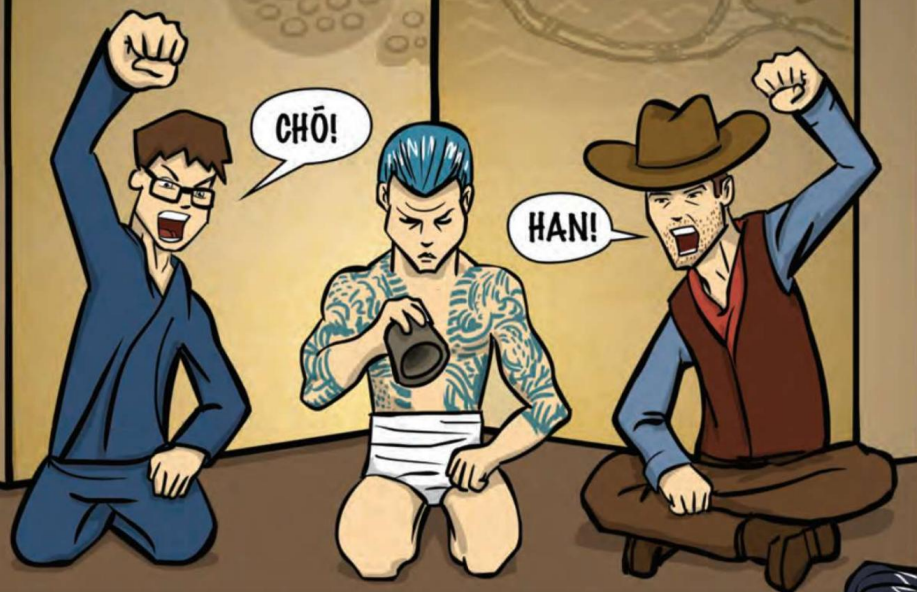
Rappresentazione di Viterbi



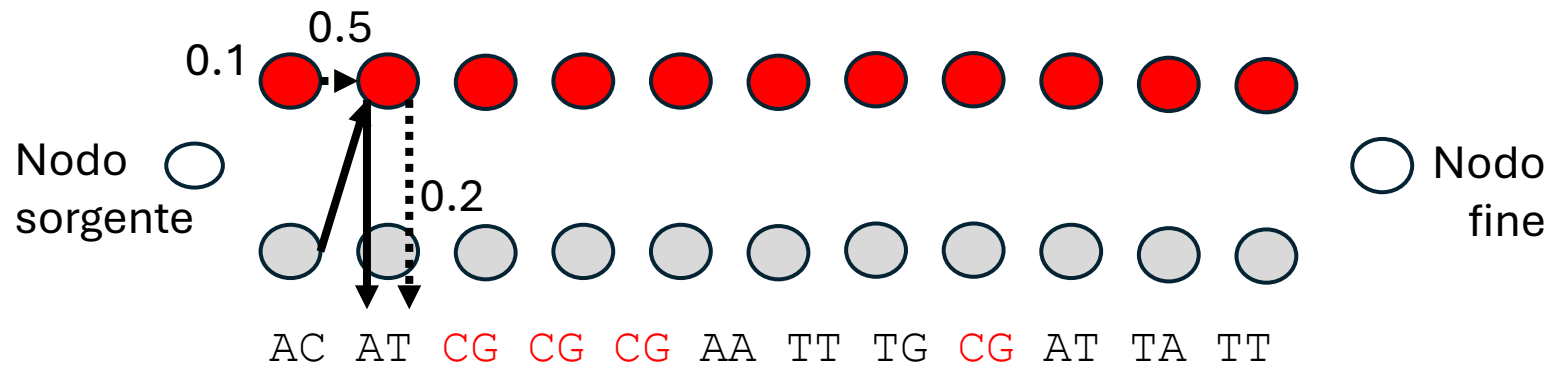
k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \times 0.9 \times 0.98 = 0.882$	
2		$0.1 \times 0.5 \times 0.98 = 0.049$	
2		$0.49 \times 0.1 \times 0.2 = 0.0098$	

Algoritmo di Viterbi

Hidden Markov Models



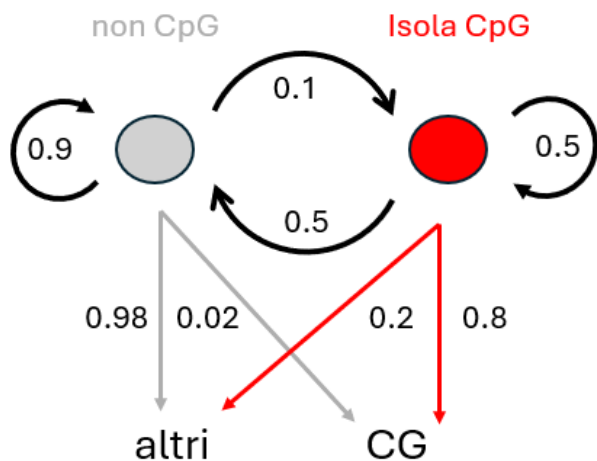
Rappresentazione di Viterbi



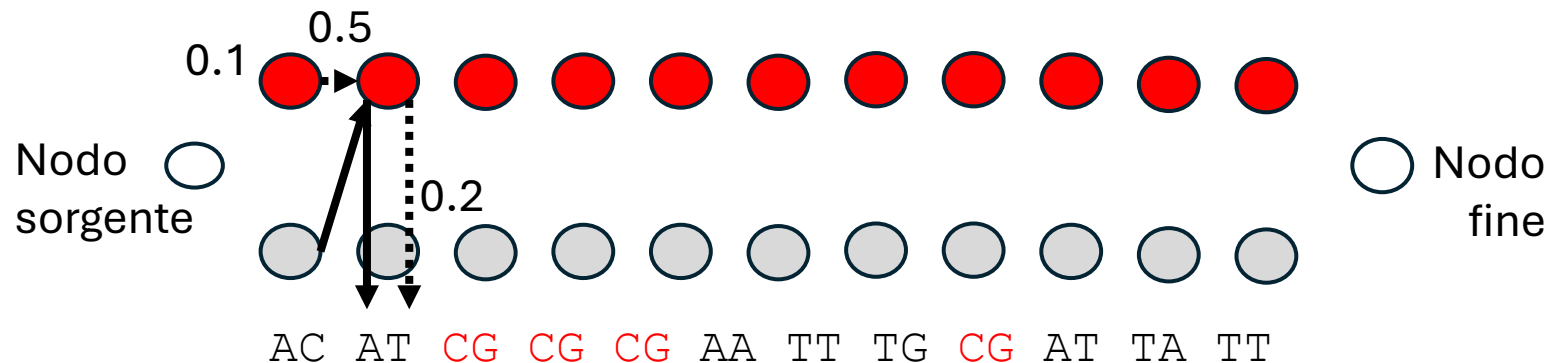
k	Stato	w	Stato precedente
1	○	0.49	-
1	●	0.1	-
2	○	$0.49 \times 0.9 \times 0.98 = 0.882$	○
2	○	$0.1 \times 0.5 \times 0.98 = 0.049$	●
2	●	$0.49 \times 0.1 \times 0.2 = 0.0098$	○
2	●	$0.1 \times 0.5 \times 0.2 = 0.01$	●

Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi

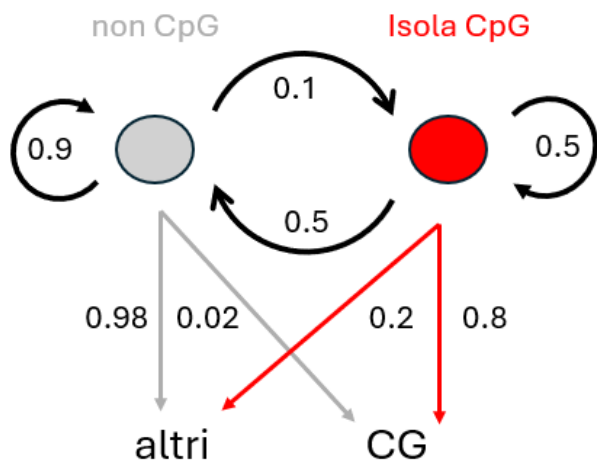
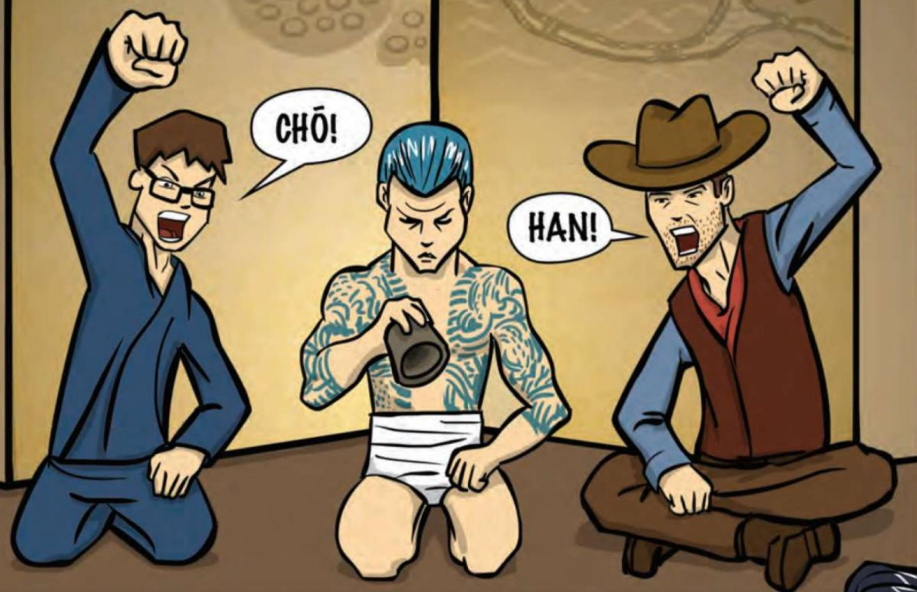


k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \cdot 0.9 \cdot 0.98 = 0.882$	
2		$0.1 \cdot 0.5 \cdot 0.98 = 0.049$	
2		$0.49 \cdot 0.1 \cdot 0.2 = 0.0098$	
2		$0.1 \cdot 0.5 \cdot 0.2 = 0.01$	

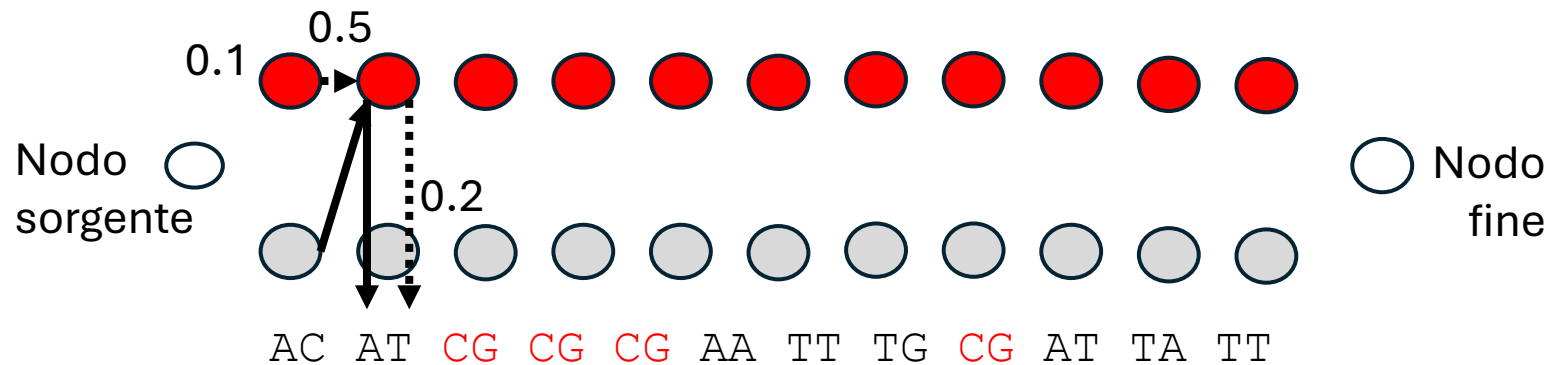
In teoria potrei continuare così fino alla fine, ma finiremmo per contare tutti i percorsi, e vorremmo evitarlo!! Allo stesso tempo, dobbiamo prima calcolare l'intera likelihood per sapere quale è il percorso più probabile (con la likelihood più alta)! Quindi semplifichiamo! Ma come?!?!?

Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi

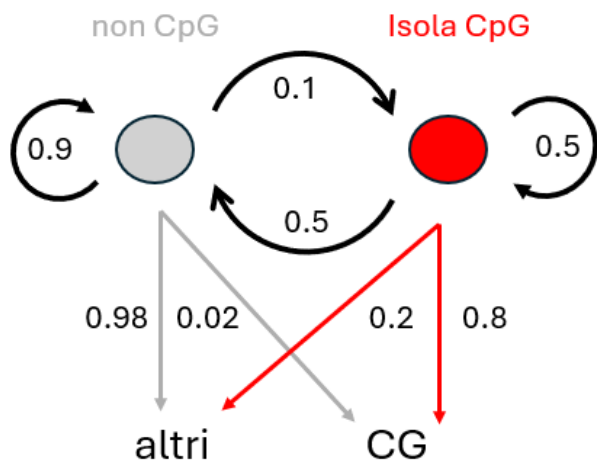


k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \times 0.9 \times 0.98 = 0.882$	
2		$0.1 \times 0.5 \times 0.98 = 0.049$	
2		$0.49 \times 0.1 \times 0.2 = 0.0098$	
2		$0.1 \times 0.5 \times 0.2 = 0.01$	

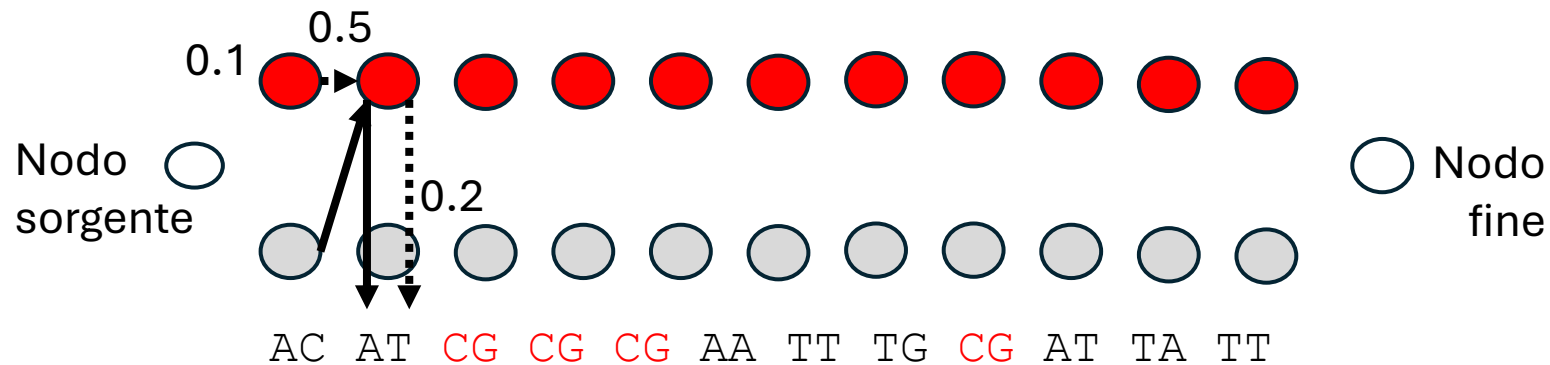
In teoria potrei continuare così fino alla fine, ma finiremmo per contare tutti i percorsi, e vorremmo evitarlo!!

Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi

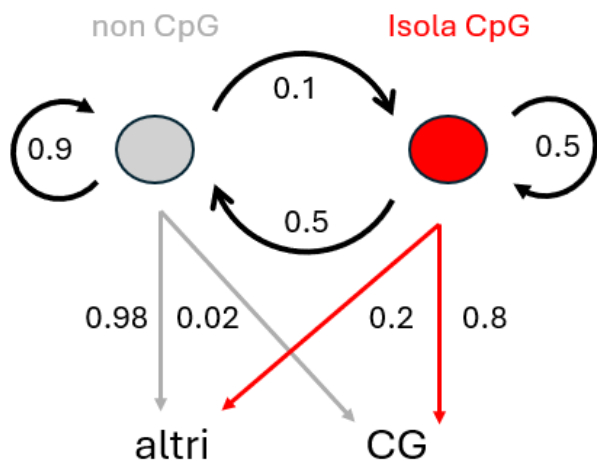
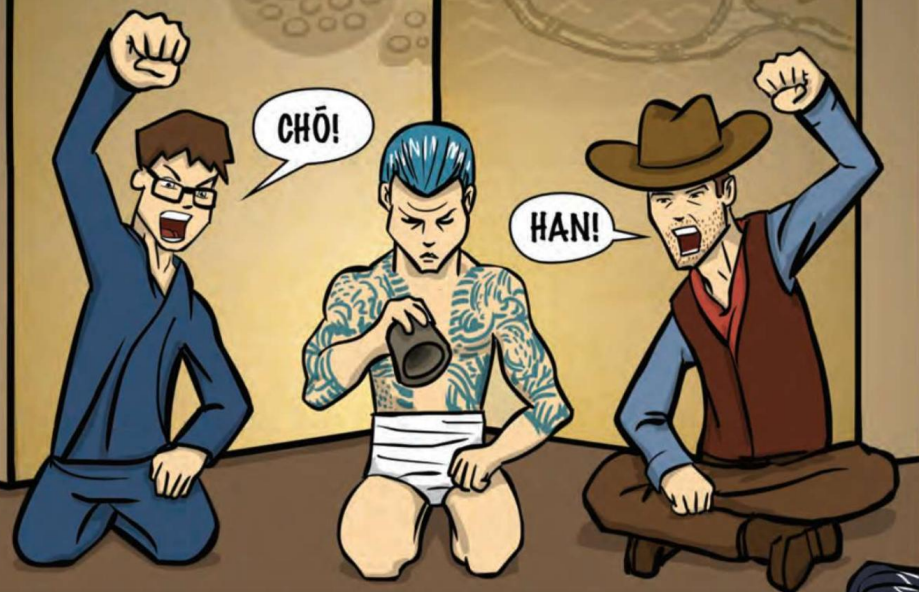


k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \cdot 0.9 \cdot 0.98 = 0.882$	
2		$0.1 \cdot 0.5 \cdot 0.98 = 0.049$	
2		$0.49 \cdot 0.1 \cdot 0.2 = 0.0098$	
2		$0.1 \cdot 0.5 \cdot 0.2 = 0.01$	

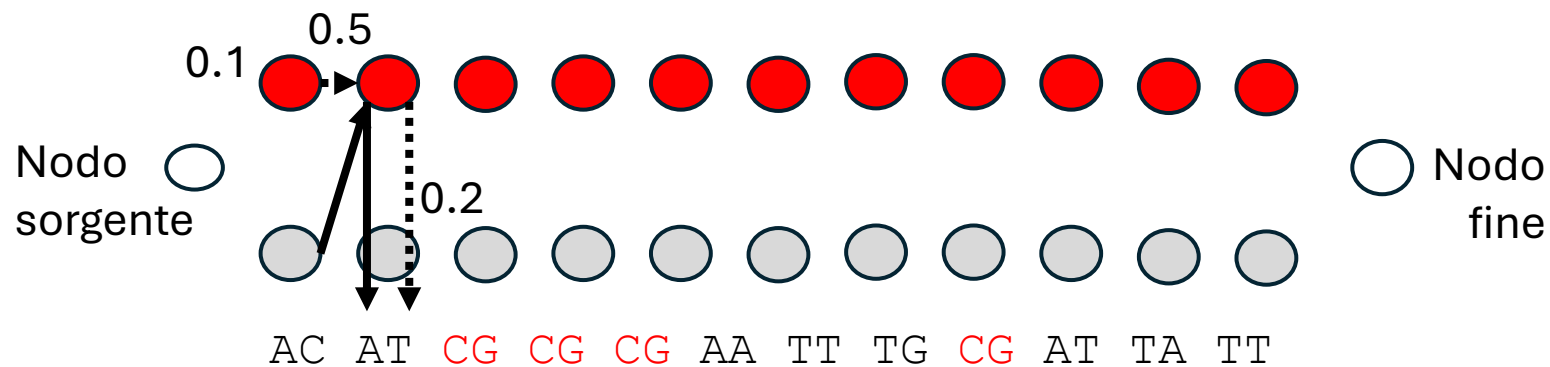
In teoria potrei continuare così fino alla fine, ma finiremmo per contare tutti i percorsi, e vorremmo evitarlo. Allo stesso tempo, dobbiamo prima calcolare l'intera likelihood per sapere quale è il percorso più probabile (con la likelihood più alta)! Quindi semplifichiamo! Ma come?!?!?

Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi



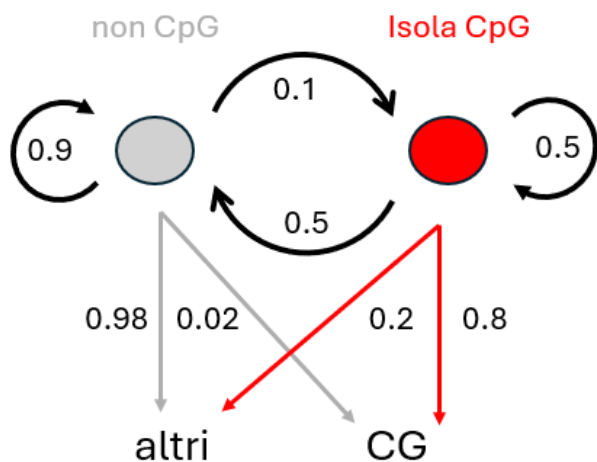
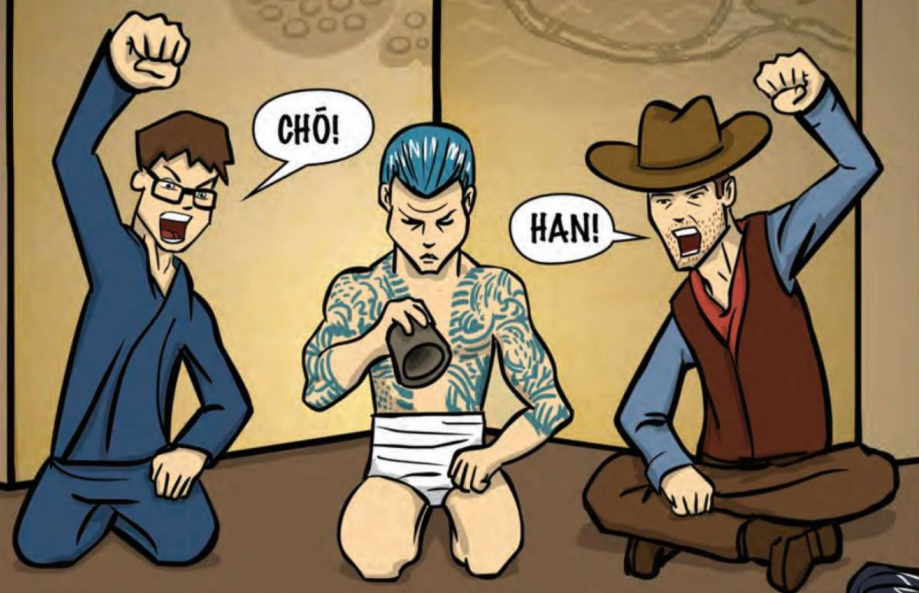
k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \cdot 0.9 \cdot 0.98 = 0.882$	
2		$0.1 \cdot 0.5 \cdot 0.98 = 0.049$	
2		$0.49 \cdot 0.1 \cdot 0.2 = 0.0098$	
2		$0.1 \cdot 0.5 \cdot 0.2 = 0.01$	

*In altre parole,
vogliamo arrivare fino alla fine per sapere quale è il percorso migliore..*

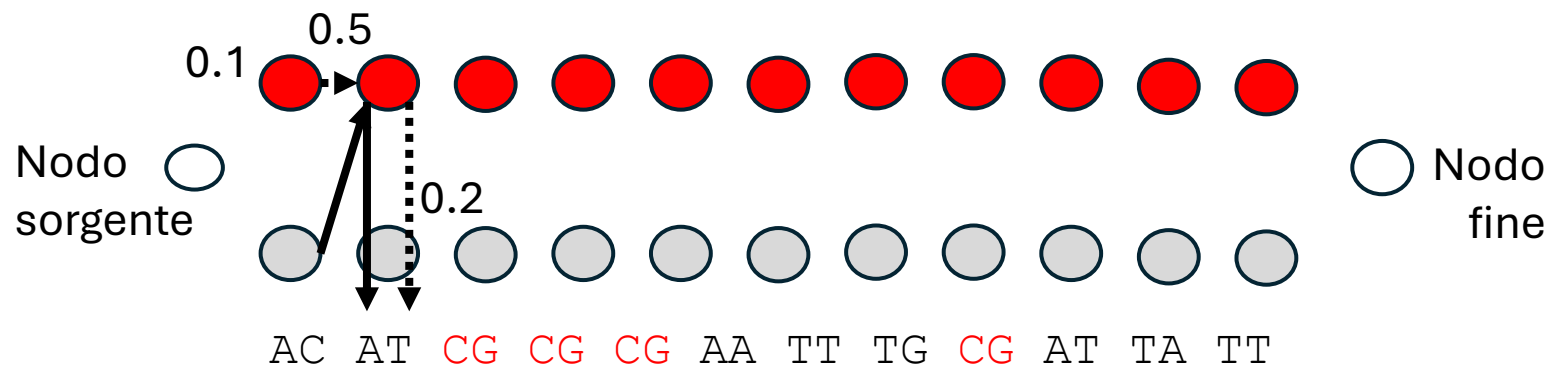
*..ma senza ricordarci tutti i possibili percorsi.
Idee?*

Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi

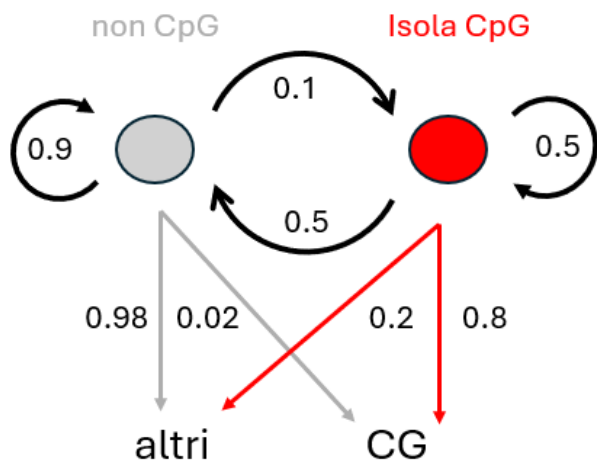
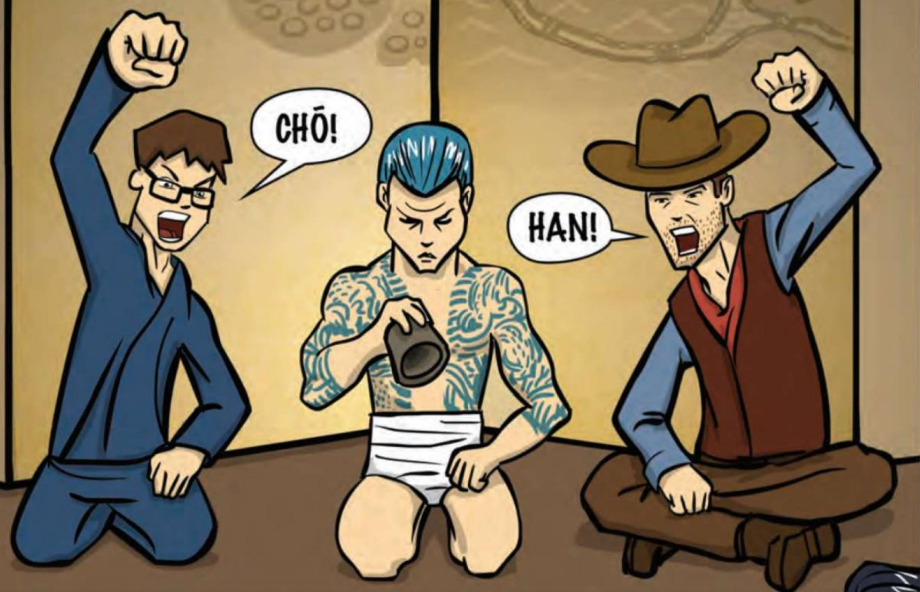


k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \times 0.9 \times 0.98 = 0.882$	
2		$0.1 \times 0.5 \times 0.98 = 0.049$	
2		$0.49 \times 0.1 \times 0.2 = 0.0098$	
2		$0.1 \times 0.5 \times 0.2 = 0.01$	

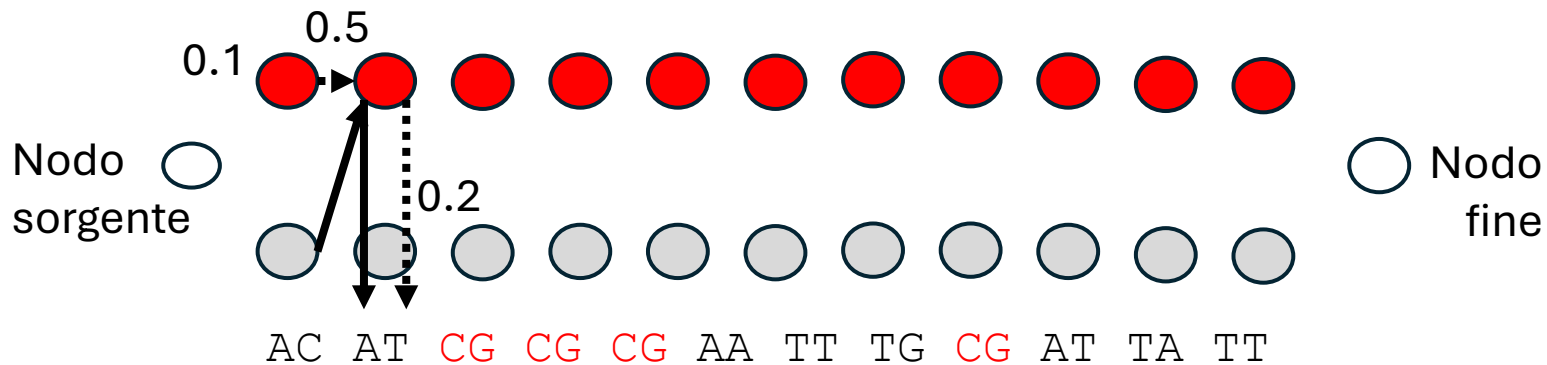
Escludiamo i sottopercorsi che appaiono subottimali!
(e che quindi sicuramente non saranno parte del percorso piu' probabile)

Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi

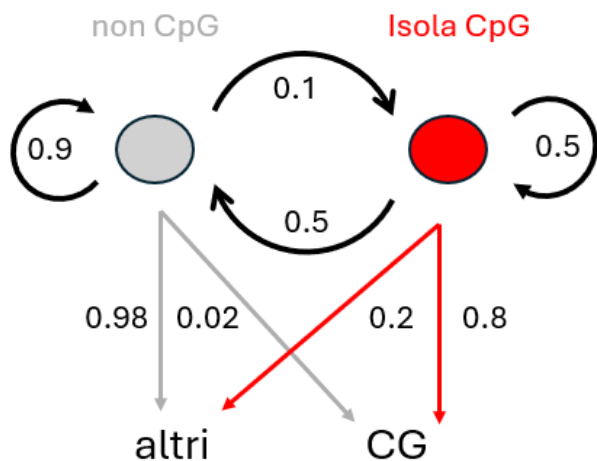
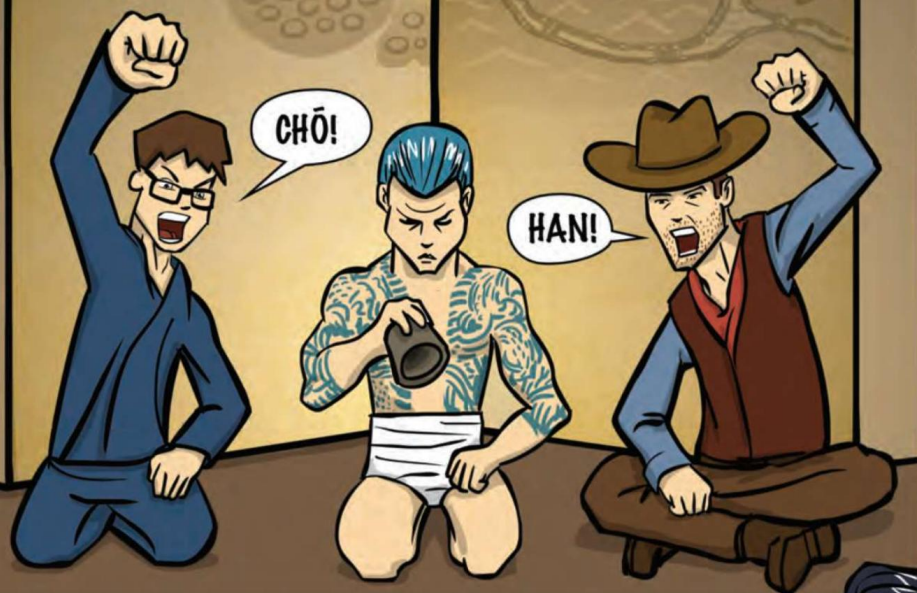


k	Stato	w	Stato precedente	Se anche il percorso passasse per nonCpG, questo sarebbe piuttosto preceduto da nonCpG, non da CpG
1		0.49	-	
1		0.1	-	
2		$0.49 * 0.9 * 0.98 = 0.882$		
2		$0.1 * 0.5 * 0.98 = 0.049$		
2		$0.49 * 0.1 * 0.2 = 0.0098$		
2		$0.1 * 0.5 * 0.2 = 0.01$		

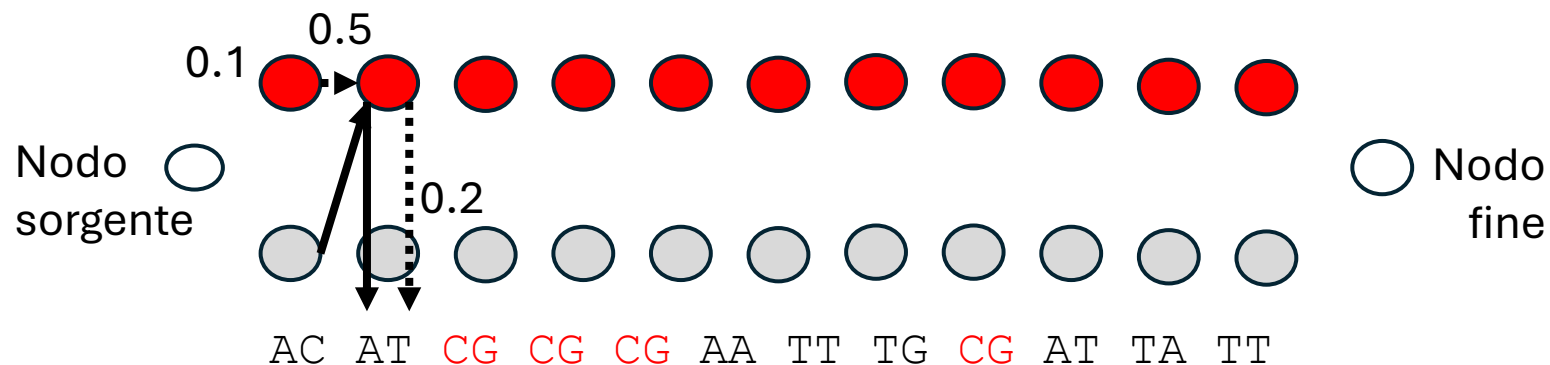
Escludiamo i sottopercorsi che appaiono subottimali!
(e che quindi sicuramente non saranno parte del percorso piu' probabile)

Algoritmo di Viterbi

Hidden Markov Models



Rappresentazione di Viterbi



k	Stato	w	Stato precedente	Se anche il percorso passasse per nonCpG, questo sarebbe piuttosto preceduto da nonCpG, non da CpG
1		0.49	-	
1		0.1	-	
2		$0.49 \cdot 0.9 \cdot 0.98 = 0.882$		
2		$0.1 \cdot 0.5 \cdot 0.98 = 0.049$		
2		$0.49 \cdot 0.1 \cdot 0.2 = 0.0098$		
2		$0.1 \cdot 0.5 \cdot 0.2 = 0.01$		Se passasse per CpG, sarebbe piu' probabilmente preceduto da CpG

Escludiamo i sottopercorsi che appaiono subottimali!
(e che quindi sicuramente non saranno parte del percorso piu' probabile)

Algoritmo di Viterbi

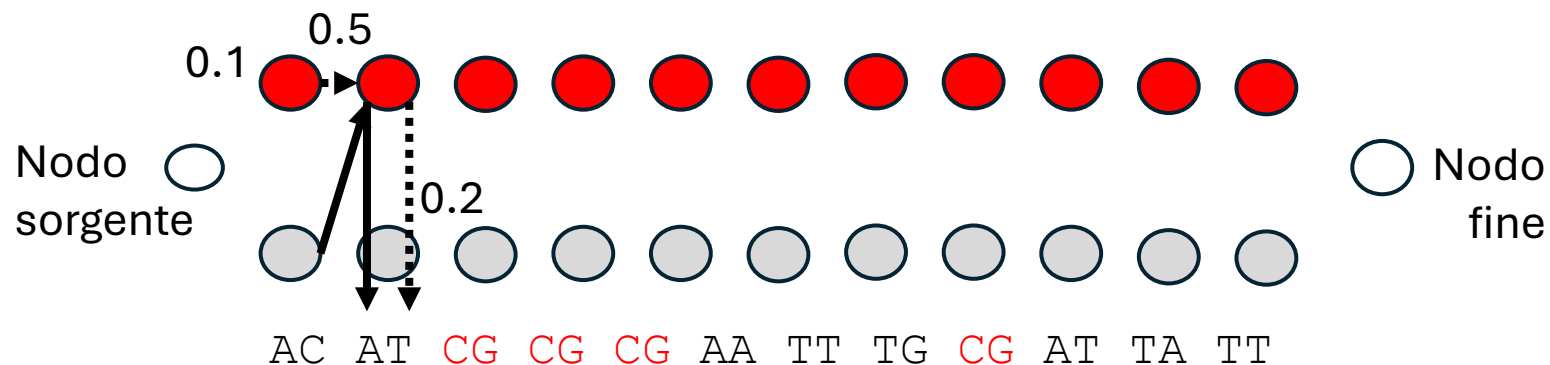
Hidden Markov Models



Ricetta dell'algoritmo di Viterbi:

1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso (w_k) e lo stato precedente più probabile.

Rappresentazione di Viterbi

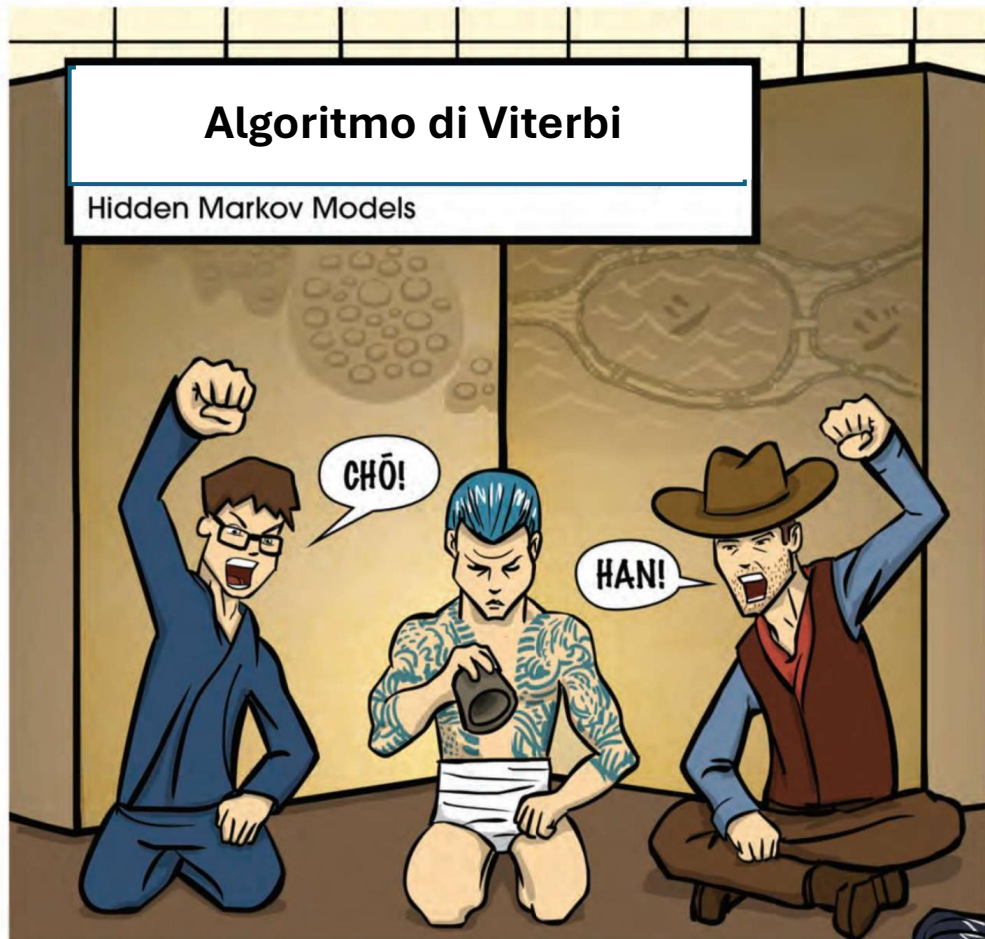


k	Stato	w	Stato precedente	Se anche il percorso passasse per nonCpG, questo sarebbe piuttosto preceduto da nonCpG, non da CpG
1		0.49	-	
1		0.1	-	
2		$0.49 \cdot 0.9 \cdot 0.98 = 0.882$		
2		$0.1 \cdot 0.5 \cdot 0.98 = 0.049$		
2		$0.49 \cdot 0.1 \cdot 0.2 = 0.0098$		
2		$0.1 \cdot 0.5 \cdot 0.2 = 0.01$		Se passasse per CpG, sarebbe più probabilmente preceduto da CpG

Escludiamo i sottopercorsi che appaiono subottimali!
(e che quindi sicuramente non saranno parte del percorso più probabile)

Algoritmo di Viterbi

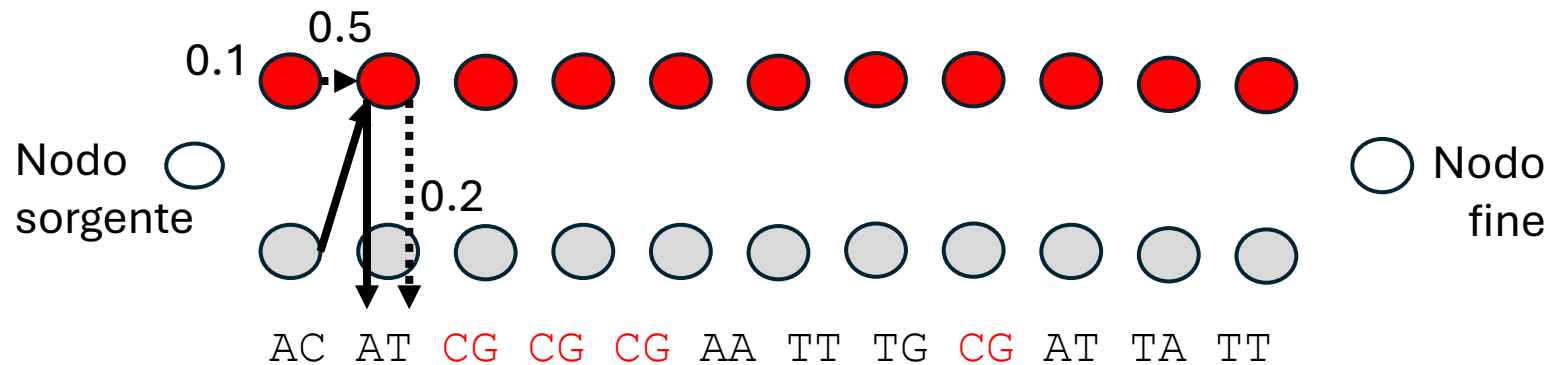
Hidden Markov Models



Ricetta dell'algoritmo di Viterbi:

1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso (w_k) e lo stato precedente più probabile.

Rappresentazione di Viterbi



k	Stato	w	Stato precedente
1		0.49	-
1		0.1	-
2		$0.49 \cdot 0.9 \cdot 0.98 = 0.882$	
2		$0.1 \cdot 0.5 \cdot 0.98 = 0.049$	
2		$0.49 \cdot 0.1 \cdot 0.2 = 0.0098$	
2		$0.1 \cdot 0.5 \cdot 0.2 = 0.01$	

Algoritmo di Viterbi

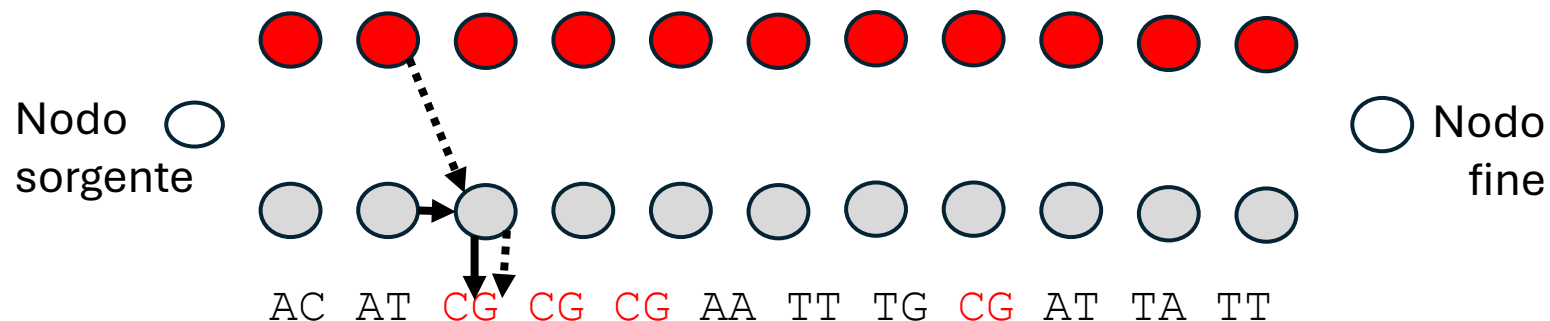
Hidden Markov Models



Ricetta dell'algoritmo di Viterbi:

1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente più probabile.

Rappresentazione di Viterbi



k	Stato	w_k	Stato precedente
1	○	0.49	-
1	●	0.1	-
2	○	$0.49 \cdot 0.9 \cdot 0.98 = 0.882$	○
2	○	$0.1 \cdot 0.5 \cdot 0.98 = 0.049$	●
2	●	$0.49 \cdot 0.1 \cdot 0.2 = 0.0098$	○
2	●	$0.1 \cdot 0.5 \cdot 0.2 = 0.01$	●
3	○	$0.432 \cdot 0.9 \cdot 0.02 = 0.0078$	○
3	○	$0.01 \cdot 0.5 \cdot 0.02 = 0.0001$	●
3	●	$0.432 \cdot 0.1 \cdot 0.8 = 0.0346$	○
3	●	$0.01 \cdot 0.5 \cdot 0.8 = 0.004$	●

Algoritmo di Viterbi

Hidden Markov Models



Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.

Rappresentazione di Viterbi

Nodo
sorgente

Nodo
fine



AC AT CG CG CG AA TT TG CG AT TA TT

k

$w_k(\text{nonCpG})$

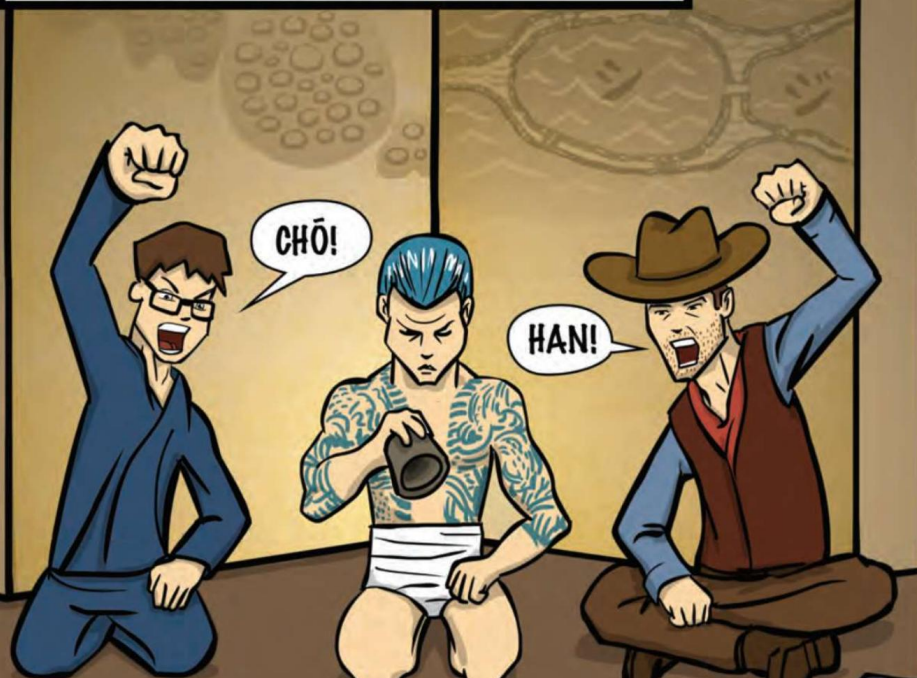
e stato
precedente

$w_k(\text{IsolaCpG})$

e stato
precedente

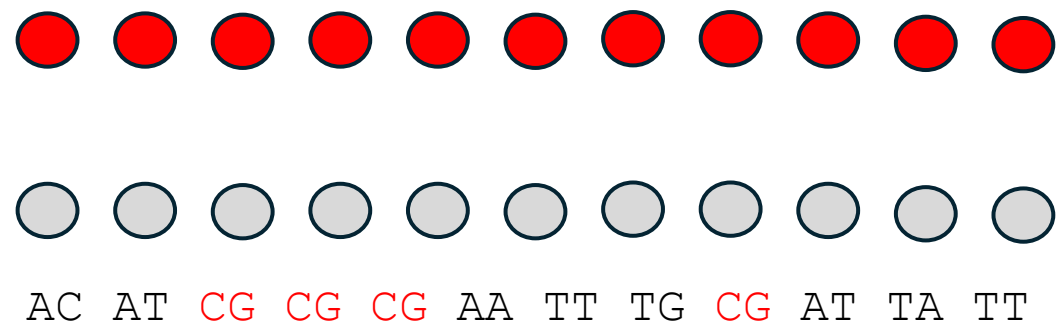
k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
1	0.49	[-]	0.1	[-]
2	0.432		0.01	
3	7.78e-03		3.46e-02	
4	3.46e-04		1.38e-02	
5	1.38e-04		5.53e-03	
6	2.71e-03		5.53e-04	
7	2.39e-03		5.53e-05	
8	2.11e-03		4.78e-05	
9	3.80e-05		1.69e-04	
10	8.27e-05		1.69e-05	
11	7.29e-05		1.69e-06	

Hidden Markov Models



- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.
- 2) Al termine, sappiamo quale percorso era il piu' probabile!

Nodo sorgente ○



○ Nodo
fine

k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
1	0.49	[-]	0.1	[-]
2	0.432		0.01	
3	$7.78e-03$		$3.46e-02$	
4	$3.46e-04$		$1.38e-02$	
5	$1.38e-04$		$5.53e-03$	
6	$2.71e-03$		$5.53e-04$	
7	$2.39e-03$		$5.53e-05$	
8	$2.11e-03$		$4.78e-05$	
9	$3.80e-05$		$1.69e-04$	
10	$8.27e-05$		$1.69e-05$	
11	$7.29e-05$		$1.69e-06$	

Algoritmo di Viterbi

Hidden Markov Models



Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.
- 2) Al termine, sappiamo quale percorso era il piu' probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!

Rappresentazione di Viterbi

Nodo
sorgente

Nodo
fine



AC AT CG CG CG AA TT TG CG AT TA TT

k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
-----	----------------------	--------------------	------------------------	--------------------

1	0.49	[−]	0.1	[−]
2	0.432		0.01	
3	7.78e-03		3.46e-02	
4	3.46e-04		1.38e-02	
5	1.38e-04		5.53e-03	
6	2.71e-03		5.53e-04	
7	2.39e-03		5.53e-05	
8	2.11e-03		4.78e-05	
9	3.80e-05		1.69e-04	
10	8.27e-05		1.69e-05	
11	7.29e-05		1.69e-06	



Algoritmo di Viterbi

Hidden Markov Models



Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.
- 2) Al termine, sappiamo quale percorso era il piu' probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!

Rappresentazione di Viterbi

Nodo sorgente ○

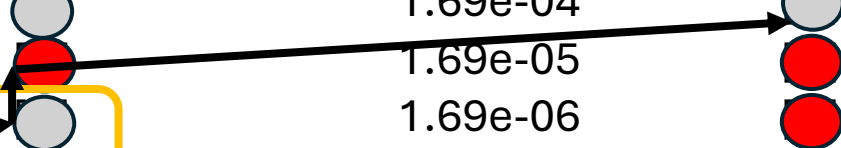
○ Nodo fine



AC AT CG CG CG AA TT TG CG AT TA TT

k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
-----	----------------------	--------------------	------------------------	--------------------

1	0.49	[−]	0.1	[−]
2	0.432	○	0.01	●
3	7.78e-03	○	3.46e-02	○
4	3.46e-04	●	1.38e-02	●
5	1.38e-04	●	5.53e-03	●
6	2.71e-03	●	5.53e-04	●
7	2.39e-03	○	5.53e-05	●
8	2.11e-03	○	4.78e-05	○
9	3.80e-05	○	1.69e-04	○
10	8.27e-05	●	1.69e-05	●
11	7.29e-05	○	1.69e-06	●



Algoritmo di Viterbi

Hidden Markov Models



Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.
- 2) Al termine, sappiamo quale percorso era il piu' probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!

Rappresentazione di Viterbi

Nodo sorgente ○

○ Nodo fine



AC AT CG CG CG AA TT TG CG AT TA TT

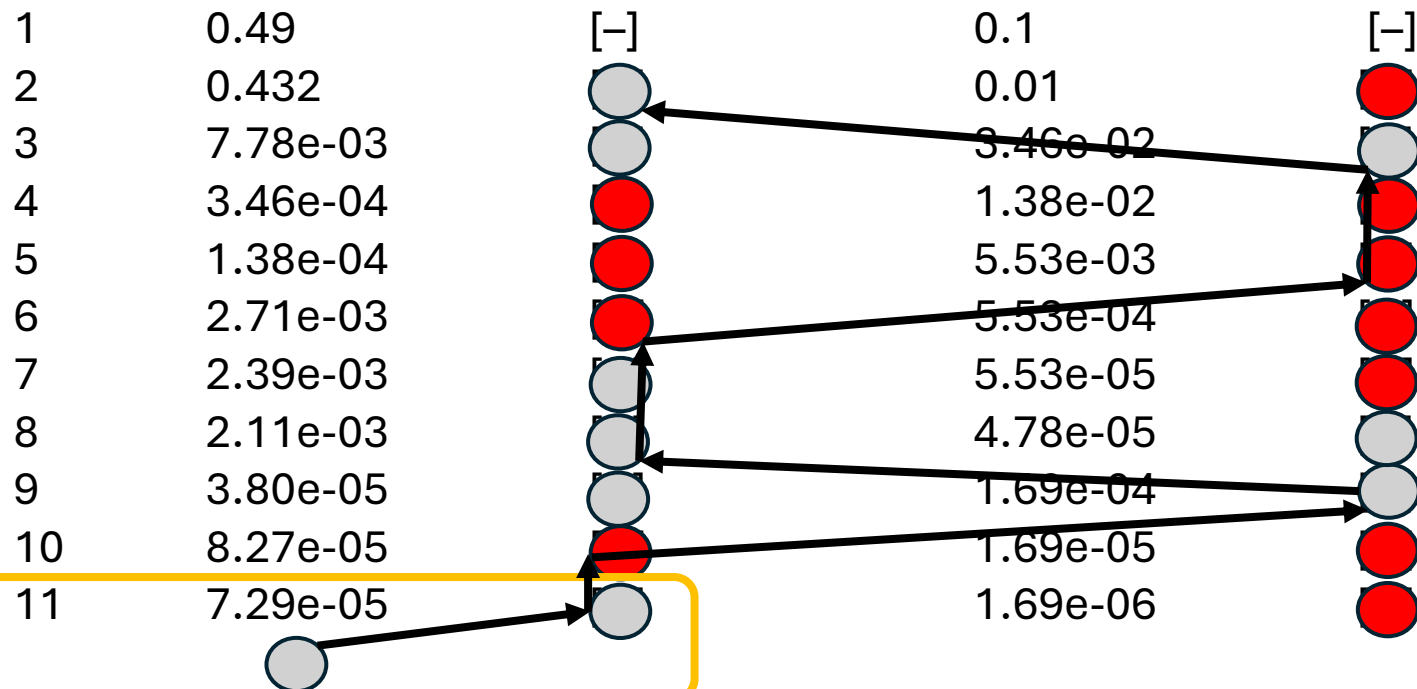
k

$w_k(\text{nonCpG})$

e stato precedente

$w_k(\text{IsolaCpG})$

e stato precedente



Algoritmo di Viterbi

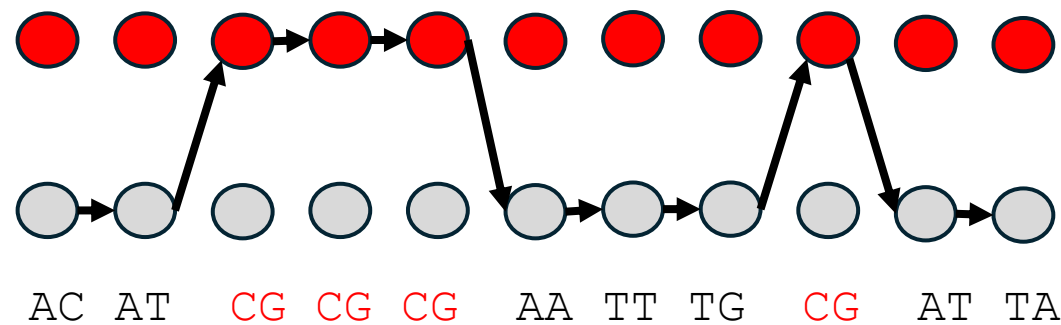
Hidden Markov Models



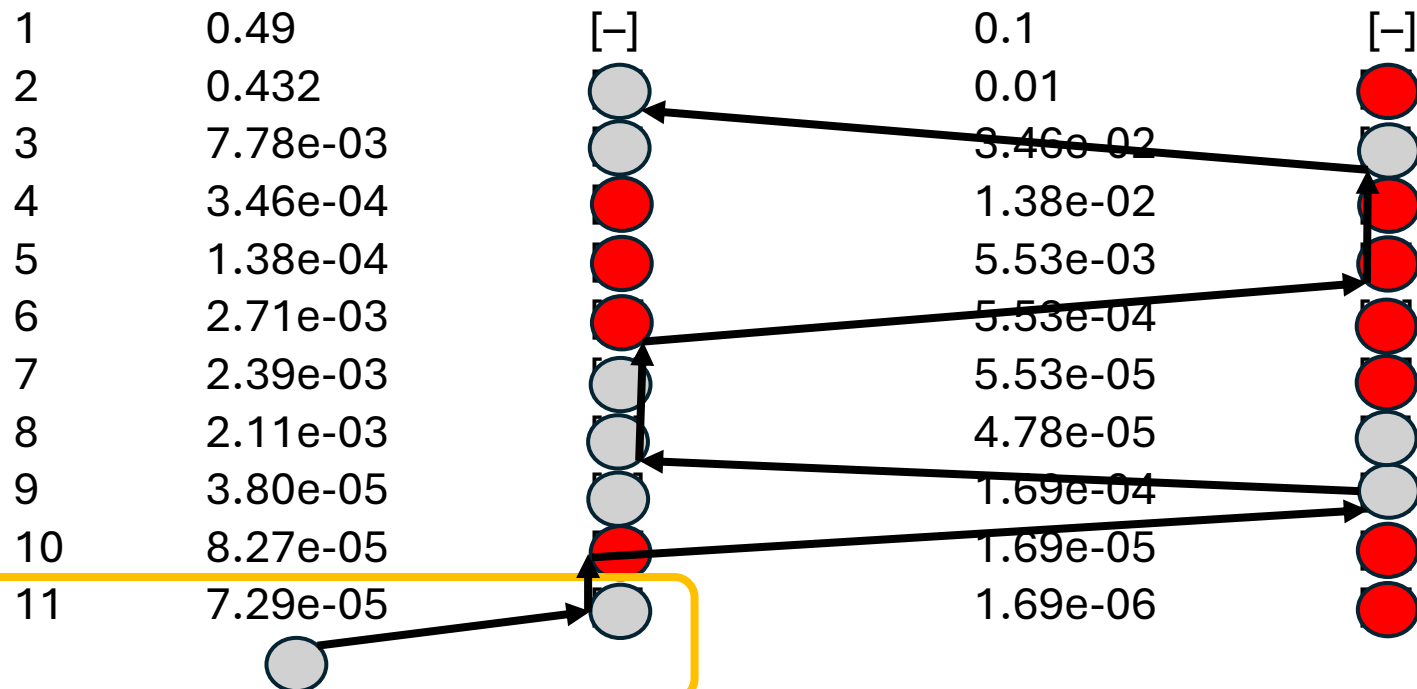
Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.
- 2) Al termine, sappiamo quale percorso era il piu' probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!

Rappresentazione di Viterbi



k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
-----	----------------------	--------------------	------------------------	--------------------



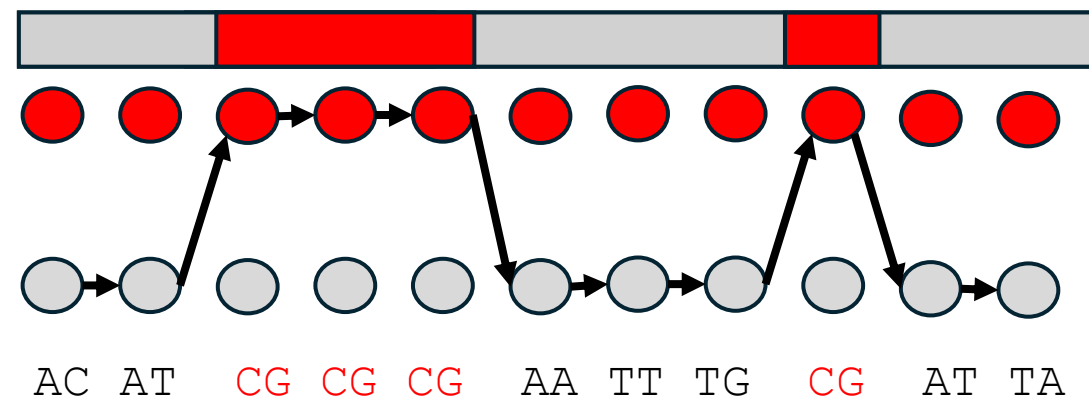
Algoritmo di Viterbi

Hidden Markov Models

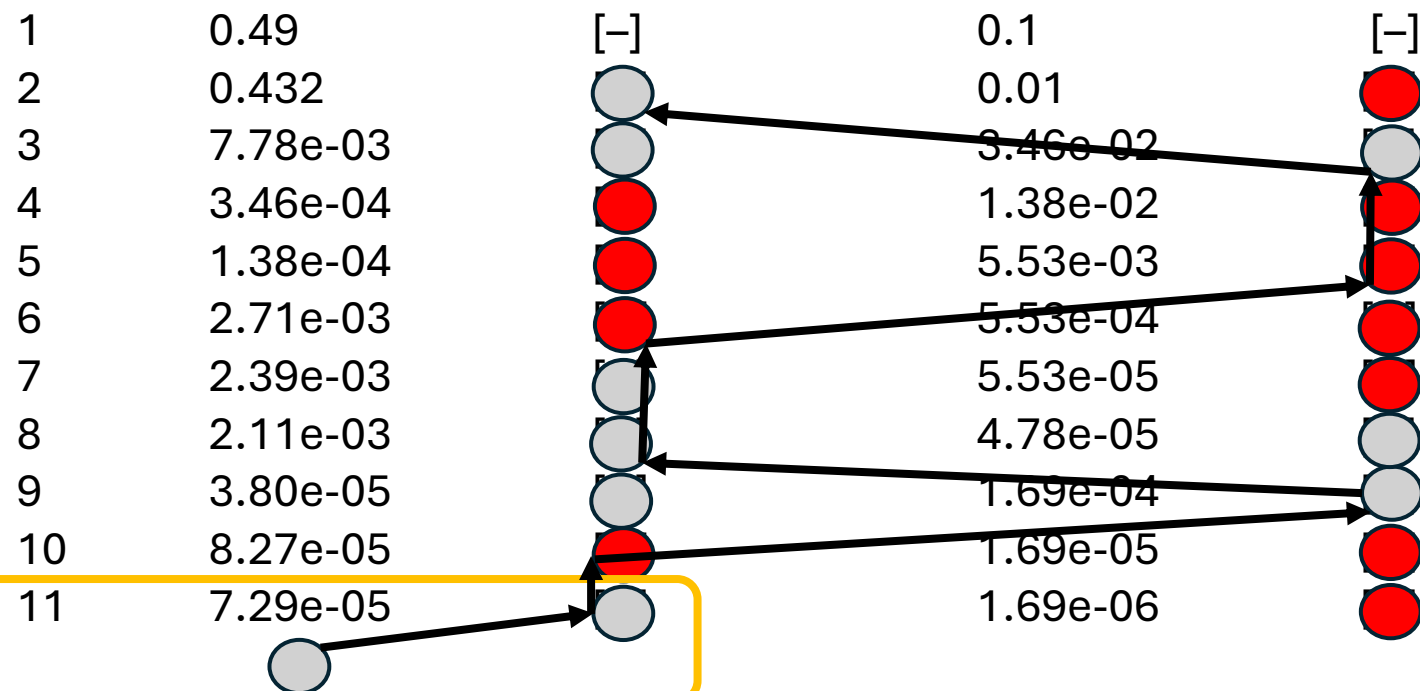


Ricetta dell'algoritmo di Viterbi:

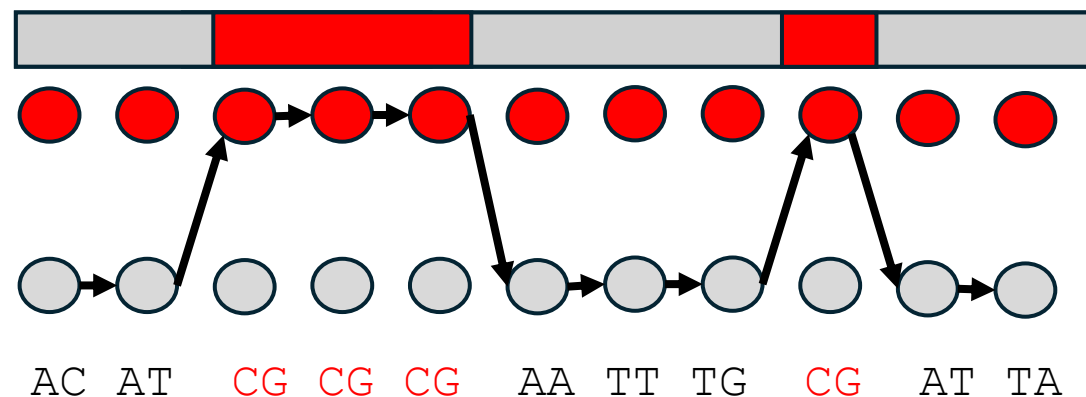
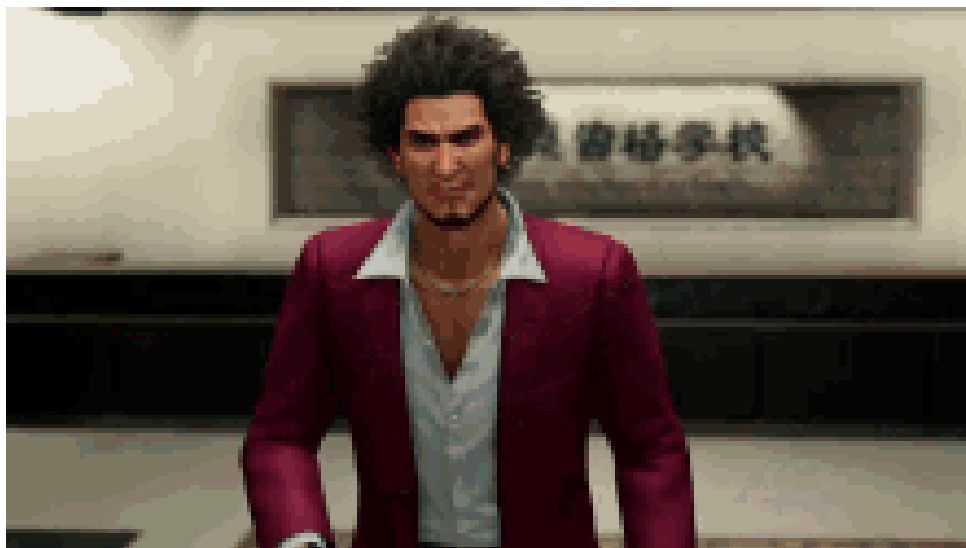
- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.
- 2) Al termine, sappiamo quale percorso era il piu' probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!



k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
-----	----------------------	--------------------	------------------------	--------------------



Algoritmo di Viterbi



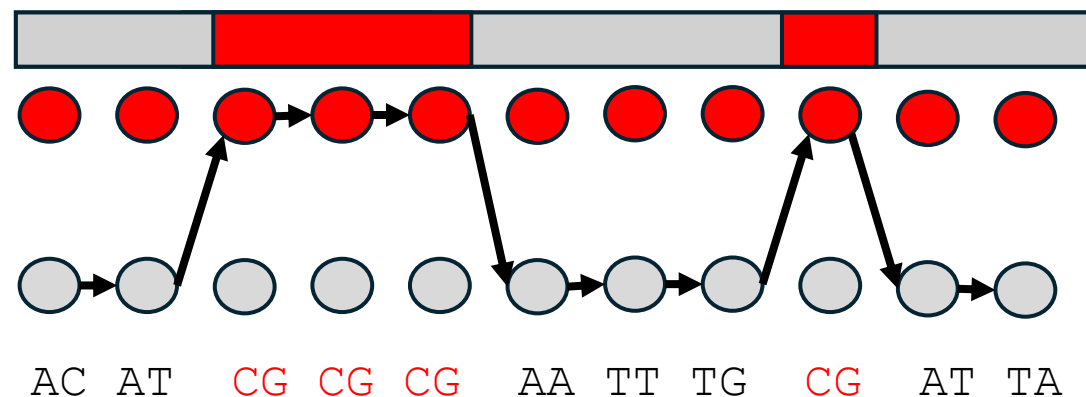
k $w_k(\text{nonCpG})$ e stato precedente $w_k(\text{IsolaCpG})$ e stato precedente

1	0.49	[−]	0.1	[−]
2	0.432	●	0.01	●
3	7.78e-03	●	3.46e-02	●
4	3.46e-04	●	1.38e-02	●
5	1.38e-04	●	5.53e-03	●
6	2.71e-03	●	5.53e-04	●
7	2.39e-03	●	5.53e-05	●
8	2.11e-03	●	4.78e-05	●
9	3.80e-05	●	1.69e-04	●
10	8.27e-05	●	1.69e-05	●
11	7.29e-05	●	1.69e-06	●

Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente piu' probabile.
- 2) Al termine, sappiamo quale percorso era il piu' probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!

Algoritmo di Viterbi



k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
-----	----------------------	--------------------	------------------------	--------------------

1	0.49	[−]	0.1	[−]
2	0.432	●	0.01	●
3	7.78e-03	●	3.46e-02	●
4	3.46e-04	●	1.38e-02	●
5	1.38e-04	●	5.53e-03	●
6	2.71e-03	●	5.53e-04	●
7	2.39e-03	●	5.53e-05	●
8	2.11e-03	●	4.78e-05	●
9	3.80e-05	●	1.69e-04	●
10	8.27e-05	●	1.69e-05	●
11	7.29e-05	●	1.69e-06	●

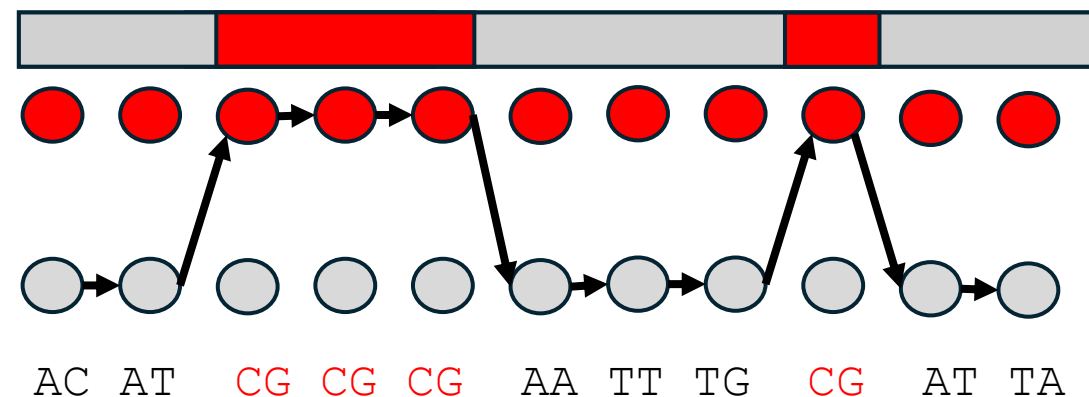
Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente più probabile.
- 2) Al termine, sappiamo quale percorso era il più probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso più probabile»?



k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
1	0.49	[-]	0.1	[-]
2	0.432	Gray	0.01	Red
3	7.78e-03	Gray	3.46e-02	Gray
4	3.46e-04	Red	1.38e-02	Red
5	1.38e-04	Red	5.53e-03	Red
6	2.71e-03	Red	5.53e-04	Red
7	2.39e-03	Gray	5.53e-05	Red
8	2.11e-03	Gray	4.78e-05	Gray
9	3.80e-05	Gray	1.69e-04	Gray
10	8.27e-05	Red	1.69e-05	Red
11	7.29e-05	Gray	1.69e-06	Red

Algoritmo di Viterbi



Vediamo un altro esempio

Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso più probabile»?

Algoritmo «greedy»

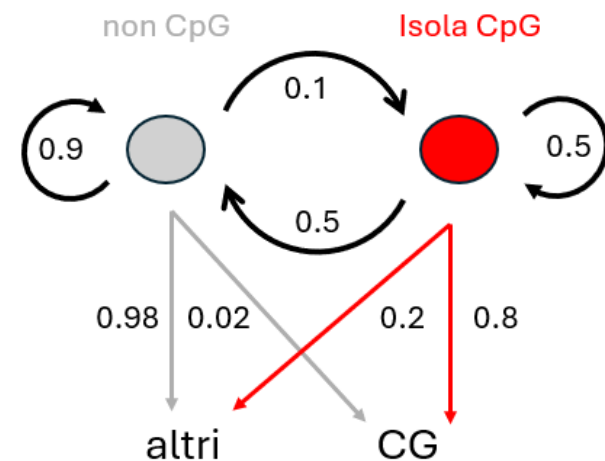
Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



AC CG AA

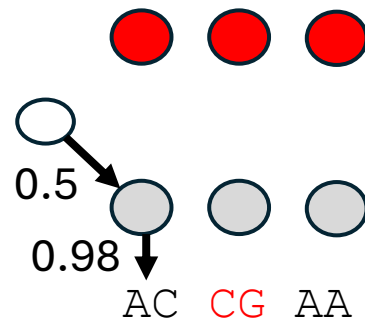


Algoritmo «greedy»

Algoritmo di Viterbi

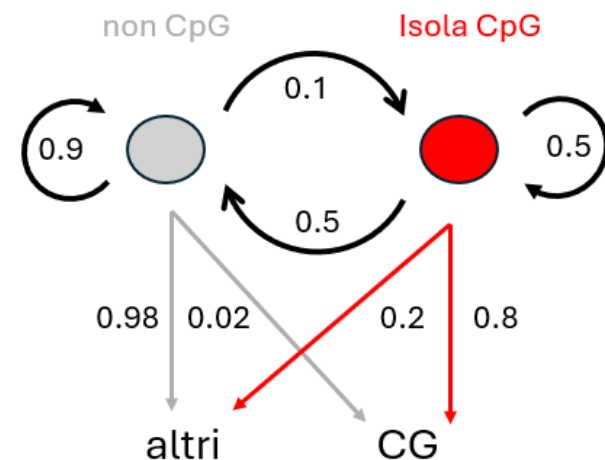


Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

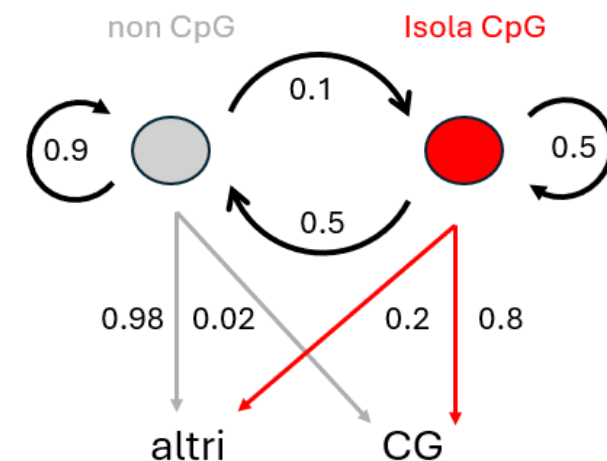
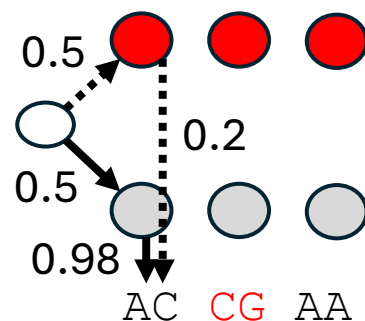
Il percorso continuo ha probabilità
 $0.5 \times 0.98 = 0.49$



Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

Il percorso continuo ha probabilità

$$0.5 \times 0.98 = 0.49$$



Il percorso tratteggiato ha probabilità

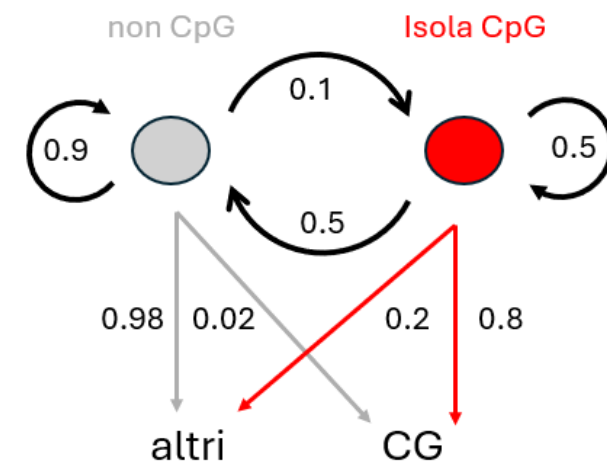
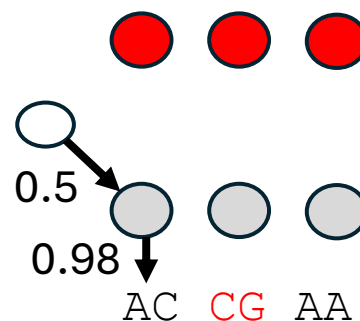
$$0.5 \times 0.2 = 0.1$$



Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

Il percorso continuo ha probabilità

$$0.5 \cdot 0.98 = 0.49$$



Il percorso tratteggiato ha probabilità

$$0.5 \cdot 0.2 = 0.1$$

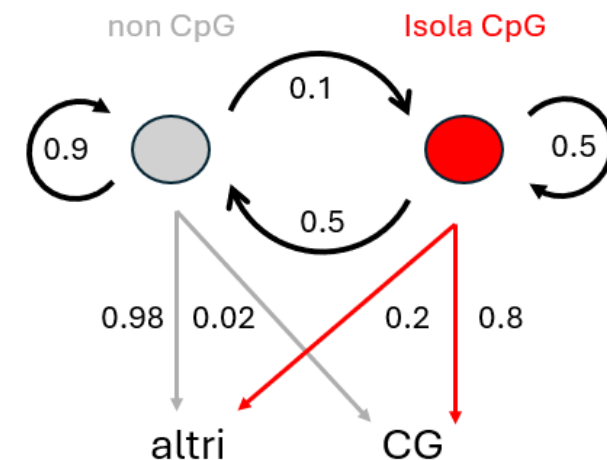
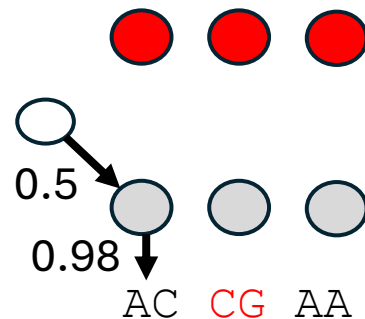


Il percorso tratteggiato è meno probabile (lo scartiamo)

Algoritmo di Viterbi



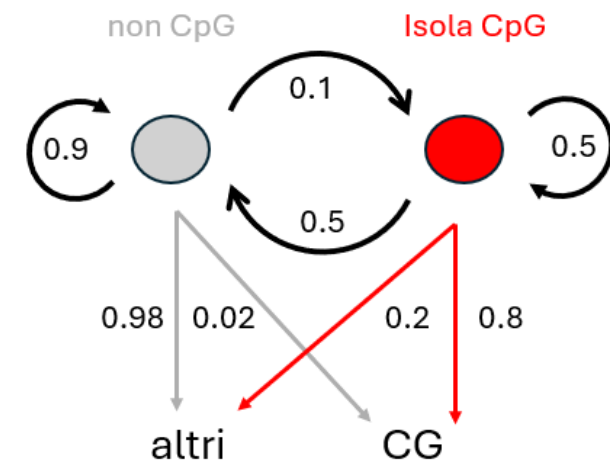
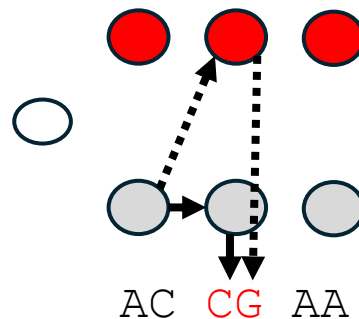
Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

Il percorso tratteggiato è meno probabile (lo scartiamo)
Posizione 1: probabilità del percorso = 0.49
→

Algoritmo di Viterbi



Algoritmo «greedy»

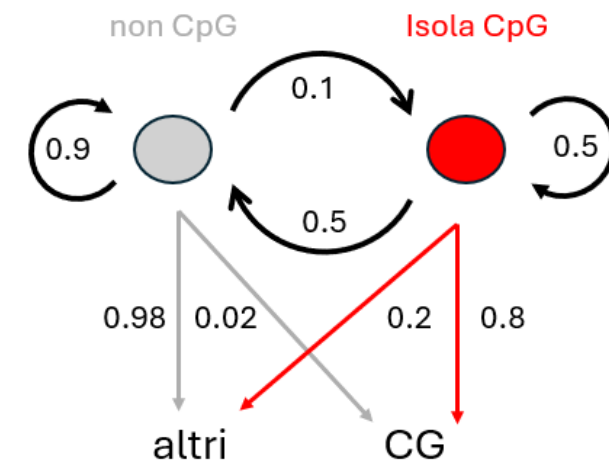
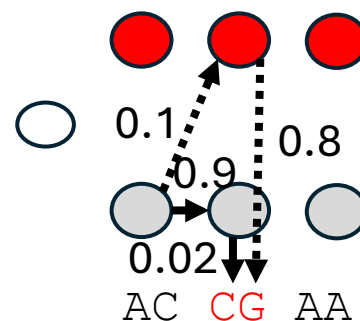
Prolunghiamo ora il percorso in posizione 2

Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

In posizione 2:

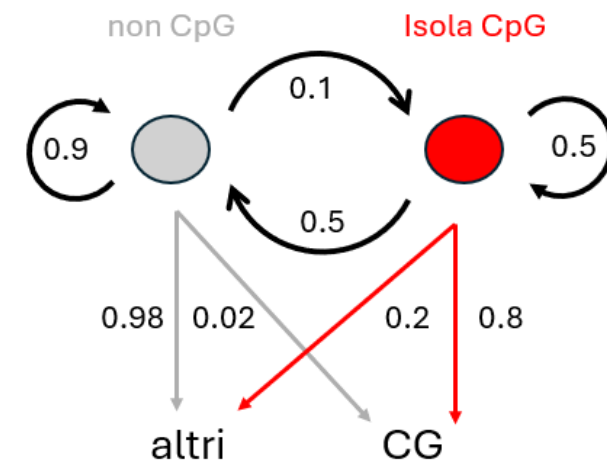
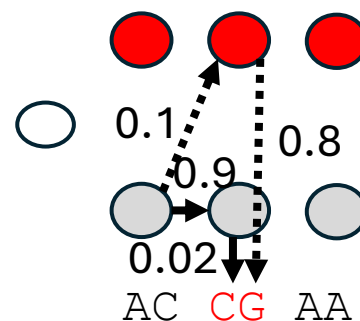
probabilità del percorso $\rightarrow 0.1 * 0.8 = 0.08$

probabilità del percorso $\rightarrow 0.9 * 0.02 = 0.018$

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

In posizione 2:

probabilità del percorso $\rightarrow 0.1 * 0.8 = 0.08$

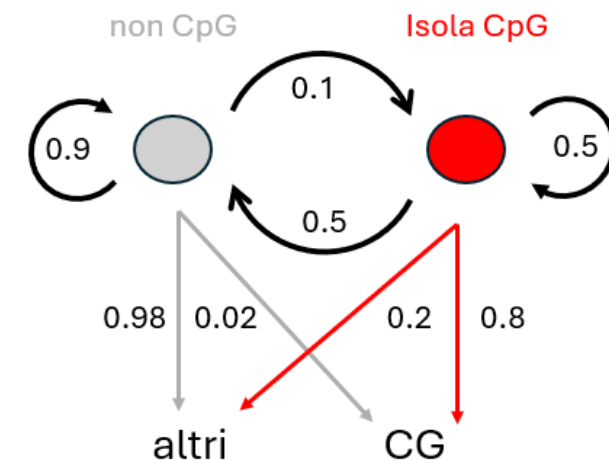
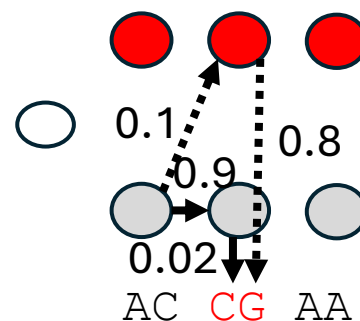
probabilità del percorso $\rightarrow 0.9 * 0.02 = 0.018$

Cosa facciamo?

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

In posizione 2:

probabilità del percorso $\rightarrow 0.1 * 0.8 = 0.08$

probabilità del percorso $\rightarrow 0.9 * 0.02 = 0.018$

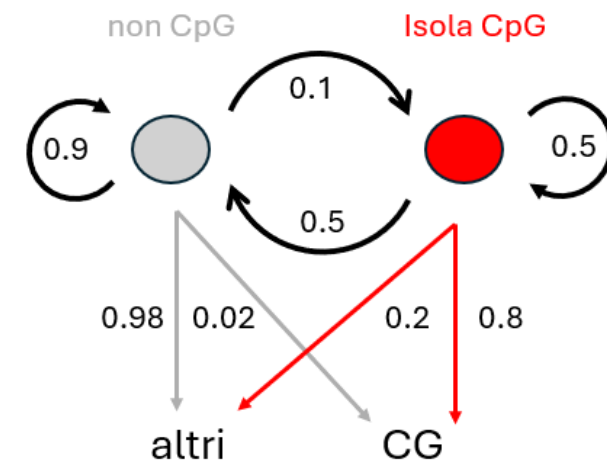
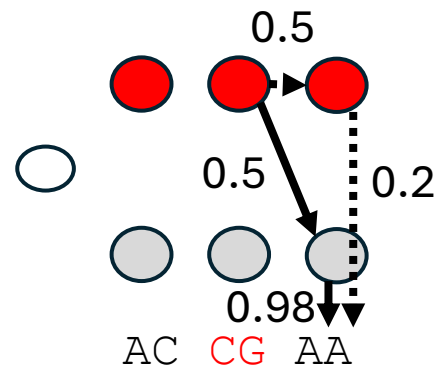
Cosa facciamo?

Scartiamo il percorso $\rightarrow$ e teniamo $\rightarrow$

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

In posizione 3:

probabilità del percorso $\rightarrow 0.5 \cdot 0.2 = 0.1$

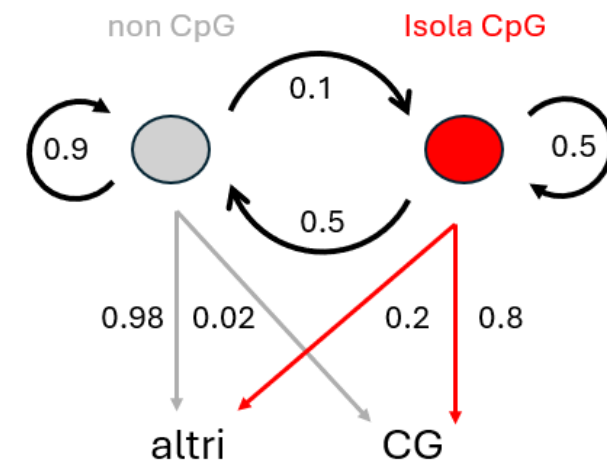
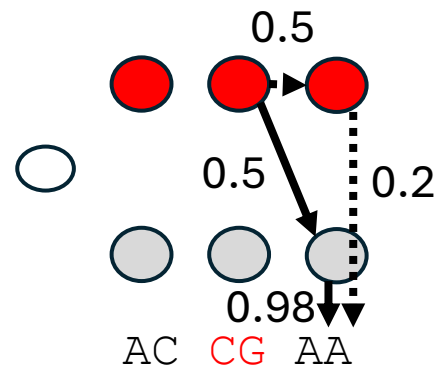
probabilità del percorso $\rightarrow 0.5 \cdot 0.98 = 0.49$

Cosa facciamo?

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

In posizione 3:

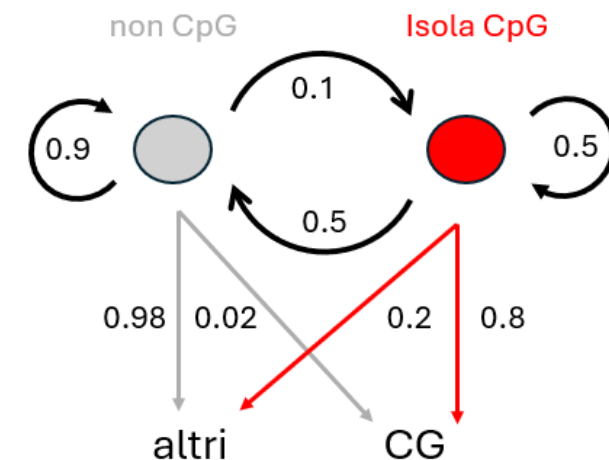
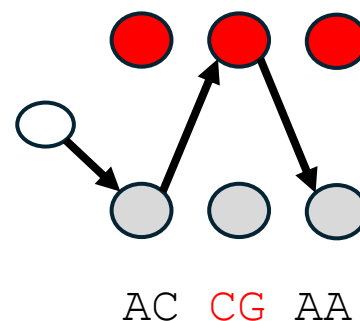
probabilità del percorso► $0.5 \cdot 0.2 = 0.1$

probabilità del percorso —► $0.5 \cdot 0.98 = 0.49$

Cosa facciamo?

Scartiamo il percorso► e teniamo —►

Algoritmo di Viterbi



Algoritmo «greedy»

Risultato:

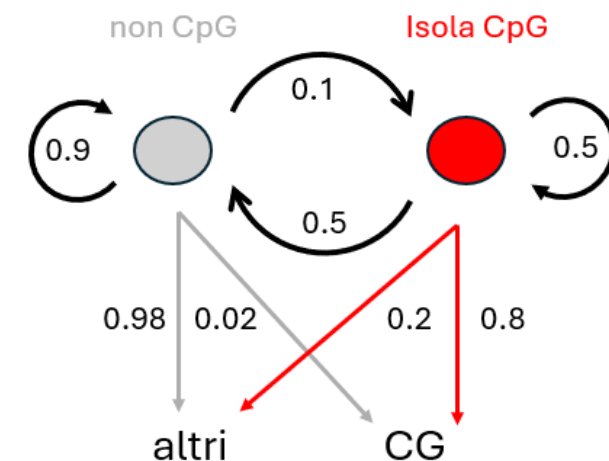
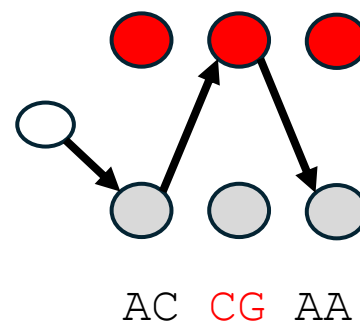
Il percorso migliore (in figura) ha likelihood:

Cosa facciamo?

Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?

Scartiamo il percorso► e teniamo —►

Algoritmo di Viterbi



Algoritmo «greedy»

Risultato:

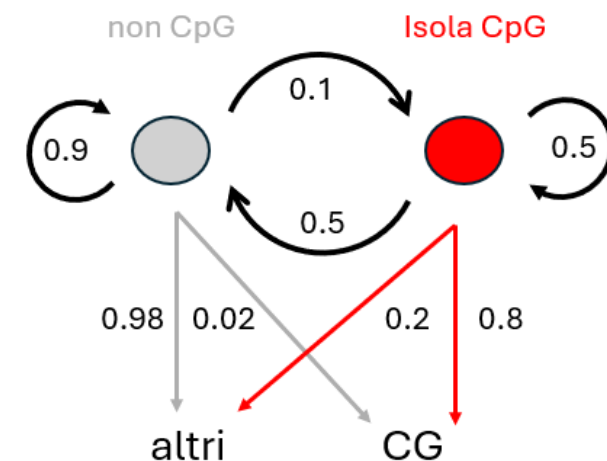
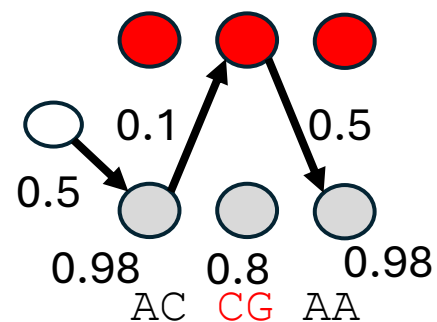
Il percorso migliore (in figura) ha likelihood:

Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

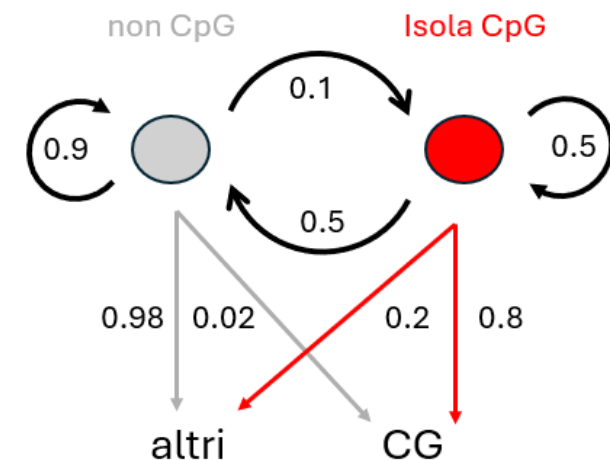
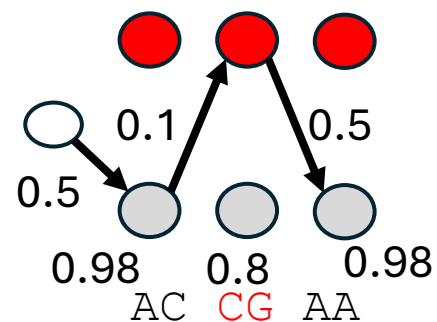
Risultato:

Il percorso migliore (in figura) ha likelihood:
 $0.5 \cdot 0.98 \cdot 0.1 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0192}$

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

Risultato:

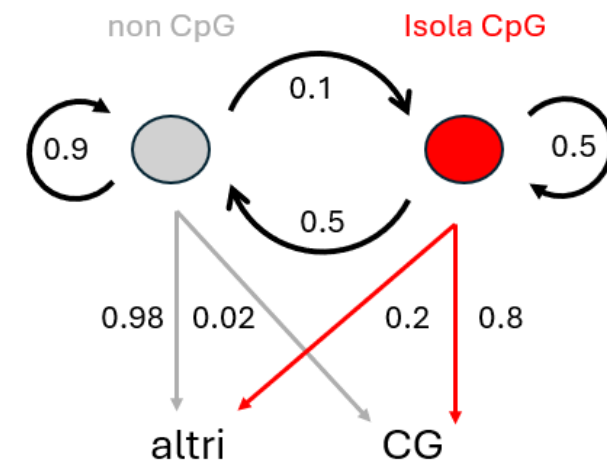
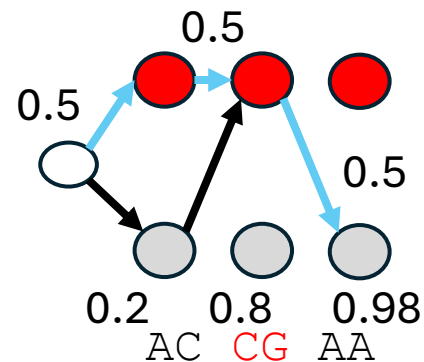
Il percorso migliore (in figura) ha likelihood:
 $0.5 \cdot 0.98 \cdot 0.1 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0192}$

E' il miglior percorso?

Algoritmo di Viterbi



Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

Risultato:

Il percorso migliore (in figura) ha likelihood:

$$0.5 \cdot 0.98 \cdot 0.1 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0192}$$



E' il miglior percorso?

Il percorso celeste ha likelihood **piu' alta**:

$$0.5 \cdot 0.2 \cdot 0.5 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0196}$$

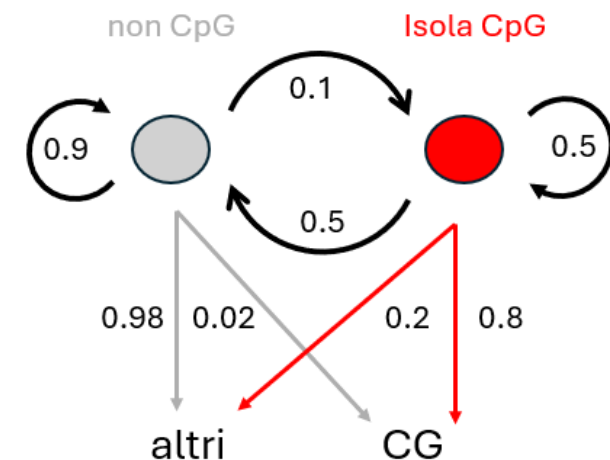
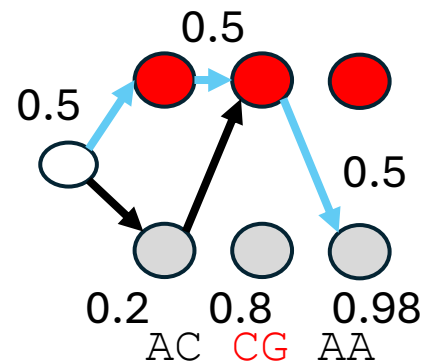


Algoritmo di Viterbi



Ve l'avevo detto! Verificate come esercizio che questo è il percorso che avremmo trovato con l'algoritmo di Viterbi!

Ma perché dobbiamo calcolare la probabilità per tutta la sequenza e poi «tornare indietro»?
Non potremmo fin dall'inizio «camminare lungo il percorso piu' probabile»?



Algoritmo «greedy»

Risultato:

Il percorso migliore (in figura) ha likelihood:
 $0.5 \cdot 0.98 \cdot 0.1 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0192}$

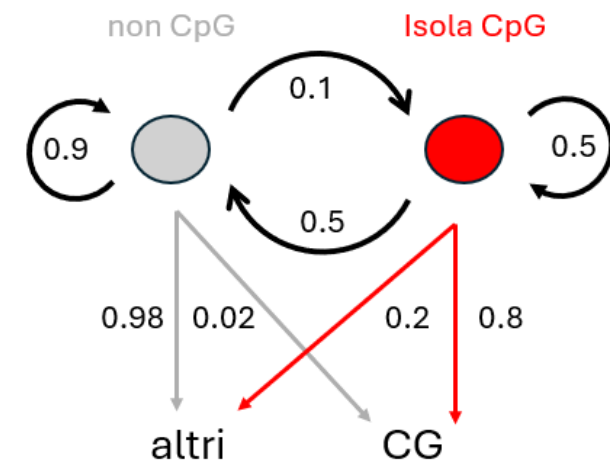
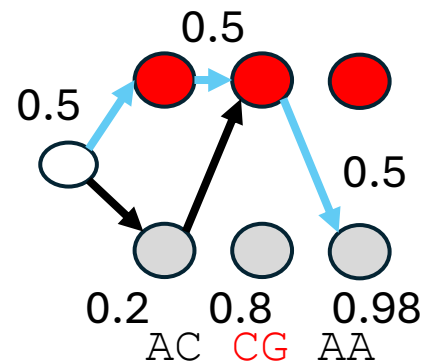
E' il miglior percorso?

Il percorso celeste ha likelihood **piu' alta**:
 $0.5 \cdot 0.2 \cdot 0.5 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0196}$

Algoritmo di Viterbi



Ma come è possibile???!!!!
In fondo seguivamo la pista piu'
probabile! C'è un inganno!
Cosa ci nascondi?!!!



Algoritmo «greedy»

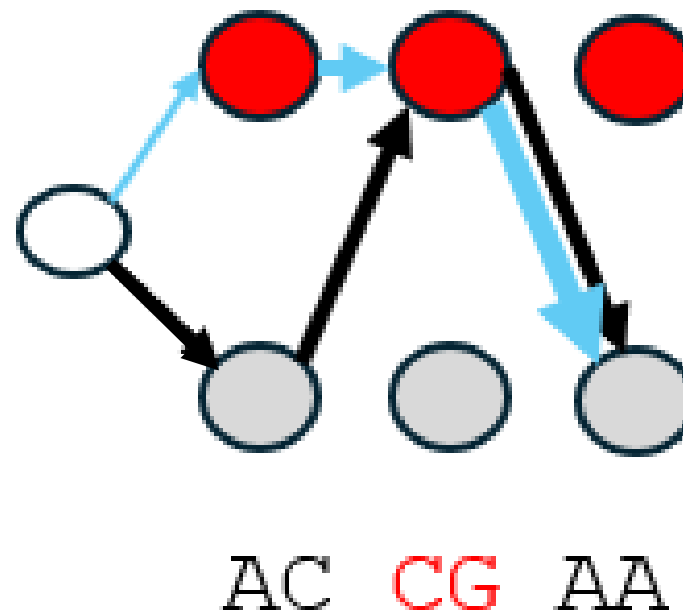
Risultato:

Il percorso migliore (in figura) ha likelihood:
 $0.5 \cdot 0.98 \cdot 0.1 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0192}$

E' il miglior percorso?

Il percorso celeste ha likelihood **piu' alta**:
 $0.5 \cdot 0.2 \cdot 0.5 \cdot 0.8 \cdot 0.5 \cdot 0.98 = \mathbf{0.0196}$

Algoritmo di Viterbi

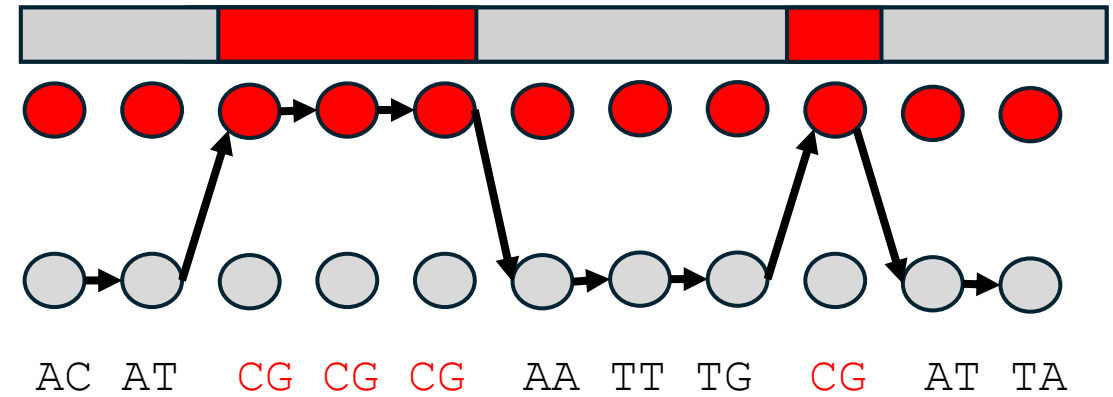


Nessun trucco. Un percorso **può sembrare piu' probabile/spesso all'inizio** (linea nera piu' spesso nel primo passaggio). Ma non sappiamo se poi un'altro percorso che passa per stati diversi diventerà piu' probabile (celeste nel secondo e terzo passaggio).

Per questo dobbiamo:

- 1) arrivare alla fine per poter decidere con sicurezza quale è il percorso piu' probabile (come fa Viterbi);
- 2) Escludere solo quelli che sicuramente sono peggiori (come fa l'algoritmo di Viterbi) ma non percorsi potenzialmente migliori che passano per stati diversi tenendone uno solo (come fa Greedy)!

Un attimo, non abbiamo finito!



**Abbiamo trovato il percorso ottimale dati i parametri
(probabilità di transizione ed emissione) decisi a priori!
Come facciamo a trovare quelli corretti??**

Ricetta dell'algoritmo di Viterbi:

- 1) calcoliamo la probabilità per il percorso salvando una tabella dove per ogni stato salviamo la probabilità del percorso che termina con quello stato in posizione k (w_k) e lo stato precedente più probabile.
- 2) Al termine, sappiamo quale percorso era il più probabile e possiamo ricostruirlo a ritroso dagli stati precedenti!!!

Niente paura, l'avevamo già visto prima!



Gruppo



Hidden Markov Models

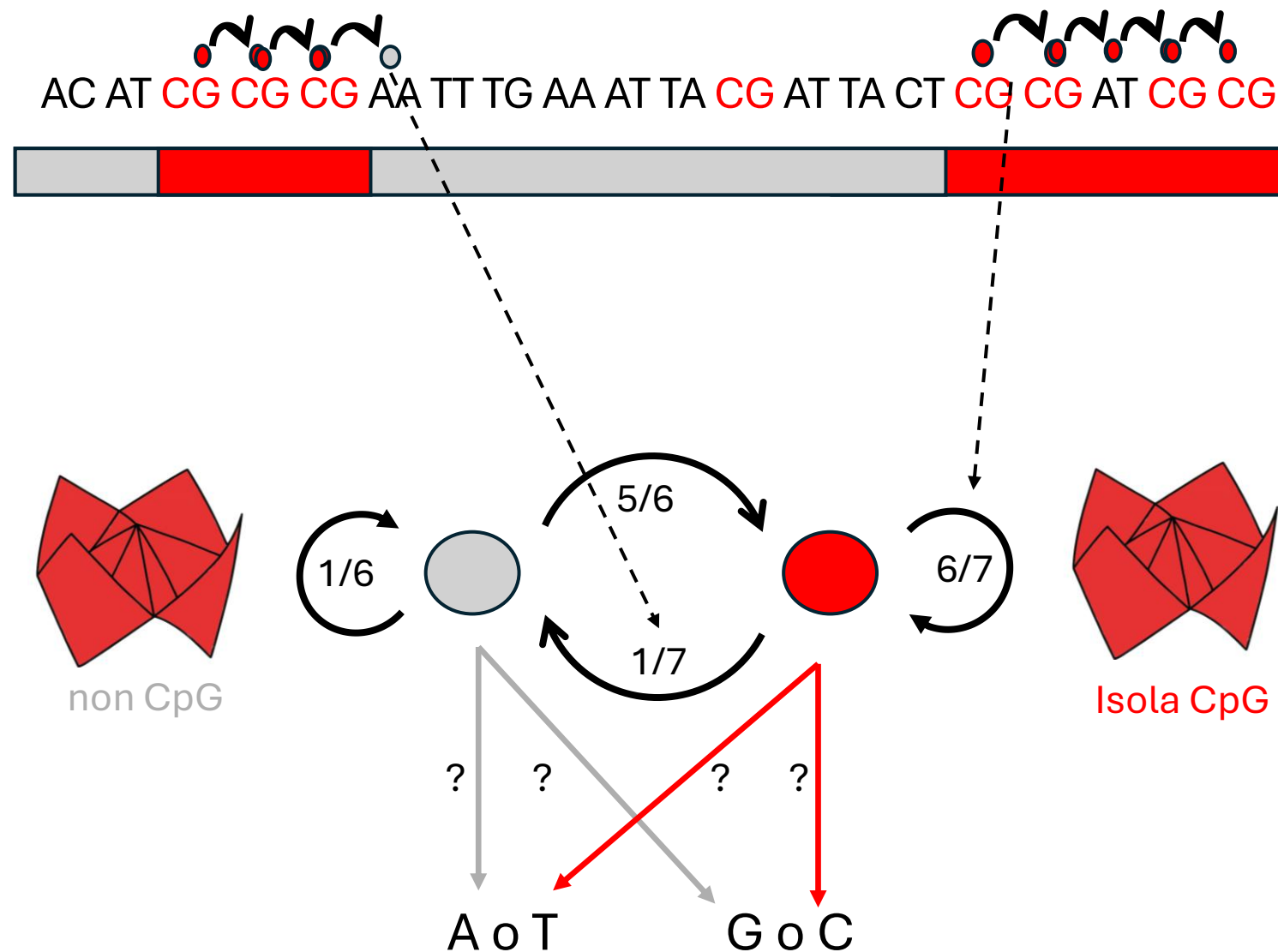


Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

How were we supposed to know that the casino was run by yakuza?!?!

Avevamo visto che dato un percorso, potevamo calcolare le probabilità di **transizione**..



Gruppo



Hidden Markov Models



Owning a Bioinformatics Time Machine isn't always fun and games. One night, we made a fateful decision to travel back to early 19th Century Tokyo to eat at the first sushi restaurant.

We took a wrong turn and found ourselves outside a gambling house. Always up for a new adventure, we decided to try our hands at a game or two.

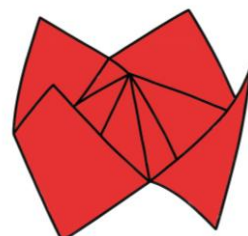
How were we supposed to know that the casino was run by yakuza?!?!

Avevamo visto che dato un percorso, potevamo calcolare le probabilità di transizione e di **emissione**!!

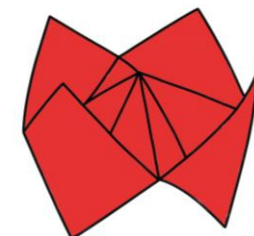
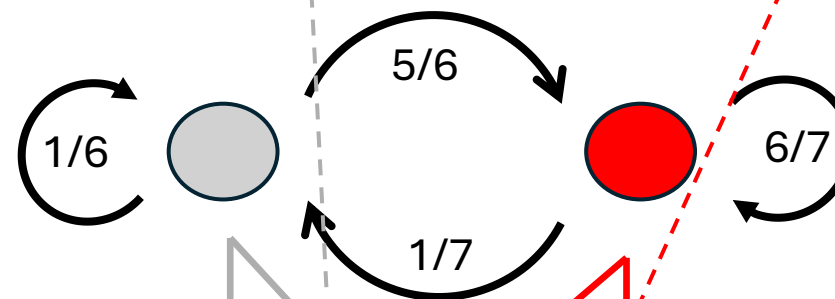
AC AT **CG CG CG** AA TT TG AA AT TA **CG** AT TA CT **CG CG** AT **CG CG**



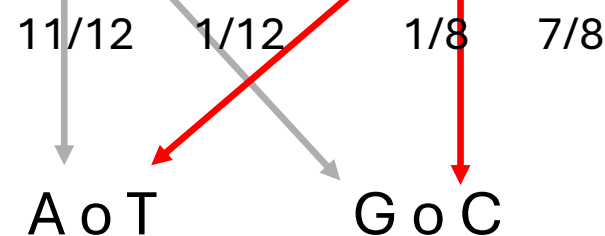
Le probabilità di emissione sono date dalla probabilità di osservare CG nei vari stati



non CpG

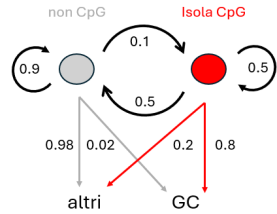


Isola CpG



Cosa sappiamo fare:

- Date delle probabilità sappiamo trovare il percorso piu' probabile



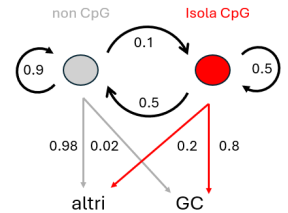
$(x, ?, \text{emissioni e transizioni}) \rightarrow \pi$



- Dato un percorso sappiamo trovare le probabilità corrispondenti

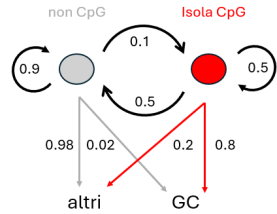


$(x, \pi, ?) \rightarrow \text{emissioni e transizioni}$



Cosa sappiamo fare:

- Date delle probabilità sappiamo trovare il percorso piu' probabile



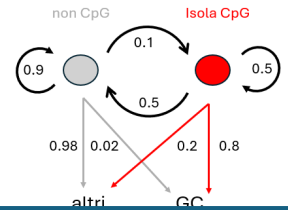
$(x, ?, \text{emissioni e transizioni}) \rightarrow \pi$



- Dato un percorso sappiamo trovare le probabilità corrispondenti



$(x, \pi, ?) \rightarrow \text{emissioni e transizioni}$



Algoritmo per risolvere una HMM con Viterbi learning

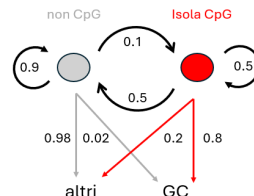
Percorso ottimale



Ci fermiamo quando convergiamo ad un valore con probabilità massima

Stimiamo probabilità di emissione) e transizione

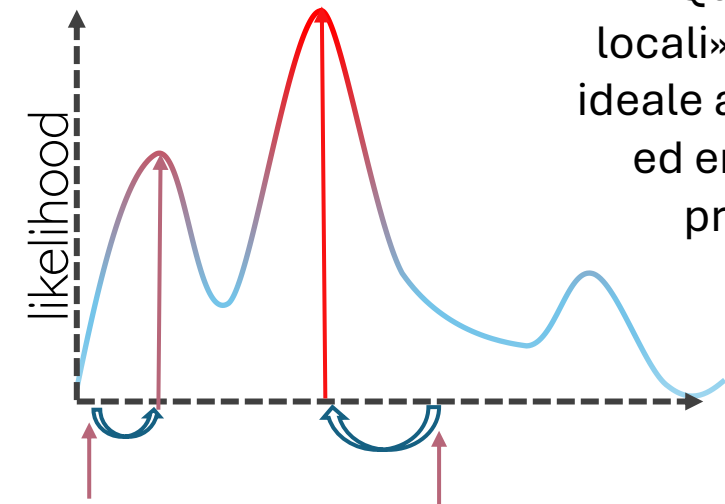
Algoritmo di Viterbi



Partiamo da probabilità di emissione e transizione casuali

Note:

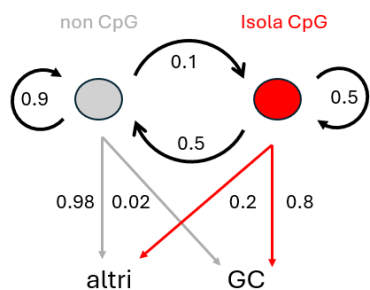
Questo algoritmo può bloccarsi su dei «massimi locali» della likelihood, cioè non trovare la soluzione ideale a partire da qualsiasi probabilità di transizione ed emissione iniziali voi diate. Quindi è necessario provare varie condizioni iniziali per assicurare la convergenza.



Valori iniziali di probabilità di transizione o emissione

Algoritmo per risolvere una HMM con Viterbi learning

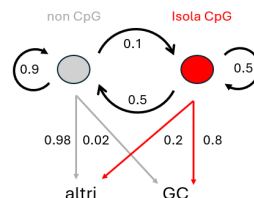
Percorso ottimale



Algoritmo di Viterbi



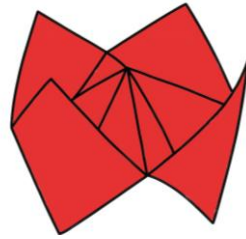
Stimiamo probabilità di emissione) e transizione



Ci fermiamo quando convergiamo la nostra likelihood smette di aumentare

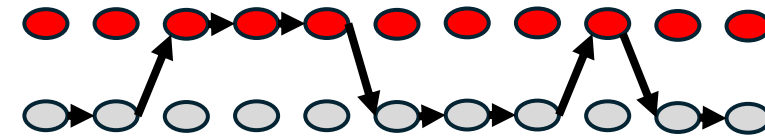
Partiamo da probabilità di emissione e transizione casuali

Conclusioni



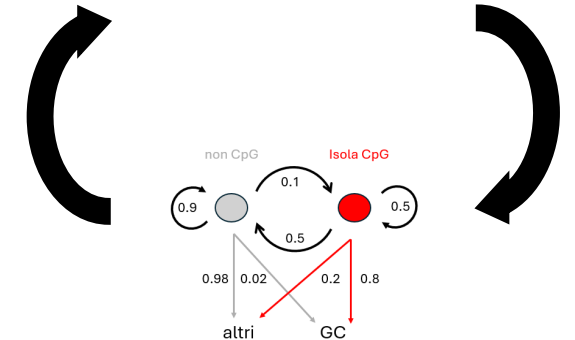
- La likelihood di un modello HMM è il prodotto di probabilità di transizione ed emissione (considerando osservazioni x e un percorso)
- Esplorare tutti i percorsi è proibitivo ma possiamo applicare alcuni algoritmi per velocizzare il processo
- Utilizzando l'algoritmo di Viterbi possiamo trovare il percorso ottimale per un modello HMM
- Applicandolo iterativamente possiamo stimare i parametri (le probabilità di transizione ed emissione) del modello

$$Pr(x, \pi) = \prod_i Pr(\text{emissione } x_i | \pi_i) Pr(\text{transizione } \pi_{i-1} \rightarrow \pi_i)$$



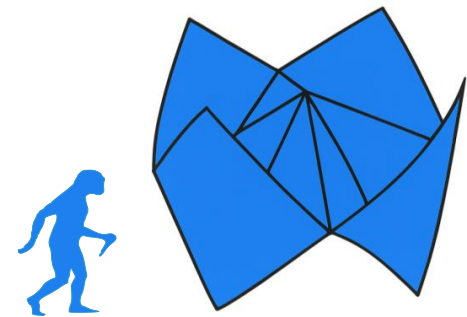
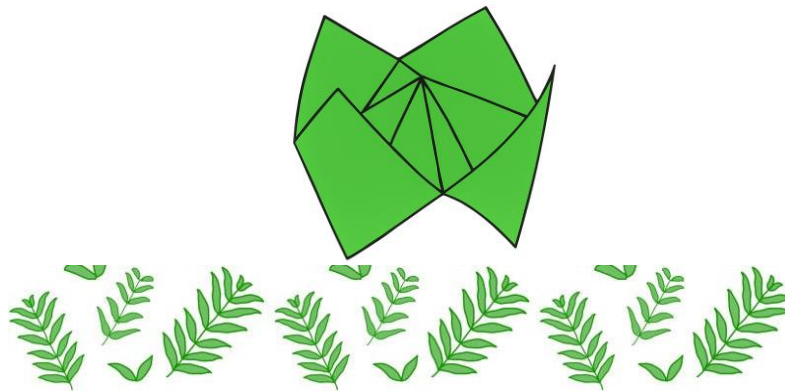
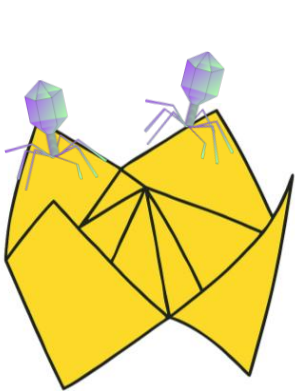
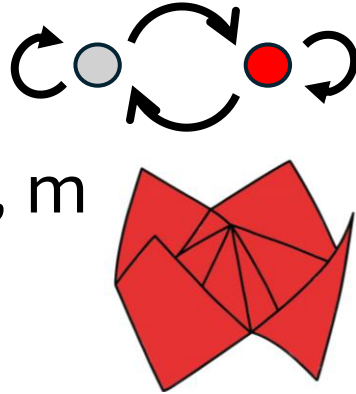
k	$w_k(\text{nonCpG})$	e stato precedente	$w_k(\text{IsolaCpG})$	e stato precedente
1	0.49	[-]	0.1	[-]
2	0.432		0.01	
3	7.78e-03		3.46e-02	
4	3.46e-04		1.38e-02	
5	1.38e-04		5.53e-03	
6	2.71e-03		5.53e-04	
7	2.39e-03		5.53e-05	
8	2.11e-03		4.78e-05	
9	3.80e-05		1.69e-04	
10	8.27e-05		1.69e-05	
11	7.29e-05		1.69e-06	

Percorso ottimale



Provate ad applicare il piccolo script awk viterbi.awk alle vostre sequenze

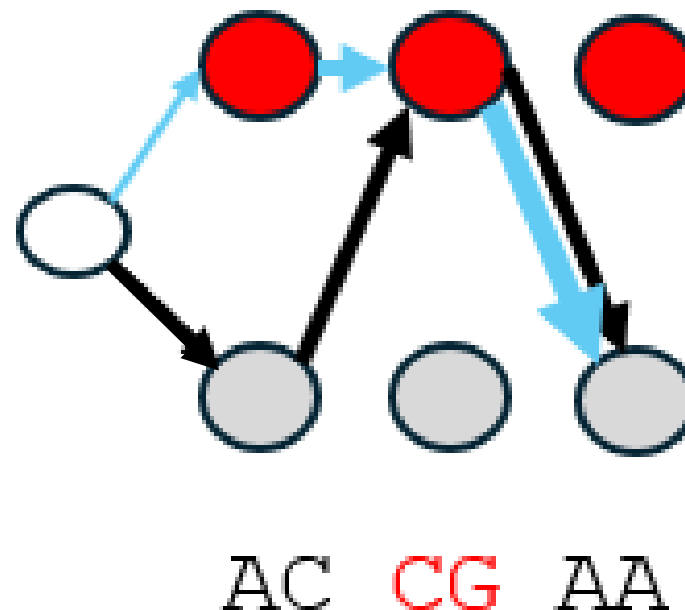
- Le probabilità sono quelle dell'esempio Isole CpG del Corso, ma potete modificarle ed esplorarle per i vostri casi
- Per farlo funzionare:
 - Esempio 1: `printf "Y\nY\nY\nX\nX\nY\n" | awk -f viterbi.awk`
 - Esempio 2: `cat FILE | awk -f viterbi.awk` #where FILE contains a X and a Y for each row



Un'ultima nota prima di andare alla pratica



L'interrogatorio di
«Algoritmo di Viterbi»
continua



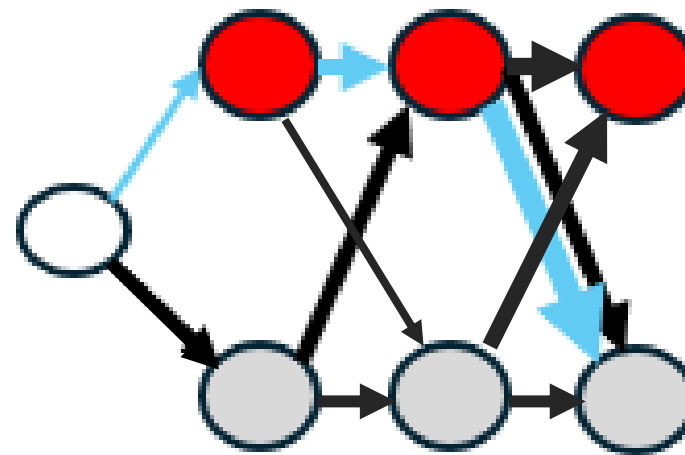
OK, «Algoritmo di Viterbi», hai trovato il percorso piu' probabile,
hai vinto tu.

Ma questo vuol dire anche che ce ne sono **vari** che magari hanno
probabilità abbastanza simili!

Ha senso tenere conto solamente del piu' probabile in assoluto?

Forse dovremmo sentire anche altri testimoni!

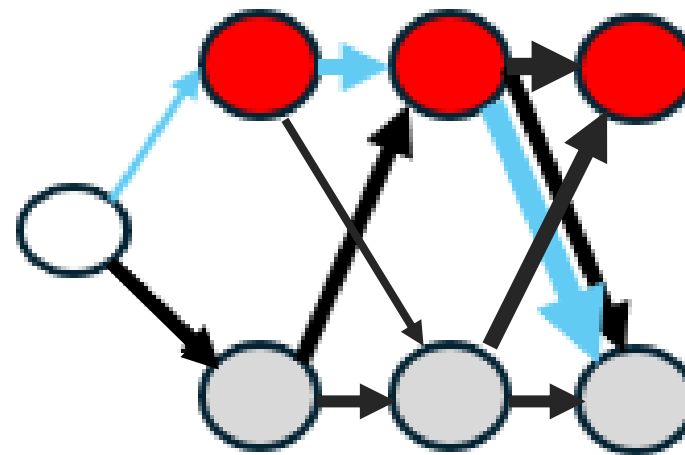
L'interrogatorio di «Algoritmo di Viterbi» continua



AC CG AA

- Il percorso **celeste** è il migliore.
- Ma esistono **molti altri percorsi**, con probabilità anche molto simili (proprio perché tanti).
- Per stimare le probabilità di transizione e di emissione possiamo essere **piu' accurati se teniamo conto di tutti questi percorsi sulla base della loro probabilità**, e non solo di uno (per quanto sia il migliore)!

L'interrogatorio di «Algoritmo di Viterbi» continua



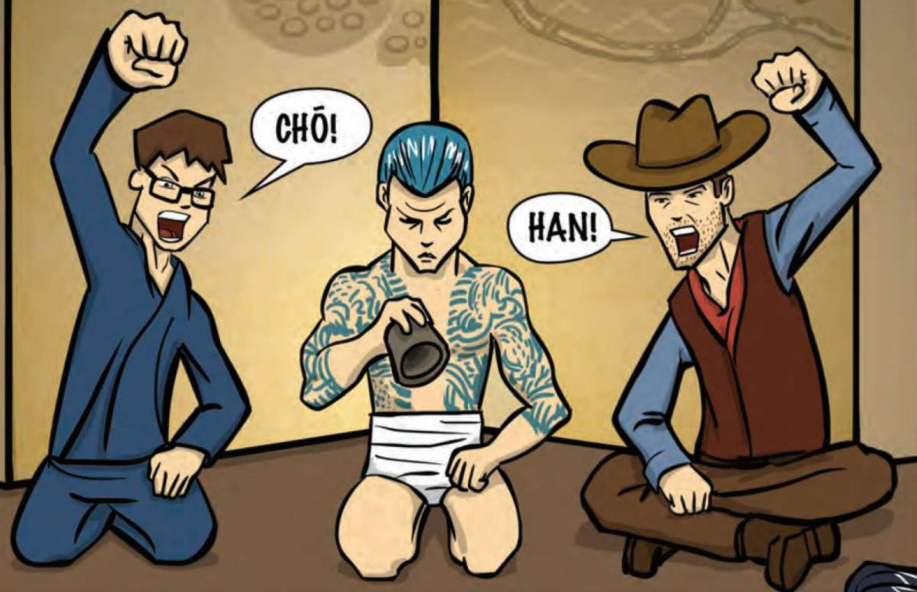
AC CG AA

- Il percorso **celeste** è il migliore.
- Ma esistono **molti altri percorsi**, con probabilità anche molto simili (proprio perché tanti).
- Per stimare le probabilità di transizione e di emissione possiamo essere **piu' accurati se teniamo conto di tutti questi percorsi sulla base della loro probabilità**, e non solo di uno (per quanto sia il migliore)!

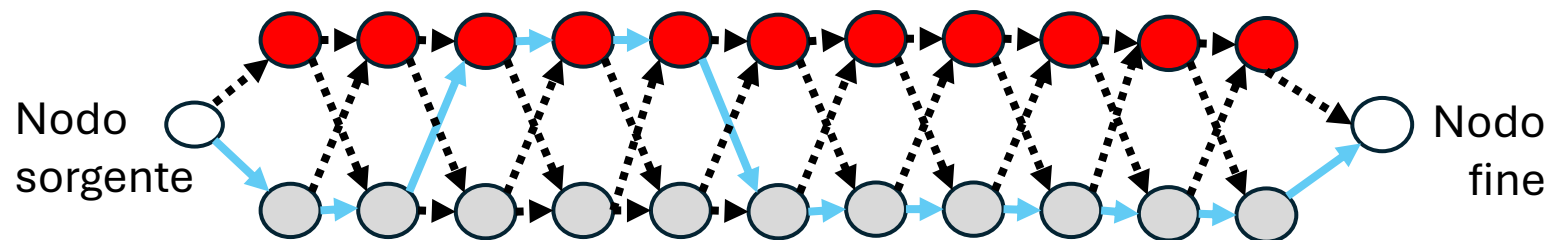
Ascoltiamo
piu' testimoni!

Baum-Welch

Hidden Markov Models



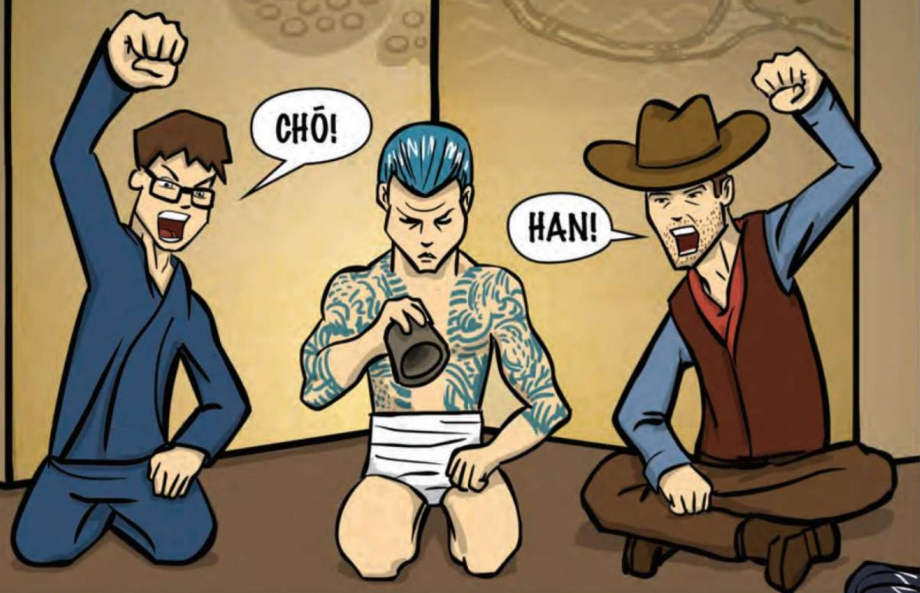
Rappresentazione di Viterbi..



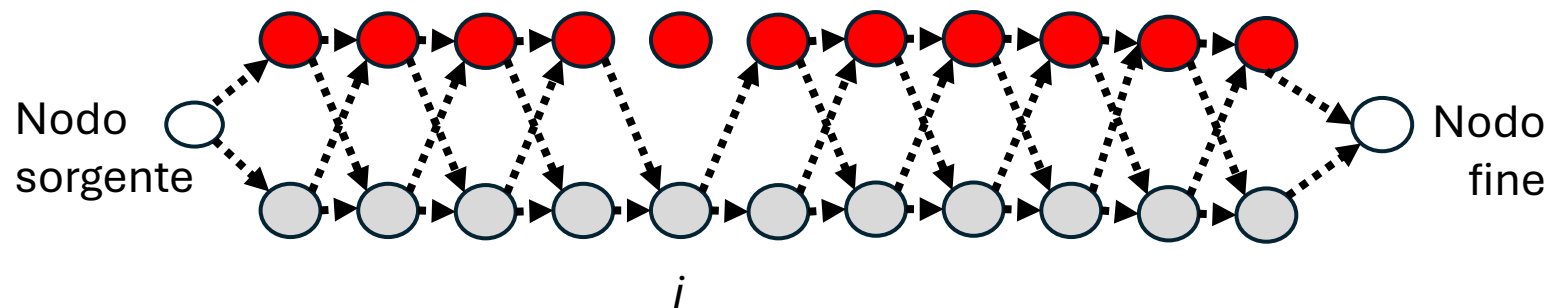
..con tutti i possibili percorsi (i testimoni), quello di piu' probabile in celeste.

Baum-Welch

Hidden Markov Models



Rappresentazione di Viterbi

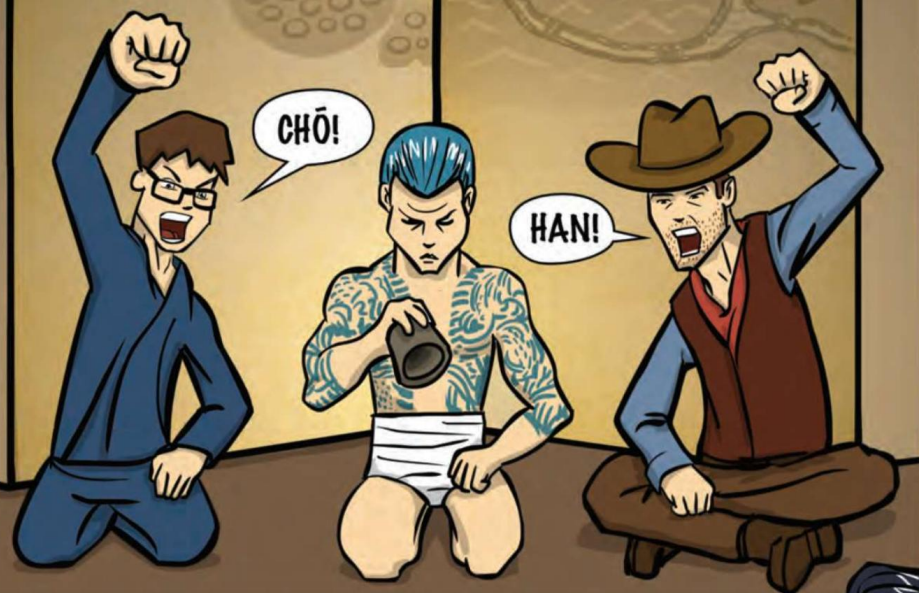


Idea intuitiva

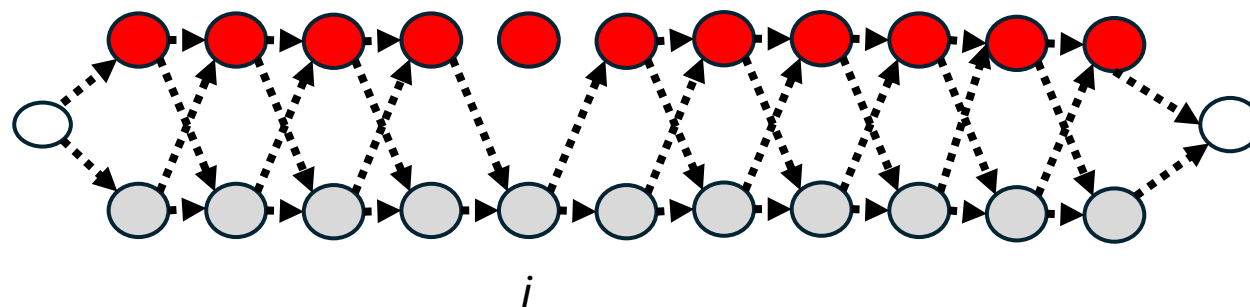
Per ogni **posizione i** posso calcolare la somma di tutte le probabilità dei **percorsi che passano in** 

Baum-Welch

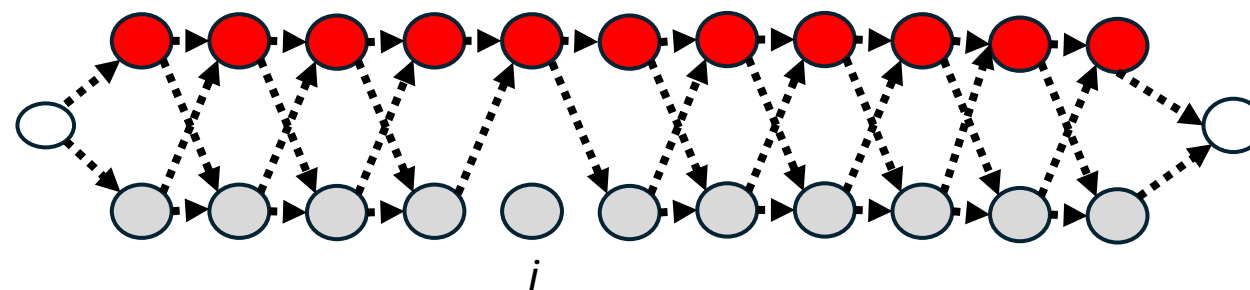
Hidden Markov Models



Rappresentazione di Viterbi



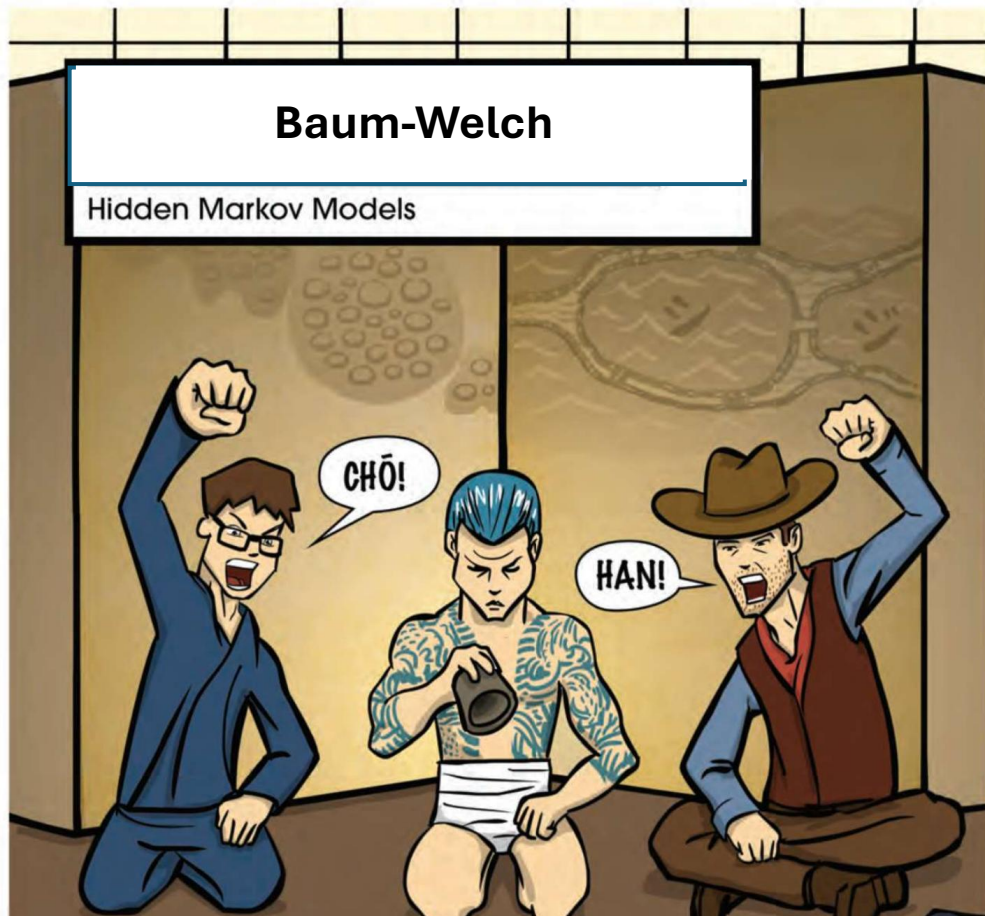
Per ogni **posizione** i posso calcolare la somma di tutte le probabilità dei **percorsi che passano in**



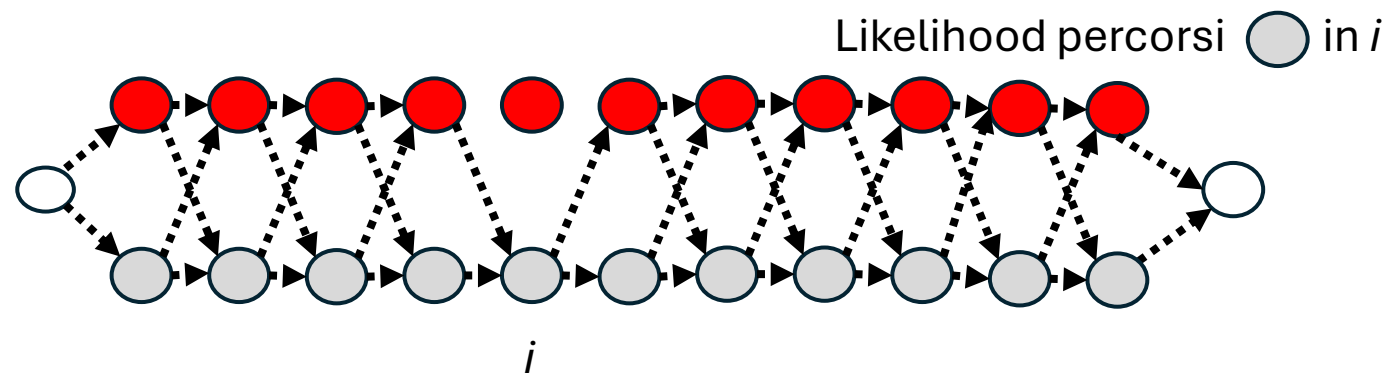
Allo stesso modo posso calcolare la somma delle probabilità di tutti i **percorsi che passano per**

Baum-Welch

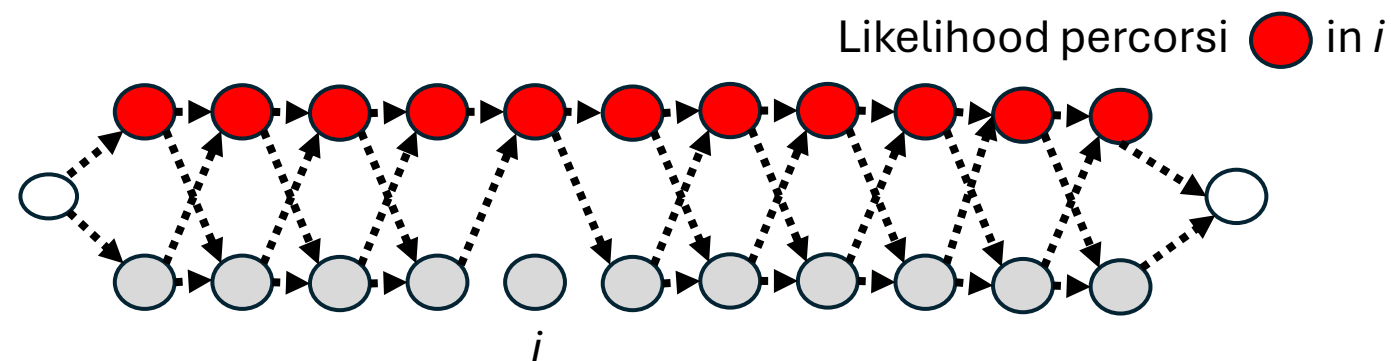
Hidden Markov Models



Rappresentazione di Viterbi



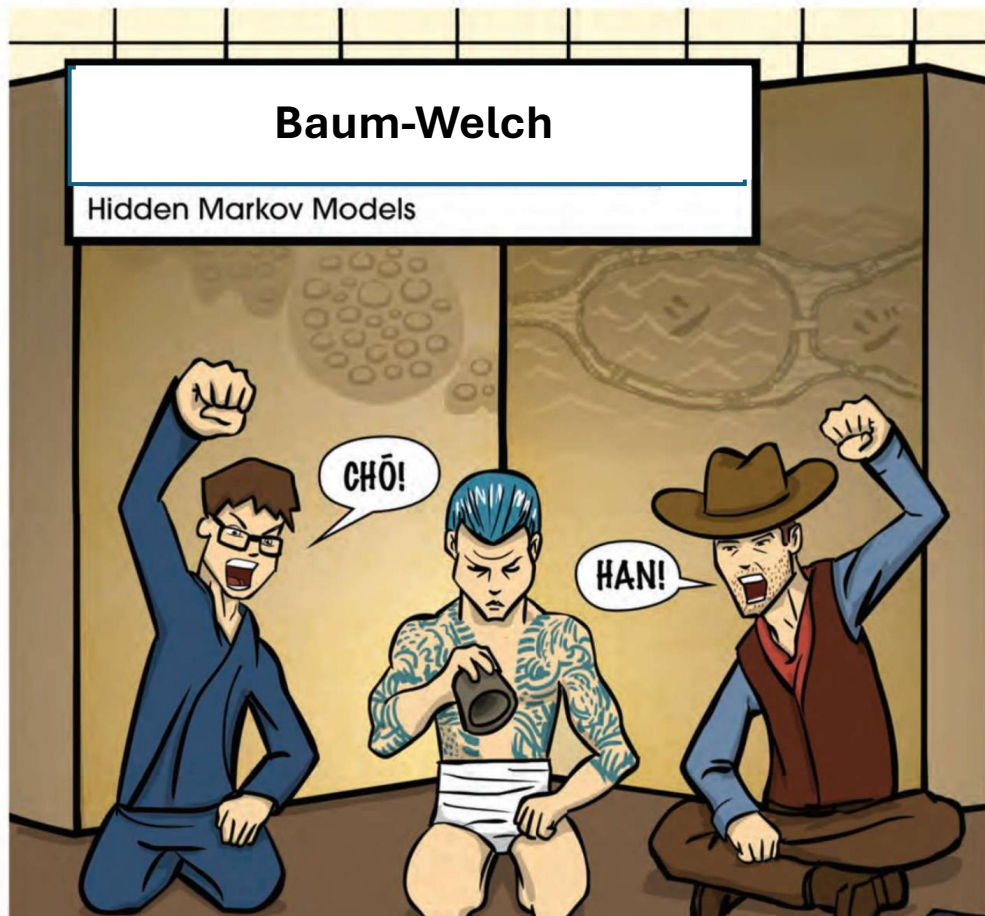
Per ogni posizione i posso calcolare la somma di tutte le probabilità dei percorsi che passano in 



Allo stesso modo posso calcolare la somma delle probabilità di tutti i percorsi che passano per 

Baum-Welch

Hidden Markov Models

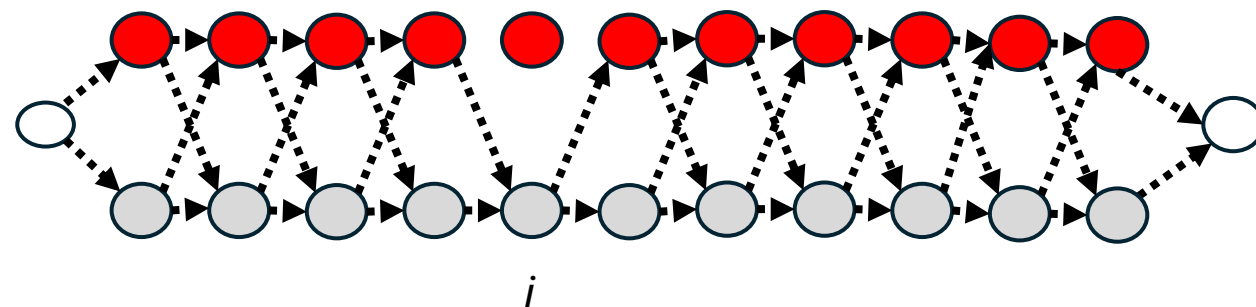


Probabilità (o
responsabilità
normalizzata)
che posizione i
sia

$$= \frac{\text{Likelihood percorsi } \text{gray circle} \text{ in } i}{\text{Likelihood percorsi } \text{gray circle} \text{ in } i + \text{Likelihood percorsi } \text{red circle} \text{ in } i}$$

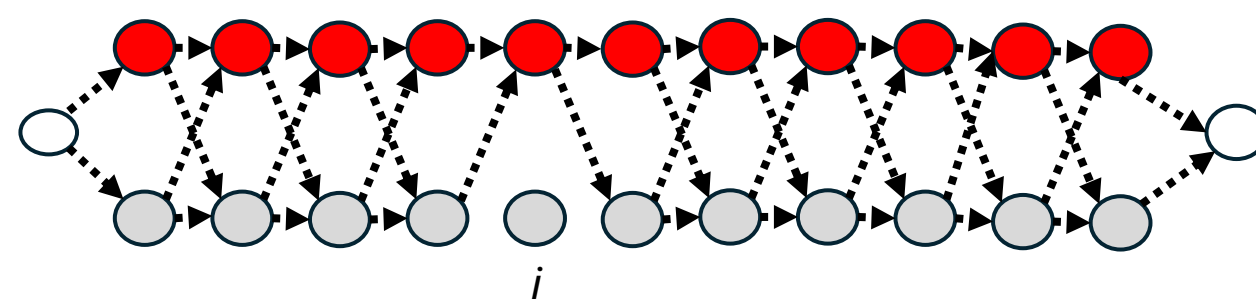
Rappresentazione di Viterbi

Likelihood percorsi in i



Per ogni posizione i posso calcolare la somma di tutte le
probabilità dei percorsi che passano in

Likelihood percorsi in i



Allo stesso modo posso calcolare la somma delle probabilità
di tutti i percorsi che passano per

Baum-Welch

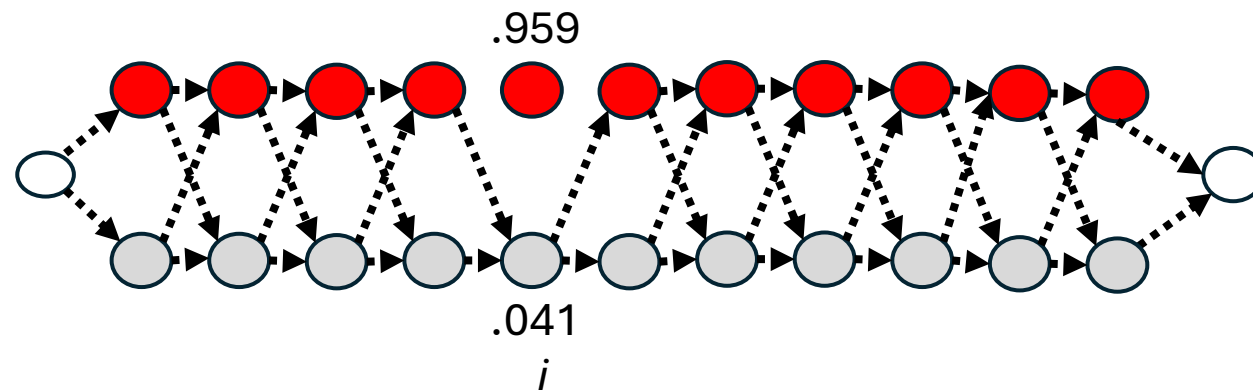
Hidden Markov Models



Probabilità (o
responsabilità
normalizzata)
che posizione i
sia

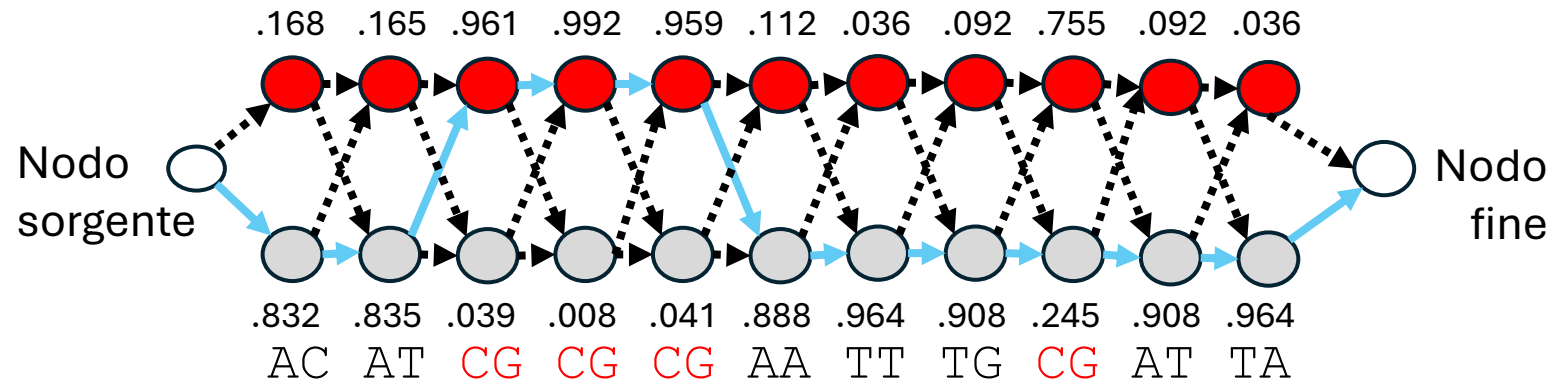
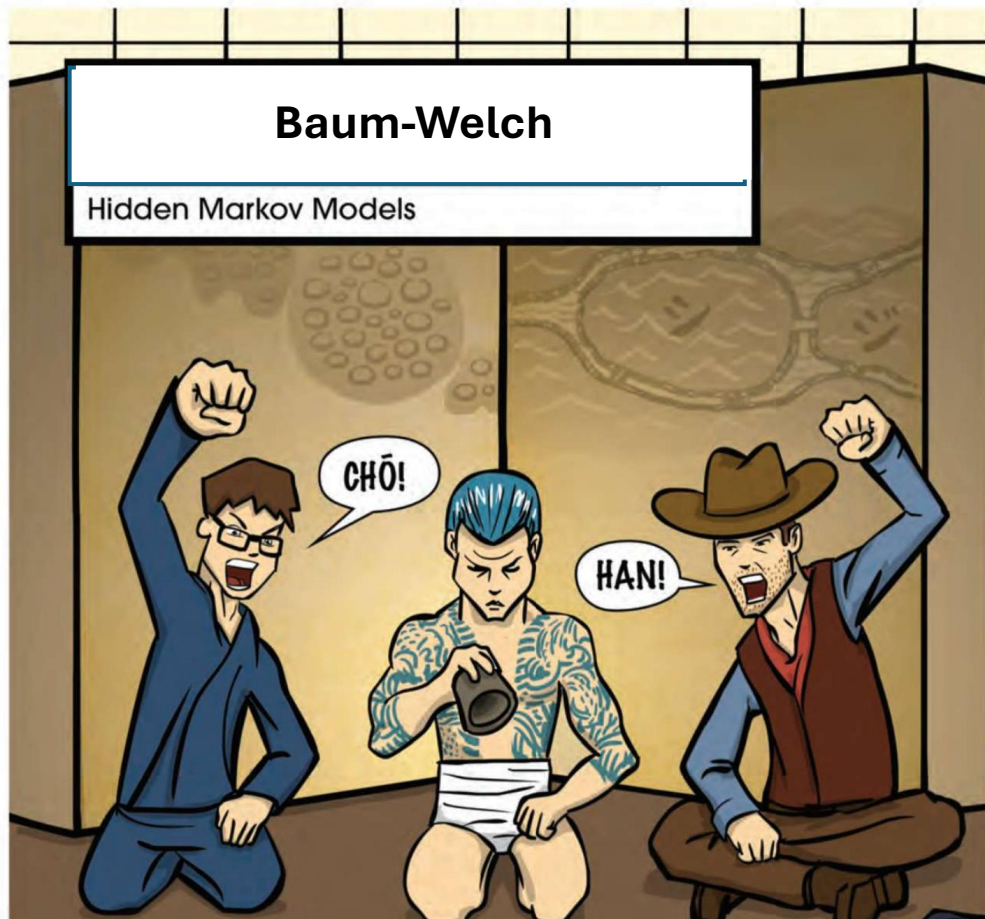
$$= \frac{\text{Likelihood percorsi } \text{grey circle icon} \text{ in } i}{\text{Likelihood percorsi } \text{grey circle icon} \text{ in } i + \text{Likelihood percorsi } \text{red circle icon} \text{ in } i}$$

Rappresentazione di Viterbi



Baum-Welch

Hidden Markov Models

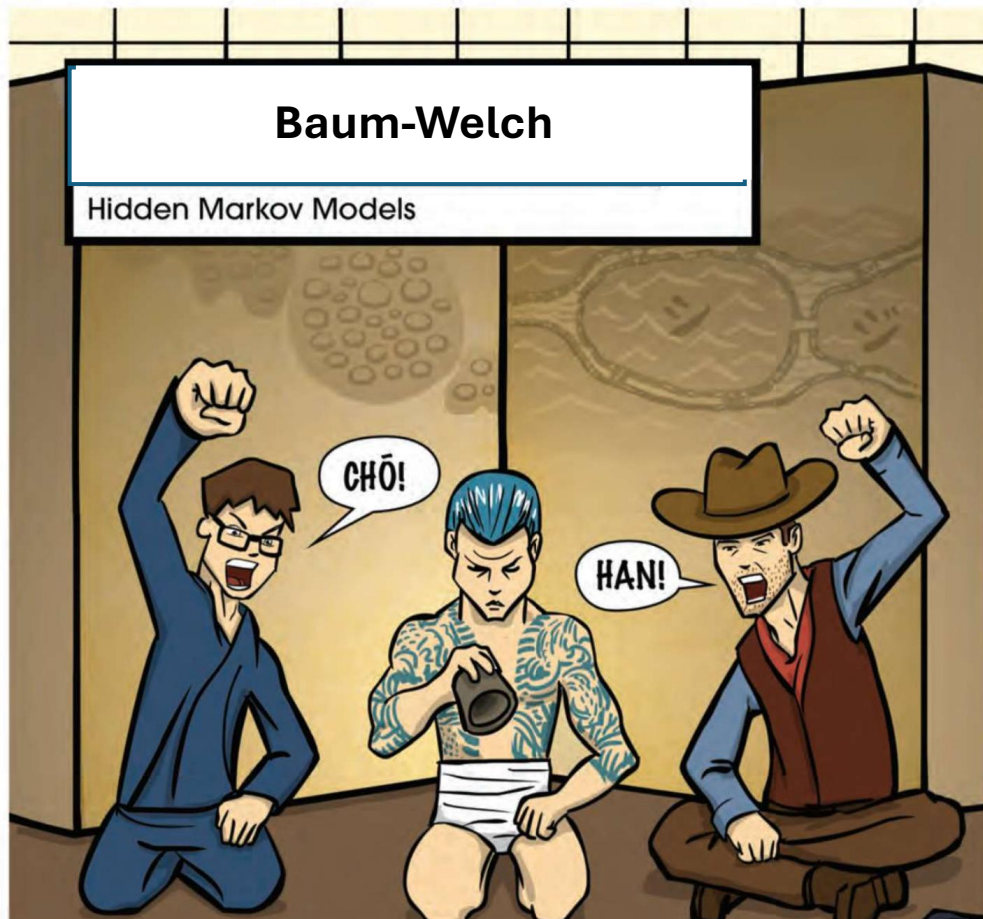


Probabilità (o
responsabilità
normalizzata)
che posizione i
sia grey circle

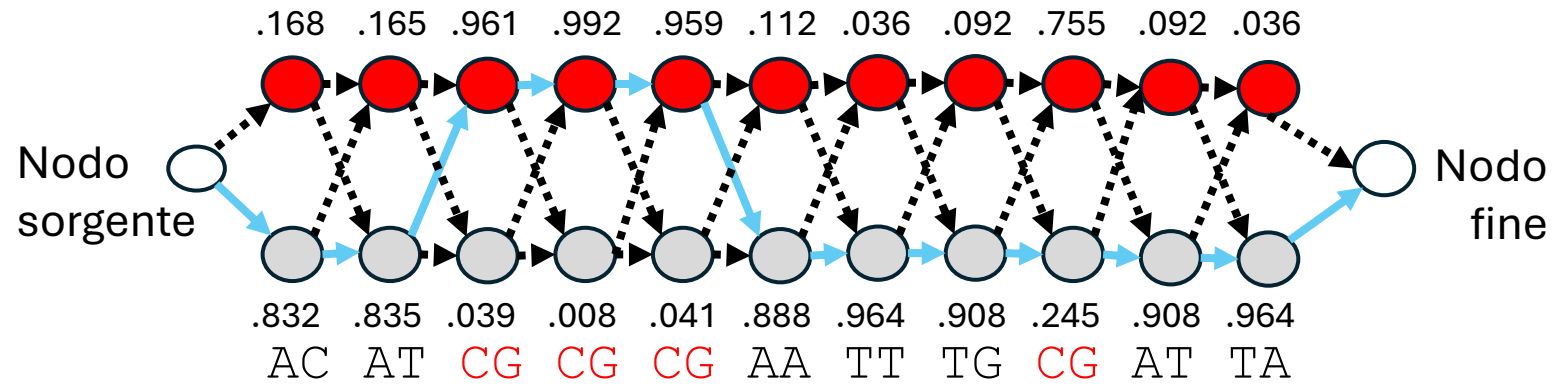
$$= \frac{\text{Likelihood percorsi } \text{grey circle} \text{ in } i}{\text{Likelihood percorsi } \text{grey circle} \text{ in } i + \text{Likelihood percorsi } \text{red circle} \text{ in } i}$$

Baum-Welch

Hidden Markov Models



Probabilità (o responsabilità **normalizzata**) che posizione i sia $\bigcirc$ =
$$\frac{\text{Likelihood percorsi } \bigcirc \text{ in } i}{\text{Likelihood percorsi } \bigcirc \text{ in } i + \text{Likelihood percorsi } \bullet \text{ in } i}$$

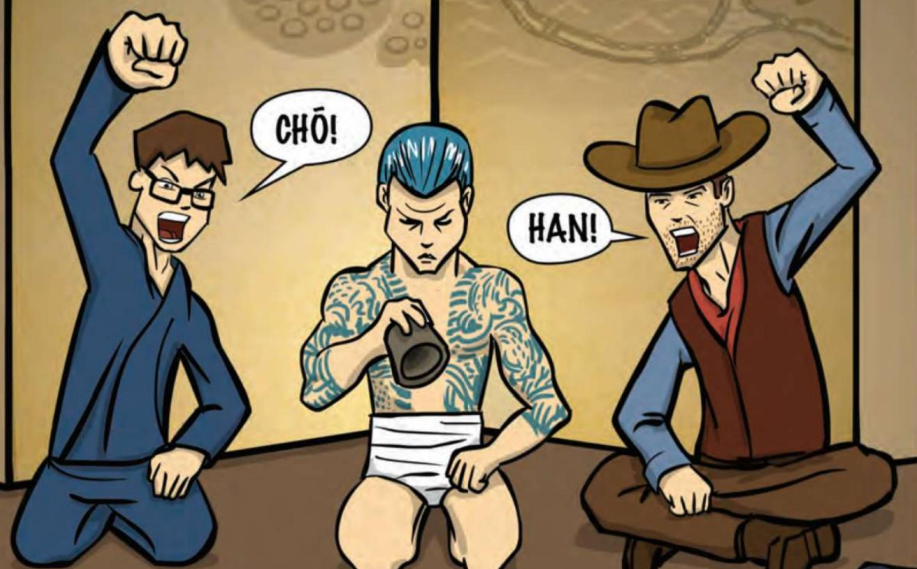


Tenere conto di tutti i possibili percorsi ci garantisce:

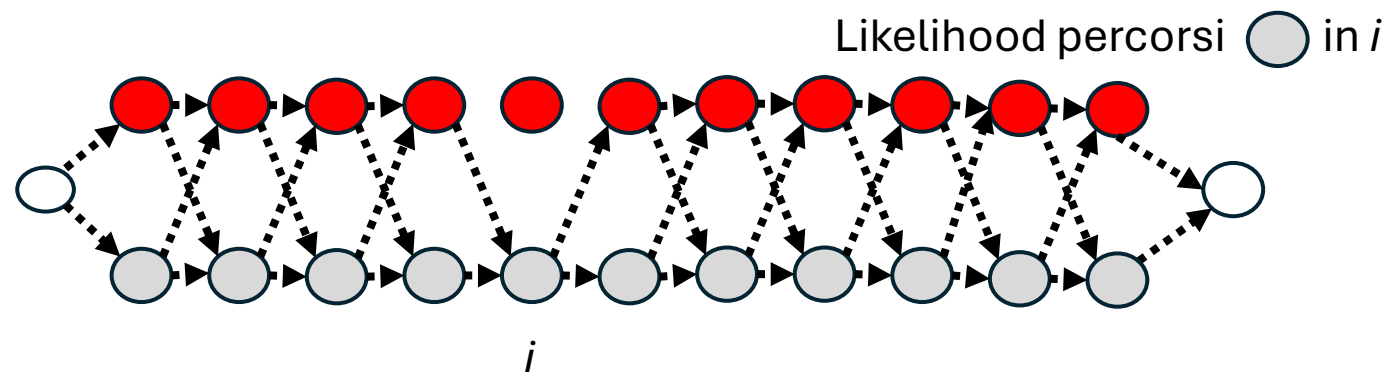
- **Probabilità esatte** di osservare i vari stati lungo il genoma;
- Una **misura di incertezza** (posizioni incerte risulteranno vicine a 0.5, mentre posizioni piu' sicure vicino allo 0% per uno stato vs 100% per l'altro)
- **Migliori stime di transizioni ed emissioni** (ora però dobbiamo fare una «media» pesata dei due stati tenendo conto delle loro probabilità invece di tener conto del singolo stato scelto da Viterbi per ogni posizione)
- In pratica quasi **tutte le HMM usate in genomica** (e non solo) **usano questo approccio**, chiamato Baum-Welch o forward/backward

Baum-Welch

Hidden Markov Models



Rappresentazione di Viterbi



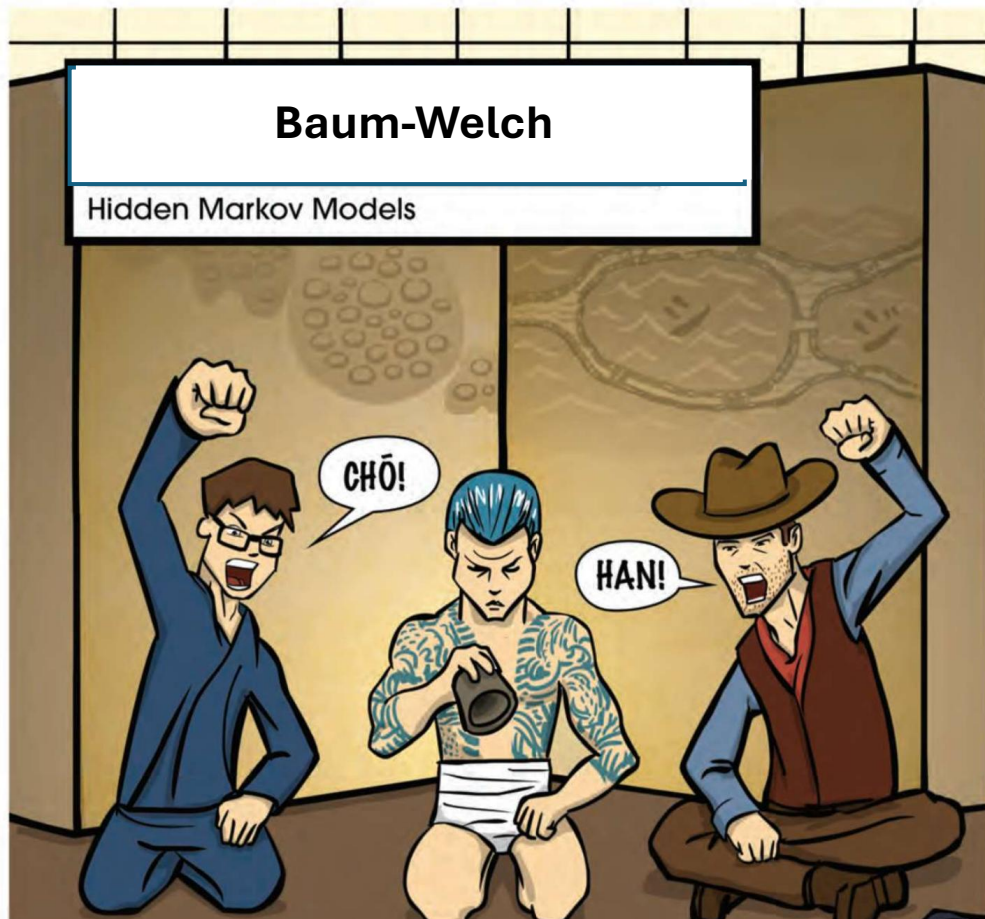
Per ogni posizione i possiamo calcolare la somma di tutte le probabilità dei percorsi che passano in 

Ma come?

(ricordiamo che non vogliamo calcolare separatamente le likelihood di tutti i percorsi visto che sono moltissimi)

Baum-Welch

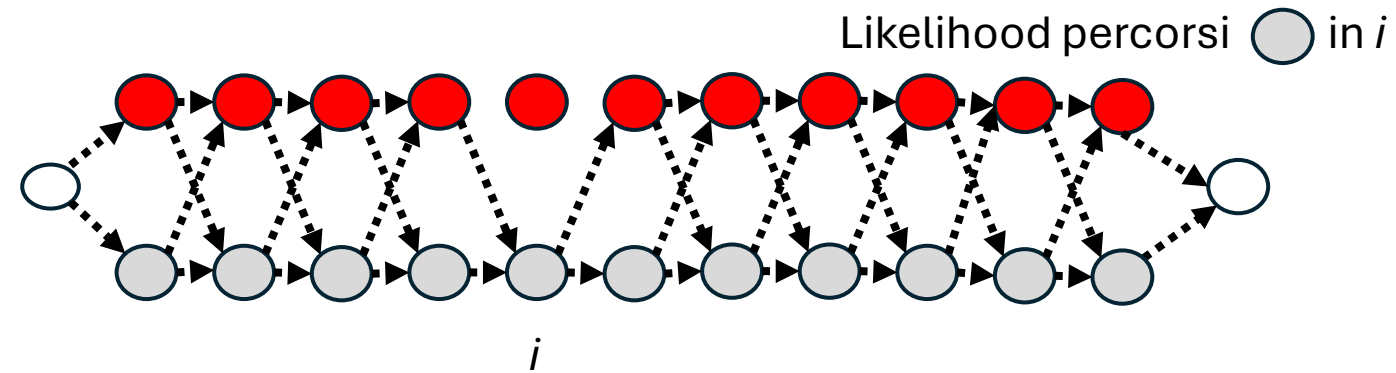
Hidden Markov Models



Nota pratica

Non preoccupatevi, non chiederò dettagli del calcolo delle probabilità all'esame. Ma Baum-Welch è molto usato quindi lo menzionerò e se volete potete approfondirlo.

Rappresentazione di Viterbi



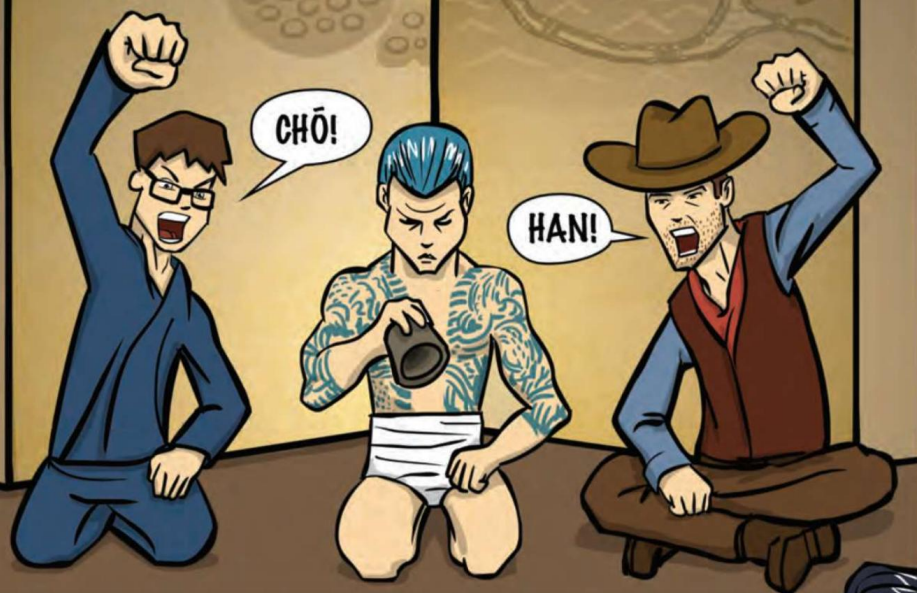
Per ogni posizione i possiamo calcolare la somma di tutte le probabilità dei percorsi che passano in 

Ma come?

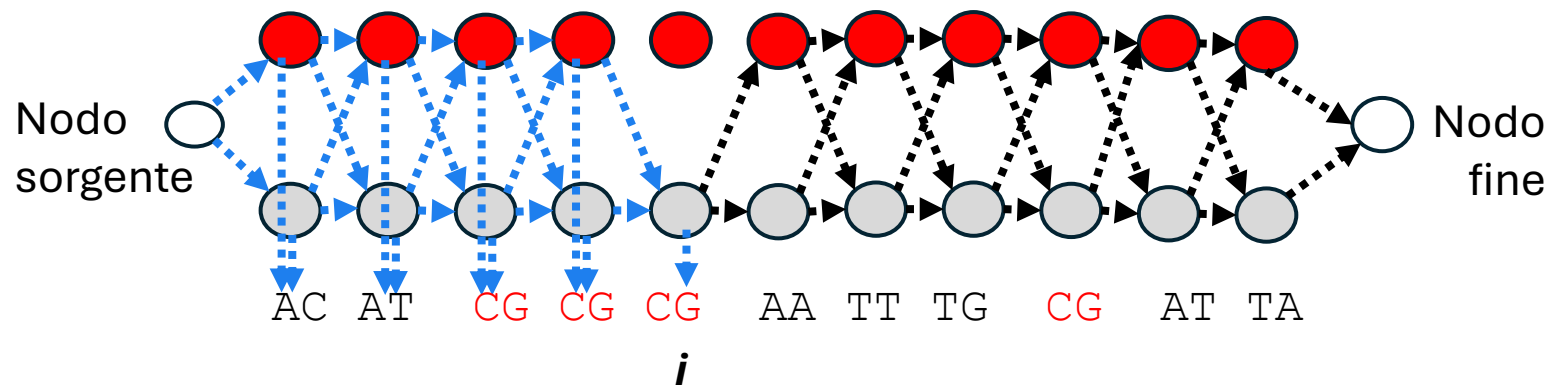
(ricordiamo che non vogliamo calcolare separatamente le likelihood di tutti i percorsi visto che sono moltissimi)

Baum-Welch

Hidden Markov Models



Rappresentazione di Viterbi

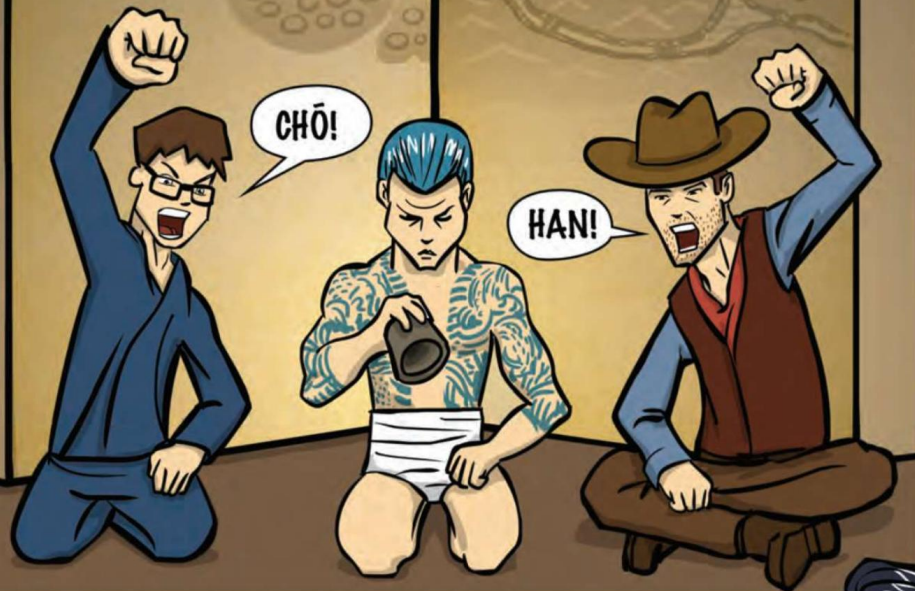


Per ogni posizione i vogliamo calcolare la somma di tutte le probabilità dei percorsi che passano in

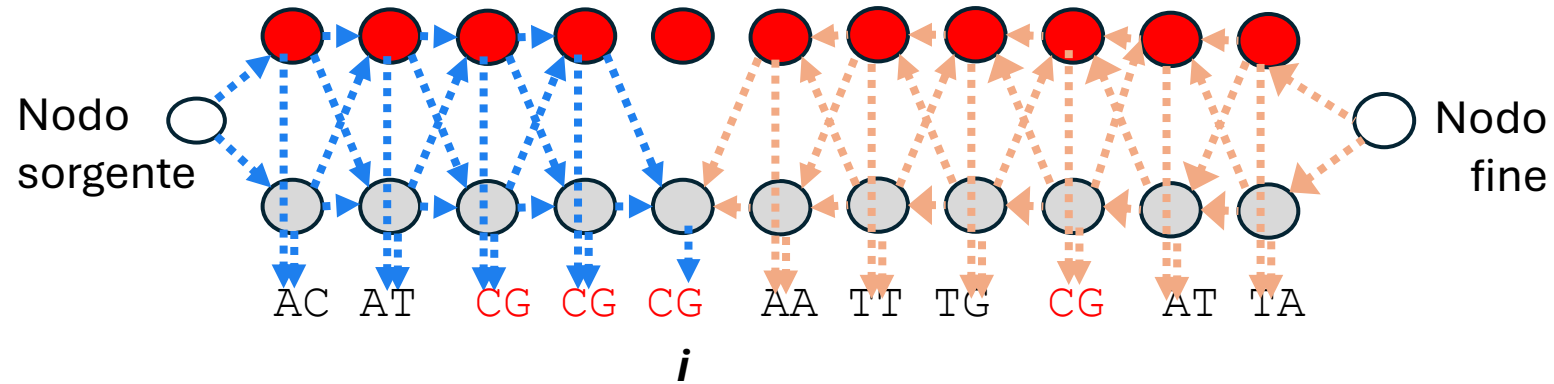
Possiamo farlo prima prendendo la somma* di tutti i percorsi fino a i , che chiameremo *forward(i , nonCPG)*

Baum-Welch

Hidden Markov Models



Rappresentazione di Viterbi



Per ogni posizione i vogliamo calcolare la somma di tutte le probabilità dei percorsi che passano in 

Possiamo farlo prima prendendo la somma* di tutti i percorsi fino a i , che chiameremo *forward*(i , nonCPG) e quella di tutti i percorsi da i alla fine, che chiameremo *backward*(i , nonCPG)

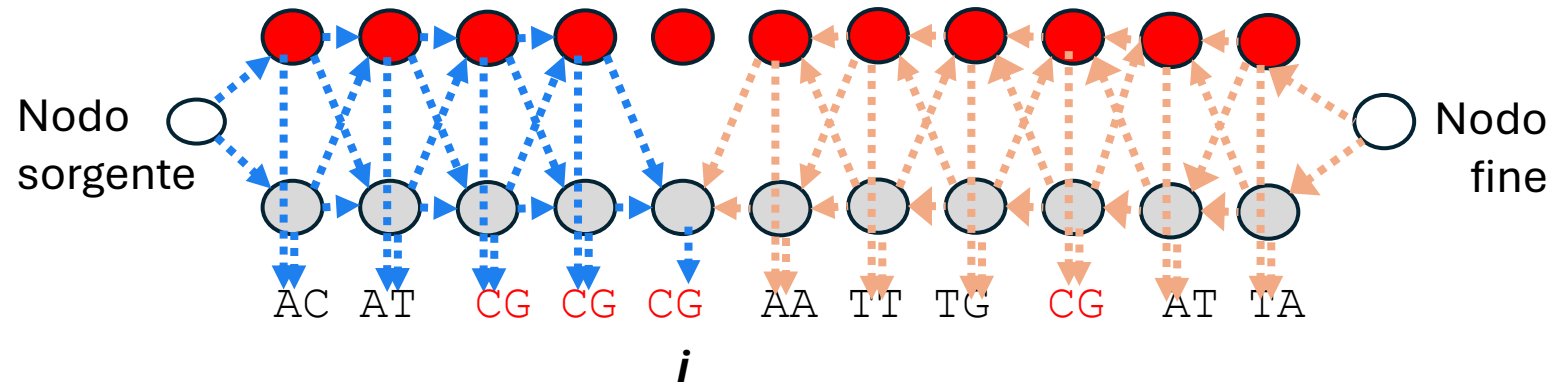
Baum-Welch

Hidden Markov Models



*I calcoli sono analoghi a quelli dell'algoritmo di Viterbi, con l'unica differenza che **sommiamo le probabilità** di tutti percorsi, invece di escluderne alcuni prendendo solo quello con la probabilità massima per ogni stato. Non li ripetiamo qui ma potete farlo per esercizio e verificare le probabilità mostrate sulle slides e sull'**esercizio supplementare** allegato (caricato su Moodle)!

Rappresentazione di Viterbi



Per ogni posizione i vogliamo calcolare la somma di tutte le probabilità dei percorsi che passano in 

Possiamo farlo prima prendendo la somma* di tutti i percorsi fino a i , che chiameremo *forward(i , nonCPG)* e quella di tutti i percorsi da i alla fine, che chiameremo *backward(i , nonCPG)*

Esercizio supplementare su Moodle

Nota pratica

Non chiederò dettagli o esercizi di calcolo delle probabilità forward o backward all'esame.

Esempio HMM CpG: Forward, Backward e Posteriori (prime 3 posizioni)

Modello HMM (toy CpG)

Stati nascosti: nonCpG, CpG

Transizioni:

$$P(\text{nonCpG} \rightarrow \text{nonCpG}) = 0.9, \quad P(\text{nonCpG} \rightarrow \text{CpG}) = 0.1$$

$$P(\text{CpG} \rightarrow \text{nonCpG}) = 0.5, \quad P(\text{CpG} \rightarrow \text{CpG}) = 0.5$$

Emissioni:

$$P(\text{AT} | \text{nonCpG}) = 0.98, \quad P(\text{CG} | \text{nonCpG}) = 0.02$$

$$P(\text{AT} | \text{CpG}) = 0.20, \quad P(\text{CG} | \text{CpG}) = 0.80$$

Probabilità iniziali: $\pi(\text{nonCpG}) = \pi(\text{CpG}) = 0.5$

Sequenza osservata (prime 3 posizioni): AT, AT, CG

(in realtà la sequenza completa è AT AT CG CG CG AT AT AT CG AT AT, ma qui mostriamo i calcoli dettagliati solo per $t=1,2,3$.)

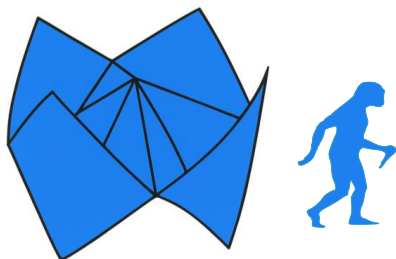
Forward α (prime 3 posizioni)

Definizione: $\alpha_t(s) = P(o_1, \dots, o_t, z_t = s)$

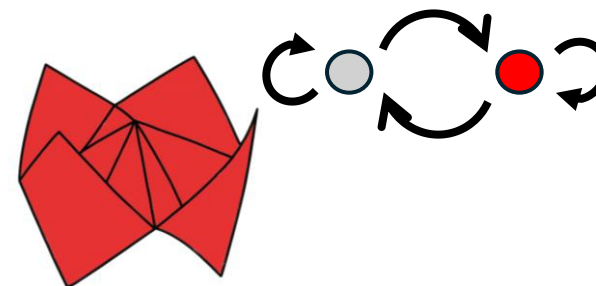
Posizione $t = 1$ (osservazione AT)

$$\begin{aligned} \alpha_1(\text{nonCpG}) &= \pi(\text{nonCpG}) \cdot P(\text{AT} | \text{nonCpG}) \\ &= 0.5 \cdot 0.98 \\ &= 0.49 \end{aligned}$$

$$\begin{aligned} \alpha_1(\text{CpG}) &= \pi(\text{CpG}) \cdot P(\text{AT} | \text{CpG}) \\ &= 0.5 \cdot 0.20 \\ &= 0.10 \end{aligned}$$



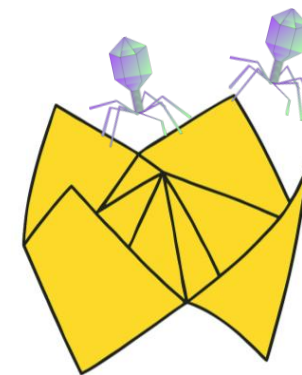
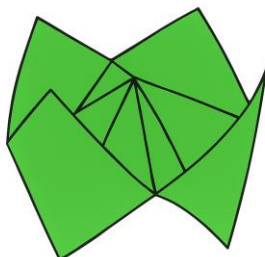
Aplotipi introgressi
da altre specie



Isole CpG

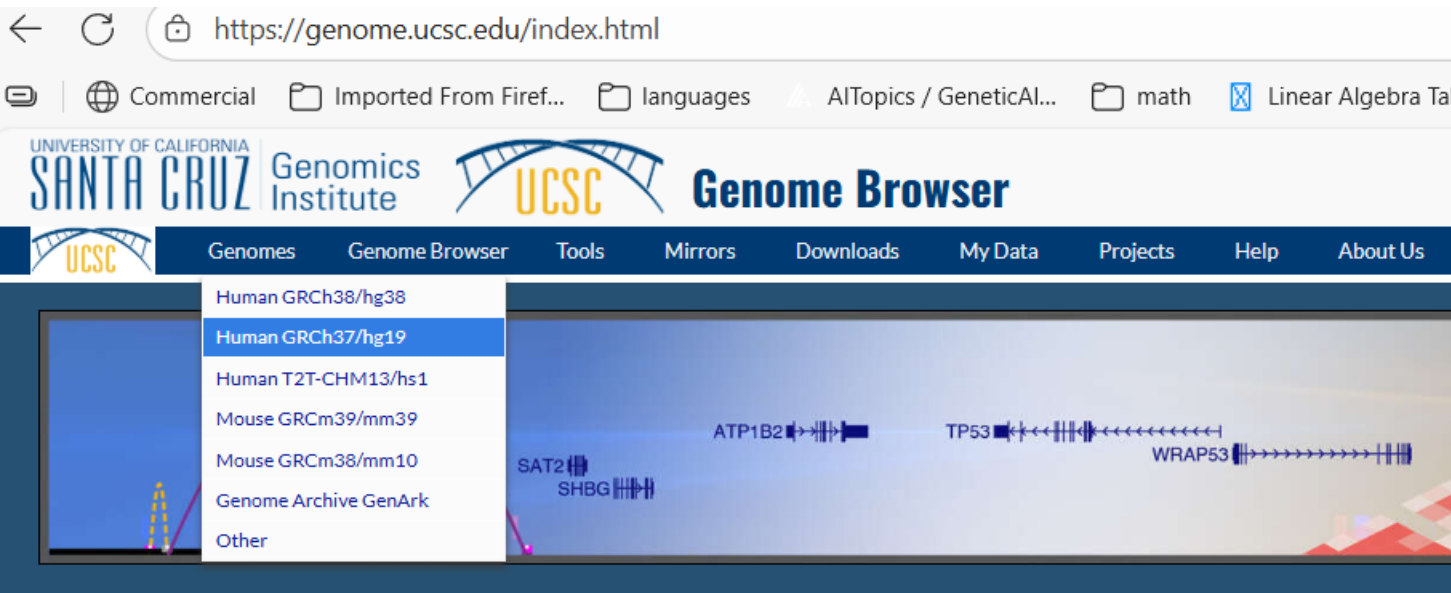
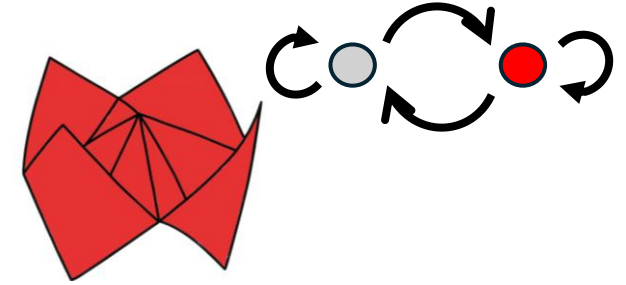
Applicazioni degli Hidden Markov Model in genomica

Ripetizioni nel genoma



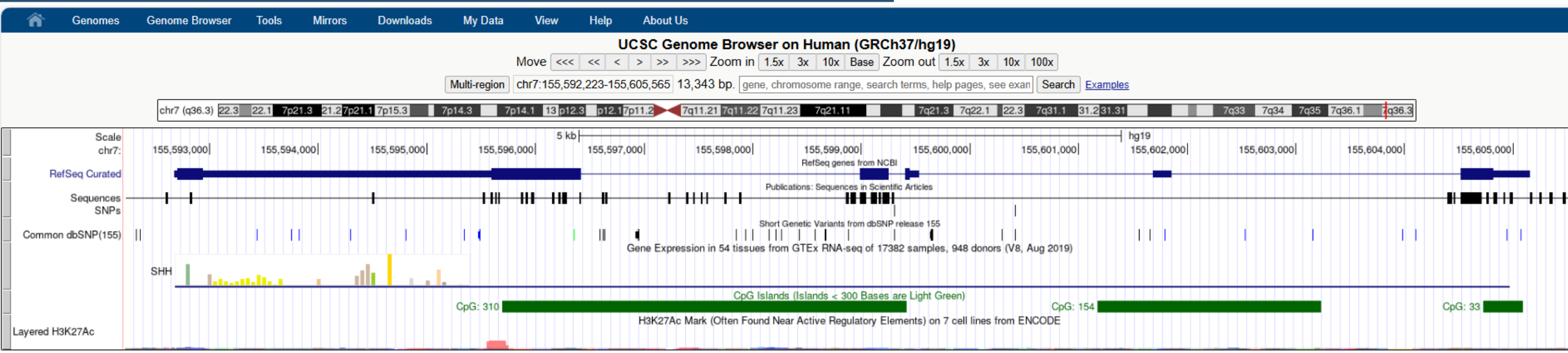
Integrazione virale

Isole CpG

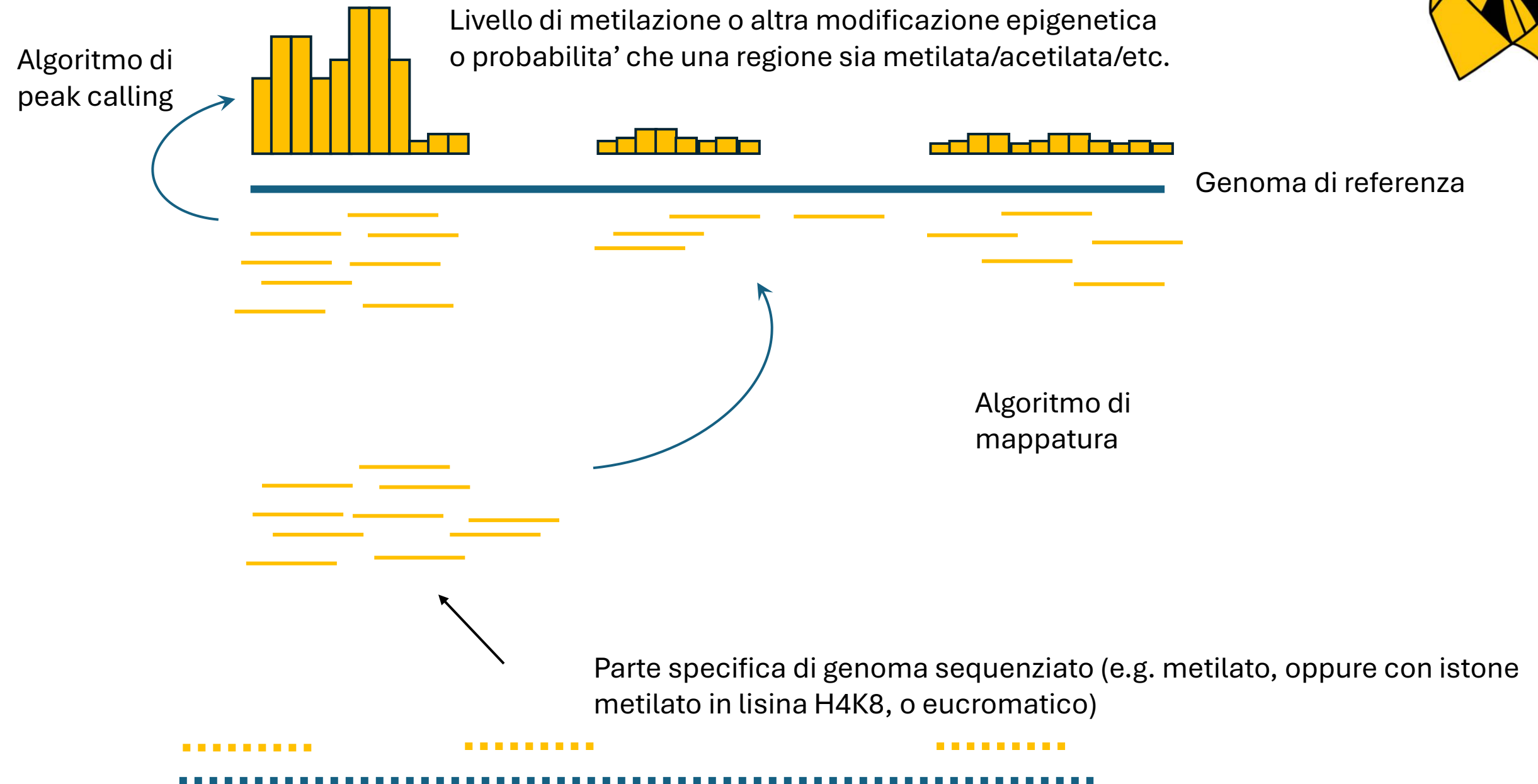


Varie HMM pubblicate per l'identificazione di isole CpG (esempio Wu et al., 2010, *Biostatistics*)

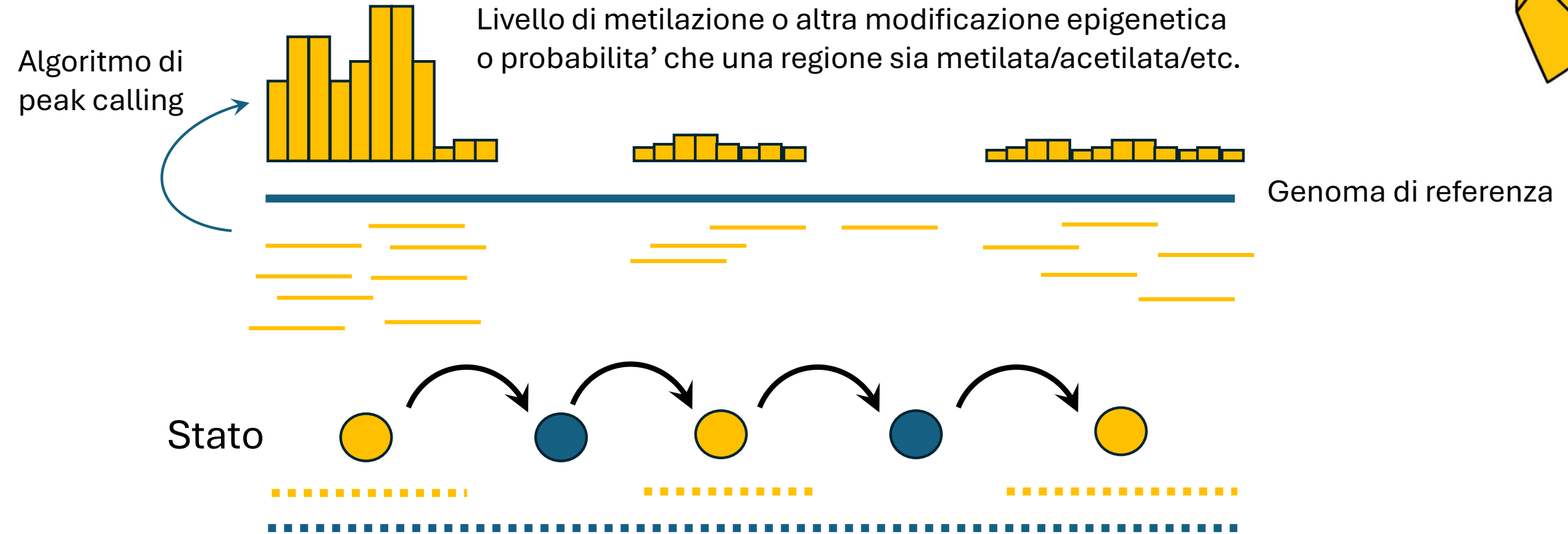
Capitolo di Robeva et al., 2013 su Moodle



Peak calling per epigenomica



Peak calling per epigenomica



Esempio:

Regione non-metilata



Regione metilata

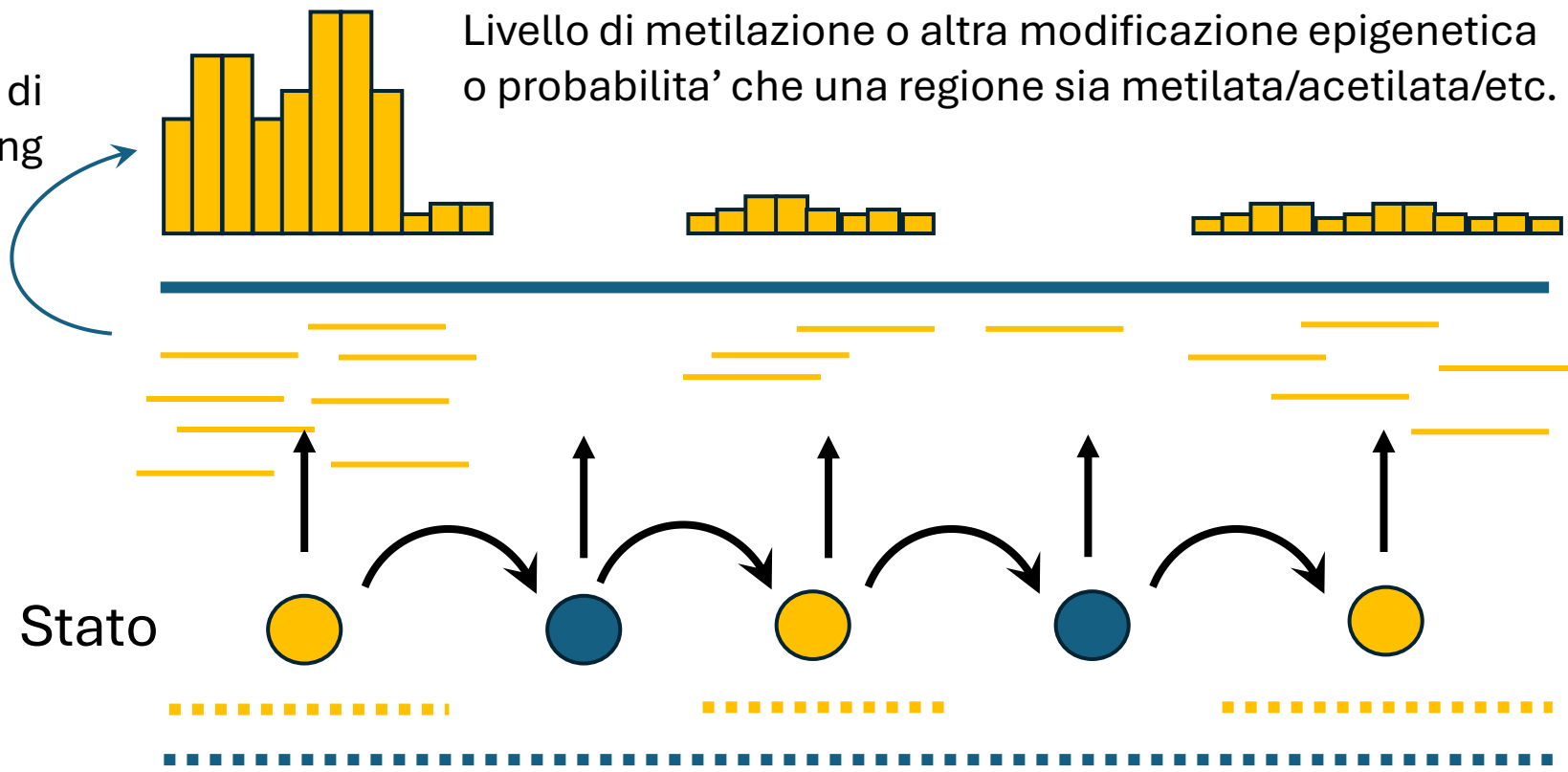


Peak calling per epigenomica



Livello di metilazione o altra modificazione epigenetica
o probabilita' che una regione sia metilata/acetilata/etc.

Algoritmo di
peak calling



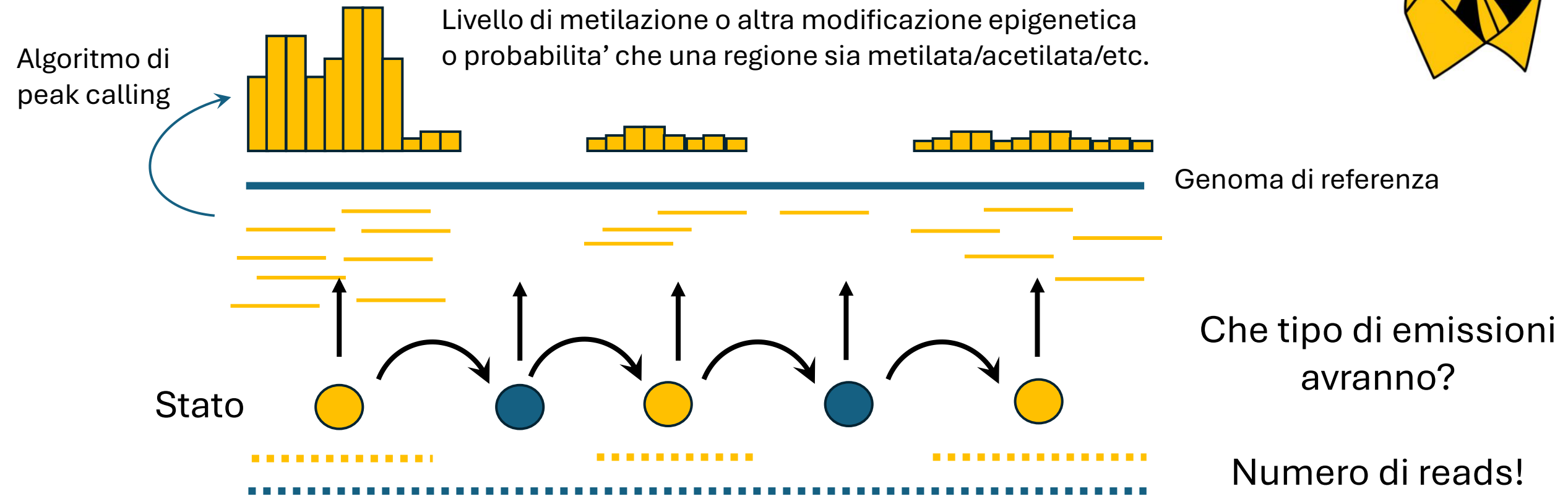
Genoma di referenza

Che tipo di emissioni
avranno?

Esempio:

- Regione non-metilata 
- Regione metilata 

Peak calling per epigenomica



Esempio:

Regione non-metilata



Regione metilata

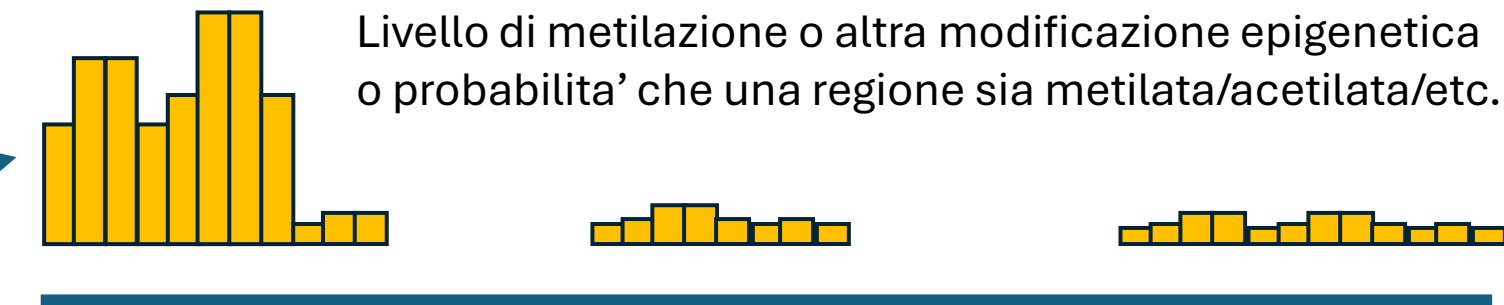


Peak calling per epigenomica

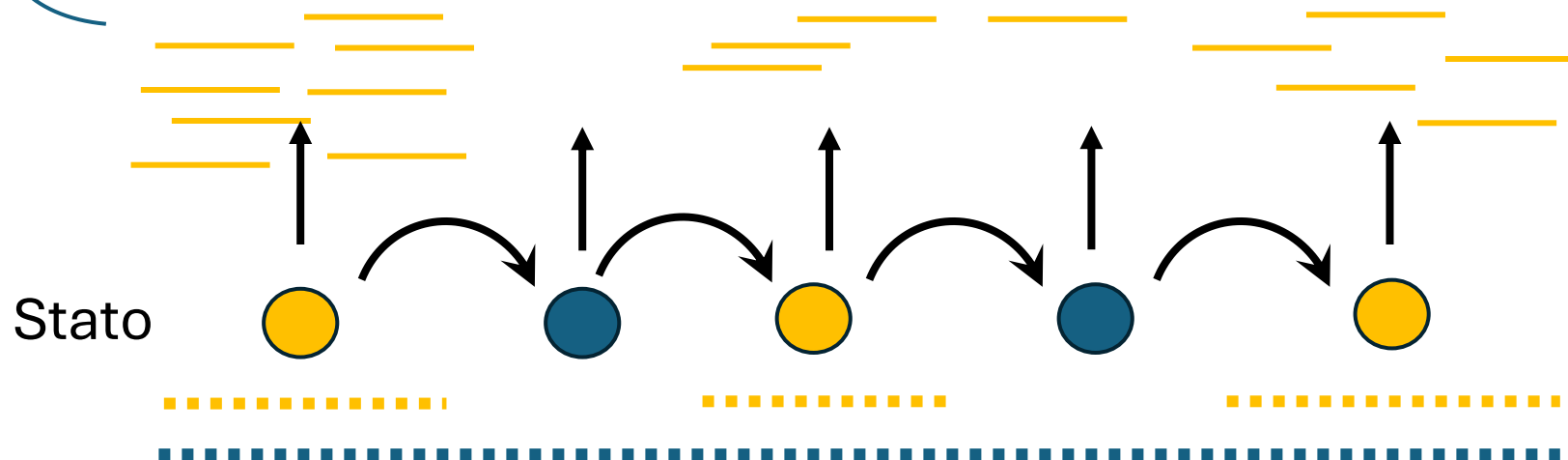


Livello di metilazione o altra modificazione epigenetica
o probabilit  che una regione sia metilata/acetilata/etc.

Algoritmo di
peak calling



Genoma di referenza



Che tipo di emissioni
avranno?

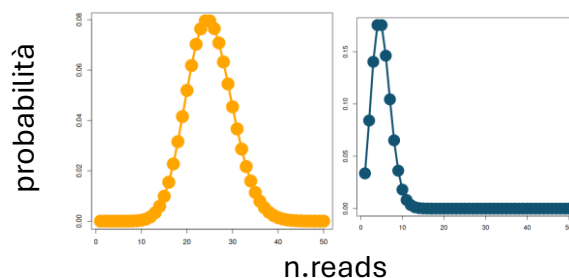
Stato

Numero di reads medio

25

5

Distribuzione
di probabilit 
(Poisson)



Assumiamo una
distribuzione di probabilit ,
ad esempio una Poisson o
una Gaussiana,
e stimiamo la media del
numero di reads,
che definir  la nostra
probabilit  per i due stati!

Numero di reads!
E come otterremo la
probabilit  di
osservarle?



Metilazione del DNA e bisulfite sequencing

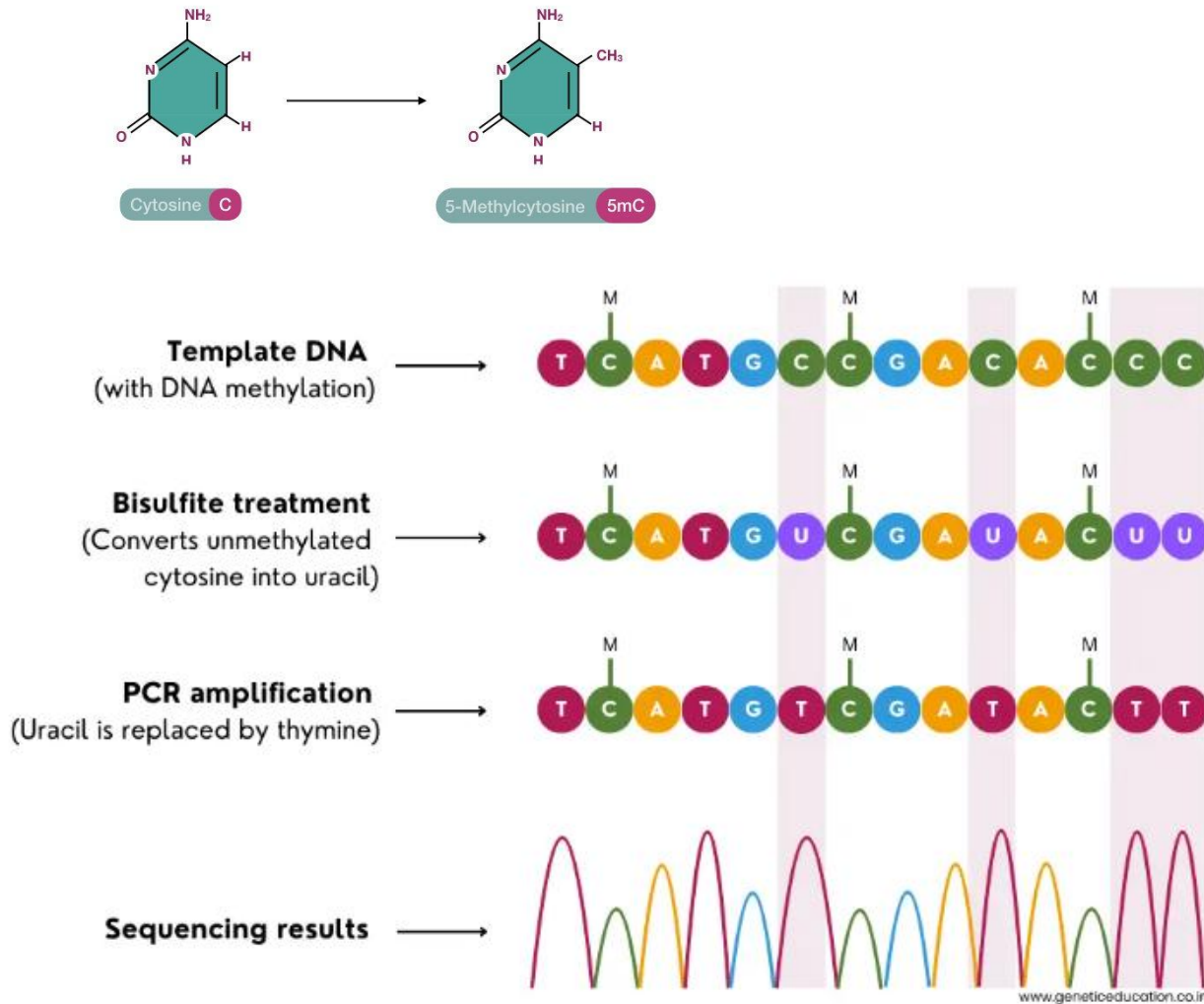


Illustration of the complete bisulfite sequencing process.

Metilazione del DNA e bisulfite sequencing

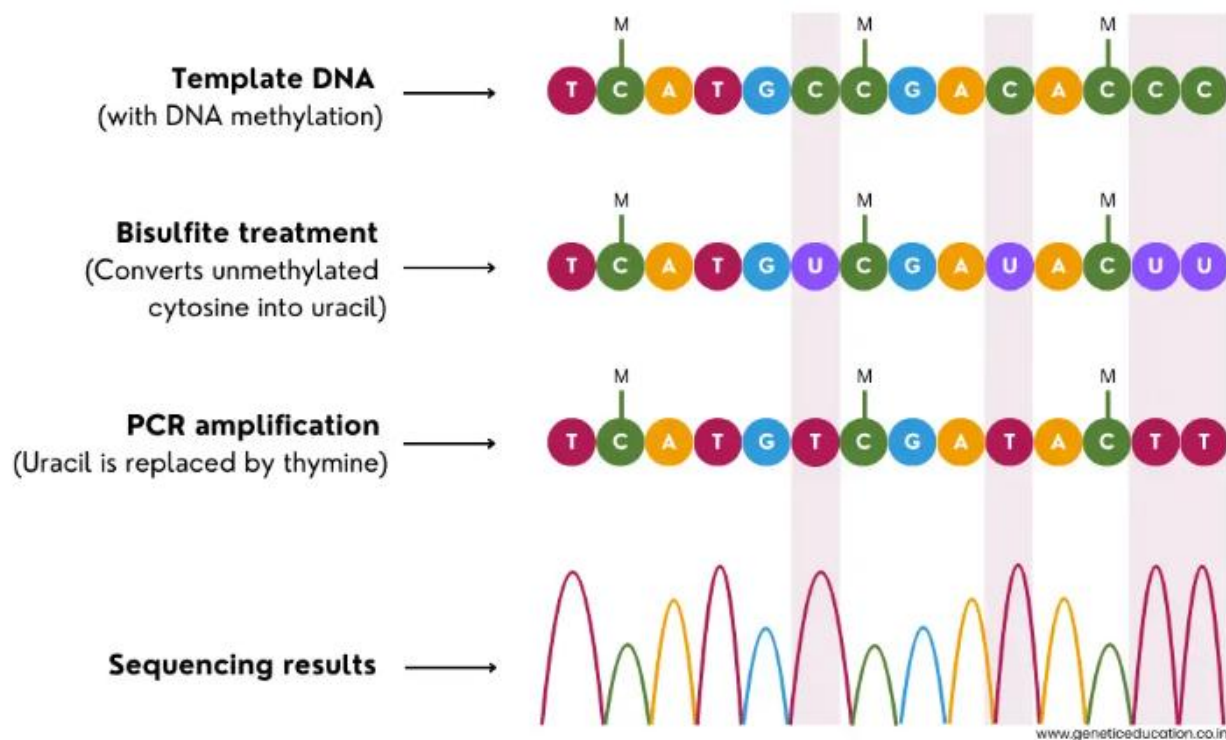
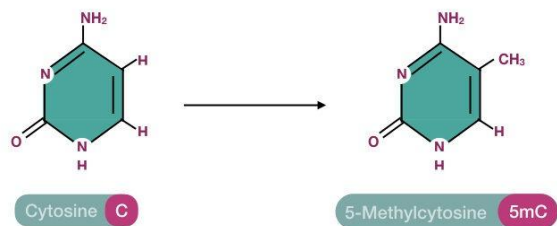
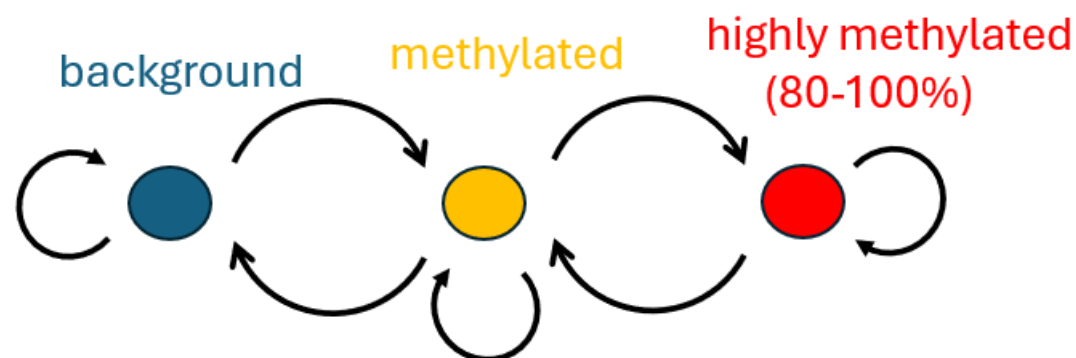


Illustration of the complete bisulfite sequencing process.

smithlabcode/
methpipe

A pipeline for analyzing DNA methylation data from bisulfite sequencing.

19 Contributors 0 Issues 70 Stars 27 Forks



<https://smithlabresearch.org/software/methpipe/>

Mapping reads: rmapbs

filtering high quality read
trimming adapters
wildcard mapping

T rich
5-A A C G T T G C
5-A A C G T C G C T G T
A rich
T T G C A G C G A G C A-3
A A C G A G C A-3

clipping mates, fixing overlap
5-A A C G T T G C
A A C G A G C A-3
5-A A C G T T G C T C G T-3

Removing duplicate reads: duplicate-remover

Randomly select one read from multiple
duplicate reads mapped to the same position

Estimating methylation levels: methcounts

-T-
-C- -T-
-C- -T-
-C- -T-
5-A A C G T C G C T C G T
T T G C A G C G A G C A-3
-C-
4 C 1 T
meth: 4/5
coverage: 5
CpG site
0 C 3 T
meth: 0/3
coverage: 3
non-CpG site

Methylation levels in regions: roimethstats

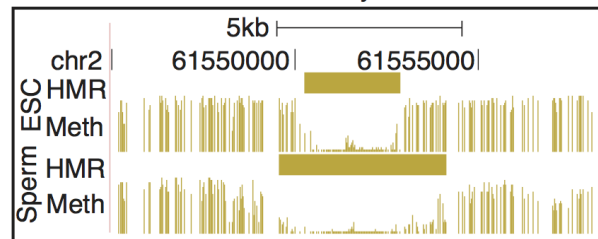
1. fraction of methylated cytosines;
2. mean methylation level;
3. methylation level averaged by coverage

Estimating bisulfite conversion rate: bsrate

1. spike-in DNA: lambda DNA
2. non-CpG cytosines
3. chloroplast or mitochondria DNA

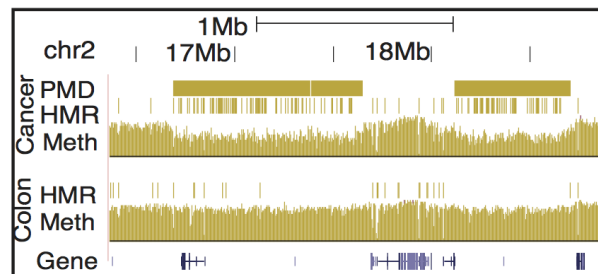
Hypo-methylated regions: hmr

mammalian methylomes



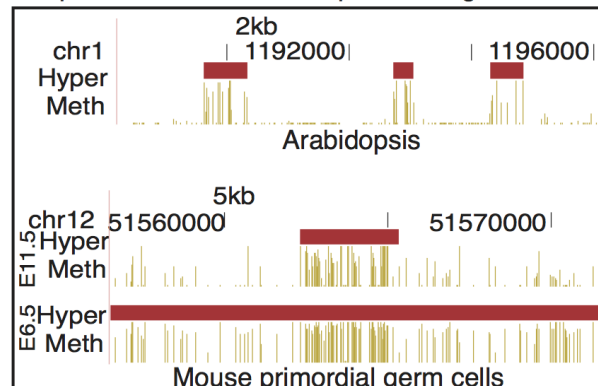
Partially methylated domains: pmd

immortalized cell lines and cancerous cells

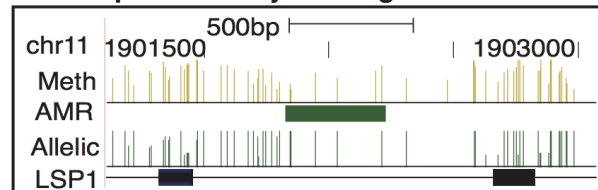


Hyper-methylated domains: hmr_plant

plant, invertebrate, and primordial germ cells



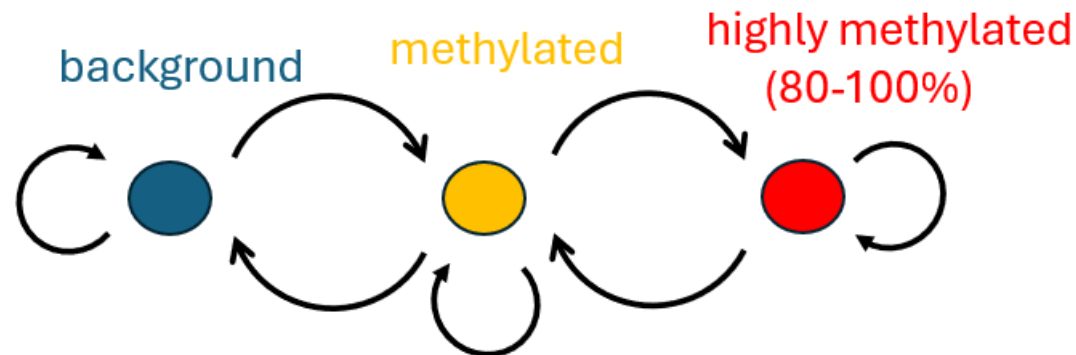
Allele-specific methylated regions: amrfinder



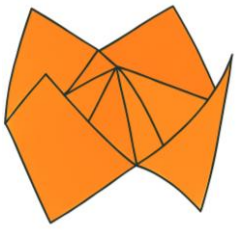
smithlabcode/ methpipe

A pipeline for analyzing DNA methylation data from
bisulfite sequencing.

19 Contributors 0 Issues 70 Stars 27 Forks

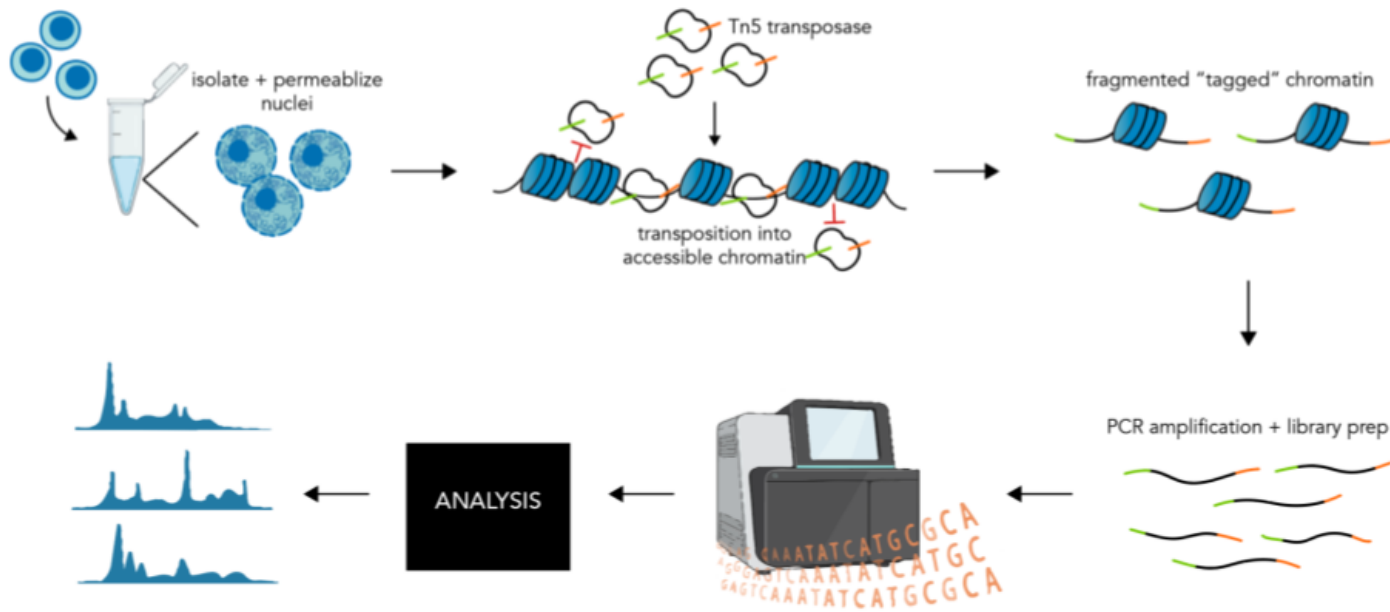


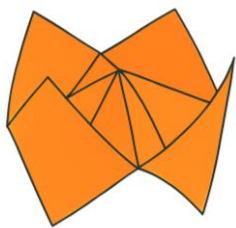
<https://smithlabresearch.org/software/methpipe/>



ATAC-seq

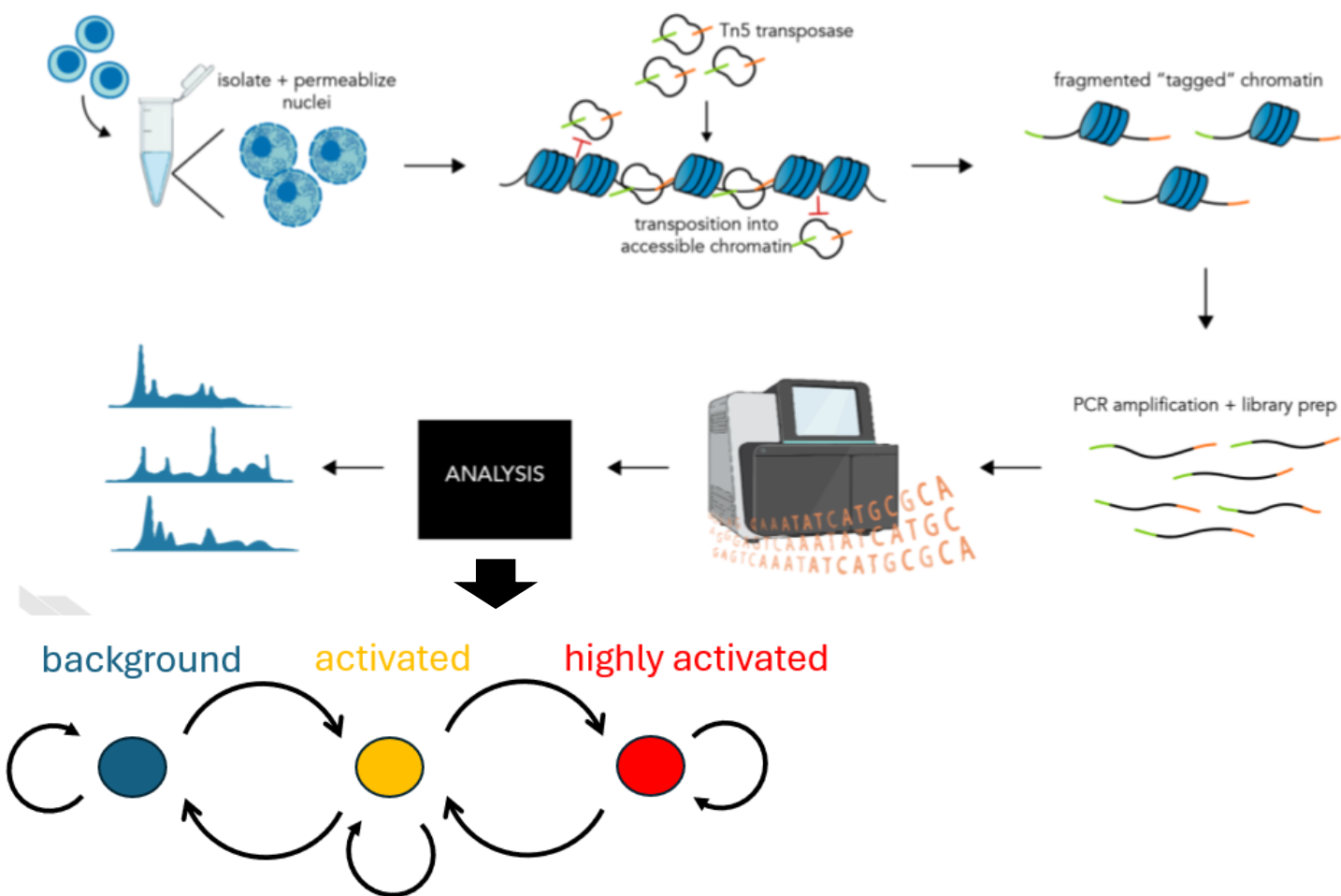
Overview of ATAC-seq





ATAC-seq

Overview of ATAC-seq

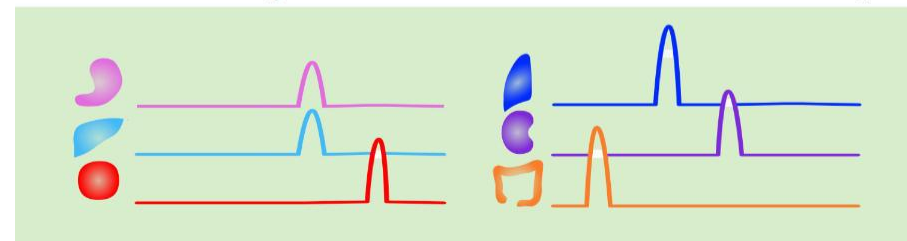
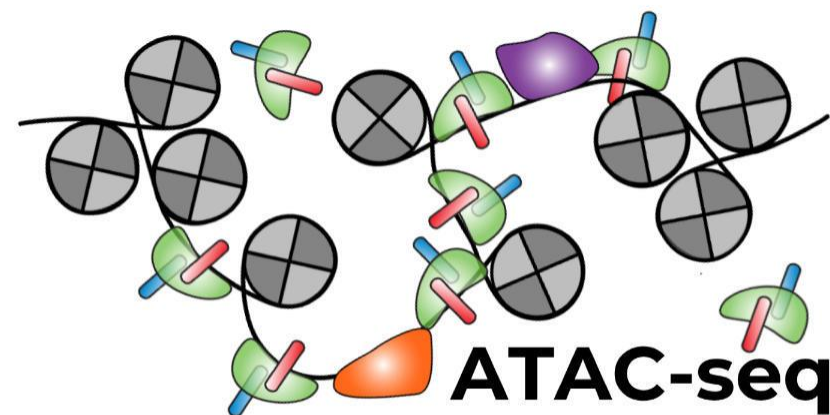


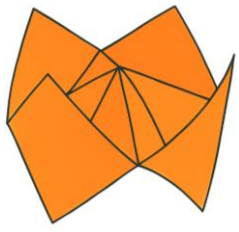
LiuLabUB/
HMMRATAC



HMMRATAC peak caller for ATAC-seq data

2 Contributors 40 Issues 99 Stars 22 Forks





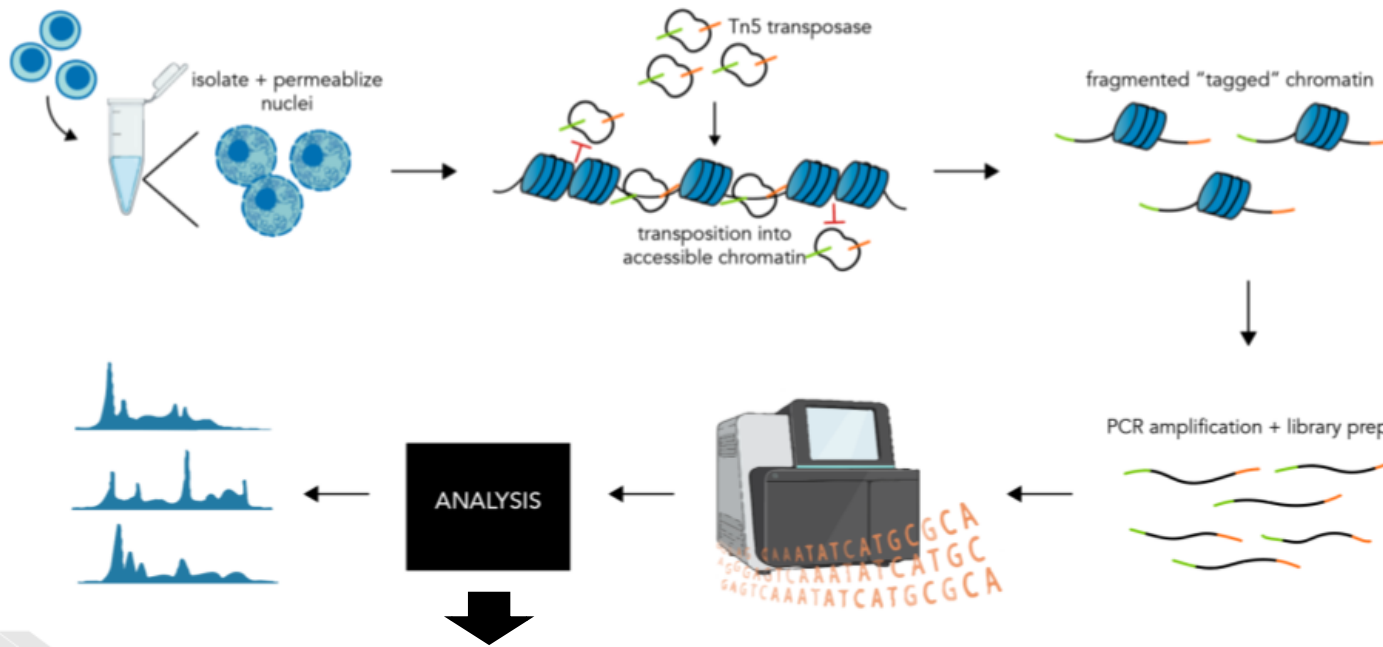
ATAC-seq

LiuLabUB/
HMMRATAC

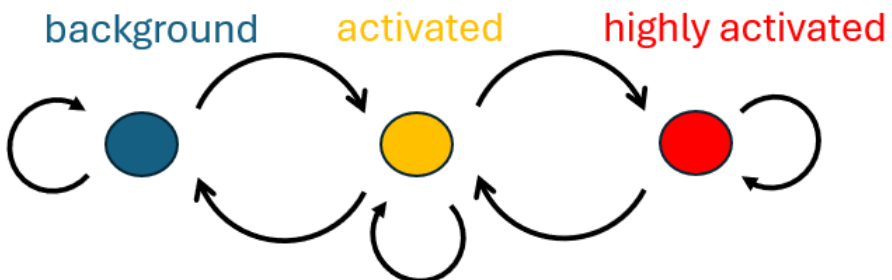
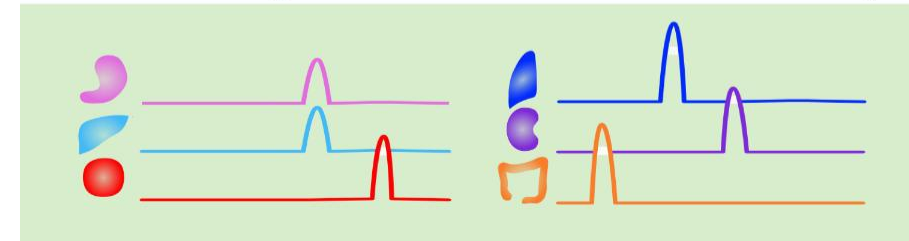
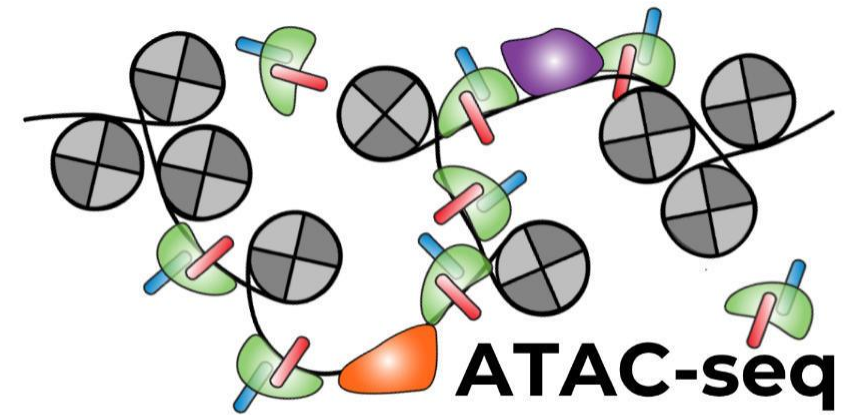


HMMRATAC peak caller for ATAC-seq data

Overview of ATAC-seq



2 Contributors 40 Issues 99 Stars 22 Forks

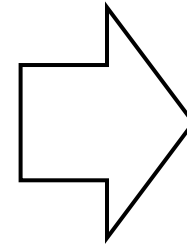


Esercitazione

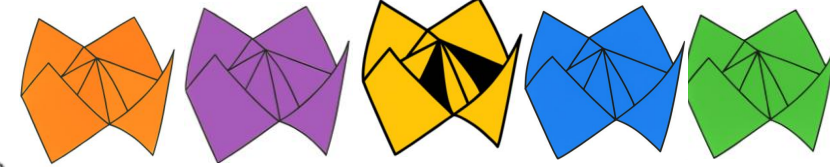
```
macs3 hmmratak -i atac_seq.bam
```

Segmentazione in domini epigenomici con CHROMHMM

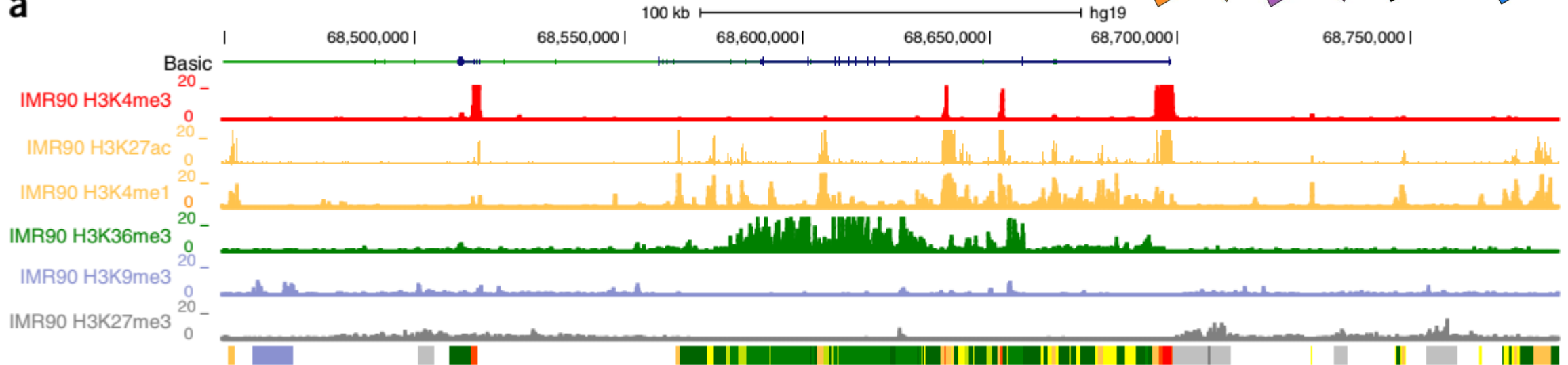
Possiamo avere moltissimi marcatori epigenomici differenti!!



Per ognuno possiamo costruire un HMM diverso.
Ma come dar senso a tutta questa informazione?

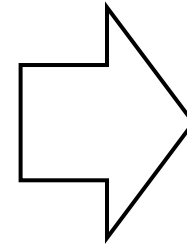


a

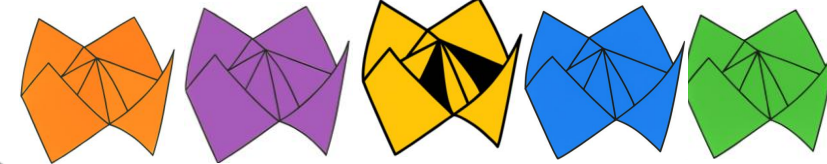


Segmentazione in domini epigenomici con CHROMHMM

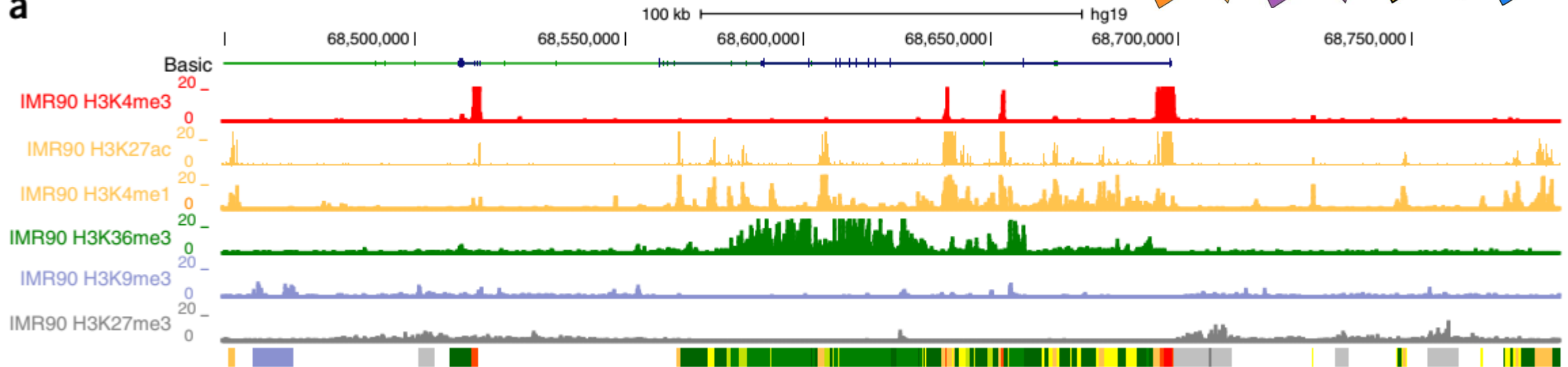
Possiamo avere moltissimi marcatori epigenomici differenti!!



Per ognuno possiamo costruire un HMM diverso.
Ma come dar senso a tutta questa informazione?



a



Segmentiamo il genoma in un **numero definito di stati** interpretabili che vadano a sintetizzare l'informazione di tutti i dati epigenomici e non. Esempio: cromatina aperta, chiusa, enhancers, geni attivi, etc.

Chromatin-state discovery and genome annotation
with ChromHMM

Jason Ernst¹⁻⁵ & Manolis Kellis^{6,7}

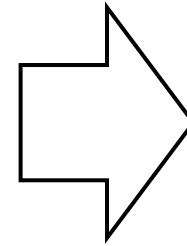
PROTOCOL

Nature Protocol 2017

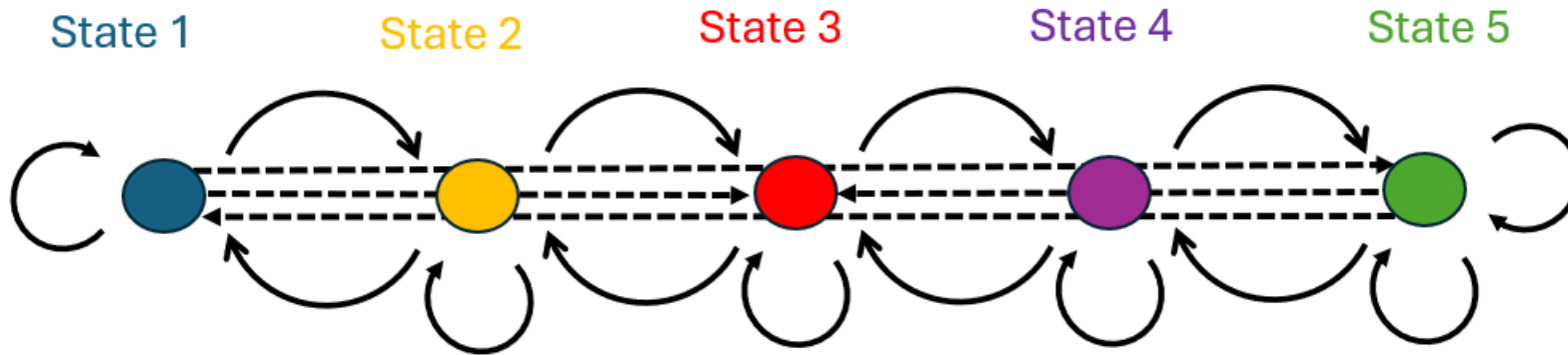
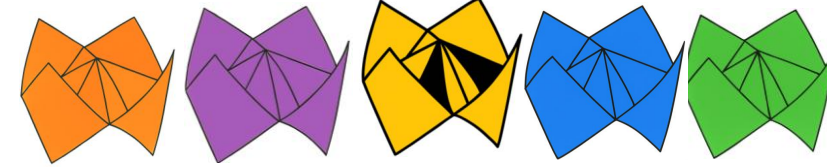


Segmentazione in domini epigenomici con CHROMHMM

Possiamo avere moltissimi marcatori epigenomici differenti!!



Per ognuno possiamo costruire un HMM diverso.
Ma come dar senso a tutta questa informazione?



Segmentiamo il genoma in un **numero definito di stati** interpretabili che vadano a sintetizzare l'informazione di tutti i dati epigenomici e non. Esempio: cromatina aperta, chiusa, enhancers, geni attivi, etc.

Chromatin-state discovery and genome annotation with ChromHMM

Jason Ernst¹⁻⁵ & Manolis Kellis^{6,7}

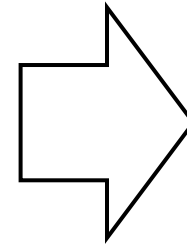
PROTOCOL

Nature Protocol 2017

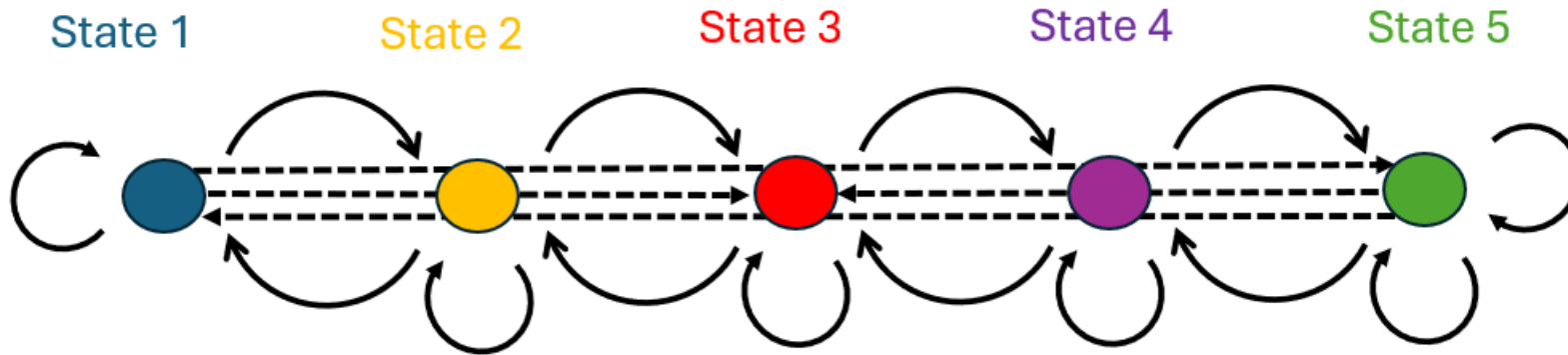
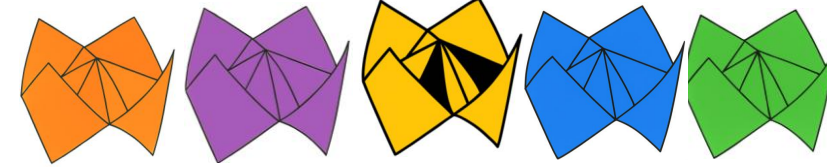


Segmentazione in domini epigenomici con CHROMHMM

Possiamo avere moltissimi marcatori epigenomici differenti!!



Per ognuno possiamo costruire un HMM diverso.
Ma come dar senso a tutta questa informazione?



Segmentiamo il genoma in un **numero definito di stati** interpretabili che vadano a sintetizzare l'informazione di tutti i dati epigenomici e non. Esempio: cromatina aperta, chiusa, enhancers, geni attivi, etc.

Chromatin-state discovery and genome annotation with ChromHMM

Jason Ernst¹⁻⁵ & Manolis Kellis^{6,7}

PROTOCOL

Nature Protocol 2017

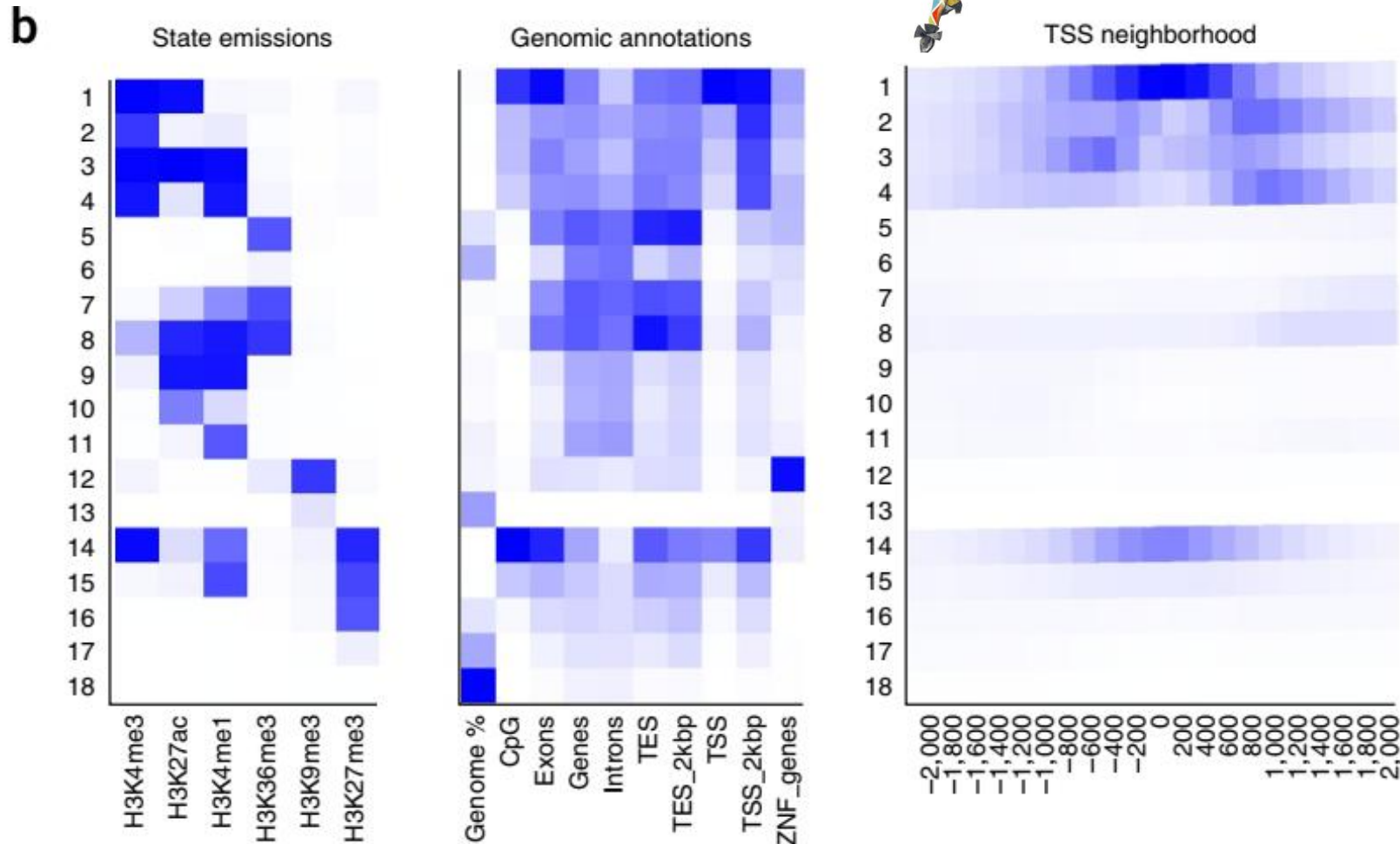
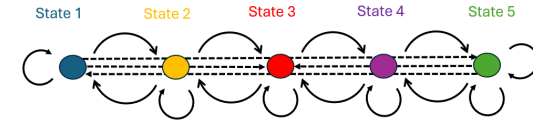


Segmentazione in domini epigenomici con CHROMHMM

Le emissioni permettono di interpretare quali segnali epigenomici sono correlati e rappresentano uno stesso stato della cromatina



Stati



State description	Abbrev.
Active TSS	TssA
Flanking TSS	TssFlnk
Flanking TSS upstream	TssFlnkU
Flanking TSS downstream	TssFlnkD
Strong transcription	Tx
Weak transcription	TxWk
Genic enhancer 1	EnhG1
Genic enhancer 2	EnhG2
Active enhancer 1	EnhA1
Active enhancer 2	EnhA2
Weak enhancer	EnhWk
ZNF genes & repeats	ZNF/Rpts
Heterochromatin	Het
Bivalent/poised TSS	TssBiv
Bivalent enhancer	EnhBiv
Repressed Polycomb	ReprPC
Weak repressed Polycomb	ReprPCWk
Quiescent/low	Quies

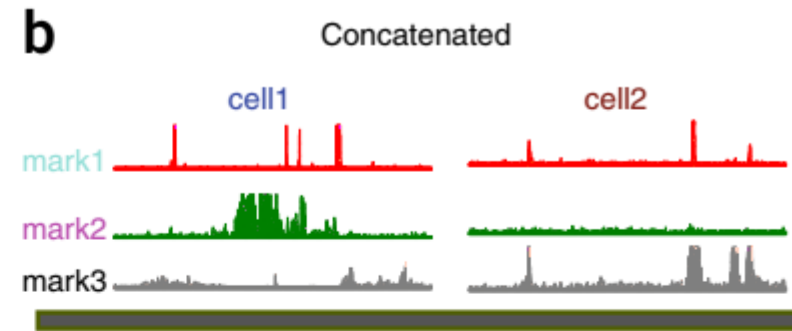
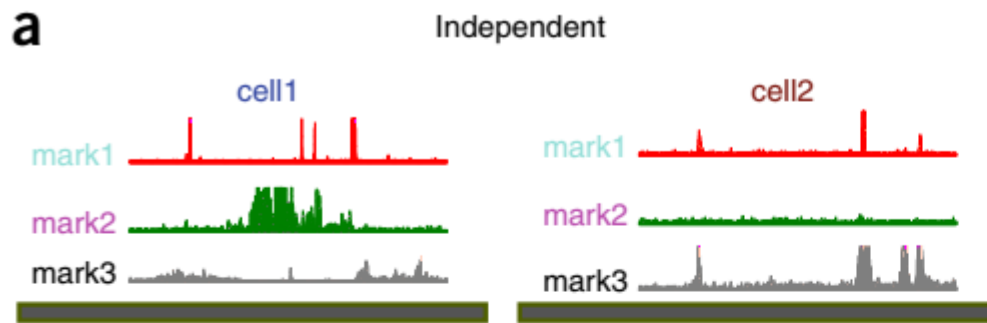
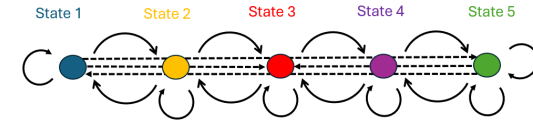


Segmentazione in domini epigenomici con CHROMHMM

CHROMHMM permette di esaminare differenze nella presenza di «stati epigenomici» tra linee cellulari o condizioni differenti



Stati



Annotation schema:



Models:

cell1 model

	mark1	mark2	mark3
State 1	0.8	0.2	0
State 2	0	0.6	0.1
State 3	0	0	0

cell2 model

	mark1	mark2	mark3
State 1	0.7	0.1	0.1
State 2	0.1	0.2	0.9
State 3	0	0.1	0

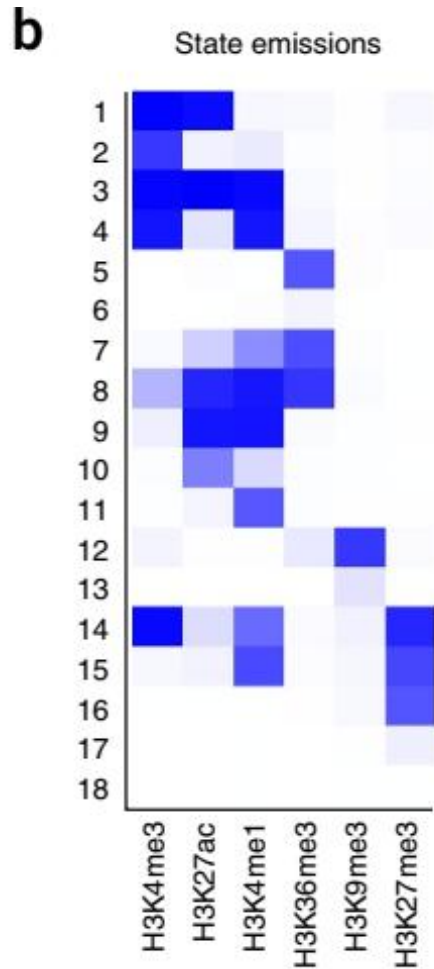
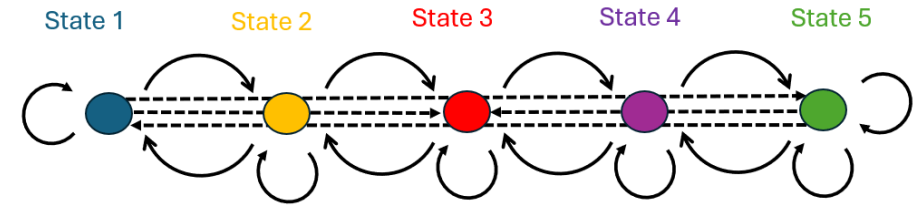
Common model

	mark1	mark2	mark3
State 1	0.9	0.1	0.1
State 2	0.1	0.5	0
State 3	0.1	0.6	0.7



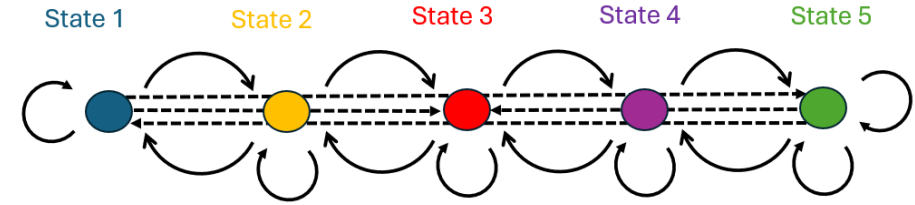
Segmentazione in domini epigenomici con CHROMHMM

Secondo voi che tipo di probabilità di emissione utilizza CHROMHMM?



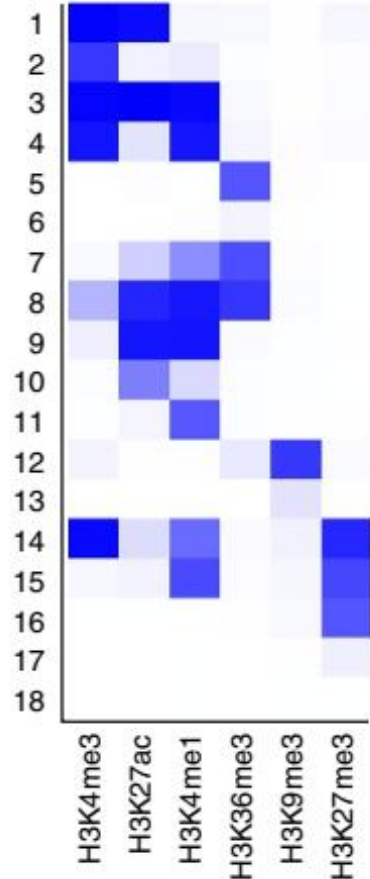
Segmentazione in domini epigenomici con CHROMHMM

Secondo voi che tipo di probabilità di emissione utilizza CHROMHMM?



b

State emissions

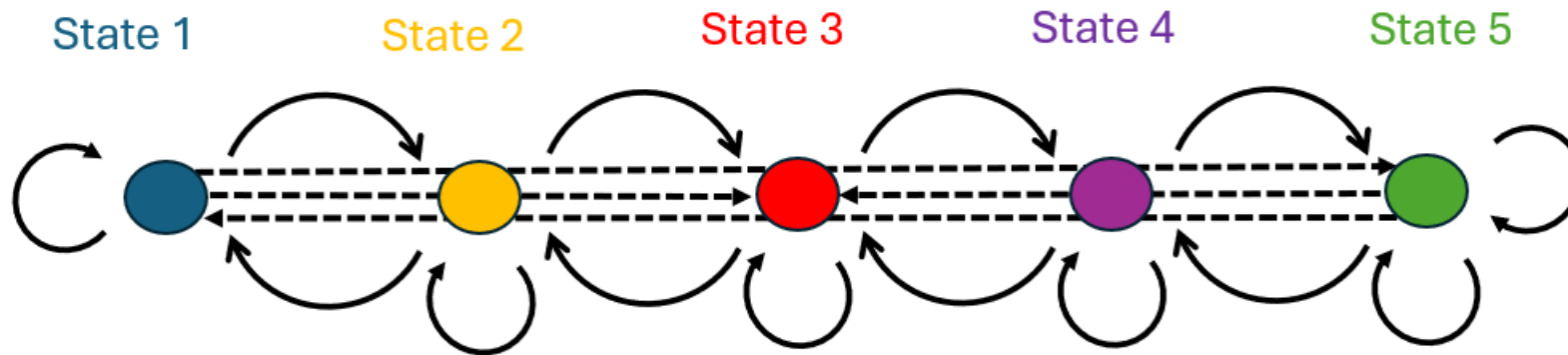


come quelle usate da noi negli esempi: CG osservata o no!

Utilizza delle emissioni binarie (marcatore presente/assenza) per semplificare i dati (non sempre comparabili) ottenuti da marcatori di diverso tipo ed esperimenti differenti.



Segmentazione in domini epigenomici con CHROMHMM

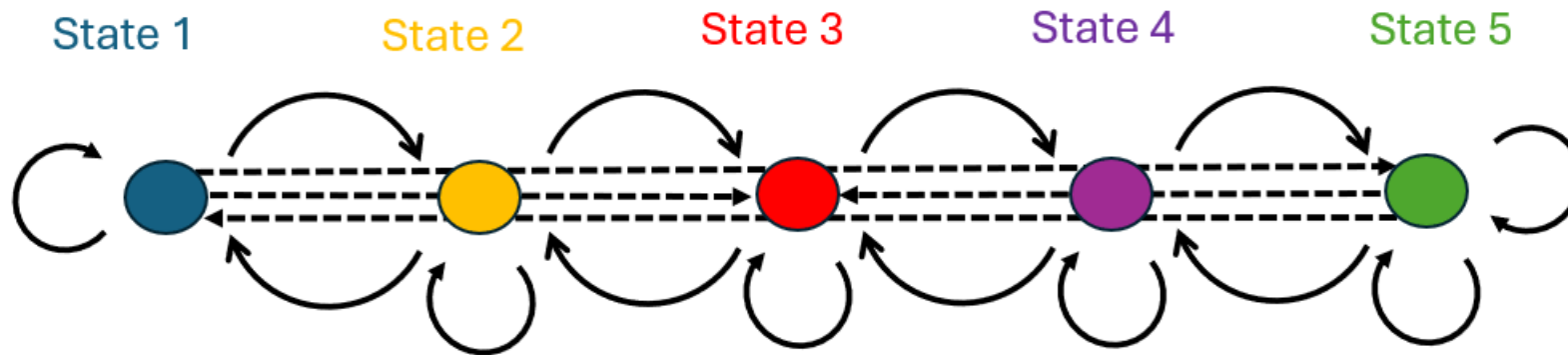


Esercitazione

```
java -jar ChromHMM.jar BinarizeBed
```



Segmentazione in domini epigenomici con CHROMHMM

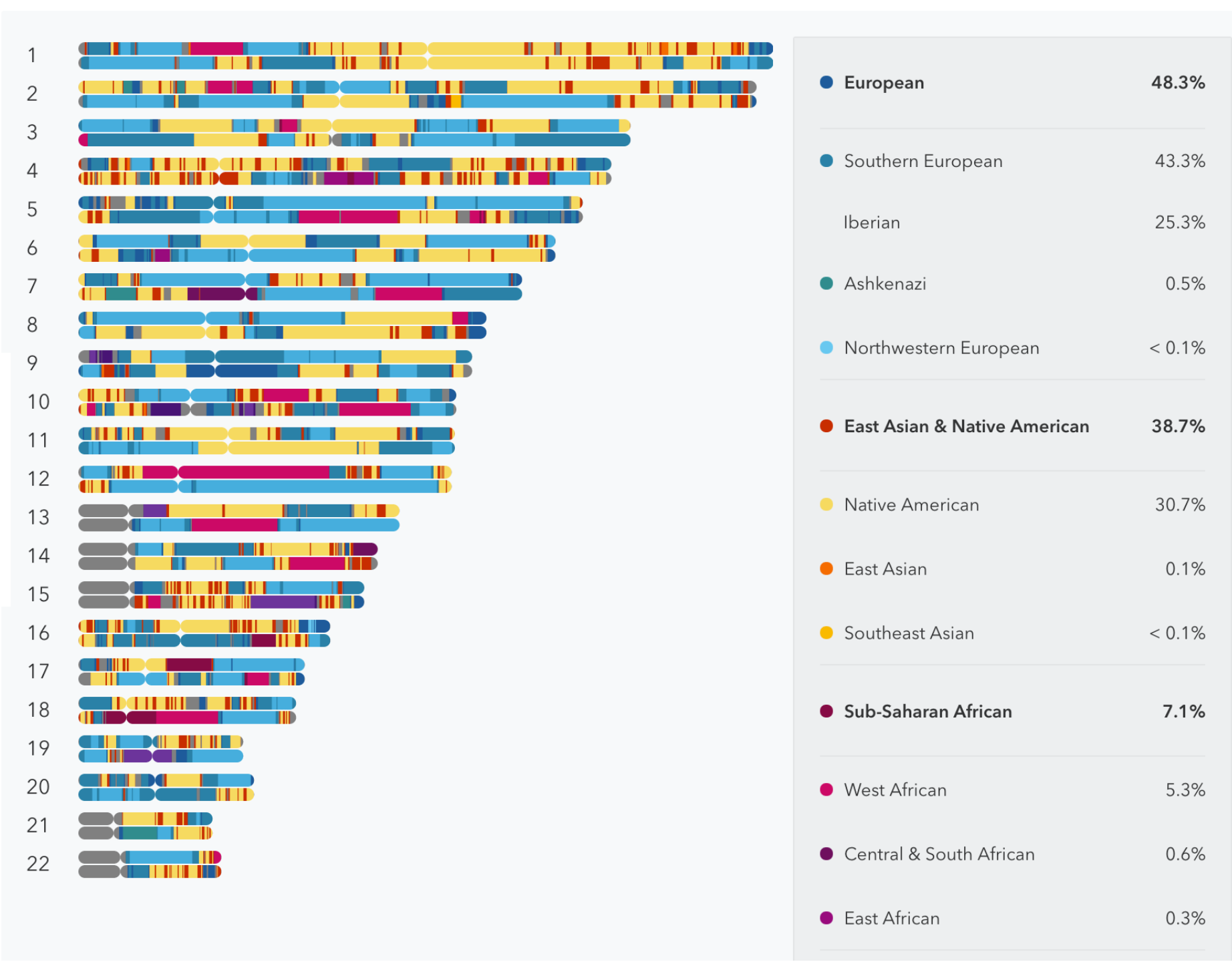
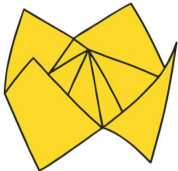


Esercitazione

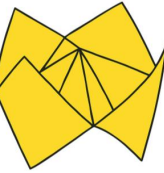
```
java -jar ChromHMM.jar BinarizeBed
```



Chromosome-painting



Imputazione di genotipi mancanti



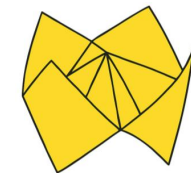
Non solo l'informazione di altri individui può essere utilizzata per migliorare l'accuratezza dei genotipi, ma anche quella di siti vicini.

Varianti vicine tra loro lungo il genoma sono spesso correlate a causa della ricombinazione

Questa informazione può essere sfruttata con l'utilizzo di Hidden Markov Models (che vedremo nelle lezioni successive)

#CHROM	POS	REF	ALT	QUAL	FILTER	INFO						
1	122	A	T	40	PASS	...	0/1	0/1	0/0	0/0	0/1	0/1
1	1422	A	C	50	PASS	...	./.	0/1	0/1	0/0	0/1	0/1
1	2456	C	C	45	PASS	...	./.	0/1	0/0	0/0	0/1	1/1
1	3563	T	G	56	PASS	...	0/1	0/1	0/0	0/0	0/1	1/1

Imputazione di genotipi mancanti



Sembra essere presente un aplotipo (una serie di alleli correlati) di **alleli Alternativi** in questa regione del genoma. Tutti gli individui con l'allele alternativo in posizione 122 e 3563 sembrano avere allele alternativo anche in posizione 1422 e 2456.

#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	 Bart	 Omer	 Marge	 Maggie	 Lisa	 Mr.Burns
1	122	A	T	40	PASS	...	0/ 1	0/ 1	0/0	0/0	0/ 1	0/1
1	1422	A	C	50	PASS	...	./.	0/ 1	0/1	0/0	0/ 1	0/1
1	2456	C	C	45	PASS	...	./.	0/ 1	0/0	0/0	0/ 1	1/1
1	3563	T	G	56	PASS	...	0/ 1	0/ 1	0/0	0/0	0/ 1	1/1

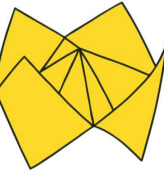
Imputazione di genotipi mancanti e phasing

Notate che questo problema è strettamente connesso al **phasing**, cioè *l'assegnazione di un allele al cromosoma di origine materna o paterna*.

In questo caso tutti gli alleli
alternativi di Bart fanno parte
di uno stesso cromosoma da
Omer

Genotipi imputati												
#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	Bart	Omer	Marge	Maggie	Lisa	Mr.Burns
1	122	A	T	40	PASS	...	0/1	1/1	0/0	0/0	0/1	0/1
1	1422	A	C	50	PASS	...	0/1	0/1	0/1	0/0	0/1	0/1
1	2456	C	C	45	PASS	...	0/1	0/1	0/0	0/0	0/1	1/1
1	3563	T	G	56	PASS	...	0/1	0/1	0/0	0/0	0/1	1/1

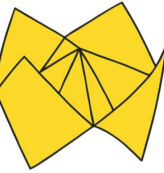
Imputazione di genotipi mancanti e phasing



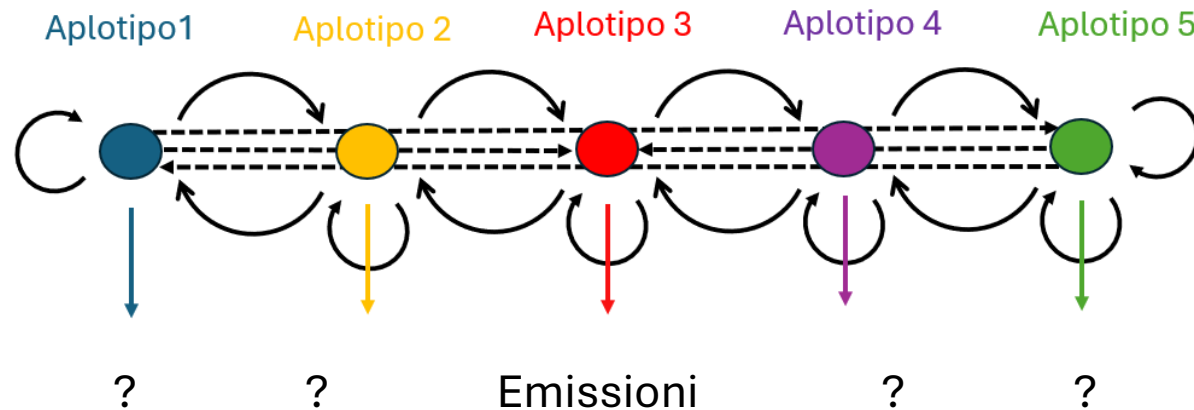
Modello HMM di Li e Stephens per phasing e imputazione (Beagle, etc.)

#CHROM	POS	REF	ALT	QUAL	FILTER	INFO	 Bart	 Omer	 Marge	 Maggie	 Lisa	 Mr.Burns
1	122	A	T	40	PASS	...	0/1	1/1	0/0	0/0	0/1	0/1
1	1422	A	C	50	PASS	...	0/1	0/1	0/1	0/0	0/1	0/1
1	2456	C	C	45	PASS	...	0/1	0/1	0/0	0/0	0/1	1/1
1	3563	T	G	56	PASS	...	0/1	0/1	0/0	0/0	0/1	1/1

Imputazione di genotipi mancanti e phasing



Modello HMM di Li e Stephens per phasing e imputazione (Beagle, etc.)



#CHROM

POS

REF

ALT

QUAL

FILTER

INFO

Bart

Omer

Marge

Maggie

Lisa

Mr.Burns

1

122

A

T

40

PASS

...

0/1

0/1

0/0

0/0

0/1

0/1

1

1422

A

C

50

PASS

...

0/1

0/1

0/1

0/1

0/1

0/1

1

2456

C

C

45

PASS

...

0/1

0/1

0/0

0/0

0/1

1/1

1

3563

T

G

56

PASS

...

0/1

0/1

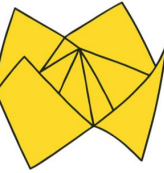
0/0

0/0

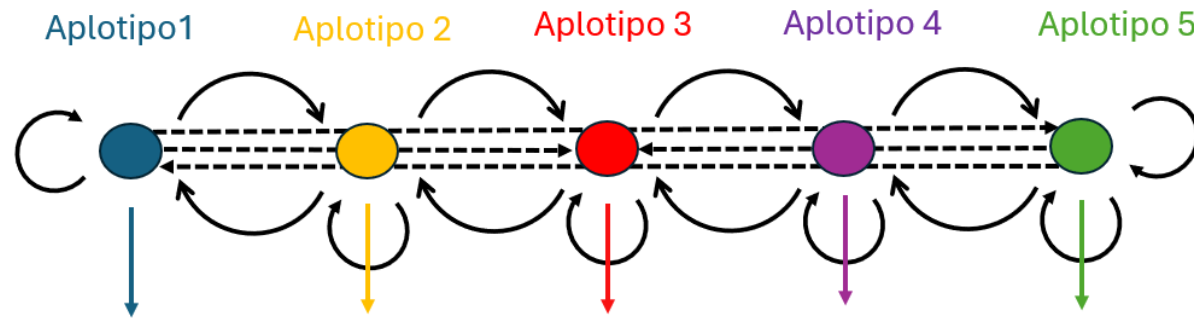
0/1

1/1

Imputazione di genotipi mancanti e phasing



Modello HMM di Li e Stephens per phasing e imputazione (Beagle, etc.)



Alleli 1 e 0 associati a ciascun aplotipo

#CHROM

POS

REF

ALT

QUAL

FILTER

INFO



Bart



Omer



Marge



Maggie



Lisa



Mr.Burns

1

122

A

T

40

PASS

...

0/1

0/1

0/0

0/0

0/1

0/1

1

1422

A

C

50

PASS

...

0/1

0/1

0/1

0/1

0/1

0/1

1

2456

C

C

45

PASS

...

0/1

0/1

0/0

0/0

0/1

1/1

1

3563

T

G

56

PASS

...

0/1

0/1

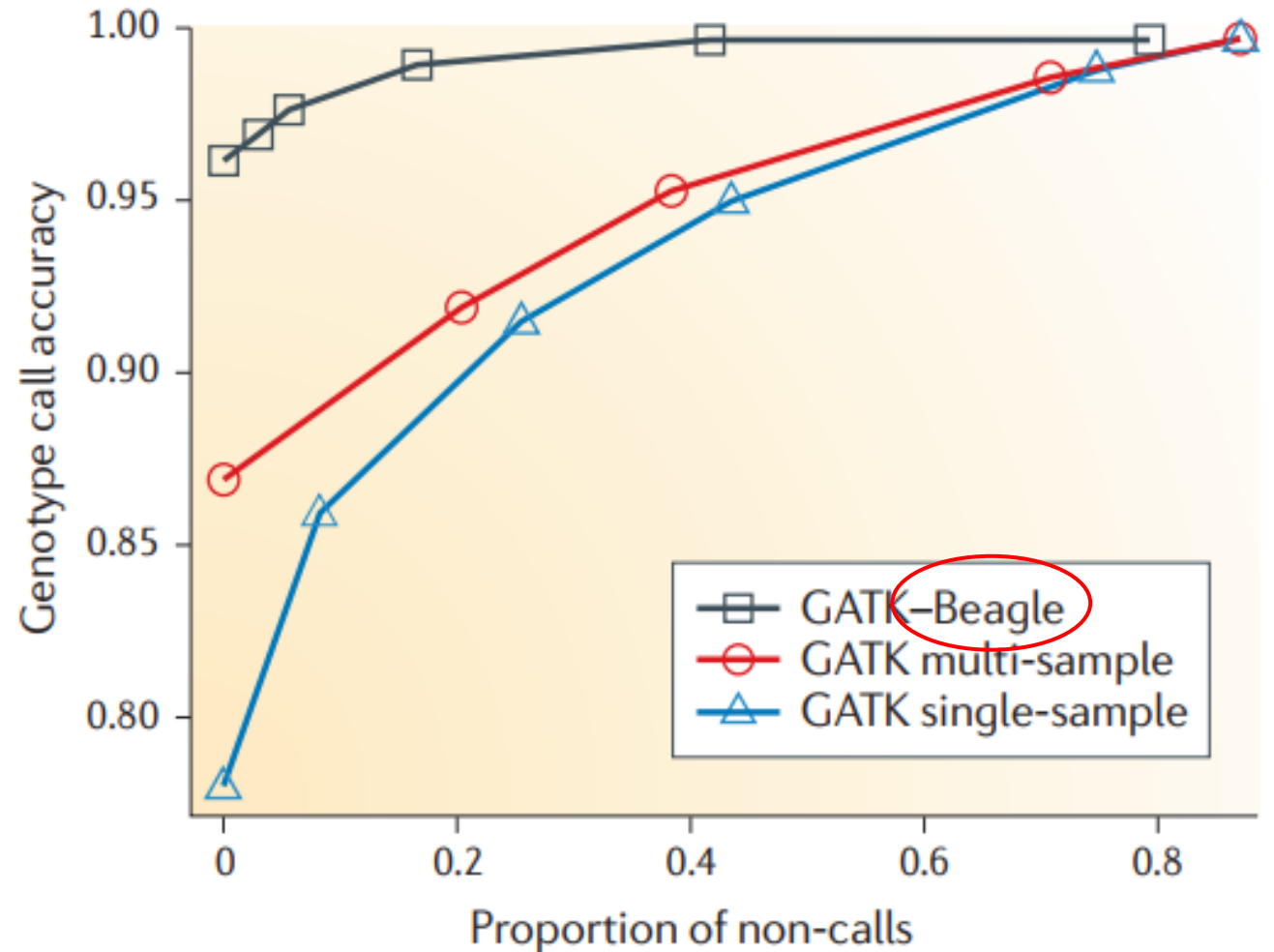
0/0

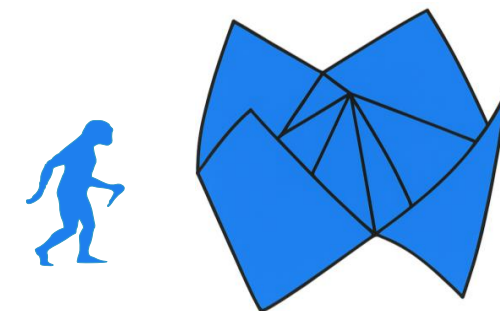
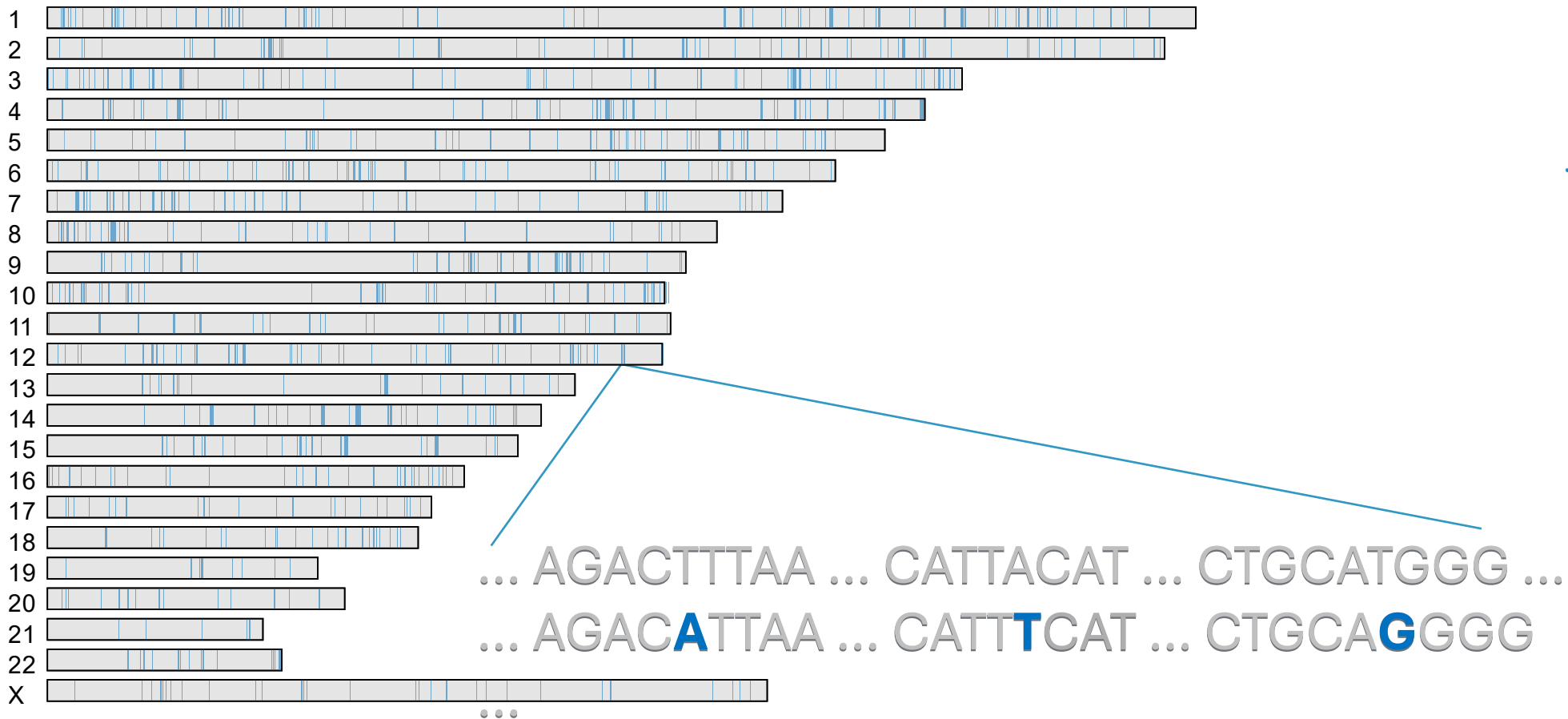
0/0

0/1

1/1

Genotipizzazione multi-sample e imputazione (qui con Beagle) migliorano l'accuratezza dei genotipi chiamati



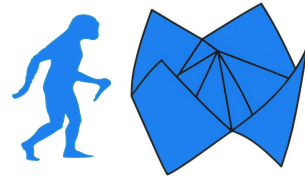


... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG
 ...
 ... AGAC**A**TTAA ... CATT**T**CAT ... CTGCA**G**GGG

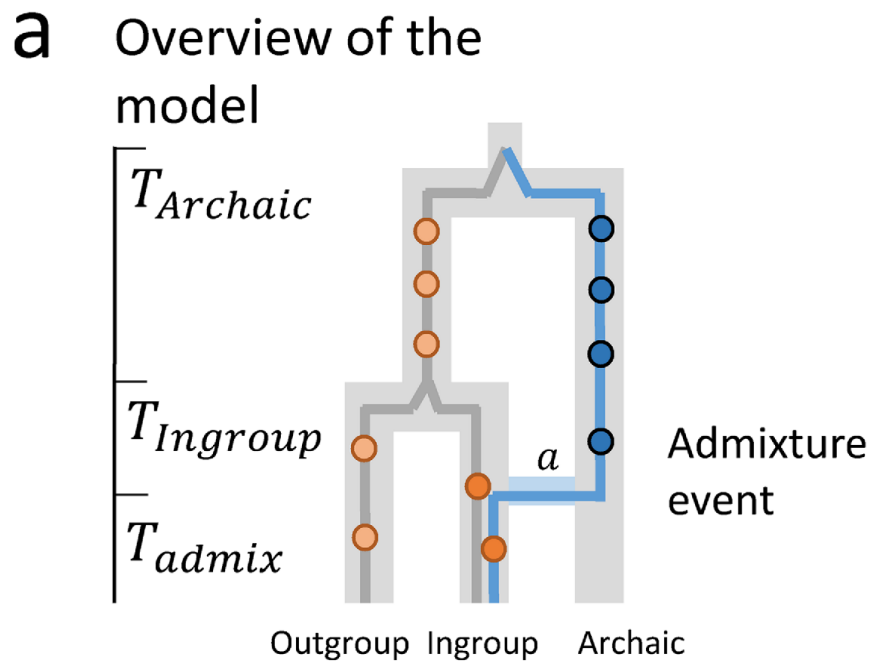
Aplotipi Neanderthal in 27,566 genomi di Islandesi



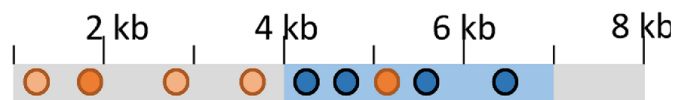
Circa 50% del genoma di Neandertal
può essere ricostruito usando 27,566
genomi di cittadini Islandesi



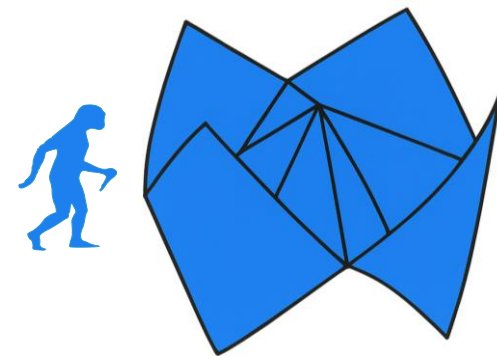
Adattato da from Skov, Call-Macia, Sveinbjörnsson , **Mafessoni** et al., 2020,
Nature



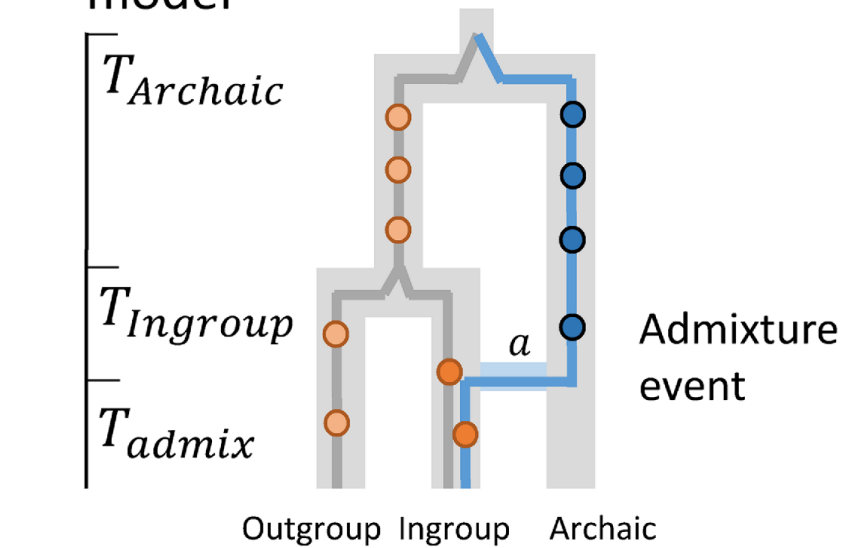
Observed variants



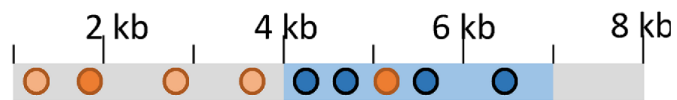
Remove variants found in outgroup (example: Africans)



a Overview of the model



Observed variants



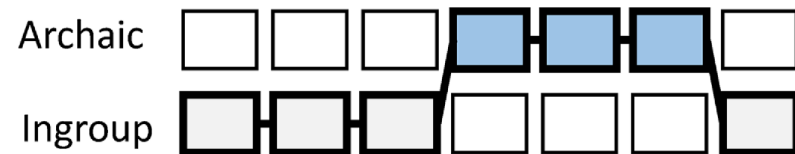
Remove variants found in outgroup (example: Africans)



Count variants in window

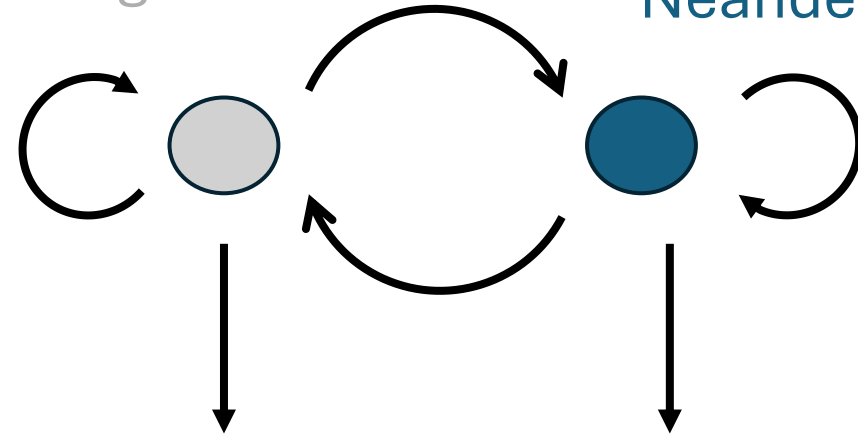


Decode sequence

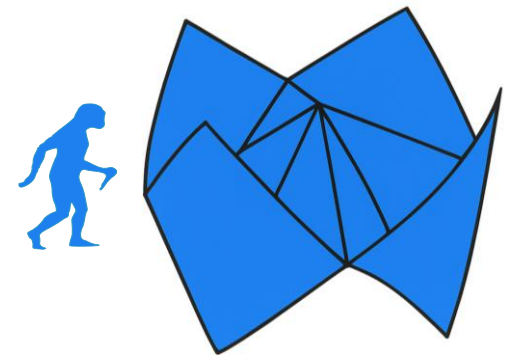


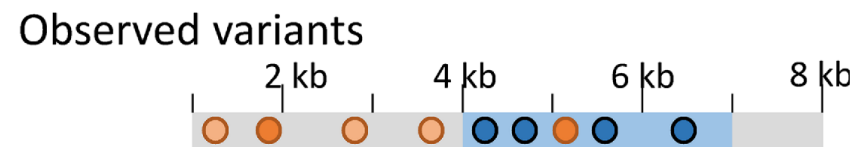
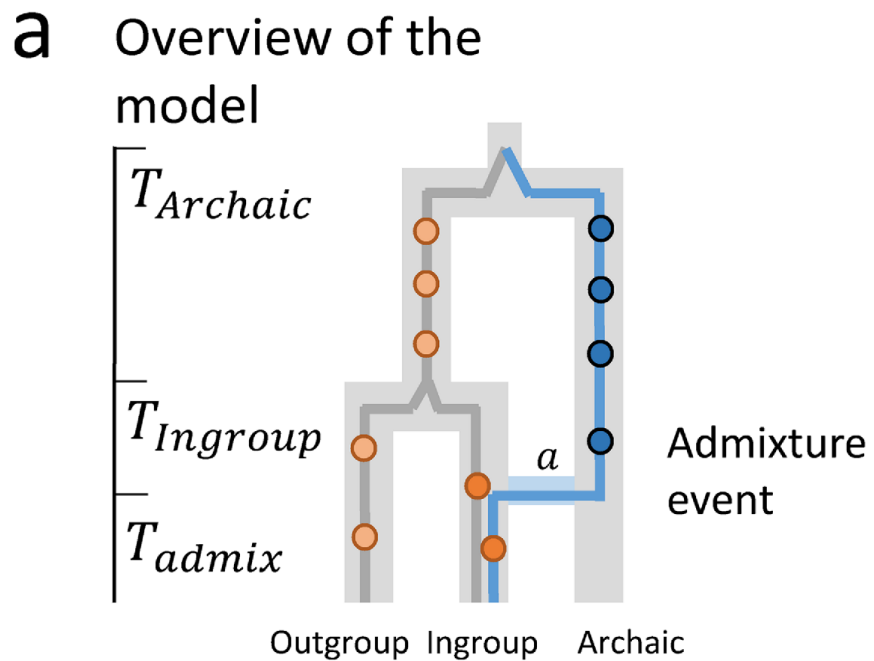
Non introgresse

Introgresse dai Neandertal



Numero di «mutazioni» in una «finestra del genoma»





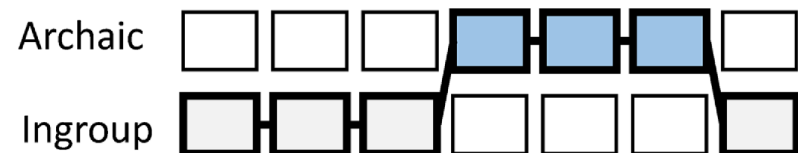
Remove variants found in outgroup (example: Africans)



Count variants in window

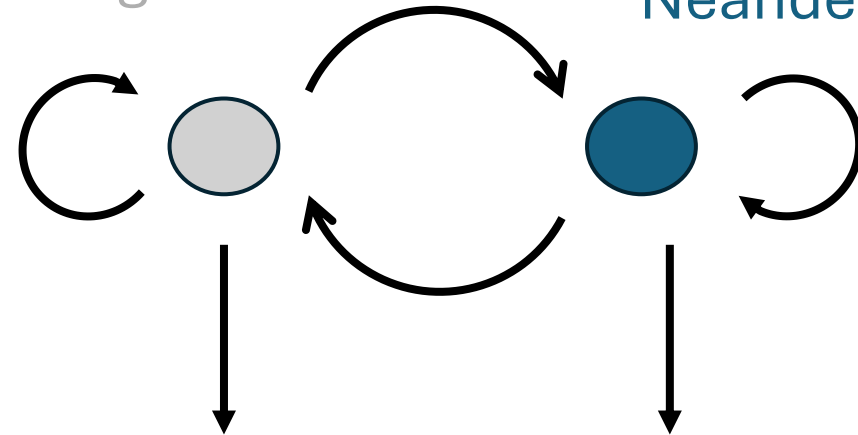


Decode sequence



Non
introgresse

Introgresse
dai
Neandertal

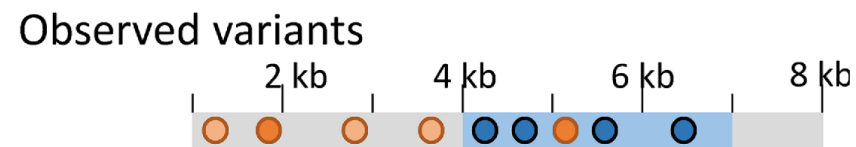
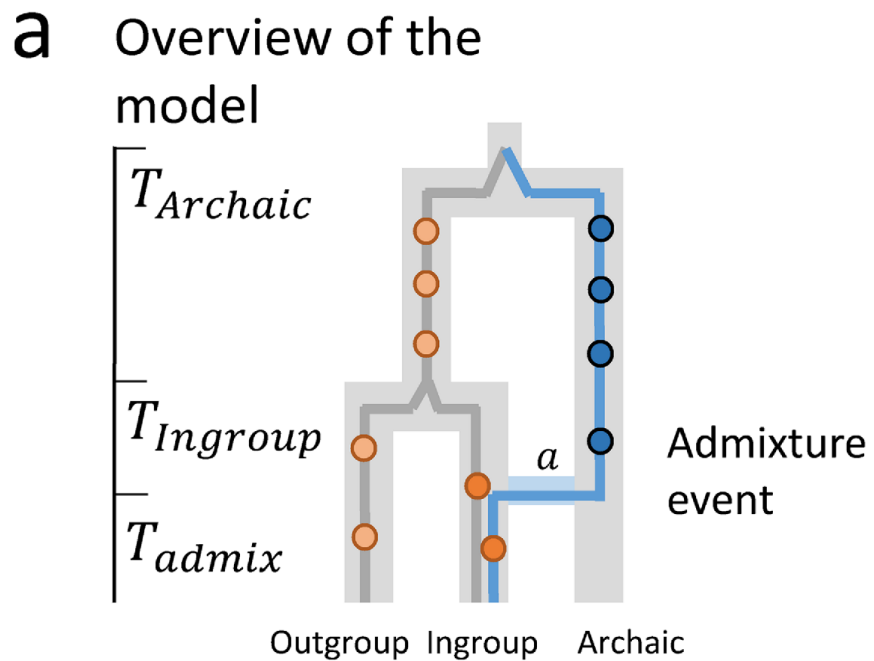


Numero di «mutazioni» in una «finestra del genoma»

Probabilità di
transizione ed
emissione sono spesso
funzioni di parametri di
nostro interesse
biologico



Stimare queste
probabilità ci da
informazioni sulla
biologia di un certo
fenomeno!



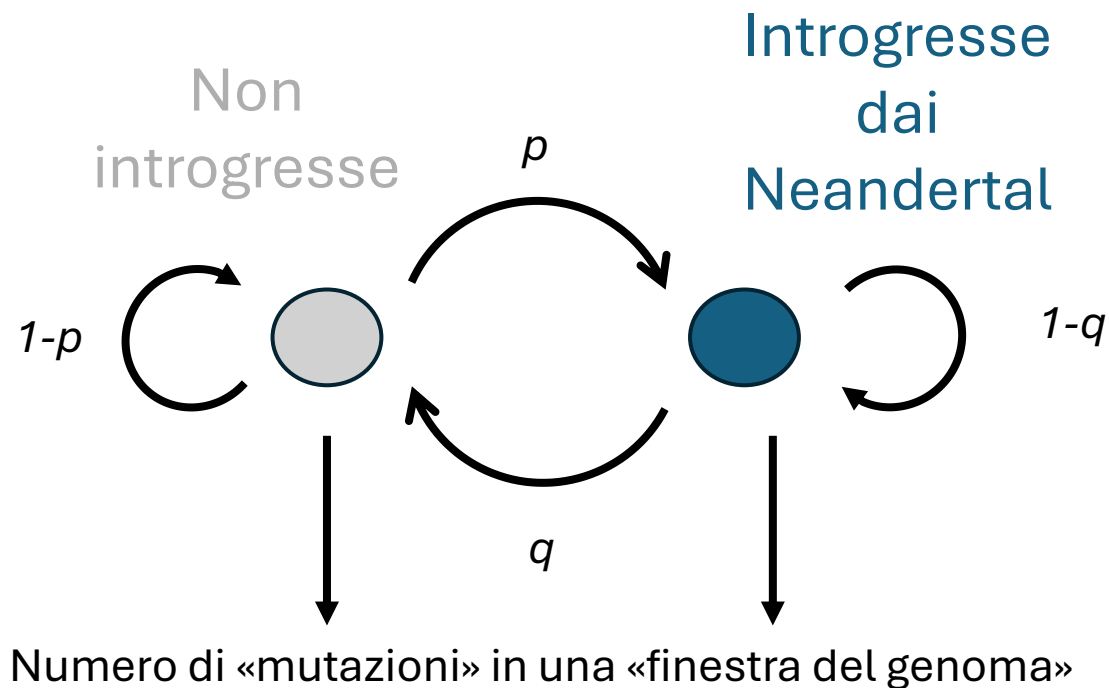
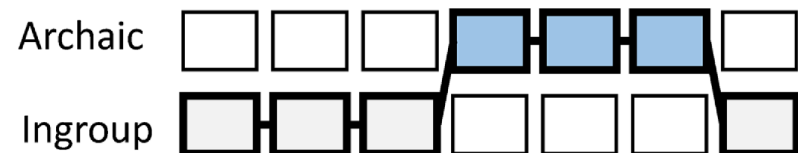
Remove variants found in outgroup (example: Africans)



Count variants in window



Decode sequence



$$q \simeq T_{admix} * \text{tasso di ricombinazione} * (1-a)$$

dove a è la percentuale di genoma introgresso

Stimare queste probabilità ci da informazioni sulla biologia di un certo fenomeno! **Esempio:** stimare il tempo passato da un evento di introgressione/flusso genico

Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)



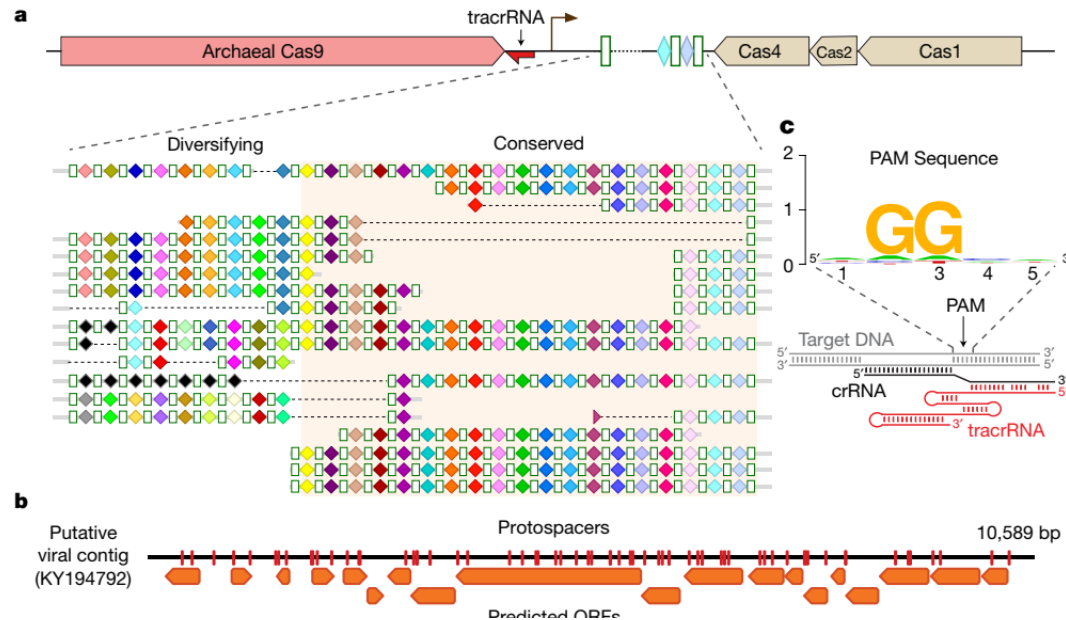
Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)

LETTER

doi:10.1038/nature21059

New CRISPR–Cas systems from uncultivated microbes

David Burstein^{1*}, Lucas B. Harrington^{2*}, Steven C. Strutt^{2*}, Alexander J. Probst¹, Karthik Anantharaman¹, Brian C. Thomas¹, Jennifer A. Doudna^{2,3,4,5,6} & Jillian F. Banfield^{1,7}



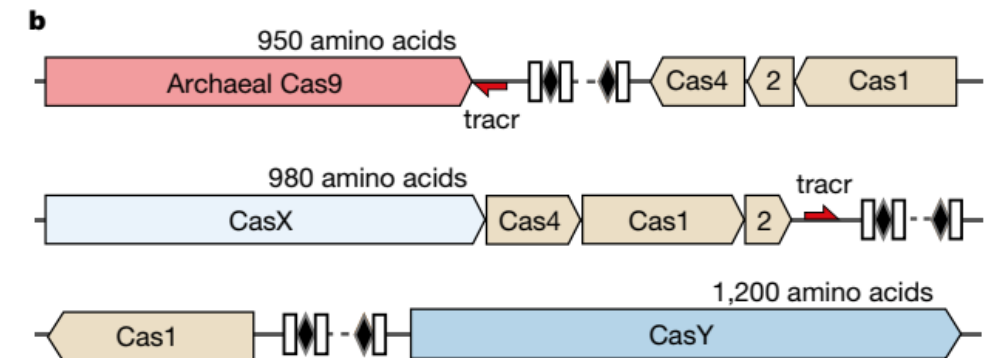
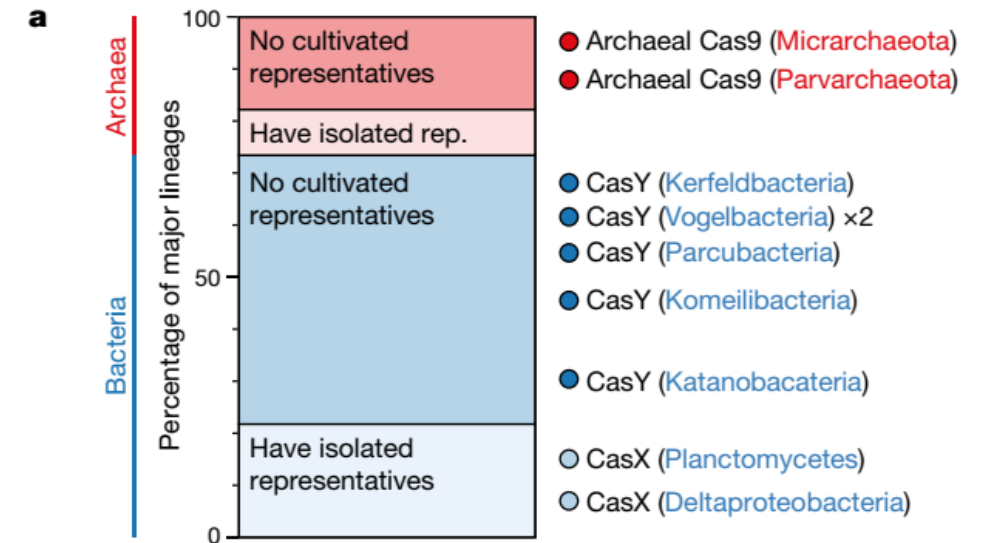
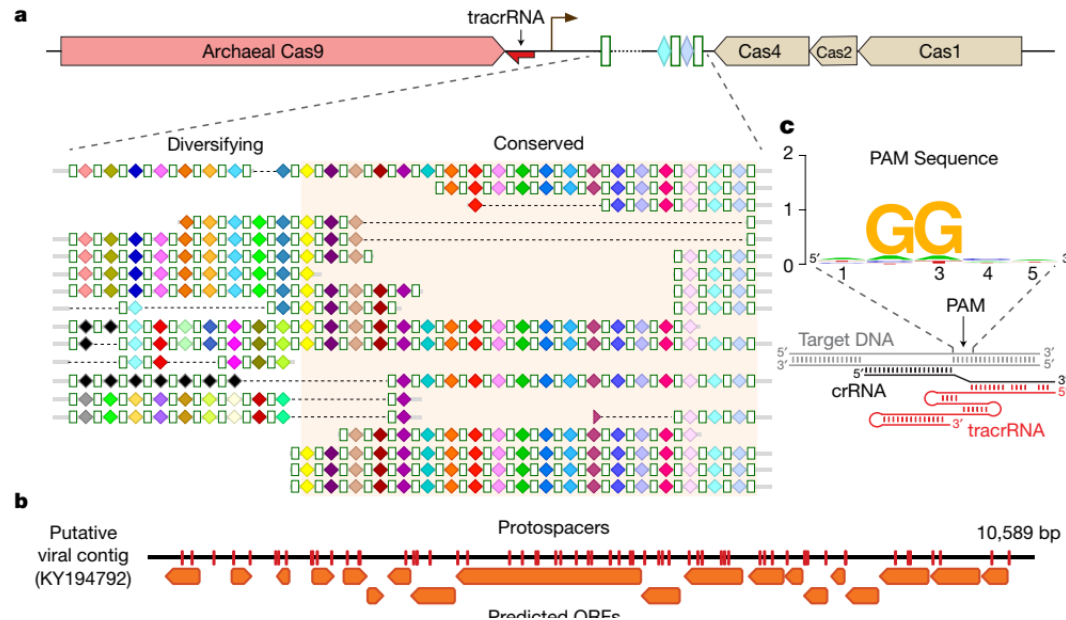
Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)

LETTER

doi:10.1038/nature21059

New CRISPR–Cas systems from uncultivated microbes

David Burstein^{1*}, Lucas B. Harrington^{2*}, Steven C. Strutt^{2*}, Alexander J. Probst¹, Karthik Anantharaman¹, Brian C. Thomas¹, Jennifer A. Doudna^{2,3,4,5,6} & Jillian F. Banfield^{1,7}



Nuove
tipologie di
Cas

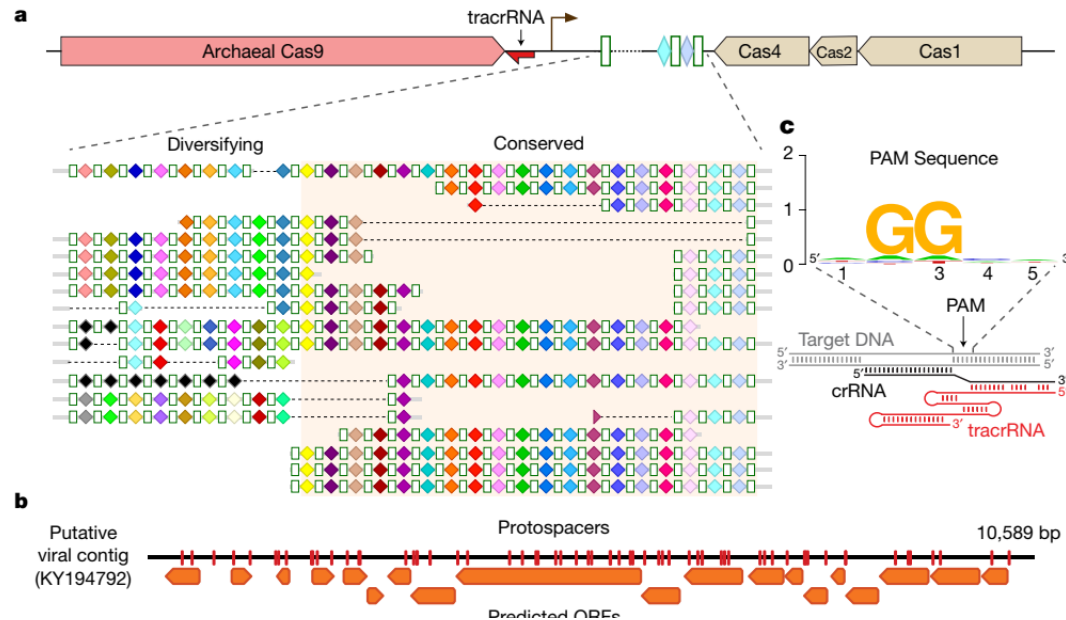
Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)

LETTER

doi:10.1038/nature21059

New CRISPR–Cas systems from uncultivated microbes

David Burstein^{1*}, Lucas B. Harrington^{2*}, Steven C. Strutt^{2*}, Alexander J. Probst¹, Karthik Anantharaman¹, Brian C. Thomas¹, Jennifer A. Doudna^{2,3,4,5,6} & Jillian F. Banfield^{1,7}



Problema:

- vorremmo trovare su nuovi genomi dei geni omologhi ad un gene di nostro interesse
- ma questi genomi sono molto distanti evolutivamente, quindi questi geni saranno molto diversi

• Cosa fareste?

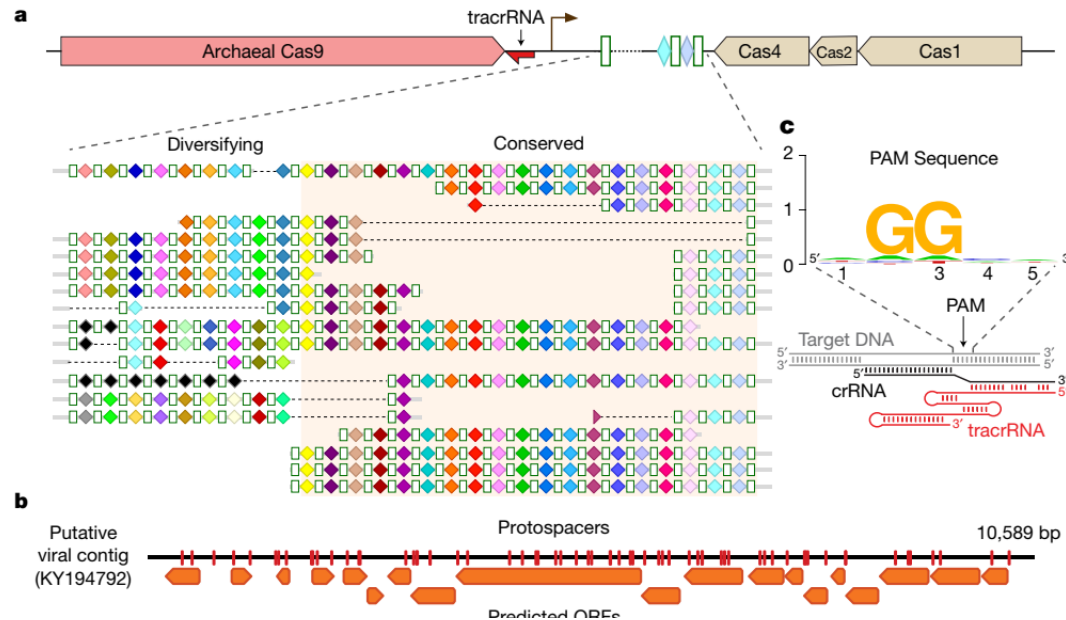
Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)

LETTER

doi:10.1038/nature21059

New CRISPR–Cas systems from uncultivated microbes

David Burstein^{1*}, Lucas B. Harrington^{2*}, Steven C. Strutt^{2*}, Alexander J. Probst¹, Karthik Anantharaman¹, Brian C. Thomas¹, Jennifer A. Doudna^{2,3,4,5,6} & Jillian F. Banfield^{1,7}



Problema:

- vorremmo trovare su nuovi genomi dei geni omologhi ad un gene di nostro interesse
- ma questi genomi sono molto distanti evolutivamente, quindi questi geni saranno molto diversi

- **Cosa fareste?**
- Potremmo mappare, ma approcci basati sulla Burrow-Wheeler sicuramente fallirebbero per le troppe differenze
- Minimizers potrebbero aiutare, ma comunque risulterebbe difficile

Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)



Problema:

- vorremmo trovare su nuovi genomi dei geni omologhi ad un gene di nostro interesse
- ma questi genomi sono molto distanti evolutivamente, quindi questi geni saranno molto diversi

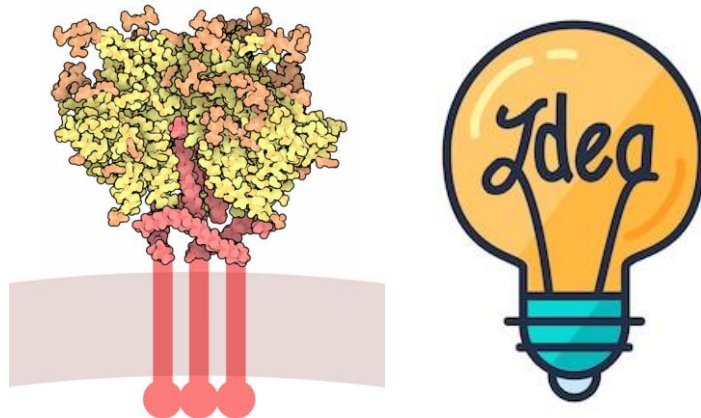
• **Idea:**

- Se già conosco la sequenza del gene in varie specie, **ho informazione di ciò può essere simile e differente.**
- Proviamo ad includere le probabilità di avere queste similarità e differenze in un Hidden Markov Model, che nasce per modellizzare sequenze!

Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)

```
VKKLGEQFR-NKTIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFN-----NSTES-----DTITL
VKKLGEQFR-NKTIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFN-----NSTDNG-----DTITL
VKKLGEQFR-NKTIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFD-----NSTESNN-----DTITL
VDKLREQFGKNKTIIFNQPSGGDLEIVMHTFNCGGEFFYCNTTQLFNSTWNS---TGNGTESYNGQENGTTTL
VDKLREQFGKNKTIIFNQPSGGDLEIVMHTFNCGGEFFYCNTTQLFNSTWNG---TNTT--GLDG--NDTITL
VDKLREQFGKNKTIIFNQSSGGDLEIVTHTFNCGGEFFYCNTTQLFNSNWTG---NSTE--GLHG--DDTITL
VKKLGEQFG-NKTIIFNQSSGGGLEIVMHSFNCGGEFFYCNTTQLFNN--TR-----NSTESNNGQGNDTTTL
VKKLREQFGKNKTIIFKQSSGGDLEIVTHTFNCAGEFFYCNTTQLFNSNWTE-----NSITGLDG--NDTITL
VGKLREQFGK-KTIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFNSTWDNSTWNSTGKDKENG-N-DTITL
```

FIGURE 10.1 A multiple alignment of a short region of gp120 proteins sampled from a single HIV-positive patient at nine different time points. Almost half of the columns (shown in darker text) are not conserved across all time points, illustrating how quickly HIV evolves, even within an individual host. Amino acids differing from the most common symbol in a column are shown in blue.



Problema:

- vorremmo trovare su nuovi genomi dei geni omologhi ad un gene di nostro interesse
- ma questi genomi sono molto distanti evolutivamente, quindi questi geni saranno molto diversi

Idea:

- Se già conosco la sequenza del gene in varie specie, **ho informazione di ciò può essere simile e differente.**
- Proviamo ad includere le probabilità di avere queste similarità e differenze in un Hidden Markov Model, che nasce per modellizzare sequenze!

Profile HMM: identificazione di geni con omologie remote (allineamento/mappatura/annotazione)

```

VKKLGEQFR-NKTIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFN-----NSTES-----DTITL
VKKLGEQFR-NKTIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFN-----NSTDNG-----DTITL
VKKLGEQFR-NKTIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFD-----NSTESNN-----DTITL
VDKLREQFGKNKTIIFNQPSGGDLEIVMHTFNCGGEFFYCNTTQLFNSTWNS---TGNGTESYNGQENGTTTL
VDKLREQFGKNKTIIFNQPSGGDLEIVMHTFNCGGEFFYCNTTQLFNSTWNG---TNTT--GLDG--NDTITL
VDKLREQFGKNKTIIFNQSSGGDLEIVTHTFNCGGEFFYCNTTQLFNSNWTG---NSTE--GLHG--DDTITL
VKKLGEQFG-NKTIIFNQSSGGGLEIVMHSFNCGGEFFYCNTTQLFNN--TR-----NSTESNNGQGNDDTTL
VKKLREQFGKNKTIIFKQSSGGDLEIVTHTFNCGGEFFYCNTTQLFNSNWTG-----NSITGLDG--NDTITL
VGKLREQFGK-KTIIIFNQPSGGDLEIVMHSFNCGGEFFYCNTTQLFNSTWDNSTWNSTGKDKENGNDTITL
    
```

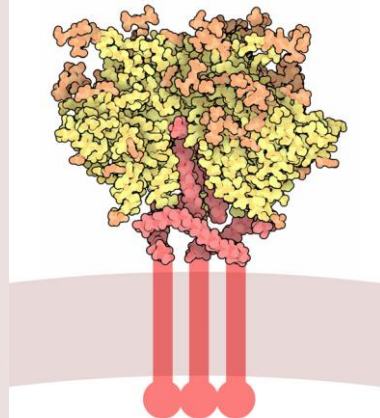
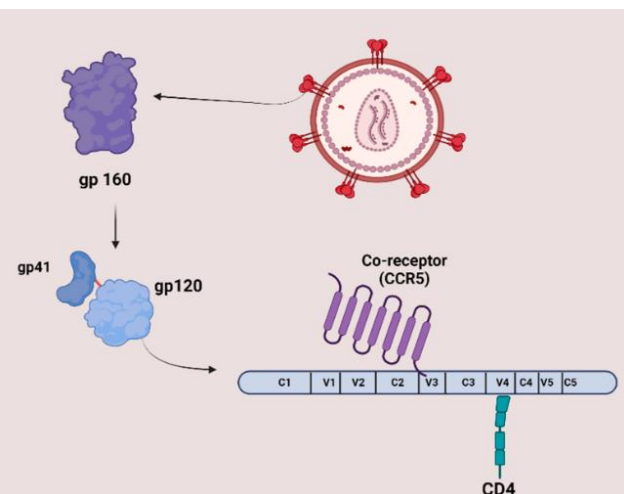
FIGURE 10.1 A multiple alignment of a short region of gp120 proteins sampled from a single HIV-positive patient at nine different time points. Almost half of the columns (shown in darker text) are not conserved across all time points, illustrating how quickly HIV evolves, even within an individual host. Amino acids differing from the most common symbol in a column are shown in blue.



Problema:

- vorremmo trovare su nuovi genomi dei geni omologhi ad un gene di nostro interesse
- ma questi genomi sono molto distanti evolutivamente, quindi questi geni saranno molto diversi

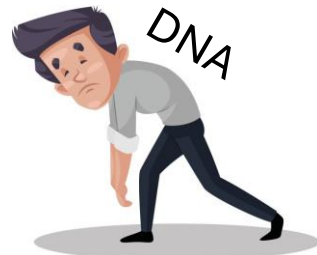
- **Idea:**
- Se già conosco la sequenza del gene in varie specie, **ho informazione di ciò può essere simile e differente.**
- Proviamo ad includere le probabilità di avere queste similarità e differenze in un Hidden Markov Model, che nasce per modellizzare sequenze!



		posizioni								
		1	2	3	4	5	6	7	8	
Alignment		A	C	D	E	F	A C	A	D	F
		A	F	D	A	—	— —	C	C	F
		A	—	—	E	F	D —	F	D	C
		A	C	A	E	F	— —	A	—	C
		A	D	D	E	F	A A	A	D	F

	posizioni								
	1	2	3	4	5	6	7	8	
Alignment	A	C	D	E	F	A C	A	D	F
	A	F	D	A	—	— —	C	C	F
	A	—	—	E	F	D —	F	D	C
	A	C	A	E	F	— —	A	—	C
	A	D	D	E	F	A A	A	D	F

Notate come per sequenze molto distanti evolutivamente sia piu' conveniente allineare le proteine piuttosto che le sequenze di DNA



posizioni

1 2 3 4 5 6 7 8

A	C	D	E	F	A C	A	D	F
A	F	D	A	-	- -	C	C	F
A	-	-	E	F	D -	F	D	C
A	C	A	E	F	- -	A	-	C
A	D	D	E	F	A A	A	D	F

Alignment



Alignment*

A	C	D	E	F		A	D	F
A	F	D	A	-		C	C	F
A	-	-	E	F		F	D	C
A	C	A	E	F		A	-	C
A	D	D	E	F		A	D	F

Escludo le posizioni troppo poco conservate (qui escludiamo colonne con piu' del 35% di spazi/indels)

Allineamento filtrato o seed

Escludo le posizioni troppo poco conservate (qui escludiamo colonne con piu' del 35% di spazi/indels)

Alignment

Alignment*

Calcolo le probabilità di emissione direttamente dai dati!

PROFILE(Alignment*)

	1	2	3	4	5	6	7	8	
A	A	C	D	E	F	A C	A	D	F
A	A	F	D	A	-	- -	C	C	F
A	A	-	-	E	F	D -	D	D	C
A	A	C	A	E	F	- -	A	-	C
A	A	D	D	E	F	A A	A	D	F
A	A	C	D	E	F		A	D	F
A	A	F	D	A	-	C	C	C	F
A	A	-	-	E	F	F	D	D	C
A	A	C	A	E	F	A	-	C	C
A	A	D	D	E	F	A	D	D	F
A	1	0	0	1/5	0	3/5	0	0	
C	0	2/4	0	0	0	1/5	1/4	2/5	
D	0	1/4	3/4	0	0	0	3/4	0	
E	0	0	0	4/5	0	0	0	0	
F	0	1/4	0	0	1	1/5	0	3/5	



Escludo le posizioni troppo poco conservate (qui escludiamo colonne con piu' del 35% di spazi/indels)

Alignment

Alignment*

Calcolo le probabilità di emissione direttamente dai dati!



PROFILE(Alignment*)

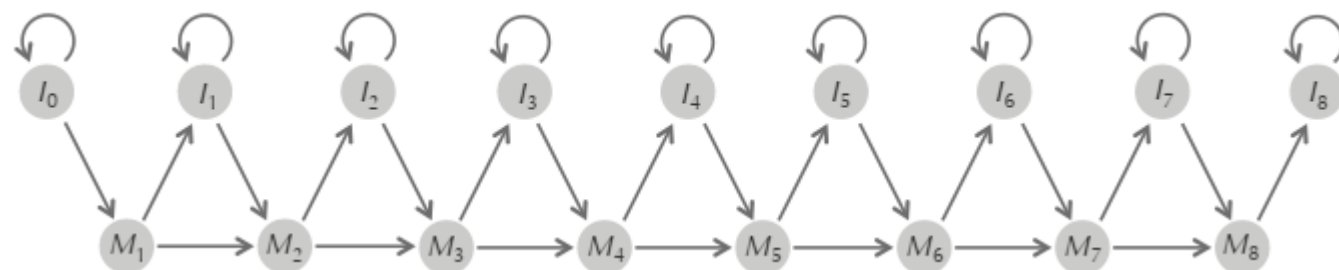
Ora posso calcolare la probabilità di osservare una sequenza «generabile» dal nostro allineamento

	1	2	3	4	5	6	7	8	
A	A	C	D	E	F	A C	A	D	F
A	A	F	D	A	-	- -	C	C	F
A	A	-	-	E	F	D -	D	D	C
A	A	C	A	E	F	- -	A	-	C
A	A	D	D	E	F	A A	A	D	F
A	A	C	D	E	F		A	D	F
A	A	F	D	A	-	C	C	C	F
A	A	-	-	E	F	F	D	D	C
A	A	C	A	E	F	A	-	C	C
A	A	D	D	E	F	A	D	D	F
A	1	0	0	1/5	0	3/5	0	0	
C	0	2/4	0	0	0	1/5	1/4	2/5	
D	0	1/4	3/4	0	0	0	3/4	0	
E	0	0	0	4/5	0	0	0	0	
F	0	1/4	0	0	1	1/5	0	3/5	



$$1 \cdot \frac{1}{4} \cdot \frac{3}{4} \cdot \frac{1}{5} \cdot 1 \cdot \frac{1}{5} \cdot \frac{3}{4} \cdot \frac{3}{5} = 0.003375.$$

	A	C	D	E	F	A	D	F
	A	F	D	A	—	C	C	F
<i>Alignment*</i>	A	—	—	E	F	F	D	C
	A	C	A	E	F	A	—	C
	A	D	D	E	F	A	D	F
Text	A	D	D	A	F	F	D	F
emission probability	1	1/4	3/4	1/5	1	1/5	3/4	3/5



Problemi???

1 2 3 4 5 6 7 8

A	C	D	E	F	A C	A	D	F
A	F	D	A	-	- -	C	C	F
A	-	-	E	F	D -	F	D	C
A	C	A	E	F	- -	A	-	C
A	D	D	E	F	A A	A	D	F

Alignment

Escludo le posizioni troppo poco conservate (qui escludiamo colonne con piu' del 35% di spazi/indels)



Alignment*

A	C	D	E	F		A	D	F
A	F	D	A	-		C	C	F
A	-	-	E	F		F	D	C
A	C	A	E	F		A	-	C
A	D	D	E	F		A	D	F

Calcolo le probabilità di emissione direttamente dai dati!



PROFILE(Alignment*)

A	1	0	0	1/5	0	3/5	0	0
C	0	2/4	0	0	0	1/5	1/4	2/5
D	0	1/4	3/4	0	0	0	3/4	0
E	0	0	0	4/5	0	0	0	0
F	0	1/4	0	0	1	1/5	0	3/5



Ora posso calcolare la probabilità di osservare una sequenza «generabile» dal nostro allineamento



Problemi???

Mi aspetto anche inserzioni,
delezioni e riarrangiamenti!!

Alignment

1	2	3	4	5	6	7	8	
A	C	D	E	F	A C	A	D	F
A	F	D	A	—	— —	C	C	F
A	—	—	E	F	D —	F	D	C
A	C	A	E	F	— —	A	—	C
A	D	D	E	F	A A	A	D	F

Escludo le posizioni troppo
poco conservate (qui
escludiamo colonne con
piu' del 35% di
spazi/indels)

Alignment*

	A	C	D	E	F		A	D	F
	A	F	D	A	-		C	C	F
	A	-	-	E	F		F	D	C
	A	C	A	E	F		A	-	C
	A	D	D	E	F		A	D	F

Calcolo le probabilità di
emissione direttamente
dai dati!

PROFILE(Alignment*)

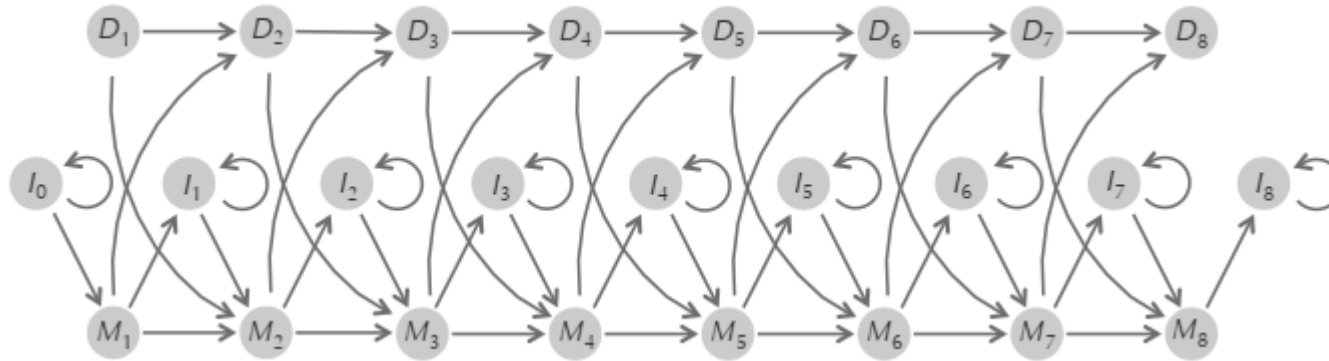
A	1	0	0	1/5	0	3/5	0	0
C	0	2/4	0	0	0	1/5	1/4	2/5
D	0	1/4	3/4	0	0	0	3/4	0
E	0	0	0	4/5	0	0	0	0
F	0	1/4	0	0	1	1/5	0	3/5



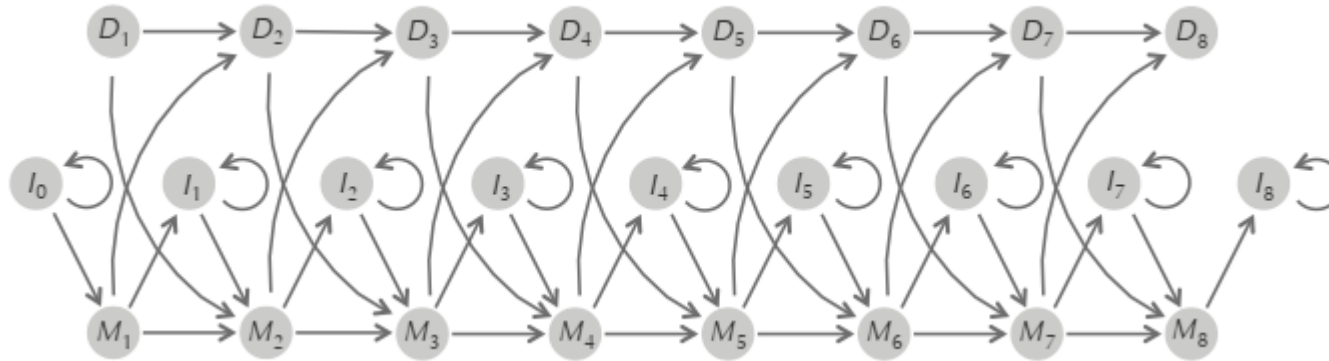
Ora posso calcolare la probabilità di
osservare una sequenza
«generabile» dal nostro allineamento



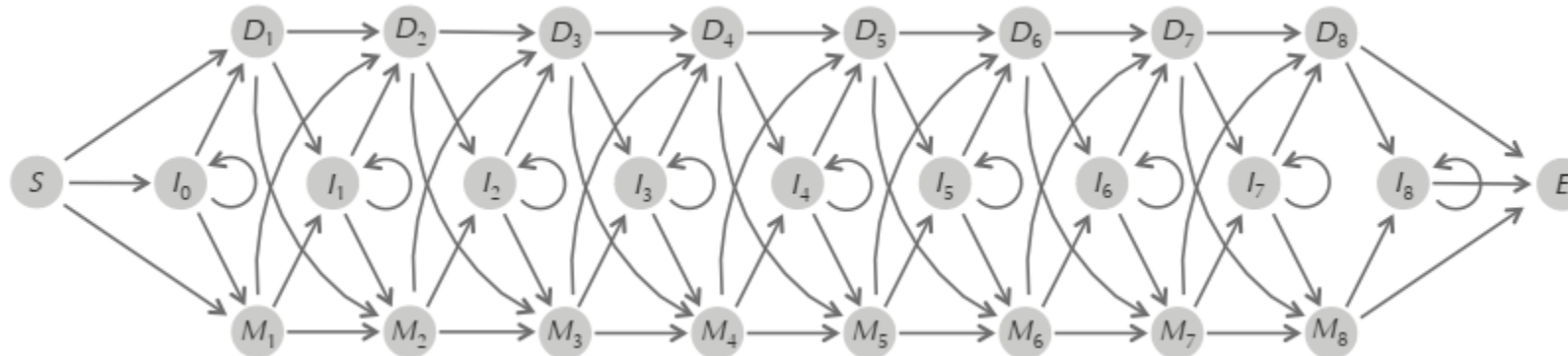
Introduciamo delezioni (D) come stati silenti (nessuna emissione)
ed inserzioni come stati che aggiungono aminoacidi aggiuntivi



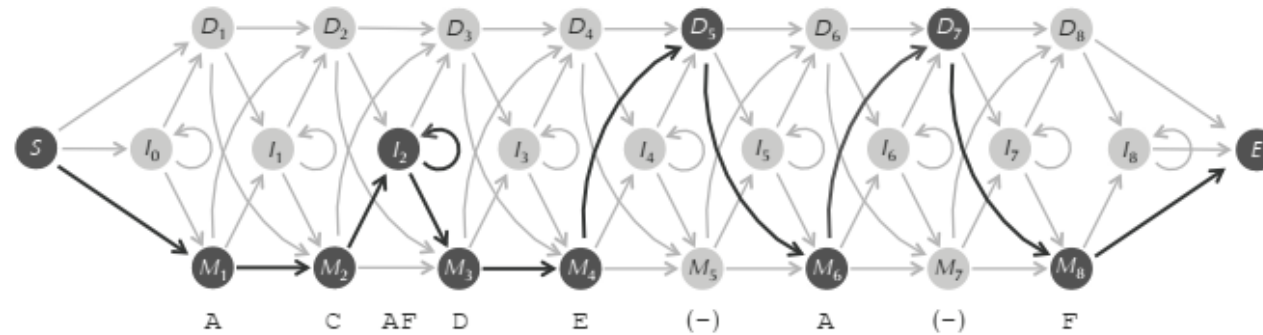
Introduciamo delezioni (D) come stati silenti (nessuna emissione)
ed inserzioni come stati che aggiungono aminoacidi aggiuntivi



Se congiungiamo delezioni ad inserzioni possiamo anche
modellizzare sostituzioni complesse

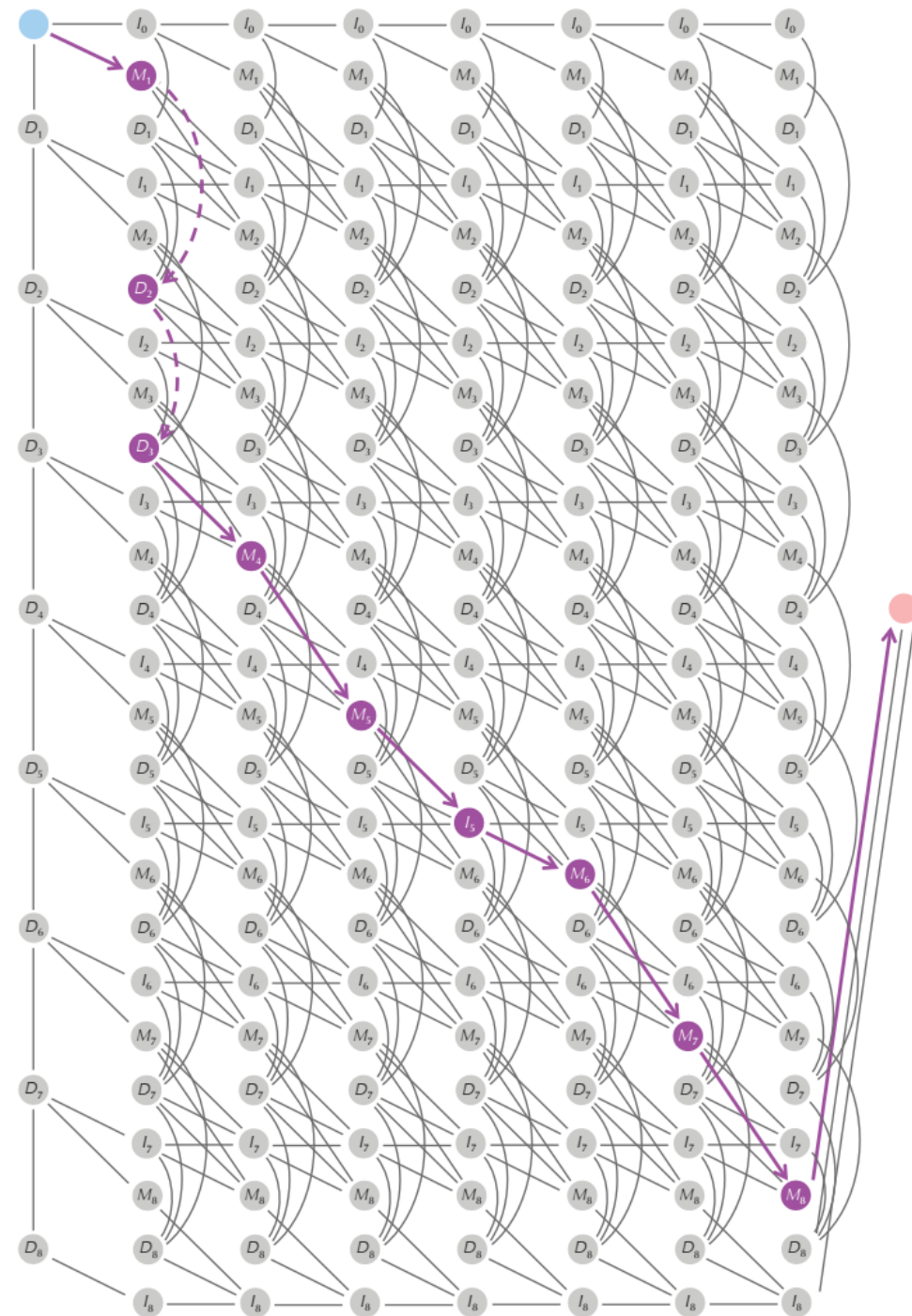


Una proteina rappresentata da una Profile HMM



Quale è la sequenza di questa proteina?

Rappresentazione di Viterbi per profile HMM



EddyRivasLab/ hmm

HMMER: biological sequence analysis using profile
HMMs



11

Contributors



16

Issues



388

Stars



85

Forks

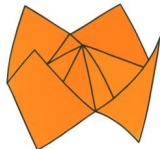


Esercitazioni

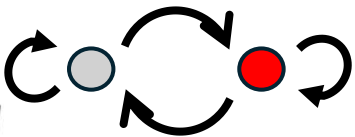
```
hmmbuild globin.hmm globin_alignment.fasta
```

```
hmmsearch --tblout hits.table \  
globin.hmm \  
uniprot_sprot.fasta
```

ATAC-seq



Profile HMM



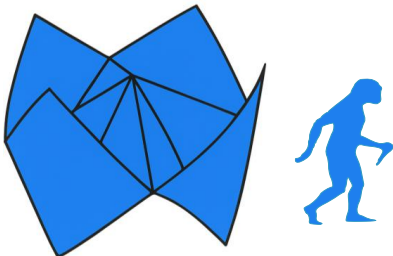
Isole CpG

applicazioni

(solo alcune)

degli Hidden Markov
Model in genomica

Aplotipi introgressi
da altre specie



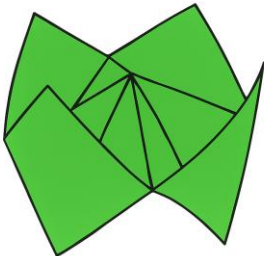
Metilazione



Segmentazione epigenomica



Ripetizioni nel genoma



Imputazione e phasing

