Markov chains and hidden Markov models

Sophie Schbath

Ecole Jeunes Chercheur.e.s du GDR BIM, Fréjus, 5 juin 2018

Introduction

Markov chains and hidden Markov chains, two probabilistic models used for very different pruposes in bioinformatics, and particularly in sequence analysis.

Markov chains are not really used to model a biological sequence ; they are rather used to set a background model for random sequences mimicing some features of an observed sequence, allowing to detect exceptional events (thanks to *p*-values/*e*-values).
$\longrightarrow$ Part 1 with an illustration on DNA motifs

Conversely, **hidden Markov chains** will often be used to model the alternation of predefined states along a process (a sequence, an alignment, a signal, etc.)
$\longrightarrow$ Part 2

Part 1 : Markov chains

Definition

A Markov chain (by default of order 1) is a sequence of random variables

$$X_1, X_2, X_3, \cdots, X_n, \cdots$$

which are generally **not independent** from each other.

More particularly, each random variable X_i depends on the previous variable X_{i-1} :

$$\mathbb{P}(X_i = b \mid X_1, X_2, \ldots, X_{i-1}) = \mathbb{P}(X_i = b \mid X_{i-1}).$$

In this lesson, we will only consider the case where the X_i's take value in a finite set - say alphabet $\mathcal{A}$, for instance $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

Transition matrix

Given the value of X_{i-1}, for instance $X_{i-1} = a$, the next random variable X_i takes its value according to the following transition probabilities :

$$\pi(a, b) = \mathbb{P}(X_i = b \mid X_{i-1} = a), a, b \in \mathcal{A}.$$

The matrix $\Pi = (\pi(a, b))_{a,b \in \mathcal{A}}$ is called the **transition matrix** of the Markov chain.

Transition matrix

Given the value of X_{i-1}, for instance $X_{i-1} = a$, the next random variable X_i takes its value according to the following transition probabilities :

$$\pi(a,b) = \mathbb{P}(X_i = b \mid X_{i-1} = a), a, b \in \mathcal{A}.$$

The matrix $\Pi = (\pi(a,b))_{a,b\in\mathcal{A}}$ is called the **transition matrix** of the Markov chain.
Example : $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$,

$$\Pi = \begin{pmatrix} 0.1 & 0.4 & 0.2 & 0.3 \\ 0.5 & 0 & 0.25 & 0.25 \\ 0.3 & 0.2 & 0.2 & 0.3 \\ 0.25 & 0.2 & 0.35 & 0.2 \end{pmatrix}$$

$$\begin{aligned} \mathbb{P}(X_i = \mathtt{c} \mid X_{i-1} = \mathtt{c}) &= 0 \\ \mathbb{P}(X_i = \mathtt{a} \mid X_{i-1} = \mathtt{t}) &= 0.25 \qquad \text{etc.} \end{aligned}$$

Note that $\sum_{b\in\mathcal{A}} \pi(a,b) = 1, \quad \forall a \in \mathcal{A}$.

Initial distribution

The distribution of the first random variable X_1 is given by the **initial distribution** of the Markov chain, say ν :

$$\nu(b) := \mathbb{P}(X_1 = b), \forall b \in \mathcal{A}.$$

Parameters $\theta = (\Pi, \nu)$ are sufficient to generate/simulate a Markov chain :

1. simulate X_1 from ν : $\mathbb{P}(X_1 = b) = \nu(b)$,
2. simulate X_2 from Π : $\mathbb{P}(X_2 = b \mid X_1) = \pi(X_1, b)$,
3. simulate X_3 from Π : $\mathbb{P}(X_3 = b \mid X_2) = \pi(X_2, b)$, etc.

Distribution of X_i

In a general Markov chain $(X_i)_i$ with parameters $\theta = (\Pi, \nu)$, the random variables X_i's are not identically distributed. Indeed,

- Distribution of X_1 : $\mathbb{P}(X_1 = b) = \nu(b)$
- Distribution of X_2 : $\mathbb{P}(X_2 = b) = (\nu.\Pi)(b)$
- Distribution of X_3 : $\mathbb{P}(X_3 = b) = (\nu.\Pi^2)(b)$
- etc

More generally $\mathbb{P}(X_i = b) = (\nu.\Pi^{i-1})(b)$.

This result directly follows from the following property : **the transition matrix in k steps is Π^k** :

$$\mathbb{P}(X_{i+k} = b \mid X_i = a) = \Pi^k(a, b)$$

Exercice : prove the above property.

Stationary distribution

However, a theorem says that if the Markov chain satisfies "*some conditions*", then **there exists a distribution** μ, such that :

$$\mathbb{P}(X_i = b) \longrightarrow \mu(b), \text{ as } i \to \infty.$$

μ is called the **stationary distribution** of the Markov chain and satisfies $\mu = \mu.\Pi$

The *conditions* for the Markov chain are

- irreducibility
- aperiodicity
- positive recurrence

Particular case : These conditions are fullfilled if the state space $\mathcal{A}$ is finite and if all transition probabilities $\pi(a, b) > 0$.

Stationary distribution (2)

Irreducibility : all states are reachable from any states :

$$\forall a, b \in \mathcal{A}, , b \in \mathcal{A}, \exists k \text{ such that } \Pi^k(a, b) > 0$$

Note : Irreducibility $\Rightarrow$ at most one stationary distribution.

Aperiodicity :

$$\forall a, b \in \mathcal{A}, gcd\{k \geq 1 \mid \Pi^k(a, b) > 0\} = 1$$

Note : $\pi(a, b) > 0 \Rightarrow$ irreducibility and aperiodicity

Positive recurrence : the expectation of the return time at state a is finite, $\forall a \in \mathcal{A}$
Note : positive recurrence $\Rightarrow$ existence of stationary distributions.
Note : $\mathcal{A}$ finite + irreducibility $\Rightarrow$ positive recurrence

Stationary Markov chain

The **stationary Markov chain** with transition matrix Π is the Markov chain with parameters $\theta = (\Pi, \mu)$, i.e. its initial distribution is set to the stationary distribution.

By doing so, **all random variables X_i are identically distributed with distribution** μ (indeed, $\mu.\Pi^{i-1} = \mu$).

In most applications of Markov models, the stationary framework is considered.

Probability of pattern occurrences

Let $X_1 X_2 X_3 \cdots X_n \cdots$ be a stationary Markov chain with transition matrix Π on the alphabet $\mathcal{A}$; μ denotes the stationary distribution.

Let $w_1 w_2 \cdots w_\ell$ be a sequence of ℓ letters from $\mathcal{A}$ (a "word"), for instance `gttacg`.

The probability to observe such word in the random sequence, say starting at position i, can be easily calculated thanks to the Bayes formula :

$$\mathbb{P}(X_i X_{i+1} \cdots X_{i+\ell-1} = w_1 w_2 \cdots w_\ell) = \mu(w_1) \prod_{j=1}^{\ell-1} \pi(w_j, w_{j+1})$$

for instance

$$\mathbb{P}(\texttt{gttacg} \text{ at position } i) = \mu(\texttt{g})\pi(\texttt{g},\texttt{t})\pi(\texttt{t},\texttt{t})\pi(\texttt{t},\texttt{a})\pi(\texttt{a},\texttt{c})\pi(\texttt{c},\texttt{g})$$

Exercice : prove the above equation.

Probabilist view : The most probably sequence ?

When the Markov model is known, i.e. the transition matrix Π is given, it is then possible to look for the observed sequence of length n the most likely to be drawn.
For that, one just has to maximise the **likelihood**

$$\mathbb{P}(X_1 X_2 \cdots X_n = x_1 x_2 \cdots x_n)$$

over all possible observed sequences $x_1 x_2 \cdots x_n \in \mathcal{A}^n$.

Remember :

$$\begin{aligned}
&\mathbb{P}(X_1 X_2 \cdots X_n = x_1 x_2 \cdots x_n) \\
&= \mu(x_1)\pi(x_1, x_2) \times \cdots \times \pi(x_{n-1}, x_n) \\
&= \mu(x_1) \prod_{a,b \in \mathcal{A}} \big(\pi(a,b)\big)^{N^{\text{obs}}(ab)}
\end{aligned}$$

where $N^{\text{obs}}(ab)$ is the count of the 2-letter word ab in $x_1 x_2 \cdots x_n$.

Statistician view : The best transition matrix ?

In statistics, the observed sequence is given and one looks for the parameters (Π) which maximize the likelihood

$$L(\Pi) = \mu(x_1) \prod_{a,b \in \mathcal{A}} \big(\pi(a,b)\big)^{N^{\text{obs}}(ab)}$$

This is the well known **maximum likelihood estimation** procedure.

In our case, the (log-)likelihood is a function of $|\mathcal{A}| \times (|\mathcal{A}| - 1)$ free parameters, and resolving the partial derivatives $= 0$ leads to :

$$\widehat{\pi}(a,b) = \frac{N^{\text{obs}}(ab)}{N^{\text{obs}}(a+)}, \forall a, b \in \mathcal{A}.$$

Similarly, the stationary distribution will be estimated by

$$\widehat{\mu}(a) = \frac{N^{\text{obs}}(a)}{n}.$$

Interpretation of the MLE

Let $N(ab)$ be the number of occurrences of the 2-letter word ab in the Markov chain $X_1X_2X_3\cdots X_n$ with transition matrix Π.
We can show that

$$\mathbb{E}N(ab) = (n-1)\mathbb{P}(ab \text{ at position } i) = (n-1)\mu(a)\pi(a,b)$$

By using the MLE's $\widehat{\pi}(a,b) = \frac{N^{\text{obs}}(ab)}{N^{\text{obs}}(a+)}$ and $\widehat{\mu}(a) = \frac{N^{\text{obs}}(a)}{n}$, we have

$$\widehat{\mathbb{E}N(ab)} = \frac{n-1}{n}\frac{N^{\text{obs}}(a)}{N^{\text{obs}}(a+)}N^{\text{obs}}(ab) \simeq N^{\text{obs}}(ab).$$

In average, the random sequences would have the same 2-letter word composition than the observed sequence.

Conclusion : The Markov model of order 1 (denoted by **M1) allows to fit the counts of all 2-letter words** of the observed sequence.

Generalization to Mm, m-order Markov model

Definition : In a m-order Markov chain, the Markov property is

$$\mathbb{P}(X_i = b \mid X_1, X_2, \ldots, X_{i-1}) = \mathbb{P}(X_i = b \mid X_{i-m}, \ldots, X_{i-1}).$$

The initial/stationary distribution is given by

$$\mu(a_1 a_2 \cdots a_m) := \mathbb{P}(X_1 = a_1, \ldots, X_m = a_m)$$

The MLE of the parameters are

$$\begin{aligned} \widehat{\pi}(a_1 a_2 \cdots a_m, a_{m+1}) &= \frac{N(a_1 a_2 \cdots a_m a_{m+1})}{N(a_1 a_2 \cdots a_m +)} \\ \widehat{\mu}(a_1 a_2 \cdots a_m) &= \frac{N(a_1 a_2 \cdots a_m)}{n - m + 1} \end{aligned}$$

Conclusion : The model **Mm fits the counts of all $(m+1)$-letter words** of the observed sequence.

Bernoulli model = M0

The Bernoulli model where the random variables X_i's are independent and identically distributed with distribution μ is a particular case of the Markov model.
It would correspond to the following transition matrix in which all lines equal the stationary distribution :

$$\Pi = \begin{pmatrix} \mu(\mathtt{a}) & \mu(\mathtt{c}) & \mu(\mathtt{g}) & \mu(\mathtt{t}) \\ \mu(\mathtt{a}) & \mu(\mathtt{c}) & \mu(\mathtt{g}) & \mu(\mathtt{t}) \\ \mu(\mathtt{a}) & \mu(\mathtt{c}) & \mu(\mathtt{g}) & \mu(\mathtt{t}) \\ \mu(\mathtt{a}) & \mu(\mathtt{c}) & \mu(\mathtt{g}) & \mu(\mathtt{t}) \end{pmatrix}.$$

The Bernoulli model will be denoted by M0 in the remainder (order 0).

Model **M0 fits the letter composition** of the observed sequence.

Illustration : DNA motifs with unexpected frequencies

The Chi example in *E. coli*

Observation : The octamer `gctggtgg` occurs 762 times in the *Escherichia coli* genome (leading strands, $n = 4\,638\,858$).
Is it expected or (significantly) unexpected ?
$\Rightarrow$ **What should we expect ?** To what to compare ?

- random sequences of length $n = 4\,638\,858$ (model M00)
- + same letter composition than *E. coli* genome (model M0)
- + same 2-letter word composition than *E. coli* genome (M1),
- etc.
- + same 7-letter word composition than *E. coli* genome (M6)

Two/three steps :

1. choose a sequence model and estimate the parameters
2. compute the estimated expected count $\widehat{\mathbb{E}N}$ of the motif
3. compute the *p*-value $\mathbb{P}(N \geq 762)$ to assess significance

Influence of the model

Results for `gctggtgg` occurring 762 times in *E. coli* genome.

model	fit	expected	score	*p*-value	rank
M00	length	70.783			
M0	bases	85.944	72.9	$< 10^{-323}$	3
M1	dimers	84.943	73.5	$< 10^{-323}$	1
M2	3-mers	206.791	38.8	$< 10^{-323}$	1
M3	4-mers	355.508	22.0	$1.4\,10^{-107}$	5
M4	5-mers	355.312	22.9	$2.3\,10^{-116}$	2
M5	6-mers	420.867	19.7	$1.0\,10^{-86}$	1
M6	7-mers	610.114	10.6	$1.5\,10^{-26}$	3

Influence of the model (2)

Expected counts (**top 5** **top 50**)

		gctggtgg 762 occ.	ggcgctgg 828 occ.	ccggccta 71 occ.
M0	bases	**85.944**	**85.524**	70.445
M1	2-mers	**84.943**	**125.919**	48.173
M2	3-mers	**206.791**	**255.638**	35.830
M3	4-mers	**355.508**	**441.226**	**14.697**
M4	5-mers	**355.312**	**392.252**	**15.341**
M5	6-mers	**420.867**	**633.453**	**27.761**
M6	7-mers	**610.114**	812.339	**25.777**

Over-represented $\not\Leftrightarrow$ significantly over-represented

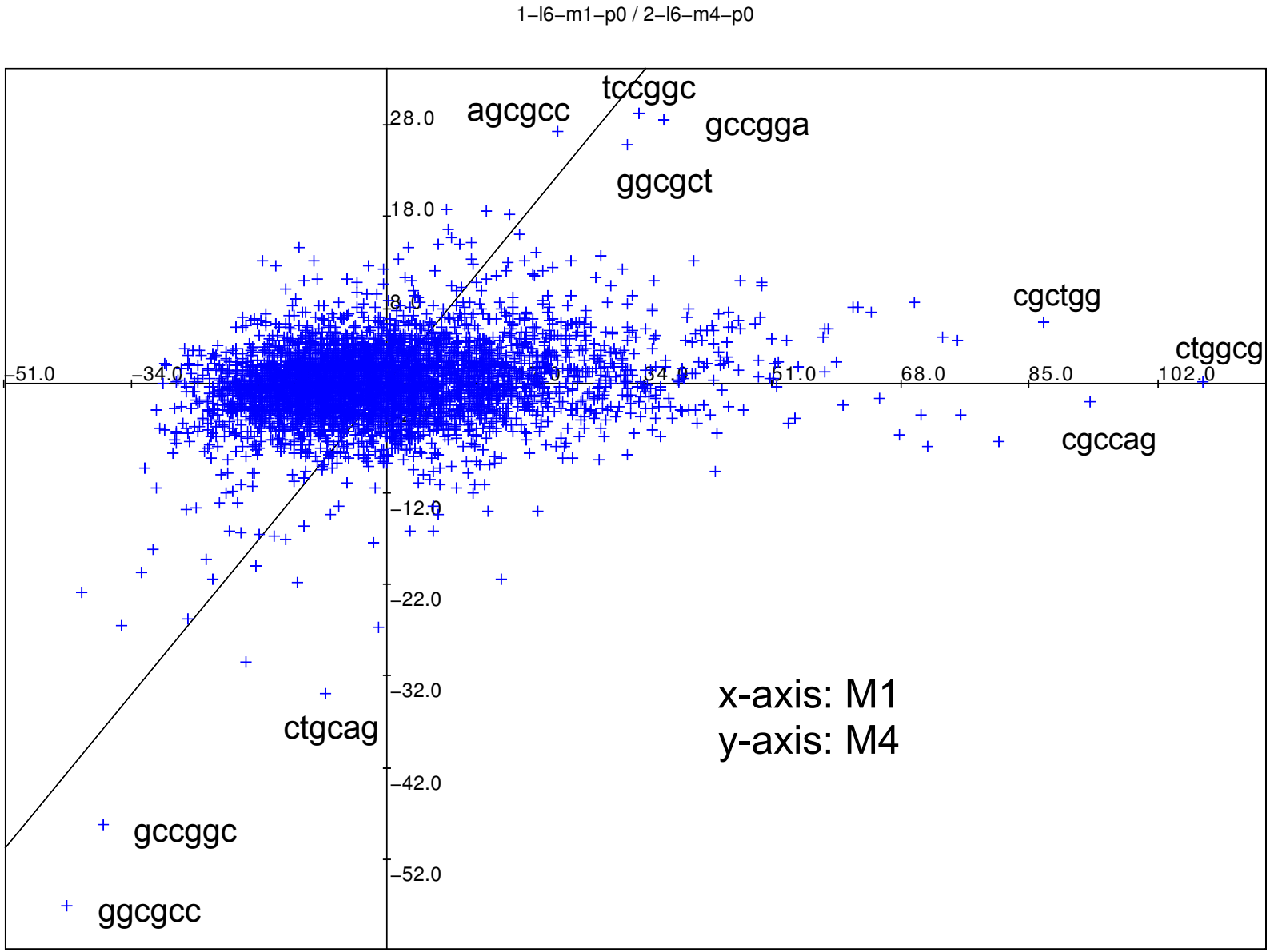
frequent $\not\Leftrightarrow$ significantly over-represented

Influence of the model (3)

The higher the order m of the Markov chain :

- the better fit to the observed sequence (in terms of kmer composition)
- the less significantly unexpected words (see next figure)

Influence of the model (4)



Influence of the model (5)

The higher the order m of the Markov chain :

- the better fit to the observed sequence (in terms of kmer composition)
- the less significantly unexpected words (see next figure)

but

- the higher the number of parameters to estimate ($\rightarrow$ needs a long sequence)
- for a particular word, the expected count is not necessarily closer to the observed count

$\Rightarrow$ **Choose the model according to the amount of data available and to the biological information you want to take into account to compute "what to expect ?"**

Part 2 : Hidden Markov models (HMM)

Limitation of Markov chain models

Markov chains from Part 1 are **homogeneous**, i.e. the transition probabilities $\pi(a, b)$ do not depend on the positions in the sequence. Depending on the application, this can be a strong assumption.

Here are some examples :

1. there exists an intrinsic 3-periodicity in coding sequences (codons)
2. compositional biases exist between exons/introns/intergenic regions due to selection pressure
3. compositional biases (%GC) exists in eukaryotes (isochores)
4. compositional biaises exist between variable/conserved regions in virus

Limitation of Markov chain models (2)

3-periodicity in coding sequences can be easily taken into account by using 3 transition matrices, one for each position inside codons.

Theoretically, one could consider an **heterogenous Markov chain** with transition matrices Π_i, $i = 1, \ldots, n$, but their estimation is usually problematic.

If the sequence of the different states (eg. exons/introns/intergenic) is known, one can use a transition matrix per state (eg. Π_{exon}, Π_{intron}, $\Pi_{\text{intergenic}}$), ...
... but generally the segmentation is part of the problem (eg. gene prediction)

$\Longrightarrow$ Hidden Markov models

Definition of an HMM

Two random processes :

1. An unobservable ("hidden") process which represents the alternating states $S_1 S_2 S_3 \cdots S_n$, $S_i \in \mathcal{S}$ and is a stationary **Markov chain of order 1**. Parameters will be denoted by Π_s and μ_s.
2. An observable process $X_1 X_2 X_3 \cdots X_n$, $X_i \in \mathcal{A}$ for which the distribution of each variable X_i ($\mathbb{P}(X_i = \cdot)$) depends on the state S_i. No more precision of these distributions.

Example 1

- $\mathbf{X} = X_1 X_2 \cdots X_n$ is a random sequence on $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.
- $\mathcal{S} = \{\text{exon, intron, intergenic }\}$ and $\mathbf{S} = S_1 S_2 \cdots S_n$ models the gene structure along the sequence $\mathbf{X}$
- $\mathbf{S}$ is a 1-order Markov chain on $\mathcal{S}$, its transition matrix Π_S is of dimension 3×3, for instance

$$\Pi_S = \begin{pmatrix} 0.8 & 0.15 & 0.05 \\ 0.1 & 0.9 & 0 \\ 0.05 & 0 & 0.95 \end{pmatrix}.$$

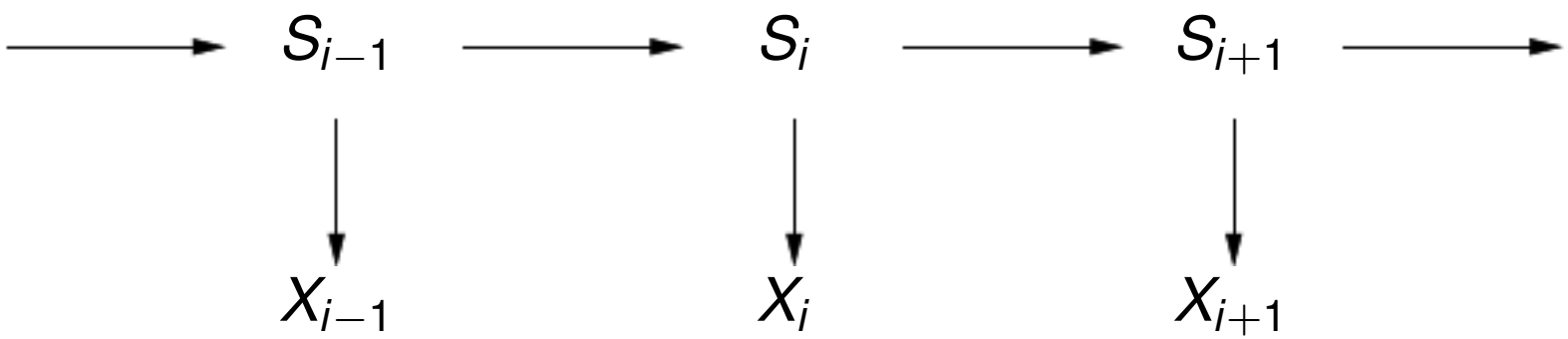
- Conditionnaly on $\mathbf{S}$, the letters X_i's are independent and distributed according to either p_{exon} or p_{intron} or $p_{\text{intergenic}}$:

$$p_u(b) = \mathbb{P}(X_i = b \mid S_i = u), \quad b \in \mathcal{A}, \quad u \in \mathcal{S}$$

for instance
$p_{\text{exon}} = (.2, .3, .3, .2)$, $p_{\text{intron}} = (.3, .2, .2, .3)$, $p_{\text{intergenic}} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$

Example 1 : dependence graph

The dependence graph of the previous HMM model is :



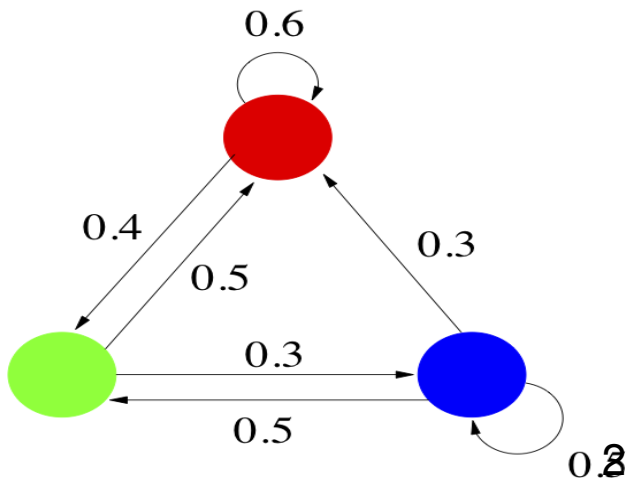
- **S** is a 1-order Markov chain on $\mathcal{S}$
- Conditionnaly on **S**, the letters X_i's are independent and $\mathbb{P}(X_i = b \mid S_i = u) = p_u(b)$.

Graphical representation of an HMM

The underlying structure of an HMM is usually represented by a graph in which nodes are the states and edges indicate non null transition probabilities between states.

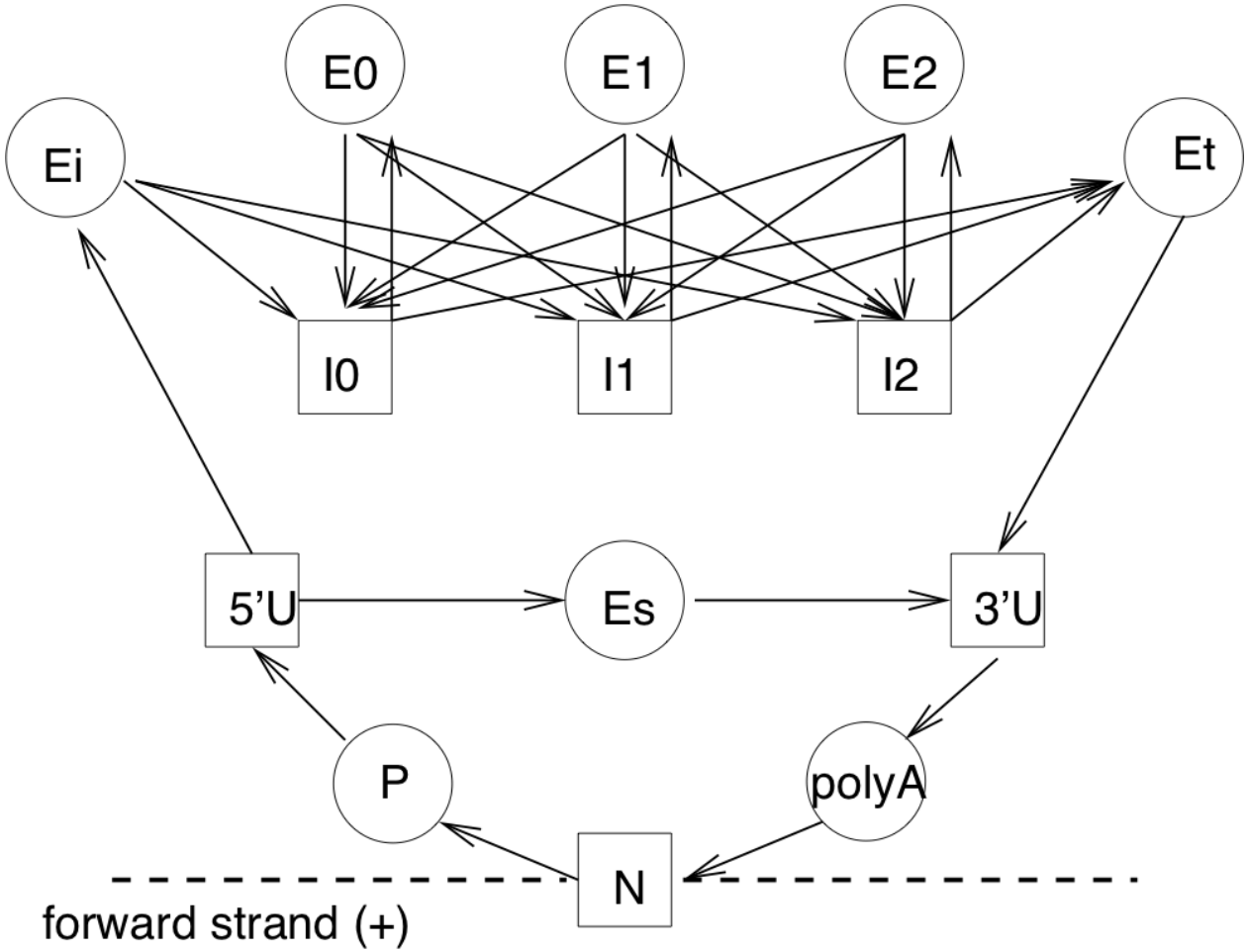
For instance, the graph below would represent an HMM with 3 states and the following transition matrix

$$\Pi_S = \begin{pmatrix} \mathbf{0.6} & 0.4 & 0 \\ 0.5 & \mathbf{0} & 0.5 \\ 0.3 & 0.5 & \mathbf{0.2} \end{pmatrix}$$

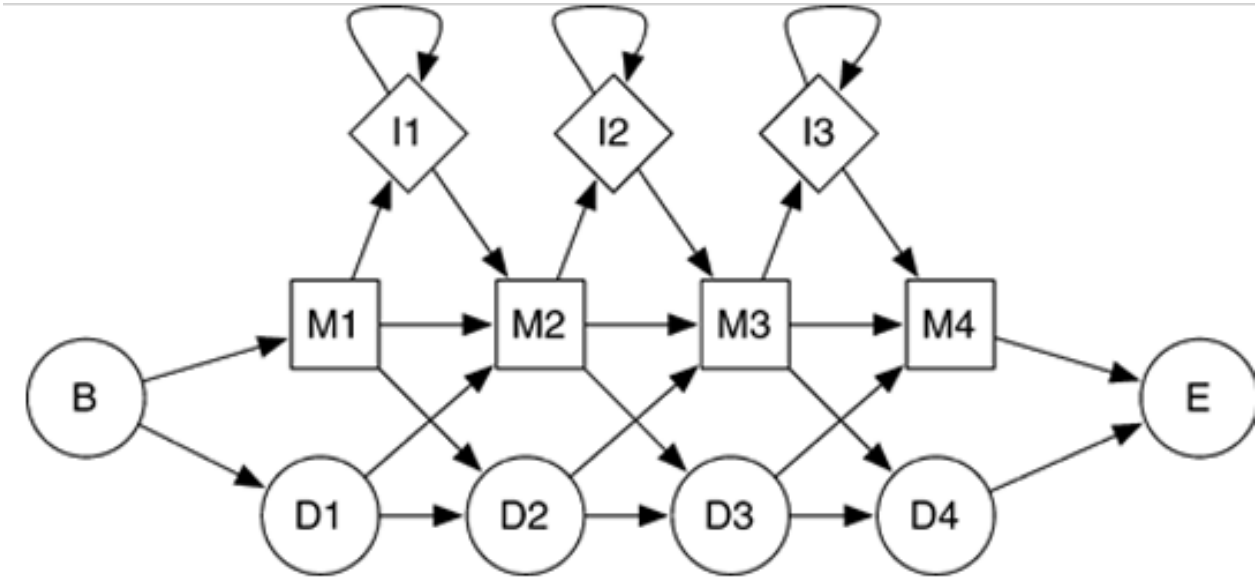


Example 2 : gene prediction

Architecture of the Genscan model for gene prediction



Example 3 : profile HMM for multiple alignment



from HMMER home page `https://bibiserv.cebitec.uni-bielefeld.de/sadr2/databasesearch/hmmer/index.html`

Segmentation property

Time spent in a given state $u \in \mathcal{S}$ is geometrically distributed :

$$\mathbb{P}(\text{time } t \text{ in state } u) = \Big(\pi_S(u,u)\Big)^{t-1}\big(1-\pi_S(u,u)\big)$$

meaning that the expected length of segments in state u is $\frac{1}{1-\pi_S(u,u)}$.

This can be a modeling limitation.

Estimation/segmentation

Given an observed sequence $x_1 x_2 \cdots x_n$, the aim is double :

1. estimate the parameters $\theta = (\Pi_S, p_u)$ for each state $u \in \mathcal{S}$ (**estimation step**)
2. recover all the hidden states $s_1 s_2 \cdots s_n$ (**segmentation step**)

For that, there are two possible approaches :

- a supervised approach, which requires a learning set, that is another observed sequence $\mathbf{x}'$ already segmented,
- an unsupervised approach.

Supervised approach

Two steps :

1. step 1 : **estimation** of $\theta = (\Pi_S, p_u)$ from a segmented sequence $\mathbf{x}'$
   $\rightarrow$ maximum likelihood method leading to $\widehat{\theta}$
2. step 2 : **segmentation** of $\mathbf{x}$ given the parameters $\widehat{\theta}$
   $\rightarrow$ Viterbi algorithm leading to states $s_1^* s_2^* \cdots s_n^*$
   (or *forward-backward* algorithm leading to $\mathbb{P}(S_i = u \mid \mathbf{X} = \mathbf{x}; \widehat{\theta})$)

In the remainder, we consider the basic HMM where the X_i's are independent conditionnaly on the hidden states.

Supervised approach : estimation step (MLE)

Here we assume that we observe the segmentation $\mathbf{s} = s_1 s_2 \cdots s_n$.

The likelihood is then $\mathbb{P}(\mathbf{X} = x_1 x_2 \cdots x_n \mid \mathbf{s}; \theta) =$

$$\begin{aligned} &= \mu_S(s_1)\pi_S(s_1, s_2)\ldots\pi_S(s_{n-1}, s_n) \times p_{s_1}(x_1)\ldots p_{s_n}(x_n) \\ &= \mu_S(s_1) \prod_{u,v\in\mathcal{S}} \pi_S(u,v)^{N(uv)} \times \prod_{u\in\mathcal{S}} \prod_{a\in\mathcal{A}} p_u(a)^{N(u,a)} \end{aligned}$$

where $N(uv)$ is the count of uv in $\mathbf{s}$ and $N(u, a)$ is the number of letter a in state u.

The likelihood maximization gives the following estimated parameters :

$$\begin{aligned} \widehat{\pi}_S(u,v) &= \frac{N(uv)}{N(u+)} \\ \widehat{p}_u(a) &= \frac{N(u,a)}{N(u)} \end{aligned}$$

Supervised approach : segmentation step (Viterbi)

Given $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and θ, we look for $(s_1^*, s_2^*, \ldots, s_n^*)$ which maximize

$$\mathbb{P}(S_1 = s_1, \ldots, S_n = s_n \,|\, \mathbf{X} = \mathbf{x}; \theta)$$

or (Bayes formula)

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, S_1 = s_1, \ldots, S_n = s_n \,; \theta)$$

Supervised approach : segmentation step (Viterbi)

Given $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and θ, we look for $(s_1^*, s_2^*, \ldots, s_n^*)$ which maximize

$$\mathbb{P}(S_1 = s_1, \ldots, S_n = s_n \,|\, \mathbf{X} = \mathbf{x}; \theta)$$

or (Bayes formula)

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, S_1 = s_1, \ldots, S_n = s_n \,; \theta)$$

Let $\mathbb{P}^* = \max_{s_1,\ldots,s_n} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, S_1 = s_1, \ldots, S_n = s_n \,; \theta)$.

$$\mathbb{P}^* = \max_{v} \underbrace{\max_{s_1,\ldots,s_{n-1}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, S_1 = s_1, \ldots, S_{n-1} = s_{n-1}, S_n = v \,; \theta)}_{Z_n(v)}$$

$$s_n^* = \arg\max_v Z_n(v)$$

Viterbi algorithm (2)

$$\mathbb{P}^* = \max_{v} \underbrace{\max_{s_1,\ldots,s_{n-1}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, S_1 = s_1, \ldots, S_{n-1} = s_{n-1}, S_n = v\,;\theta)}_{Z_n(v)}$$

$$s_n^* = \arg\max_v Z_n(v)$$

Forward recursive formula to compute $Z_i(v)$:

$$\begin{cases} Z_1(v) &= \mathbb{P}(X_1 = x_1, S_1 = v) = \mu_S(v) p_v(x_1) \\ Z_i(v) &= \max_u \Big(Z_{i-1}(u) \pi_S(u,v) \Big) p_v(x_i) \end{cases}$$

Backward recursion for s_i^* :

$$s_{i-1}^* = \arg\max_u (Z_{i-1}(u) \pi_S(u, s_i^*))$$

Unsupervised approach

If the segmentation **S** is unknown, the likelihood $\mathbb{P}(\mathbf{X} = x_1, x_2, \ldots, x_n\,; \theta)$ is not tractable.
$\Rightarrow$ EM algorithm to approximate the MLE.

Expectation-Maximization is an iterative algorithm such that the likelihood increases at each step :

- start with an initial value $\theta^{(0)}$
- iteration k, $k = 1, 2 \ldots$
  - **step E** : compute $\mathbb{P}(S_i^{(k)} = u \mid \mathbf{X} = \mathbf{x}, \theta^{(k-1)})$, $i = 1, \ldots, n$, $u \in \mathcal{S}$ (*Forward-Backward* algorithm)
  - **step M** : compute $\theta^{(k)}$ given the distribution of $\mathbf{S}^{(k)}$ by maximizing $\mathbb{E}_S[\log \mathbb{P}(\mathbf{X}, \mathbf{S}^{(k)}\,; \theta^{(k-1)}) \mid \mathbf{X}]$
- stop when $|\log \mathbb{P}(\mathbf{X} = \mathbf{x}\,; \theta^{(k+1)}) - \log \mathbb{P}(\mathbf{X} = \mathbf{x}\,; \theta^{(k)})| < \varepsilon$ ou $k > M$

EM : M step

Maximization step :

$$\pi_s^{(k)}(u,v) = \frac{\sum_i \mathbb{P}(S_i^{(k)} = u, S_{i+1}^{(k)} = v \mid \mathbf{X} = \mathbf{x}\,; \theta^{(k-1)})}{\sum_i \mathbb{P}(S_i^{(k)} = u \mid \mathbf{X} = \mathbf{x}\,; \theta^{(k-1)})}$$

$$p_u^{(k)}(a) = \frac{\sum_i \mathbf{1}\{X_i = a\}\mathbb{P}(S_i^{(k)} = u \mid \mathbf{X} = \mathbf{x}\,; \theta^{(k-1)})}{\sum_i \mathbb{P}(S_i^{(k)} = u \mid \mathbf{X} = \mathbf{x}\,; \theta^{(k-1)})}$$

For more details