BIOINFORMATIK II
PROBABILITY & STATISTICS
Summer semester 2008

The University of Zürich and ETH Zürich

Lecture 3: Markov chains.

Prof. Andrew Barbour
Dr. Nicolas Pétrélis

Adapted from a course by
Dr. D. Schuhmacher & Dr. D. Svensson.

Random processes

A **random process** (or **stochastic process**) in discrete time is a *sequence* of random variables:

$$X_0, X_1, X_2, X_3, \ldots$$

They are usually dependent.

Typically, the process describes something that *evolves* in time:
$X_i =$ the **state** of the process at time i.

Examples of random processes:

Ex. 1: Let $X_0, X_1, X_2, \ldots$ be independent random variables, $X_i \sim Bin(100, 0.25)$ for $i \geq 0$.

No dependence at all, and a trivial process.

Ex. 2: Let $X_0 = 0$, and set $X_i = X_{i-1} + Z_i$ for $i \geq 1$, where $Z_1, Z_2, \ldots$ are independent random variables with $\mathbf{P}(Z_i = -1) = \mathbf{P}(Z_i = +1) = 0.5$.

Dependence? Given the history $X_0, X_1, \ldots, X_n$ of the process (including the current state X_n), the future $X_{n+1}, X_{n+2}, \ldots$, depends on the history only through the current state X_n.

Ex. 3: Let $X_0 \sim Bin(100, 0.25)$ and let $X_i := X_0$ for all $i \geq 1$.

This is the strongest possible dependence (but a trivial process).

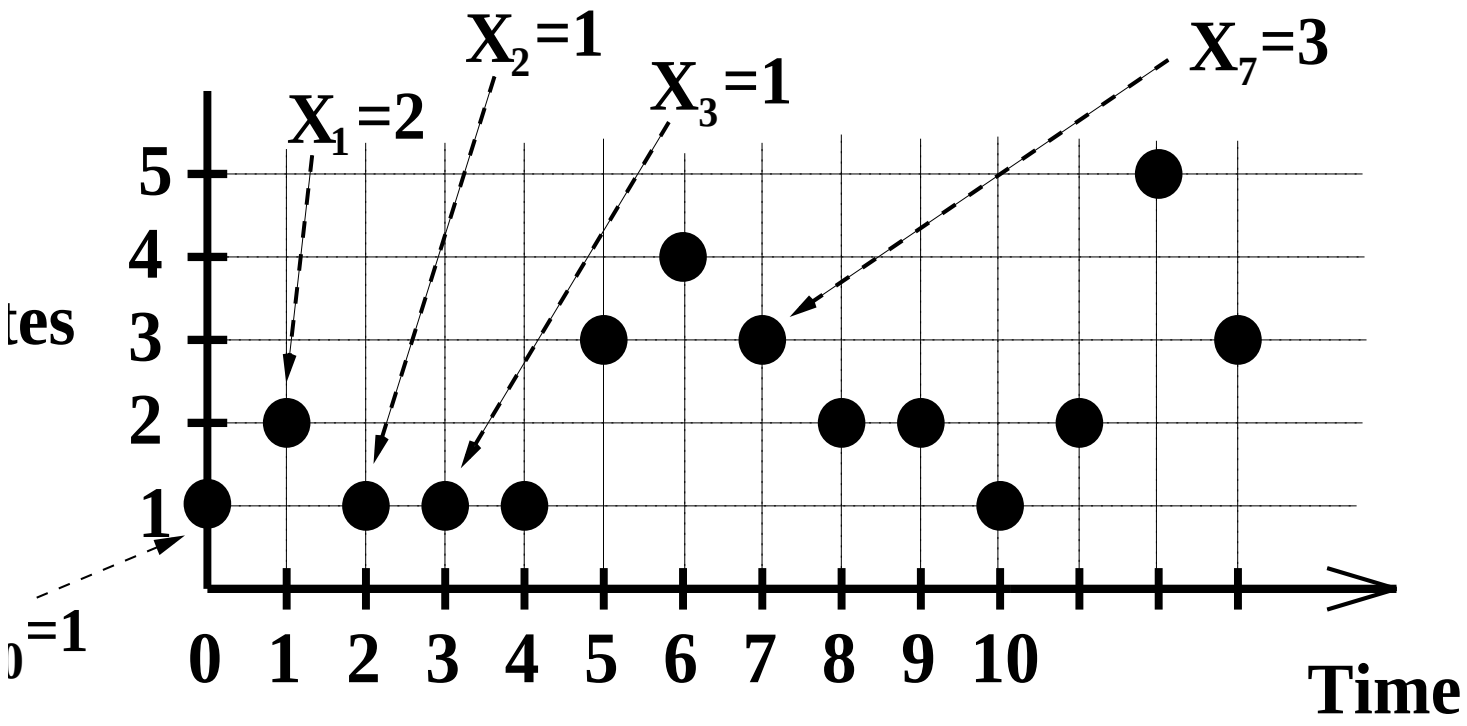
There are infinitely many examples ...

Markov chains (MC), overview:

Markov chains are of *high importance* in bioinformatics.

- A Markov chain is a particular type of **random process**, typically *evolving in time*.
- At each time point the Markov chain *visits one of a number of possible* **states**.
- A Markov chain is '*memoryless*' in the sense that '*only the current state of the process matters for the future*'

$(X_j)_{j=0}^{\infty}$ = Markov chain. At each time point the MC visits a state:



Definition: A **Markov chain** is a time-homogeneous Markovian random process which takes values in a state space S.

'Markovian' means: At each time, only the current state is important for the future:

$$\mathbf{P}(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n)$$
$$= \mathbf{P}(X_{n+1} = i_{n+1} | X_n = i_n).$$

(This is *not* true for general random processes!)

'time-homogeneous' means: The probability that a Markov chain jumps in one time step from state i to state j does not depend on time:

$$\mathbf{P}(X_{n+1} = j | X_n = i) = \mathbf{P}(X_1 = j | X_0 = i).$$

Why Markov chains? *Intuitive, mathematically tractable, well-studied topic, many good applications, ...*

Applications: *sequence evolution, scoring matrices for sequence alignments, phylogenetic tree reconstructions, much more!*

- The states visited are *not necessarily numerical*. Can e.g. be a, c, g and t.
- The concept of 'time' *can be replaced by 'space'*:
  e.g. *position in a sequence.*

.....cacagctagctacgactacgacacttttattactttattatatcagcgc.....

Example: Markov chains as a model for DNA.
Let X_j be equal to the nucleotide present at position j (counted from left to right) in a gene, say, for $j = 1, 2, \ldots, N$.

An i.i.d. random sequence of a, c, g and t's is unlikely to be a good model for the nucleotide pattern in a gene sequence

A Markov chain on $\{a, c, g, t\}$ might be a better *approximation* to reality: the probabilities for the nucl. at position $j + 1$ depends upon the nucl. at position j.

(However, in real data, more complex dependencies are found.)

Transition matrix

Let $X_0, X_1, X_2, X_3, \ldots$ be a Markov chain with **state space** S, for example $S = \{a, c, g, t\}$.

To any Markov chain, a **transition matrix P** is associated.

In the case with four possible states, P is 4×4:

$$P = \begin{pmatrix} p_{a,a} & p_{a,c} & p_{a,g} & p_{a,t} \\ p_{c,a} & p_{c,c} & p_{c,g} & p_{c,t} \\ p_{g,a} & p_{g,c} & p_{g,g} & p_{g,t} \\ p_{t,a} & p_{t,c} & p_{t,g} & p_{t,t} \end{pmatrix}.$$

Here

$$p_{i,j} = \mathbf{P}(X_{n+1} = j | X_n = i)$$

for $n \geq 0$, where $i, j \in \{a, c, g, t\}$.

The ('one-step') transition matrix P contains all the information needed to compute any *'multi-step' transition probabilities* to future states:

The *two-step probabilities* $p_{ij}^{(2)} := \mathbf{P}(X_{n+2} = j | X_n = i)$ are given by the matrix

$$P^{(2)} = P \cdot P \text{ (matrix multiplication!)}$$

(where $i, j \in S$).

E.g. with only two states (say, $S = \{0, 1\}$)

$$P^{(2)} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \cdot \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} =$$

$$\begin{pmatrix} p_{00}^2 + p_{01} \cdot p_{10} & p_{00} \cdot p_{01} + p_{01} p_{11} \\ p_{10} \cdot p_{00} + p_{11} \cdot p_{10} & p_{10} \cdot p_{01} + p_{11}^2 \end{pmatrix}.$$

How to derive the two-step transition matrix $P^{(2)}$?

Let $i, j \in S$. Then

$$\mathbf{P}(X_{n+2} = j | X_n = i) = \sum_{k \in S} \mathbf{P}(X_{n+1} = k, X_{n+2} = j | X_n = i)$$

$$= \sum_{k \in S} \frac{\mathbf{P}(X_n = i, X_{n+1} = k, X_{n+2} = j)}{\mathbf{P}(X_n = i)}$$

$$= \sum_{k \in S} \frac{\mathbf{P}(X_n = i, X_{n+1} = k, X_{n+2} = j)}{\mathbf{P}(X_n = i, X_{n+1} = k)} \cdot \frac{\mathbf{P}(X_n = i, X_{n+1} = k)}{\mathbf{P}(X_n = i)}$$

$$= \sum_{k \in S} \mathbf{P}(X_{n+2} = j | X_n = i, X_{n+1} = k) \cdot \mathbf{P}(X_{n+1} = k | X_n = i)$$

$$= \sum_{k \in S} \mathbf{P}(X_{n+2} = j | X_{n+1} = k) \cdot \mathbf{P}(X_{n+1} = k | X_n = i) = \sum_{k \in S} p_{kj} \cdot p_{ik},$$

which is the entry at position (i, j) in the matrix $P^2 = P \cdot P$.

m-step transition probabilities:

In general: the m-step transition probability $p_{ij}^{(m)} := \mathbf{P}(X_{n+m} = j | X_n = i)$ is given as the entry at position (i, j) in the matrix P^m (m-th matrix power of P).

In other words: if we denote by $P^{(m)}$ the matrix of the m-step transition probabilities, then

$$P^{(m)} = \underbrace{P \cdot P \cdots P}_{m \text{ times}} =: P^m.$$

Initial distribution

Typically, in a Markov chain, the first state X_0 is also *random*, and drawn from some **initial probability distribution**.

Let $\lambda_j = \mathbf{P}(X_0 = j)$ for $j \in S$.

E.g. $S = \{a, c, g, t\}$, then the initial probability distribution can be represented as a *row vector*

$$\vec{\lambda} = \big(\lambda_a, \lambda_c,\ \lambda_g, \lambda_t\big).$$

The probabilities for the states at time 1 are

$$\big(\mathbf{P}_{\vec{\lambda}}(X_1 = a),\ \mathbf{P}_{\vec{\lambda}}(X_1 = c),\ \mathbf{P}_{\vec{\lambda}}(X_1 = g),\ \mathbf{P}_{\vec{\lambda}}(X_1 = t)\big) = \vec{\lambda} \cdot P.$$

P and $\vec{\lambda}$ completely determine the distribution of the MC

Given the transition matrix P and the initial distribution $\vec{\lambda}$ of a Markov chain with state space S, we can compute any probability we like for that chain.

E.g. for any $i_0, i_1, \ldots, i_n \in S$

$$\mathbf{P}_{\vec{\lambda}}(X_0 = i_0, X_1 = i_1, \ldots X_n = i_n) = \lambda_{i_0} \cdot p_{i_0 i_1} \cdots p_{i_{n-1} i_n},$$

or, if we chose for the sake of simplicity again $S = \{a, c, g, t\}$,

$$\Big(\mathbf{P}_{\vec{\lambda}}(X_n = a),\ \mathbf{P}_{\vec{\lambda}}(X_n = c),\ \mathbf{P}_{\vec{\lambda}}(X_n = g),\ \mathbf{P}_{\vec{\lambda}}(X_n = t)\Big) = \vec{\lambda} \cdot P^n.$$

The MC gradually forgets about its initial distribution...

Usually, the distribution of X_n changes over time, but it stabilizes as n gets bigger and bigger:

$$\lambda_i = \mathbf{P}_{\vec{\lambda}}(X_0 = i) \neq \mathbf{P}_{\vec{\lambda}}(X_1 = i) \neq \mathbf{P}_{\vec{\lambda}}(X_2 = i) \neq \ldots$$
$$\ldots \mathbf{P}_{\vec{\lambda}}(X_N = i) \approx \mathbf{P}_{\vec{\lambda}}(X_{N+1} = i) \approx \ldots$$

for N 'large enough'.

This phenomenon can be thought of as *'the process remembers how it was started, but it gradually forgets about its initial distribution'.*

Stationary distribution, overview:

Mathematically, there is a probability distribution $\vec{\pi}$, such that $\mathbf{P}_{\vec{\lambda}}(X_n = i)$ 'converges' (sometimes in a broader sense of the word) towards π_i as $n \to \infty$.

Furthermore, if we take $\vec{\pi}$ as the initial distribution, then the distribution of X_n is equal to $\vec{\pi}$ for *any* n.

For this reason $\vec{\pi}$ is called the **stationary distribution** of the Markov chain $X_0, X_1, X_2, \ldots$

Stationary distribution

Let $\vec{\pi} = (\pi_a, \pi_c, \pi_g, \pi_t)$ be a probability distribution such that

$$\vec{\pi} = \vec{\pi} \cdot P$$

holds.

If this holds, the probability vector $\vec{\pi}$ is the **stationary distribution** (sometimes called **equilibrium distribution**) for the Markov chain.

In our situations (finite state space, 'reasonable' MC) there is always a unique stationary distribution.

Stationarity

Let $X_0, X_1, \ldots$ be a Markov chain with the initial distribution being the stationary distribution, i.e.

$$\mathbf{P}(X_0 = i) = \pi_i$$

for $i \in S$, where S is the state space of the chain.

Then the chain is *stationary* (or *'in equilibrium'*):

$$\mathbf{P}(X_n = i) = \pi_i$$

for all $i \in S$ and for all $n \geq 1$.

In other words: X_n has the same distribution for every $n \geq 0$. It is thus much easier to compute probabilities ...

Convergence to the equilibrium

Let $X_0, X_1, \ldots$ be a Markov chain with finite state space and arbitrary initial distribution $\vec{\lambda}$. In virtually all practical cases (namely if the MC is 'irreducible' and 'aperiodic'):

$$\mathbf{P}_{\vec{\lambda}}(X_n = i) \to \pi_i \quad \text{as } n \to \infty.$$

'irreducible' means: the MC never gets stuck in one part of the state space, i.e. departing from any state there is always a positive probability for the MC to reach any other state (within one *or* several time steps).

'aperiodic' means: there is no strict periodicity in the MC for returning to a fixed state, e.g. a MC on $S = \{a, c, g, t\}$ that can only reach a at every second point in time has period 2, and hence is not aperiodic.

Reversibility

Suppose that $X_0, X_1, X_2, \ldots$ is a Markov chain (irreducible, aperiodic, with finite state space) with stationary distribution $\vec{\pi}$.

If

$$\pi_i \cdot p_{ij} = \pi_j \cdot p_{ji}$$

for any two states i and j, then the Markov chain is **reversible**.

By far not all Markov chains are reversible!

***

Reversibility is often desired in phylogenetic analyses (in computing the evolutionary distance between two sequences).

‘Detailed’ versus ‘full’ balance

Compare the condition for reversibility

$$\pi_i \cdot p_{ij} = \pi_j \cdot p_{ji} \quad \text{for all } i, j \in S \tag{1}$$

with the condition for the stationary distribution, $\vec{\pi} = \vec{\pi} \cdot P$, or in ‘sum notation’

$$\sum_{j \in S} \pi_i \cdot p_{ij} = \pi_i = \sum_{j \in S} \pi_j \cdot p_{ji} \quad \text{for all } i \in S. \tag{2}$$

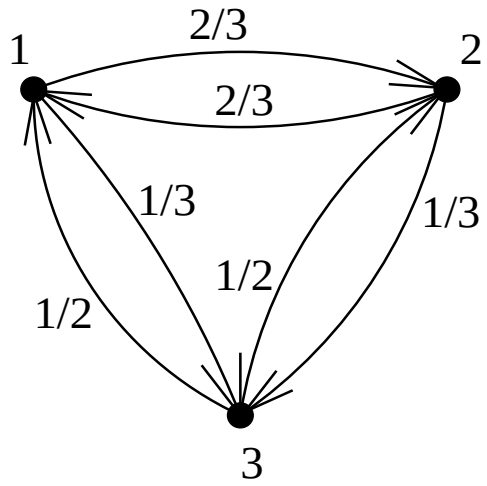
(1) is often referred to as *detailed balance condition (DBC)*, (2) as *full balance condition (FBC)*.

The DBC is much stronger, but usually also much easier to solve than the FBC. It is often a good shortcut for computing the stationary distribution of a reversible Markov chain.

Example: Let $X_0, X_1, \ldots$ be a Markov chain with state space $S := \{1, 2, 3\}$ and transition matrix

$$P := \begin{pmatrix} 0 & 2/3 & 1/3 \\ 2/3 & 0 & 1/3 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

This situation can be depicted in the form of a *transition graph*:



We try to solve the detailed balance condition $\pi_i \cdot p_{ij} = \pi_j \cdot p_{ji}$ for all $i, j \in S$. Set $\pi_1 := 1$. Then, because of the DBC,

$$\pi_2 = \frac{p_{12}}{p_{21}} \cdot \pi_1 = \frac{2/3}{2/3} \cdot 1 = 1,$$

and

$$\pi_3 = \frac{p_{13}}{p_{31}} \cdot \pi_1 = \frac{1/3}{1/2} \cdot 1 = \frac{2}{3}.$$

However, $\vec{\pi} = (\pi_1, \pi_2, \pi_3) = (1, 1, 2/3)$ is not yet a probability distribution, because $\sum_{i=1}^{3} \pi_i = 8/3 \neq 1$. But if we multiply everything with $3/8$, we get

$$\vec{\pi} = \left(\tfrac{3}{8}, \tfrac{3}{8}, \tfrac{2}{8}\right) = (0.375, 0.375, 0.25),$$

a 'probability vector' which satisfies the DBC!

$\big(\ldots \vec{\pi} = (0.375, 0.375, 0.25)\big)$

Thus, if the MC is started with this $\vec{\pi}$, we have for *any* $n \geq 0$ that

$$\big(\mathbf{P}_{\vec{\pi}}(X_n = 1), \mathbf{P}_{\vec{\pi}}(X_n = 2), \mathbf{P}_{\vec{\pi}}(X_n = 3)\big) = (0.375, 0.375, 0.25).$$

(Stationarity)

Furthermore, if the MC is started with an arbitrary initial distribution $\vec{\lambda}$, we have for n large that

$$\big(\mathbf{P}_{\vec{\lambda}}(X_n = 1), \mathbf{P}_{\vec{\lambda}}(X_n = 2), \mathbf{P}_{\vec{\lambda}}(X_n = 3)\big) \approx (0.375, 0.375, 0.25).$$

(Convergence to the equilibrium)

Higher-order Markov chains

Remember: Markov chain model for DNA

$$X_i := \textit{nucleotide at position } i \textit{ in a DNA sequence of length } N.$$

We have seen: $X_1, X_2, \ldots, X_N$ i.i.d. is unrealistic.

Better model: $X_1, X_2, \ldots, X_N$ is Markov chain (probabilities for the nucleotides at position $n+1$ depend only on the nucleotide present at the preceding position n). But in real data the dependence is usually more complex!

Better still: **Higher-order Markov chain**. (Markov chain of order k: probabilities for the nucleotides at position $n+k$ depend only on the nucleotides present at the k preceding positions $n, n+1, \ldots, n+k-1$.)

Markov chain of order k

We can write the transition probabilities of a Markov chain of order k as

$$p_{i_0 i_1 \ldots i_{k-1}; i_k} := \mathbf{P}\big(X_k = i_k | X_0 = i_0, X_1 = i_1, \ldots, X_{k-1} = i_{k-1}\big)$$

$$= \mathbf{P}\big(X_{n+k} = i_k | X_n = i_0, X_{n+1} = i_1, \ldots, X_{n+k-1} = i_{k-1}\big)$$

$$= \mathbf{P}\big(X_{n+k} = i_k | X_0 = l_0, X_1 = l_1, \ldots X_{n-1} = l_{n-1}, X_n = i_0, X_{n+1} = i_1, \ldots, X_{n+k-1} = i_{k-1}\big)$$

for any $n \geq 1$ and for any states $i_0, \ldots, i_k; l_0, \ldots, l_{n-1} \in S$.

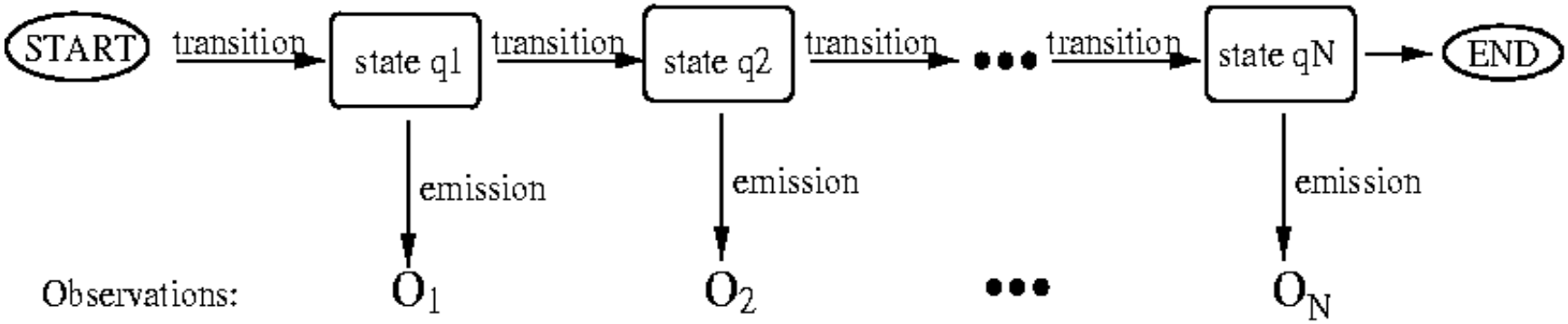
(The second equality holds because the higher-order MC is 'time-homogeneous', the third one holds because it is 'higher-order Markovian'.)

Higher-order MC vs. first order MC

Main advantage of higher-order MC:
Yield generally a better approximation of the dependence in real data. (It has been found that even DNA in nonfunctional intergenic regions tends to have high-order dependence!)

Main disadvantage:
Computationally expensive: the transition 'matrix' for an order k Markov chain model of DNA has $3 \cdot 4^k$ 'degrees of freedom', a similar model for a protein even $19 \cdot 20^k$. That means that huge amounts of data would be required for making 'reasonable' statistical statements about such a MC (e.g. for estimating the transition 'matrix'; see Lecture 4).

Hidden Markov models (HMM):



Sequence of states (= the 'path'): $\vec{q} = q_1\, q_2 \cdots q_N$, where $q_i \in S$ for $i = 1, 2, \ldots, N$, and where S is the set of possible states.
The path is usually not observable (i.e. it is *hidden*).

Sequence of observations (or 'symbols'): $\vec{o} = o_1\, o_2 \cdots o_N$, where $o_i \in A$ for $i = 1, 2, \ldots, N$, and where A is the set of possible symbols (the 'alphabet').

HMM's are important in bioinformatics, e.g. for *multiple alignments, gene finding, more.*

***

HMM's are more flexible (more realistic) than standard Markov chains, but still *mathematically feasible.*

Hidden Markov models (HMM):

- The sequence of states $\vec{q} = q_1\, q_2 \cdots q_N$ is a (standard) *Markov chain*, with transition probabilities

$$p_{kl} = \mathbf{P}(q_i = l | q_{i-1} = k),$$

where $k, l \in S$.

- In each state, a 'symbol' is emitted, symbol b with probability

$$e_k(b) = \mathbf{P}(o_i = b | q_i = k),$$

depending on in which state k the Markov chain is
(where $b \in A$, $k \in S$).

The joint probability of having a given sequence $\vec{o} = o_1\, o_2 \cdots o_N$ and a given path $\vec{q} = q_1\, q_2 \cdots q_N$ is

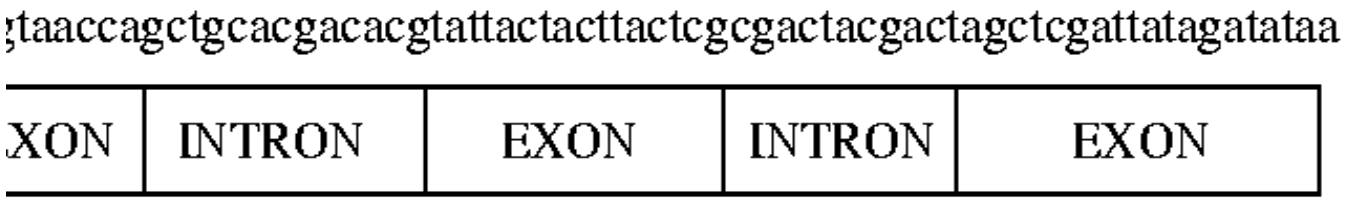
$$\mathbf{P}(\vec{o}, \vec{q}) = p_{s,q_1} \cdot e_{q_1}(o_1) \cdot p_{q_1,q_2} \cdot e_{q_2}(o_2) \cdots p_{q_{N-1},q_N} \cdot e_{q_N}(o_N) \cdot p_{q_N,f}$$

where $s =$ the start state, and $f =$ the end state.

Example of HMM application: model for nucleotide patterns in a gene.

States $S = \{0, 1\}$ where $0 =$ intron and $1 =$ exon.

Alphabet $A = \{a, c, g, t\}$.



If the state $q_i = 0$ (INTRON), emit a symbol o_i according to the probability distribution $(e_0(a), e_0(c), e_0(g), e_0(t))$.

If the state $q_i = 1$ (EXON), emit a symbol o_i according to the probability distribution $(e_1(a), e_1(c), e_1(g), e_1(t))$.

'Decoding' (finding the path)

In HMM applications the following problem often occurs:

Given the observations $\vec{o} = o_1\, o_2 \cdots o_N$ *(and given all the model parameters),* **find the path** $\vec{q} = q_1\, q_2 \cdots q_N$ *that is most likely to have generated the observations.*

For example, given the observed sequence
`ccgtactagctgtagctgtgac...atcggggctgctgccatcgatcgacgtagc`,
where are the introns and exons?

Most probable path

The *most probable path* $\vec{q}^*$ (given the observations $\vec{o}$) is

$$\vec{q}^* = \text{argmax}_{\vec{q}}\, \mathbf{P}(\,\vec{q}\,|\,\vec{o}\,).$$

Even with a moderate number of possible states, *the number of possible paths is huge*.

Therefore, calculation by exhaustive methods (testing all possible paths) becomes impossible in many cases.

But the most probable path $\vec{q}^*$ can be found recursively using a *dynamic programming algorithm*, the Viterbi algorithm (see also Lectures 6 and 7 on HMM's).

Most probable path: the Viterbi algorithm

Given: observations $o_1, o_2, \ldots, o_N$.

Define

$$\delta_n(k) = \max_{q_1, q_2, \ldots, q_{n-1}} \mathbf{P}(q_1, q_2, \ldots, q_{n-1}, q_n = k, o_1, o_2, \ldots, o_n)$$

for $k \in S$, $1 \leq n \leq N$.

INITIATION: $\delta_1(k) = p_{sk} \cdot e_k(o_1)$ for $k \in S$ ($s =$ 'start').

INDUCTION: $\delta_n(j) = \max_{i \in S} \delta_{n-1}(i) \cdot p_{ij} \cdot e_j(o_n)$ for $j \in S$, for $2 \leq n \leq N$.

BACKTRACKING: Put $q_N^* := \operatorname{argmax}_{j \in S} \delta_N(j)$
and $q_{n-1}^* := \operatorname{argmax}_{j \in S} \delta_{n-1}(j) \cdot p_{j,q_n^*}$ for $n = N, N-1, \ldots, 2$.

RESULT: Most probable path $q_1^* \, q_2^* \, \cdots \, q_N^*$.