Statistical Computing

Hidden Markov Models for Bioinformatics

- Part II -

Uwe Menzel, 2011

uwe.menzel@matstat.org

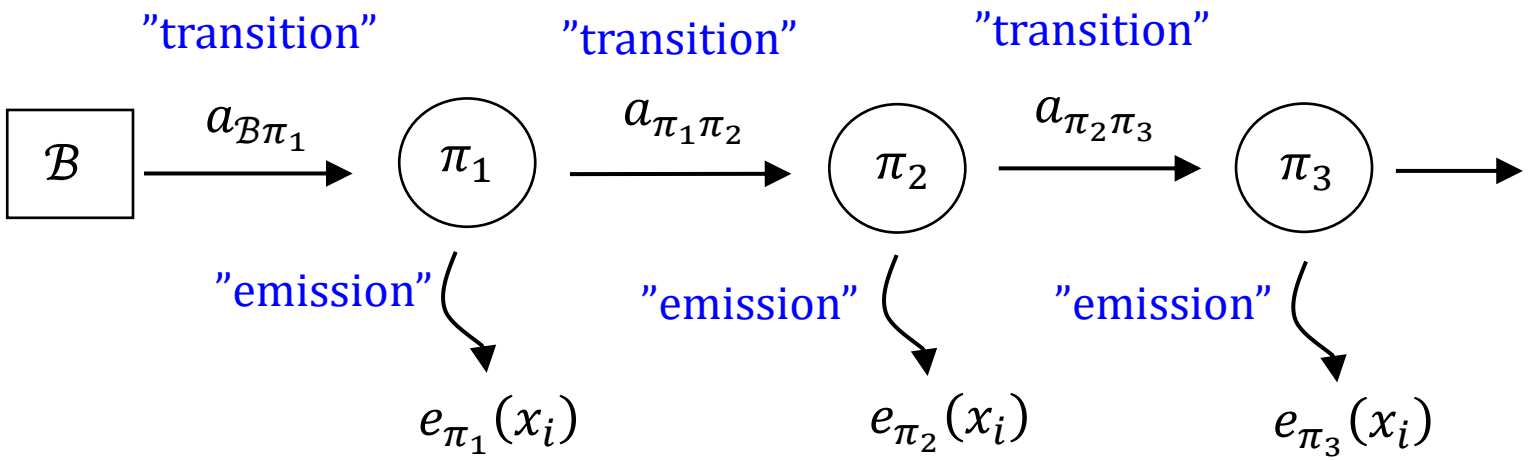
www.matstat.org

Contents Part II

The Hidden Markov Model: transition and emission probabilities

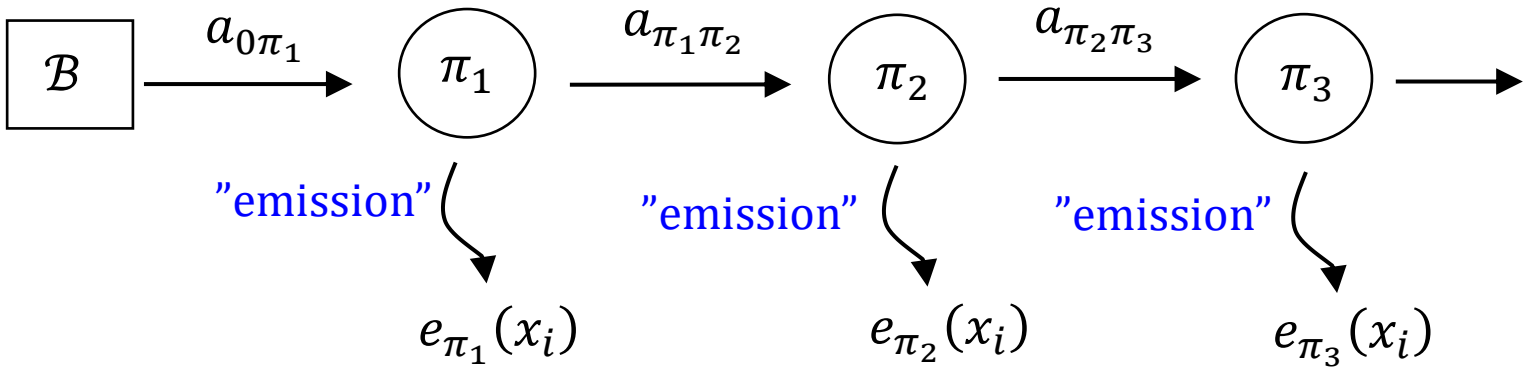
Hidden Markov Model

Now, we assume that our Markov chain is not visible for the observer (hidden), but that every state of the Markov chain can "emit" an element of a set of observable characters (symbols) with some probability. The Markov chain itself forms the state path π_i, and the emission probabilities are labeled $e_{\pi_k}(x_i)$, defining the probability that the state π_k emits the symbol x_i. For DNA, we have $x_i = \{A, C, G, T\}$. The HMM model can be illustrated like that:



Note: transitions from the (virtual) begin state will be denoted $a_{\mathcal{B}\pi_i}$ or $a_{0\pi_i}$.

Hidden Markov Model



Every state emits a symbol with a certain probability. The symbols can be observed, but not the emitting states. The jont probability of a chain of states **and** symbols is then:

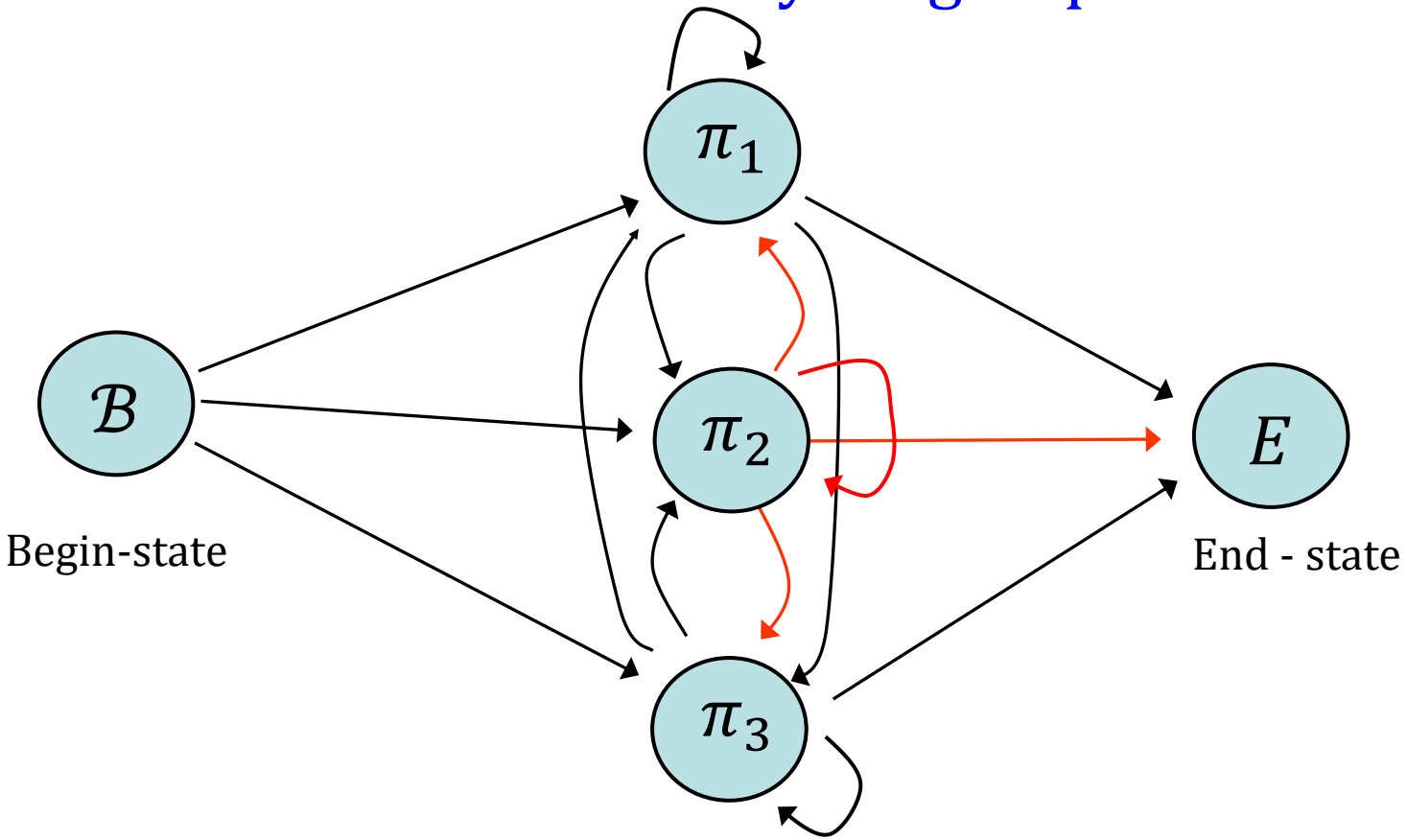
$$P(x,\pi) = a_{0\pi_1} \cdot e_{\pi_1}(x_1) \cdot a_{\pi_1\pi_2} \cdot e_{\pi_2}(x_2) \cdot a_{\pi_2\pi_3} \cdot \ldots$$

$$P(x,\pi) = a_{0\pi_1} \cdot \prod_{i=1}^{L} e_{\pi_i}(x_i) \cdot a_{\pi_i\pi_{i+1}}$$

$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$ transition probability $k \rightarrow l$ (within state path)

$e_k(b) = P(x_i = b \mid \pi_i = k)$ emission probabilities, from state k to symbol b

HMM for an arbitrary long sequence

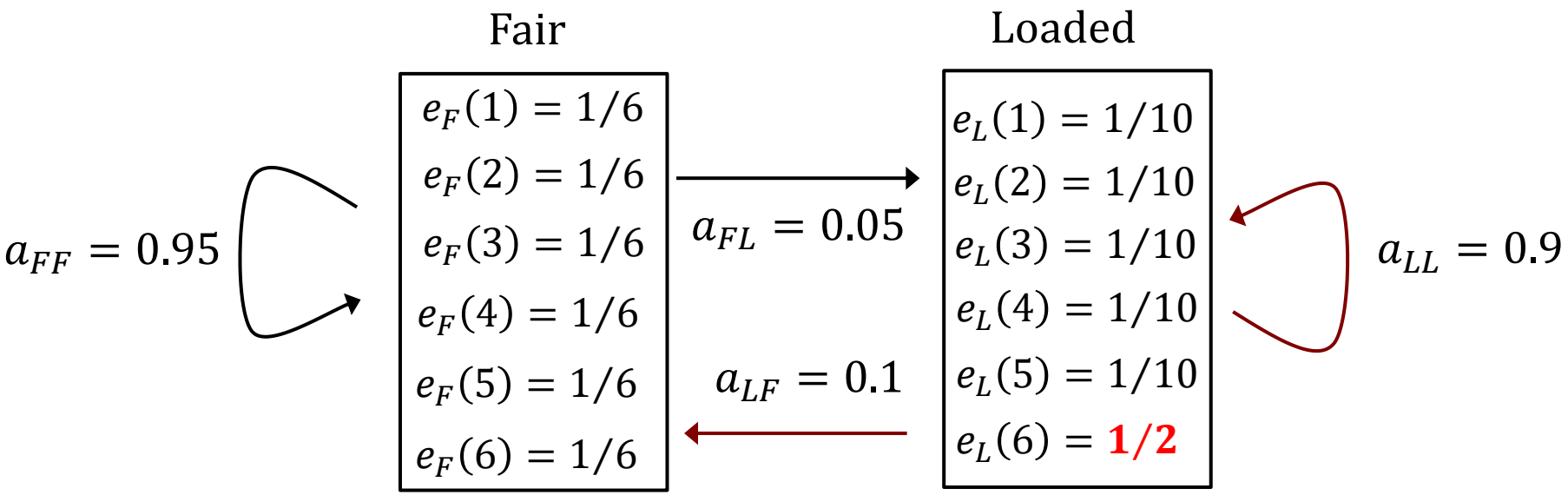


The 3 states can be traversed arbitrarily many times, emitting a symbol after each transition. The outgoing transition probabilities from each state must sum up to 1, e.g. (for state 2, in red):

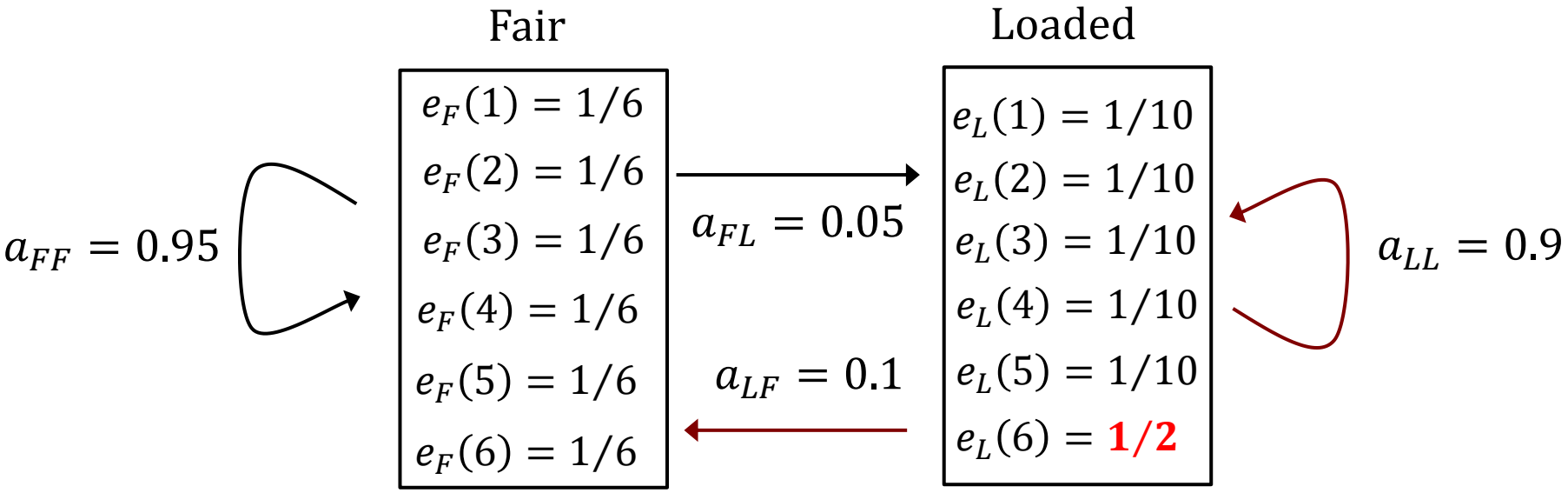
$$a_{\pi_2\pi_1} + a_{\pi_2\pi_3} + a_{\pi_2 E} = 1$$

HMM: Dishonest casino

In order to be able to cheat, the croupier uses two dice: a fair one and a loaded one. The loaded die has a higher probability to roll a six. The game can be modelled by a HMM: The two states are "Fair" and "Loaded". The guest of the casino cannot see which die is used, i.e. the states are hidden. Both states have different emission probabilities for the numbers 1 to 6 (see figure, the loaded die "emits" a six with probability 0.5). After every roll of the fair die, the probability that the croupier uses the fair die once more is 0.95 (a_{FF}). However, with probability 0.05 (a_{FL}), he switches to the loaded die (he won't do that too often to limit the risk of detection).



HMM: Dishonest casino



Examples:

- $e_F(1)$: probability that the fair (F) die rolls ("emits") a 1
- $e_L(2)$: probability that the loaded (L) die emits a 2
- a_{FF} : probability of transition Fair → Fair (after each roll)
- a_{FL} : probability of transition Fair → Loaded

The observer sees only the emissions : 3 4 2 4 6 4 6 3 4 6 6 3 6 6 3 4 6 6
The state path is hidden for the observer: F F F F F F F F L L L L L L L L L L

HMM for the recognition of CpG islands embedded in genomic DNA

States:

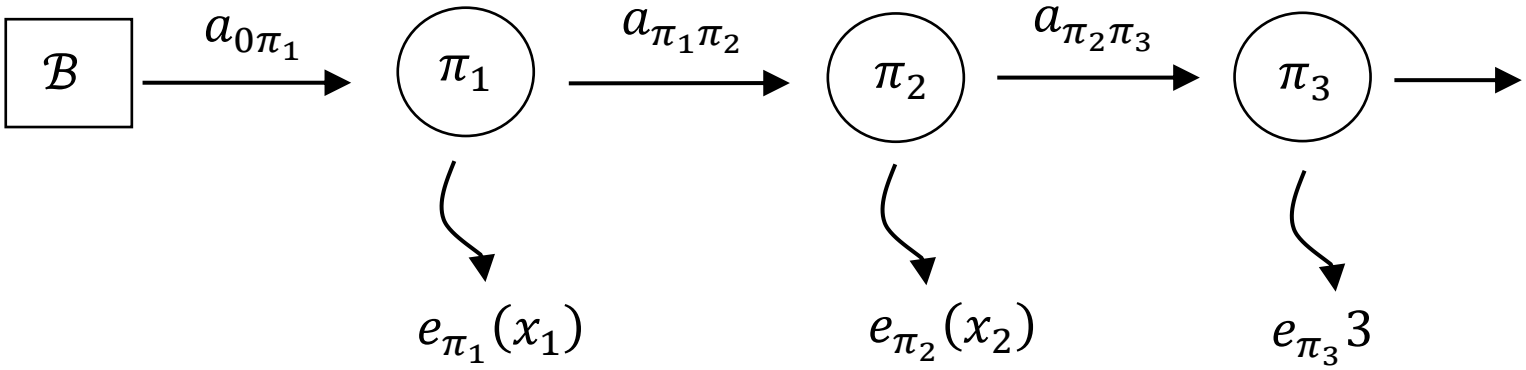
- A^+, C^+, G^+, T^+ (DNA in CpG island),
- A^-, C^-, G^-, T^- (not in island)

Symbols:

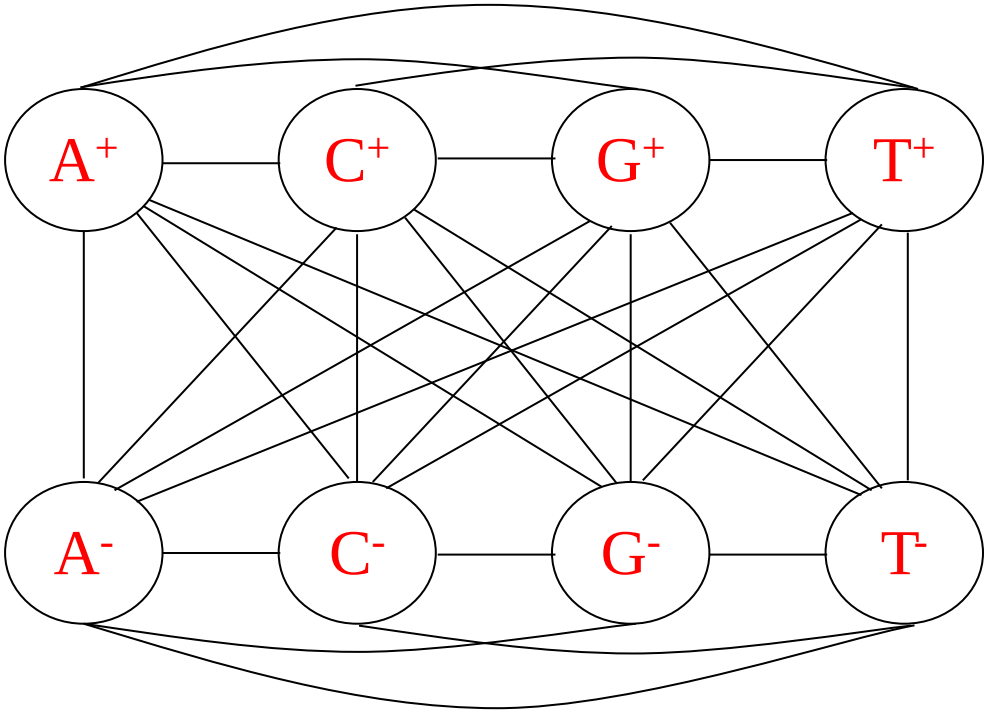
- A, C, G, T (not disclosing from which state they emerge)

Example:

- state path: $A^+, C^+, G^+, T^+, A^+, C^-, G^-, G^-, G^-, T^-$
- observation: $A,\ C,\ G,\ T,\ A,\ C,\ G,\ G,\ G,\ T$



HMM for the detection of CpG islands in long DNA- sequences



Picture: Sven Schuirer

To get the HMM model working, we need the transition and emission probabilities →

HMM for the recognition of CpG islands: Transition probabilities

From the training sets, we know the transition probabilities between adjacent nucleotides in CpG islands and in non-islands (see above):

a_{kl}	A^+	C^+	G^+	T^+	A^-	C^-	G^-	T^-
A^+	0.180	0.274	0.426	0.120	?	?	?	?
C^+	0.171	0.368	0.274	0.188	?	?	?	?
G^+	0.161	0.339	0.375	0.125	?	?	?	?
T^+	0.079	0.355	0.384	0.182	?	?	?	?
A^-	?	?	?	?	0.300	0.205	0.285	0.210
C^-	?	?	?	?	0.322	0.298	**0.078**	0.302
G^-	?	?	?	?	0.248	0.246	0.298	0.208
T^-	?	?	?	?	0.177	0.239	0.292	0.292

A simple method to connect all possible states can be found by assuming that the probability to remain in a CpG island is p, and the probability of remaining in a non-island is q →

HMM for the recognition of CpG islands: Transition probabilities

- Probability to remain in a CpG island: p
- Probability to remain in a non-island: q
- The following table of the a_{kl} ensures that every row sum is 1:

a_{kl}	A^+	C^+	G^+	T^+	A^-	C^-	G^-	T^-
A^+	$0.180 \cdot p$	$0.274 \cdot p$	$0.426 \cdot p$	$0.120 \cdot p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
C^+	$0.171 \cdot p$	$0.368 \cdot p$	$0.274 \cdot p$	$0.188 \cdot p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
G^+	$0.161 \cdot p$	$0.339 \cdot p$	$0.375 \cdot p$	$0.125 \cdot p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
T^+	$0.079 \cdot p$	$0.355 \cdot p$	$0.384 \cdot p$	$0.182 \cdot p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
A^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.300 \cdot q$	$0.205 \cdot q$	$0.285 \cdot q$	$0.210 \cdot q$
C^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.322 \cdot q$	$0.298 \cdot q$	$0.078 \cdot q$	$0.302 \cdot q$
G^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.248 \cdot q$	$0.246 \cdot q$	$0.298 \cdot q$	$0.208 \cdot q$
T^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.177 \cdot q$	$0.239 \cdot q$	$0.292 \cdot q$	$0.292 \cdot q$

HMM for the recognition of CpG islands: Transition probabilities

- Probability to remain in a CpG island: $p = 0.95$
- Probability to remain in a non-island: $q = 0.99$
- The following table of the a_{kl} ensures that every row sum is 1:

a_{kl}	A^+	C^+	G^+	T^+	A^-	C^-	G^-	T^-
A^+	0.171	0.26	0.405	0.114	0.0125	0.0125	0.0125	0.0125
C^+	0.162	0.35	0.26	0.179	0.0125	0.0125	0.0125	0.0125
G^+	0.153	0.322	0.356	0.119	0.0125	0.0125	0.0125	0.0125
T^+	0.075	0.337	0.365	0.173	0.0125	0.0125	0.0125	0.0125
A^-	0.0025	0.0025	0.0025	0.0025	0.297	0.203	0.282	0.208
C^-	0.0025	0.0025	0.0025	0.0025	0.319	0.295	0.077	0.299
G^-	0.0025	0.0025	0.0025	0.0025	0.245	0.244	0.295	0.206
T^-	0.0025	0.0025	0.0025	0.0025	0.168	0.225	0.289	0.289

HMM for the recognition of CpG islands: Transition probabilities

The transition probability from the begin state $\mathcal{B}$ (sometimes labelled 0) to some other state is assumed to be connected to the overall frequency of the corresponding symbol in the investigated genomic DNA. The freuqency value is shared equally between the "+" and the "-" state. A possible assigment could be:

$$f(A) = \mathbf{0.24}$$
$$a_{\mathcal{B}A^+} = f(A) = 0.12$$
$$a_{\mathcal{B}A^-} = f(A) = 0.12$$

$$f(C) = \mathbf{0.26}$$
$$a_{\mathcal{B}C^+} = f(C) = 0.13$$
$$a_{\mathcal{B}C^-} = f(C) = 0.13$$

$$f(G) = \mathbf{0.26}$$
$$a_{\mathcal{B}G^+} = f(G) = 0.13$$
$$a_{\mathcal{B}G^-} = f(G) = 0.13$$

$$f(T) = \mathbf{0.24}$$
$$a_{\mathcal{B}T^+} = f(T) = 0.12$$
$$a_{\mathcal{B}T^-} = f(T) = 0.12$$

$$\sum_k a_{\mathcal{B}k} = 1$$

HMM for the recognition of CpG islands:
Transition probabilities

- Probability to remain in a CpG island: $p = 0.95$
- Probability to remain in a non-island: $q = 0.99$

a_{kl}	$\mathcal{B}$	A^+	C^+	G^+	T^+	A^-	C^-	G^-	T^-
$\mathcal{B}$	0	0.12	0.13	0.13	0.12	0.12	0.13	0.13	0.12
A^+	0	0.171	0.26	0.405	0.114	0.0125	0.0125	0.0125	0.0125
C^+	0	0.162	0.35	0.26	0.179	0.0125	0.0125	0.0125	0.0125
G^+	0	0.153	0.322	0.356	0.119	0.0125	0.0125	0.0125	0.0125
T^+	0	0.075	0.337	0.365	0.173	0.0125	0.0125	0.0125	0.0125
A^-	0	0.0025	0.0025	0.0025	0.0025	0.297	0.203	0.282	0.208
C^-	0	0.0025	0.0025	0.0025	0.0025	0.319	0.295	0.077	0.299
G^-	0	0.0025	0.0025	0.0025	0.0025	0.245	0.244	0.295	0.206
T^-	0	0.0025	0.0025	0.0025	0.0025	0.168	0.225	0.289	0.289

HMM for the recognition of CpG islands: Emission probabilities

$e_k(b)$: probability for seeing symbol b when the underlying state is k

$$e_{C^+}(C) = 1 \qquad e_k(b) = P(x_i = b \mid \pi_i = k)$$

emitting state — symbol that is emitted

$e_{C^+}(C) = 1$; $e_{C^-}(C) = 1$; $e_{\pi_i}(C) = 0$ otherwise
$e_{A^+}(A) = 1$; $e_{A^-}(A) = 1$; $e_{\pi_i}(A) = 0$ otherwise
$e_{G^+}(G) = 1$; $e_{G^-}(G) = 1$; $e_{\pi_i}(G) = 0$ otherwise
$e_{T^+}(T) = 1$; $e_{T^-}(T) = 1$; $e_{\pi_i}(T) = 0$ otherwise

The states A^+, A^- can only emit the symbol A.
The states C^+, C^- can only emit the symbol C.
The states G^+, G^- can only emit the symbol G.
The states T^+, T^- can only emit the symbol T.

The state C^+ means that we have a cytosin from inside of a CpG island; while the state C^- means that we have a cytosin from other genomic regions.

HMM for the recognition of CpG islands: Emission probabilities

$e_k(b)$: probability for seeing symbol b when the underlying state is k

$$e_{C^+}(C) = 1$$

emitting state — symbol that is emitted

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

$e_k(b)$	A	C	G	T
$\mathcal{B}$	0.25	0.25	0.25	0.25
A^+	1	0	0	0
C^+	0	1	0	0
G^+	0	0	1	0
T^+	0	0	0	1
A^-	1	0	0	0
C^-	0	1	0	0
G^-	0	0	1	0
T^-	0	0	0	1