Statistical Computing

Hidden Markov Models for Bioinformatics

- Part I -

Uwe Menzel, 2011

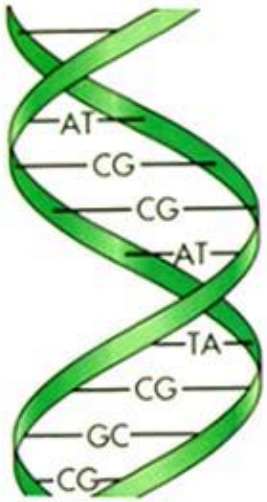
uwe.menzel@matstat.org

www.matstat.org

Contents Part I

DNA-Sequence

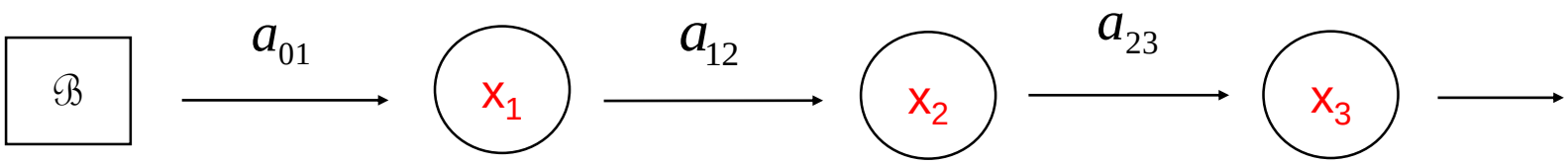
The sequence of the bases (A,T,G,C) in the DNA-molecule determines the blueprint of an organism.



What is a Markov chain and what has it to do with DNA?

Андрей Андреевич Марков (1856 – 1922)

Markov chain



Model:

- a sequence is generated by a random process

Alphabet:

- set of characters x_i building up the chain, e.g. $x_i = \{A, C, G, T\}$

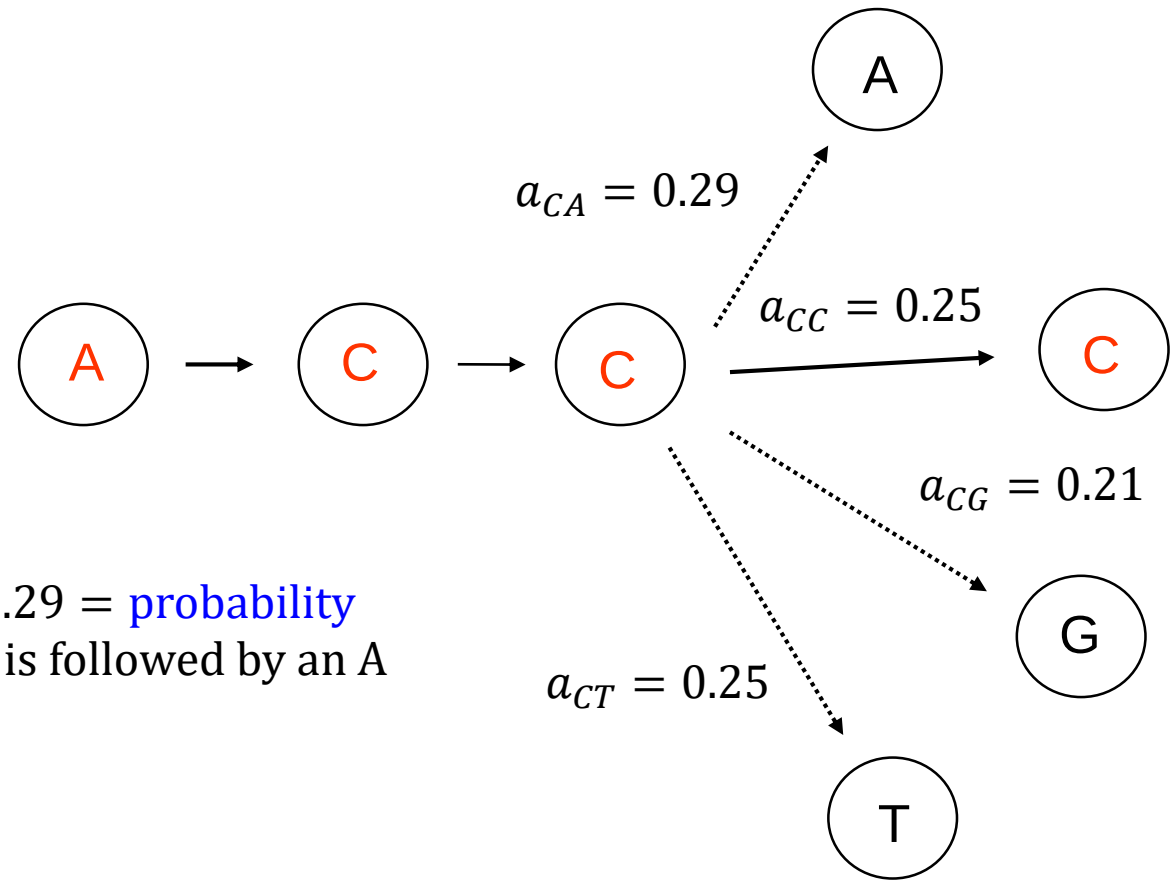
Markov property:

- the value x_{i+1} **only** depends on x_i, but not on $x_{i-1}, x_{i-2}, \ldots$.

Transition probability:

- $a_{st} = P(x_i = t \mid x_{i-1} = s)$

Markov-chain for DNA



$a_{CA} = 0.29 =$ probability
that a C is followed by an A

Drawing a long Markov chain (for DNA)

$\mathcal{B}$ = begin state
E = end state

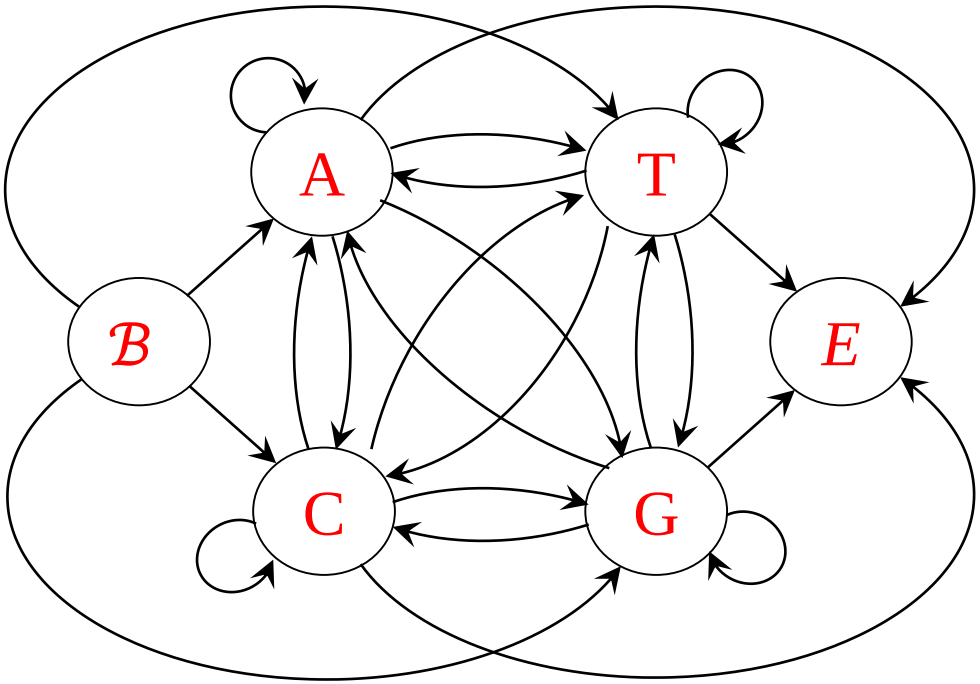


Abbildung: Sven Schuirer

Joint probability of a chain

$$P(\vec{x}) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, \dots, X_N = x_N)$$
$$= P(x_1, x_2, x_3, \dots, x_N)$$

$$P(\vec{x}) = P(X_1 = A, X_2 = C, X_3 = C, X_4 = G, X_5 = T)$$ for DNA
$$= P(A, C, C, G, T)$$

$$P(\vec{x}) = P(x_1, x_2, x_3, \dots, x_{N-1}, x_N)$$

Use multiple times: $P(x, y) = P(x \mid y) \cdot P(y)$

$$P(\vec{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2, x_1) \cdot P(x_4 \mid x_3, x_2, x_1) \cdot \dots$$

With the Markov property, this simplifies to:

$$P(\vec{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2) \cdot P(x_4 \mid x_3) \cdot \dots \cdot P(x_N \mid x_{n-1})$$

Probability of the Markov chain[1]

$$P(\vec{x}) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_2) \cdot P(x_4 \mid x_3) \cdot \ \ldots \cdot P(x_N \mid x_{n-1})$$

Let $a_{x_{i-1},x_i} = P(x_i \mid x_{i-1})$ transition probability

$$P(\vec{x}) = P(x_1) \cdot a_{x_1,x_2} \cdot a_{x_2,x_3} \cdot a_{x_3,x_4} \cdot \ \ldots \cdot a_{x_{N-1},x_N}$$

$$P(\vec{x}) = P(x_1) \cdot \prod_{i=2}^{N} a_{x_{i-1},x_i} \qquad \text{with } P(x_1) = a_{x_0,x_1}$$

x_0 is a virtual (begin) state, introduced to make the formula nicer

$$\boxed{P(\vec{x}) = \prod_{i=1}^{N} a_{x_{i-1},x_i}}$$

[1]considering homogeneous Markov-chains only

Maximum Likelihood (ML) estimators for the transition probabilities in DNA

- count the frequency of dinucleotides, c_{st}, in many genomic sequences
- normalize: divide by the sum of all outgoing transition probabilities
- a = transition probabilities; c = transition frequencies "counts")

$$\boxed{a_{st} = \frac{c_{st}}{\sum_i c_{si}} \qquad s, t \in \{A, C, G, T\}}$$

It follows that: $$\sum_t a_{st} = 1$$

The outgoing transition probabilities from each symbol sum up to 1.

For example, it holds that:

$$a_{CA} + a_{CC} + a_{CG} + a_{CT} = 1$$

Maximum Likelihood (ML) estimators for the transition probabilities in DNA

$$a_{st} = \frac{c_{st}}{\sum_i c_{si}} \qquad s,t \in \{A,C,G,T\}$$

$$c_{CG} = 100 \qquad c_{CA} = 150 \qquad c_{CT} = 50 \qquad c_{CC} = 100$$

$$a_{CG} = \frac{c_{CG}}{c_{CG} + c_{CA} + c_{CT} + c_{CC}} = \frac{100}{100+150+50+100} = 0{,}25$$

$$a_{CA} = \frac{150}{400} = 0{,}375$$

$$a_{CT} = \frac{50}{400} = 0{,}125$$

$$a_{CC} = \frac{100}{400} = 0{,}25$$

$$a_{CG} + a_{CA} + a_{CT} + a_{CC} = 1 \quad \textit{line total}$$

Matrix of transition probabilities

	A	C	G	T
A	0,300	0,205	0,285	0,210
C	0,322	0,298	0,078	0,302
G	0,248	0,246	0,298	0,208
T	0,177	0,239	0,292	0,292

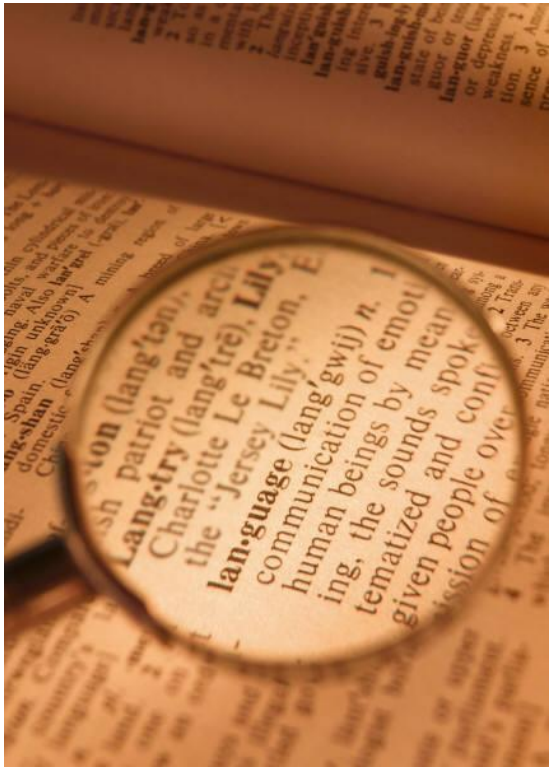
Stochastisc matrix

row sum = 1

$$P(C, A, A, G) = a_{0C} \cdot a_{CA} \cdot a_{AA} \cdot a_{AG}$$
$$= 0.25 \cdot 0.322 \cdot 0.3 \cdot 0.285 = \underline{\underline{0.00688}}$$

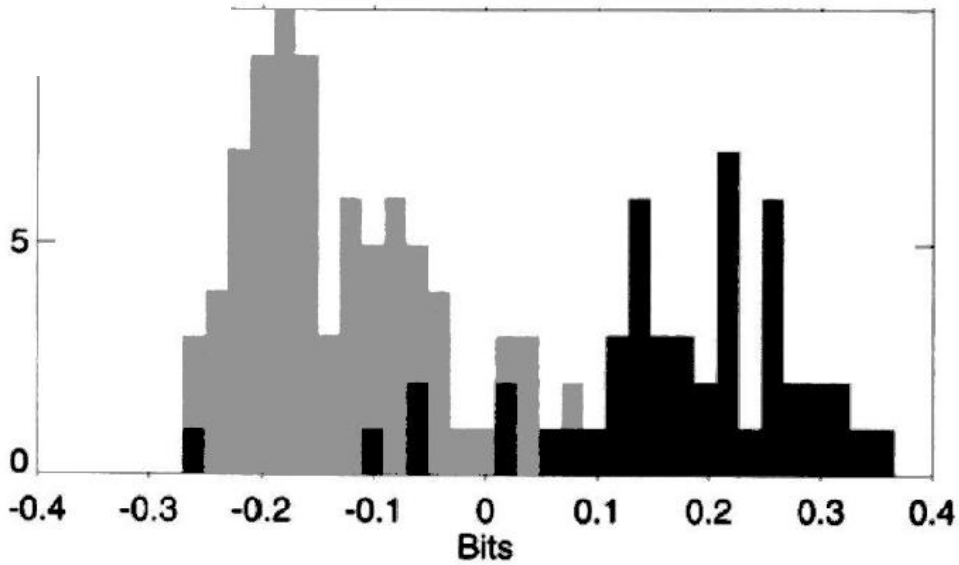
Probability of the chain C, A, A, G ; presupposed the transition probabilities presented in the table, and that the first nucleotide is a C with probability 0.25

Language recognition (Markov)



- languages differ by the frequencies of transitions between characters:
- for example, ”th” is quite frequent in English but not in Spanish

A Likelihood Ratio Test using Markov chains to determine whether a small piece of DNA is a CpG island or not



Picture from: **Durbin** et al. (Ed): Biological Sequence Analysis, Cambridge University Press, 1998

What is a CpG island ?

- What CpG frequency do we (approximately) expect ?
  - $P_{CG} \approx \frac{1}{4} \cdot \frac{1}{4} = 1/16$; more precisely $0{,}21 \cdot 0{,}21 \approx 4{,}4\%$
- actual frequency is only 0,8 % (mammalia)
- cytosine (C) in a CpG is chemically unstable:
  - methylation, deamination: $CG \rightarrow C^{meth}G \rightarrow TG$
- CpG-islands have a higher CpG percentage, compared to the rest of the genome

Exon 1 CpG Island: 12634..12767

```
 11941 ttataagatc cccctccctc taaatcctgt ccttctatca cttcatcctt CGctctcctt
 12001 taaaatgaga cagttgtcag caggaatcct gCGcaagaac acaccaccct gtttcataga
 12061 agatatctca ggtaatgtgc aaacaCGggt ttttaaaCGg agCGcatttt tctcatttgt
 12121 taatatcacc acctaaatca tctcttgcct aaaacaagga gtagaaagtg aatgaaggaa
 12181 ggaacaggtg atggtcagtg tcctttctac gcctcaaaat ttaagagttt atgtgaaaat
 12241 tcataaatat taatctcaat ccaggttaag caaaattttt tgctctcctc tttagaaatt
 12301 tctggttgcc aaagttccag aaattgcttc ctcattcctg agcctttcat tttctCGatt
 12361 tctccattat gtaaCGggga gctggagctt tgggcCGaat ttccaattaa agatgatttt
 12421 tacagtcaat gagccaCGtc agggagCGat ggcaccCGca ggCGgtatca actgatgcaa
 12481 gtgttcaagc gaatctcaac tCGtttttc CGgtgactca ttccCGgccc tgcttggcag
 12541 CGctgcaccc tttaacttaa acctCGgcCG gcCGccCGcc gggggcacag agtgtgCGcc
*12601 gggcCGCGCG gcaattggtc ccCGCGcCGa cctcCGccCG CGagCGcCGc CGcttccctt
*12661 cccCGcccCG CGtccctccc cctCGgcccc gCGCGtCGcc tgtcctcCGa gccagtCGct
*12721 gacagcCGCG gCGcCGCGag cttctcctct cctcaCGacc gaggcaggta aaCGccCGgg
 12781 gtgggaggaa CGCGggCGgg ggcaggggag cCGCGggggc CGagtgagga cccCGggcct
 12841 CGggtcccag gCGcaagggt gccCGgcCGg gCGgggtCGg gaccccagtg aggaggggcc
 12901 gggggctgcc cCGCGggCGc gtgaCGgtct CGggcctgcc CGgctgCGct ggtctcCGct
 12961 CGggtgaggc ggcttggctt CGcttttcag gttaggaaag ctccctttac tgCGCGttgg
 13021 ggggctgggg gagctggCGg agccaCGtta gggaggtCGg tggCGcCGgg gtgtctcagc
 13081 gccccctgca cccCGCGCGg gtcCGgccca gCGggCGatc gctggCGccc agggaactcc
 13141 gggagggcCG ccagCGggct cCGcaggCGc ggggCGggga ggggCGcctg ggggcCGCGg
 13201 ggctCGCGct cccCGccCGt tggcCGcccc tCGgaggcCG agatCGgggc ccagaaCGcc
 13261 ccttggcaaa gcctggCGct tcCGCGatgc ccagagggtg cttgggggga tggagagagg
 13321 ggCGccCGcc ggggtagttc CGggagcctc ggtgcctccc gcCGcagctg cagCGttcct
 13381 ccCGggaggc ggcccagccc ttcatcctCG cCGcctgagc ttctcCGagg ggggctgcag
 13441 ccttgCGgcc gttgccacCG cctggagaag CGgcccaCGc ggactgaCGg gCGggggCGg
 13501 ggcctCGggc ctCGgCGggg gCGgggtcCG gggaggcccc accctctgtt ctccaggggc
 13561 ggggagagag gagctgcagg tctgCGgcct ggccccaggt gCGatggCGg accccagctt
 13621 ggccagtcac attcctccca gtccccctgg agggagaaCG ctggccatgg ggggctccaa
 13681 ggaacaacca gcctCGgatg aCGacccttg ggtcacCGgt ctccccacct gtgCGgcagg
 13741 CGccttcaCG tttcattatt aaacaatggg gagaaatcca tgtttactgt ccttttagg
 13801 aatttttgc tcttctcttt gaggtggctg taggaaatag atttttttt taacctCGca
 13861 attccaccac ggtcacatcc atcctCGcca tCGcagagcc acagctctcc gtttttgttt
 13921 cctagcctcc agattctcac acaacacagt gcagtttcac tgctgtaatg atgaggatct
 13981 tcatggcCGc gttattttct tgttctgaga gcatcaCGgt ttaattagca gttccccata
 14041 tgatttgaag tgtttccCGt ttccttaggg aaaactcctg gtagaatagg attaaggatt
 14101 tttacaaata taattatcaa aaacatagga acagggaatt ggataaatat gttaaacttc
 14161 tggaaaaatc aacaaCGctc ttagatttgt agaagaaagg aaaaaatcac cagtggaaag
 14221 gagcaatttt acttacacaa acacagagaa ggtcttacag tgaaaaaaag ctaaccagta
```

”Training”: Finding transition probabilities for both CpG islands and non-islands

CpG-Islands

	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

model+

$$a^{+}_{CG} = 0.274$$

Non-Islands

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

model-

$$a^{-}_{CG} = 0.078$$

How to discriminate between CpG islands and non-islands

Observed sequence: X = (ATCGCGCGGC)

$$P(X \mid model+) = \prod_i a^+_{x_{i-1}x_i} = a^+_{0A} \cdot a^+_{AT} \cdot a^+_{TC} \cdot a^+_{CG} \cdot a^+_{GC} \cdot a^+_{CG} \cdot a^+_{GC} \cdot a^+_{CG} \cdot a^+_{GG} \cdot a^+_{GC}$$

$$= 0.25 \cdot 0.12 \cdot 0.355 \cdot 0.274 \cdot 0.339 \cdot 0.274 \cdot 0.339 \cdot 0.274 \cdot 0.375 \cdot 0.339$$

$$= 3.125 \cdot 10^{-6}$$ Probability of the chain under CpG island model

$$P(X \mid model-) = \prod_i a^-_{x_{i-1}x_i} = a^-_{0A} \cdot a^-_{AT} \cdot a^-_{TC} \cdot a^-_{CG} \cdot a^-_{GC} \cdot a^-_{CG} \cdot a^-_{GC} \cdot a^-_{CG} \cdot a^-_{GG} \cdot a^-_{GC}$$

$$= 0.25 \cdot 0.21 \cdot 0.239 \cdot 0.078 \cdot 0.246 \cdot 0.078 \cdot 0.246 \cdot 0.078 \cdot 0.298 \cdot 0.246$$

$$= 2.65 \cdot 10^{-8}$$ Probability of the chain under non-island model

Result: The probability of the chain is higher if we assume model "+" (CpG island model) → It is more likely that the sequence X is a CpG-Island.

Note that it might be better to carry out these calculations in log-space, to avoid underflow in computations. Products become sums in log-space making the calculations faster.

Likelihood Ratio Test for Discrimination of CpG islands and non-islands

According to the model, a sequence X is a CpG-island if:

$$P(X \mid mod +) > P(X \mid mod -)$$

$$\frac{P(X \mid mod +)}{P(X \mid mod -)} > 1$$

log likelihood ratios

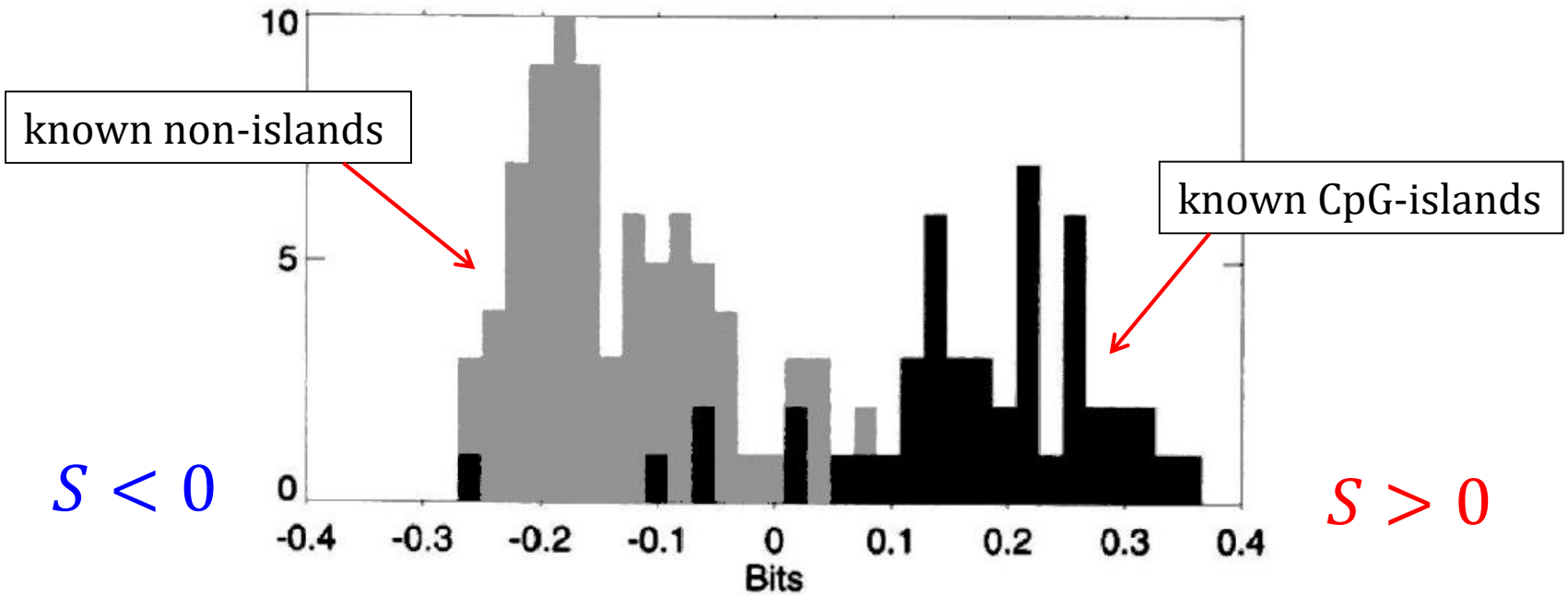
$$S = \log\left[\frac{P(X \mid mod +)}{P(X \mid mod -)}\right] = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i} > 0$$

$S =$ log odds score

If the log odds score S is bigger than 0, it is more likely that the sequence probed is a CpG island. If the log odds score is negative, it is more likely that the sequence does not emerge from a CpG island.

Does the method really work?

Calculated scores for many training sets (islands and non-islands):



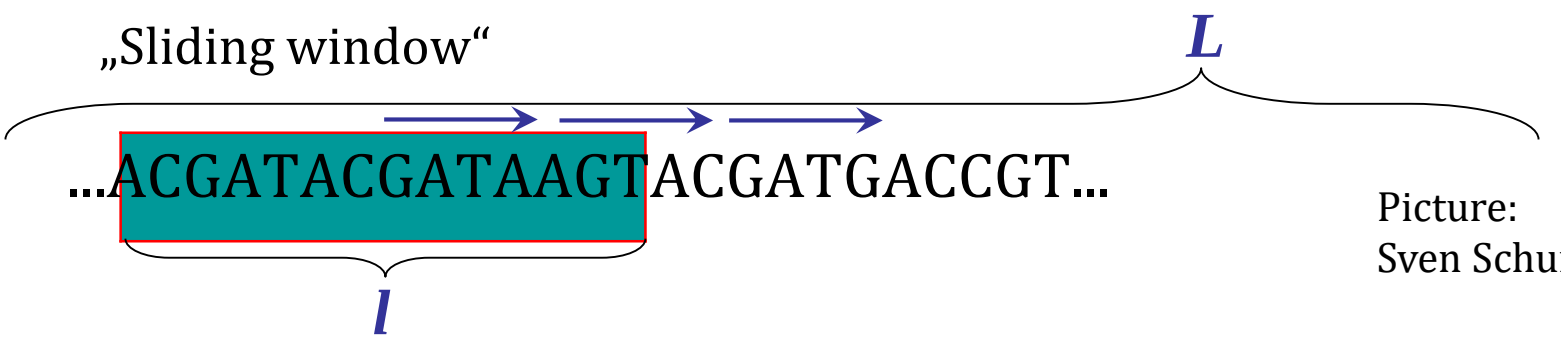
- We see that most of the known CpG islands have a positive log odds score, while most of the non-islands have a negative log odds score
- **Errors** are caused by: incorrect labels in the training sets, and problems when determining borders between CpG-islands and non-islands

Picture from: **Durbin** et al. (Ed): Biological Sequence Analysis, Cambridge University Press, 1998

Pros and Cons of the scoring model

- Given a short piece of DNA, with sufficient certainty, one can decide if it is a CpG-island or not
- You cannot identify a potential CpG-island embedded in a long genomic sequence
- The latter problem can be resolved by using a Hidden Markov Modell →

Long sequence: Finding CpG-islands with a sliding window approach



Picture:
Sven Schuirer

- Calculate log odds score S in every window of width l
- Disadvantages:
  - Runtime (?)
  - unknown size of the island → unknown window width l