



Gene Finding

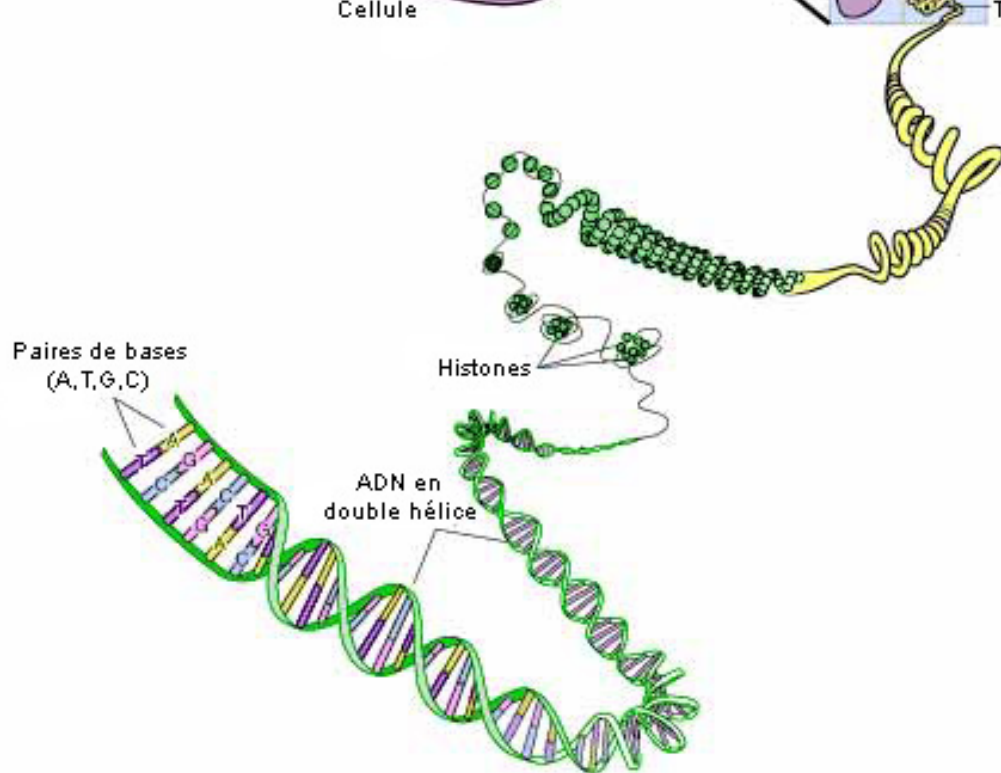
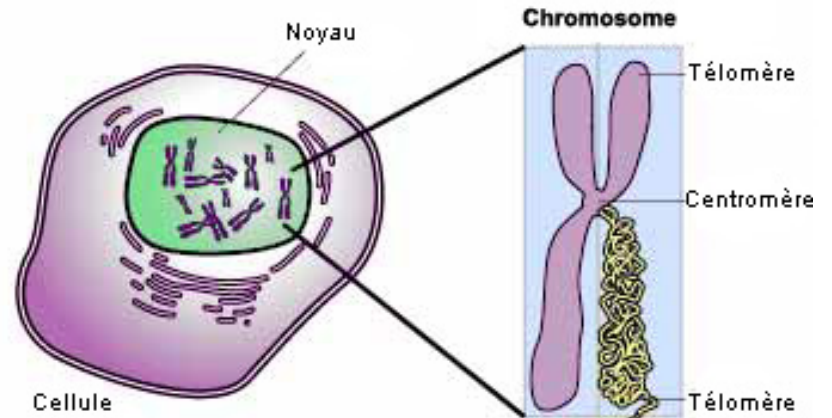
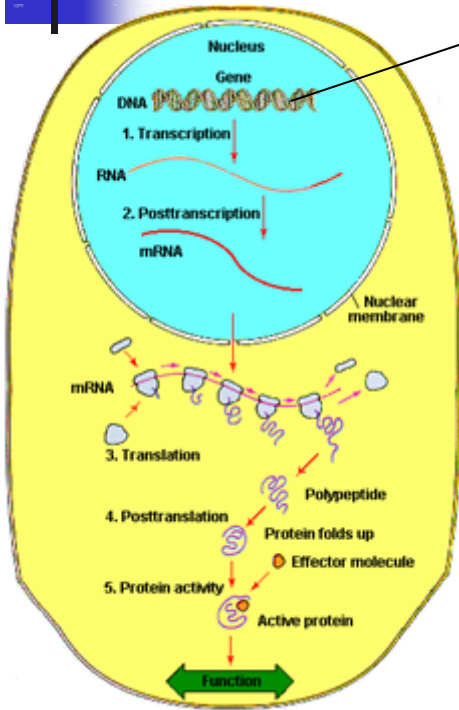
Charles Yan



Gene Finding

- Genomes of many organisms have been sequenced

Genome





Completely Sequenced Genomes

superkindom -> kindom -> phylum -> class -> order -> family -> (genera, species)

Example: eukaryote -> animal -> chordata [spinal cord] -> mammalia [suckle young] -> primate [most highly developed] -> hominidae [two-legged] -> (homo sapiens) [human, modern human]

Superkindom	Kindom	page	How many genomes	Last update
Archeabacteria	.	arch	16	jul-09-02
Bacteria	.	bac	89	jan-13-03
Eukaryote	Fungi	yeast	2	apr-14-02
Eukaryote	Protozoa	protozoan	1	jan-13-03
Eukaryote	Plant	plants	2	apr-15-02
Eukaryote	Animalia	worm	1	apr-14-02
Eukaryote	Animalia	insect	1	apr-14-02
Eukaryote	Animalia	mouse/rat	2	jan-18-06
Eukaryote	Animalia	domestic animals	2	jan-19-06
Eukaryote	Animalia	human/chimpanzee	2	jan-18-06



Gene Finding

- More than 60 eukaryotic genome sequencing projects are underway



Human Genome Project (HGP)

- To determine the sequences of the 3 billion bases that make up human DNA
 - 99% human DNA sequence finished to 99.99% accuracy (April 2003)
- To identify the approximate 100,000 genes in human DNA (The estimates has been changed to 20,000-25,000 by Oct 2004)
 - 15,000 full-length human genes identified (March 2003)
- To store this information in databases
- To develop tools for data analysis



Gene Finding

- Genomes of many organisms have been sequenced
- We need to decipher the raw sequences
 - Where are the genes?
 - What do they encode?
 - How the genes are regulated?

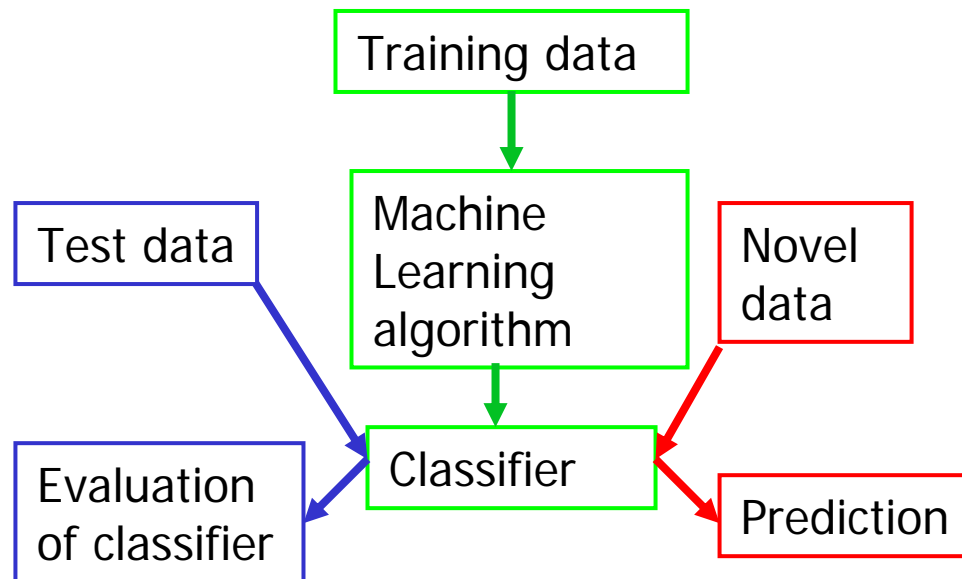


Gene Finding

- Homology-based methods, also called 'extrinsic methods'
 - It seems that only approximately half of the genes can be found by homology to other known genes (although this percentage is of course increasing as more genomes get sequenced).
- Gene prediction methods or 'intrinsic methods'
 - (<http://www.nslj-genetics.org/gene/>)

Machine Learning Approach

- Split data into a **training set** and a **test set**
- Use the training set to train a classifier
- Test the classifier on test set
- The classifier then can be applied to novel data





Data, examples, classes, classifier

ccgctttttgccagcataacggtgtcga, 1

accacgtttttgccagcatttgccagca, 0

atcatcacgatcacgaacatcaccacg, 0

...





N-fold cross-validation

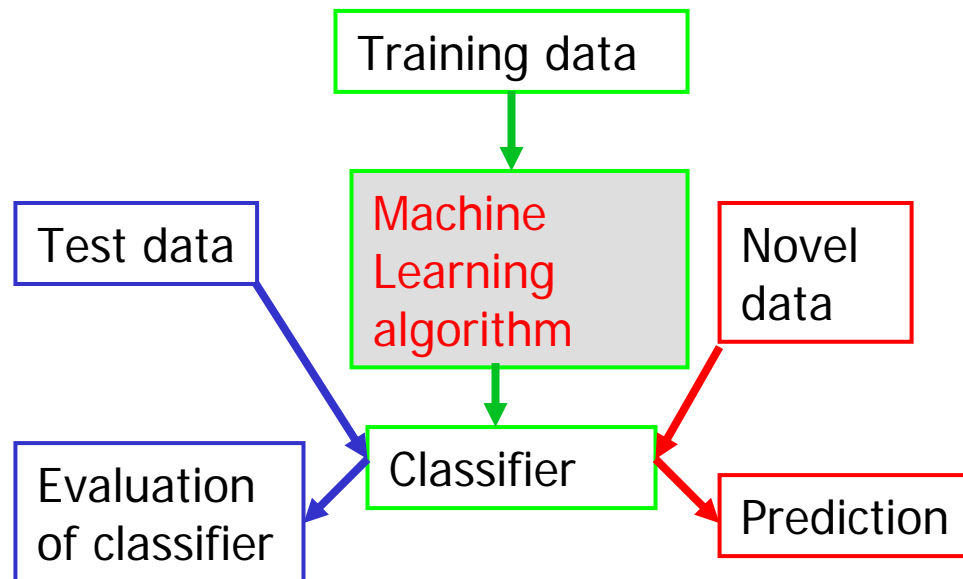
3-fold cross-validation

E.Coli K12 Genome
4,639,675



	Training Set	Test Set
Round 1	 	
Round 2		 
Round 3		 

Machine Learning Approach



Gene-finders

Name	Methods	Organism	Access
ER	Discriminant Analysis	Human, Arabidopsis	WWW
GENSCAN (seems the most accurate)	Semi Markov Model	vertebrate, caenorhabditis, arabidopsis, maize	WWW (Stanford) , WWW (MIT)
GRAIL	Neural Network	human, mouse, arabidopsis, drosophila, E.coli	WWW (ORNL or JAPAN)
GenLang	Definite Clause Grammer	Vertebrate, Drosophila, Dicot	WWW , Email
GenView	Linear combination	Human, Mouse, Diptera	WWW
GeneFinder(FGENEH, etc.)	LDA	Human, E.coli, Drosophila, Plant, Nematode, Yeast	WWW , Email
GeneID	perceptron,rules	Vertebrate	WWW , Email
GeneMark	5th-Markov	Almost all model organism	WWW , Email
GeneParser	neural networks	Human	ftp from the (WWW page)
Genie	GHMM	Human (vertebrate)	WWW
Glimmer	interpolated Markov models (IMMs)	microbial	WWW
MORGAN	Decision Tree	vertebrate	WWW
MZEF	Quadratic Discriminant Analysis	Human, mouse, Arabidopsis, Pombe	WWW , binary
NetPlantGene	Combined Neural Networks	A. thaliana	WWW
OC1	decision tree	Human	WWW
PROCRUSTES	spliced alignment	vertebrate	WWW
Sorfind	rule base	Human	ftp
VEIL	HMM	vertebrate	WWW



Gene Finding

- Finally, the machine learning method we are going to implement for the final project

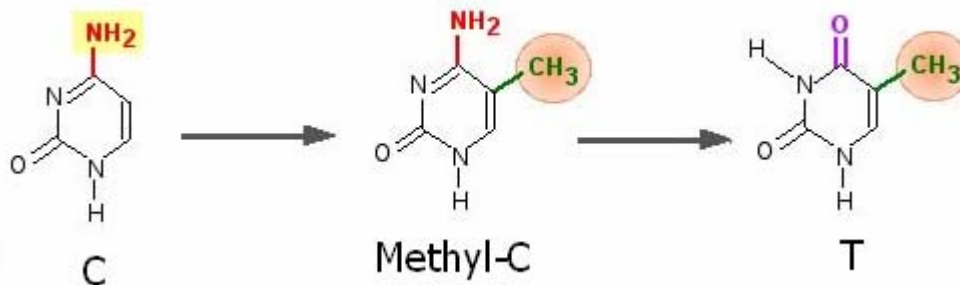


Markov Chains

- A stochastic process has the **Markov property** if the conditional probability distribution of future states of the process, given the present state and all past states, depends only upon the current state and not on any past states, i.e. it is conditionally independent of the past states (the *path* of the process) given the present state. A process with the Markov property is usually called a **Markov process**.
- A **Markov chain**, named after Andrey Markov, is a discrete-time stochastic process with the Markov property.

Markov Chains: CpG islands

- The **CG island** is a short stretch of DNA in which the frequency of the CG sequence is higher than other regions. It is also called the **CpG island**, where "p" simply indicates that "C" and "G" are connected by a phosphodiester bond.
- Whenever the dinucleotide CpG occurs, the C nucleotide is typically chemically modified by methylation.
 - C of CpG is methylated into methyl-C.
 - methyl-C mutates into T relatively easily.





Markov Chains: CpG islands

- Thus, in general, CpG dinucleotides are rarer in the genome. $F(\text{CpG}) < f(\text{C}) * f(\text{G})$.
- Methylation process is suppressed before the “starting point” of many genes.
- These regions (**CpG islands**) have more CpG than elsewhere.
- Usually, CpG islands are a few hundred to a few thousand bases long.
- Identification of CpG islands is important for gene finding.

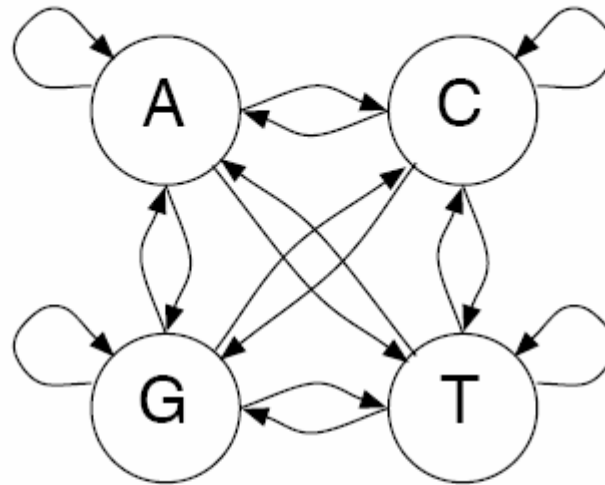
Markov Chains: CpG islands

APRT

(*Homo Sapiens*)

CCCGGGTCCGGGGGGGAAGAGCCGCTCAACGGCAGGGCCCATCCGGAGAGGCCAGCG
CCCCCGGCCGGTCCAGCCAGGCCCGCGCCTCCGCCCTGGGCTGCTCCCTCCGGGCCCT
GCACCGCCCTCCTGCTACTTGGACCGCTTCTCAGCCCTCCTCCACCCCGCGCGCCAGC
CTCCCGCGCGCAGCGTGGGGATCTCGGCCAATAAAGGAGAAAGGGCGCGGCCCGTACGGC
CGCCAGGTGCGTGGGGAGACCAGCTCAGCCCTCCTCCAGCCGCAAGGCCCGGCC
ACAGCTGCCTGGCTGCAGTCAGAAGCGTAGCCCGAGACAAGGAAGGGCGCCTTGACTCGC
ACTTTTGTCCGGTTCGAACGTTCTGCTCAGTGGTGGTGGGAATGGAGCGCGTCTTAAAA
TCGATGGCGCCTAGGAGTCCATGAAATACGGTACAGGCTTCCGGCGACGGATGCCCGCC
CCTCACCCACGCTCCGCCCTCGGGGATGCCCCACCCCTCGTGGCGGTCCCGCCCGTCCC
CGCGCAGGCGCGCTCGGGCTGCCGCTGGCTCTTCGCAACCGCGCCATGGCCGACTCCGAGC
TGCAGCTGGTTGAGCAGCGGATCCGCAGCTTCCCCGACTTCCCCACCCAGGCGTGGTAT
TCAGGTGCACGCACAGGCCCGCCCTCGTGGCGCCCGACCTGGGGCCTACCGGAATTGGG
GCTGCTGTGGTTACAGTGGCCTTGGGAGCTCAGAGAGGTTGAGACATAGGCTGGGCTCAC
ACAGCCAGGTAACAGCAGGGTGGGGTGGAGTCAGGGTCTAGGGTGGCAGCTGCCAAGCT
GTGCAACAAAGCTGTTTTCTCGGGAGGCTGAGGACCACACACCACTTCCACTCCAGGC
TGAGCTGGAGATTCAGAAAGACGCCCTGGAGCCAGGACAGAGGGTGGTGGTGGATGA
TCTGCTGGCCACTGGTGGTAAGGGTCTCCCAGCCAACTGCTGTGGCTCCAAGGGCCT
GGTGGGAGTGGACAGGACCTCGCTGTGTGACATGGGATGCAGCTTACTGTTGTCCAGAG
GGTGCCTGGTGGCCAGGCCGACACCTTCTCTCCCCATGCCTTCCCCTCCCCAACCCAGG
GGCTGGCCTGGAGCACCTGCTCTCTGCAGCCAGGCCAACTGGGGACCTCACCTCCCAT
CCCCAGGAACCATGAACGCTGCCTGTGAGCTGCTGGGCCGCTGCAGGCTGAGGTCCTGG
AGTGGCTGAGCCTGGTGGAGCTGACCTCGCTTAAGGGCAGGGAGAAGCTGGCACCTGTAC
CCTTCTTCTCTCTCCTGCAGTATGAGTGAACACAGGGCCTCCAGCCCAACATCTCCAGC

Markov Chains : CpG islands



States: A,C,G,T

Emissions: corresponding letter

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



Markov Chains : CpG islands

The probability that a sequence x is generated by a Markov chain model

$$\begin{aligned}P(x) &= P(x_1, x_2, \dots, x_n) \\&= P(x_1, x_2, \dots, x_{n-1}) \cdot P(x_n | x_1, x_2, \dots, x_{n-1}) \\&= P(x_1, x_2, \dots, x_{n-2}) \cdot P(x_{n-1} | x_1, x_2, \dots, x_{n-2}) \cdot P(x_n | x_1, x_2, \dots, x_{n-1}) \\&\dots \\&= P(x_1) \cdot P(x_2 | x_1) \dots P(x_n | x_1, x_2, \dots, x_{n-1})\end{aligned}$$

By applying many times of

$$P(X, Y) = P(X) \cdot P(Y | X)$$



Markov Chains : CpG islands

One assumption of Markov chain is that the probability of x_i only depend on the previous symbol x_{i-1} , i.e.,

$$P(x_n | x_1, x_2, \dots, x_{n-1}) = P(x_n | x_{n-1})$$

Thus,

$$\begin{aligned} P(x) &= P(x_1) \cdot P(x_2 | x_1) \dots P(x_n | x_1, x_2, \dots, x_{n-1}) \\ &= P(x_1) \cdot P(x_2 | x_1) \dots P(x_n | x_{n-1}) \\ &= P(x_1) \prod_{i=1}^{n-1} a_{x_i, x_{i+1}} \end{aligned}$$



Markov Chains : CpG islands

Training the model, i.e., estimate the transition probabilities $a_{st} = P(x_i = t \mid x_{i-1} = s)$

Maximum likelihood (ML) approach is used to estimate the transition probabilities

$$a_{st} = \frac{C_{st}}{\sum_{t'} C_{st'}}$$

Where C_{st} is the number of times that letter t followed letter s

Markov Chains : CpG islands

- A set of CpG islands (CpG model)
 - 1st row: The probabilities that A is followed by each of the four bases.
 - The sum of each row is 1
- A set of sequences that are not CpG islands (Background model)

+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	<u>0.274</u>	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	<u>0.078</u>	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292



Markov Chains : CpG islands

- Given a sequence x , does it belong to CpG islands?

Log likelihood ratio of CpG model vs background model

$$\log \frac{P(x|\text{model } +)}{P(x|\text{model } -)} = \sum_{i=1}^L \log \frac{a_{x_{i-1}x_i}^+}{a_{x_{i-1}x_i}^-}$$

If the log likelihood ratio >0 , then x belongs to CpG islands.

Markov Chains : CpG islands

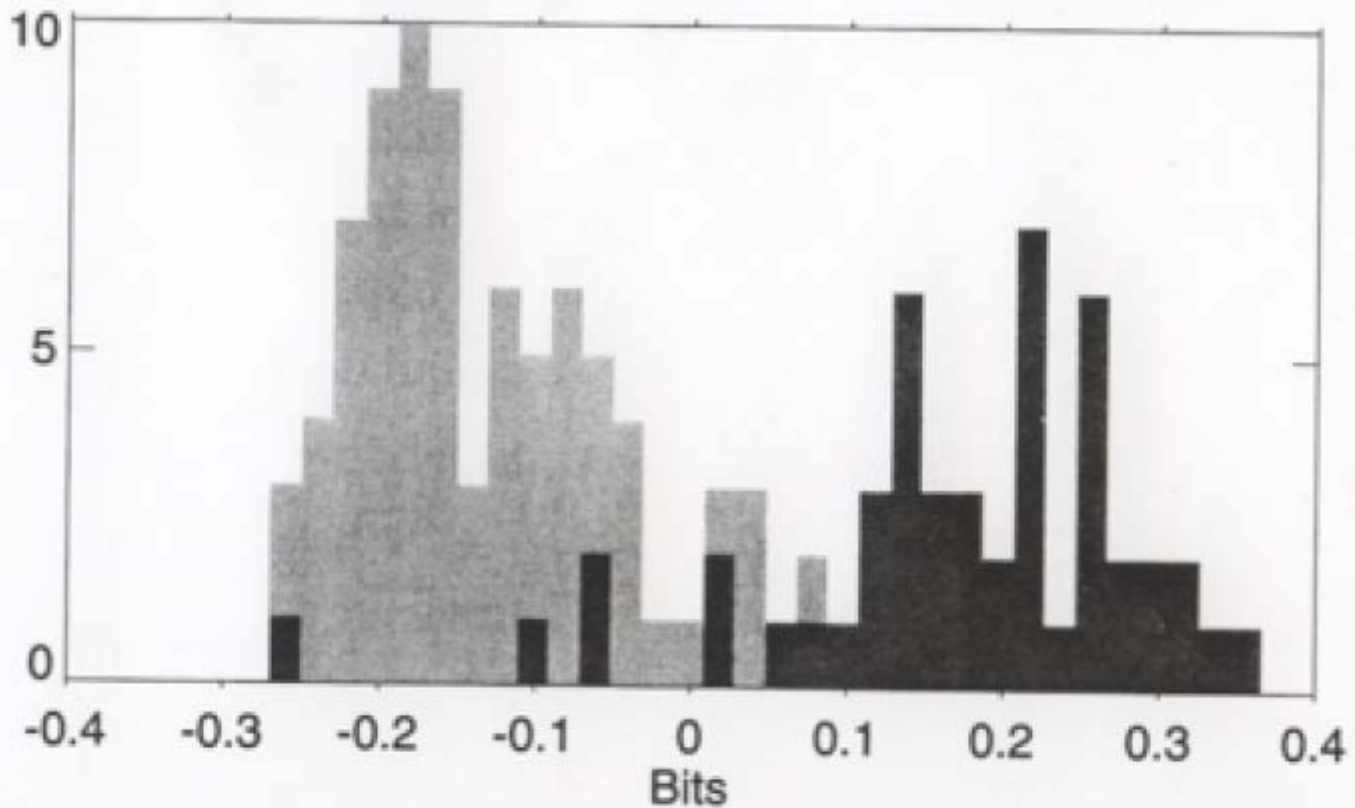


Figure 3.2 *The histogram of the length-normalised scores for all the sequences. CpG islands are shown with dark grey and non-CpG with light grey.*



Markov Chains

- 1st-order Markov chains
- Kth-order Markov chains

A sequence x_1, x_2, \dots of random variables is a *k-th order Markov chain* if, for all i :

$$P(x_i | x_1, x_2, \dots, x_{i-1}) = P(x_i | x_{i-k}, x_{i-k+1}, \dots, x_{i-1})$$

i.e., i^{th} value is independent of all but the previous k values

- Inhomogeneous Markov chains
- Semi-Markov models
- ...



Prokaryotes vs. Eukaryotes

- **Prokaryotes** are organisms without a cell nucleus.
 - Most prokaryotes are bacteria.
 - Prokaryotes can be divided into Bacteria and Archaeabacteria.
- **Eukaryotes** are organisms which a membrane-bound nucleus.



Prokaryotes vs. Eukaryotes

- Prokaryotes' genomes are relatively simple: coding region (genes) vs. non-coding region.
- Eukaryotes' genomes are complicated.

Eukaryotic genes

