Sequence models
(Bernoulli and Markov models)

Why do we need random models ?

- Any pattern discovery relies on an underlying model to estimate the random expectation.
  - This model can be simple (succession of independent and equiprobable nucleotides) or more elaborate (differences in oligonucleotide composition).
  - The choice of an inappropriate model can lead to false conclusions.
  - In practice, a sequence model can be used to generate random sequences, which will serve to validate some theoretical assumptions.
- Example: comparison of observed and expected occurrences with the binomial distribution, as applied with oligo-analysis :
  - Relies on an assumption that successive oligonucleotides are independent from each other.
  - This is clearly not the case: each k-letter word depends on the k-1 neighbour words on both sides. How far does it affect the conclusions ?
  - We could test it by generating random sequences, counting words, and fitting the distribution of observed occurrences with a binomial distribution.

Probability of a sequence segment

- What is the probability for a given sequence segment (oligonucleotide, “word”) to be found at a given position of a DNA sequence ?
- Different models can be chosen
  - **Bernoulli model**
    - Assumes independence between successive nucleotides.
    - The probability of each residue is fixed a priori (*prior residue probability*)
      - Example: $P(A) = 0.35; P(T) = 0.32; P(C) = 0.17; P(G) = 0.16$
    - Particular case: equiprobable residues
      - $P(A) = P(T) = P(C) = P(G) = 0.25$
      - Simple, but NOT realistic !
  - **Markov model**
    - The probability of each residue depends on the ***m*** preceding residues.
    - The parameter ***m*** is called the ***order*** of the Markov model
    - Remark: a Markov model of order 0 is a Bernoulli model.

Independent and equiprobable nucleotides

- The simplest model : Bernoulli with identically and independently (i.i.d.) distributed nucleotides.

  $p = P(A) = P(C) = P(G) = P(T) = 0.25$

$$P(S) = p^L$$

- The probability of a sequence
  - Is the product of its residue probabilities (independence)
  - Equiprobability: since all residues have the same probability, it is simply computed as the residue proba (*p)* to the power of the sequence length (*L*)
    - *S* is a sequence segment (e.g. an oligonucleotide)
    - *L* length of the sequence segment
    - *p* nucleotide probability
    - *P(S)* is the probability to observe this sequence segment at given position of a larger sequence
- Example
  - P(CACGTG) = $0.25^6 = 2.44e^{-4}$

Bernoulli model : independently distributed nucleotides

- A more refined model consists in using residue-specific probabilities. The probability of each residue is assumed to be constant on the whole sequence (Bernoulli schema).
- The probability of a sequence is the product of its residue probabilities.
  - $i = 1..k$ is the index of nucleotide positions
  - r_i is the residue found at position I
  - $P(r_i)$ is the probability of this residue
- Example: non-coding sequences in the yeast genome
  - P(A) = P(T) = 0.325
  - P(C) = P(G) = 0.175
  - P(CACGTG) = P(C) P(A) P(C) P(G) P(T) P(G)
    = $0.325^4 * 0.175^2$
    = $9.91E^{-5}$

$$P(S) = \prod_{i=1}^{L} P(r_i)$$

Markov chains and transition matrices

$$P(r_i \mid S_{i-m,i-1})$$

Transition matrix, order 1

	a	**c**	**g**	**t**
A	P(A|A)	P(C|A)	P(G|A)	P(T|A)
C	P(A|C)	P(C|C)	P(G|C)	P(T|C)
G	P(A|G)	P(C|G)	P(G|G)	P(T|G)
T	P(A|T)	P(C|T)	P(G|T)	P(T|T)

Transition matrix, order 2

Pref	**A**	**C**	**G**	**T**
AA	P(A|AA)	P(C|AA)	P(G|AA)	P(T|AA)
AC	P(A|AC)	P(C|AC)	P(G|AC)	P(T|AC)
AG	P(A|AG)	P(C|AG)	P(G|AG)	P(T|AG)
AT	P(A|AT)	P(C|AT)	P(G|AT)	P(T|AT)
CA	P(A|CA)	P(C|CA)	P(G|CA)	P(T|CA)
CC	P(A|CC)	P(C|CC)	P(G|CC)	P(T|CC)
CG	P(A|CG)	P(C|CG)	P(G|CG)	P(T|CG)
CT	P(A|CT)	P(C|CT)	P(G|CT)	P(T|CT)
GA	P(A|GA)	P(C|GA)	P(G|GA)	P(T|GA)
GC	P(A|GC)	P(C|GC)	P(G|GC)	P(T|GC)
GG	P(A|GG)	P(C|GG)	P(G|GG)	P(T|GG)
GT	P(A|GT)	P(C|GT)	P(G|GT)	P(T|GT)
TA	P(A|TA)	P(C|TA)	P(G|TA)	P(T|TA)
TC	P(A|TC)	P(C|TC)	P(G|TC)	P(T|TC)
TG	P(A|TG)	P(C|TG)	P(G|TG)	P(T|TG)
TT	P(A|TT)	P(C|TT)	P(G|TT)	P(T|TT)

- In a Markov model, the probability to find a letter at position *i* depends on the residues found at the *m* preceding residues.
- The tables represent the transition matrices for Markov chain models of order $m=1$ (top) and $m=2$ (bottom).
- Each row specifies one ***prefix***, each column one ***suffix***.
- The values indicate the probability to observe a given residue (suffix r_i) at position (i) of the sequence, as a function of the m preceding residues (the prefix $S_{i-m,i-1}$)
- Particular case
  - A Bernoulli model is a Markov model of order 0.

Markov model estimation ("training")

- Transition frequencies for a Markov model of order m can be estimated from the frequencies observed for oligomers (k-mers) of length k=m+1 in a reference sequence set.
- Example
  - The upper table shows dinucleotide frequencies ($k=2$) computed from the whole set of upstream sequences of the yeast *Saccharomyces cerevisiae*.
  - This table can be used to estimate a Markov model of order $m = k-1 = 1$.

Dinucleotide frequencies

Sequences	Occurrences	Frequency
S	**N(S)**	**F(S)**
AA	526,149	0.112
AC	251,377	0.054
AG	275,056	0.059
AT	414,453	0.088
CA	294,423	0.063
CC	178,324	0.038
CG	146,052	0.031
CT	275,859	0.059
GA	277,343	0.059
GC	184,367	0.039
GG	173,404	0.037
GT	239,569	0.051
TA	369,980	0.079
TC	280,475	0.060
TG	279,932	0.060
TT	521,236	0.111

$$P(r_i \mid S_{1..m}) = \frac{F_{bg}(r_i \mid S_{1..m})}{\sum_{j \in A} F_{bg}(r_j \mid S_{1..m})} = \frac{F_{bg}(S_{1..m} r_i)}{\sum_{j \in A} F_{bg}(S_{1..m} r_j)}$$

$$P(G \mid T) = \frac{F(G \mid T)}{\sum_{j \in A} F(j \mid T)} = \frac{F(TG)}{F(T*)}$$

$$= \frac{0.060}{0.079 + 0.060 + 0.060 + 0.111}$$

$$= \frac{0.060}{0.310} = 0.194$$

Transition matrix, order 1

Prefix \ Suffix	A	C	G	T	P(Prefix)	N(Suffix)
A	0.359	0.171	0.187	0.283	0.313	1,467,035
C	0.329	0.199	0.163	0.308	0.191	894,658
G	0.317	0.211	0.198	0.274	0.187	874,683
T	0.255	0.193	0.193	0.359	0.310	1,451,623
P(Suffix)	0.313	0.191	0.187	0.310		
N(Suffix)	1,467,895	894,543	874,444	1,451,117		

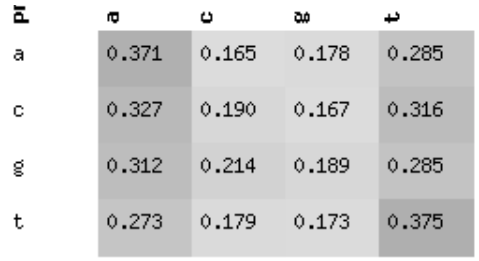
Examples of transition matrices

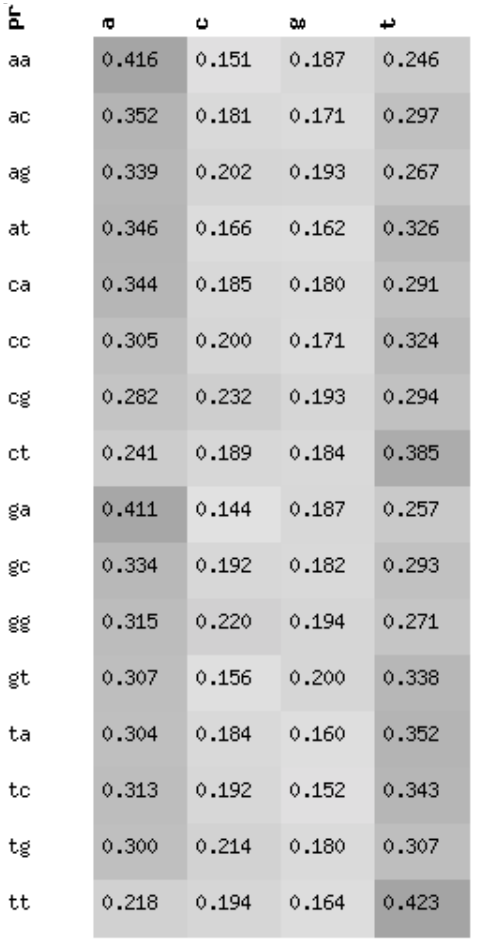
- The two tables below show the transition matrices for a Markov model of order 1 (top) and 2 (bottom), respectively.
- Trained with the whole set of **non-coding upstream sequences** of the yeast *Saccharomyces cerevisiae*.
- Notice the high probability of transitions from AA to A and TT to T.

$$P\left(r_i \mid S_{i-m,i-1}\right)$$

Pre/Suffix	A	C	G	T	P(Prefix)
a	0.371	0.165	0.178	0.285	**0.321**
c	0.327	0.190	0.167	0.316	**0.183**
g	0.312	0.214	0.189	0.285	**0.177**
t	0.273	0.179	0.173	0.375	**0.320**
Sym	**1.283**	**0.748**	**0.708**	**1.261**	
P(suffix)	**0.321**	**0.183**	**0.176**	**0.320**	

Prefix/Suffix	A	C	G	T	P(Prefix)
aa	0.416	0.151	0.187	0.246	**0.119**
ac	0.352	0.181	0.171	0.297	**0.053**
ag	0.339	0.202	0.193	0.267	**0.057**
at	0.346	0.166	0.162	0.326	**0.092**
ca	0.344	0.185	0.180	0.291	**0.060**
cc	0.305	0.200	0.171	0.324	**0.035**
cg	0.282	0.232	0.193	0.294	**0.031**
ct	0.241	0.189	0.184	0.385	**0.058**
ga	0.411	0.144	0.187	0.257	**0.055**
gc	0.334	0.192	0.182	0.293	**0.038**
gg	0.315	0.220	0.194	0.271	**0.033**
gt	0.307	0.156	0.200	0.338	**0.050**
ta	0.304	0.184	0.160	0.352	**0.087**
tc	0.313	0.192	0.152	0.343	**0.057**
tg	0.300	0.214	0.180	0.307	**0.055**
tt	0.218	0.194	0.164	0.423	**0.120**
Sum	**5.127**	**3.000**	**2.860**	**5.013**	
P(suffix)	**0.321**	**0.183**	**0.176**	**0.319**	





Markov chains and Bernoulli models

- By extension of the concept of Markov chain, Bernoulli models can be qualified as Markov models of order 0 (the order 0 means that there is no dependency between a residue and the preceding ones).
- The prior probabilities of a Makov model of order *m=0* can be estimated from the residue of single nucleotides (*k=m+1=1*) in a background sequence set.
- The table below shows the residue frequencies in the genomes of the yeast *Saccharomyces cerevisiae* and the bacteria *Escherichia coli K12*, respectively.
- Notice the strong differences between these genomes.

Markov order 0 = Bernouli

A	C	G	T	Genome
0.310	0.191	0.191	0.309	*Saccharomyces cerevisiae*
0.246	0.254	0.254	0.246	*Escherichia coli K12*

Scoring a sequence segment with a Markov model

- The example below illustrates the computation of the probability of a seqence segment ***P(S|B)*** with a background Markov model ***B*** of order 2, calibrated from 3nt frequencies on the yeast genome.

CCTACTATATGCCCAGAATT

Background model B

Transition matrix, order 2

Prefix/Suffix	A	C	G	T	P(Prefix	N(Prefix
AA	0.388	0.161	0.200	0.251	**0.112**	**525,000**
AC	0.339	0.198	0.173	0.290	**0.054**	**251,072**
AG	0.345	0.204	0.196	0.255	**0.059**	**274,601**
AT	0.311	0.184	0.182	0.323	**0.088**	**413,946**
CA	0.347	0.178	0.189	0.286	**0.063**	**293,750**
CC	0.341	0.190	0.161	0.309	**0.038**	**178,110**
CG	0.293	0.221	0.196	0.290	**0.031**	**145,876**
CT	0.229	0.195	0.205	0.371	**0.059**	**275,634**
GA	0.394	0.155	0.187	0.264	**0.059**	**277,053**
GC	0.330	0.205	0.169	0.297	**0.039**	**184,192**
GG	0.318	0.217	0.187	0.277	**0.037**	**173,266**
GT	0.285	0.175	0.204	0.336	**0.051**	**239,384**
TA	0.300	0.193	0.168	0.339	**0.079**	**369,426**
TC	0.313	0.203	0.152	0.332	**0.060**	**280,131**
TG	0.302	0.209	0.208	0.282	**0.060**	**279,783**
TT	0.210	0.208	0.189	0.392	**0.111**	**520,906**
P(Suffix)	**0.313**	**0.191**	**0.187**	**0.310**		
N(suffix)	**1,466,075**	**893,444**	**873,260**	**1,449,351**		

Sequence probability given the backgound model

$$P(S \mid B) = P(S_{1,m} \mid B) \prod_{i=m+1}^{L} P\left(r_i \mid S_{i-m,i-1}, B\right)$$

pos		P(R|W)		wR	S	P(S)
	1	P(CC)	0.038	cc	CC	3.80E-02
	2	P(T|CC)	0.309	ccT	CCT	1.17E-02
	3	P(A|CT)	0.229	ctA	CCTA	2.69E-03
	4	P(C|TA)	0.193	taC	CCTAC	5.19E-04
	5	P(T|AC)	0.290	acT	CCTACT	1.50E-04
	6	P(A|CT)	0.229	ctA	CCTACTA	3.45E-05
	7	P(T|TA)	0.339	taT	CCTACTAT	1.17E-05
	8	P(A|AT)	0.311	atA	CCTACTATA	3.63E-06
	9	P(T|TA)	0.339	taT	CCTACTATAT	1.23E-06
	10	P(G|AT)	0.182	atG	CCTACTATATG	2.25E-07
	11	P(C|TG)	0.209	tgC	CCTACTATATGC	4.69E-08
	12	P(C|GC)	0.205	gcC	CCTACTATATGCC	9.61E-09
	13	P(C|CC)	0.190	ccC	CCTACTATATGCCC	1.82E-09
	14	P(A|CC)	0.341	ccA	CCTACTATATGCCCA	6.21E-10
	15	P(G|CA)	0.189	caG	CCTACTATATGCCCAG	1.17E-10
	16	P(A|AG)	0.345	agA	CCTACTATATGCCCAGA	4.04E-11
	17	P(A|GA)	0.394	gaA	CCTACTATATGCCCAGAA	1.59E-11
	18	P(T|AA)	0.251	aaT	CCTACTATATGCCCAGAAT	4.00E-12
	19	P(T|AT)	0.323	atT	CCTACTATATGCCCAGAATT	1.29E-12

Background sequences

- The frequencies observed for a *k*-letter word in a reference sequence set (background sequence) can be used to estimate the expected frequencies of the same *k*-letter word in the sequences to be analyzed.
- Typical background models:
  - Whole genome
    - But this will bias the estimates towards coding frequencies, especially in microbial organisms, where the majority of the genome is coding.
  - Whole set of intergenic sequences
    - More accurate than whole-genome estimates, but still biased because intergenic sequences include both upstream and downstream sequences
  - Whole set of upstream sequences, same sizes as the sequences to be analyzed
    - Requires a calibration for each sequence size
  - Whole set of upstream sequences, fixed size (default on the web site)
    - Reasonably good estimate for microbes, NOT for higher organisms.

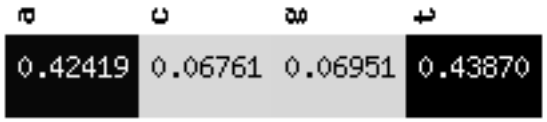
Nucleotide composition of the Plasmodium upstream sequences

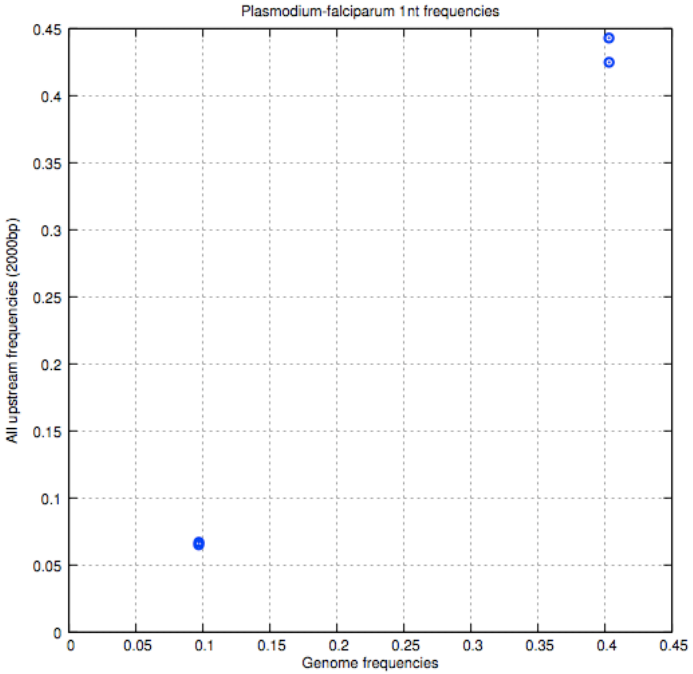
- The genome shows a strong richness in A and T residues (80%AT).
- This enrichment is even stronger in upstream non-coding sequences (86%AT).

Frequencies

Residue	Genome	Upstream (max 2kb)
a	0.40	0.42
c	0.10	0.07
g	0.10	0.07
t	0.40	0.44

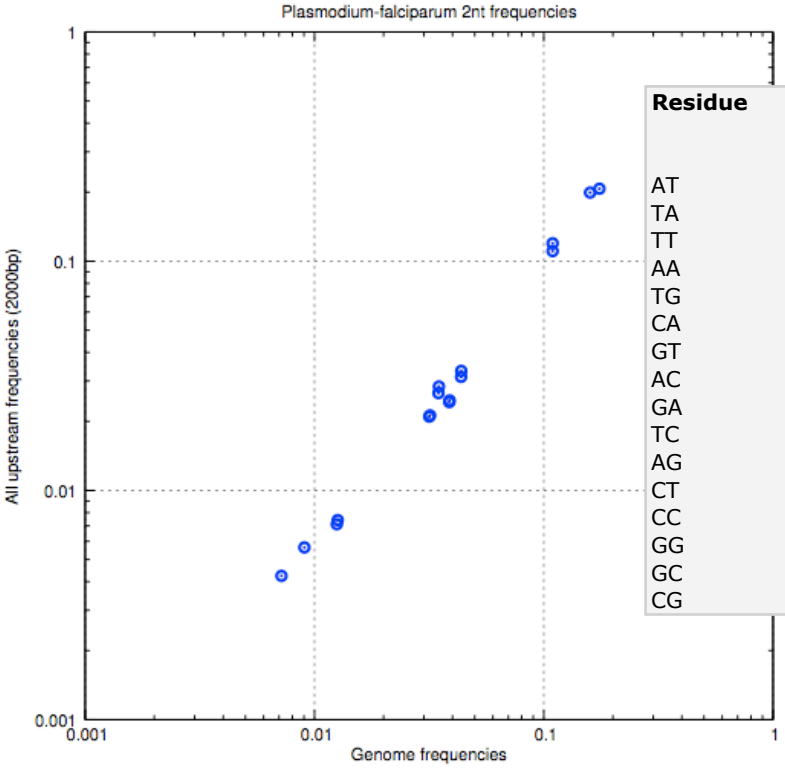
Upstream frequencies
Gray scale





Dinucleotide composition of the Plasmodium upstream sequences

- Dinucleotide frequencies reflect the AT-richness.



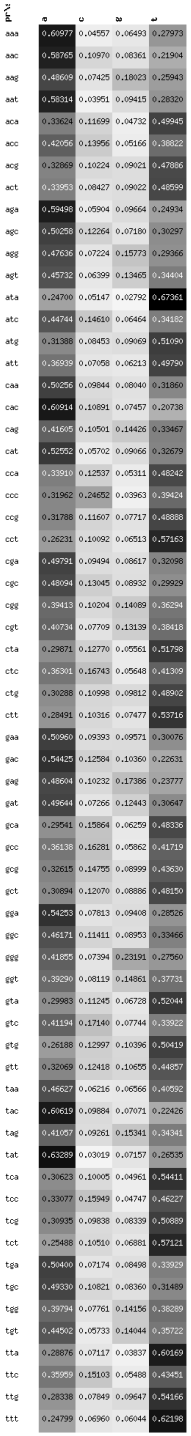
Residue	Genome	Upstream (max 2 kb)
AT	0.175	0.207
TA	0.159	0.199
TT	0.109	0.120
AA	0.109	0.111
TG	0.044	0.033
CA	0.044	0.031
GT	0.035	0.028
AC	0.035	0.027
GA	0.039	0.025
TC	0.039	0.024
AG	0.032	0.021
CT	0.032	0.021
CC	0.013	0.007
GG	0.013	0.007
GC	0.009	0.006
CG	0.007	0.004

Markov order m=1
(based of dinucleotides k=2)

pr	a	c	g	t
a	0.39983	0.06522	0.05243	0.48253
c	0.47917	0.13198	0.06745	0.32140
g	0.36686	0.08719	0.12634	0.41961
t	0.44833	0.05728	0.07758	0.41681

Transition frequencies

Markov order m=3
(based on tetranucleotides k=4)



- On the basis of oligonucleotide frequencies, one can compute Markov models, which indicate the probability to osberve a certain residue (suffix) after a certain oligonucleotide (prefix).
- The Markov model can be represented in the form of a transition table.

Markov order m=1
(based on dinucleotides k=2)

pr	a	c	g	t
a	0.39983	0.06522	0.05243	0.48253
c	0.47917	0.13198	0.06745	0.32140
g	0.36686	0.08719	0.12634	0.41961
t	0.44833	0.05728	0.07758	0.41681

Markov order m=2
(based on trinucleotides k=3)

pr	a	c	g	t
aa	0.55133	0.05771	0.06855	0.32241
ac	0.59611	0.10571	0.07782	0.22037
ag	0.45875	0.08478	0.16910	0.28737
at	0.60625	0.03631	0.08099	0.27645
ca	0.32382	0.11650	0.04995	0.50973
cc	0.36057	0.16777	0.04837	0.42329
cg	0.32114	0.10703	0.08672	0.48511
ct	0.28437	0.09985	0.07587	0.53991
ga	0.54541	0.06894	0.09101	0.29464
gc	0.49165	0.11509	0.08103	0.31223
gg	0.43450	0.07578	0.16309	0.32663
gt	0.43973	0.06254	0.13930	0.35844
ta	0.26328	0.06301	0.03410	0.63961
tc	0.38894	0.15242	0.05942	0.39921
tg	0.29671	0.08900	0.09461	0.51969
tt	0.29185	0.07502	0.06441	0.56871

Markov models show strong variations between organisms

Saccharomyces cerevisiae (Fungus)

Pr	a	c	g	t
a	0.37000	0.16588	0.17908	0.28504
c	0.32610	0.19058	0.16818	0.31514
g	0.31163	0.21456	0.18957	0.28424
t	0.27256	0.17991	0.17364	0.37389

Escherichia coli K12 (Proteobacteria)

Pr	a	c	g	t
a	0.34491	0.18156	0.17676	0.29677
c	0.30806	0.21557	0.22129	0.25507
g	0.27123	0.25972	0.21545	0.25360
t	0.24080	0.19176	0.21144	0.35599

Mycobacterium leprae (Actinobacteria)

Pr	a	c	g	t
a	0.23239	0.28694	0.25692	0.22375
c	0.24574	0.24601	0.30574	0.20252
g	0.21748	0.29238	0.25535	0.23479
t	0.18806	0.26081	0.31784	0.23329

Mycoplasma genitalium (Firmicute, intracellular)

Pr	a	c	g	t
a	0.45565	0.11743	0.13602	0.29091
c	0.39457	0.13008	0.06403	0.41132
g	0.31505	0.18738	0.12047	0.37710
t	0.32450	0.09573	0.11934	0.46044

Bacillus subtilis (Firmicute, extracellular)

Pr	a	c	g	t
a	0.38159	0.13935	0.18767	0.29139
c	0.33699	0.19499	0.16508	0.30293
g	0.34249	0.18100	0.23541	0.24110
t	0.25122	0.17199	0.19402	0.38278

Plasmodium falciparum (Aplicomplexa, intracellular)

Pr	a	c	g	t
a	0.39821	0.06446	0.05206	0.48527
c	0.47798	0.13336	0.06695	0.32171
g	0.36764	0.08587	0.12431	0.42217
t	0.44739	0.05676	0.07673	0.41912

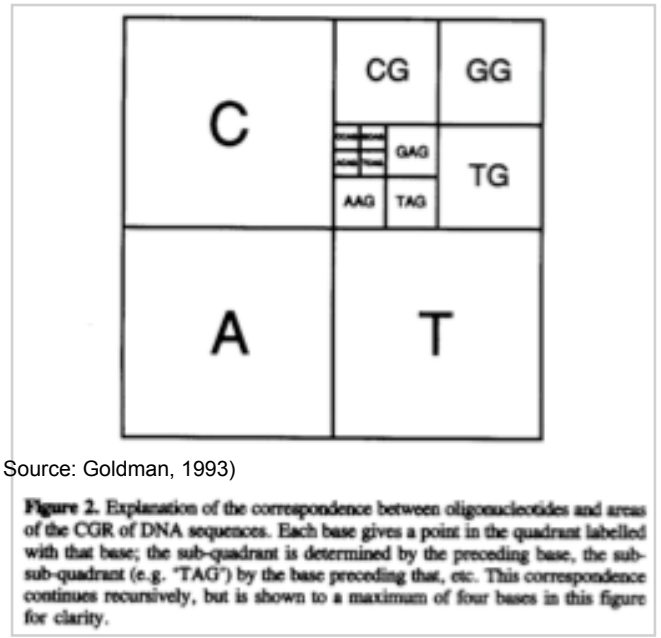
Anopheles gambiae (Insect)

Pr	a	c	g	t
a	0.34603	0.21388	0.18890	0.25119
c	0.31499	0.21232	0.24159	0.23109
g	0.26036	0.25414	0.20275	0.28275
t	0.20368	0.20710	0.24970	0.33951

Homo sapiens (Mammalian)

Pr	a	c	g	t
a	0.29760	0.19031	0.28856	0.22353
c	0.28019	0.30209	0.11692	0.30080
g	0.24408	0.24738	0.30309	0.20545
t	0.18589	0.23061	0.27491	0.30859

Chaos representation - upstream frequencies

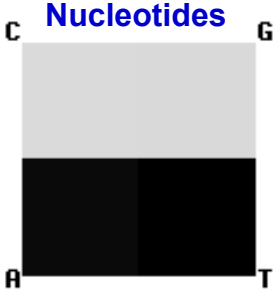


Source: Goldman, 1993)

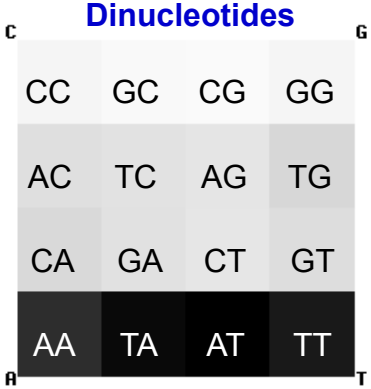
Figure 2. Explanation of the correspondence between oligonucleotides and areas of the CGR of DNA sequences. Each base gives a point in the quadrant labelled with that base; the sub-quadrant is determined by the preceding base, the sub-sub-quadrant (e.g. 'TAG') by the base preceding that, etc. This correspondence continues recursively, but is shown to a maximum of four bases in this figure for clarity.

- The chaos representation (Jeffrey, 1990) permits to visualize oligonucleotide frequencies and detect enrichment in particular ones.
- Plasmodium upstreal sequences are particularly rich for the following motifs
  - A, T nucleotides
  - Oligonucleotides made of As and Ts only (last rw pf each chaos map)
  - Poly-A and poly-T oligos (bottom corners of the maps)
  - (TA){n} motifs (the darkest boxesfrom dinucleotides to tetranucleotides.

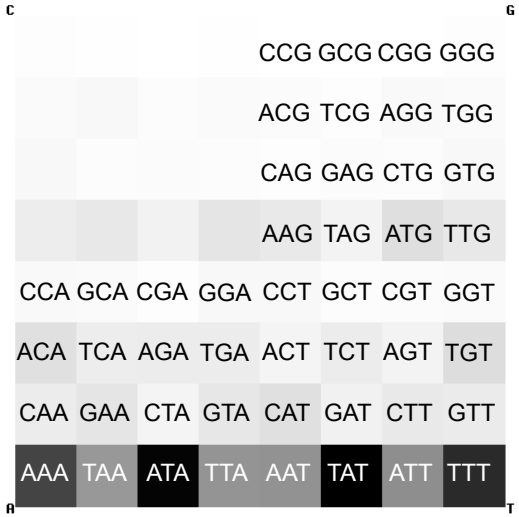
Nucleotides



Dinucleotides



Trinucleotides



Tetranucleotides

Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Nucleic Acids Res (1993) vol. 21 (10) pp. 2487-91
Jeffrey. Chaos game representation of gene structure. Nucleic Acids Res (1990) vol. 18 (8) pp. 2163-70

Hexanucleotide frequencies in Plasmodium Genome versus upstream (2Kb)

- Hexanucleotides show a very wide range of frequencies in the whole genome (X axis) as well as in the subset of upstream sequences (max 2kb, Y axis).