Regulatory sequence analysis

Sequence models

(Bernoulli and Markov models)

Jacques van Helden
Jacques.van.Helden@ulb.ac.be

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 any position of a DNA sequence ?
- Different models can be chosen :
  - Independently distributed nucleotides (Bernoulli processes)
    - Equiprobable
    - Residue-specific probability. Example: P(A or T)=0.7; P(C or G)=0.3
  - Markov chain models
  - External background

Independent and equiprobable nucleotides

- The simplest model : 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

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

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 chain 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 $S_{i-m,i-1}$)

Estimating transition frequencies from oligomer frequencies

Dinucleotide frequencies

Sequences S	Occurrences N(S)	Frequency F(S)
AA	337,835	0.073
AC	256,658	0.055
AG	237,851	0.051
AT	309,792	0.067
CA	325,118	0.070
CC	271,649	0.059
CG	346,636	0.075
CT	236,029	0.051
GA	267,234	0.058
GC	383,865	0.083
GG	270,083	0.058
GT	255,593	0.055
TA	211,948	0.046
TC	267,261	0.058
TG	322,205	0.070
TT	339,463	0.073

$$\hat{P}(R \mid W) = \frac{F_{bg}(R \mid W)}{\sum_{i \in A} F_{bg}(i \mid W)} = \frac{F_{bg}(WR)}{F_{bg}(W*)}$$

Transition matrix, order 1

S	P(A|Sp)	P(C|Sp)	P(G|Sp)	P(T|Sp)
A	0.296	0.225	0.208	0.271
C	0.276	0.230	0.294	0.200
G	0.227	0.326	0.230	0.217
T	0.186	0.234	0.282	0.298

- Transition frequencies for a Markov model of order *m* can be estimated from the frequencies observed for oligomers of length *k=m+1* in a reference sequence set.
- The tables show the dinucleotide frequencies calculated in the whole set of upstream sequences of the yeast *Saccharomyces cerevisiae (top)*, and the estimated transition frequencies.
- Example:

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

$$= \frac{0.070}{0.046 + 0.058 + 0.070 + 0.073} = 0.282$$

Markov chains and transition matrices

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

Transition matrix, order 1

	g	a	c	t
a	0.178	0.369	0.165	0.288
c	0.166	0.327	0.191	0.316
g	0.190	0.313	0.211	0.286
t	0.175	0.273	0.180	0.372

Transition matrix, order 2

	g	a	c	t
aa	0.185	0.411	0.152	0.252
ac	0.171	0.348	0.186	0.296
ag	0.193	0.337	0.201	0.269
at	0.163	0.343	0.167	0.326
ca	0.181	0.344	0.184	0.291
cc	0.168	0.313	0.198	0.321
cg	0.194	0.283	0.227	0.295
ct	0.187	0.240	0.189	0.384
ga	0.186	0.407	0.145	0.262
gc	0.180	0.331	0.194	0.295
gg	0.192	0.318	0.216	0.274
gt	0.199	0.305	0.159	0.338
ta	0.160	0.304	0.182	0.354
tc	0.151	0.313	0.192	0.344
tg	0.184	0.302	0.210	0.304
tt	0.168	0.220	0.195	0.417

- The two tables below show the transition matrices for a Markov model of order 1 (top) and 2 (bottom), respectively.
- Notice the stronge probabilityof transitions from AA to A and TT to T.

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 (CCTACTATATGCCCAGAATT) with a Markov chain of order 2, calibrated from 3nt frequencies on the yeast genome.

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

pos		P(R|W)		wR	S	P(S)
	1	P(CC)	0.039	cc	CC	3.94E-02
	3	P(T|CC)	0.301	ccT	CCT	1.19E-02
	4	P(A|CT)	0.223	ctA	CCTA	2.65E-03
	5	P(C|TA)	0.197	taC	CCTAC	5.22E-04
	6	P(T|AC)	0.291	acT	CCTACT	1.52E-04
	7	P(A|CT)	0.223	ctA	CCTACTA	3.40E-05
	8	P(T|TA)	0.313	taT	CCTACTAT	1.06E-05
	9	P(A|AT)	0.257	atA	CCTACTATA	2.73E-06
	10	P(T|TA)	0.313	taT	CCTACTATAT	8.54E-07
	11	P(G|AT)	0.208	atG	CCTACTATATG	1.77E-07
	12	P(C|TG)	0.195	tgC	CCTACTATATGC	3.45E-08
	13	P(C|GC)	0.211	gcC	CCTACTATATGCC	7.30E-09
	14	P(C|CC)	0.149	ccC	CCTACTATATGCCC	1.09E-09
	15	P(A|CC)	0.395	ccA	CCTACTATATGCCCA	4.29E-10
	16	P(G|CA)	0.196	caG	CCTACTATATGCCCAG	8.40E-11
	17	P(A|AG)	0.333	agA	CCTACTATATGCCCAGA	2.80E-11
	18	P(A|GA)	0.386	gaA	CCTACTATATGCCCAGAA	1.08E-11
	19	P(T|AA)	0.310	aaT	CCTACTATATGCCCAGAAT	3.35E-12
	20	P(T|AT)	0.335	atT	CCTACTATATGCCCAGAATT	1.12E-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.