Bioinformatique de Base
Introduction to bioinformatic sequence analysis

Costas Bouyioukos[1,2]

[1]MCF Université de Paris
[2]UMR7216 - Epigénétique et Destin Cellulaire, équipe EDCD

Université de Paris

December 9, 2019

UMR7216 Epigénétique & Destin Cellulaire

Outline

Introduction
- Course Details / Logistics

Refresh on Statistics
- Some basic concepts

Significance - Modelling Random Sequences
- Significance
- Monte Carlo

Markov Models
- zero-order Markov model
- first order Markov model

Motifs
- Profiles motifs
- Matrices

Local Score
- Calculate Local Score
- Extreme Value Theory

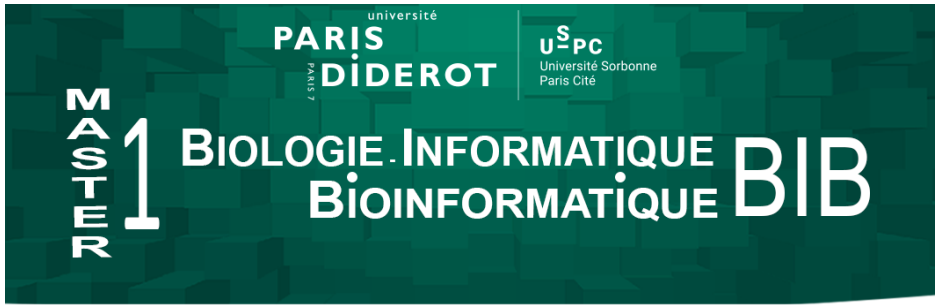
Course Logistics

- The course will run over eight weeks.
- There will be a CM every second Monday at 16h15 (various rooms).
- and a TD-TP on Monday after the CM 16h00 (room 281-89).
- Course instructors:
  - Delphine Flatters MCF, Université de Paris, MTI
  - Costas Bouyioukos, MCF Université de Paris, UMR7216

Syllabus

1. Intro Stats - Random Sequences/Shuffling
2. Markov models - Motif finding
3. Local score - Dynamic programming

Important Dates



1. Website of the course:
   - https://moodlesupd.script.univ-paris-diderot.fr/course/view.php?id=5398
   - The material will all be available AFTER the end of each course (together with resources we discuss in the class).

Intro statistics

Basic definitions

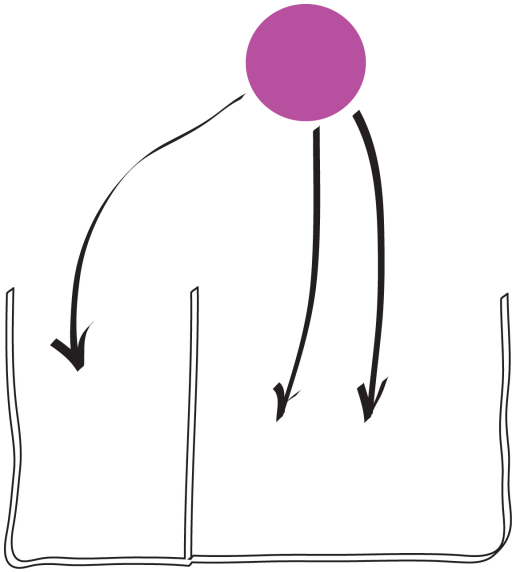
- Population mean, sample mean, median
- Variance, standard deviation, IQR
- Histogram, boxplot, qqplot, scatterplot
- Probability distributions (normal, binomial)
- Probability density function (p.d.f)
- Statistical significance, p value

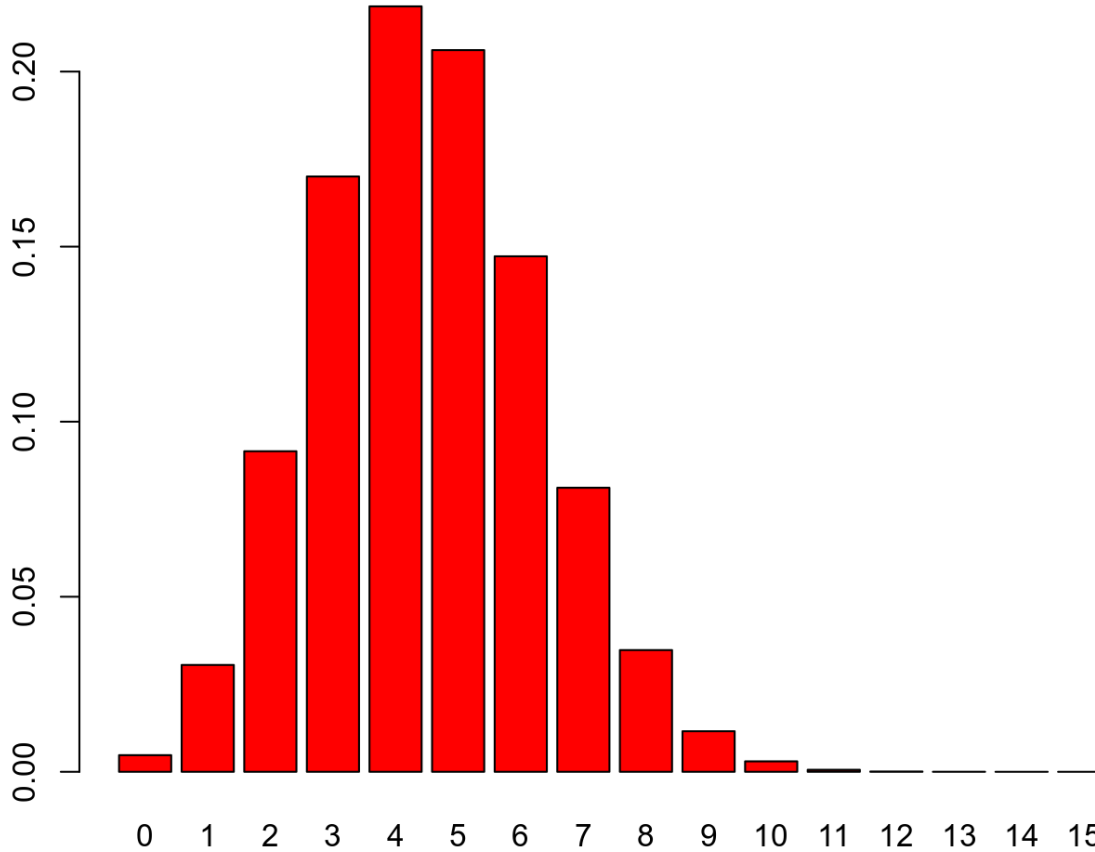
Useful Distributions

1. The normal distribution...
2. Bernoulli trials - The binomial distribution.
3. The Poisson distribution (can approximate the binomial, with one parameter)

Useful Distributions

The binomial distribution





Important Definitions

Statistic: is a single measure of some attribute of a sample, a function of a sample independent of the sample's distribution.

The... p value!!!

What is the p value???
"It is the probability that we observe the same or better statistic/observation IFF the H_0 was true!"

Biological Sequence Analysis

Is the analysis of sequences of biomolecules by computational and statistical methods.

Statistical significance

4 sequences of length 20

CGCGCGACGGGGTATAGCCC
ACGCACGCGTCGTCCAGCTC
CGGCTGCCCTCGGCGGGACC
GGGCTCGGACTGTCCAGACG

CGCGCGACGGGGTATAGCCC
ACGCACGCGTCGTCCAGCTC
CGGCTGCCCTCGGCGGGACC
GGGCTCGGACTGTCCAGACG

Question

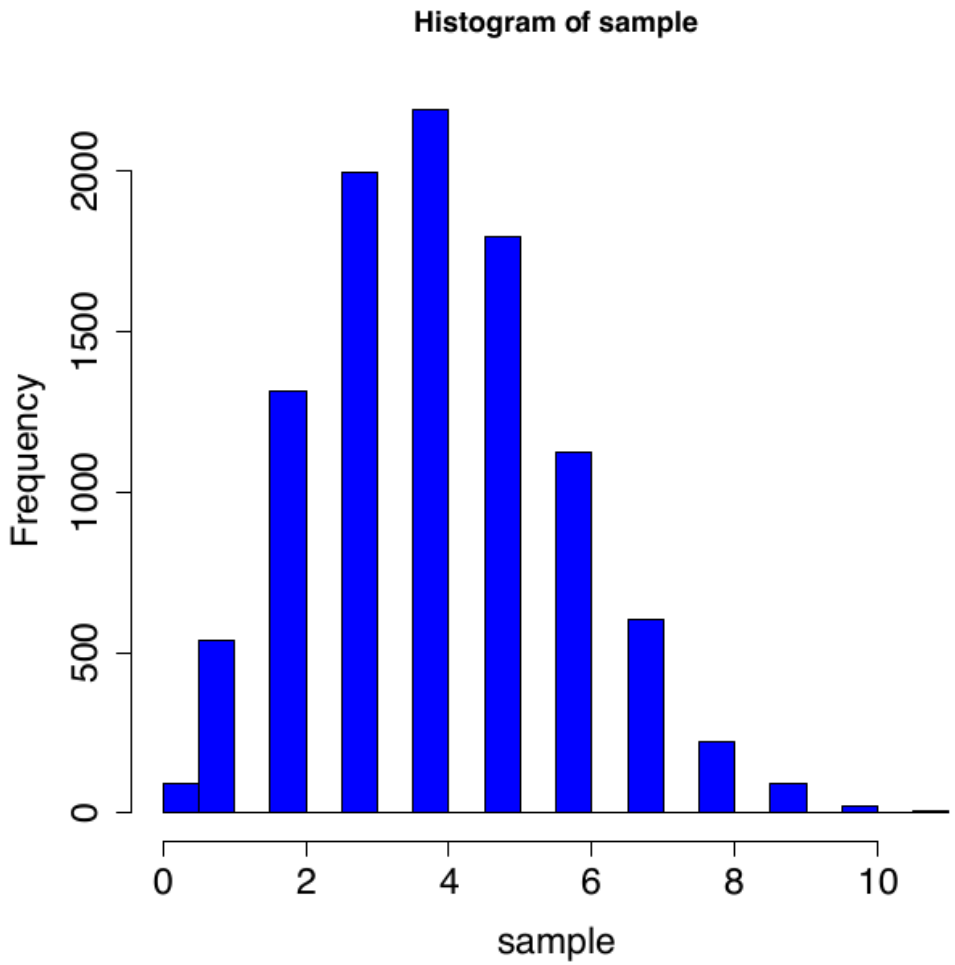
How to assess the biological relevance of this observation?

Answers

1. Negative control (work with other irrelevant sequences)
2. Statistical (background) control (work with simulated sequences)

Background Control I

Simulations (25% A, C, G, T)



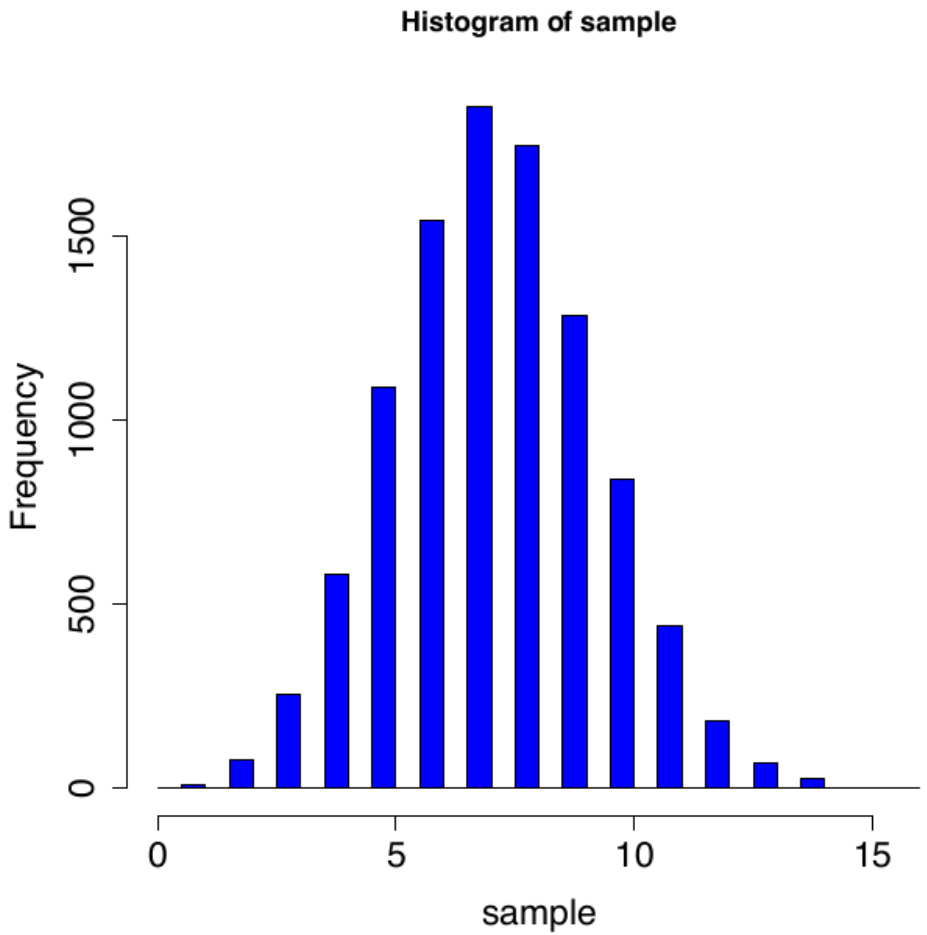
10 or more conserved:

$$\frac{27}{10,000} = 0.27\%$$

BUT, we can easily observe that the frequencies we have used are NOT representative.

Background control II

Simulations (10% A, T; 40% C, G)



10 or more conserved:

$$\frac{1,572}{10,000} = 15.72\%$$

For a more *realistic* background model, the probability to obtain sequences with 10 or more conserved sites is high.

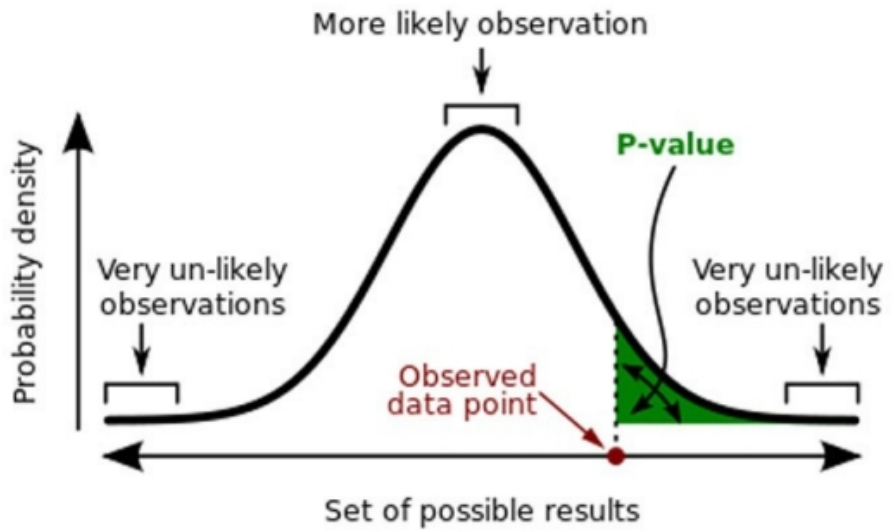
Not significant!

Assessing Significance

Hypothesis testing

What is the p value!

- We begin by defining two hypothesis
  1. H_0, the null hypothesis, the observations are a product of chance.
  2. H_a, the affirmative, the observations are caused by a real effect.
- Identify a statistic that can be used to assess the H_0
- ~~p value: The probability that the H_a hypothesis is not true.~~
- Compare the p value with a predefined threshold α.



A **p-value** (shaded green area) is the probability of an observ
(or more extreme) result assuming that the null hypothesis is t

p value: Is the probability to observe the same (or better) result, given that the H_0 is true.

Assessing significance. Empirical p values

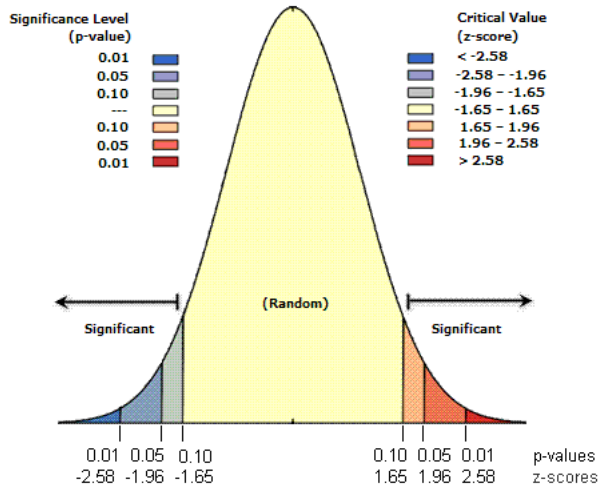
- Two ways to calculate p values.
  1. Analytically, if we know the background (H_0) distribution.
  2. **Empirically**, if we generate the background distribution.
- The course is based on these! We will study models by following the steps:
  1. Generate a sample population (background or null model)
  2. Count how many times a random observation is equal or greater than our observation.

Empirical p value

$$p_{emp} = \frac{|S_{rnd} \geq S_{obs}|}{N_{rnd}}$$

Assessing significance, Z-score

- Two ways to calculate p values.
  1. **Analytically**, if we know the background (H_0) distribution.
  2. Empirically, if we generate the background distribution.
- The course is based on these! We will study models by following the steps:
  1. Generate a sample population.
  2. *Estimate* the mean μ and the standard deviation σ of the population.



$$z = \frac{x_{obs} - \mu}{\sigma}$$

Developing models for biological sequences

Define a "naive" background model

Let $X = X_1 ... X_l$ be a random sequence over the size 4 alphabet $\mathcal{A}$ (for DNA $\mathcal{A} = a, c, g, t$). X is under the M00 model iff:

- All letters X_i are independent and identically distributed (i.i.d.).
- $P(X_i = a) = 1/k$ for all $a \in \mathcal{A}$ (uniform, all equal probability)

Remarks for the M00 model

- Simple and easy to understand
- Often used implicitly as THE random model
- However, it absolutely unrealistic

Testing the significance of the "naive" model

On the whole genome of the HIV

- Take the whole genome sequence of HIV ($l = 9718$)
- Calculate the *expected* nucleotide frequency.
- Find a suitable statistical test to calculate a statistic.

Measuring significance

letter	A	C	G	T
expected under M00	2429.5	2429.5	2429.5	2429.5
observed	3411	1773	2370	2164

- The Pearson's χ^2 statistic gives $\chi^2(3) = 604.4$
- This value is highly significant, however biologically totally uninteresting.

M0 Shuffling Model

Definition: Shuffle model

If $x = x_1 ... x_l$ is an observed sequence, a random sequence $X = X_1 ... X_l$ is under the shuffle model if $X_i = x_{S(i)}$ where S is drawn uniformly from the set of all permutations of $1, ..., l$.

Algorithm to generate shuffled sequences.

(for a sequence X of n letters, indexes $0, ..., n-1$)

for i **from** 0 **to** $n-2$ **do**

j = random integer such that $i \leq j < n$

exchange X_i with X_j

This process has $\mathcal{O}(n)$ complexity and guarantees a random permutation!

Assessing the significance with z scores

- Generate a “large” enough number of permutations (shuffled sequences)
- Compute the statistic of interest.
- *Estimate* the mean μ and the standard deviation σ of the population.
- Compute the z-score of your observation.
- The z-score is expressed in units of σ. (e.g. z-score of 2.3 means 2.3 σ from the mean)
- Find the p value associated with this z-score from the tables of the normal distribution.

Another random sequence model

zero-order Markov model

Definition: M0 model

Let $X = X_1 ... X_l$ be a random sequence over the size 4 alphabet (DNA, RNA). $\mathcal{A}$ is under the M0 model with parameter μ if:

1. All letters X_i are independent and identically distributed.
2. $P(X_i = a) = \mu(a)$ for all $a \in \mathcal{A}$

Proposition: Likelihood

The log-likelihood of the model with regards to a sequence $x = x_1 ... x_l$ is calculated by:

$$L = \sum_{a \in \mathcal{A}} F_x(a) log \mu(a)$$

Maximum likelihood estimator

Corollary

But, we do not know the parameter μ, so we can *estimate* it. The Maximum Likelihood Estimator is simply:

$$\hat{\mu} = \frac{F_x(a)}{\lambda} \forall a \in \alpha$$

Example (x = acctag)

$$\hat{\mu}(a) = \frac{2}{6}, \hat{\mu}(c) = \frac{2}{6}, \hat{\mu}(g) = \frac{1}{6}, \hat{\mu}(t) = \frac{2}{6}$$

$$L = log(\mu(a)\mu(c)\mu(c)\mu(t)\mu(a)\mu(g))$$

$$= log(\mu(a)^2\mu(c)^2\mu(g)\mu(t))$$

$$= 2log(\mu(a)) + 2log(\mu(c)) + log(\mu(g)) + log(\mu(t))$$

Generating a sequence under the M0 model

Given μ and λ follow the process

1. Calculate G the cumulative distribution of μ
2. **for each** i...λ **do:**
3. .... draw a random number $r \rightarrow U[0,1]$
4. .... x_i is the lowest α such as $G(a) > r$

Example

a	a	c	g	t
$\mu(a)$	0.32	0.33	0.17	0.18
$G(a)$	0.32	0.65	0.82	1.00

i	1	2	3	4	5	6
r	0.76	0.00	0.33	0.67	0.63	0.85
X_i	g	a	c	g	c	t

Shuffling vs. M0 model

$x = x_1 \ldots x_\ell$ and observed sequence and we compare the shuffle model to the M0 model (parameters are estimated on x):

- randomly shuffled sequences have exactly the same nucleotide frequencies than x
- random M0 sequences have on the average the same nucleotide frequencies than x
- shuffling is (slightly) slower than drawing under M0
- shuffling requires the original sequence
- shuffling is difficult (but not impossible) to extend to dinucleotide frequencies
- statistical properties of M0 are well known

$\Rightarrow$ no more shuffling from now

Take into account di-nucleotide frequencies

- Count occurrences of di-nucleotides
- Let's see what happens with C (16 in total);

Example (In the HIV1 complete genome $\ell = 9718$)

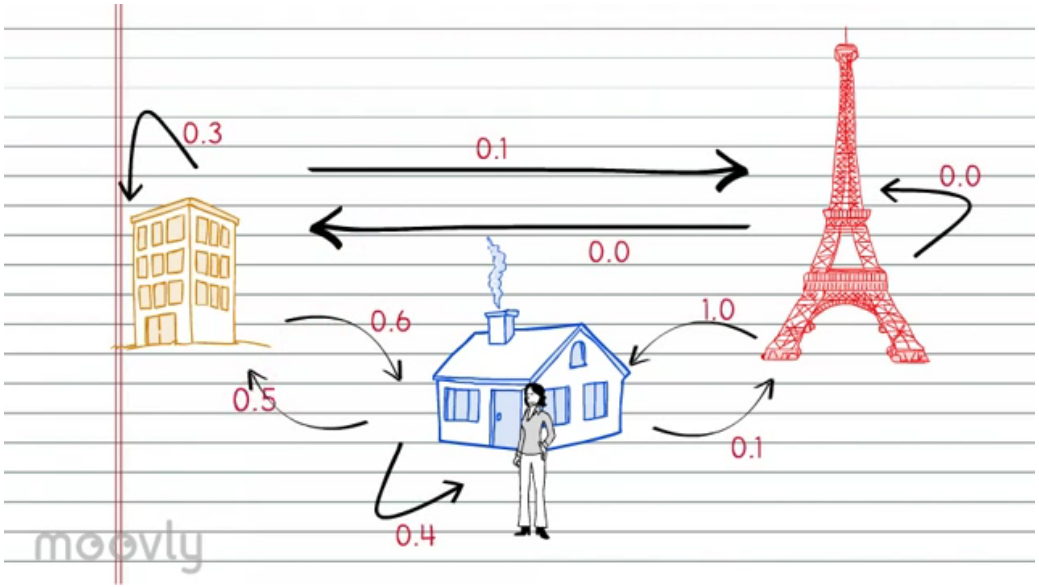
observed	a	c	g	t
a	1112	561	1024	713
c	795	413	95	470
g	820	457	661	432
t	684	342	590	548

$$\frac{F(ac)}{F(a.)} = \frac{561}{3410} = 16.45\% \qquad \frac{F(cc)}{F(c.)} = \frac{95}{1773} = 5.36\%$$

$\Rightarrow$ we need to introduce some dependence

Markov Chains, Markov Models

- Imagine all the possible events in the graph on the right.
- Out of 10 days 5 you go to work.
- 4 you stay at home
- 1 you go to the Eiffel tower.
- We call these the transition probabilities form state A to B.

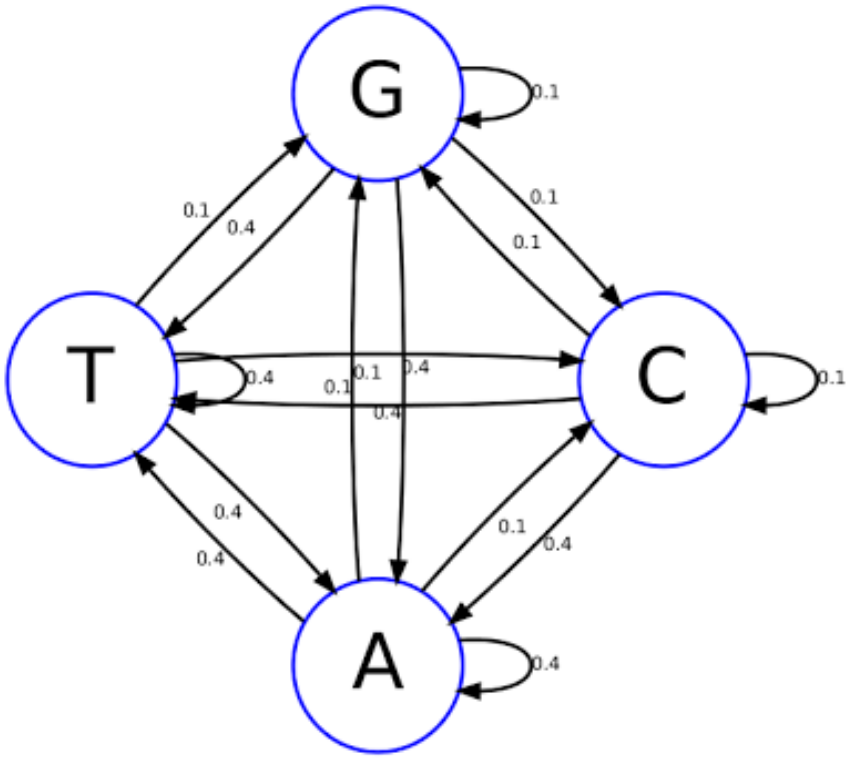


The message!

The probability to be found in a future state, DEPENDS on the actual state!

Markov Model for DNA

- To model a DNA sequence we need 4 states (A, C, G, T)
- Each state has a transition probability that connects it to the other. (i.e. the **next** nucleotide in the sequence.)
- Each move has a *transition* probability
- The easiest way to estimate the transition probability is the MLE.



Formally define the M1 Markov model

Definition M1 Markovian Model

$X = X_1 ... X_l$ is drawn according to the M1 model with starting frequency μ_1 and transition matrix π if:

- $\mathbb{P}(X_1 = a) = \mu_1(a) \forall a \in \mathcal{A}$
- $\mathbb{P}(X_1 | X_1, ..., X_{i-1}) = \mathbb{P}(X_1 | X_{i-1}) = \pi(X_{i-1}, X_i)$

Example (over the binary alphabet $\mathcal{A} = \{a, b\}$)

$$\mu_1 = \begin{pmatrix} 0.5 & 0.5 \end{pmatrix} \quad \pi = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix}$$

means that $\mathbb{P}(X_1 = a) = \mathbb{P}(X_1 = b) = 0.5$ and that

$$\mathbb{P}(X_{i+1} = a | X_i = a) = 0.9 \quad \mathbb{P}(X_{i+1} = b | X_i = a) = 0.1$$
$$\mathbb{P}(X_{i+1} = a | X_i = b) = 0.3 \quad \mathbb{P}(X_{i+1} = b | X_i = b) = 0.7$$

Likelihood and MLE

Proposition

The log-likelihood of the model considering $x = x_1 ... x_l$ is:

$$L = \log\mu_1(x_1) + \sum_{a,b \in \mathcal{A}} F_x(a,b) \log\pi(a,b)$$

Corollary

The MLE for μ_1 and π are hence:

$$\hat{\mu_1}(a) = \begin{cases} 1 \text{ if } a = x_1 \\ O \text{ otherwise} \end{cases} \quad \text{and} \quad \hat{\pi}(a,b) = \frac{F_x(a,b)}{F_x(a.)} \forall a, b \in \mathcal{A}$$

Example, the genome of HIV

Example ($x =$ tggaag... is HIV1 $\ell = 9718$)

observed	a	c	g	t	sum
a	1112	561	1024	713	3410
c	795	413	95	470	1773
g	820	457	661	432	2370
t	684	342	590	548	2164

$$
\begin{aligned}
L &= \log\left(\mu_1(\mathrm{t})\pi(\mathrm{t},\mathrm{g})\pi(\mathrm{g},\mathrm{g})\pi(\mathrm{g},\mathrm{a})\ldots\right) \\
&= \log\left(\mu_1(\mathrm{t})\pi(\mathrm{a},\mathrm{a})^{1112}\pi(\mathrm{a},\mathrm{c})^{561}\ldots\right) \\
&= \log\mu_1(\mathrm{t}) + 1112\log\pi(\mathrm{a},\mathrm{a}) + 561\log\pi(\mathrm{a},\mathrm{c}) + \ldots
\end{aligned}
$$

$$
\widehat{\pi}(\mathrm{a},\mathrm{a}) = \frac{1112}{3410} \quad \widehat{\pi}(\mathrm{a},\mathrm{c}) = \frac{561}{3410} \quad \widehat{\pi}(\mathrm{a},\mathrm{g}) = \frac{1024}{3410} \quad \ldots
$$

Extend Markov models to M_{m}

Definition M_{m} Markovian Model

$X = X_1 ... X_l$ is drawn according to the M_{m} model with starting frequency μ_m and transition matrix π if:

- $\mathbb{P}(X_1 = a_1, ..., X_1 = a_m) = \mu_1(a_1, ..., a_m) \forall a_i \in \mathcal{A}$
- $\mathbb{P}(X_i | X_{i-m}, ..., X_{i-1}) = \pi(X_{i-m}, ..., X_{i-1}, X_i)$

Corollary

The MLE for the transition matrix is:

$$\hat{\pi}(a_1, ..., a_m, b) = \frac{F_x(a_1 ... a_m b)}{F_x(a_1 ... a_m .)} \quad \forall a_1, ..., a_m, b \in \mathcal{A}$$

Model Selection, AIC, BIC

Remark

- The log-likelihood L of M_m grows with m
- For an alphabet of size k the free parameters of a M_m model are $(k-1)k^m$.
- Develop criteria for a tradeoff between m and the number of parameters.

Definition: Penalised likelihood

There are two common criteria for penalising the number of parameters.

- $\text{AIC} = -2L + 2K$ Akaike Information Criterion
- $\text{BIC} = -L + K\log(l)$ Bayesian Information Criterion

where K is the number of free parameters and l the length.

How we choose the right model?

Example (*Escherichia coli K12* $\ell = 4.6$Mb and HIV $\ell = 10$Kb)

- With the AIC:

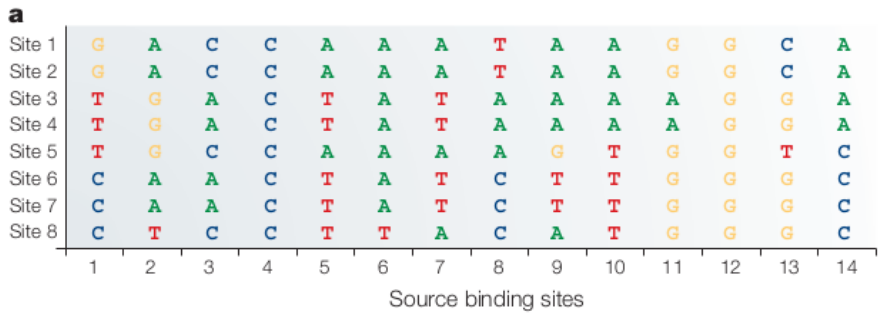
modèle	M00	M0	M1	M2	M3	M4	M5	M6	M7
HIV1	26.95	26.37	25.80	25.68	25.70	26.10	28.03	40.00	106.00
E. coli	12863	12861	12743	12626	12546	12497	12456	12435	12443

- With the BIC:

modèle	M00	M0	M1	M2	M3	M4	M5	M6	M7
HIV	26.95	26.39	25.89	26.03	27.08	31.62	50.10	128.26	459.00
E. coli	12863	12862	12743	12627	12548	12508	12497	12599	13099

Motifs, frequency tables

- Aligning together multiple biologically “interesting” sequences
- One way to describe them is the consensus sequence.
- However consensus does not allow discovery of new sequences.
- Frequency tables!



c Position frequency matrix (PFM)

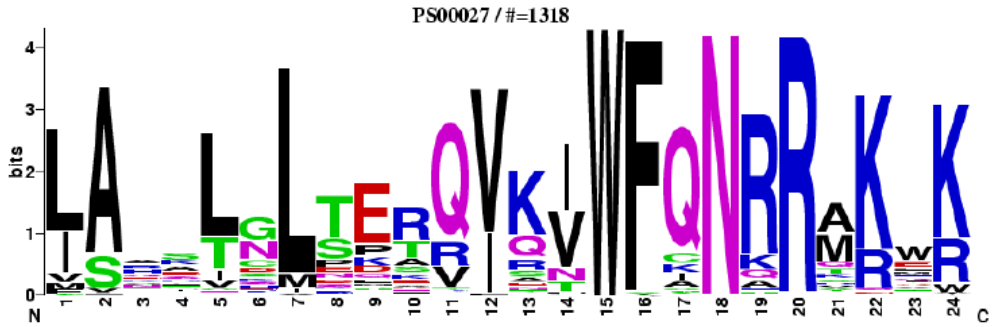
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

By counting the number of occurrences of each AA or nucleotide in each position we obtain the PFM!

Biological motifs. Examples

```
  AGCGG GSTGGTGG CCWGG CTGCAG GAATTC CAYNNNNNRTG
```

```
CPRRRGRQTYTRFQTLELEKEFHF..........NHYLTRRRRIEIAHAL.........
CPRRRGRQTYTRFQTLELEKEFHF..........NHYLTRRRRIEIAHAL.........
VSVRKKRKPYSKFQTLELEKEFLF..........NAYVSKQKRWELARNL.........
-----------------LTKYFNK..........QPYPTRREIEKLAASL.........
-----------------LTKYFNK..........QPYPTRREIEKLAASL.........
-----------------LTKYFNK..........QPYPTRREIEKLAASL.........
SGKRRRRGNLPKESVQILRDWLYEhr........yNAYPSEQEKVLLSRQT.........
SKKRRHRTTFTSLQLEELEKVFQK..........THYPDVYVREQLALRT.........
SKKRRHRTTFTSLQLEELEKVFQK..........THYPDVYVREQLALRT.........
```




```
H-D-[LIVMFY]-x-H-x-[AG]-x(2)-[NQ]-x-[LIVMFY]
```

```
A  [ 3 21 25  0  0 24  1  0 ]
C  [13  1  0  0  5  0  0  0 ]
G  [ 4  0  0  0  0  1  0  2 ]
T  [ 5  3  0 25 20  0 24 23 ]
```

Fig. 1. Various kind of biological motifs. From top to bottom: strings in IUPAC (Cornish-Bowden, 1985) alphabet (DNA), multiple alignment (proteins), sequence logo (proteins), consensus pattern (proteins), and frequency matrix (DNA). Various sources

Pseudocounts and PFMs

- Several zeros in the frequency matrix.
- That is not so representative, as sequences will have zero probability.
- For example the sequence S=GAGGTAAAC will have:

  $p(s|m) = 0.1 \times 0.6 \times 0.7 \times 1.0 \times 1.0 \times 0.6 \times 0.7 \times 0.2 \times 0.2 = 0.0007056$
- A sequence with a G in the 4^{th} position will get prob. 0 given the PFM of the example.
- We correct that by adding a

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}.$$

Log odds and Position Weight Matrix

- Taking the log odds (log of he probability ratio), we obtain the PWM.
- $M_{i,j} = \log_2(PFM_{i,j}/b_k)$
- As background we take the naive $b_k = 0.25$ model for DNA/RNA.
- Both the PFM and PWM assuming independence.
- The PWM is a statistical motif descriptor that captures the variability in sequence patterns.

A	5	0	1	0	0
C	0	2	2	4	0
G	0	3	1	0	4
T	0	0	1	1	1

$\xrightarrow{\text{Log}\left(\frac{f(\text{b,i}) + s(\text{n})}{p(\text{b})}\right)}$

A	1.6	-1.7	-0.2	-1.7	-1.7
C	-1.7	0.5	0.5	1.3	-1.7
G	-1.7	1.0	-0.2	-1.7	1.3
T	-1.7	-1.7	-0.2	-0.2	-0.2

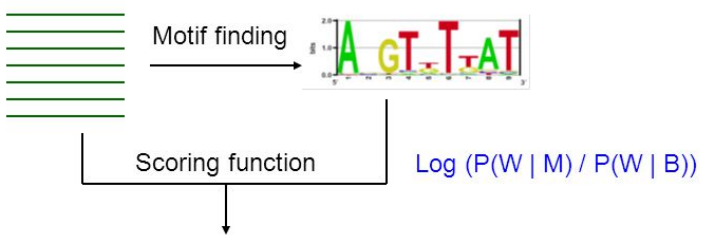
An additional reason to introduce the pseudocounts $s(n)$ is the $-\inf$ entries we obtained in the log odds.

Example, finding motifs

- We scan the sequence of interest by n-sized windows (equal to the motif)
- For each window we calculate the odd probabilities (with the PWM).
- If the score is above a predefined threshold (usually 0 for the PWM) we count.
- We calculate the occurrences in the background model.
- We compute the significance.

To search for new instances

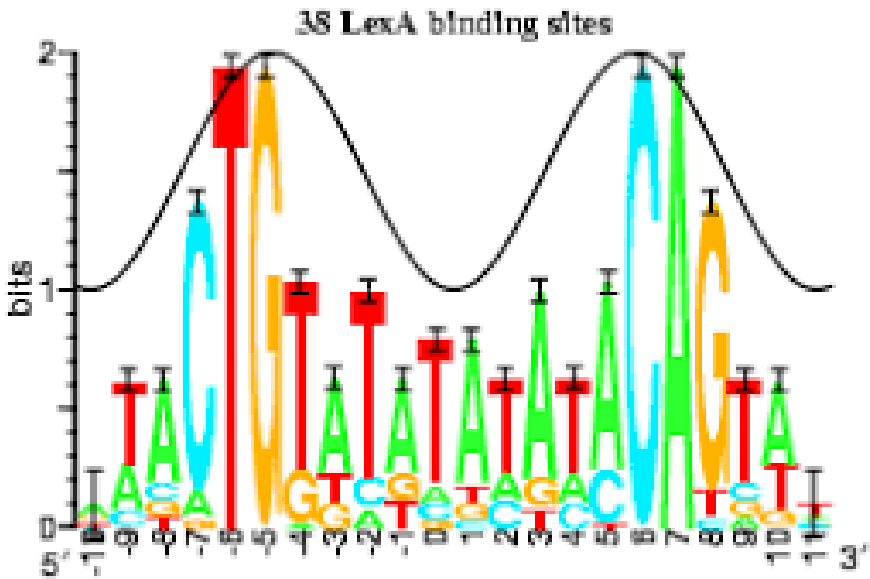
- Usually many false positives
- Score cutoff is critical
- Can estimate a score cutoff from the “true” binding sites



A set of scores for the “true” sites. Take mean - std as a cutoff.
(or a cutoff such that the majority of “true” sites can be predicted).

Motif Logos, Profile HMMs

- An intuitive representation of motifs are the sequence logos (inspired by information theory). The height of ALL letters in one position corresponds to the information content (i.e. the conservation) and the size of EACH its relative frequency.
  $I_i = \log_2(4) - \sum_1^j pfm_j \log_2(pfm_j)$
- As we use $\log_2$ they represent bits of information.



They can be used to visualise the PWM for DNA/RNA as well as for amino acids (very handy as the PWM is a 20×20 matrix.)

How to find homogeneous regions in sequences ?

Example (gc rich regions in DNA)

$$X = \text{aaa}\mathbf{g}\text{aaa}\mathbf{ggg}\text{c}\mathbf{c}\text{a}\mathbf{c}\text{a}\mathbf{gcc}\text{a}\mathbf{g}\text{aaataattt}\mathbf{c}\text{tt}$$

is there one (or more) **gc rich** region in this DNA sequence ?

Example (hydrophobic/phylic regions in proteins)

$$X = \text{YVP}\mathbf{I}\text{S}\mathbf{M}\text{Y}\mathbf{CL}\text{Q}\mathbf{WLL}\text{PV}\mathbf{LLI}\text{PKP}\mathbf{L}\text{N}\mathbf{W}\text{SDGVAS}$$

T, I, L, M, F, W and C are very **hydrophobic** amino-acids. Is there any hydrophobic region in this protein ?

The Sliding Window Approach

Idea

Given a **window size** h, for each $i = 1 \ldots \ell - h + 1$, score the feature of interest (gc content, hydrophobic content) in the **sliding window** $[i, i + h - 1]$.

Example (with $h = 5$ and score = frequency)

a a a **g** a a a **g g g c** a **c** a **c** a **g c c** a **g a** a a t a a t t t t **c** t t

1 1 1 2 2 3 4 4 4 3 3 2 3 3 4 3 4 4 3 2 2 1 0 0 0 0 1 1 1 - - - -

Y V P **I** S **M** Y **C L** Q **W L L** P V **LL I** P **K** P **L** N **W** S D G V A S

1 2 2 3 3 3 3 4 4 3 3 3 3 3 3 4 3 3 2 3 2 2 1 1 - - - - -

The local score approach

Remarks

The sliding window method is very **simple** to understand, have a low **linear complexity**, but suffers **several drawbacks**:

- how to choose the **window size** h ?
- where are exactly the **limits** of our regions of interest ?
- how to choose the **scoring function** ?

$\Rightarrow$ we need **another approach** !

Definition (Local Score)

Given a scoring function $S : \mathcal{A} \to \mathbb{R}$, the **local score** H of the sequence $X = X_1 \dots X_\ell$ is defined by:

$$H = \max_{1 \leqslant i < i' \leqslant \ell} \sum_{j=i}^{j'} S(X_j)$$

Scoring Functions can Vary

Example ($S([\mathbf{gc}]) = +1$ and $S([\mathbf{ac}]) = -1$)

$$X = \text{aaa}\mathbf{g}\text{aaa}\boxed{\mathbf{gggcacacagccag}}\text{aaataattttct}$$

Example ($S([\mathbf{gc}]) = +1$ and $S([\mathbf{ac}]) = -2$)

$$X = \text{aaa}\mathbf{g}\text{aaa}\boxed{\mathbf{gggc}}\text{acacagccagaaataattttct}$$

Example ($S([\mathbf{TILMFWC}]) = +1$ and $S(\{\mathbf{TILMFWC}\}) = -1$)

$$X = \text{YVP}\boxed{\text{IS}\mathbf{M}\text{Y}\mathbf{CL}\text{Q}\mathbf{WLL}\text{PV}\mathbf{LLI}\text{PKP}\mathbf{L}\text{N}\mathbf{W}}\text{SDGVAS}$$

Example ($S([\mathbf{TILMFWC}]) = +1$ and $S(\{\mathbf{TILMFWC}\}) = -2$)

$$X = \text{YVPIS}\mathbf{M}\text{Y}\boxed{\mathbf{CLQWLL}}\text{PV}\mathbf{LLI}\text{PKP}\mathbf{L}\text{N}\mathbf{W}\text{SDGVAS}$$

Brute force: cubic algorithm

Brut force

Score **each of the possible** segments and pick up the best score.

1. ℓ segments of length 1
2. $\ell - 1$ segments of length 2
3. ...
4. L 1 segment of length ℓ

$$\Rightarrow \text{resulting complexity is } O(\ell^3)$$

Proposition

If we denote by H_i the local score of $X_1 \dots X_i$ then we have the following **recurrence relation**:

$$H_0 = 0 \quad \text{and} \quad H_i = \max(0, H_{i-1} + S(X_i)) \quad \forall 1 \leqslant i \leqslant \ell$$

Dynamic programming: linear algorithm

Algorithm

1: $H_0 = 0$
2: **for** $i = 1 \dots \ell$ **do**
3: $\quad H_i = \max(0, H_{i-1} + S(X_i))$
4: **end for**
5: return $H = \max_{1 \leqslant i \leqslant \ell} H_i$

$\Rightarrow$ complexity is $O(\ell)$ in space and time

Example (S([gc]) $= +1$ and S([ac]) $= -1$)

X_i	-	a	a	a	**g**	a	a	a	**g**	**g**	**g**	**c**	a	**c**	a	**c**	a	**g**	**c**	**c**	a	**g**	**a**	a	a	t	a	a	t	t	t	t	**c**	t	t
H_i	0	0	0	0	1	0	0	0	1	2	3	4	3	4	3	4	3	4	5	6	5	6	**7**	6	5	4	3	2	1	0	0	0	1	0	0

we get $H = 7$ and can easily find the **corresponding segment**

Finding subsequences, Significance

How to find a segment

- Let $B = \max 0, H_1, ..., H_n$
- The corresponding segment $[a, b]$ will be:
- $b = \arg\max_j H_j$
- a the biggest j such that $H_j > 0, H_{j+1} > 0, ..., H_b > 0$
- The complexity is $\mathcal{O}(n)$

Significance

- Employ one of the random sequence models.
- Generate a large sample of random sequences.
- Calculate the empirical p value
- $$p_{emp} = \frac{|(S_{rnd} > S_{obs})|}{N_{rnd}}$$
- ... but there is a problem.

Local score summary

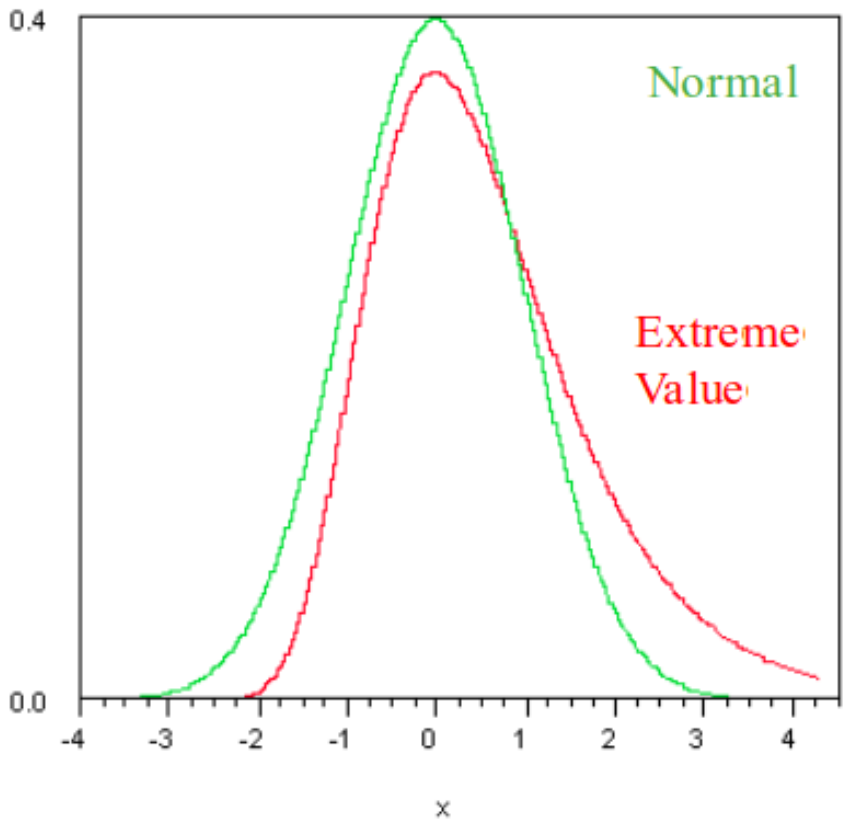
Remarks:

- **simple** and **efficient** linear algorithm
- directly point out **segments** of interest
- can be used with **complex scoring function** (ex: Kyte-Doolittle hydrophobic scale)
- all **suboptimal segments** can be found in $O(\ell)$ thanks to the algorithm from Ruzzo and Tompa (1999)

Conclusion:
far more **elegant approach** than sliding windows from a wide range of problem encountered in sequence analysis.

Extreme value theory

- Computing p values with z-scores lead to p value = 0
- The event we are observing is *rare* thus its distribution can not be capture by the normal.
- We need a better **estimate** of the rare event behaviour.
- **Extreme Value Distribution**
- Normal: $y = \frac{1}{2\pi} e^{-x^2/2}$
- Extreme Value (Gumbel): $y = e^{-(x+e^{-x})}$



p value under the EVD

What we do if p value is zero

- What is the "real" p value, $pv = 10^{-5}, 10^{-15}, 10^{-150}$???
- Zero p value does not reflect an estimate of the probability of the event.
- That can lead to very bad and sometimes catastrophic predictions (insurance, meteorology, etc.)

The cumulative function

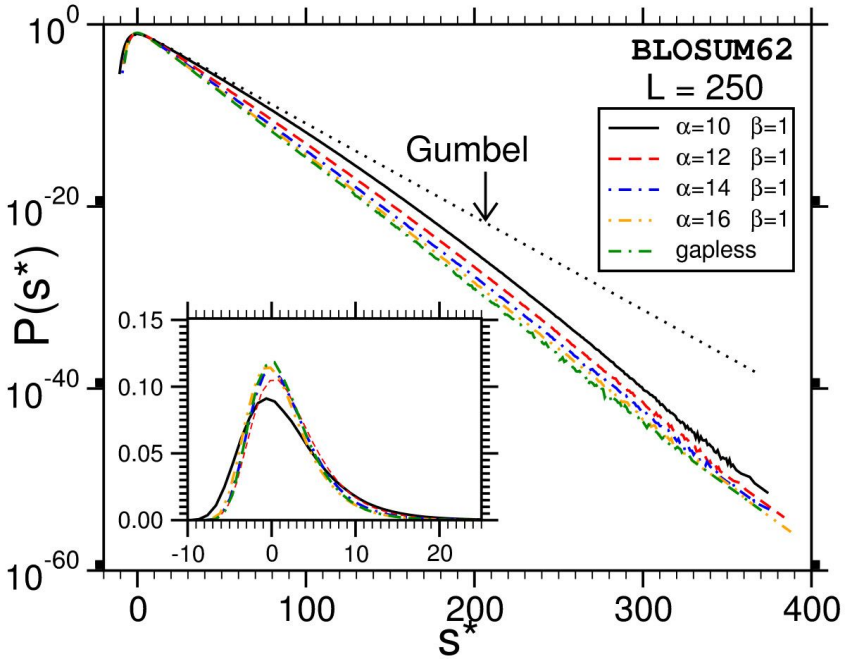
- We will only use the Gumbel distribution and employ its cumulative.
- Cumulative, **adds** the probabilities, thus allows direct access to the p value.
- The cumulative of Gumbel is:

$$F(x) = \exp^{-\exp(-z)}$$

- But how exactly we compute it:

Compute extreme event p values

- Let p value be the $P(b \leq x)$
- By definition local score $S \geq 0$ thus we can:
- Transform the Gumbel cumulative to a log/log linear model:
- $\log(-\log(P(b \leq x))) = ax + b$

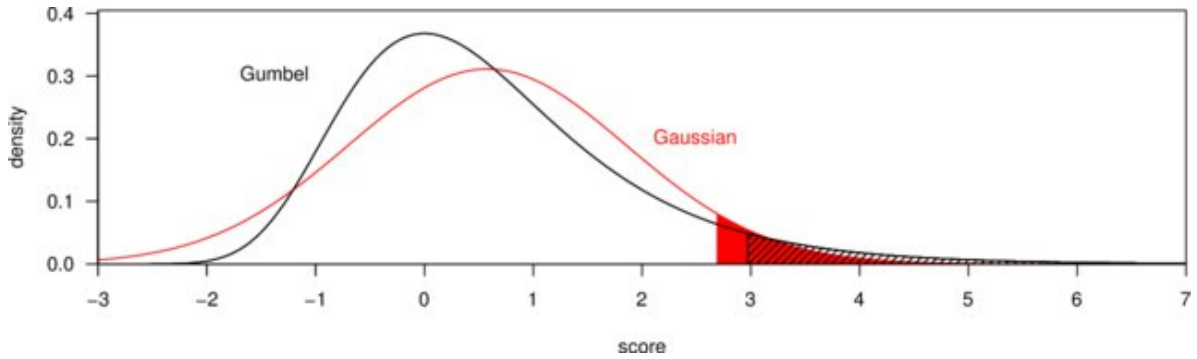


How to calculate the Gumbel p value estimation

Calculate a and b

1. $b = \frac{S_b - \hat{\mu}}{\hat{\sigma^2}}$ then sort(b)
2. $c.d.f = \frac{(1:n)}{n}$ (n in the sample size)
3. Plot $(b, log(-log(c.d.f)))$
4. Compute a and b from the above linear model by linear regression.
5. From the predicted regression line compute the p value for the given b

Relation between Gumbel and Normal distribution p-values



1. Gumbel provides better estimation of “rare events”
2. The Gaussian “z-score” largely OVER-estimates p-values.
3. Gumbel provides reasonable p-values even with low number of samples.
4. Going further: Metropolis-Hastings algorithm.

References

Introductions and material that this course is based.

- Statistical Methods, Arnaud Delrome
- Statistics for Biologists – Points of Significance
- Biological Sequence Analysis
- What is a hidden Markov model
- Sequence Logos

Further Reading

- Fundamentals of Extreme Value Theory
- Tompa and Ruzzo algorithm 1999
- Karlin Altschul statistics