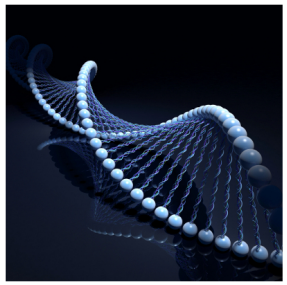
LGBIO2010: Sequence Statistics

Pierre Dupont



UCL – ICTEAM

Outline

Outline

Practical algorithms and statistical methods

Objectives

- Analyze real biological data to *help* make sense out of it
- Use appropriate algorithms and statistical methods to go beyond what can be done "manually"
  - the vast amount of available data can lead to novel insights
  - automating things avoids wasting time on repetitive tasks
  - statistics help you to discover hidden patterns
  - machine learning help you to predict on new data from past observations

Challenges

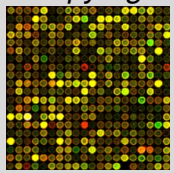
- **Computer scientist/statistician**
  - avoid to apply/design algorithms blindly without considering the underlying biology
- **Molecular biologist**
  - avoid to consider algorithms or softwares as black boxes
  - go beyond the "click on the WEB" methodology
- **Biomedical engineer**
  - reconcile both worlds

A win-win partnership

- The final validation must come from understanding the biology and from a proper field assessment (*e.g.* a clinical trial)
- New algorithms need to be designed to address biological questions
- Further research in machine learning and bio-statistics is required, *e.g.* for personalized medicine

Example

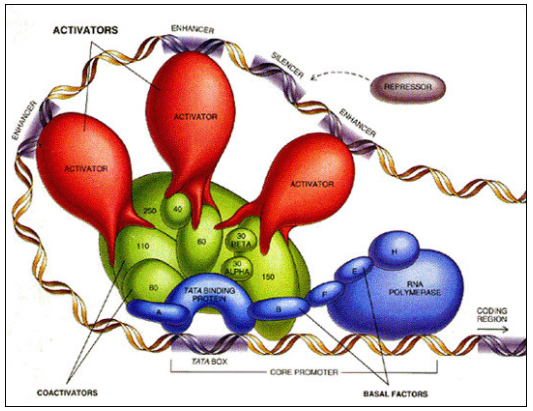
Identify biomarkers for predicting patient response to an immuno-therapy against melanoma



Outline

Abstracting the genome

All models are wrong, some are useful.
(G.E.P. Box, British statistician)




```
ATTATTTTCCCCTCCCACTCC
TCTCATCAATACAACCCCCGC
ACCCAGCACACACACACCGCT
CCATACCCCGAACCAACCAAA...
```

Base composition

Haemophilus influenzae (NC_000907)

- First full bacterial genome ever sequenced (in 1995)
- 1,830,138 bp

A	C	G	T
567623	350723	347436	564241

Well...

$$567623 + 350723 + 347436 + 564241 = 1830023 \neq 1830138$$

Base composition

Haemophilus influenzae (NC_000907)

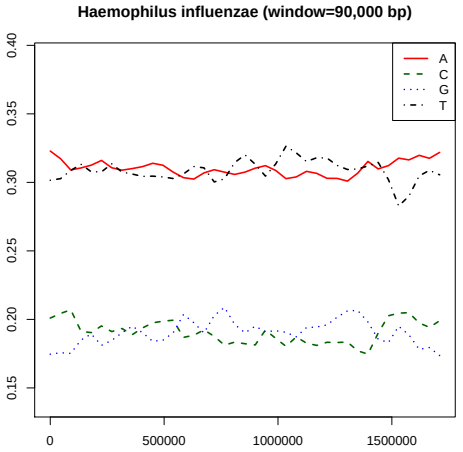
- First full bacterial genome ever sequenced (in 1995)
- 1,830,138 bp

A	C	G	T	K	M	N	R	S	W	Y
567623	350723	347436	564241	14	11	46	10	12	11	11

K	=	G or T
M	=	A or C
N	=	any base
R	=	A or G
S	=	G or C
W	=	A or T
Y	=	C or T

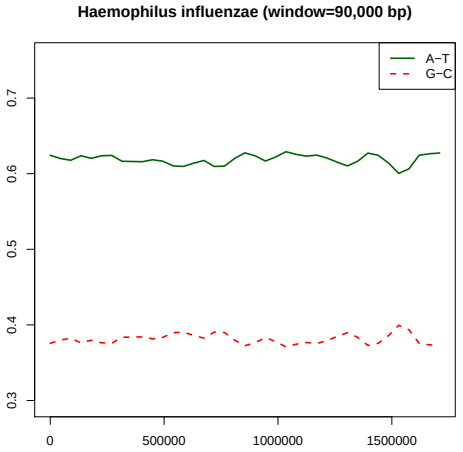
Relative base frequencies

A	C	G	T
31.02%	19.16%	18.98%	30.83%

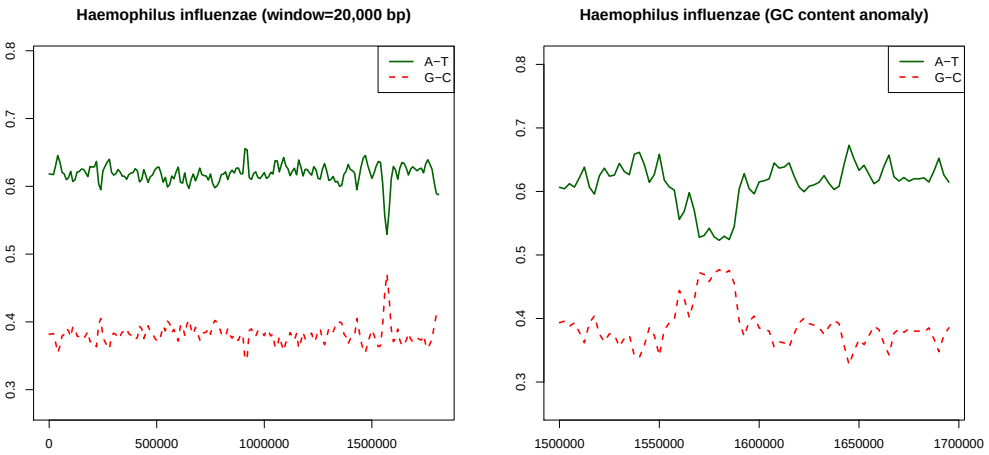


GC-content

A-T	G-C
61.8%	38.2%

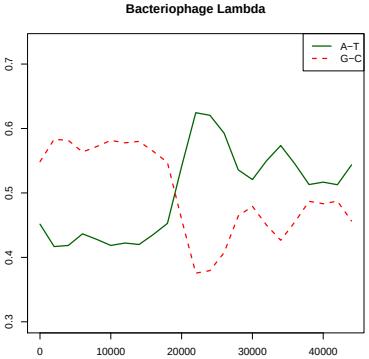


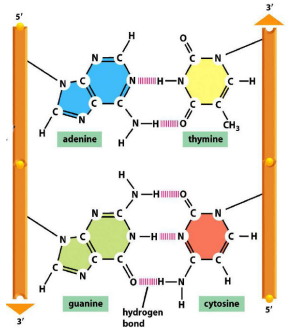
A closer look at GC-content



This anomaly is attributed to an ancient insertion of viral DNA

Bacteriophage lambda GC-content





- *Enterobacteria phage lambda*: a virus that infects bacteria
- First half is GC-rich, second half is AT-rich
- AT-rich regions denature at lower temperatures
- It is believed that the ability to quickly denature DNA facilitates the insertion in the bacterial cell being infected

Dimer frequencies

Haemophilus influenzae

A	C	G	T
31.02%	19.16%	18.98%	30.83%

	*A	*C	*G	*T
A*	**0.1202**	0.0505	0.0483	0.0912
C*	0.0665	0.0372	0.0396	0.0484
G*	0.0514	0.0522	0.0363	0.0499
T*	0.0721	0.0518	0.0656	**0.1189**

- Equally likely dimers would appear $\frac{1}{16} = 0.0625$ of the time
- `AA` and `TT` look particularly frequent
- `CC`, `CG` and `GG` look particularly rare

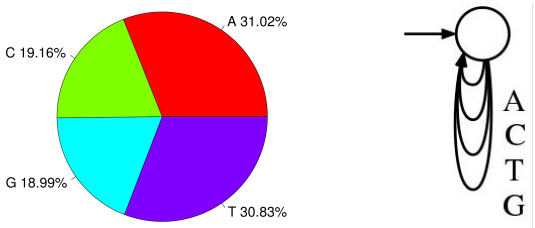
Are those statistical biases informative?

A	C	G	T
31.02%	19.16%	18.98%	30.83%

	*A	*C	*G	*T
A*	**0.1202**	0.0505	0.0483	0.0912
C*	0.0665	0.0372	0.0396	0.0484
G*	0.0514	0.0522	0.0363	0.0499
T*	0.0721	0.0518	0.0656	**0.1189**

- `AA` and `TT` look particularly frequent but `A` and `T` alone are also the most frequent
- what would be the expected frequency of dimers observed by chance?
  - need for a background model to characterize the non-uniform distribution of individual nucleotides

Multinomial background model



- Estimate the probability of each nucleotide independently
  $\hat{P}(A) = f(A) = \frac{\text{number of A's}}{\text{sequence length}} = 0.3102$
- Use those estimates in a random generative model
  `AAGTTGACATAATTTGCT...`
- The expected frequency of a dimer XY:

$$E[f(XY)] = \hat{P}(X)\hat{P}(Y) = f(X)f(Y)$$

Odd ratios

Ratio between observed frequency and expected frequency

$$\frac{f(XY)}{E[f(XY)]} = \frac{f(XY)}{f(X)f(Y)}$$

	*A	*C	*G	*T
A*	1.2490	0.8495	0.8209	0.9533
C*	1.1180	1.0119	1.0892	0.8189
G*	0.8735	**1.4348**	1.0074	0.8525
T*	0.7540	0.8762	1.1202	1.2504

- The most over-represented dimer is `GC` ($\neq$ G+C-content !)
- `CC`, `CG` and `GG` are not particularly rare
- The rarity of `TA` is quite universal
- A statistical test could tell us whether we depart significantly from 1.0

Multinomial model = random permutation

$$\frac{f(XY)}{f_{random}(XY)}$$

- the expected frequency $E[f(XY)] = f(X)f(Y)$ with a multinomial model can be replaced by the observed frequency $f_{random}(XY)$ in a random sequence generated from this model
- equivalent to a random permutation of the original sequence

Original	Permuted
`TATGGCAATTAAAAT`	`CAAGATTGATAATAT`

Odd ratios generalized to k-mers

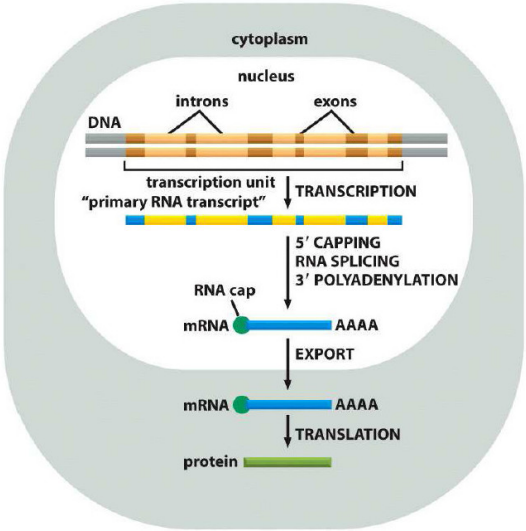
Ratio between observed frequency and expected frequency

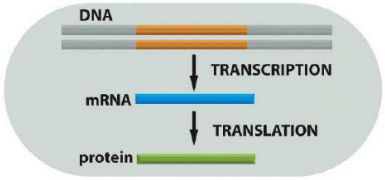
$$\frac{f(X_1 \dots X_k)}{E[f(X_1 \dots X_k)]} = \frac{f(X_1 \dots X_k)}{\prod_{i=1}^{k} f(X_i)}$$

- Useful when looking for frequent patterns of k consecutive characters
- Two most frequent 10-mers in *Haemophilus influenzae*: `AAAGTGCGGT` and `ACCGCACTTT` occur more than 500 times
  Are those 10-mers informative? Addressing this question requires:
  - a more sophisticated background model (*e.g.* permutation respecting codon structure or a set of "reference" sequences)
  - a statistical test to assess significance
  - biological validation

Outline

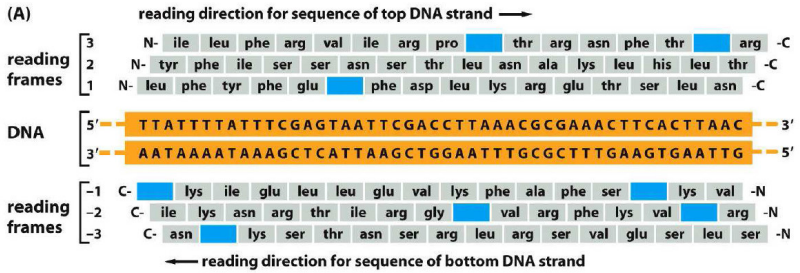
Eukaryotes versus prokaryotes





Illustrations from Molecular Biology of the Cell (© Garland Science 2008)

Open reading frames



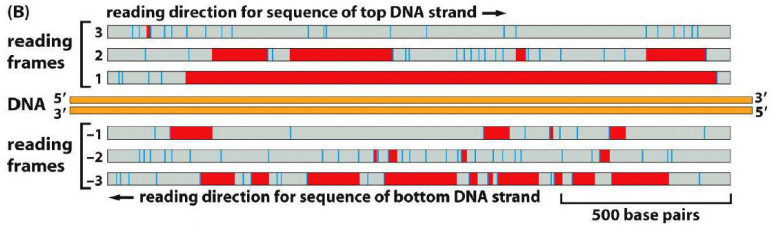


Illustration from Molecular Biology of the Cell (© Garland Science 2008)

Standard genetic code

RNA alphabet

First position (5′ end)	Second position				Third position (3′ end)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Mitochondria

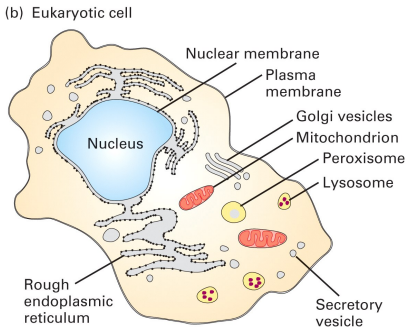


Illustration from Molecular Cell Biology, 5e (© WHFreeman 2004).

- Mitochondria include their own DNA
- The protein synthesis process is similar to the one of prokaryotes
- Some pecularities in the genetic code (hence not fully standard!)

The basic algorithm

ORF finding

repeat *along the sequence*

1. look for a first START
   (or the next START after a STOP on the same frame)
2. look for the next STOP on the same reading frame

- Consider each reading frame:
  - 3 on the forward strand
  - 3 on its reverse complement
    (the same problem, not the same solution!)
- After a START, you may find other codons for `Met` before a STOP
  - those are not true start
  - An ORF is a longest strech of DNA between a START and a STOP, without being interrupted by another STOP on the same frame

Is an ORF actually a coding gene?

1. The DNA stretch found between a START and a STOP codon might be due to chance
2. The ORF might be there but the gene not expressed
   - a trace of the past (or future) evolution
   - we are ignoring all regulations at the transcription/translation levels
3. Some gene sequences do not strictly follow the standard ORF structure
   - we are ignoring some rare exceptions in the genetic code (*e.g.* some fragment of stop codon may play the role of full STOP)

- We address question 1 through a statistical test procedure
- We leave questions 2,3 for a careful biological validation

A concrete example

- Suppose you want to assess the efficiency of a new pain-killing drug versus the current best alternative
- You consider a representative sample of patients receiving drug 1 (the new drug) and another sample receiving drug 2 (the control)
- You assess how many minutes it takes for each patient suffering from a headache to feel better after taking either drug

Sample 1	Sample 2
2	7
3	8
6	5
5	6
1	7

- You want to know whether there is any significant difference of efficiency between both drugs

Simplifying assumptions

Caution

We deliberately ignored **many important related questions**

- what is a representative sample, how big should it be and how to collect it?
- can feeling better be accurately casted into a yes/no answer?
- is the number of minutes before relief a relevant criterion and can it be accurately evaluated?
- cost?
- side-effects?
- existing patents?
- competitors?
- regulations?
- ...

Reformulate the question

1. Compute the sample means m_1, m_2 and sample variance s_1^2, s_2^2
2. Check whether the difference between both means is significant or else should be attributed to randomness in our respective samples

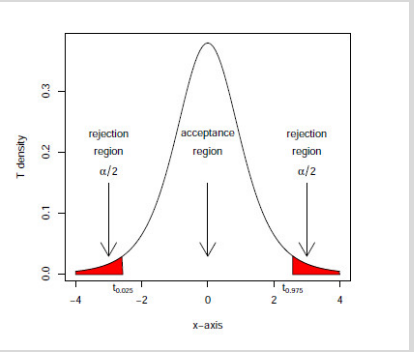
Sample 1	Sample 2
2	7
3	8
6	5
5	6
1	7
$m_1 = 3.4$	$m_2 = 6.6$
$s_1^2 = 4.3$	$s_2^2 = 1.3$
$n_1 = 5$	$n_2 = 5$

Statistical test

- Null hypothesis: $H_0 : \mu_2 = \mu_1$
- Alternative hypothesis: $H_a : \mu_2 \neq \mu_1$
- Test statistics : $T = m_2 - m_1 = \hat{\mu}_2 - \hat{\mu}_1$
- It is known that T approximately follows a Student t distribution

$$T = \frac{(\hat{\mu}_2 - \hat{\mu}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Fix a significance threshold α (*e.g.* 5%)



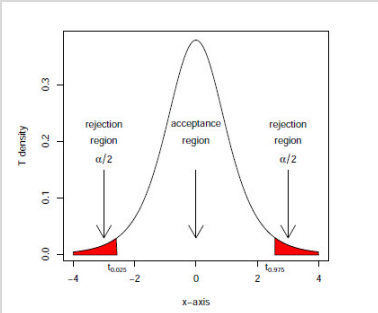
$T = \frac{(6.6-3.4)-(0)}{\sqrt{\frac{4.3}{5}+\frac{1.3}{5}}} = 3.0237 > t_{0.975} = 2.43 \Rightarrow$ **reject** H_0

$\Rightarrow$ claim the drugs are not equally effective

p-value of a test

Instead of fixing *a priori* a significance threshold $\alpha = 5\%$, one can report the p-value of the test

- p-value = the smallest α that would lead to reject the test
- here $T = 3.0237 \Rightarrow$ p-value = 0.02229
  - the true means could still be equal (= the drugs could still be equally effective), but the probability of our conclusion to be wrong is 2.2%
- the lower the more significant the result



- a p-value of max 5% (or 1%) is still often considered for the test to be deemed (highly) significant
- we will revisit this question when discussing multiple testing

Assessing the statistical significance of ORFs

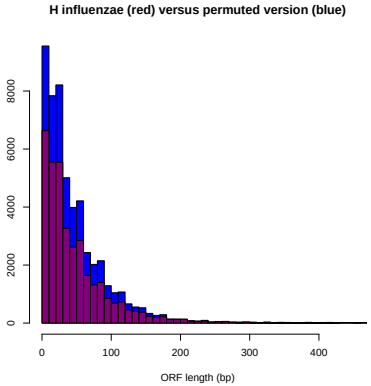
Hypothesis

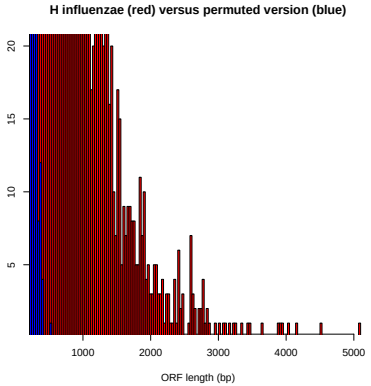
- Significant ORFs in an actual sequence should be longer than ORFs observed by chance
- The NULL model (= control) is typically made of a random permutation of the original sequence

Algorithm

1. Find all ORFs in a random permutation of the original sequence
2. Report the length distribution of random ORFs
3. Accept as significant ORFs, any ORF in the original sequence longer than a prescribed threshold
   - the maximal ORF length observed at random
   - the 99% percentile of the random ORF length distribution
     $\Rightarrow$ permutation test with a $p-$value = 1%

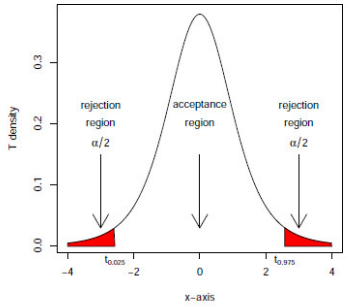
ORF finding in *H influenzae*



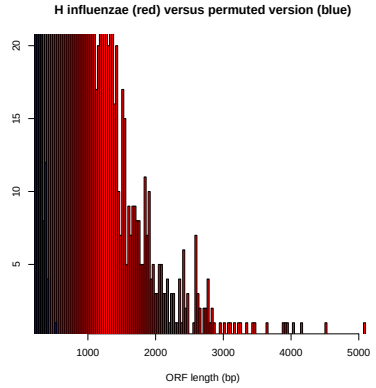


- Maximum random ORF length $= 528 \Rightarrow$ 1252 actual ORFs are longer
  99% percentile $= 204 \Rightarrow$ 2219 actual ORFs are longer
- 1765 genes are annotated (on NCBI), including some pseudo-genes or hypothetical protein coding

Statistical note



Parametric two-tailed
t-test



Non-parametric one-tailed
permutation test

Outline

Your first assignment

1. Register as a group of 2 students on Moodle
   - suggestion: make groups of mixed background and work together
2. Download the latest version of the assignment handout and read it today!
3. Get your hands on real biological data and real software
   - you might have to program a bit (at least some scripting)
4. Do not wait the last minute! Submit your report in due time

Get help on learning R if needed

Or use any other public software by yourself

1. Check Mini-project 1 : Sequence statistics on Moodle
   - watch a short R tutorial video
   - Check a brief introduction to R and go through it step by step
   - If you do know R, check at least section 4 of this brief introduction
   - Attend to a tutorial session in the SIEMENS computer room in the Reaumur building on March 07 at 10:45am
   - Make sure R is installed on your own laptop or use some INGI computer
   - Get help from `vincent.branders@uclouvain.be`
2. Check the first assignment handout and submit the result on Moodle in due time

Enjoy team work