Statistics Applied to Bioinformatics

Probabilities

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

Overview: Probabilities

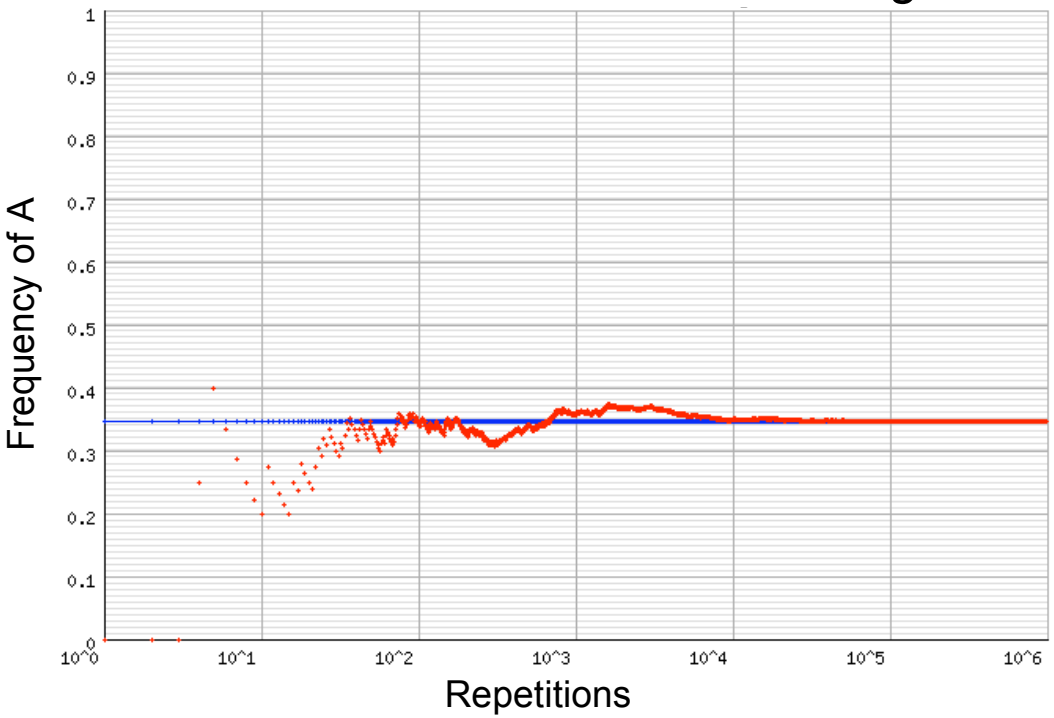
- Definitions
  - Trial, event, probability
  - Stochastic independency
  - Exclusion
- Probabilities of event combinations
  - Mutually exclusive events
  - Complementary events
  - Independent events
  - Non necessarily independent events
- Conditional probabilities

Frequential definition of probability

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

$$0 \leq P(A) \leq 1$$

Selection of random nucleotides in a genome



- Trial
  - a random position is selected in the genome of *Mycoplasma genitalium*
- Event
  - If the nucleotide at this position is a adenine (A), the trial is considered as a success
- 1,000,000 successive trials were performed
- The frequency of success is plotted as a function of the number of trials
- The frequency progressively converges towards the value of 0.35
- The probability is the value that would be obtained with an infinite number of trials

Mutually exclusive events

$$P(A_1 \wedge A_2) = 0 \Leftrightarrow P(A_1 \vee A_2) = P(A_1) + P(A_2)$$

Where $\vee$ is the logical OR

$\wedge$ is the logical AND

- E.g.: Calculating degenerate nucleotide probability
  - $P(W) = P(A \vee T) = P(A) + P(T)$
  - $P(S) = P(C \vee G) = P(C) + P(G)$

Complementary events

$$P(A_1 \vee A_2 \vee \ldots \vee A_m) = 1$$

- E.g.: coding / non-coding sequences in *Mycobacterium genitalium*
  - $P(coding) = 0.902$
  - $P(coding \vee non\text{-}coding) = 1$

    $1\ P(non\text{-}coding) = 1 - P(coding) = 0.098$

- Example: Probability of the degenerate nucleotide N
  - $P(N) = P(A) + P(T) + P(C) + P(G) = 1$

Conditional probabilities

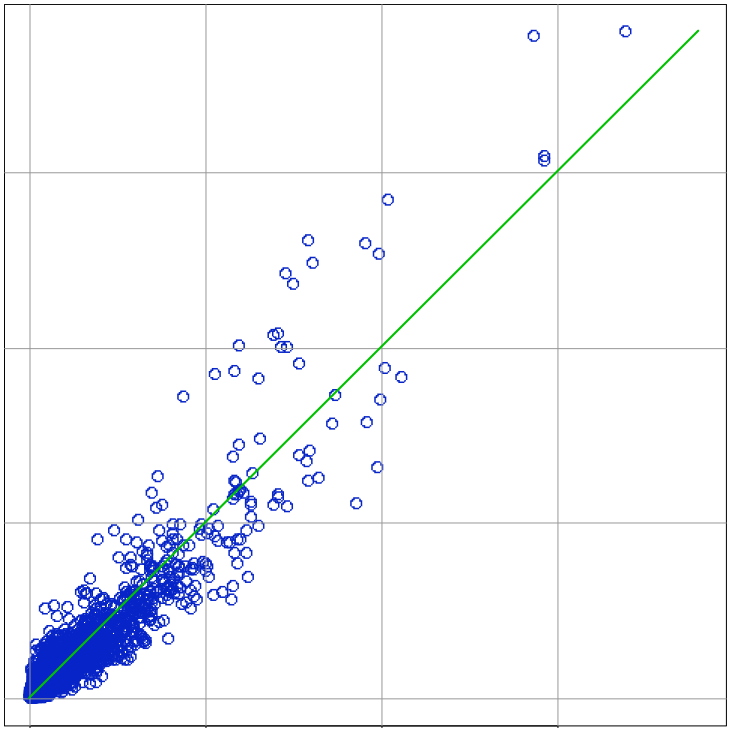
$$P(A|B) = P(A \wedge B)/P(B)$$

$P(A|B)$ is the *probability of A given B*

$$P(B|A) = P(A \wedge B)/P(A)$$

- Example
  - Hexanucleotide probabilities differ between coding and intergenic sequences
  - This has been used for intron/exon discrimination (Claverie, 1986)

Conditional probabilities - hexanucleotide frequencies



hexant	coding	non.coding	ratio
acatgt	0.000113	0.000018	6.22
acgtga	0.000103	0.000018	5.69
acgtaa	0.000102	0.000018	5.64
ctgcag	0.000178	0.000036	4.90
gatatc	0.000236	0.000055	4.32
tctgca	0.000182	0.000045	4.01
...	...	...	...
tatata	0.000123	0.000363	0.34
tcccca	0.000153	0.000472	0.32
agtggg	0.000112	0.000391	0.29
attata	0.000138	0.000536	0.26

	size	*percent*
coding	524	90.2%
non-coding	57	9.8%
genome	581	100.0%

$$P(\text{non-coding}) = 0.098$$

$$P(\text{coding}) = 0.902$$

$$P(\text{attata}|\text{coding}) = 1.38E^{-4}$$

$$P(\text{attata}|\text{non-coding}) = 5.36E^{-4}$$

Conditional probabilities - Multiplication theorem

$$P(A \wedge B) = P(A)P(B|A) = P(B)P(A|B)$$

- If one selects a random position in the genome, what is the probability of falling on a coding attata ?

$$P(B) = P(coding) = \quad 0.902$$

$$P(A|B) = P(attata|coding) = \quad 1.38E^{-4}$$

$$P(A \wedge B) = P(attata \wedge coding) = P(B)P(A|B) = \quad 1.24E^{-4}$$

- If one selects a random position in the genome, what is the probability of falling on a non-coding attata ?

$$P(!B) = P(non\text{-}coding) = \quad 0.098$$

$$P(A|!B) = P(attata|non\text{-}coding) = \quad 5.36E^{-4}$$

$$P(A \wedge !B) = P(attata \wedge non\text{-}coding) = P(!B)P(A|!B) = \quad 5.25E^{-5}$$

- If one selects a random position in the genome, what is the probability of falling on attata ?

$$P(B) + P(!B) = 1 \Rightarrow P(A) = P(A \wedge B) + P(A \wedge !B) = \quad 1.77E^{-4}$$

Conditional probabilities - Baye's theorem

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

$P(B) = P(coding) =$ 0.902

$P(A|B) = P(attata|coding) =$ $1.38E^{-4}$

$P(B|A) = P(coding|attata) = P(A|B)P(B)/P(A) =$ 0.703

$P(!B) = P(non\text{-}coding) =$ 0.098

$P(A|!B) = P(attata|non\text{-}coding) =$ $5.36E^{-4}$

$P(!B|A) = P(non\text{-}coding|attata) = P(A|!B)P(!B)/P(A) =$ 0.297

Stochastic independence

- Two events A and B are said stochastically independent when
  - $P(A|B) = P(A|!B) = P(A)$
  - $P(B|A) = P(B|!A) = P(B)$
- For stochastically independent events, the joint probability is the product of probabilities
  - $P(A \wedge B) = P(A)P(B)$
- E.g.: calculating oligonucleotide probability with a model of independent succession of nucleotides
  - $P(GATAAG) = P(G) * P(A) * P(T) * P(A) * P(A) * P(G)$
  - Note : this is not appropriate for biological sequences, where there are strong dependencies between neighbour nucleotides (see next slide)

Non necessarily independent events

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

$$\begin{aligned} P(A \vee B \vee C) = {} & P(A) + P(B) + P(C) \\ & - P(A \wedge B) - P(A \wedge C) - P(B \wedge C) \\ & + P(A \wedge B \wedge C) \end{aligned}$$

Dinucleotide frequencies – Bernoulli model

- Nucleotide frequencies observed in yeast non-coding sequences
  - $P(A) = P(T) = 0.325$
  - $P(C) = P(G) = 0.175$
- If the residues followed each other independently
- Expected dinucleotide frequencies under the hypothesis of a Bernoulli model
  - $P(AT) = P(A) * P(T) = 0.106$
  - $P(TT) = P(T) * P(T) = 0.106$
  - $P(GT) = P(G) * P(T) = 0.057$
  - $P(CT) = P(C) * P(T) = 0,057$
- Observed dinucleotide frequencies in yeast non-coding sequences
  - $F(AT) = 0.097$
  - $F(TT) = 0.119$
  - $F(GT) = 0.052$
  - $F(CT) = 0,055$

Conditional probabilities – Markov chains

- Nucleotide frequencies observed in yeast non-coding sequences
  - $P(A) = P(T) = 0.325$
  - $P(C) = P(G) = 0.175$
- Observed dinucleotide frequencies
  - $P(AT) = 0.097$
  - $P(TT) = 0.119$
  - $P(GT) = 0.052$
  - $P(CT) = 0,055$
- Conditional probabilities
  - $P(T|A) = P(AT)/P(A) = 0.097/0.325 = 0.298$
  - $P(T|T) = P(TT)/P(T) = 0.119/0.325 = 0.366$
  - $P(T|G) = P(GT)/P(G) = 0.052/0.175 = 0.297$
  - $P(T|C) = P(CT)/P(C) = 0.055/0.175 = 0.314$

Statistics Applied to Bioinformatics

Probabilities - exercises

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

Probabilities - Exercises

- Assuming a DNA sequence with independently and identically distributed nucleotides, calculate the probabilities of the following oligonucleotides.
  - A
  - AA
  - AAAA
  - AAAAAA
  - CCCCCC
  - CACACA
  - CANNTG
  - CACGTK
- Calculate the probabilities of the same oligonucleotides, assuming that nucleotides are independently distributed, but have the following prior probabilities
  - P(A) = 0.31; P(T) = 0.29 ; P(C) = 0.19; P(G) = 0.21

Probabilities - Exercises

- All hexanucleotide frequencies have been measured in a complete genome. If the pentanucleotide starting at position j of this genome is GATAA, what is the probability for the hexanucleotide at the same position to be GATAAG ? Write the formula and calculate the value. The required (and some more) intergenic frequencies are provided below.

AGATAA	0.0005523518490
CGATAA	0.0002483362066
GATAAA	0.0006012194389
GATAAC	0.0002327874281
GATAAG	0.0002733623360
GATAAT	0.0005949999274
GGATAA	0.0002788414294
TGATAA	0.0006226915617
AATAAG	0.0005998866864
ATAAGA	0.0005396166589
ATAAGC	0.0003003135522
ATAAGG	0.0003078658161
ATAAGT	0.0004167072663
CATAAG	0.0002418205280
TATAAG	0.0004486933251

Probabilities - Exercises

- The table below provides the frequencies of start and stop codons in genomic, coding and intergenic sequences respectively.

		Genomic	Coding	Intergenic
Start	**ATG**	0.01825	0.01868	0.01706
Stop	**TAA**	0.02238	0.01991	0.02900
Stop	**TAG**	0.01289	0.01246	0.01408
Stop	**TGA**	0.02012	0.02118	0.01738

- Calculate the probability to observe, in each sequence type, an open reading frame of
  - at least 30 bp
  - at least 300 bp
  - at least 1000 bp
- If we observe an open reading frame of 300bp, what is the probability to be in a coding region, knowing that 72% of the genome is coding ?