Overview of statistics II

Stéphane Aris-Brosou

September 21, 2009

uOttawa
L'Université canadienne
Canada's university

1 Objectives

2 Conditional probabilities

3 Markov chains

4 Conclusions

Context

- last lecture: hypothesis testing
- hypothesis testing is one important activity in statistics
- the other one concerns inference

Context

- last lecture: hypothesis testing
- hypothesis testing is one important activity in statistics
- the other one concerns inference

> *"Inference is the act or process of deriving a conclusion based solely on what one already knows. (...) Inferential statistics or statistical induction comprises the use of statistics to make inferences concerning some unknown aspect of a population. It is distinguished from descriptive statistics."*

(Wikipedia, 2008)

Today's objectives

- conditional probabilities
- introduction to Markov chains (MCs)
- parameter estimation

Today's objectives

- conditional probabilities
- introduction to Markov chains (MCs)
- parameter estimation

What is the probability of a DNA sequence?
(knowing that nucleotides are not independent of each other in real biological sequences...)

How can we use this formulation to learn something about the biological process?

Notations

- let x represent the data
- with discrete-state discrete-time MCs, the data are typically the state of the system at times $1, 2, \ldots, N$
- with DNA and protein sequence analysis, the data are often the amino acid or nucleotide residues that occupy positions $1, 2, \ldots, N$ of a sequence of length N.
- position i of the sequence is denoted x_i
- *e.g.*, $x = TCAGC$
- then $x_1 = T$, $x_2 = C$, $x_3 = A$, $x_4 = G$, $x_5 = C$
- NB. The state space Σ for DNA is $\{A, C, G, T\}$

Representation of a sequence

$$
\begin{aligned}
x_1 &= T \\
x_1\, x_2 &= TC \\
x_1\, x_2\, x_3 &= TCA \\
x_1\, x_2\, x_3\, x_4 &= TCAG \\
x_1\, x_2\, x_3\, x_4\, x_5 &= TCAGC
\end{aligned}
$$

- The sample space of DNA sequences of length N is the set of all possible DNA sequences of this length. This space is often denoted Ω.
- The probability space is the triplet (Ω, A, P), where A is a collection of events and P is a probability measure (also called probability function).

Conditional probability

- For events a and b:
- $P(a, b) = P(a|b)\,P(b) = P(b|a)\,P(a)$
- events a and b are independent iif:
- $P(a|b) = P(a)$
- Exercise: show that this is equivalent to
  - $P(b|a) = P(b)$
  - $P(a, b) = P(a)\,P(b)$
- NB. $P(a, b) = P(b, a)$

Consequence: Bayes' theorem

- An important consequence that will be used in phylogenetics, gene finders *etc.* (bioinformatics)

$$P(a) = \frac{P(a,b)}{P(b|a)} = \frac{P(a|b)P(b)}{P(b|a)}$$

$$P(b|a) = \frac{P(a|b)\,P(b)}{P(a)}$$

- see lecture on Bayesian modeling and Bayesian techniques (Markov chain Monte Carlo [MCMC] samplers);
- more on this in a few minutes...

Two other important results

- **Total probabilities**

  if $b = \{b_1, b_2, \ldots, b_N\}$ is a collection of mutually disjoint events in A satisfying $\Omega = \cup b_i$, $P(b_i) > 0$:

$$P(a) = \sum_i P(a|b_i)\, P(b_i)$$

- **Multiplication rule**

  Let $a_1, a_2, \ldots, a_N$ be a collection of events for which $P(a_1, a_2, \ldots, a_N) > 0$:

$$P(a_1, a_2, \ldots, a_N) = P(a_1)P(a_2|a_1)P(a_3|a_1, a_2)\ldots$$

$$\ldots P(a_N|a_1, a_2, \ldots, a_{N-1})$$

Lets take another little break

Exercise: show that, under the above conditions, $\forall a \in A$ we have:

$$P(b_i|a) = P(a|b_i)P(b_i) / \sum_j P(a|b_j)P(b_j)$$

Lets take another little break

Exercise: show that, under the above conditions, $\forall a \in A$ we have:

$$P(b_i|a) = P(a|b_i)P(b_i) / \sum_j P(a|b_j)P(b_j)$$

Hint: use Bayes' theorem, then the total probabilities theorem, and voila!

Back to DNA sequences...

- If x_1 and x_2 are independent sites, then:

$$P(x_1, x_2) = P(x_1)\,P(x_2)$$

- If x_1, x_2 and x_3 are independent sites, then:

$$P(x_1, x_2, x_3) = P(x_1)P(x_2)P(x_3)$$

- If x_i and x_j are independent $\forall(i, j)$, then:

$$P(x) = P(x_1, x_2, \ldots, x_N) =$$

- **Exercise**: given the above assumptions, what is $P(x_i|x_1, x_2, \ldots, x_{i-1})$?

Markov chains (MCs)

- $P(x_i|x_1, x_2, \ldots, x_{i-1}) = P(x_i)$ for all $i > 0$ defines an MC of order 0;
- $P(x_i|x_1, x_2, \ldots, x_{i-1}) = P(x_i|x_{i-1})$ for all $i > 1$ defines an MC of order 1;
- $P(x_i|x_1, x_2, \ldots, x_{i-1}) = P(x_i|x_{i-1}, x_{i-2}, \ldots, x_{i-k})$ for all $i > k$ defines an MC of order k;
- NB. when the order of the chain is not specified, it is usually an MC of order 1 that people are thinking about.

Parameter estimation: simpler model

- Simplification of our DNA sequence (and of the problem): we consider only the distinction between purines ($R = \{A, G\}$) and pyrimidines ($Y = \{C, T\}$);
- then: $x = TCAGC = YYRRY$
- we also assume that we have a 1^{st} order Markov chain; in this case:

$$P(x) = P(x_1 = Y)P(x_2 = Y|x_1 = Y)$$
$$\times P(x_3 = R|x_2 = Y)P(x_4 = R|x_3 = R)$$
$$\times P(x_5 = Y|x_4 = R)$$

Time-homogeneous chains

- Let E and F denote any two particular states of the Markov chain;
- a time-homogeneous MC is one in which $P(x_i = E|x_{i-1} = F)$ is identical for all possible values of i.

Transition probabilities (math context)

In the purine/pyrimidine case, we assume that for all possible values of i:

$$\begin{cases} P(x_i = R|x_{i-1} = R) &= p_{RR} \\ P(x_i = R|x_{i-1} = Y) &= p_{YR} \\ P(x_i = Y|x_{i-1} = Y) &= p_{YY} \\ P(x_i = Y|x_{i-1} = R) &= p_{RY} \end{cases} \qquad P = \{p_{ij}\} = \begin{bmatrix} p_{RR} & p_{RY} \\ p_{YR} & p_{YY} \end{bmatrix}$$

$$P(x = YYRRRY) = P(x_1 = Y)\, p_{YY}\, p_{YR}\, p_{RR}\, p_{RY}$$

Constraints & degrees of freedom

- Note that:

  $p_{RR} + p_{RY} = 1$

  $p_{YR} + p_{YY} = 1$

- this constraint means that we have two degrees of freedom left when specifying the value of four parameters: if we know p_{RY} and p_{YR}, then we know p_{RR} and p_{YY};

  $P(x|p_{RR}, p_{YY}) =$

  $$P(x_1 = Y)\, p_{YY}\,(1 - p_{YY})\, p_{RR}\,(1 - p_{RR})$$

- **problem**: how to treat $P(x_1 = Y)$?

Stationarity

- First possibility: $P(x_1 = Y)$ is known *a priori*, *e.g.*, $P(x_1 = Y) = 1$ because we know that the first base in x is a pyrimidine;
- another possibility: $P(x_1 = Y)$ contains information that will help us estimate the transition probabilities:
  - the probability that the MC occupied a given state at a particular position will depend on the initial condition and on the transition probabilities
  - if the MC is sampled long after it began, the states of the chain become almost independent of the initial state of the chain
  - if the MC is sampled for an infinite amount of time, the state of the chain will no longer depend on the initial conditions. This is **stationarity**.

Stationarity (contd)

- At stationarity then, the state of the chain will depend only on transition probabilities $P = \{p_{ij}\}$; the probability that the chain is then in state i at a given position is denoted π_i.
- For a stationary MC:

$$\pi P = \pi$$

- or, written differently:

$$\pi_i = \sum_j \pi_j p_{ij}, \forall (i,j) \in \{R, Y\}^2$$

Back to our sequence data

- we now have:
  $P(x|p_{RR}, p_{YY}) = \pi_Y \, p_{YY} \, (1 - p_{YY}) \, p_{RR} \, (1 - p_{RR})$
- this is the probability of the data, given the two parameters of the model: p_{RR} and p_{YY};
- this function is also called the **likelihood** (of the model parameters)
- **maximum likelihood** estimation is done by finding the values of the parameters that maximize the likelihood.

The Bayesian viewpoint

- From a Bayesian perspective, inference is based on a different quantity: the probability density of the parameters given the data, $P(\theta|x)$;
- this is called a **posterior distribution**
- $P(\theta|x) = \frac{P(x|\theta)\,P(\theta)}{P(x)}$
- $P(\theta)$ is a **prior distribution** and must be specified ahead of the analysis;
- $P(x)$ is a constant that can be expensive to compute.

Exercise

- Assume that p_{RR} and p_{YY} are iid. & both follow (i) uniform **or** (ii) beta distributions with parameters (α_R, β_R) and (α_Y, α_Y); express the posterior distribution $P(p_{RR}, p_{YY}|x)$

pdf beta: $p(z|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1}$

$y \in [0,1]$, $\alpha > 0$, $\beta > 0$, $\Gamma(\alpha) = (\alpha-1)! = \int_0^\infty y^{\alpha-1} e^{-y} dy$

- NB. The MAP (Maximum *a posteriori*) estimates are the values of p_{RR} and p_{YY} that maximize the posterior density;
- NB. p_{RR} and p_{YY} are no longer exactly parameters but **random variables**.

Exercise

- Assume that p_{RR} and p_{YY} are iid. & both follow (i) uniform **or** (ii) beta distributions with parameters (α_R, β_R) and (α_Y, α_Y); express the posterior distribution $P(p_{RR}, p_{YY}|x)$

pdf beta: $p(z|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1}$

$y \in [0,1]$, $\alpha > 0$, $\beta > 0$, $\Gamma(\alpha) = (\alpha-1)! = \int_0^\infty y^{\alpha-1} e^{-y} dy$

In a Bayesian analysis, we want to determine

$$P(p_{RR}, p_{YY}|x) = \frac{P(x|p_{RR},p_{YY})\, P(p_{RR},p_{YY})}{P(x)}$$

Exercise

- Assume that p_{RR} and p_{YY} are iid. & both follow (i) uniform **or** (ii) beta distributions with parameters (α_R, β_R) and (α_Y, α_Y); express the posterior distribution $P(p_{RR}, p_{YY}|x)$

pdf beta: $p(z|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} z^{\alpha-1}(1-z)^{\beta-1}$

$y \in [0,1]$, $\alpha > 0$, $\beta > 0$, $\Gamma(\alpha) = (\alpha-1)! = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$

In a Bayesian analysis, we want to determine

$$P(p_{RR}, p_{YY}|x) = \frac{P(x|p_{RR},p_{YY})\, P(p_{RR},p_{YY})}{P(x)}$$

$$P(p_{RR}, p_{YY}|x) = \frac{P(x|p_{RR},p_{YY})\, P(p_{RR},p_{YY})}{\int_{p_{RR}=0}^{1} \int_{p_{YY}=0}^{1} P(x|p_{RR},p_{YY})\, P(p_{RR},p_{YY})\, dp_{RR}\, dp_{YY}}$$

Domains where MCs are used

- pairwise alignments / MSAs
- construction of sequence databases (*e.g.*, Pfam)
- phylogenetics (in particular, very last lecture)
- motif / gene finding
- 3D structure prediction

Conclusions

- probability theory is an important aspect of bioinformatics
- Markov models make it possible to describe local structures of molecular sequences, and make inferences and predictions about parameters of interest (correlations between sites *etc.*)
- Bayesian models enable the description of more complex structures, but often at a significant computational cost.

Reading assignment for Wednesday

- text: pp.143-154 on pairwise alignments

Exercise 10

```
# to run in batch mode, type:
# R CMD BATCH 04_correctd.exe10.R
# for E. coli, the vector of equilibrium base frequencies is (1/4, 1/4, 1/4, 1/4) as on p.50
# here, I use the nucleotide coding used in your text (see pp.46-47)
# E. coli
ori_data = c(6.78, 0.05, 5.99, 0.01, 2.64, 0.03, 0.85, 4.70, 2.15, 10.04, 0.01, 1.76, 5.99,
9.06, 3.63, 1.12)
# M. genitalium
# ori_data = c(0.15, 1.20, 0.18, 0.01, 0.01, 0.39, 4.70, 1.10, 0.34, 1.07, 0.09, 0.61, 1.93,
2.28, 0.05, 0.13)
# here is how to simulate ONE particular sequence (under the null)
p_a = .25 # 0.45 #.25
p_c = .25 # 0.09 #.25
p_g = .25 # 0.09 #.25
p_t = .25 # 0.37 #.25
pi <- c(p_a, p_c, p_g, p_t) # equilibrium frequencies
seqlen <- 1000 # sequence length
x <- c(1, 2, 3, 4) # alphabet
X <- sample(x, seqlen, replace=TRUE, pi)
X
```

Exercise 10

```
# now we count dinucleotide (absolute) frequencies
counter <- numeric(16) # the ordering follows that in Table 2.2, p.50
for(k in 1:(seqlen-1)){ # reading sequence for all dinucleotides
for(i in 1:4){ # first position of dinucleotides
for(j in 1:4){ # second position of dinucleotides
if( (X[k] == i) && (X[k+1] == j) ){
counter[ (4 * (i-1) ) + j ] <- counter[ (4 * (i-1) ) + j ] + 1
} # end of if clause
} # second position of dinucleotides
} # first position of dinucleotides
} # reading sequence for all dinucleotides
counter
# we then compute the statistic X2/c (see p.49, eq.  2.22)
chi2 <- 0
for(i in 1:4){
for(j in 1:4){
obserd <- counter[ (4 * (i-1) ) + j ]
expect <- (seqlen-1)*pi[i]*pi[j]
if(i == j){
c <- 1 + 2*pi[i] - 3*(pi[i])^2
}else{
c <- 1 - 3*pi[i]*pi[j]
}
chi2 <- chi2 + ( ( obserd - expect )^2 ) / expect / c
}
}
chi2
```

Exercise 10

```
# the next step is to simulate a large number of sequences
# and put all the above steps together
Nreps <- 10000
reps <- numeric(Nreps)
pvalues <- numeric(16)
matreps <- matrix(rep(0, 16*Nreps), ncol=16, byrow=F)
for(l in 1:Nreps){
# simulate data
p_a = .25
p_c = .25
p_g = .25
p_t = .25
pi <- c(p_a, p_c, p_g, p_t) # equilibrium frequencies
seqlen <- 1000 # sequence length
x <- c(1, 2, 3, 4) # alphabet
X <- sample(x, seqlen, replace=TRUE, pi)
# count dinucleotide (absolute) frequencies
counter <- numeric(16) # the ordering follows that in Table 2.2, p.50
for(k in 1:(seqlen-1)){ # reading sequence for all dinucleotides
for(i in 1:4){ # first position of dinucleotides
for(j in 1:4){ # second position of dinucleotides
if( (X[k] == i) && (X[k+1] == j) ){
counter[ (4 * (i-1) ) + j ] <- counter[ (4 * (i-1) ) + j ] + 1
} # end of if clause
} # second position of dinucleotides
} # first position of dinucleotides
} # reading sequence for all dinucleotides
counter
```

Exercise 10

```
# compute the statistic X2̂/c (see p.49, eq.  2.22)
chi2 <- 0
for(i in 1:4){
for(j in 1:4){
obserd <- counter[ (4 * (i-1) ) + j ]
expect <- (seqlen-1)*pi[i]*pi[j]
if(i == j){
c <- 1 + 2*pi[i] - 3*(pi[i])2̂
}else{
c <- 1 - 3*pi[i]*pi[j]
}
matreps[l, (4 * (i-1) ) + j] <- ( ( obserd - expect )2̂ ) / expect / c
if( abs(matreps[l, (4 * (i-1) ) + j]) > abs(ori_data[(4 * (i-1) ) + j]) ) pvalues[(4 * (i-1) ) +
j] <- pvalues[(4 * (i-1) ) + j] + 1
}
}
}
pdf(file = "04_correctd.exe10.Rplots.pdf")
par(mfrow=c(4,4))
for(i in 1:16) hist(matreps[,i], nclass=50, main="")
par(mfrow=c(1,1))
dev.off()
pvalues/Nreps
```