# Fundamentals in Bioinformatics Oriented Statistics

**Claudio J. Struchiner**

**May, 2001**

# Outline:

**Human Cytomegalovirus CMV / palindromes**

- distributions (homogeneous Poisson process)
- hypothesis test
- parameter estimation (moments and maximum likelihood)
- Bayes' theorem

**Alignment**

- scoring, markov chains, hidden markov chains

**Phylogenetic Trees**

- bootstrap

# Human Cytomegalovirus CMV / Palindromes

- member of Herpes virus family; potentially life-threatening; incidence varies 30-80%; in immune-depressed causes pneumonia, neurological disorders, gastro-intestinal disease, mental retardation and deafness
- pattern investigation:
  complementary palindromes - sequence of letters that reads in reverse as the complement of the forward sequence (e.g., GGGCATGCCC); cluster of palindromes may indicate the origin of replication which may help finding vaccines or drugs.

- data:
  CMV DNA is 229,354 letters long (*Hemophilus influenzae*-1.8 million, human-3 billion); the longest palindrome is 18 base pairs; computer search algorithms indicate the presence of 296 palindromes 10-18 pairs long; shorter can occur too frequently just by chance and are ignored.
- switch to R and show the data; more info at www.r-project.org
- departures from an uniform random scatter of palindromes across the DNA. how do we find clusters of palindromes? how do we determine whether a cluster is just a chance occurrence or a

potential replication site?

- look for structure: examine locations, the counts, and the spacings between palindromes in nonoverlapping regions of the DNA; simulate random scatter and compare visually; use graphical methods to examine the spacings between consecutive palindromes and sums of consecutive pairs, triplets, etc spacings.
- does the interval with the greatest number of palindromes indicate a potential origin of replication? tight cluster of a few palindromes could easily go undetected if the regions examined are too large; if the regions are too

small, a cluster of palindromes may be split between adjacent intervals and not appear as a high-count interval. the counts for shorter regions will be more variable than those for longer regions.

- THEORY — Distributions
  - Homogeneous Poisson Process
    - describes events arriving along a time scale (phone calls, deaths, radiactive deacay, etc)
    - the underlying rate ($\lambda$) at which points, called hits, occur doesn't change with location (homogeneity)
    - the number of points falling in separate

regions are independent

* no two points can land exactly the same place
* $P(k$ points in a unit interval$) = \frac{\lambda}{k!}^k e^{-\lambda}$; $\lambda$ is the rate of hits per unit area
* adopting the poisson process implies: palindromes are scattered randomly and uniformily acroos the DNA; the number of palindromes in any small piece of DNA is independent of the number of palindromes in another, nonoverlaping piece; the chance that one tiny piece of DNA has a palindrome in it is the same for all tiny pieces of DNA.

– back to the example:

* estimate $\hat{\lambda}$ by the method of moments
  $\Rightarrow \hat{\lambda} = \frac{294}{57} = 5.16$ per 4000 base pairs (could also be estimated by maximum likelihood; in this case yields the same result)
* goodness of fit (switch to R)
* locations and the uniform distribution
  a poisson process on a region can be viewed as a process that first generates a random number, which is number of hits, and then generates locations for the hits according to the uniform distribution. (switch to R)
* spacings and the exponential and gamma distributions

P(the distance between the first and second hits$> t$) = P(no hits in an interval of length t)= $e^{-\lambda t}$

$\Rightarrow$ the distance between successive hits follows the exponential distribution with paramenter $\lambda$; the distance between hits that are two apart follows a gamma distribution with parameters 2 and $\lambda$; the exponential is special case of gamma with 1, $\lambda$; the $\chi^2_k$ is also a special case of the gamma with $\frac{k}{2}$ and $\frac{1}{2}$

$*$ maximum number of hits

$$
\begin{aligned}
&P(\text{maximum count over } m \\
&\text{ nonoverlaping intervals } \geq \text{k}) = \\
&= 1 - P(\text{ maximum count } < k) \\
&= 1 - P(\text{ all interval counts } < k) \\
&= 1 - [P(\text{ first interval count } < k)]^m \\
&= 1 - [\tfrac{\lambda^0}{0!}e^{-\lambda} + \cdots + \tfrac{\lambda^{k-1}}{k-1!}e^{-\lambda}]^m
\end{aligned}
$$

- More Distributions
  - Binomial: if $p$ is the probability of getting a '1' and $1-p$ is the probability of getting a '0', the probability that $k$ out of $N$ tries yield a '1' is

$$P(k \text{ '1's out of } N) = \begin{pmatrix} N \\ k \end{pmatrix} p^k (1-p)^{N-k}$$

where $\begin{pmatrix} N \\ k \end{pmatrix} = \frac{N!}{((N-k)!k!)}$, the number of ways of choosing $k$ objects from $N$.

- Gaussian: $f(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$
- Multinomial: is a generalization of the binomial to the case where the expirements have $K$ independent outcomes with probabilities $\theta_i, i = 1, \cdots, K$
  ex: rolling a die

– Dirichlet: density over probabilities

$$\mathcal{D}(\theta|\alpha) = Z^{-1}(\alpha) \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \delta \left( \sum_{i=1}^{K} \theta_i - 1 \right),$$

$\alpha = \alpha_1 \cdots \alpha_k$ are constants specifying the distribution and the $0 \leq \theta_i \leq 1$ sum up to 1. the multinomial is a distribution over its exponents $n_i$ whereas the dirichelet is a distribution over the numbers $\theta_i$ that are exponentiated; the two distributions are said to be conjugate distriburions.

$Z$ can be expressed in terms of the gamma

function

$$Z(\alpha) = \int \prod_{i=1}^{K} \theta_i^{\alpha_i - 1} \delta\left(\sum_i \theta_i - 1\right) d\theta = \frac{\prod_i \Gamma(\alpha_i)}{\Gamma\left(\sum_i \alpha_i\right)},$$

the gamma is a generalization of the factorial function to real numbers

$$\begin{cases} \Gamma(n) = (n-1)! & n \in \mathcal{N}^+ \\ \Gamma(x+1) = x\Gamma(x) & x \in \mathcal{R}^+ \end{cases}$$

example: loaded die with probability parameters $\theta = \theta_1, \cdots, \theta_6$ and sampling probabilities vectors $\theta$ from a dirichelet parameterised by $\alpha = \alpha_1, \cdots, \alpha_6$; $\alpha_i' = 10, \alpha_i'' = 2 \Rightarrow$ both distributions produce fair dice in average but a

loaded die is more likely under the second set of parameters (greater variability).

- Gamma: conjugate to the poisson; appropriate for modeling the probabilities of rates (analogously to the multinomial/dirichelet conjugate)

$$g(x, \alpha, \beta) = \frac{e^{-\beta x} x^{\alpha-1} \beta^{\alpha}}{\Gamma(\alpha)}, 0 < x, \alpha, \beta < \infty$$

- Extreme value: modeling the breaking-point of a chain (weakest link), assessing the significance of a maximum score from a set of alignments.

- Inference
  - Maximum Likelihood

infer parameters $\theta = \{\theta_i\}$ for model $M$ from a set of data $D$

$$\theta^{ML} = \hat{\theta} = \max_{\theta} P(D|\theta, M)$$

likelihood IS NOT a probability distribution or density; interesting properties like consistency; gives poor results when the data are scanty (3 rolls of a die)— switch to R

– Posterior Probability Distribution

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)}$$

sample from the posterior; choose the maximum a posteriori probability (MAP); take the

posterior mean estimator (PME)

$$\theta^{\mathrm{MAP}} = \max_{\theta} P(D|\theta, M)P(\theta|M)$$

$$\theta^{\mathrm{PME}} = \int \theta P(\theta|n)d\theta$$

– Examples of Bayes' theorem
occasionally dishonest casino: 99% fair, 1% loaded (P(six)=1/2)
pick a die at random and roll 3 times; suppose you get 3 sixes; $P(D_{\mathrm{loaded}}|3 \text{ sixes})$ ?

$$P(D_{\text{loaded}}|3 \text{ sixes}) =$$

$$= \frac{P(3 \text{ sixes}|D_{\text{loaded}})P(D_{\text{loaded}})}{P(3 \text{ sixes})}$$

$$= \frac{(0.5^3)(0.01)}{(0.5^3)(0.01) + (1/6)^3(0.99)} = 0.21$$

– Data Augmentation
Expectation Maximization (EM): maximum likelihood with missing data;
Markov Chain Monte Carlo: Gibbs - sample from the distribution obtained by keeping all variables fixed except one, i.e. the conditional distribution (bugs/winbugs; see

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

# Pairwise Alignment

- insertion/deletion/substitution: (1)what sorts of alignment should be considered; (2)the scoring system used to rank alignment; (3)the algorithm to find optimal scoring alignments; (4)the statistical methods to evaluate the significance of an alignment score.
- random model (R)

$$P(x, y|R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

match model (M)

$$P(x, y|M) = \prod_i p_{x_i y_i}$$

$p_{a,b}$ are givem by a substitution matrix (i.e. BLOSUM50) which makes a statement about the probability of observing $ab$ pairs in real alignments.

- $$\frac{P(x,y|M)}{P(x,y|R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$
$$S = \sum_i s(x_i, y_i), s(a, b) = \log\left(\frac{p_{ab}}{q_a q_b}\right)$$

$s(a, b)$ is the log likelihood ratio of the residue pair $(a, b)$ ocurring as an aligned pair, as opposed to an unligned pair

- given an alignment score, how do we decide if it is a biologically meaningful alignment giving evidence for a homology or, just the best alignment between two entirely unrelated sequences? $P(M|x,y)$ vs $P(x,y|M)$

$$
\begin{aligned}
P(M|x,y) &= \frac{P(x,y|M)P(M)}{P(x,y)} \\
&= \frac{P(x,y|M)P(M)}{P(x,y|M)P(M)+P(x,y|R)P(R)} \\
&= \frac{P(x,y|M)P(M)/P(x,y|R)P(R)}{1+P(x,y|M)P(M)/P(x,y|R)P(R)}
\end{aligned}
$$

let $S' = S + \log\left(\frac{P(M)}{P(R)}\right), \;\; S = \log\left(\frac{P(x,y|M)}{P(x,y|R)}\right)$

$\Rightarrow P(M|x,y) = \frac{e^{S'}}{1+e^{S'}}$

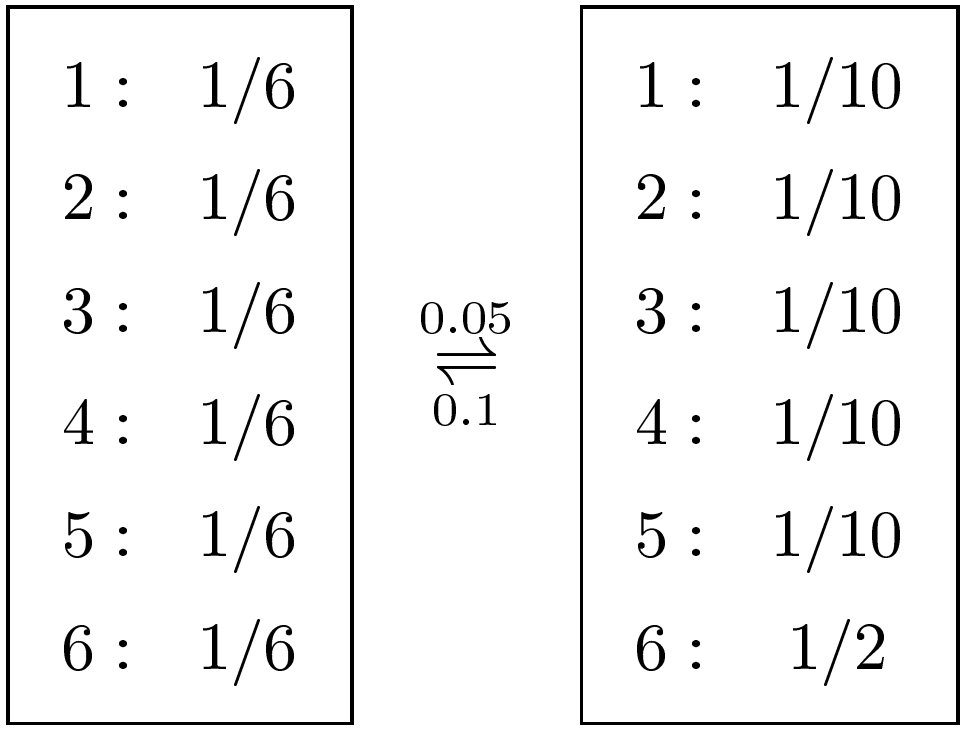
fixed prior does not take into consideration length

of search. the more you search, the more you find false positives; set the prior odds ratio in inverse proportion to the number of sequences in the database N.

- classical approach
  work out the distribution of the maximum of N match scores to independent random sequences (Extreme Value Dist); if small, the observation is considered significant.

# Markov and Hidden Markov Models in a Nutshell

- markov models: describes the probability of one residue following another residue (one state following another state), the transition probabilities: $a_{s,t} = P(x_i = t | x_{i-1} = s) \Rightarrow$ build scores based on estimates of transition probabilities
- hmm: fair vs loaded dice




the markov chain is given by the unseen sequence of fair and loaded dice that emit the observed states according to the emission probabilities in of each state.

ex: CpG islands; CG is chemichally modified by

methylation mutating into T; CpG nucleotides are rare except when the methylation is supressed giving rise to CpG islands around promoters or 'start' regions; given a short stretch how should we decide if it comes from a CpG island and how we find them?

- phylogenetic trees
- bootstrap (switch to R)