BINF 630 – Lecture 4

Introduction to Probability and Statistical Analysis

Saleet Jafri

Introduction to Probability

Suppose we flip a coin 10 times. Each time we will get either a H (heads) or T (tails).

In probability jargon there are 10 trials each with an outcome of H or T.

Assume that the probability of getting a H is p i.e. P(H) = Prob(H) = p. For a fair coin this is 0.5. P(T) = 1 – p as T is the only other possible outcome and the sum of all the probabilities must be 1.

Bionomial Distribution

Now suppose that we want to know the probability of getting k heads in n trials. We can call this event A.

$P(A) = p^k(1-p)^{n-k}\ n!/(k!(n-k)!)$

Example: k =2 n=3

There are 8 possible outcomes, 3 with the desired event – P(A)=3/8

HHH HTH HTT TTH $(1/2)^2(1/2)^1 3!/(2!1!)=3/8$

HHT THH THT TTT

Expected Values

- The expected value is the expected outcome. For example E(# of Heads) 3 coin tosses is

HHH HTH HTT TTH

HHT THH THT TTT

$E(H) = \Sigma\ H\ P(H)$

$= 0(1/8) + 1(3/8) + 2(3/8) + 3(1/8) = 1.5$

$= np = (3)(1/2)$

Expected Values

- The expected value is the expected outcome. For example E(# of Heads) 3 coin tosses is

HHH HTH HTT TTH

HHT THH THT TTT

$Var(H) = E(H^2) - [E(H)]^2$

$= np(1-p) = (3)(1/2)(1/2) = 3/4$

Conditional Probability

- P(B) is the probability that event B occurs
- P(A and B) is the probability that both and A and B occur
- P(A or B) is the probability that either A or B or both occur
- P(A|B) is the probability that occurs given that B has already occurred
- Bayes' Rule

$P(A|B) = P(A \text{ and } B)/P(B)$

comes from

$P(A \text{ and } B) = P(B)\, P(A|B)$

Bayesian Statistics

- Suppose a gene A has two possible alleles A1 and A2 and another gene B has two possible states B1 and B2
- P(B) = P(B1) + P(B2) =1 and P(A) = P(A1) + P(A2) = 1
- Suppose that we know P(B1) = 0.3, then P(B2) = 1 – 0.3 = 0.7
- Suppose that we also know that P(A1|B1) = 0.8 and P(A2|B2) = 0.7.
- Since P(A1|B1) + P(A2|B1) = 1, P(A2|B1) = 1 – 0.8 = 0.2
- Since P(A1|B2) + P(A2|B2) = 1, P(A1|B2) = 1 – 0.7 = 0.3

- How do we find the joint probabilities of A1 and B1?

Bayesian Statistics

	A1	A2	
B1	0.24	0.06	0.3
B2	0.21	0.49	0.7
	0.45	0.55	1.0

- How do fill in this table?

- Use Bayesian Statistics
- P(A1 and B1) = P(B1)P(A1|B1) = 0.3 x 0.8 = 0.24
- P(A2 and B2) = P(B2)P(A2|B2) = 0.7 x 0.7 = 0.49
- P(A1 and B2) = P(B2)P(A1|B2) = 0.7 x 0.3 = 0.21
- P(A2 and B1) = P(B1)P(A2|B1) = 0.3 x 0.2 = 0.06

Other Distributions

- Binominal Distribution
- Poisson Distribution
- Exponential Distribution
- Normal Distribution
- Uniform Distribution
- Gumbel Extreme Value Distribution

Statistical Hypothesis Testing

- Statistical hypothesis testing is used to determine whether a finding is statistically significant.
- A null hypothesis will be formulated and a significance level chosen.
- The test will see if the null hypothesis is rejected in favor of the alternative.

Distributions and Hypothesis Testing

- A. The extereme value distribution
- B. Normal Distribution
- If you want 5% of the area under the curve to be to the right of the dark line x has to be larger for the extreme value distribution.

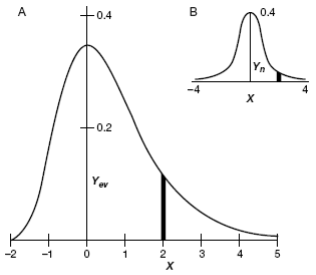

FIGURE 4.2. Probability values for the extreme value distribution (*A*) and the normal distribution (*B*). The area under each curve is 1.

Gumbel Extreme Value Distribution

- The distribution of alignment scores between random sequences follows the Gumbel Extreme Value distribution and not the normal distribution.
- While the area under both curves is 1, the extreme value distribution is skewed.
- For x=1.96 we have a 95% confidence interval (area between -1.96 and +1.96) in the normal distribution.
- For x =3 we have 95% of the area to the left in the extreme value distribution.

Confidence interval

- If the probability that the null hypothesis occurs is less than 5% we say that we reject the null hypothesis with a 95% confidence interval.

Local Alignment for Distantly Related Proteins

- The PAM matrices devised by Dayhoff works for closely related protein sequences.
- For proteins not closely related Dayhoff introduced an additional parameter t which represents the evolutionary time scale
- Hence

$$p_{AB}(t) = P(A, B \mid t)$$

$$p_{AB}(t) = P(A, B \mid t)$$

$P(B \mid A, t)$ is the probability that amino acid A is substituted by B within evolutionary distance t.

$$P(A, B \mid t) = P(B \mid A, t) P(A \mid t)$$

Assume that the amino acid distribution does not change during evolution. Then

$$P(A \mid t) = P(A) = q_A$$

$$P(A, B \mid t) = P(B \mid A, t) q_A$$

$$P(A, B \mid t) = P(B \mid A, t) q_A$$

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

Hence, P(B|A,t) can be estimated from the relative frequency of the pair (A,B) in the known alignment of two sequences s and s' with distance t and from the relative frequency q_A of the amino acid.

This allows us to generate a substitution matrix M' for all pairs of amino acids for a specific evolutionary distance t, using the matrix

$$M = (P(B \mid A, t))_{AB}$$

$$M' = (P(B \mid A, kt))_{AB}$$

$$M' = (P(B \mid A, t))_{AB} \times (P(B \mid A, t))_{AB} \times \cdots \times (P(B \mid A, t))_{AB} \quad k - times$$

$$M' = M^k$$

M is the PAM matrix (Point Accepted Mutation)

Evolutionary Time

- Two sequences s and s' have an evolutionary distance of 1 PAM unit if s was converted to s' by a series of accepted substitutions with an average of 1 accepted substitution per 100 amino acids
- Note that it is accepted substitution – backsubstitutions are ignored in this model.
- In other words a substitution matrix is 1 PAM unit if the expected number of substitutions in a typical protein is 1%.
- Hence $N=M^n$ is a n PAM matrix

Evolutionary Time

- For sequences whose distance is thought to be more than 100 PAM units distance, the PAM250 matrix is used.
- 1 PAM unit is considered to be about 10^7 years, but this can vary.

Finding M

- It is not likely to find homologous sequences s and s' that are exactly 1 PAM distant.
- A scaling factor λ was chosen so that the matrix generated by

$$P(B \mid A,t) = \lambda \frac{n_{AB}(s,s')}{q_A} \quad for\ B \neq A$$

$$P(A \mid A,t) = 1 - \sum_{B \neq A} P(B \mid A,t)$$

is 1 PAM.

Caveats

- Since it is difficult to extrapolate to distantly related proteins from closely related units other substitution matrices have been proposed such as BLOSUM.
- BLOSUM is better for distantly related proteins.

What is the E value

- The E value measures the significance of an alignment.
- Suppose you align a test sequence with 10,000 sequences one of which matches the test sequence (called the related sequence)
- All the sequences will have a alignment score associated with them.
- The E values gives an estimate of the number of unrelated sequences that have a score as high as the related sequence score.
- The lower the E value the better the alignment. E=0.05 corresponds to the 95% confidence interval. However for truly related sequences $E < 10^{-10}$.

Gumbel Extreme Value Distribution

The extreme value distribution is described by

$P(S<x) = \exp[-e^{-x}]$

$P(S>=x) = 1 - \exp[-e^{-x}]$

To account for alignment scores it is written as

$P(S>=x) = 1 - \exp[-e^{-\lambda(x-u}]$

λ is the mode (highest point) and u is the decay (or scale). These are found through parameter estimation.

Parameter Estimation

Method of Moments – The mean and standard deviation is calculated for random alignment scores for a given sequence length range. The parameters are calculated as

$\lambda = \pi/(\sigma \text{ SQRT}(6) = 1.2825\ \sigma$

$u = \mu - \gamma / \lambda = \mu - 0.4500\ \sigma$

where γ is Eulers constant

Parameter Estimation

- Maximum Likelihood Estimation – statistical methods that chooses the parameters so that they are the most likely to explain the data.
- Poisson Approximation Method – This methods aligns a few pairs of random sequences but fines a large number of possible alignments. This is very fast using dynamic programming as as soon as the scoring matrix is built all possible alignments are there. Of these a number of alignments above a certain threshold are retained. The average number of alignments above a threshold yield a value of s which can be used to derive the parameters using a Poisson distribution.

Significance of Local Alignment

The extreme value distribution can be used to calculate the significance of a local alignment.

Assume that there are two sequences about 250 amino acids long and they are aligned by Smith-Waterman using the PAM 250 substitution matrix and a high gap score to omit gaps.

Suppose the following alignment was found

FWLEV EGNSMTAPTG
FWLDVQGDSMTAPAG

Significance of a Local Alignment

A significant alignment between unrelated or random sequences will have a score of

$S = \log_2(nm) = \log_2(250*250) = 16$ bits

The score of the above alignment is 73 using the PAM 250 matrix (Fig 3.14)

This must be converted into bit units by the conversion factor units = 1/3 bits yielding 73/3 = 24.3 bits.

This greater than 16 bits by 8.3 bits so the alignment is very significant.

Significance of Local Alignment

A more complicated method by Altchul and Gish provide estimates of parameters K=0.09 and λ=0.229 for the PAM250 matrix.

Equations 23 yields

$S' = \lambda S - \ln(Kmn)$

$= 0.229*24.3 - \ln(0.09*250*250)$

$=16.72 - 8.63 = 8.09$ bits

Significance of Local Alignment

Equation 24 yields the probability

$P(s'>8.09) = 1 - \exp[-e^{-8.09}] = 3.1 \times 10^{-4}$

The probability can also be calculated by equation 26

$P(s'>8.09) = -e^{-8.09} = 3.1 \times 10^{-4}$

FASTA and BLAST

- FASTA and BLAST also calculate significance of the search results alignments
- FASTA uses alignment scores between unrelated sequences to calculate the parameters of the extreme value distribution
- BLAST calculates estimates of the statistical parameters based on the socring matrix and sequence composition.

Bayes Block Aligner

- Software was developed by Zhu et al (1998) that slides a sequence along another to find the highest ungapped regions and blocks.
- Instead of using a substitution matrix and gap scoring system a Bayesian statistical approach is used.

Bayesian Evolutionary Distance

How can we tell how closely related two sequence are?

Apply different PAM matrices starting with PAM1 and compare the odds scores. When the odds are sufficiently high it is a better esitmate fo the true evolutionary distance.

Homework

- Problem 1 on page 160