Bioinformatics – Lecture Notes

Announcements

On February 21 and 26 I will be out of town. Dr. Mehmet Candas will deliver those two lectures. He will give a talk on bioinformatics from the biologists perspective.

Class 11 – February 19, 2002 -

1. Evolutionary Problems
   a. sequence
      i) For example, the amino acid substitution rate per site per year (λ) is 5.3 x 10^{-9} for guinea pig but only 0.33 x 10^{-9} for other organisms.

      ii) The *evolutionary time* is the average time to produce one substitution per 100 amino acids

$$T_u = \frac{1}{100\,\lambda}$$

      Example – There are 2 differences in a sequence of 100 amino acids when comparing calf and pea histone H4. Since plants and animals diverged 1 billion years ago, T_u= 0.5 billion years

$$\lambda = \frac{1}{100 T_u} \approx 10^{-11}$$

      iii) probability of substitution – several way to calculate it. The best way is using the PAM matrices.

   b. Nucleotide sequences
      i) different from amino acid sequences due to redundancy in the genetic code (ie several codons can code for a particular amino acid.
      ii) Most substitutions in the 3rd position are synonomous (UC* is the RNA coding for serine – the corresponding DNA would be AG*). Since evolution should depend on function and this is conferred by the amino acid sequence, it has been suggested that the "molecular clock" should be based on the substitution rate in the third position of the codon. In fact, in the fibrinopeptides, this is as high as the amino acid substitution rate.

iii) In the definition of PAM matrices, one assumes a discrete Markov Chain, with the PAM matrix being the transition matrix for the Markov Chain.

c. Markov chains (page 38-43 text)

i) Assume that we have a process that has discrete observable states $x_1, x_2, \ldots$. When we monitor this over time we get a sequence of the states occupied $q_1, q_2, \ldots$. where q_i = any of $x_1, x_2, \ldots$. This sequence is a Markov Chain. Note that while there can be an infinite number of states, the Markov chain has a countable number of elements.

ii) Another property of a Markov process is that "history does not matter". This means that the state assumed at time t+1 depends on the state assumed on t (not on any other previous state). This is called the Markov property. Let $X = \{X_n, n = 1, 2, \ldots\}$ be a discrete time random process with state space S whose elements are $s_1, s_2, \ldots$ X is a Markov chain if for any $n \geq 0$, the probability that X_{n+1} takes on any value $s_k \in S$ is conditional on the value of X_n but does not depend on the values of $X_{n-1}, X_{n-2}, \ldots$. The one-time-step transition probabilities

$$p_{jk}(n) = \Pr\{X_n = s_k \mid X_{n-1} = s_j\} \; j,k=1,2,\ldots \; n = 1,2, \ldots$$

Since X_0 is a random variable called the initial condition,

$$p_j(0) = \Pr\{X_0 = s_j\} \quad j=1,2,\ldots$$

iii) Transition matrix – put the p_{jk} into a matrix **P**

iv) Markov Process $\leftrightarrow$ Markov Chain

v) A sequence of amino acids can be thought of as a Markov chain.

vi) Persistent – $\sum_{n=0}^{\infty} p_{i,i}(n) = \infty$

vii) Transient – not persistent

viii) Stationary Markov process – the probabilities $p_{jk}(n)$ do not depend on n, that is they are constant. Another way of saying this is an initial distribution $\boldsymbol{p}^*$ is said to be **stationary** if $\boldsymbol{p}^* P(t) = \boldsymbol{p}^*$.

ix) irreducible – every state can be reached from every other state

Question: What is the sum of each column and each row?

Question: Why does P(t+s) = P(s)P(t)

Question: If a Markov process is stationary and P is the transition matrix, what is P^2

d. Application of Markov processes to evolutionary models
   i) The PAM matrix has its substitution probabilities determined from closely related amino acid sequences, it assumes that the substitutions have occurred through one application of the transition matrix (ie no multiple substitutions and a given site) and assumes that evolutionary distance results from repeated application of the same PAM matrix.
   ii) A better evolutionary model is needed. (text p 140-144) This requires the use of a continuous Markov process rather than a discrete Markov chain. This still has the Markov property.

d) Clustering Methods