Bioinformatics – Lecture Notes

Announcements

Happy Valentine's Day

On February 21 and 26 I will be out of town. Dr. Mehmet Candas will deliver those two lectures.

Class 10 – February 14, 2002 -

1. Steiner Sequences – This is related to multiple sequence alignment. In a method of DNA sequencing called *single-molecule DNA sequencing*, a single stranded DNA molecule is cut a single base pair at a time. The freed base flows down a glass tube by an optical sensor that determines the base. This technique has errors especially near the ends of the DNA. If this method is repeated many times, many copies of erroneous DNA are generated. This method computes an alignment of these sequences to find the actual sequence. Go over the algorithm.
    a. Divide the sequences into groups of three.
    b. Using Dynamic Programming to align each set of three sequences using the cost function given before
    $W(x,y,z) = 0$ x=y=x
    1 it two are the same
    0 if all are different
    c. Find the consensus sequence by taking the majority letter in each column.
    d. Remove all gaps
    e. Repeat for each set of three

2. Genomic Rearrangements
    a. Intra-chromosomal events
        i) *inversion* – flipping a sequence 5'-3' $\rightarrow$ 3'-5'
        ii) *duplication* – repeating a sequence more than once
        iii) *tranposition* – portion of the chromosome is broken and placed elsewhere

    b. Inter-chromosomal events
        i) *reciprocal translocation* – end segments not including the centromere break off and exchange positions
        ii) *chromosomal duplication* – double the number of chromosomes
        iii) *fission* – break a chromosome into two
        iv) *fusion* – join two chromosomes

    c. Abstraction of genes

i) chromosomes are unordered sets of genes
ii) *synteny* – the partitioning of genes into chromosomes
iii) *syntenic* – genes lying in the same chromosome
iv) *syntenic distance* – If two genes have a common ancestor, how many fissions, fusions, reciprocal translocation events are necessary to transform one gene into another?
v) NP complete problem – a class of computational problems for which no efficient solution algorithm has been found (ie traveling salesman problem, graph covering problem)
vi) Example from book

3. Evolutionary Problems
    a. The problem –
        i) The fossil record suggests that modern man diverged from apes about 5-6 million years ago. Modern *Homo sapiens* emerged between 100,000-60,000 years ago
        ii) DNA and sequence alignment by Paabo support this.
        iii) Work based on mitochondrial DNA by Wilson et al suggest the modern man emerged only 200,000 years ago with the divergence into different races 50,000 years ago
            1. mitochondrial DNA circular
            2. maternal inheritance
            3. 10x faster mutation rate than nuclear DNA

    b. To understand the data we must understand the methods behind phylogenetic trees or evolutionary trees
        i) Clustering methods
        ii) Maximum likelihood methods
        iii) Quartet puzzling

    c. Preliminaries
    Taxon (*taxa* plural) or *operation taxon unit* is a entity whose distance from other entities can be measures (ie species, amion acid sequence, language, etc.)

    Comparisons are made on measurements or assumptions concerning rates of evolutionary change. This is complicated by *back mutations*, *parallel mutations*, and variations in mutation rate. We will only consider substitutions.

    d. Amino acid sequence

i) For example, the amino acid substitution rate per site per year (λ) is 5.3×10^{-9} for guinea pig but only 0.33×10^{-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.

e. Nucleotide sequences