1. ***The basics***

a) Define the term "silent mutation"

Silent mutations are DNA mutations that do not result in a change to the amino acid sequence of a protein

b) What is the "central dogma" of molecular biology?

Biological information flows in only one direction, from DNA to RNA to proteins. Replication (DNA) ->Transcription (DNA-RNA)->Translation (RNA->protein)

c) Identify the longest open reading frame in the following DNA sequence and translate it into an amino-acid sequence (note: translation table provided at the end of the exam)

TGCGTATGTATGTCAGACGGTGAGACGCTTGCGGGCTAAGCGACG

ATG TCA GAC GGT GAG ACG CTT GCG GGC TAA

M S D G E T L A G

2. Sequence alignment

a) Describe the recurrence and location of answer for global alignment between two sequences.

Recurrence:

$$OPT(i,j) = \min \begin{cases} \text{cost}(a_i, b_j) + OPT(i-1, j-1) & \text{match } a_i, b_j \\ \text{gap} + OPT(i-1, j) & a_i \text{ is not matched} \\ \text{gap} + OPT(i, j-1) & b_j \text{ is not matched} \end{cases}$$

Cost of the optimal alignment between $a_1 \ldots a_i$ and $b_1 \ldots b_j$

Written in terms of the costs of smaller problems

Bottom rightmost cell

b) Perform a global **multiple** sequence alignment on the following sequences and report the alignment and Sum-of-Pairs score. Use Seq1 as Sc in both (center of star tree). **MATCH** = +1, **MISMATCH** = -1 , **GAP** = -1.

Seq1: AGT

Seq2: ACT

Seq3: AGAT

Seq1: AGT

Seq2: ACT

Seq1: AG-T

Seq3: AGAT

AG-T

AC-T

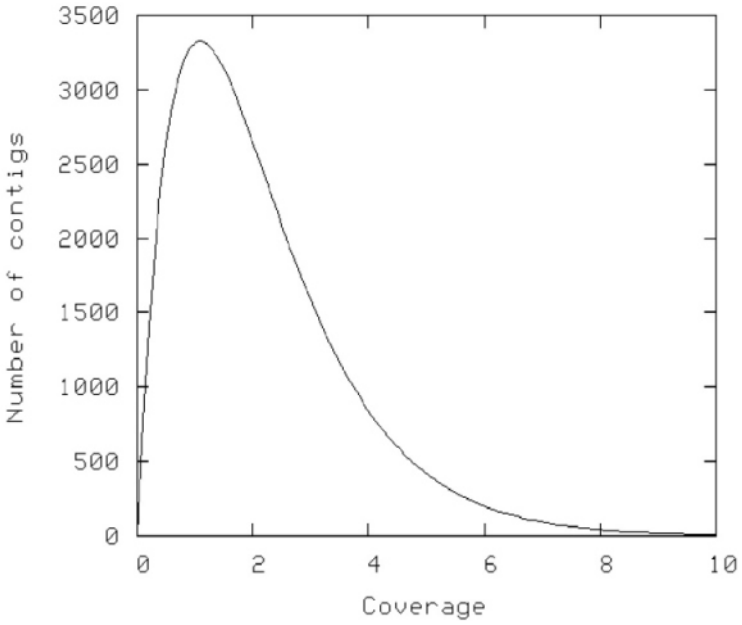
AGAT

S(-,-) = 0 (if not we double count)

+1, +1, +1, -1, +1, -1, 0, -1, -1, +1, +1, +1 = **3**

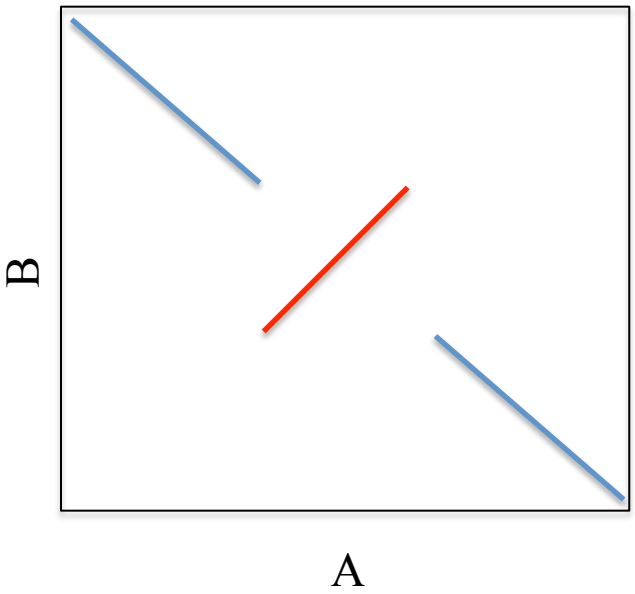
3. Genome assembly

a) The Lander-Waterman model describes the expected number of contigs (N) in a genome project as a function of the genome length G, read length L, depth of coverage c, and the overlap between sequences o. Without remembering the exact formula, sketch the rough shape of the dependency between N and c, assuming G, L, and o are fixed.



4. Genome alignment

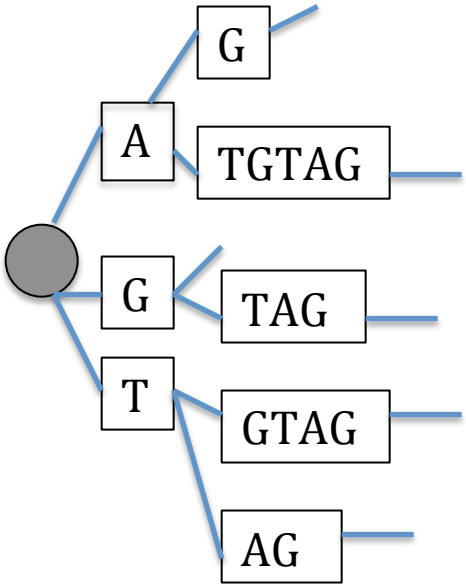
Briefly describe what is depicted in the dot plot below:



A large scale inversion event between genome A & genome B.

4. Data structures: Suffix trees

a) Given the following string, construct a suffix tree of ATGTAG



a) Label the path of the string GTAG in the above suffix tree. Give the time complexity of finding a query of length 'n'.

$O(n)$

Translation table

Ter - stop codon

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu	TCA S Ser	TAA * Ter	TGA * Ter
TTG L Leu	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile	ACC T Thr	AAC N Asn	AGC S Ser
ATA I Ile	ACA T Thr	AAA K Lys	AGA R Arg
ATG M Met	ACG T Thr	AAG K Lys	AGG R Arg
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly