1. ***The basics***

a) Define the term "silent mutation"

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

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

2. Sequence alignment

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

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.

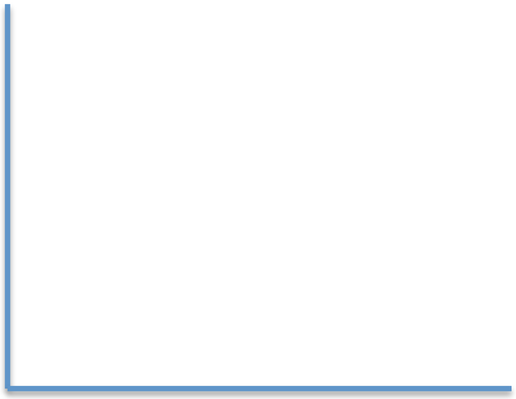
Seq1: AGT

Seq2: ACT

Seq3: AGAT

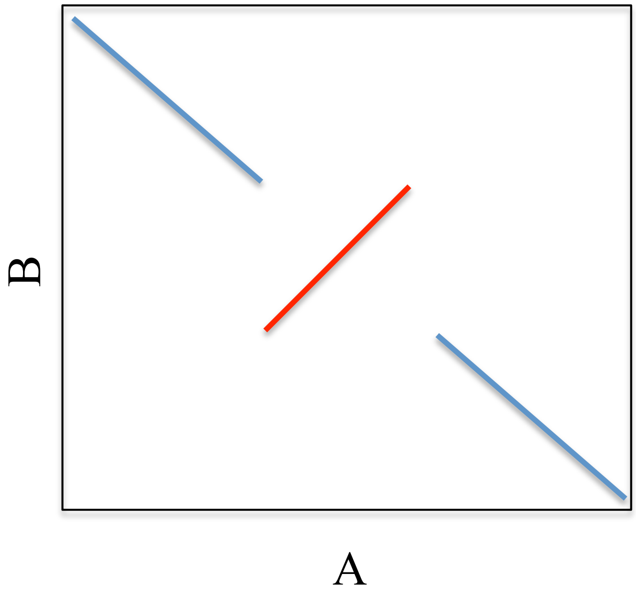
3. Genome assembly

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:



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'.

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