Santhanakrishnan Ramani
SID: 105720585

Bioinformatics & Genomics

Assignment – 3

1) a) BLASTN is more sensitive and is used in finding more distantly related sequences, reason being it uses a shorter default word size. The word size is adjustable in BLASTN and can be reduced from the default value to a minimum of 7 to increase search sensitivity. (Referred from https://blast.ncbi.nlm.nih.gov/BLAST_guide.pdf)

b) Type III restriction protein res subunit [Ignisphaera aggregans DSM17230] contains a sequence that is the most similar by overall score.

Fraction of the query is included in this alignment – 27 %

3881 – 5548 coordinates of query match and it is mapped on to the reverse strand.

c) Archaea species had the most hit to the sequence

d) Blosum80 matrix, as its main use is to group sequences that are >80% similar and it is very conservative.

e) Yes, the sequence was taken from the species Ignisphaera aggregans, when done a nucleotide blast using the sequence I got 99% query cover and 100 % ident. (Accession No. CP002098.1)

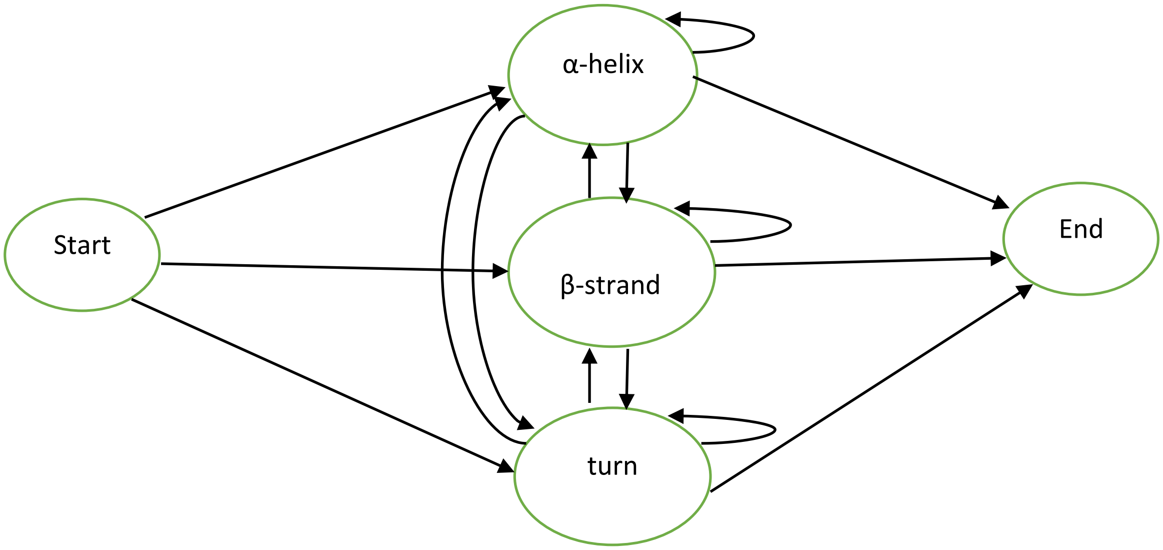
2) a) Prof. David Haussler scored his first publication in the journal "Science" in the year 2004 (PMID: 15131266)

b) They both where co-authors in one publication, and they worked on the **1 alpha, 25-dihydroxyvitamin D3** molecule (PMID: 4812038)

c) Two parts of the human body that have been found to contain archaea

- Human Gut (intestines) (PMID: 17563350)
- Oral Cavity (PMID: 15067114)

3) a) State diagram of the HMM



b) Emission parameters needed to describe this model –> N (M – 1) = 3 (20 – 1) = 57

c) Transition parameters needed to describe this model –> N (N – 1) = 3 (3 – 1) = 6

4) a) Gene - SH3TC2 (SH3 domain and tetratricopeptide repeats 2)

b) Number of exons – 17, Genomic Size - 81,025

c) DNA Sequencing was done using SOLiD (Sequencing by Oligonucleotide Ligation and Detection) system

d) The nonsense mutation (R954X) and the missense mutation (Y169H)

e) Pes cavus (highly arched feet) or pes planus (flat feet)

5) a) Mouse_lemur and Bushbaby

b) dbSNP id **rs17722293,** chromosome coordinate chr5:148402467-148402467

c) Exon 2, and first five nucleotides **TTCCA**

6) a) Viterbi algorithm is used to calculate the most likely path for a sequence.

b) Most likely path for the sequence GGCACTGAA is **HHHLLLLLL.**

	G	G	C	A	C	T	G	A	A
H	**0.15**	**0.0225**	**0.0034**	3.4e-04	**6.12e-05**	6.1200e-06	**8.808e-07**	8.8080e-08	1.2683e-08
L	0.10	0.0150	0.0023	**5.1e-04**	**6.12e-05**	**7.3440e-06**	**8.808e-07**	**1.5854e-07**	**2.8537e-08**