BF527: Applications in Bioinformatics

Please prepare a typed report to submit through Blackboard. Please include printouts of all Python code that you write. Your code should follow the guidelines laid out in class, including commenting. Partial credit will be given for nonfunctional code that is logical and well commented. This assignment must be completed on your own.

HOMEWORK #7

See Blackboard for assignment and due dates

PROBLEM 7.1 (40%):

Explore the gene expression dataset in **Gene Expression Omnibus (GEO)** with the accession **GSE4115**. In this study, the authors compared the gene expression in histologically normal bronchial epithelium in 79 samples from smokers with lung cancer with 73 samples from smokers without lung cancer. The goal of the study was to identify a diagnostic gene expression profile that can distinguish between lung cancer and non-lung cancer samples.

Use GEO to identify the number of genes differentially expressed in smokers with lung cancer, versus smokers without lung cancer. Use a Two-tailed T-test with 0.05 as the Significance Level.

Use a web tool of your choice to identify any significant pathways or biological processes that may be affected in lung cancer patients. Do you find anything interesting; does it make sense? What would you do next as a follow up experiment (bioinformatics or biology)?

PROBLEM 7.2 (50%):

Mitogen-activated protein kinase 6 (MAPK6) is an enzyme that is a member of the Ser/Thr protein kinase family. MAPK6, along with other MAP kinases, are extracellular signal-regulated kinases, which are activated through protein phosphorylation. MAPK6 is known to contain one protein kinase domain (Pkinase), located at the N-terminus. This Pkinase domain is a linear motif binding (LMB) domain, and is known to bind the following short linear motifs (SLiMs):

LMB Domain	SLiM
Pkinase	N.E.K..N
	N.Y....E
	S...D.PL
	S..SS
	S.S..S
	ST.S
	F.FP

Write a Python script that integrates information about MAPK6's interaction partners, their sequences, and the known motifs that the Pkinase binds, to determine how many of MAPK6's interactions are mediated by a LMB domain-motif interaction. The file **`HW7.2_uniprot_proteins.fasta`** (available on blackboard) is a fasta file that contains proteins that are known to interact with MAPK6. Your code should print: (1) the Uniprot ID of the binding partner; (2) the motif found in the binding partner; and, (3) the location of the motif in the binding partner's sequence. There may be more than one motif within one protein. There may also be no motifs in a protein.

PROBLEM 7.3 (10%):

Please list the three follow-up bioinformatic analyses you and your partner are planning to do for the final project.

EXTRA CREDIT (5 points):

Watch the 3 webinars hosted by **George Church** on Youtube (~20 minutes):

1) http://www.youtube.com/watch?v=mVZI7NBgcWM
2) http://www.youtube.com/watch?v=2r9DpthvNKM&feature=related
3) http://www.youtube.com/watch?v=mgXAO8pv-X4&feature=related

Discuss the potential benefits/detriments of getting your genome sequenced and the potential benefits/detriments of making your genome sequence public for all to see.

Would you get your genome sequenced? Would you make it public? Why or why not?