BF527: Applications in Bioinformatics

> **Note:** Please submit the Jupyter notebook through Blackboard. 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 6

See Blackboard for assignment and due dates

PROBLEM 6.1 (30%):

You sequence a new organism and identify the gene sequence below but have no idea what its function is. Use **at least two** publicly available web tools to predict the function of the gene product:

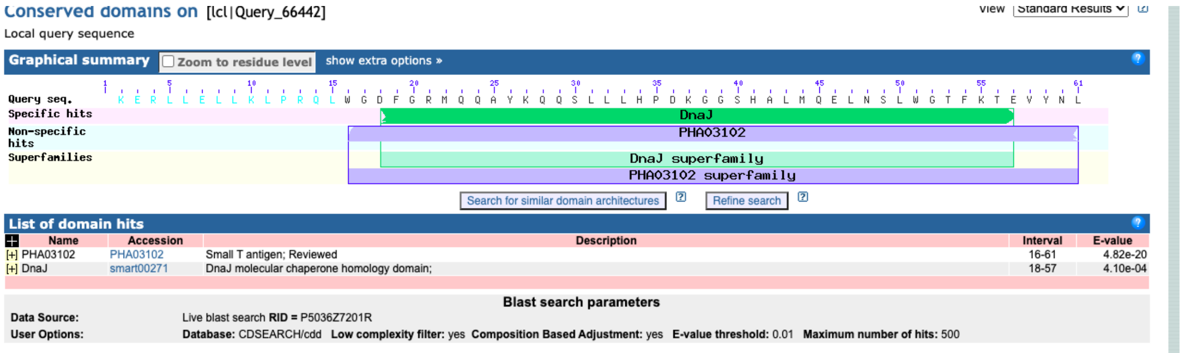
>Unknown protein DKERLLELLKLPRQLWGDFGRMQQAYKQQSLLLHPDKGGSHALMQELNSLWGTFKTEVYNL

Describe your procedure, and interpret your results by answring the following questions:

1. What does it match to, and how well does it match?
2. Are there any conserved domains? What do those domains do?
3. What is the most likely Gene Ontology terms that would be assigned to this protein?

1.Using NCBI BLAST, the unknown protein sequence matches to Chain A, LARGE T ANTIGEN [Alphapolyomavirus muris]. The Score is 98.6 bits(244),e-value is 3e-25,identity is 46/46(100%). Using Uniprot, the sequence matches to P0C567 from Murine polyomavirus (strain Crawford small-plaque) (MPyV).The identity is 100%, score is 153.069, and the e-value is 8.4e-43.

2.



Conserved domains on [lcl|Query_66442]

Local query sequence

List of domain hits

Name	Accession	Description	Interval	E-value
[+] PHA03102	PHA03102	Small T antigen; Reviewed	16-61	4.82e-20
[+] DnaJ	smart00271	DnaJ molecular chaperone homology domain;	18-57	4.10e-04

Blast search parameters

Data Source: Live blast search **RID =** P5036Z7201R

User Options: **Database:** CDSEARCH/cdd **Low complexity filter:** yes **Composition Based Adjustment:** yes **E-value threshold:** 0.01 **Maximum number of hits:** 500

The hsp70 chaperone machine performs many diverse roles in the cell, including folding of nascent proteins, translocation of polypeptides across organelle membranes, coordinating responses to stress, and targeting selected proteins for degradation. DnaJ is a member of

the hsp40 family of molecular chaperones, which is also called the J-protein family, the members of which regulate the activity of hsp70s. DnaJ (hsp40) binds to dnaK (hsp70) and stimulates its ATPase activity, generating the ADP-bound state of dnaK, which interacts stably with the polypeptide substrate.
（https://www.ebi.ac.uk/interpro/entry/InterPro/IPR001623/)

3.GO term： Biological Process protein folding (GO:0006457) response to heat (GO:0009408) Molecular Function ATP binding (GO:0005524) unfolded protein binding (GO:0051082)

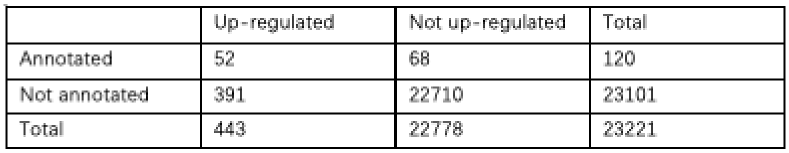
PROBLEM 6.2 (40%) There are **23221** human proteins with some Gene Ontology (GO biological process annotation. You conduct a gene expression microarray experiment and identify **443** genes that are up-regulated; you can assume that the remaining (23221–443 =) **22778** genes are not up-regulated. Using GO, you note that **52** of the up-regulated genes are annotated as participating in the “Notch signaling pathway”. There are a total of **120** proteins in GO annotated with the term "Notch signaling pathway".

- **1.** Without using a calculator or computer, do you think you have an enrichment of Notch signaling pathway members in your list of up-regulated genes? Describe your intuition for why there is or isn’t enrichment of the GO term in your list.

  > *Hint: think in terms of percentages.*

The propotion of 52 annotated genes out of 443 in up-regulated genes is 11.7%, and the proportion of 120 total annotated genes out of 23221 total genes is only 0.51%, so there is a strong overrepresentation (enrichment) of “Notch signaling pathway” in up-regulated genes.

- **2.** Fill out the contingency table that describes the overlap between up-regulated genes and “Notch signaling pathway” annotated genes. What distribution can be used to calculate the probability of observing this contingency table? Set up the formula and enter the correct numbers for calculating the probability, but ***DO NOT*** compute the probability.



	Up-regulated	Not up-regulated	Total
Annotated	52	68	120
Not annotated	391	22710	23101
Total	443	22778	23221

The hypergeometric distribution could be used to calculate the probability of seeing any given contingency table.

$$P = \frac{\binom{443}{52}\binom{22778}{68}}{\binom{23221}{120}} = \frac{443!22778!120!(23221-120)!}{23221!52!391!68!22710!}$$

- **3.** How many contingency tables would be required to calculate a p-value for assessing significant enrichment of the Notch signaling pathway in your list of up-regulated genes? Note that the maximum number of genes that can be both up-regulated and belong to the Notch signaling pathway is 120.

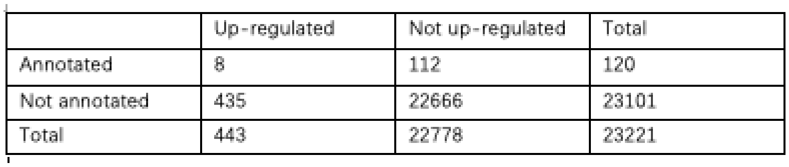
68

- **4.** Go to http://www.langsrud.com/fisher.htm and calculate the p-value (use right sided). Is finding 52 Notch signaling pathway proteins in a list of 443 up-regulated genes significant? Explain and cite the appropriate numbers.

Yes, the corresponding p-value of my table is Right : p-value = 1.9460262632525435e-57, which is far less than 0.05, and this indicates finding 52 Notch signaling pathway proteins in a list of 443 up-regulated genes is significant.

- **5. Extra Credit:** Fill out the contingency table describing up-regulated genes versus annotated genes that will have the highest probability of being observed. You do not actually need to compute anything here, use proportions to think of the most likely number of annotated genes you would detect. Explain your thought process.

Assume there are x genes annotated “Notch signaling pathway” are found in Up-regulated genes. If the genes annotated as “Notch signaling pathway” are randomly distributed, the most likely observed proportion of x relative to the upregulated genes should be equal to the proportion of upregulated genes relative to the overall genes, which is x： 443 = 443 ： 23221. Then x is approximately 8 genes. The contingency table is:



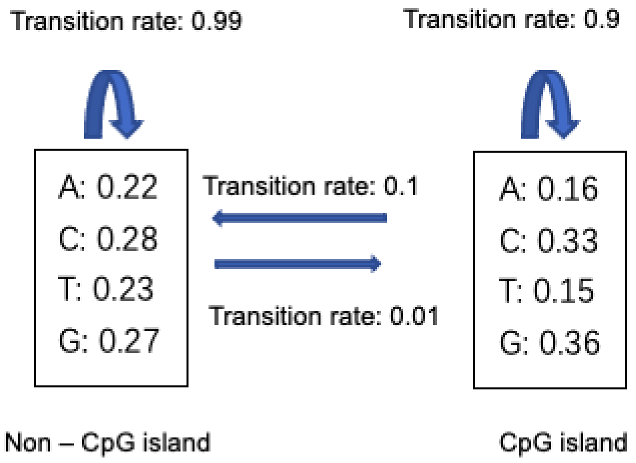
	Up-regulated	Not up-regulated	Total
Annotated	8	112	120
Not annotated	435	22666	23101
Total	443	22778	23221

PROBLEM 6.3 (30%): CpG islands or CG islands are genomic regions that contain a high frequency of C and G nucleotides. They are often found in gene promoters and are important for regulation of gene expression.

- **1.** Construct a model (*i.e.* draw a picture) representing an HMM for finding CpG islands across the genome which includes hidden states, emission states, transition probabilities, and emission probabilities. While the exact transition and emission probabilities are not known (you can make up numbers), the probabilities in your model should generally reflect the biology of CpG islands. Explain the reasoning behind your choice of probabilities.

> ***Hint:** When determining the states in your HMM, think about what it means to be inside or outside a CpG island. **This can be represented by a very simple HMM, don’t overthink things.** For the transition probabilities between hidden states, do additional research and find out what percentage

> of the genome is estimated to be in a CpG island. This percentage or frequency should at some level be reflected in your transition probabilities.*



The emissions probabilities were made up. The transition probabilities were made up based on the fact that CpG islands have been estimated to constitute 1%–2% of the mammalian genome (Antequera and Bird 1993).

- **2.** In HMM models, the transition probabilities leaving a particular hidden state should always add up to 1 (*i.e.* the probabilities of the arrows going out from a state should add up to 1). Is this the same for the probabilities entering a particular state? In other words, do the probabilities of the arrows pointing towards a state add always add up to 1? Can you explain this?

No, the sum of probabilities entering a particular state(for example, entering A-) is not always one. This is because entering one particular state is not all the results for this event, instead, the sum of probabilities of entering all possible states should be one.