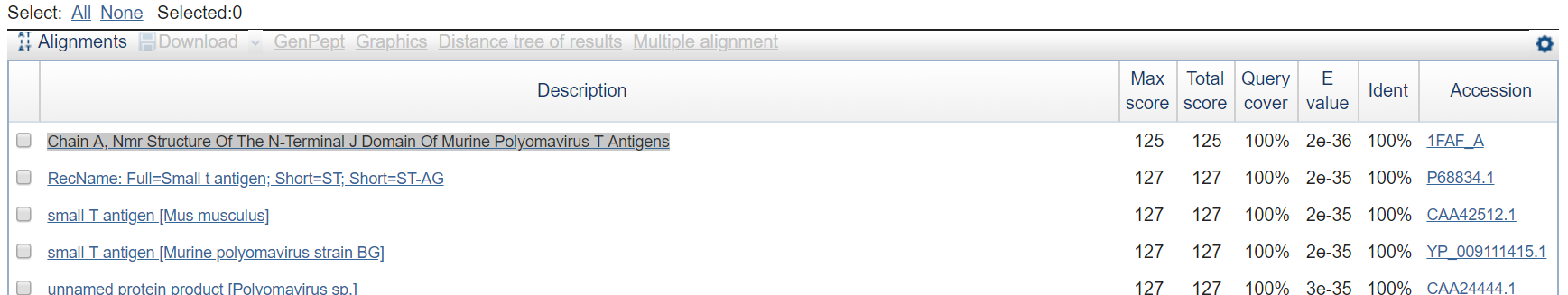
Problem 6.1:

The two tools I used to identify the sequence are:

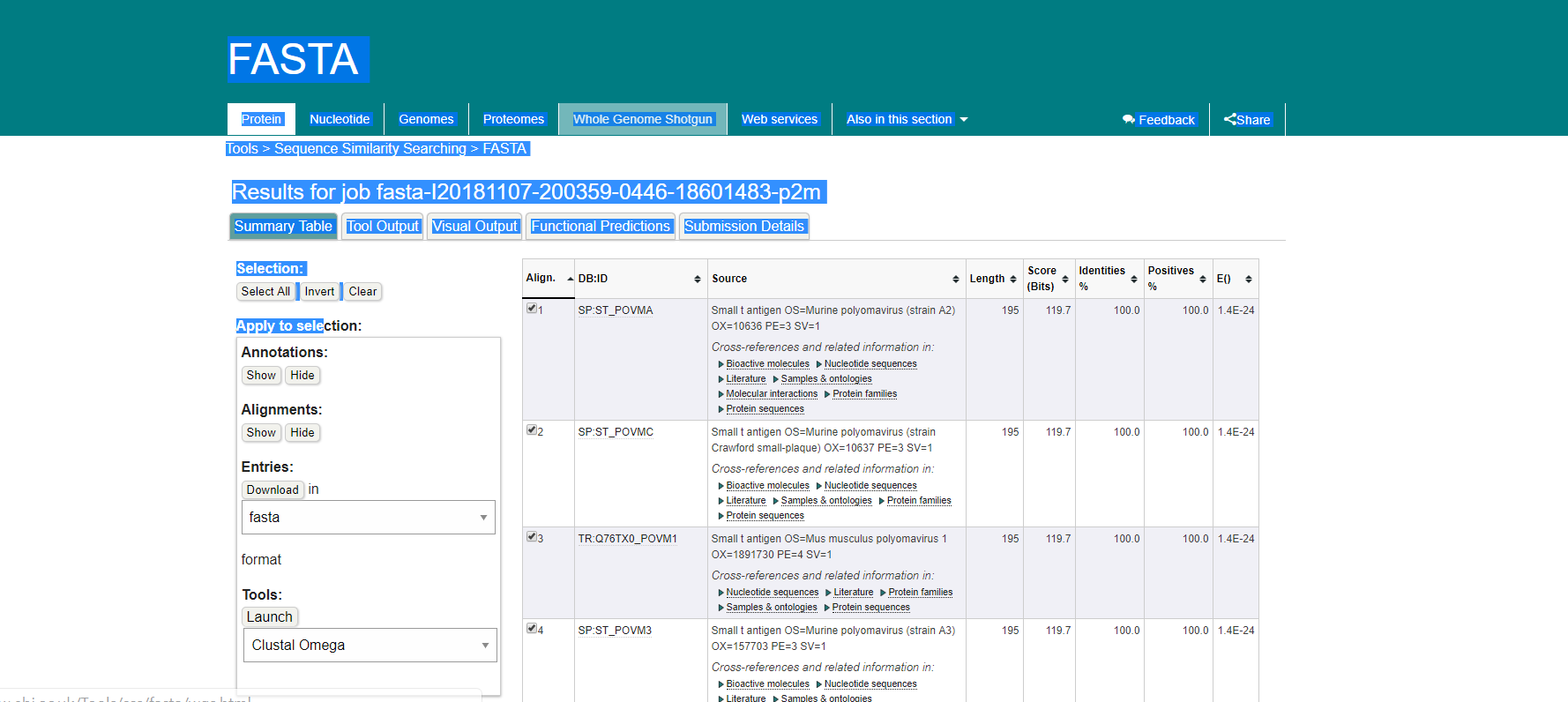
1) As given in NIH, BLAST (basic local alignment search tool) finds region of similarity between biological sequences. The program compares protein sequences to sequences in the databases and calculates the statistical significance.

So, I used **Blastp** and found it to be matching with the following protein with 100% query coverage and an expect value of 10^-36 (a very strong match)

[Chain A, Nmr Structure Of The N-Terminal J Domain Of Murine Polyomavirus T Antigens](https://blast.ncbi.nlm.nih.gov/Blast.cgi#alnHdr_11513524)

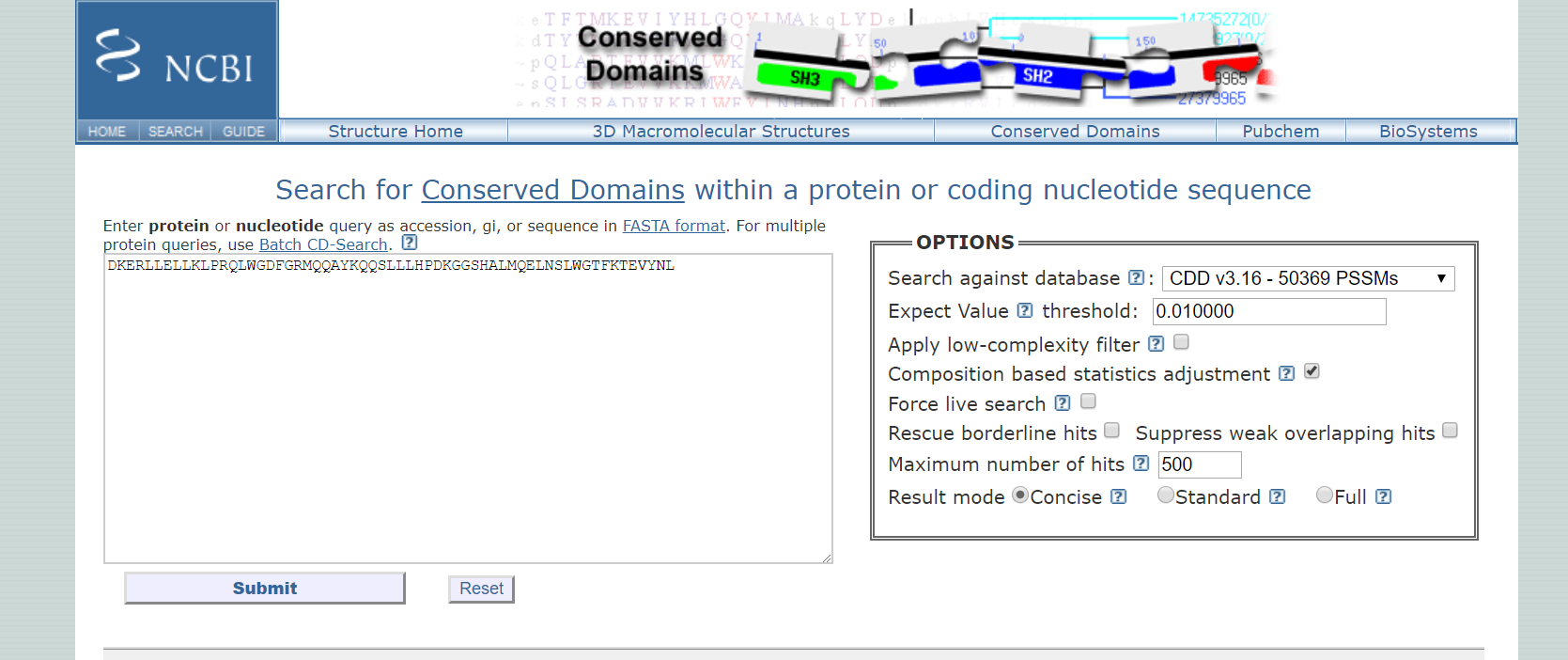


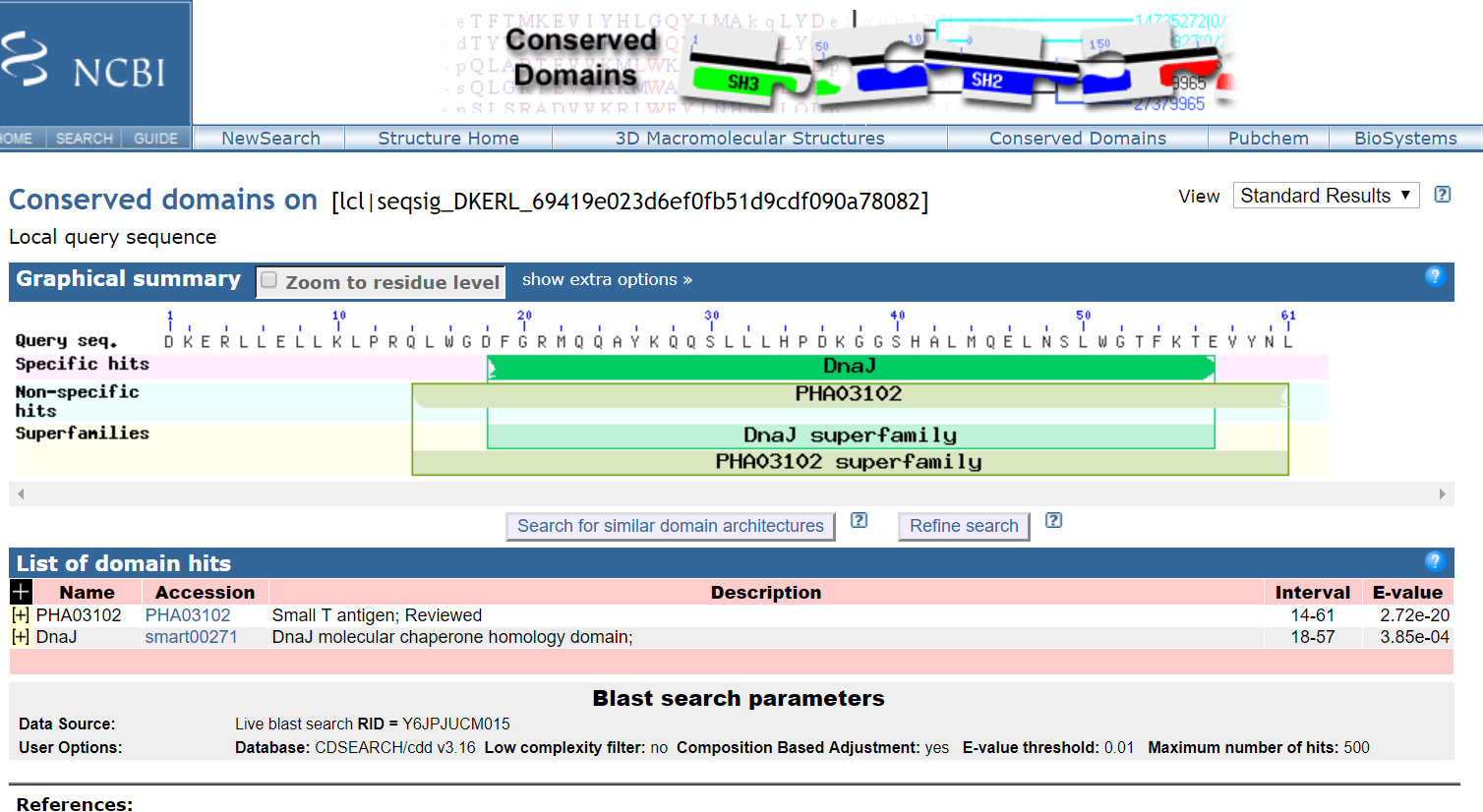
2) The **FASTA** suite of programs can also be used to do sequence similarity search against protein databases. The e value was 1.4E-24 (which is not as strong match as Blastp search)



Yes, there are conserved domain. I used NCBI CCD to find that the sequence has two conserved domains.

1. Small T antigen is a protein that encodes double stranded DNA viruses.
2. DNAj molecular chaperone homology domain is heat shock protein which has important function in the process of folding, unfolding and translation.





The GO term associated with this protein are binding, protein binding and molecular function.

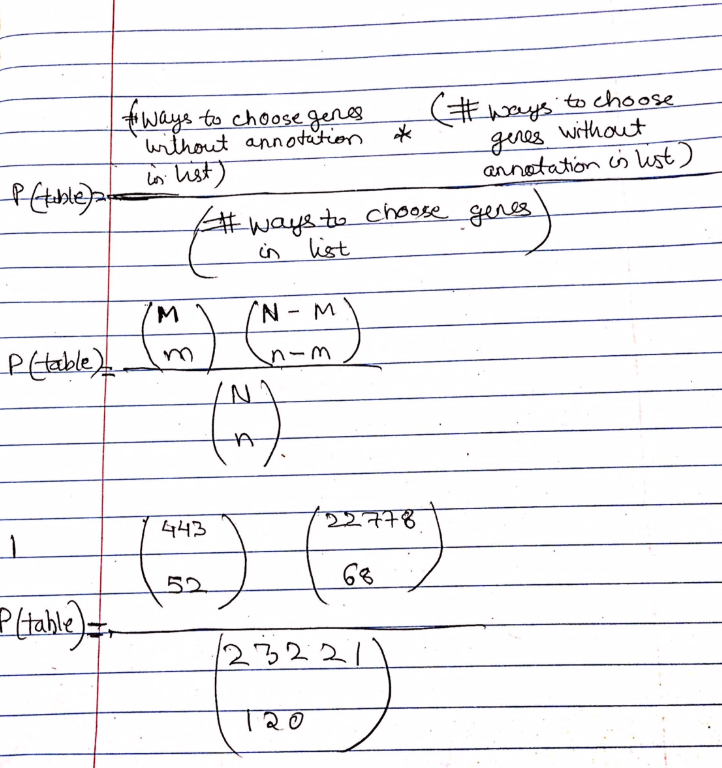
I searched through NCBI website and found that according to NCBI, the NCBI BioSystems database was designed to centralize access to existing databases and to integrate their records with associated data types (genes, proteins, small molecules, etc.).

I retrieved all GO records using NCBI Biosystems database.

Problem 6.2:

1. Yes, I think there is an enrichment of Notch signalling pathway members. The given problem state that there are 443 up-regulated genes identified. There are 52 up-regulated genes are annotated as participating in the “Notch signalling pathway”. It makes up approximately 12% of the genes.
2. Contigency table is below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Detected | Not Detected | Total |
| Has annotation | 52 (m) | 391 (M-n) | 443 (M) |
| Does not have annotation | 68 (n-m) | 22710 (N-M-n+m) | 22778 (N-M) |
| Total | 120 (n) | 23101(N-n) | 23221 (N) |

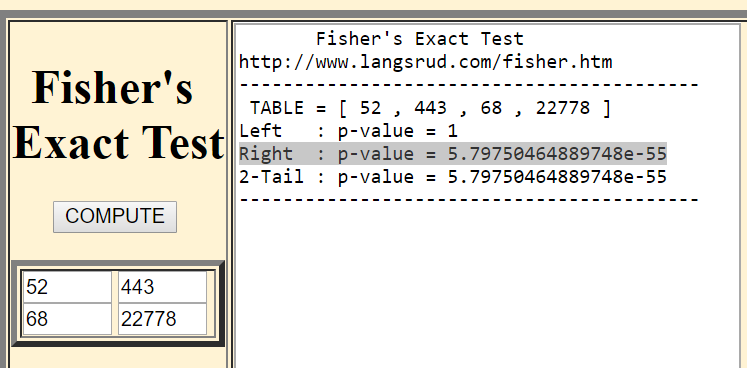
3)

There would be 69 contingency tables required to calculate a p-value for assessing significant enrichment of the Notch signalling pathway in our list of up-regulated genes.

4) The fisher exact right-side value is as below:

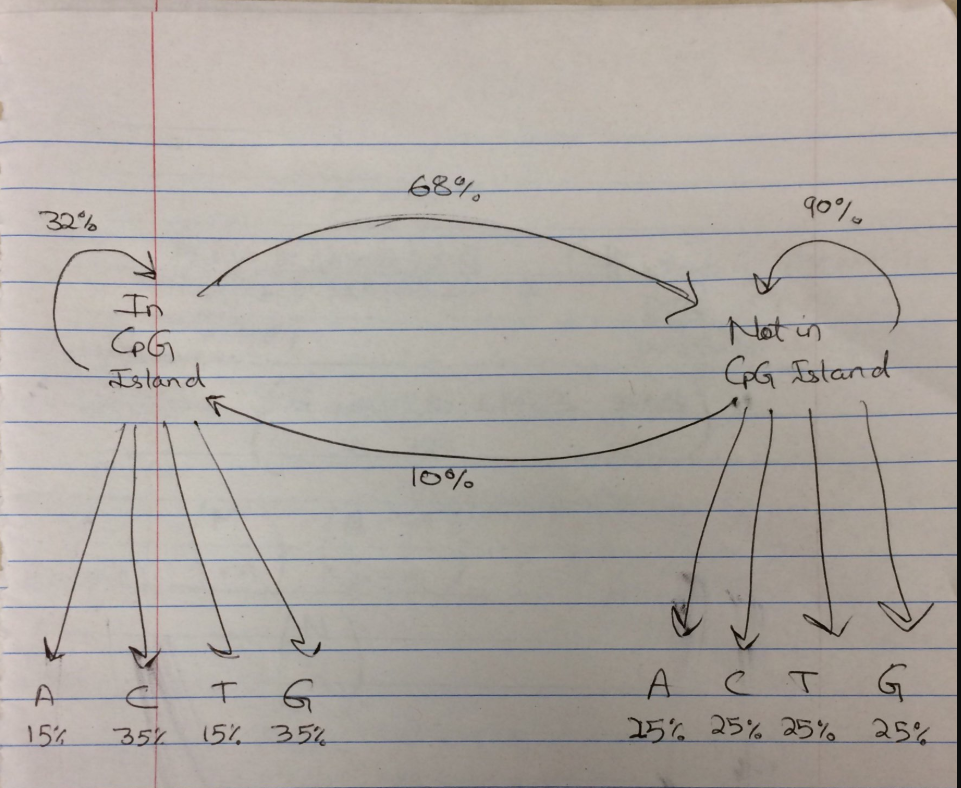
Right: p-value = 5.79750464889748e-55

It is significant as the value is very lower in comparison to alpha value of 0.05.



Problem 6.3:

1)



Explanation:

The HMM model for CpG islands consists of two hidden transition states namely, “in CpG islands” and the other “not in CpG islands”. The probability of landing up in “not in CpG islands” is more than the “in CpG islands” which is corresponds to the biological fact that the percentage of CpG island observed in human genome is 42%. Thus, the hidden transition probability of entering to the CpG islands is 42%. Also, there is small probability to switch from “not in CpG islands” to “in CpG islands”. However, it is more probable to switch from “in CpG islands” to “not in CpG islands”.

In DNA sequence we can observe the four nucleotides namely, A, G, C and T. Thus, these are our four emission states for each of the hidden transition states. The emission probability for “not in CpG islands” has equal probability to find any nucleotide. So, the same probability percentages for each nucleotide. But the emission probability for “in CpG islands” consist of higher probability for the C and G content than the rest of nucleotide. So, the C and G nucleotide has high probability than the rest of the nucleotide.

2)

No, the probability of the arrows entering a particular state will not be 1 because there is always a chance one can enter many states. But when we leave the state it is always the case so it adds up to one.