Genefinder Write-up

Oscar Zhang

Results:

See the complete output or list of proteins in the output.txt file.

It is very likely that many of the protein results are actual genes. For example the first protein on the list(also the longest protein on the list)is a multi-species protein based on results from protein blast with a score 722 It comes Salmonella as expected from our input of a Salmonella sample strain and likely of a function in the secretion system. It's also pathogenic given coming from Salmonella likely a Proteobacteria or Gammaproteobacteria. For the 2nd protein on the list, the uncertainty is greater with a much lower score and positives percentage. However, it's likely to be a glycoside hydrolase family 32 protein and not pathogenic.

Reflection:

- *What are some opportunities to use gene finding technology to benefit others?*
    - Genefinding technologies can quickly locate potentially meaningful sequences in a strain that can lead to important insights for identifying disease, viruses and discovery of new proteins or harmful pathogens. It can be used for practicing doctors when examining samples from their patients. Gene finding tech can also be used for researchers to improve the efficiency of their research by completing computationally intensive tasks such as locating the proteins in a very long dna strain.
- *What are some of the program's limitations? What are the downsides of the program's limitations if used beyond what it's explicitly good for?*
    - The program is purely based on condition checking and algorithmic programmed by human hand(me, who is not the best programmer nor biologist in the world). It fails to identify special cases in dna strands, but it might produce erroneous results with a less common strand. Furthermore, It does not check whether the input is a valid dna, meaning, we can pass in a random string and it would produce a meaningless output. It is not significantly useful beyond its explicitly stated purpose
- *How could you adapt the program to enable more advanced applications?*
    - We would need more complex algorithms to account for the variety of dna strands. One simple add-on, however, would be to connect the program to the protein blast search engine and enable the program to perform search on each candidate after finding the results. Based on the returned result from protein Blast, we can implement functions using the scores provided by Blast to locate the most meaningful interpretations of the proteins. We can also implement machine learning to allow it learn from its errors and extract more value from the vast data on dna online.