Project Information

You should ensure that you clearly answer the questions mentioned below as PART of your project report.

The Public Health Agency of Canada (PHA) is testing a new crop protection product that a Saskatoon start-up company, Biosanto, has developed. Crops sprayed with this product appear to grow more quickly with fewer infestations by certain insect species. However, the PHAC needs to ensure that the product is not a health risk to Canadians. Biosanto has indicated that the product contains live microorganisms, however no bacteria could be cultured from the samples supplied. So, they decided to complete some direct sequencing of DNA present in the samples to characterize it further. The following sequences were included amongst many eukaryotic sequences found within samples analyzed. These sequences need to be analyzed to determine:
- What organism, or organisms, likely contain such sequences? Are any organisms that contain such sequences likely to be a health risk to Canadians? Why or Why not? What followup experimentation would be advised?
As part of this analyses, you need to report on what can be determined regarding the
- Potential function of any complete genes encoded in these sequences.
- The protein subcellular localization of any deduced proteins derived from the gene sequences, and whether either of these proteins could be targets for diagnostics for detecting such bacteria.

Note that as part of this particular component of the analysis, you should become familiar with the evolutionary relationships between the genes located in these sequences and their homologs, as this will help identify how unique the genes encoded in these sequences are. You should also identify what you can determine about the structure of the given protein products of any genes, to potentially identify regions of the proteins that may be more likely localized on the cell surface (or protein surface). Note, however, that you are not required to perform any homology modeling to identify the structure. General topology predictions based on secondary structure are adequate. Also, you are not required to construct a phylogenetic tree of *all* homologs to deduce the evolutionary relationships – however, selected homologs may be used to construct a phylogenetic tree to help support any claims you make about evolutionary relationships with other proteins involved in causing disease or not.

In the end, keep in mind that there will be some analyzes that are straightforward and some that won't be. Don't dwell on those that aren't straightforward (i.e. performing extensive analyses on more complex issues), but rather use the discussion section of your report to mention what could or couldn't be done and what you would do in the future to investigate this question further, if you had had more time.

As eluded to above, in the discussion of your report, you should describe any other bioinformatics or laboratory experiments that you would recommend be additionally completed to investigate this more thoroughly. Mention the pitfalls of any methods that

were noted. Such descriptions can be very brief and do not have to be exhaustive. Just provide a sense of some follow up that would be suitable.

Sequences to be analyzed:

Seq1

gttgcgtctgcgtagcctgatggctgttattgaggataactccctgtttatccacgtcaaactgacggtaggtattgtgggaaacgc
catcggcgcccggcgtctgaatattgacctgcggcgtgccgttggcgctgcgaataatagtcggctgctgacgccccggtgcct
gtccgtcggccacaatcgccgcctgtgccgtaagcgaaattccgccaagggcgaataacaccccaaaccgtagcgcagtcag
ccggcaacaacgcggcggtgaagcggagcgacgtgcgcggcgtccggttcttccccgtccggcccgcgctatctcggccac
caccatccacatctggcgagattgattaaagataatacgatataactgcttgttcatgaattttactccacacaacactacccaccgg
cagttttgctatttgaattaatactgccaatacaggttgaatccggccgtgacgggggaggttgggaaaccggcgggtttggatag
cgggacaccgacaaacaggtcgtagctcacccctttcaggctgcctcgccagcccagagcgctccccgccaggtgttttccga
gcagtgtatcgctgccccgtccgcccacttcgccataatccattgccagatacagcgattgccctttcagcggggtttgccagtcc
agttcgttgcggctgtaccaaccccgatcagcggataaggtcaactcgccatcaaagccgcgcaccgtccagcgaccgccaat
cgagaagcgatcctggggcgtcaatgggctggcgcttatctgacgcaaatattgggtttgataagaaaagggctgagaaaaca
gtgttaacggcccggcaagcgtcgctgagaactgggtgattttgggtaaggcagtcccttcattgcggtactcttccggtgccggt
tgtgcgccaaaccagcgcatgccctgttgataagtcacccgggtatccagcgtcgcgctgccgataaagtgatgatgggacag
ccccagcttccagcccgcggtttcccggcgctgaacgtcgatttcagtgttatcaataaaattgcgagacgcccgacgataaattt
caccgttaagcagggttttctggtcggcattacggtgcagcacccggctaagctgaaccgccaggttttggcttttgccccgatatt
ggatcgcctgattcagcccggctattgtctgggtataatcgtaatgactggccgtcacgctcgccatccagtagccgaagggcaa
agagtaatagagcgtgtcattttggctgccttcccggcagacgaatgaatatcgcgtccaccggaaagataaaacagatcactta
acgccaggggattatccagaaagaaggtgagtccgccctgataacgtccggtgcttttggagccggtatcatccagataagtcg
cgactcgccagtgccgggactgacgccagtccagcacaatatcgctctcccccggttgctctgccgggacaatgttcatatctgc
ctgaacggtggggaaacgttgcagattttctaatccctgctcaatatcccgtaaatccagtaattttccttcacgcgcgggaaaggg
ggttatccgctggatatagttcccgctgtcaggggtatagcggatatgccggatttttcccggtatcatcaccagttttaaggtttgg
gtacgcaaatcctgacgaggaaccagaacccggctggtgacatagccgtgattaatgagctgatcttgcaattgggttatcagcc
ggctaatgcctttatcgcccagacaatggttttcgccctgttgggcgacaagccggaacggtaaccagtgcgggaaattttcggt
acctgccagtctgacgtgggtaatgggaaaacacggagtttccaccgcaaaggcggatgattccccggtttccggtaccgaca
ggtgaacctccgcaggcggcggggctaattgcgtctccagggcttgctggcgggcctgttgatgaatgagttgttgatctgcgat
gtcggcggcttgactacctgagactgcgctacaaaaaactattaatactacaatactgcttttcattgacttgctattccacaatgttg
aaacagcattaaccacgaatttaacttagagatcaatatttagtgttgattatcaaccgggcaactaaaatagggggttggttaagat
agcagcaaaaacaagaggttatatgacaaagcaaaaaacaggccaatgtgattattttgtaacagcattgtataaaaaacagatta
atattaatcatttaataatcaacattaataatttcaaattaattttatctttttgttttttattatgtgattaatatcacatttattatgattgatttcg

Seq2

gaaatttagggagtaagaagatggaagctattaataatatcagtagttctcaggtgaacagtgtcaataatagtgttgatattgtaac
tgacaaaatccccacaccgacaacacatgaagctgctattcaacagtcggaaaaaaccaggctcacggataaggtgaattgatt
gtaccgaaacagggtttcgatcccagtcgcttaggaaatgtcggacaagaattaaataatacgctagatatgatgtcgctgctgttt
gagattgcaaagaggatgcgtgaactgggtattttgcagcgtgaactgaaaataaagcgtcctgatgctcagaagtcacaagtg
gatgagatgcgacatagcgccaaattgatgattgctatggcagtggtttcaggcttaatgacgattggttcggggctgataggtgg
ttttttcttccgctaaaaatagtcaagcgattaagcagcagaagacgttggaaggtaatattgctggtcgtaacgaattaattgatctta
agttggaacagctaggtaagagtagcgatgcggatcgggcagcaattggtagagtatggaatgtagcacaagatgctgacaag

agtgctctgaaatctattaccttggtttttgaaagtcgtaacagtaagcagcagttcttaagttcagtgatgaaatctttagagagtatg
tccaataattcggtgcaggtagaacagggactttctcaggctaaagctaaagaacatgaagtggattcatctatttctcaacatgaa
aaacagaagagtgaagatcagatctcattgaataacaacttcatgagagaggttttgcaattgatacagcagctttatcagagtcat
gtccaagcatggcgcgcggcgactggcgtggtgtaatgctagcgaataagtctgagtgcaccgagaggtgcgctcttatcgatg
gatgctgctatttaagctcgtttttaaaatctactaatagcgttaaaggtaaaagtatagaaaagcgaaaatacccaattaagatttca
atctttgacaggcagaacgtgttaaggaaaaaaagcacgcgtctatgctctggtgtttggtttttgataatcaccggagtgacctttt
attaaaagaattgcccttgagtgacagaggcacaaaagatgatggtagtaaactggcccccttttctgcgggagtaacccgtttgt
ttgaccattcggtagttcaagttggtctgattgccttctctagcaggagtaacccatgatggatgtgatagccatcgtcaaccaagtt
ttgtctgaatttgctgttaaattcggcttgccgcacctggctttaaatcatgaaggtgtggccgc

Please remember, this report is focused on trying to give you experience performing an analysis of a sequence using a variety of tools and show that you understand appropriate use of tools and their benefits and limitations. It is not supposed to be a long report so keep your answers brief, and to the point, and I'm sure you'll do well!

Report Structure:

- Abstract (20%) – *250 words maximum*
- Main Text: Focus on Results and Discussion (50%) – "1500 words max, with < 1000 words preferred" – Clearly report what you've found, presenting results in suitable subsections. Ensure any questions mentioned in the above problem-statement are answered within your report – highlighted as necessary as clear statements or sections. Show a schematic diagram of a protein's proposed structure, including its proposed subcellular location. Show a phylogenetic tree and discuss its evolutionary history and possible function. What conclusions can you draw? What can't you conclude? Why? Show me that you understand the benefits and limitations of the methods used. What analyses – "wet lab" or bioinformatics-based – would you recommend be performed next to investigate this issue further?
- Methods: (30%) Summarize methods. Marks for using appropriate methods and using methods appropriately (i.e. give settings when relevant – if not I will assume default settings used). Point form is fine.

Note: Lab time near the end of the course will be devoted to project work, but you are welcome to start some analyses early to get a head start, hence this is provided ahead of time.