PBL in Bioinformatics – “Making Predictions”

Facilitator: Fiona Brinkman

You are a new Assistant Professor in bioinformatics at the University of Ottawa, who specializes in forensic microbial genome analysis. A Dr. C. Kent contacts you from the Department of National Defense and asks to meet with you in person. At this meeting he first asks you to sign a confidentiality agreement and then informs you of a perplexing problem they have been having.

In a new, top secret, level 3 (biosafety containment level) laboratory (located under a basement bar in Ottawa’s Market area) they have sequenced the genome of a newly discovered microbe that is a bioterrorism threat. This highly pathogenic, green-glowing Gram-negative bacterium has been named *Kryptococcus viridis.* They have been using the genome sequence to identify all possible secreted proteins, as part of efforts to identify key secreted toxins, but based on their initial proteome analyses of secreted proteins (2D gel analysis of extracellular proteins) there seems to be many more proteins secreted than their bioinformatic analysis of the genome sequence suggests.

Proteome analysis of other cellular components found no evidence that the secreted proteins subcellular fraction was contaminated with proteins from other parts of the cell. N-terminal sequencing of select proteins in the secreted fraction revealed that they were similar to sequences of specific genes in the genome, however the deduced protein sequence from the corresponding gene sequence had a longer N-terminal sequence extension. No signal peptide was computationally identified by SignalP 4.1 within these N-terminal sequence extensions. The following is one example:

Gene translation, with additional N-terminal extension not found in the processed protein marked in bold:
vvnslradetpvigkieapravenrghndlsgengnrrlnvippktraggmnkvfslkysflakgfvavselarris
vkgklksassiiispitiaivsyappslaatvnadisyqtfrdfaenkgafivgasninvydkngvlvgvldkapmp
dfsaatmntgtlppgdhtlyspqyvvtakhvngsdimsfghiqnnytvvgennhnsldikirklnkivtevapaeis
svgavngayqeggrfkafyrlggglqyikdkngnltpvytnggfltggtisalssynngqmitaptgdifnpangpl
anylnkgdsgsupermanlfaydsldkkwvlvgvlssgsehgnnwvvttqdflhqqpkhdfdktisydsekgslqwr
ynknsgvgtlsqesvvwdmhgkkggdlnagknlqftgnngeiilhdsidqgagylqffdnytvtsltdqtwtgggii
tekgvnvlwqvngvnddnlhkvgegtltvngkgvnngglkvgdgtvilnqrpddnghkqafssinissgratvi

Dr. Kent would like you to look at these sequences and determine whether such secreted proteins can be better computationally predicted to be secreted. Since the proteome analysis only identifies a subset of proteins, accurate computational prediction of secreted proteins may significantly help them identify all possible secreted proteins from this pathogen.

Brainstorm about

- *What possible problems may be occurring*
- *What further questions you’d want to ask Dr. Kent*

As you’re discussing this, list possible learning objectives that you have already identified.

Guiding Discussion Questions

Why care about predicting secreted proteins?

How are proteins secreted? Why is there a missing N-terminal sequence in the example processed protein?

What computational methods have been developed to identify secreted proteins? What is SignalP and what approach does it use?

What possible problems can occur in predictive algorithms?

Which is more important, precision or recall?

How accurate is bacterial gene prediction? What miss-predictions most commonly occur that may affect prediction of protein features like in secreted proteins?

How can one improve predictive algorithms for identifying secreted proteins? Are any of these approaches more generalizable to improving prediction of protein function or other predictions in general?

Goal (Not to be written for this initial case – only discussed verbally)

Draft a reply to Dr. Kent for a bioinformatics approach that may more accurately identify secreted proteins in an automated fashion from whole bacterial genome sequences. In your document, explain why you think he has been having trouble and how you believe your method overcomes this problem.

Keep in mind that your reply must handle the examination of confidential data.

Learning Objectives

Become familiar with the most common computational methods used for bacterial gene prediction and prediction of secreted proteins, as models for common predictive algorithm approaches.

Appreciate the pitfalls associated with such predictive algorithms, as models for issues associated with similar predictive algorithms in general.

Appreciate the impact of gene miss-predictions on downstream analyses.

Appreciate the issue of balancing precision vs recall in predictive algorithms and which may be more important, when.