Project Defense Ranjit Kumaresan
CMSC858P 109059121

Problem Statement: (taken from last year's undergrad bioinformatics course webpage. #14)
Simple Annotation Pipeline
Run a gene finder on a genome, identify a set of genes, then compare these against GenBank using blast to determine their probable function. This project will require you to learn how to use the gene finder Glimmer, and how to run batch Blast jobs against GenBank.

Proposal:
This is a Discovery project. A list of definitions and a short description of softwares to be used are first given:

Gene is a locatable region of genomic sequence corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions and/or other functional sequence regions.
Glimmer (stands for Gene Locator and Interpolated Markov ModelER) is a bioinformatics ystem for finding genes that uses the interpolated Markov model formalism.
GenBank (sequence database) is an open access, annotated collection of all publicly available nucleotide sequences and their protein ranslations.
Blast (Basic Local Alignment Search Tool) is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. A *BLAST search* enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

Definition of the problem:
As the problem statement suggests, given a genome, we use Glimmer to identify the set of genes present in it. Next, we use Blast to find out the function of the genes that we found in the genome. I also plan to implement some extensions (ideas detailed out in later passages)

Relevance of the Project:
Almost of all (of my friends) people who are in the Biology Department at UMD, have used Blast. But they have used only the web interface and some of them didn't even know about the existence of a batch processing option for Blast. Also, they all agreed that an annotation pipeline will definitely be an useful software.
The software basically can be used to give an idea of what the genomic sequence is capable of doing. For example, it is possible to identify what set of proteins are required for digesting food and extracting energy out of it. Now given a genomic sequence, the software, by using Blast can now find out the probable functions of the genes in that genome. Using this list, our software can be extended to see if given a genome and a function, whether the genome contains genes that can perform the function. It can also be extended to help in identifying the various genes that are responsible for a function. Such a software is clearly welcome to any biologist.

Relevance to the Course:
Clearly this project is about alignments (Blast) and biological sequences. We learnt about a lot of tools/software in this course, but we hadn't played around with them. I see this project as an opportunity to make myself familiar with these software and also help in furthering their usability. With no prior experience in either Blast, GenBank or Glimmer, I see this interesting project as a challenging one.