CSC 526 Project

PROJECT OVERVIEW

Protein structure prediction has been a very important and challenging research problem in bioinformatics for years. Predicting protein tertiary structure is a very challenging problem, but more tractable if using simpler secondary structure definitions. A rule-based data-mining approach called BLAST-RT-RICO (Relaxed Threshold Rule Induction from Coverings) that utilizes multiple sequence alignment information to **predict protein secondary structure** was presented in the form of two research papers:

Paper 1:
L. Lee, J. L. Leopold, and R. L. Frank, "Protein Secondary Structure Prediction Using BLAST and Relaxed Threshold Rule Induction from Coverings," *Proc. 2011 IEEE Symp. on Computational Intelligence in Bioinformatics and Computational Biology*, Paris, France, Apr 11-15, 2011, accepted for publication.

Paper 2:
L. Lee, J. L. Leopold, C. Kandoth, and R. L. Frank, "Protein Secondary Structure Prediction Using RT-RICO: a Rule-Based Approach," *Open Bioinformatics J.*, vol. 4, pp. 17-30, 2010.

The initial test results of BLAST-RT-RICO method indicate that, given a test protein, if there are homologous sequences with known secondary structures in the PDB database, BLAST-RT-RICO can generate a very accurate secondary structure prediction. But there are a lot of questions remain unanswered. This project aims to answer some of these questions.

This is a group project. The group projects will be completed by groups of **3 students**. The group grade on the project is based on the reports, program and presentation. The individual grade is based on your attendance during the last weeks of class (for other presentations) and your level of participation in the group project as evaluated by your team members.

The project is incremental. The deliverables are as follows:

Part #1. Progress report: refer to individual topic	7%
Part #2. Progress report: refer to individual topic	8%
Part #3. Project presentation	5%
Part #4. Project report or paper / program code	15%

Project groups can choose one of the following project topics:

Topic 1, Segment length (5?)

In the above two papers, a segment length of 5 was used to generate rules and perform predictions. The group is to write Perl programs to experiment with different segment lengths (3, 7, 9, 11, 13, 15) to find the most suitable segment length to produce the best prediction results (using the RS126 test data set and CB396 test

data set). In other words, write your Perl program(s) to repeat all experiments conducted in paper 1, using different segment lengths. (Estimated program running time: 2 weeks for one segment length.)

Deliverables:
Part #1. Progress report (summary: 1 pg; plan, individual student's work and timeline: 1 pg; RT-RICO algorithm half done: code)
Part #2. Progress report (work progress: 2 pgs; RT-RICO algorithm: code; web crawler program code)
Part #3. Project presentation (ppt slides)
Part #4. Research paper (graduate option), project report (undergraduate option)

Topic 2, Threshold parameter (90% confidence?)

In the above two papers, a threshold value of 90% (or 0.9) was used to generate rules and perform predictions. The group is to write Perl programs to experiment with different threshold values (0.8, 0.7, 0.6, 0.5, 0.4, 0.3, …) to find the most suitable threshold value to produce the best prediction results (using the RS126 test data set and CB396 test data set). In other words, write your Perl program(s) to repeat all experiments conducted in paper 1, using different threshold values. (Estimated program running time: 2 weeks for one threshold value.)

Deliverables:
Part #1. Progress report (summary: 1 pg; plan, individual student's work and timeline: 1 pg; RT-RICO algorithm half done: code)
Part #2. Progress report (work progress: 2 pgs; RT-RICO algorithm: code; web crawler program code)
Part #3. Project presentation (ppt slides)
Part #4. Research paper (graduate option), project report (undergraduate option)

Topic 3, Query all proteins with known secondary structure

In the above two papers, only RS126 test data set and CB396 test data set were used to perform predictions. Write a Perl program (or Perl programs) to use all available proteins with known secondary structure (from PDB database) to perform predictions. In other words, write your Perl program(s) to repeat all experiments conducted in paper 1, using all the available proteins. The points awarded for this project is proportional to the number of proteins you manage to run to generate the final Q_3 score. (Estimated program running time: 2 weeks for 500 proteins. Last year there were around 166,000 proteins with known secondary structure (PDB database). There are duplicate entries, so I estimate that there are around 60,000 to 70,000 proteins with known secondary structure.)

Deliverables:
Part #1. Progress report (summary: 1 pg; plan, individual student's work and timeline: 1 pg; RT-RICO algorithm half done: code)
Part #2. Progress report (work progress: 2 pgs; RT-RICO algorithm: code; web crawler program code)
Part #3. Project presentation (ppt slides)
Part #4. Research paper (graduate option), project report (undergraduate option)

Topic 4, BLAST score / e-value analysis for all available proteins with known secondary structure

Given a test protein, if there are homologous sequences / proteins with known secondary structures in the PDB database, BLAST-RT-RICO can generate a very accurate secondary structure prediction. It is important to know the number (or %) of proteins that have homologous sequences / proteins with known secondary structure in the PDB database.

Last year there were around 166,000 proteins with known secondary structure (PDB database). There are duplicate entries, so I estimate that there are around 60,000 to 70,000 proteins with known secondary structure.) Write a web crawler program to perform BLAST searches for all the proteins. Use these proteins as inputs, for each protein, find (and record) the related homologous sequences / proteins with known secondary structures in the PDB database (excluding the input protein of course), using both BLAST score and e-value (separately, in two different set of tests) as similarity indicators. Analyze the statistics and provide recommendations for future research direction.

Deliverables:
Part #1. Progress report (summary: 1 pg; plan, individual student's work and timeline: 1 pg; web crawler program: code)
Part #2. Progress report (work progress: 2 pgs; statistics analysis program: code)
Part #3. Project presentation (ppt slides)
Part #4. Research paper (graduate option), project report (undergraduate option)

Topic 5, Protein 3D structure prediction problem statement, free energy function.
The protein 3D structure prediction remains an extremely difficult and unresolved problem. The two main issues are the calculation of protein free energy and finding the global minimum of this energy.
The group need to find / read current scientific research papers in this area, and produce a typical (computer science) problem statement in the form of a research paper. The problem statement should include input (mathematical definition and example test data), output (mathematical definition and example test data) and problem description (including how to verify the accuracy of the prediction). The paper should also provide examples of how to obtain the test data, how to derive the final output and how to calculate the prediction accuracy score. The group must teach / introduce the problem definition to the class.

Deliverables:
Part #1. Progress report (summary: 1 pg; plan, individual student's work and timeline: 1 pg; collection of (at least 10 related research papers or books): title and actual paper)
Part #2. Progress report (work progress: 2 pgs; background statement and problem statement: research paper format)
Part #3. Project presentation (ppt slides) and teaching session
Part #4. Research paper (graduate option), project report (undergraduate option)