BMI 203
Winter 2017
Homework Assignment #3

Due Friday 02/24/2017 by 11:59pm PST

In this assignment, you will implement an alignment algorithm that uses a variety of scoring matrices, and then find an optimal scoring matrix. Note that there are two parts! You can obtain scoring matrices and sequences from https://www.dropbox.com/sh/70kyjqggwtsolyh/AAAfG6x_x7_VBNGdJFq8cuSOa?dl=0.

Part 1

Implement the Smith-Waterman algorithm and instrument the code such that it can use any scoring matrix provided (i.e. will read it in from a separate file). You may adapt code you can obtain from the Web (there are a number implementations you can find). But you must demonstrate your understanding of the algorithm with detailed comments.

Answer the following questions:

1. Consider the false positive rate (proportion of negative pairs with scores that exceed a score threshold) when the true positive rate (proportion of positive pairs with scores above the threshold) is 0.7. What's the best false positive rate that you can achieve with varying both gap opening (from 1 to 20) and extension penalties (from 1 to 5) with the BLOSUM50 matrix? What is the best gap penalty combination?

2. Using the gap penalties you determined from question 1, which of the provided scoring matrices performs the best, in terms of false positive rate (at a true positive rate of 0.7)? What are the performance rates of each of the matrices? Create a Receiver Operator Curve (ROC) graph which shows the fraction of true positives on the Y axis and the fraction of false positives on the X axis. Include on the graph data for each of the provided matrices. Please take care to make your ROC graphs square, with both X and Y axes limited to the range [0:1]. Note, you can download ROC code from here: http://www.jainlab.org/Public/ucsf-roc.zip. It is not guaranteed to be bug free but it might save you some time.

3. How does the performance change if you normalize the Smith-Waterman scores by the length of the shortest sequence in a pair (i.e. divide the raw score by the min length)? Show the ROC curves for your best matrix and for the same matrix with normalized scores. Are the false positive rates better or worse? Why do you think this is so?

Part 2

Using the best gap penalties and matrix from part 1, create an alignment for each positive pair of sequences and each negative pair of sequences. You will use these static alignments as a starting point from which to optimize a scoring matrix to maximize separation of scores of the positive and negative pairs.

1. Devise an optimization algorithm to modify the values in a starting score matrix such as to maximize the following objective function: sum of TP rates for FP rates of 0.0, 0.1, 0.2, and 0.3. The maximum value for the objective function is 4.0 (where you are getting perfect separation of positive and negative pairs even at the lowest false positive rate). You should use the gap and extension penalties derived from Part 1. Remember, you must maintain symmetry in your matrix. You can make use of real-valued scores in the matrices if desired (this is probably a good idea).

2. Beginning from the best matrix from above (that which produced the alignments), run your optimization algorithm to maximize the fitness of the new matrix. How much improvement do you see in the fitness? Show the full ROC curves for the original matrix and the optimized matrix. What happens when you now realign the sequences using the new matrix and rescore? Show the new ROC curve following realignment on the same graph as above. Qualitatively discuss how your matrix and your alignments change following optimization.

3. Beginning from the MATIO matrix, but using the same initial sequence alignments, re-run the optimization. Show the same ROC plots as for (2). Discuss the relationship between the results you see here and the results you saw for (2).

4. Describe your optimization algorithm briefly. How might you improve it?

5. What would be required in order to make a convincing case that an optimized matrix will be of general utility and will actually be beneficial for people to use in searching databases?

To complete this assignment:

- Comment code: It is OK to get code from anywhere, but intuitive descriptions showing you understand what/why all steps are doing must be included.
- Email a single pdf (name = JaneSmith_BMI203_HW2.pdf) to ryan.hernandez@ucsf.edu and tamas@tamasnagy.com answering the above questions with prose, graphs, and the optimized matrices you derived.
- A link to your Github repository
    - Make sure there is a link to the Travis build results for your repo in the README file
    - Note that only commits prior to the due date will be considered!