In-class Assignment #3

DUE May 12th, 2009

You must show all code, results, and answer questions to get full credit.

Alignment activity:

This activity compares the eyeless gene of Drosophila Melanoganster with the human gene aniridia. They are master regulatory genes producing proteins that control large cascade of other genes. Certain segments of genes eyeless of Drosophila melanogaster and human aniridia are almost identical. The most important of such segments encodes the PAX (paired-box) domain, a sequence of 128 amino acids whose function is to bind specific sequences of DNA. Another common segment is the HOX (homeobox) domain that is thought to be part of more than 0.2% of the total number of vertebrate genes.

1. Here we compare the HOX domain of human and fly. Obtain the peptide sequences (amino acid) from the GenBank database: AAD01939 (human sequence) and AAQ67266 (fly sequence).
2. Use seqdotplot to see if there are any areas that are clearly aligned. Turn in plot.
3. Try a global alignment using the function nwalign.
4. Is this a high, medium, or low valued score? Why?
5. Try a few other BLOSUM matrices (e.g. BLOSUM30, try a minimum of 3) – how does the resulting nwalign score change? Explain why.
6. Use different open and extension gap penalties in your scoring (e.g. gapopen=5, extendgap=5, try a minimum of 3) – how does the nwalign score change? Explain why.
7. Use showalignment to show the sequences alignment. Turn in alignment. Note the parameters used to make this alignment and the final score.
8. Use BLAST to verify your findings.

To assess if the score is significant the first step is to make some random sequences that are similar to that of the fly protein. One way to do this is to take random permutations of the fly sequence. This can be done with the *randperm* function. Then calculate the global alignment of these random sequences against the human protein and look at the statistical significance of the scores.

1. Permute the fly sequence 50 times, and globally align it (nwalign) to the human sequence.

2. Make a histogram of the 50 scores. Put a stem of the score you obtained in 7. on this graph. (Use “hold on” to plot on top of the histogram). Turn in this plot
3. Is the original score distant from the other scores? If yes, this means it is statistically significant. Is your score statistically significant?

You will now repeat the process of estimating the significance of an alignment this time using local alignment and a slightly different method of generating the random sequences. Instead of simply permuting the letters in the sequence, an alternative is to draw a sequence from a multinomial distribution which is estimated from the fly protein sequence. You can do this using the aacount and randseq functions; the first estimates the amino acid frequencies of the query sequence and the later randomly creates new sequences based on this distribution.

I will provide you the following code:

```
[lscore,locAlig] = swalign(human,fly,'scoringmatrix','blosum30','gapopen',5,'extendgap',5);

fprintf('Score = %g \n',lscore)

showalignment(locAlig);



localscores = zeros(n,1);

aas = aacount(fly);

for i = 1:n

   randProtein = randseq(flyLen,'FROMSTRUCTURE',aas);

   localscores(i) = swalign(human,randProtein,'scoringmatrix','blosum30','gapopen',5,'extendgap',5);

end
```

1. What is the score that is printed?
2. Now, make a histogram of all the local scores. Turn in this plot
3. Is the lscore stastically significant from the localscores?

You can consider also the coding regions of the PAX genes previously discussed.

1. Get the nucleotide sequences 'AY707088' and the 1st coding region from 'NM_001014694'
2. Convert them to amino acid sequences (humanprotein and flyprotein) using the nt2aa command.
3. Write the following code and turn in the plot:

```
[score,alignment] =
nwalign(humanProtein,flyProtein,'scoringmatrix','pam50','SHOWSCORE',
true,'gapopen',5,'extendgap',5);

fprintf('Score = %g \n',score)
```

4. Do the above step again, but now for SWALIGN. Turn in the plot.
5. How do the two dynamic programming plots look?