

This homework is due February 28 in class.

1 Hidden Markov Models for CpG Island prediction (100 points)

The Cs of most CpG dinucleotides in the human genome are methylated. Methyl-C tends to mutate to T, and so CpG dinucleotides tend to decay to TpG/CpA. This is believed to account for the fact that in bulk human DNA CpG dinucleotides occur about five times less frequently than expected (Bird 1980, Jones et al 1992). CpG islands are unmethylated regions of the genome that are associated with the 5' ends of most house-keeping genes and many regulated genes (Bird 1986, Larsen et al 1992). The absence of methylation slows CpG decay, and so CpG islands can be detected in DNA sequence as regions in which CpG pairs occur at close to the expected frequency. The fact that CpG islands can be detected in this way indicates that the corresponding germline DNA has been substantially hypomethylated for an extended period of time, and in fact about 80% of CpG islands are common to man and mouse (Antequera and Bird 1993).

About 56% of human genes and 47% of mouse genes are associated with CpG islands (Antequera and Bird 1993). Often CpG islands overlap the promoter and extend about 1000 base pairs downstream into the transcription unit. Identification of potential CpG islands during sequence analysis helps to define the extreme 5' ends of genes, something that is notoriously difficult with cDNA based approaches. Probably because they are associated with genes, CpG islands tend to be unique sequences and are therefore very useful in genome mapping projects.

In this assignment, we will develop 1st order Markov models for CpG and non-CpG regions, as well as a Hidden Markov Model (HMM) for CpG island detection. We will use both models to predict the location of CpG islands in human chromosome 22. The Markov model will have four states corresponding to the four DNA bases. The HMM model will have two hidden states – one state denoting CpG islands, and the other denoting non-CpG islands. The number of observation symbols is 4, corresponding to the four DNA bases. First, some questions to which you should provide written answers. Assume you have annotated training data marking regions of human chromosome 22 as being CpG islands.

- a. (5 points) How do you calculate maximum likelihood estimates for the transition probabilities and initial state probabilities of Markov models for CpG islands and non CpG sequences?
- b. (5 points) How can you use the Markov models to predict the location of CpG islands on human chromosome 22?
- c. (5 points) How do you calculate maximum likelihood estimates of the transition probabilities, emission probabilities and initial state probabilities of the HMM? Write down the estimation formulas.
- d. (5 points) How do you use the HMM to determine the location of CpG islands? What is the difference between Viterbi decoding and smoothing or posterior decoding? Write down their mathematical definitions and then explain how they differ.

| start | stop | symbol |
|----------|----------|-----------|
| 14430001 | 15077850 | NT_028395 |
| 15227851 | 18889431 | NT_011519 |
| 18939432 | 42215733 | NT_011520 |
| 42227734 | 43057958 | NT_011521 |
| 43107959 | 47356150 | NT_011523 |
| 47366251 | 48750436 | NT_011525 |
| 48767137 | 49087576 | NT_019197 |
| 49089177 | 49107103 | NT_113818 |
| 49126804 | 49591432 | NT_011526 |

<http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?taxid=9606&chr=22> gets you to Chromosome 22 on the NCBI browser. Chromosome 22 has nine contigs, shown above. We will use the first three contigs for this assignment.

Note the start and stop coordinates for the contigs. There are sequencing gaps. The annotation file referred to later on in this document, uses these coordinate positions, so make sure you define Matlab data structures to record them. To retrieve the sequences corresponding to these contigs, use the `getgenbank` function in Matlab with the ID name in the third column. For example,

```
>> contig1 = getgenbank('NT_028395','SequenceOnly',true)
```

This stores the entire nucleotide sequence for the first contig of Chromosome22 into a Matlab sequence array called `contig1`. You can view the GC density using the Matlab function `basecount`.

```
>> basecount(contig1)
```

You can plot the nucleotide density along the contig using the matlab function `ntdensity`

```
>>ntdensity(contig1)
```

To get a count of nucleotide pairs and triples (also called codons), use the Matlab functions `dimercount` and `codoncount`.

```
>>dimercount(contig1)
```

```
>>codoncount(contig1)
```

CpG island annotations are in the file `ncbimapBuild_36.3_22_cpg_0K.txt`, part of which is shown below. Homo sapiens Genome (Build 36.3)

```
#Chromosome: 22
```

```
#####
```

```
#Map: cpg
```

```
#Region: 1..49,691,432
```

| | |
|----------|----------|
| 14465506 | 14465802 |
| 14476299 | 14476584 |
| 14476831 | 14477068 |
| 14477323 | 14477925 |
| 14477323 | 14478281 |
| ... | ... |

The first CpG island starts at position 14465506 and ends in 14465802. By looking up the contig coordinates, you can see that this island is on the first contig. The CpG islands are not disjoint. Note that the 4th and 5th islands overlap. Partial overlaps between CpG islands can also occur. You can read the annotation file into a Matlab array as follows.

```
>>[cpgStart,cpgStop]=textread('ncbimapBuild\_36.3\_22\_cpg_0K.txt','%d %d','headerlines',6)
```

Now you will have two Matlab arrays with the starting and ending points of 1500 CpG island annotations provided by the NCBI for Human Chromosome 22.

- (10 points) Estimate the initial state probabilities and transition probabilities for 1st order Markov models of CpG islands. Randomly choose 90% of the islands to train the model. Reserve the rest of the CpG islands for model testing (part c). The function `randperm` is very useful for generating a random permutation of the 1500 islands. Run your estimator three times, and report the variation in the learned parameters over these runs.
- (10 points) Estimate the initial state probabilities and transition probabilities for 1st order Markov models of non-CpG sequences. You need to devise a method for gathering non-CpG sequences. One approach is to sample sub-sequences of the contigs before or after the CpG islands. Gather at least as many non-CpG sequences as there are CpG sequences. Run your estimator on 90% of these sequences, reserving 10% for testing. Run the estimator three times, and report the variation in the learned parameters over these runs.
- (10 points) Use the estimated Markov models from part a to test the left out islands (10% of the original set of 1500 sequences). Compute the likelihood score for these sequences with respect to the CpG Markov model as well as the non-CpG Markov model. How accurate is your Markov model based predictor? How many annotated CpG islands were completely missed by it (false negatives)? Now use the same models to predict non-CpG sequences. How many spurious islands were predicted (false positives)? What is the percentage of false positives and false negatives?
- (10 points) Use the `hmmestimate` function in the Statistics toolbox of Matlab to estimate the emission and transition probabilities of the HMM. Use 90% of the CpG islands and the non-CpG sequences for training. Run the `hmmestimate` function three times, and report on the variation in the learned parameters of the HMM between the runs.
- (10 points) Use the HMM parameters estimated from the training data. to decode the first three contigs. Use the Viterbi algorithm implemented in Matlab as `hmmviterbi` to predict the CpG islands. Compare your predictions against the known CpG islands. How accurate is the Viterbi algorithm? How many annotated CpG islands were completely missed by it (false negatives)? How many spurious islands were predicted (false positives)? There is also a grey

area – predicted islands may overlap but not entirely coincide with annotated ones; make⁴ a decision about how to report these results. Note that exact borders of CpG islands are somewhat arbitrary, and it is possible that some true CpG islands have not been annotated. What is the percentage of false positives and false negatives? How accurately are the end points predicted?

- f. (10 points) Now use posterior decoding (implemented by `hmmdecode`) to make predictions about the locations of CpG islands. Use the learned HMM parameters and posterior decoding to predict islands on all three contigs. Compare your results against the known annotations, and provide statistics on accuracy, false positive, false negative rate, and accuracy of end point prediction.
- g. (20 points) Compare the predictions based on Viterbi decoding against those made by posterior decoding or smoothing. For `contig1`, `contig2`, and the portion of `contig3` for which we have annotations, produce some graphical (or at least tabular) representation of the location of the 'real' (annotated) CpG islands, the ones predicted by Viterbi, and the ones predicted by posterior decoding. Second, summarize quantitative facts about the performance of each prediction method as outlined above. Third, compare your results against that of a simple thresholding technique used in the Matlab function `cpgisland` (playing with window size parameter, cg content parameter and the `cpgoe` parameter).

Please include all the code you write as part of your homework submission. All the data provided for this assignment is on the class website. You will need Matlab with the Statistics and Bioinformatics Toolbox to do this assignment.