

### HW 3: due Tuesday, November 16th

PART 1. Consider the two sequences

1. ATCCTGG, and
2. ATCTTTA.

. What is the log likelihood ratio (LLR) that each is a coding sequence given the following assumptions? Show your work.

1. The position independent nucleotide probabilities are:

coding			
A	C	G	T
.2	.3	.3	.2

non-coding			
A	C	G	T
.4	.1	.1	.4

2. You are using a homogeneous first-order Markov model, with initial

initial prob's				
probabilities	A	C	G	T
	.3	.2	.2	.3

and transition probabilities as shown in the matrices below, where the column designates the base observed in the previous position, and the

row designates the output base from the current position.					
	coding				
	A	C	G	T	
A	.2	.2	.4	.2	
C	.3	.4	.3	.3	
G	.3	.3	.4	.3	
T	.2	.1	.2	.2	

non-coding				
	A	C	G	T
A	.5	.2	.2	.3
C	.1	.4	.2	.1
G	.1	.2	.4	.1
T	.3	.2	.2	.5

Part II. On the course website, you will find four files called **homework3.txt**, **cell-line-training.txt**, **cell-line-validation.txt** and **cell-line-test.txt**. These files contain microarray data from the paper “Gene expression patterns in ovarian carcinomas”, *Molecular Biology of the Cell*, Vol 14, pp. 4376–4386, November 2003. The hw dataset contains neither all samples nor all gene expression values from the original paper, rather you are only working with a smaller subset of the data.

1. The file **homework3.txt** contains 10 samples; some are from ovarian cancer tumors and some are called breast cancer tumors. The claim is that these two types of tumor samples produce microarray expression vectors that are sufficiently different that they can be distinguished by clustering methods. In particular, implement hierarchical clustering using the ordinary (Euclidean) distance metric for computing inter-sample distances, and the average-linkage clustering method of computer inter-cluster distances on this dataset and output a resulting partition of the 10 samples into 2 types. Submit your program source code via **provide** as follows:

```
conbrio% provide comp150b hw3 myfilename1.here myfilename2.here ..
```

Submit the resulting cluster tree your program produces AND the two clusters you will output in hardcopy.

2. The file **cell-line-training.txt** contains 10 samples, this time labelled correctly as to whether they are ovarian or breast cancer samples. The file **cell-line-validation.txt** contains another correctly-labeled 5 additional samples. Using the  $k$ -nearest neighbors classifier (again with

ordinary Euclidean distance), assume the cell-line-training samples are correctly labelled, and pretend the cell-line-validation class labels are unknown.

a) Compute the percentage of the cell-line-validation samples correctly classified based on the training data when  $k = 1, 3, 5$  and  $7$ , and report this in a table.

b) Which value(s) of  $k$  do the best? Choose a value of  $k$  that does best, call it  $K_{opt}$ .

c) Now classify the unlabelled sample in the file **cell-line-test.txt** using  $K_{opt}$ -nearest-neighbors. Do you predict it to be a sample from breast or ovarian cancer cell lines?