

# Algorithms in Bioinformatics I, WS2002/3

## Assignment sheet # 1

Daniel Huson

October 14, 2002

### 1 The occasionally dishonest casino (1 point)

An occasionally dishonest casino uses two kinds of dice, of which 99% are fair, but 1% are loaded, so that a 6 appears 50% of the time. We pick up a die from a table at random. What are  $P(\text{six} \mid D_{\text{loaded}})$  and  $P(\text{six} \mid D_{\text{fair}})$ ? What is the probability of rolling a six from the die we picked up?

### 2 The occasionally dishonest casino, continued (2 point)

You pick up a die and roll  $[6, 6, 6]$ . What is the probability that the dice is loaded? How many sixes in a row would you need to see to be at least 90% sure that the dice is loaded?

### 3 Bayes' theorem (1 point)

Show that  $P(X, Y) = P(X \mid Y)P(Y)$  implies Bayes' theorem.

### 4 Take the test? (2 points)

A rare genetic disease is discovered. Although only one in a million people carry it, you consider getting screened. You are told that the genetic test is extremely good: it is 100% sensitive (it is always correct if you have the disease) and 99.99% specific (it gives a false positive result only 0.01% of the time). Using Bayes' theorem, explain why you might decide not to take the test.

### 5 Dot matrix tool (4 points)

Write a program that takes as input two DNA or protein sequences, and two numbers  $w$  and  $s$  and draws the dot matrix for the two sequences with window size  $w$  and stringency  $s$ . To write the program, please modify the `main` method of `Program1`, obtainable from:

`www-ab.informatik.uni-tuebingen.de/teaching/ws02/abi1/programs/program01.zip`.

The `DotPlot.java` and `SimpleFileInput.java` are used by the main program to read in sequence files in fasta-format and to display your dot diagram.

Run the program on the following data sets:

- `human.pax6.fasta` vs `mouse.pax6.fasta`
- `human.similar2pax6.fasta` vs `human.pax6.fasta`
- `DBA_HUMAN.fastas` vs `LGB2_LUPLU.fasta`
- `HumanLPLreceptor.fasta` vs itself

For each case, choose appropriate values for  $w$  and  $s$ , print and discuss the resulting diagrams and explain your choice of window size  $w$  and stringency  $s$ .

## 6 Additional useful activities

For example, read chapters 1 and 2 of “Biological sequence analysis” by Durbin et al. Download some pairs of related sequences from GenBank and analyze them using the dot matrix program.

**Due by 10am, Monday, 21 Oct 2002**