

# Algorithms in Bioinformatics I, WS2006/7

## Assignment sheet # 1

Tobias Dezulian

October 24, 2006

### Probability

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

An occasionally dishonest casino uses two kinds of dice, of which 98% are fair, but 2% 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 (1 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 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 99.98% specific (it gives a false positive result only 0.02% of the time). Using Bayes' theorem, explain why you might decide not to take the test.

# DNA Compression

## 4 Preparation (1 points)

Download the following mtDNA sequences:

- primates: gorilla, human,
- ferungulates: cat,
- rodents: rat and house mouse.

You find mitochondrial genomes here:

[http://www.ncbi.nlm.nih.gov/genomes/static/euk\\_o.html](http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html)

The NCBI taxonomy database may be helpful:

<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>

For the following tasks you first need to download the program **GenCompress** from

<http://monod.uwaterloo.ca/downloads/gencompress/>

We first give the following definitions:

- $Compress(w)$  is the output sequence that results from the application of **GenCompress** to the sequence  $w$ .  
Likewise,  $Compress(w|z)$  is the output of the conditional compression of sequence  $w$  under  $z$ .
- $|Compress(w)|$  denotes the length of  $Compress(w)$  in bytes. Likewise,  $|Compress(w|z)|$  denotes the length of  $Compress(w|z)$  in bytes.
- $|w|$  denotes the length of the canonically (2-bits per nucleotide) encoded sequence  $w$  in bytes.
- The *compression ratio*  $r_{compress}(w)$  of a sequence  $w$  is defined as:

$$r_{compress}(w) = 1 - \frac{|Compress(w)|}{|w|}$$

- $r_{compress}(w|z)$  is defined likewise. Note that **GenCompress** automatically reports these compression ratios.
- $wz$  denotes the concatenation of sequences  $w$  and  $z$ .

## 5 Compression ratios (3 points)

Using the program `GenCompress`, determine the compression ratio for each mtDNA genome  $w$ . Compute the matrix of all pair-wise conditional compression ratios  $r_{compress}(w|z)$ . Compute the matrix of all pair-wise “concatenated” compression ratios  $r_{compress}(wz)$ .

## 6 Mutual information distances (2 points)

From this information, compute the “mutual information” distance matrix using the formula

$$D(w, z) := 1 - \frac{|Compress(w)| - |Compress(w | z)|}{|Compress(wz)|}$$

**Due by 10am, Monday, 23 Oct 2006**