

# Algorithms in Bioinformatics 2, SoSe2007

## Assignment sheet # 8

Christian Rausch

June 11, 2007

### 1 The Human Mitochondrial Genome (1 Point)

Go to the NCBI homepage ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)) and search “Gemome” for “homo sapiens[organism] mitochondrion”. Download one of the human genome mitochondrion sequences in FastA format. Please recall briefly what mitochondrions are and why they are important for eukaryotic cells. According to what theory, the mitochondrial genome came into the eukaryotic cell?

### 2 WGS Simulator (9 points)

To solve this task, please download the following file:

```
www-ab.informatik.uni-tuebingen.de/teaching/ss07/albi2/assign/files/assign08.zip
```

and modify the file `albi2.assembly.WGSSimulator`.

The goal of this week’s assignment is to implement a “whole genome shotgun sequencing simulator”. This program reads as input a target DNA sequence in FastA format and produces as output a multi-FastA file that contains paired-reads sampled from the given target sequence.

The program has the following parameters:

```
-i <String> : Input file
-o <String> : Output file
-mr <int>   : mean read length
-sr <int>   : standard deviation for read length
-mc <int>   : mean clone length
-sc <int>   : standard deviation for clone length
-c <double> : x-fold sequencing coverage
-e <int>    : percent sequencing error rate
```

The program samples clones from the target sequence of the given mean clone length, normally distributed with the given standard deviation.

Each end of a clone is “sequenced”, giving rise to a pair of reads of the given mean read length, normally distributed with the given standard deviation.

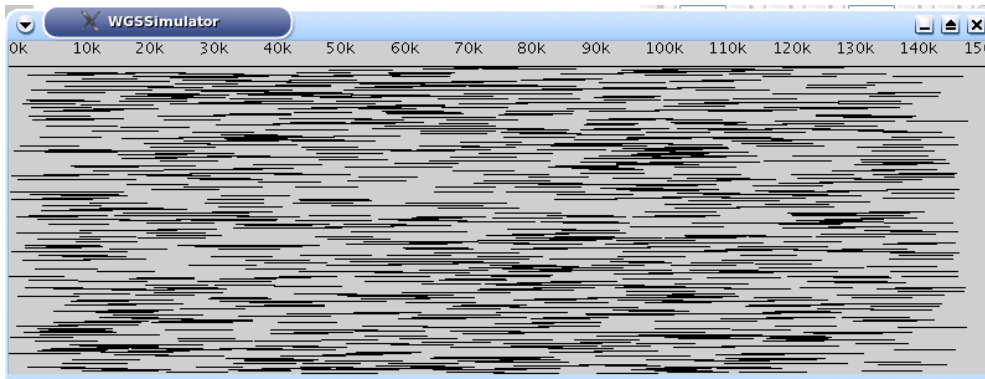
Each read contains “sequencing” errors at the given error rate, these errors are substitution errors only, not deletions or insertions.

The target sequence should be sampled at the given  $x$ -fold read coverage, that is, you must sample the correct number of reads such that any fixed position in the target sequence is contained in  $x$  reads, on average.

As mentioned above, output is in multi-FastA format, which can be generated using the class `albi2.util.FastA`. Each read is represented by a record consisting of a header and then the sequence of the read. Please format the header as shown here:

```
>name=f1047a s=19031 e=19463 l=432 ror=1 cor=0 err=9
```

Here, the name of a read consists of the letter **f**, the number of the clone, and a **a** or **z**. So, the name of two reads taken from the same clone differs only by the last letter, **a** or **z**. The values **s** and **e** give the start and end coordinates of the read in the target sequence, always from 5' to 3' end of the reads. The value **l** is the length of the read. The value **ror** is the read orientation, either 0, if it was sampled from the forward strand of the target sequence, and 1 else. The value **err** contains the number of sequencing errors in the read. Additionally, the program should produce a picture of the sampling by clones, in which horizontal lines placed at random heights in the picture indicate the positions of the sampled clones:



Please apply your program to the human mitochondrial genome (optionally to a larger genome of your choice, e.g. viral, plastid, or bacterial genome). Assume that the mean clone length is 5kb, with a standard deviation of 0.5kb, that the mean read length is 500bp, with a standard deviation of 50bp, that the coverage is 5 and the sequencing error is 1%.

Next week's assignments will assume that you have produced a working copying of the simulator.

Assignments due: **Monday, June 18, 10am**