Universität Bielefeld
Technische Fakultät

Dr. Alexander Sczyrba
Linda Sundermann

Algorithms in Genome Research
Winter 2012/2013

Exercises

Number 1, Discussion: 26. October 2012

1. Remember physical mapping by clone-probe hybridization.
   (a) What are the main assumptions when the problem is modelled as the consecutive ones problem?
   (b) Discuss experimental reasons why the assumptions do not hold in practice.
2. Solve the consecutive-ones problem for the following clone-probe hybridization matrix M (if possible).

$$M = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

3. Apply the Lander-Waterman statistics to fill out the following table for
   (a) Human Whole Genome Shotgun Sequencing ($G = 3 \times 10^9$, $L = 500$)
   (b) Human BAC sequencing ($G = 300000$, $L = 500$)

Coverage a	#reads $N = aG/L$	#nt sequences aG	%genome sequenced $(1 - e^{-a})$	mean #contigs $(G/L)ae^{-a}$	mean contig length $(e^{-a} - 1)L/a$
0.5					
1.0					
2.0					
5.0					
7.0					
10.0					
15.0					

(Hint: of course, you can also write a short program to calculate a more detailed table or plot the function as a graph!)

4. Find the shortest common superstring of the following sequences:

```
 1 ATAGCC
 2 ATATAT
 3 ATATCG
 4 CGGGAC
 5 GACATA
 6 GACTAT
 7 GCCGGT
 8 GGTATA
 9 TATATA
10 TATCGG
```

   Is the coverage uniform? If not, find a layout with a more uniform coverage.