

# Algorithms in Bioinformatics II, SS2003

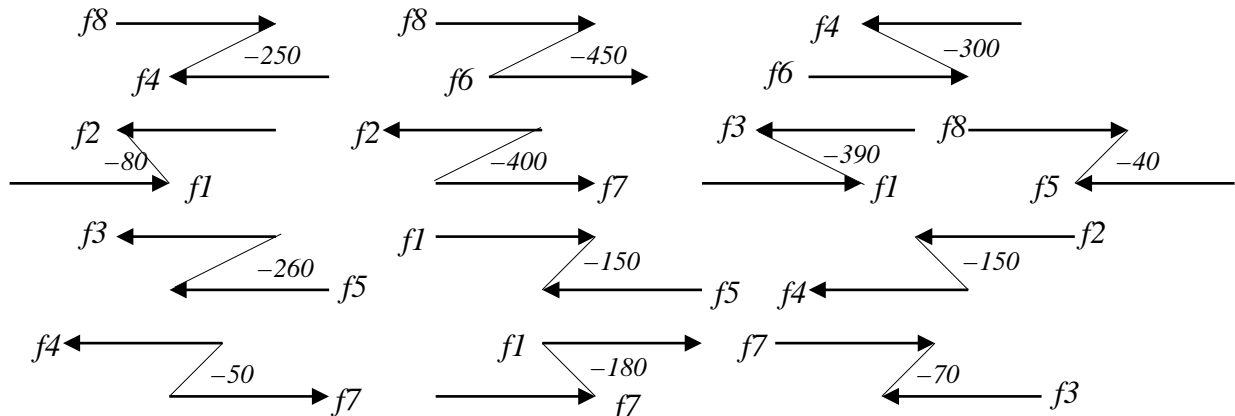
## Assignment sheet # 9

Daniel Huson

June 23, 2003

### 1 The overlap graph (3 points)

Given reads  $\mathcal{F} = \{f_1, f_2, \dots, f_8\}$ , each of approximate length 500, that overlap as follows:



Draw the overlap graph  $OG$ , labeling all edges as discussed in the lecture.

### 2 A minimal spanning tree (1 point)

Draw a minimal spanning tree in  $OG$ , containing all read edges.

### 3 The layout (1 point)

Draw the layout of the reads as given by the minimal spanning tree, indicating the approximate coordinates of the start and end of each read.

### 4 Consistent overlaps (1 point)

Are all overlaps consistent with the computed layout? If not, which overlaps are not consistent with the layout, and why?

## 5 The arrival statistic for unitigs (4 points)

Let  $R$  be the number of reads and  $G$  the estimated length of the source sequence. For a unitig of approximate length  $\rho$ , the probability of it containing  $k$  reads, that is, of seeing  $k-1$  start positions of reads in an interval of length  $\rho$ , is

$$\frac{e^{-c}c^k}{k!}, \text{ with } c := \frac{\rho R}{G},$$

if the unitig is not oversampled, and

$$\frac{e^{-2c}2^k c^k}{k!},$$

if the unitig is the result of collapsing two repeats.

Why? Please explain both formulas.

Show that the natural log of the ratio of these two probabilities is:

$$c - (\log 2)k.$$

Discuss the interpretation of this value when it is negative, is 0, and when it is positive.

## 6 Application of arrival statistic for unitigs (1 points)

Given a source sequence of length  $G = 500kb$  and a collection of  $R = 7000$  reads.

Can a unitig of approximate length  $\rho = 5000$  that consists of 250 reads safely be considered unique?

Assignments due: **Monday, June 30, 10am**