

Algorithms in Bioinformatics II, SS2003

Assignment sheet # 11

Daniel Huson

July 6, 2003

1 On l -tuples and orphans (4 points)

In a sequencing project for an unknown source sequence A of approximate length 1.5 Mb, 30000 reads of average length 500 bp have been produced, at an error rate of 1%. Questions:

- In the absence of sequencing errors and repeats, what is the number of different 25-tuples that we expect to see?
- Now, based on an error rate of 1%, give an estimation of the number of erroneous 25-tuples that we expect to see. This should make clear why error correction is necessary.
- Consider a solid 25-tuple t . How many orphans s do we expect to see as neighbors of t ?
- What is the mean multiplicity of t ?

2 The threshold M for orphans (4 points)

In the definition of an orphan, a threshold M is mentioned. For a short source sequence, $M = 1$ is a good choice. However, for a very long source sequence and high coverage, $M = 2$ or perhaps even $M = 3$ may be good.

- Why?
- To support your explanation, using a statistical model (please state explicitly), give an estimation of the probability that a given l -tuple s has a neighbor $t \in A_l$ in a random sequence A , as a function of l and the length n of A .
- In application of the previous calculation to a sequencing project in which the estimated source sequence length is N and the sequencing coverage is x , we should set $n = xN$, explain!
- Compute this probability for the following three cases (each assuming a sequencing coverage of 10):

$$|A| = 10 \text{ kb}, l = 15, \quad |A| = 1 \text{ Mb}, l = 20, \quad \text{and} \quad |A| = 150 \text{ Mb}, l = 20.$$

3 The Chinese Postman Problem (2 points)

Display a shortest closed Chinese Postman path for both of these graphs (unweighted (a) and weighted (b)) and report the number of edges/ total weight. Give an argument why the displayed tour is minimal.



Assignments due: **Monday, July 14, 10am**