Universität Bielefeld
Technische Fakultät

AG Genominformatik
B.Sc. Julia Herold

Algorithms in Genome Research

Winter 2005/2006

Exercises

Sheet 6, Discussion 05.01.2006

1. What is meant by heuristic sequence database search? In what sense is Smith-Waterman an exact method, in what sense is it heuristic? How about Blast?

2. Order the following e-values by significance, starting with the most significant one:

```
3.2e-23
5.2
12.3
.01
3.8e-66
.003
9.0e-66
1.0e-40
.005
2.8e-23
```

3. Discuss the influence of sequence database size on the significance on a Blast hit. Does the expectation of a high alignment score increase or decrease with the database size?

4. Discuss similarities and differences between EST clustering and protein clustering (data, algorithms, goal)?

5. What are q-grams? Given a DNA sequence T of length $n = 42$. How many q-grams does it contain if

   (a) q = 5

   (b) q = 2

   (c) for any q (in general)?

   Prepare to explain the idea of q-gram based filtering methods in a few precise sentences.