Computer Science 549 – Computational Biology
Prof. Steven Skiena
Fall 2005

Homework 1 Answer Key
October 11, 2005

1. Genbank and SwissProt (15 points)
   (a) refer to:
   
   &term = myoglobin
   
   (b) refer to:
   
   or
   
   http://us.expasy.org/sprot/hpi/
   
   (c) (depends)
   
   (d) refer to:
   

2. PubMed (15 points)
   (a) (depends)
   
   (b) (depends)
   
   (c) 18 results by searching for “skiena” at:
   

3. CAP-2 (40 points)
   
   (a) (depends)
   
   (b) When a sequence has repeats, it creates a lot of ambiguity during shotgun sequencing. Due to these ambiguities, sequence assembly algorithms return a large number of unordered contigs, being ineffective. Repeats of about a thousand bps are enough to create trouble with sequencing.
(c) The number of gaps increases with increase in sequence length. The number of gaps decreases with increase in coverage.

(d) The accuracy of reconstruction increases with increase in coverage. The number of gaps decreases with increase in sequencing error rate.

(e) The reconstruction accuracy differs widely over random, viral and mammalian sequences. When fragments are reconstructed into contigs, the accuracy is very good in the case of short sequences (\( \sim 1,000 \) bp), fairly good for moderately long sequences (\( \sim 20,000 \) bp) and minimal for long, mammalian sequences.

4. Multiset matching (15 points)

The algorithm should keep a counter of the number of occurrences for each character in the multiset \( S \). Firstly, scan the multiset to get the initial counter values. Secondly, go through the text \( T \), character by character. At each step, if the current substring allows the format of \( S \), decrease the corresponding counter and continue to match the next character. Otherwise, set counters to their initial values and start a new substring from the next character. Whenever all the counters reach 0, report a match and start a new substring from the next character. The first step takes \( O(m) \), the second step takes \( O(n) \) for constant-sized alphabet.

5. Short sequence occurrence (15 points)

\[ M : \text{number of all possible DNA sequences with length } n \]
\[ M = 4^n \]

\[ m : \text{number of DNA sequences (of length } n) \text{ containing an arbitrary string of length } L \]
\[ m = (n - L + 1) \cdot 4^{n-L} \]

\[ P : \text{probability that a string of length } L \text{ appears in a DNA sequence of length } n. \]
\[ P = \frac{m}{M} = \frac{(n - L + 1) \cdot 4^{n-L}}{4^n} = \frac{(n - L + 1)}{4^L} \]

By choosing appropriate probability threshold for "being very likely NOT to appear", we can get the shortest length \( l \) that is very likely NOT to occur:
\[ l = \inf(L) \approx c \cdot \log_4 n \]
where \( c \) should be around 2.

By choosing appropriate probability threshold for "being very likely to appear", we can get the longest length \( l' \) that is very likely to occur:
\[ l' = \sup(L) \approx c' \cdot \log_4 n \]
where \( c' \) should be around \( \frac{1}{2} \).