**Homework Assignment #2**

**Due date** – March 26, 2010 (Fri), 12:00PM (noon) @ Zachry 216G

**Problem 1. Local alignment (60 points)**

We want to find the best *local* alignment between two sequences \(X = x_1x_2 \ldots x_m\) and \(Y = y_1y_2 \ldots y_n\). Use the specified scoring schemes to

(i) find the \((m+1) \times (n+1)\) matrix for the *local* alignment score \(M(i, j)\)

(ii) find the maximum *local* alignment score

(iii) find the best *local* alignment between \(X\) and \(Y\).

(a) \(X = \text{LOCAL ALIGNMENT}, Y = \text{CALIGRAPHY}.\) (30 points)

Scores: \(\text{match} = +2, \text{mismatch} = -1, \text{gap} = -3.\)

(b) \(X = \text{ERROR CORRECTION CODE}, Y = \text{AUTO CORRELATION}.\) (30 points)

Scores: \(\text{match} = +3, \text{mismatch} = -1, \text{gap} = -2.\)

**Problem 2. Finding local matches. (100 points)**

Assume that we want to implement an algorithm that finds *multiple local matches* of a query sequence in a given target sequence. The specifications of the algorithm are as follows:

(i) Use the following scores: \(\text{match} = +3, \text{mismatch} = -1, \text{gap} = -2.\)

(ii) we want to find all local alignments with score \(M(i, j) \geq 5.\)

(iii) if there are two overlapping matches, we choose the one with a larger score.

(iv) the final result should contain all such *nonoverlapping* local matches.

(a) Give a “brief” sketch (step-by-step description; \(\leq 10\) lines) of your algorithm. (20 points)

(b) Write a computer program that finds all non-overlapping local matches that satisfy the above specifications. Please submit your code. (40 points)

(\textbf{Note:} You only have to find the start/end positions of the matches. You do not have to find the “alignment” itself.)
(c) Run your program to find all the local matches, where the query $X$ and the target sequence $Y$ are as follows. (40 points)

(i) $X = ABAC$
    $Y = ABCDDABABACDACDBAC$

(ii) $X = POSTSCRIPT$
    $Y = US POST OFFICE POSTERS DESCRIBING MOST CRYPTIC ACRONYMS$

Problem 3. Markov chains. (70 points)

(a) Consider the first order Markov chain shown in Fig. 1 (a). It has three states, $S$, $R$, and $C$, which represent a sunny, rainy, and cloudy day, respectively. Assume that the weather during the last three days was “sunny–cloudy–rainy”. What is the probability of having a sunny day tomorrow? (10 points)

(b) Consider again the Markov chain in Fig. 1 (a). If today’s weather is “sunny”, what is the probability of having a “rainy” day after five days and a “sunny” day after ten days? (20 points)

(Note: $X$ can be any of $\{S, R, C\}$.)

(c) When we model a family of symbol sequences using a Markov chain, we are implicitly making an assumption about its length distribution. For example, let us consider the Markov chain in Fig. 1 (b). What is the average length of the symbol sequences that are generated by the given model? (20 points)

Figure 1: Markov chains.
(d) Now, let us modify the previous model as shown in Fig. 1 (c). What is the average length of the symbol sequences that are generated by the modified model? (20 points)

**Problem 4. Generating symmetric sequences.** (70 points)

Let us consider the Markov chain shown in Fig. 2. By making one or more transitions between states, it can generate any sequence that consists of As and Bs. Given a symbol sequence, we can compute its observation probability from the transitions probabilities. For example,

\[
P(AB) = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{18}
\]
\[
P(BBB) = \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \frac{1}{54}.
\]

(a) What is the probability of generating a symmetric sequence with an even length? (30 points)

(b) What is the probability of generating a symmetric sequence with an odd length? (30 points)

(c) From (a) & (b), what is the probability that the Markov chain will generate a symmetric sequence of any length? (5 points)

(d) Let us assume that we are interested in modeling sequences with symmetric regions. Do you think using Markov chains would be a good idea? Briefly justify your answer. (5 points)

![Figure 2: Generating symmetric sequences.](image-url)