Computational Genomics
© Benny Chor and Igor Ulitsky
School of Computer Science
Tel-Aviv University
May 19, 2009

Computational Genomics: Assignment No. 3
Due June 7^{th} in Igor's mailbox

General Guidelines: This assignment is part of the final grade in the course. It should be done *independently*, either individually or in pairs, without any help from others. Duplicated and copied works will be given zero grade. Using articles, books, or web sites is perfectly acceptable as long as you include the reference in your relevant answer. Exceptionally original answers will be rewarded with bonus points. If a questions requires a description of an algorithm, you must prove its correctness and analyze its time and space complexity.

Credit: Solve 8 question out of the 9 sections for full credit. Solving all the sections will result in bonus points.

1. **Fitting a solution strategy to a real life problem** A molecular biologist has contacted you and told you that she has just discovered a novel regulatory element (sequence) found in the 3′ UTR (untranslated region) of a mouse gene. She is now interested in finding its counterparts in human transcripts. Describe the strategy that you would use to help her in each of the following scenarios. In all cases you are given a database of all the known human transcripts and want to retrieve all the transcripts that contain a matching element. (i) The element is very short (an exact sequence of 8 base pairs), but in order for it to be functional, several instances of it must be found in the 3′ UTR (not necessarily close to one another). (ii) The element is rather long (a 30 bp sequence), but some changes in it (including gaps and mismatches) can be tolerated. (iii) Same as (ii), but the element is partly degenerate: it is known in which positions a specific nucleotide is obligatory for function, and which positions can tolerate more that one nucleotide.
   You do not need to describe an algorithm in this question, but rather sketch the general strategy.

2. **Hiden Markov Models.** The legendary Mr. POKT is dining daily at his favorite cafeteria, the Lukewarm Vegie, which serves salad, lasagna or pasta (all the naturally vegetarian dishes). While always craving for something prior to leaving the office, Mr. POKT frequently changes his mind upon reaching the cafeteria. For example, if Mr. POKT craved for a salad, there is a 70% chance he'll have a salad, 20% lasagna and 10% pasta. In a similar manner, if Mr. POKT craved for a lasagna or a pasta, the probabilities are (0.9,0.05,0.05) and (0.6,0.3,0.1) (preserving the (salad\lasagna\pasta) order), respectively. What Mr. POKT craves for the next day depends on what he craved for previous day (and not on what he actually ate...): after craving for a salad,

there's a 10% chance Mr. POKT will crave for a salad the next day, 50% lasagna and 40% pasta. The transition probabilities from craving for lasagna and pasta are (0.9,0.05,0.05) and (0.7,0.2,0.1), respectively.

(a) Design an HMM describing Mr. POKT's eating behavior. Assume that the initial probabilities for (salad, lasagna, pasta) are (0.2,0.4,0.4). If in a given week Mr. POKT had: pasta, salad, lasagna, pasta and salad on Sunday to Thursday, respectively, what are the probabilities that Mr. POKT had a craving for each food type on Tuesday?

(b) What is the most probable sequence of daily cravings for this week?

3. **Predicting the Secondary Structure of a Protein.** Consider the problem of predicting the secondary structure of a protein from its sequence. Given the sequence of the protein, the problem is to partition the sequence into segments belonging to one of the following structural classes: (a) α-helix (b) β-strand or (c) coil. It is known that the secondary structure correlates with the preference for certain amino acids in the segment, and in particular with the polarity of the amino acids in it (each of the 20 amino acids is identified as hydrophobic, hydrophilic (polar), or neutral). The following features are known about these preferences:

   - α-helices frequently contain the "amphiphilic motif": a succession of two polar (p) and two hydrophobic (a) residues: "pphh" and "hhpp".
   - β-sheets frequently contain alternations of polar and hydrophobic residues: "hphp" and "phph".
   - The first and the last residue in an α-helix and a β-sheet have a special amino acid composition, designated as a "helix cap" and a "strand cap".
   - Coil sequences are not known to contain any specific amino acid preferences.
   - Both terminal ends of a protein are in a coil structure, and α-helix and β-sheet structures are always separated by coil structures.

   (a) Construct an HMM for this problem.

   (b) Describe an algorithm for segmenting a protein sequence into secondary structures.

4. Consider a Markov model with an end state, in which the transition from any state to the end state has the same probability $\tau > 0$. A sequence generated by the model is the sequence of states traversed until the end state is reached. Show that the sum of probabilities over all possible sequences of any length is 1. This proves that the model really describes a proper probability distribution over the whole space of sequences. Hint: prove first that the probability of the event "the model produces a sequence of length exactly L equals $\tau(1-\tau)^{L-1}$.

5. **Ancestral ML reconstruction** You are the computational biologist of the *Jurassic Park*. You are given a phylogenetic tree, with branch lengths, including the park animals (and only them) as internal nodes, and contemporary sequences of all their present day descendants. The sequences of the park animals are not available. You also have the *PAM* matrix (for each branch length in the tree), and you are working only on one specific protein, which is multiply aligned without gaps for all the input sequences.

   (a) Explain how to reconstruct the most plausible set of sequences of the ancient creatures, i.e, the set whose probability given the data is maximized.

   (b) In contrast to item 5a, assume all you care about is one specific creature (*T-Rex*, for instance). Fortunately, T-Rex has no known descendants, and its position in the tree is known. So in fact it is a leaf, but a "dead" one. Suggest how to reconstruct the most plausible T-Rex sequence given the contemporary sequences.

6. You are given a binary phylogenetic tree T. For each leaf (contemporary species) you are given an RNA sequence. The length of all the sequences is equal, and you are given a gapless multiple sequence alignment of all the sequences. In addition, for each contemporary species, it is known which pairs of bases form bonds in the secondary structure of the RNA (only G-C and A-U bonds are allowed). Denote this set as $P = (x_1, y_1), ..., (x_k, y_k)$. Each base can form a pair with no more than one base (some bases remain unpaired). We assume that the RNA secondary structure is conserved through evolution and it is identical in all the contemporary species. The positions of the paired bases are marked in the multiple sequence alignment. However, it is possible that, due to mutations, some bonds did not appear in some of the ancestral species. We assume that mutations in different positions are independent, except for the pairs in P. Describe an algorithm that will find the minimum cost ancestral sequences, if the cost for a base change is X and the cost for disrupting a pair in P is Y for each ancestral sequence in which the pair is absent. A toy input and a possible (but not necessarily optimal) solution are shown in Figure 1.

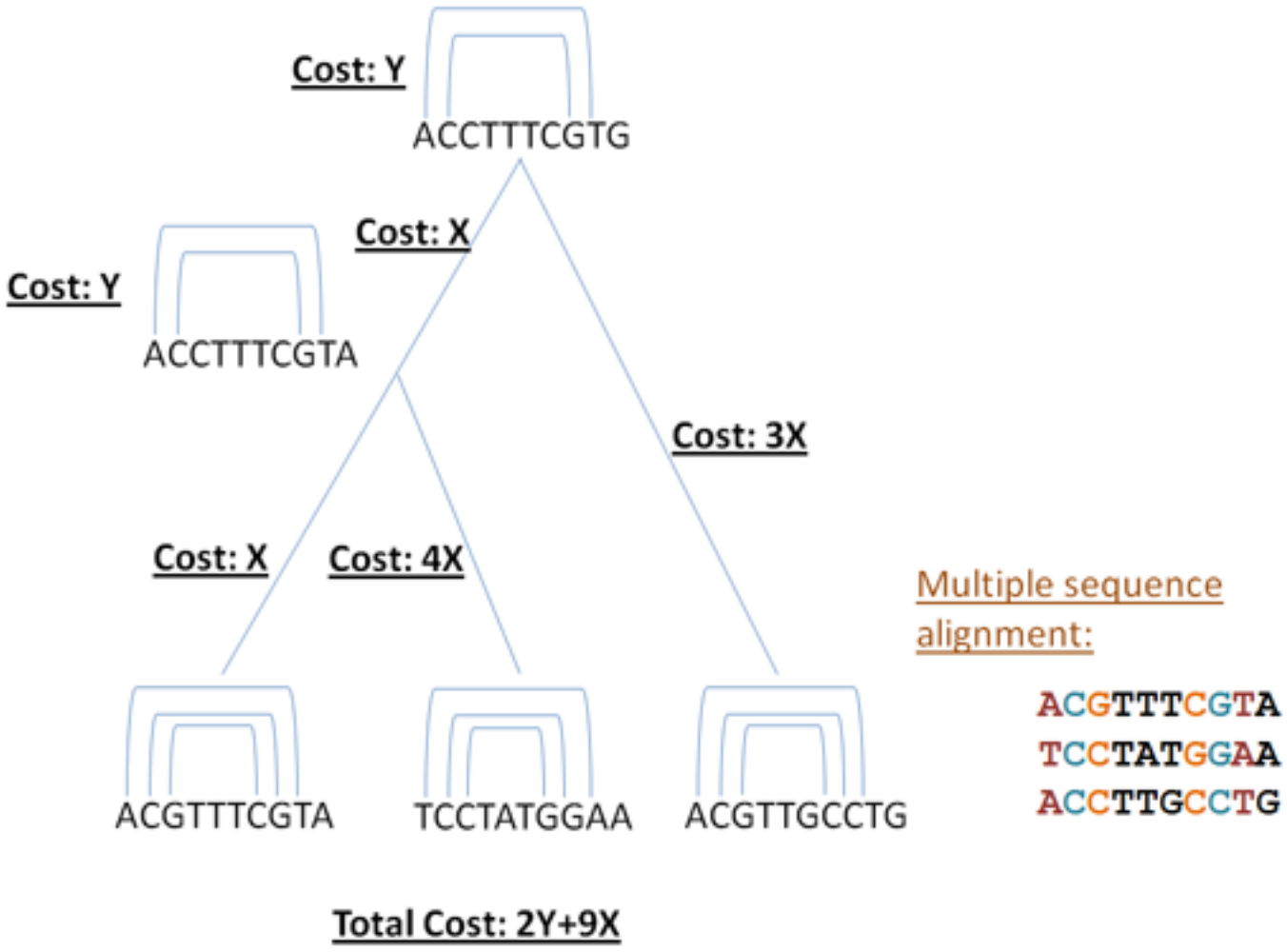

Figure 1: A toy example of problem 6. The shown reconstruction is not necessary the optimal one.