

Home Work 4

CSE/BENG/BIMM 181, Winter 2012

Given on Mon, Feb 27, 2012. Due on Mon. March 5, 2012

State your answers legibly and concisely. Your solutions will be graded on correctness, elegance, **clarity** and originality. Your proofs should **avoid getting bogged down in too much detail**. Your algorithms should be as efficient as possible (inefficient algorithms will result in lower grades). Please note that the work handed in must be your own. No late assignments will be accepted without prior arrangement. Every problem is 20 points.

The Following Problems are from textbook:

Problem 12.4 (20 pts)

The k -Means clustering algorithm randomly selects an original partition into clusters and deterministically rearranges clusters afterward. Design a randomized version of the k -Means algorithm that uses coin tossing to rearrange clusters.

Problem 8.21 (20 pts)

Develop a version of the spectral alignment algorithm that is geared to mutations rather than modifications. In this case the jumps between diagonals are not arbitrary and one has to limit the possible shifts between diagonals to mass differences between amino acids participating in the mutation.

Problem 8.26 (20 pts)

When analyzing a protein $p_1 \dots p_n$, a mass spectrometer measures the masses of both the prefix peptides $p_1 \dots p_i$ and of the suffix peptides $p_i \dots p_n$, for $1 \leq i \leq n$. The prefix-suffix mass spectrum includes the masses of all prefix and suffix peptides. For example, CSE produces the following prefix-suffix spectrum $\{103, 129, 103 + 87, 129 + 87, 103 + 87 + 129\} = \{103, 129, 190, 216, 319\}$ and it remains unknown which masses in the prefix-suffix spectrum are derived from the prefix peptides and which are derived from the suffix peptides.

The prefix-suffix spectrum may contain as few as n masses (for palindromic peptides with every suffix mass matched by a prefix mass) and as many as $2n - 1$ masses (if the overall peptide mass is the only match between suffix and prefix masses).

Reconstruct a peptide given its prefix-suffix spectrum. Devise an efficient algorithm for this problem under the assumption that the prefix-suffix spectrum of a peptide of length n contains $2n - 1$ masses.

Problem 4 (20 pts)

Show that starting from different initializations, k -means algorithm leads to different results. Construct the instance of k -means clustering problem for which Lloyd algorithm leads to particularly bad solution. Can you construct an instance for which Lloyd algorithm produces a solution that has the squared error distortion that is 2 times larger than the optimal.

Problem 5: Programming (20 pts)

Implement an algorithm for solving the exact Partial Digest Problem (presented in class on February 22) and use it to solve the problem 4.2.

Problem 4.2 (reproduced here):

Consider partial digest $L = \{1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6, 6, 6, 9, 9, 10, 11, 12, 15\}$. Solve the Partial Digest problem for L (i.e., find X such that $\Delta X = L$).