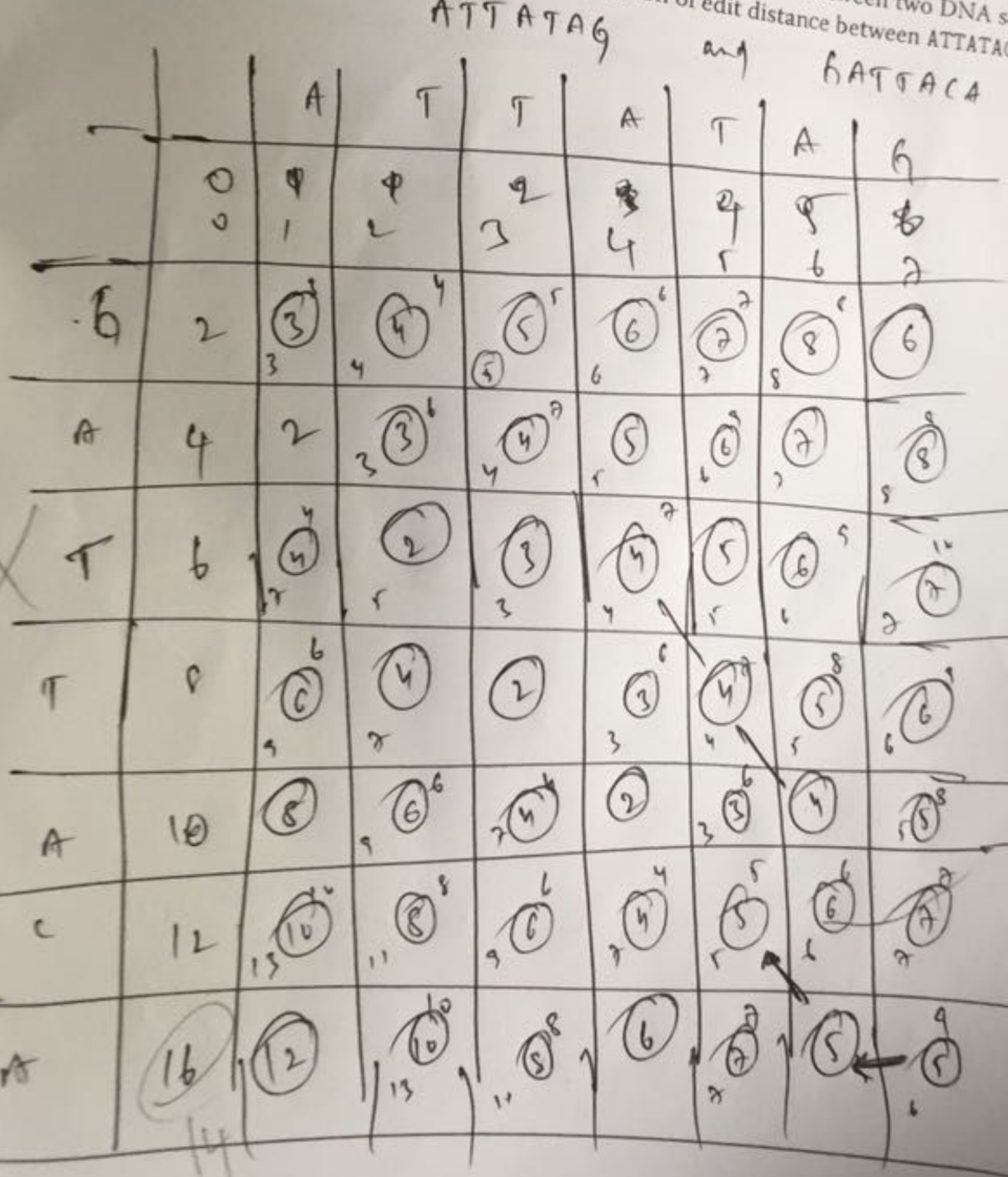
10. (8 marks) When DNA sequences are aligned to compute evolutionary distances, the insertion, deletion and substitution costs are seldom unity. A simple substitution cost matrix is usually used, where the cost of substitution of a purine (A, G) by another is unity, while the substitution of a purine by a pyrimidine (C, T) or vice-versa is twice that. The cost of an insertion and deletion are also 2. Write the formulation to compute edit distance between two DNA sequences based on this scheme, and illustrate the computation of edit distance between ATTATAG and GATTACA.



BT 3051 — Data Structures and Algorithms for Biology
Jul–Nov 2016
Quiz 2

October 25, 2016

Name: N. Ramakeo **Roll number:** B E 1 3 B D 2 1

Instructions: This quiz is 'closed book', but you can use *your own* written class notes. Possession of any other material will be construed as cheating. Answer all questions. **Keep your answers brief and to the point.** Allotted time is 50 minutes. There are a total of **6** pages in this quiz.

Total marks: 50

1. (8 marks) Write out a regular expression for each of the following cases. (Assume that you are `grep -E`, so use suitable 'escapes'):

   (a) All words beginning with a vowel and ending with a vowel

   (b) All words with at least 12 letters

   . . .

   (c) All words beginning with a t but having no t thereafter

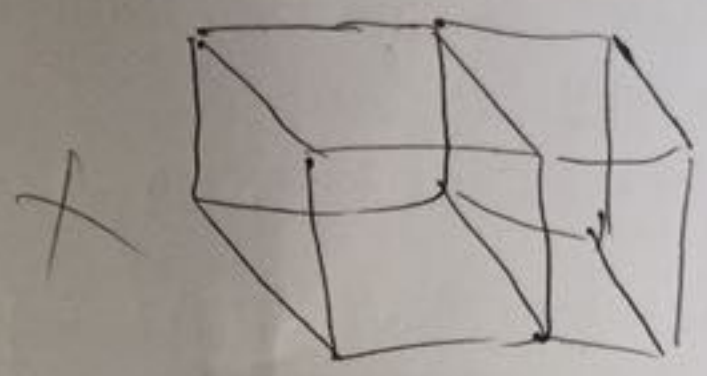
   ^at

   (d) Five letter palindromes

   (e) Stretches of repeated codons in DNA (any reading frame is acceptable)

   (f) Words with even number of 'a's

2. (3 marks) Draw a simple undirected graph G that has 12 vertices, 18 edges, and 3 connected components.



3. (3 marks) What is memoisation? How does it help?

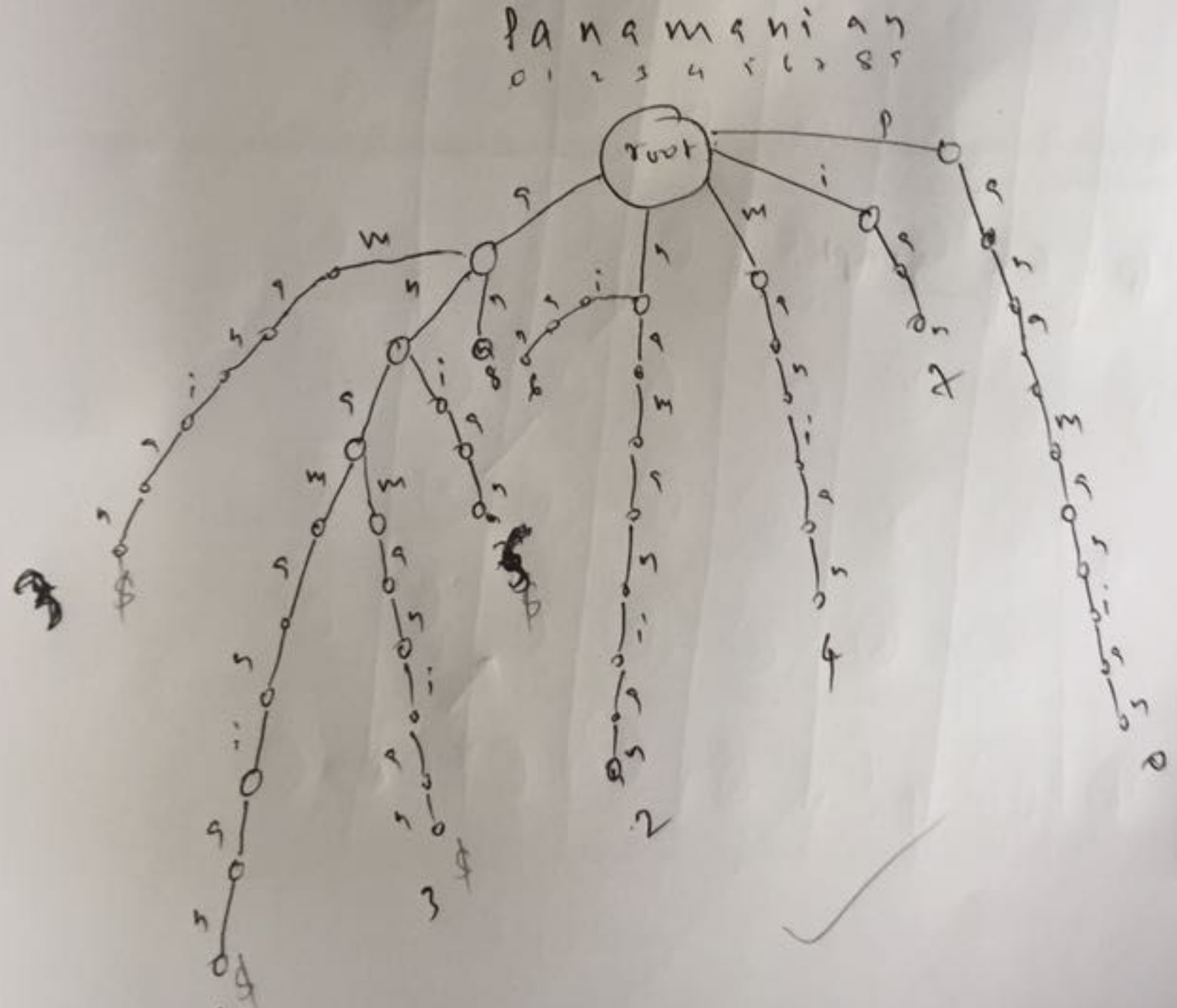
4. (4 marks) Outline the key idea underlying the Knuth-Morris-Pratt algorithm for string matching. Illustrate with an example.

For Certain alignment of the pattern, if we find
Several matching characters but then detect a mismatch, we ignore
all the information gained by the successful comparisons after
restarting with the next incremental placement of the pattern

5. (4 marks) To search a text composed of strings with a very small alphabet, such as DNA, which algorithm would you prefer: (a) Boyer–Moore (b) Knuth–Morris–Pratt? Why?

(a) Boyer - Moore

6. (5 marks) Construct a **suffix trie** for the text Panamanian. Illustrate what happens when you query for the pattern 'an' against the text.



When we query for the pattern "an" we'll get 3 matches at i, 5 and 8

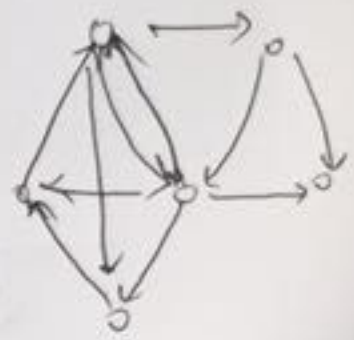
7. (4 marks) What is a *strongly connected component*, in a directed graph? Explain with an example.

there is path in either direction is called Strongly connected Component.

Example: Twitter

a can follow b, but

b may not follow a



8. (5 marks) Outline an algorithm (not code!) to create a hypercube graph. What is its complexity? What would be the complexity of the most efficient algorithm?

Hyper Cube graph:

9. (6 marks) There are 100 pairs of cuvettes containing distinct iGEM Bio-bricks (DNA samples, essentially), marked "O" and "M" for original and mutated samples. The mutated samples contain a few point mutations (single base replacements) from the corresponding original sample, introduced so as to improve the performance of the bio-brick. Unfortunately, while shipping these cuvettes to a collaborator for further analysis, the samples somehow got mispaired. However, for further analyses, it is important to re-pair the samples. Fortunately, every sample has a unique identifier, based on which the sequence can be obtained from a database. Can you suggest an algorithm to re-pair the samples, in an efficient manner? List any assumptions that you make.