Bioinformatics Example Sheet 2

Genome assembly

Exercise 1 [Problem specification]

(a) What is *genome sequencing*? How is it used in bioinformatics?

(b) What is *genome assembly*? How is it used in bioinformatics?

(c) Attempt **[2019P9Q2 (b)]**.

(d) Define the *string reconstruction problem*. How does it differ from the genome assembly problem?

Exercise 2 [Hamiltonian path]

(a) Define the *Hamiltonian path* problem.

(b) Describe the reduction of the string reconstruction problem to the Hamiltonian problem.

(c) What are the disadvantages with this reduction?

Exercise 3 [Algorithms for the Hamiltonian path problem (optional)]

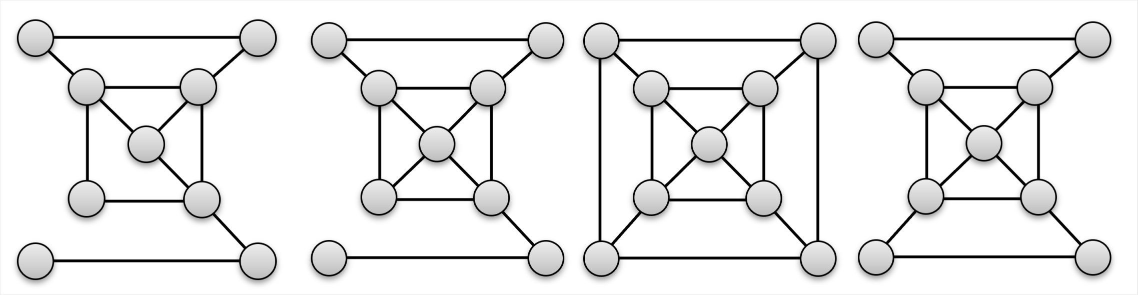
(a) Describe the brute force algorithm for the Hamiltonian path problem. What is its time complexity?

(b) Can you describe a DP approach for the Hamiltonian path problem which improves the brute force search?

Exercise 4 [Euler path modelling]

(a) Describe the *Eulerian path* problem.

(b) Find the Eulerian path in the following graphs.

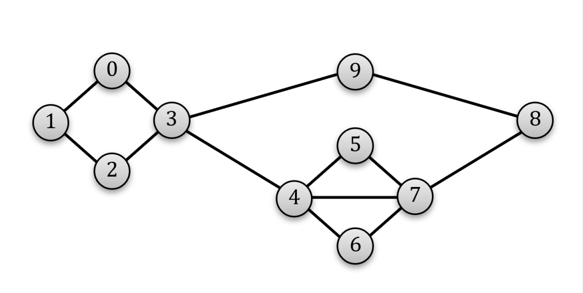


(c) Explain how the genome assembly problem is modelled as a Eulerian path problem.

(d) Consider a connected graph $G = (V, E)$ where all vertices (except for possibly two) have an even degree. Argue that G has an Eulerian path.

(e) Describe an algorithm that finds the Eulerian path in a graph or determines that no such path exists. What is the time complexity of your algorithm?

(f) Trace your algorithm in the following graph.



(g) (optional) Implement the algorithm.

Exercise 5 [De Bruijn sequence]

(a) Define a *de Bruijn sequence* binary strings of length n.

(b) Find a de Bruijn sequence for binary strings of length 3.

(c) Explain how to find such a sequence for binary strings of length n. How efficient is your algorithm? Can there exist shorter sequences? Why?

(d) Does this method extend to non-binary alphabets, e.g., $\{a, b, c\}$?

(e) Attempt **[2017P7Q4 (c)]**.

(f) Attempt **[2020P9Q2 (d)]**.

Exercise 6 [Magic trick (optional)] Explain how this magic trick works.

Further Reading 1 [Bit handling and de Bruijn sequences] Read this and this, and explain how to efficiently find the position of the most significant bit of a 32-bit integer using only a few number of arithmetic operations.

Exercise 7 [Read pairs]

(a) Explain what *read pairs* are.

(b) What problem can arise and how can it be solved computationally?

(c) Are the modelling assumptions realistic?

Clustering

Exercise 8 [The clustering problem]

(a) Define *clustering*.

(b) What is the *good clustering principle*? Do you agree with it?

(c) Attempt **[2015P7Q3 (d)]**.

Exercise 9 [k-center clustering problem]

(a) Define the *k-center clustering* problem.

(b) Does it always make sense to have a center for a cluster?

(c) Explain the *farthest-first traversal* algorithm.

Exercise 10 [A guarantee for farthest-first traversal (optional)] In this exercise, you will prove that the farthest-first traversal heuristic finds a solution which is at most 2× the optimal. Let d be the largest distance of a point to the closest of the k centers returned by the farthest-first traversal algorithm.

(a) Argue that there must be $k+1$ points with pairwise distance at least d.

(b) Create a ball of radius $r/2$ around each of these $k+1$ points. Argue that these balls do not overlap.

(c) Argue that for any k-clustering there will be one ball without a center. What can you deduce from this?

Further Reading 2 [Approximation hardness] Approximating the k-center problem in $2 - \epsilon$ is NP-Hard. For the details, see “A best possible heuristic for the k-center problem” by S. D. Hochbaum and

D. B. Shmoys (e.g. here).

Exercise 11 [Clustering metrics]

(a) Describe and compare different metrics for evaluating a clustering.

(b) How would you evaluate a clustering algorithm?

(c) Attempt **[2020P9Q2 (b)]**.

Exercise 12 [*k*-means clustering]

(a) Prove that the *center of gravity* minimises the distortion to all points.

(b) Describe *Lloyd's algorithm.*

(c) Does it always converge to the optimum? Why?

Exercise 13 [Hierarchical clustering] Describe approaches for *hierarchical clustering.*

Exercise 14 [Soft/hard clustering] What is *soft* clustering? How does it compare to *hard* clustering?

Exercise 15 [Markov Clustering algorithm]

(a) Describe the *Markov Clustering algorithm.*

(b) What is the time complexity for each step of the algorithm?

Expectation Maximisation

Exercise 16 [Expectation Maximisation]

(a) Describe the two steps of *Expectation Maximisation (EM).*

(b) Describe EM in the example with the coin-tosses given in the lecture notes.

(c) Describe EM in the context of clustering.

(d) On what kind of datasets can we apply EM? Are there any guarantees for its convergence?