02-712 Week 2

Biological Modeling and Simulation

Aidan Jan

September 2, 2025

More (Intractible) Graph Problems

Matching Problems

- Input: $G = (V, E)$, $w : E \to \mathbb{R}$
- Output: $E' \subseteq E$ such that $\nexists u, v, w \in V$ such that $(u, v) \in E'$ and $(u, w) \in E'$
- Objective: $\max \sum_{e \in E} w(e)$

Basically, a matching is a set of edges that do not share any vertices. The question is, what is the highest weighted matching on the graph?

- Interestingly, this problem is intractible for general graphs. However, it is solvable in polynomial time for bipartite graphs.

Vertex Cover

- Input: Undirected $G = (V, E)$, $B \in \mathbb{N}$.
- Objective: Does there exist $V' \subseteq V$ such that $\forall (u, v) \in E$, $u \in V'$, $v \in V'$, $|V'| \leq B$?

Basically, a vertex cover is a set of vertices such that every edge on the graph is conneccted to at least one of the vertices in the set. Find the minimum number of vertices required to create a vertex cover.

Independent Set

- Input: $G = (V, E)$, $B \in \mathbb{N}$.
- Objective: Does there exist $V' \subseteq V$ such that $\nexists (u, v) \in E$ where $u \in V'$, $v \in V'$ such that $|V'| \geq B$?

Find the largest set such that no two vertices in the set are connected by an edge.

Max Clique Problem

- Input: $G = (V, E)$, $B \in \mathbb{N}$.
- Objective: Does there exist $V' \subseteq V$ such that $\forall v_1, v_2 \in V'$, $(v_1, v_2) \in E$, and $|V'| \leq B$?

A clique is another name for a complete subgraph. (e.g., a subgraph where every vertex in it is connected to every other vertex in it via an edge.) What is the largest clique in graph G?

Graph Coloring

- Input: $G = (V, E)$, $k \in \mathbb{N}$.
- Objective: Does there exist $e : V \to \{1, \ldots, k\}$ such that $\forall(u, v) \in E$, $c(u) \neq c(v)$?

Each edge is 'colored' with a number less than k. Is it possible to color the entire graph such that no neighboring vertices are colored the same?

Max Subgraph / Max Induced Subgraph with Property Π

- Input: $G = (V, E)$, $B \in \mathbb{N}$.
- Objective (variant 1): Does there exist $G' = (V, E')$, where $E' \subseteq E$ such that G has property Π if $|E'| \geq B$?
- Objective (variant 2): Does there exist $G' = (V', E')$, where $V' \subseteq V$ and $(u, v) \in E'$ and $u \in V' \wedge v \in V'$ if $|V'| \geq B$?

Variant 1: Does the graph have property Π?
Variant 2: Does the graph contain any subgraphs with property Π?
Property Π is a nontrivial, inheritable property of a graph. (Defined separately.) Some examples of possibilities of Π:

- Bipartiteness
- Independence
- Planarity

Nontrivial = are there an infinite number of graphs that do and do not have property Π?
Inheritable = can adding edges or vertices change the result? If not, it is inheritable.

String Problems

Longest Common Subsequence

This is a problem that is solvable (runnable in $O(n^2)$), but not tractible. Usually genomes are way too big for this to be practical at all.

- Given a multiple sequences of letters (or nucleotides):
  - ABCABCABC
  - CABBCCBB
  - ABCABABC
  - etc.
- What is the longest sequence of letters that is in each sequence? The letters don't need to be consecutive. ($O(|S||T|)$)
- This is a variation of the longest common substring, where the characters must be consecutive. ($O(|S| + |T|)$)

Formal definition for longest common subsequence:

- Input: sequences $w_1, w_2, \ldots, w_k$, $B \subseteq \mathbb{N}$
- Objective: Does there exist a sequence s of $w_1, \ldots, w_k$, where $|s| \geq B$?

This problem is intractible in k. This means that if k is known (for example, you KNOW you are aligning 3 genomes), or treated as a constant, it is tractible. If k is treated as a variable, then this problem is intractible.

Shortest Common Superstring

Given a sequence of short strings:

- ACGGA
- GGACT
- ACTCCA
- CAGG

What is the shortest string that contains all of the short strings?

- For this example: ACGGACTCCAGG

This is an intractible problem.

Shortest Common Supersequence

Given a sequence of short strings:

- GCGCA
- CGATA
- ACGAAA

What is the shortest string that contains all of the sequences? (The characters don't have to be consecutive.)

- For this example: GACGCATAA

Set Problems

Minimum Test Set

- Input: set $S = \{s_1, \ldots, s_n\}$, collection $C = \{c_1, \ldots, c_m\}$, where each $c_i \subseteq S$, $B \in \mathbb{N}$
- Objective: Does there exist a $C' \subseteq C$, such that $\forall s_i, s_j \in S \exists c \in C'$ such that $|\{s_i, s_j\} \cap C| = 1$, $|C'| \leq B$?

For example, consider

- $s_1 =$ ATACC
- $s_2 =$ CTGTC
- $s_3 =$ CGGTG
- $s_4 =$ ATGTG
- $s_5 =$ AGGCC
- $c_1 = \{s_1, s_3, s_5\}$
- $c_2 = \{s_1, s_2, s_4\}$
- $c_3 = \{s_2, s_3, s_4, s_5\}$

The question is, what is the minimum of subset of collections you need to be able to narrow down a to a single set that is in all the collections?

- In this example, c_1 is the collection of sets with A as the first letter. c_2 is the set with T as the second letter. c_3 is the set with G as the third letter.
- Pretend we have sets for the forth and fifth letters too.
- Suppose you extract a bacterial genome and it matches one of the sets, but you don't know which. What is the minimum number of positions in the genome you need to look at to determine the set?

Minimum Set Cover Problem

- Input: Set S, Collection C, Bound B.
- Objective: Does there exist $C' \subseteq C$ such that $\forall s \in S \exists c \in C'$ such that $s \in C, |C'| \leq B$?

Practical explanation: suppose you have a list of antibiotics, and each one is able to kill a set of bacteria $c \in C$. You have a pool of bacteria (S) that you want to kill. What is the minimum number of antibiotics needed to kill all the bacteria?