Finding Diverse Strings and Longest Common Subsequences in a Graph

Luca Lombardo

Seminar for the course of Bioinformatics

Structure of the Presentation

1. Introduction to the problems
   - Diverse Strings Problems
   - Σ-DAG for LCSs
2. Exact Algorithms for Bounded Number of Diverse Strings
   - Max-Min Diverse String Set problem
   - Max-Sum Diverse String Set problem
3. Approximation Algorithms for Unbounded Number of Diverse Strings
   - Approximation Algorithm for Max-Min Diverse String Set problem
   - Max-Sum Farthest r-String problem
4. FPT Algorithms for Bounded Number and Length of Diverse Strings
5. Hardness Results
   - Hardness of Diverse String Set for Unbounded K
   - Hardness of Diverse String LCSs for Unbounded K

Longest Common Subsequence (LCS)

Definition

Given a set of m strings $S = \{S_1, S_2, \ldots, S_m\}$, a **common subsequence** (CS) is a sequence that appears in all m strings. A **longest common subsequence** LCS is a common subsequence of maximum length. We denote the set of all LCSs of S as $LCS(S)$.

The goal is to find a diverse set of solutions to the LCS problem under the Hamming distance.

Definition

Given two strings $X, Y \in \Sigma^r$, the **Hamming distance** between X and Y, denoted with $d_H(X, Y)$, is the number of positions at which the corresponding symbols differ.

A simple example

Add here the example of the paper

Efficient methods for finding a diverse set of solutions

More formally, let's consider the following two diversity measures for a multiset $\mathcal{X} = \{X_1, X_2, \ldots, X_K\} \subseteq \Sigma^r$ of solutions, allowing repetitions:

$$D^{\text{sum}}_{d_H}(\mathcal{X}) = \sum_{1 \le i < j \le K} d_H(X_i, X_j) \qquad \text{MAX-SUM DIVERSITY} \tag{1}$$

$$D^{\text{min}}_{d_H}(\mathcal{X}) = \min_{i<j} d_H(X_i, X_j) \qquad \text{MAX-MIN DIVERSITY} \tag{2}$$

Notation

A subset $\mathcal{X} \subseteq \Sigma^r$ is Δ-diverse w.r.t. $D^{\tau}_{d_H}$ if $D^{\tau}_{d_H}(\mathcal{X}) \ge \Delta$ for some $\Delta \ge 0$.

Where for $\tau \in \{sum, min\}$, $D^{\tau}_{d_H}$ denotes one of the two diversity measures.

Two problems

Problem 1: Diverse LCSs with Diversity measure $D^{\tau}_{d_H}$

Input: A set $S = \{S_1, S_2, \ldots, S_m\}$ of $m \geq 2$ strings over Σ, an integer $K \geq 1$ and $\Delta \geq 0$.
Question: Is there some set $\mathcal{X} \subseteq LCS(S)$ such that $\mathcal{X} = K$ and $D^{\tau}_{d_H}(\mathcal{X}) \geq \Delta$?

Problem 2: Diverse String Set

Input: $K, r, \Delta \in \mathbb{Z}$ and a Σ-DAG G for a set $L(G) \subseteq \Sigma^r$ of strings.
Question: Decide if there exists some subset $\mathcal{X} \subseteq L(G)$ such that $|\mathcal{X}| = K$ and $D^{\tau}_{d_H}(\mathcal{X}) \geq \Delta$.

Σ-Labeled Directed Acyclic Graphs (Σ-DAGs)

Definitions

- **Alphabet (Σ)**: Set of symbols.
- **String Set (Language)**: $L = \{X_1, X_2, \ldots, X_n\} \subseteq \Sigma^*$, with:
  - **Total Length**: $||L|| = \sum_{X \in L} |X|$
  - **Max Length**: $\mathrm{maxLen}(L) = \max_{X \in L} |X|$
- **r-String**: Any string X where $|X| = r$

Σ-DAG Structure

A graph $G = (V, E, s, t)$ with:

- V: vertices, E: labeled edges (v, c, w) with $c \in \Sigma$
- **Source** s and **Sink** t such that paths exist from s to all vertices
- **Size**: $\mathrm{size}(G)$, the number of its labeled edges

Σ-Labeled Directed Acyclic Graphs (Σ-DAGs)

Any path $P = (e_1, e_2, \ldots, e_n)$ of outgoing edges *spells out* a string str(P) $= c_1 c_2 \ldots c_n \in \Sigma^n$ where c_i is the label of edge e_i.

Language Representation

A Σ-DAG represents $L(G) \subset \Sigma^*$: all strings spelled from paths $s \to t$. Equivalent to an NFA over Σ with initial s, final t, and no ϵ-edges.

Remarks

- For any set L of strings, a Σ-DAG G exists s.t. $L(G) = L$ and size$(G) \leq ||L||$. Construction time: $O(||L|| \log |\Sigma|)$.
- If G represents a set L of r-strings ($L \subseteq \Sigma^r, r \geq 0$), all paths $s \to v$ have the same length $d \leq r$.

Algorithm for Max-Min Diverse String Set problem

Algorithm for Max-Min Diverse String Set problem

Algorithm for MAX-SUM DIVERSE STRING SET problem

Algorithm for MAX-SUM DIVERSE STRING SET problem

Max-Min Diverse String Set problem

MAX-SUM FARTHEST r-STRING problem

FPT Algorithms for Bounded Number and Length of Diverse Strings

FPT Algorithms for Bounded Number and Length of Diverse Strings

Hardness of Diverse String Set for Unbounded K

Hardness of Diverse String LCSs for Unbounded K