On Solving the Multiple Variable Gapped Longest Common Subsequence Problem

Marko Djukanović[1,2, 6] Nikola Balaban[2] Christian Blum[3]
Aleksandar Kartelj[4] Sašo Džeroski[5] Žiga Zebec[6]

[1]University of Nova Gorica, Nova Gorica, Slovenia
[2]Faculty of Natural Sciences and Mathematics, University of Banja Luka, Banja Luka, Bosnia and Herzegovina
[3]Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain
[4]Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
[5]Jožef Stefan Institute, Ljubljana, Slovenia
[6]Institute of Information Sciences (IZUM), Maribor, Slovenia

– Eurocast 2026: 20th International Conference on Computer Aided Systems Theory, February 23-27, 2026, Las Palmas de Gran Canaria, Spain–

Outline

- Introduction & Preliminaries
- Problem Definition
- Graph state space
- Iterative multi-source Beam Search
- Experimental Evaluation
    - General problem
    - Special problem
- Conclusions

Introduction

- Objects we deal with: sequences (strings) over finite alphabet
  - DNA/RNA over $\{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}/\mathtt{U}\}$
  - Proteins over 20 (canonical) amino acids: $\{\mathtt{A}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{P}, \mathtt{Q}...\}$
- Computational biology
  - **One of central tasks:** sequence comparison, finding common motifs between sequences
  - compare structurally but also semantically/functionality
  - sequence alignment problems
- Subsequences: reveal structural similarities → **Longest commmon subsequence problem** variants

Longest common subsequence problem (LCSP)

- Basic problem in Computational biology
- Intensively solved over last 50 years
  - theoretically as well as practically
  - Practically: many approximation algorothms, (meta-) heuristics, exact approaches, etc.

Definition (LCSP)

Input: Given a set of sequences $S = \{s_1, \ldots, s_m\}$
Task: Find a subsequence s which is common for all sequences from S of maximum possible length.

Example

Input: S={`AATTGC`, `ATTAC`}
LCS solution: s=ATTC

Literature & LCS Problem Variants

Theoretically:

- When $m = 2$ – polynomially solvable (in $O(n^2)$): Dynamic programming, Hunt-Schlimanski, ...
- When m arbitrary large – $\mathcal{NP}$-hard:
    - subject of interest within last 30 years: approximation approaches, meta-heuristics (ACO, Beam search, ...), but also exact approaches (A*, anytime approaches,...)

Problem-related variants:

- Arc-annotated LCS problem
- Constrained (restricted/imposed) LCS problem
- Repetition-free, Longest filled LCS,...
- **Gapped LCS problem**

The gaped LCS problem

Definition (A gap sequence)

Given is a sequence s and an assigned function $G_s \colon \{1, \ldots, |s|\} \mapsto \mathbb{N}$. An ordered pair (s, G_s) is called a **sequence with gaps**.

Definition (A gapped subsequence)

Sequence $\tilde{s}$ is a gapped subsequence of (s, G_s) ifff

- S is a subsequence of s
- the gapped constraint G_s is fulfilled

Literature & Motivation

content...