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], and Žiga Zebec[6]

[1] University of Nova Gorica, Nova Gorica, Slovenia
marko.dukanovic@ung.si
[2] Faculty of Natural Sciences and Mathematics, University of Banja Luka, Banja Luka, Bosnia and Herzegovina
nikola.balaban@student.pmf.unibl.org, marko.djukanovic@pmf.unibl.org
[3] Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain
christian.blum@iia.csic.es
[4] Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
kartelj@matf.rs
[5] Jožef Stefan Institute, Ljubljana, Slovenia
saso.dzeroski@ijs.si
[6] Institute of Information Sciences (IZUM), Maribor, Slovenia
ziga.zebec@izum.si

The *Longest Common Subsequence Problem* (LCSP) [1] is a well-known combinatorial optimization problem with numerous applications in computational biology, particularly in the analysis of molecular sequences. Given an arbitrary set $S = \{s_1, \ldots, s_m\}$ of input sequences over an alphabet Σ, the goal is to determine the longest possible subsequence common to all $s_i \in S$. In recent decades, many practically motivated variants of the LCSP have been proposed by considering additional structural or biological constraints such as the constrained, arc-preserving, repetition-free version, etc. This work is focused on solving the *Variable Gapped* LCSP (VGLCSP), introduced in [3] and theoretically examined in [2]. This problem extends the standard LCSP by imposing flexible distance limits ("gaps") between consecutive symbols in the resulting subsequence, allowing these gaps to vary along the input sequences. Formally, for each input sequence $s_i \in S$, a gap function $G_{s_i}\colon \{1, \ldots, |s_i|\} \mapsto \mathbb{N}$ defines the maximum distance allowed between consecutive characters of a subsequence. That is, if the characters of a subsequence s appear at the positions $i_i^1 < \ldots < i_i^{|s|}$ in s_i, we say that the gap constraint is satisfied if $i_i^x - i_i^{x-1} \leq G_{s_i}[i_i^x] + 1$ for all $x = 2, \ldots, |s|$. A common subsequence s is considered feasible if the corresponding gap constraints hold across all sequences in S. The VGLCSP represents a flexible and realistic model of sequence alignment, which is particularly useful for comparing DNA or protein sequences with variable structural distances. Although exact dynamic programming approaches exist for the two-sequence case ($m = 2$), their computational complexity becomes prohibitive for larger m. To the best of our knowledge, no effective exact or approximate approaches currently exist for the generalized VGLCSP, a clearly NP-hard problem.

Building on the well-established state-space representation of the LCSP [1], we propose a general *state graph* formulation for the VGLCSP. Each node rep-

resents a partial feasible subsequence, characterized by the vector of current positions in each input sequence and the subsequence length. Specifically, a state $v = (\mathbf{p}^{L,v}, l^v)$ is induced by a partial solution s^v if: (i) $p_i^{L,v}-1$ is the smallest index such that s^v is a subsequence of the prefix string $s_i[1]\cdots s_i[p_i^{L,v}-1]$; (ii) $l^v = |s^v|$; and (iii) all the gaps G_{s_i} are preserved with s^v. A directed arc $\alpha = (v_1 v_2)$, labeled $lett(\alpha) = a \in \Sigma$, exists between states v_1 and v_2 if $l^{v_2} = l^{v_1}+1$ and $s^{v_2} = s^{v_1}\cdot a$. To expand a state v, all letters that occur in each suffix of s_i defined by respective positions $p_i^{L,v}$ are considered. For each such letter a, the minimal positions $p_i^{L,a} \geq p_i^{L,v}$ of its appearances are identified fulfilling $p_i^{L,a} - p_i^{L,v} \leq G_{s_i}(p_i^{L,a})$. The resulting child node, if not suboptimal, is given by $w = (p^{L,a}+1, l^v+1)$. The root node $r := ((1,\ldots,1),0)$ represents the empty subsequence, while goal states correspond to non-extendable feasible subsequences. Unlike the standard LCSP, the VGLCSP admits multiple (possibly exponentially many) root nodes. For example, in sequences $s_1 =$ `ATGGAAAA` and $s_2 =$ `ATCCAAAA` with $G_{s_1} = G_{s_2} = 1$, the state $((1,1),0)$ cannot reach any state with the position vector $(5,5)$ via transitions, missing optimal subsequence `AAA`. This motivates the methodology of *Iterative Multi-Source Beam Search* (IMSBS). BS is a popular meta-heuristic approach that performs in a breadth-first-search manner, controlled by the size of beam β and a heuristic h. In the IMSBS, BS is iteratively executed with a set of promising root nodes as an initial beam, selected from a dynamically maintained global pool of root nodes, controlling diversification of the approach. Each iteration performs a complete BS, guided by a heuristic h based on letter frequencies, minimal residual substring lengths, or probability-weighted estimates.

Preliminary experiments on synthetically generated instances demonstrate that IMSBS consistently outperforms the single-root BS baseline initialized at $r = (\mathbf{1}, 0)$. The results reveal clear benefits of multi-source exploration and highlight differences between heuristic strategies in terms of efficiency and solution quality. These findings open promising directions for designing more advanced hybrid and learning-assisted approaches to tackle the generalized VGLCSP.

Acknowledgments. This publication is co- funded by the European Union's Horizon Europe research and innovation program under the Marie Sklodowska-Curie COFUND Postdoctoral Programme grant agreement No.101081355– SMASH and by the Republic of Slovenia and the European Union from the European Regional Development Fund. Co-funded by European Union. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the REA can be held responsible for them.

References

1. Finding longest common subsequences: New anytime A* search results. Applied Soft Computing **95**, 106499 (2020)
2. Adamson, D., Kosche, M., Koß, T., Manea, F., Siemer, S.: Longest common subsequence with gap constraints. In: International Conference on Combinatorics on Words. pp. 60–76. Springer (2023)
3. Penga, Y.H., Yangb, C.B.: The longest common subsequence problem with variable gapped constraints. In: Proceedings of the 28th Workshop on Combinatorial Mathematics and Computation Theory, Penghu, Taiwan. pp. 17–23 (2011)