Exploring Longest Common Subsequences and the Chvátal-Sankoff Constants

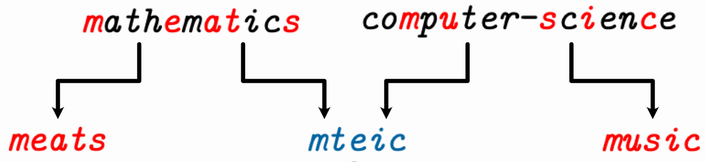
Chase Miller (CS), Andrew Salls (CS/MA), Duncan Soiffer (CS/MA)
Advisors: Professor George Heineman (CS), Professor Daniel Reichman (CS), Professor Gábor Sárközy (CS)

Acknowledgements: Thank you to James Kingsley, Director of High Performance Computing and Faculty Support, for his advice and assistance in understanding the inner workings of the Turing compute cluster and how best to take advantage of it.

Submitted for Publication (CPM 2024)

Introduction

A subsequence of a string is obtained by removing zero or more characters from it while preserving order:



Finding the "**Longest Common Subsequence**" (LCS) of a set of sequences is a fundamental problem in computer science with numerous important applications.

Computational Biology
- Sequence comparison
- Phylogenetic tree reconstruction

Computational Linguistics
- File comparison
- Plagiarism detection
- Spelling correction
- Pattern matching

Other Fields
- Anomaly detection
- (Biological) data compression
- Stratigraphy

In this project, we explore the LCS problem. In doing so, we
- **Achieve world-record lower bounds on the Chvátal-Sankoff constants**
- **Prove several new LCS properties**
- **Create an interactive tool illustrating our findings**

Computing LCS

Given two strings of length m, n, how do you compute LCS?
- Naive approach running time: $\mathcal{O}(2^m \cdot n)$
- Dynamic programming (DP) approach: 2D array stores results of partial input strings. Running time: $\mathcal{O}(m \cdot n)$

The DP algorithm is taught in thousands of classes across the U.S. to illustrate the concept!

		b	a	c	b	a
	0	0	0	0	0	0
a	0	0	1	1	1	1
b	0	1	1	1	2	2
a	0	1	2	2	2	3
c	0	1	2	3	3	3
a	0	1	2	3	3	4

Dynamic programming LCS table for strings **abaca** and **bacba**

The Chvátal-Sankoff Constants

The Chvátal-Sankoff constant for two random binary strings (γ) describes the expected length of the LCS ($\mathcal{L}$) of the strings as their lengths (ℓ) tend towards infinity:

$$\gamma = \lim_{\ell \to \infty} \frac{\mathbb{E}(\mathcal{L})}{\ell}$$

Lower and upper bounds are known, **but exact values are not!**

Determining the constants is a well-known open problem—appearing in several textbooks—with extensive work to approximate and bound them (4, 5).

World-Record Lower Bounds

Chvátal and Sankoff (1975)
$0.697844 \leq \gamma \leq 0.866595$
$\gamma \approx 0.8082$

Deken (1979)
$0.7615 \leq \gamma \leq 0.8575$

Steele (1986)
$\gamma \approx 0.8284$

Baeza-Yates, Navarro, Scheihing, and Gavaldà (1999)
$\gamma \approx 0.8118$

Boutet de Monvel (1999)
$\gamma \approx 0.812282$

Dančík and Paterson (1995)
$0.77391 \leq \gamma \leq 0.83763$
$\gamma \approx 0.81225$

Bundschuch (2001)
$\gamma \approx 0.812653$

Lueker (2009)
$0.788071 \leq \gamma \leq 0.82628$

Cox and Bukh (2019)
$\gamma \approx 0.8122$

New world-record lower bound (ours):
$$0.792052 \leq \gamma$$

We additionally computed new best lower bounds for **all but one** of the general-case constants.

Our Method
- Improve on algorithms by Lueker and Kiwi and Soto (2, 3).
- Requires *immense* compute, memory use. Naive impl. stores **>4 Terabytes every iteration** for large values.
- Overcome through **parallelization**, use of LCS **symmetries**, and **recursive sub-chunking** for sequential memory I/O.

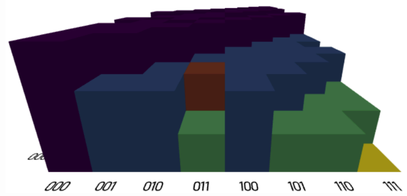
Binary Longest Inter-String Subsequence

BLISS Playground

BLISS Playground is a suite of interactive tools to aid in understanding and exploring the unique properties of the LCS problem, complete with fun puzzles!

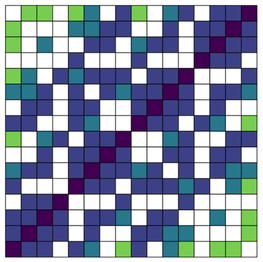
SCAN ME

LCS Edit Properties Explorer



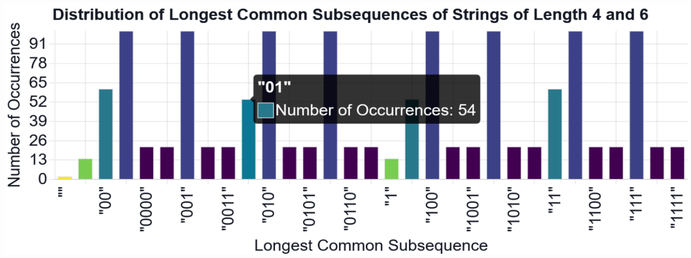
Longest Common Subsequence
x: 011
y: 0001
Update LCS

LCS Matrix Builder



Enabled Properties:
- Commutative
- Complement
- Reverse
- Slice and Prefix
- Slice and Suffix
- Slice and Circumfix

LCS Distribution Explorer



Conclusions & Future Work

Key Takeaways
- Advanced understanding of LCS through new bounds and properties
- Created BLISS for education and exploration. Check it out!

Future Work
- Improve general-case impl., further computation
- Computing upper bounds
- Advancements in AI-powered math may drive further discoveries.

References

(1) V. Chvátal and D. Sankoff. "Longest common subsequences of two random sequences". In: *J. App. Prob.* (1975).
(2) M. Kiwi and J. Soto. "On a speculated relation between Chvátal-Sankoff constants of several sequences". In: *Comb. Prob. Comp.* (2009).
(3) G. S. Lueker. "Improved bounds on the average length of longest common subsequences". In: *J. ACM* (2009).
(4) M. J. Steele. *Probability Theory and Combinatorial Optimization*. Soc. for Ind. & App. Math, 1997.
(5) M. S. Waterman. *Introduction to computational biology*. Chapman & Hall/CRC, 1995.