The Chvátal–Sankoff problem: Understanding random string comparison through stochastic processes

Alexander Tiskin

Department of Mathematics and Computer Science, St Petersburg University

LCS problem

a, b: strings of length m, n

The longest common subsequence (LCS) score:

- length of longest string that is a subsequence of both a and b
- in computational biology, unweighted alignment
- in ergodic theory, used to define the Feldman–Katok metric
- in software engineering, the diff tool

$lcs(\text{“BAABCBCA”}, \text{“CABCABA”}) = length(\text{“ABCBA”}) = 5$

LCS problem

LCS problem

LCS score for a vs b

LCS: running time

$O(mn)$ [Wagner, Fischer: 1974]

$O\left(\frac{mn}{(\log n)^c}\right)$ [Masek, Paterson: 1980] [Crochemore+: 2003]

[Paterson, Dančík: 1994] [Bille, Farach-Colton: 2008]

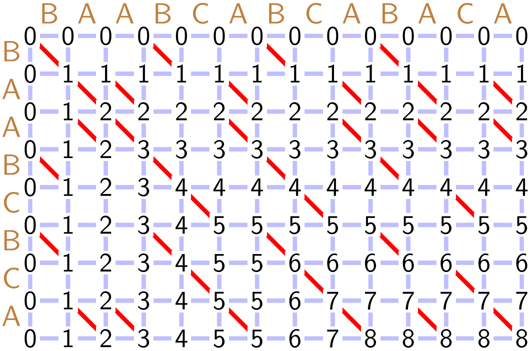
Polylog's exponent c depends on alphabet size and computation model

LCS in time $O(n^{2-\epsilon})$, $\epsilon > 0$, $m = n$: impossible unless SETH false

[Abboud+: 2015] [Backurs, Indyk: 2015]

LCS problem

LCS computation by classical dynamic programming (DP)



blue $= 0$

red $= 1$

$a =$ "BAABCBCA"

$b =$ "BAABCABCABACA"

$lcs(a, b) = 8$

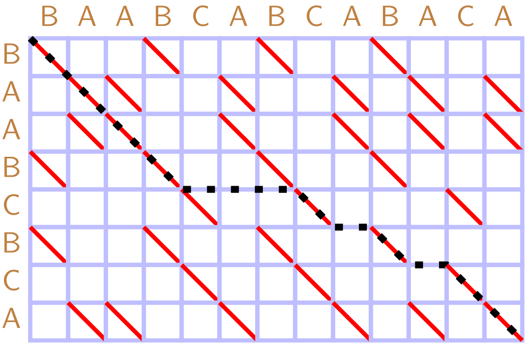
$lcs(a, \emptyset) = 0$

$lcs(\emptyset, b) = 0$

$$lcs(a\alpha, b\beta) = \begin{cases} \max(lcs(a\alpha, b), lcs(a, b\beta)) & \text{if } \alpha \neq \beta \\ lcs(a, b) + 1 & \text{if } \alpha = \beta \end{cases}$$

LCS problem

LCS as a maximum path in the LCS grid



blue $= 0$

red $= 1$

$a =$ "BAABCBCA"

$b =$ "BAABCABCABACA"

$lcs(a, b) = 8$

LCS = highest-score path top-left $\rightsquigarrow$ bottom-right

Transposition networks

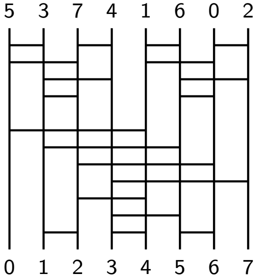
Comparison network: a circuit of comparators, each sorting a pair of values

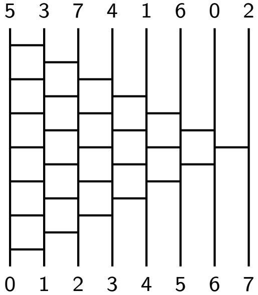
Classical model for non-branching merging, sorting, selection...

Comparison networks are visualised by wire diagrams

Transposition network: all comparisons are between adjacent wires

Sorting network: comparison/transposition network that sorts the input





Transposition networks

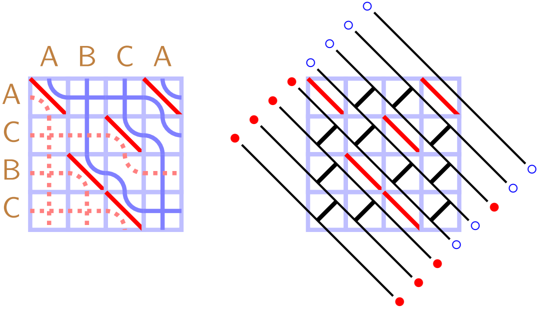
Connections to

- graph theory (expanders)
- probability (rich theory of random transposition sorting networks)
- statistical mechanics (stochastic particle interaction processes)

Applications: parallel algorithms, network design

Transposition networks

LCS: transposition network with binary anti-sorted (step) input

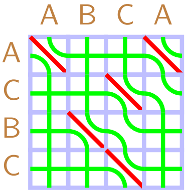


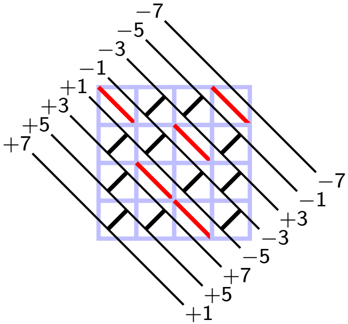
Comparators: character mismatches

Values: holes (∘) and particles (•)

Transposition networks

Semi-local LCS: transposition network with generic anti-sorted input





Comparators: character mismatches

Each value traces a strand in sticky braid (element of the Hecke monoid)

Comparing random strings

a, b: uniformly random permutation strings of length n, alphabet size n

LCS grid: n random matches, one per grid row/column

Transposition network: $n^2 - n$ random comparisons (mismatches)

Equivalent to LIS of a uniformly random permutation

$\mathbb{E}\, lcs(a, b) \sim 2n^{1/2} \qquad n \to \infty$ [Vershik, Kerov: 1977]

Comparing random strings

a, b: uniform Bernoulli sequences of length n, alphabet size $\sigma = O(1)$

LCS grid: $\approx n^2/\sigma$ random matches, one per grid row/column

Transposition network: $\approx n^2(1 - 1/\sigma)$ random comparisons (mismatches)

$\mathbb{E}\, lcs(a, b) \sim \gamma_\sigma n \qquad n \to \infty$ [Chvátal, Sankoff: 1975]

$0 \leq \gamma_\sigma n - \mathbb{E}\, lcs(a, b) \leq O\big((n \log n)^{1/2}\big)$ [Alexander: 1994]

γ_σ: Chvátal–Sankoff constants

From now on, $\sigma = 2$, $\gamma = \gamma_2$

The Chvátal–Sankoff problem: find γ; expected normalised LCS length of a pair of equally long uniformly random binary strings

More generally, find γ_σ for all $\sigma \geq 2$

Comparing random strings

Precise value of γ unknown

	$\gamma >$	$\gamma <$	
[Chvátal, Sankoff: 1975]	0.697844	0.866595	≈ 0.8082
[Deken, 1979]	0.7615	0.8575	
[Steele, 1986] (Arratia)			$\stackrel{?}{=} 2(\sqrt{2}-1) \approx 0.8284$
[Paterson, Dančík: 1994]	0.77391	0.83763	≈ 0.812
[Baeza-Yates et al.: 1999]			≈ 0.8118
[Boutet de Monvel: 1999]			≈ 0.812282
[Bundshuh: 2001]			≈ 0.812653
[Lueker: 2009]	0.788071	0.826280	(refutes Arratia)
[Bukh, Cox: 2022]			≈ 0.8122
this work	exact equations		algebraic ≈ 0.8085

Comparing random strings

Stochastic processes in discrete time:

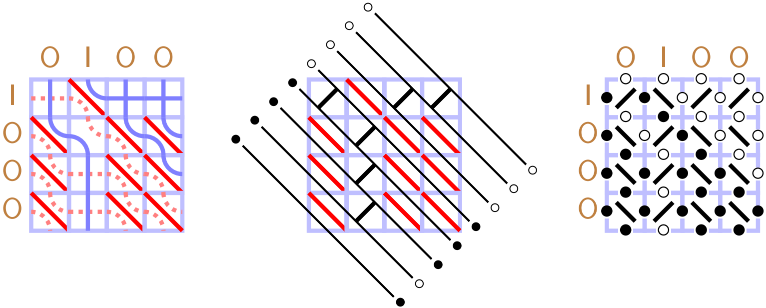
- discrete-time TASEP particle process (the "traffic jam" model)
- Young diagram corner growth model
- six-vertex model of statistical mechanics

Scaling limits well-known to exist, expressed by PDEs
[Rajewsky+: 1997; Martin, Schmidt: 2011; Borodin+: 2016]

Approaching the Chvátal–Sankoff problem:

- represent random LCS as a stochastic particle model
- local fit with an easier model by polynomial equations
- invariant distribution for both models
- global behaviour from local invariance via scaling limit PDE

Model *CS*: random LCS transposition network as a stochastic process



Evolution variants:

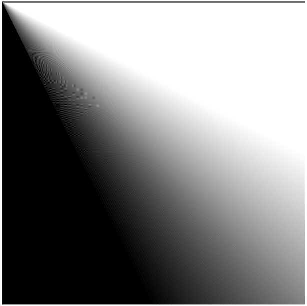
- time vertical, space horizontal, or vice versa (sequential update)
- time diagonal, space antidiagonal (sublattice-parallel update)

Comparing random strings

Scalar conservation law

$\frac{\partial}{\partial t} y + \frac{\partial}{\partial x} f(y) = 0$ fluid density $y(x, t)$ flux $f(y)$, concave

Step initial condition at $t = 0$: $y(x, 0) = \begin{cases} 1 & x < 0 \\ 0 & x > 0 \end{cases}$



Comparing random strings

Scalar conservation law (contd.)

Solution for $t > 0$:

$$y(x,t) = \begin{cases} (f')^{-1}(x/t) & f'(1)t \le x \le f'(0)t \text{ (rarefaction wave)} \\ y(x,0) & \text{otherwise (frozen area)} \end{cases}$$

Assume $0 \le y \le 1$, $f(0) = f(1) = 0$: natural for fluid density/flux

Peak flux $\tilde{f} = f(\tilde{y})$ at density $\tilde{y} = (f')^{-1}(0) = y(0,1)$

Assume f symmetric: $f(y) = f(1-y)$ $\quad \tilde{y} = \frac{1}{2}$

$\tilde{f} = f\left(\frac{1}{2}\right) =$ mass transported across origin $x = 0$ by $t = 1$

$\gamma = 1 - \tilde{f}$ in model CS

Comparing random strings

Scaling limit asymptotics for a particle-hole process

Time diagonal, space antidiagonal

Denote $\bar{z} = 1 - z$, conditional probabilities $A \mid B$ (condition in red)

Consider a small neighbourhood of $x = 0$, $t = 1$

- particle-hole symmetry: $u = \; = \;$; $\bar{u} = \; = \;$
- provides peak flux: $\tilde{f} = f$

Swap rate $p = \; = \; \mid \;$ Flux $\tilde{f} = f = \; = \; \cdot p$

To obtain $\tilde{f}$ for model CS, must study carefully dependencies between

- site values , , ,
- cell types , , as determined by characters of a, b

Comparing random strings

Model $B(1/2)$ (the Bernoulli model)

Arratia–Steele conjecture: pretend types of all cells mutually independent

[Steele: 1986; Seppäläinen: 1997; Majumdar, Nechaev: 2005; ...]

Motivation:

- cell types independent in triples (in particular, -shapes)
- ...but not in quadruples (-shape completes square uniquely)
- perhaps -shape dependence doesn't matter?

Swap rate $p =$  $=$  $= \frac{1}{2}$

Time-invariant distribution: all sites independent (more on next slide)

$\gamma^{B(1/2)} = 2(\sqrt{2} - 1) \approx 0.8284 \neq \gamma$

Conjecture disproved by upper bound [Lueker: 2009]

Model B (the generalised Bernoulli model)

Separate $p = \quad = \frac{1}{2}$ into conditional probabilities

- swap rate $p_2 = \quad$ now free to be $\neq \frac{1}{2}$
- pseudo-rates $p_0 = \quad \simeq p_3 = \quad , \; p_1 = \quad$

Swap rate balanced out by pseudo-rates to preserve $\quad = \frac{1}{2}$

Time-invariant distribution: alternating Bernoulli (AB) sequence

- doubly-infinite; space-invariant under shift $i \mapsto i + 1$ and reversal $i \mapsto -i - 1$ with simultaneous exchange of $\circ$ and $\bullet$
- all sites mutually independent

AB sequence parameter $u = \quad$ determined by swap rate p_2

Comparing random strings

Model B (contd.)

Fit (pseudo-)rates to model CS locally in a neighbourhood of $x = 0$, $t = 1$ via equations in (pseudo-)rates and the parameter of AB sequence

- time-invariance equations: 1 time step; link u with p_2 for model B
- string matching equations: 3 time steps; link models B, CS
- total probability: $\bar{u}\bar{u}p_2 + 2u\bar{u}p_0 + uup_1 = \; + 2 \; + \; \equiv \; = \frac{1}{2}$

Solve by Mathematica's `Solve`, option `Quartics -> True`

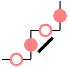
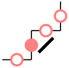
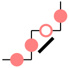
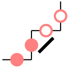
$$u = \sqrt{\tfrac{7}{3}} - \sqrt{\tfrac{23-5\sqrt{21}}{6}} - 1 \approx 0.407025$$

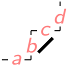
$$p_2 = -\tfrac{2}{3} + \tfrac{34}{3}u - 19u^2 - 4u^3 \approx 0.528838$$

$$\gamma^B = 1 - f^B = 1 - \bar{u}up_2 \approx 0.814050 \neq \gamma$$

Fit not perfect: AB property not expressed fully by equations

Model M (the Markov model)

Swap partial rates $p_5 =$ , $p_4 =$  $\simeq p_{13} =$ , $p_{12} =$ 

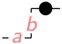
Pseudo-rates $p_{abcd} =$ 

$abcd \in \{0, \ldots, 15\} \setminus \{5, 4, 13, 12\}$

Time-invariant distribution: alternating second-order Markov (AM2) sequence

- doubly-infinite; space-invariant under shift $i \mapsto i + 1$ and reversal $i \mapsto -i - 1$ with simultaneous exchange of $\circ$ and $\bullet$
- conditioned on adjacent site pair (ξ_i, ξ_{i+1}), infinite prefix $(\ldots, \xi_{i-2}, \xi_{i-1})$ independent of infinite suffix $(\xi_{i+2}, \xi_{i+3}, \ldots)$

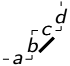
AM2 sequence parameters $u =$ , $v_a =$ 

, $w_{ab} =$ 

determined by swap partial rates p_5, $p_4 = p_{13}$, p_{12}

Model M (contd.)

Fit (pseudo-)rates to model CS locally in a neighbourhood of $x = 0$, $t = 1$ via equations in (pseudo-)rates and the parameters of AM2 sequence

- time-invariance equations: 1 time step; link u, v_a, w_{ab} with p_5, $p_4 = p_{13}$, p_{12} for model M
- string matching equations: 3 time steps; link models M, CS
- total probability: $\sum_{a,b,c,d \in \{\circ,\bullet\}}$  $\equiv \diagup = \frac{1}{2}$

Perfect fit: AM2 property expressed by polynomial equations, $\gamma = \gamma^M$

Equation coefficients 1 and 2; hence, γ is algebraic

Closed-form expression unlikely due to complexity of equations

Comparing random strings

Experiment options

"Naive" (very slow convergence)

- generate long random strings; compute LCS; repeat

Simulating model *CS* (done; slow convergence)

- initialise with AB sequence for $t = 0$
- run model *CS* to stationary state (max 20 time steps)
- bit-parallel LCS [Crochemore+: 2003] and various optimisations

Solving iteratively for model M parameters (assumes model's correctness)

Current estimate $\gamma \approx 0.8085$

Needs extra confirmation/reconciling with previous work

Conclusions

The Chvátal–Sankoff problem: expected normalised LCS length γ of a pair of equally long uniformly random binary strings

Expressed as hydrodynamic limit of stochastic particle process (model CS)

Linked with another stochastic process (model M): local fitting in a small neighbourhood of main diagonal

Flux for model M expressed by a (large) system of algebraic equations

- implies that γ is algebraic
- closed-form solution unlikely due to equations' complexity

Essentially resolves the Chvátal–Sankoff problem (with a somewhat negative flavour)

Numerical solution: several options, work in progress

Conclusions

Further work:

- distribution properties beyond expectation γ (e.g. Tracey–Widom?)
- strings of unequal length, limit shape (similar but more cumbersome)
- skewed character distribution (challenging, no -independence)
- Levenshtein distance (special case of ternary strings)
- ternary or larger alphabet (challenging, no -uniqueness)
- more than two strings (looks hopeless)