An A*-algorithm for the Unordered Tree Edit Distance with Custom Costs

Benjamin Paassen
https://gitlab.com/bpaassen/uted

HUMBOLDT-UNIVERSITÄT ZU BERLIN

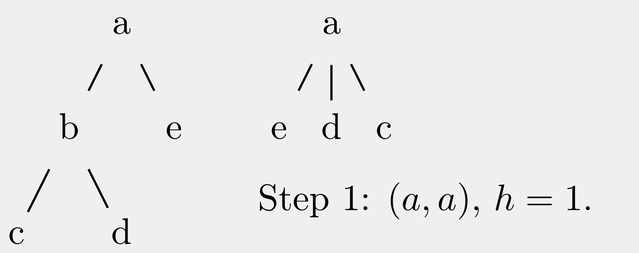
Motivation

The *unordered tree edit distance* is an intuitive metric for unordered trees (e.g. chemical molecules). While it is MAX-SNP-hard [4], it can be computed efficiently via an A* algorithm for small examples [3].

Our contribution: A* algorithm that is compatible with *custom costs*.

A* algorithm

1. Start by mapping root to root. Compute lower bound h and put on priority queue Q.
2. Poll best lower bound from Q. Consider all possible extensions, compute lower bounds, and put on Q.
3. If mapping is complete, stop.



Step 1: (a, a), $h = 1$.

Step 2:
$(a, a), (b, -)$, $h = 1 + 0$.
$(a, a), (b, e)$, $h = 1 + 4$.
$(a, a), (b, d)$, $h = 1 + 3$.
$(a, a), (b, c)$, $h = 1 + 3$.

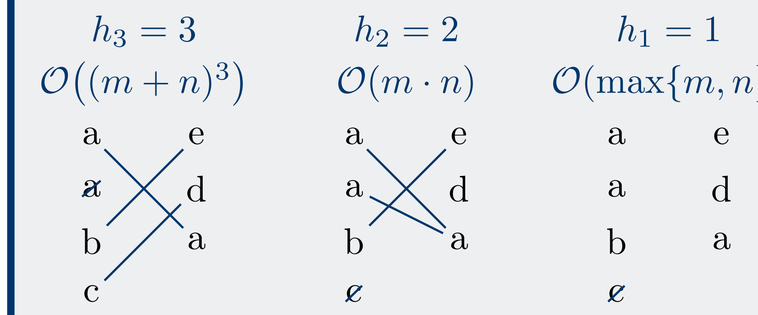
Three novel lower bounds

Main idea: Treat remaining trees as unordered sets.

h_3: optimal assignment via Hungarian algorithm.

h_2: best matching partner for each node; deletion of least matching.

h_1: only deletions.



It holds: $h_1 \leq h_2 \leq h_3 \leq$ actual edit distance.

Experimental Setup

Datasets:

Alkanes: 150 alkane molecules with boiling points by [1]. Custom cost: hydrogen count.

ZINC: 100 smallest molecules with chemical stabilities by [2]. Custom cost: electron count.

Baselines: Linear-time lower bound by Yoshino [3] for runtime; constrained UTED and ordered TED for regression.

Experimental Results

	unit costs				custom costs		
data set	h_1	h_2	h_3	yoshino	h_1	h_2	h_3
	runtime [ms]						
alkanes	8.70	12.15	10.72	9.52	**7.34**	8.21	9.92
ZINC	549.38	277.15	192.97	266.66	130.62	75.53	**68.12**
	no. of searched options						
alkanes	376	348	260	279	318	302	**246**
ZINC	24586	9164	4158	6781	6643	2655	**1379**

5-nearest neighbor regression RMSE in 15-fold crossvalidation:

	unit costs			custom costs		
data set	UTED	CUTED	TED	UTED	CUTED	TED
alkanes	0.27	0.27	0.27	**0.25**	**0.25**	**0.25**
ZINC	1.33	1.31	1.36	**1.24**	1.26	1.29

References:

[1] Claudio Gallicchio and Alessio Micheli. Tree echo state networks. *Neurocomputing*, 101:319 – 337, 2013.

[2] Matt J. Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *Proc. ICML*, 2017.

[3] Takuya Yoshino, Shoichi Higuchi, and Kouichi Hirata. A dynamic programming A* algorithm for computing unordered tree edit distance. In *Proc. IIAI*, pages 135–140, 2013.

[4] Kaizhong Zhang and Tao Jiang. Some MAX SNP-hard results concerning unordered labeled trees. *Information Processing Letters*, 49(5):249–254, 1994.

Acknowledgements: Funding by the German Research Foundation (DFG) under grant number PA 3460/2-1 is gratefully acknowledged.