University of Ljubljana
Faculty of Computer and Information Science

Hajduković Gojko

Optimal leaf ordering of phylogenetic trees

master's thesis

the 2nd cycle master's study programme
computer and information science

Supervisor: Assistant professor Tomaž Curk
Co-supervisor: Post doctoral fellow Igor Ruiz de los Mozos

Ljubljana, 2021

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Hajduković Gojko

Optimalno razvrščanje listov filogenetskih dreves

magistrsko delo

magistrski študijski program druge stopnje
računalništvo in informatika

Mentor: Assistant professor Tomaž Curk
Somentor: Post doctoral fellow Igor Ruiz de los Mozos

Ljubljana, 2021

[1]V dogovorju z mentorjem lahko kandidat magistrsko delo s pripadajočo izvorno kodo izda tudi pod drugo licenco, ki ponuja določen del pravic vsem: npr. Creative Commons, GNU GPL. V tem primeru na to mesto vstavite opis licence, na primer tekst [5].

Acknowledgments

Worth mentioning in the acknowledgment is everyone who contributed to your thesis.

Hajduković Gojko, 2021

To all the flowers of this world.

"The only reason for time is so that everything doesn't happen at once."

— Albert Einstein

Contents

List of used acronmys

acronym	meaning
CA	classification accuracy
DBMS	database management system
SVM	support vector machine
...	...

Abstract

Title: Optimal leaf ordering of phylogenetic trees

This sample document presents an approach to typesetting your BSc thesis using LaTeX. A proper abstract should contain around 100 words which makes this one way too short. A good abstract contains: (1) a short description of the tackled problem, (2) a short description of your approach to solving the problem, and (3) (the most successful) result or contribution in your thesis.

Keywords

computer, computer, computer

Povzetek

Naslov: Optimalno razvrščanje listov filogenetskih dreves

V vzorcu je predstavljen postopek priprave magistrskega dela z uporabo okolja LaTeX. Vaš povzetek mora sicer vsebovati približno 100 besed, ta tukaj je odločno prekratek. Dober povzetek vključuje: (1) kratek opis obravnavanega problema, (2) kratek opis vašega pristopa za reševanje tega problema in (3) (najbolj uspešen) rezultat ali prispevek magistrske naloge.

Ključne besede

računalnik, računalnik, računalnik

Razširjeni povzetek

To je primer razširjenega povzetka v slovenščini, ki je obvezen za naloge pisane v angleščini. Razširjeni povzetek mora vsebovati vse glavne elemente dela napisanega v angleščini skupaj s kratkim uvodom in povzetkom glavnih elementov metode, glavnih eksperimentalnih rezultatov in glavnih ugotovitev. Razširjeni povzetek naj bo strukturiran v podpoglavja (spodaj je naveden le okvirni primer in je nezavezujoč). Čez palec navadno razširjeni povzetek nanese okoli 10 odstotkov obsega celotnega dela.

I Kratek pregled sorodnih del

II Predlagana metoda

III Eksperimentalna evaluacija

IV Sklep

poljuben tekst poljuben tekst

poljuben tekst poljuben tekst

Chapter 1

Introduction

Phylogeny is the study of evolutionary relationships within a set of species or any taxa in general. These relationships are depicted in a phylogenetic trees. They can either be rooted, and have an explicit ancestral node that is known, or they can be unrooted and have no such node, inferring just relationships among species [**?**]. The leaves of these trees represent species or other taxa while internal nodes refer to (hypothetical) ancestors. Branches connecting the leaves represent evolutionary distances among taxa. A following core problems emerge from the field of phylogenetics:
Given the information about the extant taxa, how does one infer phylogenetic tree and does the constructed tree represent the true evolutionary relationships for those taxa? Given the constructed tree, how does one infer the the correct leaf ordering and how does one visually represent such a tree? This thesis concerns with studying the problem of inferring correct leaf ordering and visually representing the phylogenetic tree with respect to such an odering.

1.1 Motivation

The dendrogram is a graphical representation which can be used to depict any clustering tree. It is typically used to represent the binary tree structure

produced as an output of Hierarchical Clustering. The dendrogram serves as an useful tool in exploratory data analysis, allowing clusters of input elements to be easily inferred from the subtree structures below a certain treshold.

However, the main shortcoming of graphically representing phylogenetic trees with dendrogram lays in its directionality. Due to its directionality phylogenetic tree is represented as it is rooted tree, while in fact most of the approaches for inferring phylogenetic trees produce an unrooted tree. Such an inability leads to the phylogenetic tree being visually observed as an imbalanced tree. Other shortcomings of dendrogram arise for the trees with high number of leaves, such that it becomes visually impractical to find required leaf label since they can often be hidden in further distanced branches.

The common misinterpretation of a dendrogram lays behind the leaves being initially randomly ordered. It rises the assumption of two leaves being similar based on the proximity in such an order. However, similarity between the leaves cannot be inferred based on its proximity in the leaf ordering, since the leaves could belong to different subtree structures or being quite distant from each other. Namely, a tree with $\mathtt{n}$ leaves has 2^{n-1} different linear leaf orderings as the orientation of sub-trees can be flipped at each merge maintaining the original tree structure.

Number of methods have been proposed in order to solve the linear leaf ordering problem. Although, they vary in the criteria that is to be optimized, they show the need and the impact linear leaf ordering has on visual analysis. Bar-Joseph et al. showed the significance that linear leaf ordering has on biological analysis, in the context where genes or experiments are often hypothesized to be related when having high proximity in linear leaf ordering. Furthermore, linear leaf ordering has a particular significance on analysis of gene-expression data. Recent studies show the applicability and significance of phylogenetic analysis on gene-expression data. We aim at generalizing the leaf ordering problem onto phylogenetic trees.

Furthermore, an important feature of phylogenetic trees is its internal node degree. In the contrary to the trees produced through clustering methods, internal nodes of phylogenetic trees can be polytomies joining more than three branches. Unresolved trees are often misinterpreted with meaning often only ascribed to the vertical proximity of leaves, not taking into account its order. Cerruti et al show that linear leaf ordering gives a biological significance to unresolved phylogenetic trees.

Due to aforementioned shortcomings of graphically representing phylogenetic trees with dendrogram, number of methods have been developed in order to depict phylogenetic trees. Bachmaier et al. introduced two novel methods for representing phylogenetic trees, radial and circular layout. Although, these methods represent phylogenetic trees in 2D space, thus tackling the imbalance problem, they truthfully depict given edge lengths and label names. Shortcoming of these approaches lays in theirs inability to truthfully present ratios among leaf nodes with respect to linear leaf ordering. We aim to re-position tree nodes in order to improve the visual representation of phylogenetic trees with respect to linear leaf ordering.

1.2 Goals and Thesis structure

The purpose of this study is to generalize linear leaf ordering problem to 2D space, suitable for visually representing phylogenetic trees with respect to such leaf order. Our contribution aims at helping in visual analysis of phylogenetic trees, thus providing a correct assumption that similarity among leaves grows with respect to their proximity in leaf ordering.

Chapter 2

Related work

Throughout the years, a number of approaches have been proposed in order to solve the linear leaf ordering problem and to truthfully depict phylogenetic trees. In this chapter we first give an overview of the linear leaf ordering problem and the proposed methods. Moreover, we focus on in-depth presentation of the methods that generalize leaf ordering problem to k-ary trees. While, these approaches conceptually differ, they may yield the similar results in the terms of ordering quality. Furthermore, we observe the results of these methods and discuss its applicability to ordering leaves of phylogenetic trees and we summarize its simmilarities and differencies. Whilst, in the second part we present the methods that aim to depict phylogenetic trees in 2D space and tackle afore mentioned shortcomings of dendrograms.

2.1 Leaf ordering methods

Due to its biological significance, linear leaf ordering problem represents a well-researched topic in the field of evolutionary computation. Main methods to solve linear leaf ordering problem are grouped based on the critetion to be optimized and the optimization algorithm. Caraux et al split the leaf ordering approaches into three categories:

- **Leaf reorganization** - optimal leaf ordering is inferred by reorganizing

leaves and flipping sub-trees at each merge in order to optimize specific criterion.

- **Unidimensional scaling and seriation** - aim at placing a set of data objects along a dataset row such that the distances between points best reflect the dissimilarities between objects.
- **Identification and reorganization of classes** - aims at aggregating a set of leaves into a class and to further reorganize inferred classes.

State-of-the-art approaches are based on leaf reorganization of trees. Moreover, regardless the fact that these approaches differ in optimization criterion and the optimization algorithm they provide the same formal definition of leaf ordering problem. Provided the tree T with n leaves denoted with z1,.....,zn and by v1,.....,vn the n-1 internal nodes of T, a linear leaf ordering that preserves structure of T is defined as an ordering generated by flipping internal tree nodes. A tree with `n` leaves, has `n-1` internal nodes, thus having 2^{n-1} different leaf orderings. Leaf reorganization approaches are further divided on those yielding the optimal ordering and those yielding the sub-optimal ordering.

Bar-Joseph et al [] proposed an optimal approach that aim to solve linear leaf ordering for binary trees. Later on, the algorithm is generalized to obtain an optimal leaf ordering for k-ary trees. The proposed algorithm works in a recursive manner that resembles dynamic programming. Moreover, the optimization criterion is defined as

2.2 Visualization techniques

Matematična ali popolna indukcija je eno prvih orodij, ki jih spoznamo za dokazovanje trditev pri matematičnih predmetih.

Izrek 2.1 *Za vsako naravno število n velja*

$$n < 2^n. \tag{2.1}$$

Dokaz. Dokazovanje z indukcijo zahteva, da neenakost (2.1) najprej preverimo za najmanjše naravno število — 0. Res, ker je $0 < 1 = 2^0$, je neenačba (2.1) za $n = 0$ izpolnjena.

Sledi indukcijski korak. S predpostavko, da je neenakost (2.1) veljavna pri nekem naravnem številu n, je potrebno pokazati, da je ista neenakost v veljavi tudi pri njegovem nasledniku — naravnem številu $n + 1$. Izračun zapišemo s tremi vrsticami, ki jih končamo s piko, saj do del tega stavka:

$$\begin{aligned} n + 1 &< 2^n + 1, && (2.2)\\ &\leq 2^n + 2^n, && (2.3)\\ &= 2^{n+1}. \end{aligned}$$

Neenakost (2.2) je posledica indukcijske predpostavke, neenakost (2.3) pa enostavno dejstvo, da je za vsako naravno število n izraz 2^n vsaj tako velik kot 1. S tem je dokaz Izreka 2.1 zaključen. □

Opazimo, da je LaTeX številko izreka podredil številki poglavja.

Chapter 3

Methodology

3.1 Data

3.2 Evaluation

3.3 Algorithms

Slike in daljše tabele praviloma vključujemo v dokument kot plovke. Pozicija plovke v končnem izdelku ni pogojena s tekom besedila, temveč z izgledom strani. LaTeX bo skušal plovko postaviti samostojno, praviloma na vrh strani, na kateri se na takšno plovko prvič sklicujemo. Pri tem pa bo na vsako stran končnega izdelka želel postaviti tudi sorazmerno velik del besedila. V skrajnem primeru, če imamo res preveč plovk, se bo odločil za stran popolnoma zapolnjeno s plovkami.

Bitne slike, vektorske slike, kakršnekoli slike, z LaTeXom lahko vključimo vse. Slika 3.1 je v `.pdf` formatu. Pa res lahko vključimo slike katerihkoli formatov? Žal ne. Programski paket LaTeX lahko uporabljamo v več dialektih. Ukaz `latex` ne mara vključenih slik v formatu Portable Document Format `.pdf`, ukaz `pdflatex` pa ne prebavi slik v Encapsulated Postscript Formatu `.eps`. Strnjeno v Tabeli 3.1.

Nasvet? Odločite se za uporabo ukaza `pdflatex`. Vaš izdelek bo brez

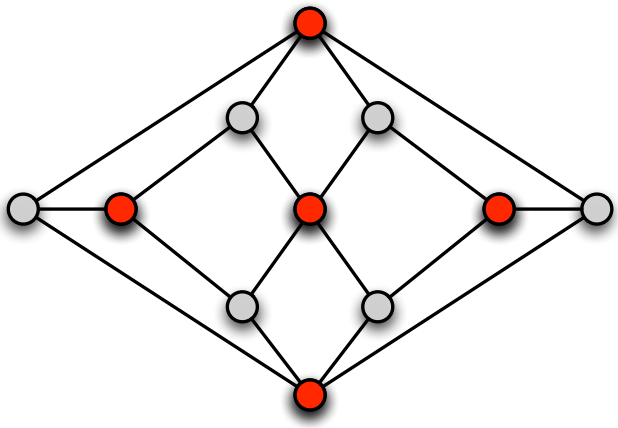

Figure 3.1: Herschelov graf, vektorska grafika.

Table 3.1

ukaz/format	`.pdf`	`.eps`	ostali formati
`pdflatex`	da	ne	da
`latex`	ne	da	da

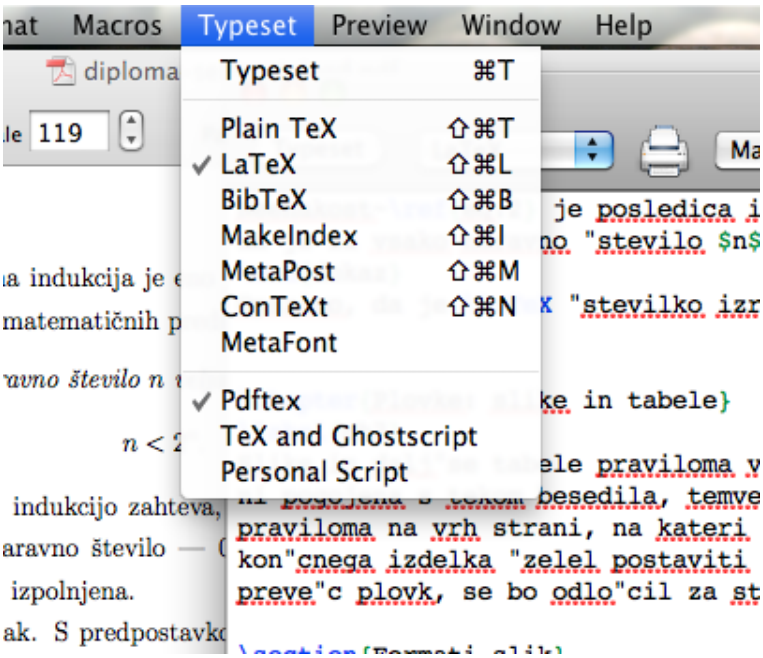

Figure 3.2: Kateri dialekt uporabljati?

vmesnih stopenj na voljo v .pdf formatu in ga lahko odnesete v vsako tiskarno. Če morate na vsak način vključiti sliko, ki jo imate v `.eps` formatu, jo vnaprej pretvorite v alternativni format, denimo `.pdf`.

Včasih se da v okolju za uporabo programskega paketa LaTeX nastaviti na kakšen način bomo prebavljali vhodne dokumente. Spustni meni na Sliki 3.2 odkriva uporabo LaTeXa v njegovi pdf inkarnaciji — `pdflatex`. Vključena Slika 3.2 je seveda bitna.

Chapter 4

Results

4.1 Notacije

Za notacijo spremenljivk ter skalarjev uporabimo običajno notacijo, t.j., spremenljivka x in skalar a. Pri notaciji matrik ter vektorjev pa se poslužujemo krepega fonta. Torej, matrika $\boldsymbol{A}$ ter vektor $\boldsymbol{v}$,

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1q} \\ a_{21} & a_{22} & \dots & a_{2q} \\ \vdots & & & \\ a_{p1} & a_{p2} & \dots & a_{pq} \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{bmatrix}.$$

4.2 Lepe tabele in psevdokoda

Psevdokoda 1 prikazuje primer delovanja genetskega algoritma, medtem ko Tabela 4.1 prikazuje primer lepe tabele brez vertikalnih črt.

Table 4.1: Primer enostavne tabele.

Ime	Vrednost	Opis
a	0.03	skalar
x	-1	spremenljivka

Algorithm 1 Psevdokoda genetskega algoritma

1: $t \leftarrow 0$
2: $InitPopulation[P(t)] \leftarrow$ inicializiraj populacijo
3: $EvalPopulation[P(t)] \leftarrow$ evaluiraj populacijo
4: **repeat**
5: $\quad P'(t) \leftarrow Variation[P(t)] \leftarrow$ generiraj novo populacijo
6: $\quad EvalPopulation[P'(t)] \leftarrow$ evaluiraj novo populacijo
7: $\quad P(t+1) \leftarrow ApplyGeneticOperators[P'(t) \in Q]$
8: $\quad t \leftarrow t+1$
9: **until** prekinitev
10: **if** rezultat dovolj dober **then**
11: $\quad$ shrani rezultat
12: **end if**

Chapter 5

Experimental validation

Kot smo omenili že v uvodu, je pravi način za citiranje literature uporaba BibTeXa [4]. Programski paket LaTeXje prvotno predstavljen v priročniku [3] in je v resnici nadgradnja sistema TeX avtorja Donalda Knutha, znanega po denimo, če izpustim njegovo umetnost programiranja, Knuth-Bendixovem algoritmu [2].

Vsem raziskovalcem s področja računalništva pa svetujem v branje mnenje L. Fortnowa [1].

Chapter 6

Conclusion

Izbira LaTeX ali ne LaTeX je seveda prepuščena vam samim. Res je, da so prvi koraki v LaTeXu težavni. Ta dokument naj vam služi kot začetna opora pri hoji.

Appendix A

Title of the appendix 1

Example of the appendix.

Bibliography

[1] L. Fortnow, "Viewpoint: Time for computer science to grow up", *Communications of the ACM*, št. 52, zv. 8, str. 33–35, 2009.

[2] D. E. Knuth, P. Bendix. "Simple word problems in universal algebras", v zborniku: Computational Problems in Abstract Algebra (ur. J. Leech), 1970, str. 263–297.

[3] L. Lamport. *LaTEX: A Document Preparation System*. Addison-Wesley, 1986.

[4] O. Patashnik (1998) BibTEXing. Dostopno na: `http://ftp.univie.ac.at/packages/tex/biblio/bibtex/contrib/doc/btxdoc.pdf`

[5] licence-cc.pdf. Dostopno na: `https://ucilnica.fri.uni-lj.si/course/view.php?id=274`