

Algorithms and Complexity in Phylogenetics

Edited by

Magnus Bordewich¹, Britta Dorn², Simone Linz³, and
Rolf Niedermeier⁴

1 Durham University, GB, m.j.r.bordewich@durham.ac.uk

2 Universität Tübingen, DE, britta.dorn@uni-tuebingen.de

3 University of Auckland, NZ, s.linz@auckland.ac.nz

4 TU Berlin, DE, rolf.niedermeier@tu-berlin.de

Abstract

Phylogenetics is the study of ancestral relationships between species. Its central goal is the reconstruction and analysis of phylogenetic trees and networks. Even though research in phylogenetics is motivated by biological questions and applications, it heavily relies on mathematics and computer science. Dagstuhl Seminar 19443 on *Algorithms and Complexity in Phylogenetics* aimed at bringing together researchers from phylogenetics and theoretical computer science to enable an exchange of expertise, facilitate interactions across both research areas, and establish new collaborations. This report documents the program and outcomes of the seminar. It contains an executive summary, abstracts of talks, short summaries of working groups, and a list of open problems that were posed during the seminar.

Seminar October 27–31, 2019 – <http://www.dagstuhl.de/19443>

2012 ACM Subject Classification Theory of computation → Parameterized complexity and exact algorithms, Mathematics of computing → Graph algorithms, Applied computing → Molecular evolution

Keywords and phrases Approximation algorithms, Evolution, Parameterized algorithms, Phylogenetic trees and networks

Digital Object Identifier 10.4230/DagRep.9.10.134

Edited in cooperation with Kristina Wicke, Universität Greifswald, DE

1 Executive Summary

Magnus Bordewich (Durham University, GB)

Britta Dorn (Universität Tübingen, DE)

Simone Linz (University of Auckland, NZ)

Rolf Niedermeier (TU Berlin, DE)

License  Creative Commons BY 3.0 Unported license
© Magnus Bordewich, Britta Dorn, Simone Linz, and Rolf Niedermeier

Disentangling the evolutionary relationships between species dates back at least to Charles Darwin and his voyage on board the Beagle. Ever since, the research area of phylogenetics focusses on the reconstruction and analysis of rooted leaf-labeled trees, called phylogenetic (evolutionary) trees, to unravel ancestral relationships between entities like species, languages, and viruses. However, processes such as horizontal gene transfer and hybridization challenge the model of a phylogenetic tree since they result in mosaic patterns of relationships that cannot be represented by a single tree. Indeed, it is now widely acknowledged that rooted leaf-labeled digraphs with underlying cycles, called phylogenetic networks, are better suited to represent evolutionary histories.



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Algorithms and Complexity in Phylogenetics, *Dagstuhl Reports*, Vol. 9, Issue 10, pp. 134–151

Editors: Magnus Bordewich, Britta Dorn, Simone Linz, and Rolf Niedermeier



Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Biological questions and applications motivate much of the research in phylogenetics. Nevertheless, most of the software that is routinely used by evolutionary biologists has its roots in theoretical research areas which include algorithms, computational complexity, graph theory, algebra, and probability theory. With a shift from phylogenetic trees towards more complex graphs, the development of new algorithms for phylogenetic networks is currently an active area of research that requires deep insight from computer science and mathematics.

The objective of the seminar was to facilitate interactions between the two research communities of (i) computational and mathematical phylogenetics and (ii) theoretical computer science with a focus on algorithms and complexity. Specifically, its goal was to advance the development of novel algorithms (with provable performance guarantee) to reconstruct and analyze phylogenetic networks that are grounded in techniques from theoretical computer science such as parameterized and approximation algorithms.

This four-day seminar brought together 27 researchers from ten countries, whose research spans theoretical computer science and algorithms, (discrete) mathematics, and computational and mathematical phylogenetics. The seminar program included six overview talks, nine research talks (one of which via Skype), a rump session for short five-minute contributions, and slots for discussions and group work on open problems. More specifically, the overview talks provided introductions to techniques and current trends in parameterized algorithms, combinatorial decompositions, and enumeration algorithms on one hand, and introductions to spaces of phylogenetic trees and networks, and the reconstruction of networks from smaller networks and trees on the other hand. Additionally, each overview talk included open questions and challenges that provided a foundation for discussions and group work throughout the week. The research talks, of which three were given by postgraduate students, covered topical streams of research, including phylogenetic split theory, the placement of phylogenetic problems in higher classes of the polynomial hierarchy, new insight into the popular so-called TREE CONTAINMENT problem, and phylogenetic diversity and biodiversity indices. Moreover, five working groups were formed on the second day of the seminar. While the research projects that were initiated in these groups are ongoing, some groups obtained first results during the seminar that were presented on the last day.

By building on initially existing synergies between the two research communities, the seminar has taken a leap towards developing new and fostering existing collaborations between both communities. Collaborative work was encouraged and put into practice over formal and informal discussions as well as three group work sessions. Since a significant number of open problems in phylogenetics require the combined expertise of experts in phylogenetics and theoretical computer science, we expect the collaborations formed at Schloss Dagstuhl to make progress on problems across the traditional discipline boundaries and, ideally, lead to joint peer-reviewed journal or conference publications.

To conclude, this seminar has acknowledged that exchange and connection between the two research communities of theoretical computer science and phylogenetics is fruitful for both sides. Techniques and methods from algorithms and complexity as well as theoretical considerations in general enable, account for, and foster new insights in problems from phylogenetics. Conversely, the specific features and problem structures appearing in the context of phylogenetic trees and networks provide novel theoretical challenges and new directions for foundational research in algorithms and computational complexity.

We thank all participants for their contributions and for openly sharing their ideas and research questions that led to a positive working atmosphere and many discussions throughout the seminar. Furthermore, we sincerely thank the team of Schloss Dagstuhl for their excellent support and communication as well as for providing an enjoyable seminar environment.

2 Table of Contents

Executive Summary

Magnus Bordewich, Britta Dorn, Simone Linz, and Rolf Niedermeier 134

Overview of Talks

Reconstructing equidistant phylogenetic networks from distance matrices <i>Allan Bai</i>	138
Deciding whether two phylogenetic networks embed the same trees is hard <i>Janosch Döcker</i>	138
Orthology relations <i>Katharina T. Huber</i>	139
New kernels for TBR distance (or, equivalently, the maximum agreement forest problem): theory and practice <i>Steven Kelk</i>	139
An introduction to fixed-parameter algorithms: basic techniques and recent ideas <i>Christian Komusiewicz and André Nichterlein</i>	140
Some open problems in phylogenetic split theory <i>Vincent Moulton</i>	140
Continuous spaces of phylogenetic trees and networks <i>Megan Owen</i>	141
Algorithmic tree and network problems in phylogenetics <i>Charles Semple</i>	141
Searching for optimal phylogenetic histories <i>Katherine St. John</i>	142
Applying SNAQ with 5-net <i>Nihan Tokaç</i>	142
Combinatorial decompositions and enumeration algorithms <i>Alexandru Tomescu</i>	143
Constructing phylogenetic networks from smaller networks <i>Leo van Iersel</i>	143
Tree Containment with polytomies <i>Mathias Weller</i>	144
Phylogenetic diversity and biodiversity indices on phylogenetic networks <i>Kristina Wicke</i>	144
Global-Local Clustering <i>Norbert Zeh</i>	145

Working Groups


Longest Common Subsequence for similar strings <i>Laurent Bulteau, Mark E. L. Jones, Rolf Niedermeier, and Till Tantau</i>	146
---	-----

Hybridization for many trees with non-identical leaf sets <i>Britta Dorn, Christian Komusiewicz, Catherine McCartin, André Nichterlein, Mathias Weller, and Norbert Zeh</i>	147
Hybridization Number for multiple multifurcating trees <i>Mark E. L. Jones and Vincent Moulton</i>	148
Maximum agreement subtrees <i>Katherine St. John, Magnus Bordewich, Simone Linz, Megan Owen, Charles Semple, and Kristina Wicke</i>	149
Constructing phylogenetic networks from trinets <i>Leo van Iersel, Vincent Moulton, Leen Stougie, and Nihan Tokac</i>	149
Open Problems	
List of open problems <i>Leo van Iersel</i>	150
Participants	151

3 Overview of Talks

3.1 Reconstructing equidistant phylogenetic networks from distance matrices

Allan Bai (University of Canterbury – Christchurch, NZ)

License  Creative Commons BY 3.0 Unported license
© Allan Bai

Main reference Magnus Bordewich, Katharina T. Huber, Vincent Moulton, Charles Semple: “Recovering normal networks from shortest inter-taxa distance information”, *Journal of Mathematical Biology*, Vol. 77(3), pp. 571–594, 2018.

URL <http://dx.doi.org/10.1007/s00285-018-1218-x>

A phylogenetic network is a rooted acyclic directed graph, where the set of all vertices with out-degree zero are called leaves. A phylogenetic network is equidistant if all paths from the root to the leaves are equal. It has been shown in the past that certain classes of phylogenetic networks are determined by the distances between the leaves, up to a certain equivalence. In this talk, I will give an overview of the known algorithms for reconstructing normal and tree-child networks. I will also present my research on reconstructing shortcut free networks using minimum distance matrices.

3.2 Deciding whether two phylogenetic networks embed the same trees is hard

Janosch Döcker (Universität Tübingen, DE)

License  Creative Commons BY 3.0 Unported license
© Janosch Döcker

Joint work of Janosch Döcker, Simone Linz, Charles Semple

Main reference Janosch Döcker, Simone Linz, Charles Semple: “Displaying trees across two phylogenetic networks”, *Theor. Comput. Sci.*, Vol. 796, pp. 129–146, 2019.

URL <https://doi.org/10.1016/j.tcs.2019.09.003>

Phylogenetic networks are frequently used to represent the ancestral relationships between a collection of extant species. Noting that each phylogenetic network N embeds a collection of phylogenetic trees, we refer to this collection as the display set of N . A well-studied and biologically relevant problem asks, given a phylogenetic network N and a phylogenetic tree T , whether T is contained in the display set of N . We study the computational complexity of several questions related to the display sets of two given phylogenetic networks. In particular, we show that deciding whether two phylogenetic networks have the same display set is computationally hard and, more specifically, that it can be placed on the second level of the polynomial-time hierarchy.

3.3 Orthology relations

Katharina T. Huber (*University of East Anglia – Norwich, GB*)

License © Creative Commons BY 3.0 Unported license
© Katharina T. Huber

Joint work of Katharina T. Huber, Guillaume Scholz

Main reference Katharina T. Huber, Guillaume E. Scholz: “Beyond Representing Orthology Relations by Trees”, *Algorithmica*, Vol. 80(1), pp. 73–103, 2018.

URL <https://doi.org/10.1007/s00453-016-0241-9>

Reconstructing the evolutionary past of a family of genes is an important aspect of many genomic studies. To help with this, simple relations on a set of sequences called orthology relations may be employed [1]. In addition to being interesting from a practical point of view, they are also attractive from a theoretical perspective in that e.g. a characterization is known for when such a relation is representable by a certain type of phylogenetic tree [2]. Perhaps not surprisingly, real biological data however hardly ever satisfies that characterization. Starting with a brief introduction into the area, we review some recent results concerning such relations [3].

References

- 1 M. Hellmuth, N. Wieseke, M. Lechner, H-P. Lenhof, M. Middendorf, and P.F. Stadler. *Phylogenomics with paralogs* PNAS, 112:2058-2063, 2015.
- 2 M. Hellmuth, M. Hernandez-Rosales, K.T. Huber, V. Moulton, P.F. Stadler, and N. Wieseke. *Orthology relations, symbolic ultrametrics and cographs*. *Journal of Mathematical Biology*, 66:39-420, 2013.
- 3 K.T. Huber, and E.G. Scholz. *Beyond representing orthology relations by trees*. *Algorithmica*, 80:73-103, 2018.

3.4 New kernels for TBR distance (or, equivalently, the maximum agreement forest problem): theory and practice

Steven Kelk

License © Creative Commons BY 3.0 Unported license
© Steven Kelk

Joint work of Steven Kelk, Simone Linz, Rim Van Wersch

Main reference Steven Kelk, Simone Linz: “A Tight Kernel for Computing the Tree Bisection and Reconnection Distance between Two Phylogenetic Trees”, *SIAM J. Discrete Math.*, Vol. 33(3), pp. 1556–1574, 2019.

URL <http://dx.doi.org/10.1137/18M122724X>

Given two phylogenetic (i.e. evolutionary) trees, the TBR-distance between them is the minimum number of “Tree Bisection and Reconnection” topological rearrangement moves required to transform one tree into another. This problem is NP-hard but was shown to be FPT in 2001 by Allen and Steel [1], who showed that polynomial-time reduction rules can be applied to reduce instances to size $28k$, where k is the TBR distance. In this talk we show that the kernelization strategy proposed by Allen and Steel [1] actually reduces the trees to size $15k - 9$, and that this is tight. The sharpened analysis is made possible by exploiting the equivalence of the TBR-distance problem to the problem of embedding the two trees parsimoniously into an undirected graph. Combining this equivalence with an older equivalence (that of “maximum agreement forests”) then yields a whole suite of new polynomial-time reduction rules which further shrink the trees to size $11k - 9$. We have also implemented the new reduction rules and describe briefly the results of preliminary experiments indicating that the new rules do, in practice, lead to further reductions in kernel size.

References

- 1 B. Allen, and M. Steel. *Subtree transfer operations and their induced metrics on evolutionary trees*. Annals of Combinatorics, 5:1-15, 2001.

3.5 An introduction to fixed-parameter algorithms: basic techniques and recent ideas

Christian Komusiewicz (Universität Marburg, DE) and André Nichterlein (TU Berlin, DE)

License © Creative Commons BY 3.0 Unported license
© Christian Komusiewicz and André Nichterlein

Main reference Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, Saket Saurabh: “Parameterized Algorithms”, Springer, 2015.
URL <https://doi.org/10.1007/978-3-319-21275-3>

Main reference Fedor V. Fomin, Daniel Lokshtanov, Saket Saurabh, Meirav Zehavi: “Kernelization: Theory of Parameterized Preprocessing”, Cambridge University Press, 2019.
URL <https://doi.org/10.1017/9781107415157>

We first review two algorithmic techniques for developing fixed-parameter algorithms: search tree algorithms and kernelization. Then we describe methods for showing fixed-parameter intractability. Finally, we discuss three more recent issues in fixed-parameter algorithms:

1. identification of good parameters via parameter hierarchies,
2. FPT-approximation, and
3. parameterized local search.

3.6 Some open problems in phylogenetic split theory

Vincent Moulton (University of East Anglia – Norwich, GB)

License © Creative Commons BY 3.0 Unported license
© Vincent Moulton

Split networks are one of the most commonly used type of phylogenetic network [2, 3]. Underlying these networks are combinatorial structures called split systems [1], which have a rich associated mathematical and computational theory. In this talk, we will give a brief introduction to split systems and split networks, and present some open problems related to these structures. These problems include:

1. What is the complexity of deciding whether or not a split system is flat [4]?
2. Is there an efficient algorithm to compute phylogenetic diversity for weakly compatible split systems [7]?
3. Does the 1-skeleton of the tight-span of a totally-split decomposable metric contain an optimal realization for the metric [5]?
4. Is a maximum linearly independent split system orderly if and only if it is circular [6]?

References

- 1 H.-J. Bandelt, A. Dress. *A canonical decomposition theory for metrics on a finite set*. Advances in Mathematics, 92:47-105, 1992.
- 2 D. Bryant, V. Moulton. *Neighbor-net: an agglomerative method for the construction of phylogenetic networks*. Molecular Biology and Evolution, 21:255-65, 2004.
- 3 D. Huson, D. Bryant. *Application of phylogenetic networks in evolutionary studies*. Molecular Biology and Evolution, 23:254-67, 2005.

- 4 M. Balvocute, A. Spillner, V. Moulton. *FlatNJ: A novel network-based approach to visualize evolutionary and biogeographical relationships*. Systematic Biology, 63:383-96, 2014.
- 5 S. Herrmann, J. Koolen, A. Lesser, V. Moulton, T. Wu. *Optimal realizations of two-dimensional, totally-decomposable metrics*. Discrete Mathematics, 338:1289-99, 2015.
- 6 V. Moulton, A. Spillner. *Order distance and split systems*.
<https://arxiv.org/abs/1910.10119>
- 7 A. Spillner, B. Nguyen, V. Moulton. *Computing phylogenetic diversity for split systems*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 5:235-44, 2008.

3.7 Continuous spaces of phylogenetic trees and networks

Megan Owen (*Lehman College – New York, US*)

- License** © Creative Commons BY 3.0 Unported license
© Megan Owen
- Joint work of** Daniel G. Brown, Gillian Grindstaff, Megan Owen
- Main reference** Daniel G Brown, Megan Owen: “Mean and Variance of Phylogenetic Trees”, Systematic Biology, Vol. 69(1), pp. 139–154, 2019.
URL <https://doi.org/10.1093/sysbio/syz041>
- Main reference** Gillian Grindstaff, Megan Owen: “Geometric comparison of phylogenetic trees with different leaf sets”, CoRR, Vol. abs/1807.04235, 2018.
URL <https://arxiv.org/abs/1807.04235>

A metric phylogenetic tree is a phylogenetic tree with lengths on its edges. Since evolutionary processes such as the coalescent depend on tree edge lengths, it is important to have a framework for analyzing both the tree topology and edge lengths together. One such framework is a continuous geometric space of phylogenetic trees, which has the metric trees as its points, and which accounts for the intrinsic properties of the trees through the geometry of the space. The most well-known such space is the Billera-Holmes-Vogtmann (BHV) treespace, and I will describe it, algorithms on it, and how it can be used to analyze tree data. I will also describe several other continuous treespaces, including an approach to analyzing trees with over-lapping leaf sets, and discuss network spaces.

3.8 Algorithmic tree and network problems in phylogenetics

Charles Semple (*University of Canterbury – Christchurch, NZ*)


- License** © Creative Commons BY 3.0 Unported license
© Charles Semple
- Joint work of** Mihaela Baroni, Magnus Bordewich, Janosch Döcker, Stefan Grünewald, Peter Humphries, Simone Linz, Vincent Moulton, Charles Semple
- Main reference** Janosch Döcker, Simone Linz, Charles Semple: “Displaying trees across two phylogenetic networks”, Theor. Comput. Sci., Vol. 796, pp. 129–146, 2019.
URL <http://dx.doi.org/10.1016/j.tcs.2019.09.003>
- Main reference** Peter J. Humphries, Simone Linz, Charles Semple: “Cherry Picking: A Characterization of the Temporal Hybridization Number for a Set of Phylogenies”, Bulletin of Mathematical Biology, Vol. 75(10), pp. 1879–1890, 2013.
URL <http://dx.doi.org/10.1007/s11538-013-9874-x>

Phylogenetic networks are a particular type of rooted, acyclic digraph and are used in computational biology to represent the non-treelike evolutionary history of extant species. Non-treelike (reticulate) processes in evolution include lateral gene transfer and hybridisation. Although evolution is not necessarily treelike at the species-level, at the level of genes, we typically assume treelike evolution. Thus phylogenetic networks are often viewed as an

amalgamation of gene trees (phylogenetic trees representing the evolutionary history of single genes). From this viewpoint, two of the most well-studied computational problems concerning phylogenetic networks is that of (i) determining the minimum number of reticulations for a network to embed a given set of conflicting trees and (ii) deciding whether or not a given network embeds a given tree. In this talk, I will present an overview of these problems and variations of them.

3.9 Searching for optimal phylogenetic histories

Katherine St. John (CUNY Hunter College – New York, US)

License  Creative Commons BY 3.0 Unported license

© Katherine St. John


Main reference Katherine St. John: “Review Paper: The Shape of Phylogenetic Treespace”, *Systematic Biology*, Vol. 66(1), pp. e83–e94, 2016.

URL <https://doi.org/10.1093/sysbio/syw025>

Evolutionary histories, or phylogenies, form an integral part of much work in biology. In addition to the intrinsic interest in the interrelationships between species, phylogenies are used for drug design, multiple sequence alignment, and even as evidence in a recent criminal trial. A simple representation for a phylogeny is a rooted, binary tree, where the leaves represent the species, and internal nodes represent their hypothetical ancestors. In this talk, we outline the optimality criteria used for evaluating phylogenetic trees and organizing the search space, the space of n -leaf trees. We classify the most popular metrics and the resulting treespaces. We examine the choice of metrics on the success of the search on finding the optimal trees, as well as the complexity of the algorithms, with emphasis on those problems that yield tractable or fixed parameter tractable algorithms.

3.10 Applying SNAQ with 5-net

Nihan Tokaç (Antalya International University, TR)

License  Creative Commons BY 3.0 Unported license

© Nihan Tokaç

Joint work of Céline Scornavacca, Nihan Tokaç

Phylogenetic networks are a necessary tool to represent the evolutionary history of species including horizontal gene transfers, hybridizations or gene flow. In [1], Solís-Lemus and Ané have inferred phylogenetic networks in a pseudolikelihood framework. In this work, the pseudolikelihood of a network is based on the likelihood formulas of its 5-taxon subnetworks instead of 4-taxon subnetworks

References

- 1 C. Solís-Lemus, C. Ané. *Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting*. *PLoS Genetics*, 12(3), e1005896, 2016.

3.11 Combinatorial decompositions and enumeration algorithms

Alexandru Tomescu (University of Helsinki, FI)

License © Creative Commons BY 3.0 Unported license
© Alexandru Tomescu

Main reference Alexandru I. Tomescu, Paul Medvedev: “Safe and Complete Contig Assembly Via Omnitigs”, in Proc. of the Research in Computational Molecular Biology – 20th Annual Conference, RECOMB 2016, Santa Monica, CA, USA, April 17–21, 2016, Proceedings, Lecture Notes in Computer Science, Vol. 9649, pp. 152–163, Springer, 2016.

URL http://dx.doi.org/10.1007/978-3-319-31957-5_11

A combinatorial decomposition is a characterization of an object in terms of same objects, but of smaller size. We introduce the classical decomposition of labeled DAGs by sources and show how it can be used for counting, approximate counting and random generation. We then introduce some classical complexity measures of an enumeration algorithm, present best- k enumeration algorithms and introduce the more recent topic of safe and complete algorithms, which was introduced to formalize the genome assembly problem. A safe algorithm is one outputting only partial solutions that are common to all solutions to a problem. A particular variant of the notion of safety has been previously studied under the title of persistency.

3.12 Constructing phylogenetic networks from smaller networks

Leo van Iersel (TU Delft, NL)

License © Creative Commons BY 3.0 Unported license
© Leo van Iersel

Joint work of Leo van Iersel, Vincent Moulton, Katharina Huber, Taoyang Wu, Celine Scornavacca, James Oldman, Sjors Koele

Main reference Katharina T. Huber, Leo van Iersel, Vincent Moulton, Céline Scornavacca, Taoyang Wu: “Reconstructing Phylogenetic Level-1 Networks from Nondense Binet and Trinet Sets”, *Algorithmica*, Vol. 77(1), pp. 173–200, 2017.

URL <http://dx.doi.org/10.1007/s00453-015-0069-8>

A common approach towards reconstructing large phylogenetic trees is to combine various phylogenetic trees on different, overlapping leaf label sets to a single phylogenetic tree on all leaf labels. A similar approach has been proposed for phylogenetic networks. In this case, the input consists of phylogenetic networks with different but overlapping leaf labels sets, and the goal is to construct a phylogenetic network that contains each of the input networks. Unfortunately, it has been shown that phylogenetic networks are in general not uniquely determined by their subnetworks. This contrasts the situation for phylogenetic trees, which are uniquely determined by their sets of contained 3-leaf trees, which are called triplets. Nevertheless, it has been shown that certain restricted classes of phylogenetic networks are uniquely determined by their sets of trinet sets, which are 3-leaf networks. For the severely restricted class of level-1 networks, it is even possible to reconstruct the network given all its trinet sets in polynomial time. However, for non-dense trinet sets, this problem is already NP-hard and the only existing algorithms are an efficient heuristic and an exponential-time exact algorithm. Open problems in this area include the questions whether there exist fixed-parameter algorithms and whether more general classes of networks are still uniquely determined by their trinet sets.

3.13 Tree Containment with polytomies

Mathias Weller University Paris-Est – Marne-la-Vallée, FR)

License © Creative Commons BY 3.0 Unported license
© Mathias Weller

Joint work of Matthias Bentert, Mathias Weller

Main reference Mathias Weller: “Linear-Time Tree Containment in Phylogenetic Networks”, in Proc. of the Comparative Genomics – 16th International Conference, RECOMB-CG 2018, Magog-Orford, QC, Canada, October 9-12, 2018, Proceedings, Lecture Notes in Computer Science, Vol. 11183, pp. 309–323, Springer, 2018.

URL http://dx.doi.org/10.1007/978-3-030-00834-5_18

Main reference Matthias Bentert, Josef Malík, Mathias Weller: “Tree Containment With Soft Polytomies”, in Proc. of the 16th Scandinavian Symposium and Workshops on Algorithm Theory, SWAT 2018, June 18-20, 2018, Malmö, Sweden, LIPIcs, Vol. 101, pp. 9:1–9:14, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2018.

URL <http://dx.doi.org/10.4230/LIPIcs.SWAT.2018.9>

In this work, we consider the Tree Containment problem, asking whether a given rooted phylogenetic tree is embeddable (leaf-label respecting topological minor) in a given phylogenetic network. We improve previously known results by presenting a linear-time algorithm for a broad class of networks, properly including the class of reticulation visible networks. We also show parameterized algorithms for a parameter that is stronger (that is, smaller in all instances) than the “level” of the network. All results work for so-called “hard polytomies” meaning high-degree nodes that represent large species fan outs. We further consider the more biologically relevant case of “soft polytomies”, where high-degree nodes represent uncertainty in the tree (branches with low support are often contracted after construction of the phylogeny) and show algorithms and NP-hardness for some classes of networks.

3.14 Phylogenetic diversity and biodiversity indices on phylogenetic networks

Kristina Wicke (Universität Greifswald, DE)

License © Creative Commons BY 3.0 Unported license
© Kristina Wicke

Joint work of Mareike Fischer, Kristina Wicke

Main reference Kristina Wicke, Mareike Fischer: “Phylogenetic diversity and biodiversity indices on phylogenetic networks”, Mathematical Biosciences, Vol. 298, pp. 80–90, 2018.

URL <https://doi.org/10.1016/j.mbs.2018.02.005>

Facing a major extinction crisis and the inevitable loss of biodiversity at the same time with limited financial means, it is often necessary to prioritize species for conservation. Existing approaches for prioritization, e.g. the Fair Proportion index, the Equal Splits index or the Shapley value, are based on phylogenetic trees and rank species according to their contribution to overall phylogenetic diversity (PD). However, in many cases evolution is not treelike and thus, phylogenetic networks have come to the fore as a generalization of phylogenetic trees, allowing for the representation of non-treelike evolutionary events, such as horizontal gene transfer or hybridization.

While phylogenetic diversity and PD indices have been studied in great detail for trees, research on how these concepts might be used in the context of phylogenetic networks is still in its infancy. In this talk, I will thus introduce phylogenetic diversity and PD indices for trees, before considering first attempts to extend these concepts from trees to networks. These attempts range from considering the treelike content of a network (e.g. the (multi)set of trees displayed by a network or the lowest stable ancestor tree associated with it) to

directly taking into account the network structure. In this talk, I will discuss some of these approaches, analyze their advantages and drawbacks and indicate some directions for future research.

3.15 Global-Local Clustering

Norbert Zeh (*Dalhousie University – Halifax, CA*)

License © Creative Commons BY 3.0 Unported license
© Norbert Zeh

Joint work of Mark E. L. Jones, Remie Janssen, Yuki Murakami, Leo van Iersel, Norbert Zeh

Given two phylogenetic trees over the same leaf set, cluster reduction identifies a pair of internal (i.e., non-root, non-leaf) nodes, one from each tree, so that both nodes have the same set of descendant leaves and then splits the input into two pairs of subtrees. Roughly, the first pair is the pair of descendant trees of the two chosen nodes; the second pair is the pair of trees obtained by giving the two chosen nodes a new label and removing their proper descendants. By applying this partition repeatedly, one obtains a partition of the two inputs trees into a collection of tree pairs that are hopefully much smaller. This partition is useful because Baroni, Semple, and Steel (2006) and Linz and Semple (2011) have shown that a maximum agreement forest (MAF; equivalent to the subtree prune-and-regraft distance) and a maximum acyclic agreement forest (MAAF; equivalent to an optimal hybridization network) of two trees can be computed from MAFs or MAAFs of the tree pairs in this partition. This has been verified to be incredibly effective for speeding up the computation of such agreement forests in practice (e.g., see Li and Zeh, 2017). However, some inputs do not decompose into small clusters because there may not be any pair of internal nodes with the same set of descendant leaves. For such inputs, cluster reduction is completely ineffective.

With the goal of extending the applicability of cluster reduction to a wider range of inputs, we introduce the notion of global-local clustering. An agreement forest (AF) is a forest that can be obtained from both input trees by cutting an appropriate subset of edges. A maximum agreement forest (MAF) is an agreement forest that can be obtained by cutting as few edges as possible. Such a forest can be found in $O(2^k * \text{poly}(n))$ time, where k is the number of edges cut and n is the number of leaves in the input. A (g, l, k) -clustering is a pair of edge sets, one per input tree, each of size at most k and such that cutting these edges produces an AF, along with a partition of each set into g “global” edges and the remaining “local edges”; this partition must have the property that each cluster in the cluster partition of the two forests obtained by cutting the global edges contains at most l local edges. We call a (g, l, k) -clustering an optimal (g, l) -clustering if k is the number of edges that need to be cut to obtain a MAF. Given the set of global edges, the local edges in all clusters can be found in $O(2^l * \text{poly}(n))$ time. The question is whether we can find the global edges in $O(f(g, l) * \text{poly}(n))$ time, that is, in time independent of k . This would allow us to find MAFs and MAAFs quickly even if k is large because g and l may be much smaller than k . This is true even if the input does not decompose into small clusters without cutting global edges first.

We give two positive and one negative answer to variations of this question. We show that given a fixed pair (g, l) , we can decide in $O(f(g, l) * \text{poly}(n))$ time whether a given input has a (g, l, k) -clustering for some k and if so, find the smallest k for which such a (g, l, k) -clustering exists. The standard approach for finding a MAF is to ask whether k edge cuts suffice to obtain an AF for increasing values of k until one obtains an affirmative

answer. We show that using this approach, we cannot decide in $O(f(g, l) * \text{poly}(n))$ time whether there exists an optimal (g, l) -clustering for a given input, essentially because we do not know whether the clustering we have obtained for a given pair (g, l) is optimal until we have tried all clusterings up to $g = k$ or $l = k$; larger values of g or l may allow us to cut fewer edges overall. On the positive side, we show that if there exists a (g, l) -clustering, even a suboptimal one, then the display graph of the two input trees has a tree width that is a function of only g and l . Thus, if the goal is simply to find a MAF in $O(f(g, l) * \text{poly}(n))$ time, we can construct the display graph along with a tree decomposition of this graph, and then employ the monadic second order logic framework of Kelk et al. (2016) to obtain a MAF in $O(f'(\text{treewidth}) * \text{poly}(n)) = O(f(g, l) * \text{poly}(n))$ time.

The main open problem is to determine whether there exist practical algorithms for finding (g, l, k) -clusterings. Neither our branching algorithm for a fixed pair (g, l) nor the tree decomposition-based algorithm of Kelk et al. is practical.

References

- 1 M. Baroni, C. Semple, and M. Steel. *Hybrids in real time*. Systematic Biology, 55:46-56, 2006.
- 2 S. Linz and C. Semple. *A cluster reduction for computing the subtree distance between phylogenies*. Annals of Combinatorics, 15:465-484, 2011.
- 3 Z. Li and N. Zeh. *Computing maximum agreement forests without cluster partitioning is folly*. In Proceedings of the European Symposium on Algorithms, pages 56:1–56:14, 2017.
- 4 S. Kelk, L. van Iersel, C. Scornavacca, and M. Weller. *Phylogenetic incongruence through the lens of monadic second order logic*. Journal of Graph Algorithms and Applications, 20:189–215, 2016.

4 Working Groups

4.1 Longest Common Subsequence for similar strings

Laurent Bulteau (University Paris-Est – Marne-la-Vallée, FR), Mark E. L. Jones (CWI – Amsterdam, NL), Rolf Niedermeier (TU Berlin, DE), and Till Tantau (Universität zu Lübeck, DE)

License © Creative Commons BY 3.0 Unported license

© Laurent Bulteau, Mark E. L. Jones, Rolf Niedermeier, and Till Tantau

The longest common subsequence problem (LCS) is a core string comparison problem, where one seeks a common pattern appearing in a (possibly large) set of (possibly long) strings. It has been well-studied in a large number of settings, in particular regarding parameterized algorithms. However, the complexity remains open for one of the most natural parameterizations : the maximum number of deletions in any input string. Similar questions are open for related problems: Shortest Common Supersequence (where the pattern is obtained by inserting characters rather than deletions), Center String and Median String (allowing all edit operations). The closest related results are an FPT algorithm for Closest String parameterized by the distance (allowing substitutions only) and FPT algorithm for LCS parameterized by the distance plus number of strings.

Using the concept of maximal common subsequence, we have developed a fixed-parameter algorithm solving LCS, with only the distance as parameter (using only linear time on the size of the input), thus answering our main open question. We expect that a similar method should apply to Shortest Common Supersequence. However, a different approach may be required for Center/Median string problems.

4.2 Hybridization for many trees with non-identical leaf sets

Britta Dorn (Universität Tübingen, DE), Christian Komusiewicz (Universität Marburg, DE), Catherine McCartin (Massey University, NZ), André Nichterlein (TU Berlin, DE), Mathias Weller (University Paris-Est – Marne-la-Vallée, FR), and Norbert Zeh (Dalhousie University – Halifax, CA)

License © Creative Commons BY 3.0 Unported license

© Britta Dorn, Christian Komusiewicz, Catherine McCartin, André Nichterlein, Mathias Weller, and Norbert Zeh

A phylogenetic network is a directed acyclic graph with a single source and whose sinks are labeled bijectively with the elements of some label set. A phylogenetic tree is a phylogenetic network that is in fact a rooted tree. A network is said to display a tree if the tree can be obtained from the network by deleting a subset of its vertices and edges and suppressing nodes of in-degree 1 and out-degree 1. A hybridization network for a set of trees is a network that displays all trees in the set. It is an optimal hybridization network if it has the minimum (undirected) cyclomatic number among all hybridization networks for the same set of trees.

The parameterized complexity of constructing an optimal hybridization network for a pair of input trees is fairly well understood. For multiple input trees, the story is more complicated. There is no known practical parameterized algorithm for constructing an optimal hybridization network for more than two trees. The problem is known to be fixed-parameter tractable because there exists a quadratic kernel [1]. For specialized network construction problems such as finding tree-child networks or temporal networks, there exist practically efficient branching algorithms (cf. [2, 3]). Even these algorithms are limited in their usefulness to practitioners because of their assumption that all input trees share the same leaf set. This is rarely the case for real-world inputs obtained by constructing phylogenetic trees from genes shared by (subsets of) a set of species. Thus, we would like to construct optimal hybridization networks for multiple input trees with overlapping but non-identical leaf sets. This type of problem has received little attention so far because non-identical leaf sets pose no challenge whatsoever for pairs of trees. For more than two input trees, the problem becomes significantly harder. It is not fixed-parameter tractable when parameterized only by the hybridization number because deciding whether a given set of triplets (trees with 3 leaves) has a network that displays them and has hybridization number at most 2 is NP-hard [5].


Two somewhat natural parameterizations involve a pair of parameters. The first one considers the hybridization number k and the number of leaves l that are missing from at least one input tree. The second one considers the hybridization number and the number of input trees. Note that in the triplet example above, both the number of leaves absent from at least one input tree and the number of trees are large (at least linear in n). Our working group proved that the hybridization number problem is fixed-parameter tractable in both of these parameterizations. For the parameterization by hybridization number and missing leaves, we prove that there exists a kernel of size $O(k^2 + l)$ by extending the existence proof of a size- $O(k^2)$ kernel for the case of identical leaf sets. For the parameterization by hybridization number and number of trees, we prove that the display graph of the input trees has a tree width that is a function of k and t . Thus, the monadic second order logic framework of [4] can be used to find an optimal hybridization network for these trees in $O(\text{tree width})$ time and thus in $O(f(k, t))$ time. While these results do not lead to practical algorithms for constructing hybridization networks on sets of trees with non-identical leaf sets, they shed light on the computational complexity of this problem.

References

- 1 L. van Iersel, S. Kelk, and C. Scornavacca. *Kernelization for the hybridization number problem on multiple nonbinary trees*. Journal of Computer and System Sciences, 82:1075–1089, 2016.
- 2 L. van Iersel, R. Janssen, M. Jones, Y. Murakami, and N. Zeh. *A practical fixed-parameter algorithm for constructing tree-child networks from multiple binary trees*. aXiv:1907.08474 [cs.DM], 2019.
- 3 S. Borst. Personal communication, 2019.
- 4 S. Kelk, L. van Iersel, C. Scornavacca, and M. Weller. *Phylogenetic incongruence through the lens of monadic second order logic*. Journal of Graph Algorithms and Applications, 20:189–215, 2016.
- 5 J. Jansson, N. B. Nguyen, and W.-K. Sung. *Algorithms for combining rooted triplets into a galled phylogenetic network*. SIAM Journal on Computing, 35:1098–1121, 2006.

4.3 Hybridization Number for multiple multifurcating trees

Mark E. L. Jones (CWI – Amsterdam, NL) and Vincent Moulton (University of East Anglia – Norwich, GB)

License  Creative Commons BY 3.0 Unported license
© Mark E. L. Jones and Vincent Moulton

In the **Hybridization Number** problem, we are given a set of rooted phylogenetic trees on a set of taxa X , and our aim is find a phylogenetic network that is as simple as possible while displaying all of those trees. (“As simple as possible” means having minimum reticulation number; that is, minimum number of edges that must be removed to turn the network into a tree). This problem is NP-hard and APX-hard, even when the input consists of two binary trees.

For two binary trees, the problem is fixed-parameter tractable with respect to the reticulation number. This result has been extended to non-binary trees and to instances with more than one input tree. In fact, FPT algorithms are known for **Hybridization Number** when either the number of trees or the maximum outdegree of any tree is bounded. However, the general case when there is an unbounded number of input trees with unbounded outdegree remains open.

Our group began attempted to find a general FPT algorithm for **Hybridization Number** on arbitrary number of trees with unbounded outdegree. Previous results have made use of the notion of a “generator” for a network which characterizes the overall structure in terms of the location of the reticulation nodes. A key observation is that a network with reticulation number k has $O(k)$ “sides” that limit the possible structure of trees displayed by this network. Existing techniques make use of this structure to show the correctness of some flavor of chain reduction rule (in which caterpillar-like substructure that are common to all trees can be reduced by deleting some taxa). This in combination with the standard subtree reduction rule has led to kernels with $O(dk^2)$ taxa (in the case of trees with maximum outdegree d) or $O(k(5k)^t)$ taxa (in the case of t trees with unbounded outdegree).

Ideally, we would be able to prove the correctness of a similar chain reduction rules for multiple large-outdegree trees (with the length of the chain likely increased), which would be enough to imply a kernel in a similar way. So far we have had no success in proving the correctness of such a rule. Nevertheless we believe it may be possible to exploit the generator structure of a k -reticulation network to enable more fine-grained reduction rules.

4.4 Maximum agreement subtrees

Katherine St. John (CUNY Hunter College – New York, US), Magnus Bordewich (Durham University, GB), Simone Linz (University of Auckland, NZ), Megan Owen (Lehman College – New York, US), Charles Semple (University of Canterbury – Christchurch, NZ), and Kristina Wicke (Universität Greifswald, DE)

License © Creative Commons BY 3.0 Unported license

© Katherine St. John, Magnus Bordewich, Simone Linz, Megan Owen, Charles Semple, and Kristina Wicke

Our working group focused on a conjecture of Martin and Thatte on lower bounds on the maximum agreement subtree distance between two trees. Steel and Székely [3] shows that for any two trees on n leave, the agreement subtree is of size $\Omega(\log(\log n))$. Martin and Thatte [1] improves the lower bound to $\Omega(\sqrt{\log n})$ and conjectured that a lower bound of $\Omega(\sqrt{n})$ if both trees are balanced. There has been related work on expected distance between trees on the same shape is $\Omega(\sqrt{n})$ [2], suggesting that the conjecture should hold.

References

- 1 D. M. Martin, and B. D. Thatte. *The maximum agreement subtree problem*. Discrete Applied Mathematics 161:13–14, 1805–1817, 2013.
- 2 P. Misra, and S. Sullivant. *Bounds on the expected size of the maximum agreement subtree for a given tree shape*. SIAM Journal on Discrete Mathematics 33:2316–2325, 2019.
- 3 M. Steel, and L. A. Székely. *An improved bound on the maximum agreement subtree problem*. Applied Mathematics Letters 22:1778–1780, 2009.

4.5 Constructing phylogenetic networks from trinets

Leo van Iersel (TU Delft, NL), Vincent Moulton (University of East Anglia – Norwich, GB), Leen Stougie (CWI – Amsterdam, NL), and Nihan Tokaç (Antalya International University, TR)

License © Creative Commons BY 3.0 Unported license

© Leo van Iersel, Vincent Moulton, Leen Stougie, and Nihan Tokaç

Main reference Leo van Iersel, Vincent Moulton: “Trinets encode tree-child and level-2 phylogenetic networks”, Journal of Mathematical Biology, Vol. 68(7), pp. 1707–1729, 2014.

URL <http://dx.doi.org/10.1007/s00285-013-0683-5>

There are many interesting problems related to constructing phylogenetic networks from subnetworks and in particular from trinets, which are 3-leaf subnetworks. While it has been shown that binary level-2 and tree-child networks are encoded (uniquely determined) by their trinets, there currently does not exist a polynomial-time algorithm for reconstructing a binary tree-child network from its trinets. Moreover, it is not clear whether binary level- k networks are encoded by their trinets for $3 \leq k \leq 11$. A counter example is known only for level-12.

In this working group, we have first focused on the question whether there exists a polynomial-time algorithm for reconstructing a binary temporal network from its trinets. The class of temporal networks form a subclass of the tree-child networks. A network is *tree-child* if every non-leaf vertex has a child that is not a reticulation. A network is *temporal* if it is tree-child and its vertices can be labeled by integers such that the label value remains unchanged along reticulation arcs and strictly increases along tree-arcs. We have a sketch of a polynomial-time algorithm for the temporal case.


Secondly, we considered the same question for the class of tree-child networks. For this case, it seems that there also exists a polynomial-time algorithm although with a worse running time than in the temporal case.

Finally, we considered the question whether level-3 networks are encoded by their trinets. Unfortunately, it seems that the techniques used for level-1 and level-2 do not generalize to level-3. Hence, we discussed different ways to try to prove the result for level-3.

5 Open Problems

5.1 List of open problems

Leo van Iersel (TU Delft, NL)

License  Creative Commons BY 3.0 Unported license
© Leo van Iersel

1. Are level- k networks, with $3 \leq k \leq 11$, uniquely determined by their trinets (3-leaf subnetworks)? And can they be reconstructed from their trinets in polynomial time?
2. Can tree-child networks be reconstructed from their trinets in polynomial time?
3. Are there FPT algorithms for constructing level-1 networks from non-dense trinet sets, with any reasonable parameter?
4. Is constructing a level- k network displaying a given dense set of triplets FPT with k as parameter?
5. Is constructing a temporal network displaying a given set of non-binary trees FPT with the number of reticulations in the network (and possibly the maximum outdegree of the input trees) as parameters?
6. Does there exist a polynomial-time algorithm for deciding whether there exists a binary tree-child network displaying a given set of (at least three) binary trees?
7. Does there exist a polynomial-time algorithm for deciding whether a given unrooted binary network can be oriented to become a rooted tree-child network? And to become a rooted stack-free network?
8. Does there exist an FPT algorithm for constructing a network with reticulation number at most k displaying a given set of non-binary trees, when the only parameter is k ?
9. Does there exist an EPT algorithm (i.e. an FPT algorithm with running time $c^k \text{poly}(n)$) for constructing a network with reticulation number at most k displaying a given set of 4 binary trees?
10. Does there exist an FPT algorithm for constructing a network with reticulation number at most k displaying a given set of binary trees with non-equal leaf-sets, when the two parameters are k and the number of leaves that do not appear in all trees? (Note that all previous problems assume equal leaf-sets.)
11. For $4 \leq r \leq 7$, does there exist a constant $f(r)$ such that any set of r -state characters is compatible if and only if every size- $f(r)$ subset is compatible?

Participants

- Allan Bai
University of Canterbury –
Christchurch, NZ
- Magnus Bordewich
Durham University, GB
- Laurent Bulteau
University Paris-Est –
Marne-la-Vallée, FR
- Janosch Döcker
Universität Tübingen, DE
- Britta Dorn
Universität Tübingen, DE
- Anne-Sophie Himmel
TU Berlin, DE
- Katharina T. Huber
University of East Anglia –
Norwich, GB
- Mark E. L. Jones
CWI – Amsterdam, NL
- Steven Kelk
Maastricht University, NL
- Christian Komusiewicz
Universität Marburg, DE
- Simone Linz
University of Auckland, NZ
- Marefatollah Mansouri
TU Wien, AT
- Catherine McCartin
Massey University, NZ
- Vincent Moulton
University of East Anglia –
Norwich, GB
- André Nichterlein
TU Berlin, DE
- Rolf Niedermeier
TU Berlin, DE
- Megan Owen
Lehman College –
New York, US
- Charles Semple
University of Canterbury –
Christchurch, NZ
- Katherine St. John
CUNY Hunter College –
New York, US
- Leen Stougie
CWI Amsterdam, NL
- Till Tantau
Universität zu Lübeck, DE
- Nihan Tokaç
Antalya International
University, TR
- Alexandru Tomescu
University of Helsinki, FI
- Leo van Iersel
TU Delft, NL
- Mathias Weller
University Paris-Est –
Marne-la-Vallée, FR
- Kristina Wicke
Universität Greifswald, DE
- Norbert Zeh
Dalhousie University –
Halifax, CA

