university of groningen / faculty of science and engineering / mathematics and applied mathematics

# Computational Stemmatology

MSc Applied Mathematics Project Proposal

**Keywords**: Computational Stemmatology, Natural Language Processing, Data Science

**Starting Date**: 16th October 2023

**Midpoint Date**: 16th March 2024

**Final Date**: 12th July 2024

**Student**: Darren Zammit s5284236

**First Supervisor**: Dr. Julian Koellermeier

**Second Supervisor**: Dr. Tsegaye Tashu

# 1 A Brief Introduction to Computational Stemmatology

Computational stemmatology is a discipline wherein one aims to reconstruct the genealogy (stemma) of a text on the basis of differences between surviving manuscripts (witnesses) [1]. Over time a text is copied which usually results in changes. One can analyse the differences between different witnesses to construct the stemma. Philologists construct stemmata to study the errors made and how scribes influenced a text which may be useful to study the evolution of cultures, values, and ideas. On top of this, the field has potential applications in plagiarism detection [10], analysing the evolution of computer viruses [11] and content-based social network analysis [12]

Graphical modelling takes a central role in constructing stemmata [1]. Contemporary methods – which are primarily analogous to those in phylogenetics, the study of the evolutionary history of species – can be split into distance-based [2], parsimony-based [3], and statistical methods [4]. In these methods every word in the text is equally important. In reality there are many different kinds of errors and changes, some more significant than others, leading to the algorithm failing to faithfully construct a stemma. For example, spelling mistakes or changing symbols should have far less impact on a script than adding or redacting text. Another issue which these methods struggle to address is that of contamination, that is when a text is copied from multiple manuscripts [1]. If different scripts were used for different chapters then this could lead to outputs which faithfully reflect the stemma of most chapters but not all if different chapters were copied from different manuscripts.

Since Lee's seminal paper [5] in applying computational phylogenetics models in stemmatology multiple advances in computational stemmatology have been made. However, most papers in computational stemmatology are authored by computer scientists and philologists with mathematicians and statisticians representing a small minority of authors [1]. Moreover, algorithms designed specifically for stemmatology, even ones whose framework is adapted for stemmatolology, are very few in number [1]. Computational methods usually rely on transforming the data into analogues of genetic data and feeding it directly into a phylogenetics program as though it were phylogenetic data.

# 2 Research Aims and Some (Tentative) Ideas

The first phase of the project will consist of a literature review of modern methods in computational stemmatology and phylogenetics. The aim is to extend existing methods or adapt novel methods from phylogenetics and test them for accuracy and robustness against contamination, missing strings and missing manuscripts on artificial data sets and texts for which the stemma is already known such as those used in [6] which have become a standard in testing methods due to the diversity of their stemmatic structures.

The RHM algorith [16], a compression based method using a minimum-information criterion and stochastic tree optimization was developed specifically for stemmatology and has shown to be very effective at constructing stemmata. This algorithm approximates the Kolmogorov complexity, defined as the smallest possible but complete description of a string such as compressing "aaa" into "3a". It is mathematically impossible to prove that any such description is the smallest possible thus one always relies on approximations. RHM approximates the dissimilarity between witnesses by using GZIP compression as the approximation. One idea would be to apply different approximations for Kolmogorov complexity and compare the validity of the different cost functions for stemmatic analysis.

Another highly successful algorithm designed specifically for stemmatology is SEMSTEM [14]. SEMSTEM leverages the Bayesian Structural Expectation Maximisation (SEM) algorithm [15] to construct the Bayesian belief network. The SEM algorithm uses maximum likelihood to compute expectations and maximalisations for missing data. Unlike other algorithms, it can produce multifurcating trees which is more compatible with stemmatology. On top of this the degree of the internal nodes is not limited either and the extant manuscripts used as input to the method can be used as labels for the internal nodes as well as the leaf nodes of the stemma. The base algorithm (SEM) infers the unknown contents of latent edges not considered by the algorithm for computing a previous tree structure based on the contents of observed nodes. This estimation is then used to derive a maximum likelihood tree, the process for which features a number of unobserved nodes. The algorithm iterates this process to produce a best estimate tree structure. One

direction we could take is to attempt learning the Bayesian belief network using other methods/concepts such as neural networks [17] or variational inference [18].

The resulting stemmata will be compared to the outputs of commonly used phylogenetic programs such as PAUP [7], Phylip [8] and Mr Bayes [9] which have been used as standard benchmarks in computational stemmatology [13]. This can be achieved by converting the textual data into genetic data which can be fed directly into the aforementioned programs [1].

In the final phase of the project we plan on working with actual philologists to generate stemmata using tailored models for previously unstudied texts. Some of the texts considered were impenetrable even to manual stemmatological methods and thus this would be the first time a stemma is produced for these manuscripts.

# 3 Timeline

We plan on starting the project on the 16th of October 2023. The student shall have 25 credits of extra courses in 2023/24. Topics in Probability and Statistics (5 credits) in term 1, Statistical Signal Processing and Mathematical Modelling of Infectious Diseases (10 credits) in term 2, Statistical Genomics (5 credits) in term 3 and the Research Seminar in Applied Mathematics (5 credits). Thus, the midterm date will be taken as the 16th of March and the end date will be taken as the 12th of July 2024.

# References


[1] Roelli, Philipp. (2020). Handbook of stemmatology : history, methodology, digital approaches. Berlin, Germany: De Gruyter.

[2] M. Spencer and C. J. Howe, "Estimating distances between manuscripts based on copying errors," Literary and Linguistic Computing, vol. 16, p. 467–484, 2001.

[3] W. M. Fitch, "Toward defining the course of evolution: minimum change for a specific tree topology," Systematic Biology, vol. 20, p. 406–416, 1971.

[4] Q. Nguyen and T. Roos, "Likelihood-based inference of phylogenetic networks from sequence data by phylodag," in International Conference on Algorithms for Computational Biology, 2015.

[5] A. Lee, 'Numerical Taxonomy Revisited: John Griffith, Cladistic Analysis and St. Augustine's Quaestiones in Heptateuchum.' Studia Patristica XX: 24-32, 1989.

[6] T. Roos, T. Heikkilä, Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets, Literary and Linguistic Computing, Volume 24, Issue 4, December 2009, Pages 417–433, https://doi.org/10.1093/llc/fqp002.

[7] D. L. Swofford, PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts. 2003.

[8] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005.

[9] F. Ronquist, and J. P. Huelsenbeck. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574. 2003.

[10] C. Liu, C. Chen, J. Han, and P. S. Yu, "GPLAG: Detection of software plagiarism by program dependence graph analysis," Proc. KDD 2006, pp. 872–881, 2006

[11] J. Sun, S. Papadimitriou, C.-Y. Lin, N. Cao, S. Liu, W. Qian, "MultiVis: Content-based social network exploration through multi-way visual analysis," Proc. SDM 2009, pp. 1063–1074, 2009

[12] S. Wehner, "Analyzing worms and network traffic using compression," J. Comp. Secur., vol. 15, pp. 303–320, 2007

[13] T. Roos, T. Heikkilä, Evaluating methods for computer-assisted stemmatology using artificial benchmark data sets, Literary and Linguistic Computing, Volume 24, Issue 4, December 2009, Pages 417–433, https://doi.org/10.1093/llc/fqp002

[14] T. Roos, and Y. Zou. 2011. "Analysis of Textual Variation by Latent Tree Structures." In Proceedings of the 11th International Conference on Data Mining (ICDM-2011), edited by Diane J. Cook, 567–576. Los Alamitos, California: IEEE Computer Society. Available at: http://www.cs.helsinki.fi/u/ttonteri/pub/icdm2011.pdf. Accessed 28 October 2015.

[15] N. Friedman 1998. "The Bayesian Structural EM Algorithm." In Uncertainty in Artificial Intelligence: Proceedings of the Fourteenth Conference (1998), edited by Gregory F. Cooper, 129–138. San Francisco: Morgan Kaufmann. Available at: www.cs.huji.ac.il/ nir/Papers/Fr2.pdf Accessed 28 October 2015.

[16] T. Roos, T. Heikkilä and P. Myllymäki. 2006. "A Compression-Based Method for Stemmatic Analysis." In ECAI 2006: Proceedings of the 17th European Conference on Artificial Intelligence: August 29 – September 1, 2006, edited by Gerhard Brewka et al., 805–806. Amsterdam: IOS Press.

[17] S. Monti, G. F. Cooper. "Learning Bayesian belief networks with neural network estimators." Advances in Neural Information Processing Systems: 1996.

[18] D. M. Blei, A. Kucukelbir and J. D. McAuliffe (2017) Variational Inference: A Review for Statisticians, Journal of the American Statistical Association, 112:518, 859-877, DOI: 10.1080/01621459.2017.1285773