Phylogenetic Trees Can Reveal Manuscript Outliers Based on Their Pitch Content

Tim Eipert & Fabian C. Moss

Institut für Musikforschung, Julius-Maximilians-Universität Würzburg, Germany

✉ { tim.eipert | fabian.moss }@uni-wuerzburg.de timeipert fabianmoss

Introduction

Digital editions of music offer the possibility of **distant reading**, which, in addition to detailed analysis of samples, allows conclusions to be drawn about the characteristics of a repertoire. There are a variety of methods from disciplines such as linguistics or bioinformatics.

This poster demonstrates how **phylogenetic analysis**, borrowed from bioinformatics, can be used to compare different sources based on the variants of overlapping monophonic medieval liturgical chants contained by them. Based on a newly created dataset, we address the question of how much the identified sources differ in content and whether this ordering is related to the place of origin of the manuscripts.

Dataset

Graduale Synopticum is an edition of the earliest sources of liturgical chants related to the proper of the mass. It is a valuable resource as it provides a synopsis of the melodic variants in an online database. However, the transcribed melodies are only available as image files. We encoded the images with the software **OMMR4all** to use them with computational methods. We get about 120 different chants in up to 7 versions. The output format is Volpiano, which uses a string of the characters *a* to *i* to encode the pitches of the melody. The full dataset is being prepared and will be published soon.

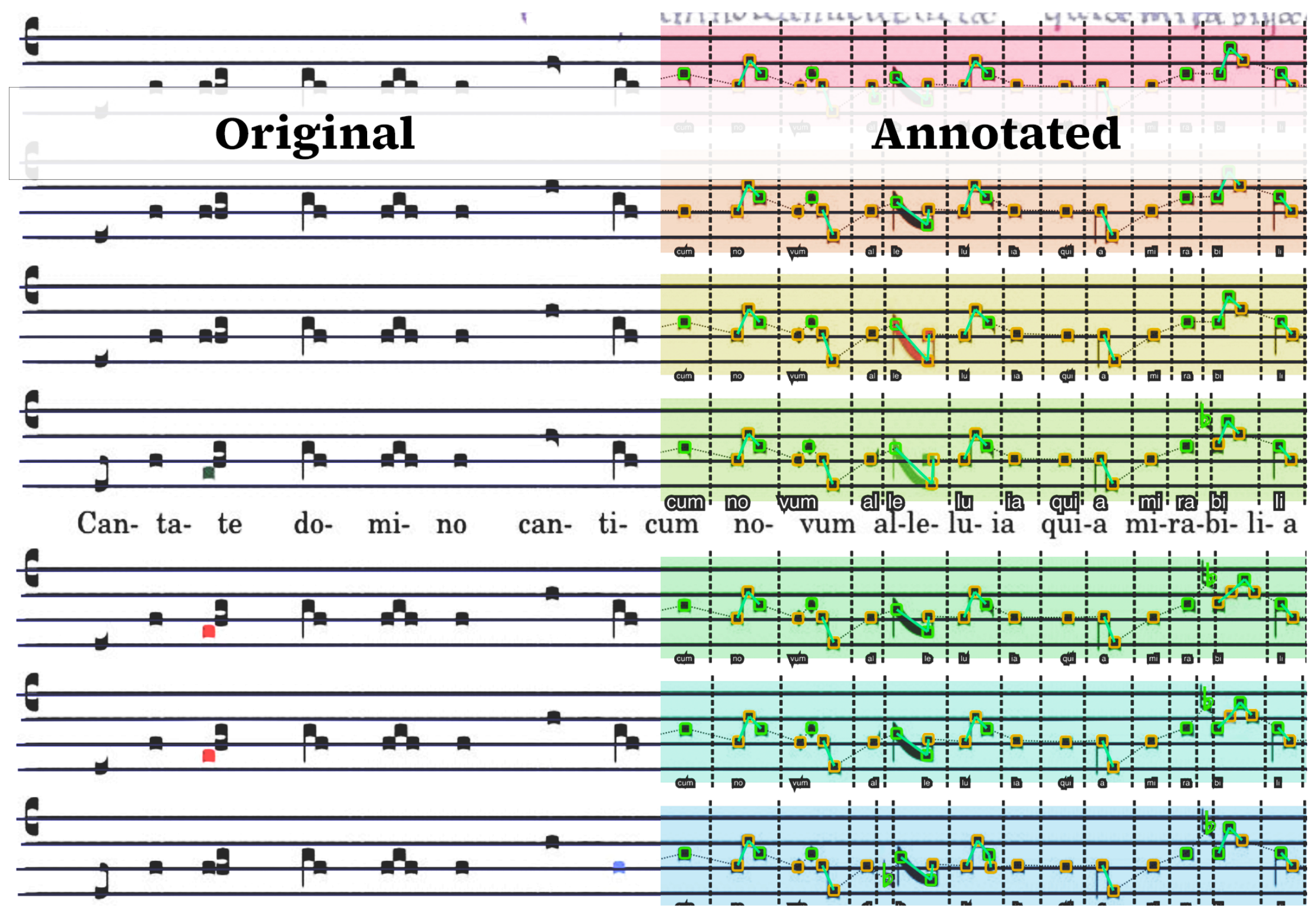
Original
Annotated
Can- ta- te do- mi- no can- ti- cum no- vum al-le- lu- ia qui-a mi-ra-bi- li- a

Figure 1: Original and annotated version of a chant transcription in the **OMMR4all** editor. The staff lines, symbols, and their connections, as well as text alignment, are annotated.

Methods

🔍 Optical Music Recognition: We use the software *Optical Medieval Music Recognition for All* (OMMR4all) [2], a tool developed in Würzburg for the automatical transcription of medieval manuscripts, which generates a Volpiano encoding of all the variants in the Graduale Synopticum.

↔ Multiple Sequence Alignment: We align the chant pitches using MAFFT [5], a software tool for multiple sequence alignment (MSA). MAFFT allows arbitrary strings with a custom substitution matrix, unlike other MSA implementations. We concatenate the aligned chants into a single dataset to obtain an overview of all the chants in the corpus. For this alignment, we apply a basic substitution matrix that assigns lower probabilities to larger melodic leaps compared to smaller steps, reflecting the typical movement in chant melodies.

🌲 Phylogenetic Tree Inference: To construct a phylogenetic tree, we use an adjusted version of MrBayes [1], a software tool based on Bayesian modeling [4]. This approach generates a probability distribution over a range of possible trees, enabling us to determine the most probable evolutionary relationships among the chants. The branch length correlates with the number of substitutions between two manuscripts. The tree summarizes information from 900 sampled trees from the posterior distribution [3].

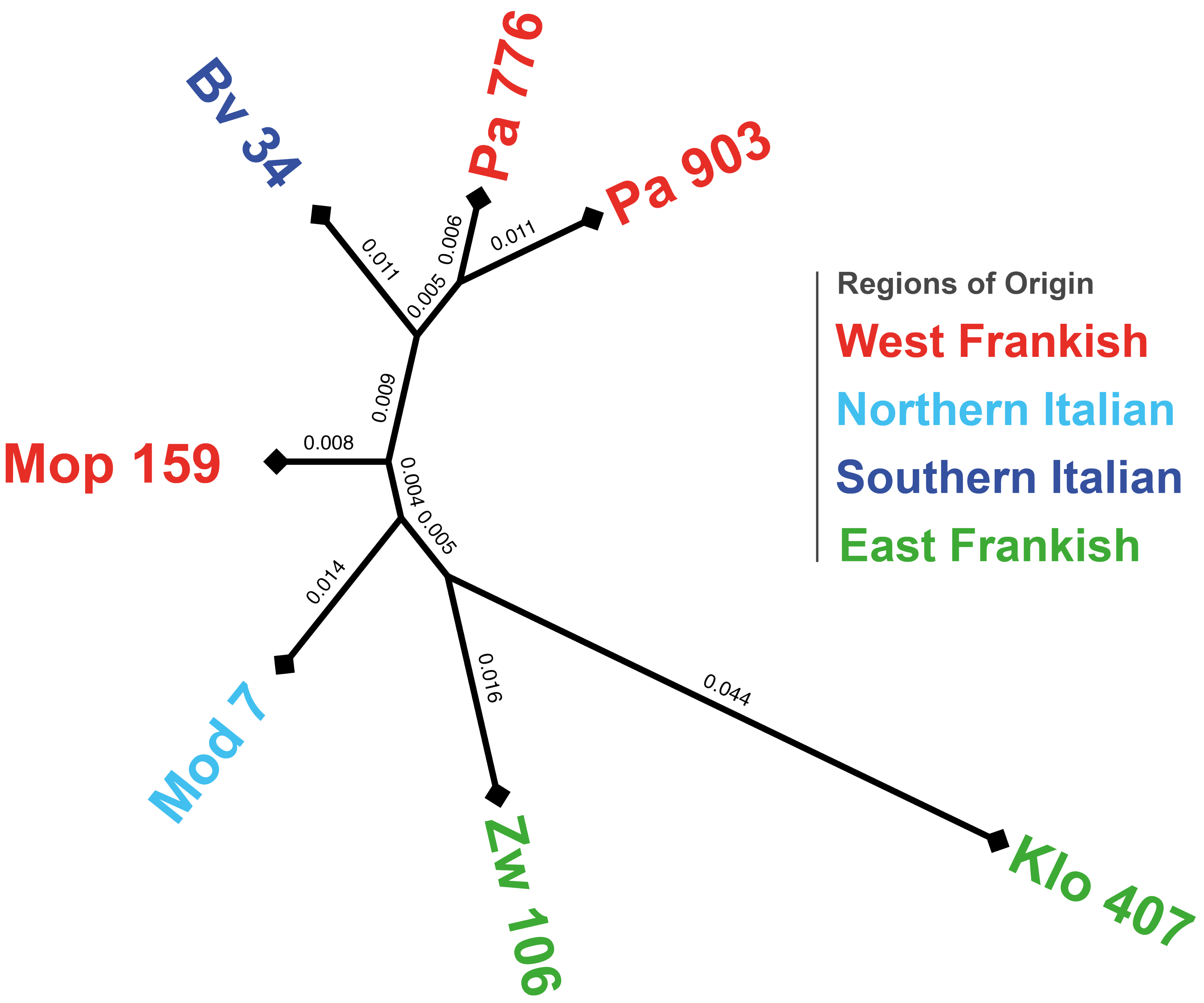

Figure 2: Maximum clade credibility tree with manuscripts (leaves) and their nearest common ancestors (branches). The branch length correlates with the number of substitutions between two manuscripts. The tree summarizes the information from 900 tree samples from the posterior distribution inferred from the dataset. [3]

Result

The phylogenetic analysis reveals that manuscripts from **similar geographical origins** tend to **cluster together**. However, the Klosterneuburg manuscript stands apart, exemplifying an East-Frankish melodic chant dialect. The analysis shows that *Pa 903* and *Pa 776*, both sources from the West Frankish territory, exhibit similarities with *Bv 34*, a source from Southern Italy. In contrast, both East Frankish manuscripts (*Zw 106* and *Klo 407*) are closest to each other in terms of content and geographical proximity. Nonetheless, they are far apart regarding the number of substitutions between the melodic variants, possibly due to the use of *Zw 106* in the context of a Cistercian monastic order. *Mod 7*, a Northern Italian source, occupies an intermediate position between the Roman/West Frankish and East Frankish traditions.

Discussion & Future Research

The assumptions underlying its construction constrain the resulting phylogenetic tree. For example, the **bifurcating structure** typical of biological entities may not accurately capture the actual development of chant variants. However, it can still provide **a helpful overview** of the variation in manuscript contents. The value of the results increases as more sources are included, allowing for a broader contextualization of additional manuscripts based on content similarity. Manuscripts that do not fit the established structure can be identified for further investigation.

References

[1] J. Hajic Jr, G. A. Ballen, K. H. Mühlová, and H. Vlhová-Wörner. "Towards Building a Phylogeny of Gregorian Chant Melodies.". In: *ISMIR*. 2023, pp. 571–578.

[2] A. Hartelt, T. Eipert, and F. Puppe. "Optical Medieval Music Recognition—A Complete Pipeline for Historic Chants". In: *Applied Sciences* 14.16 (2024), p. 7355.

[3] J. Heled and R. R. Bouckaert. "Looking for trees in the forest: summary tree from posterior samples". In: *BMC evolutionary biology* 13 (2013), pp. 1–11.

[4] J. P. Huelsenbeck and F. Ronquist. "MRBAYES: Bayesian inference of phylogenetic trees". In: *Bioinformatics* 17.8 (2001), pp. 754–755.

[5] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". In: *Nucleic acids research* 30.14 (2002), pp. 3059–3066.