A Compression-Based Method for Stemmatic Analysis

Teemu Roos [1] and **Tuomas Heikkilä** [2] and **Petri Myllymäki** [1]

1 INTRODUCTION

Stemmatology studies relations among different variants of a text that has been gradually altered as a result of imperfectly copying the text over and over again. Underlying these variants there is what we could call a 'family tree', a graph representing the process of copying the text where each new version becomes a direct descendant of the exemplar(s) from which it is copied. The aim of stemmatic analysis is to reconstruct this family tree, known as the 'stemma', based on the surviving copies of the text. Applications are mainly in humanities, especially textual criticism, but the methods can be used to study the evolution of any symbolic objects, including chain letters [2] and computer viruses. We propose an algorithm for stemmatic analysis based on a minimum-information criterion and stochastic tree optimization. The intuitive idea behind compression-based approaches is that if a text can be significantly compressed, then the compression algorithm has found regularities which can be further exploited in an analysis such as ours. Our approach is related to phylogenetic reconstruction criteria such as maximum parsimony and maximum likelihood, and builds upon algorithmic techniques developed for bioinformatics. For a more detailed version see [14, 15].

2 A MINIMUM-INFORMATION CRITERION

In his seminal work on computer-assisted stemmatology, O'Hara used a parsimony method of the PAUP software [18] in Robinson's Textual Criticism challenge [13]. For further applications of maximum parsimony and related method, see [8, 9, 17, 20] and the references therein.

Our compression-based *minimum information* criterion shares many properties of the maximum parsimony method. Both can also be seen as instances of the *minimum description length* (MDL) principle of Rissanen [12] which in turn is a formal version of Ockham's razor. The minimum-information criterion measures the amount of information, or *code-length*, required to reproduce all the manuscripts by the process of copying and modifying the text under study. In order to describe a new version of an existing manuscript, one needs an amount of information that depends on both the amount and the type of modifications made. For instance, describing a deletion of a word or a change of word order requires less information than introducing a completely new expression.

In order to be concrete, we need a precise, numerical, and computable measure for the amount of information. The commonly accepted definition of the amount information in individual objects is Kolmogorov complexity, see [10], defined as the length of the shortest computer program to describe the given object. However, Kolmogorov complexity is defined only up to a constant, and fundamentally uncomputable. Therefore, in the spirit of a number of earlier authors [1, 2, 3, 4, 6, 11, 19], we approximate Kolmogorov complexity by using a compression program (`gzip`). In particular, given two strings, x and y, the amount of information in y conditional on x, denoted by $C(y \mid x)$ is given by the length of the compressed version of the concatenated string x, y minus the length of the compressed version of x alone. One of the advantages of using a string compression method that operates directly on the text is that only minimal preprocessing is required, contrary to most of the methods referred to above.

Let $G = (V, E)$ be an undirected graph where V is a set of nodes corresponding to the text variants, $E \subset V \times V$ is a set of edges. We require that the graph is a connected bifurcating tree, meaning that (i) each node has either one or three neighbors, and (ii) the tree is acyclic. Such a graph G can be made directed by picking any one of the nodes as the root and directing each edge away from the root. Given a directed graph $\vec{G}$, the total information cost of the tree is defined as

$$C(\vec{G}) \quad = \quad \sum_{v \in V} C(v \mid \mathrm{Pa}(v)) \tag{1}$$

where the sum is over all the variants v, and $\mathrm{Pa}(v)$ denotes the parent node of v unless v is the root in which case $\mathrm{Pa}(v)$ is the empty string. For practical reasons (mainly for reduced computational cost) we make some modifications to this criterion, see the full version [15].

3 AN ALGORITHM FOR CONSTRUCTING STEMMATA

Since it is known that many of the text variants have been lost during the centuries between the time of the writing of the first versions and present time, it is not realistic to build a tree of only the available variants that we have as our data. The common way (in phylogeny) of handling this problem is to include in the tree a number of 'hidden' nodes, i.e., nodes representing individuals whose characteristics are unobserved. We construct bifurcating trees that have N observed nodes as leafs, and $N - 2$ hidden nodes as the interior nodes.

Evaluating the criterion (1) now involves the problem of dealing with the hidden nodes. Without knowing the values of v, it is not possible to compute $C(v \mid \mathrm{Pa}_i(v))$. We solve this problem by searching simultaneously for the best tree structure $\vec{G}$ and for the optimal contents of the hidden nodes with respect to criterion (1). Perhaps surprisingly, given a tree structure, finding the optimal contents is feasible. The method for efficiently optimizing the contents of the hidden nodes is an instance of dynamic programming (or 'elimination' in graphical models) and called 'the Sankoff algorithm' [5] or 'Felsenstein's algorithm' [16].

There still remains the problem of finding the tree structure, which together with corresponding optimal contents of the hidden nodes

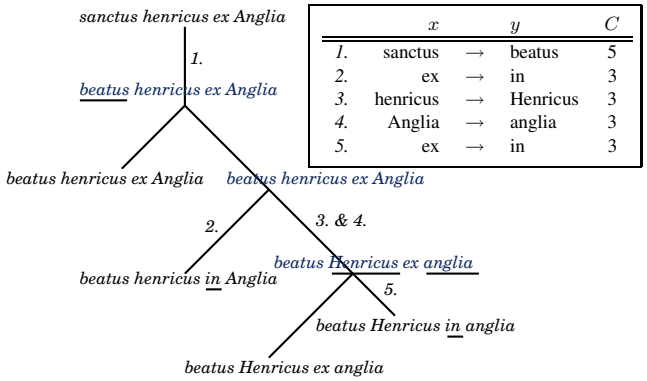
[1] Complex Systems Computation Group, Helsinki Institute for Information Technology, Finland, email: firstname.lastname@cs.helsinki.fi

[2] Dept. of History, University of Helsinki, Finland, email: tuomas.m.heikkila @helsinki.fi

minimizes criterion (1). The obvious solution, trying all possible tree structures and choosing the best one, fails because for N leafs nodes, the number of possible bifurcating trees is exponentially large (see [5]). Instead, we have to resort to heuristic search, trying to find as good a tree as possible in the time available. We use a simulated annealing algorithm that starts with an arbitrary tree and iteratively tries to improve it by small random modification, such as exchanging the places of two subtrees.

4 RESULTS AND CONCLUSIONS

We illustrate the behavior of the method by an artificial example in Fig. 1. Assume that we have observed five pieces of text, shown at the tips of the tree's branches. One of the trees — not the only one — minimizing the information cost with total cost of 44 units (bytes) is drawn in the figure. The sum of the (unconditional) complexities of the four words in the top-most string ("*sanctus henricus ex Anglia*") is equal to $8 + 9 + 3 + 7 = 27$, which happens to coincide with the length of the string, including spaces and a finishing newline. The changes, labeled by numbers 1–5 in the figure, yield $5 + 3 + 3 + 3 + 3 = 17$ units of cost. Thus the total cost of the tree equals $27 + 17 = 44$ units.



	x		y	C
1.	sanctus	→	beatus	5
2.	ex	→	in	3
3.	henricus	→	Henricus	3
4.	Anglia	→	anglia	3
5.	ex	→	in	3

Figure 1. An example tree obtained with the compression-based method for the five strings at the tips of the branches. Changes are underlined and numbered. Costs of changes are listed in the box; no cost is incurred by a transition that leaves the string unchanged, i.e., $C(y \mid x) = 0$ if $y = x$. Best reconstructions at interior nodes are shown at the branching points.

In our main experiment, reported in the full version [15], we applied the presented method to the tradition of the legend of St. Henry of Finland[3], of which some fifty manuscripts are known. Even for such a moderate number, manual stemma reconstruction is prohibitive due to the vast number of potential explanations, and the obtained stemma is the first attempt at a complete stemma of the legend of St. Henry. The relationships discovered by the method are largely supported by more traditional analysis in earlier work. Moreover, our results have pointed out groups of manuscripts not noticed in earlier manual analysis.

We are currently carrying out controlled experiments with artificial data with known 'ground-truth' solution to which the results can be compared. Outside historical and biological applications, analysis of computer viruses is an interesting future research topic.

[3] St. Henry is a key figure of the Finnish Middle Ages. According to the medieval tradition, he was the Bishop of Uppsala (Sweden), and one of the leaders of a Swedish expedition to Finland around 1155, during which he was murdered. The oldest text concerning St. Henry is his legend written in Latin by the end of the 13th century at the very latest. For identification of the sources as well as a modern edition of the legend see [7].

ACKNOWLEDGEMENTS

This work has significantly benefited from discussions with Tommi Mononen and Kimmo Valtonen at HIIT, and Prof. Paul Vitányi and Rudi Cilibrasi at CWI. This work was supported in part by IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

REFERENCES

[1] D. Benedetto, E. Caglioti, and V. Loreto, 'Language trees and zipping', *Physical Review Letters*, **88**(4), 048702–1–048702–4, (2002).
[2] C.H. Bennett, M. Li, and B. Ma, 'Chain letters and evolutionary histories', *Scientific American*, 76–81, (November 2003).
[3] X. Chen, S. Kwong, and M. Li, 'A compression algorithm for DNA sequences and its applications in genome comparison', in *Genome Informatics*, eds., K. Asai, S. Miyano, and T. Takagi, Tokyo, (1999). Universal Academy Press.
[4] R. Cilibrasi and P.M.B. Vitányi, 'Clustering by compression', *IEEE Transactions on Information Theory*, **51**(4), 1523–1545, (2005).
[5] J. Felsenstein, *Inferring phylogenies*, Sinauer Associates, Sunderland, Massachusetts, 2004.
[6] S. Grumbach and F. Tahi, 'A new challenge for compression algorithms: genetic sequences', *Journal of Information Processing and Management*, **30**(6), 875–866, (1994).
[7] T. Heikkilä, *Pyhän Henrikin legenda* (in Finnish), Suomalaisen Kirjallisuuden Seuran Toimituksia 1039, Helsinki, 2005.
[8] C.J. Howe, A.C. Barbrook, M. Spencer, P. Robinson, B. Bordalejo, and L.R. Mooney, 'Manuscript evolution', *Trends in Genetics*, **17**(3), 147–152, (2001).
[9] A.-C. Lantin, P. V. Baret, and C. Macé, 'Phylogenetic analysis of Gregory of Nazianzus' Homily 27', in *7èmes Journées Internationales d'Analyse statistique des Données Textuelles*, eds., G. Purnelle, C. Fairon, and A. Dister, pp. 700–707, Louvain-la-Neuve, (2004).
[10] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications, 2nd. Ed.*, Springer-Verlag, New York, 1997.
[11] D. Loewenstern, H. Hirsh, P. Yianilos, and M. Noordewier, 'DNA sequence classification using compression-based induction', Technical Report 95–04, DIMACS, (1995).
[12] J. Rissanen, 'Modeling by shortest data description', *Automatica*, **14**, 465–471, (1978).
[13] P. Robinson and R.J. O'Hara, 'Report on the textual criticism challenge 1991', *Bryn Mawr Classical Review*, **3**(4), 331–337, (1992).
[14] T. Roos, T. Heikkilä, R. Cilibrasi, and P. Myllymäki, 'Compression-based stemmatology: A study of the legend of St. Henry of Finland', Technical Report HIIT-2005-3, Helsinki Institute for Information Technology.
[15] T. Roos, T. Heikkilä, and P. Myllymäki. A compression-based method for stemmatic analysis, 2006. Full version available at `http://www.cs.helsinki.fi/teemu.roos/pub/ecai06full.pdf`.
[16] A. Siepel and D. Haussler, 'Phylogenetic estimation of context-dependent substitution rates by maximum likelihood', *Molecular Biology and Evolution*, **21**(3), 468–488, (2004).
[17] M. Spencer, K. Wachtel, and C.J. Howe, 'The Greek Vorlage of the Syra Harclensis: A comparative study on method in exploring textual genealogy', *TC: A Journal of Biblical Textual Criticism*, **7**, (2002).
[18] D.L. Swofford. PAUP*: Phylogenetic analysis using parsimony (*and other methods). version 4., 2003.
[19] J.-S. Varre, J.-P. Delahaye, and É. Rivals, 'The transformation distance: a dissimilarity measure based on movements of segments', in *Proceedings of German Conference on Bioinformatics*, Koel, Germany, (1998).
[20] E. Wattel and M.P. van Mulken, 'Weighted formal support of a pedigree', in *Studies in Stemmatology*, eds., P. van Reenen and M.P. van Mulken, 135–169, Benjamins Publishing, Amsterdam, (1996).