

Language trees, zipping and error estimation

Toni Giorgino *

7th November 2002

Abstract

A method was recently proposed to estimate distances between a pair of given texts. The distance estimation appeared to be reliable enough to infer a phylogenetic tree of languages, even though no error estimation has been provided. This essay reviews the method and explains its application for inferring phylogeny on a collection of heterogeneous texts. An approach for estimating the confidence of the classification is introduced and the results are discussed.

1 Background and method

A simple method for estimating a “distance” between a pair of given text was recently described in a paper [1], which attracted attention and criticism [2]. Its simplicity and the lack of assumptions on the nature of texts, however, make it an appealing technique for evaluating distance matrices. A (symmetric) distance matrix contains an entry for each pair of items, i.e., texts from different sources. The authors propose a distance D_{AB} reported in the appendix and show

*Questa è la tesina conclusiva del primo anno della scuola SAFI, scritta per il corso “Linguaggio ed Evoluzione” tenuto dai proff. Cavalli-Sforza e Wang.

three applications: language attribution, authorship attribution, and inferring a phylogenetic tree.

The latter application may be questionable. First of all, it should be noted that the resulting tree is not “phylogenetic” in principle, as it describes in a pictorial form the distances between pair of languages as they are now. The fact that a common ancestor existed in the past is not implied by the tree (even though it may be a good hint).

More remarkably, an estimate of the error of the classification is not given. Given a set of entirely random sequences, for example, a meaningless tree would be obtained anyway (hopefully, the lengths of its branches would then all be approximately equal; the phylogeny picture given in the cited article does not show those lengths).

2 The problem chosen

This essay will describe an application of the principle proposed by [1]. The method will be slightly extended to provide a confidence measure of the classification, as it will be described in the following. The goal of the present work was to attempt to classify a group of texts written in the same language presumably in different ages. Ideally, the classification would take the shape of a tree fitted to the observed data. The tree is unrooted and the fitting does not assume the presence of an evolutionary clock: the distance from the root of the tree to each of the leaves may be different.¹ This hypothesis is reasonable because, even in the questionable assumption that distances are more or less proportional to the ages of the texts, they may have “frozen” in different times in the past.

¹Tree fitting from distance matrix the cited work and this essay, was performed with the software Phylogeny Inference Package “Phylip”.

The data set which was taken for classification is a Hebrew version of the bible; details of the transcription of the version used are given in reference [3]. The Hebrew language was chosen to avoid artefacts, possibly introduced by translation; this choice is admittedly naive as it ignores phylogeny on the origin of the texts. The data were originally divided into 39 books, of varying length between 23 and 2528 verses. Books were taken as the elements to be classified, i.e., the leaves of the tree. The method described in [1] was implemented as a set of computer programs.

3 Results and error estimation

To apply the distance formula given above, all book pairs were formed and each book was taken as the source \mathbf{A} , in turn. As discussed in the appendix, each source \mathbf{A} should produce a long sequence A and a shorter one a . The long version of the text was derived by taking the book in its entirety; the short version a was generated by choosing *randomly* 50 of its verses. A requirement for the method to work, further discussed in [4], is that the long and the short version differ significantly in size. Books shorter than 200 verses were therefore excluded from the classification, as they could not provide a “long” version; the threshold was arbitrarily chosen to grant a difference in size of a factor of at least four. The selection left 24 texts, which were used as sources; distances were evaluated pairwise using the formula given above. The left side of table 1 shows the result of the classification, as a lower-triangular distance matrix.

An interesting point in the method described above is that the evaluation of a distance entry is *stochastic*. This feature appears because of the random sampling of the book to form its “short” version. Running the same distance program multiple times, therefore, results in multiple instances of the distance

matrix which are partly different from each other.

Randomness was introduced to evaluate the reliability of the classification: the program was in fact run 50 times in order to collect statistical information on each matrix entry. The method employed has some resemblance with techniques known as “bootstrapping” methods, in which pseudo-random samples are built by randomly repeating or selecting a subset of the available data [5].

The 50 distance matrices collected (samples) were averaged, and the standard deviation of each matrix element was calculated: table 1 shows it along with every entry. Qualitative inspection shows anticorrelation between the value of the distance and the error which affects it. Also, errors seem often to be either near 10% or 20%, depending on whether the corresponding distance is more or less than 5. This dichotomy is confirmed by more careful plots (not shown here).

The right side of table 1 shows a plot of the distance matrix. This plot conveys some interesting information: there are some books which are further than all the other (high ridges). When the matrix is made into a tree, these books stick out as very long branches: figure 1 left shows the tree corresponding to the averaged matrix, obtained by the Fitch-Margoliash method.

	QOH	ISA	JER
QOH			
ISA	6.910 ± 7%		
JER	6.532 ± 7%	1.248 ± 24%	
DAN	13.855 ± 8%	8.604 ± 15%	7.849 ± 14%
NEH	12.968 ± 4%	7.408 ± 8%	7.033 ± 6%
ZEC	12.957 ± 4%	7.395 ± 7%	6.784 ± 6%
2SA	7.261 ± 7%	1.773 ± 21%	1.684 ± 16%
JOB	7.761 ± 8%	2.710 ± 15%	2.373 ± 16%
1CH	7.179 ± 6%	1.773 ± 20%	1.531 ± 18%
NUM	6.903 ± 7%	1.475 ± 20%	1.198 ± 19%
ISA	6.653 ± 6%	1.401 ± 21%	0.952 ± 23%
2CH	6.962 ± 7%	1.606 ± 19%	1.236 ± 17%
PRO	9.028 ± 6%	4.059 ± 12%	3.736 ± 13%
LEV	7.349 ± 5%	2.052 ± 19%	1.596 ± 20%
EZE	6.780 ± 6%	1.277 ± 19%	0.928 ± 22%
DEU	6.896 ± 6%	1.609 ± 15%	1.255 ± 21%
GEN	6.556 ± 7%	1.209 ± 21%	0.955 ± 20%

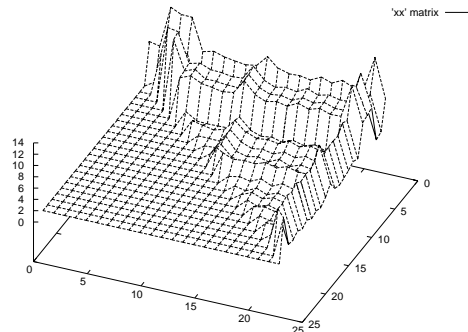


Table 1: *Left*: part of the distance matrix, with relative error shown. *Right*: Plot of the distance matrix.

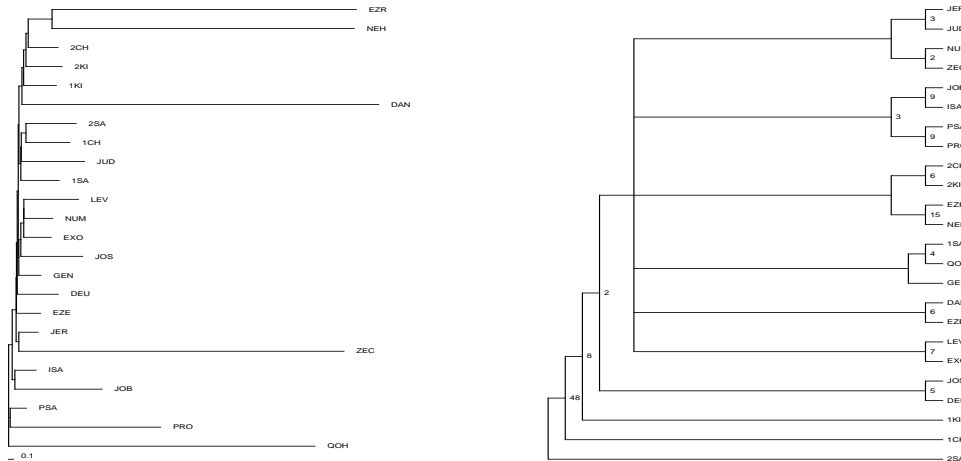


Figure 1: *Left*: the tree built out of the average distance matrix. Branch lengths are significant. *Right*: the samples' consensus tree. Numbers indicate in how many trees (out of 50) that branch was present. The number 48 is an artefact and it is not significant. Branch lengths are also not significant.

4 Consensus tree

As said before, a byproduct of the analysis performed was a fair number of distance matrices. In principle, one could obtain a tree from *each one* of the matrices collected. However, visual inspection of all the trees is unlikely to provide useful information. Instead, it would be interesting to try and see if a significant fraction of the trees share (at least partly) a common subset.

Searching for common subtrees is known as building the *consensus tree*; the algorithm used in the following is similar to Nelson's [6]. Consensus trees are useful when there is a good amount of phylogenetic data available, obtained for example by bootstrapping methods [7].

The consensus tree built out of the 50 trees collected is shown in the right side of figure 1. The result, however, is a bit dismaying: there is no subtree shared by a significant fraction of them. In particular, none of these subtrees appears 50% of the times, which would make it a *majority rule* consensus tree.

The most interesting branching, present in about one third of the cases, is the forking between EZR and NEH.² Other branches are also have some importance, but they are better seen in runs with larger statistics (not shown here).

5 Conclusions

This essay showed how the basic technique described in [1] can be modified to provide error estimates. The method proposed is applied to a data set made by bible books, and results in an ensemble of statistically independent distance matrices. The classification then proceeds in two steps: first an *average distance matrix* was calculated out of the randomized instances, together with the standard error of each element.

Two facts are discovered, which seem to indicate that the distance matrix obtained is correct, namely: (*a*) some elements have distances which are consistently large with respect to all the other and (*b*) the figures are little affected by the randomized choice of text fragments, i.e., their standard error is between 10 and 15%. The presence of elements which are remarkably further than others is also apparent from ridges in the plot of the distance matrix.

The neat separation of elements in two groups with respect to their standard error — either 10% or 20% — although of unclear origin, is also a strong indication that the method has really discovered some disomogeneity in the data set, which could maybe investigated further.

On the other hand, the availability of multiple distance matrices with the corresponding trees suggests also to apply a consensus tree method. No majority rule consensus tree could be found in the data set examined. In the data set examined, this is interpreted as the fact that the trees built from this data set

²Namely, the books of Ezra and Nechemia.

are *not* reliable.

At least two reasons may account for this failure. In the first place, the method could not be sensitive enough to differentiate writer styles. The resulting trees would then be affected by random permutations of their shorter branches. A second explanation could be that a tree is simply inappropriate to describe the distance matrix. Difference of writing style, for instance, could be much more significant than differences in ages. A tree representation is to be sought if a priori reasons exist to believe that a bifurcation process underlies the observed data. In the case of written texts this assumption may be inappropriate, and it would not be surprising if authors wrote each with his own style without significant progressive “groupings”.

The only remarkable correlation found by the consensus tree is the one between EZR and NEH. Although not surprising (the two books contain long list of names), the fact that this known correlation was actually identified provides anyway a positive feedback on the reliability of the method.

Appendix

The method relies on a well-known compression algorithm which tries to reduce redundancy in an arbitrary input sequence of symbols. The algorithm achieves reduction by taking note of repeated strings, which are replaced by back references. Given a sequence A and a shorter sequence b , the authors of [1] define the quantity $\Delta_{Ab} = L_{Ab} - L_A$ where L_x is the length of the sequence x after compression. Intuitively, Δ is a measure of how well the sequence b compresses under the “influence” of the preceding corpus A .

The distance between text extracted from two sources \mathbf{A} and \mathbf{B} is estimated with a quantity similar to Δ but symmetric with respect to its arguments: $D_{\mathbf{AB}} =$

$\frac{\Delta_{Ab}-\Delta_{Bb}}{\Delta_{Bb}} - \frac{\Delta_{Ba}-\Delta_{Aa}}{\Delta_{Aa}}$, where A and a are respectively a long and a short sequence produced from the source \mathbf{A} . Although not described in the article, $D_{\mathbf{AB}}$ may be negative or not satisfy the triangular inequality, in which case it has to be corrected [8].

References

- [1] D. Benedetto, E. Caglioti, and V. Loreto, “Language trees and zipping,” *Phys. Rev. Lett.*, vol. 88, no. 4, p. 048702, 2002.
- [2] J. Goodman, “Extended comment on language trees and zipping.” Preprint on it.arxiv.org/abs/cond-mat/0202383.
- [3] S. Gross, “Notes on the ASCII version of the Tanach.” <http://www.innerx.net/personal/tsmith/DeuTorah.html>.
- [4] A. Puglisi, D. Benedetto, E. Caglioti, V. Loreto, and A. Vulpiani, “Data compression and learning in time sequences analysis.” Preprint on it.arxiv.org/abs/cond-mat/0207321.
- [5] P. Diaconis and B. Efron, “Computer intensive methods in statistics,” *Scientific American*, pp. 116–130, May 1983.
- [6] G. Nelson, “Cladistic analysis and synthesis: principles and definitions . . .,” *Systematic Zoology*, vol. 28, pp. 1–21, 1979.
- [7] J. Felsenstein, “Confidence limits on phylogenies: an approach using the bootstrap,” *Evolution*, no. 39, pp. 783–791, 1985.
- [8] E. Caglioti Personal communication.