SANTA FE INSTITUTE – COMPLEX SYSTEMS SUMMER SCHOOL '09
WORKING PAPER

1, 2, 3, language!

Building the phylogenetic tree of languages with numbers

01/09/2009

Andrew Berdahl[1,2], Lucas Lacasa[3,4]

[1]Complexity Scienge group, Dept. of Physics and Astronomy, University of Calgary, Canada
[2]Dept. of Ecology and Evolutionary Biology, Princeton University, United States
[3]Centre de Recerca Matematica, Barcelona, Spain
[4]Dept. of Applied Mathematics, ETSIA, Universidad Politécnica de Madrid, Spain

Abstract: In this paper we make use of bioinformatics tools to build up the phylogenetic tree of languages. We had access to a large dataset gathering the numbers one to ten in over five thousand languages. In a first step, for each language we have concatenated each of the ten numbers in a string. After defining a mapping between the 26-letter alphabet and the DNA-like codons, we make use of a global alignment method to calculate the distance between pairs of strings, i.e. between languages. We finally generate the distance matrix and its associated phylogenetic tree. Specifically, we have used this method to generate the phylogenetic tree of indoeuropean languages. Despite the small size of the dataset (only ten words per language), preliminary results perfectly match the state of the art. We finally discuss some potential applications and future work, on relation to culture-based concepts such as trading or spreading of culture.

Introduction

Sequence alignment procedures are widely used in computational biosciences, in order to quantify the correlations between DNA strings. These methods rely on information theoretic measures that quantify the information that two elements share. Concretely, the so called global alignment methods estimate the informational distance between two strings of similar size, by inserting an optimal number of blanks and subsequently comparing each string element-by-element according to a pre-defined distance criterion (e.g. a Hamming distance). By doing so, one can compare the DNA of different species, calculate their pairwise distance, and derive a phylogenetic tree.

While these methods are usually tackled in biology, it is straightforward that they are not only restricted to such biological garment, since they capture correlations between symbolic strings. In this paper we make use of these and other bioinformatics tools to build up the phylogenetic tree of languages. Concretely, we will make use of a global sequence alignment to compare languages. We had access to a large dataset gathering the numbers one to ten in over five thousand languages (http://zompist.com/numbers.shtml). In a first step, for each language we have concatenated each of the ten numbers in a symbolic string.

In order to use a standard software of phylogenetic tree calculation found in *Matlab* (based on DNA allignment), we have mapped the 26-letter alphabet to DNA-like codons. Then, we have mapped each alphabetical string (numbers one to ten concatenated in a given language) to a DNA-like string. We have finally calculated a pairwise distance between strings using a standard global sequence alignment method and obtained the resulting phylogenetic tree. Specifically, we have used this method to generate the phylogenetic tree of indoeuropean languages.

Alphabet mapping

Each of the 26 letters of the alphabet (a,b,c,d…) is mapped to a 3-nucleotide string, where the nucleotides are chosen from the set {A,T,C,G}. Note that this mapping enable us to use standard softwares for DNA alignment, something that can be skipped if needed by programming our own sequence alignment algorithm.

There are many criteria susceptible to be considered in the mapping. For instance, phonetic and feature-based properties should be encoded in a realistic mapping. However, our main point in this paper is understand whether if small datasets, such as the numbers one to ten, are enough for the discrimination between languages. Accordingly, in a first approximation we have done a random mapping: each letter from the alphabet is mapped to a random combination of nucleotides (we have nonetheless checked that the mapping is injective, and that each letter maps to different nucleotide strings). In future work, we will address more

sophisticated mappings, that will encode phonetic and feature-based properties, and will compare those results with this null model.

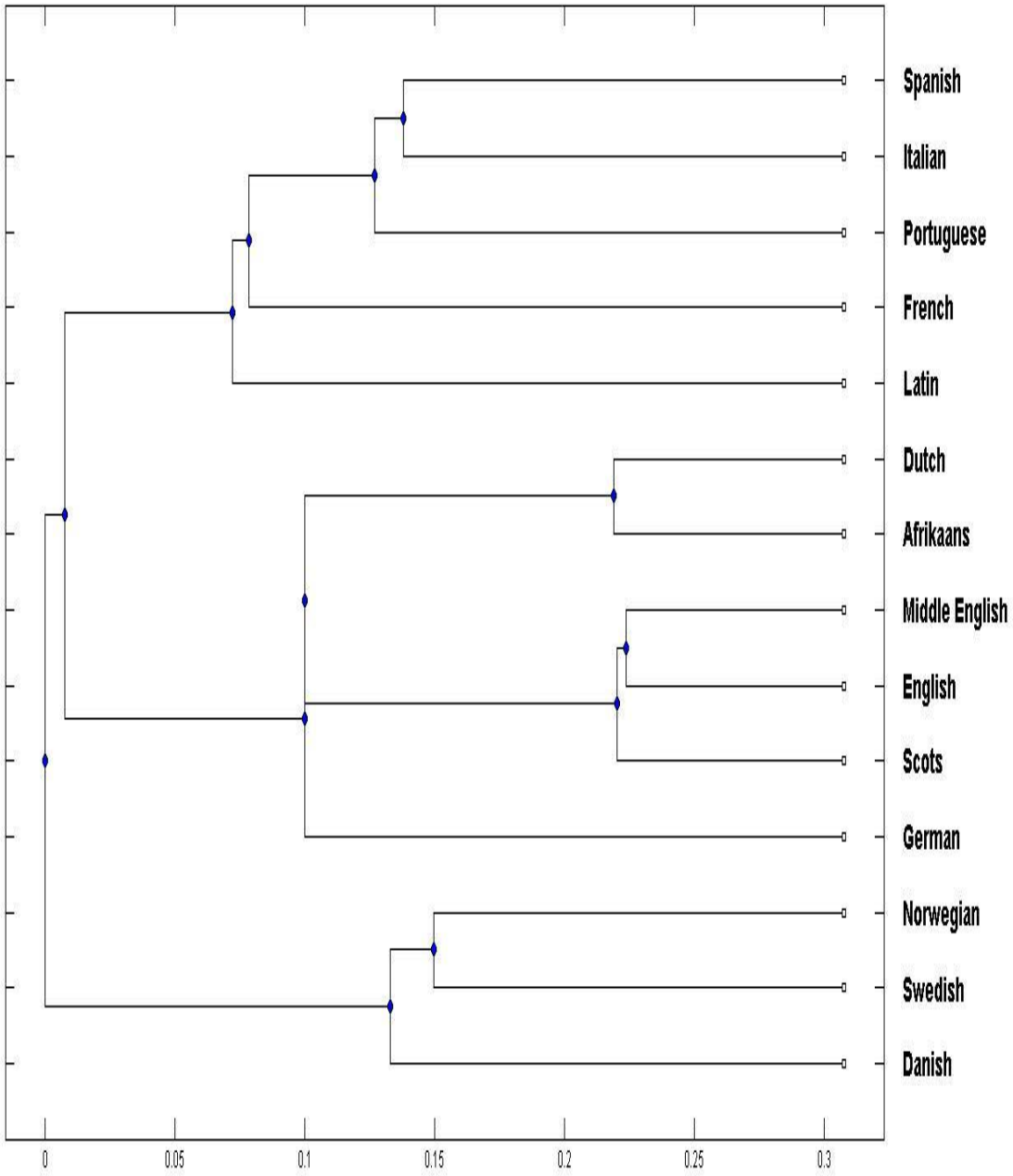
We finally have a new alphabet: each of the 26 letters is a 3-nucleotide string. In order to make global sequence alignment analysis, for each language we will concatenate each number into a single string, and then map it into a DNA-like string according to the previous mapping. As a result, each language is encoded in an ordered set of elements from {A,T,C,G} (note that the set size is not the same for all languages, while two similar languages are likely to have a similar associated set).

Results

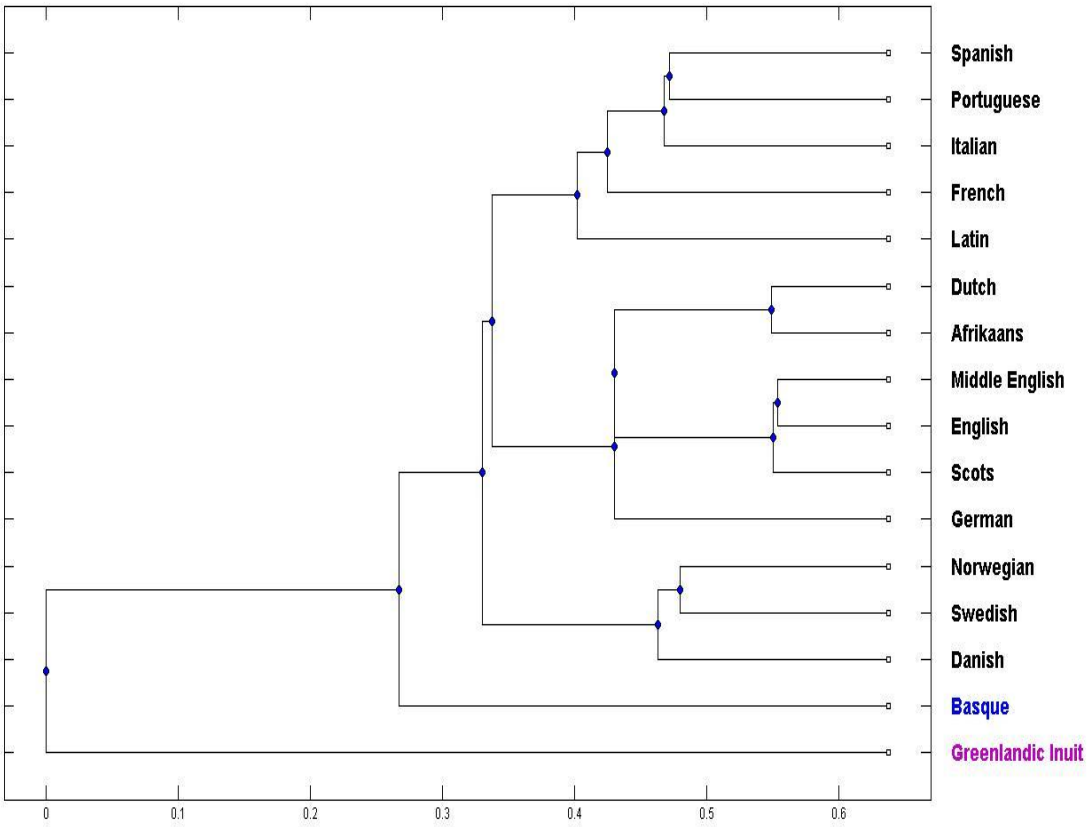
We are now ready to compare languages in a quantifiable way. We proceed as it follows:

- Select a pool of languages from the dataset.
- For each language, concatenate the numbers in a single string (e.g., for English, the resulting string is: onetwothreefourfivesixseveneightnineten).
- Map the latter string to a DNA-like string according to the predefined mapping letter→DNA codon.
-Make global sequence alignment between pairs of sequences, and derive the distance between pairs of languages.
- Construct the distance matrix between all the languages included in the pool.
- Derive its phylogenetic tree by using a standard bioinformatics tool.

In what follows we show the phylogenetic tree for two cases. In the first case, we have chosen a pool of 14 languages, all of them of an indoeuropean nature following the state of the art. We have chosen these languages since their roots are well established, so that we can validate our approach. Note that, despite the fact that our datasets (strings) are very small, the phylogenetic tree is on good agreement with the state of the art. In particular, it differentiates the latins in a subgroup (whose internal order is correct as well), the anglosaxons in another subgroup, and the north european in another subgroup.



While these preliminary results are promising, we have gone one step further, and included in the pool another two languages (Basque and Inuit) that are known to be outliers: they do not belong to the indoeuropean tree. Interestingly, Basque language is spoken in a small region of Spain and has consequently co-evolved with other indoeuropean languages, however the state of the art places it outside that tree. Below we plot the resulting phylogenetic tree. As it can be noticed, the method situates both Basque and Inuit outside the indoeuropan tree, as expected.



Concluding remarks

Despite the fact that the dataset for each language is very small (only ten words per language) and that no phonetic or feature-based properties have been encoded in the mapping of each letter to a codon, the obtained phylogenetic trees are surprisingly correct. A possible justification suggests that within a language, the numbers (concretely, the first ten numbers) have a fundamental role, and they are very representative of the associated societal culture. In this sense, it could be interesting to analyze the evolution and spreading of culture (by means of trading, for instance) in terms of such phylogenetic trees. This concepts should be analyzed in detail in further work. A refinement of the mapping criteria, an extension to different alphabets, and the calculation of other phylogenetic trees will also be at the core of future research.

Acknowledgments

The authors acknowledge professor D.E. Smith (SFI) for an interesting debate on linguistics, Dan Rockmore and all the people that made possible SFI-CSSS for such an interesting opportunity, and the support from SFI and NSF.