Modeling sequence evolution with kernel methods

Margherita Bresco · Marco Turchi · Tijl De Bie · Nello Cristianini

Published online: 24 July 2007 © Springer Science+Business Media, LLC 2007

Abstract We model the evolution of biological and linguistic sequences by comparing their statistical properties. This comparison is performed by means of efficiently computable kernel functions, that take two sequences as an input and return a measure of statistical similarity between them. We show how the use of such kernels allows to reconstruct the phylogenetic trees of primates based on the mitochondrial DNA (mtDNA) of existing animals, and the phylogenetic tree of Indo-European and other languages based on sample documents from existing languages.

Kernel methods provide a convenient framework for many pattern analysis tasks, and recent advances have been focused on efficient methods for sequence comparison and analysis. While a large toolbox of algorithms has been developed to analyze data by using kernels, in this paper we demonstrate their use in combination with standard phylogenetic reconstruction algorithms and visualization methods.

M. Bresco
Department of Mathematics and Informatics, University of Salerno, Salerno, Italy
e-mail: mbresco@unisa.it

M. Turchi
Department of Information Engineering, University of Siena, Siena, Italy
e-mail: turchi@dii.unisi.it

T. De Bie
ECS, ISIS Research Group, University of Southampton, Southampton, UK
e-mail: tdb@ecs.soton.ac.uk

N. Cristianini
Department of Statistics, University of California, Davis, USA
e-mail: nello@wald.ucdavis.edu
1 Introduction

Biological sequences evolve by slowly accumulating single mutations, but also by occasionally undergoing major rearrangements, such as duplications, inversions or transpositions of entire subsequences. A similarity measure between two biological sequences should ideally consider all such factors, as well as being computationally scalable. Often sequence alignment methods are used to estimate phylogenetic distance between two sequences, implicitly assuming no major rearrangements affected the two sequences. An alternative (and approximate) approach would be to compare all sequence segments of appropriate length, to see how many of them are in common, regardless of their position. If the length has been chosen to be suitably large, one can rule out matches due to chance, and estimate in this way sequence similarity. A sequence could then be represented by an indicator function over the set of all strings of length $p$, and two sequences could be declared similar if they have similar ‘feature functions’. In other words, we could embed DNA sequences (having a 4-symbol alphabet) in a $4^p$-dimensional space and use the distance in that space as a measure of dissimilarity.

In the same way, languages evolve over long times by slowly changing the spelling of words, by incorporating new words from other languages, as well as ‘internally’ producing new words, e.g. by combining existing ones. The remarkable stability of certain words over time has been recognized for centuries, with the Indo-European hypothesis being based on the similarity of this core set of words across many languages [15]. However, there are many other fairly invariant properties in a language. For example, simple statistical properties of languages have been known since the early days of information theory [17]. They are certainly used now in language classification systems. The (approximate) identity of many long strings (of the length of one or several words) between two documents of the same language might be used to infer semantic similarity, whereas different frequencies of very short sequences have been used to classify documents as belonging to different languages [3, 10].

In this paper we investigate how simple statistical properties of sequences (biological or linguistic) can be used to infer their similarity and possibly also their evolutionary relations. We will see how languages belonging to the same linguistic family have very similar statistical properties, as well as genomes from related species. We will use these statistical properties to embed the sequences into a vector space, to obtain their pairwise distances and hypothesize an evolutionary tree. All this is done by using kernel methods, exploiting recent techniques for the efficient computation of statistical sequence similarity.

Although a more sophisticated analysis could be possible after constructing the kernel matrix [18], we limit our examples to the reconstruction of phylogenetic trees and to the plotting of Multi Dimensional Scaling graphs. Many other analytic approaches can be directly applied, as detailed in [18].

2 Methods

In order to track the evolution of sequences, we will use standard tools from pattern analysis and from phylogenetic analysis. Our strategy will be to embed all the