Generation of hierarchically correlated multivariate symbolic sequences

With an application to the assessment of bootstrap confidence in phylogenetic analysis

M. Tumminello\textsuperscript{1,}\textsuperscript{a}, F. Lillo\textsuperscript{1,}\textsuperscript{2}, and R.N. Mantegna\textsuperscript{1}

\textsuperscript{1} Dipartimento di Fisica e Tecnologie Relative, Università di Palermo, Viale delle Scienze, 90128 Palermo, Italy
\textsuperscript{2} Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Received 11 February 2008 / Received in final form 28 April 2008
Published online 18 June 2008 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2008

Abstract. We introduce a method to generate multivariate series of symbols from a finite alphabet with a given hierarchical structure of similarities based on the Hamming distance. The target hierarchical structure of similarities is arbitrary, for instance the one obtained by some hierarchical clustering method applied to an empirical matrix of similarities. The method that we present here is based on a generating mechanism that does not make use of mutation rate, which is widely used in phylogenetic analysis. Here we use the proposed simulation method to investigate the relationship between the bootstrap value associated with a node of a phylogeny and the probability of finding that node in the true phylogeny. The results of this analysis are compared with those obtained in the literature according to an evolutionary model with a per-symbol constant mutation rate. We observe that the relationship between the bootstrap value of a node and the probability of the corresponding clade being correct is sensitive to both the length of data series and the length of the branch connecting the node to its closest ancestor in the phylogenetic tree, whereas such a relationship is only slightly affected by the topology of the true phylogeny and by the absolute value of similarity.

PACS. 89.75.-k Complex systems – 02.50.Sk Multivariate analysis – 02.10.Ox Combinatorics; graph theory

1 Introduction

Symbolic sequences are investigated in many different fields, including information theory, biological sequence analysis, linguistics, chaotic time series, and communication theory. A lot of efforts have been devoted to devise methods for generating univariate or multivariate sequences with given statistical properties [1–5]. Since pair correlations are often used to describe the dependence between variables, the problem of generating symbolic sequences with given pair correlation properties is of particular interest. Many methods have been proposed for generating symbolic sequences with given univariate correlation structure, e.g. given autocorrelation, and to generate symbolic sequences with given multivariate correlation structure, e.g. given cross correlation among pair of sequences [6–8]. In this second case one wants to generate multivariate sequences of symbols according to some given properties of pair similarities. In this paper we propose a method for generating multivariate sequences with a given similarity structure of hierarchical nature. This method is inspired by the Hierarchically Nested Factor Model (HNFM) recently introduced by us [9] to generate hierarchically organized multivariate sequences of continuously distributed random variables. The applications of the method here proposed are manifold. For example, in phylogenetic analysis the characteristics of the investigated species are coded in discrete (symbolic) variables, such as nucleotides, amino acids, discrete characters, and phylogenetic methods give a phylogenetic tree as an output. A phylogenetic tree is a tree describing the evolutionary relationship among different biological species (the leaves of the tree) that are assumed to have some common ancestors (the nodes of the tree). The structure of a phylogenetic tree is the one of a hierarchical tree, and our method gives the possibility of simulating sequences with such a structure without making any assumption on the evolutionary dynamics of the system.

As a specific application of the generation method, in this paper we consider a common problem in phylogenetic analysis, specifically the assessment through bootstrap analysis of the statistical confidence of a phylogenetic tree. Phylogeny is the study of evolutionary relations among different elements (for example, organisms or languages). There are many different methods to reconstruct a phylogenetic tree from a set of data. One of the key problems in phylogenetic analysis is the assessment of the accuracy of a given tree feature (e.g. a node or an internal branch).
Since a statistical theory of the errors of a phylogenetic method is usually difficult to achieve, a common approach to assess the accuracy of the features of a phylogenetic tree is bootstrap analysis [10]. By sampling with replacements the data matrix and by applying the tree reconstruction method to each bootstrap replica, one can obtain a confidence value of a feature by computing the bootstrap value, which is the fraction of replica trees sharing the considered feature with the original tree. A description of bootstrap applied to phylogenetic analysis is reported in Appendix A. In a seminal paper, Hillis and Bull [11] showed that the bootstrap value is an underestimation of the probability of inferring the correct feature for bootstrap values larger than 40%. By using computer simulations of evolution dynamics of sequences they showed, for example, that bootstrap values “of ≥70% usually correspond to a probability of ≥95% that the corresponding clade is real” [11]. The result of Hillis and Bull is based on a generic evolutionary model with a per-symbol constant mutation rate. While in molecular evolution this seems to be a natural starting model, in other contexts, such as language, culture or technology evolution, mutation rate and dynamical models based on it might be more vague concepts. Since our generation method is independent of any dynamical assumption, we believe it may be well suited for application in these contexts. In this paper we apply our generation method to the assessment of bootstrap confidence in phylogenetic analysis. We perform a simulation analysis similar to the one presented by Hillis and Bull in reference [11] but using our generation method. Similarly to them we find that the bootstrap value underestimates the probability that a clade inferred from sample data belongs to the true phylogeny. The paper is organized as follows. In Section 2 we present our method for generating multivariate symbolic sequences with a given hierarchical structure of similarities based on the Hamming distance. In Section 3 we present the application of the method to the assessment of bootstrap value as a measure of confidence. Section 4 concludes.

2 Method for generating hierarchically organized multivariate symbolic sequences

In this section, we introduce a method allowing to simulate multivariate series of symbols from a finite alphabet. The objective is to generate symbolic sequences with a hierarchical structure of similarities between the elements of the system. This structure may correspond, for instance, to the one revealed by a hierarchical clustering method that has been applied to an empirical matrix of similarities based on the Hamming distance. The hierarchical structure of similarities is almost arbitrary and constraints are discussed at the end of this section.

Let \( X \) be a set of series of symbols from a finite alphabet \( A = \{a_1, ..., a_p\} \). We indicate the length of each series with \( T \) and we assume that the number of series in the set is \( N \). Let us arrange the data \( X \) in such a way that each column of \( X \) corresponds to a specific series. According to the Hamming distance we define the similarity of elements \( i \) and \( j \) as

\[
s(i, j) = \frac{1}{T} \sum_{k=1}^{T} \delta(x_{ki}, x_{kj}),
\]

where \( \delta(x_{ki}, x_{kj}) = 1 \) if \( x_{ki} = x_{kj} \) and 0 otherwise. It is easy to show the following properties of \( s(i, j) \):

\[
s(i, i) = 1
\]

\[
s(i, j) \leq \frac{1}{T} \sum_{k=1}^{T} 1 = 1
\]

\[
s(i, j) \geq \frac{1}{T} \sum_{k=1}^{T} 0 = 0.
\]

These properties show that \( s(i, j) \) assumes rational values in the closed interval [0, 1]. Furthermore, it can be shown that \( s(i, j) \) is the result of a scalar product. Indeed each symbol \( a_i \) of the alphabet can be mapped into a vector of length \( p \) with all the components equal to zero but the \( i \)th component being equal to 1. Any series \( x_k \) of length \( T \) can therefore be mapped into a vector \( \tilde{x}_k \) of length \( T \cdot p \) by substituting symbols in the series with the corresponding binary mapping. We can rewrite equation (1) in terms of series \( \tilde{x}_k \) as

\[
s(i, j) = \frac{1}{T} \sum_{k=1}^{T} \tilde{x}_{ki} \cdot \tilde{x}_{kj} = \frac{1}{T} \tilde{x}_i \cdot \tilde{x}_j.
\]

The properties described in equations (2-5) imply that the matrix \( S \) of similarities \( s(i, j) \) can be interpreted as a correlation matrix, because (i) it is positive definite as the result of scalar product of equation (5), (ii) its diagonal elements are equal to 1 and (iii) all the elements \( s(i, j) \) assume values in the range [0, 1]. The latter condition indicates that similarities are described only in terms of positive numbers according to the Hamming distance. By applying a hierarchical clustering method to the matrix \( S \) of elements \( s(i, j) \) of equation (1) one obtains a filtered similarity matrix \( S^S \) and a rooted tree [12]. A rooted tree is a tree in which a special node (the root) is singled out. This node is labeled \( \alpha_1 \) in the illustrative example of Figure 1. We distinguish between leaves and internal nodes of the rooted tree. Specifically, vertices of degree 1 represent leaves (vertices labeled 1, 2, ..., 10 in Fig. 1) while vertices of degree greater than 1 are internal nodes (vertices labeled \( \alpha_1, \alpha_2, ..., \alpha_9 \) in Fig. 1). We also say that an internal node \( w \) is the parent of the node \( v \), and we use the notation \( w = g(v) \), if \( w \) immediately precedes \( v \) on the path from the root to \( v \). For example it is \( \alpha_2 = g(\alpha_7) \) in Figure 1. Analogously we say that an internal node \( w \) is a son of the node \( v \) if \( v \) is the parent of \( w \), i.e. \( v = g(w) \). In the example above \( \alpha_7 \) is the son of node \( \alpha_2 \). Beside the topological structure, rooted trees obtained through standard hierarchical clustering methods applied to a matrix of similarities based on the Hamming distance have also metric properties. In fact, clustering methods associate a