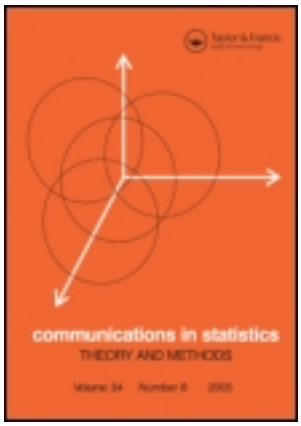

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

A Note on Distinguishing Random Trees Populations

Ana Georgina Flesia [a]
[a] CIEM-CONICET, FaMAF-UNC, Ciudad Universitaria, Córdoba, Argentina
Version of record first published: 07 Dec 2012.

To cite this article: Ana Georgina Flesia (2013): A Note on Distinguishing Random Trees Populations, Communications in Statistics - Theory and Methods, 42:2, 239-251

To link to this article: http://dx.doi.org/10.1080/03610926.2011.579371

Communications in Statistics—Theory and Methods, 42: 239–251, 2013

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2011.579371

Taylor & Francis
Taylor & Francis Group

A Note on Distinguishing Random Trees Populations

ANA GEORGINA FLESIA

CIEM-CONICET, FaMAF-UNC, Ciudad Universitaria, Córdoba, Argentina

This article addresses the problem of identifying differences between populations of trees. Recently, a sophisticated test was proposed by Busch et al. (2009), the BFFS test, a Kolmogorov type of test that maximizes the differences between the information of the samples, but it does not have a naive computation, since it involves a search over the set of trees that grows exponentially fast. An algorithm for computing the test statistic was devised in Busch et al. (2009), considering a search for a minimum cut over a transport network in a Ford Fulkerson type routine. The test was shown powerful but complex at the time to apply it in practice. On the contrary, we propose a very simple statistical test based on the distance between empirical mean trees, as an analog of the two sample Z statistic for comparing two means. Despite its simplicity, we can report that the test is quite powerful to separate distributions with different means, but it does not distinguish between different populations with the same means. In that case, the BFFS test should be applied. Nevertheless, on a real data set from proteomics, also discussed on Busch et al. (2009), our test obtained the same results, making it a valuable preliminary evaluation tool for random trees population discrimination.

Keywords Protein functionality; Random trees.

Mathematics Subject Classification 62H30; 68T05; 92B15; 92C55.

1. Introduction

Random trees have long been an important modeling tool. Tree models arise naturally when a collection of observed objects are all descended from a common ancestor via a process of duplication followed by gradual differentiation. There are two broad approaches to constructing random evolutionary trees in this setting: forwards in time "branching process" models, such as the Galton-Watson process, and backwards-in-time "coalescent" models such as Kingman's coalescent introduced in Kingman (1982). But there are other trees in the statistics and computer science literature, like *phylogenetic* or *evolutionary trees*, *probabilistic trees*, *search trees*, *tries*, and others, all of them with particularities related to the field of applications from whom have arisen. Some important references, among others, for these constructions are Holmes (2003) for the classical approach to phylogeny,

Received August 3, 2009; Accepted March 30, 2011
Address correspondence to Ana Georgina Flesia, CIEM-CONICET, FaMAF-UNC, Ciudad Universitaria, Córdoba, Argentina; E-mail: flesia@mate.uncor.edu

Sturmfels and Patcher (2005) for the new ideas of algebraic statistics, and Devroye (1998) for the classical computer science approach to probabilistic trees, tries, and its relationships with branching processes.

In this article, we consider trees that have a root and evolve forward in time in discrete generations, and each parent node has up to m offspring nodes in the next generation, as in Busch et al. (2009) (BFFS from now on). Otter (1949) and Neveu (1986) defined a tree as a subset of the nodes satisfying the condition "son present implies father present," the natural sigma algebra is the minimal one containing cylinders, sets of trees defined by the presence/absence of a finite number of nodes. The natural topology is the one generated by the cylinders as open sets. Under distances associated to this topology the space of trees is a compact metric space. In this context, BFFS prove law of large numbers for empiric samples of trees and an invariance principle on the space of continuous functions defined on the space of trees. In many cases, binary search trees, tries, and other probabilistic trees can be embedded into this set-up; see, for instance, Devroye (1998).

Let v, v^* be distributions that give mass only to finite trees. The goal is to test differences between the population laws

$$H_0 : v = v^* \quad H_A : v \neq v^* \tag{1}$$

using i.i.d. random samples with distribution v and v^*, respectively. Intuitively, if the expected mean of each population is different, a naive test for this problem will reject the null hypothesis when the distance between the empirical means associated with each sample is large enough, but it will fail if the population have different laws but the same expected mean. A Kolmogorov-type of test have been devised for this problem in Busch et al. (2009), but a direct approach to calculate effectively the test statistic is quite difficult, since it is based on a supremo defined over the space of all trees, which grows exponentially fast. In Busch et al. (2009), an algorithm for computing the BFFS test was introduced, in the context of multiple discrimination of proteins into families. This algorithm searches for a minimum cut over a network, using a Ford Fulkerson type of routine. In this article, we will report a good performance of the naive test over the same protein discrimination problem worked out on Busch et al. (2009), the problem of checking the coherence of hypothesized functionality families. We suppose that each family of proteins is related to a random tree, and the alleged members of each family form a sample of the law of the random tree that characterizes the family. We check if there is enough information in the samples to reject the hypothesis of equal populations.

In addition, we studied the naive distance-based test over simulations of Galton Watson processes, concluding that the power of the test is smaller than the BFFS test for specific cases but its computational simplicity calls for its application as a preliminary approach to the problem.

This article is organized as follows. Section 2 describes the class of metric spaces we consider. Section 3 develops the details of the simulations and the real data example from genomics. The conclusions of the article are given in Sec. 4

2. Trees, Distances and Tests

We will review the definition of BFFS tree that can be roughly thought of as a set of nodes satisfying the condition "son present implies father present." Following Busch et al. (2009), we will consider an alphabet $\mathcal{A} = \{1, \ldots, m\}$, with $m \geq 2$ integer,

representing the maximum number of children of a given node of the tree. Let $V = \{1, 11, 21, \ldots, m1, 111, 211 \ldots\}$, the set of finite sequences of elements in $\mathcal{A}$, all of them finishing with the symbol 1, which represents the *root of the tree*. The *full tree* is the oriented graph $t_f = (V, E)$ with edges $E \subset V \times V$ given by $E = \{(v, av) : v \in V, a \in \mathcal{A}\}$, where av is the sequence obtained by concatenation of v and a. In the full tree, each node (vertex) has exactly m outgoing edges (to its offsprings) and one ingoing edge (from her father), except for the root who has only outgoing edges. The node $v = a_k \ldots a_2 1$ is said to belong to the *generation* k; in this case, we write $\text{gen}(v) = k$. Generation 1 has only one node, the root.

We define a *BFFS tree* as a function $t : V \to \{0, 1\}$ satisfying

$$t(v) \geq t(av) \tag{2}$$

for all $v \in V$ and $a \in A$. Abusing notation, a tree t is identified with the subgraph of the full tree $t = (V_t, E_t)$ with

$$V_t = \{v \in V : t(v) = 1\} \quad \text{and} \quad E_t = \{(v, av) \in E : t(v) = t(av) = 1\}. \tag{3}$$

In Fig. 1, we observe a BFFS tree of depth 4. With this type of notation, the father of a node is written as a suffix in the description of the son, as it is often done in the definition of a Variable Length Markov Chain.

Let $\mathcal{T}$ be the set of all trees, and let $\phi : V \to \mathbb{R}^+$ be a strictly positive function such that $\sum_{v \in V} \phi(v) < \infty$. We define a distance between two trees in $\mathcal{T}$ as a weighted sum over the nodes that are present in a tree and absent in the other, following the formula

$$d(t, y) = \sum_{v \in V} \phi(v) |t(v) - y(v)|. \tag{4}$$

The natural sigma algebra is the minimal one containing cylinders, sets of trees defined by the presence/absence of a finite number of nodes. The natural topology

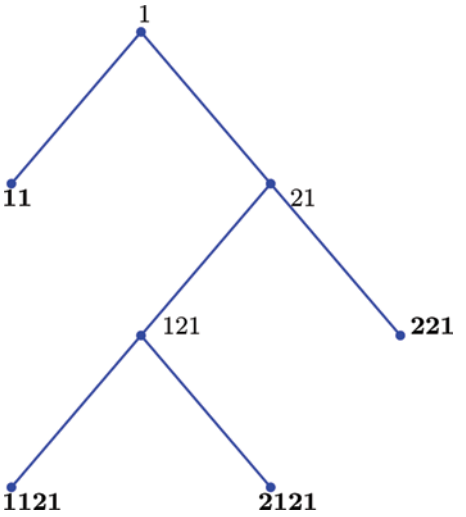

Figure 1. An example of BFFS tree of depth 4. The leaves are written in boldface. (color figure available online)

is the one generated by the cylinders as open sets. So it is easy to prove that the distance d we defined before generates the natural topology, and $(\mathcal{T}, d)$ becomes a compact metric space; see Busch et al. (2009). We denote $\mathcal{B}$ the σ-field of Borel subsets of $\mathcal{T}$, induced by the metric d.

Random trees. A *random tree* with distribution v is a measurable function

$$T : \Omega \to \mathcal{T} \quad \text{such that } \mathcal{P}(T \in A) = \int_A v(dt) \tag{5}$$

for any Borel set $A \in \mathcal{B}$, where $(\Omega, \mathcal{F}, \mathcal{P})$ is a probability space and v a probability on $(\mathcal{T}, \mathcal{B})$.

Expected value. The expected value or d-mean of a random tree T is the set (of trees) $\mathbb{E}_d T$ which minimizes the expected distance to T:

$$\mathbb{E}_d T := \arg\min_{t \in \mathcal{T}} \int_{\mathcal{T}} d(t, y) v(dy). \tag{6}$$

The set $\mathbb{E}_d T$ is not empty; see Busch et al. (2009). Any element of the set $\mathbb{E}_d T$ is also called a d-mean or d-center. Since $\mathbb{E}_d T$ depends only on the distribution v induced by T on $\mathcal{T}$, it may also be denoted as $\mathbb{E}_d(v)$.

Empiric mean trees. Let $\mathbf{T} = (T_1, \ldots, T_n)$ be a random sample of T (independent random trees with the same law as T). The empiric mean tree (empiric d-center, sample d-mean) is defined as the random set of trees given by

$$\overline{\mathbf{T}} := \arg\min_{t \in \mathcal{T}} \frac{1}{n} \sum_{i=1}^{n} d(T_i, t). \tag{7}$$

This formula may show the problem as more difficult that it is, since it is calling for a search over the whole set of trees, that grows exponentially in the number of nodes. But it is easy to prove that the empiric mean tree of a set of trees can be built by majority vote over the nodes. That means, that at least one of them can be defined as the tree whose nodes are present only if they are present in at least half of the sample.

Proposition 2.1. *Let $\mathbf{T} = (T_1, \ldots, T_n)$ be a random sample of T , and let t^* be the tree defined as the tree whose nodes are present only if they are present in at least half of the sample. Then t^* is an empiric mean tree.*

Proof. Let first notice that if $t \in \mathcal{T}$

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} d(T_i, t) &= \frac{1}{n} \sum_{i=1}^{n} \sum_{v \in V} \phi(v) |T_i(v) - t(v)| \\
&= \sum_{v \in V_t} \phi(v) \frac{1}{n} \sum_{i=1}^{n} |T_i(v) - t(v)| + \sum_{v \in \cup V_{T_i} / V_t} \phi(v) \frac{1}{n} \sum_{i=1}^{n} |T_i(v) - t(v)| \\
&= \sum_{v \in V_t} \phi(v) \frac{\#\text{ trees in the sample } v \text{ is not present}}{n} \\
&\quad + \sum_{v \in \cup V_{T_i} / V_t} \phi(v) \frac{\#\text{ trees in the sample } v \text{ is present}}{n}.
\end{aligned}$$

So, to reduce the average of distances we have to reduce both summands, keeping and adding nodes to the candidates of empiric means. The first point to notice is that the first summand is reduced when the candidate t keeps nodes that are present in many trees of the sample. If t keeps a node that is not in any tree of the sample, the first summand adds the full value of $\phi(v)$. The second summand is reduced when the tree t do not keep a node that is present only in a few trees of the sample. The cut off that balance the presence-absence relationship for each node is then $1/2$.

Remark 2.1. We should notice that if the number of trees in the sample is odd, the empiric mean is unique, but if the sample size is even, the node that it is present in exactly a half of the sample can be kept or not, without increasing the distance, so we will have at least two empiric means, one will have the least of possible present nodes, and the other the most.

Remark 2.2. In Wang and Marron (2007), BFFS trees are considered as structural trees, and attributes have been added to each node, in order to model more complex data, like human blood vessel systems. These real-life tree-like objects needs a mathematical model with richer structure, but as the authors pointed out, the simplicity of the tree model is lost, and the generalization of the measure introduced on the space did not allow an easy computation of the mean tree. The same problem was pointed out by Banks and Constantine (1998), considering binary labeled trees. Hamming type of metrics, like the BFFS metric have the capability of easy computation and should be the choice when the problem allows it. On spaces of binary labeled trees, like phylogenetic trees, a popular choice of metric is the Robinson Foulds metric, but the mean tree is not longer computable without a search over the entire (finite) space. Consensus trees have been proposed, but the idea of a unique consensus tree when merging different databases of trees has been challenged, (Bryant, 2003), and a statistical test of differences of populations is still needed; see Stockham et al. (2002) for a data mining approach to the problem based on clustering.

Example 2.1 (Galton-Watson Related Population of Trees)**.** We consider now only binary (not labeled) trees, that is, $m = 2$. The extension to an arbitrary number of offsprings m is straightforward. In a binary binomial Galton-Watson model, the offspring number is $0, 1$, or 2 with probabilities $(1-p)^2$, $2p(1-p)$, and p^2. The expected mean tree keeps a node v, if and only if $\text{gen}(v) \leq k_0$, where $k_0 = \max\{k \in \{0, 1 \dots\} : p^k \geq 1/2\}$. When $p < 1/2$, the expected mean tree is the empty tree. For instance, if $p = 0.5$ and $p^* = 0.75$, the expected mean trees are $T_p = \{1\}$ and $T_{p^*} = \{1, 11, 12\}$, the full trees of depth 1 and 2, respectively, but for $p \in [0.5, 0.70]$ the population have the same expected mean tree. This is a very simple parametric case where the maximum likelihood test has maximum power, so it is not of much use to introduce a new test in this setting, *if we knew* that we have a Galton Watson process producing our observations. We consider this example only to asses the power of the proposed test via simulation.

Example 2.2 (Variable Length Markov Chains and Related Population of Trees)**.** A Variable Length Markov Chain is a stochastic process introduced first by Rissanen (1983) in the setting of information theory, and that have been recalled lately by Bühlmann and Wyner (1999), and many others in the context of Protein Functionality Modeling; see Bryant (2003) and references therein.

In this model the probability of occurrence of each symbol at a given time depends on a finite number of precedent symbols. The number of relevant precedent symbols may be variable and depends on each specific sub-sequence. More precisely, a VLMC is a stochastic process $(X_n)_{n\in\mathbb{Z}}$, with values on a finite alphabet $\mathcal{A}$, such that

$$P[X_n = \cdot \mid X_{-\infty}^{n-1} = x_{-\infty}^{n-1}] = P[X_n = \cdot \mid X_{n-k}^{n-1} = x_{n-k}^{n-1}], \tag{8}$$

where x_s^r represents the sequence $x_s, x_{s+1}, \ldots, x_r$ and k is a stopping time that depends on the sequence $x_{n-k}, \ldots, x_{n-1}$. As the process is homogeneous, the relevant past sequences $(x_{n-k}, \ldots, x_{n-1})$ do not depend on n and are called *contexts*, and denoted by $(x_{-k}, \ldots, x_{-1})$. The set of all contexts τ can be represented as a rooted tree t, where each complete path from the leaves to the root in t represents a context. Calling p the transition probabilities associated to each context in τ given by (8), the pair (τ, p), called *probabilistic context tree*, has all information relevant to the model; see Rissanen (1983) and Bühlmann and Wyner (1999).

As an example, take a binary alphabet $\mathcal{A} = \{1, 2\}$ and transition probabilities

$$P[X_n = x_n \mid X_{-\infty}^{n-1} = x_{-\infty}^{n-1}] = \begin{cases} P[X_n = 1 \mid X_{n-2}^{n-1} = 1\ 1] = 0.7, \\ P[X_n = 1 \mid X_{n-2}^{n-1} = 2\ 1] = 0.4, \\ P[X_n = 1 \mid X_{n-1} = 2] = 0.2, \end{cases} \tag{9}$$

so that, if $x_{n-1} = 2$, then the stopping time $k = 1$ and $X_n = 1$ with probability 0.2; otherwise, the stopping time is $k = 2$ and $X_n = 1$ with probability 0.7 if both $x_{n-1} = x_{n-2} = 1$ or with probability 0.4 if $x_{n-1} = 1$ and $x_{n-2} = 2$. The set of contexts is $\tau = \{111, 211, 21\}$, when the set of all active nodes of the associated rooted tree t is $V_t = \{1, 111, 211, 21, 1\}$, since 11 is an internal node in the path of the context 111 and 211, and 1 is the root. Another example over the same alphabet is given by the transition probabilities

$$P[Y_n = y_n \mid Y_{-\infty}^{n-1} = y_{-\infty}^{n-1}] = \begin{cases} P[Y_n = 1 \mid Y_{n-1} = 1] = 0.6, \\ P[Y_n = 1 \mid Y_{n-2}^{n-1} = 2\ 2] = 0.4, \\ P[Y_n = 1 \mid Y_{n-2}^{n-1} = 1\ 2] = 0.2. \end{cases} \tag{10}$$

The set of contexts is $\eta = \{11, 121, 221\}$, when the set of all active nodes of the rooted tree y is $V_y = \{1, 11, 121, 21, 221\}$, since 21 is an internal node in the path of the context 12 and 22. The corresponding rooted trees t and y are represented in Fig. 2. Let us compute the distance between the these two trees,

$$\begin{aligned} d(t, y) &= \sum_{v\in V} \phi(v)|t(v) - y(v)| \\ &= \phi(1)|t(1) - y(1)| + \phi(11)|t(11) - y(11)| + \phi(21)|t(21) - y(21)| \\ &\quad + \phi(111)|t(111) - y(111)| + \phi(12)|t(121) - y(121)| \\ &\quad + \phi(211)|t(211) - y(211)| + \phi(221)|t(221) - y(221)| \\ &= 0 + 0 + 0 + \phi(111) + \phi(121) + \phi(211) + \phi(221) \\ &= 4 \times 0.36^3 = 0.186624, \end{aligned}$$

considering $\phi(v) = z^{gen(v)}$, $z = 0.36$.

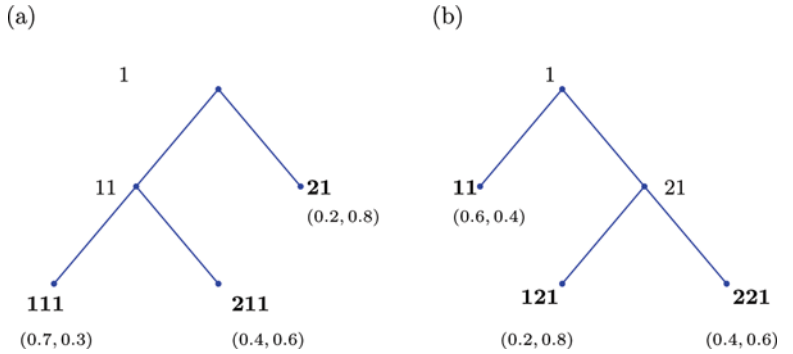

Figure 2. An example of two probabilistic context trees over the alphabet $A = \{1, 2\}$. (a) The tree t represents the pair (τ, p), where $\tau = \{111, 211, 21\}$ is the set of contexts and p are the transition probabilities given by (9). (b) The tree y represents the pair (η, q), where $\eta = \{121, 221, 11\}$ is the set of contexts and q are the transition probabilities given by (10). (color figure available online)

Now, let us suppose that we are given a sequence of symbols that have been produced by a VLMC with an unknown context tree. There are several algorithms that estimates the context tree associated to the chain using the sequence as an input. Let us fix the rule of estimation, for example, the Probabilistic Suffix Trees algorithm (PST) from Bejerano (2004). This rule is a *random tree* that generate trees in $\mathcal{T}$ following a given probability distribution v that is associated to the chain. If we have two independent samples of strings that have been hypothetically produced by two different unknown chains, we would like to derive a test that will rule if there is evidence in the samples to support that hypothesis. We should stress the fact that we are not using the probability transitions but the structure of the estimated context trees to derive the test, but we are not losing much information since such structure is indirectly related to the probability of occurrence of all possible contexts.

Testing differences of populations. We consider measures $v \in \mathcal{Q}_f$, the space of probability measures that concentrate mass on trees with a finite number of nodes. We describe the two-sample problem.

Let v, v^* be distributions in $\mathcal{Q}_f$. The goal is to test

$$H_0 : v = v^* \quad H_A : v \neq v^* \tag{11}$$

using i.i.d. random samples $\mathbf{T} = (T_1, \ldots, T_n)$ and $\mathbf{T}^* = (T_1^*, \ldots, T_m^*)$ with distribution v and v^*, respectively.

Test Based on the Distance Between Mean Trees. When the expected d-means are different, $\mathbb{E}T \neq \mathbb{E}T^*$, one expects that the distance between the empirical mean trees $\overline{\mathbf{T}}, \overline{\mathbf{T}}^*$ will be positive, for functions ϕ which do not penalize too much the first generations, as $\phi(v) = z^{\text{gen}(v)}$ with $0 < z < \frac{1}{m}$. A simple and naive test for this problem will reject the null hypothesis when the distance between the empirical means associated with each sample is large enough.

Computation. The lack of knowledge of the distribution of the distance between empirical means may be overcame using Monte Carlo randomization. If the null

hypothesis is $v = v^*$, and

$$d_k = d(\overline{\mathbf{T}}, \overline{\mathbf{T}}^*) = \sum_{v \in V} |\overline{\mathbf{T}}(v) - \overline{\mathbf{T}}^*(v)| \phi(v)$$

is the empiric distance between the mean trees of the kth pair of simulated sample, created by randomly rearranging the whole set of observations, and assigning the first n_1 observations to the first sample and the rest to the second sample, we define the quantile q_α as the value such that

$$\alpha = P(d(\overline{\mathbf{T}}, \overline{\mathbf{T}}^*) > q_\alpha).$$

This value can be approximated using the order statistics $d^{(1)}, \ldots, d^{(N)}$ and taking q_α as $d^{([N(1-\alpha)])}$ (here $[a]$ denotes the greatest integer not greater than a). For the original samples $\mathbf{T}$ and $\mathbf{T}^*$, the test will reject the hypothesis if $d(\overline{\mathbf{T}}, \overline{\mathbf{T}}^*) > q_\alpha$ at level α. The Type 2 error can be estimated analogously for each alternative hypothesis v_a.

3. Computational Examples

Simulation. To study the performance of the tests on a controlled environment we simulate several populations of trees using Galton-Watson processes and simple variations of it. We carefully choose the parameters to challenge the power of the tests.

Assume we have two random samples, each one from a Galton-Watson process with possibly different parameters p and p^*, denoted $GP(p)$ and $GP(p^*)$. We we would like to test if these samples come from the same process, that is,

$$H_0 : T \sim GP(p), T^* \sim GP(p^*) \quad p = p^* \qquad H_A : T \sim GP(p), T^* \sim GP(p^*), \quad p \neq p^*$$

In our simulation we already know the parameters of the underlying distributions v and v^*. Thus, we have performed a Monte Carlo simulation test sampling trees from a mixture of both laws at random, until we reach the size of the first sample and label it sample 1. Then continue selecting with the same mixture, until we reach the size of the second sample, and label it sample 2. We compute the test statistics with these random samples, and store it, and repeat the process 1,000 times. Then we generate a fixed number of times a sample from the distribution v, and a sample from the distribution v^*, and calculate the test statistics with them. If the true test statistic is greater than (1-α)% of the random values, then the null hypothesis is rejected at $p < \alpha$. The percentage of rejections for each value of α is considered a measure of the power of the test.

We computed the percentage of rejection over 1,000 tests of level $\alpha = 0.10, 0.05, 0.01$, when $T^*_{k,1}, \ldots, T^*_{k,n}$ is $GP(p^*)$, with $p^* = 0.6, 0.75, 0.8$, and 0.85, for sample sizes $n = 31, 51, 101, 151$, and 201. The results are reported on Table 1.

These results are in agreement with our intuitive ideas. As the sample size increases, the test is not able to reject the hypothesis of equal populations when $p = 0.5$ and $p^* = 0.6$, since their expected mean trees are equal. But when the expected mean trees are different, the test detects the difference with higher power as the sample size increases.

Table 1
Percentage of rejections over 1,000 tests, computed with with $p = 0.5$ and $p^* = 0.6$, 0.75, 0.8, 0.85, sample size $n = 31$, 51, 101, 151, and 201

$\alpha = 0.1$	$n = 31$	$n = 51$	$n = 101$	$n = 151$	$n = 201$
$p = 0.6$	5.6	2.1	0	0	0
$p = 0.75$	52.8	65	92.5	99.3	100
$p = 0.8$	86.2	93.7	99.9	100	100
$p = 0.85$	99	99.9	100	100	100
$\alpha = 0.05$	$n = 31$	$n = 51$	$n = 101$	$n = 151$	$n = 201$
$p = 0.6$	5.60	02.1	0	0	0
$p = 0.75$	52.80	47	92.5	99.3	100
$p = 0.8$	78.60	93.7	94.7	100	100
$p = 0.85$	97.80	99.8	100	100	100
$\alpha = 0.01$	$n = 31$	$n = 51$	$n = 101$	$n = 151$	$n = 201$
$p = 0.6$	0.70	2.10	0	0	0
$p = 0.75$	39.10	47.00	58.40	95.10	96.9
$p = 0.8$	51.70	76.90	94.70	95.10	96.2
$p = 0.85$	55.40	98.40	100	100	100

Variable Length Markov Chain Modeling of Protein Functionality. A central problem in computational biology is to determine the function of a new discovered protein using the information contained in its amino acid sequence. Proteins are complex molecules composed by small blocks called amino acids. The amino acids are linearly linked, forming a specific sequence for each protein. There exist 20 different amino acids represented by a one-letter code.

There are several problems related to protein functionality, but we will only point out two of them here. One is the classification of the function of a new protein with the help of a training set, and the other is clustering a group of new and known proteins into meaningful functionality families. The goal of clustering protein sequences is to get a biologically meaningful partitioning. Genome projects are generating enormous amounts of sequence data that need to be effectively analyzed. Given to the amount of available data, and the lack of proper definition, clustering is a very difficult task, so there is a need for ways of checking the validity of the partition proposed. As most databases are created by sequence alignment related methods, an impartial way of checking validity would be to apply an alignment-free, model-based methodology.

Most methods for clustering and classification need as input a similarity matrix, usually computed by sequence alignment. Model-based clustering and classification without sequence alignment is leaded by Markov Chain modeling. For example, Bejerano and Yona (2001) modeled protein sequences with stationary Variable Length Markov Chains (VLMC) in order to classify a new given protein as belonging to the family whose model has higher probability of having produced that string. This approach needs also a reliable training set in order to build an accurate estimate of the unknown context tree of the chain.

In this section, we propose checking the coherence of of selected protein families performing a simultaneous hypothesis test, as in Busch et al. (2009). Some partial results were included in Flesia and Freiman (2007).

We would like to test if several families that are members of a well known database are simultaneously significantly different. The Pfam database is known to be a good reference for protein functionality clustering, so it would provide a benchmark for assessing the performance of our approach.

We start modeling each functionality family of proteins as realizations of an unknown VLMC. But instead of learning the model using all the sequences of a given family to estimate the context tree with the Probabilistic Suffix Trees algorithm (PST), we consider this rule as a *random tree* that generate one tree in $\mathcal{T}$ per sequence. The probability distribution ν of the random tree is associated to the chain that rules the family in an unknown fashion. If we have two independent samples of strings that have been hypothetically produced by two different unknown chains, we estimate with each of them the context tree of its chain and then consider we have two independent samples of trees, each one following a distribution associated to the family. We then test if there is enough evidence in the samples to reject the hypothesis of equal distribution. If we do reject the hypothesis, we consider the two families significantly different.

Let $\mathcal{T}_3$ be the space of trees with $m = 20$ possible children per node (the symbols of the amino acid alphabet), and fixed maximum length $M = 3$ and the parameter of the distance fixed as $z = 0.36$. We test if families selected from de P-FAM database (Bateman et al., 2004), are simultaneously significantly different using the following two-step procedure.

(1) Transform the amino acid chains into trees via the Probabilistic Suffix Trees (PST) from Bejerano (2004), obtaining 10 samples of trees of maximum depth 3.
(2) Apply a Bonferroni correction to the 45 pairwise BFFS-based comparisons, meaning each test is performed with a level of significance of $\alpha = 0.05/45 = 0.001$ to get a simultaneous comparison of the 10 families, with overall level $\alpha = 0.05$.

We run all the pairwise tests at level 0.001. We also run the tests under the null hypothesis splitting each data set at random in two subsets. Table 2 shows the critical and the observed values for all pairwise tests of different families (non diagonal terms). For the null hypotheses the observed value and the p-value appear in boldface at the diagonal. Despite the crude nature of the Bonferroni method, the hypothesis of equal distribution is rejected in all cases when the samples came from different populations, confirming the coherence of the selected protein families. In the case of the same family split in halves, we can observe p-values ranging from 0.12–0.90, values that can be used also to analyze the coherence of the family.

4. Final Remarks

We proposed a naive test to compare two populations of trees with laws that do not have the same expected mean. The procedure is very simple, since it is based on the idea that the empiric mean tree of each sample, a strong consistent estimator of the expectation of the law that generates each population, should be separated in terms of BFFS distance, a Hamming type of distance with easy computation. The test will reject the hypothesis of equal populations if the distance between the empiric means

Table 2
Critical value and observed value of 45 pairwise comparisons at level $\alpha = 0.001$. Test rejects when the observed value is greater than the critical value. In boldface, observed value and p-value when testing the same population, $N = 1,000$. The distance's parameter zeta is equal to 0.36

Family	actin	adh-short	adh-zinc	ank	ATP-synt-A
actin	**(0.49, 0.71)**	(2.41, 9.85)	(3.37, 10.44)	(3.47, 9.84)	(4.62, 11.21)
adh-short		**(1.36, 0.43)**	(1.91, 4.91)	(2.44, 5.55)	(2.05, 5.27)
adh-zinc			**(1.66, 0.57)**	(2.58, 5.30)	(3.05, 6.70)
ank				**(1.86, 0.81)**	(4.27, 8.37)
ATP-synt-A					**(1.67 0.52)**
Family	beta-lactamase	cox2	cpn10	DNA-pol	efhand
actin	(3.76, 9.71)	(4.04, 11.46)	(5.52, 12.20)	(4.01, 9.73)	(3.06, 11.79)
adh-short	(2.52, 3.91)	(2.14, 6.24)	(2.42, 6.20)	(3.61, 6.44)	(1.86, 6.07)
adh-zinc	(2.64, 5.14)	(2.51, 7.38)	(2.79, 7.91)	(2.58, 5.90)	(2.23, 6.93)
ank	(3.14, 5.32)	(3.03, 7.94)	(5.04, 10.32)	(2.51, 3.16)	(2.74, 8.79)
ATP-synt-A	(3.34, 6.25)	(2.75, 4.95)	(2.61, 5.28)	(4.88, 8.98)	(2.08, 6.27)
beta-lactamase	**(1.819, 0.93)**	(2.98, 6.94)	(2.83, 6.52)	(3.16, 6.58)	(2.74, 6.49)
cox2		**(2.05, 0.09)**	(2.95, 6.99)	(3.77, 9.19)	(1.67, 6.49)
cpn10			**(1.30, 0.02)**	(6.46, 11.86)	(2.02, 3.86)
DNA-pol				**(1.81, 0.23)**	(3.58, 10.24)
efhand					**(0.65, 0.93)**

is big enough to ensure a small Type 1 error. The quantile of the distribution was derived by Monte Carlo randomization, and the power was studied through Galton Watson simulations.

In addition, we addressed a problem of functional genomics, to check the coherence of hypothesized functionality families. We suppose that each family of proteins is related to a random tree, and the alleged members of each family form a sample of the law of the random tree that characterizes the family. We check if there is enough information in the samples to reject the hypothesis of equal populations, using a Bonferroni simultaneous testing procedure. Summarizing, our framework is the following:

- each family $\mathcal{F}$ of protein domains induce a (different, hopefully) probability distribution ν on the space of trees $\mathcal{T}$;
- given two families $\mathcal{F}$ and $\mathcal{F}'$ we consider their associated signatures, i.e., the probability laws π and ν' on the space $\mathcal{T}$;
- for each family $\mathcal{F}_j$ we take a sample of protein sequences of size n_j, and for each sequence in the sample we construct the **pst** context tree estimator, as described in Bejerano (2004). We obtain a sample of size n_j of iid random elements on $\mathcal{T}$ with distribution π_j; and
- finally, for each pair of families $\mathcal{F}_j, \mathcal{F}_{j'}$ we test if both distributions π_j and $\pi_{j'}$ are the same.

This approach will not work if the two populations have the same expected mean tree, as in the case of two samples of strings that have been generated by

chains with the same context tree but different transition probabilities. A more sophisticated test was already proposed by Busch et al. (2009), a Kolmogorov type of test that maximizes the differences between the information of the samples, but it does not have a naive computation, since it involves a search over the set of trees that grows exponentially fast. An algorithm for computing the test statistics was devised, considering a search for a minimum cut over a transport network, and applying a Ford Fulkerson routine. It is not a naive approach at all, and obtains similar results as the ones shown in Sec. 3 on the protein discrimination problem. On the other hand, the power of the test, computed also over Galton Watson simulations, is higher and it may be applied even if the mean trees of the samples are the same; see Busch et al. (2009) for details.

This suggests applying this simple test as a preliminary approach, and if there is no rejection, implement the BFFS test.

Acknowledgments

We thank Antonio Galves for illuminating discussions about the discriminative power of context trees. We also thank Ricardo Fraiman for his continuos encouragement and support, and the Universidad de San Andres for its hospitality when part of this article was produced. We would like to thank Florencia Leonardi for providing the data used in our example of determination of protein functionality, which was also analyzed in Leonardi (2007). This work was partially supported by PICT 2005-31659 and Secyt grant 69/08 and 05/B352.

References

Balding, D., Ferrari, P., Fraiman, R., Sued, M. (2009). Limit theorems for sequences of random trees. *Test* 18:302–315.

Banks, D., Constantine, G. (1998). Metric models for random graphs. *J. Classifi.* 15:199–223.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., Eddy, S. R. (2004). The Pfam protein families database. *Nucl. Acids Res.* 32(90001):D138–141.

Bejerano, G. (2004). Algorithms for variable length Markov chain modeling. *Bioinformatics* 20(5):788–789.

Bejerano, G. (2003). Automata learning and stochastic modeling for biosequence analysis. PhD thesis, Hebrew University, Jerusalem, Israel.

Bejerano, G., Yona, G. (2001). Variations on probabilistic suffix trees: statistical modeling and prediction of protein families. *Bioinformatics* 17(1):23–43.

Bühlmann, P., Wyner, A. J. (1999). Variable length Markov chains. *Ann. Statist.* 27:480–513.

Busch, J. R., Ferrari, P., Flesia, A. G., Fraiman, R., Leonardi, F. (2009). Testing statistical hypothesis on random trees and applications to the protein classification problem. *Ann. Appl. Statist.* 3(2):542–563.

Bryant, D. (2003). A classification of consensus methods for phylogenies. In: Janowitz, M., Lapointe, F.-J., McMorris, F. R., Mirkin, B., Roberts, F. S., eds. *BioConsensus*. DIMACS. New York: AMS, pp. 163–184.

Devroye, L. (1998). Branching processes and their applications in the analysis of tree structures and tree algorithms. In: Habib, M., McDiarmid, C., Ramirez-Alfonsin, J., Reed, B., eds. *Probabilistic Methods for Algorithmic Discrete Mathematics*. Berlin: Springer-Verlag, pp. 249–314.

Enright, A., Van Dongen, S., Ouzounis, C. (2002). An efficient algorithm for large-scale detection of protein families. *Nucl. Acids Res.* 30(7):1575–1584.

Flesia, A. G., Freiman, R. (2007). A distance test on random trees. In: Elaskar, S., Pilotta, E., Torres, G., eds. *Proc. the I Congress On Computat. Indust. Appl. Mathe.* October 2–5, Córdoba, Argentina: Mecanica Computacional, XXVI:2016–2025.

Holmes, S. (2003). Bootstrapping phylogenetic trees: Theory and methods. *Statist. Sci.* 18:241–255.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* 13:235–248.

Leonardi, F. G. (2007). Parsimonious stochastic chains with application to classification and phylogeny of protein sequences. PhD thesis, Universidade de São Paulo, São Paulo, Brazil.

Neveu, J. (1986). Arbes et processus de Galton-Watson. *Ann. Inst. H. Poincare, Probabilities et Statistique* 22:199–207.

Otter, R. (1949). The multiplicative process. *Ann. Math. Stat.* 20:206–224.

Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Inform. Theor.* 29(5):656–664.

Sturmfels, B., Patcher, L. (2005). *Algebraic Statistics for Computational Biology*. Cambridge: Cambridge University Press.

Stockham, C., Wang, L., Warnow, T. (2002). Statistically based postprocessing of phylogenetic analysis by clustering. *Bioinformatics* 18:S285–S293.

Wang, H., Marron, J. S. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* 35:1849–1873.