Phylogenetic Trees via Hamming Distance Decomposition Tests

Cézar A. F. Anselmo & Aluísio Pinheiro
Departamento de Estatística
Universidade Estadual de Campinas - Brazil

Abstract

The paper considers the problem of Phylogenetic tree construction. Our approach to the the problem bases itself on a non-parametric paradigma seeking a model free construction and symmetry on Type I and II errors. Trees are constructed through sequential tests using Hamming distance dissimilarity measures, from internal nodes to the tips. The method has some advantages over the traditional methods. It is very fast, computationally efficient, and feasible to be used for very large datasets. Two other novelties are its capacity to deal directly with multiple sequences per group (and built its statistical properties upon this richer information) and that the *best* tree will not have a predetermined number of tips i.e. the resulting number of tips will be statistically meaningful. We applied the method in a sample of primate mitochondrial DNA sequences, illustrating that it can perform quite well even on very unbalanced design. Computational complexities are also addressed.
Keywords: Phylogenetic Tree; Hamming Distance; Dissimilarity Measures; Statistical Genetics; Non-parametric Test.

1 Introduction

In the last decades one has seen increasing interest and power of analyzing genetic data. Scientists are collecting genetic data in exponentially faster speed and synthetic measures are of great importance for either analyzing a single sample or for comparing different procedures or samples. One such measure is the phylogenetic tree.

A phylogenetic tree is a graph depicting the ancestor-descendant relationship between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree

connect the tips to their (unobservable) ancestral sequences (Hölder and Lewis(2003)). One can use such a representation to infer about temporal relationships between species or DNA sequences for questions such as: which species are more closely related; is there a common ancestor for two species; or can we group two species on the tip when facing a third species?

A survey of phylogenetic trees in both their biological aspects as well as statistical features is provided by Weir(1996). More recent works with comprehensive surveys are Salemi and Vandame (2003) and Hölder and Lewis (2003).

One can construct phylogenetic trees (for sequences or species) by several different methods. Some of the most referred on the literature are parsimony, maximum likelihood and distance matrices procedures.

Each different method has its own motivation which in one hand provides its strength but on the other hand burdens it with its weakness. They all do relly on a common feature: a single sequence for each group. That is attained by either a single specimen from each group or by the use of a *consensual* sequence as the legitimate representative of a group. Apart from the biological limitations of such dimension reduction we do worry about statistical properties of trees thereof provided.

Here, we develop a methodology that can be used for single specimens cases or multiple cases with unbalanced sample sizes. The only difference between those situations is that on the latter one will have more statistical power than on the former one.

Pinheiro et al (2003) has studied a decomposition of Hamming distance that quantifies the amount of diversity between individuals from the same pseudogroup and diversity within individuals of different groups. That decomposition is interesting from its interpretation and also because it enables one to employ U-statistics theory to prove solid statistical properties.

We propose a procedure that sequentially builds the tree from its internal nodes to its tips. The p-values from each test are computed via bootstrap resampling. Moreover, in each step the topology with larger *diversity separation* is used. We apply the methodology to a sample of primate mitochondrial DNA sequences.

The main advantages of this method reside on its flexibility, its statistical properties, its easeness of use, and its computational performance.

2 Phylogenetic Trees via Hamming Distance

The procedure we propose works for either genes or more complex data. Therefore, when referring to the differences between sequences we will use the general term *group* that should be understood in its context.

Aside from working exclusively with single sequences per group, the usual methods do not explicitly quantify dissimilarities between groups and dissimilarities within groups. The latter is not considered biologically relevant. This theoretical insignificance is numerically represented by a *consensual* sequence for each group. Therefore, measuring only the dissimilarities between groups without comparing them to the a priori neglected within groups dissimilarities generates procedures with a clear conceptual bias towards group separation.

This bias can be qualitatively put as follows. Suppose one is able to quantify meaningfully disimilarity and that there is a theoretical dissimilarity value d_0 (unknown) which is the smallest value for which separation of any two groups is biologically meaninfgul. Whenever the dissimilarity between two pseudo-groups can be exactly measured and it fails to be larger than d_0 one can not consider these two pseudo-groups different but their elements are part of a single group and their computed dissimilarity is only a measure of individual differences. Otherwise, if their dissimilarity is larger than d_0, one should consider them two legitimate groups.

Let d_L be the true theoretical between groups dissimilarity measure (unknown). Models which fail to comparatively evaluate within and between dissimilarities should favor grouping with d_L ($< d_0$) computed dissimilary with a higher probability than not separating groups whose computed dissimilarity is $d_0 + (d_0 - d_L)$. This lack of simmetry is called in Statistics a bias. The degree of bias will depend on the procedure which is taken and on the specificities of the sample and genomes being studied. The theoretical aspects of this bias are unknown and abstract (at least to our knowledge) but the conceptual risks of bias toward separation should be clear.

We use a non-parametric technique which tries to address the bias issue through a careful decomposition of dissimilarities via U-statistics theory. Moreover because we deal with nonparametrics one should expect distribution and model free results i.e. the less structured model implemented will give less power but on the other hand it will work with overall smaller Type I and Type II error probabilities under a larger family of distributions. Because we are able to extract information from each individual sequence the usual power issues with nonparametric techniques are lessened to a degree in which one has either really small p-values or really large ones. Therefore any doubt from a inferential point of view will be less dependent on the method itself and should be regarded as a characteristic of the problem.

The tree is constructed via recursive tests based on the U-statistics decomposition of Hamming distance for the sampled sequences, as proposed by Pinheiro et al. (2003). For the sake of completeness we will describe briefly the mathematical aspects of such a decomposition and its statistical implications.

Consider a general computational sequence analysis (CSA) with K sites, each one having 4 possible categories (Pinheiro et al. (2003) treats the general C categories case) in each site $k = 1, 2, \ldots, K$. Each category represents a nucleotide but that can be used for protein or codon sequences without any aditional notational burden. Let $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iK})$ be a random vector of responses where X_{ik} represents the categorical outcome c ($c = 1, 2, 3, 4$) at site k for the i-th sequence.

One defines the Hamming distance between a pair (i, i') of sequences as:

$$D_{ii'} = \frac{1}{K} \sum_{k=1}^{K} I\left(X_{ik} \neq X_{i'k}\right), \tag{1}$$

where $I(X_{ik} \neq X_{i'k})$ is zero if the respective k-th sites on the i-th and i'-th sequences are equal and one if they are different. So, $D_{ii'}$ is the proportion of sites where $\mathbf{X}_i$ and $\mathbf{X}_{i'}$ do not match.

Suppose one has G groups and each group has n_g sampled sequences, $g = 1, 2, \ldots, G$. Pinheiro et al. (2003) defines three Hamming distance related measures, $D_n(B)$, $D_n(W)$ and $\bar{D}_n(0)$), respectively by:

$$D_n(B) = \frac{1}{n-1} \left\{ \sum_{g=1}^{G} \sum_{g'=g+1}^{G} n_g n_{g'} \left(2\bar{D}_{gg'} - \bar{D}_{gg} - \bar{D}_{g'g'}\right) \right\} \tag{2}$$

$$D_n(W) = \sum_{g=1}^{G} \frac{n_g}{n} \bar{D}_{gg} \tag{3}$$

$$\bar{D}_n(0) = D_n(W) + D_n(B), \tag{4}$$

where $\bar{D}_{gg}$ is the average distance within the g-th group, and $\bar{D}_{gg'}$ is the average distance between groups g and g', where $g, g' = 1, 2, \ldots, G$ and $g' \neq g$.

$\bar{D}_{gg}$ is a U-statistic of degree 2 and $\bar{D}_{gg'}$ is a two sample U-statistic of degree $(1, 1)$. Some deterministic inequalities can then be employed to pursue tests of differences between groups, with solid asymptotic statistical results for the test statistic $D_n(B)$.

In applications, it is very hard to address the exact distribution of $D_n(B)$. It is theoretically possible to build the asymptotic distribution directly from some fairly complex functions of the data but those procedures are computationally expensive. There is also th issue of ensuring that the asymptotic approximation is good enough for the sampled data. The safer procedure is to employ bootstrap resamplig techniques and their empirical percentiles for decisions.

Suppose the phylogenetic tree is rooted and binary. Whenever two groups are considered statistically different there is only one possible tree. However, for more than

two possible groups, say G, a statistically significant difference does not guarantee that there are G different groups (the test will grasp any minute difference between any two groups as long as its power is good enough). Therefore for phylogenetic trees construction one will eventually perform some two groups tests. Moreover for more than two groups, say $\{1, 2, 3\}$, even after it is statistically inferred that there are three different groups one must decide among $((1), ((2), (3)))$, $((2), ((1), (3)))$, or $((3), ((1), (2)))$, using a *Newick*-like notation.

The construction of phylogenetic trees goes as follows. Suppose one has G possible groups for which one wants to build a three with at most G different groups. We will exemplify the construction with $G = 4$ but its theoretical aspects are not any different when G gets large albeit its computational complexity may preclude us from proceeding. There are two possible topologies (up to nominating the tips) with a total of seven different groupings.

In step 1 seven tests are performed. It is important to notice that, due to the statistical power of discrimination provided by the multiple sequences in each group, all tentative grouping present a statistically significant small value of $D_n(B)$. Therefore, an additional measure is taken into account to choose the most relevant separation. The ratio $D_n(B)/\bar{D}_n(0) = (1+D_n(W)/D_n(B))^{-1}$ works as follows. While $D_n(B)$ quantifies the overall difference between pair of groups, $D_n(W)$ measures those characteristics that single individuals within their respective groups. Therefore, is it natural to reject grouping for which the aforementioned ratio is small because the innergroup diversity which is considered tolerable and measured by $D_n(W)$ is much larger then the pseudo-groups diversity (given by $D_n(B)$). On the other hand, groups that have the largest ratio $D_n(B)/\bar{D}_n(0)$ provide the best separation among all possible groups configuration.

After step 1 is performed one has either a tree with two tips with two pseudo-groups each or a tree with one tip with one group and another with tree pseudo-groups. In the former case, one will perform the two remainig tests and the final tree will have two, three or four tips if, respetively, none of the $D_n(B)$'s in step 2 are statistically significant, exactly one of them is, or both are. In the latter situation, the three possible grouping of one against two pseudo groups will be performed. If all $D_n(B)$'s are statistically negligible one will have a final two tips tree. Otherwise, one chooses the path that maximizes the $D_n(B)/D_n(W)$ ratio, having a tree with two tips of one group each and a still unresolved tip with two pseudo-groups. Finally the last test for this tip is performed and if its $D_n(B)$ is statistically significant one has a four tip tree. Otherwise, one has a three tip tree.

The following algorithm summarizes the procedure:

(0.1) Let $\mathcal{G}$ be the set of all pseudo-groups

(0.2) Compute all distances D_{ij}, $\forall i, j \in \mathcal{G}_g$, $g = 1, 2, \ldots, G$, using equation (1), where $\mathcal{G}_g$ is the g-th pseudo-group to be tested

(0.3) Take the set of pseudo-groups $\mathcal{E} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G\}$ a partition of $\mathcal{G}$ and let $N = G$

(1) For $l = 1$ to $\lfloor N/2 \rfloor$

- for all possible partitions $\{\dot{\mathcal{E}}_1, \dot{\mathcal{E}}_2\}$ of $\mathcal{E}$, such that $\dot{\mathcal{E}}_1$ contains l pseudo-groups and $\dot{\mathcal{E}}_2$ contains the remaning $N - l$ pseudo-groups
  1. Compute $D_n(B)^{true}$ and $D_n(W)^{true}$ with equations (2) and (3)
  2. Repeat for $b = 1$ to B (the number of bootstraps)
     (a) Select randomly a sample of $|\mathcal{E}|$ sequences with replacement from $\mathcal{E}$
     (b) Take the first l sequences to be pseudo-observations from $\dot{\mathcal{E}}_1^b$ and the last $N - l$ to be from $\dot{\mathcal{E}}_2^b$
     (c) Compute $D_n(B)^b$ and $D_n(W)^b$ from formulas (2) and (3), and store its values

(2) Consider all $D_n(B)^{true}$ from **(1)** and let $\mathcal{S}$ be the set of all statistically significant ones, using the percentiles from the bootstraped $D_n(B)^b$, $b = 1, 2, \ldots, B$

1. If $\mathcal{S} = \emptyset$, there are no statistically significant groups and the procedure for $\mathcal{E}$ is over and $\mathcal{E}$ can not be divided any further
2. If $\mathcal{S} \neq \emptyset$, choose the partition of $\mathcal{E}$ as $\{\tilde{\mathcal{E}}_1, \tilde{\mathcal{E}}_2\}$ for which the ratio between $D_n(B)$ and $D_n(W)$ is maximum
3. Repeat the algorithm for $\mathcal{E} = \tilde{\mathcal{G}}_1$ and $\mathcal{E} = \tilde{\mathcal{G}}_2$, until there is not a group with two or more pseudo-groups untested

3 Application

The data sets consists of sequences of mitochondrial DNA from humans, chimpanzees, gorillas and orangutans with 438bp. The tests were performed with $B = 10000$ bootstraped resamples, taken from the original pooled sample. Four tables are presented. Table 1 shows the results of the proposed procedure with 96 sequences of which 58 are from humans, 6 chimpanzees, 15 gorilla and 7 orangutan. Chimps sequences add a special flavor because they have three sequences from the so-called isolate specimens. Unless otherwise mentioned all analysis were performed with the isolates and results without the isolates (not shown) led to the same qualitatively conclusions. For control purposes, Table 2 and 3 show the results for the Hamming distance procedure with different pseudo-groups as explained below. In each table, te most relevant numbers

(either the largest $D_n(B)/D_n(W)$ ratio in the step or all $D_n(B)$ p-values for the last possible step) are presented in bold face.

First we performed the tests procedure on the complete sample, starting with four possible groups: humans (H), chimpanzees (C), gorillas (G) and orangutans (O). Step 1 tests indicate that the best ratio is attained when humans are separated from the other primates. On a second step one expects the three primates to be separetd. Again, all three groupings are statistically significant and the aforementioned ratio selects the separation of the orangutans from the other two primates. Finally it remains to test whether it is feasible to separate chimpanzees and gorillas and that is confirmed by the computed $D_n(B)$. We should stress the fact that although a sequence of tests is taken, their respective p-values are so small (usually smaller than resolution enables one to measure), that the overall *fixed size* of the phylogenetic construction can be made as small as one desires. That means that the resulting structure is statistically sounding.

Table 1: $\{H, C, G, O\}$ Phylogenetic Tree

Grouping	p-value		Ratio
	$D_n(W)$	$D_n(B)$	
((H),(C,G,O))	.9946	.0000	**.4076**
((C),(H,G,O))	.6243	.0001	.0929
((G),(H,C,O))	.9716	.0000	.3154
((O),(H,C,G))	.9168	.0000	.2520
((H,C),(G,O))	.9909	.0000	.3687
((H,G),(C,O))	.8464	.0000	.2008
((H,O),(C,G))	.9800	.0000	.3358
Result after Step 1 - ((H),(C,G,O))			
((C),(G,O))	.9413	.0009	.2284
((G),(C,O))	.9966	.0000	.4147
((O),(C,G))	1.000	.0000	**.5584**
Result after Step 2 - ((H),((O),(C,G)))			
((C),(G))	.9975	**.0000**	.6129
Result after Step 3 - ((H),((O),((C),(G))))			

Table 2: $\{H_1, H_2, C, G\}$ and $\{H_1, H_2, C, O\}$ Phylogenetic Trees

Grouping	p-value $D_n(W)$	p-value $D_n(B)$	Ratio	Grouping	p-value $D_n(W)$	p-value $D_n(B)$	Ratio
$((H_1),(H_2,C,G))$	.5832	.0001	.0927	$((H_1),(H_2,C,O))$	.5069	.0090	.0467
$((H_2),(H_1,C,G))$	.6135	.0000	.1133	$((H_2),(H_1,C,O))$	.5372	.0021	.0672
((C),(H_1,H_2,G))	.6307	.0002	.1337	((C),(H_1,H_2,O))	.6764	.0002	.1771
((G),(H_1,H_2,C))	.9921	.0000	.4596	((O),(H_1,H_2,C))	.9333	.0000	**.4221**
((H_1,H_2),(C,G))	.9977	.0000	**.5031**	(H_1,H_2),(C,O))	.9132	.0000	3867
((H_1,C),(H_2,G))	.5240	.0043	.0500	(H_1,C),(H_2,O))	.4780	.0253	.0303
((H_1,G),(H_2,C))	.5402	.0017	.0612	(H_1,O),(H_2,C))	.4931	.0146	.0391
Step 1 - ((H_1,H_2),(C,G))				Step 1 - ((H_1,H_2,C),(O))			
((H_1),(H_2))	.4377	**.1932**	.0015	((H_1),(H_2,C))	.4587	.0642	.0197
((C),(G))	.9979	**.0000**	.6129	((H_2),(H_1,C))	.4735	.0292	.0335
				((H_1,H_2),(C))	.6909	.0001	**.3508**
Step 2 - ((H_1,H_2),((C),(G)))				Step 2 - (((H_1,H_2),(C)),(O))			
				((H_1),(H_2))	.4334	**.1828**	.0015
				Step 3 - (((H_1,H_2),(C)),(O))			

In order to illustrate that the main application is not simply a spurious numerical artifact we present also the p-values for the $D_n(W)$ and they are (as they should be) all negligible. The other *diagnostic* performed was the use of the same method for three groupings other than groups of single species. The human group was randomly divided in two groups of 29 sequences each, called H_1 and H_2. Table 2 shows the analysis for two pseudo-group constructions - $\{H_1, H_2, C, G\}$ and $\{H_1, H_2, C, O\}$. Table 3 shows the analysis for $\{H_1, H_2, G, O\}$ and $\{H_1, H_2, C, G, O\}$. Tables 2 and 3 have as a common feature that the proposed procedure is able to decrease the four (or five) pseudo-groups to a three (or four) tips phylogenetic tree. All the chosen groups are formed by single species and the five pseudo-groups procedure results in the same as in Table 1.

Table 4 shows some problems in using a single sequence per group. Phylogenetic trees are build from single sequences from each species for three groups. When C, G and O are considered the resulting tree always puts chimpanzees and gorillas apart from orangutans. However, for the remaining three cases results will strongly depend on the sequence chosen.

4 Computational Complexity

Computational performance of a phylogenetic tree construction method can be divided in two main parts: the number of trees *considered* in the analysis and *what* will be done with each of these trees. Since each consensual sequence method will have a very specific approach on this action on a single tree, a direct performance comparison can not be made. The numbers we will present in Table 5 are quite conservative towards MP and ML methods, i.e., the huge computational disadvantages those paradigmas have compared to the proposed method are quite understated. We compute an upper bound for our procedure and compare it to a lower bound for the MP and ML-based methods. One should notice that the complexity for Hamming distance are not average numbers, but the worst possible sequence of events. Moreover a linear complexity of estimation procedures is assumed for the MP and ML methods, which is quite optimistic for all but the very simplest models. Finally, the computational complexity for the Hamming distance procedure is a sum of the number of sequences and the product of a power of the number of pseudo-groups and the number of sites. The number of sequences is therefore only a memory burden but it is not otherwise worrysome in each step. On the other hand, the complexity fo MP(or ML) methods is a factorial on the number of sequences (not on the number of pseudo-groups).

Maximum parsimony methods have complexity not smaller than $(2G-3) = 1 \times 3 \times 5 \times \cdots \times (2G-3)$ for each site in consideration. Maximum likelihood methods complexity are even larger because besides the number of possible trees one deals with parameters estimation.

Our method has complexity not smaller than K, the number of sites. Since all two-sequences distances can be computed in advance, that will have a $G(G-1)/2$ complexity. Notice that the number of pseudo-groups is much smaller than the number of sequences, say n and that if one uses more than one sequence per group the complexity of MP or ML methods will be larger than $(2n-3)!!$. Some comparative figures are shown in Table 5, with lower bounds for MP and ML methods and upper bounds for the Hamming distance method.

For the application , the programs were run in C on a AMD Atlon 2100+ 1.73GHz computer. Procedures for 4 pseudo-groups, such as those in Table 2, would take 50 seconds while computations for 5 pseudo-groups, such as those in Table 6, would take 149 seconds. For instance, a 10 group program would possibly run in less than two hours on the same machine, with no major concern on the number of sequences for each group.

Table 3: $\{H_1, H_2, G, O\}$ and $\{H_1, H_2, C, G, O\}$ Phylogenetic Trees

Grouping	p-value $D_n(W)$	p-value $D_n(B)$	Ratio
((H$_1$),(H$_2$,G,O))	.5937	.0000	.0851
((H$_2$),(H$_1$,G,O))	.6378	.0000	.1064
((G),(H$_1$,H$_2$,O))	.9777	.0000	.3815
((O),(H$_1$,H$_2$,G))	.9166	.0000	.2908
((H$_1$,H$_2$),(G,O))	.9935	.0000	**.4547**
((H$_1$,G),(H$_2$,O))	.5701	.0013	.0650
((H$_1$,O),(H$_2$,G))	.5577	.0010	.0621
Step 1 - ((H$_1$,H$_2$),(G,O))			
((H$_1$),(H$_2$))	.4425	**.1924**	.0015
((G),(O))	.9999	**.0000**	.8112
Step 2 - ((H$_1$,H$_2$),((G),(O)))			

Grouping	p-value $D_n(W)$	p-value $D_n(B)$	Ratio
((H$_1$),(H$_2$,C,G,O))	.6383	.0000	.0886
((H$_2$),(H$_1$,C,G,O))	.6658	.0000	.1084
((C),(H$_1$,H$_2$,G,O))	.6344	.0001	.0928
((G),(H$_1$,H$_2$,C,O))	.9697	.0000	3154
((O),(H$_1$,H$_2$,C,G))	.9136	.0000	.2520
((H$_1$,H$_2$),(C,G,O))	.9962	.0000	**.4076**
((H$_1$,C),(H$_2$,G,O))	.5533	.0011	.0536
((H$_1$,G),(H$_2$,C,O))	.5328	.0018	0465
((H$_1$,O),(H$_2$,C,G))	.5749	.0010	.0647
((H$_2$,C),(H$_1$,G,O))	.5814	.0002	.0675
((H$_2$,G),(H$_1$,C,O))	.5256	.0024	.0466
((H$_2$,O),(H$_1$,C,G))	.5809	.0003	.0690
((C,G),(H$_1$,H$_2$,O))	.9785	.0000	3358
((C,O),(H$_1$,H$_2$,G))	.8497	.0000	.2008
((G,O),(H$_1$,H$_2$,C))	.9909	.0000	.3687
Step 1 - ((H$_1$,H$_2$),(C,G,O)			
((C),(G,O))	.9424	.0010	.2284
((G),(C,O))	.9978	.0000	.4147
((O),(C,G))	.9998	.0000	**.5584**
Step 2 - ((H$_1$,H$_2$),((C,G),(O))			
((H$_1$),(H$_2$))	.4500	**.1890**	.0015
((C),(G))	.9972	**.0000**	.6129
Step 3 - ((H$_1$,H$_2$),(((C),(G)),(O)))			

Table 4: Phylogenetic Tree with a single sequence from each species

Topology	with the 3 Isolates				without the Isolates			
	Samples				Samples			
	100	1000	10000	All	100	1000	10000	All
((H,C),G)	.6000	.6110	.5987	.5981	.3900	.2710	.2921	.2915
((H,G),C)	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
((C,G),H)	.4000	.3890	.4013	.4039	.6100	.7290	.7079	.7085
((H,C),O)	.9800	.9880	.9841	.9843	.9800	.9880	.9841	.9826
((H,O),C)	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
((C,O),H)	.0200	.0120	.0159	.0157	.0200	.0120	.0159	.0174
((H,G),O)	.9800	.9880	.9841	.9843				
((H,O),G)	.0000	.0000	.0000	.0000				
((C,G),H)	.0200	.0120	.0159	.0157				
((C,G),O)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
((C,O),G)	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
((G,O),C)	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Table 5: Computational Complexity of Phylogenetic Tree Construction

G (for Hamming) n (for ML or MP)	Worst Case for Hamming	Lower Estimate for ML or MP
2	1	1
3	4	3
4	11	15
5	26	105
10	1013	34459425
50	1.125×10^{15}	2.752×10^{76}
100	1.267×10^{30}	3.349×10^{184}
500	3.273×10^{150}	1.008×10^{1280}
1000	1.071×10^{301}	3.847×10^{2863}
10000	1.995×10^{3010}	1.601×10^{38663}

5 Discussion

We present an alternative method for the construction of phylogenetic trees. The method has some advantages over the traditional methods. For instance, since it uses Hamming distance and it is sequentially built through binary separations it is very fast and computationally efficient. It is based on non-parametric ideas what makes it less prone to model bias. Moreover its statistical properties can be asserted using its direct U-statistics representation. One can clearly relate the within and in-between dissimilarities to species and individual characteristics. Two other novelties are its capacity to deal with multiple (and different) sequences per group (and built its statistical properties upon this richer information) and that the *best* tree will not have a predetermined number of tips i.e. the resulting number of tips will be statistically meaningful. We applied the method in a sample of primate mitochondrial DNA sequences, illustrating that it can perform quite well even on very unbalanced design and that is feasible to be used for very large datasets.

6 References

Durret, R. (2002). *Probabilistic Models for DNA Sequence Evolution.* Springer-Verlag, New York.

Ewens, W. J. and Grant, G. R. (2001). *Statsitical Methods in Bioinformatics.* Springer-Verlag, New York.

Hölder, M. and Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Genetics* **4**:275-284.

Lee, A. J. (1990). *U-Statistics - Theory and Practice.* Marcel Dekker, Inc., New York.

Pinheiro, H. P., Pinheiro, A. and Sen, P. K. (2003). Comparison of genomic sequences using Hamming distance. *Journal of Statistical Planning and Inference.* (in press.)

Salemi, M. and Vandame, A.-M. (2003). *The Phylogenetic Handbook: a Practical Approach to DNA and Protein Phylogeny.* Cambridge University Press, Cambridge, 406 pages.

Weir, B. (1996). *Statistical Analysis of Genetic Data II.* Sinnauer, MA.