

Building phylogenetic trees from binary oligonucleotide fingerprint vectors

A. Figueroa¹, Z. Liu¹, R. Mancini-Jones², J. Borneman², and T. Jiang¹

Abstract—Oligonucleotide Fingerprinting of Ribosomal RNA Genes (OFRG) is a method that permits the identification of microorganisms through ribosomal RNA gene (rDNA) analysis. OFRG sorts arrayed rDNA gene clones into clusters through a series of hybridization experiments, each using a single oligonucleotide probe. This series of hybridization experiments generates a "fingerprint" for each rDNA clone. The fingerprints are binary vectors that specify whether the probes hybridized or did not hybridize to the clones. Identification of the microorganisms is accomplished by clustering fingerprints from unidentified clones with those from identified clones. The most accurate taxonomic classifications from rDNA analysis are often obtained through complete nucleotide sequence analysis. However, the direct information that we acquire from OFRG is the presence or absence of a specific substring of nucleotides in the unidentified rDNA sequence. This paper provides several ways to associate information to the set of fingerprints obtained by OFRG.

Keywords—oligonucleotide fingerprinting, phylogenetic trees.

I. INTRODUCTION

Oligonucleotide Fingerprinting of Ribosomal RNA Genes (OFRG) is a method that permits the identification of ribosomal RNA genes (rDNA) using a DNA array [15], [16]. OFRG sorts arrayed rDNA gene clones into clusters through a series of hybridization experiments, each using a single oligonucleotide probe. This series of hybridization experiments generates a *fingerprint* for each rDNA clone. These fingerprints are binary vectors that specify whether the probes hybridized or did not hybridize to the clones. For example, a clone with fingerprint vector (0, 1, 1, 0) means that in 4 hybridization experiments, the second and third probes hybridized with the clone while the first and fourth did not. The set of probes used in these hybridization experiments is called a *probe set*. Given such fingerprints, a set of rDNA clones can be partitioned into clusters such that clones with a common fingerprint vector belong to the same cluster [6].

Ideally, nucleotide sequence analysis of rRNA gene extracted from a microorganism can lead to the identification of the microorganism. However, the direct information that we acquire from OFRG is the presence or absence of a specific substring of nucleotides in the unidentified rDNA sequence.

This paper provides different ways to associate information to the set of fingerprints. The types of information considered includes: 1) Are there some previously identified rDNA sequences that have common fingerprint vectors with those being analyzed? 2) Can we build a phylogenetic tree from

the fingerprints with a topology structure similar to the tree built from the corresponding set of rDNA sequences? If so, 3) How can we associate a DNA sequence or a taxonomic name to each hierarchical clustering level? These questions are addressed in this paper.

As an application, 163 fungal rDNA gene clones derived from soil are analyzed. The nucleotide sequences of the rDNA clones were obtained using the ABI PRISM [®]BigDyeTM Terminators v3.0 Cycle Sequencing Kit. A set of 22 probes were used as the probe set. These probes have been previously shown to be a good probe set to characterize fungal communities [15]. The DNA probes are 10 nucleotides in length.

Note that because we have the nucleotide sequence for each clone, we can create the fingerprints without doing any hybridization experiment. Given a clone sequence and a probe set, an *artificial fingerprint* for a clone is a binary vector of the same length as the number of probes in the probe set, where if the i^{th} probe appears as a substring in the clone sequence then a value of 1 is given at the i^{th} position in the vector, and 0 otherwise. From now on, we will only consider artificial fingerprints and refer to them simply as fingerprints. In the above application, 75 different fingerprint vectors were obtained from the 163 fungal rDNA clones, each set of clones with a common fingerprint vector form a cluster.

In Section II-A we show that phylogenetic trees built from fingerprint vectors are quite different from phylogenetic trees built from rDNA sequences. Then, Section II-B introduces an rDNA database as a reference useful for building better phylogenetic trees. Section III describes a method to label internal nodes in a phylogenetic tree. Some simulation results are presented in Section IV.

II. A MORE EFFECTIVE METHOD TO CONSTRUCT PHYLOGENETIC TREES FROM FINGERPRINT VECTORS

The tree of life theory suggests that all organisms on the Earth have a common ancestor. Thus any set of species is related, and this relationship is called a *phylogeny*. Usually the relationship can be represented by a *phylogenetic tree*. Many different methods have been suggested to reconstruct phylogenetic trees. These methods have been classified in general as distance methods (*e.g.* [11], [13]), maximum parsimony (*e.g.* [3]), and maximum likelihood (*e.g.* [4]).

Distance methods have shown to be effective for reconstructing phylogenetic trees from a set of DNA sequences. In these methods, a crucial step is to define the pairwise distance between two DNA sequences. A lot of work has been done on this topic (*e.g.* [5], [8], [9]). A pairwise distance between

1. Dept. of Comp. Sci., Univ. of Calif., Riverside. Contact author email: jiang@cs.ucr.edu.

2. Dept. of Plant Pathology, Univ. of Calif., Riverside.

two DNA sequences is commonly used as an estimate of the total branch length between them. The most common distance between two DNA sequences x and y is the (weighted) edit distance, defined as the minimum alignment value of x and y . From now on, $dist(x, y)$ will denote the edit distance between strings x and y . $dist(x, y)$ can be easily computed in time $O(|x||y|)$ [12].

A. Phylogenetic trees from fingerprint vectors

Given a set S of DNA sequences, define a *sequence tree* as the tree structure where each leaf in the tree contains a DNA sequence from S . Given a set F of fingerprint vectors, define a *fingerprint tree* as the tree structure where each leaf in the tree contains a fingerprint vector from F . If we assume that the above methods are good methods for building phylogenetic trees on DNA sequences, and F is the resulting set of fingerprint vectors for S under a given probe set, then our goal here is to build a phylogenetic tree using the fingerprint vectors such that the fingerprint tree is as much similar as possible to the sequence tree.

In [15], [16], fingerprint trees were built using a distance method called unweighted pair group method using arithmetic averages (UPGMA) and the hamming distance between two fingerprint vectors as the distance metric. However, as it is shown in this paper, fingerprint trees and sequence trees may have only a few structures in common. Table I shows tree-to-tree distances between the fingerprint tree and the sequence tree for the 163 fungal rDNA clones. Two popular methods were used to build the trees, UPGMA and Neighbor-joining. The software package PAUP: Phylogenetic Analysis Using Parsimony [14] offers implementations for both methods. Metrics dI and d are from [7]. dI indicates the number of leaves that have to be pruned from both trees to obtain a common substructure. Metric d incorporates information on the distances between the pruned leaves on the original trees. Metric $SYMDIFF$ is from [10] and defined as the number of edges (links) in the tree that do not appear in the other tree. Note that for two identical trees, the distance between them is zero under any of the above metrics.

TABLE I
TREE-TO-TREE DISTANCES BETWEEN THE SEQUENCE TREE AND
FINGERPRINT TREE FOR A SET OF 163 FUNGAL CLONES.

	dI	d	$SYMDIFF$
UPGMA	124	124.0365	270
NJ	121	121.0337	268

As we can see in Table I, out of 163 leaves, more than 120 leaves have to be removed from each tree to have a common substructure. Hence, the sequence tree and the fingerprint tree are not similar to each other. On average, they have less than 25% structure in common.

A simplified version of these trees are trees built from clusters rather than clones. Define a *fingerprint cluster tree* as the tree structure where each leaf in the tree contains a different fingerprint vector, *i.e.*, each leaf represents a cluster of clone sequences with a common fingerprint vector. In a similar way, define a *sequence cluster tree* as a tree structure where each leaf in the tree is labelled by a DNA sequence taken from

DNA sequences that have a common fingerprint vector. Such sequence labels are representative sequences associated with each cluster.

In this paper, we choose a representative sequence for each cluster as the sequence whose total edit distance to the other sequences within the same cluster is the minimum.

Table II shows tree-to-tree distances between the sequence cluster tree and the fingerprint cluster tree for 75 clusters in 163 fungal rDNA clones.

TABLE II
TREE-TO-TREE DISTANCES BETWEEN THE SEQUENCE CLUSTER TREE AND
FINGERPRINT CLUSTER TREE FOR 75 CLUSTERS OF 163 FUNGAL CLONES.

	dI	d	$SYMDIFF$
UPGMA	56	56.0717	124
NJ	54	54.0772	120

As we can see in Table II, out of 75 leaves, more than 50 leaves have to be removed from both trees to have a common substructure. This means that less than 33% of the structure is common to both trees. Again, even for trees built from clusters, the fingerprint cluster trees and sequence cluster trees are far from each other.

In the next section we will show how we can enhance the fingerprint tree with the help of a reference database so it looks more similar to the sequence tree than those built in [15], [16].

B. More accurate phylogenetic trees using a reference database

In the previous section, we showed that trees built from fingerprint vectors can be very different from the trees built from DNA sequences. In this section, we will show how we can improve the construction of fingerprint trees with the help of a reference database so the tree looks much similar to the sequence tree.

To start, we collected a set of fungal rDNA sequences from GeneBank (NCBI) as a reference database to our clones. We call this database the *fungal rDNA database*. The following describes how we built the database.

First, we retrieved all 18s fungal genes reported in GeneBank. We call this set the *Raw data set*. Second, we looked for sequences in the Raw data set such that: 1) they contain two rDNA primers (left, TTAGCATGGAATAATR-RAATAGGA and right, TGGGATAGRGCATTGCAAT) that were used for selectively amplify fungal rDNA clones from a sample of interest, and 2) the number of bases within these primers is between 650 and 820. We call this set of sequences the *PostRaw data set*. Then, from each sequence in the PostRaw data set we took the substring limited by the primers and put it into a new set. We call this set the *Pattern data set*. Sequences in the Pattern data set will form part of the fungal rDNA database. However, many sequences similar to those in the Pattern data set were undetected in this procedure due to the fact that some sequences in the Raw data set did not contain the primers. To recover such sequences, we used the sequence similarity searching program BLAST [1]. We run BLAST on sequences in the Raw data set against those in the Pattern data set. By that, we were able to identify many

of such sequences. We call this new set the *Homo data set*. Finally, the fungal rDNA database is the union of Pattern and Homo data sets. A total of 1732 fungal rDNA sequences form the fungal rDNA database.

In the next part of this section we describe how we could make use of the fungal rDNA database to build a phylogenetic tree from a set of fingerprint vectors.

Definition 2.1: A clone is an identified clone if there is a sequence in rDNA database such that under the same probe set, both the clone and the sequence have a common fingerprint vector. Such a sequence is called an identifier sequence.

Definition 2.2: An identified cluster is a set of identified clones with a common fingerprint vector.

104 clones out of 163 were identified with the fungal rDNA database. these identified clones are from 33 clusters out of the 75 clusters. Note that for each identified cluster there is at least one identifier sequence associate with that cluster. We used these identifier sequences to build a tree for the identified clones.

Following the same idea described in the previous section, we look for a representative sequence for each identified cluster among all identifier sequences in the cluster. Then we use these representative sequences to build a tree. We call this tree the *identifier cluster tree*. Each leaf in the identifier cluster tree represents a cluster. We will show here that the identifier cluster tree is significantly more similar to the sequence cluster tree than the fingerprint cluster trees. Table III shows tree-to-tree distances between the sequence cluster tree against the identifier cluster tree and the fingerprint cluster tree for 33 clusters of 104 fungal clones.

TABLE III

TREE-TO-TREE DISTANCES BETWEEN THE SEQUENCE CLUSTER TREE AGAINST THE IDENTIFIER CLUSTER TREE AND THE FINGERPRINT CLUSTER TREE FOR 33 CLUSTERS OF 104 FUNGAL CLONES.

	IDENTIFIER CLUSTER TREE			FINGERPRINT CLUSTER TREE		
	d1	d	SYMDIFF	d1	d	SYMDIFF
UPGMA	10	10.1576	38	23	23.1304	48
NJ	9	9.1477	36	19	19.1356	42

As we can see in Table III the identifier cluster tree is considerably more similar to the sequence cluster tree than the fingerprint cluster tree. On average, they have close to 78% of the structure in common.

Note that this introduces the problem of how to include clones not identified by the fungal rDNA database in the identifier cluster tree.

III. LABELLING NODES IN A PHYLOGENETIC TREE

In Section II, we showed how to reconstruct a phylogenetic tree using fingerprint vectors and an rDNA database. We also described a way to find a representative DNA sequence per cluster. Such representative can be used as a DNA label for the cluster that it belongs to. This section describes a method to label nodes of such a tree with taxonomic names.

A. Labelling nodes in a phylogenetic tree with taxonomic names

Taxonomy is the science of classifying organisms into categories or taxa and includes such procedures as identifying and naming. This classification system starts with the most general

category name, the kingdom, and ends with the most specific name, the species. The taxonomic system in general includes seven categories, which increase in specificity: Kingdom, Phylum, Class, Order, Family, Genus, and Species. Each level of identification relates to common physiological traits and evolutionary ancestry. Let's index the categories from 1 to 7 starting from the lowest specificity, Species, and ending with the highest specificity, Kingdom.

Recall that, each leaf in the identifier cluster tree represents an identified cluster, and for each identified cluster there is a set of identifier sequences associated with it. We propose here to use the taxonomy information of these identifier sequences to label internal nodes in the identifier cluster tree. First, we add 7 more fields per record in the fungal rDNA database, one for each category. The taxonomic names for each category and for each identifier sequence were retrieved from GeneBank (NCBI)

For an identified cluster tree T and an internal node v in T , let $I(v)$ be the set of all identifier sequences associated with the leaves in T_v and $C_i(I(v))$ the set of names of category i associated with $I(v)$. Then, node v is labelled with category i if category i is the lowest specificity category such that $|C_i(I(v))| = 1$; otherwise, no label is given to v . The labelling by taxonomic name algorithm is described in Figure 1.

Algorithm: (T)
for each level of T , with the bottom level first, **do**
for each node v at the level **do**
for $i = 1$ to 7
if $|C_i(I(v))| = 1$ **then**
 $s(v) \leftarrow \text{category}(i)$; **break**;
end

Fig. 1. Algorithm for labelling internal nodes in a phylogenetic tree with taxonomic names.

Figure 2 shows a subtree of the identifier cluster tree for 104 fungal rDNA clones. This subtree contains 8 identified clusters, where for example, identifier sequences in clusters A_B01 and A_D04 have a common genus name: *Microascus*, and identifier sequences in clusters A_C01, A_H05, and B_D01 have a common class name: *Glomeromycetes*.

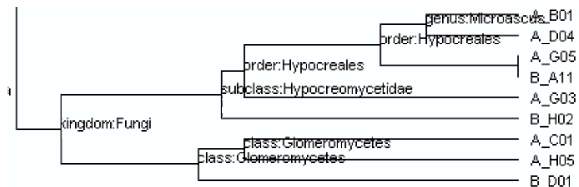


Fig. 2. Subtree of the identifier cluster tree for 104 fungal rDNA clones. 8 identified clusters are shown, where internal nodes are labelled with taxonomic names.

IV. SIMULATION RESULTS

We have tested our approach for reconstructing phylogenetic trees from fingerprint vectors with a reference database by some simulations. We describe first a simulation test.

We used a bacterial rDNA database for our tests. This database was created in a similar way as the fungal rDNA database described in Section II-B. The sequences in the

bacterial rDNA database satisfy the following conditions: 1) they are 16s bacterial genes, 2) they contain two rDNA primers (left, AGRRTTTGATYHTGGYTACG and right, AAGTCG-TAACAAAGGTAVC), and 3) the number of bases within these primers is between 1350 and 1550. For the purpose of simulations, duplicates in the bacterial rDNA database were eliminated. A total of 8572 sequences form the bacterial rDNA database.

Each test consists of a random subset of 100 sequences taken from the bacterial rDNA database. This subset was considered as the set of clones for building a phylogenetic tree and the remaining set, 8472 sequences, was considered as the reference database. For each test, a probe set of 27 probes of 10 nucleotides each were designed using the algorithm in [2].

Table IV reports tree-to-tree distances between the sequence cluster tree and the identifier cluster tree or the fingerprint cluster tree on 4 different tests. The first column contains the number of identified clusters per test, the second column contains the method used to build the trees, then two main columns show the tree-to-tree distances. The first main column contains the distances between the sequence cluster tree and the identifier cluster tree, and the second main column contains the distances between the sequence cluster tree and the fingerprint cluster tree. Table IV clearly shows that the identifier cluster tree is considerably more similar to the sequence cluster tree than the fingerprint cluster trees.

TABLE IV

TREE-TO-TREE DISTANCES BETWEEN THE SEQUENCE CLUSTER TREE AND THE IDENTIFIER CLUSTER TREE OR THE FINGERPRINT CLUSTER TREE ON SIMULATED DATA.

# IC	Method	IDENTIFIER CLUSTER TREE			FINGERPRINT CLUSTER TREE		
		d1	d	SYMDIFF	d1	d	SYMDIFF
74	UPGMA	24	24.074	74	49	49.070	120
	NJ	26	26.069	70	46	46.067	120
64	UPGMA	13	13.067	26	43	43.079	86
	NJ	23	23.078	46	36	36.084	86
73	UPGMA	24	24.065	74	54	54.070	106
	NJ	26	26.070	74	54	54.070	112
67	UPGMA	15	15.072	52	48	48.070	110
	NJ	19	19.072	58	45	45.078	110

We have also tested our method for labelling phylogenetic trees with taxonomic names by some simulations. These simulations were created in the same way as described above. In this case, we tested the accuracy of the taxonomic names by comparing the taxonomic names given by our approach with the true taxonomic names of the sequences.

Table V shows the accuracy ratio of labelled clusters on 3 different tests. The first column contains the total number of labelled clusters, then the accuracy ratio of the labelled clusters per category name is shown in columns 2 to 6. For example, in the first test a total of 55 clusters were labelled, where 17 of them have a common Genus name and 16 out of these 17 were properly right labelled by our approach.

V. CONCLUDING REMARKS

We have proposed an effective method to reconstruct phylogenetic trees from fingerprint vectors with the help of a reference database. However, as it was pointed out above, this approach is limited by the reference database. In other words,

TABLE V
ACCURACY RATIO OF LABELLED CLUSTER BY TAXONOMIC NAME ON SIMULATED DATA.

Total	Category Name					
	Kingdom	Phylum	Class	Order	Family	Genus
55	1/1	5/5	9/9	11/13	9/10	16/17
62	0/0	4/4	8/8	8/11	9/11	17/18
59	1/1	6/6	9/9	9/10	9/9	24/24

this approach only provides a way to build phylogenetic trees on identified clones. It would be interesting to incorporate in the tree clones that were not identified by the reference database. It would also be interesting to modify the algorithm of Section III-A to allow some name mismatches within a cluster. This might allow us to increase the numbers of the category names with high specificity and decrease the numbers of the category names with low specificity.

ACKNOWLEDGMENT

The research was partially supported by a UC MEXUS/CONACYT doctoral fellowship and NSF Grant CCR-9988353 to A.F., NSF Grant DBI-0133265 to Z.L., R.M., J.B and T.J., and NSF Grants CCR-0309002 and ITR-0085910 to T.J.

REFERENCES

- [1] <http://www.ncbi.nlm.nih.gov/blast/>.
- [2] J. Borneman, M. Chrobak, G. Della Vedova, A. Figueroa, and T. Jiang. Probe selection algorithms with applications in analysis of microbial communities. *Bioinformatics*, 17(Suppl. 1):S39–S48, 2001.
- [3] J. Camin and R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19:311–326, 1965.
- [4] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.
- [5] J. Felsenstein. Phylogenies from molecular sequences: inference and reliability. *Annuals Rev. Genetics*, 22:521–565, 1988.
- [6] A. Figueroa, J. Borneman, and T. Jiang. Clustering binary fingerprint vectors with missing values for dna array data analysis. *Journal of Computational Biology*, to appear, 2003.
- [7] W. Goddard, E. Kubicka, G. Kubicki, and F. McMorris. The agreement metric for labeled binary trees. *Mathematical Bioscience*, 123:215–226, 1994.
- [8] T. Jukes and C. Cantor. *Mammalian Protein Metabolism*, chapter Evolution of protein molecules, pages 21–132. Academic Press, New York, 1969.
- [9] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Molecular Evolution*, 16:111–120, 1980.
- [10] D. Penny and M. Hendy. The use of tree comparison metrics. *Systematic Zoology*, 34:75–82, 1985.
- [11] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.
- [12] D. Sankoff and J. E. Kruskal. *Time Warps, String Edits, and Macromolecules: the Theory and Practice of Sequence Comparison*. Addison Wesley, Reading, MA, 1983.
- [13] P. Sneath and R. Sokal. *Numerical Taxonomy*, pages 230–234. W.K. Freeman and Company, San Francisco, CA, 1973.
- [14] D. Swofford. *PAUP: Phylogenetic Analysis Using Parsimony version 4.0 beta 10*. Sinauer Associates, Sunderland, Massachusetts, 2002.
- [15] L. Valinsky, G. Della Vedova, T. Jiang, and J. Borneman. Oligonucleotide fingerprinting of ribosomal rna genes for analysis of fungal community composition. *Applied and Environmental Microbiology*, 68(12):5999–6004, 2002.
- [16] L. Valinsky, G. Della Vedova, A. Scupham, S. Alvey, A. Figueroa, B. Yin, R. Hartin, M. Chrobak, D. Crowley, T. Jiang, and J. Borneman. Analysis of bacterial community composition by oligonucleotide fingerprinting of rna genes. *Applied and Environmental Microbiology*, 68(7):3243–3250, 2002.