12-6-2017

Genomic Reconstruction of the Tree of Life

Sambriddhi Mainali

GENOMIC RECONSTRUCTION OF THE TREE OF LIFE

by

Sambriddhi Mainali

A Thesis

Submitted in Partial Fulfillment of the

Requirement for the Degree of

Masters of Science

Major: Computer Science

The University of Memphis

December 2017

Acknowledgement

I am grateful to my advisor, Dr. Max H Garzon. He has always been there whenever I run into trouble and guided me throughout the whole process. He consistently motivated me to take initiative and control over this thesis to make it my own work. Furthermore, he steered me in the correct direction whenever I was stumbling.

I would also like to thank professors Ramin Homayouni (Biology) and Vinhthuy Phan (Computer Science) for serving on my committee. Their comments were very valuable and substantially helped improve the presentation of the results in this thesis.

I would also like to acknowledge contributions from Dr. Fredy Alexander Colorado of the National University of Colombia. Without his guidance and input on sample selection and assessment of the results, I would not have been able to accomplish the goals in this project.

Finally, I express my gratitude towards my parents, my in-laws and my husband for their unfailing support and encouragement. Without them, this thesis would not have been completed successfully. Thank you.

Abstract

A new methodology is presented for molecular phylogenetic analysis addressing a fundamental problem in biology, namely the reconstruction of the Tree of Life (TOL). Here, phylogenies are based on patterns of hybridization similarity in their DNA. Furthermore, phylogenies are based on a set of universal biomarkers (so-called nxh chips) chosen *a priori*, independently of the target group of organisms. Therefore, this methodology enables analyses of groups with biologically distant organisms, and hence could be scaled to obtain a universal tree of life. Unlike conventional molecular methods, it produces a hypothesis in a single run, without optimizing across numerous hypotheses for consensus. Prototype hypotheses for the top two and three layers of the standard bio-taxonomy are presented in detail. The hypotheses agree with the biological Ground Truth in over 70% of the relationships. Higher quality nxh chips are likely to produce better hypotheses, but are more difficult to design.

Table of Contents

List of Figures

List of Tables

Introduction

The Field of Phylogenetics

A question that nearly every one eventually asks oneself about life is – where did I come from? Well, one might think, our existence originated from our parents, but then the question becomes, where did they come from? One of the main concerns in phylogenetics is to answer the final question, *where did life come fro*m? These questions can be traced back to the arrival of *Homo sapiens*, long before Darwin's theory of evolution came into the scene. The scientific branch of biology that addresses these questions is phylogenetics.

An excursion into phylogenetics requires some basic background in biology. DNA (deoxyribonucleic acid) can be regarded as a macromolecule that stores genetic information about an organism in the form of a blueprint that is transcribed in order to perform protein synthesis, which is the primary constituent material of all the organisms in the biome (Watson & Crick, 1953). In its simplest form, DNA can be defined as a string containing sequences of nucleotides represented by A, C, G, and T. (Watson & Crick, 1953) first described the double helical structure of these molecules and, since then, DNA has been established as the blueprint of life. It is transformed inside living cells by a transcription process into an intermediate molecule RNA (ribonucleic acid), as a single stranded molecule that consists of sequence of nucleotides represented by A, C, G, and U. This is responsible of carrying the genomic information to the biological machinery (ribosomes) that translate it into proteins that make up every actual biological organism on planet earth (Crick, 1970). Francis Crick introduced this concept in 1958 (Crick, 1958), today known as the *Central Dogma* of molecular biology, and argued that the role

of these genetic materials (DNA and RNA) is to regulate the synthesis of protein, rather than active participation. Furthermore, he put forward the idea that the Central Dogma is the process responsible for transmission of genetic information in three normal flows: a) Information can be transferred from DNA to DNA (DNA *replication*), b) information transmission from DNA to mRNA (*transcription*) and c) from mRNA to the site of protein synthesis (*translation*) (Crick, 1958; Crick 1970). The new organism reproduces via its own DNA, possibly with some changes (caused by mutations, for example), and the process starts all over. Although not considered a universal process today, the Central Dogma remains largely accepted as a primary mechanism for biological function.

The history of this type of attempts to explain evolution can be traced back to the period where Darwin proposed the theory of evolution. Darwin argued that growth in population would give rise to struggles among species for existence because of limited resources. Only those who were able to adapt would survive; many organisms would fail to survive to an age of reproduction. Hence, in order to survive, evolution is inevitable. Evolutionary theory is commonly accepted today as the most likely explanation for the variety of living organisms in existence today (herein referred to as the *biome*), all arising from some ancient and unknown common ancestor (Sober, 2009).

Since the time Darwin proposed his theory of evolution (Darwin, 1859), biologists have been hard at work trying to reconstruct the exact sequence of evolutionary changes that gave rise to today's biome, usually referred as the true *Tree of Life* (TOL), which will be referred to as the *Gold Standard* below. All organisms in the biome are organized in the form of a phylogenetic tree (Nei and Kumar, 2000). All these efforts constitute the field of *phylogenetics* today.

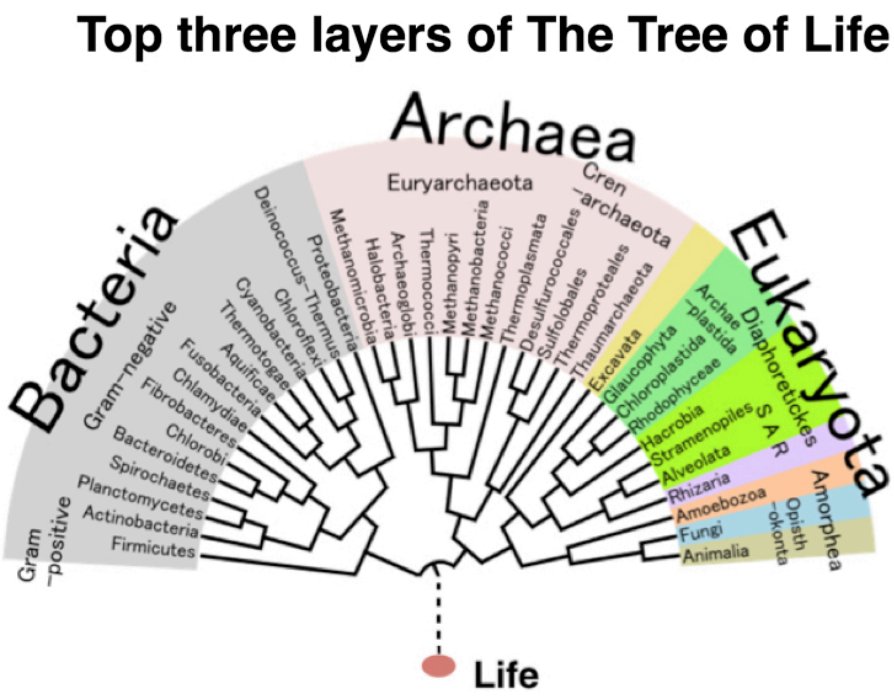

Figure 1. Classification of the three domains in the Tree of Life showing the top three layers of the biome. The predominant view in biology is that this tree currently reflects the most likely hypothesis about the evolutionary relationships among the organisms. According to Darwin Theory of Natural Selection, chains of evolutionary events would have eventually caused multiple differentiations from a primeval form of life that led to today's diversity on earth.

Fundamental Problems in Phylogenetics

The fundamental problem in phylogenetics is thus to get a fairly accurate sense of what evolutionary changes caused the common ancestor to diversify over the course of evolutionary time to give rise to the entire biome present today. The ideal approach would be to jump in a time machine, go back to the past, and record the course of changes that occurred over time going forward. In the absence of such a thing, our only option is to resort to fossil records or other available evidence (such as living descendants) in order to reconstruct a hypothetical version of the TOL, referred to herein as a *phylogeny*. But, these records are necessarily fragmentary at best, and frequently inaccessible. Thus, in practice, the fundamental problem of

phylogenetics becomes, **how to formulate a hypothesis that can estimate the true phylogeny of the TOL (the Gold Standard) and provide some solid evidence to assess the reliability of the formulated hypothesis.**

Biologists have been working to formulate such hypotheses and evidence since Darwin's times over 150 years ago. The major methods currently in use will be summarized in Chapter 2. An ongoing project called the *Open Tree of Life* (Hinchliff et al., 2015) (OTL) is a systematic integrative effort to patch together a comprehensive phylogeny from a variety of partial phylogenies formulated by a number of biologists in many projects through a variety of methods. This phylogeny will be used as reference for the assessment of the quality of the phylogenies produced in this thesis, and will be referred to as the Ground Truth hypothesis.

Bioinformatic Approaches to Molecular Phylogenetics

In the last two decades, major advances in computer science and biotechnology have revealed that the basic carriers of genetic information (such as DNA and RNA) hide a number of relevant secrets about the TOL that can be unearthed by computational analyses. The goal of this project is thus to introduce a new universal technique for phylogenetic analyses based on genomic sequences alone, and provide an assessment of its biological significance using the Ground Truth (in lieu of the inaccessible Gold Standard) as a reference.

Major Concepts in Phylogenetics

Current Methods for Phylogenetic Inference in Biology

To formulate an accurate phylogenetic hypothesis, biologists originally used apparent morphological features to group organisms into groups (clades) that presumably followed a longer common path of evolution before differentiation into species, according to the Central Dogma. However, in most of the cases, this approach does not yield consistent phylogenies, and may even be misleading. For example: bats were considered as part of the clade of birds due to their wings and their ability to fly, but not very related to mammals. Eventually, many such controversial phylogenies were created.

Since the discovery of DNA as the blueprint of life (Watson & Crick, 1953), molecular methods have proven to be more reliable and have become a preferred method for construction of phylogenies. One of the major advantages of these methods is that it enables the use of DNA as means of comparison among organisms. Furthermore, evolution, in an organism, can now be defined as the pattern of changes in DNA, at the nucleotide level. Due to the presence of these patterns, the use of mathematical and computational models to compare DNA sequences of several organisms has become possible and productive. Unlike morphological features, long sequences of nucleotides in DNA are expected to contain a huge amount of phylogenetic information, although extracting it requires extensive and profound analytical work. Therefore, molecular methods are believed to be capable of addressing some of the deficiencies in the traditional approach mentioned above (Nei and Kumar, 2000). However, recent developments in the field of phylogenetics show that standalone use of molecular methods is not sufficient for reliable reconstruction of a phylogeny. Current research attempts to combine molecular data with

fossil records, for example, produces better phylogenies by combining biomarkers, such as *cytochrome c oxidase* subunit I (COI) or ribosomal rRNA (16S rRNA), with molecular clocks recalibrated using fossil records (Burridge, et al, 2017). This thesis can also be regarded as a probe into the power of molecular (or more specifically, genomic) methods alone for phylogenetic reconstruction. To that end, it is necessary to summarize the molecular methods currently in use in phylogenetics in the remaining of this section. These methods are usually implemented and available in software suites, with PAUP being one of the preferred methods and used in this thesis for comparison in evaluation.

Maximum Parsimony

In phylogenetics, the principle of parsimony states that among all possible trees describing the phylogenetic relationship between Operational Taxonomic Units (OTUs) (Sokal and Sneath, 1963), the phylogeny requiring the minimum number of evolutionary changes is a more likely hypothesis. Therefore, the most parsimonious phylogeny should be closest to the Gold Standard. To actually apply this principle, it is necessary to define the parsimony length of a phylogenetic tree. The parsimony length can be defined as the sum of the Hamming distances of sequences labeling endpoints of edges in the tree (Kim and Warnow, 1999). Originally this method was used for analyzing phylogenetic relationship among organisms using their morphological features; later on, this method was modified to analyze molecular data as well. However, the validation of the phylogenies in this thesis will, naturally, include phylogenies based on molecular data only.

Currently, several tools are available that use the molecular data (selected DNA biomarker sequences available in the organisms of interest, usually highly conserved) in order to infer phylogenetic relationship among a group of organisms. But all of these tools require an

alignment of the biomarkers used. A distinction between informative and noninformative sites in the alignment must be obtained to proceed further. For a site to be informative, at least two different kinds of nucleotides must be present in at least two OTUs (Nei and Kumar, 2000). Then, for each possible tree for these OTUs, the total number of changes is calculated across all informative sites. The tree with the minimum number of changes among all the possible trees is selected. The process is illustrated in Figure 2 for four OTUs.

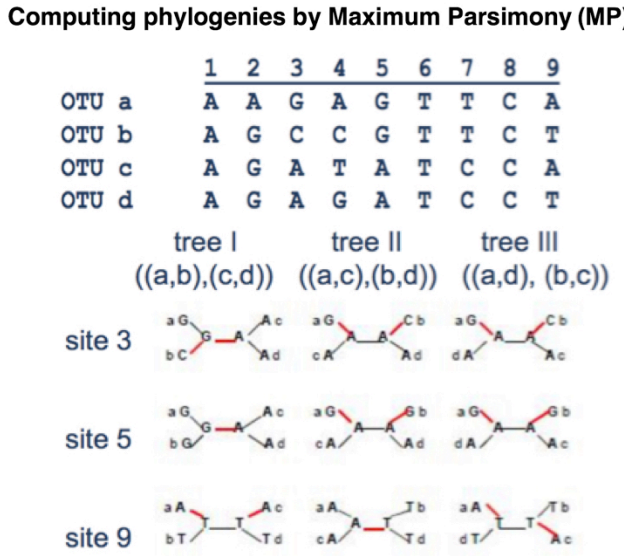

Figure 2. Parsimonious phylogenetic Inference at the molecular level with four organisms according to (Varvio, 2011). In biology, parsimony is the principle that favors the simple and most straightforward explanation of any biological phenomenon among valid competing explanations. In order to reconstruct a phylogeny using Maximum Parsimony (MP), the sequences for the group of organisms are aligned first. Next, informative and noninformative sites are distinguished, and the total number of changes occurring across all the informative sites in the group is calculated. The tree with the minimum number of changes across all informative sites is a tree of maximum parsimony.

For a small number of taxa, it is possible to perform an exhaustive search for MP trees. However, as the number of taxa increases, this method becomes infeasible and only heuristic methods can actually find reasonable estimates of phylogenetic hypothesis to estimate the Gold Standard. For example, for 4 OTUs, there are only 3 possible unrooted bifurcating trees, but even for just 10 OTUs, the number of possible unrooted bifurcating trees is 2,027,025. In general, for

n OTUs, there are $(2n-5)! / [(n-3)!2^{(n-3)}]$ possible unrooted bifurcating trees (Cho, 2012). When *n* is less than 10, one can use so-called branch-and-bound search methods; however, when *n* becomes greater than 10, heuristic search is a must (Nei and Kumar, 2000).

Many MP trees are possible for a single group of organisms, especially under uncertain evidence in the sequences. That is why biologists rather tend to combine a number of hypotheses into a single composite tree, which is called a *consensus tree*. Most commonly used consensus trees are strict-rule based, but good trees can also be obtained by a majority consensus, where a relationship is preserved if and only if a fraction of the trees (e.g., 85%) exhibit it. In a strict consensus tree, any conflicting branching patterns for a set of OTUs among all the candidate trees are resolved by creating a multifurcating branching pattern. In a p% majority based consensus tree, a branching pattern present in at least a fraction p/100 of the trees is adopted (Nei and Kumar, 2000). Generally, a majority rule-based consensus trees of 70% or above are considered to be reliable.

Although this method is a more sophisticated way of making phylogenetic inferences, there are situations when it tends to yield incorrect trees. The theoretical foundation of this method specifies that the lesser the number of evolutionary changes in a group of organisms, the more accurate the phylogeny is. This amounts to the assumption that there are no backward and parallel substitutions at each nucleotide site and the number of nucleotides in each sequence is sufficiently large to come to a conclusion. However, in practice, the nucleotides are pruned to backward and parallel substitutions and often the number of nucleotides in each sequence is small (Nei and Kumar, 2000). In this situation, the trees obtained using this method tend to be incorrect.

Maximum Likelihood

Maximum likelihood is a method used in statistics to estimate the parameters in a statistical model, given a set of observations, by finding the values for the parameters that maximizes the probability of obtaining the observations (Myung, 2003). The main concern in phylogenetic inference for a set of species using ML is to find *all* the phylogenetic trees for the biomarkers, but the problem is, too many possible trees render the calculation impossible in practice. The method of maximum likelihood selects the tree (model) that is more likely to generate the sequences of the given set of species (Kim and Warnow, 1999). This is made possible by Bayes's Theorem in probability, here given by

$$P(\text{Model}|\text{Data}) = P(\text{Model and Data})/P(\text{Data}) = P(\text{Data}|\text{Model})P(\text{Model})/P(\text{Data})$$

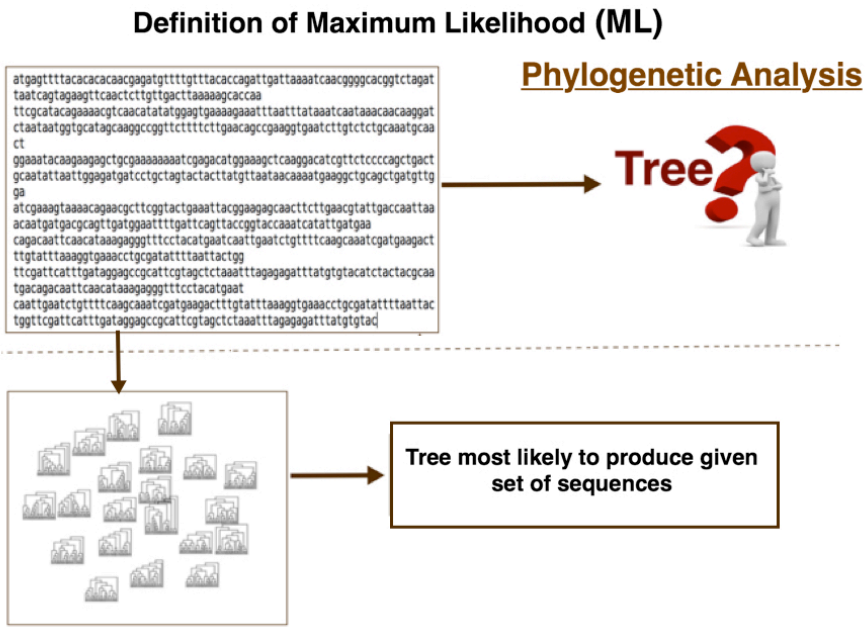

Figure 3. Maximum Likelihood (ML) estimation of a phylogeny. For a given set of biomarkers, the true phylogeny shows the actual sequence of differentiation events that resulted in these sequences. A priori, many trees are conceivable for the given set. ML selects a tree that is most likely to produce the given set of sequences.

Phylogenetic inference using ML makes several assumptions (Kosiol, Bofkin and Whelan, 2005) as follows:

1. Branches of the evolutionary tree evolve independently;
2. Evolution at particular sites in the sequence alignment is dependent on the current state only, not on the past states of evolution;
3. Reasonable topology estimates lead to reasonable parameter estimates;
4. Sites in sequence alignments change at the same overall rate;
5. The evolutionary process changes the same way going forward and backwards;
6. The rate with which transition mutations (change from a purine nucleotide to another purine or from a pyrimidine to another pyrimidine, i. e. A<-> G or C<->T) occur is relative to transversion mutations (change from a purine nucleotide to pyrimidine or *vice versa*, i.e. A<->T or G<->C) in DNA.

Like with the MP method, the computation of the likelihood of the sequences from the entire possible tree becomes infeasible as the number of taxa grows, even modestly. Therefore, effective use of ML requires the use of heuristics, such as hill climbing search (see Figure 4), hierarchical clustering, and Monte Carlo search (see Figure 5).

Hill Climbing

Hill climbing is a local search algorithm in numerical analysis, where an arbitrary solution is chosen for a problem and attempts are made by incrementally replacing an element in the solution with a neighbor to improve the estimate of the maximum value of a function (Russell and Modern, 2003), such as the likelihood of generating the given sequences.

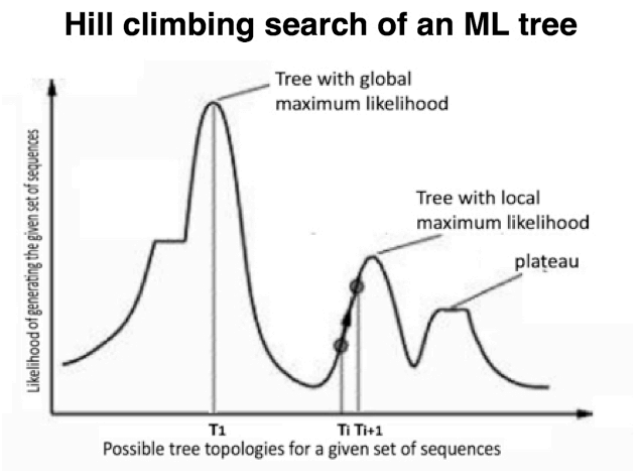

Figure 4. Hill Climbing Search of Maximum Likelihood phylogenies for a given set of sequences according to (Russell and Modern, 2003). If the *x*-axis represents all the possible tree topologies for a given set of sequences and the *y*-axis gives their likelihood, the ML tree that maximizes likelihood of generating the sequences is the ML tree. However, although it is theoretically possible to enumerate all the possible trees and compute their likelihood for the sequences exactly, it is very difficult in practice to enumerate a super-exponential number of possible topologies to realize the maximum value tree. In order to estimate it, one can pick a random initial tree and recursively use a current tree T_j to find a similar topology T_{j+1} to improve the estimate if it has higher likelihood. The process is repeated for a certain number of iterations or until improvements in likelihood become negligible, evidence that they might be close to a maximum. There is a risk that the method gets trapped in a local maximum and will not yield a good approximation in one try, so the approach is repeated a number of times.

Monte Carlo

In this search technique, repeated random samples (with replacement) are taken in order to estimate the solution of a problem.

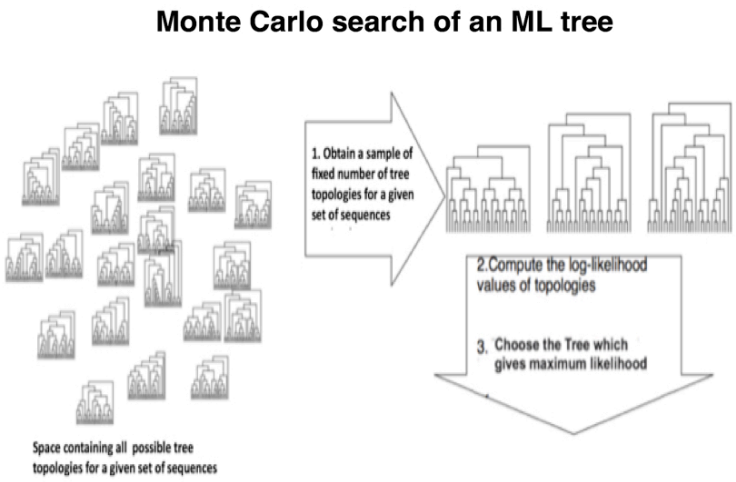

Figure 5. Monte Carlo Search of Maximum Likelihood phylogenies for a given set of organisms according to (Suzuki et al, 2004). In this approach, a sufficiently large sample is drawn from the population of all possible tree topologies for the set in order to get a close estimate of one with maximum probability of producing the original sequences. The tree with largest such likelihood in the sample is selected as an estimate for the true maximum.

Unweighted Pair Group Method with Arithmetic Mean (Distance Method)

This is a bottom-up clustering method. This method requires a distance matrix between each pair of organisms. The algorithm is illustrated in Figure 6. At each step, two closest nodes are grouped into one higher-level cluster joint by an intermediate node. The distance matrix is then updated with the new distance between all the remaining nodes from the new intermediate node. This process is repeated iteratively until only one cluster remains.

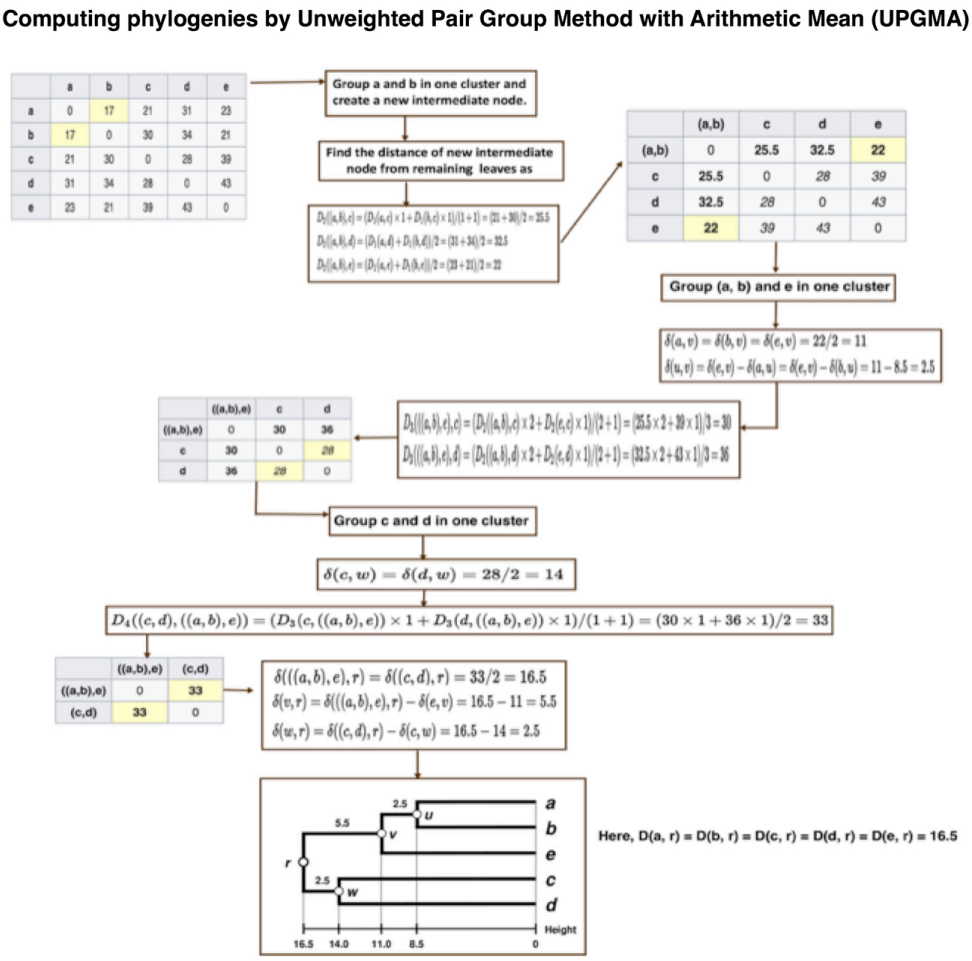

Figure 6. An example of the construction of a phylogeny using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) according to ("UPGMA", accessed 2017). This process starts with each node as an individual cluster. Then the two clusters closest to each other are joined together forming a larger cluster, joined by an intermediate node. The two subclusters are assumed to be equidistant from the intermediate node joining them. Then, the distances from the intermediate node are calculated for all the remaining clusters in the space and the distance matrix is updated. The process is repeated until a phylogeny is obtained that groups together all the leaves into a higher cluster with one root node.

This method assumes that evolution occurs at constant rate, i.e. that all the leaves are equidistant from the common ancestor. On this assumption, a rooted tree can be obtained because it is easy to infer the root of the tree, particularly when gene frequency data are available for phylogenetic reconstruction. This method produces reasonably good trees compared with other distance methods (Nei and Kumar, 2000).

Neighbor Joining

Neighbor Joining (NJ) is also a distance method based on bottom-up (agglomerative) clustering technique. It groups two nodes into a cluster that are farther apart from the rest of the nodes. Naruya Saitou and Masatoshi Nei proposed this method in 1987 (Nei and Kumar, 2000). The process of constructing phylogenies by this method is illustrated in Figure 7.

The construction of a tree begins with a star tree on the assumption that there is no clustering of taxa (Nei and Kumar, 2000). At each iterative step, this method tries to minimize the sum of the lengths of all the branches in the current topology. In order to do so, a Q-matrix is calculated which stores the distance values between each pair of taxa *(i, j)* reflecting their proximity to remaining taxa, as shown in Figure 7. An intermediate node *u* joins the pair of taxa (*a, b*), which has lowest Q-value. Then, distances *d(u, a)* and *d(u, b)* are computed. This will lead to an update in the distance matrix, which now includes the distance from *u* to each of the remaining leaves. The process is repeated until the sum of the lengths of all the branches of the tree can no longer be reduced.

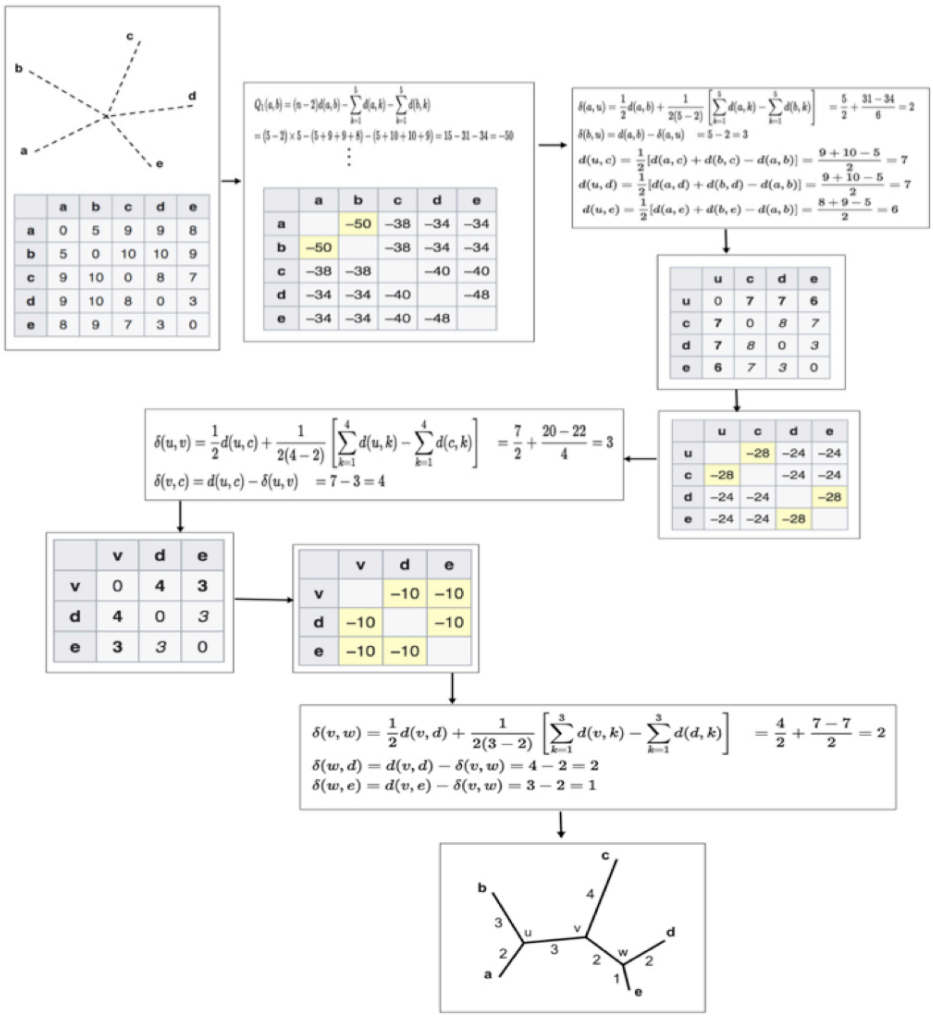

Figure 7. An example showing the construction of phylogeny using Neighbor Joining (NJ) Method, according to ("Neighbor Joining", accessed 2017). This is another distance-matrix method where two clusters that are closer to each other than to other clusters are joined into a single cluster. The method starts with star-like phylogenetic relationship among organisms. Based on a distance-matrix, a Q-matrix is calculated that illustrates how close two organisms are to each other given the context of the remaining organisms. An intermediate node joins the two organisms, which have lowest Q-value. The distances from the intermediate node to the remaining organisms are re-calculated and the distance matrix is updated. The process is repeated iteratively, until a single node remains.

This method is mainly used for DNA and protein sequences. It is very fast by comparison to Maximum Likelihood and Maximum Parsimony methods, as only a small proportion of all

possible topologies are considered (Saitu and Nei, 1987). When the given input distance matrix is correct, this method is expected to produce a reliable phylogeny.

Drawbacks of Conventional Methods

There are several problems with the conventional methods discussed above. First, they all require a multiple sequence alignment to construct a phylogeny. For smaller sequences, like COIs, that might not be the problem, but as the length of biomarker sequences increases, even for a modest number of organisms, getting the alignment may become computationally taxing and increasingly infeasible. Second, the order in which the biomarkers are given also impacts the resultant phylogeny, as the alignment will be different than when the same set of organisms are arranged differently. An important point to note here is that, due to higher dependencies in the alignment, the phylogeny for a subset of organisms in one batch may differ from that obtained when they are in another batch.

Another major disadvantage is that, as the group of organisms of interest changes, the biomarker usually has to change because it has to be available in all organisms in the target group. Biologists believe that only changes in highly conserved genes truly cause significant evolutionary steps. As a result, phylogenies based on those conserved genes are considered to be a better approximation of the Gold Standard. However, selecting different genes leads to different phylogenies and it is very difficult to find such conserved genes in *all* organisms (Garzon and Wong, 2011). Hence, a specialized choice of biomarkers is made depending on the group of organisms under study. The fundamental problem of phylogenetics would then have to remain unresolved for the TOL at large.

Metrics for Similarity Assessment of Phylogenies

Due to the availability of several methods to reconstruct phylogenies, an analysis must be done to select the most accurate phylogeny. Generally, biologists perform a comprehensive qualitative analysis of the resulting phylogenies to vet their biological accuracy and consistency with other evidence in biology. However, these analyses are hugely affected by knowledge and preferences of the researcher performing the analyses. Because of this lack of common ground in qualitative analysis, quantitative objective metrics are needed to compare phylogenetic trees. Two of the most commonly used metrics are the Robinson-Foulds and the Path-distance indices. They were used to validate the phylogenies obtained in this research.

The Robinson-Foulds (RF) distance is a widely used measure of (dis)similarity between trees that is based on the characteristics of the trees without performing any edit operations (Lin et al., 2012). The algorithm to compute the RF index between any two trees is shown in Figure 8. Removing an edge in a phylogenetic tree disconnects the tree and creates a partition of the leaves. A set P_i containing all the possible partitions present in tree T_i, can be computed. In order to compute the RF index between trees T_1 and T_2, first, P_1 and P_2 are computed and the total numbers of $P_1[i] \in T_1$ but not in T_2 and the total number of $P_2[j] \in T_2$ but not in T_1 are obtained. Finally, the RF index equals the weighted average of these two numbers. Methods available in the R phangorn package were used in order to compute normalized RF index between any two trees.

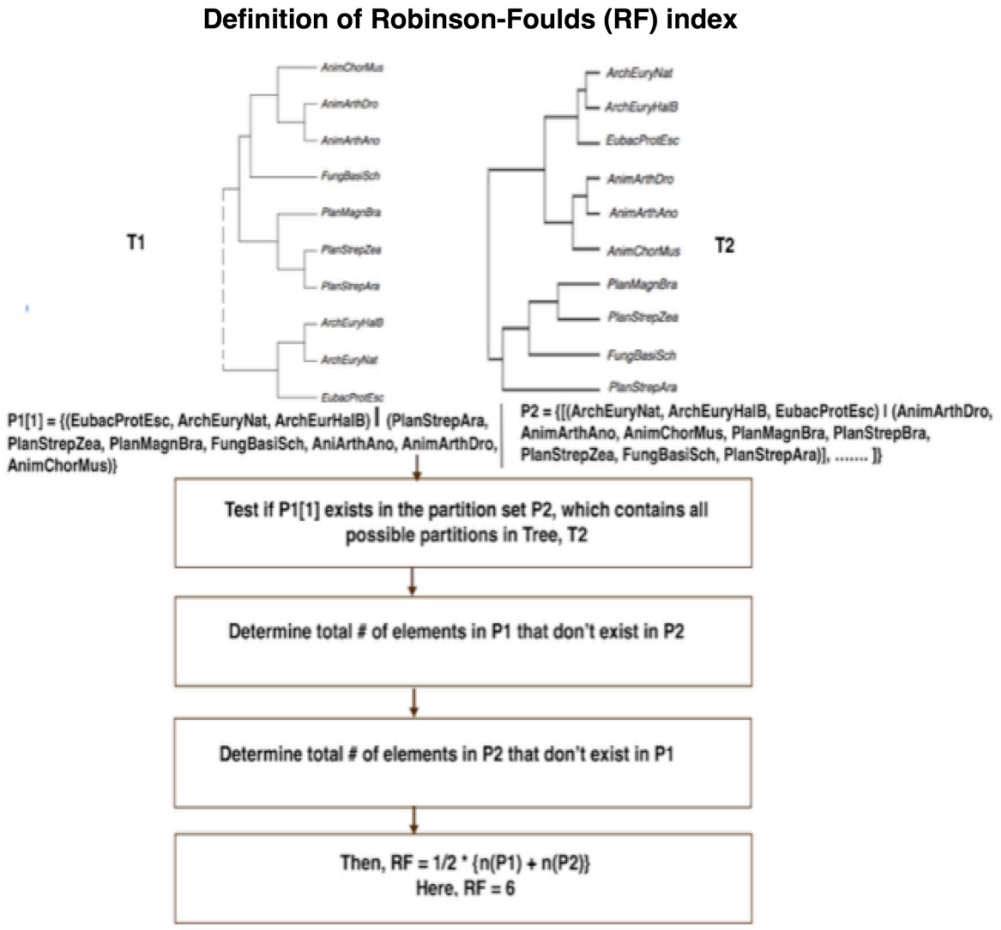

Figure 8. Calculation of the Robinson-Foulds (RF) index between two phylogenies T1 and T2 quantifying their degrees of similarity (Robinson and Foulds, 1981). A partition, P1[1] is created in T1 when deleting an arc (branch). The set containing all possible partitions of T2, P2 is searched for the presence of P1[1]. The process is repeated for all possible branches in P1 and the total number of partitions in P1 that do not have a matching partition in P2 is determined. The process is repeated to find total number of partitions in P2 that do not exist in P1. The average of these two numbers is the RF degree of similarity between the two phylogenies T1 and T2.

The major advantage of this index is that it measures the dissimilarities between any two trees as given, based on their own characteristics, without performing any edit operations (Lin, et al, 2012). Since for any tree with n nodes, only $n - 3$ nontrivial bipartitions are possible, when there is n number of species, the maximum possible value of an RF index between any two trees is $2n - 6$ (Robinson and Foulds, 1981; Steel and Penny, 1993). So, it is easy to quantify dissimilarities using the RF index. However, the change in the RF index can be unpredictable.

For example: on the one hand, the RF index might not be able to make fair discrimination between dissimilar tree topologies, but on the other, moving a leaf at the end of a tree to the other end might create a tree with maximum possible RF index to the original tree.

Because of its lack of robustness, the Path distance (PD) index was also used in this thesis. This index translates trees into higher dimensional vectors to represent the (dis)similarity between them in terms of their Euclidean distance (Steel and Penny, 1993). The algorithm for computing PD indices is shown in Figure 9. For any pair of trees, T_1 and T_2, the distance between all the possible pair of leaves *(j, k)* (in some fixed order) determines the tree uniquely and can be computed initially. Then, vectors *d(1)* and *d(2)* formed after arranging previously computed d(i, j, k) where, *i = 1* or *2*, identify the trees. Now, after translating T_1 and T_2 into d(1) and *d(2)* respectively, the PD index equals the Euclidean distance between them. In this thesis, for the sake of comparison on a common scale, all the PD indices were normalized by dividing each index with maximum value of all possible indices in the batch.

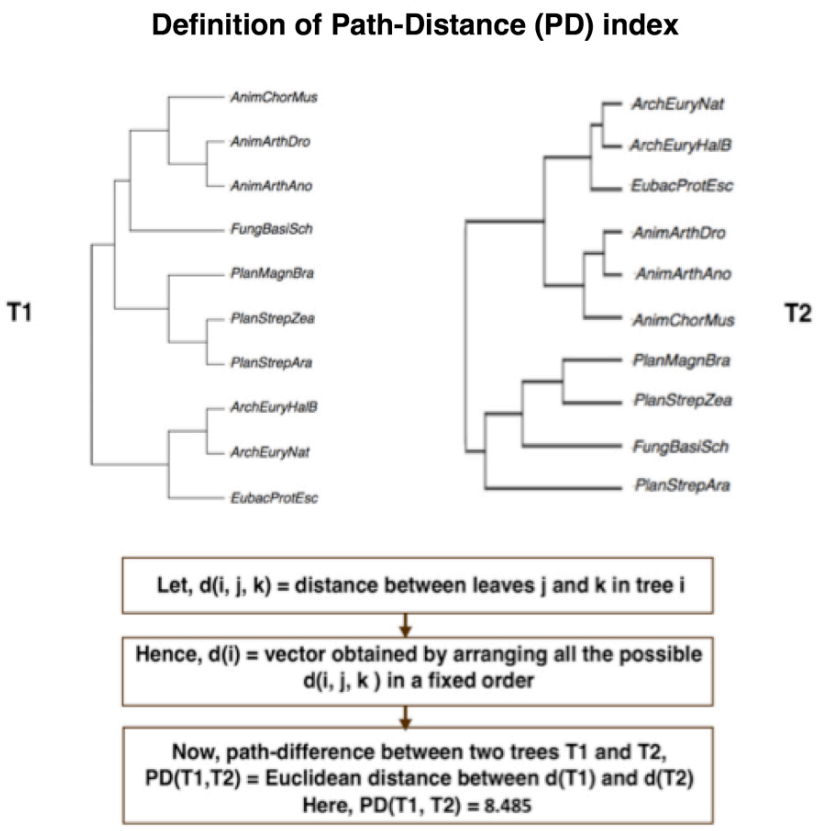

Figure 9. Calculation of the Path-distance (PD) index between phylogenies T1 and T2 as an alternative way to compute their degree of similarity (Steel and Penny, 1993). T1 and T2 are translated into vectors *d(1)* and *d(2)* by arranging the distances between all possible pairs of leaves in T1 and T2 (labeled with the given sequences) in a fixed order, respectively. Such vector determines the trees uniquely. The PD index between T1 and T2 is defined as the Euclidean distance between the vectors *d(1)* and *d(2)*.

PD indices have several features that evidence that they are useful while performing phylogenetic analysis. They require less computation time, which makes it desirable while comparing large trees (Steel and Penny, 1993). It might be more suitable while comparing dissimilar trees. However, in the course of this research, it was observed that most of the PD indices cluster closely together, giving a smaller range of. So, the choice of better estimate by comparison to the Ground Truth could not be made solely on PD indices.

Genomic Methods for Phylogenetic Inference

Next Generation Microarrays (nxh chips)

DNA – the blue print of life – plays a major role in determining morphological and metabolic behaviors of an organism, along with the kind of diseases that organism is prone to suffer from. For example: researchers believe that mutations in the genes BRCA1 and BRCA2 causes as many as 60 percent of all cases of hereditary breast and ovarian cancers in female *Homo sapiens*. But, researchers also found over 800 different mutations in BRCA1 alone (Cook-Deagon, et al, 2010). This clearly shows that huge amounts of information are being stored in DNA, which can be captured, processed, manipulated and analyzed in order to make any assessment about any organism and their relationships.

As a result, microarrays were developed as a tool to capture and mine large-scale genomic data. They are planar substrates such as glass, mica, plastic or silicon, where DNA strands are affixed to allow specific bindings of bio-samples collected from an organism (Schena, 2003). During the early 1990s, the first microarray experiments were performed using complementary DNA (cDNA) affixed to the microarrays. The length of a typical cDNA is 500-2500 base pairs and they are widely used in gene expression assays (Schena, 2003). Since 1990s, microarrays have been refined and today have become most commonly used powerful tools to capture and mine genomic and metabolomic information. The information gathered by these tools has wide applications in the fields of biology, medicine, health and scientific research.

Biologists refer to probes to indicate the biosample as shown in Figure 10 (Right) and target to indicate the microscopic element fixed on the microarray as shown in Figure 10 (Left). In this thesis, the terms "*probe*" and "*target*" are used with the reversed meaning.

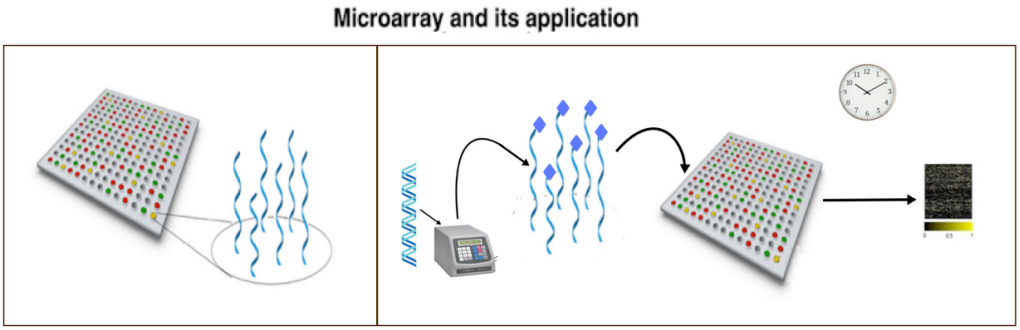

Figure 10. Microarrays are a standard technology in biological applications. Genes of interest are affixed to a solid surface, such as glass or mica. Target sequences (usually RNA or cDNA) are collected from the organism under study. These DNA strands are tagged using fluorophores (Nagl, et al, 2005) and poured on the chip. After some relaxation time to allow for hybridization to reach equilibrium, a fluorescent readout can be collected in a picture.

Despite the advantages offered by microarrays, the analysis relying on these data gives results that are hardly reproducible because of the high uncertainty of hybridization of targets to probes if such is present. No constraints are implemented in these chips to minimize cross-hybridization between probes. As a result, the results are not accurate and hence unreliable due to the lack of reproducibility of results, as argued in (Garzon and Mainali, 2017).

A second disadvantage of microarrays is that they might miss target strands if they do not hybridize to any probe, and thus miss signals that could yield useful information. For example, probes are arranged on the chip without giving any consideration to the fact that they might hybridize with themselves. If that happens, then the tagged targets will not have any chance to hybridize with the probes, resulting in missed signals.

These drawbacks of conventional microarrays have been addressed in a number of works (Garzon & Mainali, 2017; Garzon & Bobba, 2012) with the introduction of next generation microarrays, where none of the probes hybridize with each other or themselves. The problem of finding such large set of DNA oligonucleotides is known as the *Codeword Design* problem (Garzon, 2012; Garzon and Bobba, 2012) and has been proven to be very hard (technically, NP-

complete - see Garzon and Bobba, 2012) to solve in full generality. A solution is called a noncrosshybridizing (nxh) set and they would be extremely useful in the design of a next generation microarray. The size and composition of such nxh sets is therefore very difficult to establish because they ultimately rely on the structure of Gibbs Energy landscapes that govern structural properties of hybridization between oligonucleotides. But the good news is that this problem can be translated into a geometric sphere-packing problem by mapping oligos with high hybridization affinity into neighboring points in a geometric lattice in the familiar Euclidean space (Garzon and Bobba, 2012).

To address this problem, a new model was proposed, the hybridization distance (*h*-distance), that quantifies the possibility of two oligos *x* and *y* hybridizing with each other (Garzon, 2012; Garzon et al, 2012). In this model, the sphere-packing problem can be approximately solved in the DNA spaces of small oligonucleotides. An example of computation of *h*-distance between two oligos *x* and *y*, where *x* = *agc* and *y* = *tgg*, is shown in Figure 11. The *h*-distance model is an effective approximation of the Gibbs Energy that regulates hybridization in DNA. A decision made for hybridization between two strands obtained by the *h*-distance method agrees with a decision based on the Gibb's Energy Nearest Neighbor model over 80% of the time (Garzon and Bobba, 2012). The only price to pay here is *h*-distance does not distinguish between an oligo and its Watson-Crick complement. The term *pmer* (for *poligomer*) will be used to refer to such pairs of a strand and its complement, which may be identical in the case of a Watson-Crick palindrome. With that approximation, the *h*-distance can now be treated like an ordinary distance function (similar to ordinary distance) and used to quantify the amount of noise inherent in a microarray design (Garzon and Mainali, 2017; Garzon & Bobba, 2012).

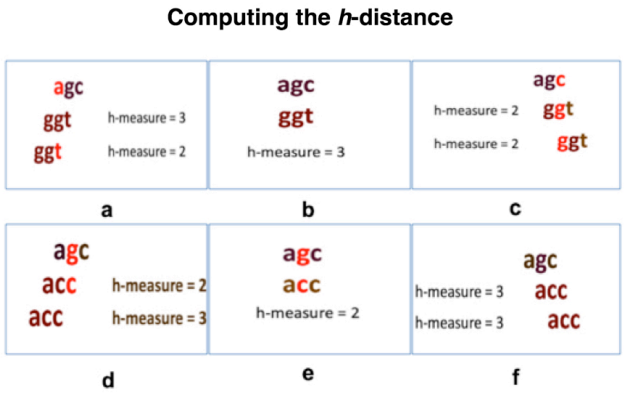

Figure 11. Computation of the *h*-distance between two oligos *x* and *y* of a common length *n*. *x* and reversed y^R are aligned in all possible frame-shifts (here five) and the number of complementary matches for each frameshift is counted. The h-measure for *x* and *y* is the defect (difference) to the length *n* of the maximum of these values. The same procedure is repeated to find the *h*-measure between *x* and its WC complement y'. The *h*-distance between *x* and *y* is the minimum of these two *h*-measures *h(x, y)* and *h(x, y')* (Garzon et al., 1997; Garzon and Bobba, 2012; Garzon, 2012).

Reliability Analysis and Design of nxh chips

The design of a next generation nxh chip is based on judicious selection of probes. First of all, a threshold τ for hybridization by *h*-distance is selected. Then, a judicious selection of probes is made in such a way that all of these probes are separated from one another with the minimal distance τ so that they will not hybridize with each other or themselves (the nxh property). A good chip design (basis) should also have sufficiently many of probes in it so that each target shred would get chance to hybridize with at least one probe. Furthermore, all targets will hybridize to at most one probe. This becomes true for an nxh chip design with hybridization threshold $\tau/2$ because, as shown in Figure 12 (Right), if a shred *z* has *h*-distance less than $\tau/2$ for two of the probes *i* and *j*, then it cannot hybridize with two different probes, due to the triangle inequality and the minimal separation τ in *h*-distance between any pair of probes.

Such a design can be implemented with standard microarrays technology, where a physical chip would consists of a number of spots corresponding to the oligos in an nxh set. Each spot consists of two bundles, one containing a fixed number of copies of an nxh oligo probe

and the other same number of copies of its Watson-Crick complement, separated by the minimal distance τ, so that they will not be able to hybridize, as illustrated in Figure 12 (Left).

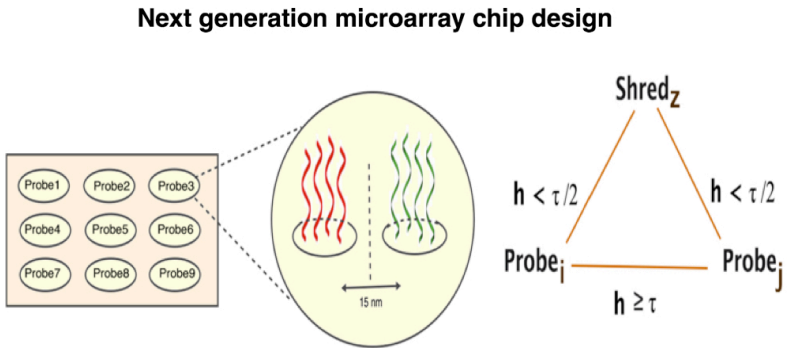

Figure 12. Noncrosshybridizing (nxh) chip design (left) in (Garzon and Mainali, 2017). The chip design consists of a number of spots in 1-1 correspondence with a so-called *basis* set of nxh oligos. All the pairs of the basis oligos in the chip are at *h*-distance at least τ from each other. Each spot consists of a fixed number of copies of a basis elements and same number of copies of their Watson-Crick complements, laid at a fixed distance to prevent cross-hybridization. A target shred *z* is assumed to be able to hybridize with a probe if and only if its *h*-distance to the probe is less than $\tau/2$. Due to triangle inequality (inherent in the metric property of *h*), copies of a random *z* cannot hybridize with two of them (right). Thus, the nxh chip minimizes the amount of noise, thus addressing a problem in standard micro-array technology (Garzon and Mainali, 2017).

With this chip design (basis) in hand, we can now capture genomic information about any organism from *any* biomarker in a so-called *digital signature*. The marker will be shredded using some standard technology, such as sonication or cleavage (Sambrook and Russell, 2006). Unlike conventional microarrays, probes in the chip will be tagged using fluorophores (Nagl, et al, 2005), because we want to eliminate the noise caused by target shreds that could hybridize with them when poured to the chip. Then, all the shreds are poured on the chip containing the probes. After allowing sufficient relaxation time for hybridization to occur, we obtain the readout for the digital signature for that particular organism. The procedure is illustrated in Figure 13.

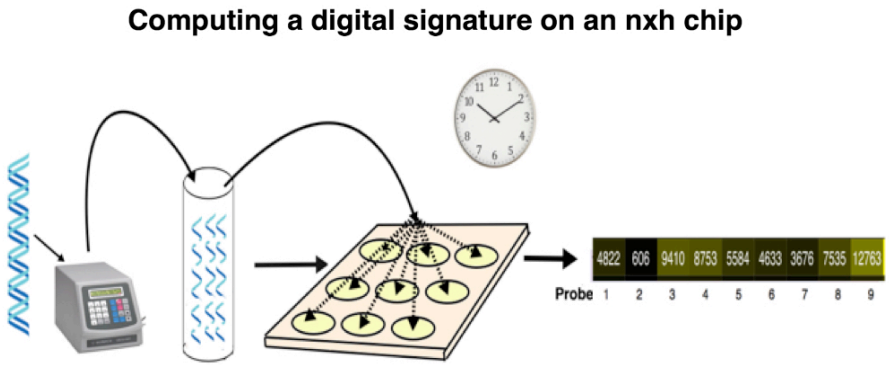

Figure 13. Calculation of the digital signature of an organism using a representative sequence in (Garzon and Mainali, 2017). A sequence of an organism is shredded using sonication. Unlike conventional microarrays, the probes in the chip are tagged using a fluorophore (Nagl, et al, 2005). The shreds containing tagged probes are poured to the chip. A signature can be collected in a photograph (readout), as with standard microarrays.

In order to quantify the quality of this chip, a random experiment can be performed. The sample space for the experiment contains *all possible* pmers with the length same as the length of probes in the basis. The experiment selects a pmer from the sample space at random. All the pmers are selected with uniform probability. The random variable X used in this quantification is the total number of probes that a random pmer sticks to. Now, the quality of any basis can be expressed in terms of two metrics, namely, the *expected value* of X and the expected value of the random variable Y counting the number of bits required to represent the values of the random variable, which is called the *Shannon Entropy*. Ideally, a basis would show a random variable with values constantly (or at least its expected value is) equal to 1, and its Shannon Entropy is equal to 0. The quality profile of several bases is shown in Figure 15.

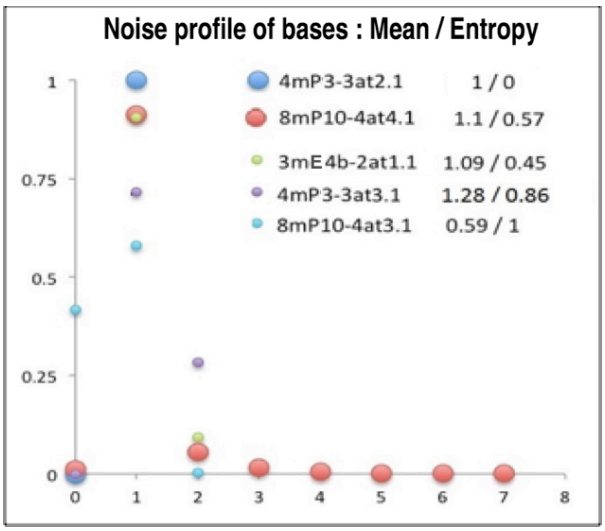

Figure 14. Reliability analysis of nxh chip based on two metrics of reliability, namely the *expected value* and *Shannon Entropy* of the random variable *X* that counts the number of probes in a basis (*x*-axis) that a pmer selected at random could stick to, shown in the *y*-axis for five nxh bases, as given in (Garzon and Mainali, 2017). The sample space consists of all the possible pmers of a common length equal to the length of the probes in the basis. The Expected value of *X* can be used to quantify the amount of noise on the chip. Ideally, we would like this random variable to be constant of value 1 (no ambiguity in the hybridization process for a noise-free chip). Alternatively, the Shannon Entropy (defined as the expected value of number of bits required to represent the values of *X*) could be used to quantify the degree of uncertainty with which a target shred sticks to a single probe. This value should equal 0 for an ideal chip design.

Even having such as perfect basis, the possibility remains that a majority of the pmers stick to only one probe in the basis. Ideally, we want equal amounts of signal being captured at each probe on the chip, when the same amount of all the possible probes are poured to a chip, i.e., random pmers hybridize evenly to all probes, and the signature obtained is not biased towards any probe in the basis. As shown in Figure 14 and Figure 15, 4mP3-3at2.1 is a perfect basis but the signature of pmers is slightly biased towards third probe. A separate experiment was performed, which is not described in this thesis, in order to quantify the impact of this bias. However, no significant impact of biased signals was observed for phylogenetic reconstruction using the bases in Figure 16 during that experiment.

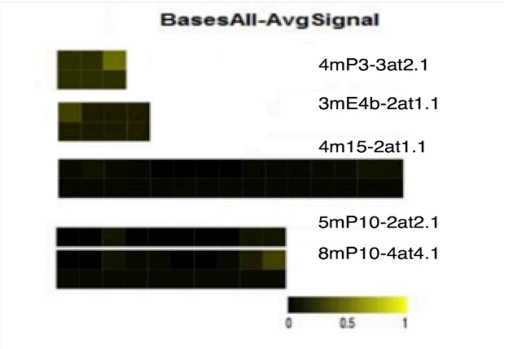

Figure 15. Distribution of pmers likelihood of hybridization across the probes in a basis, as given in (Garzon and Mainali, 2017). Ideally, one would desire an unbiased chip design, i.e., one where the likelihood is uniform for all the pmers if they were poured onto the chip in equal concentration as shown in the third row for each basis. Here, the signature of all the 4pmers on perfect basis, 4mPolar3-3 is slightly biased towards the third probe.

In order to further test the quality of the nxh chips, we ran a control experiment in which random signatures were obtained from a chip (say, due to a careless choice of probes on the chip). For that purpose, an R program was used to generate random signatures (satisfying similar constraints to signatures generated from a random set of probes) for the 13 organisms in sample 1. The phylogenetic trees based on Euclidean distance, Angle and Composite of Euclidean distance and Angle (Ang1000) between the signatures were produced using the *ape* (Paradis et al, 2004) and *phangorn* (Schliep, 2010) packages available in R (they were the same packages used in the computations below). Then, based on the RF indices and PD indices, those phylogenetic trees were compared against phylogenies on nxh chips. The process was repeated 32 times. On average, it was observed that phylogenies on the nxh chip are significantly closer to the Ground Truth than the phylogenies on random signatures, as shown in Figure 16. To further investigate the statistical significance of these differences, two *z*-tests were performed with the research hypotheses "the average of the RF (PD) indices on random set of signatures is greater than that on nxh chips". The null hypotheses were formulated as the negation of the research hypotheses. Sufficient evidence was not found in those indices to reject the null hypothesis,

except only when the similarity metric used to obtain the phylogenies was the Angle between signatures. Moreover, the phylogenies based on Angle were found not to be biologically meaningful, so they are not analyzed further qualitatively in this thesis. In conclusion, the phylogenies based on Euclidean distance and Ang1000 metrics were statistically significantly better than the phylogenies based on random set of signatures, according to RF and PD indices.

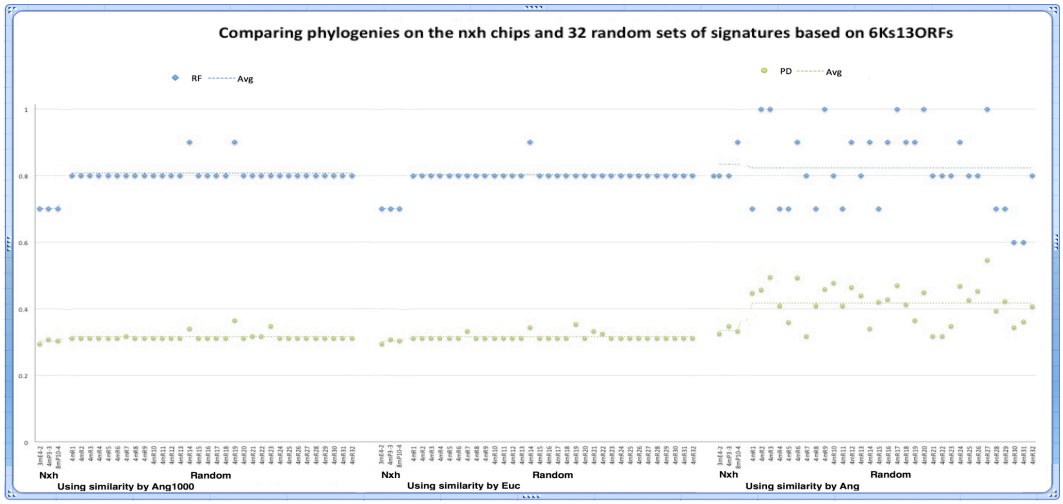

Figure 16. Quantitative assessment based on RF and PD indices, of phylogenies on nxh chips, by comparing the indices of phylogenies on random set of signatures to the Ground Truth. The phylogenies based on nxh chips are statistically significantly closer, on average index, to the biological Ground truth than the phylogenies based on random set of signatures, even more so using the RF index than the PD index, according to a *z*-test comparing the means of the two sets of phylogenies for sample 1.

The bases that were used to capture genomic information as the digital signatures of 18 organisms from sample 1, 39 organisms from sample 2 and 17 bacteria from sample 3 are given as:

Table 1. Nxh chip designs used to obtain digital signatures

Name	Length of the probes	Number of Probes	T
3mE3-2at1.1	3	3	1.1
3mE4b-2at1.1	3	4	1.1
3mE4-2at1.1	3	4	1.1
4mP3-3at2.1	4	3	2.1
4m15-2at1.1	4	15	1.1
5mP6-3at2.1	5	6	2.1
5mP10-2at2.1	5	10	2.1
8mP10-4at4.1	8	10	4.1

Assessment of quality of the phylogenetic hypothesis

Multiple sequence alignments for COIs of organisms in sample 1 were obtained using Clustal Omega (McWilliam, 2013). Since the software **PAUP** package (Swofford, 2002) is the most commonly used by biologists for phylogenetic analyses, it was downloaded from *phylosolutions.com/paup-test* and then used (with default parameters) in order to generate phylogenies based on strict consensus majority consensus by Maximum Parsimony and Maximum likelihood, based on the Decision Theoretic Framework (DT) (Darriba, et al, 2012; Fungiflora & Gascuel, 2003). These hypotheses were used to assess the biological significance of the difference of the nxh chip based hypotheses from the Ground Truth.

Since the proposed methodology does not integrate molecular clocks, the quantitative analyses (using indices described in Chapter 2) of these hypotheses are solely based on the phylogenetic relationship between organisms excluding branch length in the hypotheses. In addition, qualitative analyses of the nxh chip based phylogenies were performed and are presented in Chapter 4. Biologists have defined the phylogenetic relationship among organisms based on the complexity of life, meaning that the organisms sharing the same degree of complexity of life should be closely related. The qualitative analyses of these nxh chip based phylogenies were primarily based on how accurately the phylogenetic relationships reflected what biologists generally accept about the development of life, particular concerning the biological complexity of the organisms involved.

Table 2. Nomenclature of representative sequences for phylogenetic reconstruction

Notation	Meaning
6Ks18	18 organisms from major 6 Kingdoms from Table 3
6Ks15	15 organisms from major 6 Kingdoms extracted from Table 3
6Ks13	13 organisms from major 6 Kingdoms extracted from Table 3

Table 2. (Continued)

Notation	Meaning
6Ks39	39 organisms from major 6 Kingdoms from Table 4
COIs	Sequences for Cytochrome c oxidase subunit 1
M+COIs	Sequences for Mitochondrial genome and COIs
MORF+COIs	Coding Sequences on Mitochondrial genome and COI sequences
ORFs	Coding Sequences on nuclear genome
bac5All	Sequences for whole genome of 17 bacteria in sample 3 in Table 5
bac517	17 organisms of the kingdom Bacteria shown in Table 5 for sample 3
Ang1000	The phylogeny based on composite (1000*Angle + Euclidean distance) between signatures of organisms as metric of similarity
Euc	The phylogeny based on Euclidean distance between signatures of organisms as metric of similarity
Ang	The phylogeny based on Angle between signatures of organisms as metric of similarity
Ward	Ward algorithm used for hierarchical clustering to compute phylogeny

Data Collection and Preprocessing

The data collection was performed in two phases. In the first phase, Cytochrome C Oxidase subunit I (COI)s, Mitochondrial genomes (when the sequences were available), Coding Sequences (ORFs) on Mitochondrial genome and the whole genome (when the sequences were available) were downloaded from NCBI, the National center for Biotechnology Information (Wheeler et al, 2007), and BOLD, the Barcode of Life Data System (Hebert et al., 2003) for the organisms shown in Table 3. The distribution of species across biological taxonomy at top two layers for sample 1 is shown in Figure 17. In addition, sequences for , Coding Sequences (ORFs) on Mitochondrial genome and Cytochrome Oxidase subunit I (COI)s were downloaded from NCBI (Wheeler et al, 2007) and BOLD database (Hebert et al., 2003) in Table 4. Finally, in the third phase, sequences for whole genome of 17 bacteria in Table 5 were collected as sample 3 from NCBI (Wheeler et al, 2007).

Table 3. Sample 1 (to Phylum level, downloaded from NCBI and BOLD Database, 2017)

Kingdom	Phylum	Organism	Notation
Animalia	*Arthropoda*	*Anopheles gambiae*	AnimArthAno
	Arthropoda	*Drosophila pseudoobscura*	AnimArthDro
	Arthropoda	*Locusta migratoria*	AnimArthLoc
	Chordates	*Homo sapiens*	AnimChorHom
	Chordates	*Mus musculus*	AnimChorMus
	Cnidaria	*Hydra vulgaris*	AnimCniHyd
Fungi	*Ascomycota*	*Saccharomyces castellii*	FungAscoSac
	Basidiomycota	*Schizophyllum commune*	FungBasiSch
	Mucorales	*Rhizopus oryzae*	FungMucoRhi
Plantae	*Magnoliophyta*	*Brassica napsus*	PlanMagnBra
	Streptophyta	*Arabidopsis thaliana*	PlanStrepAra
	Streptophyta	*Zea mays*	PlanStrepZea
Protista	*Euglenozoa*	*Euglena gracilis*	ProtEuglEug
Archaebacteria	*Euryarchaeota*	*Halobacterium salinarum*	ArchEuryHalB
	Euryarchaeota	*Haloquadratum walsbyi*	ArchEuryHalQ
	Euryarchaeota	*Natrinema pellirubrum*	ArchEuryNat
Eubacteria	*Proteobacteria*	*Escherichia coli*	EubacProtEsc
	Proteobacteria	*Photobacterium profundum*	EubacProtPho

Table 4. Sample 2 (to Phylum and Class level, downloaded from NCBI and BOLD Database, 2017)

Kingdom	Phylum/Class	Species	Notation
Animalia	*Artropoda/arachnida*	*Mesobuthus martensii*	AnimArtAraMesM
Animalia	*Artropoda/insecta*	*Anopheles gambiae*	AnimArtInsAnoG
Animalia	*Artropoda/insecta*	*Apis mellifera*	AnimArtInsApiM
Animalia	*Artropoda/insecta*	*Drosophila melanogaster*	AnimArtInsDroM
Animalia	*Artropoda/insecta*	*Drosophila pseudoobscura*	AnimArtInsDroP
Animalia	*Artropoda/insecta*	*Locusta migratoria*	AnimArtInsLocM
Animalia	*Artropoda/insecta*	*Zeugodacus cucurbitae*	AnimArtInsZeuC
Animalia	*Artropoda/Merostomata*	*Limulus polyphemus*	AnimArtMerLimP
Animalia	*Chordata/aves*	*Gallus gallus*	AnimChoAveGalG
Animalia	*Chordata/aves*	*Egretta garzetta*	AnimChoAveEgrG
Animalia	*Chordata/aves*	*Columba livia*	AnimChoAveColL
Animalia	*Chordata/mammalia*	*Homo sapiens*	AnimChoMamHomS
Animalia	*Chordata/mammalia*	*Mus musculus*	AnimChoMamMusM
Animalia	*Chordata/mammalia*	*Pan troglodytes*	AnimChoMamPanT
Animalia	*Chordata/mammalia*	*Sus scrofa (pig)*	AnimChoMamSusS
Animalia	*Chordata/Sauropsida*	*Malaclemys terrapin*	AnimChoSauMalT
Animalia	*Chordata/Sauropsida*	*Python bivittatus*	AnimChoSauPytB
Animalia	*Chordata/Sauropsida*	*Alligator mississippiensis*	AnimChoSauAllM
Animalia	*Cnidaria/Hydrozoa*	*Hydra vulgaris*	AnimCniHydHydV
Fungi	*Ascomycota/ Saccharomycotina*	*Saccharomyces cerevisiae*	FungAscSacSacC
Fungi	*Ascomycota/ Saccharomycotina*	*Candida albicans*	FungAscSacCanA
Plantae	Bryophyta/ Bryopsida	Physcomitrella patens	PlanBryBryPhyP

Table 4. (Continued)

Kingdom	Phylum/Class	Species	Notation
Plantae	*Magnoliophyta/Liliopsida*	*Zea mays*	PlanMagLilZeaM
Plantae	*Magnoliophyta/Liliopsida*	*Oryza sativa*	PlanMagLilOryS
Plantae	*Magnoliophyta/ Magnoliopsida*	*Arabidopsis thaliana*	PlanMagMagAraT
Plantae	*Magnoliophyta/ Magnoliopsida*	*Brassica napsus*	PlanMagMagBraN
Plantae	*Magnoliophyta/ Magnoliopsida*	*Beta vulgaris*	PlanMagMagBetV
Archae	*Euryarchaeota/Halobacteria*	*Haloquadratum walsbyi*	ArchEurHalHalW
Archae	*Euryarchaeota/Halobacteria*	*Haloarcula hispanica*	ArchEurHalHalH
Archae	*Euryarchaeota/Halobacteria*	*Halarchaeum acidiphilum*	ArchEurHalHalA
Eubacteria	*Chlamydiae/ Chlamydiales*	*Chlamydia trachomatis*	EubacChlChlChlT
Eubacteria	*Chlamydiae/ Chlamydiales*	*Chlamydia psittaci*	EubacChlChlChlP
Eubacteria	*Chlamydiae/ Parachlamydiales*	*Parachlamydia acanthamoebae*	EubacChlParParA
Eubacteria	*Firmicutes/Bacilli*	*Staphylococcus aureus*	EubacFirBacStaA
Eubacteria	*Proteobacteria/ Gammaproteobacteria*	*Salmonella enterica*	EubacProGamSalE
Eubacteria	*Proteobacteria/ Gammaproteobacteria*	*Acinetobacter baumannii*	EubacProGamAciB
Eubacteria	*Proteobacteria/ Gammaproteobacteria*	*Klebsiella pneumoniae*	EubacProGamKleP
Eubacteria	*Proteobacteria/ Gammaproteobacteria*	*Pseudomonas aeruginosa*	EubacProGamPseA

Table 5. Sample 3 (to Genus level, downloaded from NCBI and BOLD Database, 2016)

Name	Notation
Escherichia coli CFT073	E.coliCFT073
Escherichia coli K12	E.coliK12
Escherichia coli O15-7-H7 VT2 Sakai	E.coliO15-7-H7
Neisseria gonorrhoeae FA1090	Neis.gonorrhoeae
Neisseria meningitidis FAM18	Neis.meningitidis
Pseudomonas fluorescens Pf-5	Pseu.fluorescens
Pseudomonas entomophila L48	Pseu.entomophila
Pseudomonas aeruginosa PA01	Pseu.aeruginosa
Rickettsia felis URRWXCal2	Rick.felis
Rickettsia conorii Malish 7	Rick.conorii
Salmonella enterica Paratyphi ATCC9150	Sal.entericaP
Salmonella typhimurium LT2 SGSC1412	Sal.typhimurium
Salmonella enterica Choleraes Plasmid 50	Sal.entericaC
Shigella boydii Sb227	Shig.boydii
Yersinia pestis KIM	Yers.pestisK
Yersinia pestis Antiqua	Yers.pestisA

After collecting the sequences, they were preprocessed. Each sequence in all the samples was cleansed to remove nonDNA symbols (such as fasta annotations) present in the sequences, shredded into fragments of length *n* equal to the length of the probes on the nxh chips, and finally, pmer counts were obtained using Perl code. A pmer count represents the frequency of all possible *n*-mer oligos and their Watson-Crick complement in a sequence.

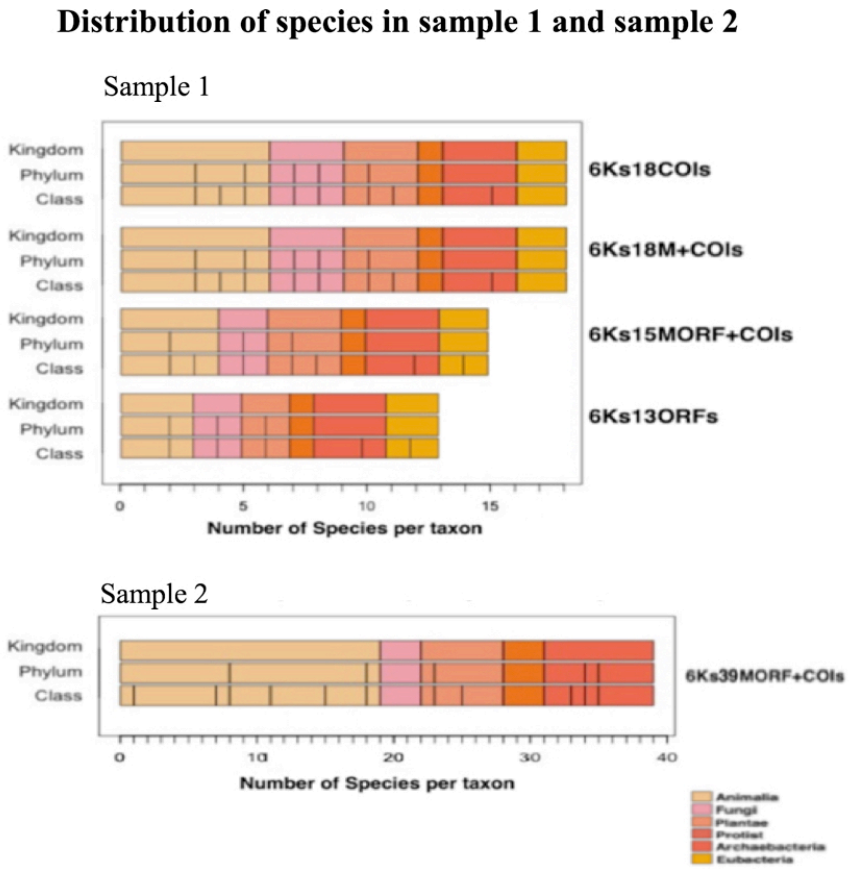

Figure 17. Distribution of species per taxon in top three levels of the taxonomy for organisms in samples 1 and 2, for the sequences in the far right. The width of each rectangle is proportional to the number of species in the taxa belonging to the kingdom, phylum or class, as indicated in each taxon.

Results

The preprocessed sequences were analyzed to obtain phylogenies. Firstly, the pmer counts were used to obtain signatures for the organisms using Perl software. Then, R code was used to obtain visualizations for digital signatures as well as dendrograms based on different metrics of similarity between these signatures. Three simple metrics were used, namely Angle, Euclidean distance, and composites of Angle and Euclidean distance (Ang1000) between the digital signature vectors. Other more sophisticated metrics such as Contrast (Garzon & Wong, 2011) were also computed for similarity metrics between these signatures. However, the analyses showed that most of the phylogenies based on Contrast metric were caterpillar trees, so they are not discussed below.

These phylogenies obtained by the *h*-distance were then compared against the OTL Ground Truth. Both qualitative and quantitative analyses were performed. Quantitative analysis was performed using modules available in the *ape* (Paradis et al, 2004) and *phangorn* (Schliep, 2010) packages in R on the RF and PD indices. These analyses were performed at three levels – Phylum across kingdoms, Class across phyla and genera in a given kingdom (*bacteria*). The heatmaps show digital signatures alongside the phylogenies on nxh chips for samples 1 and 2 (in subsection discussing qualitative analyses) to give a sense of what the signatures would look like when sequences are processed. (The raw counts shown in the heatmaps have been slightly altered throughout in order to preserve intellectual property of the sequences used as probes in the nxh bases. However, the color codes in the heatmaps are accurate and these numbers are close enough to give an good idea about the actual order of magnitude of the number of shreds that would stick to the probes on the nxh chips).

Phylogenetic Analysis at Phylum level

The phylogenies obtained using *h*-distance were compared against the Ground Truth extracted from *Open Tree of Life* (Hinchliff et al., 2015) at the phylum level for sample 1. The findings of quantitative and qualitative assessments are shown next.

Quantitative Assessment of the Phylogenies

Most of the time, the findings from quantitative analyses turned out to be in fairly close agreement with qualitative analyses obtained by conventional methods in biology on the same choice of the biomarker(s), the sequences and the metric(s) for phylogeny construction. As shown in Figure 18, COIs do not appear to be a good universal biomarker for phylogenetic reconstruction, according to the RF index of similarity. However, PD indices show that some of the phylogenies based on these sequences are of comparable quality to the customary phylogenies produced from PAUP based on COIs.

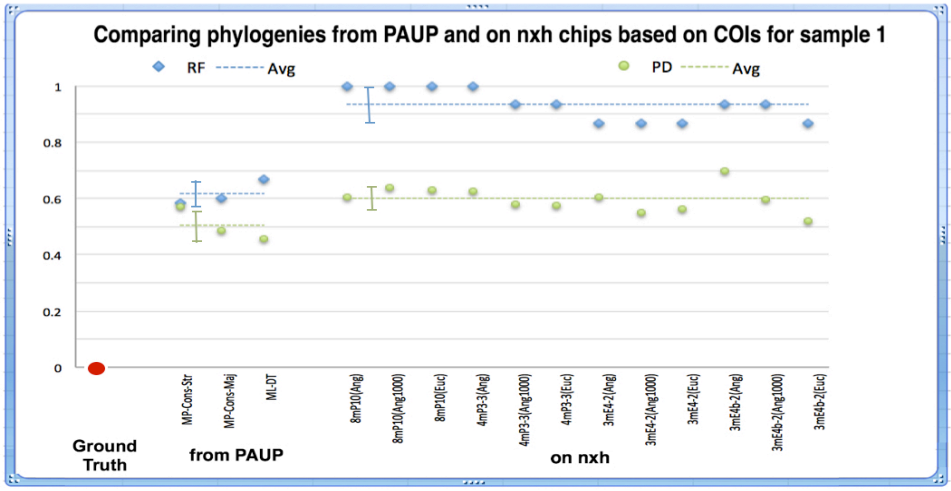

Figure 18. Quantitative assessments, based on RF and PD indices, of the phylogenies on nxh chip for 18 organisms in sample 1 based on COIs. PD indices show that some of these phylogenies are statistically just as good as phylogenies generated from PAUP, the conventional method used by biologists.

However, the assessment shown in Figure 19 shows that, on average, accuracy increases (i.e., indices decrease) when M+COIs were used instead of COIs for phylogenetic inference. RF indices on nxh chip-based phylogenies were still higher than those on PAUP based phylogenies, however, PD indices on nxh chip-based phylogenies were significantly lower than those on PAUP based phylogenies. In order to determine the statistical significance of this difference on PD index, a *t*-test was performed. The alternative hypothesis used for the test was "The means of PD indices on phylogenies from PAUP and those on nxh chips are equal." The null hypothesis was the negation of the research hypothesis. As a result, the test showed that the null hypothesis was rejected. Therefore the differences are statistically significant.

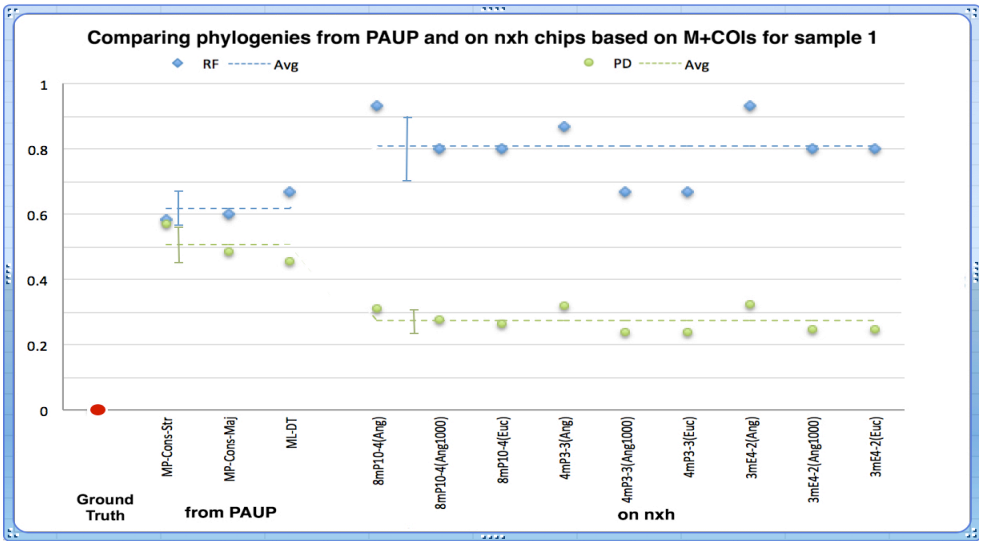

Figure 19. Quantitative assessments, based on RF and PD indices, of the phylogenies on nxh chip for 18 organisms in sample 1 based on M+COIs. RF indices show that few phylogenies on nxh chips are comparable to those from PAUP. However, PD indices show that on average, the phylogenies on nxh chips are statistically significantly closer, on average index, to biological Ground truth than the phylogenies from PAUP, according to a *t*-test performed on comparison of the means of the two samples. On the other hand these phylogenies are comparatively closer to the Ground Truth (lower indices) than those based on COIs (Fig. 18.)

In turn, significant improvement was observed in both the RF and PD indices between the Ground Truth and phylogenies on nxh chips when M+COIs markers were replaced with the

combination of signatures from ORFs of full Mitochondrial sequences and COIs, as shown in Figure 20. In addition, statistically similar results were obtained when MORF+COIs were replaced with nuclear ORFs. Although the PD indices increased slightly, there was small variation between the indices on nxh chip-based phylogenies as shown in Figure 21. A similar one-sided *t*-test was performed to determine statistical significance of the differences of averages on PD indices between phylogenies on nxh chips and those from PAUP on both of the dataset i.e., MORF+COIs and ORFs. The alternative hypothesis used for the test was "The difference between the means in PD indices on phylogenies from PAUP and those on nxh chips is positive." The null hypothesis was the negation of the alternative hypothesis. As a result, the test showed now that there is sufficient evidence to reject the null hypothesis.

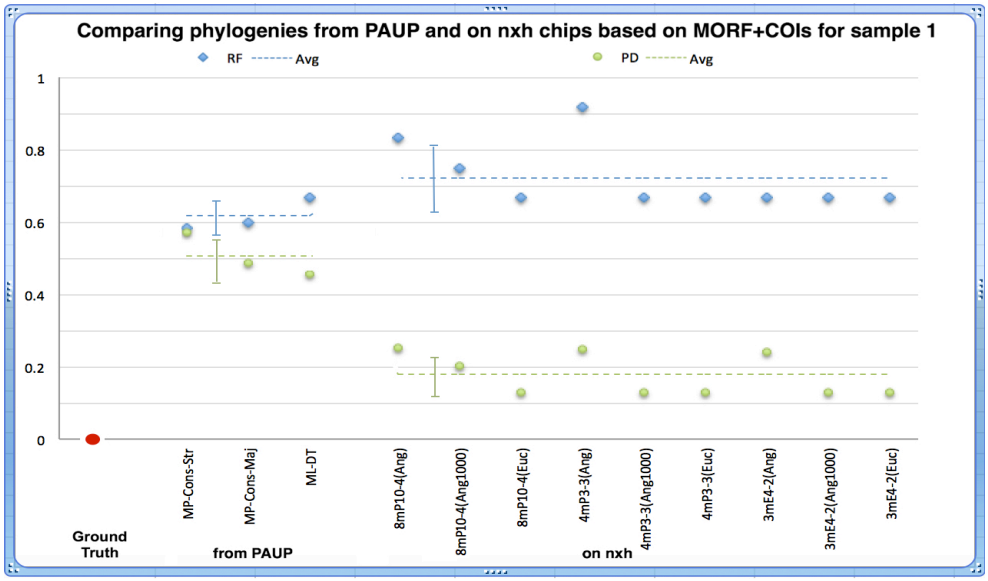

Figure 20. Quantitative assessments, based on RF and PD indices, of the phylogenies obtained by the *h*-distance method for 15 organisms in sample 1 based on MORF+COIs. PD indices show that the phylogenies on nxh chips are statistically significantly closer, on average index, to biological Ground truth than the phylogenies from PAUP, according to a *t*-test performed on comparison of the means of the two samples. On the other hand, RF indices show that phylogenies from PAUP are better than those on nxh chips even though some of them become comparable to the former ones. However, phylogenies on nxh chips based on MORF+COIs are significantly closer to the Ground Truth than those based on M+COIs in Fig. 19.

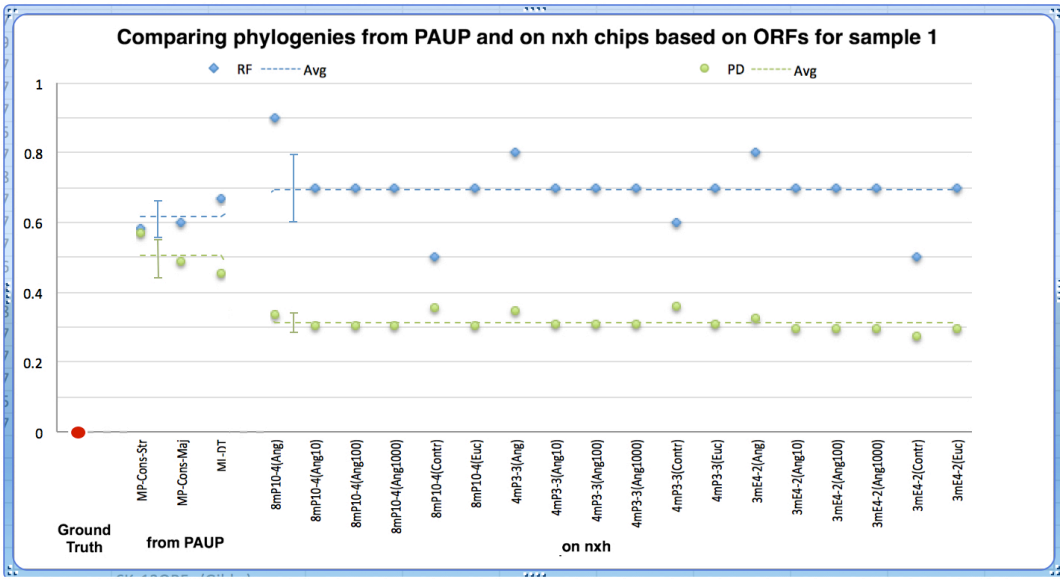

Figure 21. Quantitative assessments, based on RF and PD indices of the phylogenies on nxh chips for 13 organisms in sample 1 based on ORFs. RF indices show most of these phylogenies are statistically just as good as phylogenies using PAUP. However PD indices show that the phylogenies on nxh chips are statistically significantly closer, on average index, to the Ground truth than the phylogenies from PAUP, according to a *t*-test performed on comparison of the means of the two samples.

The question arises as to how significant these values may be for the accuracy of the phylogenies in terms of phylogenetic relationships. In order to answer that question, a few changes were made to the Ground Truth tree and new indices were computed between the Ground Truth and the altered Ground Truth. The process was repeated 14 times. From Figure 22, it is evident that PD indices between the Ground Truth and the phylogenies on nxh chips-based on MORF+COIs were comparable to indices between the Ground Truth and the altered Ground Truth when one or two changes are made to the Ground Truth tree. Therefore, the difference is due to a failure to capture only very few genetic links between the target organisms.

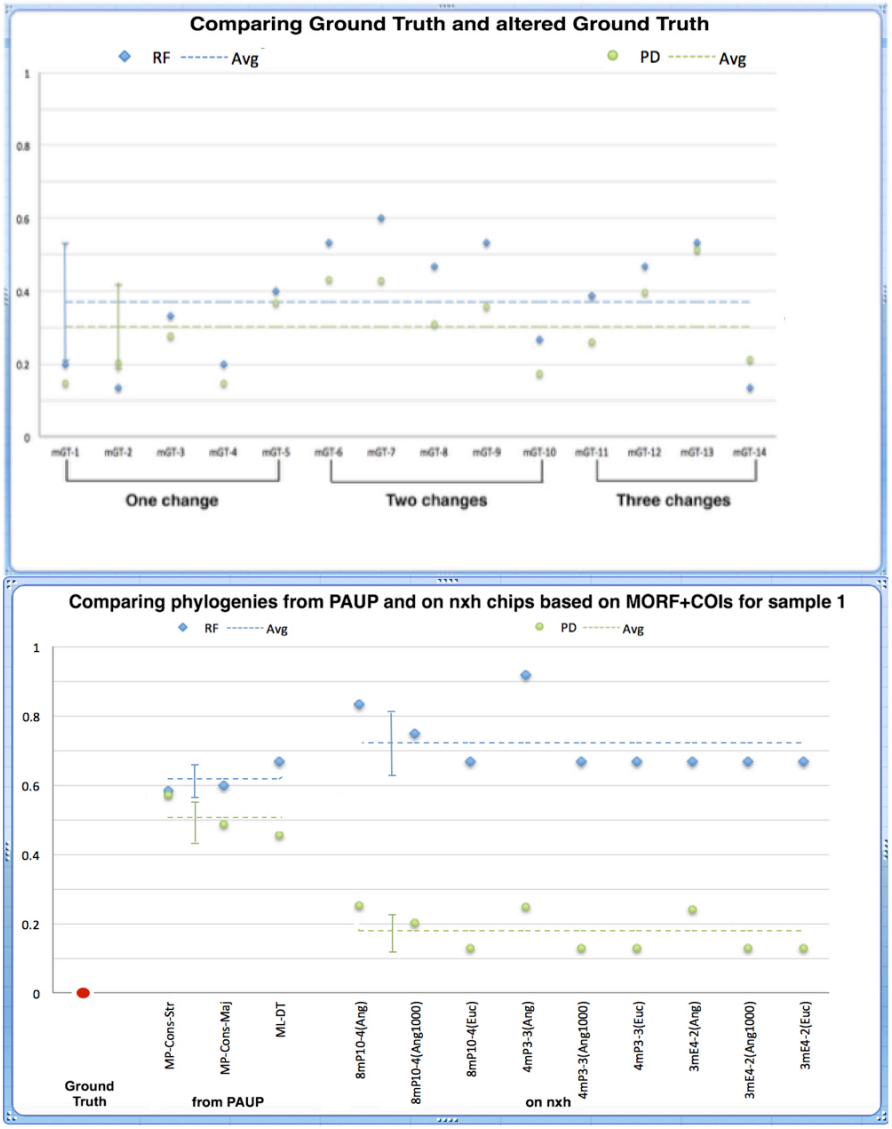

Figure 22. Assessment of the biological significance of the index differences in Figure 20 based on RF and PD indices. RF and PD indices were computed for systematic modifications of the Ground Truth (mGTs, top figure) and the phylogenies on nxh chips for 15 organisms in sample 1 based on MORF+COIs in Figure 20 (bottom figure). This chart shows that the score difference amounts to only one or two branches out of 9 branches in the Ground Truth phylogeny. That puts the accuracy of the *h*-distance method above the 85% range in terms of accuracy in branches of the Ground Truth.

Qualitative Assessment of the Phylogenies

The phylogenies based on COIs were also compared against the Ground Truth more qualitatively from the biological standpoint. In Figure 23 and Figure 24 phylogenies on nxh chips 4mP3-3at2.1 and 8mP10-4at4.1 using Ang1000 and Euclidean distance for similarity

metrics were compared against the Ground Truth. From Figures 23 and 24, it is evident that COIs were not suitable biomarkers to perform phylogenetic analyses when organisms were too diverse from one another. In Figure 23, even though most of the organisms in *Animalia* are shown to be in same clade, misclassifications such as *Archae* and *Eubacteria* being grouped with plants and animals are very critical. The situation is same with the phylogeny on basis 8mP10-4at4.1 in Figure 24.

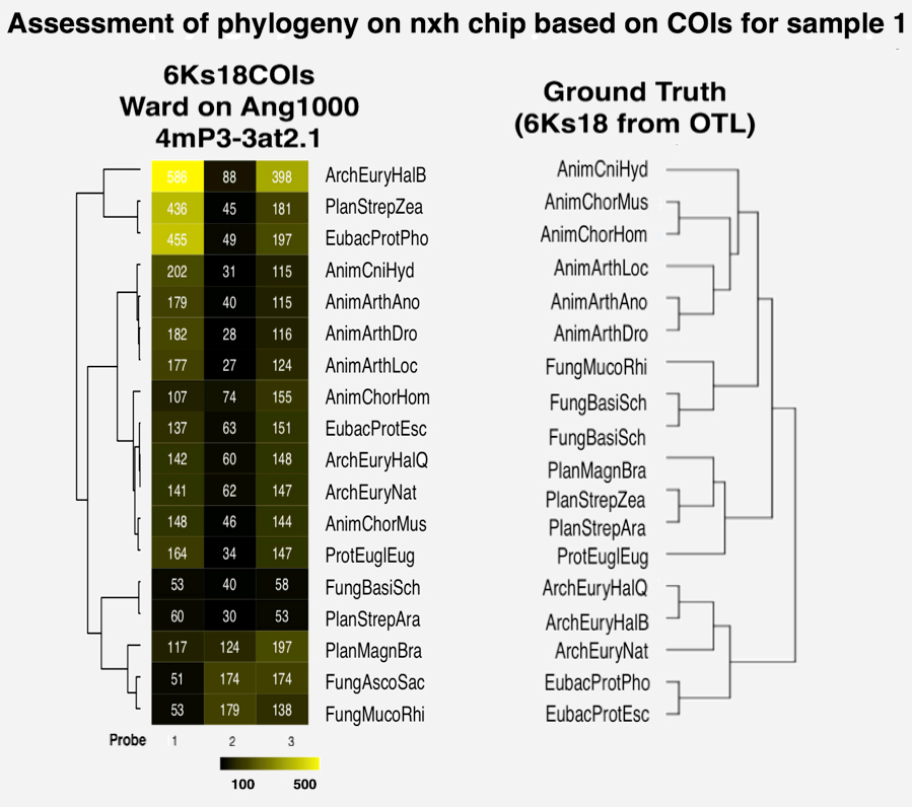

Figure 23. Left: Assessment of the phylogeny on nxh chip 4mP3-3at2.1, based on COIs for organisms in sample 1, using similarity by Ang1000. Right: Ground Truth for the same sample 1. Even though most of the *Animalia* are grouped together, this phylogeny shows some serious misclassifications, as *Fungi* and *Plantae* are grouped together, and *Archae* and *Eubacteria* are separated from their clade and grouped with *Animalia* and *Plantae*.

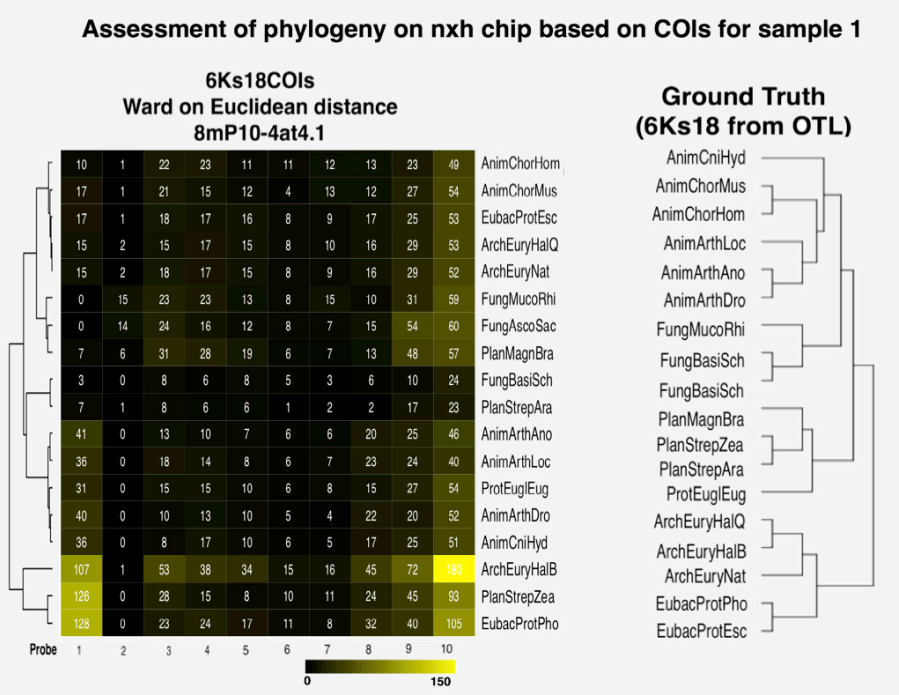

Figure 24. Left: Assessment of the phylogeny on nxh chip 8mP10-4at4.1, based on COI for organisms in sample 1, and using Euclidean distance for similarity metric. Right: Ground Truth for the same sample 1. Like phylogeny on nxh chip in Figure 23, this phylogeny shows some serious misclassification; for example, *Archae* and *Eubacteria* are grouped with *Animalia* and *Fungi*.

The findings from this analysis raise an important question: are COIs too short for their signatures to capture enough information for phylogenetic reconstruction? In order to answer that question, the sequences for all the organisms, except organisms in kingdoms *Eubacteria* and *Archae*, were replaced by whole mitochondrial genomes. The analysis in Figures 25 and 26 show that the result got much better when M+COIs were used. The phylogenies are now grouping more biologically similar organisms together than the phylogenies with COIs. However, the inability of clustering algorithm to make clear distinction between *Animalia*, *Fungi* and *Protist* was still problematic to ignore from the biological standpoint.

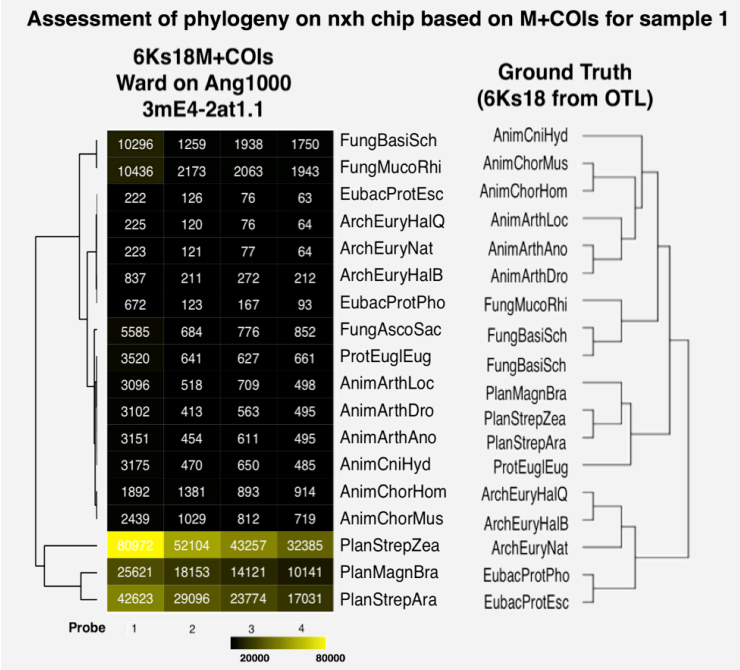

Figure 25. Left: Assessment of the phylogeny obtained on nxh chip 3mE4-2at1.1, based on M+COIs for organisms in sample 1, using Ang1000 for similarity metric. Right: Ground Truth for the same sample 1. Qualitatively, more accurate hypothesis can be gained using M+COIs than using COIs.

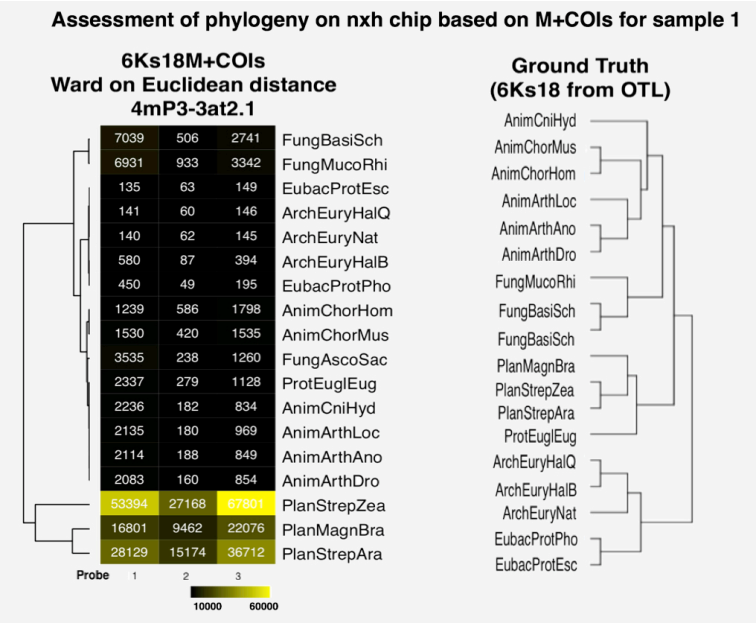

Figure 26. Left: Assessment of the phylogeny on nxh basis 4mP3-3at2.1, based on M+COIs for organisms in sample 1, and using Euclidean distance for similarity metric. Right: Ground Truth for the same sample 1. Qualitatively, more accurate hypothesis can be gained using M+COIs than using COIs using this similarity metric as well.

In an attempt to get better results, the coding sequences of mitochondrial genome and COIs were used to obtain signatures for 15 organisms in sample 1. Then, phylogenies based on those signatures were produced. As shown in Figures 27 and 28, phylogenies on bases 8mP10-4at4.1 and 4mP3-3at2.1 using Euclidean distance and Ang1000 for similarity metrics were compared against the Ground Truth. Although we could still experience a few misclassifications at the phylum level such as *Fungi* were split, the trees showed a clear distinction between organisms when coding sequences are used instead of whole genomes. Qualitatively, the feasibility of both hypotheses on Figures 27 and 28 appear to be the same.

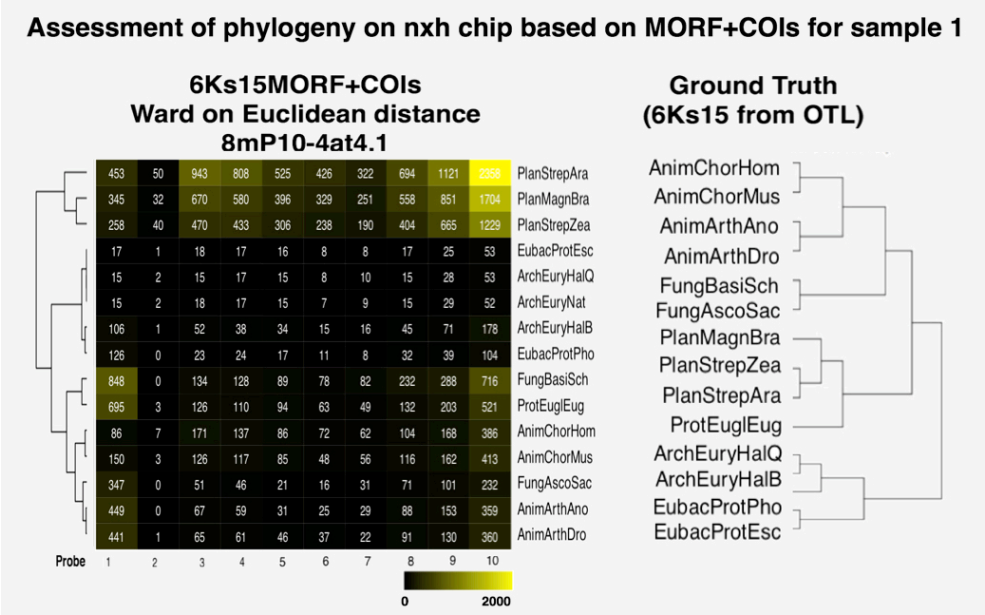

Figure 27. Left: Assessment of the phylogeny on nxh chip 8mP10-4at4.1, based on MORF+COIs for organisms in sample 1, and using Euclidean distance for similarity metric. Right: Ground Truth for the same sample 1. From this assessment, it is evident that introns in whole mitochondrial genomes can create noise in digital signatures for phylogenetics.

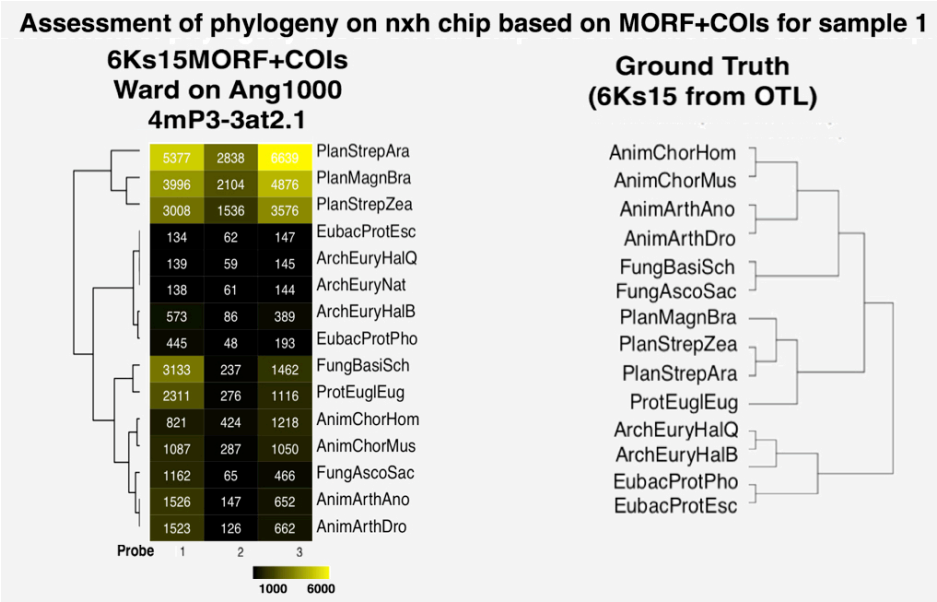

Figure 28. Left: Assessment of the phylogeny on nxh chip 4mP3-3at2.1, based on MORF+COIs sequences for organisms in sample 1, and using Ang1000 as for similarity metric. Right: Ground Truth for the same sample 1. From this assessment, it appears that introns in whole mitochondrial genome were creating noise in the signatures. Furthermore, this phylogeny seems to be depicting a similar evolutionary trend as depicted by the phylogeny on chip 8mP10-4at4.1 in Figure 27.

In conclusion, it can be said that the introns present in full genome were likely introducing noise in the signatures that lead to misclassification and trees with shorter tips. If so, phylogenies based on coding sequences in the whole nuclear genome should be better than those based on the whole genome. The problem is that preprocessing whole genomes *in silico* to obtain signatures of these organisms were expected to be computationally challenging for organisms such as *Homo sapiens*, *Brassica napsus* and many more. Nevertheless, this methodology indeed allowed phylogenies based on coding sequences in whole nuclear genome to be computed and compared against the Ground Truth.

As shown in Figures 29 and 30, some serious misclassifications occurred in phylogenies based on the bases 3mE4-2at1.1 with Ang1000 and 8mP10-4at4.1 with Euclidean distance for similarity metric, such as placing *Mus musculus* and *Archae* in the same clade. To be fair, there

were peculiarities present in the sequences. For example, only one mitochondrial genome was available for *Protist*, so this sequence was used in all analyses; sequences for some of the organisms contained long sequence of nonDNA characters; and, not all the organisms have mitochondria/nucleus. Therefore the question remains unanswered whether the error is due to the lack of evidence in the biomarkers, or is a shortcoming inherent to this methodology.

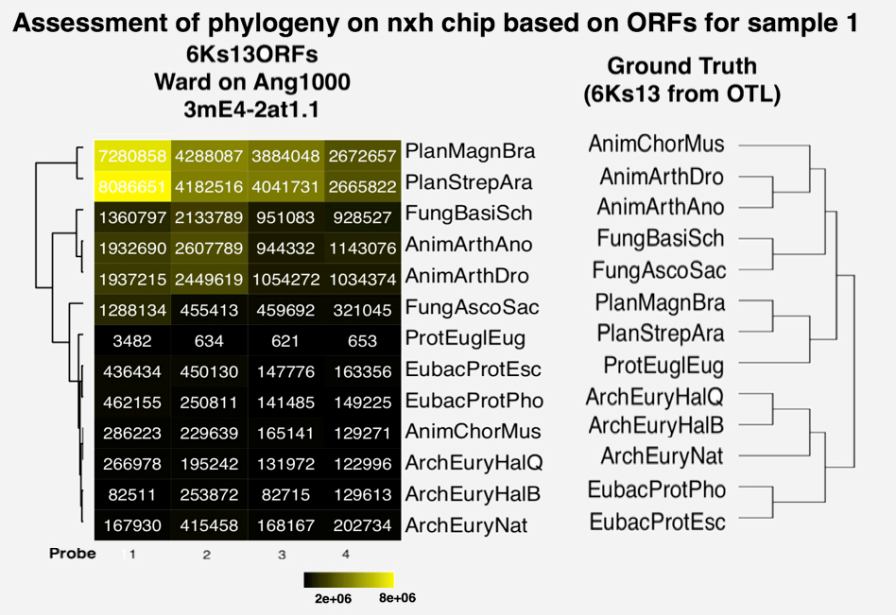

Figure 29. Left: Assessment of the phylogeny on nxh chip 3mE4-2at1.1, based on ORFs for organisms in sample 1, and using Ang1000 for similarity metric. Right: Ground Truth for the same sample 1. The phylogeny on the nxh chip shows some serious misclassification of grouping *Animalia* with *Archae* and separating two *Fungi.*

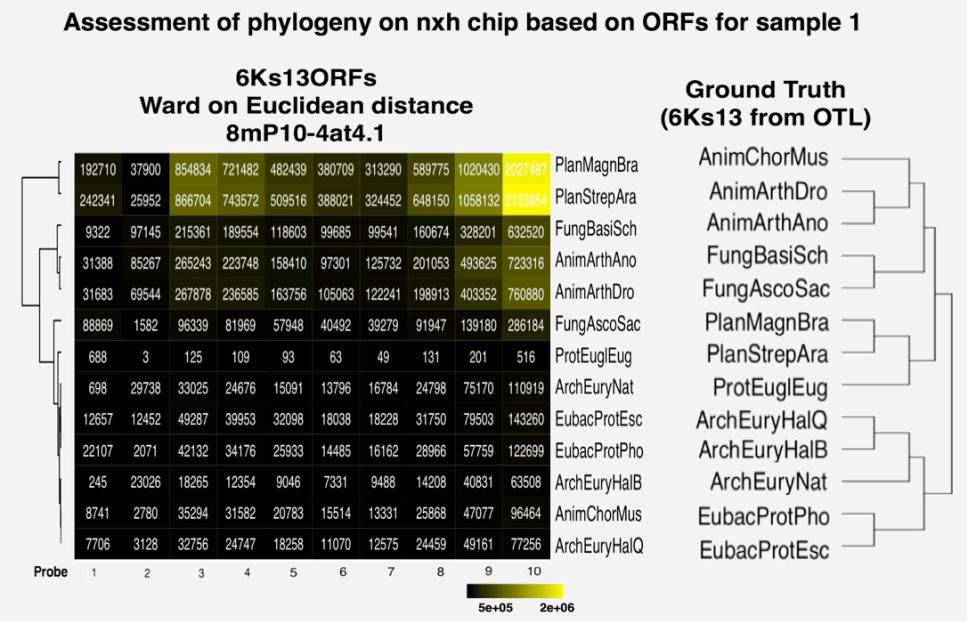

Figure 30. Left: Assessment of the phylogeny on nxh chip 8mP10-4at4.1, based on ORFs for organisms in sample 1, using Euclidean distance for similarity metric. Right: Ground Truth for the same sample 1. The phylogeny on the nxh chip shows some serious misclassifications, such as grouping one species of *Archae* and one species of *Animalia* in the same branch.

Phylogenetic Analysis at Class level

The phylogenies obtained using *h*-distance method were compared against the Ground Truth, the Ground Truth extracted from *Open Tree of Life* (Hinchliff et al., 2015) at the class level. The findings of quantitative and qualitative assessments are shown next.

Quantitative Assessment of the Phylogenies

When analyzed quantitatively at class level, both indices are significantly larger than the phylogenies at phylum level in Figure 20. The inclusion of more phylogenetic links for comparison was responsible for this change. However, both indices had much better agreement across various bases and metrics between signatures to produce phylogenies closer to the Ground Truth, as shown in Figure 31. On one hand, PD indices indicate phylogenies based on the basis 4mP3-3at2.1, using either Ang1000 or Euclidean distance for similarity metric, were

significantly closer to the Ground Truth. On the other, RF indices show that all the phylogenies on all nxh chips are comparable to the Ground Truth, except for phylogeny on 8mP10-4at4.1 using Angle for similarity metric.

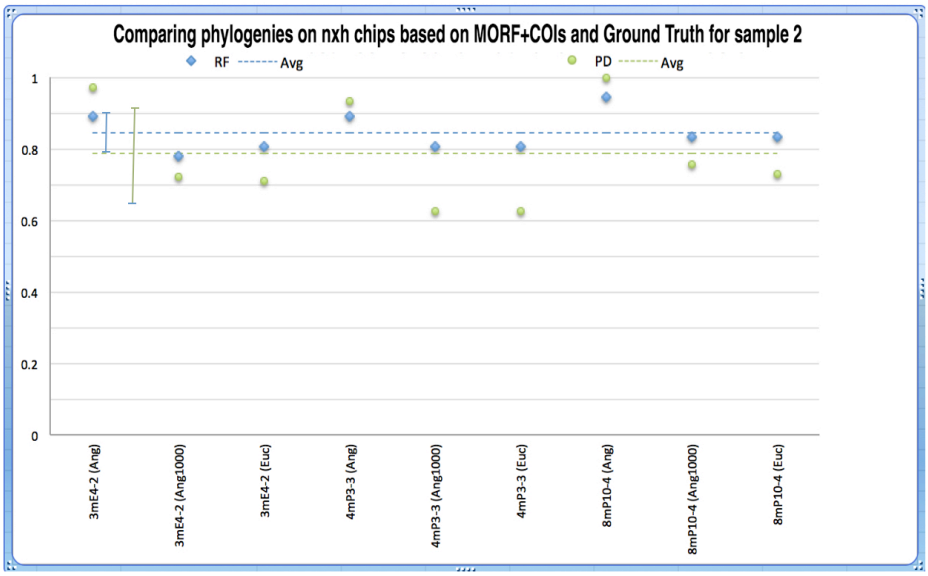

Figure 31. Quantitative assessment phylogenies on nxh chips, based on the RF and PD indices, for 39 organisms in sample 2 based on MORF+COIs. Both indices are significantly larger than in phylogenies at the phylum level in Figure 20. However, PD indices show that phylogenies on the nxh chip 4mP3-3at2.1 using either Ang1000 or Euclidean distance for similarity metric are significantly closer to the Ground Truth. On the other hand, RF indices show that the phylogenies on all nxh chips are comparable to the Ground Truth, except for phylogeny on 8mP10-4at4.1 using Angle for similarity metric.

Qualitative Assessment of the Phylogenies

The better phylogenies according to indices as shown in Figure 31 were analyzed qualitatively, with respect to the complexity of life as used by most of biologists. As shown in Figure 32 and Figure 33, most of the phylogenetic relationships represented in the Ground Truth were reproduced at the class level in phylogenies on nxh chips. There were some concerns such as *Fungi* being split and variously grouped with plants and animals and species in the *Sauropsida* class were separated as shown in Figure 32. The organisms in the *Sauropsida* class were grouped

together in the phylogeny on the nxh chip shown in Figure 33; however, *Fungi* and *Hydra* were still separated from their clade.

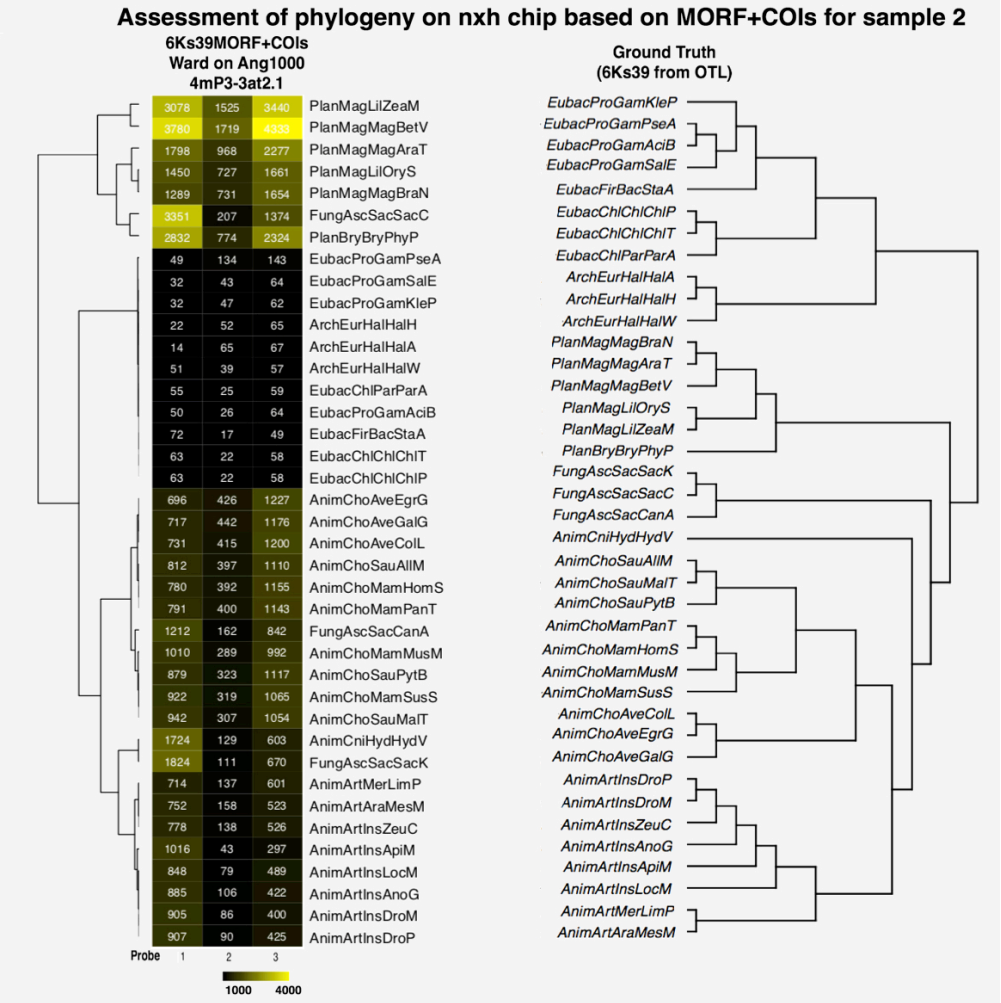

Figure 32. Left: Assessment of the phylogeny on nxh chip 4mP3-3at2.1, based on MORF+COIs for organisms in sample 2, and using Ang1000 for similarity metric. Right: Ground Truth for the same sample 2. Almost all the phylogenetic relationships were reproduced, except for a couple of misclassifications where organisms in the class *Sauropsida* and phylum *Fungi* were separated from their clades.

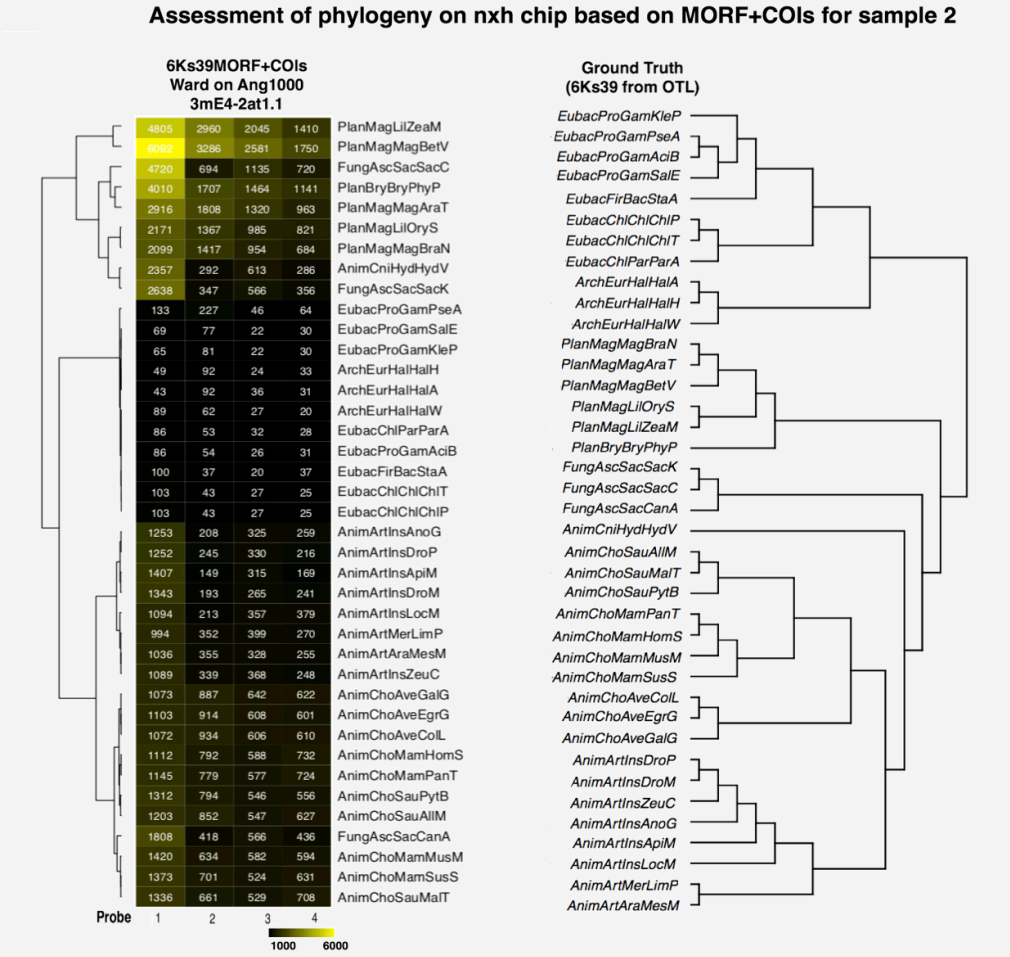

Figure 33. Left: Assessment of the phylogeny on nxh chip 3mE4-2at1.1, based on MORF+COIs for organisms in sample 2, and using Ang1000 for similarity metric. Right: Ground Truth for the same sample 2. In contrast to phylogeny on the nxh chip 4mP3-3at2.1 in Figure 32, all the organisms in the class *Sauropsida* were grouped together. However, organisms such as *Fungi* and *Hydra* were separated from their clade.

Phylogenetic Analysis at Genus level in Bacteria

The sequences of whole genomes for 17 *bacterial* organisms as shown in Table 5 for sample 3 were downloaded from NCBI (Wheeler et al, 2007). The *16S rRNA* phylogeny generated by the CSRS method (Garzon and Wong, 2011) was considered as the Ground Truth for these 17 *bacteria*. The findings of quantitative and qualitative assessments are shown in the following two sections.

Quantitative Assessment of the Phylogenies

The quantitative assessment for *h*-distance based phylogenies for 17 *bacteria* using whole genome is shown in Figure 34. All the phylogenies produced using different metrics but the same basis were statistically identical. Furthermore, both indices show that phylogenies on the basis 4m15-2at1.1 are closest to the Ground Truth. Apart from those phylogenies, both indices indicate that phylogenies on the bases 5mP10-2at2.1 and 5mP6-3at2.1 appear closer to the Ground Truth.

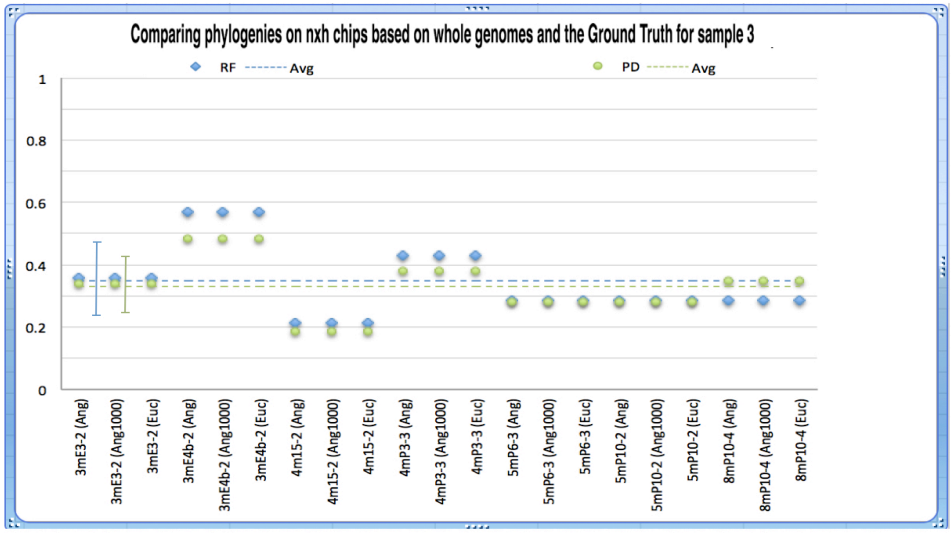

Figure 34. Quantitative assessment of phylogenies on nxh chips, based on the RF and PD indices, for 17 *bacteria* in sample 3 based on whole genome. Both indices show that phylogenies on the basis 4m15-2at1.1 are closest to the Ground Truth. Additionally, both indices also indicate that phylogenies on bases 5mP10-2at2.1 and 5mP6-3at2.1 are also statistically closer to the Ground Truth, regardless of the similarity metric used.

Qualitative Assessment of the Phylogenies

Qualitative assessment was also done for the phylogenies that appeared to be closer to the Ground Truth by the indices in Figure 34 according to the quantitative analysis. When the whole genome was used to construct the phylogenies for 17 *bacteria* in sample 3, bases 4m15-2at1.1 and 5mP6-3at2.1 produced phylogenies closer to the Ground Truth in (Garzon and Wong, 2011), as shown in Figure 35 and Figure 36. The phylogenies on 4m15-2at1.1 and 5mP6-3at2.1,

exactly reproduced the phylogenetic relationship in the Ground Truth at the genus level, except for a minor difference in the position of one species of *Salmonella*. However, the position of this species seems to be contradictory even in the Ground Truth. By contrast, as shown in Figure 33, the phylogeny on 5mP6-3at2.1 clustered all the species in the same genus closer, even all the species of *Salmonella* are in one branch. This phylogeny thus seems to be more convincing than the phylogeny shown in the Ground Truth.

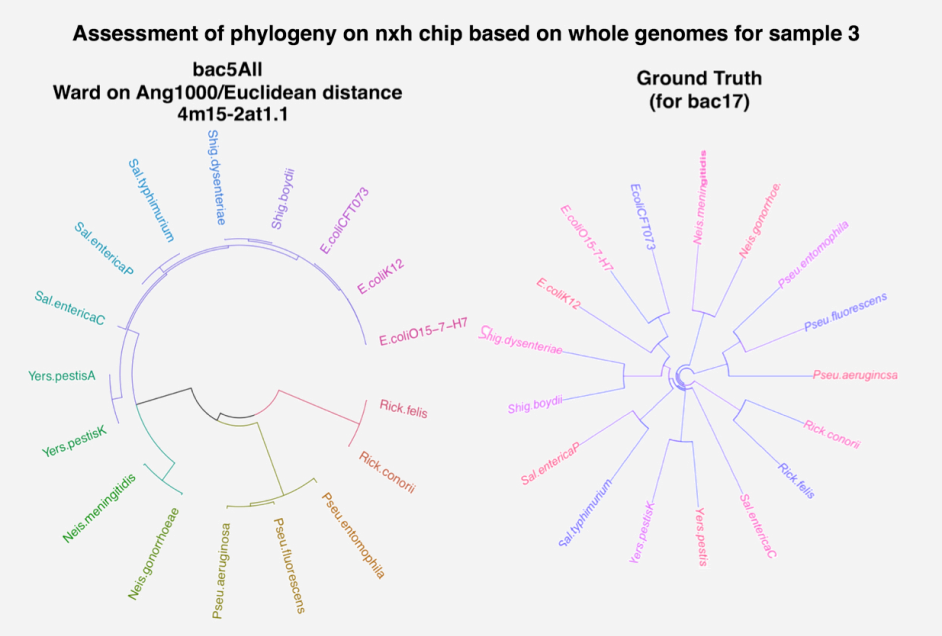

Figure 35. Left: Qualitative assessment of the phylogenies on the nxh chip 4m15-2at1.1 for 17 *bacteria* in sample 3 based on whole genome sequences using Ang1000 for similarity metric Right: The Ground Truth for the same sample is a *16S rRNA* tree generated by the CSRS method presented in (Garzon and Wong, 2011). The phylogeny on chip 4m15-2at1.1 exactly reproduced the phylogenetic relationship at the genus level in the Ground Truth, except for a minor difference in the position of one species of *Salmonella*.

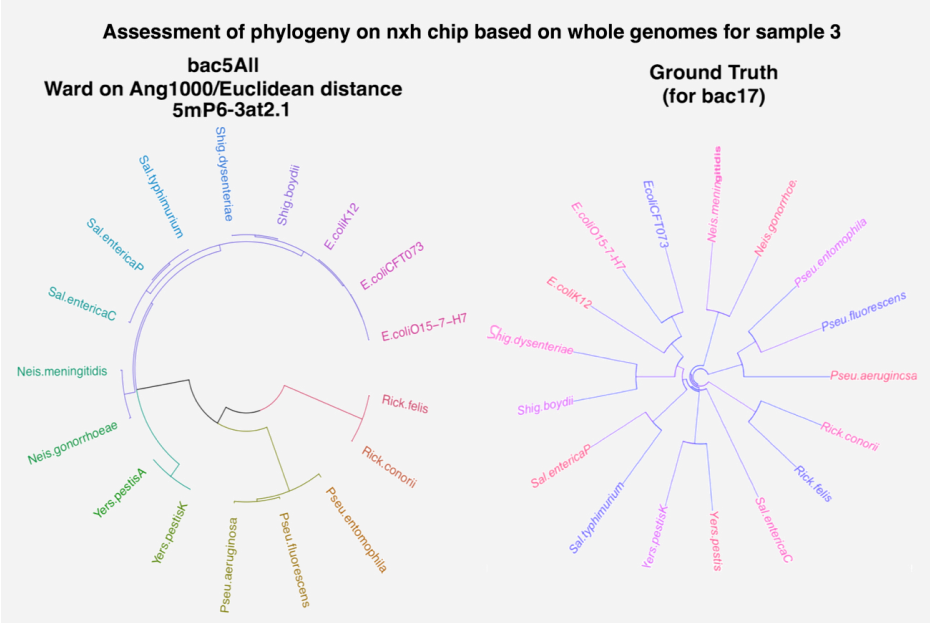

Figure 36. Left: Qualitative assessment of the phylogenies on the nxh chip 5mP6-3at2.1 for 17 *bacteria* in sample 3 based on whole genome. Using Ang1000 and Euclidean for similarity metric results in the same phylogeny. Right: The Ground Truth for the same sample 3 is again the *16S rRNA* tree generated by the CSRS method in (Garzon and Wong, 2011). The phylogenies on chip 5mP6-3at2.1 exactly reproduce all the phylogenetic relationships at the genus level in the Ground Truth, with a minor difference in the position of one species of *Salmonella*. However, this difference led to the grouping of all species in *Salmonella* in the same branch, which is biologically more convincing than the Ground Truth.

Conclusions and Future Work

Biological Significance of the Genomic Methods

Constructing a universal tree of life that could cover the entire biome has been a major goal for many biologists since Darwin proposed his theory of evolution. All attempts in that direction point towards the fundamental question in phylogenetics: where did all the organisms at the current biome come from? A reasonable approach to address the issue would be to formulate a phylogenetic hypothesis to estimate the true evolutionary TOL, with some supporting evidence. Biologists have been working over 100 years to formulate such hypothesis. The *Open Tree of Life* (Hinchliff, et al, 2015) is a most representative and systematic integrated hypothesis of such comprehensive efforts. This phylogeny was considered as the Ground Truth to serve the goal of phylogeny evaluation in this thesis.

A new methodology was introduced that produces such hypotheses as estimations of the Ground Truth in biology that bear strong supporting evidence of their validity. This method is based on the selection of a universal set of biomarkers (basis), meaning that a change in the target set of organisms will not require a change in the markers. This approach makes possible the construction of a universal tree of life *ab initio*, and possibly encompassing the entire biome. Furthermore, unlike the conventional molecular methods, multiple sequence alignments for phylogeny reconstruction are no longer required. This advantage completely removes dependency of a phylogenetic hypothesis on the order of sequence alignments, which presumably, leads to the construction of more accurate and stable phylogenies. In addition, signatures using longer sequences, even of whole genomes, can now be computed, meaning more genomic information can now be used for better phylogenetic hypothesis formulation.

However, it was observed that ORF sequences may produce better phylogenies than full sequences including introns.

Using this methodology, specific phylogenetic hypotheses are proposed in this thesis and both quantitative and qualitative assessments were made to investigate their biological significance by comparing those hypotheses against the biological Ground Truth. In the first pass, sequences such as COIs, mitochondrial genomes, ORFs on mitochondrial genomes and nuclear genomes were used to devise the phylogenetic hypotheses for organisms in sample 1. Phylogenies on bases 3mE4-2at1.1 and 8mP10-4at4.1 are closer to the Ground Truth based on ORFs sequences using Ang1000 and Euclidean distance for similarity metrics. In a second pass, sequences for ORFs on mitochondrial genome for organisms in sample 2 were analyzed to produce hypotheses at the class level. The nxh chips 3mE4-2at1.1 and 4mP3-3at2.1 produced better estimates using the Ang1000 for similarity metric. Finally, in a third pass, whole genomes of 17 *bacteria* in sample 3 were analyzed to produce hypotheses at the genus level. For *bacterial* phylogenies, the selection of similarity metric did not affect the quality of the estimation with respect to the Ground Truth. Bases 4m15-2at1.1 and 5mP3-6at2.1 produced good hypotheses. It is likely that larger nxh bases based on longer markers (such as 12- and 16-mers) may produce more accurate phylogenies. However, finding these bases is a difficult **NP**-complete problem (Garzon, 2012; Garzon and Bobba, 2012). It is also likely that when better quality sequences are homogenously available and not mixed with other sequences (e.g., MORFs with COIs), this methodology gives more accurate estimates of the Ground Truth.

Future Work

The major question that arises after investigating all the findings is how powerful this methodology is to estimate the Gold Standard to the maximum degree of accuracy. It is a fact in

this thesis that this methodology has so far failed to reconstruct the exact Ground truth hypothesis that biologists consider to be their most accurate estimate of the Gold Standard. However, most of the phylogenetic links reproduced the Ground Truth at phylum (around 66%), class level (around 70%) and genus level (over 90%) to a high percentage of accuracy in the branches. As mentioned in the Introduction, it is well known that molecular data alone cannot possibly fully reconstruct phylogenetic relationships simply because these are impacted by other factors (such as mutations, environment, geography) having a significant influence on the evolution of an organism. From this standpoint, it is then remarkable that this methodology is capable of inferring genetic relationships *on genomic data al*one that have been obtained by combination of many other (including nonmolecular) means.

A second question concerns implementation. Although technology is available to implement the proposed methodology, there remain experimental challenges to its translation to a wet lab to see how effective these models become when applied massively to the entire biome in real life. Hence, a major question is, after selecting a suitable hybridization threshold τ, how can we enforce this τ on real DNA chips operating *in vitro*? Another major challenge is target shredding. *In silico*, we can obtain perfect shredding, but in practice, technologies such as sonication and cleaving will not produce all the shreds of equal desirable uniform length *in vitro*. Given the theoretical foundation behind this work and the dynamics of actual hybridization *in vitro*, it is conceivable that this methodology might just prove robust enough in the wet lab to some variability in the length of the shreds.

References

Burridge, A. K., Hörnlein, C., Janssen, A. W., Hughes, M., Bush, S. L., Marlétaz, F., ... & Young, J. R. (2017). Time-calibrated molecular phylogeny of pteropods. *PLoS One, 12* (6), e0177325.

Cho, A. (2012). Constructing phylogenetic trees using maximum likelihood. *Scripps Senior Theses*. Paper 46. http://scholarship.claremont.edu/scripps_theses/46 (last accessed 2017).

Cook-Deegan, R., DeRienzo, C., Carbone, J., Chandrasekharan, S., Heaney, C., & Conover, C. (2010). Impact of gene patents and licensing practices on access to genetic testing for inherited susceptibility to cancer: comparing breast and ovarian cancers with colon cancers. *Genetics in Medicine*, *12*, S15-S38.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561-563.

Crick, F.H.C. (1958). On Protein Synthesis. In F.K. Sanders. *Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules*. Cambridge University Press.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8), 772-772.

Darwin, C. (1859). On the origin of the species by natural selection. *London-Murray*.

Garzon, M. H. and Mainali, S. (2017). Towards Reliable Microarray Analysis and Design. In *Proc. of the 9th Int. Conference on Bioinformatics and Computational Biology BiCOB'17*. Int. Society for Computer Applications ISCA. (6 pp).

Garzon, M. H., & Mainali, S. (2017, April). Towards a Universal Genomic Positioning System: Phylogenetics and Species IDentification. In *Proc*. of the *International Conference on Bioinformatics and Biomedical Engineering* IWBBIO'17, Lecture Notes in Bioinformatics 10209, Springer (pp 469-479).

Garzon, M. H., & Bobba, K. C. (2012, August). A geometric approach to Gibbs energy landscapes and optimal DNA codeword design. In *Proc. of the International Workshop on DNA-Based Computers*, Lecture Notes in Computer Science **7433** (D. Stefanovic and A. Turberfield, eds.), Springer, Berlin, Heidelberg, (pp. 73-85).

Garzon, M. (2012). DNA Codeword Design: Theory and applications. *Parallel Process. Lett*., **24**:2 (2014), 1-21, (pp 11-26).

Garzon, M. H., & Wong, T. Y. (2011). DNA chips for species identification and biological phylogenies. *Natural Computing 10* (1), 375-389.

Garzon, M., Neathery, P., Deaton, R., Murphy, R. C., Franceschetti, D. R., & Stevens Jr, S. E. (1997). A new metric for DNA computing. In *Proc. of the 2nd Genetic Programming Conference,* Morgan-Kaufmann*,* (pp. 472-278).

Fungiflora, O. S., & Gascuel, O. (2003). A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Syst Biol*, 52, 696704.

Hebert, P. D., Cywinska, A., & Ball, S. L. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B: Biological Sciences*, *270*(1512), 313-321.

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., ... & Gude, K. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, *112*(41), 12764-12769.

Kim, J., & Warnow, T. (1999). Tutorial on phylogenetic tree estimation. *Intelligent Systems for Molecular Biology, Heidelberg*.

Kosiol, C., Bofkin, L., & Whelan, S. (2006). Phylogenetics by likelihood: evolutionary modeling as a tool for understanding the genome. *Journal of biomedical informatics*, *39*(1), 51-61.

Lin, Y., Rajan, V., & Moret, B. M. (2012). A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *9*(4), 1014-1022.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., ... & Lopez, R. (2013). Analysis tool web services from the EMBL-EBI. *Nucleic acids research*, 41(W1), W597-W600.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90-100.

Nagl, S., Schaeferling, M., & Wolfbeis, O. S. (2005). Fluorescence analysis in microarray technology. *Microchimica Acta*, *151*(1), 1-21.

Nei, M., & Kumar, S. (2000). Molecular evolution and phylogenetics. *Oxford university press*.

Neighbor joining. (2017). In *Wikipedia, The Free Encyclopedia*. Retrieved 04:40, November 1, 2017, from https://en.wikipedia.org/w/index.php?title=Neighbor_joining&oldid=799183399

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2), 289-290.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, *53*(1-2), 131-147.

Russell, S., & Modern, P. N. A. I. A. (2003). Approach. *Prentice Hall Pearson Education Inc., Upper Saddle River, New Jersey*, *7458*, 116-119.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406-425.

Schena, M. (2003). *Microarray analysis*. *Wiley-Liss*,.

Sambrook, J., & Russell, D. W. (2006). Fragmentation of DNA by sonication. *Cold Spring Harbor Protocols*, *2006*(4), pdb-prot4538.

Schliep, K. P. (2010). phangorn: phylogenetic analysis in R. *Bioinformatics*, btq706.

Sober, E. (2009). Did Darwin write the Origin backwards?. *Proceedings of the National Academy of Sciences*, *106*(Supplement 1), 10048-10055.

Sokal & Sneath: Principles of Numerical Taxonomy, San Francisco: W.H. Freeman, 1963.

Suzuki, R., Taniguchi, T., & Shimodaira, H. (2004). An approximate maximum likelihood method for phylogenetic tree analysis based on low-temperature Markov chain Monte Carlo. *red*, *68*(60), 95.

Steel, M. A., & Penny, D. (1993). Distributions of tree comparison metrics—some new results. *Systematic biology*, *42*(2), 126-141.

Swofford, D. L. (2002). PAUP*: phylogenetic analysis using parsimony (* and other methods). Sunderland, MA.

UPGMA. (2017). In *Wikipedia, The Free Encyclopedia*. Retrieved 04:38, November 1, 2017, from https://en.wikipedia.org/w/index.php?title=UPGMA&oldid=801046652

Varvio, Sirkka-Liisa (2011). Maximum Parsimony in Phylogenetic Inference. *Phylogenetic inference and data analysis*. https://wiki.helsinki.fi/display/mathstatKurssit/Phylogeny+inference+and+data+analysis%2C+spring+2011. (last accessed 2017).

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids. *Nature*, *171*(4356), 737-738.

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., ... & Feolo, M. (2007). Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1), D13-D21.