UNIVERSITY OF CALIFORNIA, MERCED

Detecting Phylogenetic Signals From Deep Roots of the Tree of Life

A dissertation submitted in partial fulfillment of the requirements
for the degree Doctor of Philosophy

in

Quantitative and Systems Biology

by

Katherine Colleen Harris Amrine

Committee in charge:
Professor Carolin Frank, Chair
Professor David Ardell
Professor Meng-Lin Tsao
Professor Suzanne Sindi

August 2013

UNIVERSITY OF CALIFORNIA, MERCED
Graduate Division

The Dissertation of Katherine Colleen Harris Amrine is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Faculty Advisor: _____
David H. Ardell

Committee Members: _____
Chair: Carolin Frank

_____
Meng-Lin Tsao

_____
Suzanne Sindi

_____
Date

Contents

List of Figures

List of Tables

Acknowledgements

First off, I would like to thank my committee. I would like to thank my committee chair Dr. Carolin Frank, for invaluable research advice, career advice, and advice for how to thrive as a women in science. I would like to thank committee member Dr. Meng-Lin Tsao, who has been a contributor to the progress of my work for my entire graduate student career at UC Merced. I would like to thank Dr. Suzanne Sindi for providing a sounding board for ideas in a field in which my training started out as weak, and for joining my committee and crucial help in the structure and content of this dissertation.

I would like to thank my Advisor, Dr. David Ardell. The road has been bumpy, exciting, devastating, dynamic, and overall truly brilliant. I have learned how to be a scientist by balancing exploration with tangible productivity, and how to keep my priorities in check. I have enjoyed being your student, and I hope to build off the strong platform of rules for quality research that I have gained in your lab.

I would like to thank all of the Yosemite Ave. Starbucks baristas in Merced CA. If not for four months of straight work in the leather chairs in the corner with refilled iced coffee, this dissertation would be nonexistent.

I would like to thank my Merced family; Kristynn, Chelsea, Alyssa, Steve, Alisa and Julie. You were (and still are) my rocks. It takes a village to raise a child, and you have been essential in making it possible to be a mom and a graduate student (because it also seems to take a village to raise a dissertation). I would not have finished without your support and your unconditional love for my son, which often resulted in phone calls from me asking for last-minute babysitting. Kristynn, thank you for letting us commandeer your third bedroom for my last month in Merced so that I could utilize consistent childcare in the city in which I was no longer a resident. I know it wasn't easy. I can't fit all of the thanks in this acknowledgment sections, but I know we will all be friends for life. I will continue to thank each of you for years to come.

I would like to thank my immediate family. My brothers and sister, although you never knew what I was really doing, you helped me develop the drive to achieve my goals. Thank you mom, for believing in me, and ALWAYS being there, no matter what the issue. You are my best friend. Thank you dad, for never expecting anything less than perfect, because you knew that perfect was achievable. I miss you every day.

Thank you Brady and Jared for loving me unconditionally and giving me a reason to smile at the end of every good and bad day. I couldn't ask for a better life.

Finally, thank you, UC Merced. You are a school for dreamers, and you will always hold a special place in my heart.

VITA

EDUCATION

2008	Bachelor of Science in Mathematics, University of Wyoming
2008	Bachelor of Science in Molecular Biology, University of Wyoming
2011	Advanced to Candidacy, Doctor of Philosophy in Quantitative and Systems Biology, University of California, Merced

PUBLICATIONS

Amrine, K.C.H., Swingley, W.D., and D.H. Ardell (2013) tRNA signatures reveal polyphyletic origins of streamlined SAR11 genomes among the Alphaproteobacteria. *Under Review at PLoS Comp Biol*

Huzurbazar, S., Kolesov, G., Massey, S.E., **Harris, K.C.**, Churbanov, A., and D.A. Liberles (2010) Lineage-specific differences in the amino acid substitution process. Journal of Molecular Biology 396: 1410-1421.

Li, M., Gu, R., Su, C.-C., Routh, M.D., **Harris, K.C.**, Jewell, E.S., McDermott, G., and E.W. Yu (2007) Crystal structure of the transcriptional regulator AcrR from *Escherichia coli*. Journal of Molecular Biology 374:591-603.

FIELDS OF STUDY

Major Field: Quantitative and Systems Biology

Studies in Computational Biology
Professor David H. Ardell – University of California, Merced

Studies in Molecular Evolution
Professor David A. Liberles – University of Wyoming

Studies in Protein-Substrate Binding
Professor Edward W. Yu – Iowa State University

Detecting Phylogenetic Signals From Deep Roots of the Tree of Life

by

Katherine Colleen Harris Amrine

University of California, Merced, 2013

Advisor: Prof. David H. Ardell

ABSTRACT OF THE DISSERTATION

In this dissertation, it will be shown that new and unconventional approaches to phylogenetic and classification problems using systems biological data and machine learning fare well against the standard practices in computational time, power, and accuracy. First, we introduce various themes in evolutionary biology, and explain the transfer RNA (tRNA) interactome.

Then, we describe a new way to classify individual organisms based on information from whole genomes. We begin by predicting features by which proteins identify tRNAs coined "Class Informative Features (CIFs)", which form a species-specific "identity code" using a functional information calculation utilizing Information theory and conditional probability. We predict different, but related, codes for different groups of organisms. Then we train an artificial neural network to recognize which code a new, unknown genome is most related to using only primary sequence data. We apply our method to SAR11, one of the most abundant bacterial clades in the world's oceans, and hypothesized to share a phylogenetic sistering with the last alphaproteobacterial mitochondrial ancestor. We find that different strains of SAR11 are more distantly related, both to each other and to mitochondria, than previously thought.

Next, we apply the same logic to the determination to the origin of the Plastid within the Cyanobacteria. We show that using Jensen-Shannon Information Difference calculations, we retrieve a tree which phylogenetically groups Plastids with Cyanobacteria not classically thought to be associated with the cyanobacterial chloroplast ancestor, We also show evidence for refuting classical cyanobacterial topologies. We have uncovered evidence in recent literature that shows mechanistic justification for our largest CIFs.

Finally, we investigate the trend of CIFs across the bacterial tree of life, showing that CIFs maintain a relatively consistent G+C content in all genomes that can be classified by order. This work has developed a pipeline to classify any fully sequenced bacterial genome into a user-defined bacterial order. With modification to the training of the classifier and better Leave-One-Out Cross-Validation of the scoring of the data, we expect that this method will be robust to biological and statistical variations in current tree-building methods.

Chapter 1

Shifting focus in evolutionary biology – identifying a new signal for phylogenetic tree reconstruction and taxonomic classification

1.1 The evolution of bacterial classification and phylogeny

The quest for perfected methods to build phylogenetic trees and accurately classify novel bacteria are popular topics in biological science. From the days of grouping eukaryotic and prokaryotic species together by morphological traits to the first program creating a tree using maximum likelihood by Joe Felsenstein (Felsenstein, 1981). Methods exist and evolve constantly to build trees directly from aligned DNA, RNA and proteins from various samples. Many are successful at making general gene trees and species trees. Variations of these methods exist including using different models to account for biological variability (reviewed in O'Meara (2012) and Anisimova et al. (2013)).

Ever since the characterization of the central dogma of molecular biology by Crick (1958)(DNA $\rightarrow$ RNA $\rightarrow$ protein), the *products* of this universal process have been exploited for phylogenetic reconstruction at any taxonomic level. Investigating a genetic marker of true speciation events, different from the widely utilized macromolecules, remains less important. We will present a phylogenetic marker called "Class-Informative Features" or "CIFs" that avoid the common roadblocks of tree-building in biased circumstances.

1.2 The historical marker – 16S

16S ribosomal RNA is the widely accepted genetic marker used to infer speciation patterns. Originally introduced by Carl Woese and George Fox in 1977 to show that there are three domains of life, 16S phylogeny has become the standard evolutionary marker. With the discovery of the Archaea, Carl Woese foresaw the complete paradigm shift in the study of microbiology as he stated in his 1987 review *"Whatever else it is or whatever impact it may have, the study of bacterial evolutionary relationships is central to the historical account of life on this planet. We may lay no claim to a comprehensive understanding of biology until we know this history, at least in its outline."* (Woese, 1987). 16S ribosomal RNA is important in tracing evolution because species in the entire tree of life contain ribosomal RNA. As described in his 1987 review, we can view 16S rRNA as an excellent genetic marker because it has (i) multiple large domains that can evolve slowly, but possibly independently (Patel, 2001), (ii) retained its general function (Woese, 1987) (iii) and can be directly sequenced using few primers that seem to match 16S orthologs in most species (Lane et al., 1985).

Today, the ability to easily sequence 16S sequences has allowed for the massive expansion of the growing impactful field of metagenomics. We can now identify novel species and characterize known living organisms in various environments. The lack of current sequences and varying evolutionary rates to compare them creates problems in identifying cutoffs to define speciation events for calculation of sequence similarity (percent identity or %ID) (Bosshard et al., 2006; Mignard and Flandrois, 2006). Inside defined bacterial clades, 16S similarity can range from 62% to 91%. 3% diversity has been proposed as a conservative cutoff for species clusters from DNA-DNA hybridization experiments (Stackebrandt and Goebel, 1994). Some phylogenetic programs claim identity solely based on the closest % ID in a database, not taking into account that the sequence may be representative of a novel species (Janda and Abbott, 2007).

Another level of complexity springs from G+C content bias (or Nonstationarity) (Wu et al., 2012). In RNA, nonstationarity is more associated with optimal growth temperature than being vertically inherited which can create signal that places organisms with extreme environments at the root of the tree of life even using conserved 16S rRNA (Galtier and Lobry, 1997).

1.3 Complications in bacterial classification and phylogeny

1.3.1 Horizontal gene transfer

Bacteria asexually reproduce, replicating their DNA and passing it to their divided counterpart to create two organisms from one. This is vertical inheritance. The replication of the DNA that will be passed to the newly dividing cell is a source for mutation. These

mutations can lead to cell death, decreased fitness, no increase or decrease in fitness, or an increase in fitness (reviewed in Baake and Gabriel (2000)). Any time the organism survives and is able to reproduce, there is a chance that the mutation will become a common feature in the population. These mutations, which turn into substitutions, are markers that we try to trace when re-creating phylogenetic trees. In a perfect statistical model, we would be able to trace these changes to completely and correctly depict a tree of speciation throughout the history of life.

Of course, the picture is not so simple. Bacterial organisms are know to engage in homologous recombination with closely related species, and horizontally transfer their genes to distantly related species, allowing them to become part of another organism's genome. Many specific bacteria have been shown to successfully participate in homologous recombination with organisms up to 25% divergent in DNA sequence, which is a drastic increase from the 3% cutoff in eukaryotic species (Duncan et al., 1989; Vulić et al., 1997; Majewski and Cohan, 1999). If any of these horizontally inherited genes used when creating a phylogenetic tree or grouping organisms using any available sequence-based tool, an incorrect picture of speciation will be portrayed. In some organisms, their vertically inherited genes are estimated to be as much as 30% of their genomic DNA. For example, up to 24% of the bacterial *Thermotoga maritima* genome was predicted to have been obtained from archaeal lineages (Nelson, K E and Clayton, R A and Gill, S R and Gwinn, M L and Dodson, R J and Haft, D H and Hickey, E K and Peterson, J D et al., 1999). This type of bias is biologically driven.

1.3.2 Does a true tree exist?

If life evolved from one common ancestor in bacterial evolution, where reproduction is asexual, there has to be a path of vertical inheritance despite the varying rates of horizontal gene transfer across the tree of life, which often cloud the picture created by various tree-building algorithms. Some scientists disagree that there is one hierarchical tree of life (Doolittle and Bapteste, 2007) given the amount of shared information across species through homologous recombination and illegitimate recombination.

The work in this dissertation assumes a true bifurcating tree exists based on the laws of life, and attempts to fined a true marker for vertical inheritance. We recognize that the evolution of life is more complicated than vertical inheritance especially in bacteria, but the methods in this dissertation aim to resolve relationships arising from asexual reproduction.

1.4 Methods for phylogenetic tree reconstruction

1.4.1 DNA

Deoxyribonucleic Acid (DNA), or the template for all life's genetic make-up, can essentially be represented by four organic compounds which are described by two groups:

two-ringed purines (Adenine (A) and Guanine (G)) and one-ringed pyrimidines (Cytosine (C) and Threonine (T)). These bases can mutate from one to the other, and most detrimental mutations do not persist through the population. The ones that begin to be passed to descendants constitute substitutions. DNA models like Jukes-Cantor take a simplistic approach to estimate nucleotide substitutions by assuming that all substitutions are equally likely (Jukes and Cantor, 1969). More complicated models like Kimura 2 parameter and Felsenstein models allow for either unequal substitution frequencies –different for a transition (purine $\leftrightarrow$ purine or pyrimidine $\leftrightarrow$ pyrimidine) or transversion (purine $\leftrightarrow$ pyrimidine)– or unequal base frequencies (straying from assumed equal amounts of A,T,C and G), respectively (Kimura, 1980; Felsenstein, 1981). Model HKY85 (Hasegawa et al., 1985) and General reversible models allow for more varying rates, but as the number of varying rates increase, so does the complexity of the model. Adding complexity adds degrees of freedom which runs the risk of over-fitting data.

1.4.2 RNA

RNA transcribed from DNA to code for a protein, or coding RNA, can be beneficial for phylogenetic tree reconstruction due to the fact that edited RNA transcripts are easily alignable with exons (specific coding regions) all in their correct location. Applying these same models for DNA allows different constraints to be prioritized. Also, triplets of codes can be utilized more effectively with knowledge of the final product (the protein) being considered in the model.

In RNA transcribed from DNA to fold and perform a specific biochemical function, or noncoding RNA, the situation is similar to basic RNA models, but with added variables for base pairing in secondary structure. If an alignment can be made for secondary structure, this information can be exploited. Programs like RaxML (Stamatakis et al., 2004) and PHASE (Jow et al., 2003) have implemented such models well, but they are computationally expensive.

1.4.3 Proteins

Protein alignments, in theory, provide a better picture of true evolution because back and parallel substitutions are less likely to occur when more letters make up the possible alphabet (bases $\{$A,T,C,G$\}$ in DNA and amino acids $\{$A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y$\}$ in proteins). Back and parallel substitutions will make two sequences seem similar due to chance rather than evolution (discussed in detail below). Protein evolution is often modeled by substitution matrices which characterize probabilities of seeing amino acid changes based on quantifying those seen in nature. The first implementation was by Margaret Dayhoff where dot products of the matrix of amounts of amino acids seen in nature are created to model more distant evolution. Variations of the DAYHOFF (Dayhoff and Schwartz, 1978) matrix have been created, for example, the BLOSUM matrix (Henikoff and Henikoff, 1992) is used in many

programs including ClustalW (Henikoff and Henikoff, 1992; Higgins and Sharp, 1988).

1.4.4 Data compilation

Methods have been created to combine data to weed out fluctuations in detected signal by bias. The "supermatrix method", where conserved protein/DNA/RNA sequences are concatenated together to create a large alignment of features, and then a tree algorithm combines the data hierarchically based on the specific model one dictates (reviewed in de Queiroz and Gatesy (2007)). The idea behind their approach is that if there are sites that do not relay true phylogenetic signal, they will be washed out with the overwhelming amount of data. Another method, the "supertree" method, will take groups of alignments of conserved proteins/DNA/RNA and individually compute trees for each of these sets (reviewed in Bininda-Emonds (2005)). Then, once the trees have been calculated, a consensus tree can be estimated. This prescribes to the idea that if any gene was horizontally transferred, or has undergone rapid evolution, it will only result in a nominal subsection of the trees reporting the false signal. Allowing a consensus tree to persevere will overlook these non-evolutionary signals.

1.5 Bias in tree-building

Up to this point, we have touched on how bias can cause various tree-building programs to compromise true results, but the details of the types of biases which gives these variations in outcomes have not yet been discussed.

Methodological bias exist in the implementation of every tree-building algorithm that are not necessarily biologically driven, but just driven by the constraints of the alphabet used to describe the system. In DNA, we have ATC and G, and RNA, ACU and G. One can imagine, that two independent sequences that have a string of nucleotides can mutate to the same sequence, not by being passed down from a common ancestor, but by chance. This is called a parallel substitution, and is difficult to statistically account for. Sequence differences will saturate at 75% due to the fact that they will be 25% similar just by chance. This is common with only having four possibilities. Similarly, a sequence of nucleotides that looks the same as another sequence can look the same because they both evolve from the same ancestor, or because they were the same, but one changed to a new nucleotide, then changed back. This is called a back substitution, and occurs in nature. Basic models will have to take this into account if they would like to depict an accurate picture.

Another methodological bias called long branch attraction causes problems throughout the tree of life. It is methodological bias due to the fact that the input to the program dictates the output. If you place something not like all the other things in a dataset into a program, the program will place the data where it sees fit, at the root, branched basally in relation to the rest of the data. Long branch attraction occurs when a datapoint looks so diverged from the rest of the data, that a phylogenetic program will

place it near the root with a long branch (reviewed in Bergsten (2005)).

In Proteins, there can be over 20 letters used to describe elements of a peptide chain. Although parallel and back substitutions are less likely due to a larger alphabet, they still occur and create false signal.

1.6 Biological bias in biological data

We've talked about these changes, but not why they may happen and if they happen in some biological systems more than others. The fact is that environmental constraints can drive biological signals to look similar just due to the organism evolving to survive. Various constraints exist with many different theories on the importance of each. Temperature is a common theme when talking about biological bias. It is well-known that GC bonds are stronger, forming three hydrogen bonds in DNA double helices, and AT (or AU) bonds are weaker, only forming two. This would infer that when DNA and RNA are present in higher temperatures, each needs to maintain stronger bonds to keep structures from denaturing, or unfolding, sabotaging their function.

Overall, you cannot account for every type of bias, but if one can identify signal that is less biased in relation to the rest of a genome, accounting for it becomes much simpler.

1.7 The tRNA interaction network

Transfer RNAs (tRNA) are noncoding short nucleotide sequences transcribed from DNA. Their primary role is to participate in the collection of amino acids during protein syntheses and are used to decode the genetic code by creating a link from blueprint (or DNA) to product (or protein). tRNA structure consists of (1) an acceptor stem where the $3'$ and $5'$ ends meet that carries an amino acid (purple in Figure 1.1), (2) an anticodon loop at the end of the anticodon stem which interacts with codons on mRNA (dark blue in Figure 1.1), and (3) T and (4) D stems with loops that partake in tertiary interactions with each other (orange and light blue, respectively in Figure 1.1. Some tRNAs also contain a long string of nucleotides in between the anticodon stem and the T arm called the Variable loop. These tRNAs are called "type II" tRNAs. tRNA secondary structure is often depicted as a cloverleaf-like image, but they are mostly in an L-shape in their folded forms with the anticodon loop and acceptor stem at opposite ends of the L.

tRNAs are extremely conserved, and normally have roughly 74 nucleotides, definable by a rigid, universal coordinate system (Sprinzl et al., 1998). The coordinates have flexibility in areas known to have extra nucleotides with letters added on (for example 20A,20B,20C in the D loop or e11,e12,e13... in the variable loop).

There are 64 (4^3) possible codons in the genetic code made up of 3 consecutive nucleotides, resulting in 20 different possible amino acids. Not all of these combinations are actually utilized by every system, but each is a known link in the standard genetic

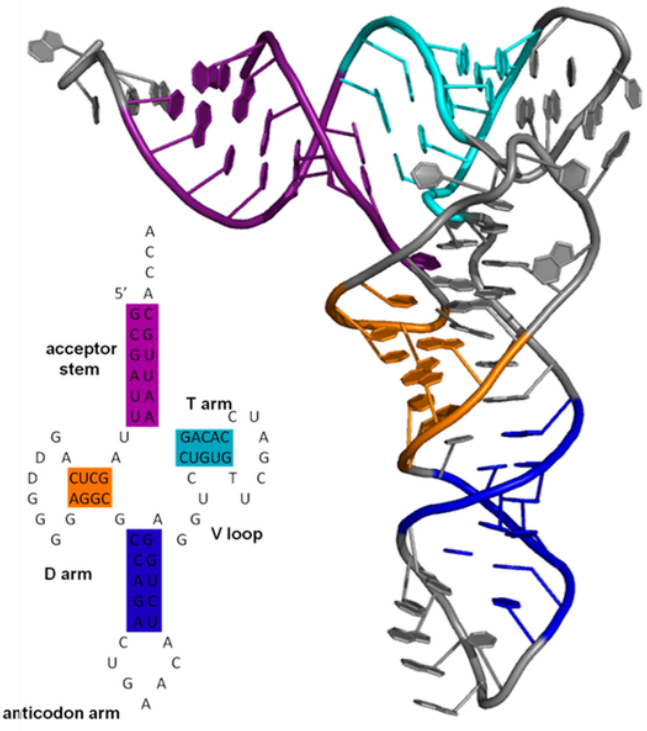


Figure 1.1: **An example of Secondary and Tertiary structure of a tRNA.** Secondary and Tertiary Structure of PDB 6tna rendered with PyMol (DeLano Scientific Freeware). (Sussman et al. JMB, 1978). Figure By Jennifer Fribourgh & Kyle Schneider:http://en.wikipedia.org/wiki/File:TRNA_all2.png

code. There are multiple codons for each amino acid although some are more favored than others. There is a standard initiator codon (AUG) that codes for a very unique methionine amino acid (f-Met) and a few stop codons (UAG, UAA and UGA) that do not have any tRNAs that naturally match, but initiate the dislocation of the ribosome to stop the decoding of an mRNA. In some cases, UGA (or the "opal" stop codon) has been shown to undergo translational recoding to code for a 21st amino acid, named Selenocysteine (Böck et al., 1991). A 22nd amino acid discovered in Archaea (termed pyrrolysine) uses the UAG stop codon (or the "amber" stop codon) (Srinivasan et al., 2002). Some tRNA genes have also mutated to adapt for nonsense mutations in the genetic code that lead to early truncation of functional proteins. These tRNA genes are called amber suppressors (Chan and Garen, 1970). To our knowledge, the tRNA interaction network is as close to a closed system with multiple moving parts that is inherited throughout the tree of life. All tRNAs must fit in the exact same spots in the ribosome, making their structure well conserved throughout the tree of life. Conversely, they must be distinct enough to be distinguished from one another in order to be charged with the correct amino acid by a tRNA aminoacyl synthetase. This product of tRNA-AA is used in translation. These constraints on structure and sequence are the fundamental properties that make the body of this dissertation possible.

If one can use statistical theories to detect the specific nucleotides that make tRNAs distinguishable from each other and with a measure of importance attached to that distinction, maps can be created to use in order to assess relatedness from two different sets of species. Due to the unique nature of this system, only the tRNAs are needed to be compared to each other utilizing simple statistics in order to define the important factors and assign weights to said factors.

1.8 Information theory

Entropy is a measure of disorder in a system. In investigations of multiple sequence alignments, the entropy we measure is calculated with four possibilities in nucleotide sequences. The entropy we calculate is measured in bits (log base 2), and is used to describe the minimum descriptive complexity of any nucleotide site.

Entropy, H, of a discrete random variable X is represented as $H(X) = -\Sigma p(x) log[p(x)]$, with $0 \leq p(x) \leq 1$ being the probability of event x in a set of possible events with $\sum p(x) = 1$. If logarithms are calculated to base 2, then entropy $H(X)$ and Information, $I(X)$, are measured in bits. Information is the calculation of the entropy of a random variable subtracted from the expected entropy calculated from a "background" multivariate distribution over the same event space. The background distribution might be calculated from genomic frequencies of bases in DNA, for instance, in which case the events x are the bases of DNA, i.e. $X \in A, C, T, G$. In applications to sequence analysis, the random variable X is usually take to be from a multinomial distribution representing variation at a single amino acid or nucleotide site from functionally or structurally analogous, or homologous macromolecules.

1.9 Machine Learning for bacterial classification

Machine Learning techniques are becoming more popular in many aspects of biological sequence analysis, especially multiple sequence alignments. Little has been done in the study of phylogenetic tree construction and strain classification. Rotteger et al. used Machine Learning to detect if a gene resulted from Lateral Gene Transfer from discordant phylogenies (Roettger et al., 2009). One recent paper used machine learning to relate proteins (Lin et al., 2013). In 2010, a method was created to use machine learning to hierarchically classify bacteria based on 16S rRNA sequences with little improvement to current methods (Slabbinck et al., 2010). We will attempt to use machine learning to relate species, and alleviate bias by detecting signal using Information theory.

In this work, we use nonlinear models for classification of unknown bacteria into defined groups with scores depicted by Class Informative Features calculated with Information theory. In order to determine that a nonlinear model was necessary for a classifier, multiple methods were evaluated. We utilize a Multilayer Perceptron Machine Learning Algorithm implemented in WEKA (Hall et al., 2009). Nonlinear model necessity can be justified by the make-up of the scores in our data. Due to the fact that we are utilizing scores that are carefully calculated weighted compositions, there is extreme overlap of signal included from one score to the other. The scoring method condenses the individual score contributions into one total score per clade, resulting in extreme generalization. The classifier allows for this, given its excellent performance during 10-fold cross-validation during training.

A Multilayer Perceptron consists of 3 essential layers each made up of nodes: one input layer that consists of a set of nodes representing the input variables, one output layer consisting of all possible outcomes, and a hidden layer that can actually consist of multiple sub-layers (see Figure 1.2). Each node in the hidden layer is modeled by a sigmoid function invisible to the user. During training, the algorithm will decide the weights of the connectors of each layer, and the number of sub-layers that make up the hidden middle layer. This method allows for probabilistic classification. The process of backpropagation, or backward propagation of errors allows the weight of a specific activation function in the neural network to be re-evaluated and modified in order to better predict the output from the input of a MLP (Rumelhart et al., 2002). The weights that are chosen minimize the error in the output.

Other machine learning algorithms we tried were linear, including a Support Vector Machine which consists of a vector space with the same number of dimensions as one less the number of outcomes. Boundaries for classification are defined by various vectors calculated during training that optimally separate the data. There is no relevant probability associated with this method. Support Vector Machines out-of-box proved unsuccessful in describing our data, no matter how we tried to represent scores. One could re-calculate the unique identity elements for each defined group to distinctly separate the data in order to try to train with a simpler method than MLP. This would eliminate information about the differing intensity (or informativeness) of shared CIFs between groups, making the

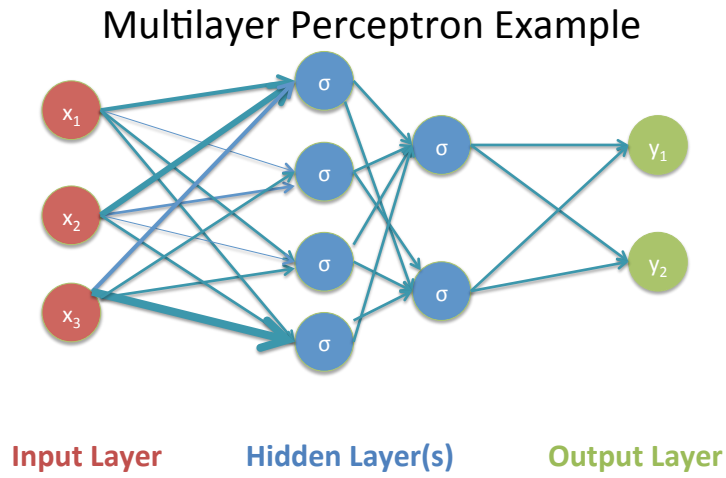


Figure 1.2: **A depiction of a Multilayer Perceptron with two hidden layers, three input values, and two outcomes.** Different thickness of arrows shows that different connectors can have different weights depending on back-propagation.

nonlinearity in our scores an essential nonlinearity.

Trying an ensemble learning method to train a classifier called Random Forests also produced noisy results, giving high error rates when implemented on scores derived from CIFs. Knowing that Random Forests train based on the mode of the dataset, our data would have to have consistent score profiles across all scored sets of data. This seems unlikely due to the possibility that CIFs may be contained in our test sets that are not contained in our training data.

Chapter 2

tRNA signatures reveal polyphyletic origins of streamlined SAR11 genomes among the Alphaproteobacteria

In revision at *PLoS Computational Biology*
Authors: *Katherine C.H. Amrine, Wesley D. Swingley, and David H. Ardell*

2.1 Abstract

Phylogenomic analyses are subject to bias from convergence in macromolecular compositions and noise from horizontal gene transfer (HGT). Accordingly, compositional convergence leads to contradictory results on the phylogeny of taxa such as the ecologically dominant SAR11 group of Alphaproteobacteria, that have extremely streamlined, A+T-biased genomes. While careful modeling can reduce bias artifacts caused by convergence, the most consistent and robust phylogenetic signal in genomes may lie in the features governing macromolecular interactions. Here we develop a novel phyloclassification method based on signatures derived from bioinformatically defined tRNA Class-Informative Features (CIFs). tRNA CIFs are enriched for features that underlie tRNA-protein interactions. Using a simple tRNA-CIF-based phyloclassifier, we obtained results consistent with bias-corrected whole proteome phylogenomic studies, rejecting monophyly of SAR11 and affiliating most strains with Rhizobiales with strong statistical support. Yet, as expected by their elevated genomic A+T contents, SAR11 and Rickettsiales tRNA genes are also similarly and distinctly A+T-rich within Alphaproteobacteria. Using conventional supermatrix methods on total tRNA sequence data, we could recover the artifactual result of a monophyletic SAR11 grouping with Rickettsiales. Thus tRNA CIF-based phyloclassification is relatively robust to base content convergence of tRNAs. Also, given the notoriously promiscuous HGT rates of aminoacyl-tRNA synthetase genes, tRNA CIF-based phyloclassification may be at least partly robust to HGT of network components. We describe how unique features of the

tRNA-protein interaction network facilitate mining of traits governing macromolecular interactions from genomic data, and discuss why interaction-governing traits may be especially useful to solve difficult problems in microbial classification and phylogeny.

2.2 Introduction

What parts of genomes are most robust to compositional convergence? What information is most faithfully vertically inherited? The key assumptions of compositional stationarity and consistency in gene histories underpin most current approaches in phylogenomics and are frequently violated (reviewed in *e.g.*Gribaldo and Philippe (2002)). HGT is so widespread that the very existence of a "Tree of Life" has been questioned (Gogarten et al., 2002; Bapteste et al., 2009). Better understanding of ancient phylogenetic relationships requires discovery of new universal, slowly-evolving phylogenetic markers that are robust to compositional convergence and HGT.

The controversial phylogeny of *Ca.*Pelagibacter ubique (SAR11) is a case in point. SAR11 make up between a fifth and a third of the bacterial biomass in marine and freshwater ecosystems (Morris et al., 2002). Adaptations to extreme environmental nutrient limitation may explain why SAR11 have very small cell and genome sizes and small fractions of intergenic DNA (Giovannoni, 2005). While some recent phylogenomic studies define a clade among SAR11, the largely endoparasitic Rickettsiales and the alphaproteobacterial ancestor of mitochondria (Williams et al., 2007; Georgiades et al., 2011; Thrash et al., 2011), others argue this placement is an artifact of independent convergence towards increased genomic A+T content, and that SAR11 belongs closer to other free-living Alphaproteobacteria such as the Rhizobiales and Rhodobacteraceae (Brindefalk et al., 2011; Rodríguez-Ezpeleta and Embley, 2012; Viklund et al., 2012). Monophyly of SAR11 was also recently rejected (Rodríguez-Ezpeleta and Embley, 2012).

Nonstationary macromolecular compositions are a known source of bias in phylogenomics (Foster, 2004; Losos et al., 2012). Widespread variation in macromolecular compositions may be associated with loss of DNA repair pathways in reduced genomes (Dale et al., 2003; Viklund et al., 2012), unveiling an inherent A+T-bias of mutation in bacteria (Hershberg and Petrov, 2010) and elevating genomic A+T content (Moran, 2002; Lind and Andersson, 2008). A process such as this has likely altered protein and RNA compositions genome-wide in SAR11, and if such effects are accounted for, the placement of SAR11 with Rickettsiales drops away as an apparent artifact (Rodríguez-Ezpeleta and Embley, 2012; Viklund et al., 2012). Consistent with this interpretation, SAR11 strain HTTC1062 shares a surprising and unique codivergence of $tRNA^{His}$ and histidyl-tRNA synthetase (HisRS) with a clade of free-living Alphaproteobacteria (Wang et al., 2007a; Ardell, 2010) that likely arose only once in bacteria (Ardell and Andersson, 2006). This synapomorphy contradicts the placement of SAR11 with Rickettsiales and mitochondria.

This work was motivated to determine whether the entire system of tRNA-protein

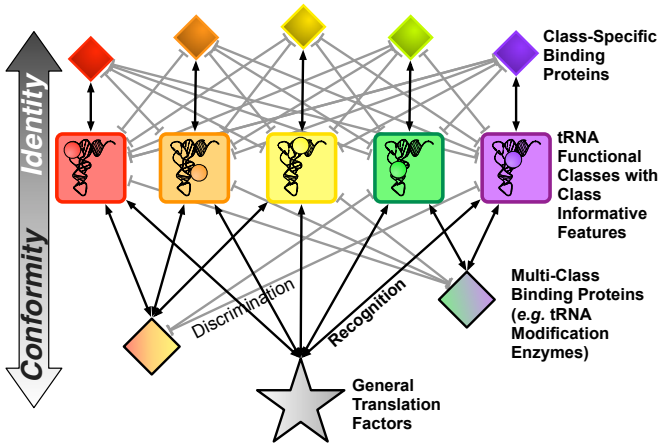


Figure 2.1: **A universal schema for tRNA-protein interaction networks.**

interactions could be exploited to address phylogeny of bacteria, particularly SAR11. The highly conserved tRNA-protein interaction network (Fig. 2.1) has special advantages for comparative systems biological study from genomic data. First, the components and interactions of this network are highly conserved. Second, bioinformatic mining of interaction-determining traits from genomic tRNA data is favorable because tRNA structures are highly conserved not just across extant taxa but also across different functional classes of tRNAs ("conformity" (Wolfson et al., 2001)). Yet each functional class of tRNA must maintain a hierarchy of increasingly specific interactions with various proteins and other factors ("identity" (Giege, 2008)). The conflicting requirements of conformity and identity allow structural comparison and contrast to predict class-informative traits of tRNAs from sequence data by relatively simple bioinformatic methods (Ardell, 2010). The features that govern tRNA-protein interactions diverge across the three domains of life (reviewed in (Giegé et al., 1998)) and also within the domain of bacteria (Ardell and Andersson, 2006).

In prior work, we developed "function logos" to predict, at the level of individual nucleotides before post-transcriptional modification, what genetically templated information in tRNA gene sequences is associated to specific functional identity classes (Freyhult et al., 2006). We now call these function-logo-based predictions Class-Informative Features (CIFs). A tRNA CIF answers a question like: "if a tRNA gene from a group of related genomes carries a specific nucleotide at a specific structural position, how much information do we gain about that tRNAs specific function?" Such information estimates are corrected for biased sampling of functional classes and sample size effects (Freyhult et al., 2006), and their statistical significance may be calculated (Ardell, 2010). Although an individual bacterial genome does not present enough data to generate a function logo, related genome data may be lumped, weakly assuming homogeneity of tRNA identity rules (although heterogeneity generally reduces signal). Function logos recover known tRNA identity elements (*i.e.* features that govern

the specificity of interactions between tRNAs and proteins) (Giegé et al., 1998), and more generally, predict features governing interactions with class-specific network partners such as amidotransferases (Bailly et al., 2006). A recent molecular dynamics study on a $tRNA^{Glu}$-GluRS (Glutaminal tRNA-synthetase) complex identified tRNA functional sites involved in intra- and inter-molecular allosteric signaling within GluRS that couples substrate recognition to reaction catalysis (Sethi et al., 2009). The predicted sites are correlated with those from proteobacterial function logos (Freyhult et al., 2007).

In this work, we show that tRNA CIFs have diverged among Alphaproteobacteria in a phylogenetically informative manner. Second, as phylogenetic markers, tRNA CIFs are more robust to compositional convergence than the tRNA bodies in which they are embedded. Using our tRNA-CIF-based phyloclassification approach, we confirm that SAR11 are polyphyletic with the majority of strains clustering with the free-living Alphaproteobacteria. Our results have implications for how to best mine genomic data for phylogenetic signals.

2.3 Results

We re-annotated alphaproteobacterial tDNA data from tRNAdb-CE 2011 (Abe et al., 2011) and other prepublication genomic data, and split them into two groups according to whether or not their source genome contained the uniquely derived synapomorphic traits previously described (Ardell and Andersson, 2006): a gene for $tRNA^{His}$ containing A73 (using "Sprinzl coordinates", (Sprinzl et al., 1998)) and lacking templated $-1G$. We could thereby partition the data into an RRCH clade (Rhodobacteraceae, Rhizobiales, Caulobacterales, Hyphomonadaceae), which present the uniquely derived $tRNA^{His}$, and the RSR grade (Rhodospirillales, Sphingomonadales, and Rickettsiales, excluding SAR11), which present "normal" bacterial $tRNA^{His}$ with C73 and genomically templated $-1G$. In all, data from 214 alphaproteobacterial genomes represented 11644 predicted tRNA sequences (8773 sequences unique within genomes and 3064 total unique sequences). Our final dataset contained 147 genomes (8597 tRNAs) for the RRCH clade, 59 genomes (2792 tRNAs) for the RSR grade, and 8 genomes (255 tRNAs) of SAR11 strains.

The unique traits of the RRCH $tRNA^{His}$ are perfectly associated to substitutions of key residues in the motif IIb tRNA-binding loops of HisRS involved in tRNA recognition (Ardell and Andersson, 2006). Seven of eight SAR11 strains exhibited the unique $tRNA^{His}$/HisRS codivergence traits in common with RRCH genomes. In contrast, strain HIMB59 presented ancestral bacterial characters in both $tRNA^{His}$ and HisRS (Supp. Fig. 2.7). These results immediately suggest that HIMB59 is not monophyletic with the other SAR11 strains, consistent with (Rodríguez-Ezpeleta and Embley, 2012).

We computed function logos (Freyhult et al., 2006) of the RRCH clade and RSR grade to form the basis of a tRNA-CIF-based binary phyloclassifier as shown schematically in Fig. 2.2. To reduce bias, we used a Leave-One-Out Cross-Validation (LOOCV) approach. For comparison, we also performed LOOCV phyloclassification

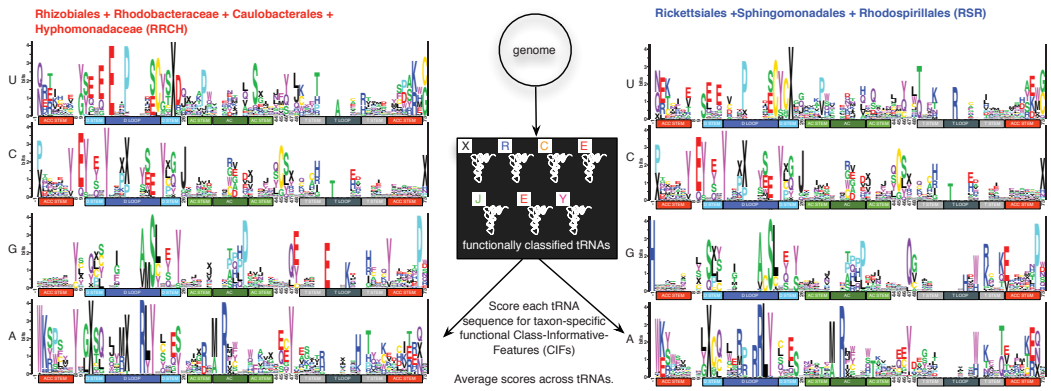


Figure 2.2: **Function logos (Freyhult et al., 2006) of tRNA CIFs in the RRCH and RSR groups of Alphaproteobacteria, and overview of tRNA-CIF-based binary phyloclassification.**

using sequence profiles of entire tRNAs, with typical results shown in Fig. 2.3B. Although the tRNA-CIF-based phyloclassifier (Fig. 2.3A) was biased positively by the much larger RRCH sample size, it achieved better phylogenetic separation of genomes than the total-tRNA-sequence-based phyloclassifier (Fig. 2.3B). The Sphingomonadales and Rhodospirillales separated in scores from the Rickettsiales in both classifiers. Most importantly, the tRNA-CIF-based phyloclassifier placed all eight SAR11 genomes closer to the RRCH clade and far away from the Rickettsiales with HIMB59 overlapping the Rhodospirillales, while the total-tRNA-sequence-based phyloclassifier placed all eight SAR11 genomes closer to the Rickettsiales. Supplementary Figure 2.8 shows the effects of different treatments of missing data in the total-tRNA-sequence-based classifier. Method "zero," shown in Fig. 2.3C, is most analogous to the method used to generate Fig. 2.3A. Method "skip" (Supp. Fig. 2.8D) shows that SAR11 tRNAs share sequence characters in common with the RSR grade that are not seen in the RRCH clade. Methods "small" and "pseudo" (Supp. Figs. 2.8A and 2.8B) show that SAR11 have sequence traits not observed in either RSR or RRCH.

Many other tRNA classes besides $tRNA^{His}$ contribute to the differentiated classification of RRCH and RSR genomes by the CIF-based binary classifier (Fig. 2.4). Other tRNA classes are also differentiated between these two groups, including $tRNA^{Cys}$, $tRNA^{Asp}$, $tRNA^{Glu}$, $tRNA^{Ile}_{LAU}$ (symbolized "J"), $tRNA^{Lys}$, $tRNA^{Tyr}$. These results extend the observations of Wang et al. (2007a) who discovered unusual base-pair features of $tRNA^{Glu}$ in the RRCH clade. In classes for which the RRCH and RSR groups are well-differentiated, HIMB59 uniquely groups with RSR while other strains group with RRCH, while for other tRNA classes, all putative SAR11 strains lie outside the RRCH and RSR distributions. This implies that more diverse alphaproteobacterial genomic data are necessary to completely resolve the phylogenetic affiliation of SAR11 strains, but strongly contradict a monophyletic affiliation of SAR11 with Rickettsiales.

The increases in genomic A+T contents in SAR11 and Rickettsiales have also

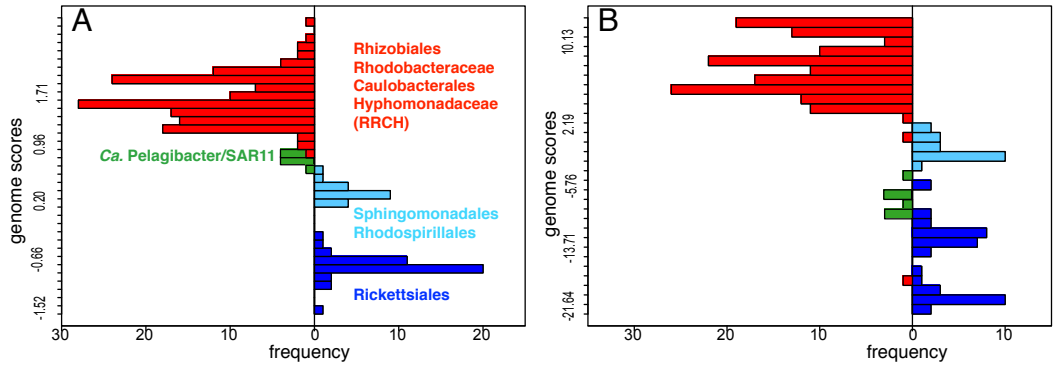


Figure 2.3: **Leave-One-Out Cross-Validation (LOO-CV) scores of alphaproteobacterial genomes under two different binary phyloclassifiers.** A. tRNA-CIF-based phyloclassifier B. Total tRNA sequence-based phyloclassifier.

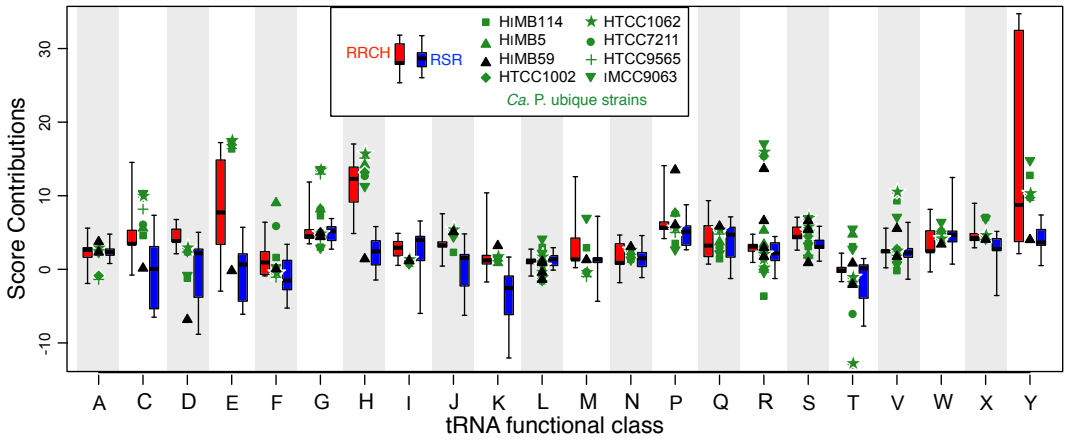


Figure 2.4: **Breakout of class contributions to scores under the tRNA CIF-based binary phyloclassifier.**

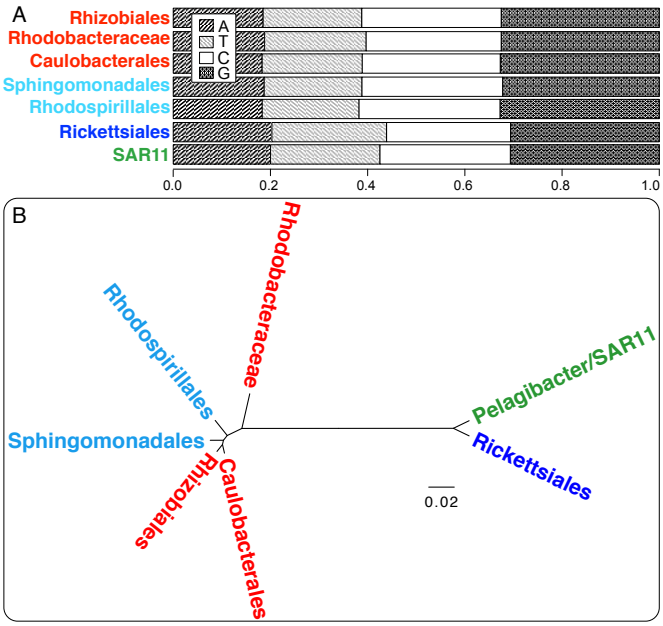


Figure 2.5: **Base compositions of alphaproteobacterial tRNAs showing convergence between Rickettsiales and SAR11.** A. Stacked bar graphs of tRNA base composition by clade. B. UPGMA clustering of clades based on Euclidean distances of tRNA base compositions under the centered log ratio transformation (Aitchison, 1986).

driven elevated A+T contents of their tRNA genes (Fig. 2.5A). Rickettsiales and SAR11 tRNA genes are both notably elevated in both A and T, and share an overall similarity in composition distinct from other Alphaproteobacteria. Hierarchical clustering of alphaproteobacterial taxa based on tRNA gene base contents closely group SAR11 and Rickettsiales together (Fig. 2.5B).

Nonstationary tRNA base content — convergence to greater A+T content — causes all eight SAR11 strains in our dataset to group with Rickettsiales using phylogenomic approaches based on total tRNA sequence evidence. In a "supermatrix" phylogenomic approach, concatenating genes for 28 isoacceptor classes from 169 species (2156 total sites) and using the GTR+Gamma model in RAxML, we estimated a Maximum Likelihood tree in which all eight putative SAR11 strains branch together with Rickettsiales (Supp. Fig. 2.9). For this analysis, in 31% of instances when isoacceptor genes were picked from a genome, we randomly picked one gene from a set of isoacceptor paralogs. However, our results did not depend on which paralog we picked. Using a distance-based approach with FastTree, we computed a consensus cladogram over 100 replicate alignments each representing different randomized picks over paralogs. As the consensus cladogram shows (Supplementary Figure 2.10) each replicate distance tree placed all eight putative SAR11 strains together with Rickettsiales. The recently introduced tRNA-specific FastUniFrac-based method for microbial classification (Widmann et al., 2010) also places all SAR11 strains together with

Rickettsiales.

However, as shown in Fig. 2.6, a multiway classifier based on tRNA CIFs bins all SAR11 strains with the Rhizobiales except for HIMB59, which bins with the Rhodospirillales, consistent with the results of (Rodríguez-Ezpeleta and Embley, 2012). These results use a Multilayer Perceptron (MLP) classifier implemented in WEKA (Hall et al., 2009) and only seven taxon-specific CIF-based summary scores. The MLP is the simplest non-linear classifier able to handle the interdependent signals in the CIF-based scores for tree-like data (Theodoridis and Koutroumbas, 1999). In a Leave-One-Out cross-classification, all other genomes scored consistently with NCBI Taxonomy except three placed in Rhodobacteraceae based on 16S ribosomal RNA evidence: *Stappia aggregata*, *Labrenzia alexandrii* and the denitrifying *Pseudovibrio sp.* JE062. None of these genomes scored strongly against Rhodobacteraceae except *Pseudovibrio*, which scored four times greater against the Rhizobiales.

To assess robustness of our results we performed two controls: we bootstrapped sites of tRNA data in each genome to be classified, and we filtered away small CIFs with Gorodkin heights < 0.5 from our models, retrained the classifier and bootstrapped sites again. Generally bootstrap support values correspond to original classification probabilities. All SAR11 strains have support values $> 80\%$ as Rhizobiales, majority bootstrap values as Rhizobiales (HIMB114 at 70% with Rickettsiales at 15% and HTCC7211 at 54% with Rickettsiales at 13%), or plurality bootstrap value as Rickettsiales (HIM5 at 48% with Rickettsiales at 18%) except HIMB59 which had a bootstrap support value of 87% to be in the Rhodospirillales.

2.4 Discussion

Our results provide strong, albeit unconventional, evidence that most SAR11 strains are affiliated with Rhizobiales, while strain HIMB59 is affiliated with Rhodospirillales. These results are entirely consistent with comprehensive phylogenomic studies that control for nonstationary macromolecular compositions in Alphaproteobacteria (Brindefalk et al., 2011; Rodríguez-Ezpeleta and Embley, 2012; Viklund et al., 2012) or a site-rate-filtered analysis (Gupta and Mok, 2007). Our CIF-based method works even though SAR11 and Rickettsiales tRNAs have converged in base content, so that total tRNA sequence-based phylogenomics gives opposite results. tRNA CIFs must be at least partly robust to compositional convergence of the tRNA bodies in which they are embedded.

It is well known that aminoacyl-tRNA synthetases (aaRS) are highly prone to HGT (Doolittle and Handy, 1998; Brown and Doolittle, 1999; Wolf et al., 1999; Woese et al., 2000; Andam and Gogarten, 2011) including in Alphaproteobacteria (Ardell and Andersson, 2006; Dohm et al., 2006; Brindefalk et al., 2006). We hypothesize that our tRNA-CIF-based phyloclassifiers are also robust to HGT of components of the tRNA-protein interaction network, consistent with Shiba and Motegi (1997), who argued that a horizontally transferred aaRS is more likely to functionally ameliorate to a tRNA-protein network into which it has been transferred rather than remodel that network

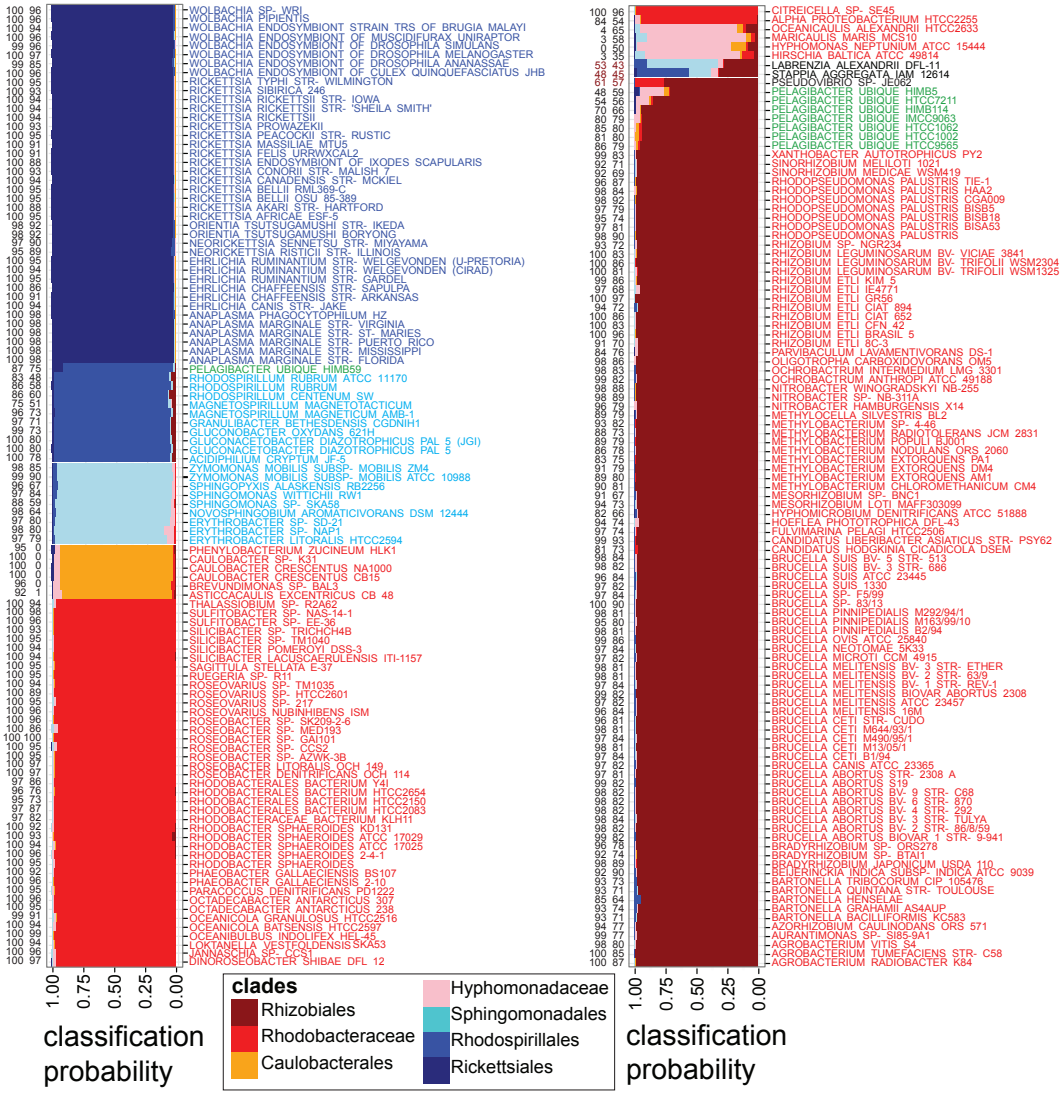


Figure 2.6: **Multiway classification of alphaproteobacterial genomes using a feature vector of seven tRNA-CIF-based summary scores and the default Multilayer Perceptron model in WEKA.** Bootstrap support values under resampling of tRNA sites against (left) all tRNA CIFs and (right) CIFs with Gorodkin heights greater than or equal to 0.5 bits and model retraining (100 replicates). All support values correspond to most probable clade as shown except for *Stappia* and *Labrenzia* for which they correspond to Rhizobiales.

to accommodate itself. HGT of aaRSs may also perturb a network so as to cause a distinct pattern of divergence (Ardell and Andersson (2006) and this work). Wang *et al.* (Wang et al., 2007a) discuss the possibility that RRCH $tRNA^{His}$ and HisRS were co-transferred into an ancestral SAR11 genome. However, this fails to explain the correlations of many other tRNA traits of SAR11 genomes with the RRCH clade reported here. Further study is needed to address the robustness of our method to component HGT.

A more distant relationship between most SAR11 strains and Rickettsiales actually strengthens the genome streamlining hypothesis (Giovannoni, 2005). If SAR11 were a true branch within Rickettsiales, it becomes more difficult to claim that genome reduction in SAR11 occurred by a selection-driven evolutionary process distinct from the drift-dominated erosion of genomes in the Rickettsiales (Andersson and Kurland, 1998; Moran, 2002; Itoh et al., 2002). By the same token, polyphyly of nominal SAR11 strains implies that the extensive similarity in genome structure and other traits between HIMB59 and SAR11 reported by (Grote et al., 2012) may have originated independently. Perhaps convergence in some traits is consistent with streamlining, which could also explain trait-sharing between SAR11 and *Prochlorococcus*, marine Cyanobacteria also argued to have undergone streamlining (Dufresne et al., 2005). Clear signs of data-limitation in our study should be taken to mean that better taxonomic sampling will improve our results and could ultimately resolve more than two origins of SAR11-type genomes among Alphaproteobacteria.

We extracted accurate and robust phylogenetic signals from tRNA gene sequences by first integrating within genomes to identify features likely to govern functional interactions with other macromolecules. Unlike small molecule interactions, macromolecular interactions are mediated by genetically determined structural and dynamic complementarities. These are intrinsically relative; a large *neutral network* (Schuster et al., 1994) of interaction-determining features should be compatible with the same interaction network. Coevolutionary divergence — turnover—of features that mediate macromolecular interactions, while conserving network architecture, has been described in the transcriptional networks of yeast (Kuo et al., 2010; Baker et al., 2011) and worms (Barrière et al., 2012) and in post-translational modifications underlying protein-protein interactions (Beltrao et al., 2012). This work demonstrates that divergence of interaction-governing features is phylogenetically informative.

It remains open how such features diverge, with possibilities including compensatory nearly neutral mutations (Hartl and Taubes, 1996), fluctuating selection (He et al., 2011), adaptive reversals (Bullaughey, 2012), and functionalization of pre-existent variation (Haag and Molla, 2005). Major changes to interaction interfaces may be sufficient to induce genetic isolation between related lineages, as discussed for the 16S rRNA- and 23S rRNA-based standard model of the "Tree of Life," in which many important and deep branches associate with large, rare macromolecular changes ("signatures") in ribosome structure and function (Winker and Woese, 1991; Roberts et al., 2008; Chen et al., 2010).

Interaction-mediating features of macromolecules may be systems biology's answer

to the phylogeny problem. Perhaps no other traits of genomes are vertically inherited more consistently than those that mediate functional interactions with other macromolecules in the same lineage. In fact, the structural and dynamic basis of interaction among macromolecular components — essential to their collaborative function in a system — may define a lineage better than any of those components can themselves, either alone or in ensemble.

2.5 Materials and Methods

2.5.1 Data

The 2011 release of the tRNAdb-CE database (Abe et al., 2011) was downloaded on August 24, 2011. From this master database, we selected Alphaproteobacteria data as specified by NCBI Taxonomy data (downloaded September 24, 2010, Sayers et al. (2010)). Also using NCBI Taxonomy, we further tripartitioned alphaproteobacterial tRNAdb-CE data into those from the RRCH clade, the RSR grade (excluding SAR11), and three SAR11 genomes2. Five additional SAR11 genomes (for strains HIMB59, HIMB5, HIMB114, IMCC9063 and HTCC9565) were obtained from J. Cameron Thrash courtesy of the lab of S. Giovannoni. We custom annotated tRNA genes in these genomes as the union of predictions from tRNAscan-SE version 1.3.1 (with `-B` option, Lowe and Eddy (1997)) and Aragorn version 1.2.34 (Laslett and Canback, 2004). We classified initiator tRNAs and $\text{tRNA}^{\text{Ile}}_{\text{CAU}}$ using TFAM version 1.4 (Tåquist et al., 2007) using a model previously created to do this based on identifications in (Silva et al., 2006). We aligned tRNAs with covea version 2.4.4 (Eddy and Durbin, 1994) and the prokaryotic tRNA covariance model (Lowe and Eddy, 1997), removed sites with more than 97% gaps with a bioperl-based utility (Stajich et al., 2002), and edited the alignment manually in Seaview 4.1 (Gouy et al., 2010) to remove CCA tails and remove sequences with unusual secondary structures. We mapped sites to Sprinzl coordinates manually (Sprinzl et al., 1998) and verified by spot-checks against tRNAdb (Jühling et al., 2009). We added a gap in the -1 position for all sequences and -1G for tRNA^{His} in the RSR group (Wang et al., 2007a).

2.5.2 tRNA CIF Estimation and Binary Classifiers

Our tRNA-CIF-based binary phyloclassifier with Leave-One-Out Cross-Validation (LOO CV) is computed directly from function logos, estimated from tDNA alignments as described in (Freyhult et al., 2006). Here, we define a *feature* $f \in F$ as a nucleotide $n \in N$ at a position $l \in L$ in a structurally aligned tDNA, where $N = \{A, C, G, T\}$ and L is the set of all Sprinzl coordinates (Sprinzl et al., 1998). The set F of all possible features is the Cartesian product $F = N \times L$. A *functional class* or *class* of a tDNA is denoted $c \in \mathcal{C}$ where $\mathcal{C} = \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y\}$ is the universe of functions we here consider, symbolized by IUPAC one-letter amino acid codes

(for aminoacylation classes), X for initiator tRNAs, and J for $\text{tDNA}^{\text{Ile}}_{\text{LAU}}$. A *taxon set of genomes* or just *taxon set* $S \in \mathcal{P}(G)$ is a set of genomes, where G is the set of all genomes, and $\mathcal{P}(G)$ is the power set of G. In this work a genome G is represented by the multiset of tDNA sequences it contains, denoted T_G. The functional information of features is computed with a map $h : (F \times C \times \mathcal{P}(G)) \longrightarrow \mathbb{R}_{\geq 0}$ from the Cartesian product of features, classes and taxon sets to non-negative real numbers. For a feature $f \in F$, class $c \in \mathcal{C}$ and taxon set $S \in \mathcal{P}(G)$, $h(f, c, S)$ is the fraction of functional information or "Gorodkin height" (Gorodkin et al., 1997), measured in bits, associated to that feature, class and taxon set. In this work, for a given taxon set S, a function logo $H(S)$ is the tuple:

$$H(S) = \{(\alpha, \beta) \mid \beta = h(\alpha, S), \forall \alpha \in (F \times \mathcal{C})\}. \tag{2.1}$$

Furthermore the set $I(S) \subset (F \times \mathcal{C})$ of *tRNA Class-Informative Features* for taxon set S is defined:

$$I(S) = \{\alpha \in (F \times \mathcal{C}) \mid h(\alpha, S) > 0\}. \tag{2.2}$$

Briefly, a tRNA Class-Informative Feature is a tRNA structural feature that is informative about the functional classes it associates with, given the context of tRNA structural features that actually co-occur among a taxon set of related cells, and corrected for biased sampling of classes and finite sampling of sequences (Freyhult et al., 2006). Let A denote a set of alphaproteobacterial genomes partitioned into three disjoint subsets X, Y and Z with $X \cup Y \cup Z = A$, representing genomes from the RRCH clade, the RSR grade, and the eight nominal *Ca.* Pelagibacter strains respectively. To execute Leave-One-Out Cross-Validation of a tRNA CIF-based binary phyloclassifier for a genome $G \in A$, we compute a score $S_C(G, S_1, S_2)$, averaging contributions from the multiset T_G of tDNAs in G scored against two function logos $H(S_1)$ and $H(S_2)$ computed respectively from two disjoint taxon sets $S_1 \subset A$ and $S_2 \subset A$, with $G \notin S_1 \cup S_2$. In this study, those sets are $X \setminus G$ and $Y \setminus G$, denoted X_G and Y_G respectively. Each tDNA $t \in T_G$ presents a set of features $F_t \subset F$ and has a functional class $c_t \in \mathcal{C}$ associated to it. The score $S_C(G, X_G, Y_G)$ is then defined:

$$S_C(G, X_G, Y_G) \equiv \frac{1}{|T_G|} \sum_{t \in T_G} \sum_{f \in F_t} h(f, c_t, X_G) - h(f, c_t, Y_G). \tag{2.3}$$

As controls, we implemented four total-tDNA-sequence based binary phyloclassifiers to score a genome G. All are slight variations in which a tRNA $t \in T_G$ of class $c(t)$ contributes a score that is a difference in log relative frequencies of the features it shares in class-specific profile models generated from X_G and Y_G. The default "zero" scoring scheme method $S^Z_T(G, X_G, Y_G)$ shown in Fig. 2.3B is defined as:

$$S^Z_T(G, X_G, Y_G) \equiv \frac{1}{|T_G|} \sum_{t \in T_G} \sum_{f \in F_t} \log_2 \frac{p^*(f|c_t, X_G)}{p^*(f|c_t, Y_G)}, \tag{2.4}$$

where

$$p^*(f|c,S) \equiv \begin{cases} \#\{f,c,S\}/\#\{c,S\} & \#\{f,c,S\} > 0 \\ 1 & \#\{f,c,S\} = 0 \end{cases}, \tag{2.5}$$

$\#\{f,c,S\}$ is the observed frequency of feature f in tDNAs of class c in set S, and $\#\{c,S\}$ is the frequency of tDNAs of class c in set S.

2.5.3 Analysis of tRNA Base Composition

We computed the base composition of tRNAs aggregated by clades using bioperl-based (Stajich et al., 2002) scripts, and transformed them by the centered log ratio transformation (Aitchison, 1986) with a custom script. We then computed Euclidean distances on the transformed composition data, and then performed hierarchical clustering by UPGMA on those distances as implemented in the program NEIGHBOR from Phylip 3.6b (Felsenstein, 2005b) and visualized in FigTree v.1.4.

2.5.4 Supermatrix and FastUniFrac Analysis

For supermatrix approaches, we created concatenated tRNA alignments from 169 Alphaproteobacteria genomes (117 RRCH, 44 RSR, 8 PEL) that all shared the same 28 isoacceptors with 77 sites per gene (2156 total sites). In cases where a species contained more than a single isoacceptor, one was chosen at random. Using a GTR+Gamma model, we ran RAxML by means of The iPlant Collaborative project RAxML server (`http://www.iplantcollaborative.org`, Stamatakis et al. (2008)) on January 23, 2013 with their installment of RAxML version 7.2.8-Alpha (executable raxmlHPC-SSE3, a sequential version of RAxML optimized for parallelization). We tested the robustness of our result to random picking of isoacceptors by creating 100 replicate concatenated alignments and running them through FastTree (Price et al., 2010). For the FastUniFrac analysis we used the FastUniFrac (Hamady et al., 2010) web-server at `http://bmf2.colorado.edu/fastunifrac/` to accommodate our large dataset. We removed two genomes from our dataset for containing fewer than 20 tRNAs, and following (Widmann et al., 2010) removed anticodon sites. Following (Widmann et al., 2010) deliberately, we computed an approximate ML tree based on Jukes-Cantor distances using FastTree (Price et al., 2010). We then queried the FastUniFrac webserver with this tree, defining environments as genomes. We then computed a UPGMA tree based on the server's output FastUniFrac distance matrix in NEIGHBOR from Phylip 3.6b (Felsenstein, 2005b).

2.5.5 Multiway Classifier

All tDNA data from the RSR and RRCH clades were partitioned into one of seven monophyletic clades: orders Rickettsiales (N = 40 genomes), Rhodospirillales (N = 10), Sphingomonadales (N = 9), Rhizobiales (N = 91), and Caulobacterales (N = 6), or families Rhodobacteraceae (N = 43) or Hyphomonadaceae (N = 4) as specified by NCBI taxonomy (downloaded September 24, 2010, (Sayers et al., 2010)). We withheld data from the eight nominal SAR11 strains, as well as from three genera *Stappia*, *Pseudovibrio*, and *Labrenzia*, based on preliminary analysis of tDNA and CIF sequence variation. Following a related strategy as with the binary classifier, we computed, for each genome, seven tRNA-CIF-based scores, one for each of the seven alphaproteobacterial clades as represented by their function logos, using the principle of Leave-One-Out Cross-Validation (LOO CV), that is, excluding data from the genome to be scored. Function logos were computed for each clade as described in (Freyhult et al., 2006). For each taxon set X_G (with genome G left out if it occurs), genome G obtains a score $S^M(G, X_G)$ defined by:

$$S_M(G, X_G) \equiv \frac{1}{|T_G|} \sum_{t \in T_G} \sum_{f \in F_t} h(f, c_t, X_G). \tag{2.6}$$

Each genome G is then represented by a vector of seven scores, one for each taxon set modeled. These labeled vectors were then used to train a Multilayer Perceptron classifier in WEKA 3.7.7 (downloaded January 24, 2012, (Hall et al., 2009)) by their defaults through the command-line interface, which include a ten-fold cross-validation procedure. We bootstrap resampled sites in genomic tRNA alignment data (100 replicates) and also bootstrap resampled a reduced (and retrained) model including only CIFs with a Gorodkin height ≥ 0.5 (Freyhult et al., 2006).

2.6 Appendix – Supplementary Data

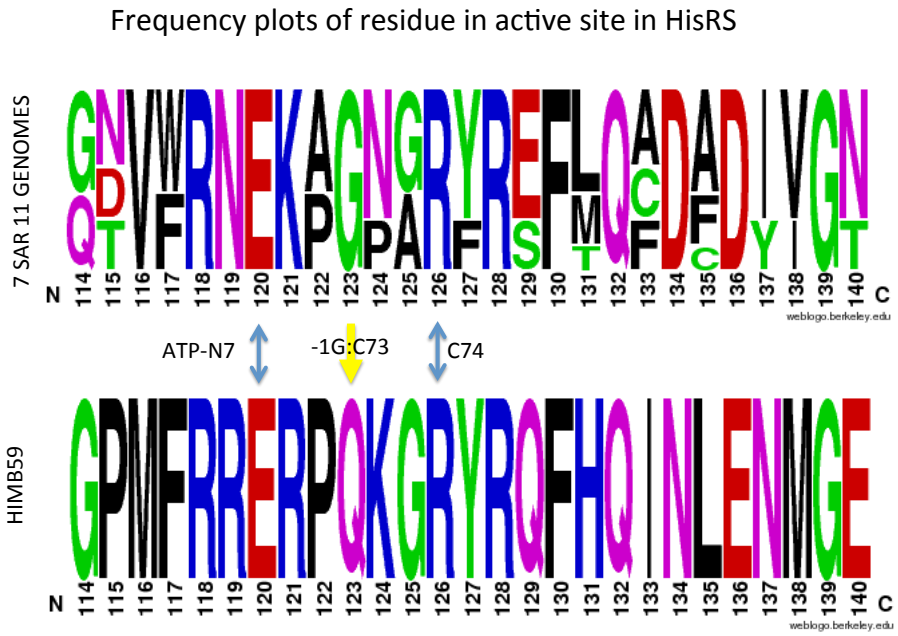


Figure 2.7: **Frequency plot logos of the motif IIb tRNA-binding loop of inferred HisRS proteins from putative SAR11 strain genomes.** These results should be compared to Figure 3 of Ardell and Andersson (2006). Seven of eight putative SAR11 genomes show derived characteristics of HisRS (shown here at top) unique to the RRCH clade, while one, HIMB59, shows ancestral characteristics common to all other bacteria. These data co-vary perfectly with $tRNA^{His}$ data and imply perfect covariation consistent with monophyly of the top seven strains with the RRCH clade, and affiliation of HIMB59 with the RSR grade.

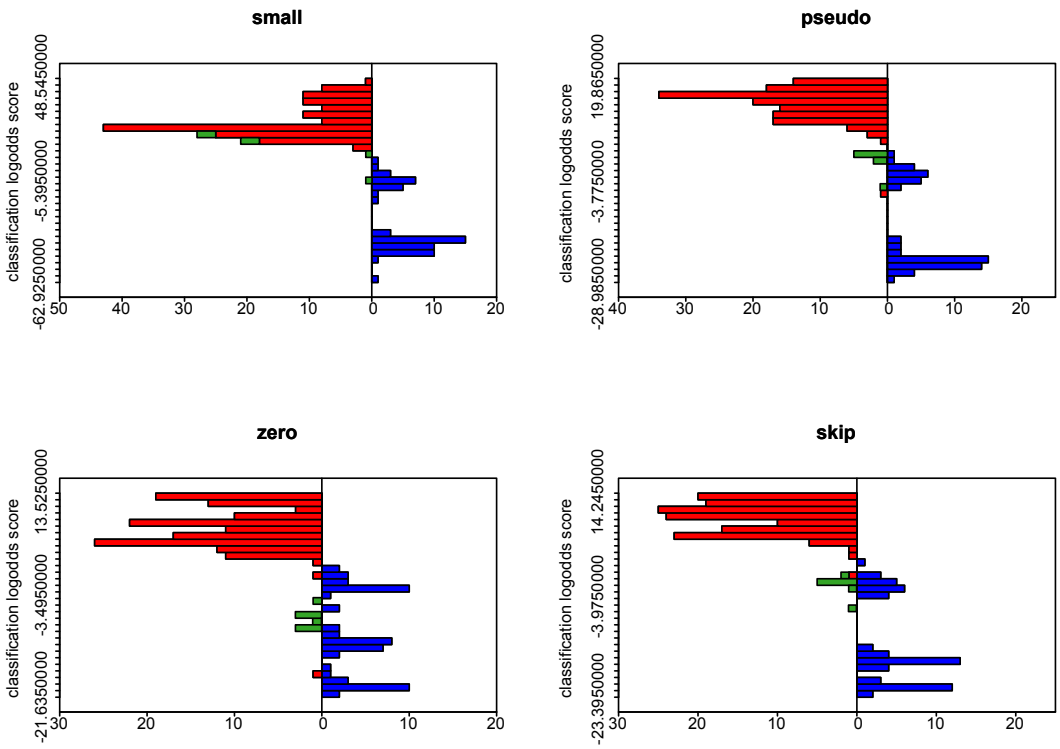


Figure 2.8: **Histograms of leave-one-out cross-validation (LOO-CV) scores of alphaproteobacterial genomes under the tRNA sequence-based binary phyloclassifier**, using four different methods for handling missing data, when a genome presents tRNA features missing from one or the other training data sets for the RRCH clade (in red) or RSR grade (in blue). Pelagibacter data is in green. Method "zero" is shown in the main text as Figure 2.3. For definitions of methods, please see the Methods and Materials section in this chapter.

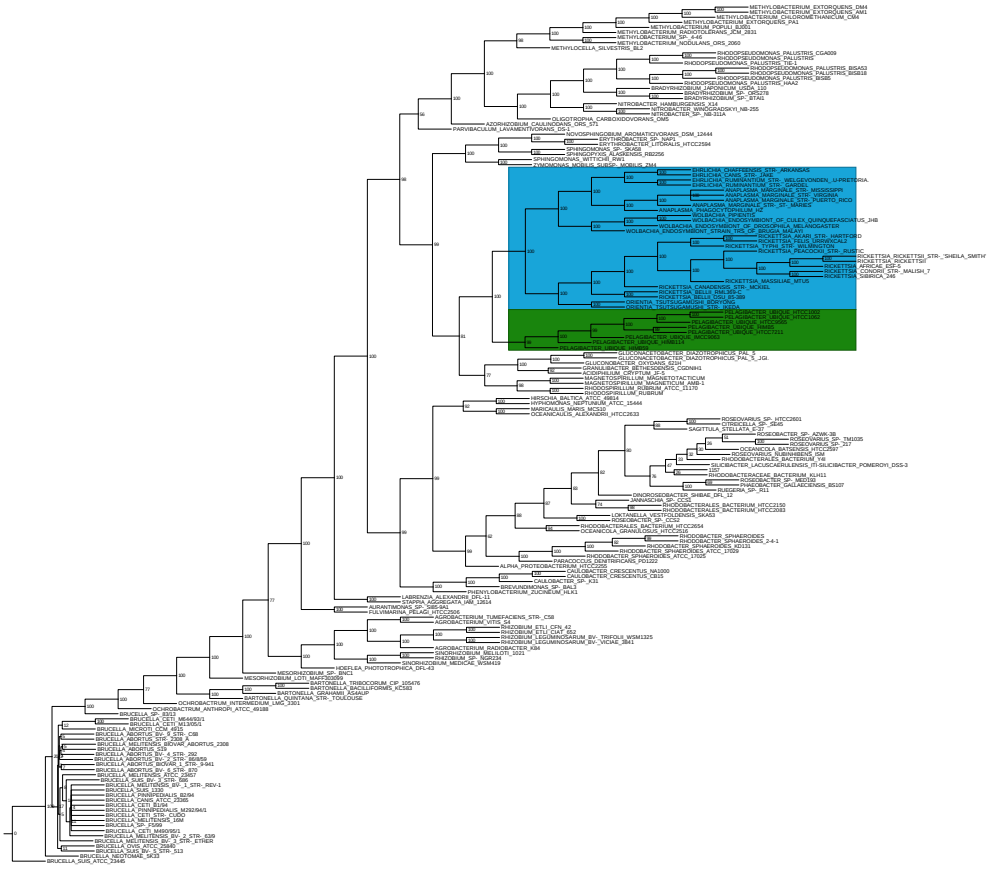


Figure 2.9: **Maximum likelihood phylogram of a concatenated supermatrix of 28 isoacceptor genes for 169 alphaproteobacterial genomes computed in RAxML using the GTR+Gamma model.** For genomes in which paralog "isodecoders" of the same isoacceptor gene, one paralog was picked randomly. This occurred in 31% of cases, where a case is one genome $\times$ isoacceptor combination. Rickettsiales genomes are boxed in blue and all eight putative SAR11 strains are boxed in green.

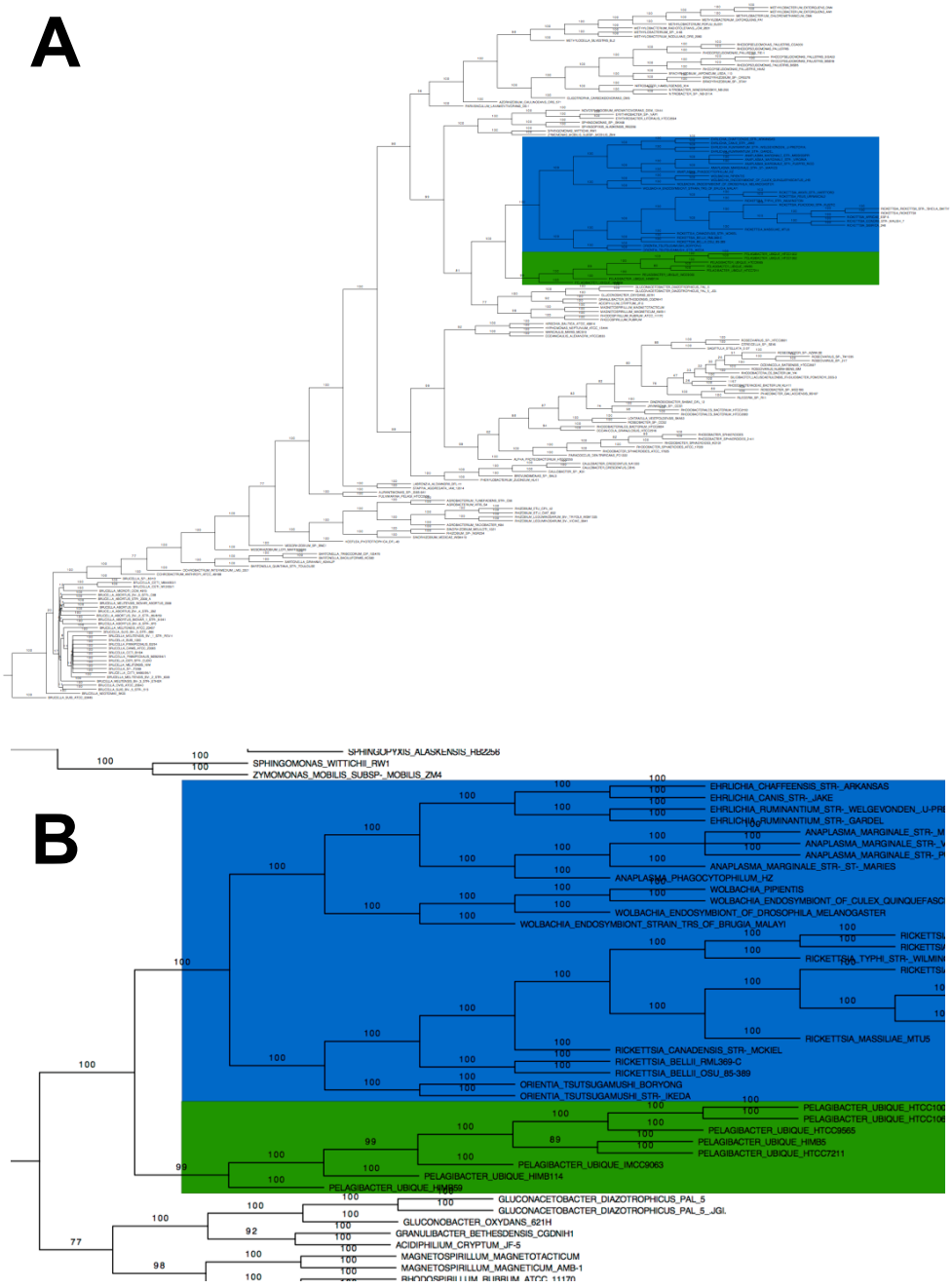


Figure 2.10: **Consensus cladogram of 100 replicates of distance- based trees computed in FastTree, each with different randomized picks of isoacceptor genes for alphaproteobacterial genomes in which paralogs for the same isoacceptor exist.** A. Complete cladogram, with Rickettsiales boxed in blue and putative SAR11 genomes, including HIMB59, in green. B. Magnification showing perfect replicate support for monophyly of Rickettsiales and the eight putative SAR11 strains.

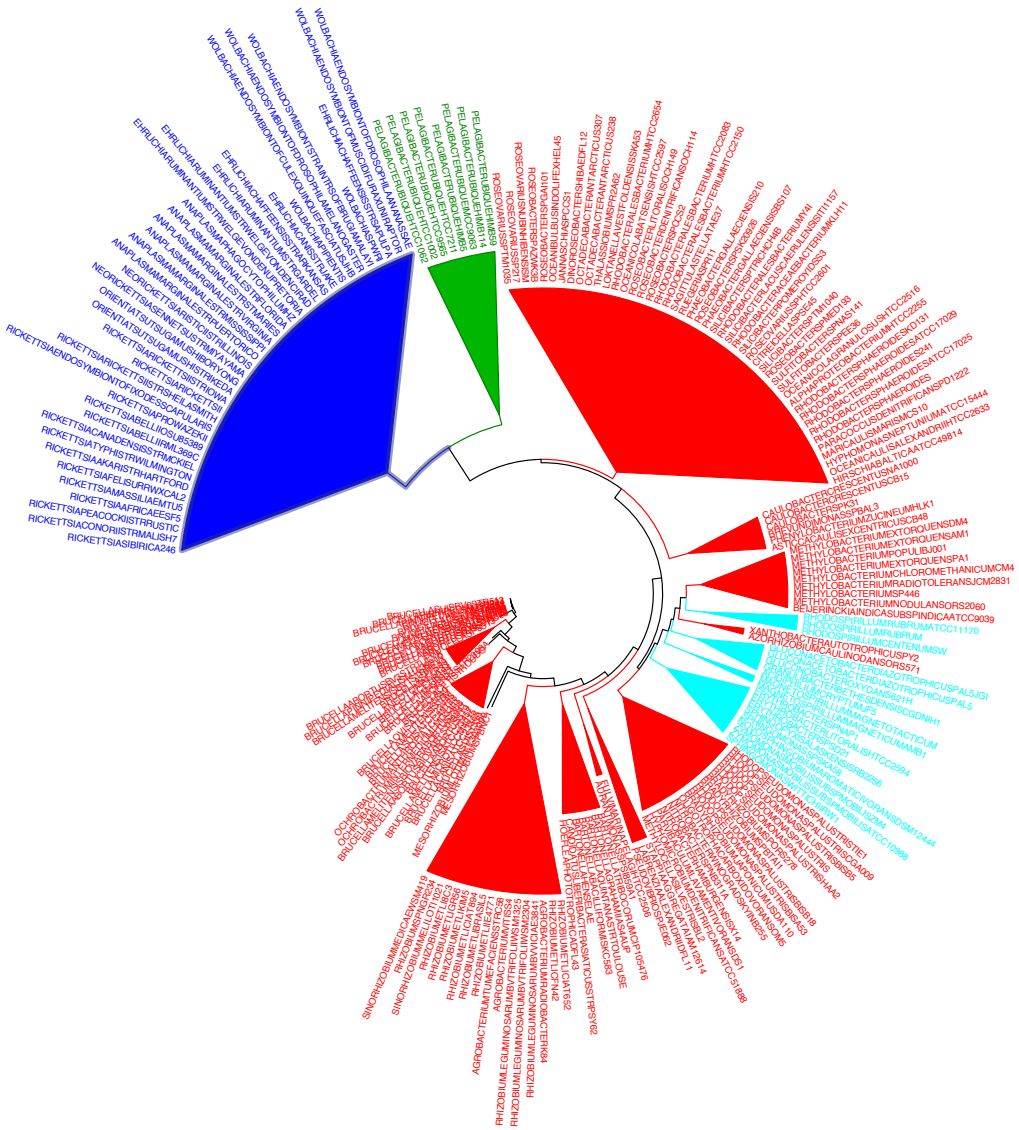


Figure 2.11: **FastUniFrac-based phylogenetic tree of Alphaproteobacteria using tRNA data as computed according to the methods of Widmann et al. (2010).** As elsewhere, blue are the RSR grade including Rickettsiales, green are SAR11, and red is the RRCH clade.

Chapter 3

tRNA Class-Informative Features locate the root of Plastids within Cyanobacteria

3.1 Abstract

While eukaryotic Plastids indisputably originated from a cyanobacterial ancestor, their exact rooting is unresolved. Such ancient evolutionary relationships challenge molecular phylogenetics because of saturation and base composition effects, genome reduction, and horizontal gene transfer. Features that target tRNAs within protein-tRNA interaction networks (called Class-Informative Features or CIFs) can be predicted bioinformatically from any group of related genomes, including Plastid genomes. Here we show that tRNA CIFs diverge slowly in a phylogenetically informative manner, and can be used to construct a bootstrapped distance-based phylogeny that supports widely accepted deep phylogenetic relationships within bacteria and Plastids. Furthermore, our results unequivocally root Plastids after the divergence of marine Prochlorococcus and Synechococcus clades ("marine Cyanobacteria"), resolving a major open question about Plastid origins. One among several derived tRNA CIFs that Plastids and their cyanobacterial sister clade share is a unique Glu-tRNA recognition element previously described only in Plastids, which putatively functions in tetrapyrrole (heme and chlorophyll) biosynthesis via the C5 pathway. tRNA CIFs are universal, slowly diverging, systems biological traits that are less prone to biases affecting traditional molecular phylogenetic data. As such, they promise to resolve other outstanding deep phylogenetic relationships in the Tree of Life.

3.2 Introduction

It is generally agreed that all cellular organisms share a common terrestrial origin (Theobald, 2010). More controversial, however, is the extent to which they have descended vertically from a common ancestor (Woese, 2002; Koonin and Wolf, 2009; McInerney et al., 2011). This controversy — of how tree-like the Tree of Life is — has

been fueled in part by limitations in methodology. Conventional molecular phylogenetics can fail to accurately resolve tree-like signals where they truly exist, if saturation effects, long-branch attraction and compositional bias artifacts are not carefully accounted for (Philippe et al., 2005). Phylogenomic studies show evidence of increasingly pervasive Horizontal Gene Transfer (HGT) early in the history of life, opening the question of what biological essence is vertically inherited if not entire genomes and how can this be accurately estimated (Abby et al., 2012; Wolf et al., 2002; Woese, 2000)? A powerful alternative to conventional methods lies in the analysis of Signatures or Rare Genetic Changes (Gupta, 2010), which suggest that more ancient tree-like signals exist to be estimated (Brochier and Philippe, 2002; Rivera and Lake, 2004). But rare genetic changes, difficult to find, are not generally universal phylogenetic markers.

Cellular binary fission would seem to imply an ancient and essential pattern of vertical inheritance characteristic to all of life's history. But what biological essence is vertically coinherited and how best can we estimate the pattern? Gene products must interact with each other through shape and dynamic complementarity in order to cofunction (Williamson, 2000). Unlike the interactions of macromolecules with small molecules, macromolecular (eg RNA-protein, RNA-RNA and protein-protein) interactions are shaped at interfaces that are entirely genetically encoded. We hypothesized that these interfaces may slowly diverge while maintaining specificity of interaction (Ardell, 2010). The coadaptations at macromolecular interfaces maintaining specificity of interactions may be the defining essence of vertical inheritance and the best source of data to estimate deep roots of the Tree of Life. Although the degree of macromolecular interactions are known to correlate inversely with rates of substitution (Fraser et al., 2002) and HGT (Jain et al., 1999), features that define macromolecular interfaces have not previously been specifically exploited for phylogeny. Such data may be robust to HGT, because while components of an interaction might transfer and even perturb a macromolecular interaction network, the "shape code" that dictates specificity of interactions (Ardell, 2010) cannot easily be transferred form one lineage to another.

In order to exploit this idea for deep phylogeny, we must be able to predict many different features governing universal macromolecular interactions from genomic data alone, and compare their presence, absence and homology across lineages. Transfer RNA genes are perhaps uniquely well-suited to this task. Different functional classes of tRNAs must conform to the same overall structure to interact with general translational factors and the ribosome, yet be structurally distinct in order to interact specifically with proteins like aminoacyl-tRNA synthetases according to their distinct functional identities. This eases sequence-based comparisons to predict Class-Informative Features (CIFs). The tRNA CIFs are (a) highly enriched for known tRNA identity elements and predict novel ones (Freyhult et al. (2006), Chapter 2), (b) coevolve with tRNA-interacting enzymes at both residue and domain levels (McClain, 1993; Giegé et al., 1993), and (c) diverge over the Tree of Life at least at the phylum-level (Giegé et al., 1998), for example between Proteobacteria and Cyanobacteria (Freyhult et al., 2007). We recently showed that a unique tRNA CIF divergence within the Alphaproteobacteria is apparently informative

about the controversial phylogenetic affinity of the prominent marine microflora *Ca.* Pelagibacter ubique (Ardell (2010); Wang et al. (2007b); Chapter 2)

Here we developed new methods to systematically analyze tRNA CIFs and applied them to study evolutionary relationships within and among Cyanobacteria and eukaryotic Plastids. Although the cyanobacterial endosymbiotic origin of eukaryotic Plastids was proposed more than a century ago (Mereschkowsky, 1905), consensus has only recently emerged on a single origin of all extant chloroplasts (except the chromatophore in *Paulinella chromatophora* (Bodył et al., 2012)). The precise rooting of Plastids within the cyanobacterial tree has been called one of the last great unresolved questions regarding this landmark evolutionary event (Keeling et al., 2004), and is challenging because of extensive divergence and massive genome reduction in Plastids, more than ten-fold reduction in gene number compared to most modern Cyanobacteria (Martin and Herrmann, 1998). While early analysis linked green eukaryotic phototrophic Plastids with prochlorophytes based on shared traits of chlorophyll b and a lack of phycobilisomes (Cavalier-Smith, 1981), sequence-based evidence suggested a more modern origin, discordantly depending on dataset and methods used (Martin et al., 1992; Archibald, 2009). In a recent phylogenomic analysis (Criscuolo and Gribaldo, 2011), Plastids were rooted after the early-branching Gleobacter violaceus and two Yellowstone Synechococcus strains, but before the major cyanobacterial divergence between the marine, unicellular Synechococcus and Prochlorococcus ("Marine" Cyanobacteria) and the remainder of the diverse cyanobacterial lineages ("Core" Cyanobacteria).

3.3 Methods and Discussion

We custom annotated tRNA gene-predictions in a collection of proteobacterial, cyanobacterial and Plastid genomes. Plastid tRNA genes, unlike mitochondrial tRNA genes, are mostly canonical in structure (Barbrook et al., 2010), although we did find some notable exceptions. Annotated $\mathrm{tRNA}^{\mathrm{Ile}}_{\mathrm{CAU}}$ contained large regions in the variable arm that mimicked those of $\mathrm{tRNA}^{\mathrm{Met}}_{\mathrm{CAU}}$ and were thus misclassified. We used conditional information (Freyhult et al., 2006) to profile tRNA class-informative base-pairs and single nucleotides within phylogenetic clades in a manner that corrects for biased class sampling (Gorodkin et al., 1997). We then developed a weighted distance metric on these profiles based on the square root of the Jensen-Shannon Divergence (Lin (1991); Endres and Schindelin (2003); Österreicher and Vajda (2003), Figure 3.1). and bootstrapped tRNA CIFs in order to generate a consensus distance-based phylogenetic tree (Desper and Gascuel, 2002; Felsenstein, 2005a) 3.2. Our tRNA CIF-based tree (Figure 3.2) contains robust support for widely accepted branching orders within Proteobacteria (Ciccarelli, 2006; Yarza et al., 2010), Cyanobacteria (Tomitani et al., 2006; Shih et al., 2013), and the eukaryotic phototrophs (Martin et al., 2002). Remarkably, our method robustly roots Plastids as a sister clade to the core Cyanobacteria, after the divergence of both the basal Gleobacter/Synechococcus and the marine Prochlorococcus/Synechococcus Cyanobacteria in 85% of our bootstrap replicates(Figure 3.2). The alternative hypothesis

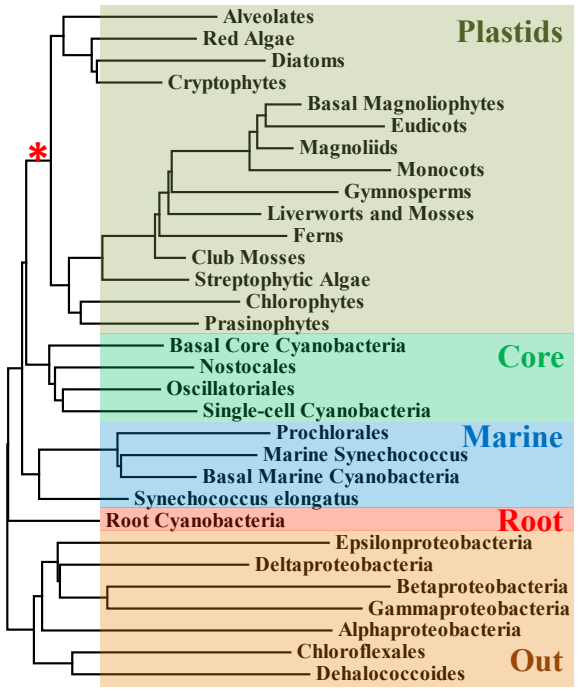


Figure 3.1: **Jensen-Shannon Divergence tree calculated for Cyanobacteria, Plastids and Proteobacteria.** All literature-supported topologies (especially Proteobacteria) are obtained, and Plastids sister with the "Core" Cyanobacteria, which branch from the marine Cyanobacteria which contains Synechococcus strains.

most recently published by Criscuolo and Gribaldo (2011), rooting Plastids shortly after the divergence of Gleobacter and Yellowstone Synechococcus, is the second best topology in our analysis by a wide margin (see Figures 3.3 and 3.4).

We used a permutation method to declare sets of significantly informative tRNA CIFs and investigate further the strength of support in our data for our main result. Remarkably, among several derived tRNA CIFs shared between Plastids and "Core" Cyanobacteria, the most informative feature in our data is a highly unusual A53-U61 base-pair in $tRNA^{Glu}$ (See Figure 3.4). This trait shared by Plastids and "Core" Cyanobacteria is unique among all tRNAs in sequenced genomes, and has been previously implicated in recognition of Glu-$tRNA^{Glu}$ by Plastid Glu-tRNA Reductase (GluTR) the first enzyme functioning in tetrapyrrole (and ultimately heme and chlorophyll) biosynthesis by the C5 pathway (Stange-Thomann et al., 1994) and was not previously observed or described in Cyanobacteria. RNAse protection assays implicate positions 53, 54, and 55 of Glu-$tRNA^{Glu}$ as sites of interaction with GluTR (Randau et al., 2004) and mutation of position 57 decouples protein and tetrapyrrole biosynthesis (Stange-Thomann

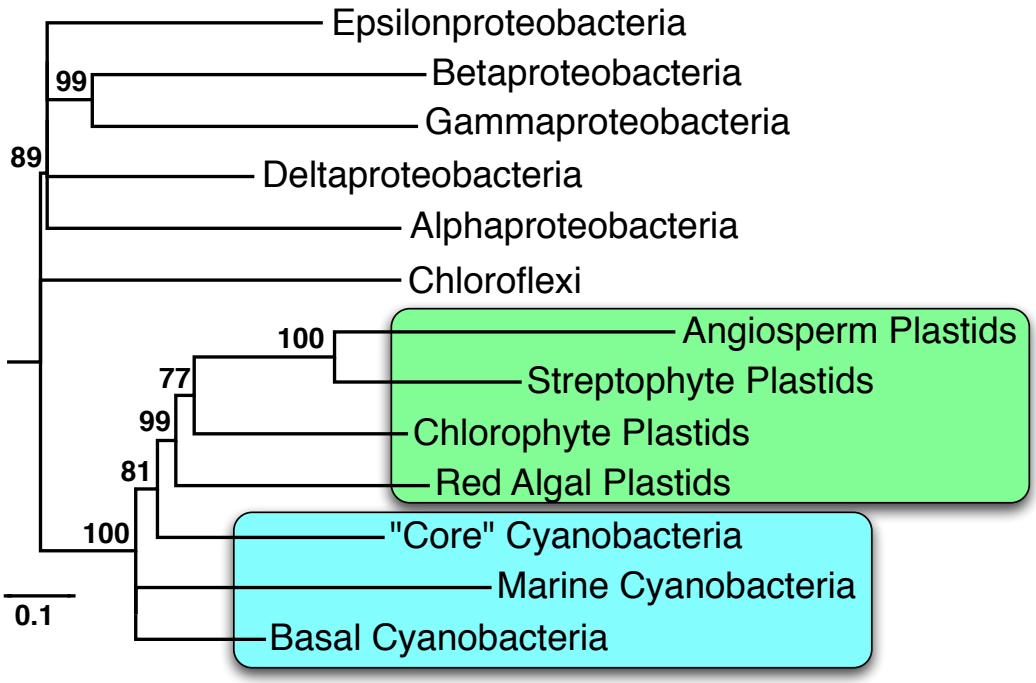
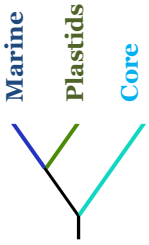


Figure 3.2: **Jensen-Shannon Divergence tree with bootstrap values.** 100 bootstrap replicates with bootstrapped tRNA sites for each replicate. There is strong support for a Core Cyanobacteria/Plastid sistering.

A **Literature-supported Topology**

CIF	Information Difference	Core + Marine Cyano	Plastid	basal Cyano	Proteobacteria + Chloroflexi
G48-J	3.644	100.00%	2.82%	0.00%	4.18%
C64-X	0.207	86.44%	3.29%	0.00%	3.92%
G50-X	0.153	86.44%	3.76%	0.00%	7.49%

B **Old-Fashioned Topology**

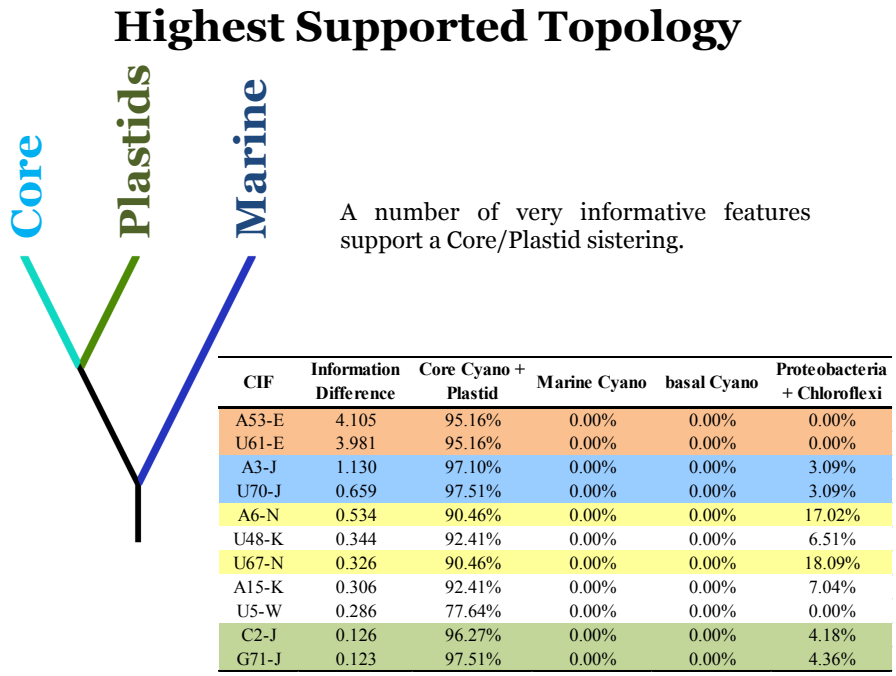


CIF	Information Difference	Marine Cyano + Plastid	Core Cyano	basal Cyano	Proteobacteria + Chloroflexi
C43-K	0.166	71.30%	24.24%	0.00%	25.70%
A39-L	0.166	83.60%	18.18%	33.33%	18.42%
A39-R	0.156	78.09%	9.09%	0.00%	12.46%

Figure 3.3: **Class Informative Feature support for (A) literature supported and (B) classic cyanobacterial ancestry.** The literature supported topology has one supporting CIF which could be explained by horizontal gene transfer and the original topology is not supported by any strong shared CIFs.

et al., 1994; Levicán et al., 2007)(See Figure 3.5). It is compelling to speculate whether this tRNA CIF may have evolved to help partition Glu-tRNA^{Glu} between the protein and tetrapyrrole biosynthesis pathways. Subtle changes in the chlorophyll biosynthetic pathway may have had a profound effect on the fitness of the cyanobacterial progenitor in its host eukaryote.

We assessed two alternative topologies around the "Core" Cyanobacteria/Plastid split (see Figure 3.3). Both alternatives yield informative features, albeit far fewer than the Plastid-core Cyanobacteria monophyly (see Figure 3.4). The top-scoring CIF linking the two main cyanobacterial clusters together in exclusion to the Plastids is base G48 in tRNA$^{Ile}_{CAU}$ (class J). It is the only highly informative feature supports a Core/Marine pairing; however, this is likely explained by an early loss in Plastids. As for the old-fashioned topology, springing from the original discovery of Chlorophyll b-containing marine Cyanobacteria (Marine Cyanobacteria/Plastid), very few weak features support a Marine/Core pairing (see Figure 3.3) which sparked debate on whether Plastids originated from these prochlorophytes. Limited CIFs supporting the two alternative topologies



CIF	Information Difference	Core Cyano + Plastid	Marine Cyano	basal Cyano	Proteobacteria + Chloroflexi
A53-E	4.105	95.16%	0.00%	0.00%	0.00%
U61-E	3.981	95.16%	0.00%	0.00%	0.00%
A3-J	1.130	97.10%	0.00%	0.00%	3.09%
U70-J	0.659	97.51%	0.00%	0.00%	3.09%
A6-N	0.534	90.46%	0.00%	0.00%	17.02%
U48-K	0.344	92.41%	0.00%	0.00%	6.51%
U67-N	0.326	90.46%	0.00%	0.00%	18.09%
A15-K	0.306	92.41%	0.00%	0.00%	7.04%
U5-W	0.286	77.64%	0.00%	0.00%	0.00%
C2-J	0.126	96.27%	0.00%	0.00%	4.18%
G71-J	0.123	97.51%	0.00%	0.00%	4.36%

Figure 3.4: **List of the top Class-Informative Features shared between the "Core" Cyanobacteria and the Plastids.** Percent conservation is also shown in the table, supporting the sistering of the "Core" Cyanobacteria and the Plastids.

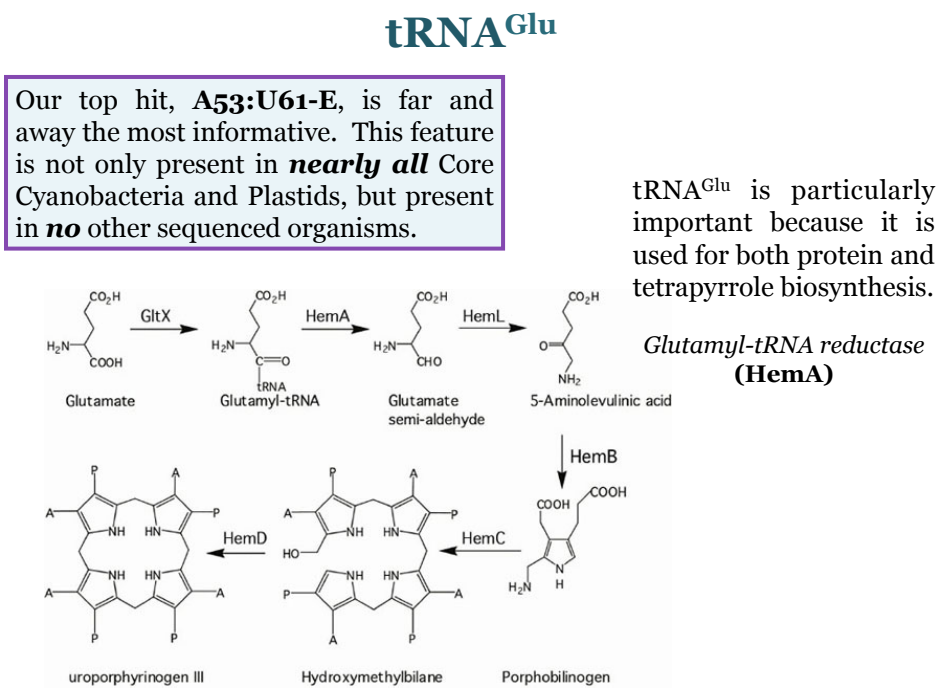


Figure 3.5: **A schematic of tetrapyrrole biosynthesis, and its relationship to a dominating CIF.**

suggests that previous attempts to reconstruct the evolutionary origin of Plastids may have been hampered by the long temporal distance to this root, akin to the difficulty in placing *Synechococcus elongatus* (Robertson et al., 2001), discussed in further detail below.

In addition, we also tested the suggested sistering of Plastids with the heterocystous Nostocales (Deusch et al., 2008), but this pairing yielded no CIFs above our threshold (data not shown). While our methods offered no support for this sistering, the localization of Plastids within the core cyanobacterial lineage supports the possibility of a nitrogen-fixing Plastid progenitor with a gene complement that might have been as comprehensive as the large Nostocales. This could also account for the apparent similarity between the gene complement in Plastids (and their host genomes) and Nostocales–higher than that for any other cyanobacterial clade (Deusch et al., 2008).

3.3.1 Resolving deep-branching species

While our algorithm relies on functional tRNA information from clades (our smallest analyzed datasets contain three species), we also scored individual species against model clades in order to clarify the evolutionary history of species that have historically been difficult to resolve. Chief among these species in the Cyanobacteria are *S. elongatus* (two closely related genomes PCC 6301 and PCC 7942) and Synechococcus *sp. PCC 7335*, which have been suggested to form a monophyletic cluster along the marine branch near the divergence of marine and "Core" clusters, though this branching location is inconsistent. Using our algorithm, both *S. elongatus* strains do, in fact, score highest against the marine cluster. However *Synechococcus sp. PCC 7335* shows clear association with the "Core" and Plastid clusters–even including the indicative A53:T61 base pair in its $tRNA^{Glu}$–a position supported by several trees presented in previous work (Criscuolo and Gribaldo, 2011; Falcón et al., 2011). While this branching position for PCC 7335 places it closest to the root of the Plastid lineage, there is no clear evidence that this strain is a direct ancestor of the progenitor species.

Cyanophora localization and red-bias

Cyanophora paradoxa is suggested to branch prior to the divergence of the red and green algal lineages (Bhattacharya and Weber, 1997; Price et al., 2012), however this position has been controversial (Deschamps and Moreira, 2009; Qiu et al., 2012). Unlike all other phototrophic eukaryotes, the Cyanophora Plastid, termed the cyanelle, contains relic cyanobacterial features, including a peptidoglycan cell wall and carboxysome-like carbon-concentrating mechanisms (Fathinejad et al., 2008). We have also recovered that red algal plastids are definitely closer related to Cyanobacteria than green algal plastids.

3.4 Conclusion

In this chapter, we present a novel method for creating phylogenetic trees which recovers all basic supported topologies (Proteobacteria and Plastids), but provides a different topology for the branching of the Plastids from the Cyanobacteria. From this work, the highest supported topology includes the "Marine" Cyanobacteria basally branching to the "Core" Cyanobacteria and the Plastids. We show little to no support for any other topology with our CIFs. Our top CIF, A53:U61-E is present in nearly all Core Cyanobacteria and plastids, and no other sequenced organisms. $tRNA^{Glu}$ is particularly important due to its dual use in both protein and tetrapyrrole biosynthesis. Our classification of the three *Synechococcus elongatus* cluster strains show that they are very strongly tied to the Core Cyanobacteria. We have also shown that strain PCC 7335 should not be included in the S. elongatus cluster. This method can be used for creating any tree with enough genomes sequenced (three genomes per grouping).

Recently, a study from the CyanoGEBA project (Shih et al., 2013) came out reporting over 50 genomes that we feel will need to be incorporated into our analysis to strengthen our results. Data curation for this project has already begun, included in the following chapter. Once the pipeline has been applied to this new data, the results will be updated and then submitted for publication.

Chapter 4

Identifying conserved traits throughout the bacterial tree of life – an exploratory analysis

4.1 Abstract

The search for a signal that is functionally conserved across the tree of life, and immune to biological bias has lead us to investigate tRNA Class-Informative features (CIFs). It has been shown in this dissertation that, in two specific cases, CIFs in combination with machine learning can be used to classify genomes that are unique in some form that cause classical phylogenetic methods to misrepresent their taxonomic relationships. In this work, we have developed an general bacterial classifier that will use tRNA CIFs to taxonomically sort bacteria into orders already populated with sequenced genomes. In our investigations, we determined that tRNA CIFs are a promising marker to classify most bacteria, and rare misclassifications can be explained. In an investigation of CIF properties to justify their value in bacterial classification, we determined that the averaging of weighted CIFs compensate for some tRNAs with a G+C bias. A site-specific analysis determined that in some sites, Information is consistent across orders. This unique trait could infer functional importance in the tRNA interaction network that can be recovered through information theory and tRNA sequence alone. We conclude that CIFs are a promising phylogenetic signal that deserve more investigation for utilization to track vertical inheritance. This work paves the way for a streamlined utilization of all sequenced genomes, and more careful training of a classifier will alleviate incurred issues.

4.2 Introduction

tRNA identity elements function as essential pieces of a vital system throughout life. In chapter 2, we demonstrated that tRNA Class-Informative features, computed with Function Logos (Freyhult et al., 2006), can be used to classify bacteria with extreme

compositional bias. In chapter three, we demonstrated that tRNA Class-Informative features can be used as phylogenetic signal to bifurcate trees with long branches using Jensen-Shannon Information Difference (Lin, 1991) as a distance metric. It is known that CIFs are distinct and meaningful based on experimental verification (reviewed in Giegé (2008)). If one randomizes class associations and calculates function logos, little to no letters are visible in the graph of CIFs.

Questions arise when investigating using CIFs for phylogenetic signal which we attempt to address in this chapter.

(1) Are the CIFs conserved? More specifically, is it extremely important for a synthetase-tRNA contact to maintain the actual signal? We would expect to see universally conserved CIFs if the actual CIF is important. Either answer to this question will provide information on how other relatively closed systems might evolve across the tree of life.

(2) Is the amount of information conserved? We would expect to see CIFs existing in all of the same sites, regardless of their make-up if the contact is the only important feature of a CIF.

(3) Can CIFs travel through different loci in a tRNA of the same class? Is it a combination of a few, or all of these factors that contribute to the evolution of informative sites in the tRNA interaction network?

(4) how noisy are CIFs?

This final chapter will investigate these concepts in order to pave the way for fine-tuning of a classifier that is robust to biological and methodological biases. Given the exciting results from Chapter 2 and Chapter 3 analyses, it proved necessary to start dissecting the properties of CIFs, and how they have evolved/are evolving.

4.3 Results and Discussion

4.3.1 G+C content across bacterial orders

In 79 orders, G+C content varies, with extremities in the bacteria in drastic environments, including the Desulfurococcales, Sulfolobales, and Thermococcales. An evaluation of transfer-RNA G+C content across the tree of life shows that they vary drastically (See Figure 4.1A). When you only look at their Information summation for each of the four states {A,U,C,G}, the signal starts to look like other orders, despite their extremities in tRNA base content (See Figure 4.1B). This suggests that the *functional* residues in tRNAs share roughly the same level of information available to charging synthetases, represented by the heights of the letters summing to roughly the same normalized frequencies in all four types of features.

4.3.2 Classifier development by order

Following the procedure in Chapter 2 with 79 groups instead of seven, we have trained a Multilayer Perceptron with scores from 2374 genomes to the 79 different groups (see

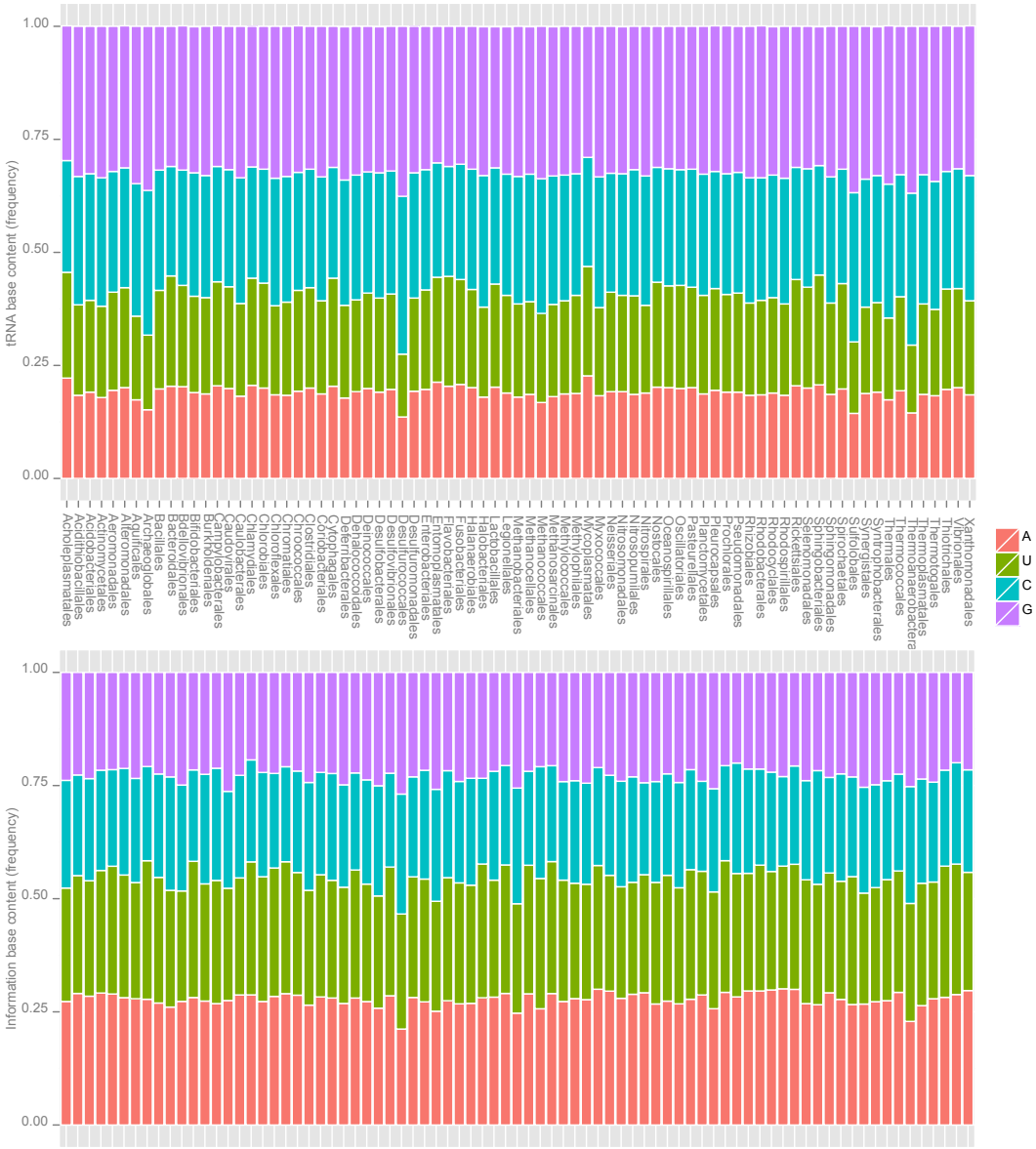


Figure 4.1: **Base content of (upper) transfer RNA and (lower) the weighted base content of their class-informative features.** In the lower plot, total information is summed in each graph, and then normalized by number of sequences.

Table 4.1: **Count of bacterial genomes and tRNAs represented in each taxonomic order in curated dataset.**

order	# genomes	# tRNAs	order	# genomes	# tRNAs
Acholeplasmatales	6	202	Methanobacteriales	8	350
Acidithiobacillales	4	224	Methanocellales	3	153
Acidobacteriales	5	241	Methanococcales	15	566
Actinomycetales	219	11197	Methanosarcinales	13	675
Aeromonadales	6	626	Methylococcales	3	135
Alteromonadales	48	4093	Methylophilales	5	234
Aquificales	12	520	Mycoplasmatales	65	2096
Archaeoglobales	5	238	Myxococcales	11	647
Bacillales	185	14625	Neisseriales	21	1302
Bacteroidales	24	1424	Nitrosomonadales	5	202
Bdellovibrionales	4	140	Nitrosopumilales	3	124
Bifidobacteriales	32	1758	Nitrospirales	4	194
Burkholderiales	95	5461	Nostocales	12	741
Campylobacterales	87	3432	Oceanospirillales	15	784
Caudovirales	4	213	Oscillatoriales	9	486
Caulobacterales	7	349	Pasteurellales	32	1830
Chlamydiales	88	3281	Planctomycetales	6	368
Chlorobiales	11	518	Pleurocapsales	3	134
Chloroflexales	5	239	Prochlorales	12	473
Chromatiales	13	604	Pseudomonadales	63	4126
Chroococcales	36	1613	Rhizobiales	97	5008
Clostridiales	113	7176	Rhodobacterales	22	1120
Coriobacteriales	8	379	Rhodocyclales	6	356
Cytophagales	12	537	Rhodospirillales	26	1535
Deferribacterales	4	169	Rickettsiales	63	2132
Dehalococcoidales	7	327	Selenomonadales	7	400
Deinococcales	8	384	Sphingobacteriales	8	462
Desulfobacterales	8	408	Sphingomonadales	14	714
Desulfovibrionales	17	1027	Spirochaetales	53	2081
Desulfurococcales	12	550	Sulfolobales	19	859
Desulfuromonadales	11	570	Synergistales	4	193
Enterobacteriales	216	16972	Syntrophobacterales	4	192
Entomoplasmatales	3	94	Thermales	12	575
Flavobacteriales	44	1780	Thermoanaerobacterales	29	1483
Fusobacteriales	6	297	Thermococcales	15	691
Halanaerobiales	5	311	Thermoplasmatales	4	184
Halobacteriales	23	1122	Thermotogales	16	757
Lactobacillales	184	11419	Thiotrichales	23	885
Legionellales	17	729	Vibrionales	23	2520
			Xanthomonadales	27	1528

table 4.1 for details). All bacterial orders, except the viruses in Caudovirales (four genomes) are primarily filled with probability density from their order, meaning that most all genomes classify with the correct order (see Figure 4.2). Misclassifications are described, and in some cases justified, in the next section, which are arguably excellent positive controls.

4.3.3 Detection of misclassified bacteria

Using a Multilayer Perceptron to train a classifier for classifying bacteria from 79 taxonomic orders, ten genomes misclassify according to their NCBI taxonomy (0.4% of all total data) . *Saccharophagus degradans 2 40* (prob 0.446), currently classified in the gammaproteobacterial order Alteromonadales (48 genomes), the only species currently defined in its genus, classifies with the gammaproteobacterial Methylococcales (3 genomes). *Teredinibacter turnerae T7901* (prob 0.697), another member of Alteromonadales (48 genomes), groups with gammaproteobacterial order Oceanospirillales (15 genomes) Both of the previously listed strains are ill sequenced in their respective genera, and thought to be relatively closely related to each other by 16S rRNA comparison (Ekborg et al., 2005; Yang et al., 2009). This could be a similar case to the *Pseudovibrio, Stappia* and *Labrenzia* strains in chapter 2, where they are so distantly related to their sequenced sister taxa that they don't classify well to anything. *Stappia* and *Labrenzia* strains are not classified into any specific order in the training set, but when tested against the training set, they classify strongly with Rhizobiales). Further, some of these genomes contain selenocysteine residues, and we do not incorporate selenocysteine class tRNAs into our analysis, which are charged by a unique synthetase (Commans and Böck, 1999). This could be implemented to improve the classifier.

Candidatus Hodgkinia cicadicola Dsem, in the order Rhizobiales, classifies as it is from order Caulobacterales. *Halothiobacillus neapolitanus c2*, a bacterium in gammaproteobacterial order Chromatiales classifies as if it is in the order gammaproteobacterial Oceanospirillales. *Leptolyngbya PCC 7376*, from cyanobacterial Oscillatoriales classifies as cyanobacterial order Chroococcales. *Chamaesiphon PCC 6605*, located in cyanobacterial order Chroococcales classifies as if it is from cyanobacterial Nostocales. Both of these genomes were recently sequenced and taxonomically grouped based on 16S rRNA sequences (Shih et al., 2013).

Aggregatibacter actinomycetemcomitans D11S 1, *Streptococcus salivarius 57 I*, *Thermoanaerobacterium saccharolyticum JW SL YS485*, and *Brachyspira hyodysenteriae WA1* which make up all of the fully sequenced members of the Caudovirales–group 1 viruses (characterized by double stranded DNA)–classify with other various groups, showing that this does not look good for virus classification, and it should be further investigated if the viruses classify with genomes in which they have shared DNA.

We then attempted to score the eight SAR11 genomes, and *Stappia*, *Pseudovibrio* and *Labrenzia* genome from Chapter 2 against the 79-way classifier. The only strain included in the creation of the classifier was *Pseudovibrio*. *Stappia* and *Labrenzia* score

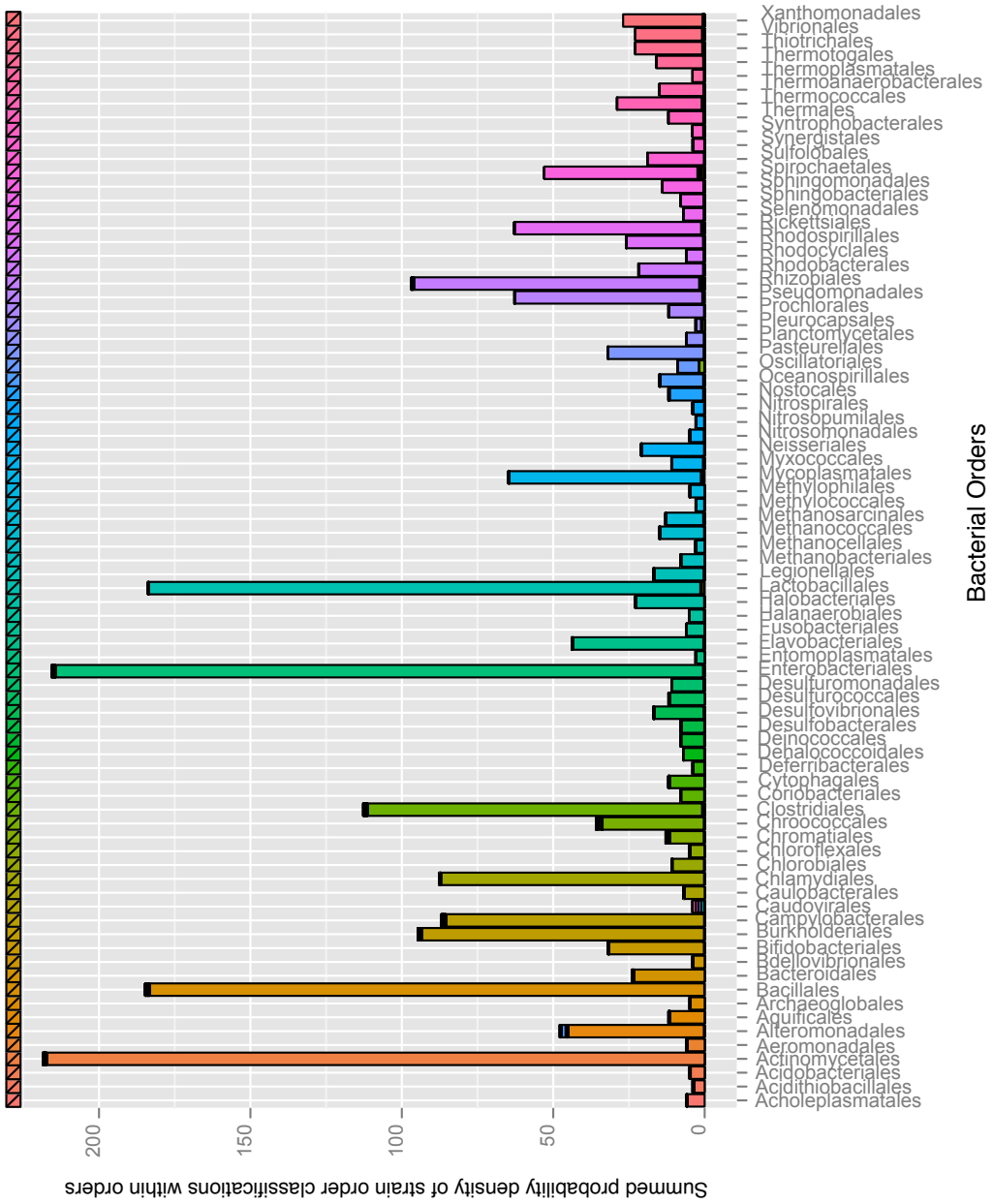


Figure 4.2: **Multiway classifier of all fully sequenced bacterial genomes in 79 orders with greater than three genomes.** A stacked bar graph in which the classification probabilities for all genomes within a given nominal bacterial order according to NCBI taxonomy have been sorted and summed by bacterial order according to the classifier model. The height of the bars are representative of the number of genomes in the dataset for each order, given that the probability density for one genome is one.

well against order Rhizobiales (probabilities 0.863 and 0.975 respectively). *Pseudovibrio* unexpectedly scores well against the Rhodobacterales (probability 0.97), and only SAR11 strain HIMB59 has a strong score against any order, and it classifies again with Rickettsiales (probability 0.94). All other SAR11 genomes score poorly against orders in and out of Alphaproteobacteria (probabilities range from 0.14-0.51). This could mean that along with a LOO-CV to compute the training dataset, a perturbation of the standard parameters for MLP training are necessary steps to fine-tuning this process for unknown, unclassified genomes.

Looking throughout the bacterial tree in all 79 orders curated in this chapter, Functional Information height distribution across sites is strikingly similar, across orders (see Figure 4.3. These site variations could be an artifact of the automated alignment, or they could be the transfer of a putative functional site in the tRNA. Given that Information is a biased estimator, the number of genomes in each order range from three to 200+, one might expect information values with wide ranges in sequences with no pattern to informative features. Looking more closely, some regions have smaller variation than others than others, like the anticodon loop across all four states. G information heights seem to be the most conserved, but larger heights are seen in cites containing nucleotides A and U.

4.3.4 Site-variation in total Information across orders

Small variation in the information heights across all four states in the anticodon loop essentially normalizes Figure 4.3, showing that in every system, the same amount of information exists in the datasets to form a functioning system. Other noticeable traits in this plot are when there are the small variability in the G+C plots vs the A+U plots. Possible evidence of some shifting in CIFs shows up in the T-stem in state U. One order has a high information peak next to a densely populated peak from other orders. Sites can clearly be identified where no information exists across orders, and looking more closely, it can be determined if there are sites in which there are large peaks in state graphs at a site where in other state graphs, there is no information.

This plot can be further dissected taxonomically to investigate class switching, site shifting, and gains and losses of CIFs across the bacterial tree of life. Employing a sliding window of information across taxa to see if windows have shifted could be employed to show CIF site shifting. Breaking out by class to see if any classes are consistently showing higher information values. If this is true, one could assess how many more conserved CIFs are in classes with more/less information.

4.3.5 the true value of CIFs

Since all heights are incorporated into the scoring scheme outlined in this dissertation, it would be wise to simulate an estimated amount of random Information that will potentially be generated with the type of data used in the CIF analyses. As you can see in

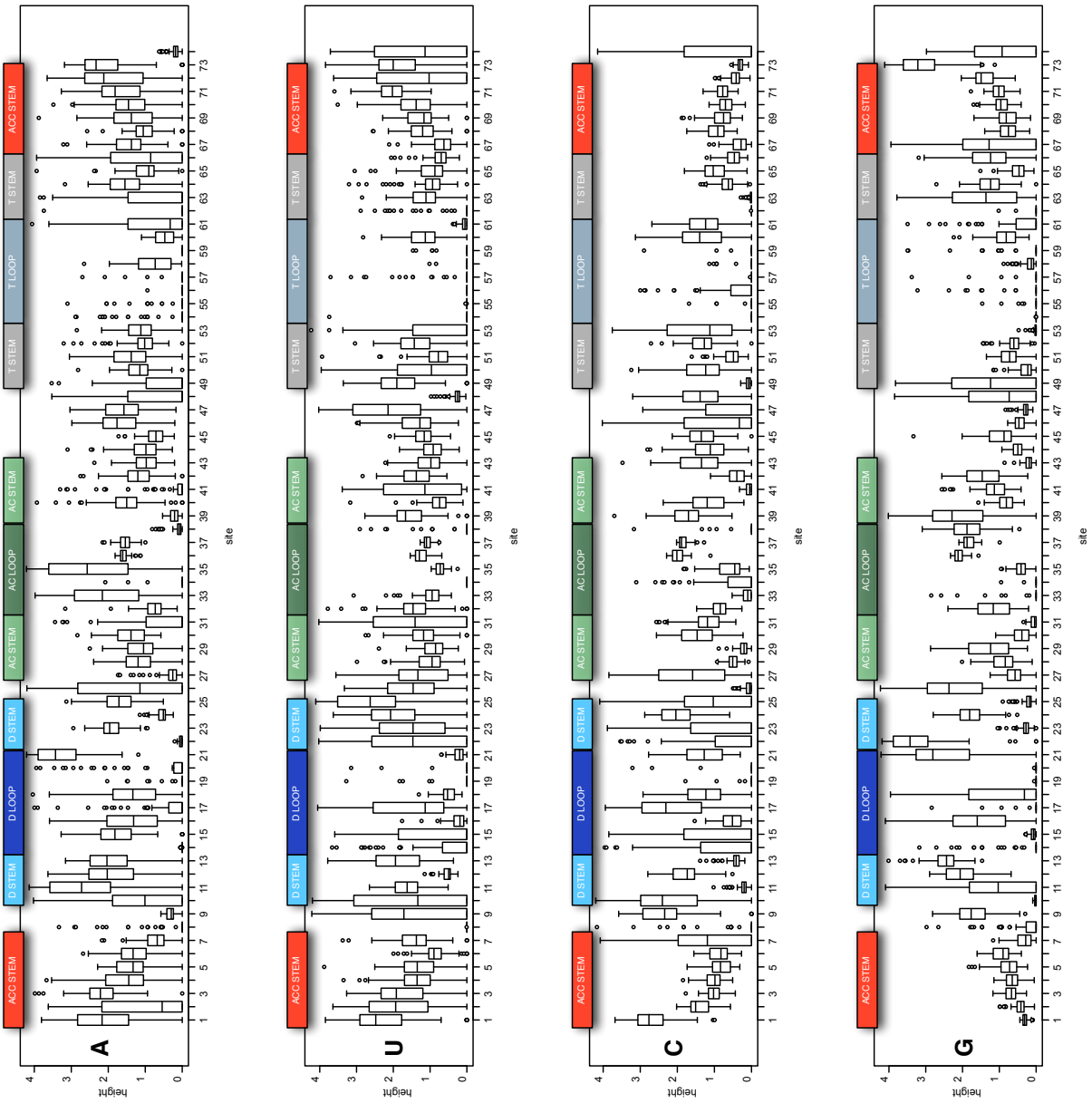


Figure 4.3: **Boxplot if Functional Information content over sites across 79 bacterial orders.** Range of boxplot at each site represents the range of information values recorded in the respective state Functional Information logo from all orders.

Fig. 4.4, randomizing class associations of the exact same subset of bacterial whole genome tRNA sequences that were used in Chapter 2 “RSR” grade (containing Rhodospirillales, Sphingomonadales, Rickettsiales) cause very little heights of Functional Information that are distinct from the background distribution expected for the data. It is clear that Class-Informative features truly do distinguish tRNAs from each other, most likely related to their function.

4.4 Materials and Methods

4.4.1 Data

Genomes from a total of 2515 bacterial strains were downloaded on June 2, 2013 from NCBI’s ftp site (Sayers et al., 2010). We custom annotated tRNA genes in these genomes as the union of predictions from tRNAscan-SE version 1.3.1 (with `-B` option, (Lowe and Eddy, 1997)) and Aragorn version 1.2.34 (Laslett and Canback, 2004). We classified initiator tRNAs and $\text{tRNA}^{\text{Ile}}_{\text{CAU}}$ using TFAM version 1.4 (Tåquist et al., 2007) using a model previously created to do this based on identifications in (Silva et al., 2006). We excluded any tRNAs that contained more than 350 bases, as done in (Freyhult et al., 2007). We aligned tRNAs with INFERNAL 1.1rc1 (Nawrocki et al., 2009) with the bacterial covariance model from RFAM (Burge et al., 2013), conservatively hand-picked the sites in Seaview 4.1 (Gouy et al., 2010) to include, cutting the alignment to 74 canonical sites (CCA tails excluded), and then separated them by order using a bioperl-based utility named *fastax* (Stajich et al., 2002). We mapped sites to Sprinzl coordinates manually (Sprinzl et al., 1998) and verified by spot-checks against tRNAdb-CE (Abe et al., 2009, 2011). We further removed sequences with more than ten gaps out of the 74 sites in an attempt to remove poorly aligned sequences. Most sequences removed were Aragorn-predicted and contained large introns. This was a conservative inclusion that is justified by the number of sequences that contain any number of gaps (See Figure 4.5). We have excluded gaps due to the fact that a gapped site may have functional importance in a tRNA, and tRNAs need most sites to function. Inclusion of gapped sequences can skew Information to present a state that mimics functional importance when in fact there is alignment error.

4.4.2 Order-Specific Data Curation

In order to assess Class Informative Features, we have initially separated data into taxonomic orders. Genomes exist for 118 bacterial orders.To make logos, we have narrowed our order sets down to orders that contain three or more genomes, leaving us with 81 orders. Two orders do not have annotated methionine tRNAs that fit our data curation pipeline, and are therefore removed (Thermoproteales–13 genomes and Methanomicrobiales–8 genomes). The final order-specific dataset contains 79 orders with 134544 tRNAs from 2374 genomes with every class (amino acid type) represented in

Figure 4.4: **Function logos from the randomization of class association in the alphaproteobacteria.** Separating tRNAs into 22 random associations, instead of their defined classes produces nearly empty Function Logos.

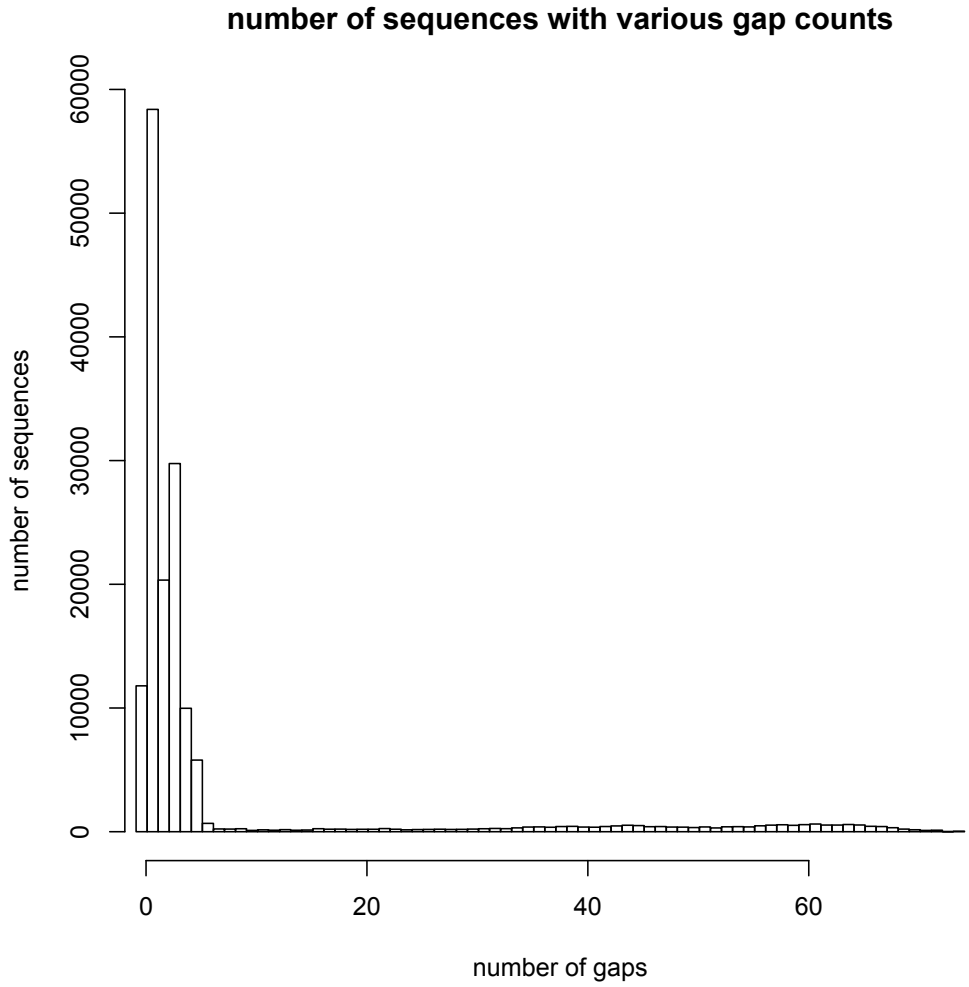


Figure 4.5: **Justification for excluding sequences with more than ten gaps.** Histogram of the number of sequences with a given amount of gaps. Each bin represents an increment of one gap. Most sequences contain five or less gaps. We chose ten to be conservative, which includes over 95% of sequences.

every order. We separated the genomes in each order by class with a possibility of fitting into 22 groups including the 20 canonical amino acids with $tRNA^{Met}_{CAU}$ and $tRNA^{Ile}_{CAU}$ as the two other possibilities. From here, we estimated Functional Information using Logofun (Freyhult et al., 2006) with exact entropy calculation of error up to 5 sequences and a forced alphabet of ACGU- (ignoring ambiguous codons).

4.4.3 Functional Information Base Composition

To calculate Functional Information base composition, Functional Information was extracted from Logofun output and summed over sites, normalized by the number of sequences in each order.

4.4.4 79-Model Classifier

Following the methods from Chapter 2, we used the sequence logos in the data curation to create models 79 groups 2374 genomes from 79 orders were classified against the models, scored as described in Chapter 2.

4.4.5 Class Randomization

Sequences from chapter two were combined into one alignment, and then randomly partitioned into 22 sets of tRNAs, equal to the size of the original classes using the Fisher-Yates shuffle implemented in Perl Fisher and Yates (1948). Function logos were then generated using logofun (Freyhult et al., 2006) with exact calculations for up to three sequences and gorodkin heights (Gorodkin et al., 1997).

4.4.6 Conclusion

We have shown that using tRNAs as a complete bacterial classifier trained as a Multilayer perceptron is promising with positive controls and excellent preliminary data. Fine-tuning of the parameters and variations of the MLP will be crucial to take the seven-model classifier from Chapter 2, to the 79-model classifier described in this chapter.

In our investigation of CIFs, we have seen cases of conserved CIFs across large portions of the tree of life, but they are subject to changes (see Chapter 2). Each tRNA class has a varying number of CIFs, some being more important than others, with a few even experimentally quantified.

We have learned that the amount of information seems to be conserved at certain sites in tRNAs, as shown by the site-specific analysis in this chapter. It is possible that as these sites undergo substitutions, that other sites in the tRNA with varying Information content across all species are compensating for the loss of specificity. The sliding window method described would be a great test for this possibility.

Overall, tRNA CIFs continue to prove to be interesting and unique phylogenetic markers that should be further investigated to fully understand their potential in phylogenetic tree-building and bacterial classification.

Chapter 5

Conclusion

5.1 Accomplishments

5.1.1 Methodology

In this dissertation, we have presented a method of biological sequence analysis uniquely and successfully provides bacterial classification using Information theory as our base. The use of tRNAs allows for fast, inexpensive computation. Information theory has proven to be accurate in prediciting tRNA identity elements (Freyhult et al., 2006) and, in Chapter 2. Identifying putative identity elements is now limited only by the amount of sequencing data available.

Using the same metric, we have deveolped methods to create phylogenetic trees that prove accurate in various parts of the tree of life including Plastids and Proteobacteria. Jensen-Shannon Divergence is a fast calculation that can be modified for any function logos contrasting different species to infer relatedness.

5.1.2 Scientific Impact

We have a much better understanding of the tRNA system trends that span the entire bacterial tree. It has been demonstrated that applying Conditional Information theory to tRNA class partitions, we can identify elements that are evolving slowly, and vary in importance to the organism based on the site in the tRNA where they reside.

We have successfully re-created topologies that have recently been proposed with more careful and biologically sound analyses (*P. ubique* in chapter 2, and *S. elongatus* in Chapter 3).

We have successfully identified tRNA features that potentially quantify the biological relevance of identity elements (*C73-H* in Chapter 2 and *A53:U61-E* in Chapter 3). We have also predicted putative features that can be verified/refuted experimentally in both Alphaproteobacteria, Cyanobacteria, and Plastids.

We have shown that scoring tRNA profiles against compiled scores made from strongly supported associations (or clade partitions), we can successfully classify bacteria

across the bacterial tree. Positive controls in the classification of viruses prove that tRNA profile scores truly do look only like their closely related species.

5.2 Next Steps

In Chapter 2, the conclusion would be stronger if order Caulobacterales contained more sequenced genomes, and if order Parviculales contained more than one sequenced genome (which could not be included in the study due to the small sample size).

The genomes of 54 new strains of Cyanobacteria were completed and published after the analysis in Chapter 3. Recalculation of the trees using Jensen-Shannon Divergence will provide more robust trees and add confidence to our conclusion. This will be done prior to publication of this work.

In Chapter 4, the Multilayer Perceptron provided interesting results with the poor classification of viruses (expected) and the misclassification of the SAR11 with poor probability scores. Going from seven classifications in Chapter 2 to 79 classifications in Chapter 4 calls for a more careful look at the implementation of Machine Learning. The perturbation of the priors in the training of the Chapter 4 Bacterial MLP will improve the quality and strength of the analysis.

A sliding window analysis of clade-specific identity profiles will allow for a better understanding of CIF movement in the tRNA, and more sequences will add depth to the classifier.

Aside from the modifications in the analysis of the CIFs, the actual calculation of the Information that makes up the score needs to be further developed to include base-pair dependent calculations to relay maximum biological relevance. Treating the sites as independently and identically distributed is counting "features" that act in one base-pair two times, when in fact biologically, they most likely act as one CIF. The Ardell lab has developed a base-pair functional information calculation that would be the very next step in fine-tuning this method.

Unfortunately, the main limitation to scientific discovery using CIF analysis lies in the lack of diverse sequencing across the bacterial tree of life. Entire genomes are not needed, but entire "tRNA-omes," or complete sets of tRNAs from any organism, are necessary for conditional information to be accurately calculated. This will be alleviated with time as more bacterial genomes are published, and potential short-term solutions could include the direct sequencing of all of the tRNAs of any organism.

Bibliography

Abby, S. S., Tannier, E., and Gouy, M. (2012). Lateral gene transfer as a support for the tree of life. *Proc Natl Acad Sci U S A*, 109(13):4962–4967.

Abe, T., Ikemura, T., Ohara, Y., Uehara, H., Kinouchi, M., Kanaya, S., Yamada, Y., Muto, A., and Inokuchi, H. (2009). tRNADB-CE: tRNA gene database curated manually by experts. *Nucleic Acids Research*, 37(Database issue):D163–8.

Abe, T., Ikemura, T., Sugahara, J., Kanai, A., Ohara, Y., Uehara, H., Kinouchi, M., Kanaya, S., Yamada, Y., Muto, A., and Inokuchi, H. (2011). tRNADB-CE 2011: tRNA gene database curated manually by experts. *Nucleic Acids Research*, 39(Database issue):D210–3.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall, New York.

Andam, C. P. and Gogarten, J. P. (2011). Biased gene transfer in microbial evolution. *Nat Rev Micro*, 9(7):543–555.

Andersson, S. G. and Kurland, C. G. (1998). Reductive evolution of resident genomes. *Trends in microbiology*, 6(7):263–268.

Anisimova, M., Liberles, D. A., Philippe, H., Provan, J., Pupko, T., and von Haeseler, A. (2013). State-of the art methodologies dictate new standards for phylogenetic analysis. *BMC evolutionary biology*, 13(1):161.

Archibald, J. M. (2009). The puzzle of plastid evolution. *Current biology : CB*, 19(2):R81–8.

Ardell, D. H. (2010). Computational analysis of tRNA identity. *FEBS Letters*, 584(2):325–333.

Ardell, D. H. and Andersson, S. G. E. (2006). TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res*, 34(3):893–904.

Baake, E. and Gabriel, W. (2000). Biological evolution through mutation, selection, and drift: An introductory review. *Ann Rev Comp Phys*, 7:203–264.

Bailly, M., Giannouli, S., Blaise, M., Stathopoulos, C., Kern, D., and Becker, H. D. (2006). A single tRNA base pair mediates bacterial tRNA-dependent biosynthesis of asparagine. *Nucleic Acids Res*, 34(21):6083–6094.

Baker, C. R., Tuch, B. B., and Johnson, A. D. (2011). Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc Natl Acad Sci U S A*, 108(18):7493–7498.

Bapteste, E., O'malley, M. A., Beiko, R. G., Ereshefsky, M., Gogarten, J. P., Franklin-Hall, L., Lapointe, F.-J., Dupré, J., Dagan, T., Boucher, Y., and Martin, W. (2009). Prokaryotic evolution and the tree of life are two different things. *Biol Direct*, 4(1):34.

Barbrook, A. C., Howe, C. J., Kurniawan, D. P., and Tarr, S. J. (2010). Organization and expression of organellar genomes. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1541):785–797.

Barrière, A., Gordon, K. L., and Ruvinsky, I. (2012). Coevolution within and between Regulatory Loci Can Preserve Promoter Function Despite Evolutionary Rate Acceleration. *PLoS Genet*, 8(9):e1002961.

Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., and Krogan, N. J. (2012). Systematic functional prioritization of protein posttranslational modifications. *Cell*, 150(2):413–425.

Bergsten, J. (2005). A review of longbranch attraction. *Cladistics*, 21(2):163–193.

Bhattacharya, D. and Weber, K. (1997). The actin gene of the glaucocystophyte Cyanophora paradoxa: analysis of the coding region and introns, and an actin phylogeny of eukaryotes. *Current genetics*, 31(5):439–446.

Bininda-Emonds, O. (2005). Supertree construction in the genomic age. *Methods in enzymology*, 5:745–757.

Böck, A., Forchhammer, K., Heider, J., and Baron, C. (1991). Selenoprotein synthesis: an expansion of the genetic code. *Trends in biochemical sciences*, 16(12):463–467.

Bodył, A., Mackiewicz, P., and Gagat, P. (2012). Organelle evolution: Paulinella breaks a paradigm. *Current biology : CB*, 22(9):R304–6.

Bosshard, P. P., Zbinden, R., Abels, S., Böddinghaus, B., Altwegg, M., and Böttger, E. C. (2006). 16S rRNA gene sequencing versus the API 20 NE system and the VITEK 2 ID-GNB card for identification of nonfermenting Gram-negative bacteria in the clinical laboratory. *Journal of clinical microbiology*, 44(4):1359–1366.

Brindefalk, B., Ettema, T. J. G., Viklund, J., Thollesson, M., and Andersson, S. G. E. (2011). A Phylometagenomic Exploration of Oceanic Alphaproteobacteria Reveals Mitochondrial Relatives Unrelated to the SAR11 Clade. *PLoS ONE*, 6(9):e24457.

Brindefalk, B., Viklund, J., Larsson, D., Thollesson, M., and Andersson, S. G. E. (2006). Origin and evolution of the mitochondrial aminoacyl-tRNA synthetases. *Mol Biol Evol*, 24(3):743–756.

Brochier, C. and Philippe, H. (2002). Phylogeny: A non-hyperthermophilic ancestor for Bacteria. *Nature*, 417(6886):244–244.

Brown, J. R. and Doolittle, W. F. (1999). Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. *J Mol Evol*, 49(4):485–495.

Bullaughey, K. (2012). Multidimensional adaptive evolution of a feed-forward network and the illusion of compensation. *Evolution*, 67(1):49–65.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A. (2013). Rfam 11.0: 10 years of RNA families. *Nucleic acids research*, 41(Database issue):D226–32.

Cavalier-Smith, T. (1981). Eukaryote kingdoms: seven or nine? *Bio Systems*, 14(3-4):461–481.

Chan, T. S. and Garen, A. (1970). Amino acid substitutions resulting from suppression of nonsense mutations. V. Tryptophan insertion by the Su9 gene, a suppressor of the UGA nonsense triplet. *Journal of molecular biology*, 49(1):231–234.

Chen, K., Eargle, J., Sarkar, K., Gruebele, M., and Luthey-Schulten, Z. (2010). Functional role of ribosomal signatures. *Biophysj*, 99(12):3930–3940.

Ciccarelli, F. D. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. *Science*, 311(5765):1283–1287.

Commans, S. and Böck, A. (1999). Selenocysteine inserting tRNAs: an overview. *FEMS microbiology reviews*, 23(3):335–351.

Crick, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–163.

Criscuolo, A. and Gribaldo, S. (2011). Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. *Molecular biology and evolution*, 28(11):3019–3032.

Dale, C., Wang, B., Moran, N., and Ochman, H. (2003). Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. *Mol Biol Evol*, 20(8):1188–1194.

Dayhoff, M. O. and Schwartz, R. M. (1978). Chapter 22: A model of evolutionary change in proteins. In *in Atlas of Protein Sequence and Structure*.

de Queiroz, A. and Gatesy, J. (2007). The supermatrix approach to systematics. *Trends in ecology & evolution*, 22(1):34–41.

Deschamps, P. and Moreira, D. (2009). Signal Conflicts in the Phylogeny of the Primary Photosynthetic Eukaryotes. *Molecular biology and evolution*, 26(12):2745–2753.

Desper, R. and Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of computational biology : a journal of computational molecular cell biology*, 9(5):687–705.

Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K. V., Allen, J. F., Martin, W., and Dagan, T. (2008). Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Molecular biology and evolution*, 25(4):748–761.

Dohm, J. C., Vingron, M., and Staub, E. (2006). Horizontal gene transfer in aminoacyl-tRNA synthetases including leucine-specific subtypes. *Journal of molecular evolution*, 63(4):437–447.

Doolittle, R. F. and Handy, J. (1998). Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Current opinion in genetics & development*, 8(6):630–636.

Doolittle, W. F. and Bapteste, E. (2007). Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci USA*, 104(7):2043–2049.

Dufresne, A., Garczarek, L., and Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol*, 6(2):R14–R14.

Duncan, K. E., Istock, C. A., Graham, J. B., and Ferguson, N. (1989). JSTOR: Evolution, Vol. 43, No. 8 (Dec., 1989), pp. 1585-1609. *Evolution*.

Eddy, S. R. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088.

Ekborg, N. A., González, J. M., Howard, M. B., Taylor, L. E., Hutcheson, S. W., and Weiner, R. M. (2005). Saccharophagus degradans gen. nov., sp. nov., a versatile marine degrader of complex polysaccharides. *International journal of systematic and evolutionary microbiology*, 55(Pt 4):1545–1549.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860.

Falcón, L. I., Magallón, S., and Castillo, A. (2011). Dating the cyanobacterial ancestor of the chloroplast. *The ISME journal*, 5(2):366.

Fathinejad, S., Steiner, J. M., Reipert, S., Marchetti, M., Allmaier, G., Burey, S. C., Ohnishi, N., Fukuzawa, H., Löffelhardt, W., and Bohnert, H. J. (2008). A carboxysomal carbon-concentrating mechanism in the cyanelles of the 'coelacanth' of the algal world, Cyanophora paradoxa? *Physiologia plantarum*, 133(1):27–32.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Felsenstein, J. (2005a). Felsenstein: PHYLIP (phylogeny inference package)... - Google Scholar. *Distributed by the author*.

Felsenstein, J. (2005b). *PHYLIP (Phylogeny Inference Package) version 3.6*. Department of Genome Sciences, University of Washington, Seattle.

Fisher, R. A. and Yates, F. (1948). *Statistical tables for biological, agricultural and medical research (3rd ed.)*. Oliver & Boyd, London.

Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic Biology*, 53(3):485–495.

Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science (New York, N.Y.)*, 296(5568):750–752.

Freyhult, E., Cui, Y., Nilsson, O., and Ardell, D. H. (2007). New computational methods reveal tRNA identity element divergence between Proteobacteria and Cyanobacteria. *Biochimie*, 89(10):1276–1288.

Freyhult, E., Moulton, V., and Ardell, D. H. (2006). Visualizing bacterial tRNA identity determinants and antideterminants using function logos and inverse function logos. *Nucleic Acids Research*, 34(3):905–916.

Galtier, N. and Lobry, J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of molecular evolution*, 44(6):632–636.

Georgiades, K., Madoui, M.-A., Le, P., Robert, C., and Raoult, D. (2011). Phylogenomic Analysis of Odyssella thessalonicensis Fortifies the Common Origin of Rickettsiales, Pelagibacter ubique and Reclimonas americana Mitochondrion. *PLoS ONE*, 6(9):e24857.

Giege, R. (2008). Toward a more complete view of tRNA biology. *Nat Struct Mol Biol*, 15(10):1007–1014.

Giegé, R. (2008). Toward a more complete view of tRNA biology. *Nature Structural & Molecular Biology*, 15(10):1007–1014.

Giegé, R., Puglisi, J. D., and Florentz, C. (1993). tRNA structure and aminoacylation efficiency. *Progress in nucleic acid research and molecular biology*, 45:129–206.

Giegé, R., Sissler, M., and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res*, 26(22):5017.

Giovannoni, S. J. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, 309(5738):1242–1245.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*, 19(12):2226–2238.

Gorodkin, J., Heyer, L. J., Brunak, S., and Stormo, G. D. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *Computer Applications In the Biosciences : CABIOS*, 13(6):583–586.

Gouy, M., Guindon, S., and Gascuel, O. (2010). Seaview version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27(2):221–224.

Gribaldo, S. and Philippe, H. (2002). Ancient phylogenetic relationships. *Theor Popul Biol*, 61(4):391–408.

Grote, J., Thrash, J. C., Huggett, M. J., Landry, Z. C., Carini, P., Giovannoni, S. J., and Rappé, M. S. (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio*, 3(5).

Gupta, R. (2010). Applications of Conserved Indels for Understanding Microbial Phylogeny. In Aharon Oren, R. T. P., editor, *Molecular Phylogeny of Microorganisms*. Horizon Scientific Press, Norfolk,UK.

Gupta, R. S. and Mok, A. (2007). Phylogenomics and signature proteins for the alpha proteobacteria and its main groups. *BMC Microbiol*, 7:106.

Haag, E. S. and Molla, M. N. (2005). Compensatory evolution of interacting gene products through multifunctional intermediates. *Evolution*, 59(8):1620–1632.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Hamady, M., Lozupone, C., and Knight, R. (2010). Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *the ISME Journal*, 4(1):17–27.

Hartl, D. L. and Taubes, C. H. (1996). Compensatory nearly neutral mutations: selection without adaptation. *Journal of Theoretical Biology*, 182(3):303–309.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2):160–174.

He, B. Z., Holloway, A. K., Maerkl, S. J., and Kreitman, M. (2011). Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. *PLoS Genet*, 7(4):e1002053.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

Hershberg, R. and Petrov, D. A. (2010). Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet*, 6(9).

Higgins, D. G. and Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, 73(1):237–244.

Itoh, T., Martin, W., and Nei, M. (2002). Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc Natl Acad Sci USA*, 99(20):12944–12948.

Jain, R., Rivera, M. C., and Lake, J. A. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*, 96(7):3801–3806.

Janda, J. M. and Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, 45(9):2761–2764.

Jow, H., Gowri-Shankar, V., and Guillard, B. (2003). PHASE: a software package for phylogenetics and sequence evolution. *University of Manchester*.

Jühling, F., Mörl, M., Hartmann, R., Sprinzl, M., Stadler, P., and Pütz, J. (2009). trnadb 2009: compilation of trna sequences and trna genes. *Nucleic acids research*, 37(suppl 1):D159–D162.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In Munro, M. N., editor, *Mammalian protein metabolism*, pages 21–132. Academic Press, N. Y.

Keeling, P. J., Archibald, J. M., Fast, N. M., and Palmer, J. D. (2004). Comment on "The evolution of modern eukaryotic phytoplankton". *Science (New York, N.Y.)*, 306(5705):2191–author reply 2191.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, 16(2):111–120.

Koonin, E. V. and Wolf, Y. I. (2009). The fundamental units, processes and patterns of evolution, and the tree of life conundrum. *Biology direct*, 4:33.

Kuo, D., Licon, K., Bandyopadhyay, S., Chuang, R., Luo, C., Catalana, J., Ravasi, T., Tan, K., and Ideker, T. (2010). Coevolution within a transcriptional network by compensatory trans and cis mutations. *Genome Res*, 20(12):1672–1678.

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A*, 82(20):6955–6959.

Laslett, D. and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research*, 32(1):11–16.

Levicán, G., Katz, A., de Armas, M., Núñez, H., and Orellana, O. (2007). Regulation of a glutamyl-tRNA synthetase by the heme status. *Proc Natl Acad Sci U S A*, 104(9):3135–3140.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Lin, T.-W., Wu, J.-W., and Chang, D. T.-H. (2013). Combining phylogenetic profiling-based and machine learning-based techniques to predict functional related proteins. *PloS one*, 8(9):e75940.

Lind, P. A. and Andersson, D. I. (2008). Whole-genome mutational biases in bacteria. *Proc Natl Acad Sci U S A*, 105(46):17878–17883.

Losos, J. B., Hillis, D. M., and Greene, H. W. (2012). Who speaks with a forked tongue? *Science*, 338(6113):1428–1429.

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, 25(5):955–964.

Majewski, J. and Cohan, F. M. (1999). Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics*, 152(4):1459–1474.

Martin, W. and Herrmann, R. G. (1998). Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol*, 118:9–17.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci U S A*, 99(19):12246–12251.

Martin, W., Somerville, C., and Loiseaux-de Goer, S. (1992). Molecular Phylogenies of Plastid Origins and Algal Evolution. *J Mol Evol*, 35:385–404.

McClain, W. H. (1993). Rules that govern tRNA identity in protein synthesis. *Journal of molecular biology*, 234(2):257–280.

McInerney, J. O., Pisani, D., Bapteste, E., and O'Connell, M. J. (2011). The Public Goods Hypothesis for the evolution of life on Earth. *Biology direct*, 6:41.

Mereschkowsky, C. (1905). About the nature and origin of chromoatophores in the vegetable kingdom. *Biol Centralbl (in German)*, 25(18):593–604.

Mignard, S. and Flandrois, J. P. (2006). 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of microbiological methods*, 67(3):574–581.

Moran, N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, 108(5):583–586.

Morris, R. M., Rappé, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A., and Giovannoni, S. J. (2002). SAR 11 clade dominates ocean surface bacterioplankton communities. *Nature*, 420(6917):806–810.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics (Oxford, England)*, 25(10):1335–1337.

Nelson, K E and Clayton, R A and Gill, S R and Gwinn, M L and Dodson, R J and Haft, D H and Hickey, E K and Peterson, J D et al. (1999). Evidence for lateral gene transfer between archaea and bacteria from genome sequence of thermotoga maritima. *Nature*, 399(6734):323–329.

O'Meara, B. C. (2012). Evolutionary inferences from phylogenies: A review of methods. *Annual Review of Ecology, Evolution, and Systematics*, 43(1):267–285.

Österreicher, F. and Vajda, I. (2003). A new class of metric divergences on probability spaces and its applicability in statistics. *Annals of the Institute of Statistical Mathematics*, 55(3):639–653.

Patel, J. B. (2001). 16S rRNA gene sequencing for bacterial pathogen identification in the clinical laboratory. *Molecular diagnosis : a journal devoted to the understanding of human disease through the clinical application of molecular biology*, 6(4):313–321.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC evolutionary biology*, 5:50.

Price, D. C., Chan, C. X., Yoon, H. S., Yang, E. C., and Qiu, e. a. (2012). Cyanophora paradoxa genome elucidates origin of photosynthesis in algae and plants. *Science (New York, N.Y.)*, 335(6070):843–847.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One*.

Qiu, H., Yang, E. C., Bhattacharya, D., and Yoon, H. S. (2012). Ancient gene paralogy may mislead inference of plastid phylogeny. *Molecular biology and evolution*, 29(11):3333–3343.

Randau, L., Schauer, S., Ambrogelly, A., Salazar, J. C., Moser, J., Sekine, S.-i., Yokoyama, S., Söll, D., and Jahn, D. (2004). tRNA recognition by glutamyl-tRNA reductase. *The Journal of biological chemistry*, 279(33):34931–34937.

Rivera, M. C. and Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431(7005):152–155.

Roberts, E., Sethi, A., Montoya, J., Woese, C. R., and Luthey-Schulten, Z. (2008). Molecular signatures of ribosomal evolution. *Proc Natl Acad Sci U S A*, 105(37):13953–13958.

Robertson, B. R., Tezuka, N., and Watanabe, M. M. (2001). Phylogenetic analyses of Synechococcus strains (cyanobacteria) using sequences of 16S rDNA and part of the phycocyanin operon reveal multiple evolutionary lines and reflect phycobilin content. *International journal of systematic and evolutionary microbiology*, 51(Pt 3):861–871.

Rodríguez-Ezpeleta, N. and Embley, T. M. (2012). The SAR11 Group of Alpha-Proteobacteria Is Not Related to the Origin of Mitochondria. *PLoS ONE*, 7(1):e30520.

Roettger, M., Martin, W., and Dagan, T. (2009). A machine-learning approach reveals that alignment properties alone can accurately predict inference of lateral gene transfer from discordant phylogenies. *Molecular biology and evolution*, 26(9):1931–1939.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (2002). Learning representations by back-propagating errors. In Polk, T. A. and Seifert, C. M., editors, *Cognitive modeling*, pages 213–220. MIT press.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., and Canese, K. e. a. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 38(Database issue):D5–16.

Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. (1994). From sequences to shapes and back: a case study in RNA secondary structures. *Proc. Biol. Sci.*, 255(1344):279–284.

Sethi, A., Eargle, J., Black, A. A., and Luthey-Schulten, Z. (2009). Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci U S A*, 106(16):6620–6625.

Shiba, K. and Motegi, H. (1997). Maintaining genetic code through adaptations of tRNA synthetases to taxonomic domains. *Trends in biochemical sciences*, 22:453–457.

Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., and Talla, e. a. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A*, 110(3):1053–1058.

Silva, F. J., Belda, E., and Talens, S. E. (2006). Differential annotation of tRNA genes with anticodon CAT in bacterial genomes. *Nucleic Acids Research*, 34(20):6015–6022.

Slabbinck, B., Waegeman, W., Dawyndt, P., De Vos, P., and De Baets, B. (2010). From learning taxonomies to phylogenetic learning: integration of 16S rRNA gene data into FAME-based bacterial classification. *BMC bioinformatics*, 11:69.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Research*, 26(1):148–153.

Srinivasan, G., James, C. M., and Krzycki, J. A. (2002). Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science (New York, N.Y.)*, 296(5572):1459–1462.

Stackebrandt, E. and Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic Bacteriology*, 44(4):846–849.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., and Dagdigian, e. a. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–1618.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A rapid bootstrap algorithm for the raxml web servers. *Systematic Biology*, 57(5):758–771.

Stamatakis, A. P., Meier, H., and Ludwig, T. (2004). New fast and accurate heuristics for inference of large phylogenetic trees. In *18th International Parallel and Distributed Processing Symposium, 2004.*, pages 193–200. IEEE.

Stange-Thomann, N., Thomann, H. U., Lloyd, A. J., Lyman, H., and Söll, D. (1994). A point mutation in Euglena gracilis chloroplast tRNA(Glu) uncouples protein and chlorophyll biosynthesis. *Proc Natl Acad Sci U S A*, 91(17):7947–7951.

Tåquist, H., Cui, Y., and Ardell, D. H. (2007). TFAM 1.0: an online tRNA function classifier. *Nucleic Acids Research*, 35(Web Server issue):W350–3.

Theobald, D. L. (2010). A formal test of the theory of universal common ancestry. *Nature*, 465(7295):219–222.

Theodoridis, S. and Koutroumbas, K. (1999). *Pattern recognition*. Academic Press, Waltham, Massachusetts.

Thrash, J. C., Boyd, A., Huggett, M. J., Grote, J., Carini, P., Yoder, R. J., Robbertse, B., Spatafora, J. W., Rappé, M. S., and Giovannoni, S. J. (2011). Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. *Sci. Rep.*, 1.

Tomitani, A., Knoll, A. H., Cavanaugh, C. M., and Ohno, T. (2006). The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc Natl Acad Sci U S A*, 103(14):5442–5447.

Viklund, J., Ettema, T. J. G., and Andersson, S. G. E. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol*, 29(2):599–615.

Vulić, M., Dionisio, F., Taddei, F., and Radman, M. (1997). Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci U S A*, 94(18):9763–9767.

Wang, C., Sobral, B. W., and Williams, K. P. (2007a). Loss of a Universal tRNA Feature. *J Bacteriol*, 189(5):1954–1962.

Wang, C., Sobral, B. W., and Williams, K. P. (2007b). Loss of a universal tRNA feature. *Journal of Bacteriology*, 189(5):1954–1962.

Widmann, J., Harris, J. K., Lozupone, C., Wolfson, A., and Knight, R. (2010). Stable tRNA-based phylogenies using only 76 nucleotides. *RNA*, 16(8):1469–1477.

Williams, K., Sobral, B., and Dickerman, A. (2007). A Robust Species Tree for the Alphaproteobacteria . *J Bacteriol*, 189(13):4578.

Williamson, J. R. (2000). Induced fit in RNA—[ndash]—protein recognition. *Nature Structural Biology*, 7(10):834–837.

Winker, S. and Woese, C. R. (1991). A definition of the domains Archaea, Bacteria and Eucarya in terms of small subunit ribosomal RNA characteristics. *Systematic and Applied Microbiology*, 14(4):305–310.

Woese, C. R. (1987). Bacterial evolution. *Microbiological reviews*, 51(2):221–271.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci USA*, 97(15):8392–8396.

Woese, C. R. (2002). On the evolution of cells. *Proc Natl Acad Sci U S A*, 99(13):8742–8747.

Woese, C. R., Olsen, G. J., Ibba, M., and Söll, D. (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev*, 64(1):202–236.

Wolf, Y. I., Aravind, L., Grishin, N. V., and Koonin, E. V. (1999). Evolution of aminoacyl-tRNA synthetases–analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res*, 9(8):689–710.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. (2002). Genome trees and the tree of life. *TRENDS in Genetics*, 18(9):472–479.

Wolfson, A., LaRiviere, F., Pleiss, J., Dale, T., Asahara, H., and Uhlenbeck, O. (2001). trna conformity. *Cold Spring Harbor Symposia on Quantitative Biology*, 66:185–194.

Wu, H., Zhang, Z., Hu, S., and Yu, J. (2012). On the molecular mechanism of GC content variation among eubacterial genomes. *Biology direct*, 7:2.

Yang, J. C., Madupu, R., Durkin, A. S., Ekborg, N. A., Pedamallu, C. S., and Hostetler, e. a. (2009). The complete genome of Teredinibacter turnerae T7901: an intracellular endosymbiont of marine wood-boring bivalves (shipworms). *PloS one*, 4(7):e6085.

Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Glöckner, F. O., and Rosselló-Móra, R. (2010). Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Systematic and applied microbiology*, 33(6):291–299.