UNIVERSITY OF CALIFORNIA, MERCED

Uncovering Deep Phylogenetic Signal in Plastid Genomes

A dissertation submitted in partial fulfillment of the requirements for
the degree Doctor of Philosophy
in
Quantitative and Systems Biology
by
Travis Joseph Lawrence

Committee in Charge:
Professor Carolin Frank
Professor David Ardell
Professor Clarissa Nobile
Professor Emily Jane Mctavish
Professor Suzanne Sindi

2018

The dissertation of Travis Joseph Lawrence is approved:

Faculty Advisor: ______________________________

David H. Ardell, Ph.D.

Committee Members: ______________________________

Chair: Carolin Frank, Ph.D.

____

Clarissa Nobile, Ph.D.

____

Suzanne Sindi, Ph.D.

____

Emily Jane Mctavish, Ph.D.

____

Date

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my gratitude to my advisor, David Ardell, for his guidance, encouragement, and dedication throughout these last five years. I will miss our discussions on science and research over tacos and beers at J&R Tacos. I would like to thank the members of my PhD committee: Carolin Frank, Clarissa Nobile, Emily Jane McTavish, and Suzanne Sindi. I would also like to thank Zhong Wang for my experience at the Joint Genome Institute, providing research advice, and for providing insight into career opportunities at National Laboratories. Finally, I want to thank my wife, Dana Carper, for her love, encouragement, and being a sounding board for my ideas. Her support and advice made this work possible.

Curriculum Vita

EDUCATION

Doctor of Philosophy in Quantitative and Systems Biology, 2013-2018
University of California, Merced

Master of Science, Biology, 2010-2013
California State University, Sacramento

Baccalaureate of Science, Biology 2006-2010
California State University, Sacramento

PUBLICATIONS

Kristen M. Valentine, Dan Davini, **Travis J. Lawrence**, Genevieve Mullins, Miguel Manansala, Mufadhal Al-Kuhlani, James M. Pinney, Jason K. Davis, Anna E. Beaudin, Suzanne S. Sindi, David M. Gravano, Katrina K. Hoyer (2018) "CD8 follicular T cells promote B cell antibody class-switch in autoimmune disease." *Journal of Immunology*, 198(1) doi: 10.4049/jimmunol.1701079

Giosa D., Felice M.R., **Lawrence T.J.**, Gulati M., Scordino, F., Giuffrè L., Passo C.L., D'Alessandro E., Criseo G., Ardell D.H., Hernday A.D., Nobile C.J., and Romeo, O. (2017). "Whole RNA-sequencing and transcriptome assembly of *Candida albicans* and *Candida africana* under chlamydospore-inducing conditions" *Genome Biology and Evolution*, 9(7):1971-1977 doi: 10.1093/gbe/evx143

Lawrence T.J. and Datwyler S.L. (2016). "Testing the Hypothesis of Allopolyploidy in the Origin of *Penstemon azureus* (Plantaginaceae)" *Frontiers in Ecology and Evolution*, 4. doi: 10.3389/fevo.2016.00060

Lawrence, T.J., Kauffman, K.T., Amrine, K.C.H., Carper, D.L., Lee, R.S., Becich, P.J., Canales, C.J., and Ardell, D.H. (2015). "FAST: FAST Analysis of Sequences Toolbox" *Frontiers in Genetics*, 6. doi: 10.3389/fgene.2015.00172

PRESENTATIONS

Conference Talks:
Lawrence, T.J. ‘tRNA Interaction Network Sheds Light on the Origin of Chloroplast” Northern California Computational Biology Student Symposium 2017

Lawrence, T.J., Ardell, D.H. “tRNA Interaction Network Sheds Light on the Origin of Chloroplast” Botany 2017

Lawrence, T.J., Datwyler, S. “Using low copy nuclear genes to test the allopolyploid origin of *Penstemon azureus*” Botany 2012

Conference Posters:
Lawrence, T.J., Ardell, D.H. “Using tRNA class informative features to determine the phylogenetic placement of Gnetophyta” Botany 2016

Lawrence, T.J., Amrine, K., Swingley, W., Ardell, D.H. “Exploring the origin of chloroplast using the tRNA-protein interaction network” Evolution 2016

Lawrence, T.J., Datwyler, S. “The allopolyploid origin of *Penstemon azureus*” Northern California Botany 2013

Lawrence, T.J., Datwyler, S. “Using low copy nuclear genes to determine the allopolyploid origin of *Penstemon azureus*” CSUPERB 2012

Abstract of the Dissertation

Uncovering Deep Phylogenetic Signal in Plastid Genomes

by
Travis Joseph Lawrence
Doctor of Philosophy, Quantitative and Systems Biology
University of California, Merced, 2018
Advisor: Prof. David H. Ardell

The overall aim of my dissertation is to show that a novel source of phylogenetic information from the plastid genome, the tRNA interaction network, coupled with machine-learning and distance-based methods, is capable of accurately reconstructing deep phylogenetic relationships. First, we review the history of the plastid genome as a source of phylogenetic information, discuss sources of systematic biases of plastid sequence data, and introduce the transfer RNA (tRNA) interaction network as a source of phylogenetic data.

Second, I determine the phylogenetic origin of plastids within the Cyanobacteria tree of life (CyanoToL). Previous studies have strongly supported contradictory conclusions, with plastids branching either early or late within the CyanoToL. I begin by predicting structural features that determine the charging potential of a tRNA with its cognate amino acid, termed tRNA Class Informative Features (CIFs) for 113 Cyanobacterial genomes within eight Cyanobacterial clades. I show that predicted tRNA CIFs differ between Cyanobacterial clades in a phylogenetically informative way that can be exploited to accurately classify Cyanobacterial genomes using a machine-learning algorithm known as a multilayer perceptron (MLP), which we have named CYANO-MLP. I then use CYANO-MLP to test competing hypotheses of the origin of plastids by classifying 440 plastids genomes. I found support for the origin of plastids among a late-branching clade of starch-producing marine/freshwater diazotrophic cyanobacteria. Finally, I show that previously used phylogenetic models are unable to accommodate systematic biases possibly explaining conflicting hypotheses.

Third, I use tRNA CIFs to determine the phylogenetic placement of gnetophytes, a small clade of plants, within the seed plant phylogeny. The location of gnetophytes has been contentious with phylogenomic studies supporting several relationships with cone-bearing seed plants (conifers). Here I use the Jensen-Shannon divergence to calculate a pairwise distance matrix between seed plant clades for plastid tRNA CIFs. Using standard distance-based phylogenetic algorithms I found support for gnetophytes as sister to conifers.

Lastly, I describe the implementation of two software packages. The first is tsfm: tRNA structure function mapper, that provides methods for predicting tRNA CIFs. The second is a suite of tools modeled after GNU Textutils named, FAST: FAST Analysis of Sequences Toolbox, for processing of molecular sequence data on the command line.

Chapter 1

Introduction

1.1 Plastid Phylogenetics and Phylogenomics

1.1.1 Development of Plastid Genes and Genomes as a Phylogenetic Marker

Shortly after the discovery of the plastid genome (Ris and Plaut, 1962) it became widely adopted as a valuable source of phylogenetic information for reconstructing evolutionary relationships of plastid-bearing eukaryotes (Vedel et al., 1978; Rhodes et al., 1981; Palmer et al., 1988) because of the conservation of gene content and relatively low rate of nucleotide substitution in non-parasitic lineages (Palmer et al., 1988; Davis et al., 2014). The plastid genome contains between 120-235 genes on average, however, plastids within parasitic host lineages can have highly reduced gene sets (Bellot and Renner, 2016; Figueroa-Martinez et al., 2017), most are a single circular molecule (Oldenburg and Bendich, 2015; Oldenburg and Bendich, 2016) and have a quadripartite organization with two inverted repeats separated by short and long single copy regions (Davis et al., 2014; Muñoz-Gómez et al., 2017). Plastid genome sizes range broadly from 11.4kb in the holoparasite *Pilostyles aethiopica*, caused by the loss of photosynthetic genes and reduction of the number and length of introns and intergenic regions (Bellot and Renner, 2016; Figueroa-Martinez et al., 2017), to 1.1Mb in the red algae *Corynoplastis japonica* mostly explained by expansion of introns and intergenic regions (Muñoz-Gómez et al., 2017). Early phylogenetic studies generated data from plastid genomes using restriction fragment length polymorphism (RFLP) in which isolated plastid DNA was digested by several restriction enzymes and the presence/absents of different DNA fragments were recorded after digestion and separation by gel electrophoresis (Perl-Treves and Galun, 1985; Shinozaki et al., 1986; Vedel et al., 1978; Jansen and Palmer, 1988; Rhodes et al., 1981; Palmer et al., 1988). These early studies were able to include large numbers of species and score markers with high reproducibility (Powell et al., 1996; Russell et al., 1997), however, they suffered from limited ability to resolve phylogenetic relationships above the family level, likely caused by homoplasy and limited number of phylogenetic markers (Palmer et al., 1988). Despite the first plastid genome being published in 1986 (Shinozaki et al., 1986) RFLPs remained the primary source of phylogenetic information from the plastid genome through the early 1990's because of the prohibitive cost of DNA sequencing. However, exponential decreases in sequencing cost starting in the mid 1990's (Check, 2014) produced a shift from RFLPs to sequencing several plastid genes with wide species sampling providing additional phylogenetic markers increasing the utility of the plastid genome to resolve evolutionary relationships (Soltis et al., 2002; Datwyler and Wolfe, 2004). The cost of sequencing precipitously dropped starting in 2006 with the advent of second-generation sequencing technology (Check, 2014) allowing investigators to fully exploit the phylogenetic information within the plastid genome sequence data by making it feasible to gather complete plastid genome sequences from hundreds of species (Davis et al., 2014). Despite the ability to easily generate plastid phylogenomic datasets several plastid evolutionary relationships remain unresolved or different plastid phylogenomic datasets have strongly supported contradictory hypotheses.

1.1.2 Complications of Plastid Genes and Genomes as a Phylogenetic Marker

The plastid genome has been and remains an invaluable source of phylogenetic information for the reconstruction of evolutionary relationships, successfully resolving countless relationships among plastid-bearing eukaryotes (Ruhfel et al., 2014; Muñoz-Gómez et al., 2017). However, the use of plastid genes and genomes in phylogenetic analyses are complicated by systematic biases caused by non-stationary compositions of genomes (Li et al., 2014), site-specific constraints of nucelotides/amino acids (Ochoa de Alda et al., 2014), and obfuscation of phylogenetic signal caused by substitutional saturation (Ruhfel et al., 2014). These complications are most significant when plastid genomes are used to reconstruct deep evolutionary relationships.

If systematic biases are not accounted for in phylogenetic models, they may affect the accuracy of phylogenetic reconstructions possibly leading to artefactual groupings in phylogenies. This can lead to persistent controversy where contradictory hypotheses with strong statistical support are recovered, which are entirely dependent on the dataset (Blanquart and Lartillot, 2008). Yet, commonly used phylogenetic models employed for deep plastid phylogenetics are time- and site-homogeneous models, which fail to accommodate non-stationary compositions and/or site-specific constraints (Blanquart and Lartillot, 2008; Jackson et al., 2018). Futhermore, it has been shown site-homogeneous models are sensitive to substitutional saturation leading to long-branch attraction artefacts (Halpern and Bruno, 1998). Although phylogenetic models exist that are capable of accommodating compositional biases (Li et al., 2014) and/or site-specific constraints (Lartillot et al., 2007; Blanquart and Lartillot, 2008) unfortunately they are computationally intractable for large phylogenomic datasets (Wang et al., 2018). The inability to adequately model systematic biases for large plastid phylogenomic datasets in a computational tractable way suggests that new phylogenetic markers that are resistant to these systematic biases are needed to accurately reconstruct difficult phylogenetic relationships using plastid genomes. Additionally, computational efficient methods to exploit the phylogenetic information in these markers will be required.

1.2 A New Source of Phylogenetic Signal from Plastid Genomes

1.2.1 tRNA Interaction Network

Transfer RNAs (tRNA) are short non-coding RNAs mainly involved in protein synthesis acting as adaptor molecules converting the information contained in the genome into proteins (Marck and Grosjean, 2002). To function as adaptor molecules during protein synthesis tRNAs must participate in two separate but equally important reactions. First, tRNAs must be recognized and aminoacylated by an aminoacyl tRNA-synthetase (aaRS) based on its amino acid identity (Goodman and Rich, 1962). The second reaction involves the recognition of an mRNA codon by the tRNA anticodon at the ribosome followed by the donation of a charged amino acid (Ogle et al., 2001).

tRNAs involved in protein synthesis must conform to the same general tRNA structure for efficient and comparable activity on general translation factors and the ribosome. This structure is routinely referred to as a clover leaf shape, despite the fact that the tertiary structure resembles an inverted capital “L”, consisting of three stem-loops, a variable loop, a base-paired stem, and an unpaired tail (Fig. 1.1). The acceptor stem contains the $5'$ and $3'$ ends (Fig. 1.1 purple) and the CCA tail (Fig. 1.1 yellow) which is typically added post-transcriptionally in Eukaryotes and in several Bacterial and Archaeal groups (Deutscher and Ni, 1982; Marck and Grosjean, 2002; Xiong and Steitz, 2004) with the site of aminoacylation occurring at the $2'$ or $3'$ hydroxyl of the terminal adenosine (Delagoutte et al., 2000; Xiong and Steitz, 2004). The anticodon loop (Fig. 1.1 blue) contains the anticodon triplet which interfaces with the codons of the mRNA through the ribosome (Fig. 1.1 gray) and is the conical location for introns (Marck and Grosjean, 2002). The D and TψC stem loops (Fig. 1.1 red and green respectively) are involved in interactions to form and stabilize the tertiary structure of the tRNA molecule (Butcher and Pyle, 2011). Lastly, the variable loop is typically between 3-21 nucleotides (Fig. 1.1 orange) and is mostly found in tRNAs that decode serine, tyrosine, and leucine codons (Marck and Grosjean, 2002).

Despite the very high structural similarity of all tRNAs engaged in protein synthesis (Fig. 1.1) each must interact productively with only one type of aaRS to be charged with its cognate amino acid and must avoid interacting productively with others (Fig. 1.2) to ensure accurate translations of the genetic code. Typically, there is one population of an aaRS in a cell for each of the 20 canonical amino acids. The charging capacity of a tRNA, termed its functional class, relies on a set of structural features called identity determinants that promote recognition by its cognate

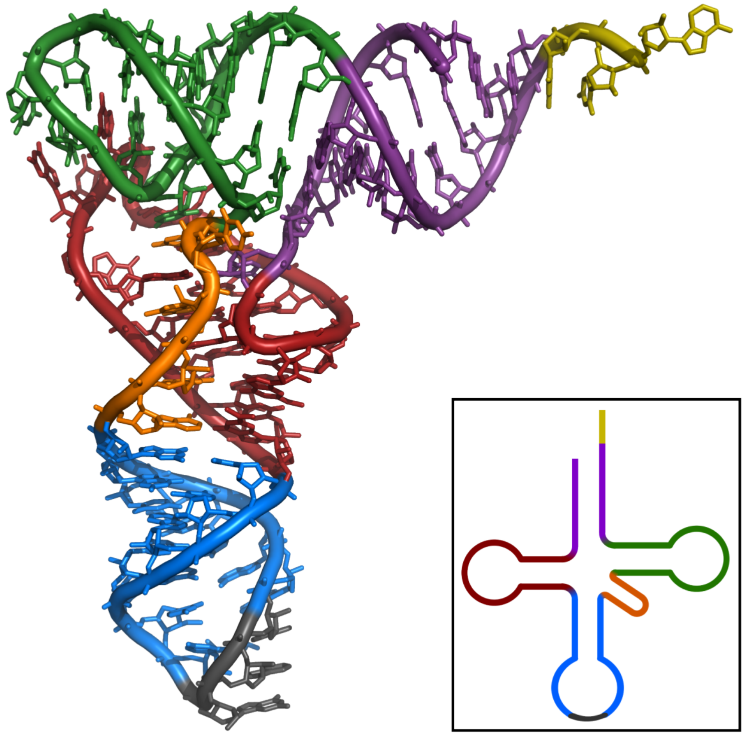

Figure 1.1: Tertiary and secondary structure of tRNA. CCA tail in yellow, acceptor stem in purple, variable loop in orange, D arm in red, anticodon arm in blue with anticodon in black, T arm in green. Source: CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=10312097

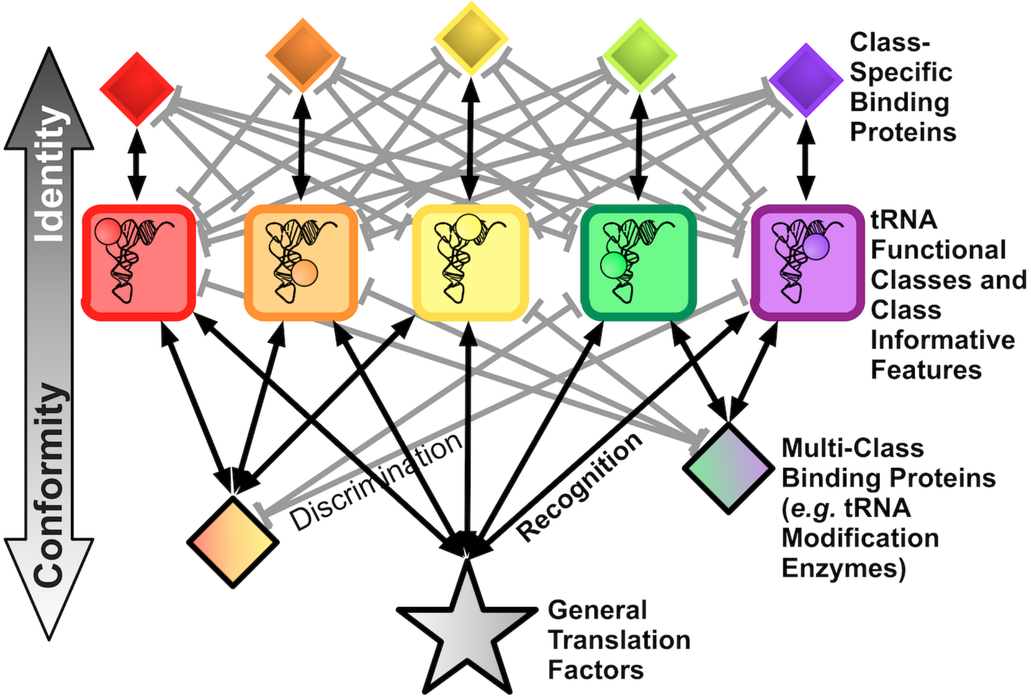
Identity
Conformity
Class-Specific Binding Proteins
tRNA Functional Classes and Class Informative Features
Multi-Class Binding Proteins (*e.g.* tRNA Modification Enzymes)
Discrimination
Recognition
General Translation Factors

Figure 1.2: Diagram of tRNA interaction network. Source: Amrine et al., 2014

aaRS and anti-determinants that discriminate against noncognate aaRSs (Freyhult et al., 2006; Giegé et al., 1998). Previously discovered tRNA identity determinants are mostly concentrated along the acceptor stem, the anticodon itself, and the discriminator base (structural position 73). Identity determinants found outside these locations tend to be species- and lineage-dependent and are distributed throughout the tRNA strucuture, however, these elements, individually, tend to be weaker determinants (Giegé et al., 1998).

Traditionally, identity determinants have been investigated experimentally mostly focusing on one functional class for a single organism. These types of investigations are labor intensive yielding limited results. Even more importantly, these investigations are unable to adequately accommodate that identity elements do not operate in isolated systems partitioned by functional class, but instead the set of determinants and anti-determinants of each functional class must act in concert within the cell to facilitate the correct charging of each tRNA functional class. To fully elucidate the set of identity elements that describe the aaRS-tRNA interaction network of an organism the interdependencies of identity elements across tRNA functional classes can not be ignored suggesting that a systems biology approach is required. The computational method introduced by Freyhult et al., 2006, which I briefly describe below and fully in Chapter 4, seems the best approach to address the interdependencies of identity elements and is utilized in this work to identify tRNA identity elements.

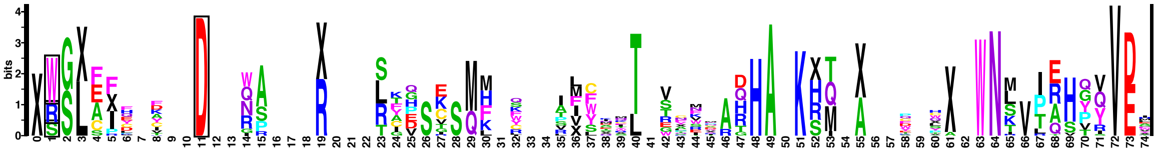

Figure 1.3: Example of a cytosine function logo produced by tsfm.

1.2.2 Estimating tRNA Identity Elements Using Information Theory

For the purpose of estimating the complete set of tRNA identity elements for a given system, termed class informative features (CIFs), a structural feature is defined as the Cartesian product of a state $x \in X$ where $X = \{A, C, G, U, -\}$ and $l \in L$ where L is equal to the length of the sequence $1 \leq l \leq L$. For example, a structural feature would be a cytosine at position 48. tRNA CIFs are determined from the creation of a function logo for each of the possible states x. A function logo is calculated in a conditional probability framework using the methods of molecular information theory (Shannon, 1948; Schneider and Stephens, 1990), which I will briefly describe.

For the purpose of using information theory to estimate tRNA CIFs the possible functional classes of tRNAs are denoted by $Y = \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y\}$. Furthermore, the functional information $I_l(Y|x)$ that a state x confers about the frequencies of different classes Y at position l is: $I_l(Y|x) = H(Y) - e(n_l(x)) - H_l(Y|x))$, where $H_l(Y|x) = -\Sigma_{y \in Y} p_l(y|x) log_2(p_l(y|x))$ is the class entropy or level of uncertainty about the functional class of tRNAs that carry state x at position l, $H(Y) = -\Sigma_{y \in Y} p(y) log_2(p(y))$ is the background entropy which depends on relative frequency of sequences belonging to different classes, $0 \leq p(y), p_l(y|x) \leq 1, \Sigma_{y \in Y} p(y) = 1, \Sigma_{y \in Y} p_l(y|x) = 1$ and $en_l(x))$ is a correction factor to correct for biases caused by small sample size.

Function logos follow the format of sequence logos (Schneider and Stephens, 1990) by plotting functional information as a stacked bar graph with sequence position on the x-axis and information on the y-axis in bits. Each element of a stack is a symbol for a functional class $y \in Y$. Symbol height is determined using the method of Gorodkin et al. (1997) so that a symbol y at position l as height: $height_l = (p_l(y|x)/p(y))/(\Sigma_{w \in Y} p_l(w|x)/p(w)) I_l(Y|x)$ and then symbols are sorted by heights with the symbols with largest height appearing on top. An example function logo is shown in figure 1.3. From this function logo it can be seen that having a cytosine at postion 11 is a strong CIF for the aspartic acid functional class with an information content of roughly 4 bits out of a theoretical maximum of 4.2 bits. This indicates that tRNAs that have the a cytosine at position 11 are very likely to belong to the aspartic acid functional class.

1.2.3 Detecting Phylogenetic Signal in tRNA CIFs

These sets of structural features that determine a tRNAs charging capacity are not static and have been shown to vary widely across the tree of life (Marck and Grosjean, 2002; Freyhult et al., 2007; Amrine et al., 2014). Furthermore, tRNA CIFs have been shown to change in a phylogenetically informative manner that may be exploited to reconstruct deep evolutionary relationships among genomes (Amrine et al., 2014). Additionally, tRNA CIFs seem resistant to compositional biases remaining AT-rich regardless of the nucleotide composition of the genome (Amrine et al., 2014). Theses attributes suggest that tRNA CIFs are an ideal phylogenetic marker for reconstructing deep evolutionary relationships. I implemented two methods to reconstruct evolutionary relationships using the phylogenetic information contained within tRNA CIFs, which I will briefly describe here and more fully in Chapter 2 and 3. The first approach is a machine learning based phyloclassifier that implements an artificial neural network algorithm to probabilistically assign genomes to predefined evolutionary clades. This method is utilized in Chapter 2 to classifier plastid genomes to cyanobacterial clades. The second approach makes use of distance based phylogenetic reconstruction algorithms using a variation of the Jensen-Shannon divergence (Endres and Schindelin, 2003) between function logos as the distance metric. This approach is applied in Chapter 3 to reconstruct the phylogenetic position of gnetophytes among seed plants.

1.3 Organization of Dissertation

The research from this dissertation are contained in four self-contained chapters written in manuscript format. In Chapter 2, “A Robust tRNA-Based Phyloclassifier Predicts a Recent Divergence of Plastids from Cyanobacteria” I test competing hypotheses about the phylogenetic position of plastids within Cyanobacterial using tRNA CIFs and a machine-learning approach. Additionally, I critically evaluate the ability of commonly used phylogenetic models to accommodate site- and lineage-heterogeneity of previously published phylogenomic datasets possibly explaining conflicting results of previous analyses. Chapter 2 is being prepared for submission. In Chapter 3, “tRNA Class Informative Features Support Gnetophytes as Sister to Conifers” I determine the phylogenetic relationship of the gnetophytes, an enigmatic group of plants, within the seed plant phylogeny using tRNA CIFs and an information distance metric. In Chapter 4, I describe the implementation of `tsfm - tRNA Structure Function Mapper`, a command line tool for estimating tRNA CIFs and calculating distance metrics between sets of function logos. Lastly, in Chapter 5 “FAST: FAST Analysis of Sequences Toolbox” I develop a set of command-line tools designed to filter, transform, annotate and analyze biological sequence data. This manuscript was published in Frontiers in Genetics on May 19 2015 (Lawrence et al., 2015). All proceeding chapters are written using the pronoun “we” referring to co-authors and myself.

1.4 References

Amrine, K. C. H., Swingley, W. D., and Ardell, D. H. (2014). tRNA signatures reveal a polyphyletic origin of SAR11 strains among alphaproteobacteria. *PLoS Computational Biology* **10**:2. Ed. by C. A. Ouzounis, e1003454. DOI: 10.1371/journal.pcbi.1003454

Bellot, S. and Renner, S. S. (2016). The Plastomes of Two Species in the Endoparasite Genus *Pilostyles* (Apodanthaceae) Each Retain Just Five or Six Possibly Functional Genes. *Genome Biology and Evolution* **8**:1, pp. 189–201. DOI: 10.1093/gbe/evv251

Blanquart, S. and Lartillot, N. (2008). A Site- and Time-Heterogeneous Model of Amino Acid Replacement. *Molecular Biology and Evolution* **25**:5, pp. 842–858. DOI: 10.1093/molbev/msn018

Butcher, S. E. and Pyle, A. M. (2011). The Molecular Interactions That Stabilize RNA Tertiary Structure: RNA Motifs, Patterns, and Networks. *Accounts of Chemical Research* **44**:12, pp. 1302–1311. DOI: 10.1021/ar200098t

Check, E. H. (2014). Technology: The $1,000 genome. *Nature* **507**:7492, pp. 294–295. DOI: 10.1038/507294a

Datwyler, S. L. and Wolfe, A. D. (2004). Phylogenetic Relationships and Morphological Evolution in Penstemon subg. Dasanthera (Veronicaceae). *Systematic Botany* **29**: pp. 165–176. DOI: http://dx.doi.org/10.1600/036364404772974077

Davis, C. C., Xi, Z., and Mathews, S. (2014). Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. *BMC Biology* **12**:1, p. 11. DOI: 10.1186/1741-7007-12-11

Delagoutte, B., Moras, D., and Cavarelli, J. (2000). tRNA aminoacylation by arginyl-tRNA synthetase: induced conformations during substrates binding. *The EMBO Journal* **19**:21, pp. 5599–5610. DOI: 10.1093/emboj/19.21.5599

Deutscher, M. P. and Ni, R. C. (1982). Purification of a low molecular weight form of rat liver arginyl-tRNA synthetase. *The Journal of biological chemistry* **257**:11, pp. 6003–6

Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory* **49**:7, pp. 1858–1860. DOI: 10.1109/TIT.2003.813506

Figueroa-Martinez, F. et al. (2017). The plastid genomes of nonphotosynthetic algae are not so small after all. *Communicative & integrative biology* **10**:1, e1283080. DOI: 10.1080/19420889.2017.1283080

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, pp. 905–16. DOI: 10.1093/nar/gkj478

Freyhult, E. et al. (2007). New computational methods reveal tRNA identity element divergence between Proteobacteria and Cyanobacteria. *Biochimie* **89**:10, pp. 1276–88. DOI: 10.1016/j.biochi.2007.07.013

Giegé, R., Sissler, M., and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research* **26**:22, pp. 5017–35

Goodman, H. M. and Rich, A. (1962). Formation of a DNA-soluble RNA hybrid and its relation to the origin, evolution, and degeneracy of soluble RNA. *Proceedings of the National Academy of Sciences of the United States of America* **48**:12, pp. 2101–9

Gorodkin, J. et al. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *Bioinformatics* **13**:6, pp. 583–586. DOI: 10.1093/bioinformatics/13.6.583

Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site- specific residue frequencies. *Molecular Biology and Evolution* **15**:7, pp. 910–917. DOI: 10.1093/oxfordjournals.molbev.a025995

Jackson, C. et al. (2018). Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Scientific Reports* **8**:1, p. 1523. DOI: 10.1038/s41598-017-18805-w

Jansen, R. K. and Palmer, J. D. (1988). Phylogenetic Implications of Chloroplast DNA Restriction Site Variation in the Mutisieae (Asteraceae). *American Journal of Botany* **75**:5, p. 753. DOI: 10.2307/2444207

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* **7**:Suppl 1, S4. DOI: 10.1186/1471-2148-7-S1-S4

Lawrence, T. J. et al. (2015). FAST: FAST Analysis of Sequences Toolbox. English. *Frontiers in Genetics* **6**: DOI: 10.3389/fgene.2015.00172

Li, B. et al. (2014). Compositional Biases among Synonymous Substitutions Cause Conflict between Gene and Protein Trees for Plastid Origins. *Molecular Biology and Evolution* **31**:7, pp. 1697–1709. DOI: 10.1093/molbev/msu105

Marck, C. and Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA (New York, N.Y.)* **8**:10, pp. 1189–232

Muñoz-Gómez, S. A. et al. (2017). The New Red Algal Subphylum Proteorhodophytina Comprises the Largest and Most Divergent Plastid Genomes Known. *Current biology : CB* **27**:11, 1677–1684.e4. DOI: 10.1016/j.cub.2017.04.054

Ochoa de Alda, J. A. G. et al. (2014). The plastid ancestor originated among one of the major cyanobacterial lineages. *Nature Communications* **5**: p. 4937. DOI: 10.1038/ncomms5937

Ogle, J. M. et al. (2001). Recognition of Cognate Transfer RNA by the 30S Ribosomal Subunit. *Science* **292**:5518, pp. 897–902. DOI: 10.1126/science.1060612

Oldenburg, D. J. and Bendich, A. J. (2015). DNA maintenance in plastids and mitochondria of plants. *Frontiers in plant science* **6**: p. 883. DOI: 10.3389/fpls.2015.00883

Oldenburg, D. J. and Bendich, A. J. (2016). The linear plastid chromosomes of maize: terminal sequences, structures, and implications for DNA replication. *Current Genetics* **62**:2, pp. 431–442. DOI: 10.1007/s00294-015-0548-0

Palmer, J. D. et al. (1988). Chloroplast DNA Variation and Plant Phylogeny. *Annals of the Missouri Botanical Garden* **75**:4, p. 1180. DOI: 10.2307/2399279

Perl-Treves, R. and Galun, E. (1985). The Cucumis plastome: physical map, intrageneric variation and phylogenetic relationships. *Theoretical and Applied Genetics* **71**:3, pp. 417–429. DOI: 10.1007/BF00251182

Powell, W. et al. (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding* **2**:3, pp. 225–238. DOI: 10.1007/BF00564200

Rhodes, P. R., Zhu, Y. S., and Kung, S. D. (1981). Nicotiana chloroplast genome. *MGG Molecular & General Genetics* **182**:1, pp. 106–111. DOI: 10.1007/BF00422775

Ris, H. and Plaut, W. (1962). Ultrastructure of DNA-containing areas in the chloroplast of Chlamydomonas. *The Journal of cell biology* **13**: pp. 383–91

Ruhfel, B. R. et al. (2014). From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC evolutionary biology* **14**: p. 23. DOI: 10.1186/1471-2148-14-23

Russell, J. R. et al. (1997). Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *TAG Theoretical and Applied Genetics* **95**:4, pp. 714–722. DOI: 10.1007/s001220050617

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**:20, pp. 6097–100

Shannon, C. E. (1948). *A mathematical theory of communication.* DOI: 10.1002/j.1538-7305.1948.tb01338.x

Shinozaki, K. et al. (1986). The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO journal* **5**:9, pp. 2043–2049

Soltis, D. E., Soltis, P. S., and Zanis, M. J. (2002). Phylogeny of seed plants based on evidence from eight genes. *American journal of botany* **89**:10, pp. 1670–81. DOI: 10.3732/ajb.89.10.1670

Vedel, F. et al. (1978). Study of Wheat Phylogeny by EcoRI Analysis of Chloroplastic and Mitochondrial DNAs. *Plant Science Letters* **13**: pp. 97–102

Wang, H.-C. et al. (2018). Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Systematic Biology* **67**:2, pp. 216–235. DOI: 10.1093/sysbio/syx068

Xiong, Y. and Steitz, T. A. (2004). Mechanism of transfer RNA maturation by CCA-adding enzyme without using an oligonucleotide template. *Nature* **430**:7000, pp. 640–645. DOI: 10.1038/nature02711

Chapter 2

A Robust tRNA-Based Phyloclassifier Predicts a Recent Divergence of Plastids from Cyanobacteria

In preparation for submission; Authors: *Travis J. Lawrence, Katherine C. H. Amrine, Wesley D. Swingley, and David H. Ardell*

2.1 Abstract

The trait of oxygenic photosynthesis was acquired by the last common ancestor of Archaeplastida — a eukaryote supergroup comprising glaucophytes, red algae, green algae and land plants — through endosymbiosis of the cyanobacterial progenitor of modern-day plastids. Although a single origin of plastids by endosymbiosis is broadly supported, the location of the root of plastids within Cyanobacteria remains controversial. Recent phylogenomic studies report contradictory evidence for plastids branching either early or late within the cyanobacterial Tree of Life. Here we describe CYANO-MLP, a general-purpose phyloclassifier of cyanobacterial genomes implemented using a Multi-Layer Perceptron that exploits phylogenetic signals in the evolving structure-function maps that we infer bioinformatically from cyanobacterial tRNA gene complements. CYANO-MLP robustly and accurately classifies cyanobacterial genomes into one of eight well-supported cyanobacterial clades. Our results with CYANO-MLP support a late-branching origin of plastids: we classify 99.32% of 440 plastid genomes into one of two late-branching cyanobacterial clades with strong statistical support, and confidently assign 98.41% of plastid genomes to one late-branching clade of starch-producing marine/freshwater diazotrophic cyanobacteria. CYANO-MLP correctly classifies the chromatophore of *Paulinella chromatophora* and rejects sisterhood of plastids with the early-branching cyanobacterial species *Gloeomargarita lithophora*. We reconcile our results with previous studies by showing that

recently applied phylogenetic models and character recoding strategies fit cyanobacterial/plastid phylogenomic datasets poorly, because of both site-heterogeneity in substitution processes and lineage-heterogeneity in compositions.

2.2 Introduction

The acquisition of a cyanobacterial endosymbiont by the last common ancestor of Archaeplastida (Adl et al., 2012) transferred the trait of oxygenic photosynthesis to eukaryotes over one billion years ago (Falcón et al., 2010). The diversity of eukaryotic photoautotrophs radiating from this event profoundly transformed the terrestrial biosphere through changes to primary biomass production, atmospheric oxygenation, and the colonization of new ecosystems (Kenrick and Crane, 1997).

Despite substantial progress and consensus on a robust Cyanobacterial Tree of Life (CyanoToL) (Ochoa de Alda et al., 2014; Shih et al., 2013; Schirrmeister et al., 2015; Latysheva et al., 2012; Blank and Sánchez-Baracaldo, 2010), the root of plastids within the CyanoToL remains controversial. Recent phylogenomic studies have strongly supported contradictory conclusions, with plastids diverging either early (Bhattacharya and Medlin, 1995; Turner et al., 1999; Shih et al., 2013; Criscuolo and Gribaldo, 2011; Ponce-Toledo et al., 2017; Sánchez-Baracaldo et al., 2017) or late (Ochoa de Alda et al., 2014; Falcón et al., 2010; Blank, 2013; Dagan et al., 2013) within the CyanoToL. However, orthogonal evidence from endosymbiotic gene transfers (Deusch et al., 2008) and eukaryotic evolution of glycogen and starch metabolic pathways (Deschamps et al., 2008; Ball et al., 2011) have more consistently supported a late origin of plastids within the CyanoToL.

Phylogenetic inferences concerning plastid origin are complicated by large evolutionary distances accumulated over at least one billion years of vertical descent, by extreme genome reduction in plastids (Martin et al., 1998; Timmis et al., 2004) and Cyanobacteria (Rocap et al., 2003; Dufresne et al., 2003; Dufresne et al., 2005), and by a number of secondary and tertiary endosymbiosis events during plastid evolution. Genome reduction alters the stationary compositions of genomes and gene products, violating the assumptions and applicability of many phylogenetic models (Philippe and Roure, 2011; Blanquart and Lartillot, 2008; Lartillot et al., 2007; Domman et al., 2015; Foster and Schultz, 2004).

Recently, we introduced a machine learning approach to the phyloclassification of genomes based on functional signatures of tRNA gene complements (Amrine et al., 2014). We demonstrated the strong recall and accuracy of a tRNA-based alphaproteobacterial phyloclassifier despite convergent non-stationary compositions of alphaproteobacterial tRNAs and likely horizontal transfers of genes for tRNAs and tRNA-interacting proteins (Amrine et al., 2014). Our tRNA phyloclassifier uses information theoretical signatures of tRNA functional features that we call Class-Informative Features (CIFs) (Freyhult et al., 2006).

In the present work, we improved our tRNA-based phyloclassifier approach and applied it to investigate the origin of plastids within the CyanoToL. Based on 5,476 tRNA gene sequences from 113 cyanobacterial genomes, our CYANO-Multi-Layer Perceptron (CYANO-MLP) phyloclassifier consistently classifies over 400 plastid genomes within the B2 and B3 sister clades of Cyanobacteria (Shih et al., 2013), which contain starch-producing marine/freshwater unicellular diazotrophic species, consistent with previously described evidence from metabolism and endosymbiotic gene transfers (Deusch et al., 2008; Deschamps et al., 2008; Falcón et al., 2010; Ochoa de Alda et al., 2014). Furthermore, we reconcile our results with prior work by showing that recently applied phylogenetic models and character recoding strategies fit the combined CyanoToL/plastid phylogenomic datasets poorly because of lineage-specific compositional biases.

2.3 Methods

2.3.1 tRNA data

We downloaded the set H of 117 cyanobacterial genomes analyzed in Shih et al., 2013, the set Gl of one genome of the cyanobacterium *Gloeomargarita lithophora*, recently argued to be a sister to plastids (Ponce-Toledo et al., 2017; Sánchez-Baracaldo et al., 2017), and the set Pc of one genome of the chromatophore of the fresh-water amoeba *Paulinella chromatophora* from NCBI. Let $C \equiv H \cup Gl \cup Pc$. For every genome $g \in C$, we annotated a set T_g of tRNA genes as the union of predictions from tRNAscan-SE v1.31 (Lowe and Eddy, 1997) in bacterial mode and ARAGORN v1.2.36 (Laslett and Canback, 2004) with default settings. We downloaded a set P of 440 complete plastid genomes containing representatives from all three lineages of Archaeplastida (Glaucocystophyta, Rhodophyta, and Viridiplantae) from NCBI. We annotated tRNA genes in plastid genomes similarly to cyanobacterial genomes, however, predictions by ARAGORN v1.2.36 (Laslett and Canback, 2004) containing introns in tRNA isotypes that have not been previously described to contain introns (Manhart and Palmer, 1990; Vogel et al., 1999; Simon et al., 2003) were discarded as likely false positives. Additionally, tRNA genes containing anticodons identified as absent in most land plant plastids (Osawa et al., 1992; Sugiura and Wakasugi, 1989; Alkatib et al., 2012) were filtered from land plant plastids.

We annotated tRNAs containing the CAU anticodon as initiator tRNA $^{\mathrm{Met}}$, elongator tRNA $^{\mathrm{Met}}$, or tRNA $^{\mathrm{Ile}}$ $_{\mathrm{CAU}}$ using TFAM v1.4 (Ardell and Andersson, 2006) with the covariance model used in Amrine et al., 2014. We aligned tRNA sequences using COVEA v2.4.4 (Eddy and Durbin, 1994) using the prokaryotic tRNA covariance model in Lowe and Eddy, 1997. We edited the alignment by first removing sites containing 99% or more gaps using FAST v1.6 (Lawrence et al., 2015), followed by removal of sequences with unusual secondary structure. Lastly, we mapped sites to Sprinzl coordinates (Sprinzl et al., 1998) and removed the variable arm, CCA tail, and sites not mapping to a Sprinzl coordinate manually using Seaview v4.6.1 (Gouy et al., 2010).

Cyanobacterial tRNAs were partitioned into sets T_g for each genome g of origin, and separately into sets T_X by cyanobacterial clade of origin, with $X \in C \equiv \{A, B1, B23, C1, C3, E, F, G\}$. These clade labels correspond to those identified in Shih et al., 2013 except for replacement of clades B2 and B3 by their union B23, and the omission of clades C2 and D for insufficient data, as those were represented by fewer than 120 tRNA sequences (Fig. 2.1 and Table 2.1). There were 113 cyanobacterial genomes remaining after exclusion of clades C2 and D.

The consolidation of tRNAs genes into these partitions for function logo calculations is justified because it has been previously shown that tRNA CIFs diverge in a phylogenetically informative way (Amrine et al., 2014; Freyhult et al., 2007) and the molecular information calculations used in function logos weights tRNA CIFs based on their conservation within the clade. We decided against consolidating clades C2 and D into their next largest monophyletic grouping because the next largest clade would have combined clades C1, C2, C3, and D. This combined clade would have been dominated by the large number of sequences in C1 (marine *Prochlorococcus*/*Synechococcus*) effectively masking the diversity of clade C3 in addition to the diversity of the smaller clades C2 and D possibly leading to a greater decrease in phylogenetic diversity represented in our function logos than if we just excluded clades C2 and D from our data analysis.

2.3.2 Genome Scoring

Following Amrine et al., 2014, we produced training data for our classifier by first calculating clade-dependent Gorodkin heights (Gorodkin et al., 1997; Amrine et al., 2014) $h^X_{f_i}$, in function logos (Freyhult et al., 2006) for all clades $X \in C \equiv \{A, B1, B23, C1, C3, E, F, G\}$ and for all features $f_i \in \{A, C, G, U\} \times SC$, where SC is the set of Sprinzl Coordinates (Sprinzl et al., 1998) from input tRNA gene sets T_X. A Gorodkin height $h^X_{f_i}$ is proportional to the gain in functional information about a tRNA after observing that its sequence contains feature f_i.

For any tRNA gene complement T_h of tRNA genes from any cyanobacterial or plastid genome h, we calculated a vector of tRNA CIF-based scores, $\mathbf{S_h} = \langle S^A_h, S^{B1}_h, S^{B23}_h, S^{C1}_h, S^{C3}_h, S^E_h, S^F_h, S^G_h \rangle$ in which score components S^X_h of genome $h \in H$ with clade $X \in C$ is the average over T_h of the sum of the Gorodkin heights (Gorodkin et al., 1997) of its features in a function logo (Freyhult et al., 2006) representation of clade X:

$$S^X_h \equiv \frac{1}{|T_h|} \sum_{t \in T_h} \sum_{f_i \in t} h^X_{f_i}, \tag{2.1}$$

where tRNA gene t is represented by a set of features We computed function logos using custom software available at https://github.com/tlawrence3/tsfm/tree/v0.9.6. Following standard recommendations (Ching et al., 2018), we standardized our score vectors before training as follows:

$$S^{X\prime}_h \equiv \left(S^X_h - \overline{S^X} \right) / \sigma^X, \tag{2.2}$$

where $\overline{S^X} = \sum_{h \in H} S_h^X / |H|$ is the average score of cyanobacterial genomes $h \in H$ against clade X, and σ^X is the standard deviation of scores of cyanobacterial genomes $h \in H$ against clade X. Denote by $\mathbf{S'_h}$ the vector of standardized scores for genome h, ordered correspondingly to $\mathbf{S_h}$.

2.3.3 Phyloclassifier training and selection

We implemented our multilayer neural network phyloclassifier using the MLPClassifier API of scikit-learn v0.18.1 (Pedregosa et al., 2011) using Python v3.5.2. We trained CYANO-MLP from the standardized score vectors $\mathbf{S'_g}$ from cyanobacterial genomes ($g \in H$) using up to 2000 training epochs stopping early if for two consecutive iterations the Cross-Entropy loss function value did not decrease by a minimum of 1×10^{-4}, with random shuffling of data between epochs. We used the rectifier activation function for hidden layer neurons, the L-BFGS algorithm for weight optimization, and an alpha value of 0.01 for the L2 regularization penalty parameter. Lastly, the softmax function was applied to the output to calculate classification probability vectors. We evaluated the performance of all permutations of neural network architectures composed of one to four hidden layers each containing eight to sixteen nodes, with the parameters values described above, using the average accuracy from leave-one-out cross validation (LOOCV) as our metric. To test the statistical significance of the average accuracy from LOOCV, we used a permutation test producing the null distribution from 100,000 datasets with clade labels randomly swapped between score vectors, followed by retraining of the classifier, and calculated average accuracy using LOOCV for each permuted dataset. To test the robustness of our classifications with respect to our tRNA CIF data, we performed 100 bootstrap replicates, re-sampling sites in our tRNA alignment data, and retrained bootstrap replicates of the CYANO-MLP model. We summarized bootstrap results for cyanobacterial genomes by the number of replicates in which the most probable classification for a genome was its clade of origin.

2.3.4 Plastid classification

For each genomic tRNA gene set T_p from each genome p in the set P of 440 genomes from plastids and the chromatophore genome of *P. chromatophora*, we produced standardized scores vectors $\mathbf{S'_p}$ using eq. 2.2. Next, we used CYANO-MLP or its bootstrap replicates to produce vectors of classification probabilities and bootstrap distributions thereof.

2.3.5 Evaluation of phylogenetic model adequacy

We examined the goodness of fit of the dataset of Shih et al., 2013 and the chloroplast-marker dataset of Ponce-Toledo et al., 2017, which both supported an early-branching position of plastids, and dataset 11 of Ochoa de Alda et al., 2014 which supported a late-branching position of plastids, with the substitution models originally used by

Shih et al., 2013 and Ochoa de Alda et al., 2014; Ponce-Toledo et al., 2017, LG+Γ (Le and Gascuel, 2008) and CAT-GTR+Γ (Lartillot and Philippe, 2004; Lartillot et al., 2007) respectively.

Posterior Predictive Analyses (PPA) were performed to test model fit for site-specific constraints using PPA-DIV (Lartillot et al., 2007) and across-lineage compositional biases using PPA-MAX and PPA-MEAN (Blanquart and Lartillot, 2008). Briefly, PPA-DIV uses the mean number of distinct amino acids observed at each site as the test statistic, where both PPA-MAX and PPA-MEAN use the deviation between individual taxon amino acid frequencies and dataset frequencies, either using the maximum squared deviation over all taxa or the average squared differences across all taxa respectively. PPA analyses were conducted using Phylobayes MPI v1.8 (Lartillot et al., 2013) using at least 1,000 replicates. Additionally, we assessed model adequacy under three amino acid recoding strategies, Dayhoff-6 (Day6) (Dayhoff et al., 1978), 6-state recoding strategy of Susko and Roger (SR6) (Susko and Roger, 2007), and the 6-state recoding strategy of Kosiol et al. (KGB6) (Kosiol et al., 2004). PPA results were interpreted using z-scores under the assumption that the test statistics follow a normal distribution. We used a z-score threshold of $\geq |5|$ as strong evidence for rejecting the model. We performed phylogenetic analyses using Phylobayes MPI v1.8 (Lartillot et al., 2013) running two MCMC chains in parallel for each analysis. Convergence of analyses was assessed using TRACECOMP and BBCOMP utilities provided with Phylobayes MPI. Convergence was assumed when the discrepancies of model parameters and bipartition frequencies between independent chains was less than 0.18. The number of cycles to discard as burn-in was determined by visually examining the traces of the log-likelihood and other model parameters for stationarity using Tracer v1.6.0.

2.4 Results

2.4.1 tRNA Data and CIF estimation

We annotated and extracted 5,476 tRNA gene sequences from the 117 cyanobacterial genomes analyzed in (Shih et al., 2013) averaging 48.46 tRNA genes per cyanobacterial genome, 14,841 tRNA gene sequences in 440 Archaeplastida plastid genomes averaging 33.73 tRNA genes per plastid genome, 44 tRNA gene sequences from the cyanobacterium *Gloeomargarita lithophora*, and 42 tRNA gene sequences from the chromatophore genome of the fresh-water amoeba *Paulinella chromatophora* (Table 2.1; Dataset S1). We estimated informative function logos for cyanobacterial clades A, B1, B23, C1, C3, E, F, and G (Fig. 2.1, S2.3-S2.10). The C1 clade is the best-sampled clade with a divergent composition, elevated in contents of G and C and diminished in contents of A and U (Table 2.1). This clade exhibited many gains of information in Uracil CIFs (Fig. 2.1B) and also Adenine CIFs (Figs. S2.3-S2.10).

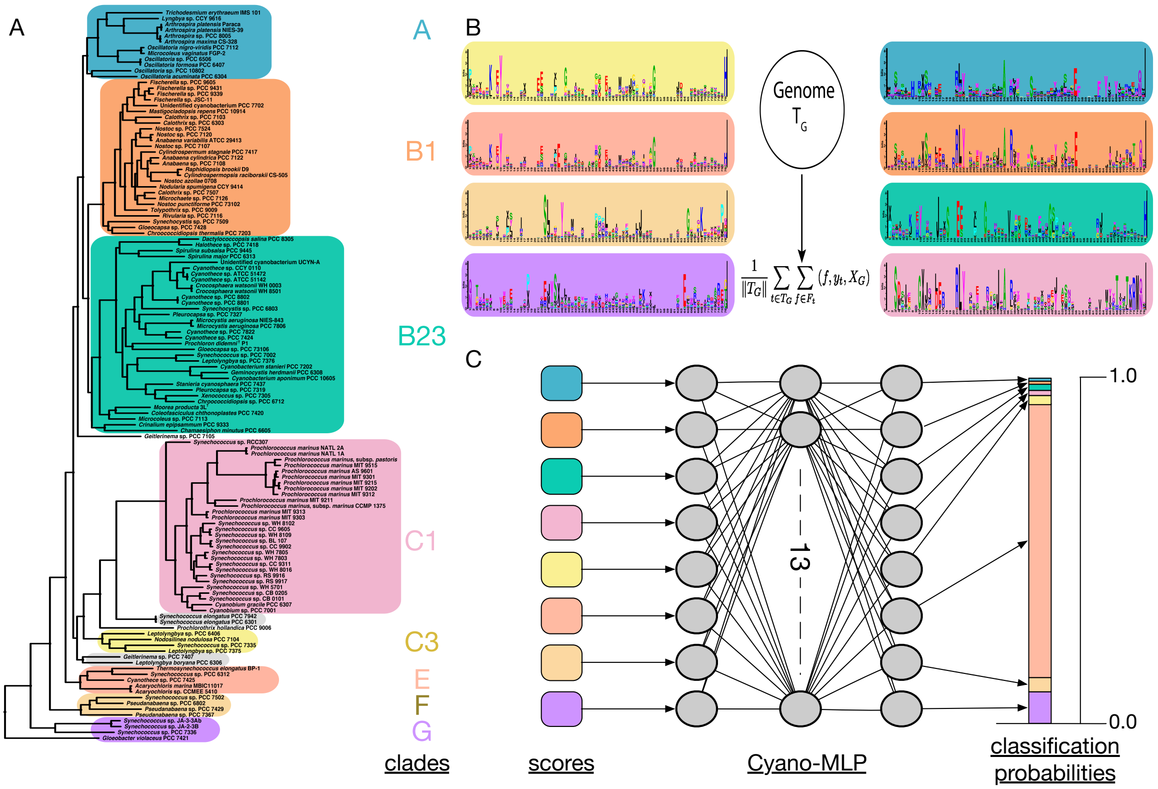
A
A
B1
B23
C1
C3
E
F
G
clades
B
Genome
T_G
$\frac{1}{\|T_G\|}\sum_{t\in T_G}\sum_{f\in F_t}(f, y_t, X_G)$
C
scores
13
Cyano-MLP
1.0
0.0
classification probabilities

Figure 2.1: Schematic overview of CYANO-MLP workflow. (A) Cyanobacterial phylogeny from (Shih et al., 2013) used to define cyanobacterial clades from which to estimate tRNA CIFs. Cyanobacterial clades are indicated by background color and annotated with labels from (Shih et al., 2013) except for clade B23, which is combination of clade B2 and B3. Clades with grey backgrounds were excluded from analysis because of limited number of genomes. (B) Uracil function logos (Freyhult et al., 2006) for each cyanobacterial clade with background colors corresponding to background colors in (A). (C) and (D) Notional diagram illustrating generation of an input score vector from the tRNA gene complement T_g of genome g, the neural network architecture of CYANO-MLP, and the classification probability vector output from CYANO-MLP, represented as a stacked bar chart.

Table 2.1: Summary statistics of the genomes and tRNA genes for Cyanobacterial clades and plastid groups. Statistics include the number of genomes (G), the number of tRNAs genes (T), average number of tRNA genes per genome (T/G), total number of nucleotides contained in tRNA genes (N), average number of nucleotides per tRNA gene (N/T), the percent of adenine (%A), thymine (%T), guanine (%G), and cytosine (%C) nucleotides contained in tRNA genes.

Clade	G	T	T/G	N	N/T	%A	%T	%G	%C
Cyanobacteria									
A	11	555	50.45	40,362	72.72	20.1	23.3	31.2	25.4
B1	27	1,395	51.67	101,640	72.86	19.7	23.3	31.6	25.4
B23	30	1,314	43.80	95,685	72.82	19.5	23.0	32.0	25.5
C1	29	1,205	41.55	87,856	72.91	18.8	21.8	32.7	26.7
C2	2	90	45.00	6,550	72.78	19.5	21.9	32.2	26.4
C3	3	142	47.34	10,346	72.86	19.1	22.8	32.3	25.7
D	2	116	58.00	8,430	72.67	19.7	23.3	31.7	25.3
E	5	266	53.20	19,378	72.85	19.3	22.8	32.1	25.8
F	4	211	52.75	15,339	72.70	20.1	23.4	31.3	25.2
G	4	182	45.50	13,218	72.63	18.7	21.7	32.8	26.8
G. lithophora	1	44	44	3,194	72.59	18.8	22.7	32.3	26.2
Plastids									
Charophyta	10	352	35.20	25,513	72.48	21.1	24.2	30.1	24.6
Chlorophyta	7	226	32.29	16,436	72.73	21.9	25.4	29.1	23.6
Cryptophyta	4	117	29.25	8,512	72.75	20.9	24.2	29.9	25.0
Heterokonta	33	992	30.06	72,229	72.81	21.2	25.1	29.6	24.1
Eudicots	191	6,593	34.52	478,337	72.55	21.4	25.2	29.7	23.7
Euglenaceae	10	276	27.60	20,070	72.72	22.3	27.1	28.4	22.3
Monilophytes	8	270	33.75	19,641	72.74	21.2	24.5	30.0	24.3
Gymnospermae	26	824	31.69	59,867	72.65	21.9	24.6	29.6	23.9
Haptophyta	4	111	27.75	8,079	72.78	21.0	25.1	29.7	24.2
Monocots	112	3,930	35.10	285,302	72.60	21.7	25.2	29.5	23.7
Magnoliids	9	315	35.00	22,856	72.56	21.3	24.9	29.7	24.0
Nymphaeales	2	68	34.00	4,934	72.56	21.4	25.1	29.8	23.7
Rhodophyta	20	624	31.20	45,438	72.82	21.9	25.2	29.1	23.8
Bryophyta	3	108	36.00	7,845	72.64	22.2	25.4	29.0	23.4
Glaucocystophyta	1	35	35	2,543	72.66	20.4	23.7	30.9	25.0
P. chromatophora	1	42	42	3,060	72.86	19.1	21.7	32.7	26.5

Table 2.2: LOOCV classification results for CYANO-MLP.

Labeled Clade	A	B1	B23	C1	C3	E	F	G
A	6 (54.55%)	1 (9.09%)	3 (27.28%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (9.09%)
B1	1 (3.70%)	26 (96.30%)	0 (0.00%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
B23	2 (6.67%)	0 (0.00%)	28 (93.33%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
C1	0 (0.00%)	0 (0.00%)	0 (0.00%)	29 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
C3	0 (0%)	0 (0%)	1 (33.33%)	0 (0%)	2 (66.67%)	0 (0%)	0 (0%)	0 (0%)
E	1 (20.00%)	0 (%)	2 (40.00%)	0 (0%)	0 (%)	2 (40.00%)	0 (0%)	0 (0%)
F	0 (0%)	1 (25.00%)	1 (25.00%)	0 (0%)	0 (0%)	0 (0%)	2 (50.00%)	0 (0%)
G	0 (0%)	1 (25.00%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (75.00%)

2.4.2 Training and Validation of a tRNA-Based Cyanobacterial Phyloclassifier

We trained our cyanobacterial phyloclassifier, termed CYANO-MLP, based on cyanobacterial clade CIFs, that takes as input tRNA gene complements from a test genome, and labels them as belonging to one of the eight cyanobacterial clades or grades shown in Fig. 2.1. Based on a systematic optimization over neural network architecture parameters, we selected a neural network consisting of a single hidden layer of 13 nodes (Fig. 2.1), which achieved an average accuracy score of 0.8673 ($p = 0.0001$), calculated using Leave One Out Cross-Validation (LOOCV; Fig. 2.2; Table 2.2). These results demonstrate the ability of CYANO-MLP to correctly classify cyanobacteria. When trained on the full cyanobacterial dataset CYANO-MLP accurately classified 100% of cyanobacteria to their respectful clades. Additionally, all cyanobacterial classifications received at least 97% bootstrap support suggesting a consistent phylogenetic signal across tRNA CIFs.

2.4.3 The *Paulinella chromatophora* Chromatophore Classifies Consistently to the Marine *Prochlorococcus / Synechococcus* Clade

The engulfment of a cyanobacterium by the ancestor of the freshwater *P. chromatophora* represents a second and more recent primary endosymbiosis event within eukaryotes (Parfrey et al., 2006; Keeling, 2004). The phylogenetic position of *P. chromatophora*'s photosynthetic organelle, called the chromatophore, as sister to the marine *Prochlorococcus*/*Synechococcus* clade (clade C1; Fig. 2.1) is uncontroversial and well-supported by several phylogenomic analyses (Bhattacharya and Medlin, 1995; Shih et al., 2013; Sánchez-Baracaldo et al., 2017), likely because of its more recent origin. Because of this, our first goal was to determine if CYANO-MLP could reproduce previous results of the phylogenetic relationship of the chromatophora to extant cyanobacterial clades. CYANO-MLP confirmed previous results classifying the chromatophore to clade C1 with 99.98% probability with 100% bootstrap support (Fig. 2.3; Table 2.3).

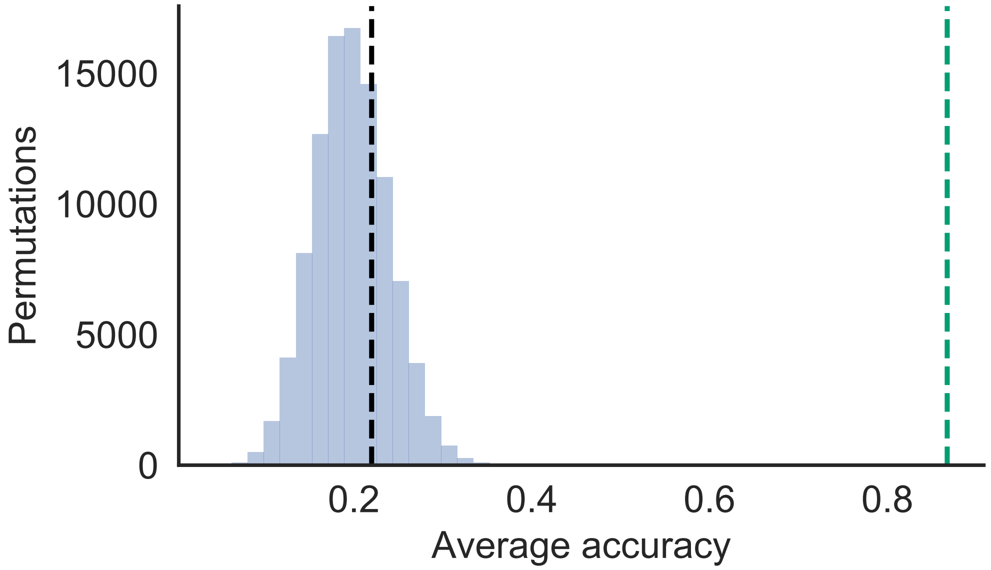

Figure 2.2: Null distribution of average accuracy using LOOCV estimated by 100,000 label swapping permutation datasets. Black dotted line is the expected average accuracy if cyanobacteria genomes were randomly classified. Green dotted line is the average accuracy using our single hidden layer phyloclassifier.

2.4.4 Plastid Genomes Robustly Phyloclassify as Late-Branching Cyanobacteria

Our main motivation was to determine the phylogenetic origin of the primary endosymbiosis event leading to Archaeplastida plastids. We classified 437/440 (99.32%) of plastid genomes to late-branching clades of Cyanobacteria with 433 plastid genomes classifying to clade B23 and 4 plastid genomes classifying to clade A with high probability (Fig. 2.3; Table S2). Plastid genomes from all three Archaeplastida lineages, Rhodophyta, Chloroplastida, and Glaucocystophyta, classified to the late-branching clade B23. Furthermore, the majority of plastid bootstrap replicates classified to late-branching clades A, B1, and B23 with the median bootstrap of plastid groups for clade B23 at or above 70, except for the Glaucocystophyta genome (Fig. 2.3,2.4,2.5,S2.1,S2.2). The remaining three genomes classified to early diverging lineages with two plastid genomes classifying to clade F and one plastid genome classifying to clade G (Fig. 2.3, Table 2.3).

2.4.5 Phyloclassification of *G. lithophora* is Consistent with its Early Divergence within Cyanobacteria

Recent phylogenomic analyses have supported plastids as sister to an early-diverging lineage with *Gloeomargarita lithophora* as its only member (Ponce-Toledo et al., 2017; Sánchez-Baracaldo et al., 2017). Unfortunately, since this lineage only contains a

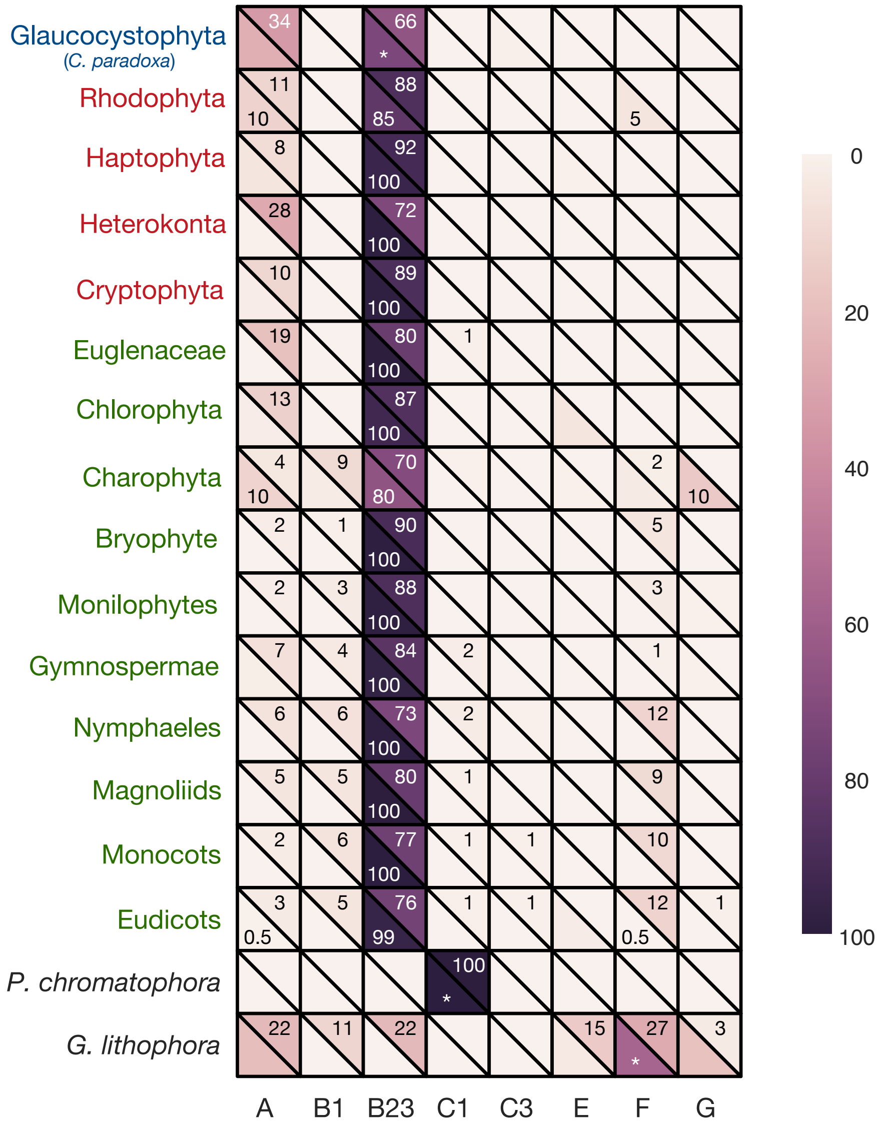

Figure 2.3: Average classification and median bootstrap results of plastid groups, the chromatophore of *P. chromatophora*, and the cyanobacterium *G. lithophora* with CYANO-MLP. Row labels are colored by the three main lineages of Archaeplastida (Red: Rhodophyta, Green: Chloroplastida, Blue: Glaucocystophyta, Black: non-Archaeplastida). (Lower Diagonals) Heatmap of the average classification probability vector for groups and single genomes. Numbers within squares indicate percentage of genomes classifying to the Cyanobacterial clade. Locations without annotations indicate that zero genomes from that group classified to the particular Cyanobacterial clade. Asterisks indicate that the group contains a single genome. (Upper Diagonals) Heatmap of median of bootstrap values over genomes for each group or single genome. Annotations are the median bootstrap value for each group or single genome. Locations lacking annotation indicate zero bootstrap replicates.

Table 2.3: Classification results for plastid genomes and the chromatophore of *P. chromatophora* using CYANO-MLP. Results are summarized by plastid groups. Number of genomes classifying to each Cyanobacterial clade and percent are shown.

Plastid Clade	A	B1	B23	C1	C3	E	F	G
Chlorophyta	0 (0%)	0 (0%)	7 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Charophyta	1 (10%)	0 (0%)	8 (80%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (10%)
Cryptophyta	0 (0%)	0 (0%)	4 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Heterokonta	0 (0%)	0 (0%)	33 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Eudicots	1 (0.52%)	0 (0%)	189 (98.95%)	0 (0%)	0 (0%)	0 (0%)	1 (0.52%)	0 (0%)
Euglenaceae	0 (0%)	0 (0%)	10 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Monilophytes	0 (0%)	0 (0%)	8 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Gymnospermae	0 (0%)	0 (0%)	26 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Haptophyte	0 (0%)	0 (0%)	4 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Monocots	0 (0%)	0 (0%)	112 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Magnoliids	0 (0%)	0 (0%)	9 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Nymphaeales	0 (0%)	0 (0%)	2 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Rhodophyta	2 (10%)	0 (0%)	17 (85%)	0 (0%)	0 (0%)	0 (0%)	1 (5%)	0 (0%)
Bryophyte	0 (0%)	0 (0%)	3 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Glaucocystophyta	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Plastid total	4(0.91%)	0 (0%)	433 (98.41%)	0 (0%)	0 (0%)	0 (0%)	2 (0.45%)	1 (0.23%)
P. chromatophora	0 (0%)	0 (0%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

single genome there were insufficient tRNA sequences to estimate informative function logos so we were unable to include this lineage in CYANO-MLP. However, we classified *G. lithophora* using CYANO-MLP to determine if it classified similarly to plastids, which would suggest that *G. lithophora* could not be rejected as sister to plastids. The majority of the classification probability of *G. lithophora* was contained within early-diverging lineages with clade F receiving 57.3%, clade G receiving 18.4%, and clade E receiving 3.2%, however, *G. lithophora* also received a 20.3% probability of belonging to the later diverging clade A (Fig. 2.3). This result supports *G. lithophora* as an early-diverging cyanobacterial lineage in agreement with recent phylogenomic analyses (Ponce-Toledo et al., 2017; Sánchez-Baracaldo et al., 2017), but rejects the sister relationship with plastids.

2.4.6 Inadequate Modeling of Systematic Biases Possibly Explains Discrepancies with Prior Work

We found site-specific amino acid constraints a critical process to describe in Cyanobacterial/plastid phylogenomic datasets (Figure 2.6A; Table 2.4). Unsurprisingly, LG, an empirical matrix model, was unable to account for site-specific constraints in all three datasets (Figure 2.6A). The inability of empirical matrix models to describe site-specific constraints has been previously reported (Lartillot and Philippe, 2004; Lartillot et al., 2007) and is known to result in long-branch attraction artefacts caused by an underestimation of homoplasy (Lartillot et al., 2007). In contrast, the CAT model, specifically designed to accommodate site-specific constraints, was able to ade-

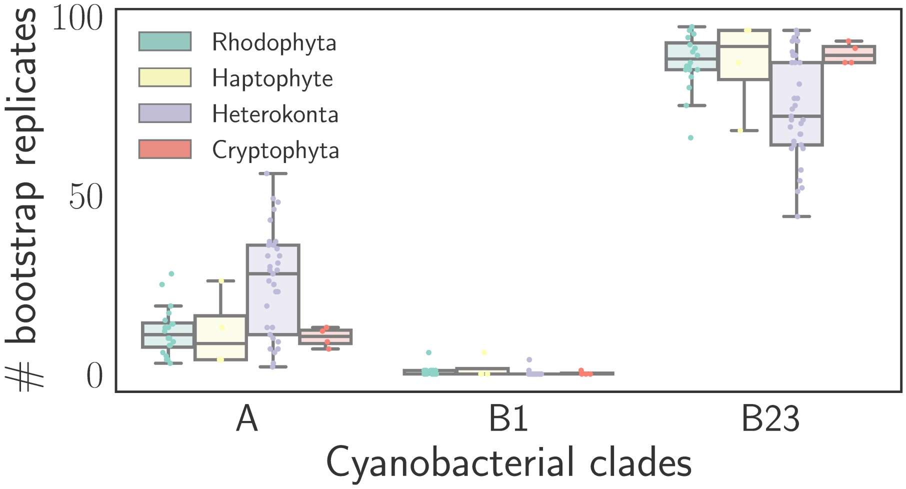

Figure 2.4: Classification results of 100 bootstrap replicates of each Rhodophyta derived plastid genome. Results are summarized by Red plastid group with boxes spanning from the 25th percentile (bottom) to the 75th percentile (top) of bootstrap replicates classifying to the indicated Cyanobacterial clade per genome with the bisecting line marking the median value. Error bars indicate the shorter of either ± the interquartile range or the span of bootstrap replicates per genome. Dots show bootstrap replicates for individual genomes. Cyanobacterial clades C1, C3, E, F, and G were omitted because a limited number of bootstrap replicates per genome classified to these clades (see Fig. S2.1).

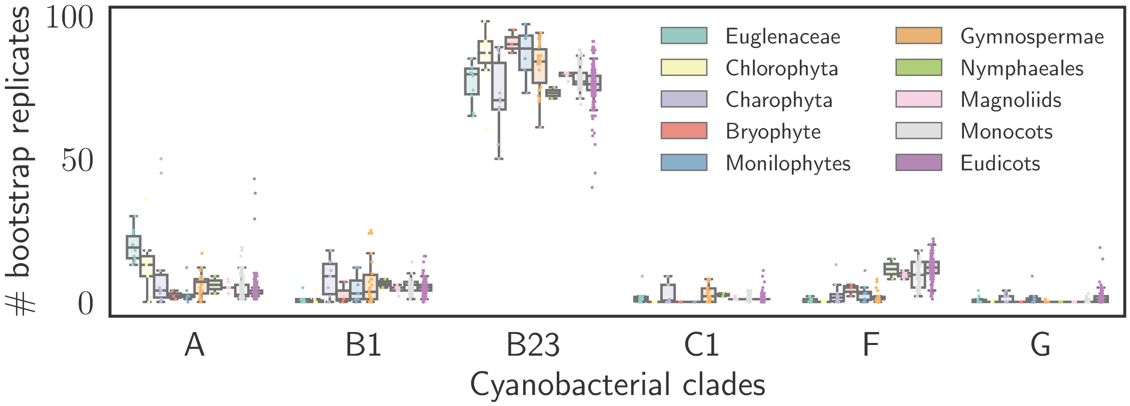

Figure 2.5: Classification results of 100 bootstrap replicates of each Chloroplastida derived plastid genome. Cyanobacterial clades C3 and E were omitted because a limited number of bootstrap replicates per genome classified to these clades (see Fig. S2.2). Results are summarized identically to Fig. 2.4.

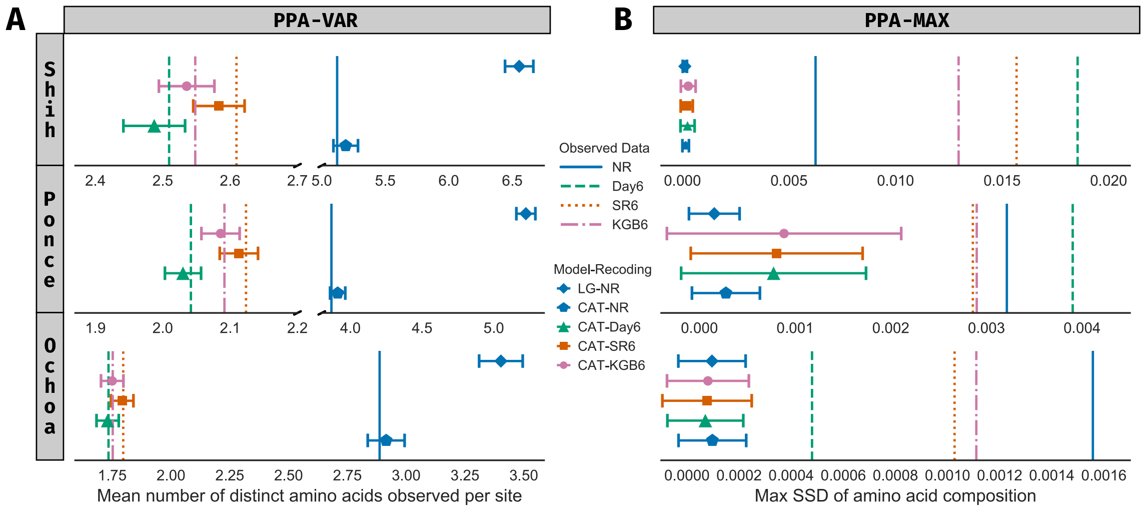

Figure 2.6: Results of posterior predictive analyses of phylogenomic datasets of Shih et al., 2013, Ponce-Toledo et al., 2017, and Ochoa de Alda et al., 2014. Rows represent phylogenomic dataset and columns indicate the posterior predictive test statistic. Observed values calculated for each test statistic is represented by vertical lines. Color and patterns of vertical lines indicate amino acid recoding strategy (NR: No recoding, DAY6: six state Dayhoff recoding, SR6: six state recoding strategy of Susko and Roger, 2007, KGB6: six state recoding strategy of Kosiol et al., 2004). Symbols specify the average value for each posterior predictive test statistic calculated from simulated datasets. Error bars depict ± 5 standard deviations. Symbol shape and color indicate phylogenetic model (LG: LG+4G, CAT: CAT-GTR+4G) and recoding strategy used for simulating data for posterior predictive test statistic calculations. If similarly colored error bars and vertical lines overlap the phylogenetic model is considered able to adequately describe the systematic bias of the dataset. (A) Results of PPA-VAR assessing the ability to describe site-specific amino acid constraints. (B) Results of PPA-MAX assessing the ability to describe lineage-specific compositional biases.

quately describe this process in each dataset (Figure 2.6A; Table 2.4). Additionally, we found that none of the model/amino acid recoding combinations adequately described lineage-specific compositional biases with all being strongly rejected ($|$ z-score $| \geq 5$) (Figure 2.6B; Table S3). It has been shown when lineage-specific compositional biases are not described adequately unrelated sequences sharing similar compositions may cluster together leading to artefacts during phylogenetic tree reconstruction (Blanquart and Lartillot, 2008).

2.5 Discussion

2.5.1 Origin of Plastids from Late-Branching Cyanobacteria

Based on the classifications of CYANO-MLP, the tRNA CIF evidence strongly supports a late-branching origin of plastids, likely within or closely related to the B23 clade of the CyanoToL, with 437/440 (99.32%) of plastid genomes classified within late-branching clades A and B23, of which, 433 classify to B23 including plastid genomes from all three lineages of Archaeplastida (Fig. 2.1,2.3; Table 2.3). Furthermore,

Table 2.4: Results of the posterior predictive analyses presented as z-scores.

Data	Model	Recoding	PPA-Div	PPA-MAX	PPA-Mean
Shih	LG+4G	None	64.3264	308.247	112.4
	CAT-GTR+4G	None	3.48296	199.277	125.469
	CAT-GTR+4G	Dayhoff6	-2.41113	276.471	76.2146
	CAT-GTR+4G	KGB6	-1.54966	183.227	75.4934
	CAT-GTR+4G	SR6	-3.4146	272.776	75.4335
Ponce	LG+4G	None	102.301	57.7857	178.858
	CAT-GTR+4G	None	4.10787	41.3374	142.932
	CAT-GTR+4G	Dayhoff6	-2.18521	16.1799	25.4514
	CAT-GTR+4G	KGB6	-1.01998	8.22212	29.0685
	CAT-GTR+4G	SR6	-1.86252	11.4145	27.5351
Ochoa D11	LG+4G	None	27.8436	56.6042	49.6017
	CAT-GTR+4G	None	1.78031	56.2285	55.5332
	CAT-GTR+4G	Dayhoff6	-0.391998	14.0614	13.6587
	CAT-GTR+4G	KGB6	-0.242077	32.7976	20.3679
	CAT-GTR+4G	SR6	-0.449035	27.7049	16.1347

these classifications are robust to bootstrap resampling of tRNA structural positions with the majority of plastid bootstrap replicates classifying to late-branching clades A, B1, and B23 with the median bootstrap of plastid genome groups for clade B23 above 70, except for the Glaucocystophyta genome which had 34 and 66 bootstrap replicates classifying to clade A and clade B23 respectively (Fig. 2.3,2.4,2.5,S2.1,S2.2). This suggests that the late-branching classification of plastids is driven by a consistent signal across the tRNA structure. Notably, our results are consistent with independent metabolic hypotheses which suggested that plastids originated from a starch-producing diazotrophic species (Deschamps et al., 2008; Ball et al., 2011), and with the hypothesis of a low-salinity origin and early diversification of photosynthetic eukaryotes (Blank, 2013; Sánchez-Baracaldo et al., 2017) based on a recent study where ancestral habitat types were reconstructed (Sánchez-Baracaldo et al., 2017).

Interpreting the predictions of a neural network in context to the hypotheses being tested and the system under study remains challenging (Ching et al., 2018). Beyond accurately classifying training data, to produce meaningful results the learned model linking input features to predictions should capture the structure of the data important to the hypotheses under investigation. To test if the classifications of CYANO-MLP were a result of learning the phylogenetic signal contained in cyanobacterial tRNA CIFs from clade labeled score vectors, we created permuted cyanobacterial datasets with clade labels randomly shuffled among score vectors to obliterate the phylogenetic structure of these datasets, retrained, and evaluated the performance of CYANO-MLP on these permuted datasets. For the majority of permuted datasets CYANO-MLP performed worse than randomly assigning genomes to clades (Fig. 2.2).

Moreover, the highest accuracy of a permuted dataset was 0.407 which was significantly lower than the 0.8673 accuracy obtained on the true training set. These results suggest that the classifications of CYANO-MLP are driven by the phylogenetic signal contained in tRNA CIFs. Next, the unique selective pressures experienced by organelle genomes may result in idiosyncratic score vectors not represented in the training set possibly misleading CYANO-MLP. However, this seems unlikely based on the classification of the *P. chromatophora* chromatophore genome to clade C1 (marine *Prochlorococcus/Synechococcus*) with 100% bootstrap support (Fig. 2.3) which is consistent with previous studies (Bhattacharya and Medlin, 1995; Shih et al., 2013; Sánchez-Baracaldo et al., 2017). Lastly, to determine if the classifications of plastids to clade B23 was an artifact caused by the absence of a closely related clade in the training data we scored and classified the early-branching cyanobacteria *G. lithophora* (Sánchez-Baracaldo et al., 2017; Ponce-Toledo et al., 2017) which lacks a closely related sister clade in the training data. Additionally, this allowed us to test recent hypotheses supporting *G. lithophora* as sister to plastids (Ponce-Toledo et al., 2017; Sánchez-Baracaldo et al., 2017). If plastid genome classifications were an artifact caused by lack of a closely related clade and/or were closely related to *G. lithophora* we would expect the classification results of plastids and *G. lithophora* to be similar. However, the classification of *G. lithophora* to the early-branching clade F suggests that plastids classifications are likely not an artifact and rejects a close relationship between *G. lithophora* and plastids (Fig. 2.3). Furthermore, the classification probabilities and bootstrap results of *G. lithophora* were more equivocal than other cyanobacteria and plastid genomes (Fig. 2.3) suggesting that dispersed classification probabilities and bootstrap replicates may indicate the absence of a closely related clade in the training data.

2.5.2 Phylogenetic Model Adequacy

Although limited sampling of cyanobacterial genomes and genes may have contributed to early conflicting results on the origin of plastids, several large-scale phylogenomic datasets have been analyzed that have strongly supported either a late- or early-branching position of plastids (Bhattacharya and Medlin, 1995; Turner et al., 1999; Shih et al., 2013; Criscuolo and Gribaldo, 2011; Ponce-Toledo et al., 2017; Sánchez-Baracaldo et al., 2017; Ochoa de Alda et al., 2014; Falcón et al., 2010; Blank, 2013; Dagan et al., 2013). When different phylogenomic datasets recover strongly supported, yet conflicting hypotheses about evolutionary relationships the reason is unlikely to be from a lack of information or errors caused by stochastic data sampling, but more likely a consequence of poor fitting phylogenetic models unable to adequately describe systematic biases of the data (Philippe and Roure, 2011; Blanquart and Lartillot, 2008; Sullivan and Swofford, 1997).

Comparable to previous studies, our posterior predictive analyses showed that both site-specific constraints and lineage-specific compositional biases are critical processes to model (Fig. 2.6; Table 2.4). We showed that site-specific constraints are well modeled using the CAT model (Fig. 2.6A; Table 2.4), however, the tested amino acid recoding strategies were unable to fully mitigate lineage-specific compositional biases (Fig. 2.6B; Table 2.4). Besides amino acid recoding strategies, Li et al., 2014 were effective in accommodating lineage-specific compositional biases by explicitly modeling shifts in composition along the tree, but site-specific constraints was not modeled. Our results suggest that to accurately reconstruct the position of plastids within the CyanoToL a model that can accommodate both site-specific constraints and lineage-specific compositional biases, such as, CAT-BP (Blanquart and Lartillot, 2008) is required. Unfortunately, the computational complexity of CAT-BP renders it intractable for large phylogenomic datasets. Notably, the only study, known to the authors, to accommodate both of these biases by using the CAT-GTR model and removing the most compositional divergent taxa, albeit using 16S nucleotide data, are consistent with our results supporting a late-branching position of plastids (Ochoa de Alda et al., 2014).

2.6 Supplementary Material

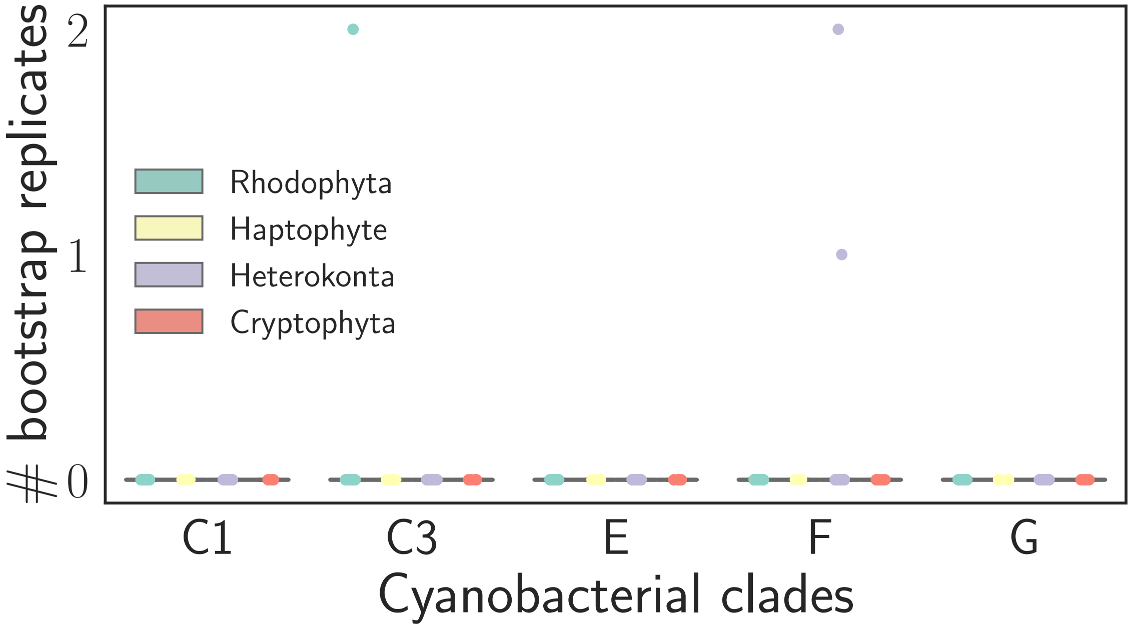

Figure S2.1: Classification results of 100 bootstrap replicates of each Rhodophyta derived plastid genome for cyanobacterial clades C1, C3, E, F, and G. Results are summarized identically to Fig. 2.4.

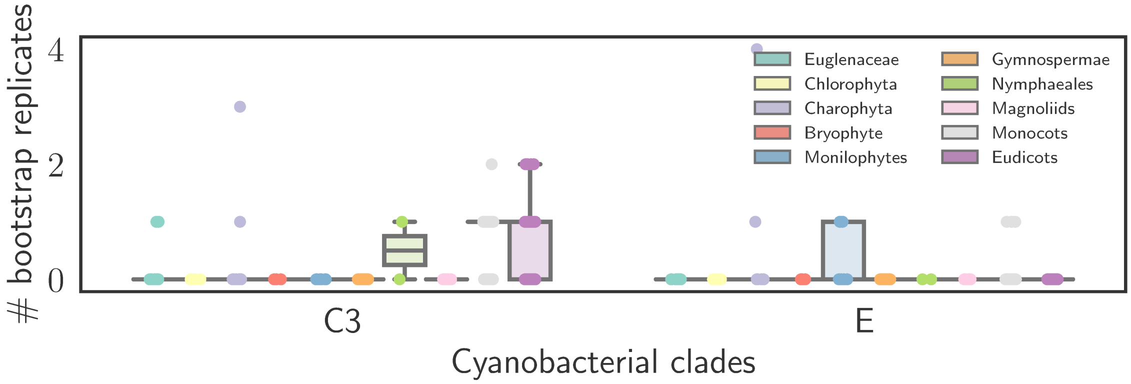

Figure S2.2: Classification results of 100 bootstrap replicates of each Chloroplastida derived plastid genome for cyanobacterial clades C3 and E. Results are summarized identically to Fig. 2.4.

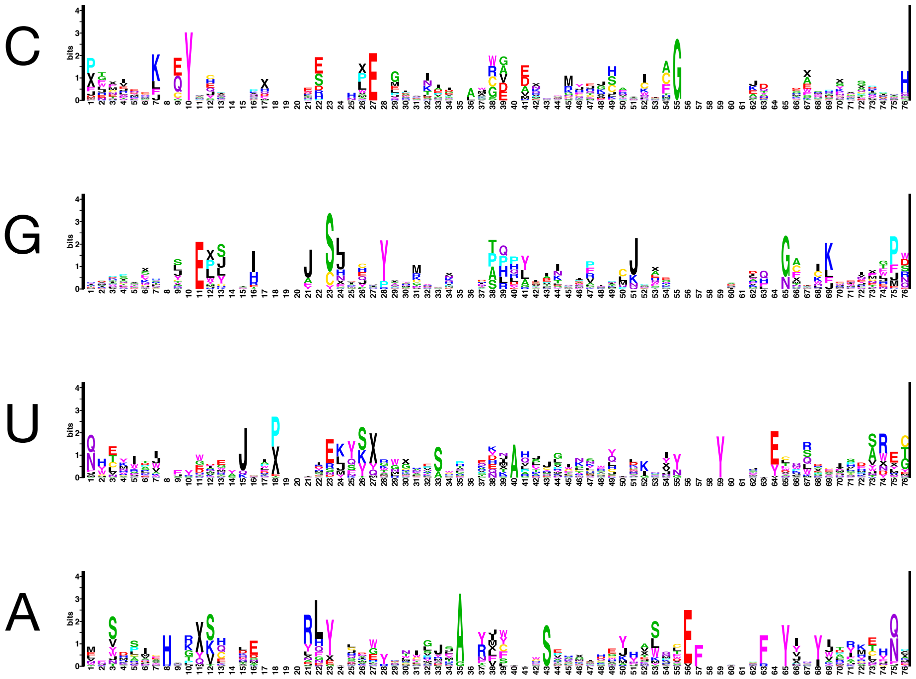

Figure S2.3: Function logos for Cyanobacterial Clade A.

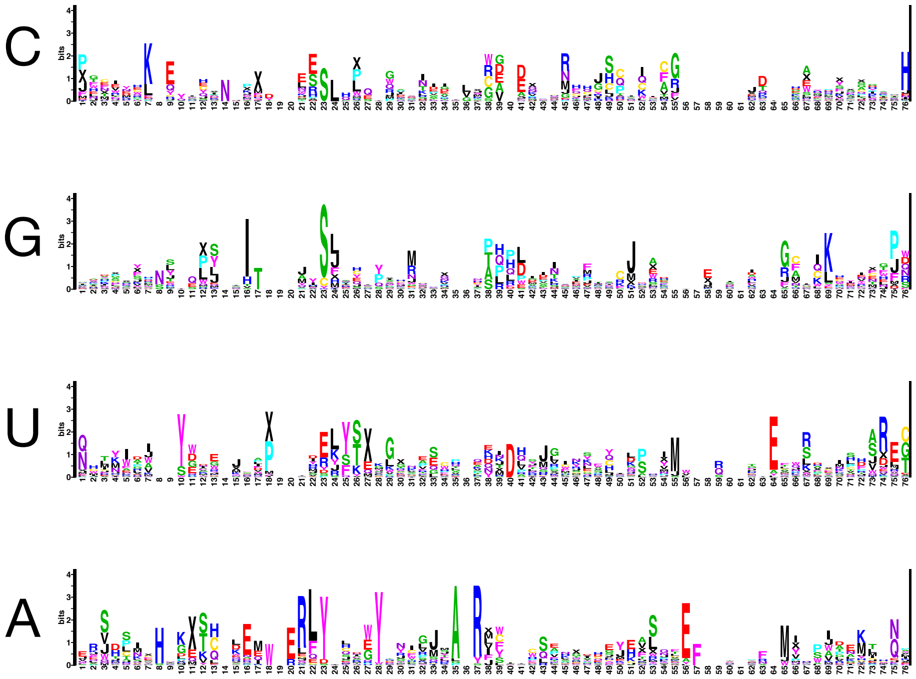

Figure S2.4: Function logos for Cyanobacterial Clade B1

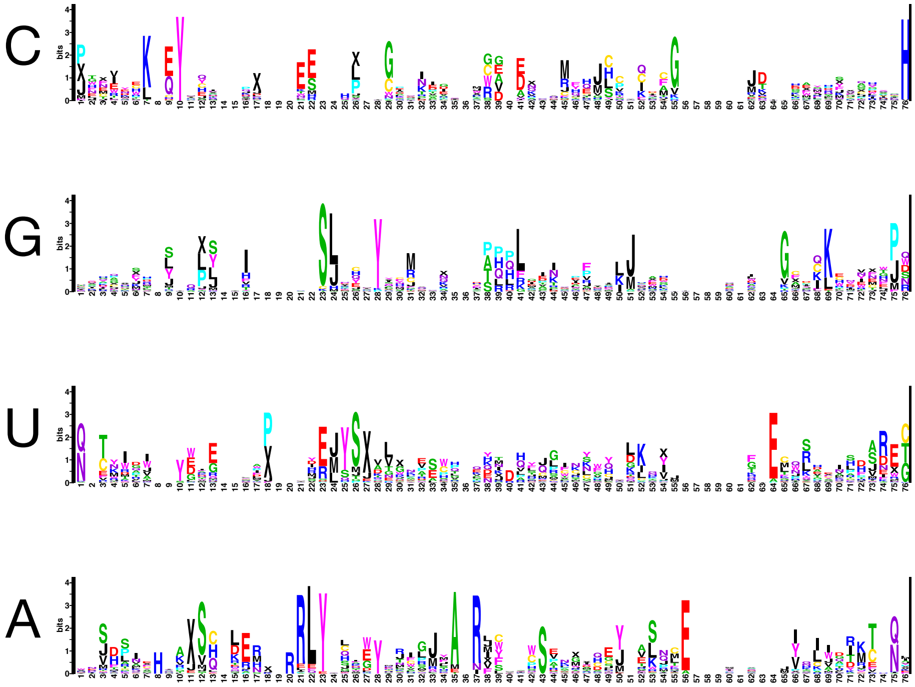

Figure S2.5: Function logos for Cyanobacterial Clade B23

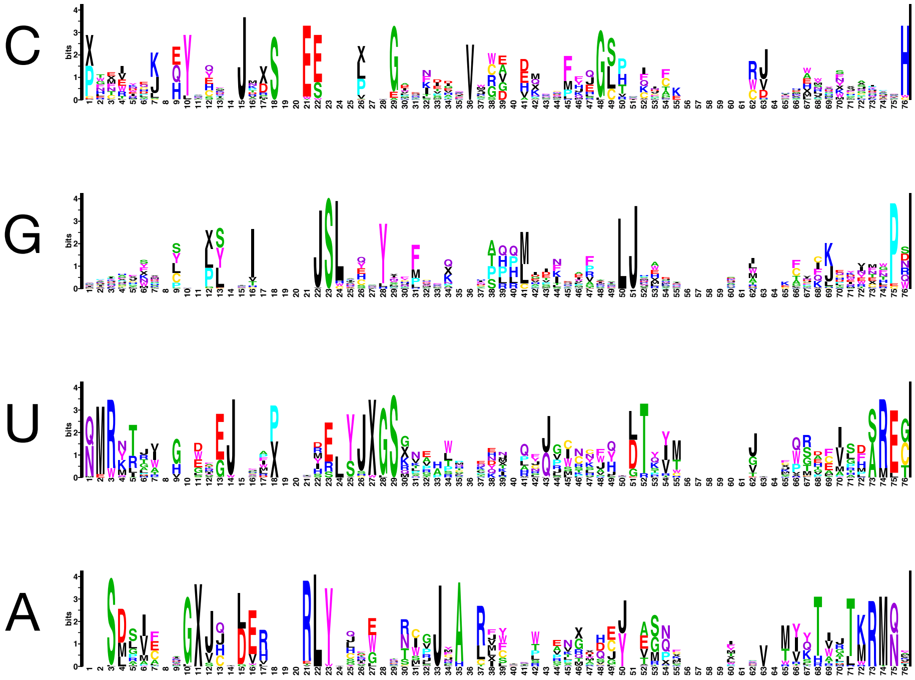

Figure S2.6: Function logos for Cyanobacterial Clade C1

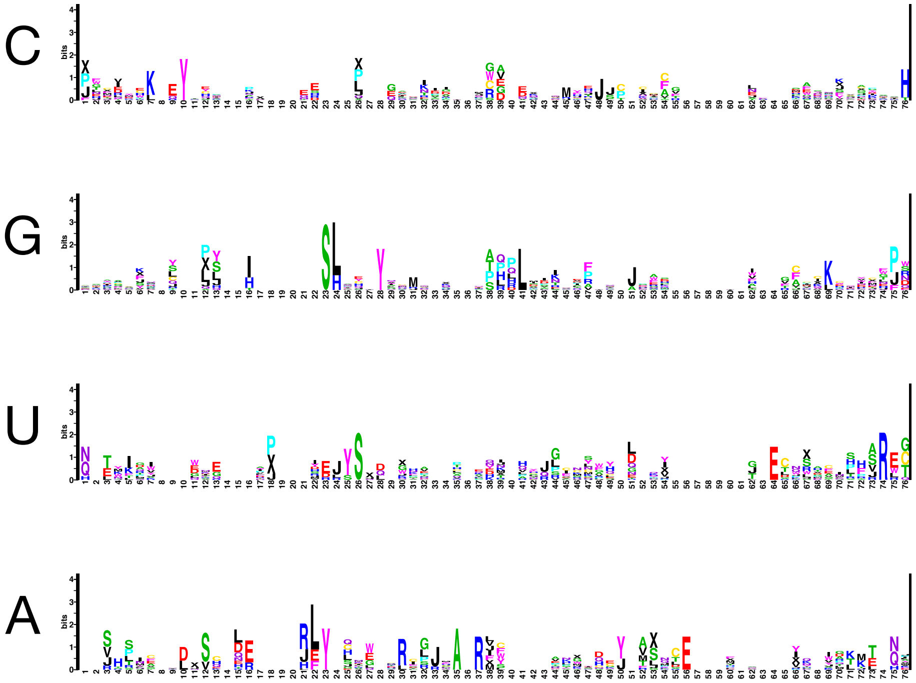

Figure S2.7: Function logos for Cyanobacterial Clade C3

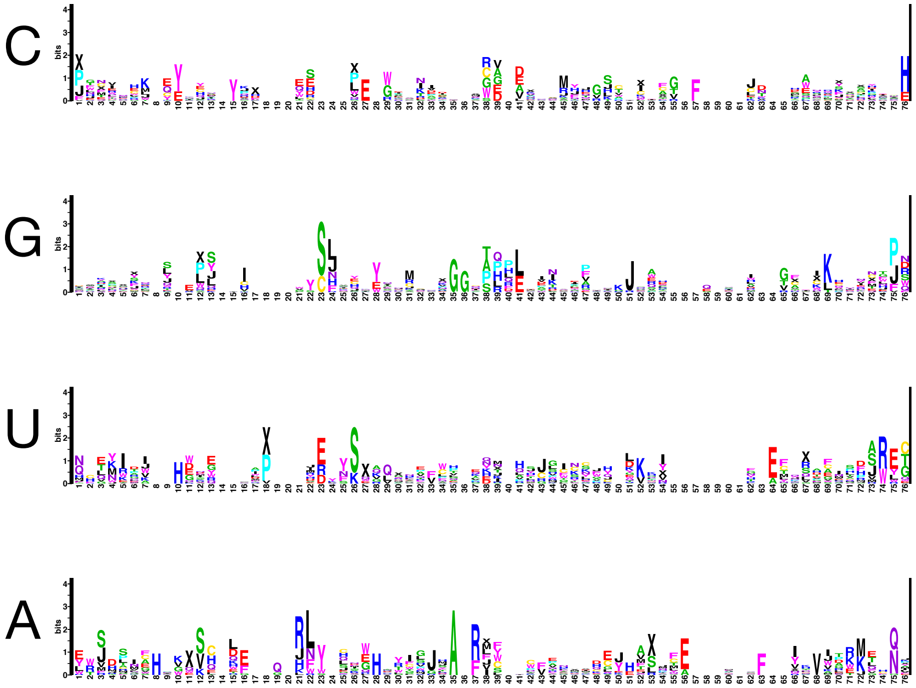

Figure S2.8: Function logos for Cyanobacterial Clade E

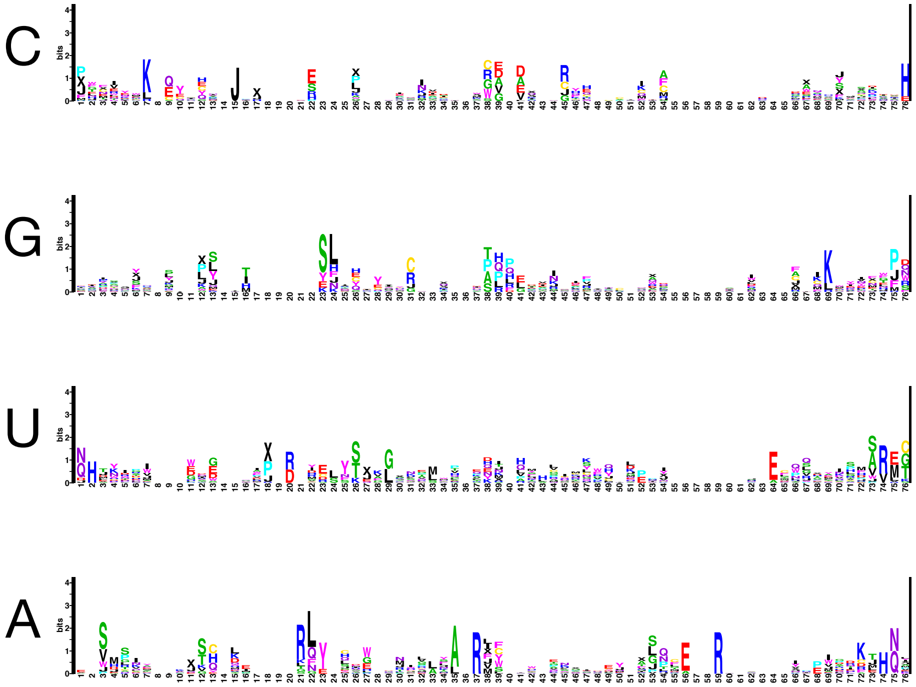

Figure S2.9: Function logos for Cyanobacterial Clade F

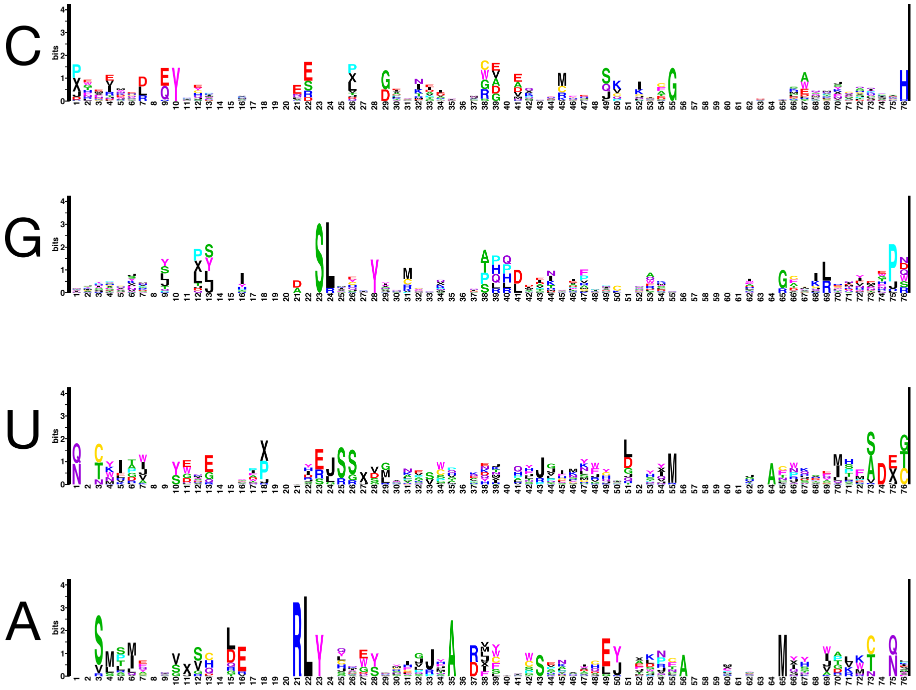

Figure S2.10: Function logos for Cyanobacterial Clade G

2.7 References

Adl, S. M. et al. (2012). The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology* **59**:5, pp. 429–493. DOI: 10.1111/j.1550-7408.2012.00644.x

Alkatib, S. et al. (2012). Evolutionary constraints on the plastid tRNA set decoding methionine and isoleucine. *Nucleic Acids Research* **40**:14, pp. 6713–24. DOI: 10.1093/nar/gks350

Amrine, K. C. H., Swingley, W. D., and Ardell, D. H. (2014). tRNA signatures reveal a polyphyletic origin of SAR11 strains among alphaproteobacteria. *PLoS Computational Biology* **10**:2. Ed. by C. A. Ouzounis, e1003454. DOI: 10.1371/journal.pcbi.1003454

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 Research* **34**:3, pp. 893–904. DOI: 10.1093/nar/gkj449

Ball, S. et al. (2011). The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *Journal of Experimental Botany* **62**:6, pp. 1775–1801. DOI: 10.1093/jxb/erq411

Bhattacharya, D. and Medlin, L. (1995). The phylogeny of plastids: a review based on comparisons of small-subunit ribosomal RNA coding regions. *Journal of Phycology* **31**:4, pp. 489–498. DOI: 10.1111/j.1529-8817.1995.tb02542.x

Blank, C. E. (2013). Origin and early evolution of photosynthetic eukaryotes in freshwater environments: reinterpreting proterozoic paleobiology and biogeochemical processes in light of trait evolution. *Journal of Phycology* **49**:6. Ed. by S. Lin, pp. 1040–1055. DOI: 10.1111/jpy.12111

Blank, C. E. and Sánchez-Baracaldo, P. (2010). Timing of morphological and ecological innovations in the cyanobacteria–A key to understanding the rise in atmospheric oxygen. *Geobiology* **8**:1, pp. 1–23. DOI: 10.1111/j.1472-4669.2009.00220.x

Blanquart, S. and Lartillot, N. (2008). A Site- and Time-Heterogeneous Model of Amino Acid Replacement. *Molecular Biology and Evolution* **25**:5, pp. 842–858. DOI: 10.1093/molbev/msn018

Ching, T. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society, Interface* **15**:141, p. 20170387. DOI: 10.1098/rsif.2017.0387

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, p. 3019. DOI: 10.1093/molbev/msr108

Dagan, T. et al. (2013). Genomes of Stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. *Genome biology and evolution* **5**:1, pp. 31–44. DOI: 10.1093/gbe/evs117

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). "A model of evolutionary change in proteins." *Atlas of Protein Sequence and Structure*, pp. 345–352

Deschamps, P. et al. (2008). Metabolic Symbiosis and the Birth of the Plant Kingdom. *Molecular Biology and Evolution* **25**:3, pp. 536–548. DOI: 10.1093/molbev/msm280

Deusch, O. et al. (2008). Genes of Cyanobacterial Origin in Plant Nuclear Genomes Point to a Heterocyst-Forming Plastid Ancestor. *Molecular Biology and Evolution* **25**:4, pp. 748–761. DOI: 10.1093/molbev/msn022

Domman, D. et al. (2015). Plastid establishment did not require a chlamydial partner. *Nature communications* **6**: p. 6421. DOI: 10.1038/ncomms7421

Dufresne, A., Garczarek, L., and Partensky, F. (2005). Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology* **6**:2, R14. DOI: 10.1186/gb-2005-6-2-r14

Dufresne, A. et al. (2003). Genome sequence of the cyanobacterium Prochlorococcus marinus SS120, a nearly minimal oxyphototrophic genome. *Proceedings of the National Academy of Sciences of the United States of America* **100**:17, pp. 10020–5. DOI: 10.1073/pnas.1733211100

Eddy, S. R. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research* **22**:11, pp. 2079–2088. DOI: 10.1093/nar/22.11.2079

Falcón, L. I., Magallón, S., and Castillo, A. (2010). Dating the cyanobacterial ancestor of the chloroplast. *The ISME Journal* **4**:6, pp. 777–783. DOI: 10.1038/ismej.2010.2

Foster, P. G. and Schultz, T. (2004). Modeling Compositional Heterogeneity. *Systematic Biology* **53**:3. Ed. by T. Schultz, pp. 485–495. DOI: 10.1080/10635150490445779

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, pp. 905–16. DOI: 10.1093/nar/gkj478

Freyhult, E. et al. (2007). New computational methods reveal tRNA identity element divergence between Proteobacteria and Cyanobacteria. *Biochimie* **89**:10, pp. 1276–88. DOI: 10.1016/j.biochi.2007.07.013

Gorodkin, J. et al. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *Bioinformatics* **13**:6, pp. 583–586. DOI: 10.1093/bioinformatics/13.6.583

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, pp. 221–224. DOI: 10.1093/molbev/msp259

Keeling, P. J. (2004). Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany* **91**:10, pp. 1481–1493. DOI: 10.3732/ajb.91.10.1481

Kenrick, P. and Crane, P. R. (1997). The origin and early evolution of plants on land. *Nature* **389**:6646, pp. 33–39. DOI: 10.1038/37918

Kosiol, C., Goldman, N., and H. Buttimore, N. (2004). A new criterion and method for amino acid classification. *Journal of Theoretical Biology* **228**:1, pp. 97–106. DOI: 10.1016/J.JTBI.2003.12.010

Lartillot, N., Brinkmann, H., and Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology* **7**:Suppl 1, S4. DOI: 10.1186/1471-2148-7-S1-S4

Lartillot, N. and Philippe, H. (2004). A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process. *Molecular Biology and Evolution* **21**:6, pp. 1095–1109. DOI: 10.1093/molbev/msh112

Lartillot, N. et al. (2013). PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment. *Systematic Biology* **62**:4, pp. 611–615. DOI: 10.1093/sysbio/syt022

Laslett, D. and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* **32**:1, pp. 11–16. DOI: 10.1093/nar/gkh152

Latysheva, N. et al. (2012). The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics* **28**:5, pp. 603–606. DOI: 10.1093/bioinformatics/bts008

Lawrence, T. J. et al. (2015). FAST: FAST Analysis of Sequences Toolbox. English. *Frontiers in Genetics* **6**: DOI: 10.3389/fgene.2015.00172

Le, S. Q. and Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution* **25**:7, pp. 1307–1320. DOI: 10.1093/molbev/msn067

Li, B. et al. (2014). Compositional Biases among Synonymous Substitutions Cause Conflict between Gene and Protein Trees for Plastid Origins. *Molecular Biology and Evolution* **31**:7, pp. 1697–1709. DOI: 10.1093/molbev/msu105

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research* **25**:5, pp. 0955–964. DOI: 10.1093/nar/25.5.0955

Manhart, J. R. and Palmer, J. D. (1990). The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. *Nature* **345**:6272, pp. 268–270. DOI: 10.1038/345268a0

Martin, W. et al. (1998). Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* **393**:6681, pp. 162–165. DOI: 10.1038/30234

Ochoa de Alda, J. A. G. et al. (2014). The plastid ancestor originated among one of the major cyanobacterial lineages. *Nature Communications* **5**: p. 4937. DOI: 10.1038/ncomms5937

Osawa, S. et al. (1992). Recent evidence for evolution of the genetic code. *Microbiological Reviews* **56**:1, pp. 229–64

Parfrey, L. W. et al. (2006). Evaluating Support for the Current Classification of Eukaryotic Diversity. *PLoS Genetics* **2**:12, e220. DOI: 10.1371/journal.pgen.0020220

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**: pp. 2825–2830

Philippe, H. and Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biology* **9**:1, p. 91. DOI: 10.1186/1741-7007-9-91

Ponce-Toledo, R. I. et al. (2017). An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. *Current Biology* **27**:3, pp. 386–391. DOI: 10.1016/J.CUB.2016.11.056

Rocap, G. et al. (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**:6952, pp. 1042–1047. DOI: 10.1038/nature01947

Sánchez-Baracaldo, P. et al. (2017). Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proceedings of the National Academy of Sciences of the United States of America* **114**:37, E7737–E7745. DOI: 10.1073/pnas.1620089114

Schirrmeister, B. E., Gugger, M., and Donoghue, P. C. J. (2015). Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. *Palaeontology* **58**:5, pp. 769–785. DOI: 10.1111/pala.12178

Shih, P. M. et al. (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **110**:3, pp. 1053–8. DOI: 10.1073/pnas.1217107110

Simon, D. et al. (2003). Phylogeny and Self-Splicing Ability of the Plastid tRNA-Leu Group I Intron. *Journal of Molecular Evolution* **57**:6, pp. 710–720. DOI: 10.1007/s00239-003-2533-3

Sprinzl, M. et al. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research* **26**:1, pp. 148–53

Sugiura, M. and Wakasugi, T. (1989). Compilation and comparison of transfer RNA genes from tobacco chloroplasts. *Critical Reviews in Plant Sciences* **8**:2, pp. 89–101. DOI: 10.1080/07352688909382271

Sullivan, J. and Swofford, D. L. (1997). Are Guinea Pigs Rodents? The Importance of Adequate Models in Molecular Phylogenetics. *Journal of Mammalian Evolution* **4**:2, pp. 77–86. DOI: 10.1023/A:1027314112438

Susko, E. and Roger, A. J. (2007). On Reduced Amino Acid Alphabets for Phylogenetic Inference. *Molecular Biology and Evolution* **24**:9, pp. 2139–2150. DOI: 10.1093/molbev/msm144

Timmis, J. N. et al. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics* **5**:2, pp. 123–135. DOI: 10.1038/nrg1271

Turner, S. et al. (1999). Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *The Journal of Eukaryotic Microbiology* **46**:4, pp. 327–338. DOI: 10.1111/j.1550-7408.1999.tb04612.x

Vogel, J., Börner, T., and Hess, W. R. (1999). Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Research* **27**:19, pp. 3866–74

Chapter 3

tRNA Class Informative Features Support Gnetophytes as Sister to Conifers

3.1 Abstract

The phylogenetic position of gnetophytes has been one of the most confounding issues in resolving relationship among seed plants. Early morphological studies supported a close relationship between flowering plants and gnetophytes. However, phylogenomic studies have suggested a close relationship to conifers supporting one of three conflicting hypotheses. The Gnepine hypothesis that supports a sister relationship with Pinaceae conifers, the Gnecup hypothesis that recovers a sister relationship with Cupressales conifers (non-Pinaceae conifers), and the Gnetifer hypothesis that supports gnetophytes as sister to all conifers. In this work, we introduce a novel approach to reconstructing phylogenies by coupling distant-based methods with an information theory distance metric to quantify differences between the evolving structure-function maps of plastid tRNA gene complements. Using the tRNA gene content from 397 plastid genomes, including representatives from all three lineages of gnetophytes, we show that the structure-function maps of plastid tRNA genes contain phylogenetic information about seed plant relationships that can be detected using distance-based methods. Furthermore, we recovered gnetophytes as sister to conifers in all our analyses providing evidence from a novel source of phylogenetic information supporting the Gnetifer hypothesis. Low bootstrap support prevents us from unequivocally supporting the Gnetifer hypotheis, but we recovered negligible support for either the Gnecup or Gnepine hypotheses.

3.2 Introduction

Extant seed plants are represented by five lineages – angiosperms (flowering plants), cycads (Cycadidae), *Ginkgo biloba*, gnetophytes, and conifers which are split into three lineages: Cupressales (Cupressaceae, Taxaceae, and Cephalotaxaceae), Araucariales (Araucariaceae and Podocarpaceae), Pinaceae. Gnetophytes are a small gymnosperm clade of about 90 species of evergreen trees, shrubs, and lianas that have been one of the most enigmatic problems in seed plant phylogenetics (Ruhfel et al., 2014; Davis et al., 2014; Burleigh and Mathews, 2004; Doyle, 2012). Early studies using morphology, first placed gnetophytes as sister or within flowering plants based on the presences of vessel elements in xylem, net-veined leaves, and the resemblance of gnetophytes' reproductive organs to simple unisexual flowers (Arber and Parkin, 1907; Arber and Parkin, 1908). Yet, later studies based on morphology supported a sister relationship or within conifers determining that vessel elements in gnetophytes and flowering plants were homoplastic, deriving from different types of vascular tissue (Eames, 1952; Bailey, 1944; Doyle, 1978). Furthermore, these studies also discovered that conifers and gnetophytes lacked scalariform pitting in their xylem, and argued that similar leaf morphology between conifers and gnetophytes was a shared derived trait (Eames, 1952; Bailey, 1944; Doyle, 1978).

The first set of phenetics studies using morphological data supported a sister relationship with gnetophytes with flowering plants (Doyle and Donoghue, 1986; Crane, 1985; Nixon et al., 1994) appearing to confirm the results of earlier morphological studies (Arber and Parkin, 1907; Arber and Parkin, 1908). However, the sister relationship of gnetophytes and flowering plants was disputed by early molecular phylogenetic studies that recovered gnetophytes as sister to all seed plants (Albert et al., 1994), sister to extant gymnosperms (Goremykin et al., 1996), or sister to conifers (Qiu et al., 1999; Chaw et al., 2000). Moreover, subsequent phylogenetic studies have been unsuccessful in eliminating uncertainty about the phylogenetic position of gnetophytes recovering conflicting topologies depending on the phylogenetic tree estimation methods and phylogenomic dataset (Bowe et al., 2000; Hajibabaei et al., 2006; Mathews, 2009; Wang and Ran, 2014; Wan et al., 2018). This inconsistency between phylogenomic datasets has resulted in five conflicting topologies being supported: 1) the Gnecup hypothesis that supports gnetophytes as sister to Cupressales conifers (Nickrent et al., 2000; Doyle, 2009; Xi et al., 2013; Ruhfel et al., 2014), 2) the Gnepine hypothesis that supports a sister relationship between gnetophytes and Pinaceae conifers (Bowe et al., 2000; Gugerli et al., 2001; Soltis et al., 2002; Burleigh and Mathews, 2004; Hajibabaei et al., 2006; Xi et al., 2013), 3) the Gnetifer hypothesis that supports gnetophytes as sister to conifers (Wickett et al., 2014), 4) sister to gymnosperms (McCoy et al., 2008; Lee et al., 2011; Li et al., 2017), or 5) sister to all seed plants (Li et al., 2017). However, recent studies and reanalysis of previous studies have converged on hypotheses supporting a close relationship between gnetophytes and conifers (Wickett et al., 2014; Wan et al., 2018). Phylogenetic inference of gnetophytes within seed plants is complicated by deep divergences between seed plant clades (Lu et al., 2014), elevated rates of nucleotide substitution within gnetophytes leading to substitutional saturation (Zhong et al., 2010; Wickett et al., 2014; Wan et al., 2018), and loss of species diversity within gymnosperms and especially within gnetophytes (Lu et al., 2014).

Here we introduce a novel approach to reconstructing phylogenies by coupling distance-based reconstruction algorithms with an information theory distance metric to quantify differences between the evolving structure-function maps of plastid tRNA gene complements. tRNAs engaged in protein synthesis must interact productively with their cognate aminoacyl tRNA-synthetase (aaRS) to be charged with the correct amino acid and must avoid interacting productively with other aaRSs to ensure accurate translations of the genetic code. These functional interactions are determined by a set of structural features called identity determinants, which promote interaction with its cognate aaaRS, and anti-determinants that discriminate against noncognate aaRSs (Giegé et al., 1998). We have previously shown that these identity determinants, which we call Class-Informative Features (CIFs) (Freyhult et al., 2006), slowly diverge in a phylogenetically informative manner (Amrine et al., 2014) making them an ideal phylogenetic marker for reconstructing deep relationships. Using the tRNA gene content of 379 seed plant and 18 fern (monilophytes) plastid genomes with our novel distance-based marker we found support for the Gnetifer hypothesis placing gnetophytes sister to conifers. Notably, we found negligible support for either the Gnecup or Gnepine hypotheses.

3.3 Methods

3.3.1 tRNA data

We downloaded 397 plastid genomes containing 18 Monilophytes, 11 Cycadidae, 27 Pinaceae, 24 Cupressales, 7 Araucariales, 7 gnetophytes, 191 Eudicot, and 112 Monocot genomes from NCBI. The plastid genomes included representatives from all three lineages of gnetophytes. For every genome we annotated tRNA genes as the union of predictions from tRNAscan-SE v1.31 (Lowe and Eddy, 1997) in organelle mode and ARAGORN v1.2.36 (Laslett and Canback, 2004) with the option to detect tRNAs with conical introns. Predictions by ARAGORN v1.2.36 (Laslett and Canback, 2004) containing introns in tRNA isotypes that have not been previously described to contain introns (Manhart and Palmer, 1990; Vogel et al., 1999; Simon et al., 2003) were discarded as likely false positives. Finally, tRNA genes containing anticodons identified as absent in most land plant plastids (Osawa et al., 1992; Sugiura and Wakasugi, 1989; Alkatib et al., 2012) were discarded.

We annotated tRNAs containing the CAU anticodon as initiator tRNA $^{\mathrm{Met}}$, elongator tRNA $^{\mathrm{Met}}$, or tRNA $^{\mathrm{Ile}}$ $_{\mathrm{CAU}}$ using TFAM v1.4 (Ardell and Andersson, 2006) with the covariance model used in Amrine et al., 2014. We aligned tRNA sequences using COVEA v2.4.4 (Eddy and Durbin, 1994) using the prokaryotic tRNA covariance model in Lowe and Eddy, 1997. We mapped sites to Sprinzl coordinates (Sprinzl et al., 1998) and removed the variable arm, CCA tail, and sites not mapping to a Sprinzl coordinate manually using Seaview v4.6.1 (Gouy et al., 2010).

3.3.2 tRNA CIF estimation

Plastid tRNAs were partitioned into sets X based on current species circumscriptions, where $X = \{Monilophytes, Cycadidae, Pinaceae, Cupressales, Araucariales, gnetophytes, Eudicots, Monocots\}$. The consolidation of tRNA genes into these partitions for tRNA CIFs estimation is justified for three reasons. First, it has been shown that tRNA CIFs diverge in a phylogenetically informative way (Amrine et al., 2014; Freyhult et al., 2007) suggesting that individual genomes in each partition should carry similar information. Second, the molecular information calculations used in function logos weights tRNA CIFs based on their conservation resulting in the up-weighting of tRNA CIFs that are consistent across genomes within the partition. Lastly, consolidation is required because approximately 120 tRNAs are required to estimate informative function logos.

To estimate tRNA CIFs we adopt the molecular information theory approach of Freyhult et al., 2006, which we briefly describe here. For the purpose of calculating tRNA CIFs the possible functional classes of tRNAs are denoted by their IUPAC one-letter amino acid code, $Z \equiv \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y\}$ and features F is the Cartesian product $n_i \in \{A, C, G, U\} \times SC$, where n_i is the set of possible nucleotides and SC is the set of Sprinzl Coordinates (Sprinzl et al., 1998). The functional information $I^X_{f_i}(Z|f_i)$ that feature $f_i \in F$ confers about the frequencies of different functional classes Z for clade X is defined as: $I^X_{f_i}(Z|f_i) = H^X(Z) - e(n(f_i)) - H^X(Z|f_i))$, where $H^X(Z|f_i) = -\Sigma_{z \in Z} p(z|f_i) log_2(p(z|f_i))$ is the class entropy or level of uncertainty about the functional class of tRNAs with feature f_i for clade X. $H^X(Z) = -\Sigma_{z \in Z} p(z) log_2(p(z))$ is the background entropy for clade X which depends on relative frequency of sequences belonging to different functional classes, and by definition: $0 \leq p(z), p(z|f_i) \leq 1, \Sigma_{z \in Z} p(z) = 1, \Sigma_{z \in Z} p(z|f_i) = 1$. $e(n(f_i)))$ is a correction factor for biases caused by small sample size which is calculated exactly for sample sizes ≤ 10 following the method of Schneider and Stephens, 1990 and estimated for sample sizes > 10 using the method of Nemenman et al., 2004. Finally, the proportion of functional information $I^X_{f_i}(Z|f_i)$ attributed to each functional class $z \in Z$ is calculated using Gorodkin heights (Gorodkin et al., 1997) where $h^z_{f_i} = (p(z|f_i)/p(z))/\Sigma_{w \in Z} p(w|f_i)/p(w))$, $h^X_{f_i}$ is the vector of Gorodkin heights $\forall z \in Z$ at feature f_i for clade X, and by definition $\Sigma h^X_{f_i} = 1$. The program tsfm v0.9.10 (https://github.com/tlawrence3/tsfm) was used to perform these calculations and to generate function logo graphics. The `-x 10` and `--entropy NSB` options were provided to tsfm to calculate the exact sample correct sample sizes ≤ 10 and to use the NSB estimator for samples sizes > 10.

3.3.3 Distance calculation

To produce a pairwise distance matrix for clades $x_{(j,k)} \in X$ we used a modified version of the Jensen-Shannon distance (JSD), an information theory metric for quantifying similarity between two probability distributions (Endres and Schindelin, 2003). We weighted the JSD calculation by the sum of the functional information $I^x_{f_i}$ of the feature f_i for clades $x_{(j,k)}$ and summed over all features $f_i \in F$ (equation 3.1). The function $H()$ in equation 3.1 is the Shannon entropy (Shannon, 1948) defined as: $H(V) = -\sum_i^n p(v_i) log_2(p(v_i))$.

$$\sum_{f_i \in F} \left(I_{f_i}^{x_j} + I_{f_i}^{x_k}\right) \sqrt{H\left(\sum_{m \in \{j,k\}} I_{f_i}^{x_m} h_{f_i}^{x_m}\right) - \sum_{m \in \{j,k\}} I_{f_i}^{x_m} H\left(h_{f_i}^{x_m}\right)} \tag{3.1}$$

In addition to the pairwise distance matrix using all features $f_i \in F$ we produced 100 bootstrap replicates by subsampling f_i with replacement.

3.3.4 Phylogenetic analysis

We reconstructed distance-based phylogenetic trees using three different algrorithms: neighbor-joining (Saitou and Nei, 1987), BIONJ (Gascuel, 1997), and minimum evolution (Rzhetsky and Nei, 1992; Rzhetsky and Nei, 1993). Monilophytes were used as the outgroup to root phylogenetic trees. We used the R package ape v5.1 (Paradis et al., 2004) in R 3.4.3 (R Core Team, 2017) to perform all phylogenetic estimations. Bootstrap replicates were summarized and mapped onto the phylogenetic tree estimated using the full dataset with the sumtrees.py utility provided in DendroPy 4.4.0 (Sukumaran and Holder, 2010).

3.4 Results

3.4.1 tRNA data and CIF estimation

We annotated and extracted 13,326 tRNA gene sequences from the 397 plastid genomes from an average of 31.74 tRNA genes per genome (T_g). The annotated and extracted tRNA genes were distributed as follows: 535 (29.72 T_g) Monilophytes, 371 (33.73 T_g) Cycadidae, 777 (28.78 T_g) Pinaceae, 671 (27.96 T_g) Cupressales, 206 (29.43 T_g) Araucariales, 243 (34.71 T_g) gnetophytes, 6,593 (34.52 T_g) Eudicots, and 3,930 (35.10 T_g) Monocots (Table 3.1). We estimated informative function logos for Monilophytes, Cycadidae, Pinaceae, Cupressales, Araucariales, gnetophytes, Eudicots, Monocots (Fig. 3.1-3.4).

3.4.2 Phylogenetic analysis

BIONJ, neighbor-joining, and minimum evolution methods recovered the same topology with slight differences in bootstrap support (Fig. 3.6-3.7). We recovered strongly supported angiosperm and gymnosperm clades in all three analyses. Additionally, all analyses recovered a moderately supported conifer clade. Gnetophytes were recovered as sister to conifers in the BIONJ, neighbor-joining, and minimum evolution analyses with 51, 46, and 44 bootstrap support respectively (Fig. 3.6-3.7; Table 3.2). The

Table 3.1: Summary statistics of the genomes and tRNA genes. Statistics include the number of genomes (G), the number of tRNAs genes (T), average number of tRNA genes per genome (T_g), total number of nucleotides contained in tRNA genes (N), average number of nucleotides per tRNA gene (N/T), the percent of adenine (%A), thymine (%T), guanine (%G), and cytosine (%C) nucleotides contained in tRNA genes.

Clade	G	T	T/G	N	N/T	%A	%T	%G	%C
Monilophytes	18	535	29.72	38,886	72.68	21.4	24.1	29.9	24.7
Cycadidae	11	371	33.73	26,901	72.51	21.1	24.2	30.0	24.7
Pinaceae	27	777	28.78	56,373	72.55	22.2	24.0	29.1	24.7
Cupressales	24	671	27.96	48,690	72.56	22.1	24.7	29.4	23.8
Araucariales	7	206	29.43	14,940	72.52	21.8	24.7	29.5	24.0
gnetophytes	7	243	34.71	17,640	72.59	21.8	24.3	29.4	24.5
Eudicots	191	6,593	34.52	478,337	72.55	21.4	25.2	29.7	23.7
Monocots	112	3,930	35.10	285,302	72.60	21.7	25.2	29.5	23.7

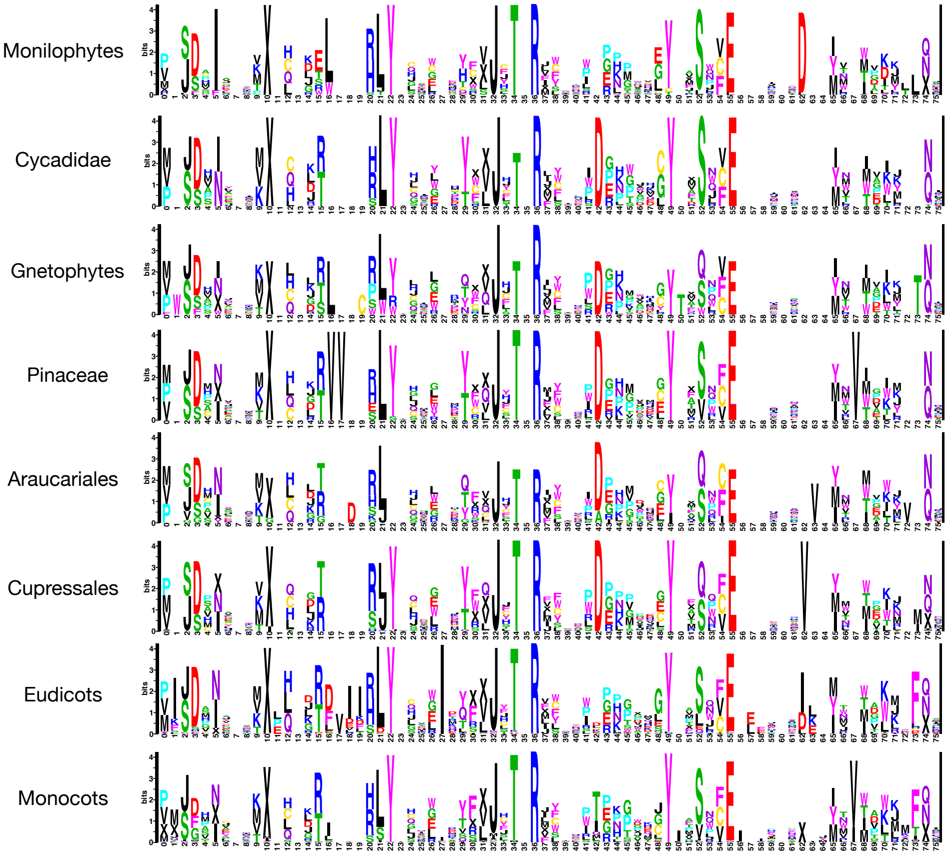

Figure 3.1: Adenine function logo for each plant clade. The y-axis of each function logo is the functional information for each position in the tRNA (x-axis) measured in bits. Stacked letters are amino acid IUPAC one-letter codes. The height of each letter represents the amount of functional information attributed to that amino acid for each position.

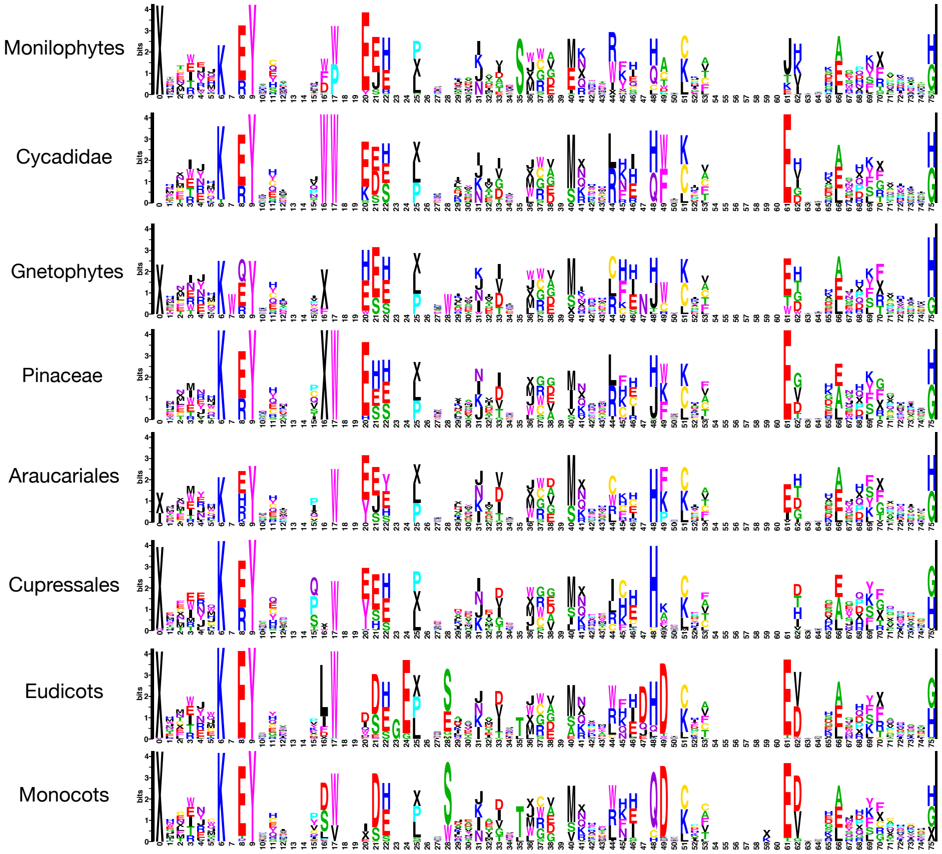

Figure 3.2: Cytosine function logo for each plant clade. The y-axis of each function logo is the functional information for each position in the tRNA (x-axis) measured in bits. Stacked letters are amino acid IUPAC one-letter codes. The height of each letter represents the amount of functional information attributed to that amino acid for each position.

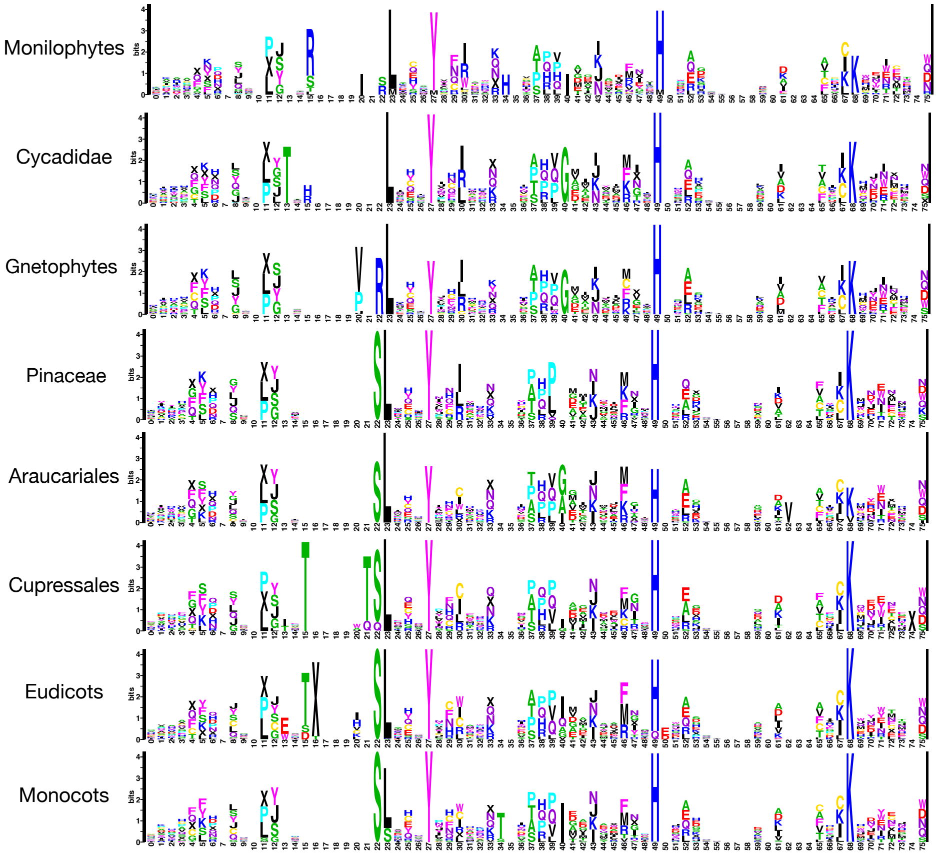

Figure 3.3: Guanine function logo for each plant clade. The y-axis of each function logo is the functional information for each position in the tRNA (x-axis) measured in bits. Stacked letters are amino acid IUPAC one-letter codes. The height of each letter represents the amount of functional information attributed to that amino acid for each position.

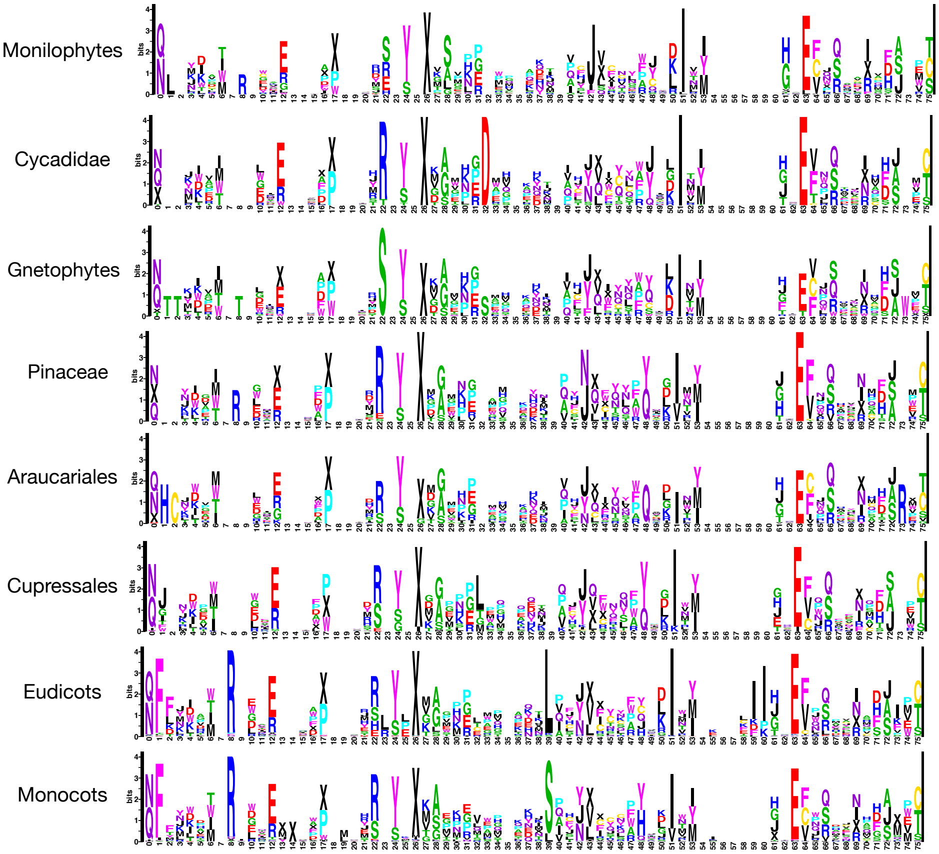

Figure 3.4: Uracil function logo for each plant clade. The y-axis of each function logo is the functional information for each position in the tRNA (x-axis) measured in bits. Stacked letters are amino acid IUPAC one-letter codes. The height of each letter represents the amount of functional information attributed to that amino acid for each position.

Table 3.2: Bootstrap replicate rooted clade frequencies. for neighbor-joining (NJ), BIONJ, and minimum evolution (ME) analyses. G - gnetophytes, P - Pinaceae, Cu - Cupressales, A - Araucariales, Cy - Cycadidae.

Bipartition	NJ	BIONJ	ME
(G, (P, (Cu, A))	46	51	44
(G, Cy)	34	34	42
(G, (Cu, A))	11	0	0
(Cu, G)	2	2	2
(P, G)	1	0	5

second most frequent bipartition in the bootstrap replicates recovered gnetophytes and cycads as sister possibly an artefact caused by the absence of *Ginkgo biloba* or the elevated rates of nucleotide substitution of gnetophytes' plastid genomes (McCoy et al., 2008) affecting the rate of tRNA CIF turnover. Notably, there was minimal bootstrap support for either the Gnepine or Gnecup hypotheses (Table 3.2).

3.5 Discussion

The exact phylogenetic relationship of gnetophytes has remained one of the most perplexing issues remaining in seed plant phylogenetics (Wickett et al., 2014; Wan et al., 2018). Using plastid tRNA CIFs, coupled with distance-based phylogenetic reconstruction methods we found support for gnetophytes as sister to conifers (Gnetifer hypothesis). Low bootstrap support of the Gnetifer hypothesis prevents us from unequivocally resolving the root of gnetophytes within seed plants, however, we found negligible support for either the Gnecup or Gnepine hypotheses. These results are similar to those of Wickett et al., 2014 that found support for the Gnetifer hypothesis using their full dataset. However, most single gene trees did not strongly support Gnetifer hypothesis, but did strongly support the monophyly of conifers rejecting both the Gnecup and and Gnepine hypotheses (Wickett et al., 2014). The lack of robust support for the phylogenetic placement of gnetophytes has been attributed to the several extinctions followed by rapid radiations experienced by conifers and gnetophytes resulting in short internal branches between conifer clades and long terminal branches (Wang and Ran, 2014). This pattern of diversification may result in limited characters reflecting the internal branching topology possibly explaining the low statistical support. Additionally, long terminal branches may cause substitutional saturation, which has been shown to cause support for the Gnecup hypothesis (Zhong et al., 2010). Our results show a similar topology of short internal branches and long terminal branches (Fig. 3.5-3.7) possibly explaining the low bootstrap support for these nodes in our trees. However, it seems unlikely although not impossible, based on combinatorial reasoning, that long terminal branches would result in saturation of tRNA CIFs leading to similar artefacts in tree reconstruction. On the other hand, the

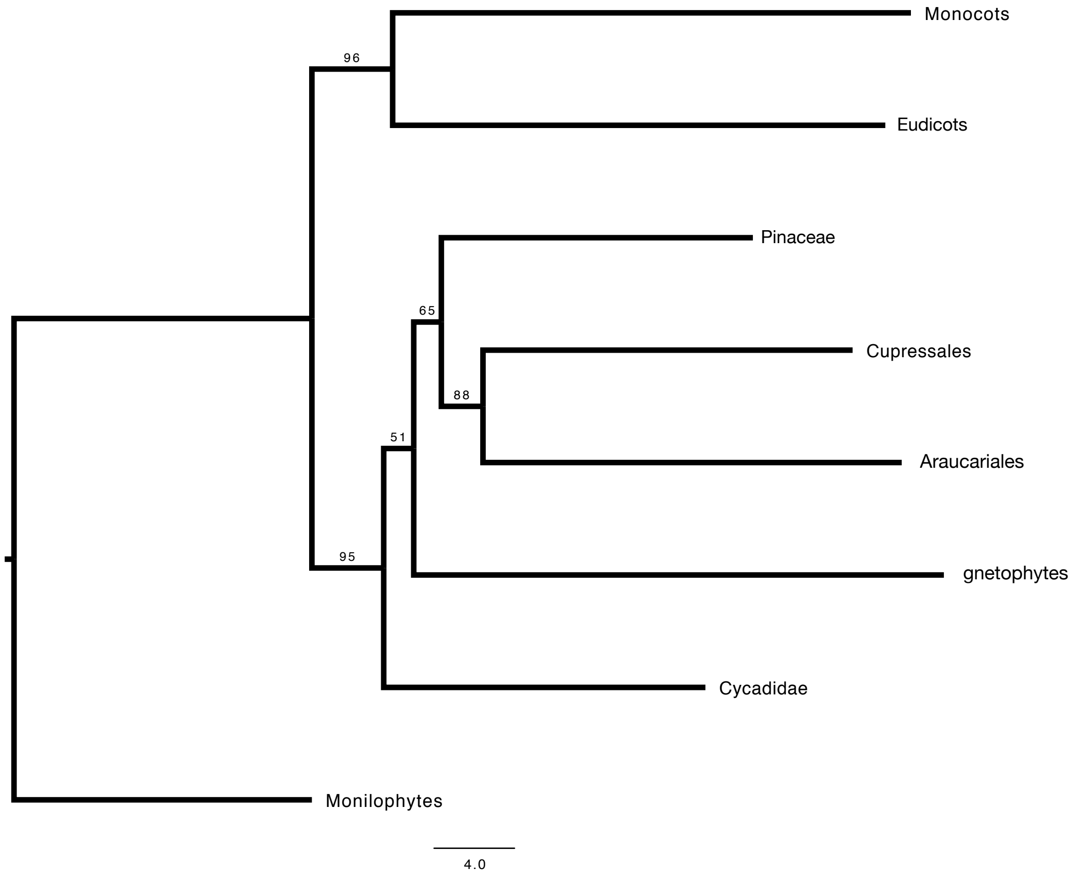

Figure 3.5: BIONJ tree annotated with bootstrap support. Branch lengths represent Jensen-Shannon distances.

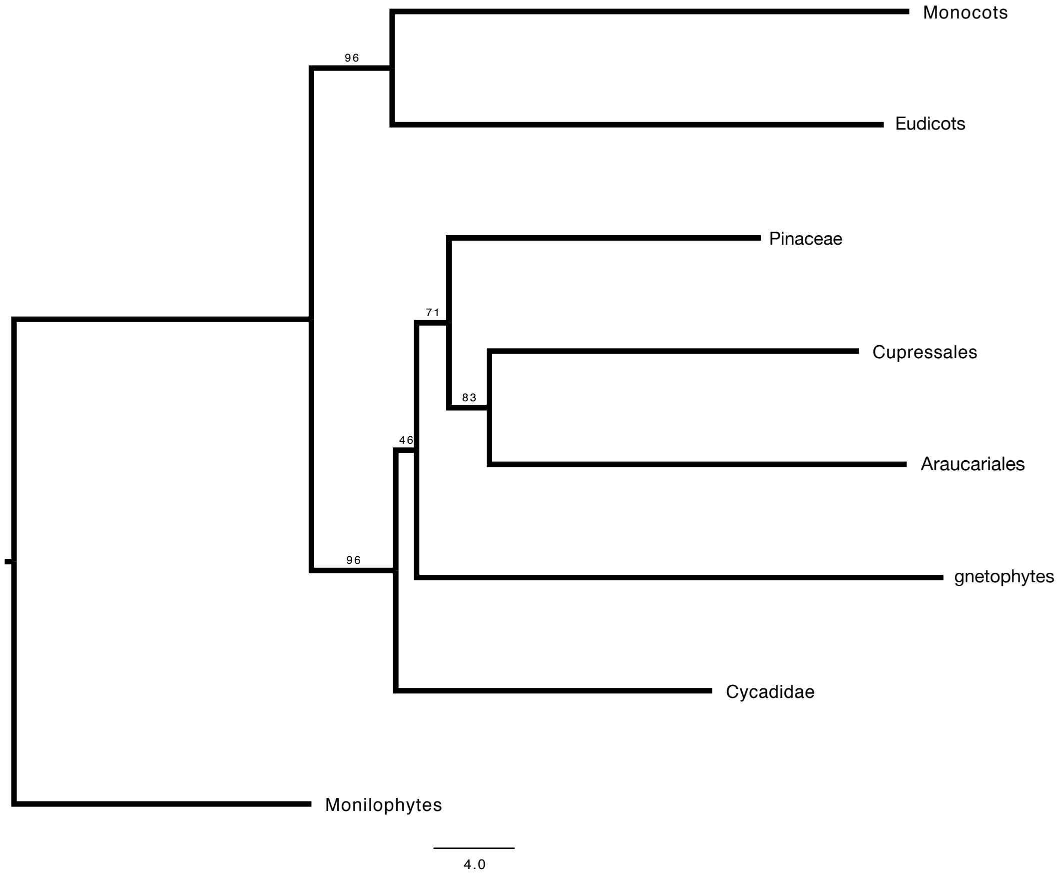

Figure 3.6: Neighbor-joining tree annotated with bootstrap support. Branch lengths represent Jensen-Shannon distances.

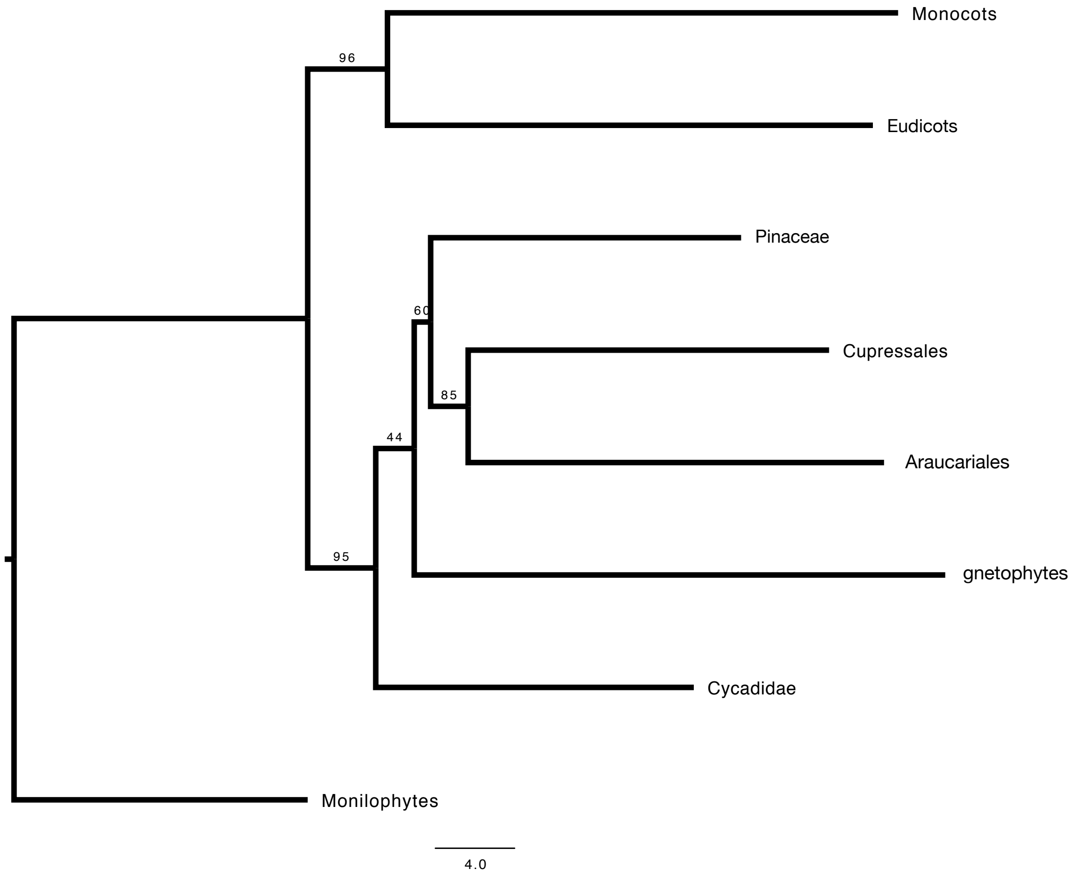

Figure 3.7: Minimum evolution tree annotated with bootstrap support. Branch lengths represent Jensen-Shannon distances.

second most common bootstrap topology in our analyses grouped cycads and gnetophytes as sister to the remaining gymnosperms (Table 3.2), which is most certainly incorrect. This may be a systematic bias caused by assuming independence of tRNA CIFs, which work in concert to ensure proper charging of tRNAs. Additionally, this could be the result of excluding *G. biloba*, which could not be included because the limited number of plastid tRNA genes did not allow for the estimation of tRNA CIFs.

The persistent uncertainty of the phylogenetic results for the placement of gnetophytes has motivated consideration of rare genomic events, which have been mainly interpreted as support for the Gnepine hypothesis. The most notable has been the loss of all 11 plastid *ndh* genes in gnetophytes and Pinaceae, which has been considered a synapomorphy supporting the Gnepine hypothesis (Braukmann et al., 2009). However, different patterns of loss of plastid- and nuclear-encoded components of the NDH complex in gnetophytes and Pinaceae suggests parallel loss of *ndh* genes (McCoy et al., 2008; Wu et al., 2011; Ruhlman et al., 2015), which is compatibility with both the Gnetifer or Gnepine hypotheses. Moreover, the loss of the plastid *ndh* gene complex is not uncommon in seed plants with multiple independent losses in Orchidaceae (Kim and Chase, 2017) and Geraniales (Ruhlman et al., 2015) further suggesting parallel loss in gnetophytes and Pinaceae.

3.6 Conclusion

Ours results provide additional evidence from a novel source of phylogenetic information supporting the Gnetifer hypothesis. However, low bootstrap support, similar to other studies (Zhong et al., 2010; Wickett et al., 2014), prevents us from making strong conclusions about the placement of gnetophytes. Future studies using tRNA CIFs to resolve the phylogenetic placement of gnetopytes would likely benefit from the inclusion of *G. biloba* by using nuclear encoded tRNA sequences, which are more numerous. However, currently there are no Cupressales nuclear genomes available which precludes testing between Gnecup, Gnepine, or Gnetifer hypotheses.

3.7 References

Albert, V. A. et al. (1994). Functional Constraints and rbcL Evidence for Land Plant Phylogeny. *Annals of the Missouri Botanical Garden* **81**:3, pp. 534–567. DOI: 10.2307/2399902

Alkatib, S. et al. (2012). Evolutionary constraints on the plastid tRNA set decoding methionine and isoleucine. *Nucleic Acids Research* **40**:14, pp. 6713–24. DOI: 10.1093/nar/gks350

Amrine, K. C. H., Swingley, W. D., and Ardell, D. H. (2014). tRNA signatures reveal a polyphyletic origin of SAR11 strains among alphaproteobacteria. *PLoS Computational Biology* **10**:2. Ed. by C. A. Ouzounis, e1003454. DOI: 10.1371/journal.pcbi.1003454

Arber, E. A. N. and Parkin, J. (1907). On the Origin of Angiosperms. *Journal of the Linnean Society of London, Botany* **38**:263, pp. 29–80. DOI: 10.1111/j.1095-8339.1907.tb01074.x

Arber, E. and Parkin, N. J. (1908). Studies on the Evolution of the Angiosperms. The Relationship of the Angiosperms to the Gnetales. *Annals of Botany* **22**: pp. 489–515

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 Research* **34**:3, pp. 893–904. DOI: 10.1093/nar/gkj449

Bailey, I. W. (1944). The Development of Vessels in Angiosperms and its Significance in Morphological Research. *American Journal of Botany* **31**:7, pp. 421–428. DOI: 10.2307/2437302

Bowe, L. M. et al. (2000). Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proceedings of the National Academy of Sciences of the United States of America* **97**:8, pp. 4092–7. DOI: 10.1073/pnas.97.8.4092

Braukmann, T. W. A., Kuzmina, M., and Stefanović, S. (2009). Loss of all plastid ndh genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. *Current Genetics* **55**:3, pp. 323–337. DOI: 10.1007/s00294-009-0249-7

Burleigh, J. G. and Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. *American journal of botany* **91**:10, pp. 1599–613. DOI: 10.3732/ajb.91.10.1599

Chaw, S. M. et al. (2000). Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. *Proceedings of the National Academy of Sciences of the United States of America* **97**:8, pp. 4086–91. DOI: 10.1073/PNAS.97.8.4086

Crane, P. R. (1985). Phylogenetic Analysis of Seed Plants and the Origin of Angiosperms. *Annals of the Missouri Botanical Garden* **72**:4, pp. 716–793. DOI: 10.2307/2399221

Davis, C. C., Xi, Z., and Mathews, S. (2014). Plastid phylogenomics and green plant phylogeny: almost full circle but not quite there. *BMC Biology* **12**:1, p. 11. DOI: 10.1186/1741-7007-12-11

Doyle, J. A. (1978). Origin of Angiosperms. *Annual Review of Ecology and Systematics*, pp. 365–392

Doyle, J. A. (2009). Seed ferns and the origin of angiosperms. *http://dx.doi.org/10.3159/1095-5674(2006)133[169:SFATOO]2.0.CO;2*. DOI: 10.3159/1095-5674(2006)133[169:SFATOO]2.0.CO;2

Doyle, J. A. (2012). Molecular and Fossil Evidence on the Origin of Angiosperms. *Annual Review of Earth and Planetary Sciences* **40**:1, pp. 301–326. DOI: 10.1146/annurev-earth-042711-105313

Doyle, J. A. and Donoghue, M. J. (1986). Seed plant phylogeny and the origin of angiosperms: An experimental cladistic approach. English. *The Botanical Review* **52**:4, pp. 321–431. DOI: 10.1007/BF02861082

Eames, A. J. (1952). Relationships of the Ephedrales. *Phytomorph* **2**: pp. 79–100

Eddy, S. R. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research* **22**:11, pp. 2079–2088. DOI: 10.1093/nar/22.11.2079

Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory* **49**:7, pp. 1858–1860. DOI: 10.1109/TIT.2003.813506

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, pp. 905–16. DOI: 10.1093/nar/gkj478

Freyhult, E. et al. (2007). New computational methods reveal tRNA identity element divergence between Proteobacteria and Cyanobacteria. *Biochimie* **89**:10, pp. 1276–88. DOI: 10.1016/j.biochi.2007.07.013

Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**:7, pp. 685–695. DOI: 10.1093/oxfordjournals.molbev.a025808

Giegé, R., Sissler, M., and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research* **26**:22, pp. 5017–35

Goremykin, V. et al. (1996). Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to rbcL data do not support gnetalean affinities of angiosperms. *Molecular biology and evolution* **13**:2, pp. 383–96

Gorodkin, J. et al. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *Bioinformatics* **13**:6, pp. 583–586. DOI: 10.1093/bioinformatics/13.6.583

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, pp. 221–224. DOI: 10.1093/molbev/msp259

Gugerli, F. et al. (2001). The evolutionary split of Pinaceae from other conifers: evidence from an intron loss and a multigene phylogeny. *Molecular phylogenetics and evolution* **21**:2, pp. 167–75. DOI: 10.1006/mpev.2001.1004

Hajibabaei, M., Xia, J., and Drouin, G. (2006). Seed plant phylogeny: Gnetophytes are derived conifers and a sister group to Pinaceae. *Molecular Phylogenetics and Evolution* **40**:1, pp. 208–217. DOI: 10.1016/J.YMPEV.2006.03.006

Kim, H. T. and Chase, M. W. (2017). Independent degradation in genes of the plastid ndh gene family in species of the orchid genus Cymbidium (Orchidaceae; Epidendroideae). *PloS one* **12**:11, e0187318. DOI: 10.1371/journal.pone.0187318

Laslett, D. and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* **32**:1, pp. 11–16. DOI: 10.1093/nar/gkh152

Lee, E. K. et al. (2011). A Functional Phylogenomic View of the Seed Plants. *PLoS Genetics* **7**:12. Ed. by M. J. Sanderson, e1002411. DOI: 10.1371/journal.pgen.1002411

Li, Z. et al. (2017). Single-Copy Genes as Molecular Markers for Phylogenomic Studies in Seed Plants. *Genome Biology and Evolution* **9**:5, pp. 1130–1147. DOI: 10.1093/gbe/evx070

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research* **25**:5, pp. 0955–964. DOI: 10.1093/nar/25.5.0955

Lu, Y. et al. (2014). Phylogeny and Divergence Times of Gymnosperms Inferred from Single-Copy Nuclear Genes. *PLoS ONE* **9**:9. Ed. by S. Buerki, e107679. DOI: 10.1371/journal.pone.0107679

Manhart, J. R. and Palmer, J. D. (1990). The gain of two chloroplast tRNA introns marks the green algal ancestors of land plants. *Nature* **345**:6272, pp. 268–270. DOI: 10.1038/345268a0

Mathews, S. (2009). Phylogenetic relationships among seed plants: Persistent questions and the limits of molecular data. *American Journal of Botany* **96**:1, pp. 228–236. DOI: 10.3732/ajb.0800178

McCoy, S. R. et al. (2008). The complete plastid genome sequence of Welwitschia mirabilis: an unusually compact plastome with accelerated divergence rates. *BMC evolutionary biology* **8**: p. 130. DOI: 10.1186/1471-2148-8-130

Nemenman, I., Bialek, W., and de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E* **69**:5, p. 056111. DOI: 10.1103/PhysRevE.69.056111

Nickrent, D. L. et al. (2000). Multigene Phylogeny of Land Plants with Special Reference to Bryophytes and the Earliest Land Plants. *Molecular Biology and Evolution* **17**:12, pp. 1885–1895. DOI: 10.1093/oxfordjournals.molbev.a026290

Nixon, K. C. et al. (1994). A Reevaluation of Seed Plant Phylogeny. *Annals of the Missouri Botanical Garden* **81**:3, p. 484. DOI: 10.2307/2399901

Osawa, S. et al. (1992). Recent evidence for evolution of the genetic code. *Microbiological Reviews* **56**:1, pp. 229–64

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**:2, pp. 289–290. DOI: 10.1093/bioinformatics/btg412

Qiu, Y.-L. et al. (1999). The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**:6760, pp. 404–407. DOI: 10.1038/46536

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria

Ruhfel, B. R. et al. (2014). From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC evolutionary biology* **14**: p. 23. DOI: 10.1186/1471-2148-14-23

Ruhlman, T. A. et al. (2015). NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. *BMC Plant Biology* **15**:1, p. 100. DOI: 10.1186/s12870-015-0484-7

Rzhetsky, A. and Nei, M. (1992). A Simple Method for Estimating and Testing Minimum-Evolution Trees. *Molecular Biology and Evolution* **9**:5, pp. 945–945. DOI: 10.1093/oxfordjournals.molbev.a040771

Rzhetsky, A. and Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution* **10**:5, pp. 1073–1095. DOI: 10.1093/oxfordjournals.molbev.a040056

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**:4, pp. 406–25. DOI: 10.1093/oxfordjournals.molbev.a040454

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**:20, pp. 6097–100

Shannon, C. E. (1948). *A mathematical theory of communication.* DOI: 10.1002/j.1538-7305.1948.tb01338.x

Simon, D. et al. (2003). Phylogeny and Self-Splicing Ability of the Plastid tRNA-Leu Group I Intron. *Journal of Molecular Evolution* **57**:6, pp. 710–720. DOI: 10.1007/s00239-003-2533-3

Soltis, D. E., Soltis, P. S., and Zanis, M. J. (2002). Phylogeny of seed plants based on evidence from eight genes. *American journal of botany* **89**:10, pp. 1670–81. DOI: 10.3732/ajb.89.10.1670

Sprinzl, M. et al. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research* **26**:1, pp. 148–53

Sugiura, M. and Wakasugi, T. (1989). Compilation and comparison of transfer RNA genes from tobacco chloroplasts. *Critical Reviews in Plant Sciences* **8**:2, pp. 89–101. DOI: 10.1080/07352688909382271

Sukumaran, J. and Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**:12, pp. 1569–1571. DOI: 10.1093/bioinformatics/btq228

Vogel, J., Börner, T., and Hess, W. R. (1999). Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucleic Acids Research* **27**:19, pp. 3866–74

Wan, T. et al. (2018). A genome for gnetophytes and early evolution of seed plants. *Nature Plants* **4**:2, pp. 82–89. DOI: 10.1038/s41477-017-0097-2

Wang, X.-Q. and Ran, J.-H. (2014). Evolution and biogeography of gymnosperms. *Molecular Phylogenetics and Evolution* **75**: pp. 24–40. DOI: 10.1016/J.YMPEV.2014.02.005

Wickett, N. J. et al. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences of the United States of America* **111**:45, E4859–68. DOI: 10.1073/pnas.1323926111

Wu, C.-S. et al. (2011). Comparative chloroplast genomes of pinaceae: insights into the mechanism of diversified genomic organizations. *Genome biology and evolution* **3**: pp. 309–19. DOI: 10.1093/gbe/evr026

Xi, Z., Rest, J. S., and Davis, C. C. (2013). Phylogenomics and coalescent analyses resolve extant seed plant relationships. *PloS one* **8**:11, e80870. DOI: 10.1371/journal.pone.0080870

Zhong, B. et al. (2010). The Position of Gnetales among Seed Plants: Overcoming Pitfalls of Chloroplast Phylogenomics. *Molecular Biology and Evolution* **27**:12, pp. 2855–2863. DOI: 10.1093/molbev/msq170

Chapter 4

tsfm - tRNA Structure Function Mapper

4.1 Abstract

Transfer RNAs (tRNAs) are short non-coding RNAs acting as adaptor molecules in the process of translating the information stored in nucleotide sequences into proteins. To ensure faithful translations of the genetic code tRNAs must interact productively with the correct aminoacyl-tRNA synthetase (aaRS) to be charged with its cognate amino acid while avoiding productive interactions with non-cognate aaRSs. A tRNAs specificity is determined by a set of identity determinants, which are defined as the nucleotide state (e.g. A,C,G,U) at structural positions along the tRNA molecule. These identity determinants are not static and have been shown to vary widely across the tree of life. Here we introduced tsfm - tRNA structure function mapper, a Python application with C extension modules implementing a previously described information theory approach for predicting tRNA identity determinants. We improve on the original approach by implementing a more accurate entropy estimator. Additionally, we expand on the original work by introducing calculations for basepair features, methods for statistical significance testing, and we implement a distance metric to quantify overall similarity of tRNA identity determinants between genomic sets of tRNAs. Moreover, given recent interest in targeting interactions between tRNAs and aaRSs for therapeutic drug development we believe that tsfm will be a valuable tool for this community by predicting potential targets for additional testing and development.

4.2 Background

Transfer RNAs (tRNA) are short non-coding RNAs mainly involved in protein synthesis acting as adaptor molecules converting the information contained in the genome into proteins (Marck and Grosjean, 2002). To operate as an adaptor molecule during protein synthesis, tRNAs are involved in two spatially and temporally separated reactions. First, tRNAs are recognized by an aminoacyl-tRNA synthetase (aaRS) that charges the tRNA by attaching a specific amino acid to its $3'$ end. The amino acid attached to a tRNA is determined by the tRNA's charging capacity referred to as its functional class (Berg and Offengand, 1958; Zamecnik et al., 1958; Goodman and Rich, 1962; McClain, 1993; Giegé et al., 1998). Mostly, there exists an aaRS enzyme specific for each amino acid that charges a cognate tRNA (Chaliotis et al., 2017). The second reaction occurs during translation and is facilitated by the ribosome, where mRNA codons are decoded by pairing with charged tRNAs with complementary anticodons followed by the donation of a charged amino acid to the nascent polypeptide chain (Söll et al., 1965; Nirenberg et al., 1965; Ogle et al., 2001).

tRNAs involved in protein synthesis must conform to the same general tRNA structure for efficient activity with general translation factors and the ribosome. This structure is routinely referred to as a clover leaf shape, consisting of three stem-loops, a variable loop, a base-paired stem, and an unpaired tail. Despite the very high structural similarity of all tRNAs engaged in protein synthesis each must interact productively with the correct aaRS to be charged with its cognate amino acid while avoiding productive interactions with others to ensure faithful translations of the genetic code (Normanly and Abelson, 1989; McClain, 1993; Giegé et al., 1998). The proper charging of a tRNA relies on a set of identity features, which are defined as the nucleotide state (e.g. A,C,G,U) at structural positions along the tRNA molecule (Giegé et al., 1998). Collectively, the features that determine the charging capacity of a tRNA are called tRNA Class Informative Features (CIFs) and contain identity determinants that promote recognition by its cognate aaRS and anti-determinants that discriminate against noncognate aaRSs (Giegé et al., 1998; Freyhult et al., 2006). Previously discovered tRNA CIFs are mostly concentrated along the acceptor stem (Giegé et al., 1998), the anticodon itself (Ho et al., 2018), and the discriminator base (structural position 73) (Crothers et al., 1972). Identity determinants found outside these locations tend to be species- and lineage-dependent and are distributed throughout the tRNA structure (Giegé et al., 1998; Freyhult et al., 2007; Ho et al., 2018).

Experimental approaches to elucidate tRNA CIFs are labor intensive. To date the tRNA CIFs required for unambiguous charging of tRNAs have only been determined biochemically for *Escherichia coli* (Ibba and Söll, 2000; Ho et al., 2018). While several tRNA CIFs are strongly conserved across lineages others vary widely across the tree of life especially within eukaryotes (Freyhult et al., 2007; Amrine et al., 2014; Ho et al., 2018) limiting the applicability of previous experimental results to novel systems. Consequently, this motivated the development of a bioinformatic approach to predicting tRNA CIFs by using the information theoretical metric entropy (Shannon, 1948; Schneider and Stephens, 1990) to quantify the association of tRNA functional classes with identity features (Freyhult et al., 2006). In this work we introduce tsfm - tRNA structure function mapper, where we reimplement and expand on the bioinformatic method introduced by Freyhult et al., 2006 in a Python application with C extension modules. We improve on the original molecular information theoretical approach to predicting tRNA CIFs by implementing the more accurate, eponymously named NSB entropy estimator (Nemenman et al., 2004) and by including basepair features in our tRNA CIF estimation. Furthermore, we have introduced new functionality providing statistical significance testing of tRNA structural features functional information content and feature-tRNA class association. We also implemented a distance metric option that quantifies overall similarity of tRNA CIFs between genomic sets of tRNAs using the Jensen-Shannon distance (Endres and Schindelin, 2003). The distance functionality is designed in a object-oriented programming approach, which is easily extended to include additional distance metrics. Lastly, given recent interest in targeting interactions between tRNAs and aaRSs for therapeutic drug development (Jain et al., 2017; Pasaje et al., 2016; Ho et al., 2018), we believe the development of tsfm is timely and will prove to be a valuable tool given its ability to estimate tRNA features critical for interactions between tRNAs and cognate aaRSs and to quantify differences between genomic sets of tRNAs.

4.3 Implementation

tsfm is written in Python with compiled C extension modules and has a command line user interface. As input tsfm expects files of structurally aligned tRNA sequences in clustal or fasta format partitioned by functional class. The naming convention expected for these files is `<dataset name>_<IUPAC AA one-letter code>.(aln|fna)` for example: `dataset1_F.aln`. We recommend using either COVE (Eddy and Durbin, 1994) or Infernal (Nawrocki and Eddy, 2013) to produce structural alignments. If predicting tRNA CIFs for basepair features a separate file containing the consensus secondary structure annotation in the extended dot bracket notation format outputted by COVE (Eddy and Durbin, 1994) or Infernal (Nawrocki and Eddy, 2013) is required. When calculating distance metrics between datasets either tRNA CIFs must be predicted for both datasets at the same time by providing alignment files for each or results from previous analyses may be used. The final output of tRNA CIF predictions and distance metrics are tab-delimited text files, and if the option is selected, function logo post-script graphics which are fully described below. Lastly, an overview of the standard workflow for tsfm is provided in figure 4.1.

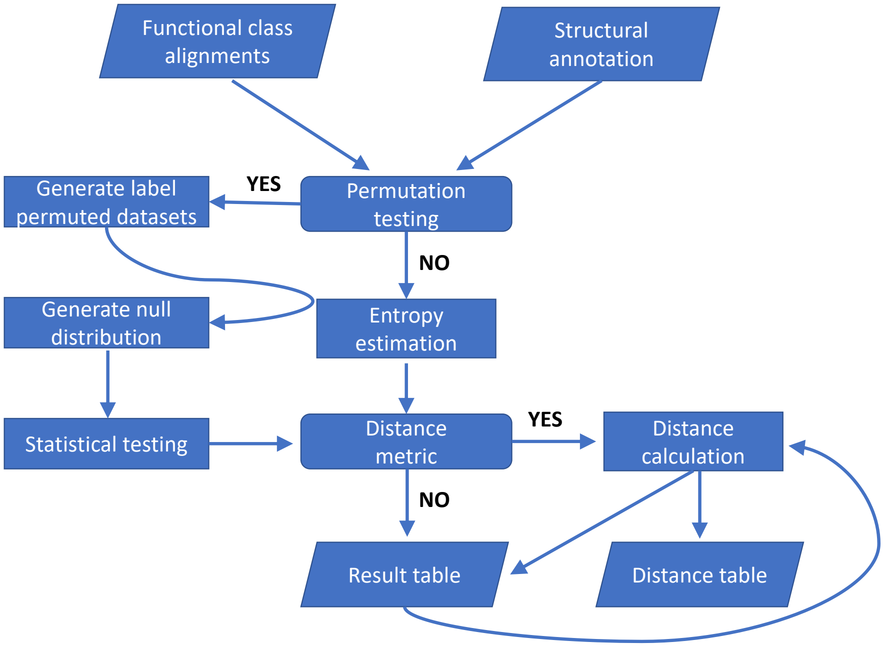

Figure 4.1: Overview of the tsfm workflow.

4.3.1 tRNA CIF prediction

Here we describe the molecular information theory approach for predicting tRNA CIFs (Freyhult et al., 2006) followed by descriptions of the methods implemented for entropy estimation in small sample sizes.

For the purpose of calculating tRNA CIFs the possible functional classes of tRNAs are denoted by their IUPAC one-letter amino acid code, $Z \equiv \{A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y\}$ and features F is the union of two Cartesian products $n_i \times SC \cup (n_i \times n_i) \times BP$, where $n_i \in \{A, C, G, U, -\}$ is the set of possible states of an alignment position, SC is the set of Sprinzl Coordinates (Sprinzl et al., 1998), and BP is the set of Sprinzl Coordinate pairs involved in basepairs of the tRNA secondary structure. The functional information $I_{f_i}(Z|f_i)$ that feature $f_i \in F$ confers about the frequencies of different functional classes Z is defined as: $I_{f_i}(Z|f_i) = H(Z) - e(n(f_i)) - H(Z|f_i))$, where $H(Z|f_i) = -\Sigma_{z \in Z} p(z|f_i) log_2(p(z|f_i))$ is the class entropy or level of uncertainty about the functional class of tRNAs with feature f_i. $H(Z) = -\Sigma_{z \in Z} p(z) log_2(p(z))$ is the background entropy which depends on relative frequency of sequences belonging to different functional classes, and by definition: $0 \leq p(z), p(z|f_i) \leq 1, \Sigma_{z \in Z} p(z) = 1, \Sigma_{z \in Z} p(z|f_i) = 1$. Finally, the proportion of functional information $I_{f_i}(Z|f_i)$ attributed to each functional class $z \in Z$ is calculated using Gorodkin heights (Gorodkin et al., 1997) where $h^z_{f_i} = (p(z|f_i)/p(z))/\Sigma_{w \in Z} p(w|f_i)/p(w))$, h_{f_i} is the vector of Gorodkin heights $\forall z \in Z$ at feature f_i, and by definition $\Sigma h^X_{f_i} = 1$.

It is known that calculating entropy from the maximum likelihood estimates of $p(z)$ and $p(z|f_i)$ from sampled frequencies will underestimate the true entropy (Basharin, 1959). We reimplement the exact method described in Schneider and Stephens, 1990 denoted above as $e(n(f_i)))$, which is correction for biases caused by small sample size. It is calculated for each sample size $1, \ldots, n$ where n is provided as an option to tsfm. The calculation of the exact method is implemented as a compiled C extension and computed in parallel over sample sizes, still it becomes prohibitively costly to calculate beyond a sample size of 17. To estimate entropy for larger sample sizes we reimplemented the Miller-Madow (Miller, 1955) estimator originally used by Freyhult et al., 2006 and an improved entropy estimator using the NSB method (Nemenman et al., 2004). The Miller-Madow (Miller, 1955) method provides an improvement over using the maximum likelihood frequencies, however, it only utilizes information from the sample size and is not dependent on the distribution of classes, which can lead to inaccurate estimates at smaller sample sizes. Consequently, we also implemented the improved NSB (Nemenman et al., 2004) estimator which uses a Bayesian approach to utilize information from the sample size and distribution of classes which has been shown to lead to better estimates at small sample sizes (Nemenman et al., 2004).

To confirm that the NSB estimator provides better entropy estimates we simulated distributions with 4.46, 3.62, 2.66, 1.86, 0.91, and 0.51 bits of entropy, randomly sampled these distributions for every sample size between 1..50, and calculated the entropy using the maximum likelihood frequency estimate, Miller-Madow (Miller, 1955), and NSB (Nemenman et al., 2004). We repeated this 100 times for each sampling depth calculating the average and standard deviation. Based on our simulations the NSB (Nemenman et al., 2004) consistently provides an improvement over the Miller-Madow (Miller, 1955) estimator at smaller sample sizes except for the highest entropy level where they performed similarly (Fig. 4.2-4.7).

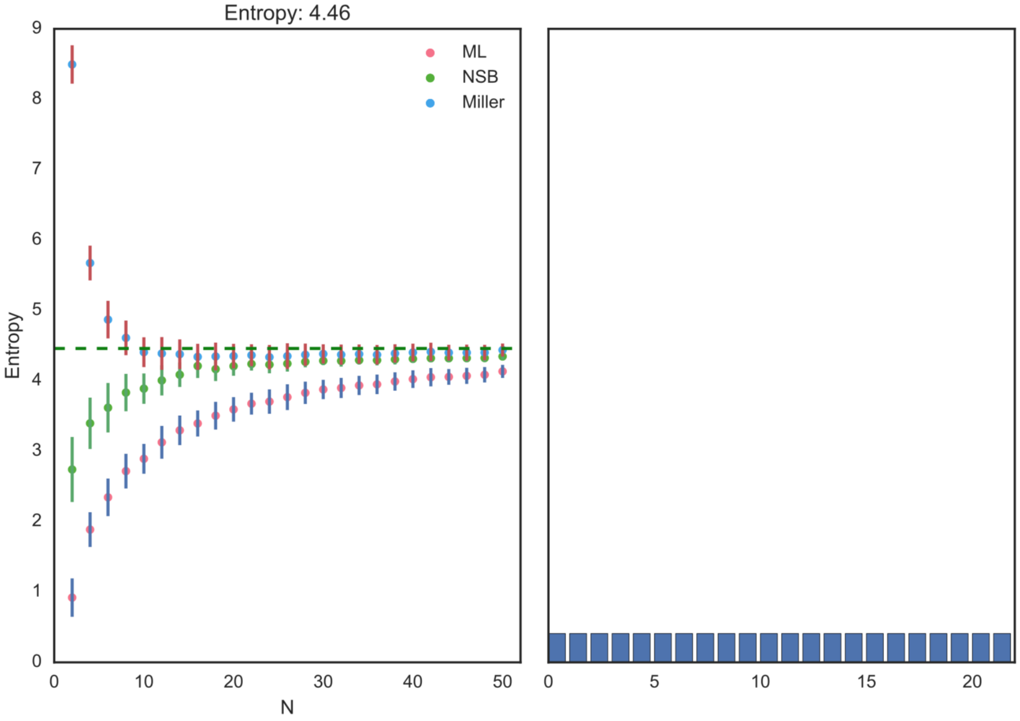

Figure 4.2: Entropy estimator comparison for a distribution with a true entropy of 4.46 bits. (Left panel) The y-axis is entropy measured in bits and the x-axis is the sampling depth. Dots represent the average entropy estimated over 100 replicates at each sampling depth for each estimator. Error bars are ± one standard deviation. The horizontal dashed green line is the true entropy for the distribution. (Right panel) A histogram representing the discrete probability distribution over 22 states that was used for random sampling.

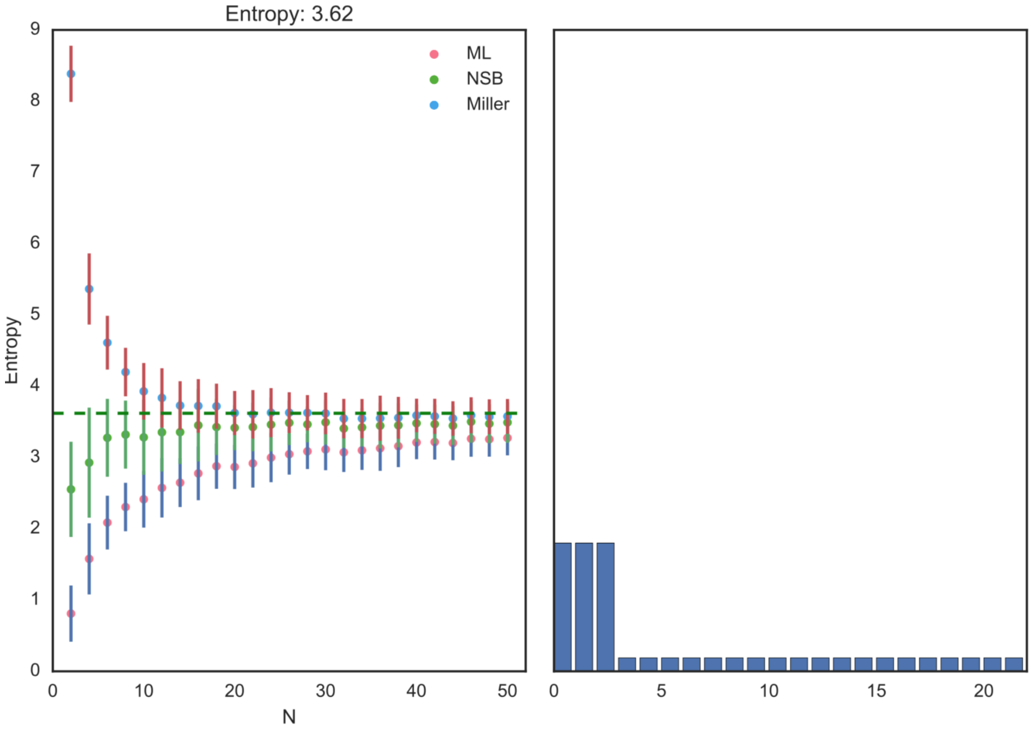

Figure 4.3: Entropy estimator comparison for a distribution with a true entropy of 3.62 bits. (Left panel) The y-axis is entropy measured in bits and the x-axis is the sampling depth. Dots represent the average entropy estimated over 100 replicates at each sampling depth for each estimator. Error bars are ± one standard deviation. The horizontal dashed green line is the true entropy for the distribution. (Right panel) A histogram representing the discrete probability distribution over 22 states that was used for random sampling.

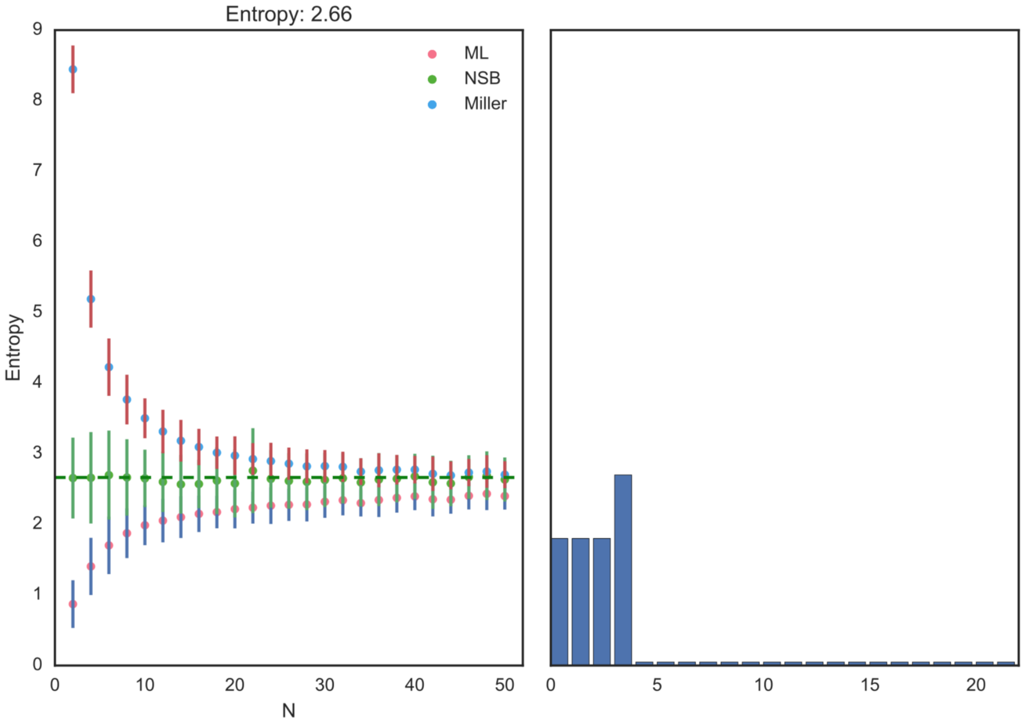

Figure 4.4: Entropy estimator comparison for a distribution with a true entropy of 2.66 bits. (Left panel) The y-axis is entropy measured in bits and the x-axis is the sampling depth. Dots represent the average entropy estimated over 100 replicates at each sampling depth for each estimator. Error bars are ± one standard deviation. The horizontal dashed green line is the true entropy for the distribution. (Right panel) A histogram representing the discrete probability distribution over 22 states that was used for random sampling.

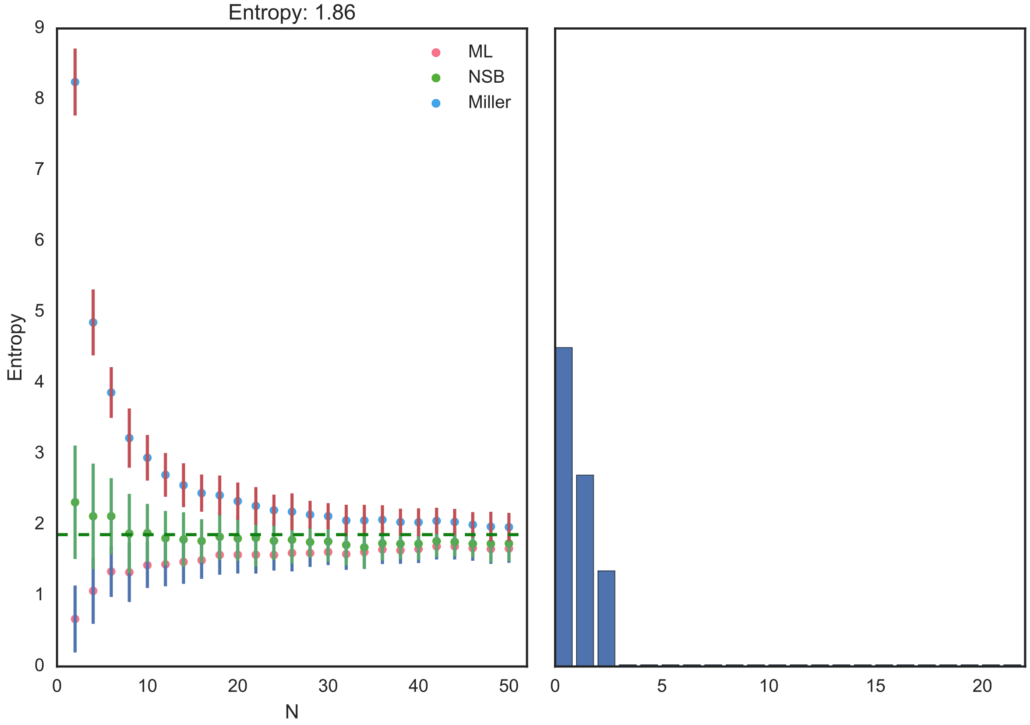

Figure 4.5: Entropy estimator comparison for a distribution with a true entropy of 1.86 bits. (Left panel) The y-axis is entropy measured in bits and the x-axis is the sampling depth. Dots represent the average entropy estimated over 100 replicates at each sampling depth for each estimator. Error bars are ± one standard deviation. The horizontal dashed green line is the true entropy for the distribution. (Right panel) A histogram representing the discrete probability distribution over 22 states that was used for random sampling.

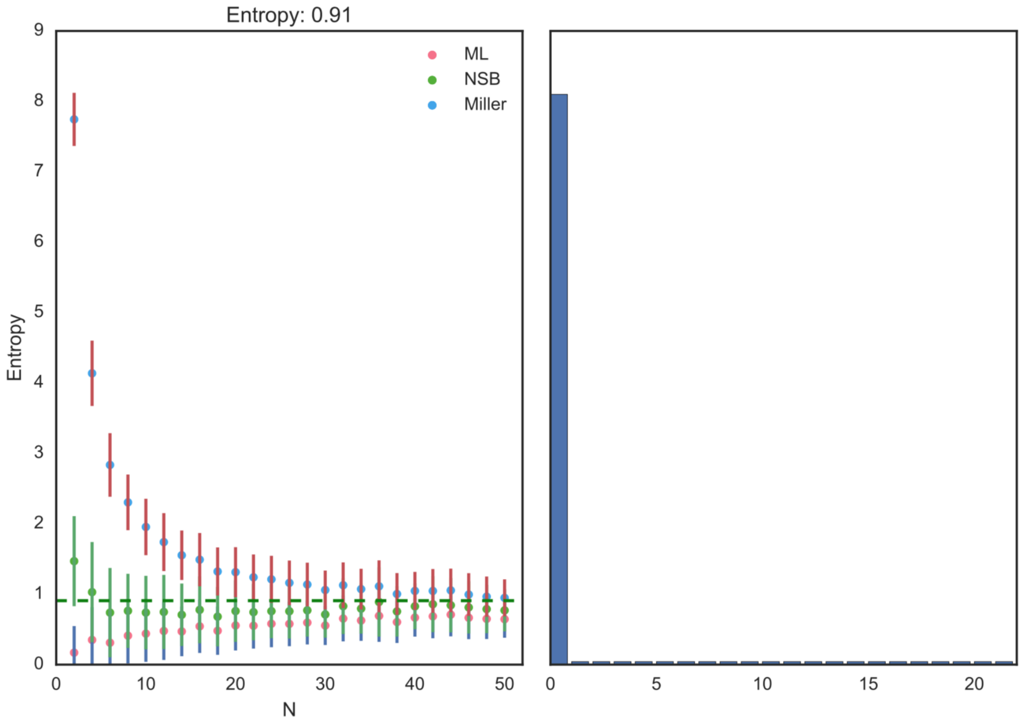

Figure 4.6: Entropy estimator comparison for a distribution with a true entropy of 0.91 bits. (Left panel) The y-axis is entropy measured in bits and the x-axis is the sampling depth. Dots represent the average entropy estimated over 100 replicates at each sampling depth for each estimator. Error bars are ± one standard deviation. The horizontal dashed green line is the true entropy for the distribution. (Right panel) A histogram representing the discrete probability distribution over 22 states that was used for random sampling.

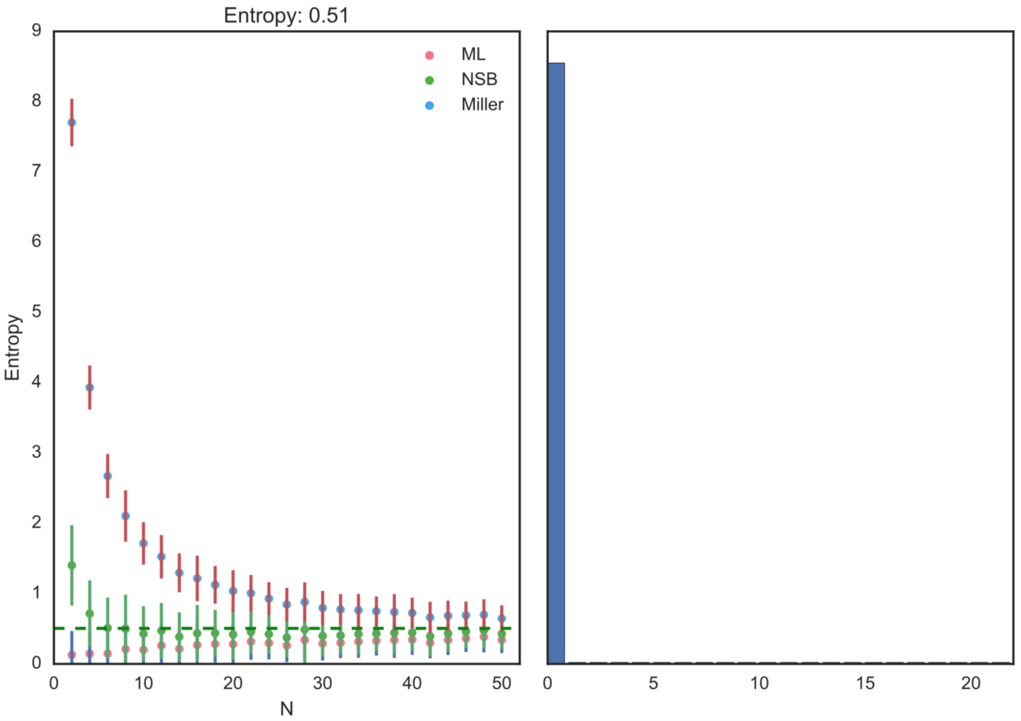

Figure 4.7: Entropy estimator comparison for a distribution with a true entropy of 0.51 bits. (Left panel) The y-axis is entropy measured in bits and the x-axis is the sampling depth. Dots represent the average entropy estimated over 100 replicates at each sampling depth for each estimator. Error bars are ± one standard deviation. The horizontal dashed green line is the true entropy for the distribution. (Right panel) A histogram representing the discrete probability distribution over 22 states that was used for random sampling.

4.3.2 Statistical testing

To provide statistical significance testing of the information content of features, $f_i \in F$, and of the information associated with a functional class, $z \in Z$ within a feature we implemented permutation based testing. The null distribution for both tests is generated by shuffling the functional class labels of input sequences followed by the calculation of the information content. The permutation p-values are calculated as $(n_{above} + 1)/(n + 1)$, where n_{above} is the number of permutations with an information value greater than the feature, f_i, or functional class, z depending on the test and n is the total number of permutations. To correct for multiple testing we provide all the options available within the multitest module of the statsmodels package, which provides several methods for controlling the family-wise error rate or the false discovery rate.

4.3.3 tRNA CIF distance metrics

To measure pairwise distances between datasets $x_{(j,k)} \in X$ where X is a set of datasets we have implemented a modified version of the Jensen-Shannon distance (JSD), an information theory metric for quantifying similarity between two probability distributions (Endres and Schindelin, 2003). We weighted the JSD calculation by the sum of the functional information $I^x_{f_i}$ of the feature f_i for datasets $x_{(j,k)}$ and summed over all features $f_i \in F$ (equation 4.1). The function $H()$ in equation 4.1 is the Shannon entropy (Shannon, 1948) defined as: $H(V) = -\sum_i^n p(v_i)log_2(p(v_i))$.

$$\sum_{f_i \in F} \left(I^{x_j}_{f_i} + I^{x_k}_{f_i}\right) \sqrt{H\left(\sum_{m \in \{j,k\}} I^{x_m}_{f_i} h^{x_m}_{f_i}\right) - \sum_{m \in \{j,k\}} I^{x_m}_{f_i} H\left(h^{x_m}_{f_i}\right)} \tag{4.1}$$

Although, we only provide a single distance metric the object-oriented programming design of tsfm easily allows the addition of new distance metrics by extending the *Distance* class.

4.3.4 Graphical output

The graphical output of tsfm follows Freyhult et al., 2006 by producing function logos by plotting functional information as a stacked bar graph with structural position on the x-axis and information on the y-axis in bits. Each element of a stack is a symbol for a functional class $z \in Z$. Symbol heights are the product of the information content for f_i and the Gorodkin height (Gorodkin et al., 1997) of z. Symbols are sorted by heights with the symbols with largest height appearing on top. An example is provided in Figure 4.8.

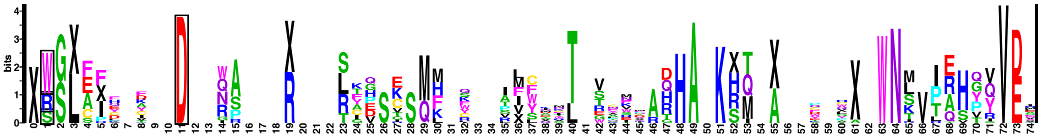

Figure 4.8: Example of a cytosine function logo produced by tsfm.

4.4 Conclusion

Here we introduced tsfm - tRNA structure function mapper, an easily installed and widely available Python application with C extension modules for predicting tRNA features that provide charging specificity using the approach introduced by Freyhult et al., 2006. We improve on the original method by implementing the more accurate NSB entropy estimator (Nemenman et al., 2004). Moreover, we also expanded on the original work by introducing calculations for basepair features, statistical significance testing, and a method to quantify overall similarity of tRNA CIFs between genomic sets of tRNAs. Given recent interest in targeting interactions between tRNAs and aaRSs for therapeutic drug development (Jain et al., 2017; Pasaje et al., 2016; Ho et al., 2018) we believe that tsfm will be a valuable tool for this community by predicting potential targets for further testing and development.

4.5 Availability and Requirements

Project name: tsfm - tRNA structure function mapper
Project home page: https://github.com/tlawrence3/tsfm
Operating system: MacOSX and Linux
Programming language: Python ($\geq$ 3.4) and C99
Other requirements: numpy, statsmodels, pandas, mpmath
License: LGPL-3.0

4.6 References

Amrine, K. C. H., Swingley, W. D., and Ardell, D. H. (2014). tRNA signatures reveal a polyphyletic origin of SAR11 strains among alphaproteobacteria. *PLoS Computational Biology* **10**:2. Ed. by C. A. Ouzounis, e1003454. DOI: 10.1371/journal.pcbi.1003454

Basharin, G. P. (1959). On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables. *Theory of Probability & Its Applications* **4**:3, pp. 333–336. DOI: 10.1137/1104033

Berg, P. and Offengand, E. J. (1958). An Enzymatic Mechanism for Linking Amino Acids to RNA. *Proceedings of the National Academy of Sciences of the United States of America* **44**:2, pp. 78–86. DOI: 10.1073/PNAS.44.2.78

Chaliotis, A. et al. (2017). The complex evolutionary history of aminoacyl-tRNA synthetases. *Nucleic acids research* **45**:3, pp. 1059–1068. DOI: 10.1093/nar/gkw1182

Crothers, D. M., Seno, T., and Söll, G. (1972). Is there a discriminator site in transfer RNA? *Proceedings of the National Academy of Sciences of the United States of America* **69**:10, pp. 3063–7

Eddy, S. R. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic Acids Research* **22**:11, pp. 2079–2088. DOI: 10.1093/nar/22.11.2079

Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory* **49**:7, pp. 1858–1860. DOI: 10.1109/TIT.2003.813506

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, pp. 905–16. DOI: 10.1093/nar/gkj478

Freyhult, E. et al. (2007). New computational methods reveal tRNA identity element divergence between Proteobacteria and Cyanobacteria. *Biochimie* **89**:10, pp. 1276–88. DOI: 10.1016/j.biochi.2007.07.013

Giegé, R., Sissler, M., and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Research* **26**:22, pp. 5017–35

Goodman, H. M. and Rich, A. (1962). Formation of a DNA-soluble RNA hybrid and its relation to the origin, evolution, and degeneracy of soluble RNA. *Proceedings of the National Academy of Sciences of the United States of America* **48**:12, pp. 2101–9

Gorodkin, J. et al. (1997). Displaying the information contents of structural RNA alignments: the structure logos. *Bioinformatics* **13**:6, pp. 583–586. DOI: 10.1093/bioinformatics/13.6.583

Ho, J. M. et al. (2018). Drugging tRNA aminoacylation. *RNA Biology*, pp. 1–11. DOI: 10.1080/15476286.2018.1429879

Ibba, M. and Söll, D. (2000). Aminoacyl-tRNA Synthesis. *Annual Review of Biochemistry* **69**:1, pp. 617–650. DOI: 10.1146/annurev.biochem.69.1.617

Jain, V. et al. (2017). Targeting Prolyl-tRNA Synthetase to Accelerate Drug Discovery against Malaria, Leishmaniasis, Toxoplasmosis, Cryptosporidiosis, and Coccidiosis. *Structure* **25**:10, 1495–1505.e6. DOI: 10.1016/j.str.2017.07.015

Marck, C. and Grosjean, H. (2002). tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA (New York, N.Y.)* **8**:10, pp. 1189–232

McClain, W. H. (1993). Rules that Govern tRNA Identity in Protein Synthesis. *Journal of Molecular Biology* **234**:2, pp. 257–280. DOI: 10.1006/jmbi.1993.1582

Miller, G. A. (1955). “Note on the bias of information estimates.” *Information Theory in Psychology II-B*. Glencoe, IL: Free Press, pp. 95–100

Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics (Oxford, England)* **29**:22, pp. 2933–5. DOI: 10.1093/bioinformatics/btt509

Nemenman, I., Bialek, W., and de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E* **69**:5, p. 056111. DOI: 10.1103/PhysRevE.69.056111

Nirenberg, M. et al. (1965). RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America* **53**:5, pp. 1161–8

Normanly, J. and Abelson, J. (1989). tRNA Identity. *Annual Review of Biochemistry* **58**:1, pp. 1029–1049. DOI: 10.1146/annurev.bi.58.070189.005121

Ogle, J. M. et al. (2001). Recognition of Cognate Transfer RNA by the 30S Ribosomal Subunit. *Science* **292**:5518, pp. 897–902. DOI: 10.1126/science.1060612

Pasaje, C. F. A. et al. (2016). Selective inhibition of apicoplast tryptophanyl-tRNA synthetase causes delayed death in Plasmodium falciparum. *Scientific Reports* **6**:1, p. 27531. DOI: 10.1038/srep27531

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**:20, pp. 6097–100

Shannon, C. E. (1948). *A mathematical theory of communication.* DOI: 10.1002/j.1538-7305.1948.tb01338.x

Söll, D. et al. (1965). Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA’s to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. *Proceedings of the National Academy of Sciences of the United States of America* **54**:5, pp. 1378–85

Sprinzl, M. et al. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research* **26**:1, pp. 148–53

Zamecnik, P. C., Stephenson, M. L., and Hecht, L. I. (1958). Intermediate Reactions in Amino Acid Incorporation. *Proceedings of the National Academy of Sciences of the United States of America* **44**:2, pp. 73–8

Chapter 5

FAST: FAST Analysis of Sequences Toolbox

5.1 Abstract

FAST (FAST Analysis of Sequences Toolbox) provides simple, powerful open source command-line tools to filter, transform, annotate and analyze biological sequence data. Modeled after the GNU (GNU's Not Unix) Textutils such as grep, cut, and tr, FAST tools such as fasgrep, fascut, and fastr make it easy to rapidly prototype expressive bioinformatic workflows in a compact and generic command vocabulary. Compact combinatorial encoding of data workflows with FAST commands can simplify the documentation and reproducibility of bioinformatic protocols, supporting better transparency in biological data science. Interface self-consistency and conformity with conventions of GNU, Matlab, Perl, BioPerl, R, and GenBank help make FAST easy and rewarding to learn. FAST automates numerical, taxonomic, and text-based sorting, selection and transformation of sequence records and alignment sites based on content, index ranges, descriptive tags, annotated features, and in-line calculated analytics, including composition and codon usage. Automated content- and feature-based extraction of sites and support for molecular population genetic statistics make FAST useful for molecular evolutionary analysis. FAST is portable, easy to install and secure thanks to the relative maturity of its Perl and BioPerl foundations, with stable releases posted to CPAN. Development as well as a publicly accessible Cookbook and Wiki are available on the FAST GitHub repository at https://github.com/tlawrence3/FAST. The default data exchange format in FAST is Multi-FastA (specifically, a restriction of BioPerl FastA format). Sanger and Illumina 1.8+ FastQ formatted files are also supported. FAST makes it easier for non-programmer biologists to interactively investigate and control biological data at the speed of thought.

5.2 Introduction

Bioinformatic software for non-programmers is traditionally implemented for user convenience in monolithic applications with Graphical User Interfaces (GUIs) (Smith et al., 1994; Rampp et al., 2006; Librado and Rozas, 2009a; Waterhouse et al., 2009; Gouy et al., 2010; Stothard, 2000). However, the monolithic application paradigm is easily outscaled by today's big biological data, particularly Next Generation Sequencing (NGS) data at gigabyte- and terabyte-scales. Better empowerment of non-programmers for genome-scale analytics of big biological data has been achieved through web-based genome browser interfaces (Markowitz et al., 2014; Cunningham et al., 2015; Rosenbloom et al., 2015). On the other hand, for smaller datasets, sequence and alignment editor applications encourage manual manipulation of data, which is error-prone and essentially irreproducible. To reduce error and increase reproducibility in the publishing of bioinformatic and biostatistical protocols it is important to facilitate the documentation and automation of data science workflows through scripts and literate programming facilities (Knuth, 1984) such as emacs org-mode (http://orgmode.org), as demonstrated in, for example (Delescluse et al., 2012) that both completely document and encode scientific workflows for machine processing of biological data.

Reproducibility in bioinformatics and biostatistics protocols is crucial to maintaining public trust in the value of its investments in high-throughput and high-dimensional measurements of complex biological systems (Baggerly and Coombes, 2009; Hutson, 2010; Baggerly and Coombes, 2011; Huang and Gottardo, 2013). In one analysis, only two of 18 published microarray gene-expression analyses were completely reproducible, in part because key analysis steps were made with proprietary closed-source software (Ioannidis et al., 2008). Furthermore, even though analytical errors are a major source of retractions in the scientific literature (Casadevall et al., 2014), peer-review and publication of scientific data processing protocols is generally not yet required to publish scientific studies. Adequate documentation of bioinformatic and biostatistical workflows and open source sharing of code upon publication (Peng, 2009) facilitates crowd-sourced verification, correction and extension of code-based analyses (Barnes, 2010; Morin et al., 2012), and reuse of software and data to enable more scientific discovery returns from public data (Peng, 2011). Peer review and publication of the data science protocols associated to scientific studies stems temptation to overinterpret results and encourages more objectivity in data science (Boulesteix, 2010). The ultimate remedy for these problems is to expand literacy in modern computational and statistical data science for science students in general (Morin et al., 2012; Joppa et al., 2013).

Web-based open-source workflow suites such as Galaxy (Blankenberg and Hillman-Jackson, 2014), Taverna (Oinn et al., 2006) and BioExtract (Lushbough et al., 2011) are a recent innovation in the direction of greater reproducibility in bioinformatics protocols for genome-scale analytics. However, the most powerful, transparent and customizable medium for reproducible bioinformatics work is only available to bioinformatics specialists and programmers through Application Programming Interfaces (APIs) such as BioPerl and Ensembl (Yates et al., 2015).

Yet workflow design suites and programming APIs require dedication and time to learn. There is a need for more bioinformatics software in between GUIs and APIs, that empowers non-programmer scientists and researchers to interactively and reproducibly control, process and analyze their data without manual interventions. Closer inspection of data and interactive construction and control of data workflows makes it so much easier to rapidly prototype error-free workflows, nipping errors in the bud that can completely confound downstream analyses. In scientific computing, the time-tested paradigm for rapid prototyping of reproducible data workflows is the Unix command-line.

In this tradition we here present FAST: FAST Analysis Sequences Toolbox, modeled after the standard Unix toolkit (Peek, 2001), now called Coreutils. The FAST tools follow the Unix philosophy to "do one thing and do it well" and "write programs to work together." (Stutz, 2000). FAST workflows are completely automated; no manual interventions to data are required. FAST falls between a GUI and an API, because it is used through a Command-Line Interface (CLI). Although the FAST tools are written in Perl using BioPerl packages (Stajich et al., 2002), FAST users do not need to be able to program Perl or know BioPerl. FAST users only need basic competence in Unix and the modest skill to compose command pipelines in the Unix shell. FAST therefore supports an emerging movement to empower non-programmer biologists to learn Unix for scientific computing. Books and courses in this emerging market include the recent "UNIX and Perl to the Rescue!" (Bradnam and Korf, 2012) and the Software Carpentry and Data Carpentry Foundations workshops (Wilson, 2014).

Unix command pipe-lines are the paradigmatic example of the "pipes and filters" design pattern that embodies serial processing of data through sequences of modular and reuseable computations. The "pipes and filters" design pattern is a special case of component-based software engineering (McIlroy, 1969) and a core paradigm in software architecture (Garlan and Shaw, 1994). The component-wise organization of FAST affords access to an infinite variety of customizable queries and workflows on biological sequence data using a small command vocabulary and combinatorial logic. Component-based software is easier to learn, maintain and extend. It also makes it easy for users to interactively develop new protocols through the modular extension and recombination of existing protocols. As shown from the examples below, non-trivial computations may be expressed on a single line of the printed page. Thus, FAST can help empower non-biologist programmers to develop and communicate powerful and reproducible bioinformatic workflows for scientific investigations and publishing.

Open-source command-line utilities for bioinformatics such as the EMBOSS package (Rice et al., 2000), the FASTX tools (Gordon, 2009) or the scripts that come with BioPerl (Stajich et al., 2002) typically offer suites of tools with simple, well-defined functions that lend themselves to scripting, but are not necessarily designed according to the Unix toolbox philosophy specifically to interoperate through serial composition over pipes. Similarly, FaBox (Villensen, 2007) is a free and open online server with functions that overlap with FAST tools, but is not designed for serial composition. On the other hand, the Unix toolbox model has been used before in more or less more specialized bioinformatics applications such as the popular SAMTools suite (Li et al., 2009) and in the processing of NMR data (Delaglio et al., 1995). A toolsuite called bp-utils, with a similar design philosophy and some overlapping functionality with FAST, has recently been released at http://diverge.hunter.cuny.edu/labwiki/Bioutils.

We have written extensive documentation for each FAST utility along with useful error messages following recommended practice (Seemann, 2013). FAST is free and open source; its code is freely available to anyone to re-use, verify and extend through its GitHub repository.

5.3 Design and Implementation of FAST Tools

5.3.1 The FAST Data Model

The Unix Coreutils paradigm allows users to treat plain-text files and data streams as databases in which records correspond to single lines containing fields separated by delimiters such as commas, tabs, or strings of white-space characters. FAST extends this paradigm to biological sequence data, allowing users to treat collections of files and streams of multi-line sequence records as databases for complex queries, transformations and analytics. FAST generalizes the GNU Coreutils model exactly because it models sequence record descriptions as an ordered collection of description fields (see below).

Another design feature of Unix tools that also characterizes the FAST tools is their ability to accept input not only from one or more files but also from what is called standard input, a data-stream supported by the Unix shell, and to output analogously to standard output. It is this facility that allows FAST tools to be serially composed in Unix pipelines that compactly represent an infinite variety of expressive bioinformatic workflows.

The default data exchange format for FAST tools is the universally recognized FastA format (Lipman and Pearson, 1985). While no universal standard exists for this format, for FAST, "FastA format" means what is conventionally called "multi-fasta" format of sequence or alignment data, largely as implementated in BioPerl in the module `Bio::SeqIO::fasta` (Stajich et al., 2002).

In the FAST implementation of FastA format, multiple sequence records may appear in a single file or input stream. Sequence data may contain gap characters. The logical elements (or fields) of a sequence record are its identifier, its description and its sequence. The identifier (indicated with id in the illustration below) and description (desc) together make the identifier line of a sequence record, which must begin with the sequence record start symbol > on a single line. The description begins after the first block of white-space on this line (indicated with <space>). The sequence of a record appears immediately after its identifier line and may continue over multiple lines until the next record starts.

In FAST, users may alter how description fields are defined in sequence records by using Perl-style regular expressions to define delimiters (indicated by <delim>). FAST uses one-based indexing of description fields.

The FAST data model is illustrated as follows:

```
>seq1-id<space>seq1-desc-field1
<delim>seq1-desc-field2<delim>...
seq1-sequence
seq1-sequence
...
seq1-sequence
>seq2-id<space>seq2-desc-field1
<delim>seq2-desc-field2<delim>...
seq2-sequence
seq2-sequence
...
seq2-sequence
```

In FAST, the sequence identifier is thought of as the 0th field of the identifier line. One-based indexing of description fields in FAST is therefore consistent with zero-based indexing in Perl and one-based indexing of sequence coordinates, making all indexing consistent and uniform in FAST.

Most FAST tools extend the field-based paradigm further by supporting tagged values in sequence record descriptions. Tagged values are name-value pairs with a format "name=value" as common in General Feature Format (GFF) used in sequence annotation (see e.g., https://www.sanger.ac.uk/resources/software/gff/) or an alternative "name:value" format that certain FAST tools themselves can annotate in-line into sequence records by appending a new field to sequence record descriptions. Support for tagged values in FAST makes it possible to operate on sequence records with unordered or heterogeneous description fields.

5.3.2 Overview of the FAST Tools

FAST utilities may be assigned to categories according to their default behavior and intended use. There are FAST tools for selection of data from sequence records, transformation of data, annotation of sequence record descriptions with computed characteristics of the data, and analysis. A complete description of all utilities included in the first major release of FAST is shown in Table 5.1.

The analysis class is distinguished from the other classes because by default, these utilities output tables of plain-text data rather than sequence record data in FastA format. Two other tools, `fasconvert` and `gbfcut`, are designed to either input or output FastA format sequence records by default. Standardization of the FAST data model allows users to serially compose FAST tools into pipelines at the Unix command-line,

Table 5.1: Utilities in first major release of FAST.

Tool/Category	Function	Coreutil analog	Operates by default upon
SELECTION			
`fasgrep`	Regex selection of records	`grep`	Identifiers
`fasfilter`	Numerical selection of records		Identifiers
`fastax`	Taxonomic selection of records		Descriptions
`fashead`	Order-based selection of records	`head`	Records
`fastail`	Order-based selection of records	`tail`	Records
`fascut`	Index-based selection and reordering of data	`cut`	Sequences
`gbfcut`	Extract sequences by regex matching on features		Features
`alncut`	Selection of sites by content		Sites
`gbfalncut`	Selection of sites by features		Sites
TRANSFORMATION			
`fassort`	Numerical or text sorting of records	`sort`	Identifiers
`fastaxsort`	Taxonomic sorting of records		Identifiers
`fasuniq`	Remove or count redundant records	`uniq`	Records
`faspaste`	Merging of records	`paste`	Sequences
`fastr`	Character transformations on records	`tr`	Identifiers
`fassub`	Regex substitutions on records		Identifiers
`fasconvert`	Convert sequence formats		Records
ANNOTATION			
`faslen`	Annotate sequence lengths		Descriptions
`fascomp`	Annotate monomeric compositions		Descriptions
`fascodon`	Annotate codon usage		Descriptions
`fasxl`	Annotate biological translations		Descriptions
`fasrc`	Annotate reverse complements		Descriptions
ANALYSIS			
`alnpi`	Molecular population genetic statistics		Sites
`faswc`	Tally sequences and characters	`wc`	Sequences

which is indicated as the “main workflow” in the overview of the project shown in Figure 5.1.

5.3.3 General Implementation and Benchmarking

The BioPerl backend of FAST 1.x is version 1.6.901 downloaded in January, 2012. `Bio::SeqIO` components were updated to version 1.6.923 on June 4, 2014 and some `Bio::Root` components were updated on July 10, 2014 (github commit 50f87e9a4d). We introduced a small number of customizations to the BioPerl code-base, primarily to enable the translation of sequences containing gaps. All of the BioPerl dependencies of FAST are isolated under its own FAST name-space. To help reduce the overall installation footprint of FAST, BioPerl dependencies of FAST scripts were analyzed with the Cava packager (http://www.cavapackager.com).

Nearly all FAST utilities process sequence records inline and therefore have linear runtime complexity in the number of sequences. Exceptions are fassort and fastail which both require some paging of data into temporary files. We performed benchmarking of FAST tools using randomly generated sequences of even composition sourced generated in Python and the Benchmark v1.15 Perl module on a MacBook Pro 2.5 Ghz Intel i7, with 8 Gb of RAM. We examined average CPU runtime over 100 replicates, comparing input sizes of 25K, 250K, or 1M sequence records of length 100, 10K, 100K, or 1M bp. Our benchmarking results show that despite data paging, `fassort` runtimes scale linearly with input size (Figure 5.2). FAST is not designed to be fastest at computing its solutions. Rather the fastness of FAST lies in how quickly an adept user can interactively prototype, develop, and express bioinformatic workflows with it.

5.3.4 Installation and Dependencies

FAST requires a working Perl installation, with official releases distributed through the Comprehensive Perl Archive Network (CPAN). A small footprint of BioPerl dependencies has been packaged together in the FAST namespace. Other CPAN dependencies may be detected and installed by the cpan package manager. A fully automated install from CPAN may on many systems be initiated by executing `perl -MCPAN -e ‘install FAST’`. A manual install follows standard Perl install procedure. After downloading and unpacking the source directory, change into that directory and execute: `perl Makefile.PL; make; make test; (sudo) make install`.

We recommend that first-time users first complete the automated install from CPAN which will handle prerequisites, and then download and open the source code directory in order to practice the example usage commands (such as those in the sequel) on sample data provided within.

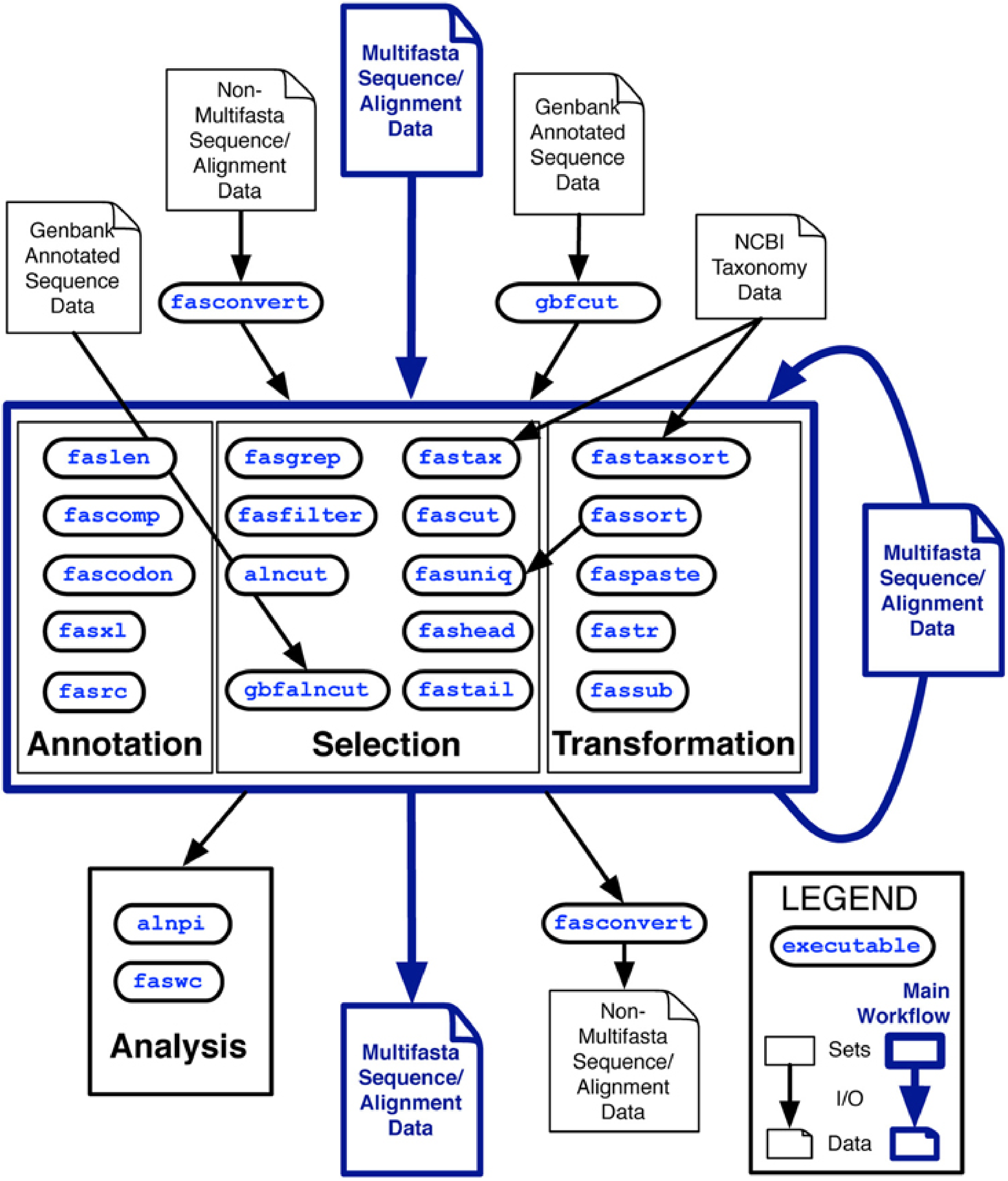

Figure 5.1: Overview of the first major release of FAST with data and workflow dependencies indicated. Inputs to FAST tools are shown at the top of the figure with outputs at the bottom. Outlined in blue is the primary working model, in which Multi-fastA sequence or alignment data is successively annotated, selected upon and transformed into new Multi-fastA data, or fed into a utility in the analysis category for tabular output of data summaries. Many of the utilities in the annotation category are also optionally capable of tabular output.

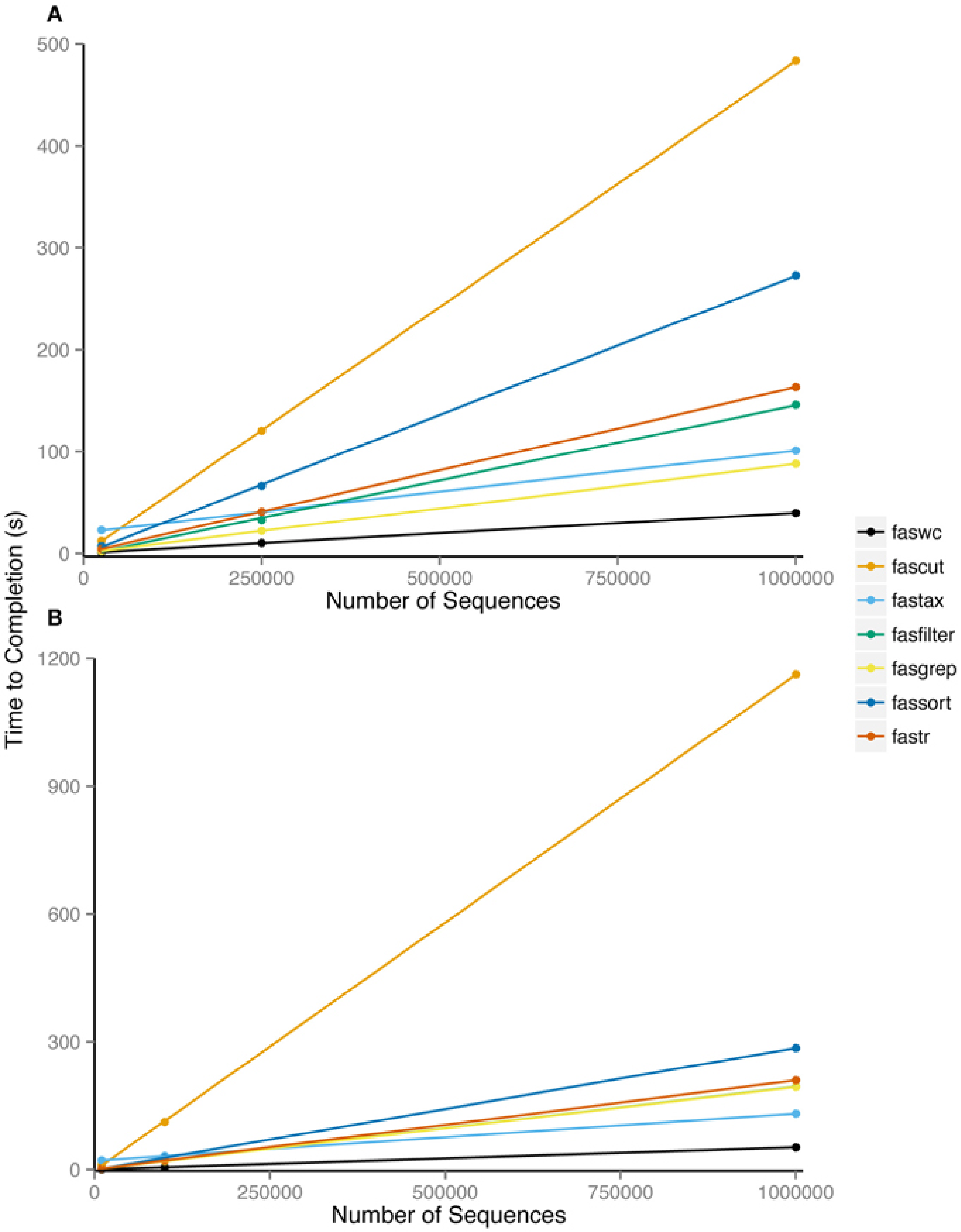

Figure 5.2: Average processor time of 100 repetitions required to complete analysis using indicated utility. Utilities were run on six datasets consisting of (A) 25,000, 2,50,000, and 10,00,000 100 bp sequences and (B) 10,000, 1,00,000, and 10,00,000 1000 bp sequences.

5.3.5 Implementation and Usage of Individual Tools

Further implementation and usage details of individual FAST tools follows. Usage examples for individual tools refer to example data that ships with the FAST source-code installer, available from CPAN. The most recent version at the time of publication is 1.06, available from http://search.cpan.org/~dhard/FAST-1.06/. However we recommend to use the most recent version of FAST. For maximum reproducibility, always cite the version number when publishing results with FAST. These usage examples should be able to run from within the installation directory after installation has completed.

fasgrep supports *regular expression*-based selection of sequence records. FAST uses Perl-style regular expressions, which are documented freely online and within Perl, and are closely related to Unix extended regular expressions. For reference on Perl regular expressions, try executing man `perlre` or `perldoc perlre`. For example, to print only protein sequences that do *not* start with M for methionine, execute:

```
fasgrep -s -v ‘‘^M’’ t/data/P450.fas
```

In the above command the `-s` option directs `fasgrep` to search the sequence data of each record. The v option directs fasgrep to print records that do not match the pattern given by its argument, which is the regular expression ˆM, in which the *anchor* “ˆ” specifies the beginning of the sequence data. `fasgrep` uses the BioPerl `Bio::Tools::SeqPattern` library to support ambiguity expansion of IUPAC codes in its regular expression arguments. Thus, to show that a segment of *Saccharomyces cerevisiae* chromosome 1 contains at least one instance of an “Autonomous Consensus Sequence” characteristic of yeast origins of replication (Leonard and McHali, 2013), look whether the following command outputs a sequence or not (note that all commands reproduced here should be entered on a single line at the Unix shell prompt):

```
fasgrep -se ‘WTTTAYRTTTW’ t/data/chr01.fas
```

which is equivalent to:

```
fasgrep -se ‘[AT]TTTA[CT][AG]TTT[AT]’ t/data/chr01.fas
```

These examples demonstrate queries on sequence data, but `fasgrep` may be directed to search against other parts of sequence records including identifiers, descriptions, fields and more.

fasfilter supports precise numerical-based selections of sequence records from numerical data in identifiers, descriptions, fields or tagged-values in descriptions. `fasfilter` supports *open ranges* such as `100-`, meaning “greater than or equal to 100,” closed ranges like `1e6-5e8` (meaning $1 * 10^6 to 5 * 10^8$) and compound ranges such as 200–400, 500-. Ranges may be specified in Perl-style (or GenBank coordinate style) like `from..to`, in R/Octave-style like `from:to` or UNIX cut-style as in `from-to`. For example, to print records with gi numbers between 200 and 500 million, try executing:

```
fasfilter -x ‘‘gi\|(\d+)’’ 2e8..5e8 t/data/P450.fas
```

This example uses the `-x` option which directs `fasfilter` to filter on the value within the capture buffer which occurs within the left-most pair of parentheses of the argument, here `(\d+)`, and `\d+` is a regular expression matching a string of one or more digits from 0 to 9. The backslash after `gi` in the first argument quotes the vertical bar character to make it literal, since the vertical bar character is a special character in regular expressions.

fascut supports index-based selections of characters and fields in sequence records allowing repetition, reordering, variable steps, and reversals. Ranges are specified otherwise similarly to `fasfilter`. Negative indices count backwards from last characters and fields. `fascut` outputs the concatenation of data selections for each sequence record. Variable step-sizes in index ranges conveniently specify first, second or third codon positions in codon sequence records, for example. Examples using this syntax appear in the sequel. To print the last ten residues of each sequence, execute:

```
fascut -10..-1 t/data/P450.fas
```

alncut implements content-based selection of sites in alignments including gap-free sites, non-allgap sites, variable or invariant sites and parsimoniously informative sites, or their set-complements, all with the option of state-frequency-thresholds applied per site. By default, `alncut` prints only invariant sites. To print the set-complement, or only variable sites, use the -v option:

```
alncut -v t/data/popset_32329588.fas
```

To print sites in which no more than two sequences contain gaps, execute:

```
alncut -gf 2 t/data/popset_32329588.fas
```

gbfcut allows annotation-based sequence-extraction from GenBank format sequence files, useful for extracting all sequences that correspond to sets of the same type of annotated features in genome data. For example, to output 5′ and 3′ Untranslated Region (UTR) sequences from a GenBank formatted sequence of a gene, we use the -k option to restrict matching to features whose “keys” match the regular expression “UTR”:

gbfcut -k UTR t/data/AF194338.1.gb

`gbfcut` can handle split features such as a coding region (CDS) that is split over several exons:

```
gbfcut -k CDS t/data/AF194338.1.gb
```

More fine-grained queries of features are possible using qualifiers defined with the `-q` option. Multiple qualifiers may be provided at once, specifying the selection of records for which all qualifiers apply (conjunction). For example, compare the output of the following two commands:

```
gbfcut -k tRNA t/data/mito-ascaris.gb
gbfcut -k tRNA -q product=Ser -q note^AGN t/data/mito-ascaris.gb
```

The second command queries for features with key “tRNA” containing at least one qualifier “/product” whose value matches the string literal “Ser” and no qualifiers of type “/note” whose values match the string literal “AGN”.

gbfalncut automates the selection of sites from alignments that correspond to one or more features annotated on one of the sequences in a separate GenBank record. This workflow eliminates the need for manual entry of coordinates and implements a useful bioinformatic query in terms of known and reproducible quantities from public data and sequence records, allowing users to query sites based on biological vocabularies of sequence features. For an example of its use see the section "Composing Workflows in FAST" in the sequel.

faspaste concatenates data from records input in parallel from multiple datastreams or files, record-by-record. The user may paste data from the standard input stream and from multiple input files, in an order defined by the arguments. Records from standard input may be used multiple times in concatenating data. Like in some implementations of the Unix tool `paste`, a hyphen input argument `-` to `faspaste` refers to the standard input stream and may be used more than once as an input argument. For maximum configurability, `faspaste` concatenates only one data field type (i.e., sequences or descriptions) at a time. Users may select which data stream will provide templates to receive concatenated data in output records. For example, to paste sequences of corresponding records from two data-files together and output them with the identifiers and descriptions of the data in the first file, execute:

```
faspaste data1.fas data2.fas
```

See the sequel for more advanced usage examples with `faspaste`.

fassort and **fasuniq** are designed to be often used together in Unix pipelines. The `fassort` utility implements numerical and textual sorting of sequence records by specific fields. The `fasuniq` utility removes (and optionally counts) records that are redundant with respect to a specific field, such as sequences or identifiers. In the implementation of `fassort`, pages of data are sorted with optimized routines in Perl `Sort::Key` that, if necessary, are written to temporary files and merged with `Sort::MergeSort`. Like its Unix Coreutil analog `uniq`, `fasuniq` compares only immediately successive input records. Therefore, users will usually want to first sort data with `fassort` before passing it to `fasuniq`. To illustrate, the following example combines and sorts input records from two instances of the same file, and then counts and removes each redundant record:

```
fassort -s t/data/P450.fas t/data/P450.fas | fasuniq -c
```

This example illustrates that the same file may be specified as an input stream more than once to any FAST command.

fastax and **fastaxsort** implement taxonomic searching and sorting of sequence records, whose records are already annotated with NCBI taxonomic identifiers using taxonomic data from NCBI taxonomy (Benson et al., 2009; Sayers et al., 2009). For example, a query of "Metazoa" would match records labeled "*Homo sapiens*", "*Drosophila melanogaster*", and "Lepidoptera" but not "*Candida albicans*" or "Alphaproteobacteria". Taxonomic selections may be logically negated and/or restricted to only those records containing valid NCBI taxonomic identifiers. Purely for historical reasons, the internal implementation of NCBI taxonomic data is custom to FAST rather than the `Bio::Taxonomy` libraries in BioPerl. A sample of data from tRNAdb-CE (Abe et al., 2014), in which data records are annotated with valid NCBI taxonomic identifiers in specific description fields, is included with the FAST installation package. After downloading datafiles "nodes.dmp" and "names.dmp" from NCBI Taxonomy, the following command filters sequences from Rhizobiales, assuming that records are labeled with their species (and strain) of origin in the third field of the description of the sample data file:

```
fastax -f 3 -S ‘‘ \| ’’ nodes.dmp names.dmp Rhizobiales \
t/data/tRNAdb-CE.sample2000.fas
```

fastr and **fassub** handle, respectively, character- and string-based transformations of sequence records. The utility fastr handles character-based transliterations, deletions and “squashing” (deletion of consecutive repeats), sequence degapping, and restriction or remapping of sequence data to strict or IUPAC ambiguity alphabets. For example, to lower-case all sequence characters, execute:

```
fastr -s ‘A-Z’ ‘a-z’ t/data/P450.fas
```

Degapping requires only the simple command:

```
fastr --degap t/data/P450.clustalw2.fas
```

The utility `fassub` allows more arbitrary substitutions on sets of strings matched to Perl regexes, implemented through direction of the Perl `s///` substitution operator on specific fields. Capture buffers may be used to refer to matched data in substitutions, for example, to reverse the order of genus and species in a file in which scientific names occur in descriptions enclosed with square brackets:

```
fassub -d ‘(\w+)(\w+)’ ‘[21]’ t/data/P450.fas
```

fascomp, **fasxl** and **fascodon** provide for annotation and analytics of compositions, translations, and codon usage frequencies of sequence records (with start and stop codons counted distinctly, in the last case). All genetic codes included in BioPerl, ultimately from NCBI Entrez, are supported.

alnpi outputs molecular population genetic statistics cited in Table 5.2 for each alignment on input. It can output a set of statistics for each alignment on input in plain text or LaTeXformat. `alnpi` also supports sliding window and pairwise analysis of input data. Data and command examples are provided to reproduce the tables and sliding window analyses of statistics published in Ardell et al., 2003. Purely for historical reasons, `alnpi` does not use the perlymorphism routines in the BioPerl library `Bio::PopGen` (Stajich et al., 2002). However, all of the code for these calculations has been reviewed and compared against calculations produced from DNASP (Librado and Rozas, 2009b) as described previously (Ardell, 2004).

5.4 Composing Workflows in FAST

Here we show how to interactively prototype a pipeline that computes the sliding window profile of Tajima’s *D* of Figure 4A in Ardell et al., 2003 from a publicly available datafile. The datafile associated to this figure is an NCBI PopSet with accession ID 32329588 containing an alignment of a fully annotated ciliate gene (accession AF194338.1) against several partially sequenced allelic variants. One of the variants with accession ID AY243496.1 appears to be partly non-functionalized. First to see this data, we view it in the pager less (press “q” to quit and “space” to page):

```
less t/data/popset_32329588.fas
```

A key feature of the Unix shell allows users to recall previous commands in their so-called history, usually by typing the “up-arrow”, for possble re-use and editing. To

Table 5.2: Molecular population genetic statistics in FAST.

Statistic	Symbol	Citation
Number of sequences	n	
Number of alleles/distinct sequences	k	
Number of segregating sites	S	
Fraction of segregating sites	s	
Average number of pairwise differences		(Nei and Li, 1979)
Nucleotide diversity	π	(Nei and Li, 1979)
Watterson estimator	θ_W	(Watterson, 1975)
Expected number of alleles	$E(K)$	(Ewens, 1972)
Tajima's D	D	(Tajima, 1989)
Fu and Li's D*	D^*	(Fu and Li, 1993)
Fu and Li's F*	F^*	(Fu and Li, 1993)
		(Simonsen et al., 1995)
Fu and Li's Eta S	η_S	(Fu and Li, 1993)
Fu and Li's Eta	η	(Fu and Li, 1993)

check the number of sequences and characters in the alignment, execute:

```
faswc t/data/popset_32329588.fas
```

To compute our population genetic statistics we wish to remove the annotated reference sequence, the deactivated allele, and one additional sequence from analysis, which we can do using `fasgrep`, and verify that it reduced data by the correct number of records (six) by piping to `faswc` (the command is broken over two lines here but may be entered as one line on the Unix prompt):

```
fasgrep -v ‘‘(AF194|349[06])’’ t/data/popset_32329588.fas | faswc
```

We can check the identifier lines by modifying the end of this pipeline:

```
fasgrep -v ‘‘(AF194|349[06])’’ t/data/popset_32329588.fas | grep \>
```

Sequencing ambiguities and gap-characters can introduce noise and uncertainty in the execution and documentation of bioinformatic workflows. For some computations, for example in molecular population genetics, one may want to be conservative and remove ambiguity- and gap-containing sites from an alignment. We can check for ambiguities in our data by outputing a composition table:

```
fasgrep -v ‘‘(AF194|349[06])’’ t/data/popset_32329588.fas | \
fascomp --table
```

To remap ambiguities to gap characters, with the intent of removing all sites containing either ambiguities or gaps, we may use `fastr` to remap all non-strict DNA characters to gap `(-)` and verify the result using `fascomp` again:

```
fasgrep -v ‘‘(AF194|349[06])’’ t/data/popset_32329588.fas | \
fastr --strict -N - | fascomp --table
```

Now, with confidence in our remapping, we extract exclusively gap-free sites from the alignment using `alncut`, and verify that we reduced alignment size with `faswc`:

```
fasgrep -v ``(AF194|349[06])'' t/data/popset_32329588.fas | \
fastr --strict -N - | alncut -g | faswc
```

Finally, we pass the verified pipeline output to `alnpi` for sliding-window analysis of Tajima's D in overlapping windows of width 100 and step size 25:

```
fasgrep -v ``(AF194|349[06])'' t/data/popset_32329588.fas | \
fastr --strict -N - | alncut -g | alnpi --window 100:25:d
```

5.5 Further FAST Workflow Examples

5.5.1 Selecting Sites from Alignments by Annotated Features

Another example, that reproduces a published result from (Ardell et al., 2003), demonstrates the utility of combining `gbfalncut` with `alnpi`, allowing users to select sites from alignments corresponding to features annotated on one of the sequences in a separate GenBank file. For example, to calculate a Tajima's D statistic for 5′ UTRs, corresponding to the the last line in Table 5.1 of that work, execute:

```
gbfalncut -k t/data/AF194338.1.gb 5.UTR t/data/popset_32329588.fas \
| fasgrep -v ``(AF194|349[06])'' | fastr --strict -N - | alncut -g \
| alnpi
```

5.5.2 Selecting Sequences by Encoded Motifs

An advantage of the annotation approach in FAST is the ability to select and sort sequences by attributes computed and annotated into data by utilities upstream in the pipeline. For example, to select protein-coding genes from a file `cds.fas` whose translations contain the N-glycosylation amino acid motif (Kornfeld and Kornfeld, 1985), one could execute:

```
fasxl -a cds.fas | fasgrep -t xl0 ``N[^P][ST][^P]'' | \
fascut -f 1..-2
```

The first command in the pipeline translates each sequence and appends the translation to the description with the tag "xl0" (indicating translation in the zeroth reading frame). The second command in the pipeline uses a regular expression to represent the N-glycosylation amino acid motif pattern as the value of a "name:value" pair in the description with tag "xl0", hence processing the annotations produced by `fasxl`. The regex argument to `fasgrep` is quoted to protect the argument from interpretation by the shell. The last command in the pipeline removes the last field in the description, restoring records as they were before they were annotated by `fasxl`.

5.5.3 Sorting Records by Third Codon Position Composition

Another example illustrates the powerful expression of ranges in `fascut`. An optional "by" parameter in ranges allows increments or decrements in steps larger than one.

To extract third-position bases from codon sequence records, compute and annotate their compositions into record descriptions, ultimately sorting records by their third-position adenosine contents, do:

```
fascut 3:-1:3 cds.fas | fascomp | fassort -nt comp_A
```

5.5.4 More Advanced Merging of Data Records

More advanced usage of `faspaste` requires Unix pipelines. For example to join both descriptions and sequences from two data-files, execute:

```
faspaste data1.fas data2.fas | faspaste -d - data2.fas
```

The hyphen second argument `(-)` to the second instance of `faspaste` refers to the input received from standard input through the pipe. This example works because by default, `faspaste` uses ("mutates") records from the data stream named in its first argument to receive the data concatenated from all records.

To prepend the first sequence of one file repeatedly to every sequence in another file, execute:

```
fashead -n 1 t/data/fasxl_test4.fas | faspaste -r - \
t/data/fasxl_test4.fas
```

To prepend the first sequence of one file repeatedly to every other sequence in another file, using identifiers and descriptions from the second file in the output, execute:

```
fashead -n 1 t/data/fasxl_test3.fas | faspaste -r -R 2 - \
t/data/fasxl_test4.fas
```

5.6 Further Documentation and Usage Examples

Upon installation, FAST generates and installs a complete `man` page for each FAST utility, which should be accessible by one or both of the following commands:

```
man fasgrep
perldoc fasgrep
```

In addition, a FAST Cookbook has been contributed by the authors and is available with the source code distribution or from the project GitHub repository at https://github.com/tlawrence3/FAST.

5.7 Concluding Remarks and Future Directions

Planned additions in future versions of FAST include `fasrand` and `alnrand` for automated sampling, permutations and bootstrapping of sequences and sites, respectively, and `fasgo` and `fasgosort` for selection and sorting of records by Gene Ontology categories (The Gene Ontology Consortium, 2015).

5.8 References

Ardell, D. H. (2004). SCANMS: adjusting for multiple comparisons in sliding window neutrality tests. *Bioinformatics* **20**:12, pp. 1986–1988. DOI: 10.1093/bioinformatics/bth187

Ardell, D. H., Lozupone, C. A., and Landweber, L. F. (2003). Polymorphism, Recombination and Alternative Unscrambling in the DNA Polymerase alpha gene of the ciliate *Stylonychia* lemnae (Alveolata; class Spirotrichea). *Genetics* **165**:4, pp. 1761–1777

Baggerly, K. A. and Coombes, K. R. (2009). Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology. *The Annals of Applied Statistics* **3**:4, pp. 1309–1334

Baggerly, K. A. and Coombes, K. R. (2011). What Information Should Be Required to Support Clinical "Omics" Publications? *Clinical Chemistry* **57**:5, pp. 688–690. DOI: 10.1373/clinchem.2010.158618

Barnes, N. (2010). Publish your computer code: it is good enough. *Nature* **467**:7317, p. 753

Blankenberg, D. and Hillman-Jackson, J. (2014). "Analysis of Next-Generation Sequencing Data Using Galaxy". *Stem Cell Transcriptional Networks*. Ed. by B. L. Kidder. Vol. 1150. Methods in Molecular Biology. Springer New York, pp. 21–43. DOI: 10.1007/978-1-4939-0512-6_2

Boulesteix, A.-L. (2010). Over-optimism in bioinformatics research. *Bioinformatics* **26**:3, pp. 437–439. DOI: 10.1093/bioinformatics/btp648

Bradnam, K. and Korf, I. (2012). *UNIX and Perl to the Rescue!: a Field Guide for the Life Sciences (and Other Data-rich Pursuits)*. Cambridge: Cambridge University Press

Casadevall, A., Steen, R. G., and Fang, F. C. (2014). Sources of error in the retracted scientific literature. *The FASEB Journal* **28**:9, pp. 3847–3855. DOI: 10.1096/fj.14-256735

Cunningham, F. et al. (2015). Ensembl 2015. *Nucleic Acids Research* **43**:D1, pp. D662–D669. DOI: 10.1093/nar/gku1010

Delaglio, F. et al. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR* **6**:3, pp. 277–293

Delescluse, M. et al. (2012). Making neurophysiological data analysis reproducible: Why and how? *Journal of Physiology-Paris* **106**:3-4, pp. 159–170. DOI: 10.1016/J.JPHYSPARIS.2011.09.011

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology* **3**:1, pp. 87–112

Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**:3, pp. 693–709

Garlan, D. and Shaw, M. (1994). An introduction to software architecture. *Comput. Sci. Dep.*

Gordon, A. (2009). FASTX Toolkit

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, pp. 221–224. DOI: 10.1093/molbev/msp259

Huang, Y. and Gottardo, R. (2013). Comparability and reproducibility of biomedical data. *Briefings in Bioinformatics* **14**:4, pp. 391–401. DOI: 10.1093/bib/bbs078

Hutson, S. (2010). Data handling errors spur debate over clinical trial. *Nature medicine* **16**:6, p. 618

Ioannidis, J. P. A. et al. (2008). Repeatability of published microarray gene expression analyses. *Nat Genet* **41**:2, pp. 149–155

Joppa, L. N. et al. (2013). Troubling Trends in Scientific Software Use. *Science* **340**:6134, pp. 814–815. DOI: 10.1126/science.1231535

Knuth, D. E. (1984). Literate programming. *The Computer Journal* **27**:2, pp. 97–111

Kornfeld, R. and Kornfeld, S. (1985). Assembly of Asparagine-Linked Oligosaccharides. *Annual Review of Biochemistry* **54**:1, pp. 631–664. DOI: 10.1146/annurev.bi.54.070185.003215

Li, H. et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352

Librado, P. and Rozas, J. (2009a). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**:11, pp. 1451–1452. DOI: 10.1093/bioinformatics/btp187

Librado, P. and Rozas, J. (2009b). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**:11, pp. 1451–1452. DOI: 10.1093/bioinformatics/btp187

Lipman, D. J. and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**:4693, pp. 1435–1441

Lushbough, C. M., Jennewein, D. M., and Brendel, V. P. (2011). The BioExtract Server: a web-based bioinformatic workflow platform. *Nucleic Acids Research* **39**:suppl 2, W528–W532. DOI: 10.1093/nar/gkr286

Markowitz, V. M. et al. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Research* **42**:D1, pp. D560–D567. DOI: 10.1093/nar/gkt963

McIlroy, D. (1969). “Mass-produced software components”. *Proceedings of the 1st International Conference on Software Engineering.* Ed. by J. Buxton, P. Naur, and R. Randell. Garmisch-Pattenkirchen, pp. 138–155

Morin, A. et al. (2012). Shining Light into Black Boxes. *Science* **336**:6078, pp. 159–160. DOI: 10.1126/science.1218263

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**:10, pp. 5269–73

Oinn, T. et al. (2006). Taverna: lessons in creating a workflow environment for the life sciences. *Concurrency and Computation: Practice and Experience* **18**:10, pp. 1067–1100. DOI: 10.1002/cpe.993

Peek, J. (2001). *Why Use a Command Line Instead of Windows?* http://www.linuxdevcenter.com/pub/a/linux/2001/11/15/learnunixos.html

Peng, R. D. (2009). Reproducible research and Biostatistics. *Biostatistics* **10**:3, pp. 405–408. DOI: 10.1093/biostatistics/kxp014

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science* **334**:6060, pp. 1226–1227. DOI: 10.1126/science.1213847

Rampp, M., Soddemann, T., and Lederer, H. (2006). The MIGenAS integrated bioinformatics toolkit for web-based sequence analysis. *Nucleic acids research* **34**:Web Server issue, W15–9. DOI: 10.1093/nar/gkl254

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16**:6, pp. 276–277. DOI: 10.1016/S0168-9525(00)02024-2

Rosenbloom, K. R. et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Research* **43**:D1, pp. D670–D681. DOI: 10.1093/nar/gku1177

Seemann, T. (2013). Ten recommendations for creating usable bioinformatics command line software. *GigaScience* **2**:1, p. 15. DOI: 10.1186/2047-217X-2-15

Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**:1, pp. 413–429

Smith, S. W. et al. (1994). The genetic data environment an expandable GUI for multiple sequence analysis. *Computer applications in the biosciences : CABIOS* **10**:6, pp. 671–5

Stajich, J. E. et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome research* **12**:10, pp. 1611–1618. DOI: 10.1101/gr.361602

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques* **28**: pp. 1102–1104

Stutz, M. (2000). *Linux and the Tools Philosophy.* http://www.linuxdevcenter.com/pub/a/linux/2000/07/25/LivingLinux.html

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:3, pp. 585–595

The Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Research* **43**:D1, pp. D1049–D1056. DOI: 10.1093/nar/gku1179

Villensen, P. (2007). FaBox: an online toolbox for fasta sequences. *Molecular Ecology Notes* **7**:6, pp. 965–968. DOI: 10.1111/j.1471-8286.2007.01821.x

Waterhouse, A. M. et al. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**:9, pp. 1189–1191. DOI: 10.1093/bioinformatics/btp033

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical population biology* **7**:2, pp. 256–76

Wilson, G. (2014). Software Carpentry: lessons learned. *F1000Research* **3**: DOI: 10.12688/f1000research.3-62.v1

Yates, A. et al. (2015). The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics* **31**:1, pp. 143–145. DOI: 10.1093/bioinformatics/btu613

Chapter 6

Conclusion

6.1 Scientific Impact

In this dissertation, we have successfully demonstrated that plastid tRNA CIFs contain phylogenetic information about deep evolutionary relationships that can be detected using machine-learning and distance-based methodology. In Chapter 2, we were successful in training a machine-learning algorithm to accurately classify Cyanobacterial genomes and applied this algorithm to test hypotheses of the origin of plastids. We found strong support for the origin of plastids among a late-branching clade of starch-producing marine/freshwater diazotrophic cyanobacteria. This has added to our understanding of this early evolutionary event that had a profound impact on the Tree of Life. In Chapter 3, we provided evidence from a novel source supporting a sister relationship between conifers and gnetophytes helping resolve the seed plant phylogeny. In Chapter 4, we describe the implementation of an open-source command line application for predicting tRNA CIFs from genomic sets of tRNAs, testing statistical significance of tRNA CIFs, and quantifying similarities in tRNA CIFs between genomic datasets. Currently, this is the only program for predicting and analyzing tRNA CIFs, greatly increasing the accessibility of these methods. This will hasten the discovery and experimental validation of tRNAs CIFs that are relevant to several fields of scientific inquiry including the development of therapeutic drugs targeting the tRNA interaction network. Lastly, in Chapter 5 we introduce a suite of open-source command line tools for processes and analyzing biological sequence data that encourages reproducible bioinformatic workflows.

6.2 Next Steps

In Chapter 2, increasing the sampling of the smaller Cyanobacterial clades would likely improve the accuracy of classifying genomes from these clades. The majority of misclassified genomes belong to these smaller clades, consequently any improvement of their classification would greatly increase the overall accuracy of CYANO-MLP. However, the small clades might be a reflection of low species diversity making deeper sampling impossible. Another approach would be to construct a reduced, but representative phylogenomic dataset that we could analyze with a sophisticated phylogenetic model that can fully describe the systematic biases of Cyanobacteria/plastid datasets. We are currently testing the computational feasibility of several datasets with the phylogenetic model CAT-BP.

In Chapter 3, the conclusions would be stronger if we could have included *Ginkgo biloba* in the phylogenetic analysis, because of its position as sister to cycads. However, the extant diversity of only a single species within the *Ginkgo* clade makes it impossible to predict plastid tRNA CIFs using our methods. We could include *G. biloba* if we predicted tRNA CIFs from the nuclear genome, which tends to contain enough tRNAs for CIF prediction. Unfortunately, nuclear genomes are available for all seed plant clades except Cupressales, which is pivotal for testing competing hypotheses of the location of gnetophytes.

In Chapter 4, we plan on simplifying the process for adding custom distance metrics to tsfm for quantify similarity between datasets. For Chapter 5, we are porting the suite of utilities from perl to C++ to provide significant speed improvements. This will increase the applicability of the FAST utilities to gigabase and terabase datasets.