THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF BIOLOGY

DATING THE EVOLUTION OF PROKARYOTES

JAANKI DAVE
SPRING 2015

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees in Biology and Biological Anthropology
with honors in Biology

Reviewed and approved* by the following:

S. Blair Hedges
Professor of Biology
Thesis Supervisor

Richard Cyr
Professor of Biology
Faculty Reader

Stephen W. Schaeffer
Professor of Biology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

While many studies have been published regarding the divergence times of eukaryote species, the same cannot be said about prokaryote species. Thus, the full picture of the evolution of prokaryotes has not yet been revealed. The European consortium SILVA first published an undated tree ("All-Species Living Tree") of prokaryotes in 2008 from 16S rRNA data. Using the March 2014 release of the All-Species Living Tree, calibrations used in recent astrobiological research, and a new timing method published last year, this project aimed to compile a timetree of prokaryote evolution while comparing different timing methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Blair Hedges and Dr. Julie Marin for all of their guidance and patience throughout this project. I would also like to thank Oscar Murillo of Arizona State University for all of his technical support.

Chapter 1

Introduction

Since Darwin postulated the tree of life, many phylogenetic studies have been conducted on this concept and resulted in a working timetree of life (1) that continues to be developed. A timetree shows estimates for nodal dates (the time of the last common ancestor) as well as branching topology (2). Timetrees are valuable tools in understanding the integration of large molecular data sets with fossil and biogeographic evidence, estimating ancestral character states and evolutionary rates, the association between causal historical processes and biological outcomes, and the development of a universal scheme for biological classifications. (2)

While eukaryotes have been widely represented in this tree, with 1378 families (1) Eubacteria and Archaebacteria have not been studied as comprehensively, due to complications such as horizontal gene transfer. Many phylogenetic trees of eukaryote species can be rooted easily with a more distantly related species—something that poses a challenge when creating a timetree for Eubacteria and Archaebacteria, which both diverged more deeply in the tree of life.

Adding a timescale to development of current species has been done through the use of molecular clocks. The molecular clock method estimates divergence time of lineages based on rates of change in nucleotide (or protein) sequences (3). Sequences change at different rates—this allows for the use of molecular clock methods at different timescales (3). Relaxed clock methods accommodate lineage specific rate variation and allow for more species and genes to be included. The relaxed clock methods also allow for the use of maximum and minimum calibrations when estimating divergence time. This is useful especially when using fossil

calibrations, which are always minimum time estimates, leading to minimum divergence times (1).

However, molecular clock methods are used infrequently for dating bacteria for several reasons. The first is that it is difficult to establish a reliable phylogeny. Horizontal gene transfer, in which genetic material is transferred between species, is an important force that can complicate bacterial phylogeny. Vertical descent passes genes through processes such as recombination in sexually reproducing populations (4). Horizontal transfer, on the other hand, does not follow this pattern, and thus influences molecular phylogenies by exchanging genetic information across species, muddling distinct genetic differences between them. Phylogenies generated for different molecules may agree broadly, but are not usually completely identical (5). Lateral gene transfer is found in different types of genes, from core metabolic functions, conserved biosynthetic pathways, and transcription and translation machinery (5). Ribosomal RNA (rRNA) is also not immune to this form of exchange. Strong functional constraints, though, leave stretches of conserved sequence in rRNA (5). Ribosomal RNA, as a mosaic, can reflect the mixed character of the entire genome (5).

There is also difficulty in selecting reliable calibration points. Unlike in eukaryotes, there is limited fossil evidence for the early history of bacteria. Early prokaryotes would have existed together in interacting groups to form biofilms (6). Ancient biofilms that are found in microbially induced sedimentary structures or stromatolites however, have not changed significantly through 3.5 billion years (6). Many fossils used to determine calibration points are from the Pharenerozoic eon (7), a time that is too recent to be considered useful in timing bacterial evolution.

Rooting the common ancestor of Archaebacteria and Eubacteria is complex. Many previous studies have assumed that prokaryotes evolved in the Precambrian (3).Major prokaryote groups existed by the Neoproterzoic (1000-550 million years ago), indicating that their major metabolic activities (anoxygenic photosynthesis, oxygenic photosynthesis, methanogenesis, and aerobic methanotrophy) (3) were functional at that time. Therefore, biomarkers based on the geological record, such as the proliferation of oxygen producing plants, can be used.

Although rRNA is also subject to lateral gene transfer, several studies have found that for classification purposes the small subunit (SSU) may be the most parsimonious and accurate way to establish genealogical relationships (8). Thus, rRNA is used to identify bacteria, assign its taxonomy and conduct phylogenetic analysis of its diversity (9).

The goal of this project was to create a prokaryote timetree describing the time of divergence of prokaryote species) by dating the SSU rRNA phylogenetic tree of approximately 10,000 species of bacteria.

Chapter 2

Current Knowledge

After life began on Earth, the superkingdoms of Eubacteria and Archaebacteria split approximately 4200 million years ago, according to one molecular clock analysis (10). Life on Earth has a monophyletic (single) origin, with evidence in the great similarity in genomes of all organisms, and the fact that cells seeded from planetary ejecta would probably not survive impact shock, heat and radiation of passing through the atmosphere (10). Endosymbiosis, in which certain organelles found in eukaryotic cells have prokaryotic origins, adds to the blending of prokaryotic and eukaryotic genomes (10).

As noted by Zuckerkandl and Pauling (11), hemoglobins change relatively constantly, as seen with amino acid substitution rates of different species and dates from fossils. The differences between the DNA of two separate species could be represented as a function of time since their divergence. The concept of the 'molecular evolutionary clock' was thus created. Most of these differences are selectively neutral, meaning they do not impact fitness of the organism. Since the genetic code is redundant with many codons encoding the same amino acid, many changes do not impact function. Kimura's neutral theory of molecular evolution states that most evolutionary changes at the molecular level are not due to natural selection, but rather through random drift of neutral mutants (12).

The number of changes in a sequence versus time changes as a linear relationship because of multiple single site mutations (13). This can lead to homoplasies (traits found in

multiple species due to other causes besides common ancestry (14)), and so to an underestimation of differences and distance between distantly related sequences (13). However, larger data sets (compared to number of nucleotides) and adequate models of sequence evolution can overcome this and generate improved phylogenies (1).

Individual gene trees are influenced by paralogy (genes descending from a common ancestor that have resulted from duplication and have diverged since then (15)), gene loss, lineage sorting and lateral gene transfer. Supertree processes have been used to generate a single tree representative of many input trees to avoid the issues listed above.

Very few timetrees are available (Battistuzzi et al 2004, Battistuzzi and Hedges 2009, Battistuzzi 2009, Jun et al 2010, and Hedges 2014) to investigate the evolution of prokaryotes. Methods for constructing these timetrees and their results are discussed in detail below.

Sheridan et al 2010 (13) investigated bacterial evolution of a small number of species. Use of SSU rRNA shows genetic diversity better than protein sequences. SSU rRNA is not influenced by the selective pressures shaping the protein sequences. Lateral transfer of SSU rRNA is also rare (13).

Times were estimated using average cyanobacterial distance and time estimation obtained from 2-methylhopanoid molecular fossils (13). Average cyanobacterial distance was determined by calculating the “average distance between two chloroplast sequences and common ancestor of all cyanobacteria” and the “average distance between three free-living cyanobacterial sequences and common ancestor of cyanobacteria” (13). The tree was rooted with the last common ancestor of Bacteria and Archaea.

A time estimate of cyanobacterial lineage in 2-methylhopanoid molecular fossils yielded a date of 2.65 Ga. The last common ancestor of Bacteria, Archaea, and Eurkarya was dated back to 4.29 Ga using cyanobacterial calibrations.

Phylogenetic relationships between bacteria were also investigated by Battistuzzi, Feijao and Hedges 2004 (16) by using amino acid sequences from 32 common proteins. Most proteins used were information storage and process proteins, with others classified for use in cellular processes and metabolism (16). Calibrations were determined using node constraints such as the age of Earth, the Solar System and the Great Oxidation Event, the proliferation of O_2 in Earth's atmosphere (16). Archaebacteria and Eubacteria were dated separately. The date of the last common ancestor could not be determined. A small number of duplicated genes were used to root the tree of life.

Results indicated that fossil calibrated dates were younger than molecular calibration. By comparing branch lengths of cyanobacteria in a protein tree and a 16S rRNA tree, no obvious bias or rate change was found. The discrepancy found between molecular and fossil dates was unclear. Hyperthermophiles were shown to branch basally—perhaps due to their high G-C concentrations (16).

Using a core protein tree from developed from 25 protein coding genes from 218 species, Battistuzzi and Hedges 2009 (17) further explored bacterial evolution. Besides the core protein tree, an rRNA tree of 189 species was also constructed. This rRNA tree combined sequences from the small and large subunit of the ribosome using a modified LogDet analysis to correct for the GC bias.

To estimate the divergence times, calibration points were taken from geologic and biomarker records. The tree's root was set at 4.2 Ga, the earliest habitable time for life based on

ocean boiling impact probabilities. The earliest continents formed 4.0 Ga. The earliest methanogens were estimated at 3.46 Ga and the earliest oxygen at 2.3 Ga. The divergence between Cholorbia and Bacteroidetes as well as the divergence of Gammaproteobacteria and Betaproteobacteria was found to be 1.64 Ga. This study also revealed a major dichtomoy (Terrabacteria and Hydrobacteria) in Eubacteria.

Further studies by Battistuzzi and Hedges were included in the book *The Timetree of Life* (7, 18). Separate analyses were conducted for Eubacteria and Archaebateria. The Eubacterial timetree had 197 species and was constructed using ML phylogeny. The origin of Chromiataceae and the divergence of Chlorobi and Bacteroidetes, both dated at 1640 Ma from biomarker evidence, were used as calibration points. The last ocean-vaporizing event, 4200 Ma, was also used as a calibration point. Most divergences in Eubacteria were found to be closely spaced in time.

A Bayesian timing method was used to generate the Archaebacterial tree. 12 families and 1 phlyum were used. The origin of methanogenesis 3460 Ma was used as a calibration point. The first divergence based on the last ocean vaporizing event about 4200 Ma was used as another calibration point.

Chapter 3

Relevance

Construction of a timetree of Eubacteria and Archaebacteria is a valuable addition to the Time Tree of Life (1). It also has the potential to contribute astrobiology. The first organisms on Earth were prokaryotes and their evolutionary history can reveal the geological and climatological processes that occurred in the early history of our planet.

The time of symbiotic events can also be estimated more accurately. Sheridan et al 2010 mentions the endosymbiosis of mitochondria, when eukaryotic cells captured alpha-proteobacterium, as well as the endosymbiosis of chloroplast, when eukaryotic cells captured cyanobacterium (13). With these advances, cells could use and produce oxygen, leading to the current state of life.

Results obtained by Battistuzzi, Feijiao and Hedges 2004 (16) were consistent with methane greenhouse theory. The time estimate between 4.11 Ga and 3.78 Ga suggests methanogens were on Earth during the Archean Period. The common ancestor of Eubacteria groups was phototropic. As the phototropic metabolism evolved, it was passed along via horizontal gene transfer. The common ancestor of Eubacteria groups had the machinery and genetic capability to use light as an energy source. The ability to photosynthesize, however, is restricted to Eubacteria within prokaryotes (16). These groups include proteobacteria, greens sulfur bacteria, green filamentous bacteria, gram positive heliobacteria and cyanobacteria.

Colonization of land happened many times in many lineages. Terrabacteria includes Actinobacteria, Cyanobacteria, Firmicutes and *Deinococcus-Thermus*. The first steps of gaining

oxygenic photosynthesis had acquisition of productive pigments that dealt with the stress of desiccation and solar radiation (17). Since ultraviolet radiation can cause biological damage, organisms that survive successfully on land need to develop pigments to survive (17). Carotenoids are an example of a photoprotective pigments developed against reactive UV-light created oxygen species (17). Actinobacteria, Cyanobacteria, Firmicutes and *Deinococcus-Thermus* are also highly resistant to dehydration, an extremely important trait when living in a non-aquatic environment.

The generation of a bacterial timetree therefore has many impacts not only in the field of evolutionary biology, but also in astrobiology and knowledge related to the early history of the Earth.

Chapter 4

Methods

For this project, the March 2014 release of the 'All Species Living Tree' was used (19). This version, with 10,271 Eubacteria and Archaebacteria species, also contained updated species accession numbers. 597 new species were added and 93 species were omitted because they had been submitted to EMBL after Silva115 had been released in August 2013.

The SILVA consortium used small subunit ribosomal RNA sequences of at least 900 bases from 10,271 bacteria species to generate a large phylogenetic tree (9). Use of SSU rRNA was important for inferring monophyly (9). There was concern that a single gene may not be representative of the prokaryote phylogeny. Crossover or horizontal gene transfer could blur results. SSU sequences are the "gold standard" for reconstructing phylogenetic relationships.

The sequences were obtained from the SILVA database and EMBL. Usable sequences were chosen if they fulfilled the minimum standards for the SSUParc database. Nearly full length sequences were used and compared to validly published species, with duplicates being deleted (9). SSU alignment was used to reconstruct phylogeny, using both the primary gene sequence and secondary structure (based on nucleotide pairing on functional helices) (9). Sequences were then aligned using SINA and manually checked for the best available SSU/LSU entry of a species and correct taxonomic assignment. Secondary structure was used for optimal placement- this defined which pairs of alignment columns would be expected to bond (9). The genealogy was then created based on the maximum likelihood algorithm RAxML which utilizes a many model approach (20).

To estimate the divergence times, three programs were used in this project: MEGA RelTime, treePL, and PATHd8 (21, 22, 23). Calibration points (Table 1), as well their minimum and maximum times (Table 2), were determined from times established by Battistuzzi and Hedges 2009. These were then assigned to corresponding twenty-seven nodes (node numbers in Table 1 and Table 2) in the SILVA tree. These calibration points were then used in all of the analyses run.

Table 1. Calibration nodes used (from Battistuzzi and Hedges 2009 (7, 18))

Node Number	*Node*	*Calibration Time (mya)*
1	Enterobacteriaceae/Pasteurellaceae	634
2	Vibrionaceae/ (1)	751
3	Coxiellaceae/Legionellaceae	1420
4	Comamondaceae/ (Burkholderia/Alcaligenaceae)	872
5	Rhodocyclaceae/(4)	1028
6	Nitrosomonadaceae/Hydrogenophilaceae	1055
7	Rhizobiaceae/Phyllobacteriaceae	509
8	Rhodobacteraceae/ (Caulobacteraceae/(Bradyrhizobiaceae/((Bartonellaceae/Brucellaceae)/(7))))	1498
9	Erythrobacteraceae/(8)	1613
10	Rhodospirillaceae/Acetobacteraceae	1432
11	Campylobacteraceae/Helicobacteraceae	1104
12	Bacteroidaceae/Porphyromonadaceae	616
13	Chlorobiaceae/(Crenotriachaceae/(12))	2099
14	Deinococcus-Thermus/Actinobacteria	2739

15	Cellulomonadaceae/Mircobacteriaceae	937
16	(Actinobacteria/Deinococcus-Thermus)/((Cyanobacteria/Chloroflexi)/(Firmicutes))	2908
17	Lactobacillaceae/Streptococcaceae	1392
18	Mycoplasmataceae/Entomoplasmataceae	523
19	Acholeplasmataceae/(18)	1860
20	Thermotogaceae/(Eubacteria)	4179
21	Methanosarcinaceae/Methanospririllaceae	2216
22	Halobacteriaceae/(21)	2430
23	Archeoglobaceae/(22)	2799
24	Thermoplasmataceae/Pircophilaceae	992
25	Methanocaldococcaceae/Methanococcaceae	1676
26	Methanopyraceae/(Methanobacteriaceae/(24))	3313
27	Crenarchaeota/Euryarchaeota	4187

Table 2. Calibration nodes used with minimum and maximum times (from Battistuzzi and Hedges 2009 (7,18))

Node Number	*Calibration Time (mya)*	*Minimum Calibration Time (mya)*	*Maximum Calibration Time (mya)*
1	634	538	739
2	751	653	854
3	1420	1297	1539
4	872	759	993
5	1028	911	1150
6	1055	932	1184

7	509	421	609
8	1498	1351	1644
9	1613	1465	1761
10	1432	1283	1580
11	1104	943	1271
12	616	506	734
13	2099	1932	2261
14	2739	2570	2897
15	937	794	1084
16	2908	2755	3041
17	1392	1223	1573
18	523	425	638
19	1860	1680	2040
20	4179	4141	4197
21	2216	2034	2394
22	2430	2256	2596
23	2799	2656	2936
24	992	829	1174
25	1676	1475	1875
26	3313	3232	3388
27	4187	4163	4199

MEGA RelTime (21) was used to generate a timetree based on the SILVA phylogeny, the twenty-seven calibrations, and the aligned sequences from SILVA. Calibrations can be seen below in Table 1. The beta version of MEGA RelTime 6 was used.

Reltime results are estimates of relative times of divergence for all nodes in a phylogenetic tree (21). Once relative times were generated by MEGA, RelTime converter was used to generate divergence dates. The file input into the RelTime converter was an Excel file with calibrated times columns for each calibration point. This column was calculated using the following formula: (Relative time node X * time (mya) of calibrated node)/(relative time of the calibrated node)

Two additional programs, treePL and PATHd8 were also used to generate timetrees and their results were compared to MEGA RelTime results by lineage-through-time plots (LTT) created in R.

treePL is a program developed by Stephen A. Smith and Brian C. O'Meara. treePL calculates divergence time estimates using penalized likelihood on large phylogenies (22). The calibrations listed in Table 1 were used to generate a dated tree.

PATHd8, developed by Tom Britton et al, was the third program used to generate a bacterial timetree. PATHd8 generalizes the mean path length (MPL) method. The relative age of a specific node is estimated by the average distance from that node to all its leaves compared to the average distance from the root to all the leaves (23). For this analysis, times were generated using both the minimum and maximum calibration times provided by Battistuzzi and Hedges (7, 18).

Chapter 5

Results

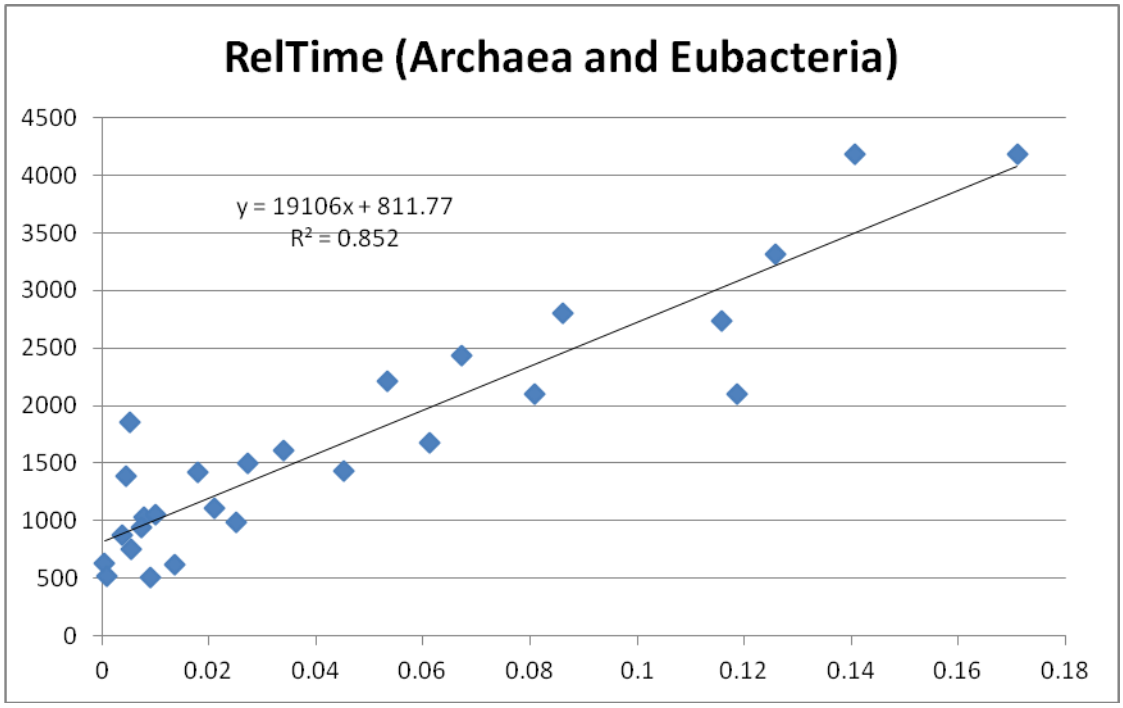

Figure 1. Relative times obtained with the program RelTime vs Calibration Times

Figure 1 shows the plot of relative times of the calibration nodes versus the times from Battistuzzi and Hedges. The R^2 value was found to be 0.852, and the slope of the line describing the relationship between the relative times and the calibration times was 19106. This value could have been used a multiplier to determine divergence times from the relative times produced. However, as seen in Figure 1, there was a large cluster of calibration nodes closer to the y-axis that was possibly the source of variation. To try to fix this, the data points were normalized and plotted, as seen in Figure 2.

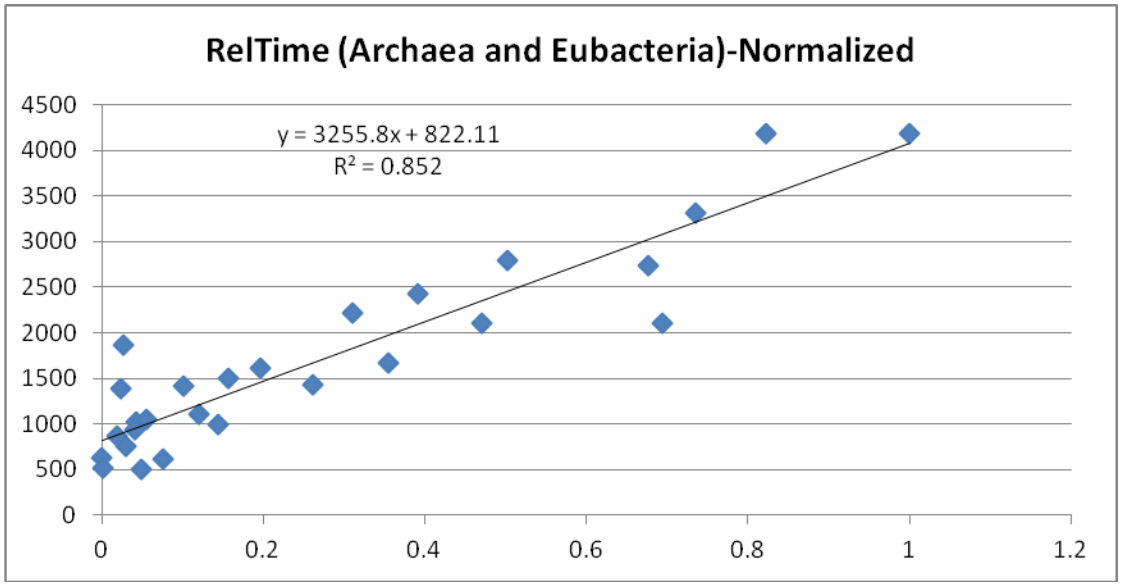

Figure 2. Relative times obtained with the program RelTime vs Calibration Time-Normalized.

. Each point was calculated using the following formula: (RelTime-Minimum Reltime)/(Maximum Reltime-Minimum Reltime). This was done to standardize and normalize the data. This also did not alter the cluster of calibration nodes.

The calibration nodes from the *Time Tree of Life* (7, 18) were compared with times established by Sheridan et al 2010 (13). This was done to compare existing similarities between the nodes. The times (Table 2) were plotted to generate a regression, as seen in Figure 3.

Table 3 Nodes compared

Node	*Sheridan (in Ma)*	*Battistuzzi/Hedges (in Ma)*
Aquifex	3460	4179
Thermotoga group	3300	4189
Chlamydia/Pirellula group	2700	2897
Campylobacter/Helicobacter	1370	1104
Desulfovibrio group	2380	2421
Chlorflexus group	3100	2761
Deinococcus group	2850	2739
Fusobacterium group	2380	3306
Euryarchaeota/Crenoarchaeota	3460	4187
Archeoglobus/Thermoplasma	3330	3160
Archeoglobus/Haloferax	3240	2799
Bifidobacterium/Arthrobacter /Streptomyces	1870	1579

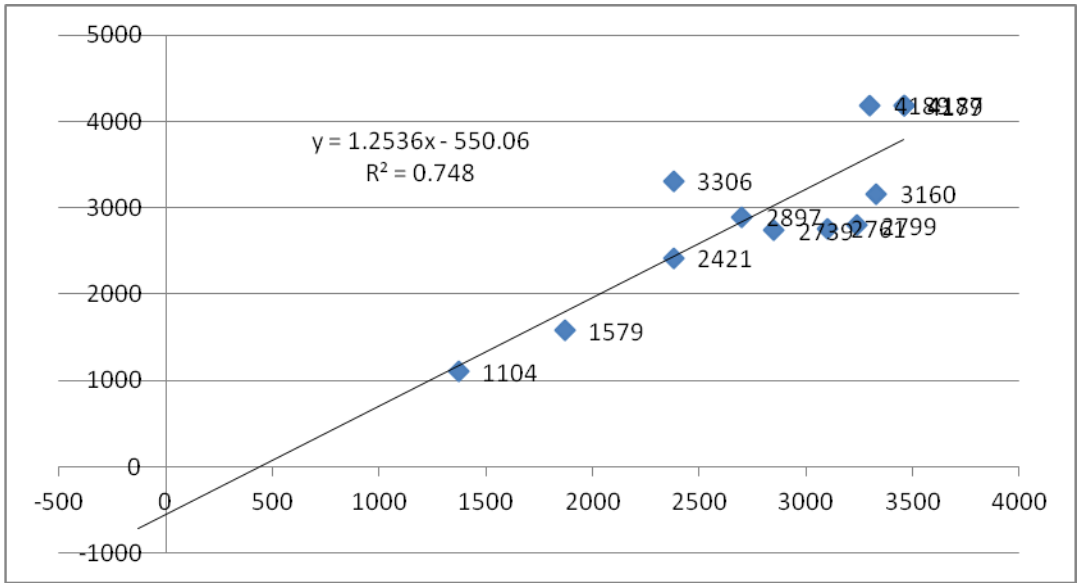

Figure 3. Sheridan et al 2010 established dates vs. Battistuzzi and Hedges 2009 dates.

We found a significant correlation (R^2=0.748, p-value=0.6506) between the dates estimated by Sheridan 2010 (13) and Battistuzzi and Hedges 2009 (7, 18). The significant correlation indicated that times determined from proteins (as was done in Battistuzzi Hedges (13)) were similar to times determined from SSU rRNA data.

Figures 4-7 are lineage-though-time plots for each tree generated. Lineage-through-time plots were generated as well to demonstrate the evolution of diversification through time. As seen in the plots of all the methods used (Figures 4-7), a progressive increase is occurring. These plots show similar increasing diversity, and so demonstrate that the timing methods used were consistent with each other.

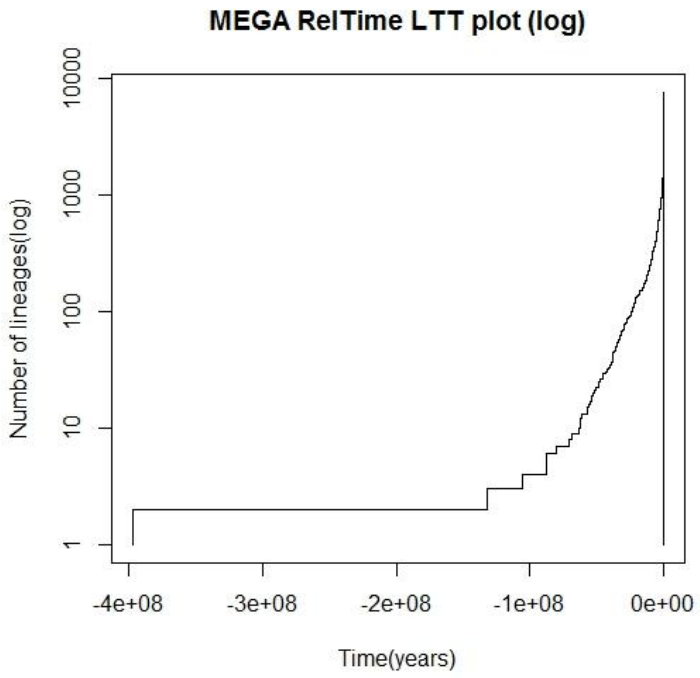

Figure 4. Lineage-through-time plot of MEGA RelTime results

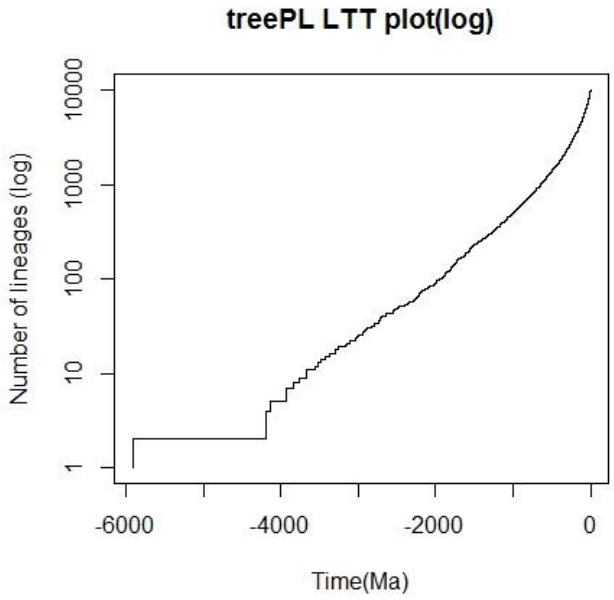

Figure 5. **Lineage-through-time plot of treePL results**

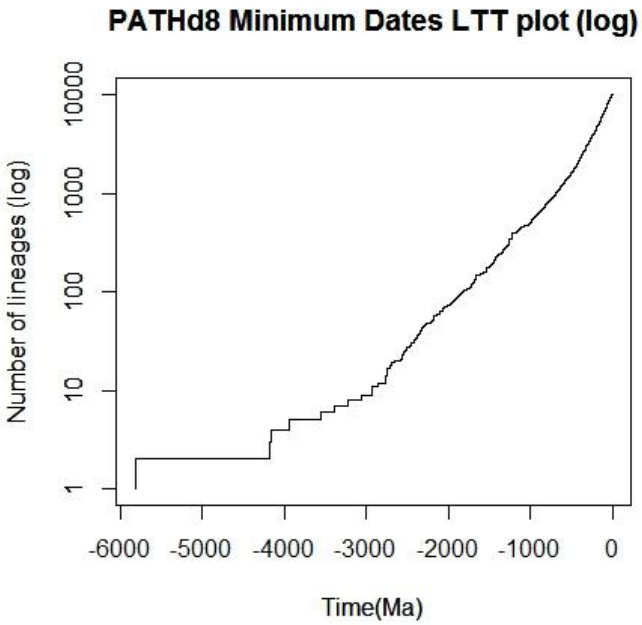

Figure 6. **Lineage-through-time plot of PATHd8 (using minimum dates) results**

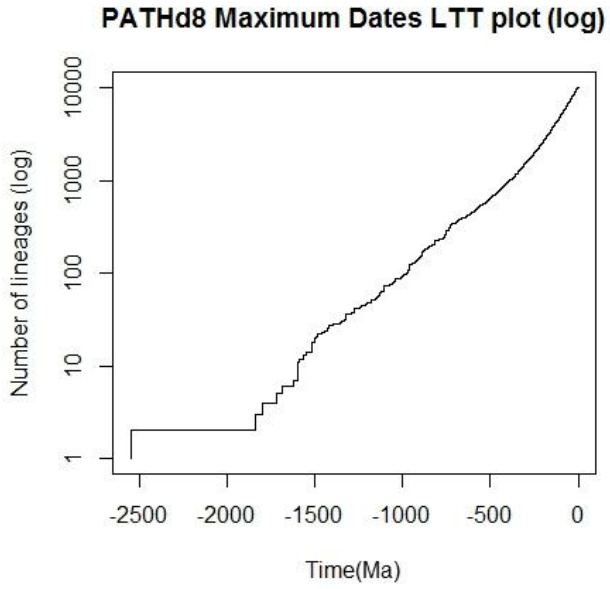

Figure 7. **Lineage-through-time plot of PATHd8 (using maximum dates) results**

To overcome the bias of the large number of younger calibration points, the slopes of the regression lines from the cluster of younger calibration points and older calibration points were added together. This is shown in Figures 8 and 9.

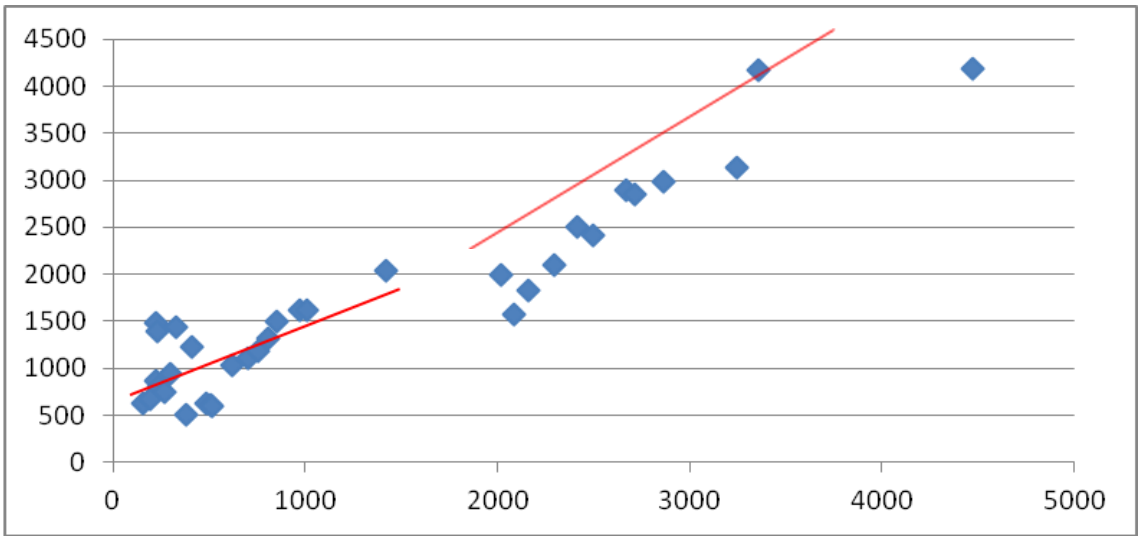

Figure 8. **Separated slopes for both clusters of calibration points.**

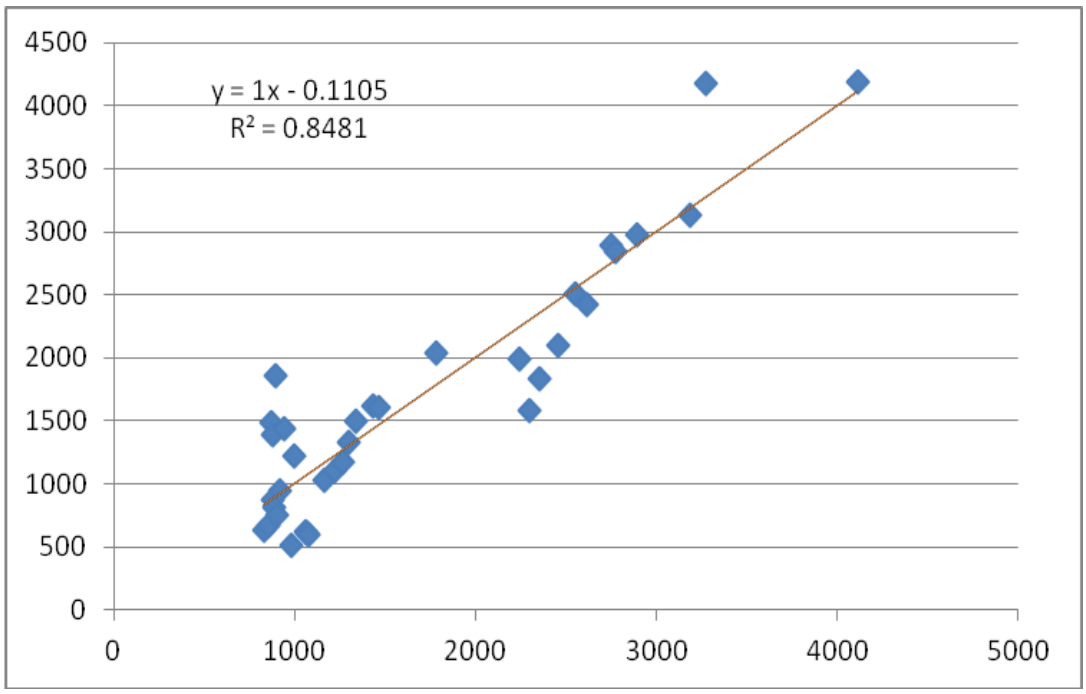

Figure 9. **Slopes of older and younger calibrations added together**

Chapter 6

Discussion

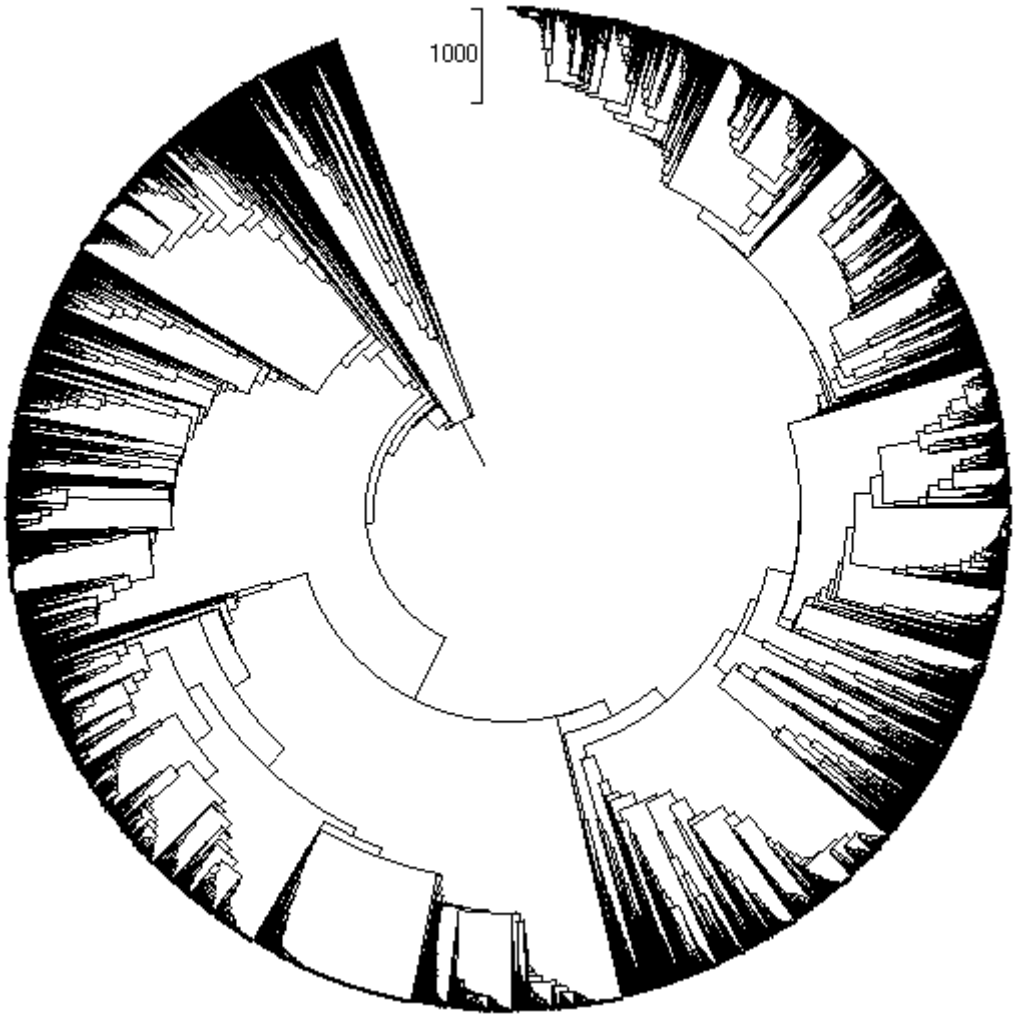

Figure 10. **Circular tree made using treePL with taxa names removed**

Figure 10 shows a circular timetree generated by treePL, demonstrating the scope of this project. Times generated indicated that bacterial divergence times were younger than expected.

Regressions show that the y-intercept (the start of life) to be at about 3000 Ma. The younger calibration points used in this analysis were pushing older nodes towards older estimated times, altering the slope of regression line. The large cluster of younger calibration points had a greater impact on the slope than the older calibration points since there were more younger calibration points than older ones. To correct for this bias, the slopes from the older and younger calibration point clusters were added together. However, this did not alter the results significantly, meaning that the issue in this analysis lies outside the greater number of younger calibration points.

The calibration dates from Sheridan 2010 (13) were compared to the calibration dates taken from Battistuzzi and Hedges (7, 18). As shown in Figure 3, the dates were similar to each other, indicating that the estimated times from both studies matched despite the use of SSU rRNA by Sheridan 2010 (13) and protein sequences by Battistuzzi and Hedges (7,18).

Although the results from this analysis did not yield a timetree for 10,000 species, they did illustrate the challenges of creating a comprehensive bacterial timetree. With such a large data set, many programs are pushed to their computing limit. The SILVA data set consisted of 10,271 unique species (19). Additionally, this phylogeny is being updated frequently to reflect clarifications in taxonomy as well as newly discovered species.

Unlike timetrees constructed for other phyla, bacteria have no ancestral group available for rooting. For this analysis, Thermatoga and Eubacteria were split with Archaebacteria as an outgroup. The lack of outgroup has led previous studies to rely on geological estimates of certain events, such as the Great Oxidation Event, to develop proper times. Calibration dates for Archaebacteria are taken from the origin of methanogenesis based on isotopically light carbon (3460 million years ago), and a maximum of 4200 million years ago for the divergence based on

the last ocean-vaporizing impact (7). Another calibration used is the maximum time for life on Earth (a maximum of 4600 million years ago) (18).

However, the geological record of prokaryotes is not robust. Biomarkers for carbon assimilation and lipid biosynthesis do not allow for many opportunities to compare the molecular timescale (3). Based on Battistuzzi and Hedges findings, most time estimates for major adaptations in Eubacteria match geological evidence (3).

The version of MEGA RelTime (version 6) used in this analysis made use of a linear molecular clock. This will be corrected in future versions of the program, but was a source of biased results in this analysis. Estimates of relative time are shown to be linearly related to the true time of divergence (20). There is however, heterogeneity in molecular evolutionary rates. There are major rate differences among groups of prokaryotes, as well as between prokaryotes and eukaryotes (3).

SSU rRNA can be used to determine phylogenetic relationships between prokaryote species, as has been done with the SILVA data set. The analyses run in this project were done to add a time scale to the increasing diversification of prokaryote species. As mentioned in Chapter 3, the biological meaning of this data has many implications.

A timetree of prokaryotes would be a valuable addition to the tree of life. An expanded tree life, containing 50,632 species, has recently been published by Hedges et al (24). Species diversity is increasing, and the rate of diversification in eukaryotes (which composed 99.7% of the species) is mostly constant (24). In this study, divergence times of closely related species of prokaryotes were about 50 to 100 times older than vertebrates, arthropods and plants (24). With an expanded prokaryote data set, the tree of life could better represent the full diversity of life.

Chapter 7

Conclusion

Although the timetree of life contains great information about eukaryotes, divergence dates for prokaryote species are not as concrete. Using previously successful molecular clock methods, this project aimed to generate estimated times of divergence among many species of prokaryotes.

Many hurdles exist when attempting to generate timetree of prokaryotes. Horizontal or lateral gene transfer can impact the bacterial exchange of genetic information, altering organisms' gene differences. Gaps in fossil evidence also make it difficult to establish calibration times and rooting of the timetree. The root of the Eubacteria and Archaebacteria timetree reaches back to the start of life on Earth. Calibration points are then based on geological markers of atmospheric change and prokaryote metabolic changes, whether they are resulting from oxygen proliferation or impact events.

Although this project was not successful in generating a prokaryote timetree of over 10,000 species, it was able to show that analysis of the SILVA tree was problematic. With further adjustments to the programs utilized, as well as a re-evaluation of the SILVA data set, a proper, comprehensive timetree of prokaryote species can be generated.

BIBLIOGRAPHY

1. Hedges S.B., Kumar S. Discovering the timetree of life. Pp. 3-18 in The Timetree of Life, S.B. Hedges and S. Kumar, Eds. (Oxford University Press, 2009).
2. Avise J.C. Timetrees: beyond cladograms, phenograms, and phylograms. Pp 19-25 in The Timetree of Life, S.B. Hedges and S. Kumar, Eds. (Oxford University Press, 2009).
3. Hedges S.B., Battistuzzi F.U., and Blair J.E. Molecular timescale of evolution in the Proterozoic. Pp 199-229 in Neoproterozoic Geobiology and Paleobiology. S. Xiao and A.J. Kaufman, Eds. (Springer, 2006)
4. Zhaxybayeva O. and Doolittle W.F. 2011. Lateral gene transfer. *Current Biology*. 21(7): R242-246.
5. Gogarten J.P., Doolittle W.F., and Lawrence J.G. 2002. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*. 19(12): 2226-2238.
6. Noffke N., Decho A.W., and Stoodley P. 2013. Slime through time: the fossil record of prokaryote evolution. *Palaios*. 28: 1-5.
7. Battistuzzi, F.U., S.B. Hedges. Archaebacteria. Pp. 101-105 in The Timetree of Life, S.B. Hedges and S. Kumar, Eds. (Oxford University Press, 2009).
8. Yarza P, Ludwig W, Euzeby J, Amann R, Schleifer K, Glockner F.O., and Rosello-Mora R. 2010. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses. *Systematic and Applied Microbiology*. 33:291-299.
9. Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Glockner F.O. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web based tools. *Nucleic Acids Research*. 1-7.
10. Hedges S.B. Life. Pp 89-98 in The Timetree of Life, S.B. Hedges and S. Kumar, Eds. (Oxford University Press, 2009).
11. Zuckerkandl E., Pauling L. 1965. Evolutionary divergence and convergence in proteins. *Evolving genes and proteins*. 97:97-166.
12. Kimura M. 1968. Evolutionary rate at the molecular level. *Nature*. 217 (5129): 624-626.
13. Sheridan, P.P., K.H. Freeman, and J.E. Brenchley. 2003. Estimated minimal divergence times of the major Bacterial and Archaeal phyla. *Geomicrobiology* 20:1-14.
14. West-Eberhard M.J. Pp 353-376. *Developmental plasticity and evolution*. Oxford University Press, 2003.
15. Studer R.A., and Robinson-Rechavi M. 2009.How confident can we be that orthologs are similar, but paralogs differ? *Trends in Genetics*. 25(5): 210-216.

16. Battistuzzi F.U., Feijao A. and Hedges, S.B. 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototropy, and the colonization of land. *BMC Evolutionary Biology* 4(1)
17. Battistuzzi F.U., and S.B. Hedges. 2009. A major clade of prokaryotes with ancient adaptations to life on land. *Molecular Biology and Evolution*. 26 (2): 335-343.
18. Battistuzzi, F.U., S.B. Hedges. Eubacteria. Pp. 106-115 in The Timetree of Life, S.B. Hedges and S. Kumar, Eds. (Oxford University Press, 2009).
19. SILVA. (2014). LTPs115. Retrieved from www.arb-silva.de
20. Yarza P., Ludwig W., Euzeby J., Amann R., Schleifer K., Glockner F.O., and Rosello-Mora, R. 2010. Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analysis. *Systematic and Applied Microbiology* .33(6): 291-299.
21. Tamura K., Stecher G., Peterson D., Filipski A, and Kumar S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*. 30(12):2725-2729.
22. Smith S.A., O'Meara B.C. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*. 28(20): 2689-2690.
23. Britton T., Anderson C.J., Jacquet D., Lundqvist S., Bremer K. 2007. Estimating divergence times in large phylogenetic trees. *Systematic Biology*. 56(5): 741-752.
24. Hedges S.B, Marin J., Suleski M., Paymer M., and Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*. 32(4):835-845.

ACADEMIC VITA

Jaanki Dave
238A Atherton Hall, University Park, PA 16802
jrd5506@psu.edu

EDUCATION:

Manheim Central High School, Class of 2012

The Pennsylvania State University, Class of 2015 (Biology (Vertebrate Physiology) and Biological Anthropology)

The Commonwealth Medical College, Class of 2019

HONORS AND AWARDS:

President's Freshman Award

Dean's List

J Henry & Minnie Hitz Scholarship

The William F. Brossman Family Scholarship

Christopher R Dyckman and Susan Scott Scholarship in Biology

LEADERSHIP:

TEDxPSU (Content Team 2013, 2014, Director of Content 2015)

Schreyer Honors Orientation mentor (2013, 2014)

Schreyer Honors College Literary Committee (Executive Board 2013-2015)

SHC-Radboud University "Future of Healthcare" Think Tank (2014-2015)

Stop Hunger Now Crowdfunding Campaign (Leadership Team-Content Writer-2015)