PREDICTING RNA MUTATION USING 3D STRUCTURE

Stephen B. Dinda

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

December, 2011

Committee:

Dr. Craig L. Zirbel, Advisor

Dr. Andy Garcia,
Graduate Faculty Representative

Dr. Neocles B. Leontis
Dr. James Albert
Ribosomes are protein-synthesizing nano-machines found in all organisms. They are RNA-based enzymes (ribozymes) and among the most ancient nano-assemblies in nature. Recent advances in x-ray crystallography have revealed atomic-resolution structures of a handful of ribosomes from diverse branches of life. Strikingly, although the ribosomal RNA (rRNA) sequences themselves have diverged significantly, the 3D structures of the core regions are highly conserved. We inferred consensus phylogenetic trees using the MrBayes phylogenetics package [19]. We added new modules to FR3D (Find RNA 3D) program suite [43] to take the consensus phylogenetic trees and estimate the mutation rate for each nucleotide in the rRNA conserved cores using dynamic programming described in [11]. We used FR3D to annotate structural features and to determine geometrical relationships between nucleotides belonging to the structurally conserved core regions of the rRNA molecules, as determined in previous work [46]. We report the results of extensive investigation of various explanatory variables, constructed from recurrent structural features observed in the RNA, to model the mutation rates. Of the 45 created variables, calculated for each nucleotide, we identified a parsimonious set of ten that models 36% of the observed variation of mutation rates over 3961 sites. We then performed a factor analysis to determine commonalities of the terms, and hierarchical regression on the main factors to adjust for difference in mutation rate in the 5S, 16S, and 23S chains.
This dissertation is dedicated to Stephen F. Dinda III, my father, who has studied war and politics so that I had the liberty to study mathematics and philosophy.
ACKNOWLEDGMENTS

(i) I would like to first acknowledge my mother who taught me the value of an education. It was by her influence that this work was possible.

(ii) I would like to acknowledge the rest of my family: grandma, Uncle Leo, Aunt Marge, Smokey, Peter, and Michael. It was your support and belief in me that made this work possible.

(iii) I would like to extend my deepest thanks to my advisor Dr. Craig L. Zirbel. Not only have you made me a better scientist and researcher, but a better person as well. You were a wonderful mentor to me and I will forever be indebted to you.

(iv) I would like to thank my friends at BGSU: Marilee Murray, Ian Deters, John Barrow, A. J. Wilhelm, Adam Combs, Warren McGovern, Stephanie McGovern, Melanie Henthorn, Kate Overmoyer, Rachelle Barr, Tumpa Bhattacharyya, Alice Cook, David Sleasman, Jill Carr, Ed Whipple, and Pat Pauken for enriching my life so much.

(v) Thanks to my teachers that have inspired me along the way. First, I wish to thank Mr. Russ Nalepa who inspired me to study math while teaching me so much more than just math. I would also like to thank Mrs. Delores Dershaw, Mrs. Patricia Vlock, Mrs. Caroline Mulhall Dr. Nancy Boudreau, and Dr. Richard N. “Herb” McGrath.

(vi) Thanks to the members of my dissertation committee: Dr. Neocles Leontis, Dr. Jim Albert, and Dr. Andy Garcia. Your support and expertise during the dissertation process was much appreciated.

(vii) Thanks to Nate Iverson fr creating the BGSU LaTeX class file.

(viii) Lastly I wish to pay respect to my alma maters Youngstown State University and Ursuline High School whose mottos are still close to my heart: Annius Liberatus and Soli Deo Gloria.
Table of Contents

CHAPTER 1: Introduction 1

1.1 Summary of work ........................................... 1
1.2 Previous work on modeling site-specific mutation rates ............... 2
1.3 Modeling site-specific mutation rates ................................ 4
1.4 Overview of results ......................................... 6

CHAPTER 2: Background and assumptions 11

2.1 RNA background ............................................ 11
   2.1.1 The ribosome and RNA nucleotides ...................... 11
   2.1.2 Introduction to RNA 3D structure ....................... 12
   2.1.3 Application of FR3D to our work ....................... 13
2.2 Randomness in the replication process ................................ 14
   2.2.1 Introduction to point substitutions of DNA .............. 14
   2.2.2 Mutations of rRNA ..................................... 15
   2.2.3 Mutations and the 3D structure .......................... 16
2.3 Modeling assumptions on single-site mutations ....................... 17
   2.3.1 Discrete versus continuous time ......................... 17
   2.3.2 Markov process ...................................... 18
   2.3.3 Evolution according to a Markov process ............... 19
2.4 Phylogenetics ............................................. 20
2.4.1 A brief introduction to a multiple sequence alignment . . . . . . . . . . 20
2.4.2 Introduction to phylogenetics . . . . . . . . . . . . . . . . . . . . . . 20
2.4.3 Heuristic approach to phylogenetics . . . . . . . . . . . . . . . . . . . 21
2.4.4 Probabilistic models of mutation . . . . . . . . . . . . . . . . . . . . 22
2.4.5 Substitution models for multiple nucleotide mutations . . . . . . . . 28
2.4.6 Maximum likelihood approach to tree estimation . . . . . . . . . . . . 30
2.4.7 Applying the probabilistic models of mutation . . . . . . . . . . . . . 31
2.4.8 Synthetic data simulation . . . . . . . . . . . . . . . . . . . . . . . . 32
2.4.9 Known difficulties of inferring a phylogeny . . . . . . . . . . . . . . 33

CHAPTER 3: Estimation of site-specific mutation rates 35

3.1 Multiple sequence alignments of bacterial rRNA . . . . . . . . . . . . . . 35
3.1.1 Introduction to multiple sequence alignments . . . . . . . . . . . . . . 35
3.1.2 Uses for alignments . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
3.2 Cleaning datasets for use . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37
3.2.1 Reducing the dataset for computational efficiency . . . . . . . . . . . 37
3.2.2 Determining a conserved core of ribosomal structures . . . . . . . . . 37
3.3 MrBayes inference of a phylogeny and site-specific mutation rates . . . . . 39
3.4 Estimating site-specific mutation rates . . . . . . . . . . . . . . . . . . . . 40
3.4.1 MrBayes estimates of site-specific mutation rates . . . . . . . . . . . 40
3.4.2 Modeling site-specific rates . . . . . . . . . . . . . . . . . . . . . . . . 41
3.4.3 Mathematical model for a mutation rate . . . . . . . . . . . . . . . . . 42
3.4.4 A maximum likelihood estimation of mutation rates . . . . . . . . . . 44
3.4.5 Calibration of branch lengths . . . . . . . . . . . . . . . . . . . . . . . 46
3.4.6 Making a maximum likelihood estimation of rates . . . . . . . . . . . 46
3.4.7 Performance of the ML algorithm . . . . . . . . . . . . . . . . . . . . . 47
3.4.8 Results of a simulation study . . . . . . . . . . . . . . . . . . . . . . . 49
3.4.9 Mutation rates estimated by maximum likelihood . . . . . . . . . . . 50
CHAPTER 4: Defining and calculating terms from the 3D structure

4.1 RNA 3D structures ........................... 52
4.2 Introduction to non-basepair interaction type variables ................. 53
4.3 Distance from the center of the ribosome .......................... 53
  4.3.1 Definition of DistanceFromCenter of the Ribosome .............. 53
  4.3.2 Exploring the DistanceFromCenter variable ..................... 54
4.4 Sphere-type variables ................................ 55
  4.4.1 Introduction to the Sphere Variable .......................... 55
  4.4.2 Standardizing the radius for sphere-type variables ............. 56
  4.4.3 NucleotideSphereCount ................................ 56
  4.4.4 Near or not near the plane of the nucleotide ................... 58
  4.4.5 DifferentSubUnitRNASphere ................................. 62
  4.4.6 DifferentSubUnitRNANearPlaneCount .......................... 65
  4.4.7 SmallerRNASphere ................................ 68
  4.4.8 Near or not near the plane of the nucleotide ................... 70
  4.4.9 AmnoAcidSphereCount ................................ 75
  4.4.10 Near or not near the plane of the nucleotide ................... 77
  4.4.11 Nucleotides potentially interacting with an amino acid ....... 81
4.5 Base stacking ................................ 82
  4.5.1 Classification of Base Stacking ............................. 82
  4.5.2 The counting approach ................................ 83
  4.5.3 Stacking Discrepancy Index ................................. 86
4.6 Base-phosphate interactions ................................ 89
  4.6.1 Introduction ................................ 89
  4.6.2 Counting approach ................................ 90
  4.6.3 Geometric approach ................................ 91
4.7 Near base-phosphate interactions ........................................... 95
  4.7.1 Introduction ............................................................. 95
4.8 Base-ribose interactions ................................................. 96
  4.8.1 Definition ............................................................... 96
  4.8.2 Exploring the BaseRiboseCount variable ......................... 97
  4.8.3 Exploring the NearBaseRiboseCount variable .................... 98
  4.8.4 Geometric approach to base-ribose interactions ............... 100
4.9 Introduction to basepairing interactions ............................ 103
4.10 cWW interactions ....................................................... 104
  4.10.1 Definition ............................................................. 104
  4.10.2 Exploring the cWWCount variable ............................... 104
4.11 non-c-WW interactions .................................................. 105
  4.11.1 Definition ............................................................. 105
  4.11.2 Exploring the noncWWCount variable ......................... 106
  4.11.3 Exploring the BasepairCount variable ......................... 107
4.12 Near basepairing interactions ........................................ 109
  4.12.1 Exploring the NearcWWCount variable .......................... 109
  4.12.2 Exploring the NearestonecWW variable ...................... 110
  4.12.3 Exploring the NearBasepairCount variable ................... 112
4.13 NumberInterfacingNucleotides ........................................ 114
  4.13.1 Definition ............................................................. 114
  4.13.2 Exploring the InterfacingNucleotidesCount variable ........ 114
4.14 Number of long range interactions .................................. 115
  4.14.1 Definition ............................................................. 115
4.15 Length of a network of interactions .................................. 116
  4.15.1 Definition ............................................................. 116
  4.15.2 Exploring the LengthOfNetworkCount variable ............... 117
4.16 Extensibility ................................................................. 119
  4.16.1 Definition ............................................................ 119
  4.16.2 Exploring the ExtensibleCount variable ...................... 119
4.17 Isosteric Constraining Effect (ICE) ................................. 120
  4.17.1 Introduction .......................................................... 120
  4.17.2 Isodiscrepancy matrix ............................................ 121
  4.17.3 Defining ICE in relation to isostericity ...................... 122
  4.17.4 Example of ICE ..................................................... 123
  4.17.5 Summary of creation of ICE ..................................... 123
  4.17.6 Exploring the ICE variable ..................................... 124
  4.17.7 NearICE .............................................................. 124
  4.17.8 Exploring the NearICE variable ............................... 125
4.18 Communication of constraints to a nucleotide .................. 127
  4.18.1 Introduction ........................................................ 127
  4.18.2 Building the TrueCommunicatorCount variable .............. 127
  4.18.3 Building the TrueCommunicatorCount variable .............. 127
  4.18.4 Exploring the TrueCommunicatorCount variable ............ 128
  4.18.5 Building the TrueCommunicatorICE variable ................ 129
  4.18.6 Building the TrueCommunicatorICE variable ................ 129
  4.18.7 Exploring the TrueCommunicatorICE variable .............. 130
4.19 Communication of near interaction constraints on a nucleotide 131
  4.19.1 Introduction to AllCommunicatorCount variables .......... 131
  4.19.2 Terms used to make the variable ............................. 131
  4.19.3 Building the AllCommunicatorCount variable .............. 132
  4.19.4 Exploring the AllCommunicatorCount variable ............. 132
  4.19.5 Terms used to make the AllCommunicatorICE variable .... 134
  4.19.6 Building the AllCommunicatorICE variable ............... 134
CHAPTER 5: Modeling site-specific mutation rates

5.1 Overview ......................................................... 138

5.2 A best subset of terms to model site-specific mutation rates .................. 139

5.3 Using the model ................................................. 144

5.4 Multilevel modeling approach ................................ 147

5.4.1 Multilevel modeling applied to site-specific mutation rates ............... 147

5.4.2 A Multilevel model for site-specific mutation rates ...................... 148

5.5 Application of factor analysis to the terms of the model ...................... 151

5.6 Conclusions drawn from the models .................................. 153

CHAPTER 6: Future Work

6.1 Developing new predictors ..................................... 154

6.2 Further Estimation Mutation Rates ................................ 156

6.3 Distribution of nucleotides .................................. 156

6.4 Estimation of the mutation rates of more than one nucleotide at a time ... 158

BIBLIOGRAPHY ............................ 159
CHAPTER 1

Introduction

1.1 Summary of work

Every living organism has small molecular machines called ribosomes in every cell which use information from the organism’s DNA to produce proteins. Ribosomes are made of three chains of RNA, known as 5S, 16S, and 23S, which fold up into a very specific 3D structure to make the ribosome work. The specific nucleotide sequence in these chains differs from organism to organism and has presumably been changing slowly over the past millions of years as species diverged from one another. Even though mutations cause changes in the chain of nucleotides that compose the structured rRNA, we infer that the 3D structure is very slow in changing by comparison because the 3D structure is what allows the ribosome to perform its function properly. For example in a helix, a common RNA structural motif, a GC basepair changing to a CG basepair is known to happen often. Therefore the individual nucleotides can change while maintaining the 3D structure.

In this work we create a regression, or statistical model, to link site-specific mutation rates in RNA sequences to local 3D structural properties of ribosomal RNA. To do this, we use a multiple sequence alignment of ribosomal RNA sequences from many different bacteria, and used MrBayes [19], [36] to determine the underlying phylogenetic tree and branch lengths
that relate the organisms. Using the tree and the branch lengths we used the dynamic programming framework of [11] to individually estimate the site-specific mutation rate for each nucleotide according to a Jukes-Cantor model of nucleotide mutation [20]. We added programs to FR3D [43], and we built a set of predictors concerning the local 3D properties around each nucleotide from the 3D structures of *Escherichia coli*, (*E.c.*), and *Thermus thermophilus*, (*T.th.*). From these predictors for each nucleotide, we find a subset of most predictive terms that we use to create a regression model of nucleotide mutation rates. We go on to apply a hierarchical component to the regression model to adjust for differences in mutation rates between the three RNA chains in ribosomal RNA, and finally we apply factor analysis to help classify the effect from the terms in the model. The main predictors of nucleotide mutation rate are discussed briefly in Section 4 of this chapter and extensively in Chapter 4.

### 1.2 Previous work on modeling site-specific mutation rates

Mutations occurring to the DNA including that which codes for rRNA is believed to occur at a constant rate [40]. However, we assume that each nucleotide found in rRNA mutates according to its own mutation rate because the mutation must still allow for the functionality of the ribosome. [51] showed that allowing for each site of rRNA to mutate according to its own rate improves a model for nucleotide mutation. Therefore, since the rate at which DNA mutates is constant across sites while the rate at which the rRNA mutates is not, the interactions that the rRNA nucleotides are making must influence the rate at which each nucleotide can mutate. We also see in other work that when each nucleotide is assumed to mutate according to its own rate improves the performance of the model.

[53] introduced an idea of relating the amount of sequence variation observed in a multiple sequence alignment of homologous, ribosomal RNA (rRNA) sequences from bacteria to
physical attributes of the crystallized structure of a bacterium, E.c. This idea takes root in the fact that the 3D structure of rRNA is very highly conserved, but that the nucleotides that make up these structures can vary. While the 3D structure is very highly conserved, there is not perfect conservation over organisms. For example E.c and T.th. are believed to have diverged from their last common ancestor more than one billion years ago, yet more than 88% of the structure is similar enough so that we compare the structure of one against the other. Therefore, [53] believed that the interactions that rRNA nucleotides make in the 3D structure affect the rate at which they can mutate. [53] look at every column of the multiple sequence alignment such that the nucleotide was in a basepairing region of the 3D structure that aligned when E.c. and T.th. 3D structures were compared. [53] referred to these as the conserved core of the rRNA. The measure of variation of each column of the alignment was defined to be the proportion of nucleotides that matched the nucleotide in E.c. at that position. Using the alignment as described above, the conservation percentage was successfully related to the 3D structure near each nucleotide. The model given was:

\[
\text{ConservationPercentage} = 26.3 \\
+ 58.1 \times \text{conservation percentage of stacking partners} \\
+ 8.0 \times \text{number of non-cWW pairs it forms} \\
+ 7.4 \times \text{number of BPh interactions in which it is the H-bond donor} \\
+ 2.7 \times \text{near a protein (1) or not (0)} \\
+ 1.8 \times \text{number of BPh interactions in which it is the phosphate acceptor} \\
+ 0.3 \times \text{number of cWW pairs it forms}
\] (1.1)

From this model, we see evidence that the 3D structure interactions do affect the rate at which nucleotides can mutate, and we are able to quantify what effect each of the interactions has on the conservation percentage. In addition, we see that base-phosphate interactions as introduced as being base specific had an effect on the conservation percentage. This model
provided an initial step in modeling conservation percentage as a function of the 3D structure, but raised questions about what aspects of the 3D structure affect nucleotide mutation rate.

1.3 Modeling site-specific mutation rates

The strength of the model for \[53\] is that they account of each factor of the 3D structure of \(E.c.\) in the presence of others in the model of conservation percentage. However, the model above has obvious weaknesses. The first weakness we point out is that conservation percentage is merely a proxy for the site-specific mutation rate; which measures the speed at which each nucleotide mutates relative to a standard unit of time. Calculating the site-specific mutation rates depends upon first determining the underlying phylogeny, or evolutionary distance in time that the organisms are from each other, which can be inferred from the multiple sequence alignment. Another weakness of the model is that the conserved core for the alignment includes only basepaired nucleotides found in the 3D structure of \(E.c.\) and \(T.th.\) instead of using all of the nucleotides from each structures that can be compared. This is a weakness because it unnecessarily removes valid observations that would help in determining the underlying phylogeny, skews any resulting model through sampling bias, and leaves out many nucleotides giving a smaller dataset with which to work. To address this weakness, we use R3DAlign \[34\] to align the 3D structures. With R3DAlign, an alignment of two 3D structures is done through a comprehensive comparison of nucleotides that compose the 3D structure. This allows for all nucleotides of \(E.c.\) and \(T.th.\) to be used to identify a larger conserved core than one obtained by focusing only on the basepairing regions of the two structures. This greatly increases the number of nucleotides that are identified to be in the conserved core, with 3961 nucleotides out of a possible 4488 nucleotides being classified as the conserved core, thereby addressing the concerns above. The final weakness of \[53\] that we point out is that the model has only a few predictors of nucleotide mutation, with the conservation percentage of stacking partners not really a predictor of conservation per-
percentage. In this work we will introduce many predictors and see that the predictors of [53] are not having the largest effect on site-specific mutation.

To accomplish the goal of this work, modeling the site-specific mutation rates using the local physical attributes of the 3D structure, we will need to calculate the site-specific mutation rates individually as well as obtaining the physical attributes of the 3D structure. We will give the background needed to estimate site-specific mutation rates in Chapter 2 and discuss estimating the site-specific mutation rates in Chapter 3. We will discuss how to extract meaningful information about the local environment of each nucleotide from the structure of *E.c.* and *T.th.* in Chapter 4, and use this information to model the site-specific mutation rates in Chapter 5. We will give a brief overview of these chapters here.

To calculate the site-specific mutation rates, we must have a model of nucleotide mutation and an understanding of previous work in this field. This will be provided in Chapter 2 including models for nucleotide mutation. In Chapter 3 we will discuss in detail how we use MrBayes [19] to infer the underlying phylogeny, and how we use this phylogeny and the dynamic programming idea given in [11] to estimate the mutation rates for a user-defined generator matrix.

The goal of this work is to model the mutation rate by the local structural details for each nucleotide. We explain how we define and calculate the 45 predictors of mutation rate for each nucleotide in Chapter 4. These predictors include those in [53] such as an indicator if a nucleotide is making a cWW basepair, the count of non-cWW interactions, the number of base-phosphate interactions, and so on. The only term from [53] that we do not consider is the conservation percentage of stacking partners. In this work, we introduce many predictors based upon the 3D structure not considered in [53], and some predictors that refine the terms of [53]. Some of the novel considerations of the 3D structure include the effect of basepairing interactions on site-specific mutation rate through the lens of isostericity, which will be defined in Chapter 2. In addition to this, we expand upon the basepairing concepts to include considering the constraining effects that the partner pairing to a nucleotide
communicates to the nucleotide, adjusted for the interaction in a term that we call All-CommunicatorICEColumn, explained further in Chapter 4. One more aspect of basepairing interactions that we consider is the length of a chain of nucleotides which connect through basepair interactions. In this work, we introduce a geometric approach to consider the effect of stacking, base-phosphate and, base-ribose interactions. We also introduce predictors based upon novel concepts of aspects of the 3D structure such as the Euclidean distance from the center of the ribosome and the local density of nucleotides or amino acids that are within a sphere around a nucleotide. In all, we consider 45 predictors of nucleotide mutation, most of which are novel. These are defined and briefly explored through some exploratory data analysis methods such as scatterplots, boxplots, and correlation coefficient with the site-specific mutation rates in Chapter 4.

In Chapter 5 we consider 45 predictors, each of which is computed for each nucleotide of site-specific mutation rates created from the 3D structure information. From this large set of predictors we present a model for the site-specific mutation rate of a nucleotide as a function of a small number of components of the 3D structure. We then apply factor analysis to the terms selected to be in the model to determine the commonality of the effect that these terms have on nucleotide mutation rate. We interpret these factors, and then end the chapter by applying a hierarchical component to the regression model to compensate for and to help understand the difference in the rate of nucleotide mutation for the different chains, 5S, 16S, and 23S.

1.4 Overview of results

We calculated each of the 45 predictors for every nucleotide in the conserved core, and determined a predictive set of 10 terms from the 45 to serve in our model of nucleotide mutation. We present the model for nucleotide mutation for the E.c structure as found in Chapter 6 here, in the order of their $R^2$ contribution given that all other terms are in the
model, and we explain the variables afterward.

\[
\log(E.c. \text{ MutationRate}) = -0.4849 \\
- 0.0918 \times \text{NucleotideSphereNearPlaneCount} \\
- 0.0491 \times \text{SmallRNANearPlaneCount} \\
- 0.0399 \times \text{AminoAcidSphereCount} \\
- 1.3603 \times \text{ICE} \\
+ 0.0718 \times \text{AllCommunicatorICEColumn} \\
+ 0.5735 \times \text{cWWCount} \\
+ 0.0100 \times \text{DistanceFromCenter} \\
- 1.8954 \times \text{SDI} \\
- 0.3957 \times \text{BRGeometry} \\
- 0.3558 \times \text{BPhGeometry} \quad (1.2)
\]

This model will serve as the base model for \textit{E.c.} We will also apply a hierarchical component to the model to adjust for the effect of being in the 5S, 16S, or 23S chain of the ribosome. Also, we will see through the factor analysis of the terms in the model presented here and in Chapter 6 that there are three forces that have the largest effect on nucleotide mutation.

We will conclude this chapter by briefly interpreting the variables in the model, and will leave the detailed explanation and basic exploration including the correlation with mutation rates and descriptive statistics for each term to Chapter 4.

The first group of terms that we consider deal with the local density surrounding a nucleotide. The first term in this group that we consider is \text{AminoAcidSphereCount}. The logic behind this term is that we expect that nucleotides that are near where proteins interact and bind to the ribosome would be slower to mutate in order to preserve the structure of the binding site. This variable counts the number of amino acids, the building blocks of proteins,
in a sphere of 16 Å around the center of each nucleotide. From the model above, the more amino acids in the sphere, the slower the nucleotide is to mutate on average. Next in the group of local density, we consider NucleotideSphereNearPlaneCount. This term is similar to AminoAcidSphereCount with two differences. The first is that we consider the number of RNA bases in a sphere surrounding a given nucleotide instead of amino acids. These bases are counted from any of the 5S, 16S, or 23S chains. Where the “NearPlane” component comes in is that we split the number of bases in a sphere into two parts, those near the plane of the base and those not near the plane of the base. We only count nucleotides in the sphere that are in the near-plane portion. The more bases of nucleotides that are near the plane to any given nucleotide, the lower the expected mutation rate. The next term of the local density group is SmallRNASphereNearPlaneCount. This term is like NucleotideSphereNearPlaneCount in that we count the number of nucleotides near the plane of the given nucleotide. However, this term differs from NucleotideSphereNearPlaneCount in that the nucleotides that are counted come from tRNA or mRNA structures as they interact with the ribosome. Since these structures have fewer nucleotides than even the 5S structure, we call them SmallRNA. The final local density term that we consider is the DistanceFromCenter. From the model above, the further from the center of the ribosome a nucleotide is in Euclidean distance from the center of the ribosome the faster is the expected mutation rate of the nucleotide. One explanation is that the active site is nearer the center of the ribosome, and that nucleotides that mutations further from the active site of the ribosome may still allow for functionality of the ribosome.

The next group of terms deals with the basepairing geometry of the nucleotide. These terms are explained greater detail in Chapter 4. We begin the brief discussion of these terms by explaining is the geometry of basepairing interactions themselves. [46] used a quantity called the Isosteric Discrepancy Index, IDI, as a way to explain the geometry of the nucleotides that could be substituted during a mutation to a basepair while preserving the type of basepairing interaction. We use this concept in the term ICE, short for Isosteric...
Constraining Effect. This quantity measures how easily the nucleotide can substitute when it is part of a basepairing interaction, based on the geometry of the nucleotides that compose the basepairing interaction, while preserving the interaction after the mutation event occurs. When a nucleotide makes multiple basepairing interactions simultaneously, we add the contribution from each basepair. We find that the harder it is to replace this nucleotide due to its basepairing interaction or interactions, the lower the expected mutation rate. The next term that we consider is the simple indicator variable cWWCount. This term takes the value 1 if the nucleotide is making a cWW basepairing interaction, and 0 otherwise. This term is included because nucleotides that make a cWW pair and nothing else have a higher mutation mutation rate than those that do not. Therefore this term, while predictive in its own right, is a corrective factor for ICE. Next, we consider the term AllCommunicator. This term, like ICE, is defined for nucleotides that are part of one or more basepairing interactions. However, unlike ICE, AllCommunicator looks at the interactions of the partner nucleotide and the constraint against mutation that it communicates back to the nucleotide. The communication effect is then modified by a score based on the geometry of the interaction between the nucleotide and its partner. This term is computed as a linear combination of some of the other predictors multiplied by a regression coefficient. Since the regression coefficient is negative, this makes the communicator have a positive value since large communicator values result in a lower expected mutation rate.

The last group of terms deals with the non-basepairing geometry of the interactions a nucleotide makes. These terms are explained in greater detail in Chapter 4. To start our brief discussion here, we begin with BPhGeometry. BPhGeometry is defined for base-phosphate interactions. Each nucleotide, A, C, G, and U has unique places where they can donate a hydrogen atom to make a base-phosphate interaction. This term measures how easily a nucleotide making a base-phosphate interaction can be substituted while preserving the interaction. Next we consider the term BRGeometry. This term is defined for base-ribose interactions. Base-ribose interactions are hydrogen bonds between a base and the ribose
sugar. Similarly to BPGeometry, BRGeometry measures how easily the base in a base-ribose interaction can be substituted while preserving the interaction based on the fact that the each nucleotide A, C, G, and U, have unique places where they can donate a hydrogen atom to make a base-ribose interaction. The last non-basepairing term that we consider is Stacking Discrepancy Index (SDI). This term is defined for nucleotides that are making at least one stacking interaction. To compute this term, we consider the nucleotide and each of its stacking partners individually. We then compare the geometry of the observed stacking partners to stacking partners across other 3D structures. Based on the discrepancy of the nearest geometry of each nucleotide we build a score of how difficult the nucleotide is to replace.
CHAPTER 2

Background and assumptions

2.1 RNA background

2.1.1 The ribosome and RNA nucleotides

The ribosome is one of the world’s oldest machines, present in every living organism. The ribosome contains over 100 moving parts, each of them flexible. The ribosome is composed of RNA nucleotides, known as ribosomal RNA or rRNA for short, and proteins. Each nucleotide is made up of three major pieces. It is the nitrogenous base A, C, G, or U connected to a ribose sugar that is connected to a phosphate group. The nucleotides A and G are known as purines while the nucleotides C and U are known as pyrimidines. Purines are larger than are pyrimidines. The purines A and G are very similar in size to each other. Similarly, the pyrimidines C and U are similar in size to each other. rRNAs are non-protein-coding RNAs that constitute the ribosome, a universal cellular molecular machine that carries out template-directed synthesis of proteins, the primary molecular workhorses of cells. The rRNAs are highly structured molecules that carry out most of the recognition and catalytic functions of the ribosome, including binding of mRNA (the template), correct selection and binding of cognate tRNAs matching the next codon of the mRNA, and catalysis of peptide bond formation to add the next amino acid to the protein chain being synthesized. Proteins
are complex macromolecules that are composed of amino acids. The function of the ribosome is to assembles amino acids into proteins. This work focuses on the mutations that occur to the DNA that codes for the rRNA nucleotides that become fixed in a population over time. Therefore we limit our discussion to rRNA and mutations of these nucleotides.

2.1.2 Introduction to RNA 3D structure

The specific sequence of nucleotides that compose an RNA molecule is called the primary structure of the RNA. Some RNA molecules fold into a 3D shape. RNA molecules that fold on themselves into a 3D structure are known as structured RNA. Structured RNA molecules perform their biological function by virtue of the 3D shape into which the molecule folds.

The information about the 3D structure is obtained through a nobel award winning process known as x-ray crystallography. Using this method takes time, and it is difficult to get a resolved 3D structure to study so the 3D structure information of the rRNA that compose the ribosome is known for only a few organisms. There has been much work to infer the 2D structure information, or the mapping of the Watson-Crick interactions within the rRNA without the geometric information that we have in the 3D structure, from the primary sequence information. Some examples of methods of inference are mFold by [54] and SimulFold by [29]. mFold makes use of Turner energy parameters [47] to minimize the amount of free energy within the rRNA, whereas SimulFold uses a Bayesian approach to predict the 2D and more. As [29] describe in their paper: “SimulFold employs a Bayesian Markov chain Monte Carlo method to sample from the joint posterior distribution of RNA structures, alignments, and trees.”

It is well known that RNA helices are prevalent in structured RNA molecules. RNA helices are composed of complementary strands of RNA in which the pairs AU, UA, CG, and GC make Watson-Crick basepairs as in DNA. RNA 3D crystal structures show a variety of other interactions that are base specific. Non-Watson-Crick basepairs are planar interactions which involve edges of the RNA bases other than the Watson-Crick edge. Including the cis
Watson-Crick (cWW) / Watson-Crick family, there are 12 geometric families of basepairs. In addition, a nucleotide can make multiple basepairing interactions at the same time. These interactions can be a combination of cWW and non-cWW basepairs, or multiple non-cWW basepairs with or without a cWW pair. These types of interactions have been studied recently in [46], where a comprehensive understanding of the mutually isosteric basepairs is given. We will use the study of [46] to quantify the effects of basepairing interactions on site-specific mutation rates as will be discussed in Chapter 4. In addition to basepairing interactions, base stacking interactions are known to be prevalent in structured RNA molecules. From the 3D structure we are able to obtain information about the geometry of these stacking interactions. Bases in structured RNA can also make strong interactions with the phosphate oxygen of other nucleotides in base-phosphate interactions, or the oxygen of the ribose backbone in base-ribose interactions. By using the 3D structure information we are able to classify these interactions and also base-specific information about these interactions as will be discussed in Chapter 4.

2.1.3 Application of FR3D to our work

FR3D is the abbreviation for Find RNA 3D. It is the dissertation work of Michael Sarver [43], and has been under continual development ever since. FR3D is a suite of computer programs that can take as an input a Protein Data Bank, PDB, file corresponding to the 3D structure of any RNA 3D structure, then read and compute information relevant to the RNA 3D structure. For this work, the input that we give to FR3D is the 3D structure corresponding to rRNA of the ribosome. FR3D then classifies interactions such as basepairing, stacking, base-phosphate, and base-ribose interactions. As part of the programs we add to FR3D, FR3D also computes the information for distance from the center as well as for the sphere variables all of which will be defined in Chapter 4.

To classify basepairs, FR3D calculates six parameters describing the relative locations of two bases. Three parameters are contained in a vector that has three components, the $x$, $y$, $z$.
and z shift. The fourth parameter is an angle of rotation in the plane of the second base. The fifth parameter is the vertical component of the vector normal to the second base. The final parameter is the vertical “gap” at the interacting edge between the bases.

FR3D uses simple limits for these parameters. Pairs of bases whose parameters fall within these limits are annotated as FR3D basepairs, like cWW, tHW, tSS, etc. For each category, an “exemplar” pair is chosen. This is an actual instance which is at the “center” of the collection of basepairs in that category. Pairs whose parameters fall outside of all of the classification limits have their distance to all exemplars calculated. If the minimum distance is below a certain cutoff and if the relative directions of the normal vectors are the same (like both normal vectors point up in the instance and in the exemplar), then the nearest exemplar determines the “near” category that the pair will be annotated with. We used FR3D to analyze the 3D crystal structures of the E.c. and T.th. ribosome. We used files from the PDB with PDB ID’s 2QBD (16S rRNA) and 2QBE (23S and 5S rRNA) for E.c and 3I8H (16S rRNA) and 3I8I (for 23S and 5S rRNA) for T.th. These files provide us with complete information about the ribosome for E.c. and T.th.

2.2 Randomness in the replication process

2.2.1 Introduction to point substitutions of DNA

Mutations of any nucleotide in the ribosome are possible, but the mutations that are deleterious to the biological functioning of the organism have a lower chance of being passed along; while the advantageous mutations have a higher chance of being passed along. Neutral mutations which are mutations that neither help nor hinder the organism in any dramatic way. The mutations too are passed along. In comparing the variation seen in the multiple sequence alignments to the 3D structure information, we see that nucleotides that are part of interactions that we classify and nucleotides not making any type of interaction that we classify can mutate. Indeed, not only is it possible that the nucleotides within these regions will
change, but also that the size of the region itself may change with one or more nucleotides being inserted or deleted within these regions. Mutations that insert nucleotide, or delete nucleotides are commonly known as indel events. For more information about how indel events are modeled, we refer the reader to [48] and [49] which serves as the cornerstone of most modern research, and to the work of Ian Homes and his collaborators in [30] and [17] [4] who used the idea of transducers to model indel mutation events. However, the type of mutation event that we will consider in our work is point substitutions. A point substitution is the event in which a single nucleotide is replaced by a different nucleotide, not deleted and not the result of an indel event. For the focus of this work, point substitutions can be thought of as errors in the copying of nucleotides that compose the regions of DNA that code for rRNA. To give an example of the frequency of point substitutions, in human DNA, this type of error is somewhat rare occurring at the rate of one error in every 100,000 nucleotides. While this is a fairly low base error rate, over the entire DNA in one human cell, there are 3 billion basepairs which would amount to 120,000 mistakes every time the cell would copy itself before the cell employs its proofreading and error checking methods. [40]

2.2.2 Mutations of rRNA

In this work, we estimate mutation rates of bacterial genes (regions of DNA) coding for ribosomal RNAs (rRNA). Mutations to rRNA occur at the level of DNA and are permanent changes in a cell’s DNA. They are passed down to progeny cells. The mutations that occur in DNA can occur in either protein coding regions or in non-coding regions. Protein-coding regions are the regions of DNA that are responsible for some types of messenger RNAs, mRNAs, produced by transcription that go on to be translated into proteins. Non-protein-coding RNAs are extremely prevalent. Even mRNA transcripts, have non-protein-coding regions - the 5'- and 3'-UTRs (untranslated regions) and the introns, which are removed during splicing by the spliceosome. Non-protein-coding RNAs function as RNAs without coding also for proteins. rRNAs are non-protein-coding RNAs as are rRNA’s.
For the purpose of this work we consider only point mutations that have been functional enough to be passed from generation to generation since the time of mutation and to be the present in the ribosome that is representative of the population of that organism. Since the shape of the rRNA molecule changes due to mutation and the rRNA performs its function by virtue of its 3D structure, then the functionality of the RNA may very well be compromised. If the functionality of the ribosome is compromised, then the ribosome might not be able to serve the biological function for the organism potentially not allowing the organism to perform its metabolic functions which may result in killing the organism, or may lead to other organisms of the population out-competing the organism with the mutation. In either case this would prevent the mutation from becoming fixed in the population. Therefore we infer that the 3D structure is very slow to change relative to the nucleotides that compose the structure.

2.2.3 Mutations and the 3D structure

It is well known that many nucleotides in RNA form helices, and that helices are prevalent in structured RNA molecules. RNA helices are composed of complementary strands of RNA in which the pairs AU, UA, CG, and GC make Watson-Crick basepairs as in DNA. The occasional GU basepair is observed, but other base combinations are rare in helices. Thus, we say that the Watson-Crick basepair is base specific. In order to maintain a Watson-Crick basepair in a particular location in the 3D structure, compensating mutations must be made in order to maintain a similar structure in the helix. We see many examples of this in multiple sequence alignments. But we also see cases in which one nucleotide in a Watson-Crick basepair makes one or more additional interactions. For example, a nucleotide may make a base-phosphate interaction which prevents it from mutating, and thus also prevents its basepairing partner from mutating, otherwise such a critical interaction could not be formed. In this way, the local 3D structure limits the rate at which nucleotides can mutate. We give a depiction of the 2D and the 3D structure for each of E.c. and T.th. in Figure
2.1 and Figure 2.2. From these figures, we see that in the 2D structure, the nucleotides that compose basepairing interactions vary, the 3D structure is very similar between the two organisms.

2.3 Modeling assumptions on single-site mutations

2.3.1 Discrete versus continuous time

In our model, mutations take place in continuous rather than discrete time. While it is true that mutations occur at discrete points in time during replication from the parent molecule to the child molecule, bacteria may reproduce in as little as 15 or 20 minutes per organism [39]. This amount of time is very short when we consider that organisms in the multiple sequence alignment are millions or billions of years apart. Mutations are registered in the
populations during this replication process and how they are passed from one generation to the next. Since there are usually many organisms of a bacterial population, there are many replications occurring within a population over a very short amount of time. Therefore for mutations that occur at the population level, we model these as occurring in continuous time.

### 2.3.2 Markov process

We model the passing of mutations in a population of organisms to subsequent generations through the process of replication. In addition, we model that mutations only be passed on by the organisms that are currently reproducing. Therefore ancestral organisms that are no longer replicating or existing cannot affect the states of the nucleotides of future generations. To express this assumption mathematically we denote $X(t+s)$ to be the states of the ribosomal nucleotides observed in the rRNA at time $t+s$, $X(t)$ to be the states, b, at time $t$, and $X(u) = a(u)$ to be the states $a(u)$ for a given time $0 \leq u < t$. The Markov assumption can be expressed as $P(X(t+s) = c|X(t) = b, X(u) = a(u), 0 \leq u < t) = P(X(t+s) = c|X(t) = b)$, where $a, b, c$ are states of the ribosomal nucleotides at time $t+s, t$ and $u$ respectively. To better understand the mathematical expression of the assumption,
consider the states, $a(u)$, $b$, $c$ to be the state of a single nucleotide at their respective times. The equality means that knowing that $X(u) = a(u)$ for $0 \leq u < t$ does not provide us with any additional information about what the state, $c$, of the nucleotide will be at time $t + s$, and that all of the information that we can use is the state of the nucleotide, $b$, at the time $t$.

### 2.3.3 Evolution according to a Markov process

We now illustrate the Markov model, as defined above, of how mutations are passed along from a population over time. To do this we consider one population of organisms which are replicating. Over time, these organisms eventually split into two populations that are replicating independently of one another. We call the last organism that represents the population before the split the “last common ancestor”. The two populations now mutate according to their respective lineage over time. Over this time the organisms in each of the two populations continue to mutate and pass along these mutations through replication. If a mutation is so prevalent that it becomes prominent in a population, it can become “fixed” in the population. Eventually the organisms that represent each of the populations become so different from the last common ancestor that the organism that represents the population can be considered a different species. We call the length of time from the last common ancestor to each of the new species a branch length. For this example, we denote the time represented by these branch lengths from the last common ancestor to each of the new species as $t_1$, and $t_2$ respectively. Subsequent speciation events form a binary tree of decent from the common ancestor.
2.4 Phylogenetics

2.4.1 A brief introduction to a multiple sequence alignment

A multiple sequence alignment represents the correspondence between a set of homologous biological molecules, whether protein, DNA or RNA. In a multiple sequence alignment, the rows of the alignment represent the primary sequence of the organisms that are represented in the alignment. The columns of a multiple sequence alignment are the nucleotides from the organisms that correspond to each other with “gap” characters inserted in the primary sequences to account for indel events. The rows and columns of each multiple sequence alignment form a data matrix for our work. The organisms represented by the alignment are meaningful the organisms are descended from a common ancestor. We obtained three alignments; one alignment for each of 5S, 16S, and 23S rRNA.

2.4.2 Introduction to phylogenetics

The focus of this work is to model the site-specific mutation rates by the 3D structure, which means that we must first estimate the site-specific mutation rate for each rRNA nucleotide. Phylogenetics will be key in accomplishing this task. Phylogenetics can be understood as the study of the evolutionary relationship of organisms over time. Focused to this work phylogenetics is used to infer the evolutionary relationship of a set of organisms, such as we have in our study in the form of a multiple sequence alignment.

A common way to model this relationship is through a phylogenetic tree. A phylogenetic tree will contain information about how closely the sequences are related relative to one another with similar organisms nearer to each other on the phylogenetic tree. In a phylogenetic tree, the “root” of the tree will be the last common ancestor of all of the organisms represented. From this ancestor come descendant organisms connected by branches. The descendant organisms serve as ancestral organisms for the next descendant organisms, and so on until the organisms that we observe today, known as the leaves of the tree. This model
is commonly referred to as the tree topology. A complete tree will also contain information about how distant each organism is from one another and to their last common ancestor, which are commonly referred to as branch lengths. A tree that has two descendants for each ancestral sequences is known as a bifurcating tree. It should be noted that any phylogenetic tree can be expressed as a bifurcating tree by using branch lengths of zero.

2.4.3 Heuristic approach to phylogenetics

The background provided in this section is a brief introduction to some popular techniques for finding a phylogenetic tree topology. For a more complete background we will refer the reader to chapters four and eleven of *Inferring Phylogenies* [11]. Two of the most popular techniques are: Neighbor Joining methods such as UPGMA [44], [31], and searches based on Nearest Neighbor Interchanges (NNI) [11]. We begin with discussing a Neighbor Joining, NJ, approach. This approach is based upon a distance matrix, where the entries of the distance matrix are simply the pairwise number of differences between the sequences. In this process the “best” tree minimizes the number of mutations of the neighboring organisms in the tree. For this type of tree, the sums of the branch lengths from one sequence to another sequence should equal the entries in the distance matrix. This approach is guaranteed to recover the true tree if the distance tree is the same as the true tree [11]. However, there is no guarantee that this tree is the “best” tree. There are many packages that are freely available to infer the phylogenetic tree by heuristic methods. Some commonly employed heuristic packages are using NNI-type techniques are: Mesquite [27], ClearCut [9], and FastTree [41], [42]. Since these methods are heuristic approaches, and are much faster than maximum likelihood methods, they perform well for large multiple sequence alignments.
2.4.4 Probabilistic models of mutation

2.4.4.1 Generator matrix for individual nucleotides

We now concentrate our attention on the parameters of a probabilistic model for individual nucleotide mutation. We model each nucleotide as a Markov process taking values in the set A, C, G, U. The generator matrix contains the parameterization of the substitution rates from one state to another, and the matrix containing the stationary or equilibrium frequencies of the states; also called the limiting distribution that we will call $\mu_\infty$. An important property of the generator matrix is that the rows of the matrix each sum to zero. Also, the generator matrix has the property that all of the off-diagonal entries take non-negative entries while the entries on the diagonal take negative values. The diagonal entries can be thought of as the rate that the nucleotide remains in the current state, while the off-diagonal entries represent the rate at which a nucleotide will leave its current state for the state represented by the column of the matrix.

2.4.4.2 Jukes Cantor 1969 generator matrix

The first generator matrix for nucleotide substitution was proposed by Jukes and Cantor in 1969, commonly referred to as JC69, [20] we will use it as an example to illustrate how a generator matrix works. In the parameterization of the JC69 generator, the vector containing $\pi_A, \pi_C, \pi_G, \pi_U$, the limiting distribution of the nucleotides, $\mu_\infty$ is uniform. Thus under this generator matrix when a mutation occurs the nucleotide will be able to mutate to any of the other nucleotides with equal probability. [8] The form of the generator is given in the Table 2.1, where $\alpha$ is the single parameter of mutation rate.

2.4.4.3 Felsenstein 1981 generator matrix

In some cases, the JC69 model can be too simplistic to accurately model the behavior of nucleotides because the proportion of each type of nucleotide represented in a molecule
Table 2.1: JC69 generator matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3α</td>
<td>α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>C</td>
<td>α</td>
<td>3α</td>
<td>α</td>
<td>α</td>
</tr>
<tr>
<td>G</td>
<td>α</td>
<td>α</td>
<td>3α</td>
<td>α</td>
</tr>
<tr>
<td>U</td>
<td>α</td>
<td>α</td>
<td>α</td>
<td>3α</td>
</tr>
</tbody>
</table>

Table 2.2: F81 generator matrix

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>d_A</td>
<td>π_Cα</td>
<td>π_Gα</td>
<td>π_Uα</td>
</tr>
<tr>
<td>C</td>
<td>π_Aα</td>
<td>d_C</td>
<td>π_Cα</td>
<td>π_Uα</td>
</tr>
<tr>
<td>G</td>
<td>π_Aα</td>
<td>π_Cα</td>
<td>d_G</td>
<td>π_Uα</td>
</tr>
<tr>
<td>U</td>
<td>π_Aα</td>
<td>π_Cα</td>
<td>π_Gα</td>
<td>d_U</td>
</tr>
</tbody>
</table>

may not be uniformly distributed. For example, the nucleotide G composes 30% of the 23S ribosomal RNA molecule in *E. c.* , while the nucleotide U composes only 18% of the molecule. For reasons such as this, one may wish to alter the generator matrix to match the frequency each nucleotide contributes in composing the molecule. In 1981, Felsenstein proposed such a model, referred to as the F81 model [10], the generator is given in Table 2.2 , where again $\alpha$ is the mutation rate, $d_A = -(\pi_C + \pi_G + \pi_U) \cdot \alpha$, $d_C = -(\pi_A + \pi_G + \pi_U) \cdot \alpha$, $d_G = -(\pi_A + \pi_C + \pi_U) \cdot \alpha$, and $d_U = -(\pi_A + \pi_C + \pi_G) \cdot \alpha$. In the generator matrices given above, we see that each of the rows sums to zero, satisfying a property to be a generator matrix. Notice that the F81 model is simply an extension of the JC69 generator, the only difference lies in that we consider $\mu_\infty$ to have entries which are the distribution of each of the nucleotides as reflected in the data.

2.4.4.4 General time reversibility

Each of the models above are examples of the more general class of models for nucleotide substitution commonly known as general time reversible models. In these models, at equilibrium the rate at which one nucleotide mutates to another is equivalent in the sense that the mutation rate for nucleotide $i$ to mutate to nucleotide $j$ is the same rate for nucleotide
$j$ to mutate to nucleotide $i$.

### 2.4.4.5 Mathematical definition of general time reversibility

Mathematically, the assumption of general time reversibility may be expressed

$$\pi_i \text{Prob}(j|i, t) = \pi_j \text{Prob}(i|j, t).$$

This property means that the rate at which a nucleotide moves to state from state $i$ to state $j$ over time $t$ is the same as the rate of moving from state $j$ to state $i$ over time $t$. Models that satisfy this property are very popular because one can treat either the sequence for which nucleotide $i$ is a member, or the sequence for which nucleotide $j$ is a member as the ancestor of the other. Furthermore since there is not an ancestry structure placed upon the sequences, this allows the practitioner to perform any phylogenetic inference free of having to create a root for the phylogenetic tree, but rather may leave such trees unrooted.

### 2.4.4.6 Biological implications of general time reversibility

A case can be made for general time reversibility models being a valid model from a biological perspective because of the way that mutations occur as described under the framework of [40]. The process of mutation does not discriminate between a purine to become a pyrimidine or a pyrimidine to become a purine. Because most of the proofreading that occurs in transcription does not necessarily take this size constraint into account, from a blend of mathematical convenience and biological assumptions, a general time reversibility model is a pretty good compromise.

### 2.4.4.7 General Time Reversible Model

For models satisfying the general time reversibility property above, one could define the most general time reversible model of nucleotide substitution which will be referred to as the general time reversible model, GTR for short. This model was created in 1986 by Tavare [47], and can be specified as in Table 2.4
Table 2.3: GTR86 generator matrix

Here * are the quantities chosen such that each row within the matrix sums to zero. Here there is a mutation rate parameter, $\alpha$ for each type of mutation under time reversibility.

2.4.4.8 Summary of GTR

As stated above, the GTR model derives its name from the property that in the substitution parameterization, substitutions from nucleotide $i$ to nucleotide $j$ happen at the same rate for nucleotide $j$ to nucleotide $i$ at equilibrium. These models are well known and very popular in phylogenetics, and are used in the phylogenetics package MrBayes. However, these models treat mutations from purine to pyrimidine as the same as a mutation from a pyrimidine to a purine. There is no biological reason for this to be exclusively how nucleotide substitution occurs, but rather are used for their mathematical convenience.

2.4.4.9 Introduction to non-GTR models

We now define isostericity of RNA basepairs similarly to how it was considered in [46]. Isostericity is a biological concept dealing with how nucleotides in a baseparing interaction can substitute while preserving structure of the baseparing interaction. For example, consider an GC Watson Crick, cWW, basepair. If the C is to mutate, but we impose the restriction that after the mutation the cWW basepair must be preserved. For the substitution to easily preserve the interaction, GC pair will mutate to a CG pair. A near isosteric substitution is a substitution such as a GC pair mutating to a GU pair. The new pair contains a purine and a pyrimidine just as the old pair, but since the C and the U are different, the geometry of the pair will change and therefore classified as near isosteric. A non isosteric change is a
Figure 2.3: Examples of Watson Crick basepairs

GC cWW basepair changing to a GA cWW basepair. The geometry of the new pair is much different than the original basepair, and for this reason is classified as nonisosteric. Figure 2.3 provides an example of an isosteric, a near isosteric, and a nonisosteric substitution of a CG basepair.

Similarly if we consider a mutation to a single nucleotide, a substitution of a purine to the other purine will occur at a different rate than a purine to a pyrimidine because of the difference in the size of the nucleotides. Similarly, a pyrimidine will mutate to the other pyrimidine at a different rate than the rate that a pyrimidine will mutate to a purine. If we place the assumption that the mutation of a purine to the other purine is the same as the pyrimidine to the other pyrimidine is equal, then we have the K3P model[22]. We give the parameterization of the model in Table 2.4
In Table 2.4, $\alpha$ and $\beta$ are mutation rate parameters. $\alpha$ is the mutation rate parameter that models transitions, or mutations that stay within the class of nucleotide such as purine to purine or pyrimidine to pyrimidine mutations, while $\beta$ is the mutation rate parameter that models transversions, or mutations that do not stay within the class of nucleotide such as purine to pyrimidine or pyrimidine to purine mutations. Notice that this substitution model does not assume that the mutation rate for a purine to change to a pyrimidine is the same as a pyrimidine to mutate to a purine. Therefore, this model is not considered to be time reversible because it can be seen that it does not satisfy the definition of time reversibility as given above. However, this model could be thought of as being semi-time reversible in the sense that it does assume that within purine or within pyrimidine mutations do occur at the same rate.

2.4.4.11 Extension of K3P

Like its GTR counterpart, this model can be generalized to a 12 parameter model that has a substitution parameter for each possible substitution [11]. However, such a model would likely not be employed over a GTR model since a GTR model is not as parameter rich as a non-GTR model, and in practice because the user can in essence adjust the substitution rates found in the generator by the use of the matrix $\mu_\infty$.

<table>
<thead>
<tr>
<th></th>
<th>$A$</th>
<th>$C$</th>
<th>$G$</th>
<th>$U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>*</td>
<td>$\pi_{C\beta}$</td>
<td>$\pi_{G\gamma}$</td>
<td>$\pi_{U\beta}$</td>
</tr>
<tr>
<td>$C$</td>
<td>$\pi_{A\alpha}$</td>
<td>*</td>
<td>$\pi_{G\alpha_{CG}}$</td>
<td>$\pi_{U\gamma}$</td>
</tr>
<tr>
<td>$G$</td>
<td>$\pi_{A\gamma}$</td>
<td>$\pi_{C\beta}$</td>
<td>*</td>
<td>$\pi_{U\beta}$</td>
</tr>
<tr>
<td>$U$</td>
<td>$\pi_{A\alpha}$</td>
<td>$\pi_{C\gamma}$</td>
<td>$\pi_{G\alpha}$</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 2.4: K3P81: non time reversible generator matrix
2.4.4.12 Relationship with generator matrices

If indeed there is an effect on the way that nucleotides can mutate based upon their geometry then a GTR model may not be able to capture this and a non-GTR model may be needed. The exact parameterization of a generator matrix may not be known, and for researchers who are not certain which is the best model of nucleotide substitution for their ends may want to use PALM by [5].

2.4.5 Substitution models for multiple nucleotide mutations

Most models of more than one nucleotide mutation are centered around modeling the mutation of basepairs, and in particular the modeling of cWW basepairs. With 65.17% of the nucleotides in the conserved core of the E.c. molecule and 65.30% of the nucleotides in the conserved core of the T.th. molecule being part of a cWW basepair, it’s no wonder why substitution models for two nucleotide substitution outperform models that only consider single nucleotide substitution. We present the RNA16A model of nucleotide substitution from the Phase package of Gowri-Shankar and Jow [15], but we refer the reader not only to Phase, but also to [52], and [29] [33], for more information about the performance of models for more than one nucleotide substitution.

2.4.5.1 Phase16A model for substitution of two nucleotides

The Phase RNA16A model is the most general model that Phase employs. As an example of how one could model two nucleotide substitution, the RNA16A model is very tuned to cWW interactions simply because they are the most frequent as noted above. In addition to the sixteen (16) stationary frequency parameters, the Phase RNA16A model adds five (5) additional parameters to adjust the rate at which base pairs are mutating. Of these five parameters, four of them deal with cWW-type substitutions that is AU/UA, CG/GC, GU/UG pairs, with the GU/UG pairs being known as wobble pairs [52], with the fifth dealing with noncWW-type substitutions. With this parameterization, Phase uses an MCMC approach
written in C++ to estimate the phylogenetic tree and substitution matrix. We give the parameterization of mutation rates below.

\( \alpha_s \): A parameter for a single nucleotide mutation from A to G or G to A or C to U or U to C such that the initial pair and the pair resulting from the mutation are both one of the six cWW pairs mentioned above. \( \alpha_d \): A parameter for a double nucleotide mutation such that one of the six cWW pairs above is substituted for another of the pairs above such that the purine (A/G) and the pyrimidine (C/U) do not change positions. So examples of such mutations would be a mutation from AU to a GC a mutation from a CG to a UA. \( \beta \): A parameter for a single nucleotide mutation to or from the six pairs listed above \( \alpha_d \). \( \gamma \): A parameter for a double nucleotide mutation from one of the six pairs listed above to a mismatched pair, or a single nucleotide pair from a mismatched pair to a pair listed above. \( \epsilon \): A parameter for a single nucleotide mutation from a mismatched pair to another mismatched pair.

Note: All other substitutions not listed above are set to have a rate set to 0 in the generator matrix. These substitutions would include a double substitution from any of the six above to a “mismatch” pair. For example a mutation from GC to a CU pair can only happen through two single nucleotide substitutions. Also a mismatch to mismatch substitution like a mutation from AC to a CU can only happen through two single nucleotide substitutions, and finally a mismatch to one of the six above substitutions like a mutation from CU to a GC can only happen through two single nucleotide substitutions.

### 2.4.5.2 Codon model and beyond

It is also the case that researchers model three substitutions at a time. This is usually done for codons in the modeling of protein evolution. MrBayes along with other packages have predefined generator matrices to model such substitutions. For more information about modeling amino acids and proteins, or three nucleotides at a time, we refer the reader to the work of [24], or [36].
2.4.6 Maximum likelihood approach to tree estimation

While heuristic approaches provide quick and sensible tree topologies, we prefer a probabilistic way to evaluate and ultimately compare topologies, branch lengths and all of the other evolutionary parameters. In 1981 Joseph Felsenstein devised such a probabilistic approach [10]. The result of this work incorporates the substitution model of [20] for sequence mutations as well as the method of [10] to infer the likelihood of a tree topology, branch lengths, and evolutionary parameters. To give the formulaic expression of [10], first consider the small phylogenetic tree, Figure 2.4

Let \( l_1, l_2, \) and \( l_3 \) be the observed leaves. Let \( y \) be the unobserved last common ancestor between \( l_2, \) and \( l_3 \) and \( x \) be the unobserved last common ancestor between \( y \) and \( l_1 \). Let \( \tau_1, \tau_2, \tau_3, \tau_4 \) be the branch lengths connecting the nodes of the tree, let \( \theta \) be all remaining phylogenetic parameters such as the parameterization of stationary state frequencies, the site-specific mutation rate parameters, which we discuss later in this chapter and extensively in Chapter 3, and the proportion of invariant sites, which we will not discuss here but refer the reader to [19] for further information about setting these parameters. The likelihood of the tree topology, \( T \), of [10] may be expressed as

\[
L(T) = \prod_{x \in A,C,G,U} \prod_{y \in A,C,G,U} P(l_3|x, \tau_3, \theta) \cdot P(y|x, \tau_4, \theta) \cdot P(l_2|y, \tau_2, \theta) \cdot P(l_1|y, \tau_1, \theta). \tag{2.1}
\]
To give an example of how to compute a component of Equation 2.1, consider the probability $P(y|x, \tau_4, \theta)$. This quantity is the probability of observing the sequence $y$ given the last ancestor $x$ was reproducing along a lineage for the amount of time $\tau_4$. This probability, $P(y|x, \tau_4, \theta)$, is taken from the $x, y$ entry of the matrix $\exp(A\tau_4)$, where $A$ is the “generator” matrix; which takes the parameterization of one of the models or a similar model as defined above. In Chapter 3, we will extensively discuss how to estimate the generator matrix for each nucleotide and a way to speed the computation of calculating the likelihood.

2.4.7 Applying the probabilistic models of mutation

2.4.7.1 Markov Chain Monte Carlo (MCMC)

Even with improvements to the procedure for finding the maximum likelihood, as discussed in [11], the estimation process remains computationally expensive, and finding the optimal combination of phylogenetic parameters and trees is well known to be NP-Hard. Therefore the estimation was done through Markov Chain Monte Carlo estimation or MCMC for short. While this process does not directly find the maximum likelihood estimate, it proposes combinations of a tree and evolutionary parameters and tends to move toward such combinations that are more likely. For more information on MCMC in the context of phylogenetics, we refer the reader either to [28], or to [11] for a more biological context. MCMC is the technique that we will employ in this work, and more specifically the phylogenetic package MrBayes [19].

2.4.7.2 MCMC employed by MrBayes

We now discuss the MCMC algorithms employed by MrBayes. MrBayes uses a Metropolis-Hastings, MCMC algorithm to sample the space of tree topologies, branch lengths, and other phylogenetic parameters listed above to try to find the optimal tree topology. The algorithm works in this setting by taking an initial a tree topology, branch lengths from an exponential prior distribution, and the phylogenetic parameters: stationary parameters from a Dirichlet
prior distribution, site-specific mutation rates from a gamma prior distribution, and the proportion of invariant sites, sites that are modeled to not mutate, from a uniform prior distribution. MrBayes then proposes a second topology with branch lengths and evolutionary parameters by making some modifications to the initial tree such as swapping branches or moving an ancestral node. These small perturbations to the last tested tree topology also include a new set of branch lengths by sampling from an exponential distribution, as well as a new set of phylogenetic parameters. The likelihood is computed for both topologies with their respective parameters. If the likelihood associated with the proposed tree has a higher likelihood than the likelihood associated with the “initial” tree then the proposed tree becomes the “initial” tree for the next iteration. If the proposed tree has a smaller likelihood than the initial tree then the proposed tree is accepted with probability equal to the ratio of the likelihood of the proposed tree to the likelihood of the initial tree. This process continues until a predefined number of simulations is reached. With this method, we attempt to sample the space of trees, branch length, and phylogenetic parameters to find the combination of these terms that maximize the log likelihood along the lines of Equation 2.1.

2.4.8 Synthetic data simulation

Before we applied MrBayes to our data, we ran a simulation study to determine how well MrBayes would recover a phylogenetic tree from a synthetic sequence alignment with the underlying phylogeny known. We conducted a simulation that had 128 sequences of length 1280, with no gaps or missing data, to closely approximate the size of the multiple sequence alignments. We generated the sequences from a binary tree of evolution, meaning that we started with one common ancestor whose sequence was all A’s that gave rise to two children that gave rise to four grand children, and so forth until we arrive at the 128 sequences. Each site of the sequence mutated at a different rate, $\alpha$ which was set to $\alpha = \frac{10k}{1280}$, where $k = 1, 2, \ldots, 1280$ is the index of the site. We set the number of generations to be 128000,
sampled every 1000 generations, with a burn in of 25% of the samples. As a result of this simulation, MrBayes recovered the underlying phylogeny of synthetic datasets exactly giving us reason to be confident in applying it to our multiple sequence alignment. We will defer the discussion of the ability of MrBayes to recover the site-specific mutation rate until Chapter 3.

2.4.9 Known difficulties of inferring a phylogeny

There are numerous difficulties in the inference of a phylogeny. To begin, many of the parameters such as the tree topology, branch lengths, global and site-specific mutation rate are interdependent. That is, the estimated mutation rate for any site will depend upon the tree topology and in particular the placement of the observed sequences with respect to one another. Furthermore the problem is that the space of all possible trees and parameters is so large that finding the best combination of a tree and parameters is computationally prohibitive. However, it is important to find a good combination of the tree topology and branch lengths because if there is much variation in the nucleotides that are placed near each other in the phylogenetic tree then this could lead us to believe that mutation is happening over a relatively short period of time and that the mutation rate for that particular site is relatively high. Conversely, if the tree is constructed so that the variation observed is between molecules placed far away from each other we could be led to believe that the mutation rate of a given site is relatively low. If we should observe much variation among sequences which are relatively close in evolutionary time, we could be led to believe that this is due to a relatively high mutation rate at this position in the molecule. While conversely, the same variation over a long period of evolutionary time could led us to believe that the mutation rate is relatively lower.

Another difficulty in inferring a phylogeny and branch lengths is identifying the evolutionary history within the phylum. For example, two organisms from the Thermus lineage, T. aquaticus and T. maritima have parts of their rRNA that appears relatively similar so
that it gives the impression that these organisms are closely related. However, this does not reflect the true evolutionary closeness but rather the result of “thermophilic convergence”, or a high GC basepairing count, a response of organisms to the hot environment in which they live [50]. The difficulty of modeling a phylogenetic tree within the phyla of bacteria is reiterated in [38] since the bacteria within a phyla have not been extensively studied. In Chapter 3 we discuss how we account for some of these concerns.
CHAPTER 3

Estimation of site-specific mutation rates

3.1 Multiple sequence alignments of bacterial rRNA

3.1.1 Introduction to multiple sequence alignments

A multiple sequence alignment represents the correspondence between a set of homologous biological molecules, whether protein, DNA or RNA. In a multiple sequence alignment, the rows of the alignment represent the biological sequences from organisms that are represented in the alignment. The columns of a multiple sequence alignment of RNA are the nucleotides from the the organisms that correspond to each other [46]. The rows and columns of each multiple sequence alignment form a data matrix for our work. The organisms represented by the alignment are meaningful when they have an evolutionary relationship, or phylogeny, and are descended from a common ancestor. We obtained three alignments; one alignment for each of 5S, 16S, and 23S rRNA. The 16S and 23S rRNA alignments were obtained from the European Ribosomal Database, and the 5S alignment was obtained from the Posnan database.
Figure 3.1 shows a very small part of the 16S rRNA alignment as it appears in the multiple sequence alignment. In Figure 3.1, we can see that there is variation in the sequences that compose the ribosomal RNA of organisms. This variation in the alignments provides us with much evolutionary information about the organisms. We can use an alignments for two different, but very much related purposes. The first use is to determine an appropriate phylogenetic tree, thereby establishing evolutionary relationships. For this purpose, we keep as many columns as we can because it allows us access to the most information to help establish any evolutionary relationships. The other use is to be able to estimate column-specific, or structure specific mutation rates known as site-specific mutation rates and structure-specific rates respectively, which are the estimated rates at which each nucleotide or structure mutates. Later, we will build statistical models of these mutation rates using 3D structure information. For this purpose, we only use the columns that have nucleotides as entries in at least one of the 3D structure molecules.
3.2 Cleaning datasets for use

To estimate the site-specific mutation rates, we used the three alignments described above. These alignments consist of homologous, RNA sequences aligned that are aligned with gaps so that length of each sequence within each of the alignments is the same.

The size of the three multiple sequence alignments are are 101 sequences of length 149 for the 5S structure, 717 sequences of length 6562 for the 16S structure, and 136 sequences of length 9344 for the 23S structure. The datasets contain a variety of organisms from different bacterial phyla. We consider each multiple sequence alignment to be a data matrix with the number of rows equal to the number of organisms represented in the multiple sequence alignment. The number of columns of a multiple sequence alignment is the total number of characters that are needed to relate all organisms that compose the rRNA chain including gaps. We will discuss how to reduce the number of columns to a much more meaningful set shortly.

3.2.1 Reducing the dataset for computational efficiency

We begin cleaning the dataset by reducing the number of uninformative sequences, with respect to this work. Many sequences in the 16S dataset contained many gap ‘-’, unknown ‘N’, or missing ‘?’ characters. These sequences will not provide much phylogenetic information and due to the large computational demands of inferring a phylogeny are removed if they have more than one percent of any gap, unknown, or missing character in the sequence. Using the criteria listed above, the 16S data set was reduced to 136 sequences. The multiple alignments of 5S and 23S bacterial rRNA were left unaffected.

3.2.2 Determining a conserved core of ribosomal structures

A multiple sequence alignment containing all columns corresponding to the 3D structure is the most information that we can have. However, some of the columns of such a dataset will
not be usable. Since the goal of our work is to model site-specific mutation rates as a function of 3D structure, we must ensure that the structure that we use is comparable across bacteria. Since E. c. and T. th are believed to have a distant common ancestor since E. c. is from the phylum Proteobacteria, and T. th is from the phylum Deinococcus-Thermus. Therefore, if we determine a “conserved core” of nucleotides that are comparable between the E. c. 3D structure and the T. th. 3D structure, then we infer that other bacteria will have the same comparable structure. We used R3DAlign [34] to align the E. c. 3D structure to the T. th. 3D structure. We define the conserved core between the structures to be the set of nucleotides in one structure that are aligned to a nucleotide in the other structure. By using the nucleotides from this core, we have a good balance between being assured that nucleotides are making comparable interactions across bacterial 3D structures, while maintaining a large dataset to develop a meaningful model for mutation rates. After determining the nucleotides that compose the conserved core, we are left with 110, 1378, and 2473 columns in the 5S, 16S, and 23S alignments respectively. This is 94.02%, 90.07%, and 87.05% of the E. c. 3D structures.

We will use the variable \( k \) to denote the index of the column in an alignment, for each column \( k = 1, 2, \ldots, n \), for each alignment. To estimate site-specific mutation rates, we will model each column of the alignment \( k \) corresponding to a position in the 3D structure of a rRNA molecule to have an underlying site-specific mutation rate \( \alpha^{(k)} \). In this study, we ignore insertions and deletions, more commonly known as or indel events and denoted by gaps, as well as missing and unknown characters in the multiple sequence alignment. MrBayes ignore these events because they do not contribute to the underlying phylogeny. We ignore them because they do not provide any information regarding the rate a nucleotide mutates to a different nucleotide.
3.3 MrBayes inference of a phylogeny and site-specific mutation rates

The estimation of site-specific mutation rates depends upon accurate inference of the tree topology and the branch lengths, which are all interdependent. The explanation how the phylogenetic trees, the branch lengths and global mutation rate are estimated is discussed in Chapter 2. In order to keep the focus on the estimation of site-specific rates we will not reiterate that discussion here, but rather discuss how we used MrBayes to infer the phylogeny. Using the alignments as described above, for the 23S alignments, we employed MrBayes version 3.2 to simultaneously estimate the phylogenetic tree and evolutionary parameters such as branch length, global mutation rate, using the F81 model, and the mutation rate for each column in the alignment. MrBayes uses the result of [51] that site-specific mutation rates are approximated well by a discretized gamma distribution, split up into regions or bins as defined by their center. The most bins that MrBayes permits is 19, and that is what we used in our estimation process.

MrBayes was run on the 23S multiple sequence alignment using a Mac Book Pro containing 2.33 GHz dual core Intel processor with 2Gb of 667 MHz RAM. To obtain convergence of a phylogenetic tree, as defined in the MrBayes manual of having standard deviation of the split frequencies less than 0.05, the dataset had to be reduced. Convergence was obtained relatively quickly taking only 556,000 simulations.

The 5S and 16S alignments were run on the Glenn Cluster at the Ohio Super Computing Center (OSC) in Columbus, Ohio using the parallel version of MrBayes [3]. The 5S was run using eight processors for a total of 3,000,000 simulations. Convergence was obtained. While the 5S multiple sequence alignment could have easily been run on the Mac Book Pro, for expediency of obtaining the phylogenetic tree and parameters, the OSC was used for this problem and required less than a day. It was particularly hard to obtain convergence of a phylogenetic tree for the 16S multiple sequence alignment, despite the reduction in the size of
the dataset to 136 rows, 1378 columns from 717 rows and 1561 columns, and the use of eight chains on the Glenn Cluster at the OSC for the maximum allotted time of four days using eight processors. Despite using the maximum allotted time on the Glenn Cluster, the run was extremely close to, but did not achieve convergence in the strictest sense with the final standard deviation of split frequencies being 0.059. Despite not achieving convergence to the standard described in the MrBayes manual, the traceplot and convergence criterion for each of the site-specific rates provided evidence that we were safe in using these estimates. Each of the three runs used a burn-in period of 25% of the runs. Because of the issues of inferring a phylogeny described in [38], we had our 16S tree validated by Jim Brown at North Carolina State University who found them to be a credible model of the evolutionary history.

3.4 Estimating site-specific mutation rates

3.4.1 MrBayes estimates of site-specific mutation rates

We had MrBayes estimate site-specific mutation rates. We give some basic graphs to visualize the distribution of the site-specific mutation rates for the \( E.c. \) and the \( T.th. \). The histograms are of the log mutation rate since we model the log of the mutation rates in Figure 3.2.

We will give a few descriptive statistics about the data on the log scale. The mean is -1.2676, the standard deviation is 1.7220, the median is -1.081, and the interquartile range is 2.8243. Of the 3961 estimated site-specific rates, 20 sites are estimated to take the minimum value, 0.0169. These sites correspond to places in the multiple sequence alignment where there is 100% conservation. In addition, one site is estimated to take the largest value of 6.1288. The distribution is nearly 20% for each nucleotide and gap character in the alignment for that site.
3.4.2 Modeling site-specific rates

We used MrBayes for two major purposes. They were to estimate the site specific mutation rates, but also to estimate the phylogenetic tree with all of the accompanying parameters that go along with this such as branch lengths. The estimation process of simultaneously estimating the phylogenetic tree as well as the site-specific mutation rates is every computationally intensive. MrBayes is designed to estimate the phylogenetic tree very well gives a rough idea of the site specific mutation rates. We now separate the two estimation processes by using MrBayes to estimate the phylogenetic tree with branch lengths, and separately estimate the site specific mutation rates for each of the sites, taking full advantage of the phylogenetic tree as computed by MrBayes. In Figure 3.3, we give same columns of the 16S as in Figure 3.1 after rearranging the rows of the multiple sequence alignment according to the phylogenetic tree. Using this rearrangement of the sequences we can estimate site-specific mutation rates directly.

Figure 3.2: Distribution of the natural log of the MrBayes estimated site-specific mutation rates from 5S, 16S, and 23S.
3.4.3 Mathematical model for a mutation rate

To obtain estimates of site-specific mutation rates, we consider a likelihood function similar to that of computing the likelihood for a tree topology. The likelihood to compute uses the framework of the [10] model used to compute the likelihood of tree topologies given the branch lengths, and phylogenetic parameters, but changed to model the likelihood of a site-specific mutation rate as given by [51]. We consider the site-specific mutation rates to be a function of the tree topology $T$, set of branch lengths $\tau$, and the set of observed sequences representing the modern day organisms, called the leaves of the tree, $L$. To better understand how this framework is applied to the problem of estimating site-specific rates, we provide the following example, and the corresponding Figure 3.4.
In this example, x and y are the states of the unobservable ancestral sequences at that site, *t*₁ - *t*₄ are the branch lengths, and *l*₁ - *l*₃ are the states of the observed sequences at the site. The likelihood is then computed as

\[
L(\alpha|T, \tau, D) = \prod_{x \in A, C, G, U} \prod_{y \in A, C, G, U} P(d_1|\alpha, x, t_1) \\
\times P(y|\alpha, x, t_4)P(l_2|\alpha, y, t_2)P(l_3|\alpha, y, t_3)
\]

(3.1)

Since finding the value of \(\alpha\) that maximizes \(L(\alpha|T, \tau, D)\) is the same as finding the value of \(\alpha\) that maximizes the value of \(\log(L(\alpha|T, \tau, D))\), we choose to find the value of \(\alpha\) that maximizes \(\log(\alpha|T, \tau, D)\) for the computational convenience that it provides. So for our example, we solve the equivalent problem of finding

\[
\arg \max_{\alpha} \log(\alpha|T, \tau, D) = \arg \max_{\alpha} \left( \sum_{x \in A, C, G, U} \sum_{y \in A, C, G, U} \log(P(l_1|\alpha, x, t_1)) + \log(P(y|\alpha, x, t_4)) + \log(P(l_2|\alpha, y, t_2)) + \log(P(l_3|\alpha, y, t_3)) \right)
\]

(3.2)

In estimating the site-specific mutation rate, \(\alpha\), it is the goal to find the argument of \(\alpha\) that maximizes this likelihood function. For the estimation process of \(\alpha\), MrBayes models the rates according to a **molecular clock**. The **molecular clock** modeling assumption for single nucleotide mutation is that the nucleotide mutates at a consistent rate over time. Each nucleotide is modeled to be mutating according to its own molecular clock. There has been much previous work showing that the rate of mutation across sites in the ribosome is not uniform [25] [32] [26] [33] [45]. Therefore the assumption that mutations occur according to a molecular clock does not always model the underlying biology well. However, the scope of this work is to model the rate of single nucleotide mutation as a function of the 3D structure.
This means obtaining one estimate of nucleotide mutation. This model is reasonable because the goal of this work is to relate the rate at which any nucleotide is mutating according to its 3D structure. The estimate that would best represent the rate of site-specific mutation is the estimate that is the consensus rate given the tree and the observed leaves. Relaxing the molecular clock assumption would a major complication in our ability to obtain the site-specific mutation rates. We illustrate two of the major complications here. The complication is that by relaxing the molecular clock assumption, each nucleotide mutates at a separate rate along each lineage. In the example above, this alters the problem of finding the one argument that maximizes the likelihood function to finding a consensus argument of $\alpha$ from the four mutation rates obtained from each lineage. In addition, if we examine the estimation of one of these four mutation rates, we see that we must estimate them two at a time. For example, we see that computing the mutation rates for the portion of the tree from $y$ down would be to find:

$$\arg \max_{\alpha_2 \alpha_3} \log(\alpha_2 \alpha_3 | T, \tau_2, \tau_3, y, D)$$

$$= \arg \max_{\alpha_2 \alpha_3} \sum_{y \in A,C,G,U} \log(P(d_2 | \alpha_2, y, t_2)) + \log(P(d_3 | \alpha_3, y, t_3))) \quad (3.3)$$

This would be very difficult to do especially if $d_2$ and $d_3$ are the same nucleotide. In general over a larger tree, if the two descendant nucleotides are the same then we are very likely to infer that the ancestral nucleotide matches the descendant nucleotide. This makes it hard to recover the history of mutation well.

### 3.4.4 A maximum likelihood estimation of mutation rates

Notice that in Equation 3.2 we must cycle over all of the possible states for the ancestral nucleotides. In this expression of the log-likelihood, this means that we must repeat calcula-
tions from site $y$ on down the tree. For example, the contribution to the likelihood if $y = A$ on down to $l_2$ and $l_3$ is the same for when $x = A, C, G$, or $U$. However, under the assumption that mutations occur according to a Markov process, the nearest common ancestor to an organism is the only source of mutation for a given organism. Under this assumption, Felsenstein addressed this inefficiency by creating a dynamic programming algorithm to economize on the number of computations needed. [11] The idea of this is to move the summations as far right as possible. For our example, we rewrite Equation 3.2 as

$$\arg \max_{\alpha} \log(\alpha|T, \tau, D) = \arg \max_{\alpha} \sum_{x \in A, C, G, U} \log(P(l_1|\alpha, x, t_1)) + \sum_{y \in A, C, G, U} \log(P(y|\alpha, x, t_4)) + \log(P(l_2|\alpha, y, t_2)) + \log(P(l_3|\alpha, y, t_3))$$

(3.4)

This result allows for the computation of $y$ on down the tree to be computed, and for this contribution to be simply stored instead of being recalculated each time.

The likelihood function that we use for this estimation process is built on the framework of [10]. Within this framework, we model nucleotide mutations according to the one parameter Jukes-Cantor model of mutation for each nucleotide in the conserved core of the multiple sequence alignment. We model according to the Jukes-Cantor model because the other models of nucleotide substitution have many more parameters, and for some sites, none of the parameters would be observed. For example consider a column that has all $G$’s. For this column, we say that the rate of nucleotide mutation according to the Jukes-Cantor model is as low as possible. However in a model of nucleotide mutation that has two parameters, the rate of mutations from a purine, $A$ or $G$, to a pyrimidine, $C$ or $U$, and vice versa is modeled by a parameter. In this case, we have no hope of modeling the rate at which a pyrimidine mutates to a purine. In addition, for the scope of this work, we are interested in the rate at which the nucleotide mutates and not necessarily to what it mutates so a more complex
model is not needed. Therefore we use the JC69 model of single nucleotide substitution. In addition, this model does not suffer due to a lack of data. For example, if there is a column that has 50% C, and 50% U, it is difficult to meaningfully estimate the rate a C or a U will be substituted for an A or a G with a model that is more complex than the JC69 model, but much easier to meaningfully estimate the rate of nucleotide substitution.

3.4.5 Calibration of branch lengths

Since we have inferred the phylogeny for three different datasets of 5S, 16S, and 23S rRNA independently, we must attempt to ensure that the inferred branch lengths are all on similar scales. Therefore we found the distance in branch lengths from the E.c. sequence to the T.th. sequence on each of the three phylogenetic trees. Since the 5S, 16S, and 23S sequences vary greatly in length from each other, and with more columns of nucleotides we believe that we would get more reliable phylogenetic trees, we took the weighted average of the distances of the three distances from E.c. to T.th. and defined that to be the “true” distance from E.c. to T.th. Using this distance, we then assumed that the branch lengths of a tree varies proportionally to the branch lengths of a tree with the “true” distance from E.c. to T.th. For example suppose that the true distance in branch lengths from E.c. to T.th. was 1.0, but for the 16S tree it is 1.1. We then divide all of the branch lengths on the 16S tree by 1.1 in order to calibrate the tree.

3.4.6 Making a maximum likelihood estimation of rates

With the mathematical formulation of the likelihood for the site-specific rates solidified, we now turn our attention to how we search for the value of $\alpha$ over a positive interval that maximizes the log-likelihood value for each site. Our goal is to estimate each site-specific mutation rate by determining the value for a mutation rate that maximizes a likelihood function for the mutation rate $\alpha$ given a tree topology $T$, the set of branch lengths $\tau$, and the observed character for each sequence, the set of which we call $D$, under the molecular
clock modeling assumption as described above, to within a tolerance $\epsilon$ that is set by the user of this algorithm. We begin the process by requiring that every mutation rate takes a positive value. Since the mutations observed in rRNA occur in DNA, and every DNA nucleotide is subject to change, we model that the mutation rate of all rRNA nucleotides is positive. Thus if a mutation rate could take values other than positive values, this model would not be defined. This is a similar assumption that [19] use since MrBayes incorporates the result of [51], which assumes that site-specific mutation rates follow a Gamma distribution. Unlike MrBayes, we do not make any assumption upon the underlying distribution of the site-specific rates.

3.4.7 Performance of the ML algorithm

We sweep over a positive interval, $[0.01, 10.01]$ at equally spaced points according to the user defined tolerance $\epsilon$. In this method we make an exhaustive sweep over the interval until the first $\alpha$ which results in a smaller likelihood than the value of $\alpha$ that was higher, or second to last to be computed is taken as the most likely value of $\alpha$. This process however is very slow being only able to compute about 21 sites per hour using an error tolerance of $\frac{1}{25000}$. With this tolerance to estimate the site specific mutation rates for all sites in our study of 3961 sites would take approximately four and a half days to estimate.

3.4.7.1 A quicker way to search for the ML estimator

Since the process as described above takes so long, we wish to devise a method of searching over the interval $[0.01,10.01]$ in a more efficient way. From the exhaustive search, we observed that the log-likelihood function appeared to be a unimodal function of $\alpha$. The idea is to pick $n$ points evenly spaced along the interval with two points being the endpoints of the interval. We then calculate the likelihood for each of these points, and determine which of the $n$ points maximizes the likelihood with respect to $\alpha$, call this point $n^*$. Since we work under the assumption that the likelihood function is unimodal with respect to $\alpha$, we know
that the true maximum likelihood must exist on the interval \([n^*_-, n^*_+]\), where \(n^*_-\) is the closest equally spaced point less than \(n^*\), and \(n^*_+\) is the closest equally spaced point greater than \(n^*\), provided that \(n^*\) is not an endpoint of the interval; we will discuss that case later. Therefore we reduce our search from the original interval to the interval \([n^*_-, n^*_+]\), which is only \(\frac{2}{n}\) of the original interval. We then take the new, smaller interval and repeat the process until the length of the interval is less than or equal to the predefined tolerance. In the case that the \(n^*\) is an endpoint, then we set the new interval to be \([n^*, n^*_+]\) if \(n^*\) is the left endpoint, and \([n^*_-, n^*]\) if \(n^*\) is the right endpoint. In this case, where \(n^*\) is an endpoint, the new interval will be \(\frac{1}{n}\) of the original interval.

3.4.7.2 Finding the optimal parameters for the programs

In the procedure described above, the number of computations is the product of the number of points that represent the subintervals into which the interval is split, which we will called \(n\), and the number of iterations for this process, which we call \(m\). The number of computations needed is the function \(c(m, n) = mn\). Our goal will be to minimize \(c(m, n)\) with respect to \(m\) and \(n\). In the worst case scenario every value of \(n^*\) will be a value other than an endpoint. This is the worst case because the reduction of the interval will be the smallest which requires the most computations until the interval is less than \(\epsilon\). Under this assumption, every time that we create a smaller interval we are left with \(\frac{2}{n}\) of the original interval. Therefore if we consider \((\frac{2}{n})^m = \epsilon\), the largest interval satisfying the tolerance \(\epsilon\), we minimize \(c(m, n)\) subject to \((\frac{2}{n})^m = \epsilon\). Solving for \(m\) we see that \(m = \frac{\log(\epsilon)}{\log(\frac{2}{n})}\). We then can express \(c(m, n)\) solely as a function of \(n\). So we minimize \(c(n) = \frac{n \cdot \log(\epsilon)}{\log(\frac{2}{n})}\). Setting \(\frac{d}{dn}[c(n)] = 0\), we see that \(n = 2e \approx 5\).

This means that \(m = -\log(\epsilon)\) is a good choice. To verify that this value of \(n\) minimizes \(c(n)\), we note that \(\frac{d^2}{dn^2}[c(n)]|_{n=2e} = -\frac{\log(\epsilon)}{2e}\) which is less than zero provided that \(0 < \epsilon < 1\). So in our estimation of site specific mutation rates, we set \(n = 5\).
3.4.8 Results of a simulation study

To test if the ML algorithm is estimating site specific mutation rates well, we performed a simulation study based upon synthetic multiple sequence alignment data similar to that which we created above. Here, we perform the study by generating synthetic data, inferring the phylogeny, estimating the site-specific mutation rates 30 times, then we compare the mean of the estimate for the mutation rate of each site to the true mutation rate, the mean MrBayes estimate for the site, and to the mean entropy based on the nucleotides observed at each site. We find that the average estimates obtained through the maximum likelihood estimation are highly correlated to the true mutation rate, to the average MrBayes estimates, and somewhat correlated to the the average observed entropy of the sites. We see that the correlation between the average estimates obtained through maximum likelihood marginally out performs the average MrBayes estimates. The mean squared error of the average mutation rates estimated by maximum likelihood versus the true rate is 0.41 while the the mean squared error of the average of the mutation rates estimated by MrBayes versus the true mutation rate is 4.42. In addition, the correlation of the estimates obtained through maximum likelihood to the true rates was 0.9560 while those obtained by MrBayes was 0.9422. We believe that difference is due in large part to the average of the maximum likelihood estimates are able to detect the larger mutation rates better than average of the MrBayes estimate. In Figure 3.5 we give a graphical comparison of each of the average of the estimated mutation rates by maximum likelihood versus the average of the MrBayes mutation rates. The MrBayes mutation rates however are more highly correlated to the average observed entropy. The correlation was 0.5105 for the estimated obtained through maximum likelihood estimation, and was 0.6772 for those obtained through MrBayes. From this simulation study we feel confident in using the mutation rates estimated by maximum likelihood instead of by MrBayes.
3.4.9 Mutation rates estimated by maximum likelihood

In a manner similar to exploring the mutation rates generated by MrBayes, we will provide descriptive statistics to the mutation rates that we have estimated from RNA using maximum likelihood.

We will give a few descriptive statistics about the data on the log scale. The mean is 0.7215, the standard deviation is 1.104, the median is 0.2562, and the interquartile range is 0.9520. Of the 3961 estimated site-specific rates, five sites are estimated to take the minimum value, 0.0100. These sites correspond to places in the multiple sequence alignment where there is 100% conservation. In addition, one site is estimated to take the largest value of 8.599. The distribution is nearly 15% for each nucleotide and 40% gap characters in the alignment for that site.

3.4.10 Hardware used

Since the programs that we ran to compute the estimates of the site specific rates outside of MrBayes do not exist in any package available at the OSC, we ran the estimation program
Figure 3.6: Distribution of the natural log of the site-specific mutation rates estimated by maximum likelihood from 5S, 16S, and 23S on the Mac Book Pro mentioned above. These programs are written in Matlab.
CHAPTER 4

Defining and calculating terms from the 3D structure

4.1 RNA 3D structures

While the ribosome is one of the most conserved structures, mutations still occur giving rise to sequence variability that we modeled by the mutation rates above. Despite sequence variation and mutations, the ribosome much be able to maintain the 3D structure well enough that the RNA performs its vital function in the organism. In [53], it was shown that the structures of *E.c.* and *T.th.* are highly conserved in that the nucleotide interactions found in the structure are very similar between *E.c.* and *T.th.* Since *E.c.* and *T.th.* are placed far apart in evolutionary history with respect to the species represented in our dataset, we assume that many of the nucleotide interactions that occur must be preserved in order to maintain the functionality of the structure despite the variation that we observe, and that the other species represented by sequences in our study also have these nucleotide interactions preserved despite the fact that we do not have the 3D structures for them. Therefore for our study we will use the crystal structures for each *E.c.* and *T.th.* in our attempt to explain the mutation rate that we have observed.
4.2 Introduction to non-basepair interaction type variables

When thinking about interactions between nucleotides on the molecular level, one might think in the standard Watson Crick sort of way like DNA forming a helix. However, it is the case in rRNA that nucleotides also create or are a part of interactions that do not require the base of another nucleotide. In this section, we explore some of these non-basepairing interactions and how they relate to site specific mutation rates. The interactions that will be discussed in this section are: whether the nucleotide is near an amino acid, or amino acids, the number of stacking partners a nucleotide has and their geometry, the number of times its base is used to interact with the phosphate in the backbone of the another nucleotide, known as a base-phosphate interaction, and its geometry, the number of times that the phosphate within the backbone of a nucleotide interacts with the base of another nucleotide, the number of times its base is used to interact with the ribose in the backbone of the another nucleotide, known as a base-ribose interaction, and its geometry, and finally the local density of nucleotides surrounding a given nucleotide. We begin by discussing stacking interactions and then the base and phosphate interactions within the context of base phosphate interactions as defined by [53]. We will conduct a brief exploration into the association or the correlation between each of the potential predictors and the site specific mutation rate, but will forgo the

4.3 Distance from the center of the ribosome

4.3.1 Definition of DistanceFromCenter of the Ribosome

We begin with introducing our most straightforward potential predictor of site specific nucleotide mutation is the Euclidean distance that each nucleotide is from the center of the ribosome, with respect to nucleotides, as measured in angstroms. That is if the point \((x_c, y_c, z_c)\)
denotes the center of the ribosome, and \((x_1, y_1, z_1)\) be the point that denotes the center of a nucleotide, then the value computed as the distance for that nucleotide is given by 
\[
\sqrt{(x_c - x_1)^2 + (y_c - y_1)^2 + (z_c - z_1)^2}.
\]

### 4.3.2 Exploring the DistanceFromCenter variable

We begin exploring the relationship between DistanceFromCenter and the log of the mutation rates for each of \(E.c.\) and \(T.th.\) molecules through the Spearman correlation coefficient. This coefficient is 0.2800 and 0.2839 for \(E.c.\) and \(T.th.\) respectively. We continue our exploration of DistanceFromCenter through a simple Scatterplot for each of \(E.c.\) and \(T.th.\).

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of \(E.c.\) and \(T.th.\). For \(E.c.\) the mean DistanceFromCenter is 75.290 with standard deviation 24.390; while for \(T.th.\) the mean DistanceFromCenter is 74.870 with standard deviation of 24.190. The median for \(E.c.\) and \(T.th.\) is 77.380 and 76.940 with interquartile range of 33.130 and 32.270 respectively. Lastly the spread of \(E.c.\) is 142.50 with minimum value of 4.9560 and maximum value of 147.50; while the spread of \(T.th.\) is 142.60 with minimum value of 2.4640 and maximum value of 145.10. We conclude this exploration with a histogram to give the distribution of DistanceFromCenter for each of
4.4 Sphere-type variables

4.4.1 Introduction to the Sphere Variable

While the distance that a nucleotide is away from the center of the structure is a powerful predictor of mutation rates, it does not have a very clear biological meaning in that it is not relating the function of the ribosome and the mutation rates in some biological sense. From Distance we learned that nucleotides closer to the center of ribosome seem to mutate more slowly than the nucleotides that are further from the center. Near the center of the ribosome, we find more nucleotides on average. This lead us to think that it was not the distance that a nucleotide was from the center of the ribosome, but rather the local density that was constraining against mutation. This idea gave rise to the sphere-type variables that we describe here.
4.4.2 Standardizing the radius for sphere-type variables

The idea of this class of variable is to build a sphere of a specified radius around a nucleotide, and count the number of elements in the sphere. To properly build this variable we need to determine the optimal radius for each sphere type. To do this we run a computational study in which we chose a radius for the sphere-type variable and then computed the $R^2$ for modeling the mutation rate as a log linear model of only the variable and the intercept term for each of E.c. and T.th. Since $R^2$ is a unimodal function with respect to radius the variable, we find the radius that maximizes the $R^2$ value for each of E.c. and T.th. In doing this, we obtain different radii for each of E.c. and T.th. Since the $R^2$ is slow to change with respect to changes in the radius for all sphere types, we take the consensus radius of E.c. and T.th. by taking the average of the two individual, optimal radii.

4.4.3 NucleotideSphereCount

4.4.3.1 Defining the NucleotideSphereCount Variable

The idea of NucleotideSphereCount is to count how many other nucleotides are “near” the nucleotide. For every nucleotide we build a sphere around the center of the nucleotide with a set radius and then count the number of nucleotides that have their center within this sphere. That is if $(x_i, y_i, z_i)$ is the point that denotes the center of nucleotide $i$ for which we wish to build a sphere around, then for all other nucleotides with center $(x_j, y_j, z_j)$, we check to see if $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} < r$, for a predefined radius, $r$, for $j \neq i$. The value of NucleotideSphereCount for nucleotide $i$ then is the number of nucleotides $j$ such that $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} < r$.

4.4.3.2 Determining the optimal radius

By our definition of NucleotideSphereCount, it begs the question, how do we choose a radius, $r$, for the sphere surrounding each nucleotide. To do this, we ran a computational study in
Figure 4.3: Scatterplot of NucleotideSphereCount vs. log mutation rates for *E. c.* and *T. th.*

which we chose a radius for the sphere and then computed the $R^2$ for modeling the mutation rate as a log linear model of only NucleotideSphereCount and the intercept term for each of *E. c.* and *T. th.* The best radius from each of the datasets was a radius of 16 angstroms, and 18 angstroms for *E. c.* and *T. th.* respectively yielded in the highest $R^2$ value. Thus we set the radius to be 17 for both structures. However, it is the case that the $R^2$ value is a relatively flat function of the radius with small changes to the $R^2$ value corresponding to small changes in the radius.

4.4.3.3 Exploring the NucleotideSphereCount variable

We begin exploring the relationship between NucleotideSphereCount and the log of the mutation rates for each of *E. c.* and *T. th.* molecules through the Spearman correlation coefficient. This coefficient is -0.3314 and -0.3475 for *E. c.* and *T. th.* respectively. We continue our exploration of NucleotideSphereCount through a simple Scatterplot for each of *E. c.* and *T. th.*

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E. c.* and *T. th.*. For *E. c.* the mean NucleotideSphereCount is 26.800 with standard deviation 6.4510; while for *T. th.* the mean NucleotideSphereCount is 27.360 with
Figure 4.4: Histogram of NucleotideSphereCount for E.c. and T.th.

standard deviation of 6.6650. The median for E.c. and T.th. is 27 and 27 with interquartile range of 9 and 1 respectively. Lastly the spread of E.c. is 45 with minimum value of 3 and maximum value of 48; while the spread of T.th. is 45 with minimum value of 3 and maximum value of 48. We conclude this exploration with a histogram to give the distribution of NucleotideSphereCount for each of E.c. and T.th..

4.4.4 Near or not near the plane of the nucleotide

We further classified the nucleotides in the NucleotideSphere by if they were in a similar plane to the nucleotide or if they were not in a plane similar to the nucleotide. By doing this we will answer the question if local density of nucleotides that are in a plane similar to itself or nucleotides that are not in a plane similar is the physical attribute that effects nucleotide mutation. To begin we must define what constitutes a nucleotide to be considered in a plane similar to the nucleotide. In addition to classifying the interactions of the 3D structure of a molecule, FR3D also gives a rotation matrix that would rotate the nucleotide from its current orientation in the 3D structure to a standard orientation. To determine if two nucleotides, say nucleotide $i$, and nucleotide $j$, are in a similar plane, we consider the following: First we compute the difference of the two centers to obtain a vector, $p$, that points from nucleotide $i$
to nucleotide $j$. We then apply the rotation matrix that orients nucleotide $i$ from its current position to the standard orientation to the vector $p$, giving us the vector $r$. For the sake of calculation we define $u$ to be the normalized vector $r$. That is $u = \frac{r}{\|r\|}$. Next we turn our attention to the vertical component of $u$ by taking the dot product with the normal vector $\langle 0, 0, 1 \rangle$. Looking at the triangle formed, we see that $\cos(\theta) = \frac{n \cdot u}{\|u\|}$, or rather without loss of generality $\cos(\theta) = \frac{n \cdot u}{\|u\|}$ so that $\theta$ is in the interval $[0, 90]$. Since cosine is a decreasing function of $\theta$ over this interval, if $\theta$ is greater than a predetermined cutoff angle of depression from the normal vector, then we conclude that nucleotide $i$ and nucleotide $j$ are in a similar plane, and not in a similar plane otherwise.

### 4.4.4.1 Determining the cutoff angle

Similar to choosing the radius of the NucleotideSphere, this begs the question as to how to choose the angle of depression, $r$, for the sphere surrounding each nucleotide. To do this, we ran the same computational study as NucleotideSphere in which we chose the angle $\theta$ for determining how close to the plane nucleotide $j$ has to be to the plane of nucleotide $i$ to be considered to be in a similar plane, while keeping the radius fixed at the value stated above. We then computed the $R^2$ for modeling the mutation rate as a log linear model of only NucleotideSphereNearPlaneCount, and the intercept term for each of E.c. and the T.th. The best angle for this term was 57 degrees for ec. and 58 degree for T.th.

### 4.4.4.2 Exploring the NucleotideSphereNearPlaneCount variable

We begin exploring the relationship between NucleotideSphereNearPlaneCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.4161 and -0.4095 for E.c. and T.th. respectively. We continue our exploration of NucleotideSphereNearPlaneCount through a simple Scatterplot for each of E.c. and T.th. We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean Nu-
Figure 4.5: Scatterplot of NucleotideSphereNearPlaneCount vs. log mutation rates for *E. c.* and *T. th.*

count is 9.5100 with standard deviation 4.8980; while for *T. th.* the mean NucleotideSphereNearPlaneCount is 10.060 with standard deviation of 4.9890. The median for *E. c.* and *T. th.* is 9 and 1 with interquartile range of 7 and 7 respectively. Lastly the spread of *E. c.* is 28 with minimum value of 0 and maximum value of 28; while the spread of *T. th.* is 35 with minimum value of 0 and maximum value of 35. We conclude this exploration with a histogram to give the distribution of NucleotideSphereNearPlaneCount for each of *E. c.* and *T. th.*

4.4.4.3 Exploring the NucleotideSphereNotNearPlaneCount variable

By constructing the NucleotideSphereNearPlaneCount variable by the cutoff angle, as a byproduct we have also constructed the NucleotideSphereNotNearPlaneCount variable. We begin exploring the relationship between NucleotideSphereNotNearPlaneCount and the log of the mutation rates for each of *E. c.* and *T. th.* molecules through the Spearman correlation coefficient. This coefficient is -0.04332 and -0.08872 for *E. c.* and *T. th.* respectively. We continue our exploration of NucleotideSphereNotNearPlaneCount through a simple Scatterplot for each of *E. c.* and *T. th.*.

We continue our exploration of this explanatory variable by examining basic descriptive
Figure 4.6: Histogram of NucleotideSphereNearPlaneCount for E.c. and T.th.

Figure 4.7: Scatterplot of NucleotideSphereNotNearPlaneCount vs. log mutation rates for E.c. and T.th.
Figure 4.8: Histogram of NucleotideSphereNotNearPlaneCount for E.c. and T.th.

statistics for each of E.c. and T.th.. For E.c. the mean NucleotideSphereNotNearPlaneCount is 17.290 with standard deviation 4.0390; while for T.th. the mean NucleotideSphereNotNearPlaneCount is 17.300 with standard deviation of 4.1780. The median for E.c. and T.th. is 17 and 17 with interquartile range of 5 and 6 respectively. Lastly the spread of E.c. is 3 with minimum value of 0 and maximum value of 3; while the spread of T.th. is 33 with minimum value of 0 and maximum value of 33. We conclude this exploration with a histogram to give the distribution of NucleotideSphereNotNearPlaneCount for each of E.c. and T.th.

4.4.5 DifferentSubUnitRNASphere

4.4.5.1 Definition

DifferentSubUnitRNASphere can be thought of exactly as NucleotideSphere, using the same way of computing the number of nucleotides in the sphere except the nucleotides in this sphere for a given nucleotide must come from a different subunit structure of 5S or larger, not in the one that the nucleotide is a member to be counted. For example if the sphere is centered around a nucleotide from the 16S chain, then the number of nucleotides that are counted come from either the 5S chain or the 23S chain, but the nucleotides from 16S are
not counted in the chain.

4.4.5.2 Determining the cutoff

Similar to NucleotideSphere, it begs the question, how do we choose a radius, \( r \), for the sphere surrounding each nucleotide. To do this, we ran the same computational study as NucleotideSphere in which we chose a radius for the sphere and then computed the \( R^2 \) for modeling the mutation rate as a log linear model of only DifferentSubUnitRNA Sphere and the intercept term for each of \( E.c. \) and the \( T.th. \). The best radius from each of the datasets was a radius of 8 angstroms, and 20 angstroms for \( E.c. \) and \( T.th. \) respectively yielded in the highest \( R^2 \) value. Thus we set the radius to be 14 for both structures. This of course is a large difference in the size of the sphere, yet we will find it is because this term is not highly predictive

4.4.5.3 Exploring DifferentSubUnitRNA Sphere

We begin exploring the relationship between DifferentSubUnitRNA Sphere Count and the log of the mutation rates for each of \( E.c. \) and \( T.th. \) molecules through the Spearman correlation coefficient. This coefficient is -0.03562 and -0.03685 for \( E.c. \) and \( T.th. \) respectively. We continue our exploration of DifferentSubUnitRNA Sphere Count through a simple Boxplot for each of \( E.c. \) and \( T.th. \).

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of \( E.c. \) and \( T.th. \). For \( E.c. \) the mean DifferentSubUnitRNA Sphere Count is 0.01464 with standard deviation 0.2030; while for \( T.th. \) the mean DifferentSubUnitRNA Sphere Count is 0.01919 with standard deviation of 0.2262. The median for \( E.c. \) and \( T.th. \) is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of \( E.c. \) is 6 with minimum value of 0 and maximum value of 6; while the spread of \( T.th. \) is 5 with minimum value of 0 and maximum value of 5. We conclude this exploration with a histogram to give the distribution of DifferentSubUnitRNA Sphere Count for each of \( E.c. \) and \( T.th. \).
Figure 4.9: Boxplot of DifferentSubUnitRNASphereCount vs. log mutation rates for *E.c.* and *T.th.*

Figure 4.10: Histogram of DifferentSubUnitRNASphereCount for *E.c.* and *T.th.*
4.4.6 **DifferentSubUnitRNANearPlaneCount**

We further classified the nucleotides in the DifferentSubUnitRNASphere by if they were in a similar plane to the nucleotide or if they were not in a plane similar to the nucleotide. By doing this we will answer the question if local density of nucleotides that are in a plane similar to itself or nucleotides that are not in a plane similar is the physical attribute that effects nucleotide mutation. To begin we must define what constitutes a nucleotide to be considered in a plane similar to the nucleotide. In addition to classifying the interactions of the 3D structure of a molecule, FR3D also gives a rotation matrix that would rotate the nucleotide from its current orientation in the 3D structure to a standard orientation. To determine if two nucleotides, say nucleotide $i$, and nucleotide $j$, are in a similar plane, we consider the following: First we compute the difference of the two centers to obtain a vector, $p$, that points from nucleotide $i$ to nucleotide $j$. We then apply the rotation matrix that orientates nucleotide $i$ from its current position to the standard orientation to the vector $p$, giving us the vector $r$. For the sake of calculation we define $u$ to be the normalized vector $r$. That is $u = \frac{r}{||r||}$. Next we turn our attention to the vertical component of $u$ by taking the dot product with the normal vector $<0,0,1>$. Looking at the triangle formed, we see that $\cos(\theta) = \frac{u \cdot n}{||u||}$, or rather without loss of generality $\cos(\theta) = |\frac{u \cdot n}{||u||}|$ so that $\theta$ is in the interval $[0,90]$. Since cosine is a decreasing function of $\theta$ over this interval, if $\theta$ is greater than a predetermined cutoff angle of depression from the normal vector, then we conclude that nucleotide $i$ and nucleotide $j$ are in a similar plane, and not in a similar plane otherwise.

4.4.6.1 **Determining the cutoff**

Similar to DifferentSubUnitRNASphere, it begs the question, how do we choose the angle of depression, $r$, for the sphere surrounding each nucleotide. To do this, we ran the same computational study as DifferentSubUnitRNASphere in which we chose the angle $\theta$ for determining how close to the plane nucleotide $j$ has to be to the plane of nucleotide $i$ to be considered to be in a similar plane, while keeping the radius fixed at the value stated above.
We then computed the $R^2$ for modeling the mutation rate as a log linear model of only DifferentSubUnitRNANearPlaneCount, and the intercept term for each of E.c. and the T.th. The best angle for this term was 81 degrees for E.c. and 87 degree for T.th.

### 4.4.6.2 Exploring DifferentSubUnitRNANearPlaneCount

We begin exploring the relationship between DifferentSubUnitRNANearPlaneCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.0398 and -0.03828 for E.c. and T.th. respectively. We continue our exploration of DifferentSubUnitRNANearPlaneCount through a simple Boxplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean DifferentSubUnitRNANearPlaneCount is 0.003282 with standard deviation 0.0572; while for T.th. the mean DifferentSubUnitRNANearPlaneCount is 0.004292 with standard deviation of 0.06913. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 1 with minimum value of 0 and maximum value of 1; while the spread of T.th. is 2 with minimum value of 0 and maximum value of 2. We conclude this exploration
4.4.6.3 Exploring the DifferentSubUnitRNANotNearPlaneCount variable

We begin exploring the relationship between DifferentSubUnitRNANotNearPlaneCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.03606 and -0.03126 for E.c. and T.th. respectively. We continue our exploration of DifferentSubUnitRNANotNearPlaneCount through a simple Boxplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean DifferentSubUnitRNANotNearPlaneCount is 0.01136 with standard deviation 0.1655; while for T.th. the mean DifferentSubUnitRNANotNearPlaneCount is 0.0149 with standard deviation of 0.1756. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 5 with minimum value of 0 and maximum value of 5; while the spread of T.th. is 4 with minimum value of 0 and maximum value of 4. We conclude this exploration with a histogram to give the distribution of DifferentSubUnitRNANotNearPlaneCount for E.c. and T.th.
Figure 4.13: Boxplot of DifferentSubUnitRNA\textsubscript{NotNearPlaneCount} vs. log mutation rates for \textit{E.c.} and \textit{T.th.} each of \textit{E.c.} and \textit{T.th.}.

### 4.4.7 SmallerRNA\textsubscript{Sphere}

#### 4.4.7.1 Definition

SmallerRNA\textsubscript{Sphere} can be thought of exactly as Nucleotide\textsubscript{Sphere}, using the same way of computing the number of nucleotides in the sphere except the nucleotides in this sphere for a given nucleotide must come from a different subunit structure of tRNA or mRNA. For example if the sphere is centered around a nucleotide from the 16S chain, then the number of nucleotides that are counted come from either a tRNA or mRNA chain, but the nucleotides from 16S, as well as 5S and 23S, are not counted in the chain.

#### 4.4.7.2 Determining the cutoff

Similar to Nucleotide\textsubscript{Sphere}, it begs the question, how do we choose a radius, \( r \), for the sphere surrounding each nucleotide. To do this, we ran the same computational study as Nucleotide\textsubscript{Sphere} in which we chose a radius for the sphere and then computed the \( R^2 \) for modeling the mutation rate as a log linear model of only SmallerRNA\textsubscript{Sphere} and the
Figure 4.14: Histogram of DifferentSubUnitRNANotNearPlaneCount for E.c. and T.th.

intercept term for T.th. For this term, we are limited by the x-ray crystallography of E.c. For 3D structures that fit our criteria, we do not have a crystallized form of E.c. that contains tRNA or mRNA. Therefore we find the best radius for the T.th. molecule, which was a radius of 30 angstroms, and then import the values of this term for T.th. to E.c. We realize that this is not an optimal way to test this term, but given our limitations on the data, it is a decent approximation.

4.4.7.3 Exploring SmallerRNASphere

Since the values of SmallerRNASphere are the same between E.c. and T.th., we need only explore for either E.c. or T.th. We begin exploring the relationship between SmallerRNASphere and log of the mutation rates through the Pearson correlation coefficient which is -0.1767. We continue our exploration of DifferentSubUnitRNASphere as it relates to the log mutation rates though a simple scatterplot for the molecules.

We begin exploring the relationship between SmallRNASphereCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.2829 and -0.2829 for E.c. and T.th. respectively. We continue our exploration of SmallRNASphereCount through a simple Scatterplot for each of E.c. and
We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th. For E.c. the mean SmallRNASphereCount is 3.3570 with standard deviation 8.2170; while for T.th. the mean SmallRNASphereCount is 3.3570 with standard deviation of 8.2170. The median for E.c. and T.th. is 0 and 0 with interquartile range of 1 and 1 respectively. Lastly the spread of E.c. is 7 with minimum value of 0 and maximum value of 7; while the spread of T.th. is 7 with with minimum value of 0 and maximum value of 7. We conclude this exploration with a histogram to give the distribution of SmallRNASphereCount for each of E.c. and T.th.

4.4.8 Near or not near the plane of the nucleotide

We further classified the nucleotides in the SmallerRNASphere by if they were in a similar plane to the nucleotide or if they were not in a plane similar to the nucleotide. By doing this we will answer the question if local density of nucleotides that are in a plane similar to itself
or nucleotides that are not in a plane similar is the physical attribute that effects nucleotide mutation. To begin we must define what constitutes a nucleotide to be considered in a plane similar to the nucleotide. In addition to classifying the interactions of the 3D structure of a molecule, FR3D also gives a rotation matrix that would rotate the nucleotide from its current orientation in the 3D structure to a standard orientation. To determine if two nucleotides, say nucleotide \( i \), and nucleotide \( j \), are in a similar plane, we consider the following: First we compute the difference of the two centers to obtain a vector, \( p \), that points from nucleotide \( i \) to nucleotide \( j \). We then apply the rotation matrix that orientates nucleotide \( i \) from its current position to the standard orientation to the vector \( p \), giving us the vector \( r \). For the sake of calculation we define \( u \) to be the normalized vector \( r \). That is \( u = \frac{r}{||r||} \). Next we turn our attention to the vertical component of \( u \) by taking the dot product with the normal vector \( <0, 0, 1> \). Looking at the triangle formed, we see that \( \cos(\theta) = \frac{n \cdot u}{||n||} \), or rather without loss of generality \( \cos(\theta) = \frac{n \cdot u}{||u||} \) so that \( \theta \) is in the interval \([0,90]\). Since cosine is a decreasing function of \( \theta \) over this interval, if \( \theta \) is greater than a predetermined cutoff angle of depression from the normal vector, then we conclude that nucleotide \( i \) and nucleotide \( j \) are in a similar plane, and not in a similar plane otherwise.

Figure 4.16: Histogram of SmallRNASphereCount for E.c. and T.th.
4.4.8.1 Determining the cutoff

Similar to SmallerRNASphere, it begs the question, how do we choose the angle of depression, \( r \), for the sphere surrounding each nucleotide. To do this, we ran the same computational study as SmallerRNASphere in which we chose the angle \( \theta \) for determining how close to the plane nucleotide \( j \) has to be to the plane of nucleotide \( i \) to be considered to be in a similar plane. We then computed the \( R^2 \) for modeling the mutation rate as a log linear model of only SmallerRNANearPlaneCount, and the intercept term for each of E.c. and the T.th. The best angle for this term was 16 degrees for T.th., and thus for E.c.

4.4.8.2 Exploring SmallerRNANearPlaneCount

We begin exploring the relationship between SmallerRNANearPlaneCount and log of the mutation rates for each of the E.c. and the T.th. molecules through the Pearson correlation coefficient which is -0.3082, and -0.2761 respectively. We continue our exploration of SmallerRNANearPlaneCount as it relates to the log mutation rates though a simple scatterplot for each of the E.c. and the T.th. molecules.

We begin exploring the relationship between SmallRNANearPlaneCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.2815 and -0.2815 for E.c. and T.th. respectively. We continue our exploration of SmallRNANearPlaneCount through a simple Scatterplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th. For E.c. the mean SmallRNANearPlaneCount is 3.2710 with standard deviation 8.0160; while for T.th. the mean SmallRNANearPlaneCount is 3.2710 with standard deviation of 8.0160. The median for E.c. and T.th. is 0 and 0 with interquartile range of 1 and 1 respectively. Lastly the spread of E.c. is 68 with minimum
value of 0 and maximum value of 68; while the spread of $T.th.$ is 68 with minimum value of 0 and maximum value of 68. We conclude this exploration with a histogram to give the distribution of SmallRNANearPlaneCount for each of $E.c.$ and $T.th.$

4.4.8.3 Exploring SmallerRNANotNearPlaneCount

We begin exploring the relationship between SmallRNANotNearPlaneCount and the log of the mutation rates for each of $E.c.$ and $T.th.$ molecules through the Spearman correlation
Figure 4.19: Boxplot of SmallRNANotNearPlaneCount vs. log mutation rates for E.c. and T.th. coefficient. This coefficient is -0.120 and -0.120 for E.c. and T.th. respectively. We continue our exploration of SmallRNANotNearPlaneCount through a simple Boxplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th. For E.c. the mean SmallRNANotNearPlaneCount is 0.08558 with standard deviation 0.5370; while for T.th. the mean SmallRNANotNearPlaneCount is 0.08558 with standard deviation of 0.5370. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 8 with minimum value of 0 and maximum value of 8; while the spread of T.th. is 8 with with minimum value of 0 and maximum value of 8. We conclude this exploration with a histogram to give the distribution of SmallRNANotNearPlaneCount for each of E.c. and T.th.
Figure 4.20: Histogram of SmallRNANotNearPlaneCount for *E.c.* and *T.th.*

### 4.4.9 AminoAcidSphereCount

#### 4.4.9.1 Definition

AminoAcidSphereCount can be thought of exactly as NucleotideSphere, using the same way of computing, but here we count the number of amino acids from nearby ribosomal proteins in the sphere for a given nucleotide instead of the number the nucleotides in this sphere. Nucleotides are not counted at all for this term.

#### 4.4.9.2 Determining the cutoff

Similar to NucleotideSphere, it begs the question, how do we choose a radius, $r$, for the sphere surrounding each nucleotide. To do this, we ran the same computational study as NucleotideSphereCount in which we chose a radius for the sphere and then computed the $R^2$ for modeling the mutation rate as a log linear model of only AminoAcidSphereCount and the intercept term for each of *E.c.* and the *T.th.* The best radius from each of the datasets was a radius of 17 angstroms, and 19 angstroms for *E.c.* and *T.th.* respectively yielded in the highest $R^2$ value. Thus we set the radius to be 18 for both structures. This is a small difference in the size of the sphere, and we will find this term predictive of mutation for both molecules.
4.4.9.3 Exploring the AminoAcidSphereCount variable

We begin exploring the relationship between AminoAcidSphereCount and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is -0.1953 and -0.1926 for *E.c.* and *T.th.* respectively. We continue our exploration of AminoAcidSphereCount through a simple Scatterplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean AminoAcidSphereCount is 8.5630 with standard deviation 9.2120; while for *T.th.* the mean AminoAcidSphereCount is 8.9090 with standard deviation of 9.3610. The median for *E.c.* and *T.th.* is 6 and 6 with interquartile range of 14 and 14 respectively. Lastly the spread of *E.c.* is 54 with minimum value of 0 and maximum value of 54; while the spread of *T.th.* is 52 with minimum value of 0 and maximum value of 52. We conclude this exploration with a histogram to give the distribution of AminoAcidSphereCount for each of *E.c.* and *T.th.*.
4.4.10 Near or not near the plane of the nucleotide

We further classified the nucleotides in the AminoAcidSphere by if they were in a similar plane to the nucleotide or if they were not in a plane similar to the nucleotide. By doing this we will answer the question if local density of nucleotides that are in a plane similar to itself or nucleotides that are not in a plane similar is the physical attribute that effects nucleotide mutation. To begin we must define what constitutes a nucleotide to be considered in a plane similar to the nucleotide. In addition to classifying the interactions of the 3D structure of a molecule, FR3D also gives a rotation matrix that would rotate the nucleotide from its current orientation in the 3D structure to a standard orientation. To determine if two nucleotides, say nucleotide $i$, and nucleotide $j$, are in a similar plane, we consider the following: First we compute the difference of the two centers to obtain a vector, $p$, that points from nucleotide $i$ to nucleotide $j$. We then apply the rotation matrix that orientates nucleotide $i$ from its current position to the standard orientation to the vector $p$, giving us the vector $r$. For the sake of calculation we define $u$ to be the normalized vector $r$. That is $u = \frac{r}{||r||}$. Next we turn our attention to the vertical component of $u$ by taking the dot product with the normal vector $<0,0,1>$. Looking at the triangle formed, we see that $\cos(\theta) = \frac{u \cdot n}{||u||}$, or rather without loss of generality $\cos(\theta) = \frac{u \cdot n}{||u||}$, so that $\theta$ is in the interval $[0,90]$. Since cosine is a
decreasing function of $\theta$ over this interval, if $\theta$ is greater than a predetermined cutoff angle of depression from the normal vector, then we conclude that nucleotide $i$ and nucleotide $j$ are in a similar plane, and not in a similar plane otherwise.

4.4.10.1 Determining the cutoff

Similar to AminoAcidSphere, it begs the question, how do we choose the angle of depression, $r$, for the sphere surrounding each nucleotide. To do this, we ran the same computational study as AminoAcidSphere in which we chose the angle $\theta$ for determining how close to the plane nucleotide $j$ has to be to the plane of nucleotide $i$ to be considered to be in a similar plane. We then computed the $R^2$ for modeling the mutation rate as a log linear model of only AminoAcidNearPlaneCount, and the intercept term for each of E.c. and the T.th. The best angle for this term was 57 degrees for E.c. and 71 degrees for T.th.

4.4.10.2 Exploring the AminoAcidNearPlaneCount variable

We begin exploring the relationship between AminoAcidNearPlaneCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.06271 and -0.07904 for E.c. and T.th. respectively. We continue our exploration of AminoAcidNearPlaneCount through a simple Scatterplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean AminoAcidNearPlaneCount is 4.3580 with standard deviation 5.2110; while for T.th. the mean AminoAcidNearPlaneCount is 4.5380 with standard deviation of 5.3630. The median for E.c. and T.th. is 2 and 3 with interquartile range of 7 and 8 respectively. Lastly the spread of E.c. is 31 with minimum value of 0 and maximum value of 31; while the spread of T.th. is 31 with minimum value of 0 and maximum value of 31. We conclude this exploration with a histogram to give the distribution of AminoAcidNearPlaneCount for each of E.c. and T.th..
Figure 4.23: Scatterplot of AminoAcidNearPlaneCount vs. log mutation rates for *E.c.* and *T.th.*

Figure 4.24: Histogram of AminoAcidNearPlaneCount for *E.c.* and *T.th.*
4.4.10.3 Exploring the AminoAcidNotNearPlaneCount variable

We begin exploring the relationship between AminoAcidNotNearPlaneCount and the log of the mutation rates for each of \textit{E.c.} and \textit{T.th.} molecules through the Spearman correlation coefficient. This coefficient is -0.09444 and -0.1036 for \textit{E.c.} and \textit{T.th.} respectively. We continue our exploration of AminoAcidNotNearPlaneCount through a simple Scatterplot for each of \textit{E.c.} and \textit{T.th.}.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of \textit{E.c.} and \textit{T.th.}. For \textit{E.c.} the mean AminoAcidNotNearPlaneCount is 4.0420 with standard deviation 5.1160; while for \textit{T.th.} the mean AminoAcidNotNearPlaneCount is 4.1090 with standard deviation of 5.0460. The median for \textit{E.c.} and \textit{T.th.} is 2 and 2 with interquartile range of 6 and 6 respectively. Lastly the spread of \textit{E.c.} is 32 with minimum value of 0 and maximum value of 32; while the spread of \textit{T.th.} is 32 with minimum value of 0 and maximum value of 32. We conclude this exploration with a histogram to give the distribution of AminoAcidNotNearPlaneCount for each of \textit{E.c.} and \textit{T.th.}. 

Figure 4.25: Scatterplot of AminoAcidNotNearPlaneCount vs. log mutation rates for \textit{E.c.} and \textit{T.th.}.
4.4.11 Nucleotides potentially interacting with an amino acid

4.4.11.1 Definition

This term is a carryover from the work of [53], and a special case of AminoAcidSphere. We define a nucleotide as potentially interacting with an amino acid if the amino acid is within 3.5 angstroms of the nucleotide. This term is an indicator taking the value one if the amino acid is within 3.5 angstroms of the nucleotide and zero otherwise. Therefore we compute this variable by constructing a sphere of radius 3.5 around the nucleotide, but here we count the number of amino acids that are in the sphere potentially interacting with the nucleotide. We call this variable NTAminoInteractionCount.

4.4.11.2 Exploring the NTAminoInteractionCount variable

We begin exploring the relationship between NTAminoInteractionCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.04988 and -0.05117 for E.c. and T.th. respectively. We continue our exploration of NTAminoInteractionCount through a simple Boxplot for each of E.c. and T.th.
Figure 4.27: Boxplot of NTAminoInteractionCount vs. log mutation rates for E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean NTAminoInteractionCount is 0.01338 with standard deviation 0.1424; while for T.th. the mean NTAminoInteractionCount is 0.0154 with standard deviation of 0.1457. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 3 with minimum value of 0 and maximum value of 3; while the spread of T.th. is 2 with minimum value of 0 and maximum value of 2. We conclude this exploration with a histogram to give the distribution of NTAminoInteractionCount for each of E.c. and T.th..

4.5 Base stacking

4.5.1 Classification of Base Stacking

FR3D, [REF 2008 FR3D paper], classifies base stacking interactions into three broad categories according to the faces of the bases that are in contact. Each base has two distinct faces. FR3D classifies the bases according to the 3’ and 5’ sides of the nucleotide. In an RNA helix, the face that faces the 3’ end of the chain is called the 3’ face, and the other face is called the 5’ face. In a helix, successive bases on one strand stack with the 3’ face of one
base interacting with the 5’ face of the next base in the chain. This is called s35 stacking, or s53 stacking based upon which way the interaction is being examined. The other types of stacking, s33, and s55 also occur in RNA. Cross-strand stacking in a helix occurs with the 5’ faces in contact; while s33 stacking occurs primarily in RNA motifs.

4.5.2 The counting approach

We will look at accounting for stacking interactions in two ways. The first that we consider we called the counting approach. In this approach, we look simply at the number of stacking interactions for a particular nucleotide.

4.5.2.1 Exploring the StackingCount variable

We begin exploring the relationship between StackingCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is 0.1214 and 0.1732 for E.c. and T.th. respectively. We continue our exploration of StackingCount through a simple Boxplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th. For E.c. the mean StackingCount is 2.2150 with standard
deviation 0.7881; while for T.th. the mean StackingCount is 2.2250 with standard deviation of 0.8031. The median for E.c. and T.th. is 2 and 2 with interquartile range of 1 and 1 respectively. Lastly the spread of E.c. is 4 with minimum value of 0 and maximum value of 4; while the spread of T.th. is 4 with minimum value of 0 and maximum value of 4. We conclude this exploration with a histogram to give the distribution of StackingCount for each of E.c. and T.th..

4.5.2.2 Exploring the NumberNearStackingPairs

FR3D can classify those basepairs that are close, but do not quite satisfy the criterion to be considered as a stacking pair. This variable, NumberNearStackingPairs, can be thought of just like the number of stacking pairs as above for nucleotides that have partner nucleotides that are close to satisfying the criterion to be considered as a stacking pair. We begin exploring the relationship between NearStackingCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.09111 and -0.1093 for E.c. and T.th. respectively. We continue our exploration of NearStackingCount through a simple Boxplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive
Figure 4.30: Histogram of StackingCount for *E.c.* and *T.th.*

Figure 4.31: Boxplot of NearStackingCount vs. log mutation rates for *E.c.* and *T.th.*
statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean NearStackingCount is 0.2340 with standard deviation 0.4851; while for *T.th.* the mean NearStackingCount is 0.2663 with standard deviation of 0.5110. The median for *E.c.* and *T.th.* is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of *E.c.* is 5 with minimum value of 0 and maximum value of 5; while the spread of *T.th.* is 3 with minimum value of 0 and maximum value of 3. We conclude this exploration with a histogram to give the distribution of NearStackingCount for each of *E.c.* and *T.th.*

### 4.5.3 Stacking Discrepancy Index

Similarly to other variables, simply counting the number of stacking partners does not tell much about what biological reason this term has on mutation rates. Therefore our second way that we consider stacking interactions is through a look at the geometry of the stacking partners. It is from this geometric relationship that we define a more biologically relevant predictor that we call the Stacking Discrepancy Index or SDI for short. The geometry of the stacking in an RNA helix is very common, and all pairs of bases stack on one another in RNA helices quite nicely and with the same geometry. But some stackings are more unusual, and are only made by certain bases. We don’t know why. Some base stacking geometries
Table 4.1: MinIDI matrix for s35 stacking partners A907(rows) A908(columns) in 1S72

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.000</td>
<td>0.1928</td>
<td>0.1462</td>
<td>0.2485</td>
</tr>
<tr>
<td>C</td>
<td>0.2182</td>
<td>0.1835</td>
<td>0.2049</td>
<td>0.2570</td>
</tr>
<tr>
<td>G</td>
<td>0.2708</td>
<td>0.1885</td>
<td>0.2091</td>
<td>0.1186</td>
</tr>
<tr>
<td>U</td>
<td>0.2468</td>
<td>0.2416</td>
<td>0.1430</td>
<td>0.3614</td>
</tr>
</tbody>
</table>

only occur in certain RNA motifs, and the nature of these motifs may be that only certain base combinations can appear there, in order to make the basepairs required for the motif.

4.5.3.1 Defining SDI

To understand SDI in depth, consider the following example. Suppose that we observe a certain stacking arrangement in an RNA motif, and we would like to know what base substitutions are likely for that particular nucleotide and its stacking partner. A way to estimate the substitution probabilities is to examine similar arrangements in other organisms. As with basepairing and base-backbone interactions, we imagine that we change one or both of the bases, without moving the backbone atoms, and consider the resulting stacking arrangement. We take the given base stacking, suppose it is an s35 stacking. Find all other s35 stacking pairs from other organisms, and rank them according to their IDI, geometric discrepancy between the observed stacking interaction and a candidate interaction, as defined in [53]. Now go through the list from lowest IDI to highest, until each base combination (AA, AC, AG, etc.) is found, and record the minimum IDI from the given base stacking at which each base combination first appears. This will give an indication whether the indicated base stacking can occur with each different base combination. We look an example of how to transform IDI into an estimate of the substitution probabilities. First we use FR3D to calculate the IDI matrix for a A907 and its stacking partner A908 from 1S72.

We then turn the geometric discrepancies recorded by the MinIDI matrix into a matrix of estimated substitution probabilities through the process of [53] We obtain the following matrix:
Table 4.2: IDI based estimated substitution matrix for s35 stacking between A907(rows) A908(columns) in 1S72

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0654</td>
<td>0.0630</td>
<td>0.0640</td>
<td>0.0616</td>
</tr>
<tr>
<td>C</td>
<td>0.0624</td>
<td>0.0633</td>
<td>0.0627</td>
<td>0.0613</td>
</tr>
<tr>
<td>G</td>
<td>0.0606</td>
<td>0.0631</td>
<td>0.0626</td>
<td>0.0645</td>
</tr>
<tr>
<td>U</td>
<td>0.0616</td>
<td>0.0618</td>
<td>0.0641</td>
<td>0.0578</td>
</tr>
</tbody>
</table>

To translate the information from above into something about how stacking effects the rate of mutation, we the created the SDI term. This term was defined for each nucleotide represented by the first of the stacking nucleotides as the standard deviation of the sum of the four rows of the matrix substitution matrix for each nucleotide. If a nucleotide does not have a stacking partner, its SDI is zero, otherwise for each stacking partner, we compute the standard deviation of the IDI based estimated substitution matrix. We do this in order to measure how easily the nucleotide that we are studying can be substituted while still preserving the stacking interaction. If the nucleotide is easily substituted then we would expect that the probability that any of the four nucleotides is used as the first nucleotide is roughly 0.25. This would result in an SDI contribution of nearly zero. Conversely, if the nucleotide does not easily substitute then to a different nucleotide then we would expect that one row sum will be near one, while the others would be near zero. This will result in an isostericity score near $\frac{\sqrt{3}}{2}$.

### 4.5.3.2 Exploring the SDI variable

We begin exploring the relationship between SDI and the log of the mutation rates for each of *E. c.* and *T. th.* molecules through the Spearman correlation coefficient. This coefficient is -0.1355 and -0.08429 for *E. c.* and *T. th.* respectively. We continue our exploration of SDI through a simple Scatterplot for each of *E. c.* and *T. th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E. c.* and *T. th.*. For *E. c.* the mean SDI is 0.07781 with standard deviation
0.07962; while for T.th. the mean SDI is 0.0787 with standard deviation of 0.07689. The median for E.c. and T.th. is 0.06909 and 0.06992 with interquartile range of 0.07702 and 0.07238 respectively. Lastly the spread of E.c. is 0.6509 with minimum value of 0 and maximum value of 0.6509; while the spread of T.th. is 0.6163 with minimum value of 0 and maximum value of 0.6163. We conclude this exploration with a histogram to give the distribution of SDI for each of E.c. and T.th..

### 4.6 Base-phosphate interactions

#### 4.6.1 Introduction

In this section we examine the effects that two types of interactions that basepairs are a part and their influence on nucleotide mutation. These interactions we have named as base and phosphate interactions. This base phosphate interaction was first explored deeply by [53]. In their work, they investigated the effects of this interaction upon the conservation percentage of nucleotides in the multiple sequence alignment that we use here. Here, we define the components base and phosphate which compose the base phosphate interaction and give some justification as to why we believe them to be relevant biological terms in
our modeling process. Like previous sections, we also do some basic exploration into the relationship with these variables and site specific mutation rates.

Base-phosphate interactions are hydrogen bonds between a hydrogen on a base and an oxygen attached to the phosphorus on the backbone. As was discussed by Zirbel et al (2009) since position of the hydrogen bonds is specific to each nucleotide. If a nucleotide is using its base in a base phosphate interaction, and if the base-phosphate interaction is to be preserved after a mutation, the nucleotide that replaces the original nucleotide must make able hydrogen bonds such that the interaction is preserved.

### 4.6.2 Counting approach

Like StackingCount, FR3D has the ability to, [REF 2008 FR3D paper], classify base phosphate interactions. We start with simply counting the number of base-phosphate interactions.

#### 4.6.2.1 Exploring the BPhBaseCount variable

We begin exploring the relationship between BPhBaseCount and the log of the mutation rates for each of \( E.c. \) and \( T.th. \) molecules through the Spearman correlation coefficient. This
We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of \textit{E.c.} and \textit{T.th.}. For \textit{E.c.} the mean BPhBaseCount is 0.1096 with standard deviation 0.3282; while for \textit{T.th.} the mean BPhBaseCount is 0.1182 with standard deviation of 0.3440. The median for \textit{E.c.} and \textit{T.th.} is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of \textit{E.c.} is 3 with minimum value of 0 and maximum value of 3; while the spread of \textit{T.th.} is 3 with minimum value of 0 and maximum value of 3. We conclude this exploration with a histogram to give the distribution of BPhBaseCount for each of \textit{E.c.} and \textit{T.th.}.

4.6.3 Geometric approach

4.6.3.1 Definition

We now consider a refinement to the base variable. From the figures above, we believe that there may be an effect on the site specific mutation rates due a nucleotide making a base-phosphate interaction. We are left with the question, what is it about the number of
base-phosphate interactions that could them harder to substitute to a different nucleotide as compared to a nucleotide that is not making such an interaction. While there appears to be an association between the simple count of the number of times the nucleotide uses its base in a base phosphate interaction and the site specific mutation rate, a simple count does not have as much biological relevance as we would like. As a refinement to the way that we think about base-phosphate interactions, we would instead like to consider how hard it would be to replace the nucleotide donating it’s base to the base phosphate interactions relative to the geometry observed in the 3D structure. The figure below is a geometric representation of where each base can form a base-phosphate interaction, where each base is in a standard orientation.

In the diagram depicting the places that a nucleotide can make a base-phosphate inter-
action, the markings 0BPh - 9BPh represent the positioning of where the nucleotide may make a hydrogen bond with the phosphate of another nucleotide; where 0BPh is a self base phosphate interaction and will be ignored for our consideration because this interaction is being made by almost all nucleotides, and there is no reason to think that it would be relevant from a biological point of view because the positioning does not greatly differ between nucleotides, and there has been no evidence to support its relevance from a statistical point of view. However, we believe 1BPh - 9BPh to be of relevance because the more times that a nucleotide uses its base for base phosphate interactions, we might expect that will constrain the nucleotide against the mutation of the nucleotide using its base.

From the diagram it can be seen that for a G making a 1BPH to change to an A making a 2BPH much more easily than it can change into a C or a U without substantially altering the orientation of the nucleotide. In this section it is our goal to capture the information about the geometry of the variable ultimately relate this information to the mutation rates.

4.6.3.2 computation of the variable

To capture the geometric information, we take which type of base-phosphate interaction that we observe in the 3D structure and compute the distance, in angstroms, to the nearest BPH interaction in the standard position that could replace the base-phosphate interaction observed in the 3D structure. We take this value to be the representation of the geometric information of the 3D structure of the base-phosphate interaction. In the case that a nucleotide is making more than one base-phosphate interaction, then the value of this variable is simply the sum of the distances computed as described above. In the final model we will call this variable BPhGeometry.

4.6.3.3 example

To give an example of how this term is computed. We consider an A making a 2BPh interaction. The distance for which a different nucleotide would have to compensate is given
<table>
<thead>
<tr>
<th>Distance</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>1.0296</td>
<td>2.4534</td>
<td>2.7110</td>
</tr>
</tbody>
</table>

Table 4.3: Distances of other nucleotides substituting for the A making a 2BPh

From the table, the contribution to the BPhGeometry is 1.0296

4.6.3.4 Exploring the BPhGeometry variable

We begin exploring the relationship between BPhGeometry and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.163 and -0.1537 for E.c. and T.th. respectively. We continue our exploration of BPhGeometry through a simple Scatterplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean BPhGeometry is 0.1171 with standard deviation 0.3848; while for T.th. the mean BPhGeometry is 0.1269 with standard deviation of 0.4030. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 3.1430 with minimum value of 0 and maximum value of 3.1430; while the spread of T.th. is 3.1430 with minimum value of 0 and maximum
value of 3.1430. We conclude this exploration with a histogram to give the distribution of BPhGeometry for each of E.c. and T.th..

4.7 Near base-phosphate interactions

4.7.1 Introduction

Just as we define the near stacking and near basepair interactions, FR3D has the capability to classify near base-phosphate interactions. For interactions that are classified as near basepairing interactions, this variable is computed in the exact same way as is the Base variable computed for those nucleotides that are classified as making a true base-phosphate interaction.

4.7.1.1 Exploring the NearBPhGeometry variable

We begin exploring the relationship between NearBPhGeometry and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.09174 and -0.07903 for E.c. and T.th. respectively. We continue our exploration of NearBPhGeometry through a simple Scatterplot for each of E.c. and T.th..
We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean NearBPhGeometry is 0.06098 with standard deviation 0.2751; while for *T.th.* the mean NearBPhGeometry is 0.0641 with standard deviation of 0.2785. The median for *E.c.* and *T.th.* is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of *E.c.* is 3.1430 with minimum value of 0 and maximum value of 3.1430; while the spread of *T.th.* is 2.9070 with minimum value of 0 and maximum value of 2.9070. We conclude this exploration with a histogram to give the distribution of NearBPhGeometry for each of *E.c.* and *T.th.*.

### 4.8 Base-ribose interactions

#### 4.8.1 Definition

Base-phosphate interactions are hydrogen bonds between a hydrogen on a base and an oxygen attached to the phosphorus on the sugar ring. These interactions we have named as base and ribose interactions. Like base-phosphate interactions, the position of the hydrogen bonds is specific to each nucleotide. If a nucleotide is using its base in a base-ribose interaction, then if the base-ribose interaction is to be preserved after a mutation, the nucleotide that
Figure 4.41: Histogram of NearBPhGeometry for *E. c.* and *T. th.*

Figure 4.42: Example of a base-ribose interaction complete with sugars

replaces the original nucleotide must make able hydrogen bonds such that the interaction is preserved.

### 4.8.2 Exploring the BaseRiboseCount variable

We begin exploring the relationship between BaseRiboseCount and the log of the mutation rates for each of *E. c.* and *T. th.* molecules through the Spearman correlation coefficient. This coefficient is -0.1643 and -0.1502 for *E. c.* and *T. th.* respectively. We continue our exploration of BaseRiboseCount through a simple Boxplot for each of *E. c.* and *T. th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E. c.* and *T. th.*. For *E. c.* the mean BaseRiboseCount is 0.1502 with standard deviation 0.4013; while for *T. th.* the mean BaseRiboseCount is 0.1310 with standard deviation of 0.3784. The median for *E. c.* and *T. th.* is 0 and 0 with interquartile range of 0
and 0 respectively. Lastly the spread of *E.c.* is 3 with minimum value of 0 and maximum value of 3; while the spread of *T.th.* is 3 with minimum value of 0 and maximum value of 3.

We conclude this exploration with a histogram to give the distribution of BaseRiboseCount for each of *E.c.* and *T.th.*

### 4.8.3 Exploring the NearBaseRiboseCount variable

Similar to base stacking, FR3D can also classify near base-ribose interactions. We begin exploring the relationship between NearBaseRiboseCount and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is 0.05479 and -0.02515 for *E.c.* and *T.th.* respectively. We continue our exploration of NearBaseRiboseCount through a simple Boxplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean NearBaseRiboseCount is 0.5037 with standard deviation 0.5869; while for *T.th.* the mean NearBaseRiboseCount is 0.4395 with standard deviation of 0.5911. The median for *E.c.* and *T.th.* is 0 and 0 with interquartile range of 1 and 1 respectively. Lastly the spread of *E.c.* is 3 with minimum value of 0 and maximum value of 3; while the spread of *T.th.* is 4 with minimum value of 0 and
Figure 4.44: Histogram of BaseRiboseCount for *E. c.* and *T. th.*

Figure 4.45: Boxplot of NearBaseRiboseCount vs. log mutation rates for *E. c.* and *T. th.*
maximum value of 4. We conclude this exploration with a histogram to give the distribution of NearBaseRiboseCount for each of \textit{E.c.} and \textit{T.th.}.

### 4.8.4 Geometric approach to base-ribose interactions

We will now define the term BRGeometry which is the geometric consideration of base-ribose interactions. This consideration is exactly the same as the consideration of the BPhGeometry. This is because the nucleotide making a base-ribose interaction is using a hydrogen to interact with the ribose of another nucleotide. Exactly like the base-phosphate interaction the nucleotide can only use its hydrogen in the same positions as the base-phosphate interactions. Therefore we compute BRGeometry the same way as BPhGeometry.

#### 4.8.4.1 Exploring the BaseRiboseGeometry variable

We begin exploring the relationship between BaseRiboseGeometry and the log of the mutation rates for each of \textit{E.c.} and \textit{T.th.} molecules through the Spearman correlation coefficient. This coefficient is -0.1709 and -0.1532 for \textit{E.c.} and \textit{T.th.} respectively. We continue our exploration of BaseRiboseGeometry through a simple Scatterplot for each of \textit{E.c.} and \textit{T.th.}.
We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of \( E.c. \) and \( T.th. \). For \( E.c. \) the mean BaseRiboseGeometry is 0.09773 with standard deviation 0.3520; while for \( T.th. \) the mean BaseRiboseGeometry is 0.09157 with standard deviation of 0.3440. The median for \( E.c. \) and \( T.th. \) is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of \( E.c. \) is 4.2670 with minimum value of 0 and maximum value of 4.2670; while the spread of \( T.th. \) is 4.2670 with with minimum value of 0 and maximum value of 4.2670. We conclude this exploration with a histogram to give the distribution of BaseRiboseGeometry for each of \( E.c. \) and \( T.th. \).
4.8.4.2 Exploring the NearBaseRiboseGeometry variable

We begin exploring the relationship between NearBaseRiboseGeometry and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is 0.01353 and -0.05857 for E.c. and T.th. respectively. We continue our exploration of NearBaseRiboseGeometry through a simple Scatterplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th. For E.c. the mean NearBaseRiboseGeometry is 0.1477 with standard deviation 0.4353; while for T.th. the mean NearBaseRiboseGeometry is 0.1512 with standard deviation of 0.4358. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0.01648 and 0.0118 respectively. Lastly the spread of E.c. is 4.5030 with minimum value of 0 and maximum value of 4.5030; while the spread of T.th. is 4.5030 with with minimum value of 0 and maximum value of 4.5030. We conclude this exploration with a histogram to give the distribution of NearBaseRiboseGeometry for each of E.c. and T.th.
4.9 Introduction to basepairing interactions

For the remainder of the chapter, we will turn our attention discuss the influence that basepairing interactions of nucleotides have on the mutation rate. This section will define and then examine types of basepairing interaction. We place these basepairing interactions, as classified from FR3D, into two major classes which are cis Watson Crick basepairs and non-cis Watson Crick basepairs, cWW and non-cWW respectively for short. We will go on to briefly discuss isostericity and use the concept of isostericity to define a term called ICE for isosteric constraining effect which is a function of isostericity and the cWW and non-cWW basepair interactions that a nucleotide makes. Finally in this section, we will discuss how the non-basepairing and basepairing interactions that are communicated as constraints to a nucleotide.
4.10 cWW interactions

4.10.1 Definition

By far the most common nucleotide basepairing interactions are cWW basepairing interactions. Of the 3948 nucleotides in each of the conserved core datasets, there are 2573 nucleotides in the E.c. molecule that are making a cWW basepair and 2578 nucleotides in the T.th. molecule that are making a cWW basepair. It is known that cWW pairing is essential to the RNA molecule structure which certainly effects the functionality of the molecule. Each nucleotide A, C, G, and U has three sides it can use to create a nucleotide basepairing interaction. These edges are the Watson Crick, the Hoogsteen and the Sugar edges. A cWW interaction like the one depicted above is defined to be a nucleotide interaction of cis-acting type that occurs along the Watson Crick edge of both nucleotides. All combinations of basepairing nucleotides are possible with the exception of the GG interaction. Also, a nucleotide may have only one cWW interaction with one other nucleotide. In RNA, the most common cWW interactions are between the nucleotides CG and UA, (or GC AU), while UG (GU) interactions are also common but much less so than the other two types. An interaction of cWW type between a nucleotide G and a nucleotide C has three hydrogen bonds that link the two bases together, while the other common interaction of cWW type is between the nucleotide A and the nucleotide U has two hydrogen bonds to create the cWW interaction. While cWW interactions are by far the most common nucleotide basepairing interaction, they are not the only type of basepairing interactions that can occur. We define the variable cWWcount as the number of cWW interactions that a nucleotide is making.

4.10.2 Exploring the cWWCount variable

We begin exploring the relationship between cWWCount and the logarithm of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is 0.3168 and 0.3230 for E.c. and T.th. respectively. We continue our exploration
of cWWCount through a simple Boxplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean cWWCount is 0.6516 with standard deviation 0.4765; while for T.th. the mean cWWCount is 0.6544 with standard deviation of 0.4756. The median for E.c. and T.th. is 1 and 1 with interquartile range of 1 and 1 respectively. Lastly the spread of E.c. is 1 with minimum value of 0 and maximum value of 1; while the spread of T.th. is 1 with minimum value of 0 and maximum value of 1. We conclude this exploration with a histogram to give the distribution of cWWCount for each of E.c. and T.th..

### 4.11 non-cWW interactions

#### 4.11.1 Definition

Nucleotides that create interactions of a non-cWW type are defined simply as a basepairing interactions between nucleotides that is of a different type than a cWW interaction. This means that bases that use other edges to create a basepairing interaction such as one nucleotide uses its sugar while the other nucleotide uses its Hoogsteen edge to create the...
basepairing interaction or if one nucleotide uses its sugar edge while the other nucleotide uses its Watson Crick Edge. Also, if both nucleotides use their Watson Crick edge to make the basepair, but the interaction is through trans-acting nucleotides then this too is considered to be a non-cWW interaction.

4.11.2 Exploring the noncWWCount variable

We begin exploring the relationship between noncWWCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.3626 and -0.3513 for E.c. and T.th. respectively. We continue our exploration of noncWWCount through a simple Boxplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean noncWWCount is 0.3083 with standard deviation 0.5552; while for T.th. the mean noncWWCount is 0.2921 with standard deviation of 0.5557. The median for E.c. and T.th. is 0 and 0 with interquartile range of 1 and 0 respectively. Lastly the spread of E.c. is 3 with minimum value of 0 and maximum value of 3; while the spread of T.th. is 4 with minimum value of 0 and maximum value of 4. We conclude this exploration with a histogram to give the distribution of noncWWCount.
4.11.3 Exploring the BasepairCount variable

We could also combine the number of cWW pairs and noncWW pairs. We call this term BasepairCount. We begin exploring the relationship between BasepairCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.07816 and -0.04945 for E.c. and T.th. respectively. We continue our exploration of BasepairCount through a simple Boxplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean BasepairCount is 0.9599 with standard deviation 0.5243; while for T.th. the mean BasepairCount is 0.9465 with standard deviation of 0.5474. The median for E.c. and T.th. is 1 and 1 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 3 with minimum value of 0 and maximum value of 3; while the spread of T.th. is 4 with minimum value of 0 and maximum value of 4. We conclude this exploration with a histogram to give the distribution of BasepairCount for each of E.c. and T.th..
Figure 4.54: Histogram of noncWWCount for *E.c.* and *T.th.*

Figure 4.55: Boxplot of BasepairCount vs. log mutation rates for *E.c.* and *T.th.*
4.12 Near basepairing interactions

Like stacking interactions, FR3D is also able to classify near basepair interactions. We consider near cWW and noncWW type interactions.

4.12.1 Exploring the NearcWWCount variable

We begin exploring the relationship between NearcWWCount and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is -0.05717 and -0.02006 for *E.c.* and *T.th.* respectively. We continue our exploration of NearcWWCount through a simple Boxplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean NearcWWCount is 0.02398 with standard deviation 0.1547; while for *T.th.* the mean NearcWWCount is 0.05579 with standard deviation of 0.2382. The median for *E.c.* and *T.th.* is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of *E.c.* is 2 with minimum value of 0 and maximum value of 2; while the spread of *T.th.* is 2 with minimum value of 0 and maximum value of 2. We conclude this exploration with a histogram to give the distribution of NearcWWCount.
for each of E.c. and T.th..

4.12.2 Exploring the NearnoncWW variable

We begin exploring the relationship between NearnoncWWCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.2191 and -0.2215 for E.c. and T.th. respectively. We continue our exploration of NearnoncWWCount through a simple Boxplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean NearnoncWWCount is 0.3133 with standard deviation 0.6232; while for T.th. the mean NearnoncWWCount is 0.3527 with standard deviation of 0.6560. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 1 respectively. Lastly the spread of E.c. is 5 with minimum value of 0 and maximum value of 5; while the spread of T.th. is 5 with minimum value of 0 and maximum value of 5. We conclude this exploration with a histogram to give the distribution of NearnoncWWCount for each of E.c. and T.th..
Figure 4.58: Histogram of NearcWWCount for E.c. and T.th.

Figure 4.59: Boxplot of NearnoneWWCount vs. log mutation rates for E.c. and T.th.
4.12.3 Exploring the NearBasepairCount variable

Similar to true basepair count, we can consider near basepairs. We begin exploring the relationship between NearBasepairCount and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is -0.221 and -0.2182 for *E.c.* and *T.th.* respectively. We continue our exploration of NearBasepairCount through a simple Boxplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean NearBasepairCount is 0.3373 with standard deviation 0.6469; while for *T.th.* the mean NearBasepairCount is 0.4085 with standard deviation of 0.6918. The median for *E.c.* and *T.th.* is 0 and 0 with interquartile range of 1 and 1 respectively. Lastly the spread of *E.c.* is 5 with minimum value of 0 and maximum value of 5; while the spread of *T.th.* is 5 with minimum value of 0 and maximum value of 5.

We conclude this exploration with a histogram to give the distribution of NearBasepairCount for each of *E.c.* and *T.th.*.
Figure 4.61: Boxplot of NearBasepairCount vs. log mutation rates for E.c. and T.th.

Figure 4.62: Histogram of NearBasepairCount for E.c. and T.th.
4.13 NumberInterfacingNucleotides

4.13.1 Definition

This term is a very special case of the counts of the number of basepairing interactions that a nucleotide makes. We do not discriminate as to whether the basepairing interaction type is a cWW basepair or a noncWW basepair, but rather we count an interaction that a nucleotide makes if the interaction is with a nucleotide from another structure.

4.13.2 Exploring the InterfacingNucleotidesCount variable

We begin exploring the relationship between InterfacingNucleotidesCount and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is -0.03514 and -0.04933 for *E.c.* and *T.th.* respectively. We continue our exploration of InterfacingNucleotidesCount through a simple Boxplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean InterfacingNucleotidesCount is
0.001515 with standard deviation 0.04492; while for T.th. the mean InterfacingNucleotidesCount is 0.00303 with standard deviation of 0.06349. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 2 with minimum value of 0 and maximum value of 2; while the spread of T.th. is 2 with minimum value of 0 and maximum value of 2. We conclude this exploration with a histogram to give the distribution of InterfacingNucleotidesCount for each of E.c. and T.th..

4.14 Number of long range interactions

4.14.1 Definition

We define an interaction of any type, such as basepair, base-phosphate, or stacking for example, to be a long range interaction if it crosses over a region of RNA that contains at least three nested cWW basepairs. Exploring the NumberLongRangeInteractionCount variable We begin exploring the relationship between NumberLongRangeInteractionCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.2507 and -0.2607 for E.c. and T.th. respectively. We continue our exploration of NumberLongRangeInteractionCount through a simple Boxplot
Figure 4.65: Boxplot of NumberLongRangeInteractionCount vs. log mutation rates for E.c. and T.th.

for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean NumberLongRangeInteractionCount is 0.5771 with standard deviation 1.0090; while for T.th. the mean NumberLongRangeInteractionCount is 0.6130 with standard deviation of 1.0480. The median for E.c. and T.th. is 0 and 0 with interquartile range of 1 and 1 respectively. Lastly the spread of E.c. is 7 with minimum value of 0 and maximum value of 7; while the spread of T.th. is 8 with minimum value of 0 and maximum value of 8. We conclude this exploration with a histogram to give the distribution of NumberLongRangeInteractionCount for each of E.c. and T.th..

4.15 Length of a network of interactions

4.15.1 Definition

This variable is the number nucleotides that are connected by a chain of basepair interactions. For example if there are three nucleotides A, B, and C such that A is basepairing with B, B basepairing with C, but A and C are not basepairing with each other, each of the three
nucleotides would be given the value of three.

4.15.2 Exploring the LengthOfNetworkCount variable

We begin exploring the relationship between LengthOfNetworkCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.01991 and -0.0214 for E.c. and T.th. respectively. We continue our exploration of LengthOfNetworkCount through a simple Boxplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean LengthOfNetworkCount is 1.1170 with standard deviation 0.9209; while for T.th. the mean LengthOfNetworkCount is 1.1120 with standard deviation of 1.0930. The median for E.c. and T.th. is 1 and 1 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 7 with minimum value of 0 and maximum value of 7; while the spread of T.th. is 9 with minimum value of 0 and maximum value of 9. We conclude this exploration with a histogram to give the distribution of LengthOfNetworkCount for each of E.c. and T.th..
Figure 4.67: Boxplot of LengthOfNetworkCount vs. log mutation rates for *E.c.* and *T.th.*

Figure 4.68: Histogram of LengthOfNetworkCount for *E.c.* and *T.th.*
4.16 Extensibility

4.16.1 Definition

We define extensibility to be a portion of a molecule in which is composed entirely of nucleotides that are either not making a basepairing interaction or if they are making a basepairing interaction, it is only a nested cWW basepair. This portion is more free to change in length compared to other portions of the molecule. This region of the structure is usually a stem composed of nucleotides that if is making an interaction it makes only a cWW pair, with no long range interactions. While in most cases, a basepair we observe in the 3D structure will be maintained in the same basepairing family, these regions experience whole basepair indel events because of the lack of constraining factors within this region. Therefore we defined an explanatory variable called extensibility as an indicator taking the value one if the nucleotide in the 3D structure satisfied the conditions for a nucleotide to be in an extensible region as defined above and zero otherwise. We give an example of such a region:

4.16.2 Exploring the ExtensibleCount variable

We begin exploring the relationship between ExtensibleCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This
coefficient is 0.1448 and 0.08776 for E.c. and T.th. respectively. We continue our exploration of ExtensibleCount through a simple Boxplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean ExtensibleCount is 0.06514 with standard deviation 0.2468; while for T.th. the mean ExtensibleCount is 0.05352 with standard deviation of 0.2251. The median for E.c. and T.th. is 0 and 0 with interquartile range of 0 and 0 respectively. Lastly the spread of E.c. is 1 with minimum value of 0 and maximum value of 1; while the spread of T.th. is 1 with minimum value of 0 and maximum value of 1. We conclude this exploration with a histogram to give the distribution of ExtensibleCount for each of E.c. and T.th..

4.17 Isosteric Constraining Effect (ICE)

4.17.1 Introduction

This section is dedicated to the formulation of the Isosteric Constraining Effect, ICE for short, placed upon a nucleotide as a result of the basepairing interactions for which a nucleotide is involved using the concept of isostericity as was defined above. In previous work of Zirbel et
al (2009), simple counts of the number of cWW and non-cWW pairs were used to predict the conservation percentage. However, not every cWW or non-cWW basepair should be treated equally because some basepairs are harder to substitute under the framework of isostericity. We first give a brief explanation of isostericity.

### 4.17.2 Isodiscrepancy matrix

We will continue the example above to explain how we use the framework of isostericity to compute an isosteric constraining effect. For a basepair such as the CG basepair above, using FR3D we construct a 4x4 matrix where the entries are the probabilities, as determined from isostericity, for each basepair combination to substitute for the CG basepair while preserving the Watson Crick interaction between the two nucleotides. The rows of the matrix represent the first nucleotide in the basepair, and the columns of the matrix represent the second nucleotide of the matrix. In the context of the CG example above, if the C mutates, yet the G remains fixed and the resulting interaction is still a Watson Crick basepair, then we would look down the third column in the isosteric matrix for the CG Watson Crick basepair given below. It can be seen, that the CG mutating to a GG is very unlikely, and mutation to an AG basepair is less likely than mutation to an UG basepair.
Table 4.4: Geometric Substitution Score Matrix for a CG cWW basepair based on an IDI Matrix

<table>
<thead>
<tr>
<th>FirstNT/SecondNT</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0233</td>
<td>0.0481</td>
<td>0.0349</td>
<td>0.1373</td>
</tr>
<tr>
<td>C</td>
<td>0.0537</td>
<td>0.0165</td>
<td>0.1411</td>
<td>0.0364</td>
</tr>
<tr>
<td>G</td>
<td>0.0348</td>
<td>0.1388</td>
<td>0.0071</td>
<td>0.0581</td>
</tr>
<tr>
<td>U</td>
<td>0.1396</td>
<td>0.0356</td>
<td>0.0657</td>
<td>0.0289</td>
</tr>
</tbody>
</table>

4.17.3 Defining ICE in relation to isostericity

Using isosteric matrices such as the one above, we wish this use this to build an isostericity score for each nucleotide making a basepairing interaction. In our formulation of the isosteric score from the basepairing interaction we will always set the basepair such that we build the score for the first nucleotide so that the score for each nucleotide is computed in a consistent manner. Returning to the CG example above, when the score is computed for the C, we consider the pair to be CG, and when the score is computed for the G, we consider the pair to be GC. The score will reflect how interchangeable the first nucleotide is in the basepair interaction. We define this score as the standard deviation of the sum of the four rows of the matrix. We do this because if every nucleotide can substitute for the C in some Watson Crick basepair(s), then we would expect that the probability that any of the four nucleotides is used as the first nucleotide is roughly 0.20. This would result in an isostericity score near zero. Conversely, if the first nucleotide in a basepair interaction within a family of basepair interactions does not easily substitute then to a different nucleotide then we would expect that one row sum will be near one, while the others would be near zero. This will result in an isostericity score near $\frac{\sqrt{2}}{2}$. If a nucleotide is part of more than one basepair interaction, then the isostericity score is computed as described above for each interaction for which the nucleotide is a part. After all isostericity scores are computed for a nucleotide, these scores are summed and the result is called the Isosteric Constraining Effect, or ICE for short.
Table 4.5: Geometric Substitution Score Matrix for a CG tWW basepair based on IDI

<table>
<thead>
<tr>
<th>FirstNT/SecondNT</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0267</td>
<td>0.0264</td>
<td>0.0177</td>
<td>0.0452</td>
</tr>
<tr>
<td>C</td>
<td>0.0407</td>
<td>0.0618</td>
<td>0.3543</td>
<td>0.0579</td>
</tr>
<tr>
<td>G</td>
<td>0.0177</td>
<td>0.0445</td>
<td>0.0303</td>
<td>0.0253</td>
</tr>
<tr>
<td>U</td>
<td>0.0985</td>
<td>0.0755</td>
<td>0.0427</td>
<td>0.0347</td>
</tr>
</tbody>
</table>

4.17.4 Example of ICE

If the C in our example was only in a Watson Crick basepairing interaction, and making the interaction with the G, the row sums of the matrix above are 0.2435, 0.2478, 0.2388, and 0.2698. This results in a contribution to ICE of 0.0137. The ICE for this nucleotide suggests the nucleotide may substitute almost as freely as if it was not part of a basepair interaction. Now, consider a new example, where a CG make a trans Watson Crick basepair. The isosteric matrix is given below:

From the substitution matrix above, the CG basepair from the trans Watson Crick basepairing family have very different row sums which are: 0.1160, 0.5147, 0.1179, and 0.2514. This results in a contribution to ICE of 0.1875, which is fairly high for a nucleotide making only one interaction.

4.17.5 Summary of creation of ICE

As it is defined above ICE takes non-negative values, taking the value zero if the nucleotide does not make any basepair interactions, or zero if all of the interactions the nucleotide makes have an isostericity score of zero; with positive values assigned to nucleotides making interactions that have a high isostericity score, or nucleotides making many interactions with moderate isostericity score. We expect ICE to be a relevant predictor of nucleotide mutation rates because if a nucleotide has a high ICE, then we expect that this nucleotide will not be substituted easily for a different nucleotide. Below we give a histogram to gain an understanding of the distribution of the values of ICE.
4.17.6 Exploring the ICE variable

We begin exploring the relationship between ICE and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is -0.1564 and -0.2024 for E.c. and T.th. respectively. We continue our exploration of ICE through a simple Scatterplot for each of E.c. and T.th..

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th. For E.c. the mean ICE is 0.2948 with standard deviation 0.2090; while for T.th. the mean ICE is 0.2866 with standard deviation of 0.2183. The median for E.c. and T.th. is 0.2500 and 0.2482 with interquartile range of 0.09919 and 0.07007 respectively. Lastly the spread of E.c. is 1.5010 with minimum value of 0 and maximum value of 1.5010; while the spread of T.th. is 1.7930 with minimum value of 0 and maximum value of 1.7930. We conclude this exploration with a histogram to give the distribution of ICE for each of E.c. and T.th.

4.17.7 NearICE

NearICE is defined just as ICE but for nucleotides making a near-type interaction.
4.17.8 Exploring the NearICE variable

We begin exploring the relationship between NearICE and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is -0.2227 and -0.2295 for *E.c.* and *T.th.* respectively. We continue our exploration of NearICE through a simple Scatterplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean NearICE is 0.1341 with standard deviation 0.2757; while for *T.th.* the mean NearICE is 0.1574 with standard deviation of 0.2894. The median for *E.c.* and *T.th.* is 0 and 0 with interquartile range of 0.2025 and 0.2500 respectively. Lastly the spread of *E.c.* is 2.3090 with minimum value of 0 and maximum value of 2.3090; while the spread of *T.th.* is 2.6130 with minimum value of 0 and maximum value of 2.6130. We conclude this exploration with a histogram to give the distribution of NearICE for each of *E.c.* and *T.th.*.
Figure 4.74: Scatterplot of NearICE vs. log mutation rates for *E.c.* and *T.th.*

Figure 4.75: Histogram of NearICE for *E.c.* and *T.th.*
4.18 Communication of constraints to a nucleotide

4.18.1 Introduction

In previous sections, we have explored the role of interactions that nucleotide make with or without other nucleotides and how these interactions relate to the site specific mutation rate. However, the interactions that a nucleotide makes does not explain all of the variation of the site specific mutation rates that we see. The interactions of the interacting nucleotide might too play a role in constraining a nucleotide against mutation. We now explain how we take into account the interactions that the interacting nucleotides make, and how they relate to the site specific mutation rates. We called this variable that takes into account the interactions of the interacting variables, the communicator of a nucleotide for the communication of the constraint passed to a nucleotide from the interacting nucleotides. We have considered four versions of this communicator idea. They are: TrueCommunicatorCount, TrueCommunicatorICE, AllCommunicatorCount, and AllCommunicatorICE.

4.18.2 Building the TrueCommunicatorCount variable

The interactions of the interacting nucleotides, as classified by FR3D, that were taken into account in the communicator variable are: the number of times that the interacting nucleotide makes a true cWW or noncWW interaction, with the interaction between the interacting nucleotide and the nucleotide that we are studying being removed, the number of times that base of the interacting nucleotide is used in a base-phosphate interaction, the number of stacking partners the interacting nucleotide has, and the number of BaseRibose interactions of the interacting nucleotide.

4.18.3 Building the TrueCommunicatorCount variable

To build TrueCommunicatorCount we had to select how to weight each of the terms relative to each other for the variables above. We ran a linear regression on the log mutation rates as
a function of the terms above and used the regression coefficients to serve as the weights for each of the corresponding terms above to build the communicator variable. We then used this weights to compute the linear combination of these variables to define the TrueCommunicatorCount value for each nucleotide in the conserved core for each 3D structure.

4.18.4 Exploring the TrueCommunicatorCount variable

We begin exploring the relationship between TrueCommunicatorCount and the log of the mutation rates for each of E.c. and T.th. molecules through the Spearman correlation coefficient. This coefficient is 0.1248 and 0.2130 for E.c. and T.th. respectively. We continue our exploration of TrueCommunicatorCount through a simple Scatterplot for each of E.c. and T.th.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of E.c. and T.th.. For E.c. the mean TrueCommunicatorCount is -0.3337 with standard deviation 0.7148; while for T.th. the mean TrueCommunicatorCount is -0.1202 with standard deviation of 0.6968. The median for E.c. and T.th. is -0.2075 and 0 with interquartile range of 0.5397 and 0.4258 respectively. Lastly the spread of E.c. is
6.2780 with minimum value of -4.854 and maximum value of 1.4240; while the spread of \( T.th. \) is 7.0470 with minimum value of -5.124 and maximum value of 1.9230. We conclude this exploration with a histogram to give the distribution of TrueCommunicatorCount for each of \( E.c. \) and \( T.th. \).

### 4.18.5 Building the TrueCommunicatorICE variable

The interactions of the interacting nucleotides, as classified by FR3D, that were taken into account in the communicator variable are: the ICE score for the interacting nucleotide with the interaction between the interacting nucleotide, and the nucleotide that we are studying being removed, BasePhosphateGeometry, SDI, and BaseRiboseGeometry.

### 4.18.6 Building the TrueCommunicatorICE variable

To build TrueCommunicatorICE we had to select how to weight each of the terms relative to each other for the variables above. We ran a linear regression on the log mutation rates as a function of the terms above and used the regression coefficients to serve as the weights for each of the corresponding terms above to build the communicator variable. We then used this weights to compute the linear combination of these variables to define the
Figure 4.78: Scatterplot of TrueCommunicatorICE vs. log mutation rates for \textit{E.c.} and \textit{T.th.}

TrueCommunicatorICE value for each nucleotide in the conserved core for each 3D structure.

\textbf{4.18.7 Exploring the TrueCommunicatorICE variable}

We begin exploring the relationship between TrueCommunicatorICE and the log of the mutation rates for each of \textit{E.c.} and \textit{T.th.} molecules through the Spearman correlation coefficient. This coefficient is 0.2416 and 0.1936 for \textit{E.c.} and \textit{T.th.} respectively. We continue our exploration of TrueCommunicatorICE through a simple Scatterplot for each of \textit{E.c.} and \textit{T.th.}.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of \textit{E.c.} and \textit{T.th.}. For \textit{E.c.} the mean TrueCommunicatorICE is -0.974 with standard deviation 0.9516; while for \textit{T.th.} the mean TrueCommunicatorICE is -0.8881 with standard deviation of 0.8888. The median for \textit{E.c.} and \textit{T.th.} is -0.6694 and -0.6145 with interquartile range of 0.6884 and 0.6045 respectively. Lastly the spread of \textit{E.c.} is 7.4910 with minimum value of -7.491 and maximum value of 0; while the spread of \textit{T.th.} is 6.6300 with minimum value of -6.630 and maximum value of 0. We conclude this exploration with a histogram to give the distribution of TrueCommunicatorICE for each of \textit{E.c.} and \textit{T.th.}. 

4.19 Communication of near interaction constraints on a nucleotide

4.19.1 Introduction to AllCommunicatorCount variables

In addition to the true interactions that FR3D classifies, FR3D also classifies near-type variables. We consider both true and near-type interactions for both the interaction between nucleotides, and to communicate the constraint the interacting nucleotide communicates back to the nucleotide that we are studying.

4.19.2 Terms used to make the variable

The interactions of the interacting nucleotides that were taken into account in the AllCommunicatorCount contain all of those from TrueCommunicatorCount, but include the following near-type interaction variables. These are: the number of times that the interacting nucleotide makes a NearcWW or NearnoncWW interaction, with the interaction between the interacting nucleotide and the nucleotide that we are studying being removed as in the Communicator variable, the number of times that base of the interacting nucleotide is used
in a Nearbase phosphate interaction, the number of NearStackingPartners of the interacting nucleotide, and the number of NearBaseRibose interactions of the interacting nucleotide.

### 4.19.3 Building the AllCommunicatorCount variable

To build the AllCommunicatorCount we had to select how to weight each of the terms relative to each other for the variables above. We ran a linear regression on the log mutation rates as a function of the terms above and used the regression coefficients to serve as the weights for each of the corresponding terms above to build the communicator variable. We then used this weights to compute the linear combination of these variables to define the AllCommunicatorCount value for each nucleotide in the conserved core for each 3D structure.

### 4.19.4 Exploring the AllCommunicatorCount variable

We begin exploring the relationship between AllCommunicatorCount and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is 0.3410 and 0.2900 for *E.c.* and *T.th.* respectively. We continue our exploration of AllCommunicatorCount through a simple Scatterplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean AllCommunicatorCount is -0.553 with standard deviation 0.9253; while for *T.th.* the mean AllCommunicatorCount is -0.4069 with standard deviation of 0.8107. The median for *E.c.* and *T.th.* is -0.1227 and -0.07548 with interquartile range of 0.8935 and 0.9349 respectively. Lastly the spread of *E.c.* is 7.8230 with minimum value of -7.136 and maximum value of 0.6879; while the spread of *T.th.* is 6.2650 with minimum value of -4.749 and maximum value of 1.5160. We conclude this exploration with a histogram to give the distribution of AllCommunicatorCount for each of *E.c.* and *T.th.*.
Figure 4.80: Scatterplot of AllCommunicatorCount vs. log mutation rates for *E.c.* and *T.th.*

Figure 4.81: Histogram of AllCommunicatorCount for *E.c.* and *T.th.*
4.19.5 Terms used to make the AllCommunicatorICE variable

The interactions of the interacting nucleotides that were taken into account in the AllCommunicatorCount contain all of those from TrueCommunicatorCount, but include the following near-type interaction variables. These are: the number of times that the interacting nucleotide makes a NearcWW or NearnoncWW interaction, with the interaction between the interacting nucleotide and the nucleotide that we are studying being removed as in the Communicator variable, the number of times that base of the interacting nucleotide is used in a Nearbase-phosphate interaction, the number of NearStackingPartners of the interacting nucleotide, and the number of NearBaseRibose interactions of the interacting nucleotide.

4.19.6 Building the AllCommunicatorICE variable

To build the AllCommunicatorCount we had to select how to weight each of the terms relative to each other for the variables above. We ran a linear regression on the log mutation rates as a function of the terms above and used the regression coefficients to serve as the weights for each of the corresponding terms above to build the communicator variable. We then used this weights to compute the linear combination of these variables to define the AllCommunicatorCount value for each nucleotide in the conserved core for each 3D structure.

4.19.7 Exploring the AllCommunicatorICE variable

We begin exploring the relationship between AllCommunicatorICE and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is 0.3616 and 0.3465 for *E.c.* and *T.th.* respectively. We continue our exploration of AllCommunicatorICE through a simple Scatterplot for each of *E.c.* and *T.th.*.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of *E.c.* and *T.th.*. For *E.c.* the mean AllCommunicatorICE is -1.552 with standard deviation 1.5710; while for *T.th.* the mean AllCommunicatorICE is -1.602 with
standard deviation of 1.5650. The median for *E.c.* and *T.th.* is -0.8356 and -0.9378 with interquartile range of 1.6740 and 1.8420 respectively. Lastly the spread of *E.c.* is 11.730 with minimum value of -11.73 and maximum value of 0; while the spread of *T.th.* is 11.040 with minimum value of -11.04 and maximum value of 0. We conclude this exploration with a histogram to give the distribution of AllCommunicatorICE for each of *E.c.* and *T.th.*.

4.19.8 Building the AllCommunicatorICEColumn variable

To build AllCommunicatorICEColumn we weighted the communicated constraint by the variability of the nucleotide that we are studying with the interacting nucleotide remaining fixed. This was done by taking the same Isostericity matrices that we used in creating the ICE score, but here we took the standard deviation of the column corresponding to the interacting nucleotide.

4.19.9 Exploring the AllCommunicatorICEColumn variable

We begin exploring the relationship between AllCommunicatorICEColumn and the log of the mutation rates for each of *E.c.* and *T.th.* molecules through the Spearman correlation coefficient. This coefficient is 0.3768 and 0.3465 for *E.c.* and *T.th.* respectively. We continue
our exploration of AllCommunicatorICEColumn through a simple Scatterplot for each of \textit{E.c.} and \textit{T.th.}.

We continue our exploration of this explanatory variable by examining basic descriptive statistics for each of \textit{E.c.} and \textit{T.th.}. For \textit{E.c.} the mean AllCommunicatorICEColumn is -1.708 with standard deviation 1.6940; while for \textit{T.th.} the mean AllCommunicatorICEColumn is -1.602 with standard deviation of 1.5650. The median for \textit{E.c.} and \textit{T.th.} is -0.9328 and -0.9378 with interquartile range of 1.9030 and 1.8420 respectively. Lastly the spread of \textit{E.c.} is 12.130 with minimum value of -12.13 and maximum value of 0; while the spread of \textit{T.th.} is 11.040 with minimum value of -11.04 and maximum value of 0. We conclude this exploration with a histogram to give the distribution of AllCommunicatorICEColumn for each of \textit{E.c.} and \textit{T.th.}.
Figure 4.84: Scatterplot of AllCommunicatorICEColumn vs. log mutation rates for E.c. and T.th.

Figure 4.85: Histogram of AllCommunicatorICEColumn for E.c. and T.th.
CHAPTER 5

Modeling site-specific mutation rates

5.1 Overview

In this chapter we model nucleotide mutation rate with the 3D structure terms that we have defined in Chapter 4. The goal of this model is to provide us insight into how the local 3D structure affects mutation rates. In addition, we would like this model to be easy to interpret thereby giving future researchers intuition about site-specific mutation rates if they have knowledge of the local 3D structure near the nucleotide. For these reasons we consider only main effects models and defer any exploration of models containing interaction terms or transformations of the predictor terms for future work. In this work, we have the 3D structure information for ribosomes of \textit{E.c.} and for \textit{T.th.} We will build a model for the site-specific mutation rates of the nucleotides of the conserved core using the 3D structure information of each \textit{E.c.} and \textit{T.th.} We would like to see a consensus of 3D structure terms for each of the \textit{E.c.} and \textit{T.th.} models because this will give further evidence that the effect of the 3D structure terms have on the mutation rate might extend to all rRNA since the last common ancestor of \textit{E.c.} and \textit{T.th.} is inferred to be so far in the past that most if not all 3D structures of bacteria will have a similar effect on site-specific mutation rates.

We begin this chapter by explaining how we select a parsimonious subset of terms to
model nucleotide mutation rates from the pool of 45 predictors using log-linear regression of nucleotide mutation rate. We will see that the subset largely consists of interactions and the local density of nucleotides and amino acids near the nucleotide. We will show that replacing “count” terms with terms that represent the geometry of interactions perform nearly as well while providing more insight into why we see the effect that we do. After producing the model of nucleotide mutation rates we then apply hierarchical regression to tailor the model to each of the 5S, 16S, and 23S structures, and perform a factor analysis on the terms to understand the commonality, if any, between the predictors.

5.2 A best subset of terms to model site-specific mutation rates

Currently, there are numerous ways to select a “best” subset of variables for use in a regression model. We begin this discussion by discussing several common methods that can be found in [7] that we did not choose before explaining the process that we did use.

The first method that we discuss is one that uses the residual mean squared error from fitted models to determine the best number of predictors for a parsimonious model. The method works by computing the residual mean squared error, $s^2(p)$, for all combinations of $p$ terms, $p = 1, 2, \ldots P$, where $P = 45$ in our case. The combinations are then compared to the model that includes all $P$ terms. The “best” subset of terms is the smallest set of $p$ terms that has $s^2(p)$ as close as possible to $s^2(P)$.

Next, we discuss Mallows $C_p$ statistic. Like the residual mean squared error, Mallows $C_p$ statistic compares models of a subset of $p$ terms to the model with all $P$ terms. The statistic is given by $C_p = \frac{RSS_p}{s^2} - (n - 2p)$, where $RSS_p$ is the residual sum of squares of a given model with $p$ terms, including the intercepts, $s^2$ is assumed to be a reliable estimate of the asymptotic residual squared error of the model, which is obtained by finding the residual sum of squares from the model containing all $P$ terms, and $n$ is the number of observations.
Using this criterion, the “best” subset of terms selected is the smallest such subset that has $C_p \approx P$.

Next we discuss the Akaike Information Criterion (AIC) [2]. This criterion automates the process of selecting a “best” subset by computing the likelihood of a model but adjusting the score downward based upon the number of terms that are in the model. The statistic is given by $AIC = 2p - 2\ln(L)$, where $p$ is the number of terms in the model and $L$ is the likelihood of the model. Using this criterion, the best subset of $p$ predictors is the one that minimizes the $AIC$ value. Similar to the AIC there is the Bayesian Information Criterion (BIC) [12]. The BIC criterion is given by $BIC = -2\ln(L) + p\ln(n)$, where $L$ is the likelihood of the model, $p$ is the number of parameters, and $n$ is the number of observations. The best subset of predictors then is the subset that corresponds to the smallest value of $BIC$.

The final method that we discuss is that of $R^2$ shrinkage as a result of additional terms being included in the model [21]. A common method employed is adjusted $R^2_{adj}$. In this method, the $R^2_{adj}$ statistic is given by $R^2_{adj} = \frac{n-1}{n-p-1}R^2$, where $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, $\hat{y}_i$ is the estimated value of $y_i$, and $\bar{y}$ is the mean of the observations $y$. The best subset of predictors is the set that maximizes the $R^2_{adj}$ value. Applied to our

The last procedure that we discuss is stepwise regression. For each step, the procedure works by testing the term most highly correlated with the response variable that is not in the model in the presence of the $p - 1$ terms that are already in the model in addition to the intercept term, for $p = 1, 2, \ldots, P$. This term is selected to be in the model if it meets a predefined criterion, such as having a p-value that is less than 0.05. If no term enters the model then the process stops. If the term does enter the model, then the $p - 1$ terms that were in the model before the $p^{th}$ term are tested in order of largest p-value to smallest p-value removing the terms that do not satisfy the criterion to remain in the model, for example to have a p-value of 0.05, until all terms that remain in the model satisfy the criterion [7]. Therefore a researcher may set the steps so that the only terms that are in the model at any given step to have a p-value less than 0.05. If we set the p-value criterion in this way, not
only does this allow us to reduce the number of 45 predictors to a best subset of predictor terms, but we also reduce it to a set that is statistically significant at the 0.05 level [6].

We found that the best subset obtained from any of the criteria: AIC, adjusted $R^2$, or stepwise regression each overfit the model. So we chose to use stepwise regression because we are able to set the p-value criterion to produce a “best” subset consisting only of statistically significant terms for modeling site-specific mutation rates from the set of 45 predictors. This is a better initial step because it along with the procedures listed above reduces the 45 predictors down to a “best” subset, but also ensures that each of the terms are statistically significant. We ran this regression on each of the $E.c.$ and the $T.th.$ datasets to determine an initial parsimonious subset for each. We set the parameters for entering and leaving the model to be 0.05. This yielded about 20 terms for each of the $E.c.$ and the $T.th.$ datasets.

We say that the stepwise regression model is overfitting predictors to the model because many of the 20 predictors are not contributing much to the power of the model of explaining the variation of the data, measured by the $R^2$ value. After about the first ten terms, for each of the $E.c.$ and the $T.th.$ models, the contribution of additional terms to the $R^2$ value was extremely small so we then use the $R^2$ value as a criterion to further reduce the number of statistically significant terms to a parsimonious subset. To remove the terms in the model that are not contributing much, we use a criterion that terms would be entered into the model in the order that they enter the model under the stepwise regression procedure until the $R^2$ value of the next term to be added would contribute less than 1% of an increase to the overall $R^2$ value. We believe that the biological effect of the terms that are adding less than one percent are being accounted for by other terms already in the model, and thus not useful in modeling site-specific mutation. To give an example, we consider the number of times the phosphate is used in a base-phosphate interaction. The contribution by adding this term to the model resulted in an increase in the $R^2$ value less than one percent. It is likely that any meaningful signal for this variable is accounted for already by the measure of local density such as the NucleotideSphereNearPlaneCount.
Restricting the terms in the model to those meeting the criterion described above resulted in a nine term model for E.c. and a ten term model for T.th. In both E.c. and T.th. NucleotideSphereNearPlaneCount, AminoAcidSphereCount, SmallRNASphereNearPlaneCount, AllCommunicator, noncWWCount, BRGeometry, BPhCount, DistanceFromCenter and SDI were determined to enter into a predictive subset. This accounts for all nine of the terms in the E.c dataset. The term NearICE is the last remaining term in the T.th. dataset. This suggests that there is a strong consensus between E.c. and T.th. in the 3D structure terms having the largest effect on mutation rates.

However our task is not solved by simply taking the terms that are the most predictive, we must now determine if more biologically relevant terms, such as the measure of constraint based upon the geometry of interactions can be substituted for the less biologically relevant terms such as simple counts of interactions while still predicting mutation rates. We do this because terms like non-cWWCount for example count all of the non Watson Crick basepairing interactions. However current understanding of basepairing interactions tells us that some bases are harder to substitute for other bases while preserving the interaction [46]. Similarly for base-phosphate interactions [53]. Therefore, we wish to replace these terms by the more biologically relevant geometric predictors to gain a better insight into the affect that the local geometry of a nucleotide has on site-specific mutation rates.

Here we compare models that include BPhGeometry instead of BPhCount, and AllCommunicatorICEColumn instead of AllCommunicatorCount. In addition, we include the terms AllICE instead of noneWWCount for the E.c model and AllICE instead of noneWWCount and NearICE in the T.th. model. With these replacements the $R^2$ value drops from 0.3732 to 0.3490 in the E.c model and from 0.3507 to 0.3307 in the T.th. model. This is a drop of about 6% and 5% respectively. However the terms that we use tell us far more about how the local 3D structure of a nucleotide is constrained against mutation and provides a better model for nucleotide evolution than the model with the simple counts.

Before going on we make one modification to the model that adjusts for cWW basepairing
Table 5.1: Regression coefficients, standard error, and predictive contribution of each term with all others in the model for the *E. c.* structure, listed in decreasing order of partial $R^2$.

<table>
<thead>
<tr>
<th>Term</th>
<th>coefficient</th>
<th>standard error</th>
<th>Partial $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>$-0.4849$</td>
<td>$0.1442$</td>
<td></td>
</tr>
<tr>
<td>NucleotideSphereNearPlaneCount</td>
<td>$-0.0918$</td>
<td>$0.0058$</td>
<td>0.0393</td>
</tr>
<tr>
<td>SmallRNA SphereNearPlaneCount</td>
<td>$-0.0491$</td>
<td>$0.0032$</td>
<td>0.0385</td>
</tr>
<tr>
<td>AminoAcidSphereCount</td>
<td>$-0.0399$</td>
<td>$0.0027$</td>
<td>0.0347</td>
</tr>
<tr>
<td>Isosteric Constraining Effect (ICE)</td>
<td>$-1.3603$</td>
<td>$0.1349$</td>
<td>0.0162</td>
</tr>
<tr>
<td>AllCommunicatorICEColumn</td>
<td>$0.1718$</td>
<td>$0.0174$</td>
<td>0.0155</td>
</tr>
<tr>
<td>Watson Crick pair indicator (cWWCount)</td>
<td>$0.5735$</td>
<td>$0.0594$</td>
<td>0.0149</td>
</tr>
<tr>
<td>DistanceFromCenter</td>
<td>$0.0100$</td>
<td>$0.0011$</td>
<td>0.0131</td>
</tr>
<tr>
<td>Stacking Discrepancy Index (SDI)</td>
<td>$-1.8954$</td>
<td>$0.3257$</td>
<td>0.0054</td>
</tr>
<tr>
<td>BaseRiboseGeometry (BRGeometry)</td>
<td>$-0.3957$</td>
<td>$0.0721$</td>
<td>0.0048</td>
</tr>
<tr>
<td>BasePhosphateGeometry (BPhGeometry)</td>
<td>$-0.3558$</td>
<td>$0.0679$</td>
<td>0.0044</td>
</tr>
</tbody>
</table>

Table 5.1: Regression coefficients, standard error, and predictive contribution of each term with all others in the model for the *E. c.* structure, listed in decreasing order of partial $R^2$.

Interactions. Since we observed that nucleotides involved in a cWW interaction mutate more frequently than nucleotides from other types of interactions, we add the indicator term cWWCount that takes the value 1 if the nucleotide is part of a cWW interaction and zero otherwise. Adding this term to the model increases the $R^2$ value from 0.03490 to 0.3574 in the *E. c.* model and from 0.3307 to 0.3454 in the *T. th.* model. Therefore the model that we present for predicting mutation rates is one that includes the intercept term and these following ten terms: NucleotideSphereNearPlaneCount, AminoAcidSphereCount, SmallRNA SphereNearPlaneCount, DistanceFromCenter, cWWCount, ICE (*E. c.*)/ AllICE (*T. th.*), AllCommunicator, BPhGeometry, BRGeometry, and SDI. We give the summary of regression coefficients and their contribution to the model for the *E. c.* model in Table 5.1 and the *T. th.* model in Table 5.2. In each of these tables, these terms are given in order of their partial $R^2$ value, which is the amount of variation which is explained by the term given that all of the other terms are in the model.

From Table 5.1 and Table 5.2 we see that these terms are statistically significant factors in modeling site-specific mutation rates since the minimum ratio of estimate to standard error is greater than 4, with ratios larger than 2 being significant. What is surprising about these models is that the local density terms, which are NucleotideSphereNearPlaneCount,
<table>
<thead>
<tr>
<th>Term</th>
<th>coefficient</th>
<th>standard error</th>
<th>Partial $R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>−0.7810</td>
<td>0.1526</td>
<td></td>
</tr>
<tr>
<td>NucleotideSphereNearPlaneCount</td>
<td>−0.0844</td>
<td>0.0059</td>
<td>0.0349</td>
</tr>
<tr>
<td>SmallRNASphereNearPlaneCount</td>
<td>−0.0389</td>
<td>0.0032</td>
<td>0.0253</td>
</tr>
<tr>
<td>AminoAcidSphereCount</td>
<td>−0.0405</td>
<td>0.0027</td>
<td>0.0360</td>
</tr>
<tr>
<td>All Isosteric Constraining Effect (AllICE)</td>
<td>−0.5606</td>
<td>0.1213</td>
<td>0.0142</td>
</tr>
<tr>
<td>AllCommunicatorICEColumn</td>
<td>0.1776</td>
<td>0.0264</td>
<td>0.0188</td>
</tr>
<tr>
<td>Watson Crick pair indicator (cWWCount)</td>
<td>0.5733</td>
<td>0.0608</td>
<td>0.0236</td>
</tr>
<tr>
<td>DistanceFromCenter</td>
<td>0.0111</td>
<td>0.0011</td>
<td>0.0138</td>
</tr>
<tr>
<td>Stacking Discrepancy Index (SDI)</td>
<td>−1.8509</td>
<td>0.3435</td>
<td>0.0037</td>
</tr>
<tr>
<td>BaseRiboseGeometry (BRGeometry)</td>
<td>−0.2986</td>
<td>0.0748</td>
<td>0.0025</td>
</tr>
<tr>
<td>BasePhosphateGeometry (BPhGeometry)</td>
<td>−0.2718</td>
<td>0.0668</td>
<td>0.0028</td>
</tr>
</tbody>
</table>

Table 5.2: Regression coefficients, standard error, and predictive contribution of each term with all others in the model for the *T.th.* structure, listed in decreasing order of partial $R^2$.

SmallRNASphereNearPlaneCount, and AminoAcidSphereCount, are the highest contributors to reducing the rate of mutation for a nucleotide instead of the terms that account for basepairing interactions, which are ICE/AllICE and AllCommunicatorICEColumn. Also, it should be noted that the Variance Inflation Factor (VIF) diagnostic was run on these terms to check for a problem of multi-collinearity. The result of this diagnostic was that no problem of multi-collinearity was detected as the maximum VIF was less than 1.5.

### 5.3 Using the model

To explain how this model of nucleotide mutation can be used, we focus on the model for the *E.c.* structure. This model for *E.c* is given by:
\[
\log(\text{Mutation Rate}) = -0.4849 - 0.0918 \cdot \text{NucleotideSphereNearPlaneCount} - 0.0491 \cdot \text{SmallRNASphereNearPlaneCount} - 0.0399 \cdot \text{AminoAcidSphereCount} - 1.3603 \cdot \text{ICE} + 0.1718 \cdot \text{AllCommunicatorICEColumn} + 0.5735 \cdot \text{cWWCount} + 0.0100 \cdot \text{DistanceFromCenter} - 1.8954 \cdot \text{SDI} - 0.3957 \cdot \text{BRGeometry} - 0.3558 \cdot \text{BPhGeometry}
\] (5.1)

However, this model may be more insightful when expressed as:
Mutation Rate = \exp(-0.4849) \\
\cdot \exp(-0.0918 \cdot \text{Nucleotide} \text{Sphere} \text{Near} \text{Plane} \text{Count}) \\
\cdot \exp(-0.0491 \cdot \text{Small} \text{RNA} \text{Sphere} \text{Near} \text{Plane} \text{Count}) \\
\cdot \exp(-0.0399 \cdot \text{Amino} \text{Acid} \text{Sphere} \text{Count}) \\
\cdot \exp(-1.3603 \cdot \text{ICE}) \\
\cdot \exp(0.1718 \cdot \text{All} \text{ Communicator} \text{ICE} \text{Column}) \\
\cdot \exp(0.5735 \cdot \text{cWW} \text{Count}) \\
\cdot \exp(0.0100 \cdot \text{Distance} \text{From} \text{Center}) \\
\cdot \exp(-1.8954 \cdot \text{SDI}) \\
\cdot \exp -0.3957 \cdot \text{BR} \text{Geometry} \\
\cdot \exp -0.3558 \cdot \text{BPh} \text{Geometry} \\
(5.2)

When expressed in this form, the multiplicative nature of the model is highlighted. When all of the explanatory variables take the value zero then the expected mutation rate is simply $e^{-0.4849}(e^0)^{10} \approx 0.6158$. This value can be thought of expected mutation rate for a nucleotide for one unit of evolutionary time for a nucleotide void of interactions, but in the center of the ribosome. This cannot happen biologically, so we point out that the intercept term is not biologically meaningful. To further understand this model, consider an example with a term that has a positive coefficient such as cWWCount. For example, if the nucleotide we used in the example above have all of the same traits except this nucleotide is making a cWW pair, then the cWWCount term takes the value one. For this case the expected mutation rate the nucleotide increases by $e^{0.5735}$ or approximately 1.7745 times higher. Conversely, if there is an explanatory variable that has a negative coefficient then the presence of this factor decreases the expected mutation rate. For example if we compare two nucleotides such that
the first has no amino acids in its sphere, while the second nucleotide has one more amino acid in its sphere with all other terms held constant between the two, then the expected mutation rate is $e^{-0.0399} \approx 0.9609$ times the mutation rate of the nucleotide that does not.

5.4 Multilevel modeling approach

5.4.1 Multilevel modeling applied to site-specific mutation rates

A multilevel model, also called hierarchical regression model, is a very useful statistical tool with many applications. We will highlight two of the common applications of multilevel modeling. The first instance where multilevel modeling is applied is a situation in which the data has a clear hierarchical structure. For example, if we wanted to model how individuals across the United States voted in the last Presidential election by the state city, and precinct in which they voted there would be a clear hierarchical structure to the data that would have an effect on the the way that individuals vote and its effect should be taken into account. In this instance, the multilevel model works to identify the variation due to the state, and city of the individual. This type of structure lends itself readily to a multilevel model because there is a clear structure to the data with known effects at the different levels. For example, the state of Ohio will sometimes vote republican, but the city of Youngstown votes democratic, while the city of Cincinnati votes republican. Therefore there are two sources of variation. If a researcher wanted to control for the regional influence in the city of the state while looking for the demographics of the individual then a multilevel model would work well as the model fit would adjust the parameters for the state and city effect accordingly. For more information on the applications and constructions of these models we refer the reader to [14], which provides a good background of this modeling process and how to apply it.

The second instance that we highlight is when there is a clear grouping of the data that forms a partition of the dataset such as the multiple sequence alignments of different rRNA chains that we have in this work. In this study, each observation comes from one of the 5S,
16S, or 23S multiple sequence alignments; this forms a partition of the entire dataset. In this setting a multilevel model is helpful in understanding the effect of the chain on modeling the site-specific mutation rates. For example, one predictor that might have varying effects on the site-specific mutation rate over the different structures is the distance from the center of the ribosome since the average distance of a nucleotide from the center of the ribosome is 99.30, 86.58, and 67.92 Angstroms for the 5S, 16S, and 23S structures respectively. In addition to this variation, each multiple sequence alignment contains different sequences and a different number of sequences, which provides another source of variation. We adjusted for this in the estimating of the site-specific mutation rates in Chapter 3 by calibrating the branch lengths, but we anticipate that there is still variation due to having different sequences in each alignment as well. By applying a multilevel model, we account for some of this variability. We do this to account for the two sources of variability due to the different structures without having to fit a separate model for each of the three structures. Mathematically this is done by placing an underlying distribution on either the intercept and the slope coefficient terms over the three structures. We will choose to allow the slope and the intercept coefficients to vary across the structures. This is known as a varying intercept, varying slope multilevel model.

### 5.4.2 A Multilevel model for site-specific mutation rates

The model that we present in this work is a varying intercept, varying slope hierarchical model grouped by the different structures. The model is given by

\[
\log(\hat{y}_i) = \hat{a}_{j[i]} + \sum_{p=1}^{P} \hat{\beta}_{p,j[i]} x_{p,i}
\]  

(5.3)

In Equation 5.3, \( y_i \) is the site-specific mutation rate. \( a_{j[i]} \) is the intercept term for the \( j^{th} \) group, for which individual \( i \) is a member, \( x_{p,i} \) is the \( p^{th} \) predictor for site-specific mutation rate \( i \), and \( \hat{\beta}_{p,j[i]} \) is the estimated regression coefficient within the \( j^{th} \) group corresponding to
Table 5.3: Underlying parameters for the normal distribution for each of the hierarchical model for E.c and T.th grouped by organism

<table>
<thead>
<tr>
<th>Term</th>
<th>E.c $\hat{\mu}$</th>
<th>E.c $\hat{\sigma}$</th>
<th>T.th $\hat{\mu}$</th>
<th>T.th $\hat{\sigma}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.8017</td>
<td>0.1870</td>
<td>-1.0975</td>
<td>0.1572</td>
</tr>
<tr>
<td>NucleotideSphereNearPlaneCount</td>
<td>-0.0906</td>
<td>0.0095</td>
<td>-0.0615</td>
<td>0.0289</td>
</tr>
<tr>
<td>SmallRNASphereNearPlaneCount</td>
<td>-0.0446</td>
<td>0.0059</td>
<td>-0.0326</td>
<td>0.0055</td>
</tr>
<tr>
<td>AminoAcidSphereCount</td>
<td>-0.0398</td>
<td>0.0027</td>
<td>-0.0381</td>
<td>0.0037</td>
</tr>
<tr>
<td>ICE (E.c)/ AllICE (T.th.)</td>
<td>-1.3456</td>
<td>0.1455</td>
<td>-0.4833</td>
<td>0.1423</td>
</tr>
<tr>
<td>AllCommunicatorICEColumn</td>
<td>0.1744</td>
<td>0.0204</td>
<td>0.1591</td>
<td>0.0410</td>
</tr>
<tr>
<td>cWWCount</td>
<td>0.5732</td>
<td>0.0597</td>
<td>0.5372</td>
<td>0.0689</td>
</tr>
<tr>
<td>DistanceFromCenter</td>
<td>0.0132</td>
<td>0.0012</td>
<td>0.0119</td>
<td>0.0033</td>
</tr>
<tr>
<td>SDI</td>
<td>-1.8257</td>
<td>0.3809</td>
<td>-1.2950</td>
<td>0.7248</td>
</tr>
<tr>
<td>BRGeometry</td>
<td>-0.4149</td>
<td>0.0890</td>
<td>-0.2744</td>
<td>0.0831</td>
</tr>
<tr>
<td>BPhGeometry</td>
<td>-0.3978</td>
<td>0.0970</td>
<td>-0.1549</td>
<td>0.1382</td>
</tr>
</tbody>
</table>

By estimating the underlying distribution, this allows us to consider the effect the predictors depending upon from which structure the nucleotide is. We now go on to give the adjusted regression parameters for each of the E.c and T.Th. molecules below:

From Table 5.4 and Table 5.5, we see that within the chain groups, the ratio of the coefficient estimates to their corresponding standard error is quite large for each ratio found in the Table for both organisms. This suggests that within each chain the 3D structure term is statistically significant for site-specific mutation rates. From the standard deviation
Table 5.4: Regression coefficients and standard errors for the hierarchical model for *E. c.* grouped by chain

<table>
<thead>
<tr>
<th>Term</th>
<th>$\hat{\beta}_{5S}$</th>
<th>$\hat{\sigma}_{5S}$</th>
<th>$\hat{\beta}_{16S}$</th>
<th>$\hat{\sigma}_{16S}$</th>
<th>$\hat{\beta}_{23S}$</th>
<th>$\hat{\sigma}_{23S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.7658</td>
<td>0.0978</td>
<td>-0.9871</td>
<td>0.0332</td>
<td>-0.6521</td>
<td>0.0240</td>
</tr>
<tr>
<td>NucleotideSphereNearPlaneCount</td>
<td>-0.0440</td>
<td>0.0067</td>
<td>-0.1032</td>
<td>0.0023</td>
<td>-0.0804</td>
<td>0.0016</td>
</tr>
<tr>
<td>SmallRNASphereNearPlaneCount</td>
<td>-0.0462</td>
<td>0.0044</td>
<td>-0.0363</td>
<td>0.0015</td>
<td>-0.0513</td>
<td>0.0011</td>
</tr>
<tr>
<td>AminoAcidSphereCount</td>
<td>-0.0398</td>
<td>0.0002</td>
<td>-0.0394</td>
<td>0.0001</td>
<td>-0.0401</td>
<td>0.0000</td>
</tr>
<tr>
<td>ICE</td>
<td>-1.3643</td>
<td>0.0510</td>
<td>-1.2489</td>
<td>0.0173</td>
<td>-1.4236</td>
<td>0.0125</td>
</tr>
<tr>
<td>AllCommunicatorICEColumn</td>
<td>0.1744</td>
<td>0.0095</td>
<td>0.1529</td>
<td>0.0032</td>
<td>0.1855</td>
<td>0.0023</td>
</tr>
<tr>
<td>cWWCount</td>
<td>0.5758</td>
<td>0.0072</td>
<td>0.5595</td>
<td>0.0025</td>
<td>0.5843</td>
<td>0.0018</td>
</tr>
<tr>
<td>DistanceFromCenter</td>
<td>0.0132</td>
<td>0.0001</td>
<td>0.0131</td>
<td>0.0000</td>
<td>0.0133</td>
<td>0.0000</td>
</tr>
<tr>
<td>SDI</td>
<td>-1.8918</td>
<td>0.1800</td>
<td>-1.4843</td>
<td>0.0612</td>
<td>-2.1011</td>
<td>0.0442</td>
</tr>
<tr>
<td>BRGeometry</td>
<td>-0.3978</td>
<td>0.0464</td>
<td>-0.5030</td>
<td>0.0158</td>
<td>-0.3438</td>
<td>0.0114</td>
</tr>
<tr>
<td>BPhGeometry</td>
<td>-0.3565</td>
<td>0.0620</td>
<td>-0.2162</td>
<td>0.0211</td>
<td>-0.4286</td>
<td>0.0152</td>
</tr>
</tbody>
</table>

Table 5.5: Regression coefficients and standard errors for the hierarchical model for *T. th.* grouped by chain

<table>
<thead>
<tr>
<th>Term</th>
<th>$\hat{\beta}_{5S}$</th>
<th>$\hat{\sigma}_{5S}$</th>
<th>$\hat{\beta}_{16S}$</th>
<th>$\hat{\sigma}_{16S}$</th>
<th>$\hat{\beta}_{23S}$</th>
<th>$\hat{\sigma}_{23S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.0795</td>
<td>0</td>
<td>-1.0795</td>
<td>0</td>
<td>-1.0795</td>
<td>0</td>
</tr>
<tr>
<td>NucleotideSphereNearPlaneCount</td>
<td>-0.0116</td>
<td>0.0211</td>
<td>-0.1012</td>
<td>0.0043</td>
<td>-0.0713</td>
<td>0.0025</td>
</tr>
<tr>
<td>SmallRNASphereNearPlaneCount</td>
<td>-0.0322</td>
<td>0.0040</td>
<td>-0.0258</td>
<td>0.0015</td>
<td>-0.0388</td>
<td>0.0013</td>
</tr>
<tr>
<td>AminoAcidSphereCount</td>
<td>-0.0337</td>
<td>0.0020</td>
<td>-0.0410</td>
<td>0.0004</td>
<td>-0.0399</td>
<td>0.0003</td>
</tr>
<tr>
<td>AllICE</td>
<td>-0.3535</td>
<td>0.0697</td>
<td>-0.5154</td>
<td>0.0102</td>
<td>-0.5810</td>
<td>0.0149</td>
</tr>
<tr>
<td>AllCommunicatorICEColumn</td>
<td>0.1017</td>
<td>0.0254</td>
<td>0.1926</td>
<td>0.0056</td>
<td>0.1832</td>
<td>0.0035</td>
</tr>
<tr>
<td>cWWCount</td>
<td>0.4800</td>
<td>0.0369</td>
<td>0.5513</td>
<td>0.0084</td>
<td>0.5802</td>
<td>0.0066</td>
</tr>
<tr>
<td>DistanceFromCenter</td>
<td>0.0067</td>
<td>0.0028</td>
<td>0.0131</td>
<td>0.0008</td>
<td>0.0157</td>
<td>0.0006</td>
</tr>
<tr>
<td>SDI</td>
<td>-0.2340</td>
<td>0.5891</td>
<td>-1.5066</td>
<td>0.1600</td>
<td>-2.1444</td>
<td>0.1312</td>
</tr>
<tr>
<td>BRGeometry</td>
<td>-0.2214</td>
<td>0.0285</td>
<td>-0.2875</td>
<td>0.0078</td>
<td>-0.3143</td>
<td>0.0061</td>
</tr>
<tr>
<td>BPhGeometry</td>
<td>0.0512</td>
<td>0.1107</td>
<td>-0.2060</td>
<td>0.0304</td>
<td>-0.3101</td>
<td>0.0237</td>
</tr>
</tbody>
</table>
terms, $\hat{\sigma}_j$, in Table 5.5, we see that the effect that these predictors have on site-specific mutation rate does not significantly vary by chains. However, there is some variation in the coefficients across the chains, which should be considered when using this model to predict the site-specific mutation rates for each organism by adjusting the parameter coefficient for the structure in which the nucleotide is found.

5.5 Application of factor analysis to the terms of the model

To better understand the terms of the original model above we use factor analysis to discern the commonalities of these terms that affect nucleotide mutation rate. Factor analysis is a technique that finds common factors and then expresses the variables in terms of these factors [35]. Mathematically this means that we find a number of factors for which each is expressed as a linear combination of the terms above. Factor analysis is the appropriate approach, more so than Principal Component Analysis, because we want to account for the commonality among the terms to gain an understanding about the similarity of the effect that they are having on site-specific mutation rate.

To begin the process of choosing the number of factors we will use, we employed the common technique used to obtain the dimensionality needed for the dataset of choosing the number of factors that have an eigenvalue larger than 1. There were four factors that have corresponding eigenvalues bigger than one, suggesting to us that the number of factors to be used was four. However, the eigenvalue corresponding to the fourth factor was 1.03, just barely reaching the criterion of having an eigenvalue bigger than 1. In addition, the only variable loading onto this factor was SmallRNASphereNearPlaneCount. Therefore we selected only three factors. We chose to use an oblique transformation. Since the transformations are invariant, we choose this transformation method of rotation to fit the variables to the factors well [35]. The results of the factor analysis are as follows. Running the factor
term 

<table>
<thead>
<tr>
<th>Term</th>
<th>E.c. factor 1</th>
<th>E.c. factor 2</th>
<th>E.c. factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NucleotideSphereNearPlaneCount</td>
<td>0.3307</td>
<td>0.3476</td>
<td>−0.5666</td>
</tr>
<tr>
<td>SmallRNASphereNearPlaneCount</td>
<td>0.1692</td>
<td>−0.1368</td>
<td>0.2435</td>
</tr>
<tr>
<td>AminoAcidSphereCount</td>
<td>−0.0897</td>
<td>−0.1791</td>
<td>−0.6518</td>
</tr>
<tr>
<td>ICE</td>
<td>−0.1382</td>
<td>0.8077</td>
<td>0.0640</td>
</tr>
<tr>
<td>AllCommunicatorICEColumn</td>
<td>0.1506</td>
<td>−0.8095</td>
<td>0.1203</td>
</tr>
<tr>
<td>cWWCount</td>
<td>−0.7735</td>
<td>−0.0734</td>
<td>0.0768</td>
</tr>
<tr>
<td>DistanceFromCenter</td>
<td>−0.0927</td>
<td>−0.0454</td>
<td>0.7791</td>
</tr>
<tr>
<td>SDI</td>
<td>0.5971</td>
<td>0.0910</td>
<td>0.1217</td>
</tr>
<tr>
<td>BRGeometry</td>
<td>0.3483</td>
<td>0.2808</td>
<td>−0.0503</td>
</tr>
<tr>
<td>BPhGeometry</td>
<td>0.6570</td>
<td>−0.2472</td>
<td>−0.0425</td>
</tr>
</tbody>
</table>

Table 5.6: Factor loading matrix from terms in the model for E.c.

<table>
<thead>
<tr>
<th>Term</th>
<th>T.th. factor 1</th>
<th>T.th. factor 2</th>
<th>T.th. factor 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>NucleotideSphereNearPlaneCount</td>
<td>0.2642</td>
<td>0.3370</td>
<td>0.6190</td>
</tr>
<tr>
<td>SmallRNASphereNearPlaneCount</td>
<td>0.1398</td>
<td>−0.0494</td>
<td>−0.3088</td>
</tr>
<tr>
<td>AminoAcidSphereCount</td>
<td>−0.0575</td>
<td>−0.0729</td>
<td>0.6845</td>
</tr>
<tr>
<td>AllICE</td>
<td>0.0490</td>
<td>0.9294</td>
<td>0.0343</td>
</tr>
<tr>
<td>AllCommunicatorICEColumn</td>
<td>−0.0375</td>
<td>−0.9031</td>
<td>−0.0886</td>
</tr>
<tr>
<td>cWWCount</td>
<td>−0.7378</td>
<td>−0.2274</td>
<td>−0.0098</td>
</tr>
<tr>
<td>DistanceFromCenter</td>
<td>−0.1523</td>
<td>−0.0259</td>
<td>−0.7487</td>
</tr>
<tr>
<td>SDI</td>
<td>0.6291</td>
<td>0.0048</td>
<td>−0.1180</td>
</tr>
<tr>
<td>BRGeometry</td>
<td>0.3320</td>
<td>0.2039</td>
<td>0.0867</td>
</tr>
<tr>
<td>BPhGeometry</td>
<td>0.6819</td>
<td>−0.2495</td>
<td>0.0588</td>
</tr>
</tbody>
</table>

Table 5.7: Factor loading matrix from terms in the model for T.th.

analysis on the E.c. and the T.th. model yielded the factor loading matrix of Table 5.6.

The common way to interpret a factor loading table is to assign variable to factors that have a loading greater than or equal to 0.30 in absolute value [35]. Based upon the loadings of the terms in the Table 5.6, we classify the factors as: local density, basepairing interaction geometry, and non-basparing interaction geometry. Local density contains the terms: NucleotideSphereNearPlaneCount, SmallRNASphereNearPlaneCount, AminoAcidSphereCount, and DistanceFromCenter. Non-basepairing interaction geometry contains the terms: cWWCount, SDI, BRGeometry, and BPhGeometry. Basepairing interaction geometry includes: ICE, and AllCommunicatorICEColumn.
5.6 Conclusions drawn from the models

From the models above, it can be inferred that the more complex the network of interactions a nucleotide is making then the slower is it expected to mutate, and that the effect of each of the types of interaction have a similar effect across the 3 chains of the ribosome. The regression model and the hierarchical model suggest that the mutation rate of a nucleotide is influenced greatly by three types of structural aspects. We see that these effects can be thought of largely as three major factors. In order of the importance observed from this study, these factors are how densely packed other units of the ribosome are that surround the nucleotide, the geometry of the basepairing interactions, and the geometry of the non-basepairing interactions of a nucleotide. From our analysis, we would expect to find the nucleotides that are mutating the slowest to be near the center of the ribosome, or near other rRNA nucleotides near the plane, or near the location where tRNA and mRNA come to interact with the ribosome. To a lesser effect, we would expect that the mutation rate of a nucleotide is lower based upon how hard it is to replace the nucleotide making the interaction while preserving the basepairing, stacking, and base-phosphate interaction. To this end, the harder the geometry of the interaction makes it to replace the nucleotide while preserving the interaction, the lower the expected mutation rate of the nucleotide. Conversely, one should expect to find nucleotides that have the highest mutation rate to be nearer the edge of the ribosome not making many interactions with other rRNA nucleotides, and not being near a protein or the place where tRNA or mRNA interact with the ribosome.
CHAPTER 6

Future Work

6.1 Developing new predictors

Several of the variables that we have used in this work are statistically significant, but not very biologically meaningful. For example the sphere-type variables that measure the local density of nucleotides and amino acids near a nucleotide are the most powerful predictors of site-specific mutation rates in the model, but do not provide any specific biological understanding as to why. The significance of this type of variable spurs interest as to what is the underlying biology preventing the nucleotides with high local density from mutating? Is there a better predictor that also contains more biological meaning? Sticking with the number of nucleotides in a sphere, another troublesome thing about sphere is the radius of 17 Å. This distance is troublesome because 17 Åis too large of a distance to be able to focus easily on one type of biological effect that may be affecting mutation-rates strongly. So clearly there is more to be discovered within the local region of the nucleotide.

After the sphere-type variables the count-type variables were most statistically significant for predicting site-specific mutation rates, but they too are not completely biologically meaningful. The count-type variables include cWWCount, noncWWCount, BaseCount, RiboseCount, StackingCount, and their near counterparts. It was seen in Chapter 5 that using
the count-type variables outperformed the more biologically meaningful variables. Yet this is surprising because many different types of interactions are grouped together in a “Count” variable. For example, there are 12 types of noncWW interactions. Our understanding of these interaction types allows for nucleotides in some types to mutate to a different nucleotide while preserving the interaction much more easily than do other types [46]. However, our predictor that accounts for this, ICE and cWWCount (to adjust for nucleotides only making a cWW interaction), still underperforms noncWWCount and cWWCount. We observe similar results for BPhGeometry compared to BaseCount as well as AllCommunicatorICE compared to AllCommunicatorCount, but we do not understand why.

Similarly we have, as a predictor, the distance that the nucleotide is from the center of the ribosome. This variable too is significant, and must be a proxy for some other more biologically more meaningful variable. Perhaps it is that nucleotides nearer the center of the ribosome are more “locked in” from their evolutionary history, and that nucleotides along the periphery are still trying to adapt to their environment. Whatever variable for which distance is a proxy remains a mystery at this time. While using the conserved core does make these proxy variables less significant, they remain in the model and as a proxy until more biologically meaningful variables can be found.

Next, and perhaps where much improvement can be gained, we have only begun to scratch the surface with looking at how the stacking of nucleotides affect the mutation rate. There is much literature giving the importance of the stacking of nucleotides and their relation to the folding of the RNA into its secondary structure [13], [54]. In [54], special attention is paid to the negative entropy of the stacking partners of the rRNA and how it can fold, drawing heavily from [13]. Mfold does very well at predicting secondary structure, and suggests that stacking is a large player in the ability of the ribosome to mutate. However we have not been able to find the aspects of stacking that relate it to mutation rates in a powerful way. The current way that we have is to look at the stacking interaction in the 3D structure and then to find the closest candidate stacking pair for each of the 16 possible stacking partners.
while maintaining the interaction type. Unfortunately this was not predictive either.

6.2 Further Estimation Mutation Rates

In this work, we laid the foundation of modeling site specific mutation rates, but there is still much work to be done in understanding how a ribosome mutates over time, and the algorithms employed to do the estimation. The first aspect upon which this work can be improved is the method that we use to compute the site specific mutation rates. This method does not depend upon the site specific rates being calculated in any order. Therefore these rates could be calculated in parallel rather than in serial. The benefit of this way of calculating the site specific rates is that we could calculate the site specific rates for other datasets in the future faster, or better if used in conjunction with a tree estimation program, which might lead to the creation of a better phylogenetic tree, or phylogenetic parameters. Second, there is previous work that suggests that a molecular clock is not the best model of nucleotide mutation [25], [26]. In future work, we would like to use the maximum likelihood estimation procedure to estimate very small trees in the full phylogenetic tree and attempt to estimate a more complete history of the mutation rate for any given site.

6.3 Distribution of nucleotides

While it is obvious that much is left unknown about the ribosome and the rate at which each nucleotide mutates over evolutionary time, there are other aspects of mutations that are equally unknown and warrant future attention. The first larger project that we consider is trying to estimate the probability of mutation of the nucleotide observed in a 3D structure to an A, C, G, or U. This project is of particular interest because it is the next logical step in modeling the mutation of the ribosome. We did some initial studies into predicting the entropy of the bases in each column of the alignment, defined as $\sum_{i \in \{A,C,G,U\}} p_i \log(p_i)$, where $p_i$ is the proportion of the bases that are of type $i$, and gaps were ignored. Then
using the same variables that we used to predict mutation rates, we predicted entropy. We believe that entropy is a logical place to start this project because it is a similar dependent variable to conservation percentage. If we gain an understanding about the way that the observed frequency of nucleotides appears in the columns of a multiple alignment this will build the foundation for more sophisticated estimation processes such as we did moving from conservation percentages to site-specific mutation rates. However none of the models that we fit to the observed entropy yielded predictive terms of consequence. The terms that predicted the best were those that were already geared toward a distribution of nucleotides such as ICE and BPhGeometry because underlying their respective computation is a matrix that has information about how easily a nucleotide can mutate to another nucleotide or stay the same by taking row or column sums as appropriate to gather a distribution of nucleotides, but we do not go on to take the standard deviation of the $1 \times 4$ vector. We refer the reader to Chapter 4 for more detail of how this information might be obtained.

The current working idea to make progress in this project is to instead build a model that yields the expected distribution for a particular site and then check that against the observed distribution for that site. We did preliminary work into this idea. Taking the same row summing procedure that was used in ICE and BPhGeometry, we then took the $1 \times 4$ vector from each of the two variables ICE and BPhGeometry and added them together weighted by their R squared value for predicting entropy with just that term and an intercept term. Since there was no multicollinearity problem with ICE and BPhGeometry for predicting entropy there would be no problem due to an interaction of these terms to weight the variables this way. Since NucleotideSphere was the only other to have some predictive power for entropy, we contrived a way to build an expected distribution from this variable. This way was to build a linear model of the proportion for each of nucleotides that remained the same, that changed to the isosteric mutation and to the other two nucleotides. We then took the normalized vector of expected probabilities of these nucleotides, weighted by the R squared contribution to entropy, and added it to the existing vector. The final distribution vector
was then normalized. Taking the MSE of the expected vector compared to the observed distribution we see that our model has much room for improvement as the MSE was 0.32 for the distribution vector at each site in the multiple alignment. If the weighting process was not used, then the MSE was 0.40 for the distribution vector at each site in the multiple alignment. This MSE is quite large since the total distribution adding to 1.0 will have a margin of error of 0.40 on average.

6.4 Estimation of the mutation rates of more than one nucleotide at a time

The final project that we will discuss is to extend the ideas in this work and in this chapter to the mutation of more than one nucleotide at a time. There is much literature that supports that nucleotides do not evolve just independently, but rather can evolve as a pair or as a system of more than two nucleotides. When models allow for the mutation of systems of nucleotides, they outperform models that are for independent evolution of nucleotides [52], [29]. Therefore further improvement lies with modeling the mutation of larger structures. The programs that we have written to accomplish the work of this dissertation are flexible enough to handle much larger structures, up to nine nucleotides mutating at once, but a meaningful parameterization of a larger generator matrix remains a challenge for us that we have yet to overcome.
BIBLIOGRAPHY


http://www.statmethods.net/stats/regression.html


