How much is Transcription-associated Mutagenesis Driving tRNA Microevolution?

Hector Banos Ling Wang Corinne Simonti Annalise Paaby
Christine Heitsch

March 26, 2025

Abstract

Transfer RNAs (tRNAs) are among the most highly conserved and frequently transcribed genes. Recent studies have identified transcription-associated mutagenesis (TAM) as a significant contributor to sequence variation around tRNA loci. However, the extent to which TAM drives allelic variation in tRNAs remains unclear, largely due to the confounding effects of strong selection pressures to maintain their structural integrity. This complexity arises because TAM-induced mutations primarily involve nucleotide transitions, which tend to preserve base-pairing stability.

To address this dichotomy at the population level, we analyzed tRNA allelic variation in contemporary *Caenorhabditis elegans* strains. We propose a model of tRNA microevolution driven by TAM and demonstrate that the observed secondary structure characteristics align with our predicted TAM-biased patterns. Furthermore, we developed a continuous Markov substitution model that incorporates TAM-specific mutational biases. This TAM-biased model fits the *C. elegans* tRNA data more effectively than standard models, such as the general time-reversible (GTR) model.

Based on these results, we conclude that TAM plays a significant role in shaping tRNA allelic variation within populations. This finding is consistent with recent experimental studies on tRNA fitness in yeast but challenges prior theoretical and computational analyses that emphasize RNA base-pairing as a primary determinant in genotype-phenotype systems.

1 Introduction

Transfer RNAs (tRNAs) are ancient and indispensable genes essential for protein synthesis, renowned for their evolutionary conservation (Tang et al., 2009; Zhang and Ferré-D'Amaré, 2016). This conservation is widely attributed to the structural integrity necessary for their function (Westhof et al., 2022), maintained by strong selection for base-pair preservation. As a result, the preservation of secondary structure is frequently used as a proxy for tRNA functionality, making tRNAs a classic model for exploring genotype-phenotype relationships.

However, the evolutionary forces shaping tRNA sequences remain poorly understood, particularly at the population level (Ishimura et al., 2014). While secondary structure preservation is assumed to dominate in macroevolutionary contexts, microevolutionary studies contradict this view. Recent *in vitro* experiments show that single mutations in loop regions can significantly affect fitness, while single mutations in the acceptor stem are surprisingly well-tolerated (Li et al., 2016). These findings challenge the thermodynamic model's assumption that loop regions experience significantly less selective pressure than paired positions (Aguirre et al., 2011; Fontana and Schuster, 1998; Reidys et al., 2001). This suggests that tRNA evolution is influenced by factors beyond structural preservation, including additional selective pressures or mutational biases.

One such bias, transcription-associated mutagenesis (TAM) (Jinks-Robertson and Bhagwat, 2014), has been identified as a major contributor to mutation rates in highly transcribed regions, including the nuclear tRNA flanking regions (Thornlow et al., 2020). However, the authors cautioned: *"it is difficult to discern whether this increased prevalence* [TAM's mutational signature] *is due to TAM or selection to preserve the structural integrity of the tRNA. To preserve tRNA secondary structure, we expect transition mutations to be more common than transversions, as they are less likely to disrupt stem helices."*

In this study, we address these challenges by analyzing tRNA genes in *Caenorhabditis elegans*. Using a novel continuous Markov substitution model, we distinguish TAM's mutational signature from selection for base-pair preservation. Aligned with the experimental results in (Domingo et al., 2018; Li et al., 2016), the data, supported by our model, shows that paired sites experience higher substitution frequencies than unpaired ones under TAM. Additionally, we observed that most substitutions, whether preserving base-pairing or not, decrease thermodynamic stability, suggesting that TAM compromises structural integrity rather than maintaining it. These results highlight how TAM shapes tRNA evolution in ways beyond structural constraints under the thermodynamic model.

2 Background

Markov models are widely used to study evolutionary processes, particularly DNA substitutions (Felsenstein, 2003; Yang, 2006). While these models are traditionally applied to gene tree inference across species, our approach departs from this by using them to describe allele formation within a population.

These models use different parameters to capture the various mutational mechanisms and selection pressures shaping DNA substitutions. Hence, the Markov exchangeability matrix for the substitution from i to j (denoted $i \to j$) for $i, j \in \{A, C, G, T\}$ may differ depending on the combination of nucleotides. In particular, transitions are expected to occur much more frequently than transversions since they preserve the purine ($A \leftrightarrow G$) versus pyrimidine ($C \leftrightarrow T$) distinction.

Moreover, for noncoding RNA genes, it is likely that transitions are further enhanced as they allow Watson-Crick base pairs G–C or C–G (denoted GC) and A–

U/U–A (AU)[a] to inter-convert via the wobble G–U/U–G (GU) pairing:

$$\begin{array}{ccccc} \mathrm{G} & & \mathrm{G} & \rightleftharpoons & \mathrm{A} \\ \mathrm{C} & \rightleftharpoons & \mathrm{U} & & \mathrm{U} \end{array} . \tag{1}$$

Note that of the three types of canonical base pairings, GC is the most thermodynamically favorable, followed by AU, and then GU. Hence, based on the pairing conservation, one would expect $C \to T$ transitions to be favored in GC pairs, $A \to G$ ones in AU, and both $G \to A$ and $T \to C$ in GU with the last as the most preferred, and the first as the least, due to thermodynamic stability.

Finally, for tRNA genes, it has been established (Thornlow et al., 2018) that TAM plays a significant role in the distribution of mutations in their flanking regions. The signature of this mutational bias is enhanced $C \to T$ and $G \to A$ transitions on the coding strand, denoted in red above, but at very different rates. A similar pattern was also seen in the tRNA genes themselves, but the extent to which this was TAM's mutational signature versus selection for pairing preservation remained uncertain. Disentangling these two factors is the focus of our results here.

2.1 Secondary structure conservation in RNA macroevolution

A noncoding RNA molecule like tRNA is composed of a single-stranded sequence which folds hierarchically (Tinoco and Bustamante, 1999) into a functional structure. The intra-sequence base pairings of the secondary structure are the critical scaffold which is then organized by tertiary interactions into the final 3D conformation. Crucially, it is the pattern of base pairing, and not necessarily the nucleotide content, that is evolutionarily conserved. This selection pressure is so strong that the occurrence of compensatory mutations in base pairs, identified under multiple sequence alignment, is the gold-standard for RNA structural inference (Cannone et al., 2002; Eddy and Durbin, 1994; Griffiths-Jones et al., 2005).

More precisely, a *secondary structure* for an RNA sequence R of length n is a set of paired indices $S(R) = \{(i, j) \mid 1 \leq i < j \leq n\}$ such that the nucleotides in positions i and j form a canonical base pair, i.e. either Watson-Crick (GC, AU) or wobble (GU). Suppose that R' is a *one-point mutant* of R, meaning that it differs from R by a single substitution. Suppose further that $S(R)$ is also a valid secondary structure for R'. Then either the substitution changed an unpaired nucleotide or it changed a paired one as described in Equation (1). In these cases we say that the substitution *preserves pairing potential*. Otherwise, it necessarily disrupts the pairing involving the mutated nucleotide.

It is critical to emphasize that a substitution being pairing potential preserving says nothing about the secondary structure of R'. The mutational constraint only requires that R' *can* assume the same set of pairings as $S(R)$, not that it *does*. It is, however, a necessary condition for R' to be a *neutral neighbor* of R.

Using RNA secondary structure formation as a model genotype-phenotype system (Cowperthwaite and Meyers, 2007; Schuster et al., 1994), the theory of neutral

[a]Uracil (U) will be used when discussing RNA structures and thymine (T) for their DNA genes.

networks was developed as a framework for understanding structural fitness landscapes. In this context, evolution is modeled as a series of one-point mutations, and a substitution is considered "neutral" if it does not alter the secondary structure. The conservation of pairing (and not just its potential preservation) is most often determined by free energy minimization (Zuker and Stiegler, 1981) under the nearest neighbor thermodyanmic model (NNTM). Hence R' is a neutral neighbor of R if it differs by a single substitution and retains the same minimum free energy (MFE) structure, i.e. the one-point mutation must be *MFE-preserving*.

Under this NNTM optimization interpretation of fitness neutrality, disruption of canonical base pairs is highly deleterious. Hence the macroevolution expectation of secondary structure conservation implies that substitutions should follow these patterns:

(U) A bias against mutations at paired sites.

(P1) A bias towards $C \to T$ in GC pairs and $A \to G$ in AU ones if a Watson-Crick pairing is mutated.

(P2) A bias towards transitions if a wobble pairing is mutated, with $T \to C$ clearly favored over $G \to A$.

2.2 Experimental insights into tRNA microevolution

Transfer RNA has one of the most direct genotype-phenotype relationships possible due to the tight coupling of sequence/structure/function. In order to read the mRNA codon and deliver the amino acid, a tRNA sequence must fold into a canonical three-dimensional 'L' shape. This essential structure is remarkably conserved across all three domains of life (Daewoo Pak and Burton, 2017). Although variations exist (Berg and Brandl, 2021), the 'cloverleaf' secondary structure is the core scaffold in this critical arrangement.

As illustrated[b] in Figure 1, the intra-sequence base pairing forms four runs of stacked base pairs, with three hairpin loops and the central multibranch loop. Three nucleotides in the anticodon loop target the complementary codon in the mRNA molecule while the corresponding amino acid, attached at the genomic tag, is loaded on the opposing acceptor stem. As pictured, the variable portion (M3) of the multiloop is unpaired, but it can sometimes be much longer (Berg and Brandl, 2021) in which case it typically includes pairings.

As a model genotype-phenotype system, tRNA folding has attracted considerable theoretical attention (Aguirre et al., 2011; Fontana and Schuster, 1998; Reidys et al., 2001) in macroevolution studies. These studies use secondary structure thermodyanmic predictions, such as the `RNAfold` function of the ViennaRNA 2.0 package (Lorenz et al., 2011) which outputs a single MFE structure, as a proxy for evolutionary fitness. In this neutral-neighbor context, any change in the base pairing pattern, i.e. the positions which are paired, is deemed deleterious.

More recently, however, tRNA has been the focus of *in vitro* experiments (Domingo et al., 2018; Li et al., 2016; Yona et al., 2013) which use sequencing technology to explore fitness at the microevolution scale. All three studies investigated a particular

[b]Sequence pictured is *Bacillus subtilis* tRNA^{Asp} with structures adapted from Liu (2013).

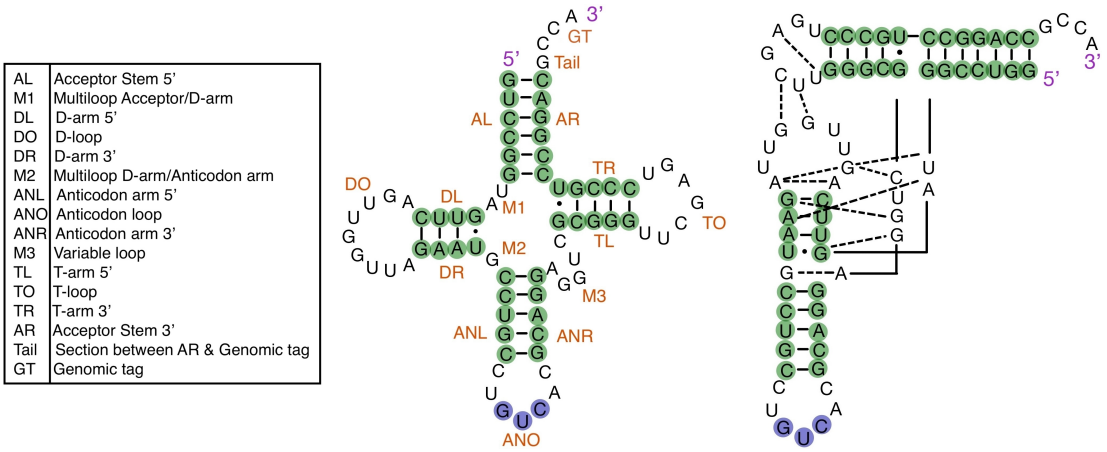


Figure 1: Representative tRNA secondary structure and tertiary interactions. From left to right: Table lists the structural components, including the genomic tag, which are then labeled on the secondary structure. The four-armed cloverleaf is closed by the acceptor (A) stem, and contains three hairpin stem-loop structures: the D, anticodon (AN), and T arms. Paired nucleotides (green) as well as the anticodon (blue) are highlighted. Most pairings are Watson-Crick (GC/AU, dash) but two wobble ones (GU, dot) are present. Various tertiary interactions (dashed lines) stabilize the 3D structure. The overall L-shape is critical to ribosome binding, and hence protein biosynthesis.

Saccharomyces cerevisiae Arginine gene: $\text{tRNA}^{\text{Arg}}_{\text{CCU}}$, which is distinctive[c] in having a single copy with anticodon CCU. Hence, because there is only one instance in the genome, changes in growth rate can be directly attributed to changes in this gene.

By deleting the gene entirely, Yona et al. (2013) demonstrated that within 200 generations the wild-type growth rate could be recovered. Investigating further they confirmed that this was achieved by mutating one of the 11 $\text{tRNA}^{\text{Arg}}_{\text{UCU}}$ copies available, and not always the same one. Hence, a single $C \to T$ substitution in the wobble position of the anticodon was sufficient to recover the wild-type fitness. Moreover, they then confirmed that such anti-codon switching can be found across all three domains of life, leading to the conclusion that this is a wide-spread adaptation mechanism. Hence, while the anticodon is generally expected to be under very strong selection pressure, it can and will evolve rapidly to meet new translation demands.

In contrast, Li et al. (2016) comprehensively characterized all one-point mutants of $\text{tRNA}^{\text{Arg}}_{\text{UCU}}$. They classified 1% as beneficial, 42% as deleterious, and the remaining majority (57%) as "(nearly) neutral." Generally, the least mutable positions were in loops, and the most in stems. As expected, mutations in the anticodon itself significantly decreased fitness, but so did ones in 3 T-loop locations as well as 2 T-stem and 1 D-stem positions. Conversely, the acceptor stem seemed to tolerate all single

[c]Although there are 20 standard amino acids and 64 distinct codons, most eukaryotic nuclear genomes encode hundreds of tRNAs.

substitutions, as did the variable loop.

This robustness to one-point mutants in key structural components was echoed by Domingo et al. (2018) who investigated all existing variants of $\text{tRNA}^{\text{Arg}}_{\text{UCU}}$ found across yeast species. They identified 14 mutations across 10 sites in 6 other genomes: 11 in the acceptor stem (5 in AL, 6 in AR) and 2 in the variable loop (M3) with 1 from the anticodon stem (ANL). All but one of the AL/AR mutations came together in compensatory pairs, of which 1 was a GU, 3 AU, and 1 GC. The remaining AR mutation formed a GU pair as did the ANL one. They then generated all possible combinations of these substitutions and tested their fitness in growth competition. All 14 one-point mutants exhibited fitness comparable to the wild-type *S. cerevisiae* gene and the extant variants.

Both studies found that fitness decreased rapidly with subsequent one-point mutations, with significant epistatic interactions also identified. The epistasis had a strong negative bias except — unsurprisingly — for compensatory mutations in paired positions. All of this suggests that, while conservation of secondary structure is critical at the macroevolutionary level, the situation may be much more nuanced and flexible across a population.

2.3 Markov models for DNA macroevolution

DNA sequences evolving in time under nucleotide substitutions are typically modeled using a continuous-time Markov process. These models are parameterized by a nucleotide distribution vector $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ and 4×4 instantaneous rate matrix Q. Then the joint distribution of nucleotides in the ancestral and descendant sequences over time t is given by $P = \text{diag}(\pi)\exp(Qt)$ where the entry p_{ij} represents the probability of observing a substitution $i \to j$ at a site.

In general, the models are distinguished by the constraints on π and Q. All forms require that $\sum \pi_i = 1$ with $\pi_i \geq 0$, that $q_{ij} > 0$ for $i \neq j$, and that $q_{ii} = -\sum_{i \neq j} q_{ij}$. In other words, π must be a probability distribution, all off-diagonal entries of Q are positive, and the diagonal ones make the row sums equal zero. Additionally, the rate matrix used is typically normalized by multiplying all original entries of Q by

$$\left(-\sum_{i=1}^{4} q_{ii}\pi_i\right)^{-1}.$$

This results in the expected number of substitutions being one per unit of time. The non-normalized form of Q is called an exchangeability matrix.

The earliest, and most mathematically tractable models, assume that π is uniform, and the 12 free entries of Q come from a set of either 1, 2 or 3 parameters (Jukes and Cantor, 1969; Kimura, 1980). When there are two, one is for transitions (which preserve the purine/pyrimidine division) and the other for transversions. The addition of a third parameter is used to distinguish transversions which are amino/keto preserving from the weak/strong symmetry.

More biologically realistic models allow nonuniform π and provide a larger number of exchangeability parameters. The most general form, associated with the *continuous*

general Markov model (GM), has no additional constraints, and hence 12 free parameters. However, this great flexibility often results in over-fitting, and is seldom needed to capture the substitution process accurately.

The most common substitution model used for phylogenetic inference is the *general time-reversible* (*GTR*) one (Tavaré, 1986). Its rate matrix has 6 free parameters satisfying $\mathrm{diag}(\pi)Q = Q^T\mathrm{diag}(\pi)$. This equality holds when the substitution process has the same parameters going forward in time as backward. When π is uniform, the matrix is symmetric; otherwise $q_{ij}/q_{ji} = \pi_j/\pi_i$. As useful as the assumption of time-reversibility is for macroevolution analyses, it is not clear that it will be appropriate for the *C. elegans* tRNA population data. Although we considered *GTR* as a possibility, our TAM-biased Markov substitution model is built from the strand-symmetric rate matrix. We note that a model can be both strand-symmetric and time-reversible only if $\pi_A = \pi_T$ and $\pi_C = \pi_G$.

The strand-symmetric (*sym*) model (Casanellas and Sullivant, 2005) also has 6 free parameters in its rate matrix. However, the equality now enforced is on the exchangeability of $i \to j$ and its Watson-Crick complement, i.e. $C \to T$ and $G \to A$ must have the same parameter. The premise is that, for many mutational mechanisms, it is not possible to determine whether a substitution initially occurred in the coding strand or the template one. We consider such mutations as background noise, and take this matrix as the foundational one for our new TAM-biased model.

2.4 Microevolution signature of mutational bias

Transfer RNAs are not only highly conserved but also highly transcribed. As such (Thornlow et al., 2018), they are subject to a process called *transcription-associated mutagenesis* (TAM). Critically, TAM's characteristic signature breaks the strand-symmetry assumption, as well as the time-reversibility one.

During RNA transcription, a hybrid DNA–RNA complex forms, leaving the non-template DNA strand exposed and vulnerable to mutagens (Jinks-Robertson and Bhagwat, 2014). The formation of non-canonical structures by the unwound DNA further promotes mutagenesis by inhibiting repair pathways and interfering with replication processes (Gaillard and Aguilera, 2016). Thus, the more frequently a gene is transcribed, the more vulnerable its coding strand becomes to mutations.

The characteristic signature for TAM is a mutational bias associated with activation-induced cytidine deaminase. In particular, TAM is identified with deamination of the noncoding strand, along with DNA repair mechanisms responding to this deamination. Consequently, an increase in $C \to T$ and $G \to A$ mutations on the coding strand is expected (Green et al., 2003; Jinks-Robertson and Bhagwat, 2014) — with the former occurring at a significantly higher rate.

Thornlow et al. (2018) observed that the 20 bases upstream/downstream of human tRNA genes are significantly more divergent than untranscribed reference regions. By analyzing these deviations, they identified that TAM's mutational signature is especially prevalent in tRNA flanking regions. They also detected a similar pattern, although at a much lower frequency, within tRNAs themselves. However, it was not clear the extent to which this was due to TAM versus selection for pairing preservation.

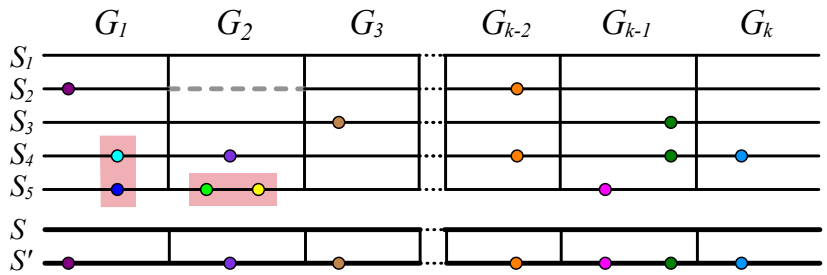


Figure 2: Schematic of mutational variation sequence construction. From top to bottom: Each line S_i represents the k concatenated genes G_j for an individual i in the population $N = \{1, \ldots, 5\}$. Every gene is not necessarily present in each individual; dashed gray lines denote a missing gene. Colored dots indicate nucleotide substitutions with respect to the consensus sequence S (shown below S_5). Highlighted in red, we see that G_2 in S_5 violates condition (i) and G_1 violates (ii) in S_4 and S_5. By treating alleles violating these two conditions as missing genes, the remaining allelic variation can be encoded in the new compound sequence S' by substituting into S each distinct-site one-point mutation.

3 Materials and Methods

Our results demonstrate that the microevolution signature for TAM can be distinguished from the macroevolution selection pressure to preserve pairings. To do this, we introduce a continuous Markov model that describes the distribution of alleles under mutational bias across a population. We then apply this new model to analyze the distribution of tRNA genes across extant *C. elegans* strains.

Typically (Felsenstein, 2003; Yang, 2006), Markov models are used for phylogenetic inference across species. This is effective since the populations have diverged enough that it is possible to reconstruct their evolutionary relationships in a gene tree. In contrast, we seek not to characterize the relationship among the *C. elegans* alleles but only to describe the distribution of mutations. In doing so, we show that they are consistent with our microevolution TAM-bias substitution model but not with macroevolution expectations of thermodynamic neutrality for secondary structure conservation.

3.1 Model development

Our model formulation is possible under a strong purifying selection and high mutation rate over a short evolutionary timescale — exactly the situation for tRNA microevolution. As we show below, these conditions allow us to capture nearly all of the observed allelic variation via a continuous Markov substitution process.

3.1.1 Modeling allele formation within a population

Let N be a population of n individuals and, for each individual $i \in \{1, 2, ..., n\}$, let S_i be the concatenated collection of k genes under similar selection pressure, e.g. all tRNAs in extant *C. elegans* strains. Let $\mathcal{S}$ be the corresponding sequence for the most recent common ancestor of N, and denote the consensus sequence for the S_i as S.

We assume that the k genes are highly conserved and the evolutionary time from $\mathcal{S}$ is relatively short. Thus, the sequences S_i are all close to each other, and also to $\mathcal{S}$. In this scenario, although not necessarily others (Czech et al., 2018; Trudeau et al., 2016), it is reasonable to approximate $\mathcal{S}$ by S.

To formulate our microevolution population model, we impose the following two requirements. Moreover, we show in *Results* that, for the tRNA microevolution situation being considered, these conditions are actually fairly mild.

(i) An allele differs from S by one nucleotide substitution.

(ii) Two alleles in distinct individuals cannot differ in the same site.

Alleles meeting (i) are, by definition, one-point mutants, and are often the individual steps in macroevolution models (Saks et al., 1998). We refer to those satisfying (ii) as *distinct-site* mutants. Under these two conditions, we can encode all mutational variation from the compliant alleles in a single compound sequence.

Specifically, we introduce the *mutational variation sequence* S' which is constructed by incorporating all observed substitutions in alleles meeting conditions (i) and (ii) into the consensus sequence S. As illustrated in Figure 2, this guarantees a one-to-one correspondence between sites where S' differs from S and all distinct-site, one-point mutations observed in the population.

By integrating all those mutations into a single sequence, we can model allele formation without requiring a phylogenetic tree. This is particularly useful as inferring the true evolutionary relationship between individuals within a population can be very challenging (De Maio et al., 2015). Instead, we describe it simply through a nucleotide substitution process from S to S'.

Moreover, since multiple mutations within the same gene occur in different alleles, we strengthen the site independence for the Markov model. Hence, we parameterize our model with a 4×4 rate matrix for nucleotide substitutions, rather than the more complicated dinucleotide ones (Allen and Whelan, 2014) which model RNA pairing covariation for macroevolution analyses.

3.1.2 Modeling substitution domain bias under TAM

As described in the *Background*, TAM's characteristic signature results in an asymmetric increase in $C \to T$ and $G \to A$ mutations on the coding strand. We hypothesize that the exchangeability asymmetry caused by TAM results in a substitution domain bias, which we call the *TAM bias*.

To formulate this mathematically, let $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ be the nucleotide distribution of the consensus sequence S (an approximation of $\mathcal{S}$, the ancestral sequence). With no mutational bias, one would expect that the substitution domains follow π.

To adjust for TAM, we introduce the *TAM bias vector* $\tau = (\tau_R, \tau_Y) \in [0, 1]^2$. We hypothesize that the nucleotides in S that mutated to form S' follow the *TAM bias*

distribution given by:

$$\pi^\tau = (\pi^\tau_A, \pi^\tau_C, \pi^\tau_G, \pi^\tau_T), \tag{2}$$

where

$$\pi^\tau_A = \pi_A - \tau_R, \quad \pi^\tau_C = \pi_C + \tau_Y, \quad \pi^\tau_G = \pi_G + \tau_R, \quad \text{and} \quad \pi^\tau_T = \pi_T - \tau_Y.$$

In other words, the TAM bias vector τ adjusts the substitution domain distribution to account for differences in transition rates within purines (R) and within pyrimidines (Y). Hence, π^τ_A is the proportion of A's in S that mutated to another nucleotide to form S' relative to all observed distinct-site, one-point substitutions.

3.1.3 Modeling a TAM-biased substitution process

Our proposed TAM-biased distribution describes how the substitution domains are altered from the root nucleotide distribution in S. However, to model allele formation within a population under TAM, we need to account for the mutation process itself. For this, we present a continuous Markov substitution model that incorporates the TAM domain bias to describe the production of S' from S.

Recall that for a Markov process over an edge of length t, the joint distribution of nucleotides in the ancestral and descendant sequences is given by $P = \text{diag}(\pi)\exp(Qt)$. In our context, the entry p_{ij} of P represents the probability of a site in S' differing from S by a substitution from i to j. To parameterize this new model P, we must propose a TAM-biased instantaneous rate matrix Q. We note that the rate matrices are presented in their exchangeability form for simplicity, but all computations are done under normalization as described in the *Background*.

We begin with a strand-symmetric model (*sym*) which assumes that a substitution is indistinguishable from its Watson-Crick complement. This accounts for the background noise, and its matrix has 6 free parameters:

$$Q_{sym} = \begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & * & \alpha & \beta & \gamma \\ \text{C} & \delta & * & \epsilon & \boxed{\lambda} \\ \text{G} & \boxed{\lambda} & \epsilon & * & \delta \\ \text{T} & \gamma & \beta & \alpha & * \\ \hline \end{array} \tag{3}$$

where $\alpha, \beta, \gamma, \delta, \epsilon, \lambda \geq 0$.

However, a characteristic of TAM is the asymmetry in transitions observed in the coding strand. Hence, we break the strand symmetry between substitutions $A \to G$ with $T \to C$ (circled in Q_{sym}), and $G \to A$ with $C \to T$ (squared in Q_{sym}), yielding a

transition-asymmetric rate matrix

$$Q_{tran} = \begin{array}{c|cccc} & \text{A} & \text{C} & \text{G} & \text{T} \\ \hline \text{A} & * & \alpha & \beta & \gamma \\ \text{C} & \delta & * & \epsilon & \boxed{\lambda} \\ \text{G} & \boxed{\theta} & \epsilon & * & \delta \\ \text{T} & \gamma & \mu & \alpha & * \end{array} \tag{4}$$

with 8 free parameters. Hence, Q_{sym} is a special case of Q_{tran} with $\theta = \lambda$ and $\mu = \beta$.

Next, we incorporate the TAM bias model for the domain distribution by adding restrictions to the Q_{tran} rate matrix entries. More specifically, we require that

$$\pi_A - \tau_R = \frac{s(1)}{Z},\ \pi_C + \tau_Y = \frac{s(2)}{Z},\ \pi_G + \tau_R = \frac{s(3)}{Z}, \text{and } \pi_T - \tau_Y = \frac{s(4)}{Z} \tag{5}$$

where $s(i) = \sum_{j \neq i} p_{ij}$ is the sum of the off-diagonal entries in each row of P with $Z = \sum_{i \neq j} p_{ij}$ as the normalizing factor. Thus, the left-hand sides come from Equation (2), whereas the right ones follow from the continuous Markov model.

We remark that these equations do not increase the number of model parameters. The entries of π are needed to compute P for every rate matrix, and hence do not count against any particular model. Additionally, the TAM bias vector $\tau = (\tau_R, \tau_Y)$ is estimated from the flanking regions, which are subject to TAM but not strong selection pressure. Hence, since τ_R and τ_Y are independent from the tRNA alleles, they are not counted as model parameters either.

As a consequence, we improve the model parameterization under the constraints in Equations (5) by identifying linear dependencies among the Q_{tran} rate matrix entries using a computer algebra package. More precisely, given the high conservation of tRNAs, we expect a very slow substitution process, implying that the entries of Qt are quite small. Thus, we approximate $\exp(Qt) \approx I + Qt$, where I is the 4×4 identity matrix. Under this approximation, we represent Equations (5) as a homogeneous linear system which can be solved using Macaulay2 (Grayson and Stillman, 2009).

As a result, we obtain the linear dependencies given below for the transversions δ and γ as well as the transition θ. We then define Q_{TAM} to be the rate matrix obtained from Q_{tran} using these relationships. This yields a total of 5 free parameters for the new TAM-biased rate matrix: the 3 transitions β, λ, and μ along with the 2 transversions α and ϵ.

We note that the choice of dependent variables among the matrix parameters was determined by Macaulay2; any alternate formulation of the relationships would yield an equivalent model parameterization although a different form for the Q_{TAM} rate matrix. The linear dependencies used here (which were manually derived from Macaulay2's initial output shown in the Appendix) are:

$$\begin{aligned} \delta &= -\epsilon - \lambda + (\mu - \beta)\Omega \\ \theta &= \lambda + (\mu - \beta)\Theta \\ \gamma &= -\alpha + \mu\Gamma + \beta\Psi \end{aligned}$$

where

$$\xi = \Lambda_4 - \Lambda_1 \text{ and } \Omega = \frac{-1-\Lambda_2}{\xi},\ \Theta = \frac{\Lambda_2 - \Lambda_3}{\xi},\ \Gamma = \frac{\Lambda_1 - 1}{\xi},\ \Psi = \frac{1-\Lambda_4}{\xi}$$

for

$$\Lambda_i = \frac{\tau(i)}{\pi_i} \text{ with } \tau(i) = \begin{cases} \tau_R & \text{if } i \text{ is odd} \\ \tau_Y & \text{if } i \text{ is even} \end{cases}.$$

Observe that the coefficients Ω, Θ, Γ, and Ψ are real numbers defined solely by the TAM bias vector τ and the root nucleotide distribution π. As a function of π and τ, they are unbounded and defined almost everywhere (except for a set of measure zero, that is, a negligible subset of the parameter space).

We remark that $\Omega \neq 0$. Hence, the inequality $\mu \neq \beta$ must hold; otherwise $\delta < 0$, which violates the definition. Likewise $\Theta \neq 0$ for almost all π and τ. But since we have already concluded that $\mu \neq \beta$ when Ω is defined, then generically $\theta \neq \lambda$ as well. This shows that satisfying Equations (5) requires breaking the transition strand symmetries.

We show in *Results* that among the five rate matrices considered — strand-symmetric (*sym*), transition-asymmetric (*tran*), and TAM biased (*TAM*) along with general Markov (*GM*) and general time-reversible (*GTR*) — the new TAM-biased Markov substitution model best fits the allelic distribution in our *C. elegans* tRNA dataset.

3.2 Data analysis

Our dataset was generated from two publicly available databases: the Genomic tRNA Database (GtRNAdb, v2.0) (Chan and Lowe, 2008, 2015) and the *C. elegans* Natural Diversity Resource (CeNDR, release 20180527) (Cook et al., 2016). The starting point was the reference genome (release WBcel235/ce11) (Harris et al., 2019) for the universal laboratory strain N2.

GtRNAdb identifies 581 functional tRNAs in the N2 nuclear genome, and CeNDR provided variant data for 330 *C. elegans* wild isolates collected from around the globe. We then retrieved 677 single nucleotide variants (SNVs) for the tRNA genes from the hard-filter variant call format (vcf) file using custom Python scripts.

Secondary structures for all tRNA sequences were obtained from `tRNAscan-SE` (Chan and Lowe, 2019) using the default parameters. As illustrated in Figure 1, nucleotides were classified as "paired" or "unpaired" according to these structures. We note that, although pictured, the genomic tag was not included in the analysis.

There were 16 SNVs which were found by `tRNAscan-SE` to occur in an intron; these were excluded from further analysis. Hence our *C. elegans* microevolution dataset consists of 661 variants for 581 tRNA genes across a population of 331 strains.

We also collected data for the 40 nucleotides upstream and 40 downstream of each SNV and N2 gene. Since these tRNA flanking regions are unwound during transcription, they are subject to TAM (Simonti et al., 2020; Thornlow et al., 2018). However, they are not under strong selection pressure, making them ideal for estimating the TAM bias vector $\tau = (\tau_R, \tau_Y)$. Under our TAM bias hypothesis, subtracting the distribution of observed substitution domains from the initial nucleotide distribution should yield $\pi - \pi^\tau = (\tau_R, -\tau_Y, -\tau_R, \tau_Y)$. As reported in *Results*, the positive and negative values are in good agreement, yielding a robust estimate for τ.

3.2.1 Testing TAM-biased distributions

Our model for TAM bias from Equation (2) proposes that the substitution domains are distributed as π^τ where $\pi^\tau_A = \pi_A - \tau_R$, $\pi^\tau_C = \pi_C + \tau_Y$, etc. To validate this hypothesis for the TAM-biased rate matrix, we compute a p-value using the `xmonte` function from the XNomial R package (Engels, 2015). This function assesses whether the observed substitutions conform to the estimated multinomial distribution. Specifically, it examines 3,000 random outcomes to estimate the probability that a sample deviates from the expected distribution by at least as much as the observed data.

We later apply the same approach to test for the preservation of pairing potential. To do this, we classify the nucleotides of each tRNA gene as either paired (p) or unpaired (u) according to the `tRNAscan-SE` secondary structure. This separates π, the nucleotide distribution for S, by pairedness into

$$\Pi = \begin{pmatrix} A_p & C_p & G_p & T_p \\ A_u & C_u & G_u & T_u \end{pmatrix} \tag{6}$$

where X_p and X_u denote the proportion of paired and unpaired nucleotides respectively for $X \in \{A, C, G, T\}$. Marginalizing the rows of Π reconstitutes π, while column marginalization yields the pairedness distribution, denoted $\Pi_{pu} = (\Pi_p, \Pi_u)$.

We also consider the distribution of base pairs, denoted $\rho = (\rho_{GC}, \rho_{AT}, \rho_{GT})$, where ρ_{GT} is the proportion of GU pairings in the tRNA secondary structures. It follows that

$$\rho_{GC} = \frac{C_p}{G_p + A_p}, \quad \rho_{AT} = \frac{A_p}{G_p + A_p}, \quad \text{and} \quad \rho_{GT} = \frac{G_p - C_p}{G_p + A_p} = \frac{T_p - A_p}{G_p + A_p}.$$

Under our TAM bias hypothesis, the distribution of nucleotides by pairedness for the substitution domains would be

$$\Pi^\tau = \begin{pmatrix} A^\tau_p & C^\tau_p & G^\tau_p & T^\tau_p \\ A^\tau_u & C^\tau_u & G^\tau_u & T^\tau_u \end{pmatrix} = \begin{pmatrix} \pi^\tau_A \left(\frac{A_p}{\pi_A}\right) & \pi^\tau_C \left(\frac{C_p}{\pi_C}\right) & \pi^\tau_G \left(\frac{G_p}{\pi_G}\right) & \pi^\tau_T \left(\frac{T_p}{\pi_T}\right) \\ \pi^\tau_A \left(\frac{A_u}{\pi_A}\right) & \pi^\tau_C \left(\frac{C_u}{\pi_C}\right) & \pi^\tau_G \left(\frac{G_u}{\pi_G}\right) & \pi^\tau_T \left(\frac{T_u}{\pi_T}\right) \end{pmatrix} \tag{7}$$

where, for instance, A^τ_p is the proportion of one-point mutants in S' that differ from S by a substitution which changed a paired A. Row and column marginalization of Π^τ yield π^τ and $\Pi^\tau_{pu} = (\Pi^\tau_p, \Pi^\tau_u)$ respectively.

Analogously, ρ^τ_{GT} represents the proportion of substitutions that changed either the G or the T in what would have been a wobble GU pairing. Consequently, the components of ρ^τ are

$$\rho^\tau_{GC} = C^\tau_p + G^\tau_p \left(\frac{\rho_{GC}}{\rho_{GC} + \rho_{GT}}\right), \quad \rho^\tau_{AT} = A^\tau_p + T^\tau_p \left(\frac{\rho_{AT}}{\rho_{AT} + \rho_{GT}}\right), \quad \text{and} \tag{8}$$
$$\rho^\tau_{GT} = G^\tau_p \left(\frac{\rho_{GT}}{\rho_{GC} + \rho_{GT}}\right) + T^\tau_p \left(\frac{\rho_{GT}}{\rho_{AT} + \rho_{GT}}\right).$$

3.2.2 Assessing Markov substitution rate matrices

To compare the five possible exchangeability matrices considered, we computed for each the three most common information criteria for model selection (Kalyaanamoor-

thy et al., 2017): the Akaike (AIC) (Akaike, 1974), the corrected Akaike (AICc) (Burnham and Anderson, 2002), and the Bayesian (BIC) (Schwarz, 1978). We found that the differences between the first two were always negligible due to the length of the consensus sequence relative to the number of mutations. Hence, we focus on the AIC and BIC.

Loosely speaking, the lower the estimate, the 'better' the model. Each criterion depends on the log-likelihood (LH) of the model being 'penalized' by the number of parameters. The LH function is given by

$$L(Q,t) = \sum_{i=1}^{n} \log P(\boldsymbol{x}_i|\pi, Q, t),$$

where π is the nucleotide distribution of S, n is the sequence length of S, $\boldsymbol{x}_i$ is the site pattern of S and S' at site i, Q is the rate matrix, t are the expected number of substitutions between S and S', and $P(\boldsymbol{x}_i|, \pi, Q, t)$ is the conditional probability of $\boldsymbol{x}_i$ given π, Q, and t.

To compute the LH function and estimate the maximum likelihood estimator (MLE) for all parameters, we again use the approximation $\exp(Qt) \approx I + Qt$. We then used MATLAB's (MathWorks, 2024) `fmincon` function to estimate Q and t MLEs using 1000 different random starting points.

We also computed $\hat{P} = \mathrm{diag}(\pi)\exp(\hat{Q}t)$ using the MLE estimates for each parameterization and compared the results to the data. The corresponding p-values for such comparisons were then obtained using the `xmonte` function.

3.2.3 Analyzing thermodynamic neutrality

We also compared observed distributions against a background of neutral neighbors. Recall that these are alleles which are both one-point mutants and MFE-preserving. For consistency, we consider only observed alleles which are also neutral neighbors. The necessary thermodynamic predictions, i.e. the MFE score and its associated secondary structure, are computed with `RNAfold` (Lorenz et al., 2011) under default parameters.

More precisely, let P be the set of all distinct-site, one-point mutant alleles used to generate S', and R the set of tRNA genes in S. For $p \in P$, let $\alpha(p)$ be its corresponding $r \in R$. Then $p \in P$ is MFE-preserving, and hence is a neutral neighbor, when it has the same `RNAfold` predicted secondary structure as $\alpha(p)$.

Let $M \subseteq P$ be MFE-preserving, i.e. all observed neutral neighbors. For $m \in M$, let N_m be the set of all neutral neighbors of $\alpha(m)$, and $\mathcal{N} = \bigcup_{m \in M} N_m$. Since $m \in N_m$, M is a subset of $\mathcal{N}$. We determine M and $\mathcal{N}$ computationally, and consider how they differ.

First we consider whether the domains were paired or not, according to the classification from the `tRNAscan-SE` secondary structure for $\alpha(m)$. Next we consider the thermodynamic stability of m and $\alpha(m)$ relative to N_m over all of M.

This is done by ranking the set $N_m \cup \{\alpha(m)\}$ by increasing MFE score with the lowest ranked as 1. The ranking is dense (García-Lapresta and Martínez-Panero, 2024); ties share the same rank, and the next rank is assigned the next integer. For instance, if three sequences tie for the 5th lowest MFE score, all are assigned rank 5 with the

next one(s) ranked 6. Rankings are then normalized by dividing by the total number of ranks for $\alpha(m)$. This defines a function $f_m : N_m \cup \{\alpha(m)\} \to (0, 1]$ for each $m \in M$ where values near 0 are the most thermodynamically stable while those near 1 are the least so. We will compare $f_m(m)$ and $f_m(\alpha(m))$ over all of M.

Finally, we consider the positional distribution of substitutions in M and $\mathcal{N}$. The tRNA structural regions are defined in Figure 1 and determined for $\alpha(m)$ based on the `tRNAscan-SE` secondary structure.

3.2.4 Comparing with experimental fitness

We consider the mutability of each site across the 531 alleles in our *C. elegans* dataset. Since not all tRNAs have the same length, the sites in each allele were normalized to the average length of the corresponding substructure component from the 15 (excluding the genomic tag) listed in the table in Figure 1. The number of mutations for each normalized site can then be presented in a frequency histogram.

We compared this distribution to the mean site-fitness scores obtained by Li et al. (2016) for the *S. cerevisiae* single-copy arginine-CCU tRNA. The sites in the tRNA_{Arg}^{CCU} were also normalized as above before reporting the average of their 3 one-point mutant fitness scores. For display purposes, this mean value was linearly re-scaled from (0.685,1.017) to (1,10).

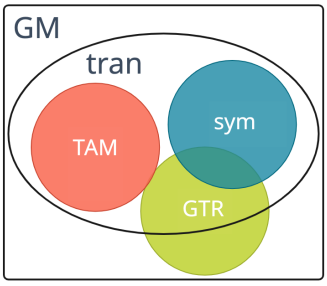
4 Results and Discussion

4.1 Model validation

We find good agreement between model hypotheses and population data. Specifically, the new mutational variation sequence S' reasonably describes the observed tRNA distribution, and the consensus sequence S is a suitable approximation to the most recent common ancestor of those *C. elegans* alleles. Moreover, the flanking regions yield a robust estimate for the TAM bias vector τ, and the tRNA substitution domains align with π^τ, the proposed TAM biased distribution. Finally, of the Markov substitution models considered, the new *TAM* matrix best describes the data.

Our full dataset consisted of 661 variants for 581 tRNA genes. Of those alleles, 70 have an indel and cannot be described by a substitution model. Among the 591 remaining, no two differ at the same position, but 60 have two or more substitutions. Hence, there are 531 distinct-site, one-point mutant alleles, and the compound sequence S' encodes 80.3% of the total tRNA mutational variation in the *C. elegans* population.

We consider the evolution of S' from S, the consensus sequence for the 531 genes. This is a good approximation to their most recent common ancestor $\mathcal{S}$ because the distribution of alleles is very sparse. In particular, nearly all consensus tRNA ($484/531 = 91.1\%$) occur in nearly all strains (at least $298/331$ so $\geq 90.0\%$). The remaining 47 consensus sequences have more frequently occurring variants, including missing entirely. However, of the 45 present in $\geq 45\%$ of strains, only 2 have a variant which occurs nearly as often ($< 15\%$ difference). Hence, in all but 4 tRNA genes, the consensus clearly dominates.



	TAM	*tran*	*sym*	*GM*	*GTR*
free parameters	5	8	6	12	6
p-value	0.63	0.82	0.39	1	0
AIC difference	0	2	19	3	117
BIC difference	0	28	28	63	125

Figure 3: Comparison of rate matrix parameterizations for Markov substitution model: general Markov (*GM*), strand-symmetric (*sym*), transition-asymmetric (*tran*), and TAM-biased (*TAM*) along with general time-reversible (*GTR*). The Akaike and Bayesian information criteria (AIC, BIC) are reported as the difference from *TAM*'s score as this was the lowest, i.e. best.

The TAM bias vector τ was estimated from the tRNA flanking regions. They are subject to TAM, but not the same selection pressure, and contained 2070 mutations. This yielded $\pi - \pi^\tau = (0.0454, -0.1111, -0.0456, 0.1110) = (\tau_R, -\tau_Y, -\tau_R, \tau_Y)$. The close match in positive and negative estimates supports our TAM bias hypothesis; averaging yields $\hat{\tau} = (0.0455, 0.1110)$.

The nucleotides A/C/G/T in S are distributed as $\pi = (0.181, 0.258, 0.323, 0.237)$[d]. Under our TAM bias model, the expected distribution for substitution domains is $\pi^\tau = (0.135, 0.370, 0.369, 0.125)$ while $\hat{\pi}^\tau = (0.117, 0.369, 0.375, 0.139)$ is observed. Under statistical testing, described in *Materials and Methods*, we fail to reject the null hypothesis with a p-value of 0.510 whereas repeating the analysis with $\tau = (0, 0)$ yields $p << 0.001$. Hence we conclude that the tRNA substitution domains are described by our proposed TAM bias distribution.

This domain distribution is then used to parameterize a new rate matrix, denoted *TAM*, for the continuous Markov substitution process model. As illustrated in Figure 3, *TAM* is a special case of *tran*, which was introduced as a generalization of the standard *sym*. We also consider the full *GM* as well as the frequently-used *GTR*.

The calculated p-values, along with other model selection criteria, are also given in Figure 3. With $p << 0.001$, it is clear that *GTR* does not fit the data. Having rejected the null hypothesis, and in view of the significantly lower AIC and BIC scores, we conclude that time-reversibility — a common macroevolution assumption — is not an appropriate expectation for this microevolution model.

Among the remaining four matrices with $p > 0.05$, the model selection criteria support *TAM* as the best fit for the substitution process which model the formation of S' from S. *TAM* has the lowest AIC, although the difference with *tran* is on the borderline of meaningful. (Recall that differences of less than 2 units in AIC and 6 in BIC are generally considered negligible.) However, BIC is definitive; *TAM* had the lowest (best) score by a wide margin. Moreover, there is no reason to prefer *tran* over

[d] All data was analyzed at 10-digit precision. Due to rounding, reported values may not add up exactly.

	Nucleotides by pairedness $\Pi/\Pi^\tau/\hat{\Pi}^\tau$	Pairedness $\Pi_{pu}/\Pi^\tau_{pu}/\hat{\Pi}^\tau_{pu}$	Base pairs $\rho/\rho^\tau/\hat{\rho}^\tau$
Int	$\begin{pmatrix} 0.071 & 0.189 & 0.213 & 0.096 \\ 0.110 & 0.070 & 0.110 & 0.141 \end{pmatrix}$	$(0.570, 0.430)$	$(0.662, 0.251, 0.087)$
Exp	$\begin{pmatrix} 0.054 & 0.270 & 0.243 & 0.051 \\ 0.082 & 0.100 & 0.125 & 0.075 \end{pmatrix}$	$(0.618, 0.382)$	$(0.785, 0.148, 0.067)$
Obs	$\begin{pmatrix} 0.045 & 0.281 & 0.269 & 0.060 \\ 0.072 & 0.089 & 0.105 & 0.079 \end{pmatrix}$	$(0.655, 0.345)$	$(0.807, 0.126, 0.066)$
p-val	0.472	0.081	0.510

Table 1: Comparison of three expected pairing distributions under TAM bias model given initial ones from S and estimated $\hat{\tau}$ with observed alleles encoded in S'.

TAM as the latter provides more specific biological insights about the substitution domains that the former does not.

Finally, we note that the *GM* p-value likely indicates overfitting, and yet the information criteria support *TAM* as a better fit to the *C. elegans* tRNA data.

4.2 Assessing preservation of pairing potential

As part of the model validation, we confirmed that the observed substitution domains follow a TAM-biased distribution. In particular, C's are overrepresented in $\hat{\pi}^\tau$ by 43.0% while G's increased by 16.0% over the initial nucleotide distribution π. We now assess whether three related distributions — nucleotides by pairedness (Π), pairedness (Π_{pu}), and base pairs (ρ) — also follow TAM-biased distributions. Details are give in Table 1. In all three cases, we fail to reject the null hypothesis for the expected and observed substitution domains.

In fact, as seen in the first and third columns, the proportion of C and G nucleotides in base pairs which are mutated is consistently *higher* than expected under our model. This is surprising as it is inconsistent with expectations for the preservation of base pair formation.

One expectation for pairing preservation is that unpaired nucleotides should be more mutable. This is clearly violated as paired positions are mutated almost twice as often (0.655 versus 0.345). This is consistent with a TAM-biased domain distribution, and represents an increase of 14.9% over the initial Π_{pu} one. As seen when $\hat{\Pi}^\tau$ is compared to Π, this is entirely due to more substitutions in paired C's and G's which increased by 48.8% and 26.2% respectively over the initial distribution.

One might then expect that mutated GC pairs are predominantly converted to GU pairings under the $C \to T$ transition. As seen in Figure 4, this is indeed the most common substitution. However, it happens only 33.5% of the time compared to the destabilizing $G \to A$ transition and the four relevant transversions. Likewise, an AU pairing is potentially preserved as a GU one in only 31.8% of substitutions. However, these proportions are actually lower than when the corresponding unpaired

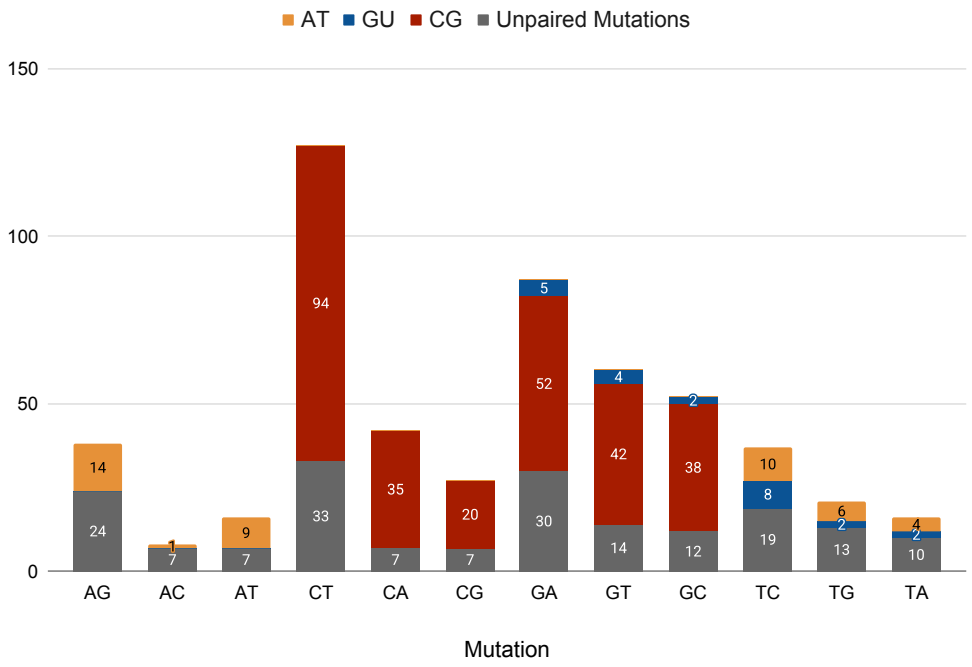


Figure 4: Number of observed substitutions, categorized by pairedness with pairings further subdivided by type. Substitutions ordered first by nucleotide domain and then transitions followed by transversions which preserve amino/keto and then weak/strong.

domain combinations are considered: 47.1% for C,G and 45.3% for A,U. Hence, in this microevolution setting, we do not see strong support for selection favoring the formation of GU pairs from either GC or AU ones.

Nor do we see enhanced conversion of GU pairings into GC/AU ones, as might also be expected. There are 974 wobble pairings in S (8.7% of 11,208) but GU is only 6.6% of the 348 pairings mutated to form S'. This is surprising as both transitions preserve pairing potential, and moreover would yield a base pair which is more thermodynamically stable.

Hence, we see a bias towards mutations in paired positions, particularly GC pairings, rather than away. Of the Watson-Crick pairings that are mutated, only 1/3 could form a wobble pairing. And there is no enhanced conversion of wobble pairings under mutation. All this suggests that the allelic distribution is driven more by TAM than preservation of secondary structure.

4.3 Comparison with neutral neighbors

When considering the distribution of alleles encoded in S', we found that TAM's signature was stronger than selection for base pair preservation. In particular, 42.7% of observed one-point mutations disrupt pairings. Yet, it is well-established that secondary structure conservation is a critical factor in tRNA macroevolution — so much so that MFE-preservation is considered a reliable proxy for fitness (Aguirre et al., 2011; Fontana and Schuster, 1998; Reidys et al., 2001). Here we show that even when our dataset is restricted to the neutral neighbors, clear deviations from macroevolution expectations are seen.

Neutrality under thermodynamic optimization is a strong constraint. Although 57.3% of substitutions from S to S' preserve pairing potential, only 37.3% are MFE-preserving. Interestingly, nearly all the difference is from the unpaired positions. The net result is that mutations still occur in paired domains slightly more often than unpaired: 57.1% versus 42.9%. This is less than the substitution domains for the full dataset (65.5% paired) but considerably more than those for the related neutral neighbors (32.6% paired).

Moreover, we still find that GC pairings are mutated more than expected from the initial (restricted) distribution. This is again consistent with our TAM-bias model, but not with thermodynamic stability. In contrast, the initial distribution, which closely resembles the full one, favors GC over AU and GU pairings roughly 8:3:1. Hence, we see the thermodynamic stability expected by macroevolution (Forster et al., 2006; Waldispühl et al., 2008) in the initial distribution, which is then being diminished by TAM-biased microevolution.

To quantify this bias further, recall that M denotes the MFE-preserving alleles from S' and $\alpha(m)$ the corresponding gene in S for each $m \in M$. The set of all related neutral neighbors for m is $\mathcal{N}_m$, and $\mathcal{N} = \bigcup_{m \in M} N_m$. For each $m \in M$, we now consider the relative thermodyanmic stability f_m for m and $\alpha(m)$ in comparison to N_m. As described in *Materials and Methods*, the f_m value ranges from 0 to 1 with higher being less favorable.

As seen in Figure 7 in *Appendix*, the values for the genes $\alpha(m)$ are fairly narrowly clustered with a mean of 0.415 and standard deviation of 0.084. In contrast, those for the alleles m are distributed over the full range with a (mean, std) = (0.563, 0.258). Hence, although GC pairs are clearly favored in the consensus genes, they are not the most thermodynamically stable in their neutral neighborhood. Interestingly, this may be indicative of mutational robustness (Gabzi et al., 2022) and/or evolvability (Wagner, 2023). In contrast, the observed neutral neighbor alleles tend to be among the less energetically favorable possibilities, once again demonstrating a clear difference from macroevolution expectations.

Finally, we consider the positional distributions for M and $\mathcal{N}$ across the different tRNA substructures. As illustrated in Figure 5, the latter closely resembles findings for neutral neighbors from macroevolution studies (Fontana and Schuster, 1998; Reidys et al., 2001). However, there are key differences for the former microevolution domain distribution.

Observed MFE-preserving mutations occur more frequently in paired domains, most notably in the acceptor stem (AR/AL) but also the 3' side of the T-arm (TR). The

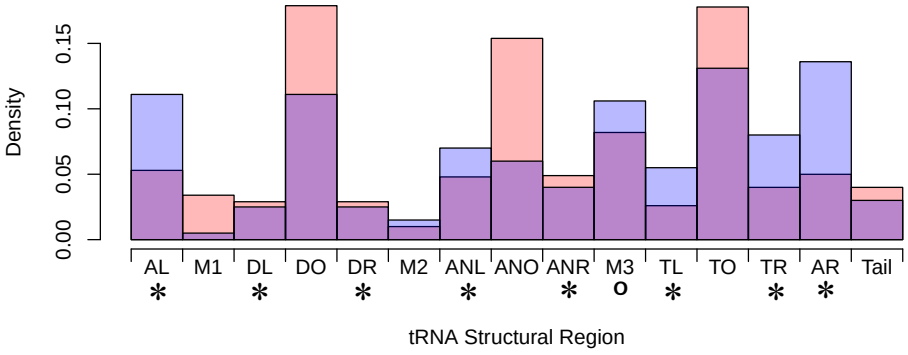


Figure 5: Distribution of observed MFE-preserving substitutions from M (blue) compared to all related neutral neighbors in $\mathcal{N}$ (red). Components of the tRNA secondary structure are labeled as in Figure 1 with paired regions marked by $*$. Note the enhancement of paired domains in M, particularly the acceptor stem, while mutations in hairpin loops, especially the anticodon one, are depressed. Conclusions about the variable arm (M3) are confounded since it can contain both kinds of sites depending on the specific tRNA.

variable arm portion of the multibranch loop (M3) is also enhanced, but conclusions are confounded since it can contain paired as well as unpaired sites depending on the particular tRNA. In contrast, substitutions in all three hairpin loops (DO, ANO, TO) are depressed, most notably for the anticodon.

The three anticodon nucleotides are an essential component of a functional tRNA molecule, so it is hardly surprising if they are less mutable than might otherwise be expected under thermodynamic predictions. However, anticodons are also known to evolve rapidly to meet new demands (Yona et al., 2013), and hence were included in our TAM-biased model. Although the anticodon itself represents $< 1/2$ of its hairpin loop, it was not *a priori* clear that there would be a discernible difference at this level of structural resolution.

4.4 Agreement with experimental fitness assays

We have demonstrated that the distribution of alleles in *C. elegans* tRNA is described by our new TAM-biased Markov substitution model but deviates substantially from macroevolution expectations, even when restricted to neutral neighbors. We now show that this distribution, and by extension our conclusions, accords with *in vitro* fitness tRNA studies (Domingo et al., 2018; Li et al., 2016; Yona et al., 2013).

Recall that their experiments use the single-copy arginine-CCU gene from *S. cerevisiae*, so are insulated from the confounding effects when multiple copies of the same

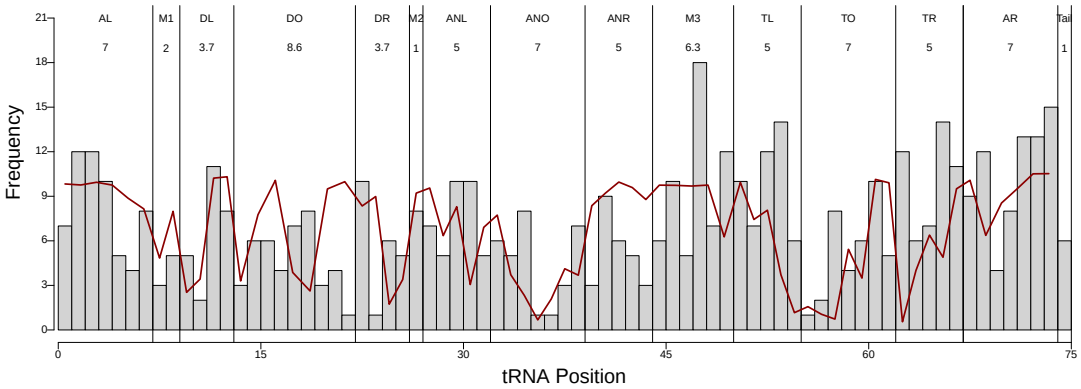


Figure 6: Comparison of substitution frequency over 531 *C. elegans* tRNA alleles with mean site-fitness (Li et al., 2016) across all one-point mutants for the *S. cerevisiae* arginine-CCU tRNA. Substructure components are labeled as in Figure 1 along with their average length, which was used for the site normalization.

isoacceptor are present. Nonetheless, as seen in Figure 6 and detailed below, we find good agreement between the site-fitness distribution for all one-point mutants obtained by Li et al. (2016) and the location of substitutions in *C. elegans* alleles. Hence, while some of these observed mutations may be deleterious ones not yet purged by selection, most are expected to be "(nearly) neutral" (Li et al., 2016) — as measured by microevolution growth competition experiments if not under thermodynamic optimization predictions.

To begin, we see that all *C. elegans* sites are mutated, but that lower frequency ones are often at (or near) a *S. cerevisiae* fitness minimum. In particular, the lowest frequency/fitness sites are seen in hairpin loops, specifically the anticodon and the 5' end of the T-loop. The former agrees with both selection to preserve the anticodon as well as experimental evidence that it can evolve rapidly when needed (Yona et al., 2013), while the latter is conserved as part of the internal promoter B-box region (Li et al., 2016). There are also proximal frequency/fitness lows in the D-stem which are not (yet) explained by functional conservation.

Conversely, regions with higher average frequency, i.e. both sides of the acceptor stem as well as the variable portion of the multiloop, also have high average fitness. This also echoes results from Domingo et al. (2018), whose dataset is built from 14 existing mutations of which 13 are from these regions.

Broadly, then, these two very different datasets agree; the acceptor stem and certain other paired sites are more mutable than expected under the (thermodynamic) neutral neighbor hypothesis, while the hairpin loops are much less so. These findings — aligned with our TAM-biased model — underscore the surprising robustness of many paired positions to mutations while also highlighting the influence of biochemical and other constraints beyond the secondary structure on the loop composition.

5 Conclusions

It is well-established that tRNAs are subject to strong purifying selection that maintains their structural integrity over macroevolutionary time scales. This is most apparent in the conservation of secondary structure, as evidenced by covarying base pairs. It is even visible at the microevolution time scale in the mutation rate drop compared to their flanking regions. However our results reveal that the distribution of alleles in extant *C. elegans* strains is consistent with our model of TAM bias but not with the expectations of thermodynamic neutrality for secondary structures.

This highlights a disconnect between scales. At the microevolutionary level, tolerable mutations often disrupt secondary structure, as reinforced by *in vitro* experiments. Yet, at the macroevolutionary level, selection pressure preserves canonical pairings. The resolution of this micro/macro dichotomy is likely to involve a variety of factors including the role of compensatory mutations (Kern and Kondrashov, 2004; Meer et al., 2010) and the interaction of isoacceptors (Yona et al., 2013) as well as the multi-copy nature of most tRNA genes (Bloom-Ackermann et al., 2014).

Beyond this, our new population model may be applicable in other contexts. The key assumptions are high gene conservation and short evolutionary time to the most recent common ancestor. These enable the use of a consensus sequence as the ancestral approximation and the encoding of the allelic distribution in a single compound mutational variation 'descendant'.

It may also be possible to integrate other mutational biases into a Markov model through careful rate matrix parameterization. This naturally leads to questions of model identifiability, whose resolution may reveal new biological insights.

6 Acknowledgments

This work was supported by the NSF-Simons Southeast Center for Mathematics and Biology (SCMB) through the grants National Science Foundation (NSF) DMS-1764406 and Simons Foundation/SFARI 594594. H.B. was also partially supported by NSF grant DMS-2331660. C.H. was also partially supported by NIH grant R01GM126554.

References

Aguirre, J., Buldú, J. M., Stich, M., and Manrubia, S. C. (2011). Topological structure of the space of phenotypes: The case of RNA neutral networks. *PLoS ONE*, 6(10):e26324.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, AC-19(6):716–723.

Allen, J. E. and Whelan, S. (2014). Assessing the state of substitution models describing noncoding rna evolution. *Genome Biology and Evolution*, 6(1):65–75.

Berg, M. D. and Brandl, C. J. (2021). Transfer RNAs: diversity in form and function. *RNA Biology*, 18(3):316–339. PMID: 32900285.

Bloom-Ackermann, Z., Navon, S., Gingold, H., Towers, R., Pilpel, Y., and Dahan, O. (2014). A comprehensive tRNA deletion library unravels the genetic architecture of the tRNA pool. *PLOS Genetics*, 10(1):1–16.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.

Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Müller, K. M., Pande, N., Shang, Z., Yu, N., and Gutell, R. R. (2002). The comparative RNA Web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, 3(1):2.

Casanellas, M. and Sullivant, S. (2005). *The Strand Symmetric Model*, page 305–321. Cambridge University Press.

Chan, P. P. and Lowe, T. M. (2008). GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Research*, 37(suppl1):D93–D97.

Chan, P. P. and Lowe, T. M. (2015). GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Research*, 44(D1):D184–D189.

Chan, P. P. and Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA genes in genomic sequences. In *Gene Prediction: Methods and Protocols*, pages 1–14. Springer New York, New York, NY.

Cook, D. E., Zdraljevic, S., Roberts, J. P., and Andersen, E. C. (2016). Cendr, the caenorhabditis elegans natural diversity resource. *Nucleic Acids Research*, 45(D1):D650–D657.

Cowperthwaite, M. C. and Meyers, L. A. (2007). How mutational networks shape evolution: Lessons from RNA models. *Annual Review of Ecology, Evolution, and Systematics*, 38(Volume 38, 2007):203–230.

Czech, L., Barbera, P., and Stamatakis, A. (2018). Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics*, 35(7):1151–1158.

Daewoo Pak, R. R.-B. and Burton, Z. F. (2017). tRNA structure and evolution and standardization to the three nucleotide genetic code. *Transcription*, 8(4):205–219. PMID: 28632998.

De Maio, N., Schrempf, D., and Kosiol, C. (2015). PoMo: An allele frequency-based approach for species tree estimation. *Systematic Biology*, 64(6):1018–1031.

Domingo, J., Diss, G., and Lehner, B. (2018). Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature*, 558(7708):117–121.

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

Engels, B. (2015). *XNomial: Exact Goodness-of-Fit Test for Multinomial Data with Fixed Probabilities*. R package version 1.0.4.

Felsenstein, J. (2003). *Inferring phylogenies*. Sinauer Associates.

Fontana, W. and Schuster, P. (1998). Shaping space: the possible and the attainable in RNA genotype–phenotype mapping. *Journal of Theoretical Biology*, 194(4):491–515.

Forster, R., Adami, C., and Wilke, C. O. (2006). Selection for mutational robustness in finite populations. *Journal of Theoretical Biology*, 243(2):181–190.

Gabzi, T., Pilpel, Y., and Friedlander, T. (2022). Fitness landscape analysis of a tRNA gene reveals that the wild type allele is sub-optimal, yet mutationally robust. *Molecular Biology and Evolution*, 39(9):msac178.

Gaillard, H. and Aguilera, A. (2016). Transcription as a threat to genome integrity. *Annual Review of Biochemistry*, 85(Volume 85, 2016):291–317.

García-Lapresta, J. L. and Martínez-Panero, M. (2024). Two characterizations of the dense rank. *Journal of Mathematical Economics*, 111:102963.

Grayson, D. R. and Stillman, M. E. (2009). Macaulay2, a software system for research in algebraic geometry. Available at `http://www2.macaulay2.com`.

Green, P., Ewing, B., Miller, W., Thomas, P. J., Green, E. D., and Program, N. C. S. (2003). Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genetics*, 33(4):514–517.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33(suppl1):D121–D124.

Harris, T. W., Arnaboldi, V., Cain, S., Chan, J., Chen, W. J., Cho, J., Davis, P., Gao, S., Grove, C. A., Kishore, R., Lee, R. Y. N., Muller, H.-M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Rodgers, F. H., Russell, M., Schindelman, G., Auken, K. V., Wang, Q., Williams, G., Wright, A. J., Yook, K., Howe, K. L., Schedl, T., Stein, L., and Sternberg, P. W. (2019). WormBase: a modern model organism information resource. *Nucleic Acids Research*, 48(D1):D762–D767.

Ishimura, R., Nagy, G., Dotu, I., Zhou, H., Yang, X.-L., Schimmel, P., Senju, S., Nishimura, Y., Chuang, J. H., and Ackerman, S. L. (2014). Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science*, 345(6195):455–459.

Jinks-Robertson, S. and Bhagwat, A. S. (2014). Transcription-associated mutagenesis. *Annual Review of Genetics*, 48(Volume 48, 2014):341–359.

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

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14(6):587–589.

Kern, A. D. and Kondrashov, F. A. (2004). Mechanisms and convergence of compensatory evolution in mammalian mitochondrial trnas. *Nature Genetics*, 36(11):1207–1212.

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

Li, C., Qian, W., Maclean, C. J., and Zhang, J. (2016). The fitness landscape of a tRNA gene. *Science*, 352(6287):837–840.

Liu, X. (2013). *Molecular Recognition of Inhibitors, Metal Ions and Substrates by Ribonuclease P.* Phd thesis, University of Michigan, Ann Arbor, MI. Available at `https://deepblue.lib.umich.edu/handle/2027.42/100013`.

Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26.

MathWorks, T. (2024). MATLAB version: 24.2.0.2871072 (r2024b).

Meer, M. V., Kondrashov, A. S., Artzy-Randrup, Y., and Kondrashov, F. A. (2010). Compensatory evolution in mitochondrial trnas navigates valleys of low fitness. *Nature*, 464(7286):279–282.

Reidys, C., Forst, C. V., and Schuster, P. (2001). Replication and mutation on neutral networks. *Bulletin of Mathematical Biology*, 63(1):57–94.

Saks, M. E., Sampson, J. R., and Abelson, J. (1998). Evolution of a transfer RNA gene through a point mutation in the anticodon. *Science*, 279(5357):1665–1670.

Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. (1994). From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1344):279–284.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Simonti, C., Paaby, A., Lachance, J., Wang, L., Heitsch, C., and Baños, H. (2020). tRNA genes in the system C. elegans. *In progress.*

Tang, D. T., Glazov, E. A., McWilliam, S. M., Barris, W. C., and Dalrymple, B. P. (2009). Analysis of the complement and molecular evolution of tRNA genes in cow. *BMC Genomics*, 10(1):188.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on Mathematics in the Life Sciences*, 17.

Thornlow, B. P., Armstrong, J., Holmes, A. D., Howard, J. M., Corbett-Detig, R. B., and Lowe, T. M. (2020). Predicting transfer RNA gene activity from sequence and genome context. *Genome Research*, 30(1):85–94.

Thornlow, B. P., Hough, J., Roger, J. M., Gong, H., Lowe, T. M., and Corbett-Detig, R. B. (2018). Transfer RNA genes experience exceptionally elevated mutation rates. *Proceedings of the National Academy of Sciences*, 115(36):8996–9001.

Tinoco, I. and Bustamante, C. (1999). How RNA folds. *Journal of Molecular Biology*, 293(2):271–281.

Trudeau, D. L., Kaltenbach, M., and Tawfik, D. S. (2016). On the potential origins of the high stability of reconstructed ancestral proteins. *Molecular Biology and Evolution*, 33(10):2633–2641.

Wagner, A. (2023). Evolvability-enhancing mutations in the fitness landscapes of an RNA and a protein. *Nature Communications*, 14(1):3624.

Waldispühl, J., Devadas, S., Berger, B., and Clote, P. (2008). Efficient algorithms for probing the rna mutation landscape. *PLOS Computational Biology*, 4(8):1–22.

Westhof, E., Thornlow, B., Chan, P. P., and Lowe, T. M. (2022). Eukaryotic tRNA sequences present conserved and amino acid-specific structural signatures. *Nucleic Acids Research*, 50(7):4100–4112.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press.

Yona, A. H., Bloom-Ackermann, Z., Frumkin, I., Hanson-Smith, V., Charpak-Amikam, Y., Feng, Q., Boeke, J. D., Dahan, O., and Pilpel, Y. (2013). trna genes rapidly change in evolution to meet novel translational demands. *eLife*, 2:e01339.

Zhang, J. and Ferré-D'Amaré, A. R. (2016). The tRNA elbow in structure, recognition and evolution. *Life*, 6(1).

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148.

7 Appendix

The output obtained from Macaulay2 by solving the linear system described in Section 3.1.3.

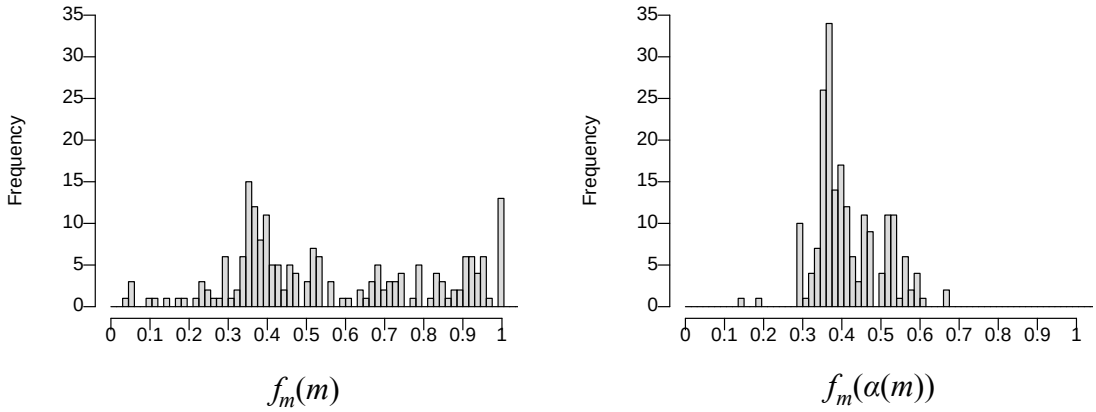


Figure 7: Relative thermodynamic stability f_m for the 198 observed neutral neighbors (left) and corresponding tRNA gene (right). Approximately half of the observed substitutions in the *C. elegans* population are disproportionately less stable than other related neutral neighbors, but no such skew is seen for the consensus tRNA genes approximating the ancestral sequence.

$$
\begin{aligned}
\lambda =& \frac{1}{\pi_T \pi_G \pi_G \pi_T \pi_A^{\tau}}((\theta - \epsilon + \alpha)\pi_G + \alpha\tau_R)\pi_C^4 + ((3\theta - 2\epsilon + 2\alpha + \beta)\pi_G^2 + ((3\theta + \gamma - 2\epsilon + 3\alpha)\pi_T + (\theta - \epsilon + 2\alpha + \beta)\tau_R \\
& - 3\theta + 2\epsilon - 2\alpha)\pi_G + \tau_R((\gamma + 2\alpha)\pi_T - 2\alpha))\pi_C^3 + ((3\theta - \epsilon + \alpha + 2\beta)\pi_G^3 + ((6\theta + 2\gamma - 3\epsilon + 3\alpha + 3\beta)\pi_T \\
& + (2\theta - \epsilon + \alpha + 2\beta)\tau_R - 6\theta + 2\epsilon - 2\alpha - 2\beta)\pi_G^2 + ((3\theta + 3\gamma - \epsilon + 2\alpha)\pi_T^2 + ((2\theta + 2\gamma - \epsilon + 2\alpha + 2\beta)\tau_R \\
& + 2\tau_R(\pi_T - 1)((\gamma + \frac{\alpha}{2})\pi_T - \frac{\alpha}{2}))\pi_C^2 + ((\theta + \beta)\pi_G^4 + ((3\theta + \gamma - \epsilon + 3\beta)\pi_T + (\theta + \beta)\tau_R - 3\theta - 2\beta)\pi_G^3 \\
& + ((3\theta + 3\gamma - \epsilon + 2\beta)\pi_T^2 + ((2\theta + \gamma - \epsilon + 2\beta)\tau_R - 6\theta - 2\gamma + \alpha\tau_Y + (\tau_Y - 3)\beta + \epsilon)\pi_T \\
& + (-2\theta - 2\beta)\tau_R + 3\theta + \beta)\pi_G^2 + ((\theta + 2\gamma)\pi_T^3 + ((\theta + 2\gamma + \beta)\tau_R - 3\theta + (\tau_Y - 3)\gamma + \alpha\tau_Y)\pi_T^2 \\
& + ((-2\theta - 2\gamma - 2\beta)\tau_R - \alpha\tau_Y + 3\theta + \gamma)\pi_T + (\theta + \beta)\tau_R - \theta)\pi_G + \pi_T\gamma\tau_R(\pi_T - 1)^2)\pi_C \\
& + \pi_T\pi_G\tau_Y(\pi_T + \pi_G - 1)(\gamma\pi_T + \beta\pi_G) \\
\mu =& \frac{-1}{\pi_G \pi_T \pi_A^{\tau}}(2\pi_T^3\gamma + ((3\gamma - \alpha + \beta)\pi_G + (3\pi_C + \tau_R - \tau_Y - 3)\gamma + \alpha\pi_C)\pi_T^2 + ((\gamma - \alpha + \beta)\pi_G^2 + ((2\pi_C + \tau_R - \tau_Y - 2)\gamma \\
& + \beta\pi_C - \beta\tau_Y + (1 - \tau_R)\alpha - \beta)\pi_G + (\pi_C - 1)(\pi_C + \tau_R - \tau_Y - 1)\gamma + \alpha\pi_C(\pi_C - \tau_Y - 1))\pi_T - \tau_Y(\pi_C + \pi_G - 1) \\
& (\alpha\pi_C + \beta\pi_G)) \\
\delta =& \frac{-1}{\pi_G \pi_T \pi_A^{\tau}}((\theta + \beta)\pi_G^3 + ((2\theta - \epsilon + \alpha + \beta)\pi_C + (2\theta + \gamma + \beta)\pi_T + (\theta + \beta)\tau_R - 2\theta - \beta)\pi_G^2 + ((\theta - \epsilon + \alpha)\pi_C^2 \\
& + ((2\theta + \gamma - \epsilon + \alpha)\pi_T + (\theta - \epsilon + \alpha + \beta)\tau_R - 2\theta + \epsilon - \alpha)\pi_C + (\theta + \gamma)\pi_T^2 + ((\theta + \gamma + \beta)\tau_R - 2\theta - \gamma)\pi_T \\
& + (-\theta - \beta)\tau_R + \theta)\pi_G + \tau_R(\pi_C + \pi_T - 1)(\gamma\pi_T + \alpha\pi_C))
\end{aligned}
$$

The dependencies of the Q_{TAM} matrix entries shown in Section 3.1.3 where derived manually from these.

H. Banos Department of Mathematics, California State University San Bernardino,

San Bernardino, CA, USA.
School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 30332, USA.

L. Wang School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA.

C. Simonti School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA.

A. Paaby School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA.

C. Heitsch School of Mathematics, Georgia Institute of Technology, Atlanta, GA, 30332, USA.