Neutral models of *de novo* gene emergence suggest that gene evolution has a preferred trajectory

Bharat Ravi Iyengar[1,†], Erich Bornberg-Bauer[1,2]

[1]Institute for Evolution and Biodiversity, University of Münster,
Hüfferstrasse 1, 48149 Münster, Germany

[2]Department of Protein Evolution, Max Planck Institute for Biology Tübingen,
Max-Planck-Ring 5, 72076 Tübingen, Germany

[†] Corresponding author: b.ravi@uni-muenster.de

Abstract

New protein coding genes can emerge from genomic regions that previously did not contain any genes, via a process called *de novo* gene emergence. To synthesize a protein, DNA must be transcribed as well as translated. Both processes need certain DNA sequence features. Stable transcription requires promoters and a polydenylation signal, while translation requires at least an open reading frame (ORF). We develop mathematical models based on mutation probabilities, and the assumption of neutral evolution, to find out how quickly genes emerge and are lost. We also investigate the effect of the order by which DNA features evolve, and if sequence composition is biased by mutation rate. We rationalize how genes are lost much more rapidly than they emerge, and how genes with long ORFs preferentially arise in regions that are already transcribed. Our study not only answers some fundamental questions on the topic of *de novo* emergence but also provides a modeling framework for future studies.

Introduction

Organisms evolve new traits by expanding their functional genome. Evolution of new genes is one of the ways by which new traits can emerge. The definition of a gene is complicated, and has been changing constantly (Gerstein *et al.*, 2007). We use the following working definition of a gene: a gene is a region of the genome that gives rise to a functional product, that is an RNA or a protein. The next big challenge is to define what a function is, and how to classify gene functions. One of the easiest definition of a gene product's function is its ability to improve the survival of the organism under one or more environmental conditions (Keeling *et al.*, 2019). Proteins perform the most diverse kinds of molecular functions ranging from catalysis of biochemical reactions to formation of cellular structures (Berg *et al.*, 2002), whereas RNAs that do not encode proteins like mRNAs, or participate in protein synthesis like rRNAs and tRNAs, are mostly involved in regulation of gene expression (Iyengar *et al.*, 2014; Statello *et al.*, 2021).

In this study we focus on the evolution of genes that encode new proteins. New protein coding genes frequently arise via duplication of existing protein coding genes. These duplicated genes then genetically and functionally diverge from their parent genes by accumulating mutations (Long *et al.*, 2003; Näsvall *et al.*, 2012). Recently it has been shown that new protein coding genes can arise independent of gene duplication, from genomic sequences that did not previously encode any protein. This phenomenon is called *de novo* gene emergence (Tautz and Domazet-Lošo, 2011; Zhao *et al.*, 2014; Schmitz and Bornberg-Bauer, 2017; Vakirlis *et al.*, 2017; Van Oss and Carvunis, 2019). *De novo* protein coding genes can emerge from intergenic sequences, non-coding RNA genes, introns, and even regions that partially overlap with existing protein coding genes (Tautz and Domazet-Lošo, 2011; Zhao *et al.*, 2014; Vakirlis *et al.*, 2017; Van Oss and Carvunis, 2019; Prabh and Rödelsperger, 2019). To express a protein, a genomic sequence should be transcribed as well as translated. Therefore, the gene needs sequence features that enable both these processes.

The primary requirement for transcription is recruitment of RNA polymerase. This process is facilitated by DNA sequences called (Lenhard *et al.*, 2012). A part of the promoter, called the core promoter is the region that determines the start of transcription (Haberle and Stark, 2018). Transcription can also be initiated without the requirement of a defined promoter sequence, and this process can occur throughout the genome Clark *et al.* (2011). Even though pervasive transcription is widespread in eukaryotic genomes, most RNAs are quickly eliminated by RNA degrading enzymes (Schmid and

Jensen, 2018). Most eukaryotic mRNAs and long non-coding RNAs, are polyadenylated at their 3′ termini, which makes them resistant to degradation. Polyadenylation, which also marks the end of transcription (Richard and Manley, 2009), is facilitated by sequences known as polyadenylation signals (Proudfoot, 2011). In prokaryotes, the end of transcription is determined by sequences known as terminators (Santangelo and Artsimovitch, 2011). Termination at an appropriate site is another fundamental requirement for a productive transcription.

The second major requirement for protein expression is translation of the transcribed mRNA. The most fundamental requirement for translation is the presence of an open reading frame (ORF). Usually, an mRNA needs additional features to initiate protein synthesis. These features include ribosome binding sites in prokaryotes (Omotajo *et al.*, 2015), and Kozak consensus sequences in eukaryotes (Kozak, 1986; Acevedo *et al.*, 2018; Noderer *et al.*, 2014).

Although gain of transcription and translation features guarantees *de novo* birth of a protein coding gene, it does not ensure that the gene would persist in the genome for many generations. The newly born gene can lose the features as easily as it gained them unless it has been fixed in the genome, for example via evolutionary selection. Specifically, if the protein synthesized by the *de novo* gene provides a fitness advantage to the host organism, it will undergo positive selection (Keeling *et al.*, 2019). Simultaneously, the protein should have low toxicity and cost of synthesis, to survive purifying selection. A common mechanism by which protein mediated toxicity occurs is by misfolding and aggregation of proteins (Bucciantini *et al.*, 2002; Hartl, 2017).

In this study we develop probabilistic models of *de novo* gene emergence in eukaryotes. We focus on *de novo* genes emerging from non-genic sequences, also known as "proto-genes" (Carvunis *et al.*, 2012; Van Oss and Carvunis, 2019). We model *de novo* gene emergence as a two step process. In the first step, a non-genic DNA sequence gains transcriptional features that allows it to express an untranslated (non-coding) RNA. In the second step this non-coding RNA gene acquires translational features that allows it to express a protein. Alternatively, translational features can emerge in the non-genic DNA before transcriptional features. Specifically, we calculate the probabilities by which polyadenylation signals and ORFs emerge, and are lost, based on the rates at which different kinds of DNA mutations naturally occur (mutation bias). With these probabilities we estimate whether transcriptional and translational features preferentially evolve in a specific order. Using a similar approach, we calculate how random mutations affect protein composition. In our models, we assume that the proteins expressed from these proto-genes provide no fitness advantage and are not toxic

to the host organism. Thus our models are based on neutral evolution. Our models predict that there is indeed a preferred sequence of DNA feature evolution during *de novo* emergence, even under the assumption of neutrality.

Results

In this work, we developed mathematical models to estimate the rates and probabilities of *de novo* gene emergence, as well as gene loss. A proto-gene emerges from non-genic DNA, when the latter mutates to gain sequence features necessary for transcription and translation (Carvunis *et al.*, 2012; Van Oss and Carvunis, 2019). Both transcription and translation are complex processes involving many biomolecular complexes that work in concert. Here, we focus on the minimal requirements for these processes to occur. Experimental data show that transcription in eukaryotes can be initiated genome wide, especially in regions proximal to enhancers, and more specifically, in regions depleted of nucleosomes (Andersson and Sandelin, 2020). Thus transcription of many proto-genes can be initiated at a basal transcription rate, and may not require specific core promoters. Although transcription may be initiated without a core promoter, a stable RNA product often requires a poly-A tail that also facilitates RNA export to cytoplasm (Richard and Manley, 2009; Stewart, 2019). This assumption is supported by experimental data, where the majority of detectably expressed proto-genes have polyadenylated RNAs (Neme and Tautz, 2016; Witt *et al.*, 2019; Majic and Payne, 2020; Blevins *et al.*, 2021). Thus, we define a poly-A signal as the primary requirement for transcription such that it should exist but only at the end of a transcribed region. For translation, we only require the gene to have an open reading frame (ORF). We only focus on intronless genes in this study.

Using simple probability models, we calculate the likelihood of finding the sequence features that facilitate *de novo* gene emergence by random chance, and the rate at which these features are gained and lost due to random mutations. These probability models are essentially described by two kinds of probability. First is the probability of finding a DNA sequence feature (such as the poly-A signal or an ORF; Methods section 2.1). This probability depends on the nucleotide distribution that can be roughly approximated by the GC-content. The second kind of probability is the transition probability, that is the probability of a DNA sequence mutating to another (Methods section 2.2). Transition probabilities that describe the gain and loss of DNA sequence features, depend on mutation rate, mutation bias, and nucleotide composition. Our model of DNA sequence evolution is in principle similar to a theoretical model described in a previous

study (Behrens and Vingron, 2010). In this study, the authors have calculated the time required for different transcription factor binding sites to emerge in gene promoters. We extend this principle to calculate not only the probability of sequence emergence but also the likelihood that an emerged sequence will remain intact or be lost (Methods section 3).

In our model we primarily focus on *de novo* gene emergence at one genomic locus. We assume that transcription is initiated at this locus with a probability of 0.12. We based this estimate on the fraction of intergenic open chromatin in *Drosophila melanogaster* that is occupied by *cis*-regulatory elements (Methods section 5). We used a default GC-content of 42% in our calculations which is reasonably close to the total genomic GC-content of human (41%, Merchant *et al.*, 2007), *Drosophila melanogaster* (41.6%, Gramates *et al.*, 2022), and *Saccharomyces cerevisiae* (38%, Wood *et al.*, 2002). Because GC content can vary between different genomic loci, we also performed of our calculations with other values of GC-content. We note that nucleotide composition in any DNA region may not be uniformly distributed. Therefore, we also included calculations using a more realistic approximation of nucleotide distribution, using frequencies of short DNA sequences in the intergenic genome of *D.melanogaster*. Specifically, we calculated the frequency of all DNA trimers for codon probabilites, and of all DNA hexamers for poly-A signal probabilities (Methods section 5).

We used a mutation rate of 7.8×10^{-9} mutations per nucleotide position per generation, which corresponds to spontaneous mutation rate estimated in a *D. melanogaster* population that was subjected to periodic population bottlenecks (mutation accumulation line, Schrider *et al.*, 2013). We used mutation bias data (Table 1) from the same study (Schrider *et al.*, 2013) which was in agreement with another previous study on human pseudogenes (Zhang and Gerstein, 2003).

Using a similar approach, we predict how the nucleotide distribution and random mutations shape protein composition and evolution.

Overall, our models define a null hypothesis under which protein coding DNA sequences evolve neutrally through mutation pressure alone.

How likely is gene loss?

If a gene is found in one population but not populations, then the most parsimonious assumption is that this gene emerged only in one specific population. We asked what is the chance that the gene was present in an ancestral population but was lost in all

lineages except one. To this end we calculated the probability of gene emergence and gene loss (Methods section 3). Specifically, the probability of gene gain can be defined as the sum of three probabilities. First is the probability that an ORF already exists and is not lost due to mutations ($P_{ORF\text{-}stay}$, Methods section 4), while a mutation causes transcription to emerge ($P_{RNA\text{-}gain}$, Methods section 5). Second is the probability that an ORF emerges ($P_{ORF\text{-}gain}$, Methods section 4) in a region of DNA that is already transcribed and continues to be transcribed ($P_{RNA\text{-}stay}$, Methods section 5). Third is the probability that neither of the two features already exist and both emerge at the same time due to mutations (this probability is very small and is negligible). We found that the probability of transcription gain is dependent on the existence of an ORF, such that it is higher when the ORF is already present. This dependence exists because an ORF does not have stop codons (`TAA`, `TAG` and `TGA`) in its sequence. This in turn, restricts the number of positions where a poly-A signal, that has a `TAA` in its sequence, can exist (Methods section 6). Because of the same reason, the probability of ORF gain is higher when the DNA is already transcribed. Therefore, we additionally calculated the conditional probabilities of transcription gain ($P_{RNA\text{-}gain\,|\,ORF}$) and ORF gain ($P_{ORF\text{-}gain\,|\,RNA}$) given the condition that an ORF and a transcript already exists, respectively. We thus define the total probability of gene gain ($P^{*}_{gene\text{-}gain}$) as:

$$P^{*}_{gene\text{-}gain} = P_{RNA\text{-}gain\,|\,ORF} \times P_{ORF\text{-}stay} + P_{ORF\text{-}gain\,|\,RNA} \times P_{RNA\text{-}stay} + P_{RNA\text{-}gain} \times P_{ORF\text{-}gain} \quad (1)$$

Equation 1 describes the total probability of gene gain that includes the probability that the gene was not previously present. To more precisely address the question of whether a gene that was absent in an ancestor but emerges in a descendant, we calculated the corresponding conditional probability of gene gain:

$$P_{gene\text{-}gain} = P^{*}_{gene\text{-}gain}/(1 - P_{ORF} - P_{RNA}) \quad (2)$$

We note that the difference between the total probability ($P^{*}_{gene\text{-}gain}$) and the conditional probability ($P_{gene\text{-}gain}$) of gene gain, is very small.

Next, we calculated the probability of gene loss, given the gene (transcription and ORF) is already present($P_{gene\text{-}loss}$). It is the sum of the probabilities that transcription ($P_{RNA\text{-}loss}$, Methods section 5) or the ORF is lost due to mutations ($P_{ORF\text{-}loss}$, Methods section 4). Specifically:

$$P_{gene\text{-}loss} = P_{RNA\text{-}loss} + P_{ORF\text{-}loss} \quad (3)$$

The probability that a gene is lost n times independently, is $(P_{gene\text{-}loss})^n$. To find out how many independent gene loss events are as likely as a single gene gain event, we calculate the ratio of logarithms (log-log ratio) of gene gain and gene loss for proto-genes with an ORF ranging from 30 – 300 codons in length.

$$\text{Gene losses per gene gain} = \log(P_{gene\text{-}gain})/\log(P_{gene\text{-}loss}) \quad (4)$$

For example, a log-log ratio value of 2 would indicate that two independent gene loss events are as likely as a single gene gain event. We calculated this ratio for genes with different ORF lengths, because ORF gain and loss probabilities depend on the length of the ORF. We found that proto-genes with a GC-content of 42% and ORFs longer than 40 codons, can be lost two times independently in the time frame of one gene gain event (log-log ratio = 2, Figure 1A). For proto-genes with GC-contents of 34% and 50%, the minimum number of codons required for a log-log ratio of 2, are 42 and 33 respectively. To better understand how GC-content determines the likelihood of gene gain relative to gene loss, we calculated the log-log ratio of the two probability values for proto-genes with different GC-content (Figure S4), and that contain ORFs of different lengths (codons). We found that the effect of GC-content on the log-log ratio depends on the length of the ORF. Specifically, the log-log ratio steadily increases with GC-content ranging from 30% to 60%, when the proto-gene contains an ORF with 30 codons. When the ORF has 60 – 90 codons, the log-log ratio initially decreases with GC content and then increases. For ORFs containing more than 120 codons, the log-log ratio decreases with increasing GC-content (within the analysed range). Because GC-content may not accurately describe the nucleotide composition of a locus, we calculated gene gain and loss probabilities using frequencies of DNA trimers and hexamers from the intergenic genome of *D. melanogaster*. Specifically, we used the frequencies of hexamers to estimate the stationary, gain and loss probabilities of poly-A signals (which is also a hexamer) and the RNA. Likewise, we used the frequencies of trimers to estimate the stationary, gain and loss probabilities of different codons and the ORF. Using these values, we found that protogenes with more than 49 codons can be independently lost twice, in the time frame of one gene gain event (orange line, Figure 1A).

Natural selection and effective population size (N_e) determine the rate at which new genes spread in a population (Kimura, 1962). To understand the combined effect of

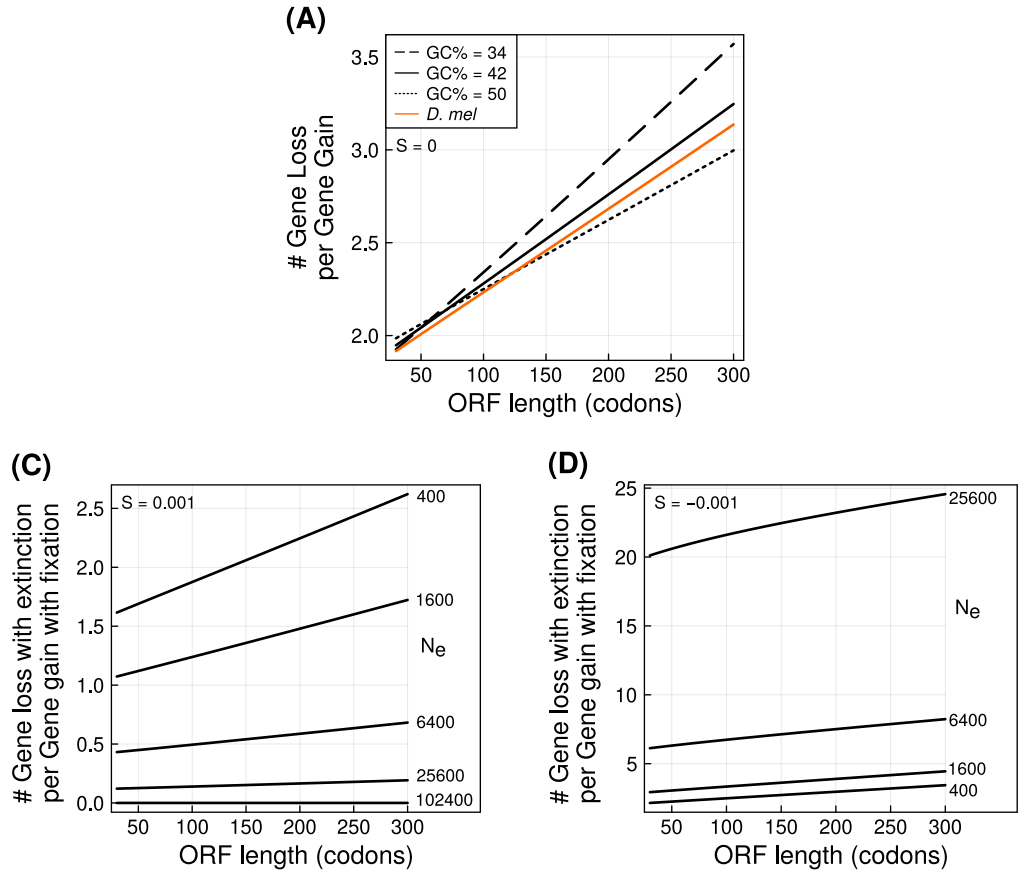

Figure 1: Genes are more likely to be independently lost twice, than being born once. The vertical axis in all panels shows the number of independent gene loss events with extinction (Equation 3) that can occur relative to one gene gain event with fixation (Equation 4, Table 2), under **(A)** no selection, **(B)** weak positive selection and **(C)** weak negative selection on the expression of the protein from a proto-gene. Horizontal axes show the number of codons in the ORF. Black lines in panel **(A)** denote the probability values estimated from overall GC content (34%, 42% and 50%) and the orange line denotes the probability values estimated from trimer and hexamer frequencies in *D. melanogaster* intergenic genome. In panels **(B)** and **(C)** we use the probabilitiy values from *D. melanogaster* intergenic genome. N_e denotes effective population size and s denotes the selection co-efficient.

mutations, and selection we calculated two fixation probabilities for populations consisting different number of diploid individuals. First is the probability that any one individual in a population gains a gene ($2N_e \times P_{gene\text{-}gain}$), and this gene eventually spreads in entire the population via selection and population dynamics (P_{fix}, Table 2). The second probability describes the event that all individuals in a population have two copies of the proto-gene, one copy of the gene is lost in an individual ($2N_e \times P_{gene\text{-}loss}$), and the gene eventually goes extinct from the entire population due to population dynamics (P_{ext}, Table 2). We calculated the log-log ratio of the two probabilities under three scenarios (Table 2), wherein the gene provides no fitness advantage (neutral, $s = 0$), the gene is marginally beneficial (positive, $s = 0.001$), and the gene is marginally deleterious (negative, $s = -0.001$). We note that gene gain and loss probabilities due to

mutation are so small that it is unlikely that two or more individuals gain or lose a gene simultaneously. Under the neutral scenario, the probability of fixation of an allele is same as the probability of the mutation that gives rise to that allele (Figure 1A). Extinction of a marginally beneficial gene is more likely than its fixation only when populations are small or when the gene has a long ORF (Figure 1B). For example, a gene of any length is less likely to go extinct than it is fixed, in populations with at least 6400 individuals. Furthermore, a beneficial gene can be independently lost in more than two small populations with 400 individuals, only when it has more than 133 codons. When a gene is mildly deleterious, its is very likely to go extinct such that independent loss in more than two populations is more likely than its fixation in one population (Figure 1C). We further note that time to fixation of a rare allele (for example, a newly emerged protogene) would be much smaller ($4N_e$ generations; Kimura, 1962) than the time to gain (10^{14} – 10^{20} generations) or lose it (10^{7} – 10^{8} generations).

Overall, our analysis suggests that a proto-gene expressed in only one population may not necessarily mean that it emerged for the first time in this population. This is also true for a model of gene emergence that requires a core promoter (TATA-box or Inr) for initiating transcription (Figure S1).

Despite low probability of gene gain at one particular locus ($\sim 10^{-15}$ per generation), the overall rate of gene gain throughout the genome could be much higher. To this end we calculated the total gain rate of genes of any length, at any locus in *D. melanogaster* intergenic genome. We found this rate to be 2.8×10^{-5} new genes per generation, that corresponds to approximately 7.5×10^{-5} new genes per year (assuming *D. melanogaster* generation time of two weeks; Fernández-Moreno *et al.*, 2007). We performed an analogous analysis on the rate of gain of new transcripts and found this rate to be 1.53 transcripts per year. This estimate is slightly higher than the rate of transcript gain predicted from *D. melanogaster* population transcriptomics data (0.13 – 0.34; Grandchamp *et al.*, 2023).

Does *de novo* gene emergence follow a preferred trajectory of events?

For a proto-gene to emerge from a non-genic DNA sequence, both transcription and ORF need to emerge. That is, probability of gene emergence is equal to the product of probabilities of transcription gain and ORF gain. Thus it may appear that the order of the occurrence of these two events does not matter. However, gene emergence is much more likely when one of the two features already exists (Equation 1) and therefore it has two possible trajectories – ORF emerges first (ORF-first) or transcription

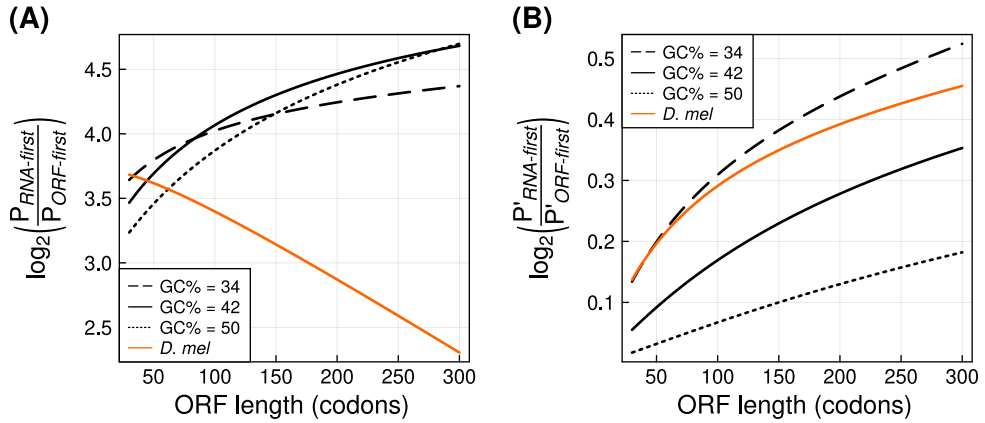

Figure 2: Proto-genes preferentially emerge RNA-first. The vertical axis shows the $\log_2$ transformed ratio of the probabilities of the RNA-first and the ORF-first trajectories, described as **(A)** a single-step process (Equations 5 & 6) and **(B)** a two-step process (Equations 7 & 8). A positive value suggests that RNA-first trajectory is more likely than ORF-first trajectory, and *vice versa*. In both panels, the horizontal axis denotes the number of codons in the ORF. Black lines denote the probability values estimated from overall GC content (34%, 42% and 50%) and the orange line denotes the probability values estimated from trimer and hexamer frequencies in *D. melanogaster* intergenic genome.

emerges first (RNA-first). Furthermore, emergence of an ORF is more likely when the proto-gene is already transcribed, and *vice versa* (Methods section 6). To understand whether *de novo* gene emergence has a preferred trajectory, we calculated two probabilities. First, the probability that an RNA exists and mutations cause a gain of an ORF but no disruption of transcription. This probability denotes the trajectory where RNA emerges first ($P_{RNA\text{-}first}$, Equation 5).

$$P_{RNA\text{-}first} = P_{RNA\text{-}stay} \times P_{ORF\text{-}gain\,|\,RNA} \tag{5}$$

The second probability ($P_{ORF\text{-}first}$, Equation 6) that denotes the trajectory where ORF emerges first, is the probability that an ORF already exists and mutations cause a gain of transcription but do not disrupt the ORF.

$$P_{ORF\text{-}first} = P_{ORF\text{-}stay} \times P_{RNA\text{-}gain\,|\,ORF} \tag{6}$$

We calculated the log-transformed ratio of $P_{RNA\text{-}first}$ and $P_{ORF\text{-}first}$, such that a positive value of the ratio would mean that ORF gain in an already transcribed DNA (RNA-first) is more likely than transcription gain for an untranscribed ORF (ORF-first). In other words, the RNA-first trajectory is more feasible. Likewise, a negative value of the ratio would suggest that the ORF-first trajectory is more feasible. We found that

proto-genes with ORFs of all the investigated lengths, preferentially emerge RNA-first (Figure 2A). More specifically, the likelihood of the RNA-first trajectory relative to that of the ORF-first trajectory increases with increasing ORF length. This is true for loci with three different values of GC content (34%, 42%, 50%). However, our complementary analysis using DNA oligomer frequencies from *D. melanogaster* intergenic genome shows that the likelihood of RNA-first trajectory relative to that of ORF-first trajectory, reduces with increasing ORF length. Nonetheless, RNA-first trajectory is still the more feasible trajectory of *de novo* gene emergence. In a more stringent scenario of *de novo* emergence, where a core promoter (TATA box or Inr) is necessary for transcription, proto-genes containing short ORFs can emerge ORF-first (Figure S2)

A more stringent definition of the RNA-first trajectory would describe a two-step process. In the first step, an RNA emerges in an untranscribed region of DNA but an ORF does not emerge. In the second step an ORF emerges, while transcription stays intact (Equation 7). Likewise, a ORF-first trajectory can be defined by a two-step probability. In the first step ORF emerges in an untranscribed DNA region, and in the second step transcription emerges, while the ORF remains intact (Equation 8).

$$P'_{RNA\text{-}first} = \underbrace{P_{RNA\text{-}gain} \times (1 - P_{ORF} - P_{ORF\text{-}gain})}_{first\ step} \times \underbrace{P_{RNA\text{-}stay} \times P_{ORF\text{-}gain|RNA}/(1 - P_{ORF})}_{second\ step} \quad (7)$$

$$P'_{ORF\text{-}first} = \underbrace{P_{ORF\text{-}gain} \times (1 - P_{RNA} - P_{RNA\text{-}gain})}_{first\ step} \times \underbrace{P_{ORF\text{-}stay} \times P_{RNA\text{-}gain|ORF}/(1 - P_{RNA})}_{second\ step} \quad (8)$$

Even with this stringent definition, *de novo* emergence via the RNA-first trajectory is more probable than via the ORF-first trajectory (Figure 2B). For all the three GC-content values, as well the specific analysis with *D. melanogaster* genome, the two-step RNA-first trajectory becomes increasingly more probable than the two-step ORF-first trajectory, with increasing ORF length.

Would extensive transcription loss suggest negative selection of toxic proteins?

Proto-genes that do not provide any fitness benefit to an organism may not be fixed in populations via natural selection. These genes may be lost due to mutation pressure. Some newborn proto-genes can also encode toxic proteins, that may aggregate or interfere with physiology in some other way. These genes would thus be eliminated

from the population genomes via negative selection. We note that an ORF is lost if the start codon is mutated, the stop codon is mutated to an amino acid encoding codon (non-stop mutation), or if an amino acid encoding codon is mutated to a stop codon (premature stop/non-sense mutation). However, it is likely that a non-stop mutation or a non-sense mutation, can still result in translation of a protein (extended or truncated, respectively). Furthermore, non-stop mutations can also lead to cellular toxicity (Choe *et al.*, 2016). Thus ORF loss does not ensure elimination of toxic proteins, which in turn suggests that transcription loss may more effectively inactivate the associated genes.

To understand which is the most probable mechanism of gene loss, we compared the probabilities of ORF loss ($P_{ORF\text{-}loss}$) and transcription loss ($P_{RNA\text{-}loss}$). We found that ORF loss is more probable than transcription loss, especially so when the ORFs are long (Figure 3A). If transcription is strictly dependent on a core promoter (TATA-box or Inr), small genes (30 codons) with low GC content (>34%) can be preferentially lost by loss of transcription (Figure S3). When the organismal fitness is affected only by the protein product of a proto-gene, then both RNA loss and ORF loss would affect the fitness equally. Thus the likelihood of gene extinction through ORF loss or RNA loss depends solely on the probability of these events.

Our analysis of gene loss mechanism considers the loss of one feature (for example transcription) without requiring the other feature to remain intact (for example the ORF). With more stringent analysis where we consider the loss of only one feature but not the other, we find a the that exclusive ORF loss is more likely than exclusive RNA

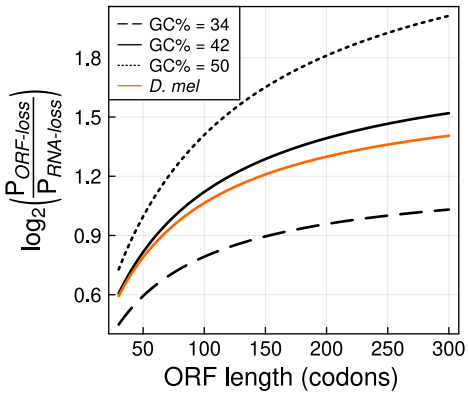

Figure 3: ORF loss in proto-genes is more probable than transcription loss. The horizontal axis denotes the number of codons in the ORF. The vertical axis shows the $\log_2$ transformed ratio of ORF loss and RNA loss probabilities, such that a positive value means ORF loss is more probable than RNA loss, and *vice versa*. Black lines denote the probability values estimated from overall GC content (34%, 42% and 50%) and the orange line denotes the probability values estimated from trimer and hexamer frequencies in *D. melanogaster* intergenic genome.

loss (Figure S5). The difference between the likelihood of the two gene loss mechanisms is more remarkable when we consider exclusive loss of one feature.

Overall, our analysis suggests widespread transcription loss may be indicative of a negative selection on protein expression.

Does mutation bias shape protein composition?

In previous sections, we showed that mutation bias affects the rate of *de novo* gene emergence and loss. We next turned our attention to whether this bias affects the composition (and thereby the chemistry) of proteins encoded by proto-genes. To this end, we first asked if the expected frequency of different amino acids is uniform. For a GC-content of 42% and with DNA trimer frequencies from *D. melanogaster* intergenic genome, we found that it is not uniform. Amino acids like leucine (L) and serine (S) have a higher expected frequency than other amino acids. On the other hand, amino acids like methionine (M) and tryptophan (W) are less probable (Figure 4A). This non-uniformity is to a great extent determined by number of degenerate codons for an amino acid. However, nucleotide composition also determines the frequency of an amino acid. For example, leucine, arginine and serine, all have six codons each. However, leucine is more likely to be encoded than serine in a random stretch of DNA, given a uniform GC-content of 42% (Figure 4A). The same is true for the frequencies of these amino acids estimated from the DNA trimer distribution. Overall, our analysis is roughly in agreement with many previous studies (Ohta and Kimura, 1971; Shen *et al.*, 2006; Gardini *et al.*, 2016). However, notable differences exist between our analysis and that of the previous studies because the latter primarily focused on characterized proteins and not random proto-gene derived protein sequences.

Next, we aimed to find out if some amino acid substitutions are more probable than the others. To this end, we calculated the gain probability (Equation 9) for every amino acid. Specifically for every amino acid x, we calculated the average probability that it substitutes any of the other nineteen amino acids, due to mutations. More precisely, the total probability that an amino acid y will mutate to an amino acid x is the product of the expected frequency of amino acid y (P_y), and the substitution rate from x to y ($\mu_{y\to x}$). The gain probability for an amino acid x is then defined as:

$$G_x = \sum_{b\neq a} P_y \times \mu_{y\to x} \tag{9}$$

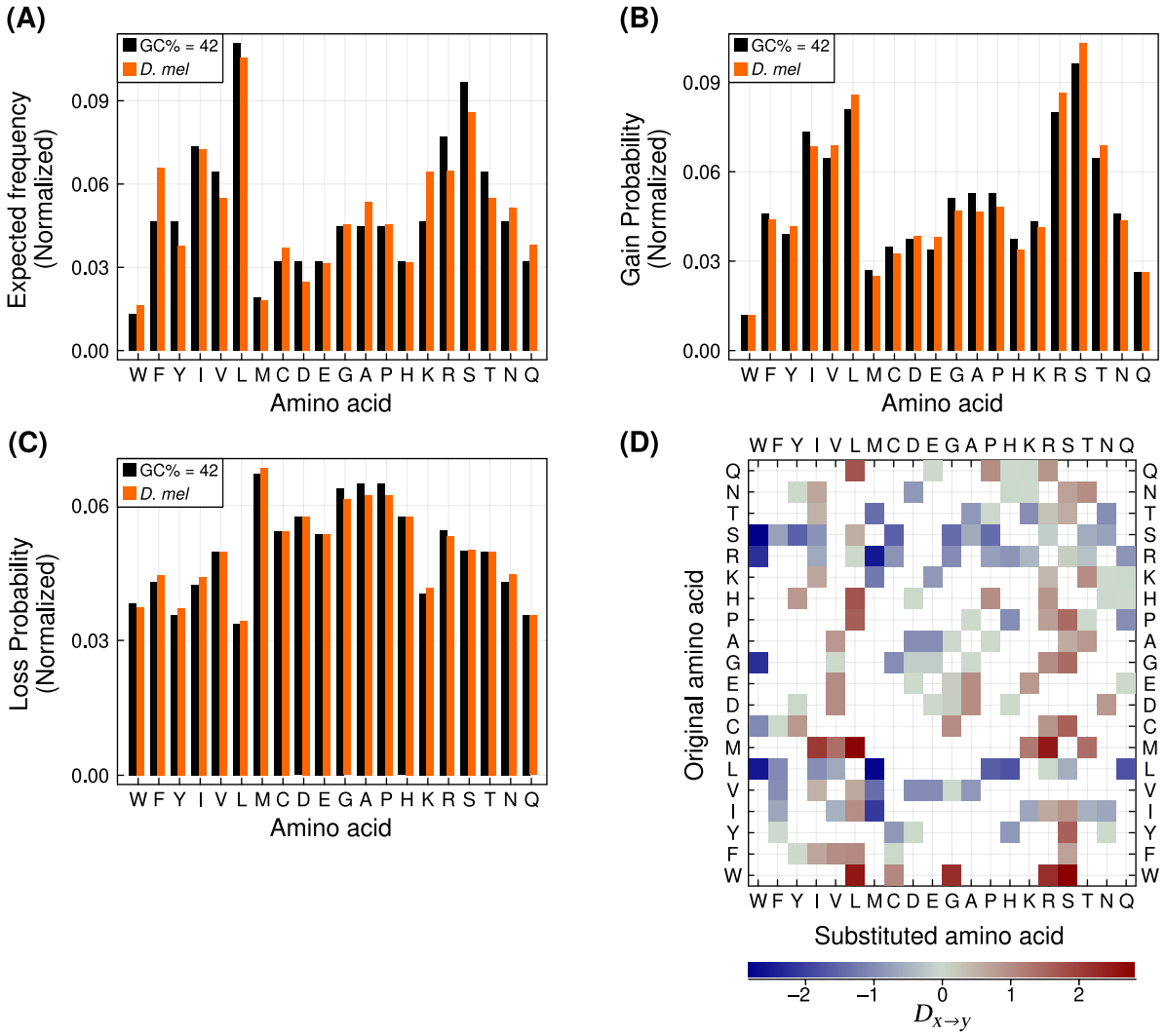

Figure 4: Mutation bias and GC-content shape protein composition. Panels **(A)**, **(B)** and **(C)** show expected frequency, gain probability (Equation 9), and loss probability (Equation 10) for different amino acids (horizontal axes). The vertical axis in all these three subplots show the normalized values of these measures such that their sum over all amino acids is equal to one. Panel **(D)** shows the directionality (colored tiles, $D_{x\to y}$, Equation 11) of the substitution from an amino acid (vertical axis) to another (horizontal axis). Positive values of $D_{x\to y}$ are shown in shades of red, whereas negative values are shown in shades of blue. We only show substitutions that can be realized by a single nucleotide mutation but we include all possible nucleotide mutations in the directionality calculations for these substitutions. In all panels, amino acids are ordered based on their chemical similarity (Kim *et al.*, 2009).

We found that amino acid gain probabilities were also non-uniform across different amino acids (Figure 4B) but they had a strong positive correlation with the expected frequency (GC = 42%, Pearson $\rho = 0.955, P < 10^{-10}$; *D. melanogaster* trimer frequencies, Pearson $\rho = 0.841, P < 10^{-10}$).

Next, we calculated how easily is an amino acid mutated to any other amino acid. Specifically, the conditional loss probability for an amino acid (L_x), given the amino

acid already exists, is defined as:

$$L_x = \sum_{x \neq y} \mu_{x \to y} \tag{10}$$

The loss probability was also not uniformly distributed (Figure 4C) but it did not significantly correlate with the expected frequency of the amino acids (GC = 42%, Pearson $\rho = -0.255, P = 0.2818$; *D. melanogaster* trimer frequencies, Pearson $\rho = -0.356, P = 0.1232$).

Next, we investigated if some amino acid substitutions are more common than the others. As expected, amino acid substitutions that require more than one nucleotide change were much less probable than those needing just a single nucleotide change. Therefore, we focused our analysis on amino acid substitutions whose total probabilities were more than that of a double nucleotide mutation (note that this procedure identifies amino acid substitutions that are possible via single nucleotide mutations but it does not exclude multinucleotide mutations from the calculation of amino acid substitution probabilities). Although single nucleotide mutations are more probable than double mutations, the different single mutations do not occur at the same rate due to mutation rate bias. Thus an amino acid substitution (for example K→E) may not be as likely to occur as the reverse substitution (E→K). That is substitutions between a pair of amino acids may have a directionality. Although many previous studies have explored the likelihood of different amino acid substitutions (Ohta and Kimura, 1971; Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992; Gonnet *et al.*, 1992; Jones *et al.*, 1992; Whelan and Goldman, 2001; Schneider *et al.*, 2005; Kosiol *et al.*, 2007; Le and Gascuel, 2008), none to our knowledge have focused primarily on the directionality of these substitutions. To verify if such a directionality exists, we calculated the log transformed ratio of forward and reverse substitution probability, $D_{x \to y}$, for every pair of amino acids, x and y. A positive value of $D_{x \to y}$ means that x to y substitution is more likely than y to x substitution, and *vice versa*.

$$D_{x \to y} = \log_2 \left(\frac{\mu_{x \to y}}{\mu_{y \to x}} \right) \tag{11}$$

As we mentioned in the previous paragraph, we restricted this analysis to substitutions that can be achieved via single nucleotide mutations. We found that most substitutions indeed had a preferred direction ($D_{xy} \neq 0$, Figure 4D). The median absolute value of D_{xy} was 0.905 and the maximum value was 2.836. That is, in 38 out of 75 amino acid

pairs we analysed, one direction of substitution was at least 1.87 times more probable than the other. For example, E→K substitution is 1.8 times more probable than K→E substitution. This may appear as a small number but many such unidirectional substitutions occurring in an evolving protein could indicate a directional selection. We note that that actual frequencies of specific substitutions depend on the initial distribution of different codons in the ORF sequence.

A previous study found that mutations that cause hydrophobic amino acids to appear on protein surfaces, can lead to evolution of protein dimers (Hochberg *et al.*, 2020). The authors also suggested that such an evolutionary process may be widespread because mutation bias tends to facilitate emergence of hydrophobic amino acids. They based this argument on another study that estimated mutation rate bias in bacteria (Hershberg and Petrov, 2010). The authors (Hochberg *et al.*, 2020) further suggested that hydrophobic amino acids are more frequently found in random protein sequences.

We asked if our model also makes similar predictions. To this end, we first calculated the percentage of amino acids in a protein sequence that are expected to be hydrophobic. A popular hydrophobicity scale of amino acids is based on solubility of an amino acid in water or ethanol (Kyte and Doolittle, 1982). However, this scale does not classify tryptophan as a hydrophobic amino acid. Another hydrophobicity scale estimated by a different study (Wimley and White, 1996) is more biologically realistic and is based on rates of transfer of different amino acids between water and a hydrophobic medium. This study analysed two hydrophobic media – a lipid bilayer and octanol, and also considered the effect of peptide bonds in the calculations. Therefore we used the hydrophobicity scale by Wimley and White (1996) for classifying amino acids such that hydrophobic amino acids have a negative free energy change of transfer from water to hydrophobic media. Conversely, a hydrophilic amino acid would have a positive value of the free energy change. Based the octanol hydrophobicity scale, we classified the following amino acids as hydrophobic – cysteine (C), phenylalanine (F), isoleucine (I), leucine (L), methionine (M), valine (V), tryptophan (W) and tyrosine (Y). With the lipid bilayer hydrophobicity scale, valine is not classified as hydrophobic, possibly because it does not integrate well in the bilayer. Therefore we used the octanol scale for our calculations.

We found that ∼40% of amino acids in a protein sequence are expected to be hydrophobic (with a GC-content of 42% as well as with *D. melanogaster* trimer distribution). This finding is broadly in agreement with the previous study (Hochberg *et al.*, 2020). We note that cytosolic proteins may require more than 42% of their constituent amino acids to be hydrophobic, in order to fold efficiently (Dill, 1985). Next, we asked if a protein

sequence in general tends to be hydrophobic. Specifically, we calculated expected hydrophobicity ($\bar{\alpha}$, Equation 12) of a protein sequence, based on expected frequency of different amino acids (P_x, Figure 4) and their hydrophobicity values (α_x).

$$\bar{\alpha} = \sum_x P_x \alpha_x \tag{12}$$

We found that the expected hydropathy is higher than zero (0.38 with GC% = 42; 0.36 with *D. melanogaster* trimer frequencies), which suggests that random protein sequences are not on an average, hydrophobic.

To find out if random mutations indeed cause hydrobphobic amino acids to occur in protein sequences, we calculated the probability that any mutation substitutes a non-hydrophobic amino acid with a hydrophobic amino acid ($P_{\alpha\text{-}gain}$, Equation 13). Likewise, we calculated the probability that any mutation causes a hydrophobic to non-hydrophobic amino acid substitution ($P_{\alpha\text{-}loss}$, Equation 14).

$$P_{\alpha\text{-}gain} = \sum_{x \mid \alpha_x > 0} \sum_{y \mid \alpha_y \geq 0} P_y \times \mu_{y \to x} \tag{13}$$

$$P_{\alpha\text{-}loss} = \sum_{x \mid \alpha_x > 0} \sum_{y \mid \alpha_y \geq 0} P_x \times \mu_{x \to y} \tag{14}$$

Next, we calculated the ratio of $P_{\alpha\text{-}gain}$ and $P_{\alpha\text{-}loss}$, and found that it is slightly more than 1 (1.114 with GC% = 42; 1.015 with *D. melanogaster* trimer frequencies). This suggests that gain of a hydrophobic amino acid is slightly more probable than its loss. We performed a complementary analysis where we calculated the average change in hydrophobicity due to any random mutation ($\bar{\Delta\alpha}$), defined by:

$$\bar{\Delta\alpha} = \frac{\sum_x \sum_y \mu_{x \to y} \times (\alpha_y - \alpha_x)}{\sum_x \sum_y \mu_{x \to y}} \tag{15}$$

We found the average change in hydrophobicity to be slightly less than zero (-0.0798), which suggests that on an average mutations cause a very small increase in hydrophobicity. We reiterate that hydrophobic amino acids have a negative value hydrophobic-

ity and thus a negative change in hydrophobicity also denotes a shift towards a more hydrophobic protein. To focus exclusively on non-hydrophobic to hydrophobic amino acid substitutions, we excluded amino acid pairs where both amino acids are either hydrophobic or are non-hydrophobic. That is, we only analysed pairs of amino acids that have opposite signs of hydrophobicity. With this focused analysis, we find that the average change in hydrophobicity is comparable (-0.0793). For both the analyses, we note that the actual change in hydrophobicity will depend on the composition of the ancestral protein sequence.

Overall our analyses show that random mutations may preferentially cause hydrophobic amino acids to accumulate in protein sequence but this effect may be very small.

Discussion

In this work, we addressed some fundamental questions about *de novo* emergence of protogenes. Broadly, we asked how quickly these genes can emerge and be lost, whether their birth and death has a preferred trajectory of mutational events, and how the composition of their protein sequence is determined. To this end, we developed mathematical models and used them to address specific questions. These models are based on mutation probability and mutation rate bias, and represent neutral evolution where DNA sequences are not evolving under natural selection. Thus our models provide an opportunity to test the alternative hypothesis that some proto-genes are evolving under selection.

We show that when a protein product of a proto-gene doesn't affect organismal fitness, the gene can be lost much more rapidly than they are gained, such that it can be independently lost two to three times in the time required for it to emerge (Figure 1A). In a hypothetical case, a proto-gene may be present in a population (A) but is absent in two other geographically isolated populations (B and C). Two evolutionary scenarios could explain this observation. In the first scenario, the proto-gene simply emerged for the first time in population A. The second scenario posits that the proto-gene was present in the ancestral population but was lost in both population B and population C. Our models suggest that the second scenario is more likely than is assumed. This inference can be further extended to the level of closely related species where the divergence times are relatively shorter than the timescale of multiple gain and loss events within a lineage. Future phylogenetic studies on proto-genes should consider this finding for inferring the dynamics of gene gain and loss. Proto-genes that encode proteins that are beneficial to an organism's fitness are likely to be protected in the genome against gene

loss (Lee *et al.*, 2019). Our models indeed show that loss of even moderately beneficial genes is unlikely in large populations (Figure 1B).

Within the limitations of our model assumptions, we answer the long standing question of which trajectory of mutational events leads to *de novo* emergence. That is, whether ORF emerges first or transcription emerges first (Schmitz and Bornberg-Bauer, 2017). We show that, in the absence of selection, genes are likely to emerge in existing transcripts (transcription emerges first, Figure 2). Long non-coding RNAs may be a potential source of proto-genes but this evolutionary process may be constrained by a different set of factors. A previous study has shown that new transcripts frequently arise in regions overlapping existing genes. These new transcripts can harbor ORFs which in turn leads to *de novo* gene emergence (Blevins *et al.*, 2021). This is an example of RNA-first trajectory. However, the ORFs present in these genes may be under indirect selection pressure that acts on the overlapping gene sequences. Therefore, they may be less frequently mutated or lost.

Although our model is based on the assumption of neutral evolution, one can speculate the effects of positive selection on the proto-gene. For example, if an ORF encodes a protein that is beneficial to the organism, then selection would cause the fixation of the gene as soon as the it emerges via RNA-first trajectory. On the other hand, the selection would have no effect on an ORF that is not transcribed, as long as it is does not harbor a functional DNA sequence (such as regulatory elements), and is not under indirect selection (for example, if it overlaps with an existing gene). The ORF may mutate before transcription emerges and the corresponding mutated protein may not be beneficial anymore. Thus positive selection may further increase the likelihood of the RNA-first trajectory.

We show that transcription loss is less likely than the loss of long ORFs (Figure 3). However, it may be difficult to find direct evidence for such a phenomenon. Specifically, it is difficult to infer from data if transcription was lost or never emerged. If the protein encoded by an ORF is indeed toxic, it would be silenced or purged rapidly in microevolutionary time scales. One way of inferring negative selection from transcription loss may involve correlation of ORF sequence variation and transcription status. Mutations within an ORF sequence are much more probable than mutations that disrupt transcription or an ORF. This is so because the entire ORF sequence can have a higher number of mutable sites than a promoter. If an ORF variant is untranscribed in many populations (species or taxa), it is possibly toxic.

A large volume of work exists on the likelihood of amino acid substitutions and many of these models are widely used in phylogenetics studies (Ohta and Kimura, 1971;

Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992; Gonnet *et al.*, 1992; Jones *et al.*, 1992; Whelan and Goldman, 2001; Schneider *et al.*, 2005; Kosiol *et al.*, 2007; Le and Gascuel, 2008). However, commonly used amino acid substitution models such as PAM and BLOSUM (Dayhoff *et al.*, 1978; Henikoff and Henikoff, 1992) do not focus on directionality of mutations (the substitution matrices are symmetric). Our aim was to highlight how proteins can evolve under mutation pressure and that mutation bias causes some substitutions to occur more frequently than the others. Furthermore, we do not base our substitution probability calculations on known protein sequences but rather on random protein sequences that are more likely to to be encoded in proto-genes. Using our models, we show that for a pair of amino acids, one direction of substitution can be significantly more likely than the other. If a substitution occurs more frequently than it is expected to occur in our neutral model, then directional evolution can be a possible explanation for the observation. We also show that random mutations can lead to a small increase in a protein's hydrophobicity. Hydrophobic amino acids can facilitate folding (Dill, 1985), but can also cause proteins to aggregate (Hochberg *et al.*, 2020). Thus the effect of hydrophobicity increasing mutations on a protein's stability and toxicity, is dependent on where they occur in the protein structure.

Like all theoretical models, our models are not without limitations. First, we assume that nucleotides are uniformly distributed according to GC-content. This is not true as nucleotide composition can vary significantly across the genome. Since our model primarily focuses on gene emergence at one locus, it can be used to predict evolutionary probabilities for different genomic loci. Furthermore, several predictions of our models are not qualitatively affected by GC-content. More realistic nucleotide composition distribution estimated from *D. melanogaster* intergenic sequences, also qualitatively agree with GC-content based predictions. Thus, most predictions are qualitatively robust to small differences in parameters. Second, in our models, we assumed uniform mutation rate throughout the genome. However, recent studies have shown that mutation rates are also not uniform across the genome, and can vary significantly (Monroe *et al.*, 2022). Our models can be adapted to investigate specific genomic loci, where estimates of the local nucleotide composition and mutation rate can more accurately determine the *de novo* gene evolution in these loci. Third, our models are based on parameters such as mutation rate and mutation bias, which we obtained from published data on human and *D.melanogaster* genomes (Zhang and Gerstein, 2003; Schrider *et al.*, 2013). The accuracy of the models' prediction will depend on the accuracy of parameters. However, mutation rate itself only determines the values of different probabilities. Our results, which are based on probability ratios, will not significantly change with a different value of mutation rate. Mutation bias may have more significant effect on the predictions than mutation rate. Existing data suggest that different species have a different

spectrum of mutations (Cano *et al.*, 2022), especially for species that are highly distant from each other. Thus our inferences based on our chosen parameters will not universally apply to all species. Fourth, we assume that occurrence of a point mutation is independent of another mutation. This is not the case for tandem mutations, where multinucleotide mutations occur more frequently than their expected frequency under the assumption of mutational independence (Harris and Nielsen, 2014). Fourth, we did not specifically model the effect of many regulatory sequence features such as enhancers, transcription factor binding sites, 5′ and 3′ untranslated regions, and Kozak consensus sequence, that can influence the evolution of proto-genes. For example *de novo* emergence is very likely to occur near enhancers (Majic and Payne, 2020). To model this phenomenon in our analyses, we assumed that DNA loci feasible for transcript emergence are close to known *cis*- regulatory elements (Gallo *et al.*, 2005; Halfon *et al.*, 2007; Gallo *et al.*, 2010; Rivera *et al.*, 2018; Keränen *et al.*, 2022). Specifically, we assumed that transcription can be initiated at a probability that is equal to the proportion of such regulatory elements in the intergenic genome. This is a rough estimate but it can be updated when more specific data becomes available. For example, ATAC-seq data can provide reliable estimates of nucleosome depleted regions, that can act as transcription initiation sites (Buenrostro *et al.*, 2013). This analysis can be supplemented with data from techniques like native elongating transcript sequencing (NET-seq; Churchman and Weissman, 2012) and global run-on sequencing (GRO-seq; Lopes *et al.*, 2017), that can detect nascent transcripts. Our modeling framework also allows incorporation of any number of sequence features in the calculations as long as their sequences can be defined (Methods section 3). Finally, we ignore insertions, deletions and transpositions as mechanisms of mutation. Insertions and deletions (indels) are reported to be less frequent than substitutions (Schrider *et al.*, 2013), and both indels and transpositions cause larger sequence alterations than point mutations. Our modeling methodology also allows incorporation of these mechanisms as a source of mutations.

A famous aphorism about theoretical models says "all models are wrong but some are useful" (Box, 1979). Our models are not an exception. They may not be 100% accurate but they make useful predictions that can lead to a more focused experimental validation. Many of our predictions are qualitatively robust to small parameter changes. Some predictions are also in agreement with experimental data. For example, our prediction of transcript gain rate is only slightly inflated than the estimated rates using an empirical analysis of *D. melanogaster* populations (Grandchamp *et al.*, 2023). This difference could have possibly resulted from our relaxed requirements for a promoter, or other unknown factors that may have affected population evolution (for example, the actual effective population size and environmental conditions; Lange *et al.*, 2021).

When supplemented with experimental data analysis our models can be made more complex to accommodate diverse molecular mechanisms driving gene evolution, and can provide more accurate predictions. Therefore our work opens up an opportunity for theoretical and computational biologists to design analyses using our modeling framework, that more accurately describe their system of interest.

Materials and Methods

1 Mutation probabilities

Substitution	Probability(μ)
A:T→T:A	0.056
A:T→G:C	0.243
A:T→C:G	0.074
G:C→A:T	0.483
G:C→T:A	0.075
G:C→C:G	0.069

Table 1: Mutation bias probabilities for different nucleotide mutations. A:T denotes an A-T base pair in a double stranded DNA. Thus A→G mutation on one DNA strand would cause a T→C mutation on the complementary strand. We describe the other mutations in the same way.

We calculated nucleotide substitution probabilities based on mutation rate and mutation rate bias data. Specifically, we used a mutation rate (λ) of 7.8×10^{-9} mutations per nucleotide position per generation (Schrider *et al.*, 2013). We derived our mutation bias parameters from two published studies, the first on *Drosophila melanogaster* (Schrider *et al.*, 2013), and the second on humans (Zhang and Gerstein, 2003). Table 1 shows the exact values of mutation bias probabilities that we used in this study.

2 Probabilities of finding, gaining, and losing a DNA sequence

2.1 Probability of finding a DNA sequence

We calculated the probability of finding a DNA sequence based on global nucleotide frequency distributions, given by the GC-content. Specifically, the probability of finding either a G or a C is:

$$S = \frac{0.5 \times \text{GC\%}}{100}$$

The probability of finding an A or a T is: $W = 0.5 - S$. Using these values, we calculated the probability of finding a DNA sequence motif by chance. For example, the probability of finding the sequence ATG would be: $W \times W \times S$.

We also estimated the probability of finding specific DNA sequences in a reference genome. Specifically, we calculated the frequencies of all 64 trimers and all 4096 hexamers in the genomic regions of *D. melanogaster* that exist in open chromatin, and do not contain any known genes or regulatory elements (see Methods section 5).

2.2 Probability of gaining a DNA sequence

We calculated the probability of gaining a DNA sequence due to mutations using GC-content, mutation rate and mutation bias. Specifically, we calculated the probability that a DNA sequence does not exist, and it emerges due to specific nucleotide mutations. More precisely, this probability is the product of two other probabilities. The first is the probability of finding a DNA sequence (x) that is not the sequence of interest. The second probability is that this sequence x mutates to the sequence of interest. To explain this calculation better, we use the example of CTA mutating to ATG. The first probability, that is the probability of finding CTA by chance is SW^2. CTA mutates to ATG via two nucleotide mutations (C→A and A→G). Thus the probability of this DNA change would be the probability of two nucleotide mutations (λ^2) multiplied by two mutation bias probabilities (G:C→T:A and A:T→G:C). Overall, the chance of CTA mutating to ATG would be:

$$SW^2 \times \lambda^2 \times \mu_{\text{G:C}\to\text{T:A}} \times \mu_{\text{A:T}\to\text{G:C}}$$

Next, we calculated the probability that every nucleotide triplet that is not ATG, mutates to ATG. This can happen via one, two, or three nucleotide mutations. The sum of all these mutation probabilities is the probability of ATG gain.

Using the same principle we calculated the gain probability of any DNA sequence motif (of any length or composition). We excluded insertions and deletions as a mechanism of gain of small DNA sequences that we analysed in this study.

2.3 Probability of losing a DNA sequence

We calculated the loss of a DNA sequence motif using the same principle we used for calculating gain probabilities. However, we defined loss probability as a conditional probability, that is we assume that the DNA sequence of interest already exists in the genome. For example, the loss probability of a specific ATG sequence would be the sum of probabilities of ATG mutating to any of the other 63 nucleotide triplets (via one, two, or three nucleotide mutations). We use conditional loss probabilities by default, because usually one is interested in finding out how quickly an existing DNA sequence can erode.

We used this method to calculate the loss probability of any DNA sequence motif, and we excluded insertions and deletions from this calculation.

3 Probabilities of finding, gaining, and losing DNA features

Usually, a specific function is encoded in DNA by several DNA sequences. For example, translation stop is encoded by three codons (TGA, TAG, TAA). We use the term DNA features to mean a set of DNA sequences that are associated with the same function. For every such DNA feature set, there is a complementary set of non-features, that is DNA sequences that are not associated with the feature's function. For example, the non-feature set of stop codons would be all the other 61 codons.

The probability that a DNA feature exists, is the sum of probabilities of every DNA sequence in that set (Methods section 2.1).

The probability that a DNA feature is gained via mutations, is the sum of probabilities of every non-feature sequence mutating to any feature sequence. If F denotes the feature set, and $\mu_{y \to x}$ denotes the probability of a DNA sequence y mutating to a DNA sequence x (see Methods section 2.2), then:

$$P_{\textit{feature-gain}} = \sum_{x \in F} \sum_{y \notin F} P_y \times \mu_{y \to x} \tag{16}$$

The probability that a DNA feature is lost via mutations is a conditional probability that given a feature exists, it mutates to any of the non-feature sequences.

$$P_{\textit{feature-loss}} = \frac{\sum_{x \in F} \sum_{y \notin F} P_x \times \mu_{x \to y}}{\sum_{x \in F} P_x} \tag{17}$$

Because a feature set usually has many DNA sequences, a mutation can change a feature sequence such that the resulting sequence is also a part of the feature set. Thus we defined the probability ($P_{\textit{feature-stay}}$) that a feature does not erode because of mutations. Specifically, it is the sum of two probabilities. First is the probability that no mutation occurs in the DNA sequence (P_0), and the second probability describes the event where the mutated sequence remains a part of the feature set.

$$P_{\textit{feature-stay}} = P_0 + \sum_{x \in F} \sum_{\substack{y \in F \\ y \neq x}} \mu_{x \to y} \tag{18}$$

The probability that no mutation occurs (P_0) in a DNA sequence of length k is described by Poisson distribution.

$$P_0 = 1 - e^{-k\lambda}$$

Because the mutation rate is biased (Table 1), the probability that no mutation occurs in a DNA sequence depends on its composition.

The probability that an `A` or a `T` mutates ($\lambda_{\texttt{AT}}$), is thus described as:

$$\lambda_{\texttt{AT}} = 2 \times \lambda \times (\mu_{\texttt{A:T}\to\texttt{T:A}} + \mu_{\texttt{A:T}\to\texttt{G:C}} + \mu_{\texttt{A:T}\to\texttt{C:G}})$$

Likewise, the probability that a `G` or a `C` mutates ($\lambda_{\texttt{GC}}$) is:

$$\lambda_{\mathtt{GC}} = 2 \times \lambda \times (\mu_{\mathtt{G:C}\to\mathtt{A:T}} + \mu_{\mathtt{G:C}\to\mathtt{T:A}} + \mu_{\mathtt{G:C}\to\mathtt{C:G}})$$

(Note that the general mutation rate, λ, is an average of $\lambda_{\mathtt{AT}}$ and $\lambda_{\mathtt{GC}}$.)

Thus the probability that a sequence of length k, containing m number of `A` and `T`, does not mutate is:

$$P_0 = (1 - e^{-m\lambda_{\mathtt{AT}}}) \times (1 - e^{-(k-m)\lambda_{\mathtt{GC}}})$$

We also calculated all the above-defined probabilities ($P_{\textit{feature-gain}}$, $P_{\textit{feature-loss}}$ and $P_{\textit{feature-stay}}$), using DNA trimer and hexamer frequencies from *D. melanogaster*. In this case the probability of finding a nucleotide sequence depends on the trimer/hexamer distributions instead of GC-content, but the probability of mutational changes are only dependent on mutation bias. Trimers and hexamers would contain codons and polyA signals, respectively.

4 Probabilities of finding, gaining, and losing an ORF

4.1 Probability of finding an ORF

A reading frame is a nucleotide sequence with a length that is a multiple of three. A reading frame that begins with a start codon (`ATG`), and ends with one of the three stop codons (`TAG`, `TGA`, `TAA`) is an open reading frame (ORF). This necessarily means that there are no stop codons within the sequence. Thus the probability of finding an ORF containing k codons including start and stop codons (P_{ORF}) is:

$$P_{ORF}(k) = P_{ATG} \times P_{stop} \times (1 - P_{stop})^{k-2} \tag{19}$$

Here, P_{ATG} and P_{stop} are the probabilities of finding a start codon, and a stop codon by chance, respectively.

4.2 Probability of gaining an ORF

As we defined in the previous section, an ORF has three requirements (start codon, stop codon, and no premature stop codon in the sequence). Thus an ORF can emerge due to mutations via three mechanisms. In each of these mechanisms, one requirement is initially absent whereas the other two are present. Then mutations cause the missing requirement to emerge while not disrupting the other two requirements. More specifically, the ORF can be gained via the following three mechanisms:

1. Gain of a start codon ($P_{\textit{ATG-gain}}$) while a stop codon continues to exist at the end of a reading frame ($P_{\textit{stop-stay}}$), and there is no emergence of stop codon within the reading frame ($1 - P_{\textit{stop}} - P_{\textit{stop-gain}}$).

2. Gain of a stop codon ($P_{\textit{stop-gain}}$), while a start codon continues to exist at the beginning of a reading frame ($P_{\textit{ATG-stay}}$), and there is no emergence of stop codon within the reading frame.

3. Loss of a premature stop codon, at any of the $k-2$ codon positions within the reading frame ($P_{\textit{stop-gain}}$). At the same time start and stop codons remain undisrupted by mutations, and no stop codon emerges at any of the other $k-3$ positions.

Thus we define the probability of ORF gain ($P_{\textit{ORF-gain}}$) as:

$$\begin{aligned} P_{\textit{ORF-gain}}(k) = \quad & P_{\textit{ATG-gain}} \times P_{\textit{stop-stay}} \times (1 - P_{\textit{stop}} - P_{\textit{stop-gain}})^{k-2} \\ & + P_{\textit{ATG-stay}} \times P_{\textit{stop-gain}} \times (1 - P_{\textit{stop}} - P_{\textit{stop-gain}})^{k-2} \\ & + P_{\textit{ATG-stay}} \times P_{\textit{stop-stay}} \times P_{\textit{stop-loss}} \times (k-2) \times (1 - P_{\textit{stop}} - P_{\textit{stop-gain}})^{k-3} \end{aligned} \tag{20}$$

4.3 Probability of ORF loss

ORF can be lost when any of its three requirements are lost. We thus define the conditional probability of ORF loss as:

$$P_{\textit{ORF-loss}}(k) = P_{\textit{ATG-loss}} + P_{\textit{stop-loss}} + (k-2) \times \frac{P_{\textit{stop-gain}}}{1 - P_{\textit{stop}}} \tag{21}$$

The last term in this equation describes the conditional probability of stop-gain, given the assumption that no stop codon exists within the ORF.

4.4 Probability that ORF remains intact

We assumed that an ORF of a certain length remains intact if none of the necessary features are lost. However, the ORF sequence can mutate to cause non-synonymous changes in the translated protein sequence. This condition applies to all the three ORF probabilities described above. We define the probability that an ORF remains intact ($P_{ORF\text{-}stay}$) as:

$$P_{ORF\text{-}stay}(k) = P_{ATG\text{-}stay} \times P_{stop\text{-}stay} \times (1 - P_{stop} - P_{stop\text{-}gain})^{k-2} \tag{22}$$

5 Probabilities of finding, gaining, and losing transcription

5.1 Probability of transcription

In this study we defined a transcription model that is not dependent on specific promoter sequences. Instead we defined a probability that transcription is initiated at a genomic locus. To this end, we first extracted the genomic coordinates of intergenic sequences in *D. melanogaster* genome (release 6.49, FlyBase; Gramates *et al.*, 2022). Using the data on the genomic locations of open chromatin in *D. melanogaster* (Milon *et al.*, 2014), we obtained open chromatin regions within intergenic genome. Next, we calculated the fraction of this open-intergenic genome occupied by *cis*- regulatory elements (Gallo *et al.*, 2005; Halfon *et al.*, 2007; Gallo *et al.*, 2010; Rivera *et al.*, 2018; Keränen *et al.*, 2022). A significant proportion of these elements include enhancers. Enhancers, that can promote transcription in their neighboring genomic regions (Andersson and Sandelin, 2020), are also known to promote *de novo* gene emergence (Majic and Payne, 2020). Thus we reasoned that proportion of *cis*- regulatory elements in open intergenic chromatin would roughly approximate the likelihood of transcription initiation ($P_{RNAinit}$). We do not model the exact sequences because data is not yet available on exact sequence requirements and mutational robustness of enhancers and other *cis*-regulatory elements. Thus poly-A signal is the primary sequence determinant of transcription probabilities. Relatedly, open intergenic chromatin regions, that are not interrupted by *cis*- regulatory elements are the potential sites for *de novo* gene emergence.

A poly-A signal marks the end of a transcript. The transcribed DNA region should thus have only one poly-A signal at its end. A transcribed DNA of length l nucleotides, contains $l-5$ hexamers (6nt long sequences). Each of these hexamers can be a poly-A signal with a probability, P_{polyA}. Probability that none of these hexamers is a poly-A signal is $(1-P_{polyA})^{l-5}$. Since we focus on protein coding genes in this work, we describe the length of transcript in terms of the length of protein it can encode. Assuming that untranslated regions have no effect on protein synthesis, the length of a transcript in number of nucleotides should be at least three times the length of the encoded protein in number of amino acids. Thus the probability that a DNA region produces a transcript that can harbor an ORF with k codons ($P_{RNA}(k)$) is:

$$P_{RNA}(k) = P_{RNAinit} \times P_{polyA} \times (1-P_{polyA})^{3k-5} \tag{23}$$

5.2 Probability of gaining transcription

As we mentioned in the previous section, two requirements need to be met to produce a transcript of any given length. First, a poly-A signal needs to be present at the end of the DNA region to be transcribed, and second, no poly-A signal should exist within this DNA region. Transcription can be gained if one of these requirements are already met and the third required feature emerges due to mutations. At the same time mutations should not destroy the existing features. It is possible that both the required features are missing and they emerge simultaneously due to mutations. However such an event is highly improbable. Thus transcription gain can occur via two mechanisms. In the first mechanism a poly-A signal is gained at the end of a sequence ($P_{polyA\text{-}gain}$), while no poly-A sites are present at any $l-5$ (or $3k-5$) sites within the sequence, and none emerge ($P_{nopolyA\text{-}stay} = 1 - P_{polyA} - P_{polyA\text{-}gain}$). In the second mechanism, a poly-A signal is present at the end of the sequence, and is not lost due to mutations. Meanwhile, one poly-A signal at any of the $3k-5$ sites within the sequence is lost ($P_{polyA\text{-}loss}$), whereas the other $3k-6$ sites do not encode a poly-A signal and continue to remain so.

The combined probability of these two transcription gain mechanism is defined as:

$$P_{RNA\text{-}gain}(k) = P_{RNAinit} \times \begin{pmatrix} P_{polyA\text{-}gain} \times (P_{nopolyA\text{-}stay})^{3k-5} \\ + P_{polyA\text{-}stay} \times (P_{nopolyA\text{-}stay})^{3k-6} \times (3k-6) \times P_{polyA\text{-}loss} \end{pmatrix} \tag{24}$$

5.3 Probability of losing transcription

Transcription is lost when either the poly-A signal is lost or if one is gained within the transcribed sequence. The probability of transcription loss ($P_{RNA\text{-}loss}$), given that already transcription exists, is thus defined as:

$$P_{RNA\text{-}loss}(k) = P_{polyA\text{-}loss} + (3k - 5) \times P_{polyA\text{-}gain} \tag{25}$$

5.4 Probability that transcription remains intact

Transcription remains intact if neither the poly-A signal is not lost and none is gained within the sequence. Probability of this event ($P_{RNA\text{-}stay}$) is defined as:

$$P_{RNA\text{-}stay}(k) = P_{RNAinit} \times P_{polyA\text{-}stay} \times (P_{nopolyA\text{-}stay})^{(3k-5)} \tag{26}$$

6 Interdependence of transcription and ORF probabilities

All the four poly-A signal variants contain a `TAA` in their sequence (position 3 – 5), which is also a stop codon. Since stop codons cannot exist inside an ORF, presence of an ORF reduces the possible number of poly-A signal sites in a DNA region. Specifically, a poly-A signal can exist in any of the three reading frames but it can overlap with the ORF only in one frame. Thus, out of $3k - 5$ positions for a 6-mer (length of a poly-A signal) in an ORF with k codons, k positions cannot harbor a poly-A signal (Equation 23). Furthermore a poly-A signal cannot overlap with the start codon, thus reducing two more positions from all sites where a poly-A signal can exist (one position is already counted – if a poly-A signal overlaps with the first codon in the second frame, then the second codon becomes a stop codon). Overall, the possible number poly-A sites on an ORF with k codons is $2k - 5$. Thus for a gene to emerge via gain of transcription, the DNA should not contain poly-A sites at these $2k - 5$ positions. As a consequence, the probability of transcription gain for a specific transcript length, is higher when the DNA region contains an ORF than when it does not (Equation 24). We have ignored the probability of poly-A signal overlapping with a stop codon because it makes a very small difference in the probability of RNA gain. This is so because at

least one stop codon (TAA) allows a poly-A signal to overlap with it in three frames.

Conversely to the effect of an existing ORF on transcription gain, stop codons are less probable in an existing transcript that does not contain any poly-A signal sequences within its sequence, than in an untranscribed DNA region. Specifically, the probability that a 6-mer contains a stop codon in the third position is same as the probability of finding a stop codon (P_{stop}). If poly-A signals are excluded from these 6-mers, then the probability of finding a stop codon is $P_{stop} - P_{poly\text{-}A}$. Conversely, lack of poly-A signals will increase the probability of amino acid coding codons and thereby their probability of mutating into a stop codon. Furthermore, the likelihood that an ORF is undisrupted my premature nonsense mutations, is lower when the DNA is already transcribed. Specifically:

$$P_{stop\text{-}gain\,|\,RNA} = \frac{P_{stop\text{-}gain} \times (1 - P_{stop} + P_{poly\text{-}A})}{1 - P_{stop}}$$

Overall the evolutionary dynamics of transcription and ORF gain, are not independent of each other. That is, presence of one feature makes the gain of the other more likely. The dependence becomes more prominent with increasing number of codons in the ORF of the proto-gene.

7 Calculation of fixation probabilities

We estimated gene fixation probabilities using Kimura's model (Kimura, 1962). Specifically we defined the probabilities of fixation and extinction of a proto-gene as shown in Table 2:

Case	Gain of gene (Fixation, P_{fix})	Loss of gene (Extinction, P_{ext})	Selection coefficient (s)
No selection on protogene	$\frac{1}{2N_e}$	$\frac{1}{2N_e}$	–
Selection on protogene	$\frac{1-e^{-2s}}{1-e^{-4N_e s}}$	$\frac{1-e^{2s}}{1-e^{4N_e s}}$	± 0.001

Table 2: Gene fixation/extinction probabilities with or without selection. Here N_e denotes effective population size and s denotes the selection co-efficient denoting the fitness advantage provided by the protein encoded in the proto-gene (positive valued for a beneficial protein, and negative valued for a deleterious protein).

8 Probability of amino acid substitutions

We calculated the probability of amino acid substitutions based on nucleotide substitution in the corresponding codons. In this case, multiple codons that code for the same amino acid constitute a feature set (Methods section 3). Similarly when analysing hydrobhobic to non-hydrophobic substitutions, all codons that represent hydrophobic amino acids form a feature set (and *vice versa*).

Data availability

We performed all calculations using Julia programming language, and all scripts are freely available on GitHub: BharatRaviIyengar/DeNovoEvolution. Specifically, our model is implemented in the script `DeNovoProb.jl` which in turn uses the script `nucleotidefuncts.jl` for some basic functions.

Acknowledgments

We thank Pete Czuppon (IEB, University of Münster), for discussions on mathematical models.

References

Acevedo, J. M., Hoermann, B., Schlimbach, T., and Teleman, A. A. 2018. Changes in global translation elongation or initiation rates shape the proteome via the Kozak sequence. *Scientific Reports*, 8(1): 4018.

Andersson, R. and Sandelin, A. 2020. Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, 21(2): 71–87.

Behrens, S. and Vingron, M. 2010. Studying the evolution of promoter sequences: A waiting time problem. *Journal of Computational Biology*, 17(12): 1591–1606.

Berg, J. M., Tymoczko, J. L., and Stryer, L. 2002. *Biochemistry*. W.H. Freeman, New York.

Blevins, W. R., Ruiz-Orera, J., Messeguer, X., and others 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nature Communications*, 12(1): 604.

Box, G. 1979. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pages 201–236. Academic Press.

Bucciantini, M., Giannoni, E., Chiti, F., and others 2002. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416(6880): 507–511.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12): 1213–1218.

Cano, A. V., Rozhoňová, H., Stoltzfus, A., McCandlish, D. M., and Payne, J. L. 2022. Mutation bias shapes the spectrum of adaptive substitutions. *Proceedings of the National Academy of Sciences*, 119(7): e2119720119.

Carvunis, A.-R., Rolland, T., Wapinski, I., and others 2012. Proto-genes and de novo gene birth. *Nature*, 487(7407): 370–374.

Choe, Y.-J., Park, S.-H., Hassemer, T., and others 2016. Failure of RQC machinery causes protein aggregation and proteotoxic stress. *Nature*, 531(7593): 191–195.

Churchman, L. S. and Weissman, J. S. 2012. Native elongating transcript sequencing (NET-seq). *Current Protocols in Molecular Biology*, 98(1).

Clark, M. B., Amaral, P. P., Schlesinger, F. J., and others 2011. The reality of pervasive transcription. *PLoS Biology*, 9(7): e1000625.

Dayhoff, M., Schwartz, R., and Orcutt, B. 1978. Model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, volume 5, pages 345–352. National Biomedical Research Foundation, Silver Spring MD.

Dill, K. A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6): 1501–1509.

Fernández-Moreno, M. A., Farr, C. L., Kaguni, L. S., and Garesse, R. 2007. Drosophila melanogaster as a model system to study mitochondrial biology. In *Methods in Molecular Biology*, pages 33–49. Humana Press.

Gallo, S. M., Li, L., Hu, Z., and Halfon, M. S. 2005. *REDfly*: a regulatory element database for *Drosophila*. *Bioinformatics*, 22(3): 381–383.

Gallo, S. M., Gerrard, D. T., Miner, D., and others 2010. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in drosophila. *Nucleic Acids Research*, 39(Database): D118–D123.

Gardini, S., Cheli, S., Baroni, S., and others 2016. On nature's strategy for assigning genetic code multiplicity. *PLOS ONE*, 11(2): e0148174.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., and others 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17(6): 669–681.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. 1992. Exhaustive matching of the entire protein sequence database. *Science*, 256(5062): 1443–1445.

Gramates, L. S., Agapite, J., Attrill, H., and others 2022. FlyBase: a guided tour of highlighted features. *Genetics*, 220(4).

Grandchamp, A., Czuppon, P., and Bornberg-Bauer, E. 2023. High turnover of *de novo* transcripts in *Drosophila melanogaster*. *bioRxiv*.

Haberle, V. and Stark, A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 19(10): 621–637.

Halfon, M. S., Gallo, S. M., and Bergman, C. M. 2007. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in drosophila. *Nucleic Acids Research*, 36(Database): D594–D598.

Harris, K. and Nielsen, R. 2014. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Research*, 24(9): 1445–1454.

Hartl, F. U. 2017. Protein misfolding diseases. *Annual Review of Biochemistry*, 86(1): 21–26.

Henikoff, S. and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22): 10915–10919.

Hershberg, R. and Petrov, D. A. 2010. Evidence That Mutation Is Universally Biased towards AT in Bacteria. *PLOS Genetics*, 6(9): 1–13.

Hochberg, G. K. A., Liu, Y., Marklund, E. G., and others 2020. A hydrophobic ratchet entrenches molecular complexes. *Nature*, 588(7838): 503–508.

Iyengar, B. R., Choudhary, A., Sarangdhar, M., and others 2014. Non-coding RNA interact to regulate neuronal development and function. *Frontiers in Cellular Neuroscience*, 8: 47.

Jones, D. T., Taylor, W. R., and Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, 8(3): 275–282.

Keeling, D. M., Garza, P., Nartey, C. M., and Carvunis, A.-R. 2019. Philosophy of Biology: The meanings of 'function' in biology and the problematic case of de novo gene emergence. *eLife*, 8: e47014.

Keränen, S. V. E., Villahoz-Baleta, A., Bruno, A. E., and Halfon, M. S. 2022. REDfly: An integrated knowledgebase for insect regulatory genomics. *Insects*, 13(7): 618.

Kim, Y., Sidney, J., Pinilla, C., Sette, A., and Peters, B. 2009. Derivation of an amino acid similarity matrix for peptide:MHC binding and its application as a bayesian prior. *BMC Bioinformatics*, 10(1).

Kimura, M. 1962. ON THE PROBABILITY OF FIXATION OF MUTANT GENES IN a POPULATION. *Genetics*, 47(6): 713–719.

Kosiol, C., Holmes, I., and Goldman, N. 2007. An empirical codon model for protein sequence evolution. *Molecular Biology and Evolution*, 24(7): 1464–1479.

Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, 44(2): 283–292.

Kyte, J. and Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1): 105–132.

Lange, J. D., Bastide, H., Lack, J. B., and Pool, J. E. 2021. A population genomic assessment of three decades of evolution in a natural *Drosophila* population. *Molecular Biology and Evolution*, 39(2).

Le, S. Q. and Gascuel, O. 2008. An improved general amino acid replacement matrix. *Molecular Biology and Evolution*, 25(7): 1307–1320.

Lee, Y. C. G., Ventura, I. M., Rice, G. R., and others 2019. Rapid evolution of gained essential developmental functions of a young gene via interactions with other essential genes. *Molecular Biology and Evolution*, 36(10): 2212–2226.

Lenhard, B., Sandelin, A., and Carninci, P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, 13(4): 233–245.

Long, M., Betrán, E., Thornton, K., and Wang, W. 2003. The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, 4(11): 865–875.

Lopes, R., Agami, R., and Korkmaz, G. 2017. GRO-seq, a tool for identification of transcripts regulating gene expression. In *Methods in Molecular Biology*, pages 45–55. Springer New York.

Majic, P. and Payne, J. L. 2020. Enhancers facilitate the birth of *de novo* genes and gene integration into regulatory networks. *Molecular Biology and Evolution*, 37(4): 1165–1178.

Merchant, S. S., Prochnik, S. E., Vallon, O., and others 2007. The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. *Science*, 318(5848): 245–250.

Milon, B., Sun, Y., Chang, W., and others 2014. Map of open and closed chromatin domains in drosophila genome. *BMC Genomics*, 15(1).

Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., and others 2022. Mutation bias reflects natural selection in Arabidopsis thaliana. *Nature*, 602(7895): 101–105.

Näsvall, J., Sun, L., Roth, J. R., and Andersson, D. I. 2012. Real-time evolution of new genes by innovation, amplification, and divergence. *Science*, 338(6105): 384–387.

Neme, R. and Tautz, D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*, 5.

Noderer, W. L., Flockhart, R. J., Bhaduri, A., and others 2014. Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Molecular Systems Biology*, 10(8): 748.

Ohta, T. and Kimura, M. 1971. Amino acid composition of proteins as a product of molecular evolution. *Science*, 174(4005): 150–153.

Omotajo, D., Tate, T., Cho, H., and Choudhary, M. 2015. Distribution and diversity of ribosome binding sites in prokaryotic genomes. *BMC Genomics*, 16(1): 604.

Prabh, N. and Rödelsperger, C. 2019. *De Novo*, divergence, and mixed origin contribute to the emergence of orphan genes in *Pristionchus* nematodes. *G3 Genes|Genomes|Genetics*, 9(7): 2277–2286.

Proudfoot, N. J. 2011. Ending the message: poly(A) signals then and now. *Genes & Development*, 25(17): 1770–1782.

Richard, P. and Manley, J. L. 2009. Transcription termination by nuclear RNA polymerases. *Genes & Development*, 23(11): 1247–1269.

Rivera, J., Keränen, S. V. E., Gallo, S. M., and Halfon, M. S. 2018. REDfly: the transcriptional regulatory element database for *Drosophila*. *Nucleic Acids Research*, 47(D1): D828–D834.

Santangelo, T. J. and Artsimovitch, I. 2011. Termination and antitermination: RNA polymerase runs a stop sign. *Nature Reviews Microbiology*, 9(5): 319–329.

Schmid, M. and Jensen, T. H. 2018. Controlling nuclear RNA levels. *Nature Reviews Genetics*, 19(8): 518–529.

Schmitz, J. and Bornberg-Bauer, E. 2017. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA . *F1000Research*, 6(57).

Schneider, A., Cannarozzi, G. M., and Gonnet, G. H. 2005. Empirical codon substitution matrix. *BMC Bioinformatics*, 6(1).

Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. 2013. Rates and Genomic Consequences of Spontaneous Mutational Events in Drosophila melanogaster. *Genetics*, 194(4): 937–954.

Shen, S., Kai, B., Ruan, J., and others 2006. Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences. *Physica A: Statistical Mechanics and its Applications*, 370(2): 651–662.

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. 2021. Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology*, 22(2): 96–118.

961 Stewart, M. 2019. Polyadenylation and nuclear export of mRNAs. *Journal of Biological Chemistry*,
962 294(9): 2977–2987.

963 Tautz, D. and Domazet-Lošo, T. 2011. The evolutionary origin of orphan genes. *Nature Reviews*
964 *Genetics*, 12(10): 692–702.

965 Vakirlis, N., Hebert, A. S., Opulente, D. A., and others 2017. A molecular portrait of de novo
966 genes in yeasts. *Molecular Biology and Evolution*, 35(3): 631–645.

967 Van Oss, S. B. and Carvunis, A.-R. 2019. De novo gene birth. *PLOS Genetics*, 15(5): 1–23.

968 Whelan, S. and Goldman, N. 2001. A general empirical model of protein evolution derived
969 from multiple protein families using a maximum-likelihood approach. *Molecular Biology and*
970 *Evolution*, 18(5): 691–699.

971 Wimley, W. C. and White, S. H. 1996. Experimentally determined hydrophobicity scale for
972 proteins at membrane interfaces. *Nature Structural & Molecular Biology*, 3(10): 842–848.

973 Witt, E., Benjamin, S., Svetec, N., and Zhao, L. 2019. Testis single-cell RNA-seq reveals the
974 dynamics of de novo gene transcription and germline mutational bias in drosophila. *eLife*, 8.

975 Wood, V., Gwilliam, R., Rajandream, M.-A., and others 2002. The genome sequence of
976 Schizosaccharomyces pombe. *Nature*, 415(6874): 871–880.

977 Zhang, Z. and Gerstein, M. 2003. Patterns of nucleotide substitution, insertion and deletion in
978 the human genome inferred from pseudogenes. *Nucleic Acids Research*, 31(18): 5338–5348.

979 Zhao, L., Saelao, P., Jones, C. D., and Begun, D. J. 2014. Origin and spread of *de Novo* genes in
980 *Drosophila melanogaster* populations. *Science*, 343(6172): 769–772.