Hema Dalavayi & Max Garzon

On the Randomness of Evolutionary Events along Lineages: Two Case Studies

Faculty Sponsor

Dr. Max Garzon

Abstract

Evolution first came into biology through the seminal work by Charles Darwin in the 1800s. Evolutionary events have often been seen as random, with genetic changes caused by factors like transcription, radiation, and environmental influences. This paper aims to explore whether evolutionary events are entirely random by examining three prediction problems for fishes of the Family *Cichlidae* (C), and bacteria from the genus *Helicobacter* (H). Predictions concern the divergence time of an organism (when its species first appeared on Earth) in a biological lineage, the difference in divergence time between two organisms, and a proxy for the next organism in the lineage. Neural network models were trained using machine learning techniques that yield relative errors as low as 5%, 6%, and 57% in their predictions, respectively. These results suggest that while evolution is influenced by random factors, some specific events may be predictable, adding evidence to the argument that evolutionary events are not entirely random.

Introduction

As pointed out by Dobzhansky, “nothing makes sense in biology except in the light of evolution” (Dobzhansky, 1950). The theory of evolution is a cornerstone of modern biology. The idea of evolution first came into biology through the seminal work by Charles Darwin (Darwin, 1859). Darwin proposed that species evolve over time through the process of natural selection, where individuals with traits suitable for their environment have a better chance to survive and reproduce, by passing these traits on to future generations. This idea revolutionized biology by providing a framework that explains how organisms on Earth change over time despite (perhaps drastic) changes in their environment. Darwin’s theory challenged the theological views that were prevalent during that time. First, many theologists believed that life on Earth was static and that they were unchanged and fixed as perfect creations by God (Paley, 1802). Paley argued that each species has a fixed purpose in life as designed by a divine Creator, with every part of an organism functioning to serve a specific role in the natural world. This view supported the belief that species were immutable and had remained unchanged since their creation. Unlike Paley’s view of a static world, Darwin argued that species change over time in response to environmental pressures. This challenged the longstanding belief in the static world and incurred a lot of criticism from both the religious and the scientific communities. However, after much debate and controversy over the years, this theory of evolution was gradually accepted by the scientific community and has become prevalent today.

Evolutionary events have often been assumed to be entirely random by biologists because genetic changes (e.g., variations and mutations) depend on changes occurring in an organism’s DNA due to many internal factors affecting individuals. Also, evolution is also influenced by environmental factors such as geography, climate and available resources. Now, statements of this sort are not very meaningful without a noncircular and rigorous definition of randomness. There are two important definitions of randomness, statistical and computational. According to Wikipedia [9], statistical randomness refers to a sequence of numbers that contains no recognizable patterns or regularities, making it impossible to predict the next number based on the previous elements in the sequence. Likewise, computational randomness refers to sequences where the elements cannot be predicted or generated by any algorithm or computer program. In other words, it is unpredictable to the extent that no computer, no matter how powerful, can predict the next element from the previous ones. Although important for our research question, the purpose of this paper is not to provide a precise definition of randomness, hence for the sake of this study, random will refer to outcomes that are unpredictable. Therefore, although Darwin never said so explicitly, most biologists have come to assume that evolutionary events are unpredictable, as it is accrued by random genetic mutations and interactions of an organism with its environment that accumulate over space and time.

On the other hand, recent studies have shown that genetic variations follow noticeable patterns though they are indeed influenced by factors such as the environment, natural selection, and evolution. For instance, (Wortel et al, 2023) states that two key factors influencing evolution are genetic elements, such as mutation bias and epistatic interactions, as well as ecological factors. Genetic factors, particularly in large populations, provide more variation for natural selection to act upon, while rapid environmental changes complicate predictions due to their complexity. This research laid the groundwork for exploring the predictability of evolution through genomic data. In addition, a case study by (Mas et al, 2020) made observations on how natural selection affected allele frequency changes in the plant of the species Mimulus guttatus. The study demonstrated that genetic models could effectively predict evolutionary outcomes when fitness measurements were available, suggesting that evolutionary changes are not entirely random but rather guided by specific genetic and ecological pressures. This fact aligns with the evolving view that, while genetic mutations are random, selection pressures can steer evolutionary processes in predictable directions in a population. These discoveries suggest that evolutionary events could be influenced by more than just random mutations pointing to the presence of underlying patterns in DNA that guide evolutionary changes.

The primary aim of this paper is to examine at a deeper and narrower level whether evolutionary outcomes are entirely random. While evidence in the previous paragraph suggests that some aspects of evolution can be predicted based on genetic variations on populations, this paper aims to explore how un/predictable evolutionary changes may be for individual organisms or species and furthermore, based only on their DNA. It will investigate, in particular, the extent to which DNA contains necessary information to make certain features of evolutionary events predictable, contrary to what is expected in a purely random process. The question is precisely addressed by solving 3 prediction problems for two (2) lineages: fishes of the family of Cichlidae (C) and bacteria from the genus HelicoBacter (H). Here, a lineage L is a sequence of individual organisms (x_1, x_2, …, x_N) that have evolved from a single common ancestor x_0. Such prediction problems have not been considered because evolution in biology has been regarded as essentially unpredictable. Cichlids are eukaryotes known for their diversity and adaptability, making them very appropriate for studying evolutionary questions. Analogously, HelicoBacter is a genus of prokaryotes in the domain of bacteria, that has been studied for their role in human health and their potential for rapid evolutionary changes due to their short generation times. For this study, a prokaryote and a eukaryote lineage will be considered to formulate precise definitions of what is being predicted before the prediction is made, to avoid the circularities typical in the subject, as follows.

Defining a prediction problem PREDICTION (L, f) precisely for a lineage L requires a specification in advance of what is being predicted based on what information (Garzon et al, 2022). The evolutionary relationship in a lineage L such as Cichlidae fishes (C) and HelicoBacter (H) is modeled as a function f that maps organisms in a lineage L in a population to what is being predicted, usually a numerical value. The predictions concern the divergence time of a species, i.e. how long ago in the Earth's historical past did the species of an organism appear on Earth since the appearance of its immediate previous ancestor(s) in the lineage. The organisms in the taxon L will be represented by segments of their genomes (e.g. certain genes) and given in a feature vector (x_1, x_2 ... $x_{t)}$ of DNA sequences as proxies to represent each organism. The prediction problem is defined by a series of instances, each with the input necessary on which to base a prediction, and corresponding questions regarding what the prediction is about. A solution to the problem is some single device (e.g. a computer program) providing the right answers to the questions being asked for every single instance, regardless of how the correct answers may be changing from one instance to the next.

(a) [DIVTIME] PREDICTING DIVERGENCE TIME (L, f)

INSTANCE: a vector $\mathbf{x}_t = (x_1, x_2, \ldots, x_t)$ representing the previous species' DNA in the lineage up to time t.
QUESTION: What is the divergence time of organism $x_{t+1}= f(\mathbf{x}_t)$, i.e., when the next organism $x_{t+1}= f(x)$ in the lineage appears on Earth?

In biological phylogenetics, this time is usually measured in terms of millions of years ago (Mya) and is usually estimated using traditional methods like fossil records or carbon dating for a given organism (there is an inherent inaccuracy in these estimates given the nature of the methods used, e.g., radioactive decay in the fossil (Carleton, 2018).)

(b) [DDIV] PREDICTING DIFFERENCE IN DIVERGENCE TIME (L, f)

INSTANCE: Two organisms (x_{t-1}, x_t) in L
QUESTION: What is the difference in divergence times?

(c) [NEXTP] PREDICTING NEXT PROXY (L, f)

INSTANCE: a vector $\mathbf{x}_{[t-4,\ t-1]}$ = (x_{t-4}, x_{t-3},x_{t-2}, x_{t-1}) representing the 4 previous species' in the lineage L.
QUESTION: What is the next element xt= f($\mathbf{x}$($_{t-4,\ t-1}$) in the lineage at time t+1?

In this prediction problem, the goal is to predict the genomic signature of the next organism, not necessarily its exact DNA sequence. (With the predicted genomic signature, the exact original DNA sequence cannot be reconstructed).

In terms of difficulty, PREDICTING DIVERGENCE TIME (L, f) was the simplest problem of the three because given only the DNA of one organism, the solution needs to extract/ predict a number representing the divergence time. At first sight, this problem appears impossible to solve. Second, PREDICTING DIFFERENCE IN DIVERGENCE TIME (L, f) is a step deeper than the previous one because, in principle, a model does have to analyze the DNA of two different organisms to predict the difference in the divergence times, either individually or jointly. Finally, the most difficult problem was PREDICTING NEXT PROXY (L, f) because the model had to analyze the DNA in numerical form, for four (4) different organisms and predict the genomic signature of the next organism in the lineage, as it occurred in the Earth's history.

This paper is organized as follows. Section 2 provides a high-level characterization of data science and machine learning techniques to be used to provide solutions to these problems. Section 3 describes the solutions obtained for each problem and a summary and an assessment of the results. Finally, Section 4 presents a discussion of the findings from the research.

Methods

The traditional/conventional approach to addressing biological problems relies on qualitative observations or current or past life and their recordings to understand life, e.,g. as taxonomies of life as it exists now, or hypotheses about the predecessors of that life in the reproductive cycle. However, solving prediction problems requires further analyses of the observations (data) to extract useful information (Garzon et al, 2022). This methodology has been refined by the emerging fields of data science and machine learning that will be used in this paper to handle these prediction problems. This methodology is structured into three distinct phases: observation of the phenomenon behind the problems and data collection, data preprocessing, and the development of machine learning models, such as neural networks, to obtain the solutions.

Data Science and Machine Learning

Data science is a field developed to address computational problems in ways that differ from traditional sciences, focusing on observing and gathering data about a phenomenon of interest rather than relying on deep, complex analyses to gain a deep understanding of a phenomenon, the conventional approach in science. With the advent of computer science and the information age, tools have been developed to analyze these huge amounts of data regardless of their specific domains (Garzon et al, 2022). This technique has enabled the identification of intricate patterns and trends across vast amounts of data which would have been a nearly impossible task to detect using the traditional methods. Thus, data science methods and tools have become a pivotal tool in solving many kinds of problems by leveraging machine learning techniques to identify patterns, make predictions, and optimize processes based on just data across a wide range of industries.

For each of the three problems defined above, it is essential to collect appropriate data to address them. In the data science approach, a computational tool (i.e. neural network) is then used to process the data, identify patterns, and generate accurate predictions based on the input. Out of the many types of tools available in machine learning to solve a problem, a neural network is particularly suited for tasks involving complex patterns in a dataset. A neural network (NN) is a model inspired by the human brain and consists of a set of (artificial) neurons, which are simple units that process local information. Each neuron (i) has a range of activation values Ai, along with a vector Wi of synaptic weights wi,j, where each weight represents the strength of the connection between two neurons i and j. In addition, each neuron is endowed with an activation function with a domain of real numbers and a codomain Ai of activation values that the neuron can assume and output after applying the activation function $\sigma_{i.,}$ like the sigmoid function $\sigma(y) = \frac{1}{1+e^{x}}$. Typically, in a neural network, each neuron computes the net input by taking the sum of the products of the activation values (a_j) and the synaptic weight $w_{i,j}$ of the neuron. This net input is then passed through the activation function $\sigma_{i,}$ which outputs a value within a specified range, such as [0, 1] or [-1, 1], depending on the function used. The resulting output becomes the activation value for the next time step. This process continues until the output layer is reached, where the final activation values produce the final output (e.g. a prediction associated with the inputs clamped to a pre-designated set of input neurons.)

Among the infinite number of possible neural networks, the correct network capable of solving a given problem is usually identified by a learning algorithm. In supervised learning, this algorithm enables the network to adjust its weights and activation functions by learning from labeled data, i.e., data containing the appropriate answers in advance. In this approach, the network is "supervised" by the labeled output (a.k.a target label) during the training phase. This means that for each of the input vectors in the training set, the correct output is already known so the goal of the network is to learn the relationship between the input and output by adjusting its weights so that with that knowledge it can make more accurate predictions on unseen data later. The neural network is deemed to make a correct prediction when its output matches the labeled data. If not, then the learning algorithm adjusts the parameters (weights) of the NN so it will a better chance to produce the correct output. Before starting the training phase, the learning algorithm splits entire dataset into a training and testing set, based on a parameter fixed by the researcher (typically 70%/30% or 80%/20% for training/testing), allowing the NN to be tested to evaluate its performance after the training is completed. The learning algorithm requires a structure of the neural network to be trained (the architecture) such as the number of layers, the type of activation function to be used and the data. After that, the learning algorithm starts with random weights to get a candidate network to improve during training. Then the learning algorithm processes each data point by feeding it as an input feature vector to the candidate network, which then does through its internal working as explained above and produces an output (this is known as the forward pass.) If the candidate network produces an output that does not match the labeled data, that means the model is not performing well, and the learning algorithm needs to adjust the weights.

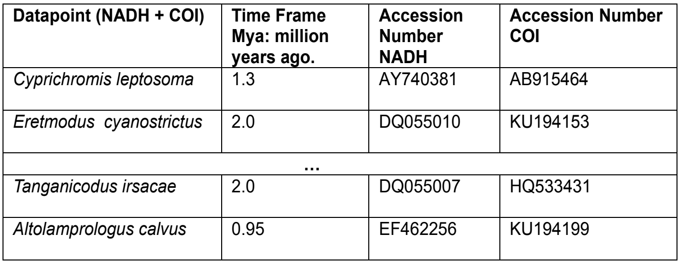
Among the many types of learning algorithms available in supervised learning, the most efficient and popular is the backpropagation algorithm, which optimizes the weights of the neural networks based on the error difference between the predicted and actual outputs. In the forward pass, if the output does match the labeled data, then the network proceeds to the next data point and continues the process. On the other hand, if the output does not match the labeled data, then the network takes on a backward pass where the error is propagated backward through the network. This backward pass works in a way such that the changes in the weights of the neurons are proportional to how much they have contributed to the incurred error. During this pass, the error is repeatedly propagated to one layer behind, penalizes the neurons and prompts change in their weights until it reaches the input layer. The learning algorithm repeats the whole process with other data points to complete one epoch. The user can set a stopping condition for the learning algorithm, such as specifying the number of iterations (epochs) to run or setting an accuracy threshold for accumulated errors, at which point the learning process stops.

Once this training phase is complete, the learning algorithm shifts to the testing phase where the trained NN is evaluated using the testing dataset that it has not seen before. The NN processes each input and makes a prediction. Then the prediction is matched by the learning algorithm against the target label and if it matches, then the prediction is marked as correct, else an error computed as the difference. Once the neural network has made all the predictions on the testing dataset, an accuracy measure is obtained. If the network's performance meets the desired criterion of quality, it is ready to effectively solve the defined problem; else the training is repeated with different parameters to obtain a better performing model.

Data Gathering and Pre-Processing

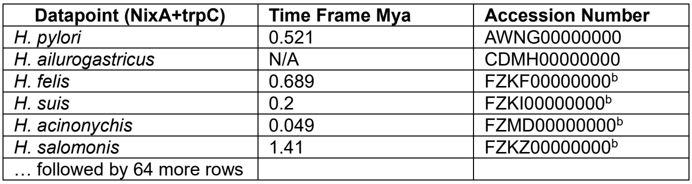
DNA encodes for critical information required to develop and sustain life in every living organism. Therefore, DNA sequences are the most appropriate data to obtain for solving the problem due to the deep structure of the DNA spaces. The lineages data was downloaded from GenBank (Sayers et al, 2022) as DNA sequences. Each data point was a DNA sequence of a specific gene. For the Cichlidae, the genes NADH (dehydrogenase) or COI (Cytochrome Oxidase I) of an organism in the lineage were used. NADH dehydrogenase is part of the mitochondrial genome which plays a crucial role in cellular respiration and is known for evolving at a moderate rate. The COI gene is also part of the mitochondrial genome and is known for evolving at a faster rate compared to other genes, making it useful for studying evolutionary events, such as the divergence times.

For the Cichlidae and the HelicoBacter lineages, the sequences were downloaded for their specific genes using the accession numbers (unique identifiers) (Garcia and Colorado, 2024). For the Cichlidae-NADH, DNA sequences for 44 NADH genes and 36 COI genes were downloaded. Likewise, for the genus HelicoBacter(H), the DNA sequences of 69 organisms for genes NixA and trpC across 6 different species. The divergence times for each organism in both lineages have been provided by (Garcia and Colorado, 2024). **Tables 1** and **2** summarize these data sets.



Datapoint (NADH + COI)	Time Frame Mya: million years ago.	Accession Number NADH	Accession Number COI
Cyprichromis leptosoma	1.3	AY740381	AB915464
Eretmodus cyanostrictus	2.0	DQ055010	KU194153
...			
Tanganicodus irsacae	2.0	DQ055007	HQ533431
Altolamprologus calvus	0.95	EF462256	KU194199

Table 1.
Typical data points from various organisms
in the Cichlidae Fish (C) lineage.*



Datapoint (NixA+trpC)	Time Frame Mya	Accession Number
H. pylori	0.521	AWNG00000000
H. ailurogastricus	N/A	CDMH00000000
H. felis	0.689	FZKF00000000[b]
H. suis	0.2	FZKI00000000[b]
H. acinonychis	0.049	FZMD00000000[b]
H. salomonis	1.41	FZKZ00000000[b]
… followed by 64 more rows		

Table 2.
Typical data points from 6 species of organisms
in the HelicoBacter (H) lineage.*

*The accession numbers were used to obtain the DNA sequences for the organism from the GenBank [8]

To solve a problem in data science using neural networks, the DNA sequences cannot be directly given as inputs to a neural network because it only takes numerical feature vector values as inputs. Therefore, these DNA sequences were transformed into feature vectors by selecting a non-cross-hybridizing (nxh) basis and using the genomic signature of the DNA proxy for individuals in the lineage as a feature vector. Using the deep structure of DNA spaces (Garzon et al, 2022), three nxh bases were selected to transform a DNA sequence into numerical feature vectors. Given that there are multiple probes of length m in each nxh basis, each DNA sequence of an organism was shredded into nonoverlapping fragments of length m. Now, each shred is compared with each probe to see whether they hybridize or not. Finally, for each probe, the count of number of shreds hybridized with the probe is normalized by dividing it with the total number of shreds. Finally, this results in a numerical vector for each probe that represents the relative frequency of hybridization between the DNA sequence and the probes in the nxh basis. This vector, called the genomic signature, captures essential information in the DNA sequence of an organism. (Garzon et al, 2022). For the Cichlidae-COI and NADH genes, 4mP3 and 5mP6 were used to convert the DNA sequences into numerical vectors. Here, 4mP3 refers to a probe of length 4 and P3 refers to the length of the vector. Table 3 summarizes the dimensionalities of the datasets.

Lineage - Genes	**nxh basis**	**Feature vectors (length)**	**Size of the dataset**
Cichlidae - NADH	4mP3	3 → 2 *	44
	5mP6	6	44
	4mP3+5mP6	8	44
Cichlidae - COI	4mP3	3→2 *	36
	5mP6	6	36
	4mP3+5mP6	8	36
HelicoBacter- NixA+trpC	4mP3	3→2 *	69
	5miC3Mg	3→2 *	69

Table 3.
Summary of feature vector lengths and
number of data points for lineages C and H.

* The original vectors have been reduced [3 → 2] for training purposes.

Training Neural Network Solution

With these tools in hand, it was possible to train a neural network (NN) to solve the target problems. During the backpropagation training phase, the feature vectors in Section 2.2 were given as inputs to the neural network (the number of input neurons to the NN was changed to match the length of the feature vectors.) The training phase halted once the predefined stopping conditions were met. After training the NN multiple times, the NN's weights converged after 10,000 epochs. The NN solutions are shown in Figures 1, 2, 3.

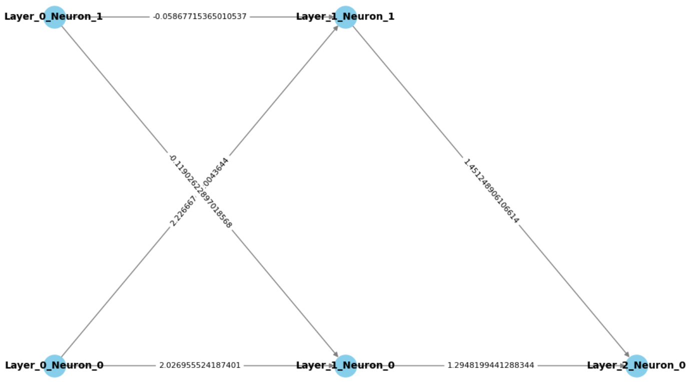


Figure 1.
Neural Network solution for
the problem of PREDICTING DIVERGENCE TIME (L, f).

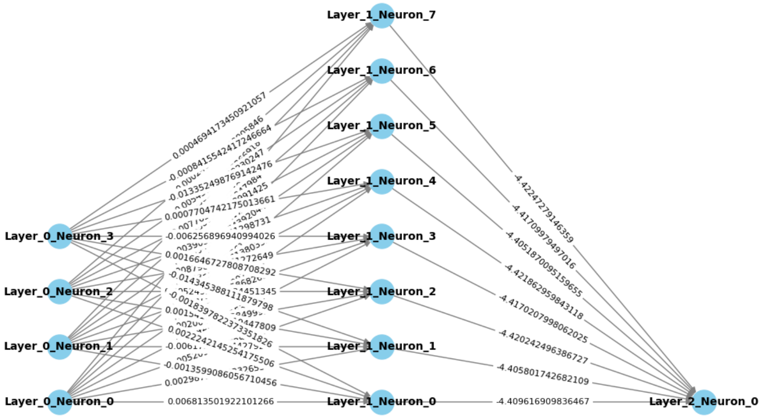


Figure 2.
Neural Network solution for the problem of
PREDICTING DIFFERENCE IN DIVERGENCE TIME (L, f).

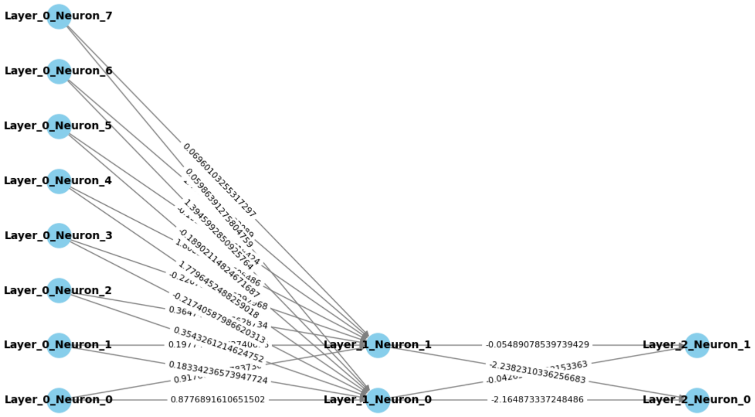


Figure 3.
Neural Network solution for the problem of
PREDICTING NEXT PROXY (L, f).

To assess the robustness and reliability of the NN model, validation techniques were applied to evaluate its performance and ensure its generalization ability on unseen data. Once the training phase was completed, the learning algorithm moved to the testing phase of the trained NN, where unseen data was given as input to generate predictions. Then the predictions were compared with the labels (target values) and an accuracy assessment was made. If the accuracy threshold set by the user was reached, the NN model was accepted as a solution to the prediction problem.

Results and Assessments

The solutions to the problems are trained NN models that predict the target features for arbitrary DNA proxies in the next lineage, including other data points not in the tribe of Cichlidae for the problem. This section explains how these models were obtained and provides an assessment as to how they might answer the major question whether evolution in a lineage is entirely random.

Predicting Divergence Time (L, f)

Firstly, the NN model used to solve this problem was developed by pre-processing the DNA sequences obtained for the two lineages as described in Section 2.2, followed by initializing the network architecture and training using backpropagation. During preprocessing, nxh bases 4mP3 and 5mP6 were used to convert the DNA sequences of the NADH and COI genes for the Cichlidae (C) lineage into numerical vectors of length 3 and 6 respectively. Similarly, nxh bases 4mP3 and 5miC3Mg were used for the HelicoBacter dataset to convert the NixA+trpC genes into numerical vectors of length 3 and 3 respectively. The dataset was split as 70%/30% for training/ testing but sometimes the split had to be adjusted to 80%/20% to obtain models with better predictions. The target label for this problem is the divergence time for each organism.

Since the target labels are continuous quantities (time) the metric for evaluation of the quality of the NN solutions is usually chosen to be the average percentage in the relative error across, given by

$$\mathrm{RE} = \frac{|actual - predicted|}{actual} * 100$$

for the predictions as a consolidated measure of how well the NN performed across all predictions in the testing phase. The results are shown in **Table 4**, together with the choices for preprocessing and the NN models.

Lineage	*nxh base*	*Arch. (hidden layers)*	*Min / Avg /Max RE (%)*
Cichlidae Fish - NADH	4mP3	[2]	0 / 31 / 43.64
	4mP3+5mP36	[9,6,3,1]	0 / 7 / 70
	5mP36	[6,5,1]	0 / 19 / 52
Cichlidae Fish - COI	4mP3	[2]	0 / 7 / 22
	5mP36	[6,3,2,1]	0 / 5 / 41
	4mP3+5mP36	[9,6,3,1]	0 / 7 / 12
HelicoBacter-NixA+trpC	4mP3	[2]	0 / 27 / 46
	5miC3Mg	[2]	0 / 28 / 47

Table 4.
Relative Error (RE) of the solutions for Problem DIVTIME.
The best prediction is given by the COI proxy for lineage C on the concatenation of signatures on 4mP3 and 5mP3 nxh bases, while 4mP3 performs unsatisfactorily for both C and H.

Predicting Difference in Divergence Time (L, f)

The NN models to solve this problem were obtained using the same process described in Section 3.1. The only difference was that a different architecture was chosen to better fit the input features for the two proxies used to predict the difference in divergence times. The feature vectors were obtained by concatenating the two feature vectors of two different input organisms from the datasets, providing the neural network with enough information to make predictions. The target label was the difference in their divergence times. **Table 5** summarizes the results.

Lineage	***nxh base***	***Arch. hidden layers)***	***Min / Avg / Max RE (%)***
Cichlidae Fish - NADH	4mP3	[8]	0 / 8 / 34 *
Cichlidae Fish - COI	4mP3	[3]	0 / 6 / 10 *
HelicoBacter-NixA+trpC	4mP3	[3]	0 / 50 / 80 *

Table 5.
Relative Error (RE) of the solutions for Problem [DDIV].
The best prediction is given by the COI proxy for lineage C on the signatures of nxh basis 4mP3, while the same nxh basis performs inadequately for H.

*Some large value outliers were excluded in the RE calculation due to the significant difference in divergence time between the newer and older species, which led to a high RE value from the original divergence.

Predicting Next Proxy (L, f)

Likewise, the feature vectors for this problem are obtained by concatenating the feature vectors of the organism and its three ancestors in the lineage. Since the first three organisms do not have three ancestors, these datapoints were ignored. Predicting the actual DNA sequence of the next organism in the lineage is a very difficult problem because it involves many other factors (e.g., environmental) than just genetic inheritance obtained from its ancestors. Thus, in this problem, the aim was reduced to predicting the genomic signature of the next organism than the entire DNA sequence. To enable the predictions, the 3D feature vectors (x,y,z), from the nxh bases 4mP3 and 5miC3Mg, were reduced to a 2D feature vector (x',y') by a geometric transformation (rotations and translations). This is possible since the feature vectors are normalized, i.e., lie on a plane in 3D Euclidean space given by the condition x+y+z=1. These transformations adjusted the orientation and position of the points but only require two coordinates. Thus, the 3D vectors are reduced to 2D vectors while hopefully retaining the underlying information in the original genomic signatures. As a result, the NN architecture will require only two output neurons for the two features in the genomic signatures being predicted. After this, the learning takes place as explained earlier. Once the required accuracy threshold is reached during training, the overall relative error for the predictions were calculated. The results are summarized in **Table 6**.

Lineage	nxh base	*Arch. hidden layers)*	*Min / Avg / Max RE (%)*
Cichlidae Fish - NADH	4mP3	[3,4]	0 / 189 / 520
HelicoBacter- NixA+trpC	4mP3	[2]	0 / 95 / 500
HelicoBacter -NixA+trpC	5miC3Mg	[2]	0 / 57 / 900

Table 6.
Relative Error (RE) of the solutions for Problem [NEXTP].
The best prediction is given by the NixA+trpC proxies for lineage H on the signatures of nxh basis 5miC3Mg , while the signatures of the nxh basis 4mP3 performs unsatisfactorily for C.

To assess the significance of these results for the three problems, one must consider a few factors. First, the difficulty of the problems, as discussed in their statements above. Second, when the DNA sequences were transformed into numerical vectors through the dimensionality reduction process, some crucial information must have been inevitably lost and therefore impacted the models' ability to predict the labels. Nonetheless, the neural networks were indeed able to make the most of the information given to them and make predictions out of that data, albeit imperfectly and thus, the DNA must contain some information for the NN to extract the target features. On the other hand, this fact raises deeper questions about the reason for the errors: Is the loss of information just in the pre-processing (from DNA to genomic signatures.) Does evolution lose some information in the way it is encoding the divergence time for example?

Secondly, it is important to evaluate the magnitude of the relative errors (REs) and their significance for the research question. In the problem [DIVTIME], the lowest relative error was 5%. In terms of millions (M) of years, a 5% error is just 50,000 years. Given the scale of evolutionary time, this level of error is therefore acceptable, and this NN could be used for such datasets producing predictions of divergence time with a high predictability. Similarly, for the second problem [DDIV], a 6% average RE (about 60,000 years) is a relatively close prediction of the difference in the divergence times at a time scales of 36M years. Even the traditional methods used to do time-series analysis using calibrated radiocarbon dates, which often have irregular uncertainties (Carleton, 2018). Finally, problem [NEXTP] was the most challenging to solve, as the NN had to forecast the genomic signatures for the next species with a minimum RE of 57% is surprisingly good, but a maximum RE of 189% is not surprising. Alternatively, it is possible that the target labels are themselves in error with respect to the gold standard of the historical record.

All things considered, it is evident that there is substantial evidence in these results to answer the question of whether evolution is entirely random. The fact that the neural network was able to predict the divergence times and the difference in the divergence time with a high degree of correctness despite a number of possibly faulty assumptions, implies that there are underlying patterns and predictability in the evolutionary process. While randomness likely plays a role in evolution, this unpredictability could stem from genetic mutations that occur randomly or from inherent limitations in the data. But the ability of such models to make informed predictions on some of the events is strong evidence that deterministic factors are contributing to shaping evolution. This challenges the assumption that evolutionary events are purely random and points to the possibility of them being a more organized process influenced by both random and predictable elements. The results of this paper also raise a deeper question: Are there random events in the evolutionary process that are provably inherently unpredictable?

Discussion and Conclusion

This paper has addressed a fundamental research question whether evolutionary events are entirely random, following up on recent studies that it might not be. The working definition of randomness was provided in the introduction as a sequence of numbers which contains no recognizable patterns or regularities that it is impossible to predict the next number in the sequence either by humans or using a computer program. The main takeaway is that the results provide quantitative evidence indicating that evolutionary events involve predictable elements along with random processes. Neural network's abilities to provide an approximation, with low relative error, of the divergence time of organisms or the (incomplete but significant) information about the genetic composition of their descendants in a lineage, i.e., genomic signatures of the DNA, suggests that there are indeed deterministic factors in evolutionary processes. Since the Min RE= 0 in all three problems, there were instances of individuals in each where the neural network made perfect predictions that matched the given label, although there were also individuals for which RE > 0 because the average RE was positive. This would appear to be in direct opposition to the conventional view of evolutionary events being essentially random.

This study could have used a larger and more representative dataset. But due to difficulties obtaining reliable data, the data set used in this study is small and narrowly focused. Nonetheless, this restriction is a necessary first step to set up the stage for a more comprehensive future study aimed at elucidating the relative contribution of environmental factors to the randomness in evolutionary processes.

On the other hand, acknowledging the limitations is necessary for these results. First, this study only included two lineages, which limits the generalizability of the results to a broader range of evolutionary scenarios. Second, a small or narrowly focused dataset may not contain enough variability to fully represent genetic diversity present in larger divergent groups, potentially leading to overfitting or missed patterns in the models. Third, it is clear in biology that evolution is influenced by both genetic and environmental factors, but the models were trained using only DNA sequences without attempting to incorporate additional information about the environment (largely unknown at divergence times), which could have provided a more comprehensive understanding of evolutionary patterns, and the effect of variables such as climate and geography.

Despite these limitations, the quantitative analyses of the predictions through low relative errors suggest a significant degree of predictability for these two lineages. Based on the results, it can be confidently stated that evolutionary events contain predictable features. Of course, that does not mean that this evidence can conclude that evolutionary events are entirely predictable. Since the study was conducted on only two lineages, while many others remain untested, the findings cannot be generalized to arbitrary lineages, although the methods are based on the deep structure of DNA that is pervasive through the entire biome. In some sense, this study really raises more questions than it answers, but these questions about predictability are of enormous interest given the uncertain future of many species due to the increasing fragility of our planet (e.g. due to climate changes.) Much more comprehensive and difficult studies are needed to confirm that evolution is not entirely random across the full range of evolutionary events. For example, can we ascertain the existence of a specific evolutionary event and strong evidence that it is random?

Acknowledgment

We would like to acknowledge Drs. Luis Fernando Garcia and Alexander Colorado of the National U of Colombia for their invaluable assistance in suggesting appropriate datasets, the specific genes and DNA sequences and collecting and furnishing the divergence times for the various organisms. Their support was essential to the success of this study.

References

1. Paley, W. (1802). Natural theology: Or, Evidence of the existence and attributes of the deity, collected from the appearances of nature (2nd ed.). R Faulder. https://doi.org/10.1037/11747-000

2. Darwin, Charles (1859). On the Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life. London: John Murray, 1859.

3. Dobzhansky, T. (1950). Mendelian populations and their evolution. The American Naturalist,84:819, 401—418.

4. Wortel MT, Agashe D, Bailey SF, Bank C, Bisschop K, Blankers T, Cairns J, Colizzi ES, Cusseddu D, Desai MM, van Dijk B, Egas M, Ellers J, Groot AT, Heckel DG, Johnson ML, Kraaijeveld K, Krug J, Laan L, Lässig M, Lind PA, Meijer J, Noble LM, Okasha S, Rainey PB, Rozen DE, Shitut S, Tans SJ, Tenaillon O, Teotónio H, de Visser JAGM, Visser ME, Vroomans RMA, Werner GDA, Wertheim B, Pennings PS. Towards evolutionary predictions: Current promises and challenges. Evol Appl. 2022 Dec 9;16(1):3-21. doi: 10.1111/eva.13513. PMID: 36699126; PMCID: PMC9850016.

5. Mas A, Lagadeuc Y, Vandenkoornhuyse P. Reflections on the Predictability of Evolution: Toward a Conceptual Framework. iScience. 2020 Oct 27;23(11):101736. doi: 10.1016/j.isci.2020.101736. PMID: 33225244; PMCID: PMC7666346.

6. Garzon, M., Yang, C., Venugopal, D., Kumar, N., Jana, K., & Deng, L. (2022). Dimensionality Reduction in Data Science. 1-265. Springer Nature. https://doi.org/10.1007/978-3-031-05371-9

7. Carleton WC, Campbell D, Collard M. Radiocarbon dating uncertainty and the reliability of the PEWMA method of time-series analysis for research on long-term human-environment interaction. PLoS One. 2018 Jan 19;13(1):e0191055. doi: 10.1371/journal.pone.0191055. PMID: 29351329; PMCID: PMC5774753.

8. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, Madej T, Marchler-Bauer A, Lanczycki C, Lathrop S, Lu Z, Thibaud-Nissen F, Murphy T, Phan L, Skripchenko Y, Tse T, Wang J, Williams R, Trawick BW, Pruitt KD, Sherry ST. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2022 Jan 7;50(D1):D20-D26. doi: 10.1093/nar/gkab1112. PMID: 34850941; PMCID: PMC8728269.

9. Wikipedia contributors. (2025, January 1). Statistical randomness. Wikipedia. https://en.wikipedia.org/wiki/Statistical_randomness