What does the tree of life look like as it grows?

Toward a dynamic theory of phylogenetics

Kevin Hudnall
kahudnall@ucdavis.edu
Biosystems Engineering Graduate Group

Abstract

A model of phylogenetic space capable of representing the change in biological form ($\partial\varepsilon$) with respect to the change in time (∂t) must be equipped with two metrics, one for form (d_ε) and one for time (d_t). Such representations are lacking in biology, suggesting that the challenge of constructing a dynamic representation of phylogenetic space has not been fully acknowledged. Here, I develop such a representation as a stochastic iterated function system. Information measures are then used to obtain a metric of biological form d_ε. A few comments are then given about how to obtain the metric of time d_t.

Introduction

Motivation

Dynamical thinking in evolution has typically been concerned with population genetics, selection in ecology (e.g., the logistics equation), and evolutionary game theory (Nowak, 2006). Little attention has been paid to the theoretical and topological underpinnings of how phylogenetic space changes over time. Though biologists have been making phylogenetic trees since before Darwin (Hitchcock, 1840), the need for dynamic representations of phylogenetic space seems to have been overlooked.

Why it is interesting

We all experience time as a biological phenomenon: we are born, we grow, we die, and then presumably that is the end of time for each of us. To assume time is something other than biological is to engage in metaphysical speculation. The question of how to represent the dynamic tree of life is not some quaint biological enterprise; it is a question of the nature of time itself, and therefore, a question central to physics and all of science.

Synopsis of project and results

The dynamic model of phylogenetic space developed here is based on five principles: the principle of sufficient dimensions (need both d_ε and d_t defined in order to have $\frac{\partial\varepsilon}{\partial t}$), the random principle (biological systems are stochastic processes), the nested principle (a species tree constitutes a genus node), and the organism-population duality principle (populations, not organisms, evolve).

A stochastic iterated function system that generates a structure capturing these five principles is then given. The stochasticity of the structure makes it well-suited for the application of information measures. The form-wise distance between two taxa i and j is captured as the information distance.

Simulation results meant to show the coherence of the model are given. These include lineage-specific entropy calculations, pathwise mutual information calculations (i.e., the mutual information between "parent" and "child"), cross-path mutual information calculations (i.e., the mutual information between distantly related species), and cross-path information distance calculations. An interesting result is the calculation of channel capacities along lineages, which provides a means of identifying which areas of the system are evolutionarily conserved, and which are rapid proliferators.

Background

Requirement 1: sufficient dimensions

Phylogenetic models with distance metrics defined on the space come in two forms: *chronograms*, where branch lengths are proportional to time intervals, and *phylograms*, where branch lengths are proportional to biological form (Bleidorn, 2017) (**Figure 1**).

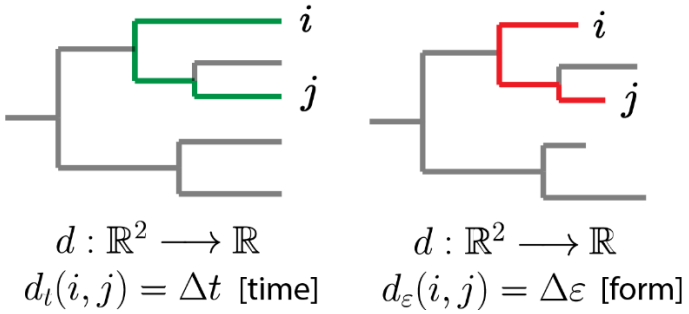

Figure 1 (LEFT) A hypothetical chronogram showing temporal distance between taxa i and j. The metric $d_t(i,j)$ returns a scalar with units of time. (RIGHT) A hypothetical phylogram showing form-wise distance between taxa i and j. The metric $d_\varepsilon(i,j)$ returns a scalar with units of biological form.

Both chronograms and phylograms are individually inadequate to represent the dynamic $\frac{\partial \varepsilon}{\partial t}$ for the simple reason that each has insufficient dimensions. For a chronogram we have t and therefore potentially ∂t, but we lack ε and therefore $\partial \varepsilon$. For a phylogram we have ε and therefore potentially $\partial \varepsilon$, but we lack t and therefore ∂t.

In order to represent the dynamic case, we need both t and ε on the same space with their corresponding metrics d_t and d_ε defined on that space.

Requirement 2: the random principle

The random principle is central to Darwin's descent-with-modification (Darwin, 1859). This is arguably the main distinction between Darwin's mechanism of evolution and Jean Baptiste Lamarck's mechanism of evolution (Lamarck, 1809). Darwin's theory is really the observation of what happens when reproducing populations are subject to a random generator under conditions of finite environmental carrying capacity. The random generator is the engine of modern evolutionary theory, and the foundation of modern biology as a whole. While the mechanisms of evolution are many, they all have as their basis random mutations in the genome during cell reproduction. Our model of phylogenetic space must capture this stochasticity of biological systems.

Requirement 3: the nested principle

While randomness is a hypothesis of Darwin's theory, the "nested principle" is a logical implication of Darwin's theory.

We can state the principle as follows:

> A family tree T^F consists of a set of families $\{F_i \ldots F_\#\}$ that contains a family F_k constituted by a genus tree T^G that consists of a set of genera $\{G_i \ldots G_\#\}$ that contains a genus G_k constituted by a species tree T^S which consists of species $\{S_i, .., S_\#\}$ that contains a species S_k, and so on. The inclusions, with k representing an individual taxon, go as:

$$\ldots T^F = \{F_i \ldots F_\#\} \ni F_k = T^G = \{G_i \ldots G_\#\} \ni G_k = T^S = \{S_i, .., S_\#\} \ni S_k \ldots$$

Where, to keep notation from blowing up, # denotes some unspecified number.
The nested principle is the statement that a species tree is the **convex hull** of a genus node. Stated simply, downscale trees constitute upscale nodes. This allows for a variable of biological form (denoted ε) along a path i, giving tree location (denoted ε_i). This effectively equates biological form with pathwise scale. The contraction of form along a path (denoted $\partial\varepsilon_i \downarrow$) expands upscale nodes to downscale trees, while the expansion of time (denoted $\partial t \uparrow$) separates nodes. **Figure 2** visually illustrates the nested principle.

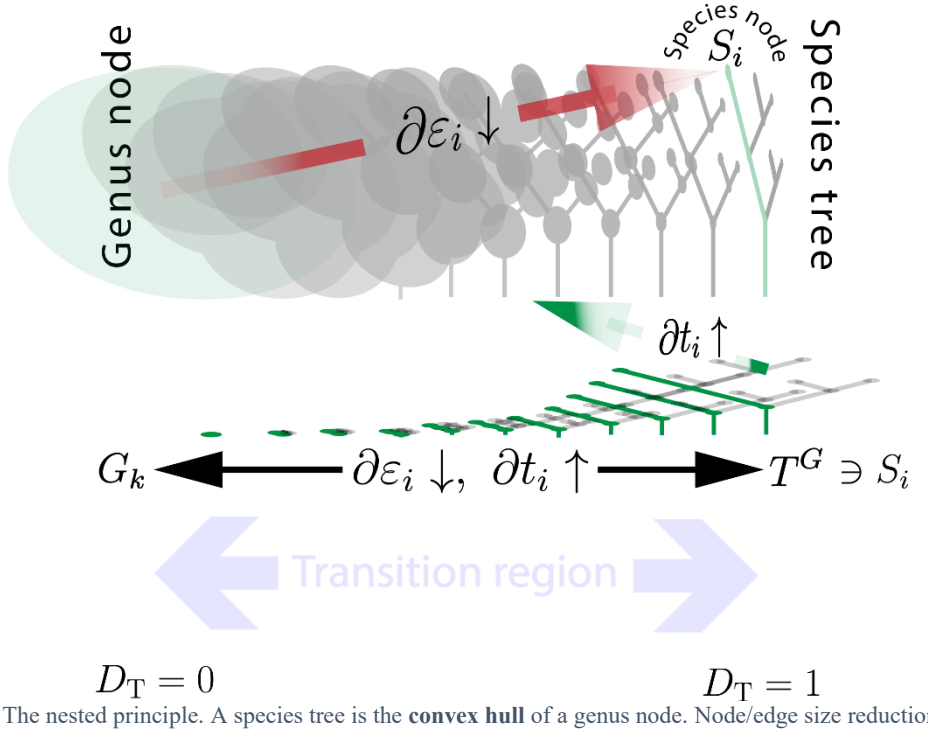

Figure 2 (TOP) The nested principle. A species tree is the **convex hull** of a genus node. Node/edge size reduction is used to represent scale ε (i.e., biological form). The contraction of scale expands a node to a tree. **Evolution** ($\partial\varepsilon$) diversifies form. (MIDDLE) The requirement of sufficient dimensions. As scale contractions diversify form, the corresponding branch elongations represent time elapse. (BOTTOM) Biological transitions. Representing the nested principle means incorporating transition regions into our model of phylogenetic space. The transitions occur as the topological dimension D_T changes from 0 (that of a point) to 1 (that of a line). These transitions allow for the incorporation of the organism-population duality principle.

Requirement 4: organism-population duality

Dynamic phylogenetics is not primarily concerned with the dead. For the dead, $\partial \mathrm{t} = \partial \varepsilon = 0 \Rightarrow \frac{\partial \varepsilon}{\partial t} = \frac{0}{0}$, and thus, there is no dynamic. Only for the living is $\partial t \neq 0$ and $\partial \varepsilon \neq 0$.

The central obstacle to formulating a dynamic phylogenetics lies in the *Population-Organism* duality. The mantra in biology is:

> Populations, not organisms, evolve. Selection operates on populations, not individuals (Nowak, 2006; or any number of references).

But the living things, the things for which ∂t is nonzero, the things generating the growth of the tree of life, are organisms. How do we bridge the gap between growth of the organism and evolution of the species? We need a unified theory of biology that explains both the growth of the organisms and evolution of the species.

The nested principle provides such a bridge by directly incorporating "biological phase transitions" into out model of phylogenetic space (**Figure 2**). A more thorough development of these ideas is beyond the scope of this report, and so omitted here.

Dynamical system

System description

The dynamic tree of life is modeled as a stochastic iterated function system, with biological form ε assigned to be the scale variable, and time t assigned to be branch lengths/The general idea is to generate a random tree using the Galton-Watson branching process, and then to randomly scale that tree (this is ε_0), where the scale is distributed uniformly on the unit interval. This is the first system iterate, and it generates the first tree $\mathcal{T}_0$ at a scale of ε_0.

For each leaf (terminal node) of tree $\mathcal{T}_0$, a new tree $\mathcal{T}_{0_i}$, is generated using the Galton-Watson process, and then this new tree is randomly scaled (this is ε_{0_i}), where the scale ε_{0_i} is distributed uniformly over the interval $(0, \varepsilon_0)$. In other words, every child tree is generated at a scale less than the scale of its parent leaf. The new tree $\mathcal{T}_{0_i}$ at scale ε_{0_i} is shifted so that it replaces the leaf of tree $\mathcal{T}_0$. Here i ranges from 1 to the number of leaves in tree $\mathcal{T}_0$. **Figure 3** gives a snapshot of the system.

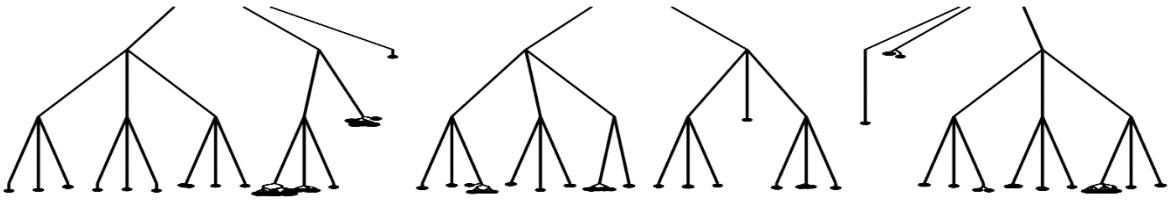

Figure 3 A snapshot of the system showing the nesting of trees within leaves.

It is claimed that this principle meets the requirement of sufficient dimensions, satisfies the random principle, the nested principle, and the has a mechanism, namely transition regions where the topological dimension D_T changes from 0 to 1, by which the organism-population duality can be approached (see **Figure 2**).

Equations of motion: stochastic iterated function system

It seems incorrect to claim that the iterated function system is the "equations of motion". "Motion" here should be understood as the lineage specific change in biological form with respect to the change in time, $\frac{\partial \varepsilon}{\partial t}$. The iterated function system does not by itself obtain for us such an expression. It does however, allow us to define metrics d_ε and d_t on the space generated by the IFS.

In this report, I have explained how to obtain the metric d_ε, from which we can obtain the expression $\partial\varepsilon$. The metric of time, d_t from which we would obtain the expression ∂t, is not deduced in this report. Obtaining an expression for d_t is complicated by the multifractal nature of the structure. Since the structure is a multifractal, the length of any branch (i.e., the quantity Δt or ∂t) depends on observation scale (i.e., the species from which the measurement is made). An expression for ∂t is omitted from this report. Therefore, what has been given in this report is really half of the dynamic – that is, half of the equations of motion.

Stochastic iterated function system

Auxiliary functions

First we need to define two functions, one that returns the root coordinates on the unit square of any system subtree, and another that returns the number of leaves in any subtree.

Let R: $[0,1] \times [0,1] \longrightarrow [0,1] \times [0,1]$ such that

$$R\left(\mathcal{T}_{0_{:i}}\right) = (a, b), \text{ the coordinates of the root node of tree } \mathcal{T}_{0_{:i}}.$$

Let L: $[0,1] \longrightarrow \mathbb{Z}$ such that

$$L\left(\mathcal{T}_{0_{:i}}\right) = \text{the number of leaves (terminal nodes) of tree } \mathcal{T}_{0_{:i}}, \text{ and}$$

Let l_{i_j} denote the jth leaf of tree $\mathcal{T}_{0_{:i}}$.

Ratio list

The ratio lists give the scale variables, which are those assigned to biological form.
For tree $\mathcal{T}_{0_{:i}}$ in the $K-1^{\text{th}}$ system iterate, the ratio list is given as:

$$\left(\varepsilon_{0_{:i_1}}, \ \ldots, \ \varepsilon_{0_{:i_{L\left(\mathcal{T}_{0:i}^{K-1}\right)}}}\right)$$

Where each $\varepsilon_{0_{:i_j}}$ is a random variable with $\varepsilon_{0_{:i_j}} \sim U\left(0, \varepsilon_{0_{:i}}\right)$, and $\varepsilon_1 \sim U(0,1)$.

Function list

The leaves l_{i_j} of tree $\mathcal{T}_{0_{:i}}$ in the K^{th} system iterate E_K, are comprised of $L\left(\mathcal{T}_{0_{:i}}^{K-1}\right)$ executions of function $T_{0_{:i_j}}^{K}$:

$$\left(T_{0_{:i_1}}^{K}, \ldots, T_{0_{:i_{L\left(\mathcal{T}_{0_{:i}}^{K-1}\right)}}}^{K}\right)$$

Where $i \in \mathbb{Z}$ with range $1 \leq i \leq L(\mathcal{T}^{K-1})$

Here, $T_{0_{:i_j}}^{K} : [0,1] \times [0,1] \longrightarrow [0,1] \times [0,1]$ is given by:

$$T_{0_{:i_j}}^{K} = \boldsymbol{Z} * \varepsilon_{0_{:i_j}} + ||R\left(T_{0_{:i_j}}^{K}\right) - l_{i_j}||_2 , l_{i_j} \in \mathcal{T}_{0_{:i}}^{K-1}$$

Where $\boldsymbol{Z}$ is a sequence generated by the Galton-Watson branching process:

$$Z_{p+1} = \begin{cases} \xi_1^{p+1} + \cdots + \xi_{Z_p}^{p+1}, & Z_p > 0 \\ 0, & Z_p = 0 \end{cases}$$

With $\xi_r^p \in \mathbb{Z}\ i.i.d.$ Nonnegative random variables, and $p, r \geq 0$.

Description of model correctness

The trees generated by the Galton-Watson process capture the stochasticity of biological branching processes. Since branch lengths give time intervals, and since the scale variable ε is also a random variable, the model captures the stochasticity of evolutionary change. That is, the model generates a system in which form is randomly diversified in time and in magnitude, while still maintaining lineage-specific characteristics.

Methods

The iterated function system was simulated using MATLAB R2020b. The maximum offspring and the maximum generations for the Galton-Watson process were each set to five. Since the systems tend to blow up fast, only seven iterations were executed.

Since each $\varepsilon_{0_{:i_j}}$ is a random variable with $\varepsilon_{0_{:i_j}} \sim U\left(0, \varepsilon_{0_{:i}}\right)$, we have that $\varepsilon_{0_{:i_j}} < \varepsilon_{0_{:i}}$ for all $\varepsilon_{:}$.
Therefore, we have:

$$\mathrm{P}\left(\mathcal{E}_{0_{:h_{i_j}}} = \varepsilon_{0_{:h_{i_j}}} \middle| \mathcal{E}_{0_{:h_i}} = \varepsilon_{0_{:h_i}}, \mathcal{E}_{0_{:h}} = \varepsilon_{0_{:h}}, \ldots, \mathcal{E}_0 = \varepsilon_0\right) = P\left(\mathcal{E}_{0_{:h_{i_j}}} = \varepsilon_{0_{:h_{i_j}}} \middle| \mathcal{E}_{0_{:h}} = \varepsilon_{0_{:h}}\right).$$

So every path in the system is a Markov chain with respect to the scale variable ε.

Entropy calculations

Since the scale random variable $\varepsilon_{0:i_j}$ is distributed uniformly over $(0, \varepsilon_{0:i})$, the entropy is simply:

$$H\left[\varepsilon_{0:i_j}\right] = log\left(\varepsilon_{0:i}\right).$$

Pathwise joint entropy calculations

The probability of child tree given the parent node is:

$$P\left(\varepsilon_{0:i_j} \middle| \varepsilon_{0:i}\right) = \frac{P\left(\varepsilon_{0:i_j}\right)}{P\left(\varepsilon_{0:i}\right)}$$

Therefore, the pathwise joint entropy is given by:

$$H\left[\varepsilon_{0:i_j}, \varepsilon_{0:i}\right] = log\left(\frac{\varepsilon_{0:i_j}}{\varepsilon_{0:i}}\right)$$

Pathwise mutual information calculations

From the entropy expressions, pathwise mutual information is given as:

$$I\left[\varepsilon_{0:i_j}; \varepsilon_{0:i}\right] = H\left[\varepsilon_{0:i_j}\right] + H\left[\varepsilon_{0:i}\right] - H\left[\varepsilon_{0:i_j}, \varepsilon_{0:i}\right].$$

Local channel capacity calculations

A "local channel" is a transition along a lineage from a leaf to a tree. (A "global channel" is a transition from system-wide iteration K to $K+1$.) **Figure 4** depicts a local channel.

Figure 4 Local channel capacity. The leaf i expands to a tree $\mathcal{T}_{i_j}$.

This is also the expression for the local channel capacity $\mathcal{C}_{ij}$ since

$$max_{p\left(\varepsilon_{0:i_j}\right)} I\left[\mathcal{E}_{0:i_j}; \mathcal{E}_{0:i}\right] = I\left[\varepsilon_{0:i_j}, \varepsilon_{0:i}\right].$$

To normalize the channel capacities by the number of leaves produced, the local channel capacity density is given as:

$$\rho_{ij} = \frac{\mathcal{C}_{ij}}{L\left(\mathcal{T}_{i_j}\right)}.$$

Cross-path joint entropy calculations

In addition to calculating the pathwise information measures, it is also desirable to calculate the mutual information and information distance between any two leaves in the system – that is, between two leaves that have an arbitrarily distanced common ancestor (**Figure 5**).

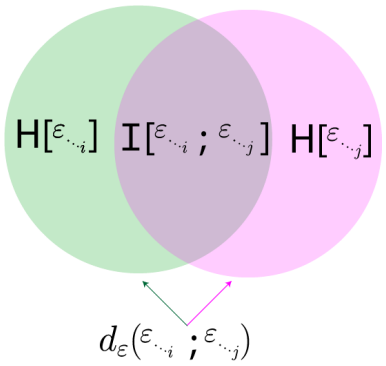

Figure 5 If i and j are any two leaves in the system, their mutual information and information distance can be determined. Information distance d_ε ultimately will be used as the metric of biological form (see **Figure 1**).

To calculate the joint cross-path entropy calculations (i.e., the entropy between any two leaves in the system) we must determine the joint probability of the scale variables of each leaf. This probability is given as:

$$P\left(\varepsilon_{0:i_j}, \varepsilon_{0:i_k}\right) = P\left(\varepsilon_{0:i_j}, \varepsilon_{0:i_k} \middle| \varepsilon_{0:i}\right) = P\left(\varepsilon_{0:i_j} \middle| \varepsilon_{0:i}\right) \cap P\left(\varepsilon_{0:i_k} \middle| \varepsilon_{0:i}\right) = \frac{P\left(\varepsilon_{0:i_j}\right)}{P\left(\varepsilon_{0:i}\right)} \cap \frac{P\left(\varepsilon_{0:i_k}\right)}{P\left(\varepsilon_{0:i}\right)}.$$

Therefore, the cross-path joint entropy is calculated as:

$$H\left[\varepsilon_{0:i_j}, \varepsilon_{0:i_k}\right] = log\left(min\left\{\frac{\varepsilon_{0:i_j}}{\varepsilon_{0:i}}, \frac{\varepsilon_{0:i_k}}{\varepsilon_{0:i}}\right\}\right)$$

Cross-path mutual information calculations

Just as in the pathwise case, the cross-path mutual information between any two leaves is calculated as:

$$I\left[\varepsilon_{0:i_j}; \varepsilon_{0:i_k}\right] = H\left[\varepsilon_{0:i_j}\right] + H\left[\varepsilon_{0:i_k}\right] - H\left[\varepsilon_{0:i_j}, \varepsilon_{0:i_k}\right]$$

Cross-path information distance calculations

The cross-path information distance is given by:

$$d_\varepsilon\left(\varepsilon_{0:i_j}, \varepsilon_{0:i_k}\right) = H\left[\varepsilon_{0:i_j}, \varepsilon_{0:i_k}\right] - I\left[\varepsilon_{0:i_j}; \varepsilon_{0:i_k}\right].$$

Example Results

Pathwise results

Figure 6 gives example information measure results for a system with seven iterations resulting in 20,643 leaves in the final iterate.

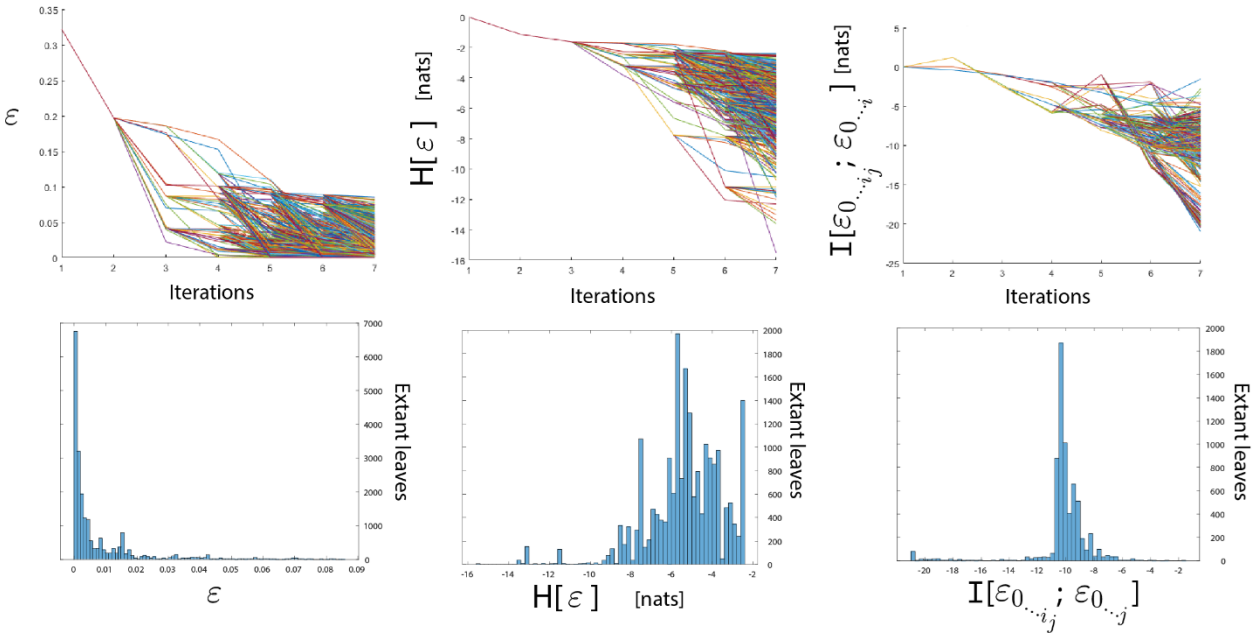

Figure 6 (LEFT COLUMN) The top plot gives the scale variable for each of the 20,646 paths in the system, and the bottom plot gives the distribution. (MIDDLE COLUMN) The top plot gives the entropy values for all 20,646 paths in the system, and the bottom plot gives their distribution. (RIGHT COLUMN) The top plot gives the pathwise mutual information (i.e., mutual information between parent leaf and child tree) for all 20,646 paths in the system, and the bottom plot gives their distributions.

Figure 7 gives example channel capacity density results for a system of seven iterations with 4,430 leaves, 1,018 trees in the final iterate.

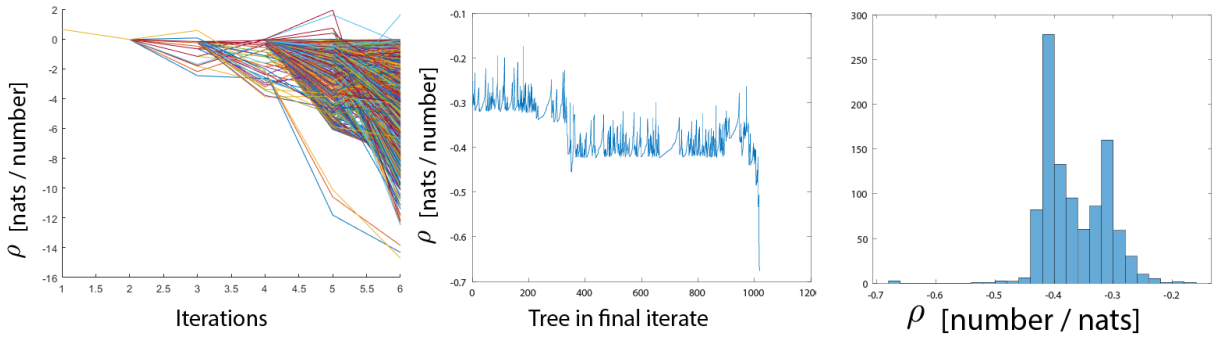

Figure 7 (LEFT) Channel capacity densities for all 4,430 system paths. (MIDDLE) Average channel capacity densities along paths to tree in final iterate. The sharp drop around the 1000th tree indicate the section of the system that is evolutionarily conserved. (RIGHT) Distribution of the channel capacity densities for all 4,430 system paths.

Cross-path results

Cross-path mutual informations and cross-path information distances were calculated by forming all pairwise combinations for a system of 6,616 leaves in the final seventh iterate. Histograms were then generated to see the distributions. **Figure 8** gives the result for the mutual informations, and **Figure 9** gives the result for the information distances.

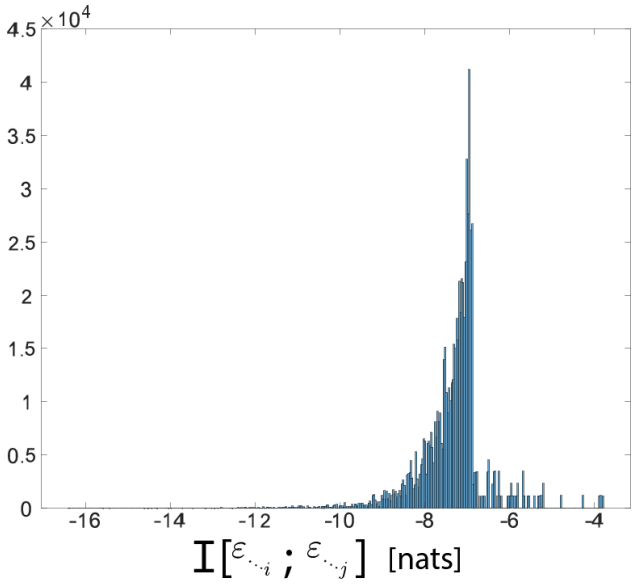

Figure 8 Mutual information for all pairwise combinations of 6,616 leaves in the final seventh iterate. *Self-comparisons are omitted.*

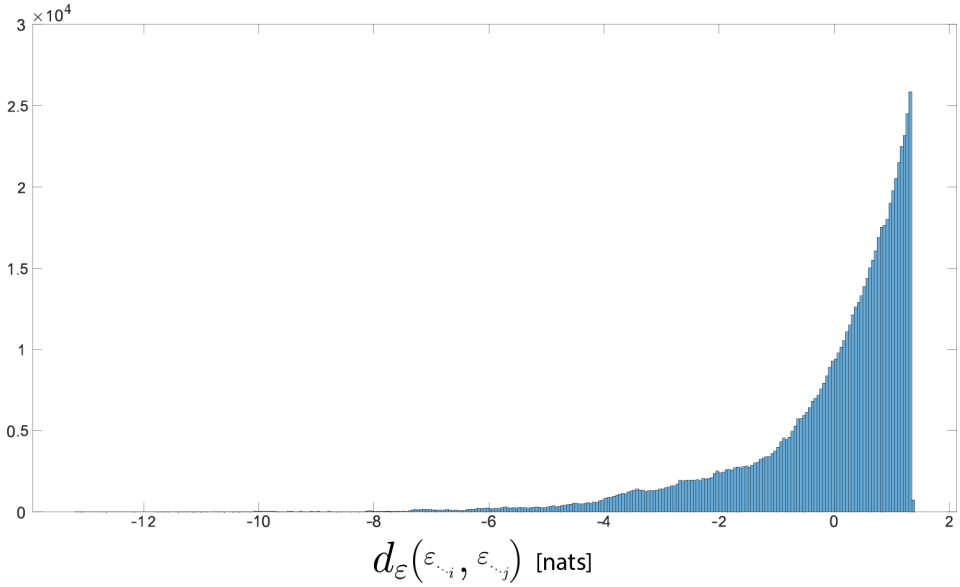

Figure 9 All unique pairwise combinations of information distances for 6,616 leaves in the final seventh iterate. Self-comparisons are omitted.

Conclusion

The system developed in this report accomplishes the tasks it set out to accomplish. A space with sufficient dimensions, that is both biological form and time, was constructed. The system is stochastic, satisfying the random principle. The system also satisfies the nested principle, thereby giving meaning to the assignment of biological form ε to scale. In doing so, the system provides a path forward for dealing with the organism-population duality principle. This path forward comes from the transition regions inherent in the system and their corresponding changes in topological dimension. The result is a system on which information measures can be clearly defined and calculated. The information measures allow for the metric of biological form to be defined as the information distance between scale variables of leaves.

The metric of time was not dealt with in this report. The multifractal nature of the structure means that such a metric must be a function of scale. In order to develop such a metric, the spectrum of fractal dimensions across the structure must first be attained. Determining this spectrum and the metric of time are the future directions of the project.

Bibliography

Bleidorn, C. (2017). *Phylogenomics*. Springer International Publishing. https://doi.org/10.1007/978-3-319-54064-1

Cover, T, and Thomas, J. (2006) Elements of Information Theory, 2nd Edition | Wiley. (n.d.).

Darwin, C.R. (1859). (1st ed) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. London: John Murray.

Hitchcock, E. (1840). (1st ed.) Elementary Geology. Amherst: J. S. & C. Adams.

Lamarck, J. B. (1809). Philosophie zoologique. Paris: Dentu.

Nowak, M. A. (2006) Evolutionary Dynamics. Belknap/Harvard.