Exposition on Phenotype Spaces: A Topological Model of Evolutionary Proximity

Crista Moreno

January 17, 2017

The construction of a topological space for the secondary structure of RNA molecules.

Outline

Genotype-Phenotype Relationship

Genome is the genetic material of an organism or the inherited instructions that are vital to life.

Genotype-Phenotype Relationship

Genome is the genetic material of an organism or the inherited instructions that are vital to life.

Phenotype is the observable characteristics of an organism or the realization of the code written in the genome.

Genotype-Phenotype Relationship

Genome is the genetic material of an organism or the inherited instructions that are vital to life.

Phenotype is the observable characteristics of an organism or the realization of the code written in the genome.

Genotype is the collection of inherited genes responsible for the phenotype of a particular individual or organism.

Genotype-Phenotype Relationship

Genome is the genetic material of an organism or the inherited instructions that are vital to life.

Phenotype is the observable characteristics of an organism or the realization of the code written in the genome.

Genotype is the collection of inherited genes responsible for the phenotype of a particular individual or organism.

Example

The collection of genes responsible for eye color for a particular individual is the genotype and the observable eye coloration is the phenotype.

RNA: genotype sequence, primary structure

Definition

RNA (ribonucleic acid) is a nucleic acid made up of the four nucleotides: guanine (G), cytosine (C), adenine (A), and uracil (U) and is essential to life.

RNA: genotype sequence, primary structure

Definition

RNA (ribonucleic acid) is a nucleic acid made up of the four nucleotides: guanine (G), cytosine (C), adenine (A), and uracil (U) and is essential to life.

Example

GUGAUGGAUU AGGAUGUCCU ACUCCUUUGC UCCGUAAGAU
AGUGCGGAGU UCCGAACUUA CACGGCGCGC GGUUAC

Example

GUGAUGGAUU AGGAUGUCCU ACUCCUUUGC UCCGUAAGAU AGUGCGGAGU UCCGAACUUA CACGGCGCGC GGUUAC

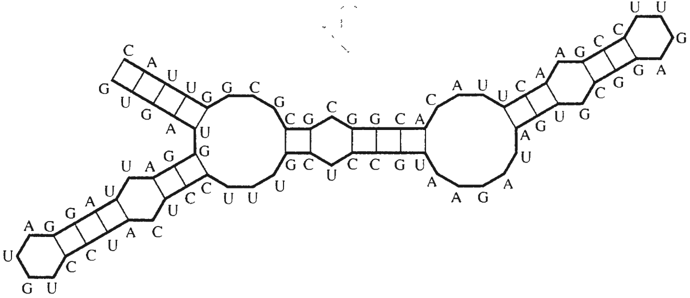


Figure 1 : RNA molecule folded and bonded chain of nucleotides.

RNA Shape Bonding Diagram

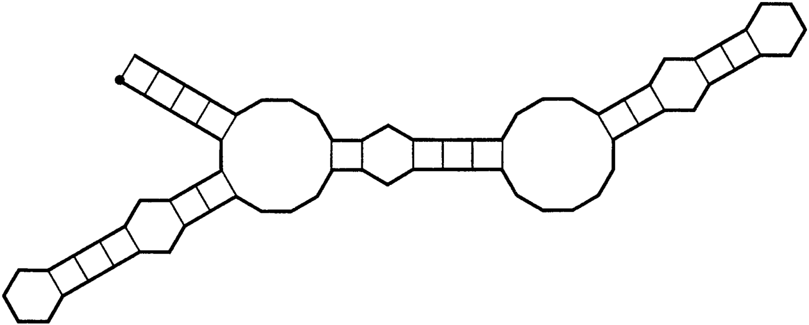


Figure 2 : RNA molecule bonding diagram (**secondary structure** or **shape**).

Different nucleotide sequences may generate the same secondary structure or shape.

RNA genotype and phenotype relationship

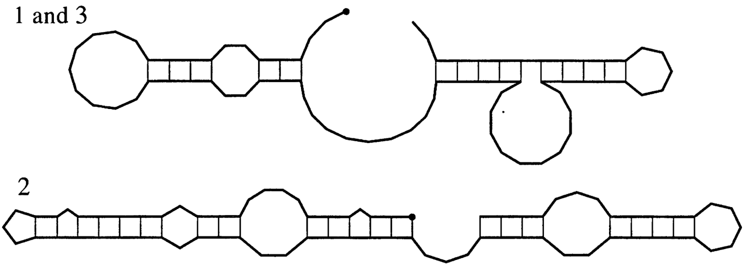


Figure 3 : Different bonding diagrams for different nucleotide sequences.

1. GGGCAGUCUC CCGGCGUUUA AGGGAUCCUG AACUUCGUCG CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG
2. GGGCAGUCUC CCGGCGUUUA AGGAAUCCUG AACUUCGUCG CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG
3. GGGCAGUCUC CCGGCCUUUA AGGGAUCCUG AACUUCGUCG CUCCCAUCCA AUCAGUCCGC CUCACGGAUG GAGUUG

Definition

A **point mutation** is a mutation from one genotype sequence to another obtained by changing a single entry in the genotype sequence.

Definition

A **point mutation** is a mutation from one genotype sequence to another obtained by changing a single entry in the genotype sequence.

Example

GCCGCGCC
GCGGCGCC

Consider all the 10 nucleotide sequences consisting of only guanine and cytosine

$$\underline{?}\ \underline{?}\ \underline{?}\ \underline{?}\ \underline{?}\ \underline{?}\ \underline{?}\ \underline{?}\ \underline{?}\ \underline{?}$$

Consider all the 10 nucleotide sequences consisting of only guanine and cytosine

? ? ? ? ? ? ? ? ? ?

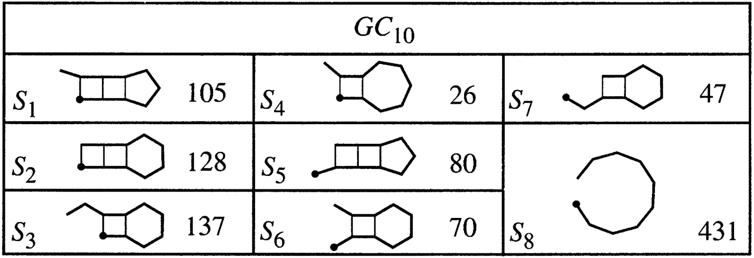


Figure 4 : The eight RNA shapes $\{\sigma_1, \ldots, \sigma_8\}$ that result from $2^{10} = 1024$ possible 10 nucleotide sequences consisting of only guanine (G) and cytosine (C).

The genotype-phenotype relationship is a many-to-one relationship.

Mathematical Model

Folding of the RNA Sequence to its Corresponding Shape

The folding process is modeled as a surjective function

$$\phi : S \to \Sigma$$

from the set S of all sequences of fixed length onto the set Σ of all minimum free energy secondary structures corresponding to that length.

Folding of the RNA Sequence to its Corresponding Shape

The folding process is modeled as a surjective function

$$\phi : S \to \Sigma$$

from the set S of all sequences of fixed length onto the set Σ of all minimum free energy secondary structures corresponding to that length.

Definition

Let $\alpha, \beta \in \Sigma$, then we say that β is **accessible** from α, $\alpha \mapsto \beta$ if there exists a pair $a, b \in S$ such that a and b differ by a point mutation, and $\phi(a) = \alpha$ and $\phi(b) = \beta$.

Folding of the RNA Sequence to its Corresponding Shape

The folding process is modeled as a surjective function

$$\phi : S \to \Sigma$$

from the set S of all sequences of fixed length onto the set Σ of all minimum free energy secondary structures corresponding to that length.

Definition

Let $\alpha, \beta \in \Sigma$, then we say that β is **accessible** from α, $\alpha \mapsto \beta$ if there exists a pair $a, b \in S$ such that a and b differ by a point mutation, and $\phi(a) = \alpha$ and $\phi(b) = \beta$.

Definition

The set of all genotype sequences in S that result in a particular RNA shape $\sigma \in \Sigma$ is called the **neutral network** or neutral set of σ and is denoted $N(\sigma) = \{s | \sigma = \phi(s)\}$.

Construction of Point Mutation Probabilities

Model

Let $\alpha, \beta \in \Sigma$

- $m_{\alpha,\beta}$ is the number of point mutations that change a sequence from $N(\alpha)$ to $N(\beta)$

Model

Let $\alpha, \beta \in \Sigma$

- $m_{\alpha,\beta}$ is the number of point mutations that change a sequence from $N(\alpha)$ to $N(\beta)$
- $m_{\beta,\alpha}$ is the number of point mutations that change a sequence from $N(\beta)$ to $N(\alpha)$

Model

Let $\alpha, \beta \in \Sigma$

- $m_{\alpha,\beta}$ is the number of point mutations that change a sequence from $N(\alpha)$ to $N(\beta)$
- $m_{\beta,\alpha}$ is the number of point mutations that change a sequence from $N(\beta)$ to $N(\alpha)$
- Note that $m_{\alpha,\beta} = m_{\beta,\alpha}$

Model

Let $\alpha, \beta \in \Sigma$

- $m_{\alpha,\beta}$ is the number of point mutations that change a sequence from $N(\alpha)$ to $N(\beta)$
- $m_{\beta,\alpha}$ is the number of point mutations that change a sequence from $N(\beta)$ to $N(\alpha)$
- Note that $m_{\alpha,\beta} = m_{\beta,\alpha}$
- $m_{\alpha,\star}$ is the number of point mutations that change a sequence to a sequence in any other neutral network.

Model

Let $\alpha, \beta \in \Sigma$

- $m_{\alpha,\beta}$ is the number of point mutations that change a sequence from $N(\alpha)$ to $N(\beta)$
- $m_{\beta,\alpha}$ is the number of point mutations that change a sequence from $N(\beta)$ to $N(\alpha)$
- Note that $m_{\alpha,\beta} = m_{\beta,\alpha}$
- $m_{\alpha,\star}$ is the number of point mutations that change a sequence to a sequence in any other neutral network.

Definition

The **mutation probability**, $p_{\alpha,\beta}$ is defined by

$$\boxed{p_{\alpha,\beta} = \frac{m_{\alpha,\beta}}{m_{\alpha,\star}}}$$

Notice that $p_{\alpha,\beta}$ may not necessarily be equal to $p_{\beta,\alpha}$ since $m_{\alpha,\star}$ may not equal $m_{\beta,\star}$. Hence, the mutation probability is asymmetric and does not serve as a metric for mutation probabilities.

Definition Metric (Distance Function)

Definition

Let X be a non-empty set. A real-valued function d defined on $X \times X$, i.e. ordered pairs of elements in X, is called a **metric** or **distance function** on X iff it satisfies, for every $a, b, c \in X$ the following axioms:

(M1) $d(a, b) \geq 0$ and $d(a, a) = 0$

(M2) (Symmetry) $d(a, b) = d(b, a)$

(M3) (Triangle Inequality) $d(a, c) \leq d(a, b) + d(b, c)$

(M4) If $a \neq b$, then $d(a, b) > 0$

Table 1 : Mutation probabilities for the RNA shapes $\sigma_1, \ldots, \sigma_8$.

	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
σ_1		0.13	0.15	0.08	0.07	0.09	0.04	0.44
σ_2	0.11		0.15	0	0.11	0.18	0.05	0.40
σ_3	0.12	0.15		0.03	0.09	0.06	0.05	0.50
σ_4	0.29	0	0.14		0.07	0.09	0.06	0.35
σ_5	0.08	0.15	0.12	0.02		0.08	0.08	0.47
σ_6	0.12	0.29	0.09	0.03	0.09		0.03	0.35
σ_7	0.08	0.12	0.13	0.03	0.13	0.05		0.46
σ_8	0.18	0.19	0.24	0.04	0.15	0.11	0.09	

Construction RNA Shape Space

Algorithm 1 Construction RNA Shape Space

for $i = 1, \ldots, 8$ **do**
  Let $R_i = \{\sigma_i \cup \{\sigma_j | p_{i,j} > 1/7\}\}$
end for

Mutation Probabilities

Table 2 : Construction of $\mathscr{R}_{1/7}$ from Table 1 Mutation probabilities.

	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
σ_1	✓		✓					✓
σ_2		✓	✓			✓		✓
σ_3		✓	✓					✓
σ_4	✓			✓				✓
σ_5		✓			✓			✓
σ_6		✓				✓		✓
σ_7							✓	✓
σ_8	✓	✓	✓		✓			✓

$$\mathscr{R}_{1/7} = \{\{\sigma_1, \sigma_3, \sigma_8\}, \{\sigma_2, \sigma_3, \sigma_6, \sigma_8\}, \{\sigma_2, \sigma_3, \sigma_8\}, \{\sigma_1, \sigma_4, \sigma_8\}, \{s_2, s_5, s_8\}, \{\sigma_2, \sigma_6, \sigma_8\}, \{\sigma_7, \sigma_8\}, \{\sigma_1, \sigma_2, \sigma_3, \sigma_5, \sigma_8\}\}$$

Mutation Probabilities

Table 2 : Construction of $\mathscr{R}_{1/7}$ from Table 1 Mutation probabilities.

	σ_1	σ_2	σ_3	σ_4	σ_5	σ_6	σ_7	σ_8
σ_1	✓		✓					✓
σ_2		✓	✓			✓		✓
σ_3		✓	✓					✓
σ_4	✓			✓				✓
σ_5		✓			✓			✓
σ_6		✓				✓		✓
σ_7							✓	✓
σ_8	✓	✓	✓		✓			✓

$$\mathscr{R}_{1/7} = \{\{\sigma_1, \sigma_3, \sigma_8\}, \{\sigma_2, \sigma_3, \sigma_6, \sigma_8\}, \{\sigma_2, \sigma_3, \sigma_8\}, \{\sigma_1, \sigma_4, \sigma_8\}, \{s_2, s_5, s_8\}, \{\sigma_2, \sigma_6, \sigma_8\}, \{\sigma_7, \sigma_8\}, \{\sigma_1, \sigma_2, \sigma_3, \sigma_5, \sigma_8\}\}$$

(Note:$\mathscr{R}_{1/7} = \{R_i\}_1^8$ is not a topology.)

Definition Topological Space

Definition

Let X be a nonempty set. A collection $\mathcal{T}$ of subsets of X is a **topology** on X if and only if $\mathcal{T}$ satisfies the following axioms:

O1: $\emptyset$ (empty set) and X are in $\mathcal{T}$.

O2: The union of the elements of any subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

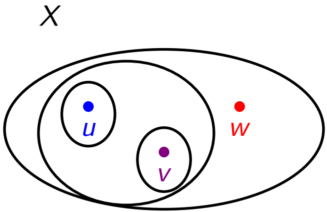
O3: The intersection of the elements of any finite subcollection of $\mathcal{T}$ is in $\mathcal{T}$.

The members of $\mathcal{T}$ are then called $\mathcal{T}$-**open sets**, or simply **open sets**. The pair $(X, \mathcal{T})$ is called a **topological space**.

Example Topological Space

Example

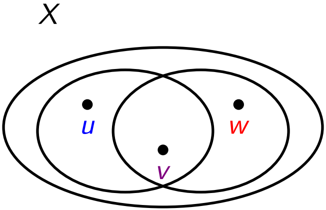
$X = \{u, v, w\}$, $\mathcal{T} = \{X, \emptyset, \{u\}, \{v\}, \{u, v\}\}$



Non-Example Topological Space

Example

$X = \{u, v, w\}$, $\mathcal{T} = \{X, \emptyset, \{u, v\}, \{v, w\}\}$ is **NOT** a topology on X. Why?



Construction of the minimal basis for the phenotype space of RNA shapes.

Algorithm 2

Extend $\mathscr{R}_{1/7}$
Let $\mathscr{T}_{1/7}$ be the **minimal topology** on GC_{10} containing $\mathscr{R}_{1/7}$.
for σ_i **do**
  take all of the sets R_j that contain σ_i
  let B_i be their intersection.
end for
$\mathscr{B}_{1/7} = \{B_i\}_{i=1}^{8}$ is the minimal basis for $\mathscr{T}_{1/7}$

Construction of the minimal basis for the phenotype space of RNA shapes.

Algorithm 2

Extend $\mathscr{R}_{1/7}$
Let $\mathscr{T}_{1/7}$ be the **minimal topology** on GC_{10} containing $\mathscr{R}_{1/7}$.
for σ_i **do**
    take all of the sets R_j that contain σ_i
    let B_i be their intersection.
end for
$\mathscr{B}_{1/7} = \{B_i\}_{i=1}^{8}$ is the minimal basis for $\mathscr{T}_{1/7}$

This resulting topological space is referred to as a **phenotype space**.

Definition Basis

Definition

Let X be a set and $\mathscr{B}$ be a colection of subsets of X. We say $\mathscr{B}$ is a **basis** (for a topology on X) if the folowing statements hold:

(i) For each $x \in X$, there is a $B \in \mathscr{B}$ such that $x \in B$.

(ii) If B_1 and B_2 are in $\mathscr{B}$ and $x \in B_1 \cap B_2$, then there exists B_3 in $\mathscr{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Definition Basis

Definition

Let X be a set and $\mathscr{B}$ be a colection of subsets of X. We say $\mathscr{B}$ is a **basis** (for a topology on X) if the folowing statements hold:

(i) For each $x \in X$, there is a $B \in \mathscr{B}$ such that $x \in B$.

(ii) If B_1 and B_2 are in $\mathscr{B}$ and $x \in B_1 \cap B_2$, then there exists B_3 in $\mathscr{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Example

The singleton sets form a basis for the discrete topology. (All subsets of the space are open sets.)

Definition Basis

Definition

Let X be a set and $\mathscr{B}$ be a colection of subsets of X. We say $\mathscr{B}$ is a **basis** (for a topology on X) if the folowing statements hold:

(i) For each $x \in X$, there is a $B \in \mathscr{B}$ such that $x \in B$.

(ii) If B_1 and B_2 are in $\mathscr{B}$ and $x \in B_1 \cap B_2$, then there exists B_3 in $\mathscr{B}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Example

The singleton sets form a basis for the discrete topology. (All subsets of the space are open sets.)

Let $\mathscr{B}$ be a basis on a set X. The **topology** $\mathscr{T}$ **generated by** $\mathscr{B}$ is obtained by defining the open sets to be the empty set and every set that is equal to a union of basis elements.

Basis $\mathscr{B}_{1/7}$ for topology $\mathscr{T}_{1/7}$

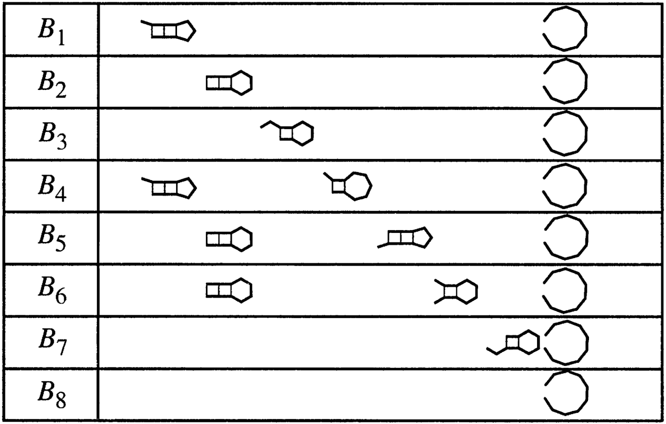


Figure 5 : Basis $\mathscr{B}_{1/7}$ for topology $\mathscr{T}_{1/7}$

Basis $\mathscr{B}_{1/7}$ for topology $\mathscr{T}_{1/7}$

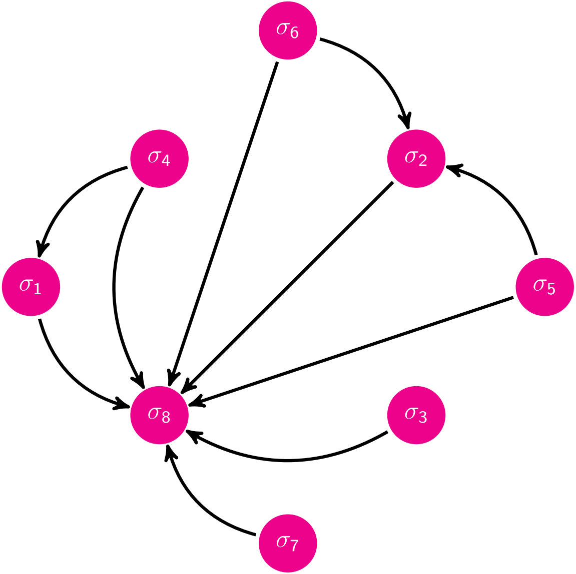


Figure 6 : Basis $\mathscr{B}_{1/7}$ represented by a directed graph.

This gives a formal way of defining continuous and discontinuous evolutionary changes.

Basis $\mathscr{B}_{1/10}$ for topology $\mathscr{T}_{1/10}$

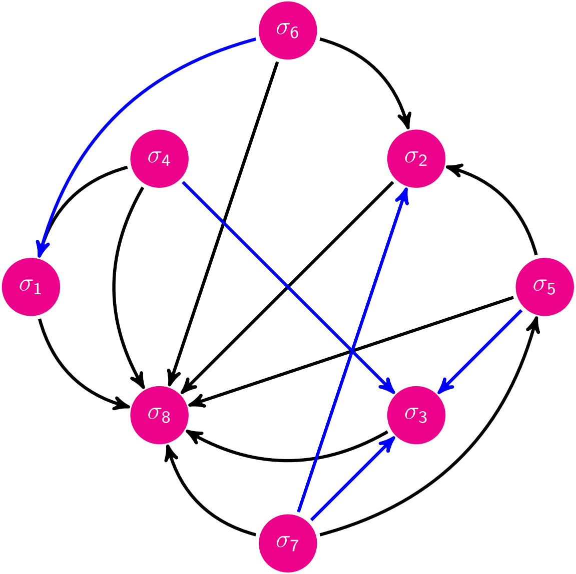


Figure 7 : Basis $\mathscr{B}_{1/10}$ represented by a directed graph.

References

1. Fontana, W., and Schuster, P. "Continuity in Evolution: On the Nature of Transitions" *Science* **280** (1998), 1451-1455.
2. Munkres, *Topology*, 2nd Edition (1999)
3. Figures (1 - 6) Adams, Franzosa, *Introduction to Topology Pure and Applied.* Pearson Prentice Hall Upper Saddle River, 2008.
4. W.F. and P.S. Shaping Space: the possible and the attainable in RNA genotype-phenotype mapping. *Journal of Theoretical Biology*, 194(4):491-5515, 1998.

Thank you!

Questions?