The Role of Topology in the Study of Evolution

Avery Broome

August 31, 2015

Abstract

In this paper, we will attempt to understand the role topology plays in analyzing RNA secondary structures by providing an overview and interpretation of the paper *RNA Shape Space Topology*. The main problem that the authors of this paper are trying to solve is that of defining a continuous and discontinuous transition between two RNA secondary structures. To that end, they turn to topology. Once they have defined a topology, they can then define functions that describe the evolution of RNA secondary structures and have a rigorous idea of what it means for those functions to be continuous and discontinuous. The use of topology ultimately provides a sound mathematical basis for understanding and determining accessibility between secondary structures.

Contents

1 Correspondence Between Biological and Topological Terms

First we must take care to define the terms used in the paper and to understand how they correspond to the usual concepts of topology. Before we

begin that task, however, we will explain some terms and concepts which may be unfamiliar to the non-biologist.

Definition 1.1. Secondary Structure
The secondary structure of an RNA molecule is simply the shape of the molecule in 3-D space, rather than the placement of nucleotides or atoms in relation to each other. An equivalent term is the phenotype.

Definition 1.2. Genotype
The genotype of an RNA molecule is the composition of the strand of RNA, or more precisely the exact sequence of base pairs.

Definition 1.3. Neutral Set
The neutral set $\mathrm{S}(x)$ of a secondary structure x is the set of all RNA sequences $\big($the sequence of nucleotides$\big)$ capable of folding into x.

Definition 1.4. Shape Space
The shape space is the set of all RNA phenotypes produced by any possible RNA genotype.

Definition 1.5. Boundary
The boundary of $\mathrm{S}(x)$ is denoted $\delta\mathrm{S}(x)$ and is the set of all point mutations of elements of $\mathrm{S}(x)$. The union of $\mathrm{S}(y)$ and $\delta\mathrm{S}(x)$, or the set of all neighbors of the neutral set of x that fold into y, is denoted $\mathrm{D}\big(\mathrm{y} \leftarrow x\big)$.
Note: "fold into" refers to the process by which secondary structures are formed: a sequence of nucleotides gets folded into a certain shape.
"neighbors" refers to all sequences obtained by taking a single point mutation of the sequences in the neutral set of x.

Definition 1.6. Accessibility
The phrase "x accessible from y," where x and y are secondary structures, means that if y undergoes a single point mutation in its genotype, it can obtain the structure of x. It translates to the two structures being topologically close.

Now that these terms have been defined, we may begin to identify how they correspond to topological structures.

- The set the topology is on is the shape space V.
- The subbasis is given as the set $\mathcal{N}= \big\{N\big(x\big) \mid x \in \mathrm{V}\big\}$ where $N\big(x\big) = \big\{y \in \mathrm{V} \mid y \leftarrow x \big\}$.
- The unique non-redundant basis is given as the set $\mathcal{B} = \big\{B\big(x\big) \mid x \in \mathrm{V}\big\}$, where $B\big(x\big) = \underset{w|x\in N(w)}{\cap} \mathrm{N}\big(w\big)$.

2 An Intuitive Understanding

Now that we have the necessary vocabulary, we will explore an intuitive understanding of what accessibility means and what constructing an accessibility topology means. Let's say we have two structures, a "red" structure and a "blue" structure. As explained in the definitions above, each of these structures has a neutral set, or network, of sequences associated with it, a "red" network and a "blue" network. This network is the set of all sequences folding into the "red" or "blue" structure. Accessibility $\big($say, accessibility of blue from red$\big)$ is the probability that if you chose a random point on the red network, you would be able to take a single step and reach the blue network. In this example, we can now define a neighborhood for "red." All we need to do is choose a threshold of accessibility. In other words, choose a threshold value for the probability just discussed; if the probability is above the threshold, then the "blue" structure is in the neighborhood of the "red" structure. With a definition for neighborhood, we also have a pre-topology and can arrive at a topological definition of continuity. [1]

3 Rigor About the Basis and Subbasis

Before we go on, we must ensure that we are on sound mathematical footing. Since no proofs have been given in the paper that the given basis and subbasis are indeed a basis and a subbasis, we will provide them here. For our own satisfaction, we will also provide a proof that the finite topology has a unique non-redundant basis

Theorem 1. $\mathcal{B} = \Big\{\ B\big(x\big) = \underset{w|x\in N(w)}{\cap} N\big(w\big) \mid x \in \mathrm{V}\Big\}$ is a basis of τ, the topology on V.

Proof. We need to show:

1. For arbitrary $x \in \mathrm{V}$, there exists a basis element $B\big(x\big)$ such that $x \in B\big(x\big)$
2. If $x \in B_{x_1}\big(x\big) \cap B_{x_2}\big(x\big)$, there exists a $B_{x_3}\big(x\big)$ such that $x \in B_{x_3}\big(x\big)$, which is contained in the intersection of $B_{x_1}\big(x\big)$ and $B_{x_2}\big(x\big)$

First show condition 1:
The basis element $\mathrm{B}\big(x\big)$ is the intersection of all $\mathrm{N}\big(w\big)$ that contain x, so their intersection will contain x. Thus for any x, we simply choose a w such

[1] Walter Fontana, personal communication, July 29 2015

that x is in $\mathrm{N}(w)$ and then we have a $\mathrm{B}(x)$ that contains x. This makes sense in the biological context because for any possible structure x, there must exist some w such that x is accessible from w. This means it is possible to simply choose a w such that x is in $\mathrm{N}(w)$.

Now show condition 2:

To show this condition, we'll first show that if $x \in B_{x_i}$, where B_{x_i} is a basis element, then $B_x \subseteq B_{x_i}$. (Note that we already know by definition of B_x that $x \in B_x$)

If we take as given that $x \in B_{x_i}$, it follows from the definition that:

$$x \in \bigcap_{x_i \in N(w)} N(w)$$

So

$$x \in \mathrm{N}(w) \text{ for arbitrary } w \text{ such that } x_i \in \mathrm{N}(w)$$

Said another way, this means:

$$\text{If } x_i \in \mathrm{N}(x), \text{ then } x \in \mathrm{N}(x)$$

Which implies

$$\text{the set } \{y \mid x_i \in \mathrm{N}(w)\} \subseteq \{w \mid x \in \mathrm{N}(w)\}$$

So it follows that

$$\bigcap_{x_i \in N(w)} N(w) \subseteq \bigcap_{x \in N(w)} N(w)$$

By definition

$$B_x \subseteq B_{x_i}$$

The rest of the proof follows. We know by the givens that $x \in B_{x_1}, B_{x_2}$. We know there is always a set B_x that contains x; let B_x be the B_{x_3} mentioned in the statement of the proof. Then we know from what we just proved that because $B_x \in B_{x_1}, B_{x_2}$, there is a B_{x_3} containing x such that $B_{x_3} \in B_{x_1}(x) \cap B_{x_2}(x)$. Note that this is true even if B_{x_1} or B_{x_2} is B_x. Then we'd only have to show that B_x is contained in, for example, B_{x_1}. $\square$

Theorem 2. $\mathcal{S} = \{\mathrm{N}(x) : x \in \mathrm{V}\}$is a subbasis

Proof. We have to show that $\underset{x\epsilon V}{\cup} N\big(x\big) = \mathrm{V}$.
We know $\mathrm{N}\big(x\big) = \big\{w\ \epsilon\ \mathrm{V} \mid w \leftarrow x\big\}$ So $\mathrm{N}(x)$ is the set of all structures w accessible from x. Since V is the set of all possible structures, all structures in V must be accessible from some x, otherwise they wouldn't be possible. Once we take the union of $\mathrm{N}(x)$ for all x in V, we will have the set of V, since all the elements in V will have been produced. Thus S is a subbasis. □

Theorem 3. A finite topological space X has a unique non-redundant basis.

Proof. Suppose that $\mathcal{B}$ and $\mathcal{B\prime}$ are non-redundant bases that give the same topology. We will show that $\mathrm{B} \subseteq \mathrm{B\prime}$ and that $\mathrm{B\prime} \subseteq \mathrm{B}$ by symmetry, which will show $\mathcal{B} = \mathcal{B\prime}$
Let U be an open set in B. We know U is open, so we can express U as

$$\mathrm{U} = \underset{V_i\epsilon B\prime}{\cup} V_i \text{ for some sets } V_i \in \mathrm{B\prime}$$

We need to show U is contained in B′
If U isn't contained in $\mathcal{B\prime}$, then

$$V_i \neq \mathrm{U}$$

The fact that V_i is open means

$$V_i = \underset{u_j\epsilon B}{\cup} u_j$$

That means

$$u_j \subset V_i \subset \mathrm{U}$$

But

$$V_i \neq \mathrm{U}, \text{ so } u_j \neq \mathrm{U}$$

This contradicts the definition of non-redundant. Therefore

$$\mathrm{U}\ \epsilon\ \mathrm{B\prime} \text{ and so } \mathrm{B} \subseteq \mathrm{B\prime}$$

By symmetry, we know

$$\mathrm{B\prime} \subseteq \mathrm{B}$$

This means

$$\mathrm{B} = \mathrm{B\prime}$$

If any two non-redundant bases are equal, then there is only one non-redundant basis. □

4 The Fontana-Schuster Topology

The authors present different ideas of accessibility, but the one they give the most credence to is the Fontana-Schuster topology.

The idea of accessibility upon which the Fontana-Schuster topology is based has three possible scenarios of accessibility. A secondary structure y is accessible from a secondary structure x if:

1. y was produced by x from shortening a stack of x
2. y was produced by x from lengthening a stack of x
3. y was produced by x by eliminating one stack of x

The "stack" of x merely refers to an actual stack of base pairs. Thus, shortening a stack would be eliminating some base pairs and elongating it would be adding some. The topology based on these conditions of accessibility would be the one defined before, generated by the basis $\mathcal{B} = \{B(x) \mid x \in \mathrm{V}\}$. However, the definition of $N(x)$ is made more precise. While $N(x)$ is the set of all secondary structures y such that y is accessible from x as it was before, "accessible" now means that one of the three conditions listed above is true.

A possible reason that this is considered more realistic is because the Fontana-Schuster topology bears resemblance to other topologies based on reasonable threshold values. In this paper, threshold values are set on $|\mathrm{D}(y \leftarrow x)|$. If the number of neighbors of x that fold into y is above a certain limit, we can say x is accessible from y. This makes sense; if, say, there was one neighbor of x that folded into y, it would certainly be unlikely that x would ever change to y by a single point mutation. The fact that the Fontana-Schuster topology gives similar results to threshold topologies when reasonable values are chosen for thresholds indicates that it may more accurately represent the accessibility relations in shape space than other topologies presented in this paper.

A reasonable question at this point would be "so what?" Now that we have identified a promising topology for modeling accessibility among secondary structure, what will we do with it? We have a topology on V, so we can define functions that map into V and determine if they are continuous. The functions we will consider will be f, φ and θ. The first is the folding map f. This function maps from the set of all RNA sequences to the shape space V. The set of all sequences that map to a particular secondary structure u is denoted as $\mathrm{S}(u)$. f is not injective, as there may be more than one element

in S(u), but it is surjective. Clearly, every possible secondary structure u must have at least one RNA strand that can be folded into that particular shape. This function is important because once we have this function, we can define a domain for it. The size of the domain in this case is the number of RNA sequences that fold into a certain structure.

The second function, which is defined in the "Trajectories and Transitions" section of the paper, is φ. It maps an interval $\left[\alpha,\, \beta\right]$, where $\alpha = t_0 < t_1 < \ldots < t_m < t_{m+1} = \beta$, in the real numbers to V. Certain properties are also defined for the function:

1. φ is constant on every open interval from t_k to t_{k+1} except on those intervals including α and β. Those are of course half closed.

2. φ does not take any of these intervals to the same structure in V.

3. Any point t_k within the interval $\left[\alpha,\, \beta\right]$ is taken to the same value by φ as the interval is that preceded it or the interval that comes after it.

The biological concept this function represents is that of evolution over time. Each element in the interval $\left[\alpha,\, \beta\right]$ represents a point in time, and so each interval represents a different stage of evolution. This context makes the first property of φ logical. Evolution works very slowly, so given a certain finite interval of time, there would be no change in the RNA secondary structure.

The second and third conditions are also easily comprehensible when viewed through a biological lens. Clearly time intervals would be chosen such that each interval represented a different stage in RNA evolution, so none of the intervals would translate to the same RNA structure under φ. The third condition is also necessary to make a plausible evolutionary model. One singular point in time could not correspond to an entirely different RNA structure than the intervals of time around it. This would mean that that structure had lasted for an infinitesimally short period of time, which is not plausible from an evolutionary standpoint.

The final function, θ, maps an interval $\left[t_k,\, t_{k+2}\right]$ to a two value subset of V $\left\{x,y\right\}$ in such a way that $\theta(t_k,\, t_{k+1})$ is equal to $\left\{x\right\}$ while $\theta(t_{k+1},\, t_{k+2})$ is equal to $\left\{y\right\}$. This function represents the concept of a transition between two phenotypes. As such, a definition of continuity has been given for the transition function, and this determines whether one phenotype is accessible from another:
A transition from x to y is continuous if and only if:

1. $x \,\epsilon\, \mathrm{B}(y)$ and $y \,\epsilon\, \mathrm{B}(x)$

2. $x \in \mathrm{B}(y)$, $y \notin \mathrm{B}(x)$ and $\theta(t_1) = y$

3. $x \notin \mathrm{B}(y)$, $y \in \mathrm{B}(x)$ and $\theta(t_1) = x$

For a directed transition, the definition is much simpler. A directed transition from x to y is continuous if and only if there is a directed edge from x to y in the directed graph Y representing the topology. These two functions combined are quite important and integral for a mathematical understanding of RNA evolution. They provide us with a way to precisely define accessibility rather than relying on vague intuition. In this model, a continuous transition from x to y means that y $\leftarrow$ x. The converse, that a discontinuous transition means that y is not accessible from x, is usually but not always true. Occasionally, there are discontinuous evolutionary trajectories, but these happen infrequently. Thus, these two functions provide us at least with a way of definitively proving whether one structure could have evolved from another. This is why topology is so important for the study of evolution; it has the vocabulary to precisely define what "close" or "accessible" means and the tools to determine whether two objects are "close."

5 Acknowledgements

It is my pleasure to thank my mentor, Drew Moore, for providing me with invaluable assistance in understanding algebraic topology and for supporting me in a somewhat unusual choice of paper topic.

References

[1] James R. Munkres. *Topology.* 2000.

[2] Jan Cupal, Stephan Kopp, Peter F. Stadler. *RNA Shape Space Topology.*