RNA Shape Space Topology

Jan Cupal[a], Stephan Kopp[a,b], Peter F. Stadler[a,c,*]

[a]Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 4277 52737 Fax: **43 1 4277 52793 E-Mail: studla@tbi.univie.ac.at
URL: http://www.tbi.univie.ac.at/~studla
[*]Address for correspondence

[b]Los Alamos National Laboratory, TSA/DO-SA
Los Alamos, NM 87545-1663, USA
Mailstop: TA-0, SM-1237, MS M997
Phone: ** (505) 665-0911 Fax: ** (505) 665-7464 Email: skopp@lanl.gov

[c]The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA
Phone: (505) 984 8800 Fax: (505) 982 0565 E-Mail: stadler@santafe.edu

Abstract

The distinction between continuous and discontinuous transitions is a long-standing problem in the theory of evolution. Continuity being a topological property, we present a formalism that treats the space of phenotypes as a (finite) topological space, with a topology that is derived from the probabilities with which of one phenotype is accessible from another through changes at the genotypic level. The shape space of RNA secondary structures is used to illustrate this approach. We show that evolutionary trajectories are continuous if and only if they follow connected paths in phenotype space.

Keywords: Genotype-Phenotype Map — RNA Folding — Finite Topological Space — Continuity in Evolution.

AMS Subject Classification: 54G99, 92B05

LA-UR 99:1324

1. Introduction

While variation is introduced by mutation or recombination at the genotypic level, it is the phenotype that is subject to selection. A thorough understanding of the relationships between genotypes and phenotypes is therefore a necessary prerequisite for a complete theory of both biological and artificial evolution.

In the simplest case — evolving RNA molecules — genotype and phenotype are two aspects of the same molecule. The specific sequence of nucleotides is the genotype, the three-dimensional shape of the molecule its phenotype. Conventional biophysics considers sequence-structure relations of biopolymers primarily with respect to the folding problem: given is a sequence; which structure does it form under the specified experimental conditions? Many problems in current molecular biology and biotechnology [33], however, raise questions that cannot be answered satisfactorily by this approach. Instead, one has to consider global properties of the *folding function* that maps the set of all possible sequences onto the set of all possible structures.

Such problems are, for example, the sensitivity of structures against mutations, or the inverse folding problem: given a structure u, which sequences do fold into this structure under the specified conditions? Naturally, one may ask how the *neutral set* $S(u)$ consisting of all sequences folding into the structure u is embedded in sequence space, and how the neutral sets $S(u)$ and $S(v)$ of two different structures u and v are located relative to each other. A long series of publications, starting with [11], answers these questions in great detail from RNA secondary structures, see e.g., [12, 23, 35, 20, 21]. More recently, similar surveys used knowledge-based potentials [3, 2] or lattice models [9, 5] for exploring protein space.

The properties of the folding function are of course intimately connected to the structure of both sequence space and shape space. The genetic operators acting on the genotypes imply the structure of sequence space in very natural way. Assuming point mutations, for instance, sequence space becomes a generalized hypercube, i.e., a graph in which two genotypes are neighbors if they differ by a single mutation. Recombination spaces are discussed in [7, 18, 39, 38].

The structure of shape space is much less obvious. In many cases a distance measure among the shapes based on geometrical similarity will be sufficient [12]. In order to understand the sequence of phenotypic changes along an evolutionary trajectory, however, it is necessary to know which phenotypes are accessible from which genotypes. Accessibility can then be used to define a relation of "nearness" among phenotypes, independently of their geometric or biophysical similarities [14, 15]. In these contributions, a notion of "continuity" is introduced and the evolutionary transitions are classified as continuous or discontinuous based on how easily one shape can be accessed from a previous one.

Building upon these ideas, it is the purpose of this work to develop a rigorous mathematical framework in which shape space is *explicitly* constructed as a (finite) topological space assuming that the "natural" topology of shape space is determined by the accessibilities shapes along evolutionary trajectories.

This contribution is organized as follows: After a brief description of the RNA secondary structure model, we discuss the construction of accessibility relations. The subsequent three sections introduce the main elements of topologies on finite sets and explore some facets of the topologies associated with undirected graphs and partially ordered sets. Most of the mathematical technicalities underlying section 4 can be found in the appendix. The topologies of a few RNA shape spaces are discussed in some detail in section 7. Section 8 gives a precise definition of evolutionary trajectories. The main result of this section states that a trajectory is continuous if and only if it follows a connected path in a certain graph derived from the accessibility relations.

2. The Sequence-Structure Map of RNA

RNA secondary structures provide a discrete, coarse grained concept of structure similar in complexity to lattice models of proteins. In contrast to the latter, however, RNA secondary structures are a faithful coarse graining of the 3D structures. Secondary structures are routinely used to display, organize, and interpret experimental findings, they are oftentimes conserved over evolutionary times scales, and *in vitro* selection experiments with RNA more often than not yield families of selected sequences that share distinctive secondary structure features.

From the mathematical point of view, a secondary structure is a list of base pairs $[i, j]$ with $i < j$ such that for any two base pairs $[i, j]$ and $[k, l]$ with $i \leq k$ holds: (i) $i = k$ if and only if $j = l$, and (ii) $k < j$ implies $i < k < l < j$ or $k < l < i < j$. The first condition simply means that each nucleotide can take part in at most one base pair. The second condition forbids knots and pseudo-knots. Secondary structures form a special type of *outer-planar* graphs, i.e., they can be drawn in the plane in such a way that all vertices (which represent the nucleotides) are arranged on a circle, and all edges (which represent the bases pairs) lie inside the circle and do not intersect. Secondary structures are conveniently represented as strings with the alphabet $\{$`(`, `.`, `)`$\}$, where unpaired bases are denoted by a dot and each base pair corresponds to a pair of matching parentheses, see Figure 1. As a consequence of their simple graph-theoretical form, RNA secondary structures can be predicted from the sequence information by means of a dynamic programming algorithm [45, 23] that makes use of a set of experimental determined energy parameters [16, 25, 42].

For a given chain length n, there are α^n different RNA sequences, where the alphabet size is $\alpha = 2$ for **GC** sequences and $\alpha = 4$ for natural RNA sequences. An exact enumeration of all possible secondary structure graphs, however, yields less than 1.87^n different secondary structures. In an exhaustive computational study all **GC** sequences with a chain length up to $n = 30$ were folded [20, 21]. These data indicate that only about 1.65^n secondary structure graphs are actually realized as the minimum energy structure of some RNA sequence.

Typically, a secondary structure u is therefore obtained from many different sequences. The set $S(u)$ of these sequences is called the *neutral set* of u. In other words, $S(u)$ is the pre-image of u w.r.t. the *folding map* f. The size-distribution of neutral sets is far from uniform [35]. In fact, there are a relatively small number of *common structures* with large neutral sets, while the overwhelming number of structures is formed by only a very small number of different sequences. Let $\mathcal{S}_n$ denote the number of distinct secondary structure graphs that are formed by RNA sequences of length n. Then we say that a structure u is common if

$$|S(u)| \geq \alpha^n / \mathcal{S}_n \,, \tag{1}$$

i.e., if $S(u)$ is larger than the average.

Data from both large samples of long sequences ($n \gg 30$) [35, 33] and from exhaustive folding of all short sequences [20, 21] support the observation that the fraction of sequences folding into common structures increases with chain length and approaches 100% in the limit of long chains. As argued in [33], only common structures play a role in natural evolution and in evolutionary biotechnology.

The most important feature of the neutral sets of common structures is that they form connected *neutral networks* in sequence space. Furthermore, the neutral networks of any two common structure come close to each other in some part of sequence space. These facts were established both in large scale computer simulations [35, 36] and by means of random graph model [31, 30]. They seem to be generic feature of the sequence structure map of both nucleic acids and polypeptides.

3. Accessibility Relations for RNA Secondary Structures

The peculiar features of the sequence-structure map cause an evolving population to diffuse on the neutral networks of a particular secondary structure u until a superior phenotype u' is encountered. At this point the population jumps from $S(u)$ to the neutral network to $S(u')$, where the diffusion process starts again [13, 24, 14, 28].

The question therefore becomes which phenotypes (secondary structure) are easily accessible from a given neutral networks, and which ones are hard or impossible to get to in a single step. In other words, we have to define a criterion for deciding whether a structure y is accessible from the neutral networks of structure x, in which case we write $y \leftarrow x$.

In the simplest case, we may say that y is accessible from x, if it is possible to jump from $S(x)$ to $S(y)$ by means of a single point mutation, i.e., if there are two sequences s and s' such that (i) s and s' differ by a single mutation, and (ii) $f(s) = x$ and $f(s') = y$. A random graph theory developed in [31, 30] predicts that any common structures should be accessible from each other. We shall see in section 7 that this is at least approximately the case.

Since sequence space is so large that not all possible sequence are ever realized in the course of simulation run (or during the history of evolution), Fontana & Schuster [14] argue that one should consider more restrictive condition for accessibility. Following their argument, we will say that y is accessible from x, only if the probability is high enough that a population located on the neutral network $S(x)$ will find a sequence folding into y. Let us write $\partial S(x)$ for the set of all sequences that are obtained as point mutations of elements from the neutral set $S(x)$ and that are not themselves members of $S(x)$. The set $\partial S(x)$ is known as the *boundary* of $S(x)$ in sequence space. We may now consider the distribution of structures that are formed by the sequences in $\partial S(x)$. For any two structures $x, y \in V$ define

$$D(y \leftarrow x) = S(y) \cap \partial S(x) \tag{2}$$

as the set of all neighbors of x that fold into y. Accessibility may now be based on the size of $D(y \leftarrow x)$ relative to $S(x)$, $S(y)$, or their boundaries. Probably the most natural quantity is the ratio probability $|D(y \leftarrow x)|/|\partial S(x)|$ that a non-neutral mutant of an x-sequence folds into the shape y.

In [14] two related measures are used: The *neighborhood frequency* $\nu(y, x)$ is the fraction of sequences folding into x that have at least one neighbor folding into y. Conversely, the *frequency of occurrence* $\vartheta(y, x)$ is the fraction of mutants of

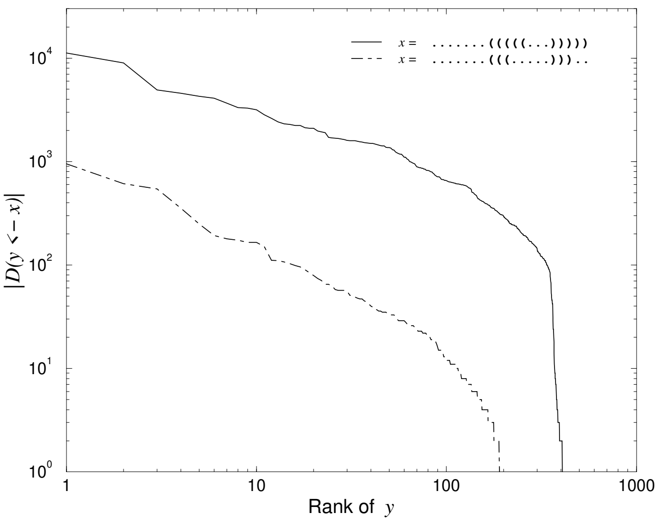

Figure 1. Size distribution of the sets $D(y \leftarrow x)$ for the most frequent and the least frequent common structures of the $\mathbf{GC}_{20}$ space. There are 1610 different secondary structures in this case.

x-sequences that fold into y. In symbols:

$$\begin{aligned}\nu(y,x) &= \frac{|S(x) \cap \partial S(y)|}{|S(x)|} = \frac{|D(x \leftarrow y)|}{|S(x)|} \\ \vartheta(y,x) &= \frac{|S(y) \cap \partial S(x)|}{(\alpha-1)n\,|S(x)|} = \frac{|D(y \leftarrow x)|}{(\alpha-1)n\,|S(x)|}\end{aligned} \tag{3}$$

The denominator in the expression for $\vartheta(y,x)$ is the total number of mutants of sequences in $S(x)$ including those that fold into x. It was introduced in [14] mainly because $|\partial S(x)|$ is unknown for long sequences.

A simple possibility for defining accessibility is to set a threshold value $\varepsilon > 0$. Hence $y \leftarrow x$ iff $|D(y \leftarrow x)| \geq \varepsilon |\partial S(x)|$. A related class of accessibility relations arises by ranking the neighbors of x with respect to $|D(y \leftarrow x)|$ and accepting only a fixed number highest-ranking structures as accessible. A threshold could also be introduced individually for each x depending on the form of the size distribution of the sets $|D(y \leftarrow x)|$ or $|D(x \leftarrow y)|$, see Figure 1.

For technical convenience, and because it captures the possibility correct replication, we require that the *accessibility relation* $\leftarrow$ is reflexive, i.e., $x \leftarrow x$ for all $x \in V$. Accessibility relations are not necessarily symmetric, that is, $y \leftarrow x$ in general does not imply $x \leftarrow y$. For instance, if $S(y)$ is much smaller than $S(x)$, than y will not be accessible from x. The two extremal cases of symmetric and anti-symmetric accessibility relations will be discussed in some more detail in sections 5 and 6.

4. Accessibility Topologies

Let $N(x) = \{y \in V | y \leftarrow x\}$ be the set of the structures that are accessible from x (including x itself). Any particular definition of accessibility hence translates into a finite collection $\mathcal{N} = \{N(x)|x \in V\}$ subsets of the shape space V, such

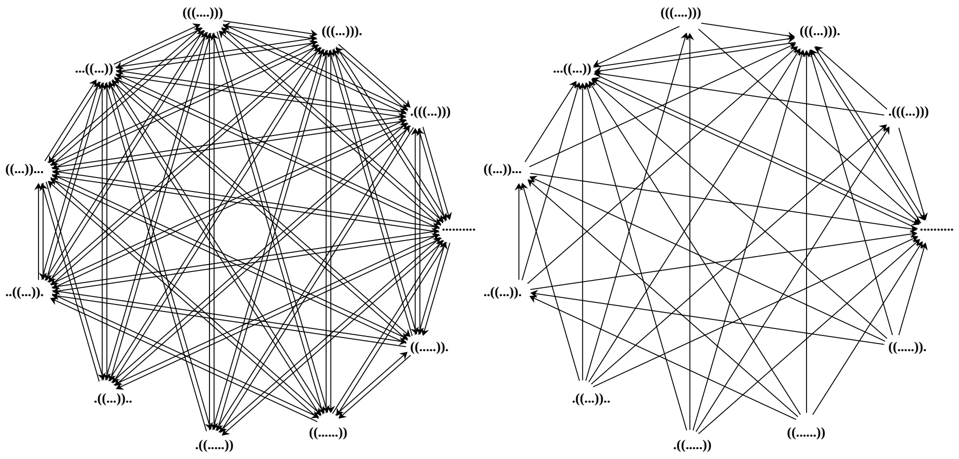

Figure 2. Accessibility relations for the smallest non-trivial RNA shape space $\mathbf{GC}_{10}$. L.h.s.: accessibility graph Γ; r.h.s.: non-redundant basis $\mathcal{B}$ of the accessibility topology represented by its directed graph $\vec{\Upsilon}$, see section 4 for details.

that $x \in N(x)$ for all $x \in V$. Note that $(V, \mathcal{N})$ is a hypergraph [4]. A more convenient representation for our purposes is the directed graph $\vec{\Gamma}$ with vertex set V and an edge $x \to y$ for x to $y \neq x$ if and only if $y \in N(x)$. Since $x \in N(x)$ by definition, there is one-to-one mapping between directed graphs with vertex set V and accessibility relations on V. As an example, the accessibility relation for the smallest non-trivial RNA space, $\mathbf{GC}_{10}$, is shown in Figure 2.

A *topology* on V is a collection τ of subsets of V, called the *open sets* in V with the following properties

(i) $\emptyset$ and V are open sets.
(ii) The union of an arbitrary number of open sets is open.
(iii) The intersection of a finite number of open sets is open.

A set $\mathcal{N} \subseteq \tau$ is a *subbasis* of τ if τ is generated by the unions and finite intersections of the sets contained in $\mathcal{N}$. A set $\mathcal{B} \subseteq \tau$ is a *basis* of τ if all open sets can be written as unions of elements of $\mathcal{B}$.

In appendix A we review the theory of finite topological spaces. We encounter an extreme simplifications which explains why finite topologies have received very little attention apart from serving as simple counterexamples [40]. The crucial property of finite topologies is the existence of a unique non-redundant basis $\mathcal{B} = \{B(x)|x \in V\}$, where $B(x)$ is the intersection of all open sets containing x, see lemma 1 in the appendix. Given an accessibility relation $\mathcal{N}$ (that is, a subbasis of a topology τ on V), we can explicitly construct the unique non-redundant basis of τ:

$$\mathcal{B} = \left\{ B(x) = \bigcap_{y:x\in N(y)} N(y) \,\middle|\, x \in V \right\} \tag{4}$$

A directed graph $\vec{\Upsilon}$ can be associated with the basis $\mathcal{B}$ in the following way: The vertex set of $\vec{\Upsilon}$ is V, and for any $x \neq y$ there is a directed edge $x \to y$ (from x to y) if and only if $y \in B(x)$. An example is given if Figure 2.

Directed graphs provide a data structure that is handled efficiently by publicly available software such as `LEDA`. It pays therefore to translate the most important topological concepts such as connectedness or various separation properties into the language of graph theory of $\vec{\Upsilon}$. The main works towards this end is described in appendix A, section A2 and proposition 1.

Theorem 1. *The finite topological space (V, τ) is connected if and only if its digraph $\vec{\Upsilon}$ is (weakly) connected. $W \subset V$ is component of V if and only if W induces a (weakly) connected component of $\vec{\Upsilon}$.*

Proof. Recall that a directed graph is (weakly) connected if the underlying undirected graph is connected.
The topological space (V, τ) is not connected if and only if there is a partition $V_1 \cup V_2 = V$ such that $V_1 \cap V_2 = \emptyset$ and both V_1 and V_2 are open, i.e., $V_i = \cup_{x \in V_i} B(x)$. Equivalently, for all $x \in V_1$ and all $y \in V_2$ hold $x \notin B(y)$ and $y \notin B(x)$. In other words, (V, τ) is the (topological) sum of V_1 and V_2 iff there are no directed edges in $\vec{\Upsilon}$ connected vertices from V_1 with vertices form V_2. The second claim follows immediately. □

Theorem 2. *Let (V, τ) be a finite topological space and let $\vec{\Upsilon}$ be the directed graph associated with its basis. Then the topological separation axioms are equivalent to the following properties of $\vec{\Upsilon}$:*

(T1) *$\vec{\Upsilon}$ contains no edges.*
(T0) *$\vec{\Upsilon}$ contains no bidirectional edge, i.e., there is no pair of vertices x and y such that $x \to y$ and $y \to x$.*
(T3) *$\vec{\Upsilon}$ is a union of disjoint cliques. Equivalently, $B(x) \cap B(y) = \emptyset$ or $B(x) = B(y)$.*
(T5) *$B(x) \cap B(y) = \emptyset$, $B(x)$, or $B(y)$, i.e., $\mathcal{B}$ is a hierarchy.*
(T4) *For each connected component V' of $\vec{\Upsilon}$ there is a point $x \in V'$ such that there is a directed edge from x to all other vertices in V', i.e., $B(x) = V'$*

Furthermore, τ is the discrete topology iff $\vec{\Upsilon}$ has no edges, while τ is the indiscrete topology iff $\vec{\Upsilon}$ is a complete graph.

Proof. The conditions for (T0), (T1), (T3), the discrete and the indiscrete topology are trivial rewritings of proposition 1 in the appendix. □

5. Topologies from Graphs

If the accessibility relation is *symmetric* then $\vec{\Gamma}$ turns into an undirected graph. We shall write Γ to emphasize this fact. A *clique* of Γ is a maximal induced complete subgraph. The clique-topology τ_Q on V is obtained by regarding the set of all cliques of Γ as a subbasis, i.e., by declaring the cliques of Γ as open sets. Note furthermore that a single vertex forms a clique if and only if it is isolated and that a pair x, y of vertices is a clique if and only if the edge (x, y) is not contained in a triangle. Each edge $(x, y) \in E$, the edge set of Γ, is contained in some clique. Trivially, if Γ is a complete graph, then $V = \mathcal{N}(x)$ and hence τ is indiscrete. Also, V is the only clique of a complete graph, thus τ_Q is indiscrete as well.

Theorem 3. *The adjacency topology τ and the clique topology τ_Q coincide for all undirected graphs.*

Proof. Consider the set Q_x of all cliques containing x. Since each edge containing x is contained in at least one of the elements of Q_x we we have $N(x) \subseteq Y_x :=$

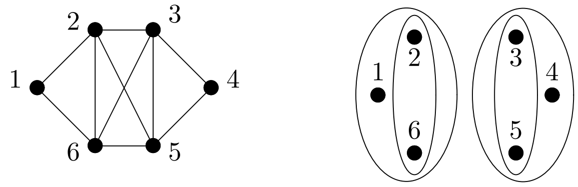

Figure 3. The topology associated with Γ is non-trivial. The cliques of Γ are the two triangles $(1,2,6)$ and $(3,4,5)$, and the K_4 formed by $(2,3,5,6)$. There are only two non-empty intersections of cliques, namely the edges $(2,6)$ and $(3,5)$, which are disjoint. The topology τ therefore consists of $\{2,6\}$, $\{3,5\}$, $\{1,2,6\}$, $\{3,4,5\}$, $\{2,3,5,6\}$, $\{1,2,3,5,6\}$, $\{2,3,4,5,6\}$, V, and $\emptyset$. The basis is $\mathcal{B} = \{\ \{1,2,6\}, \{2,4,5\}, \{2,6\}, \{3,5\}\ \}$. The topology of Γ satisfies the separation axiom $\mathsf{T5}$.

$\cup_{X\in Q_x} X$. On the other hand, a clique in Q_x contains only x itself and neighbors of x, thus $X \subseteq N(x)$ for all $X \in Q_x$. Thus $Y_x = N(x)$, i.e., all elements of the subbasis $\mathcal{N}$ of the adjacency topology τ are contained in the clique topology τ_Q. Now consider a clique X containing x and construct the set $Z = \cap_{y\in X} N(y)$. By construction, Z is an element of the adjacency topology. Furthermore we have $x \in N(y)$ for all $y \in X$ since X is subset of $N(x)$, i.e., $X \subseteq Z$. Suppose there is a vertex $z \in Z \setminus X$. Then z is by definition a neighbor of each $y \in X$, i.e., $X \cup \{z\}$ induces a complete subgraph of Γ, and hence X cannot be a clique of Γ. Thus $Z = X$, i.e., each clique of Γ is an element of the adjacency topology. ☐

The following theorem shows that a large class of graphs has as trivial adjacency topology. On the other hand, the example in Figure 3 illustrates that there are undirected graphs with a topology that is neither discrete nor indiscrete.

Theorem 4. *Suppose $\Gamma \neq K_2$ is triangle-free and does not contain pendant vertices. Then τ is discrete. If Γ contains a pendant vertex, then τ is not discrete.*

Proof. (i) Suppose $(x,y) \in E$ is not contained in a triangle and x is not a pendant vertex, i.e., x has vertex degree ≥ 2. Then $\{x,y\}$ is a clique and hence a member of τ. Since the degree of x is at least 2 there is a vertex z such that $x \in N(z)$. We have $y \notin N(z)$, otherwise x,y,z would form a triangle. Thus $N(z) \cap \{x,y\} = \{x\} \in \tau$.
(ii) Let x be a pendant vertex, and let y be the unique neighbor of x. By the same argument as above there is neighbor $z \in N(y)$ such that $x \notin N(z)$. Hence $N(x) \cap N(z) = \{y\}$. On the other hand $N(x) \cap N(q)$ is either empty or $\{y\}$ for all $q \neq x,y$. Clearly $N(x) \cap N(y) = N(x)$, hence $\{x\}$ cannot be generated as a (repeated) intersection of elements of the subbasis $\mathcal{N}$. Therefore $\{x\} \notin \tau$. ☐

An *equivalence relation* is reflexive, symmetric, and transitive. The graph Γ associated with such an accessibility relation is a collection of complete sub-graphs, hence the resulting topological space is $\mathsf{T3}$.

Very little can be said in general about the topologies of directed graphs. A vertex $x \in V$ is a *source* if there is no arrow ending in x, i.e., if $x \in N(y)$ implies $y = x$. It is clear that $N(x)$ is the smallest open set containing x in this case. Conversely, x is a *sink* if there is no arrow beginning in x, i.e. if and only if $N(x) = \{x\}$. (An isolated vertex is both a source and a sink.) It follows immediately that the adjacency topology of a connected directed graph $\vec{\Gamma}$ with at least 2 vertices is non-trivial if $\vec{\Gamma}$ contains a sink and non-discrete if $\vec{\Gamma}$ contains a source.

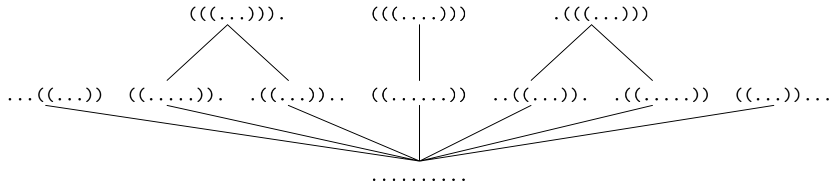

Figure 4. Hasse diagram of the substructure relation on $\mathbf{GC}_{10}$. A structure y is "smaller" than x, $y \prec x$, if y is drawn below x and there is a path from x to y that goes "downhill" in each step. The basis elements of the substructure topology consist of x and all its substructures.

6. The Substructure Topology

For each secondary structure (graph) x we denote by P_x the set of base pairs, i.e., the set of edges that do not belong to the backbone of the molecule. We say y is a substructure of x if x and y have the same chain length n and if $P_y \subseteq P_x$, and we write $y \preceq x$. Note that, for technical convenience, we regard x as a trivial substructure of itself.

Clearly, the relation $\preceq$ is a *partial order* on V, i.e., anti-symmetric, reflexive, and transitive relation on V [8]. It has a unique minimal element, namely the open structure. The maximal elements are those structures that have the maximal number of base pairs including any prescribed set of pairs. These structures are exactly the *suboptimal structures* that are computed by Zuker's suboptimal folding algorithm [44]. Figure 4 shows the so-called Hasse diagram of the substructure relation for the $\mathbf{GC}_{10}$ shape space.

The substructure relation is related to accessibility in the following way [14]. First we note that the set $S(x)$ cannot be located anywhere in sequence space. Only sequences that *can* form x, that is sequences that have one of the six possible combinations **GC**, **CG**, **AU**, **UA**, **GU**, or **UG** at the two sequence positions that form a base pair in x, are even candidates for membership in $S(x)$. These sequences form the set $C(x)$ of *compatible* sequence of x [31]. Clearly we have $C(x) \subseteq C(y)$ if and only if $y \preceq x$. The random graph theory of neutral networks [31] predicts that $S(x)$ is a dense random graph of $C(x)$ for all common structures x. Thus, for any two common structures x and y, the neutral sets $S(x)$ and $S(y)$ come close to each other almost everywhere in $C(x) \cap C(y)$. Consequently, most sequences in $S(x)$ should have mutants that fold into some substructure of x. On the other hand, if x is substructure of z, then $C(z)$ is smaller than $C(x)$, and hence it should be relatively hard in general to find a mutant of an x-sequence that actually folds into z. Based on this argument, the structures accessible from x are exactly its substructures.

Of course we have neglected all thermodynamic considerations showing that structures with very few base pairs are unstable and therefore not common. Furthermore, if x and y differ only by a single base pair, then their compatible sets have a similar size, so that a transition from x to y, $x \preceq y$, would not all that unlikely. Finally, it is unlikely that a single point mutation will disrupt many base pairs at once. We have to expect therefore that the substructure topology will not be particularly realistic. Its simplicity, on the other hand, makes it worth while to consider it in some more detail, before we turn to a more realistic approach.

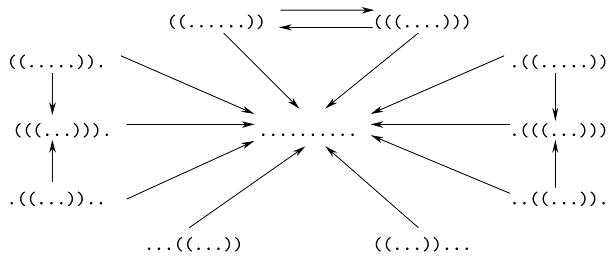

Figure 5. Basis of the Fontana-Schuster topology on $\mathbf{GC}_{10}$.
Note that $B(\texttt{(((...))).}) = N(\texttt{((.....)).}) \cap N(\texttt{.((...))..})$ is a proper subset of $N(\texttt{(((...))).})$.

Let $(V, \preceq)$ be a partially ordered set. An associated topology is obtained by declaring the sets

$$N(x) = \{y \in V \mid y \preceq x\} \tag{5}$$

open. Let $H(x, y)$ be the set of maxima of $N(x) \cap N(y)$. Then $N(x) \cap N(y) = \cup_{z \in H(x,y)} N(z)$. As a consequence $\{N(x) | x \in V\}$ is not only a subbasis but a basis of the *partial order topology.* It is non-redundant since $N(x)$ is obviously the smallest open set containing x. Since $N(x) \neq N(y)$ for all $x \neq y$, the partial order topology satisfies $\mathsf{T0}$.

If $(V, \preceq)$ has a unique minimal element z, then all open sets contain z, i.e., the topological space is hyper-connected and therefore connected. Since the open structure is the unique minimal element of the substructure relation, we conclude that *the substructure topology of RNA secondary structures forms a hyper-connected* $\mathsf{T0}$*-space.*

7. RNA Shape Space Topologies: Examples

Let us return to a more realistic model of accessibility among RNA shape spaces. Fontana and Schuster [14, 34] argue that a secondary structure y is accessible from x in one of the following three cases: (i) y is obtained from x by shortening a stack of x; (ii) y is obtained from x by elongating a stack of x, or (iii) y is obtained from x be eliminating a single stack of x. We shall refer to the resulting topology as the *Fontana-Schuster topology* of RNA shape space. As the example in Figure 5 shows the subbasis specified above is not necessarily a basis.

For relatively small RNA sequence space, up to a chain length $n \leq 30$ in the case of the 2-letter **GC**-alphabet, it is feasible to explicitly construct adjacency topologies by folding all sequences and explicit determination of their neutral networks. The data reported here are a re-evaluation of the original data published in [20, 21]. For each neutral set $S(x)$ we explicitly compute $\partial S(x)$ by producing all (single) point mutations of all sequences in $S(x)$. Boundary sequences folding into the same structure are collected in the sets $D(y \leftarrow x)$. In practice we only record $|D(y \leftarrow x)|$ instead of the list of sequences themselves. Given a particular definition of accessibility, for instance in terms of a threshold on $|D(y \leftarrow x)|$, see Figure 6, it is straightforward to extract $N(x)$ for each structure. Finally we use equ.(4) to construct the non-redundant basis of the accessibility topology.

It does not come as a surprise that the topology depends strongly on the underlying definition of accessibility. As an illustrative example we define $x \leftarrow y$

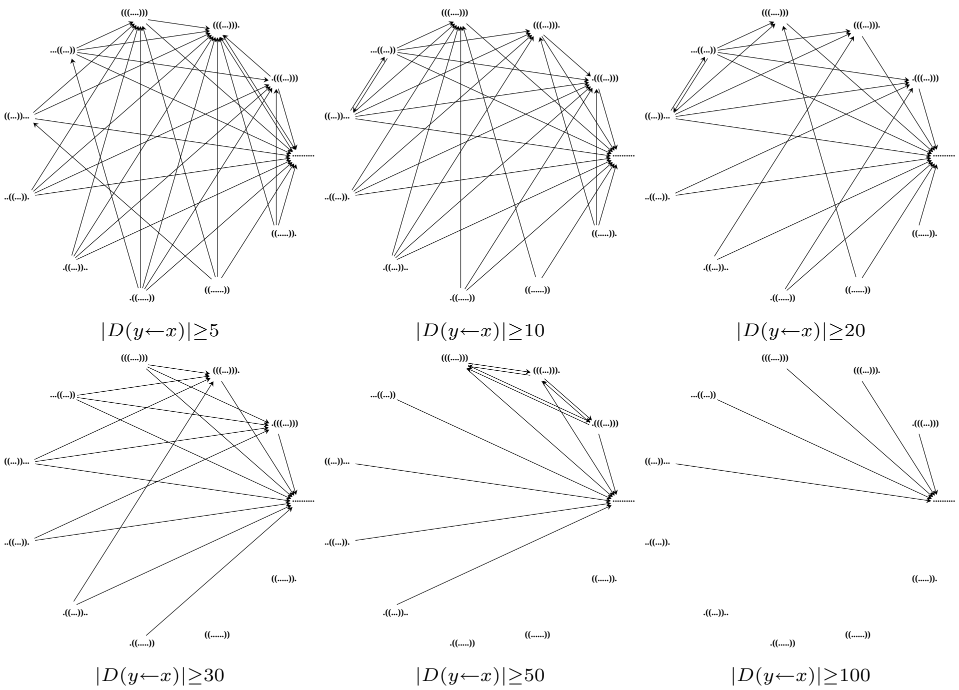

Figure 6. Accessibility topologies for $\mathbf{GC}_{10}$ as a function of a threshold for the number $|D(y \leftarrow x)|$ of mutations leading from x to y.

if $|S(y) \cap \partial S(x)| = |D(y \leftarrow x)|$ is larger than a threshold value. Clearly, if the threshold value is large enough (namely larger than $\max|S(x)|$), we obtain the discrete topology. On the other hand, as shown in Figure 2, if we only require $D(y \leftarrow x) \neq \emptyset$, we find a highly connected topological space. The transition between these two extremes is exemplified in Figure 6. The rarest structure become isolated first because both their neutral sets and their sets of neighbors are small.

Recall that a topology τ' is *finer* than τ if each open set in τ is also open in τ'. Clearly, the subbasis sets $N(x)$ shrink with increasing threshold values. One might expect that this is also true for the basis sets $B(x)$, and indeed this is often the case. A close inspection of the example in Figure 6, however, shows that an additional bidirectional edge between `((...))...` and `...((...))` arises when the threshold is raised from 5 to 10. Hence more stringent conditions on the subbasis of accessible structures does not always lead to a finer topology.

It should be noted that none of the topologies shown in Figure 6 is a close match to the Fontana-Schuster topology on the same space. This might be an artifact of the short chain length, however. Numerical experiments [14, 15, 34] indicate that accessibility topologies with a suitable threshold for accessibility and the structure-based Fontana-Schuster topology become quite similar for longer sequences.

The most important question, of course, is whether the topologies of larger sequence spaces is similar to the small examples discussed so far. For moderate size chains, $n \leq 20$ exhaustive computations are feasible. The results for the "full" accessibility topologies (no threshold values) are summarized in table 1.

Table 1. RNA Accessibility Topologies

n	$\|\mathcal{S}\|$	Size of Components									
10	11	11									
11	20	20									
12	31	31									T0
13	48	42	5	1							T0
14	73	71	2								T0
15	116	116									T0
16	195	134	1	1	1	1	1	1	1	1	...	T0
17	340	202	1	1	1	1	1	1	1	1	...	T0
18	582	255	2	1	1	1	1	1	1	1	...	T0
19	973	298	2	1	1	1	1	1	1	1	...	T0
20	1610	318	5	2	2	2	2	2	1	1	...	T0

Apparently, RNA accessibility topologies are T0-spaces in general. On the other hand, the spaces are not connected for $n \geq 16$. They decompose into a single *giant component* and a substantial number of very small components and/or isolated points. The isolated points and small components are composed predominately of rare structures.

Instead of considering all structures one might want to restrict the shape space to the *common structures*. It is straight forward to construct an accessibility topology on the set of common structures. The results are qualitatively similar to the accessibility topology on the full shape space. In general we find a giant component and a substantial number of isolated points. While there are a few cliques that contradict the T0 axiom for $n \leq 20$, we find for $n = 21$ a T0-space with 460 vertices, 107 of which form the giant component. All other structures are isolated or belong to one of 8 pairs.

A comparison of RNA accessibility topologies (Table 1) and the Fontana-Schuster topologies (Table 2) reveals the expected qualitative similarities. Quantitatively, however, there are differences. In particular, the giant components are substantially larger in the FS topologies.

8. Trajectories and Transitions

An "evolutionary trajectory" can be regarded as a time-series of V-elements, for instance the time series of the dominating shapes in a population. The following definition makes this notion precise:

Definition 1. *Let V be a finite topological space and let $[\alpha, \beta] \subset \mathbb{R}$ be a finite interval, $\alpha = t_0 < t_1 < \cdots < t_m < t_{m+1} = \beta$, $m \geq 0$. A* trajectory *is a function $\varphi : [\alpha, \beta] \to V$ with the following properties:*

(i) φ is constant on the intervals $I_0 = [\alpha, t_1)$, $I_m = (t_m, \beta]$, and $I_k = (t_k, t_{k+1})$ for $1 \leq k < m$. In the special case $m = 0$ we require that φ is constant on $[\alpha, \beta]$.

(ii) $\varphi(I_k) \neq \varphi(I_{k-1})$ for $1 \leq k \leq m$.

(iii) $\varphi(t_k)$ equals either $\varphi(I_k)$ or $\varphi(I_{k-1})$ for $1 \leq k \leq m$.

We say that φ is a proper *trajectory if $\varphi(t_k) = \varphi(I_{k-1})$ for all k. A trajectory* follows *a sequence $\{x_i \in V, 0 \leq i \leq m\}$ if $\varphi(I_k) = x_k$ for $0 \leq k \leq m$.*

Table 2. Fontana Schuster Topologies

n	$\|\mathcal{S}\|$	Size of Components										
10	11	11										T0
11	20	20										T0
12	31	31										T0
13	48	47	1									T0
14	73	70	3									T0
15	116	111	1	1	1	1	1					T0
16	195	175	1	1	1	1	1	1	1	1	...	T0
17	340	268	1	1	1	1	1	1	1	1	...	T0
18	582	455	1	1	1	1	1	1	1	1	...	T0
19	973	703	1	1	1	1	1	1	1	1	...	T0
20	1610	1076	1	1	1	1	1	1	1	1	...	T0
21	2613	1634	1	1	1	1	1	1	1	1	...	T0
22	4258	2357	1	1	1	1	1	1	1	1	...	T0
23	6936	3814	5	4	4	3	3	3	3	3	...	T0
24	11348	5890	6	5	5	4	4	4	3	3	...	T0
25	18590	9043	15	12	10	10	9	9	5	5	...	T0
26	30501	13895	18	16	13	12	9	6	6	5	...	T0
27	49949	21325	26	21	18	14	12	12	9	9	...	T0

The condition that φ is constant for almost all times reflects the fact that a finite time is required for any evolutionary change. It is clear that the "concatenation" of two trajectories $\varphi : [\alpha, \beta] \to V$ and $\psi : [\beta, \gamma] \to V$ with $\varphi(\beta) = \psi(\beta)$ is a trajectory $\psi * \varphi : [\alpha, \gamma] \to V$. Note that a continuous trajectory is always a path in the topological sense. Clearly, $\psi * \varphi$ is continuous if and only if both φ and ψ are continuous. We may decompose a trajectory therefore into parts that have only a single intermediate point, which we call *transitions*:

Definition 2. *Let V be a finite topological space, $x \neq y \in V$, $t_0 < t_2 \in \mathbb{R}$. A transition from x to y is a function $\theta : [t_0, t_2] \to \{x, y\}$ (with the induced topology) such that $\theta([t_0, t_1)) = \{x\}$ and $\theta((t_1, t_2]) = \{y\}$ for some $t_1 \in (t_0, t_2)$. We say that θ is a proper transition if $\theta(t_1) = x$.*

Theorem 5. *A transition from x to y is continuous if and only if either*

(i) $x \in B(y)$ and $y \in B(x)$, or
(ii) $x \in B(y)$, $y \notin B(x)$ and $\theta(t_1) = y$, or
(iii) $x \notin B(y)$, $y \in B(x)$ and $\theta(t_1) = x$.

A directed transition from x to y is continuous if and only if $y \in B(x)$, that is, if there is a directed edge from x to y in $\vec{\Upsilon}$.

Proof. The induced topology on the two-element subset $\{x, y\} \subseteq V$ consists of $\emptyset$, $\{x, y\}$, $B'(x) = B(x) \cap \{x, y\}$ and $B'(y) = B(y) \cap \{x, y\}$. There are only four possible cases:

(i) $B'(x) = B'(y) = \{x, y\}$, i.e., the induced topology on $\{x, y\}$ is indiscrete and hence any function into $\{x, y\}$ is continuous.

(ii) $B'(y) = \{x, y\}$ and $B'(x) = \{x\}$. Thus $\{y\}$ is a closed set. Hence the transition is continuous if $\theta^{-1}(\{y\}) = [t_1, t_2]$ is closed and $\theta^{-1}(\{x\}) = [t_0, t_1)$ is open in $[t_0, t_2]$.

(iii) $B'(x) = \{x, y\}$ and $B'(y) = \{y\}$ is analogous.

(iv) $B'(x) = \{x\}$ and $B'(y) = \{y\}$, i.e., the topology on $\{x, y\}$ is discrete and hence there is no non-constant continuous function into $\{x, y\}$.

The claim for proper transitions now follows immediately. □

Corollary 1. *Let V be finite topological space. Then there exists a continuous transition from x to y if and only if $B(x) \cap B(y) \neq \emptyset$.*

Corollary 2. *If φ is a continuous trajectory then it follows a connected path (in the graph-theoretical sense) within the undirected graph Υ.*

Corollary 3. *A proper trajectory is continuous if and only if it follows a directed path (in the graph-theoretical sense) in $\vec{\Upsilon}$.*

Corollary 3 links the abstract discussion of the previous sections with the numerical work reported in [14, 15], where a transition from x to y is *called* "continuous" if and only if y is accessible from x. If the notion of a *continuous function* could be extended to a generalization of topological spaces in which the intersection of two open sets is not necessarily open any more, then corollary 3 would turn Fontana's characterization of "continuous transitions" from a definition into a theorem. Generalized convergence spaces [6, 26], limit space [10], or pre-topologies [22] might be suitable starting points for such an extended theory. The simulation results in [14, 15, 34] indicate that, given a suitable topology on the space of phenotypes, evolutionary trajectories are continuous most of the time, interrupted by rare discontinuous transitions.

Rare structures appear mostly as isolated points or within very small components both in the accessibility topologies and in the Fontana-Schuster topology. Hence there are no continuous trajectories leading to these structures and hence they are inaccessibile by evolution. We submit that it is therefore justified to neglect these structures in the context of RNA evolution.

9. Discussion

"The variational properties of the phenotype are fundamental to its evolution by natural selection. Whether or not adaptive changes *can* be produced depends critically on the genotype-phenotype map. Despite its ubiquity in evolutionary phenomena, the genotype-phenotype map has until recently not been seen as a unifying conceptual framework for many different areas of evolutionary biology, including dissociability in development, morphological integration, developmental constraints, biological versatility, fluctuating asymmetry, the Baldwin effect, epistasis, canalization, heterochrony, genetic variance/covariance matrices, identification of quantitative trait loci, and the adaptive landscape" [41].

The genetic representation of a character determines the variability of the phenotype, and hence the *accessibility* of a particular phenotypic variant. In this context the concept of developmental constraints (*sensu* Maynard-Smith *et. al.* [27] and Schwenk [37]) can be understood as the limits of variability of traits caused by their representation or coding in the genome. The study of natural variation can at least give hints on the pattern of variability as for instance the study of osteological variation suggests the existence of constraints [1, 32]. Using computational models of genotype-phenotype maps, on the other hand, such constraints can be studied directly, and their impact on evolutionary adaptation both in terms of phenotypes and dynamics can be evaluated explicitly [14, 15].

In this contribution we have developed a rigorous mathematical framework that allows at least some discussion of evolution at the phenotypic level. Following the approach of [14, 15] we regard the set of phenotypes into a topological space. The topology is defined in terms of an accessibility relation among the phenotypes which in turn is derived from the genotype-phenotype map. The language of topology allows us, for instance, to make the notion of continuous and discontinuous transitions precise without resorting to our intuition about which phenotypic changes might seem "continuous" to us.

We have illustrated our exposition using examples from RNA folding. The genotype-phenotype map of RNA — the phenotypes being represented by secondary structures — is at this point the by far best understood computational model [35]. It is reassuring that within this model phenotypic accessibility, as defined in section 3 based on the "closeness" of sequences that fold into different phenotypes, is related to structural similarities.

In general we do not have access to entire sequence-structure map, and hence we cannot construct the accessibility topology of phenotype space. In a computational model, we will have to resort to sampling mutants. In a forthcoming study we shall explore to what extent a shape space topology can be reconstructed by sampling. Alternatively, phenotypic similarities could be used to introduce a meaningful topological structure that (hopefully) resembles accessibility. The Fontana-Schuster topology on RNA shape spaces, which was designed to emulate the accessibility relationships around the tRNA phenotype [14] may serve as an example for this approach. At least for short sequences, however, we find substantial differences: structural similarities are no guarantee for accessibility. Transitions probabilities between the neutral networks of different structures have been described in much detail in J. Weber's PhD thesis [43]. In particular, there is a relationship between the probabilities of transitions between the secondary structures x and y, the size of their neutral sets $S(x)$ and $S(y)$, and a certain structure distance measure which is based on a permutation representation of the secondary structure graph [29]. It is conceivable that it can be used to derive a shape space topology that is closer to a suitable accessibility topology.

Acknowledgments. Stimulating discussions with Walter Fontana, Christian Reidys, Peter Schuster, and Andreas Wagner are gratefully acknowledged. This work supported in part by Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* Proj. No. P13093-GEN. This paper is LA-UR 99:1324.

Appendix A: Finite Topological Spaces

A.1. Uniqueness of the Basis. Many of the standard notions of topology, and in particular their inter-dependencies degenerate when X is a finite space. The reason is the the existence of a unique smallest neighborhood $B(x)$ for each point $x \in V$, which is defined as intersection of all open sets containing x.

Lemma 1. *The family $\mathcal{B} = \{B(x)|x \in V\}$, where $B(x)$ is the intersection of all open sets containing x, forms the unique non-redundant basis of the topology τ.*

Proof. Since V is finite, $B(x)$ is an open set. It is clear from the definition that for each $z \in B(x) \cap B(y)$ hold $B(z) \subseteq B(x) \cap B(y)$, i.e., the intersection of two elements of $\mathcal{B}$ contains another basis element, and the union of elements of $\mathcal{B}$ is open. Thus $\mathcal{B}$ is a basis of τ. Uniqueness can be verified as follows: suppose there is another basis $\mathcal{B}'$. Then, by definition $B(x)$ is the union of elements of $\mathcal{B}'$. Hence either $B(x) \in \mathcal{B}'$, or $B(x)$ is a union of open sets each of which is strictly smaller than $B(x)$. This is impossible since $B(x)$ is the smallest open set containing x. Thus $B(x) \in \mathcal{B}'$ for all x and hence $\mathcal{B}' = \mathcal{B}$. □

In the following we briefly review how the wealth of topological properties is reduced in finite spaces. For definitions not repeated here we refer the reader to *Counterexamples in Topology* [40]. The main purpose of this appendix is to translate the most important properties of finite topological spaces into conditions on the collection $\mathcal{B}$ of basis sets.

We begin by observing that all finite topological spaces are *compact*, hence all the generalized notions of compactness become trivial as well.

A.2. Connectedness. Two open sets X_1 and X_2 form a *separation* of a topological space (V, τ) if they are disjoint and their union is V. In this case X_1 and X_2 are closed as well. A topological space is *connected* if it has no separation. It is obvious that $B(x)$ is connected for all x. Thus all finite topological spaces are *locally connected*, that is, there is connected neighborhood basis for each $x \in V$, namely $\{B(x)\}$.

A *path* from x to y is a continuous map $f : [0,1] \to V$ such that $f(0) = x$ and $f(1) = y$. If in addition f is invertible it is called an *arc*. Trivially, there are no arcs in finite topological spaces. A set $A \in V$ is *path connected* if for any two distinct points $x, y \in A$ there is path $f : [0,1] \to A$ from x to y.

Lemma 2. *$B(x)$ is path connected for all $x \in V$.*

Proof. Let $y \in B(x)$. We have to distinguish two cases. (i) $x \in B(y)$, i.e., $B(x) = B(y)$. Then there is no open set that contains x but not y or *vice versa* and hence no closed set containing one but not the other. Thus the induced topology on $\{x, y\}$ is indiscrete, and hence any function $f[0,1] \to \{x, y\}$ is continuous. (ii) $x \notin B(y)$. Then $\{y\}$ is open but not closed, and $\{x\}$ is closed but not open in $\{x, y\}$. Thus any function of the form $f(t) = x$ for $t \in [0, p]$ and $f(t) = y$ for $t \in (p, 1)$ is continuous, and of course a path. The concatenation of two paths is again a path, and thus $B(x)$ is path connected. □

In other words a finite topological space is always locally path connected, since for $\{B(x)\}$ is path connected neighborhood basis for each $x \in V$. A general theorem states that a topological space is path connected if it is connected and locally path connected. Thus we have the

Corollary 4. *A finite topological space is path connected if and only if it is connected.*

A space is *strongly* connected, iff all real valued continuous functions $V \to \mathbb{R}$ are constant. It is clear that on a connected finite space we cannot have a non-constant continuous function into $\mathbb{R}$. Thus connectedness and strong connectedness are the same thing in finite topological spaces.

A topological space is called *ultra-connected* if it contains no disjoint closed sets.

Lemma 3. *A finite topological space (V, τ) is ultra-connected if and only if there is an $x \in V$ such that $B(x) = V$.*

Proof. If $B(x) = V$ then the non-empty complements of elements in $\mathcal{B}$ all contain x. Consequently x is contained in all their intersections, i.e., in all closed sets. Thus there are non disjoint closed sets and V is ultra-connected. Conversely, we have to show that if any two closed sets intersect, then there is point x that is common to all of them: This is trivial if there are only two closed sets. Suppose the assertion is true for k sets. Let $U_k = C_1 \cap C_2 \cap \cdots \cap C_k$, and suppose C_{k+1} intersects C_1 through C_k, but $U_k \cap C_{k+1} = \emptyset$. Then $C_1 \cap C_k$ is a closed set that is disjoint from the closed set U_k, contradicting ultra-connectedness. □

It is *hyper-connected* if it contains no disjoint open sets. Of course ultra-connected or hyper-connected spaces are connected.

A.3. Separation Axioms. The hierarchy of *separation axioms* simplifies considerably as we shall see below. It will therefore be sufficient to consider the following separation axioms:

(T0) For all $x, y \in V$ there is an open set U such that $x \in U$ and $y \notin U$ or $y \in U$ and $x \notin U$.
(T1) For all $x, y \in V$ there is are open sets U and W such that $x \in U$, $y \notin U$, $y \in W$, and $x \notin W$.
(T3) For each $x \in V$ and each closed set A not containing x there are disjoint open neighborhoods U of x and O_A.
(T4) For any two disjoint closed sets A and B there are disjoint open neighborhoods O_A and O_B.
(T5) For any two separated sets A and B there are disjoint open neighborhoods O_A and O_B.

Other separation and disconnectedness properties degenerate in finite space. In addition, a number of implications between these properties hold in finite spaces that are not true in more general cases.

It is well known that a finite topological space is discrete if and only if it satisfies (T1). As a consequence each of the following properties characterizes the discrete topology if V is finite:

- the separation properties (T2) or (T$2_{1/2}$), i.e., Hausdorff and completely Hausdorff;
- the existence of an Urysohn function for any two points;
- the higher separation properties regular (T0 and T3), Tychonoff (T0 and T$3_{1/2}$), normal (T1 and T4), and completely normal (T1 and T5);
- the properties extremally disconnected, totally separated, and totally disconnected.
- the fact that V is a metric (or metrizable) topological space.

A space is *zero-dimensional* if it has a basis consisting of sets that are both open and closed. In finite zero-dimensional spaces each set is both open and closed, i.e., τ is a partition topology. A zero-dimensional space always satisfies the separation axiom (T3). Equivalently, finite zero-dimensional spaces are exactly the finite pseudo-metric spaces. The following lemma shows below that (T3) and zero-dimensional are equivalent properties in finite spaces.

Lemma 4. *A finite* T3*-space is zero-dimensional, i.e., it has a partition topology.*

Proof. The ($\mathsf{T3}$) axiom is equivalent to the following requirement: Each open set contains a closed neighborhood around each of its points. Thus $B(x)$ contains a closed neighborhood $C(x)$ of x. Hence there is an open set U such that $x \in U \subseteq C(x) \subseteq B(x)$. Since U is open and $x \in U$ we have $B(x) \in U$ and hence $B(x) = C(x)$, i.e., $B(x)$ is both open and closed. Hence every finite $\mathsf{T3}$-space is zero-dimensional. □

A subset $A \subseteq V$ is called *dense-in-itself* if it is contained in its derived set, i.e., in the set of all its limit points. Thus a set A is dense-in-itself if and only if for all $p \in A$ and all open sets O containing p we have $(O \cap A) \setminus \{p\} \neq \emptyset$. In a finite space this is equivalent to requiring that $(B(p) \cap A) \setminus \{p\} \neq \emptyset$. A topological space (V, τ) is *scattered* if there are no dense-in-itself subsets of V. A scattered space is always ($\mathsf{T0}$). In finite spaces the converse is also true:

Lemma 5. *A finite* $\mathsf{T0}$*-space is scattered.*

Proof. Let (V, τ) be a $\mathsf{T0}$-space. Suppose $B(x) \neq \{x\}$ for all $x \in V$. Consider a point p for which $|B(p)| \geq 2$ is minimal. Then for all $x \in B(p)$ we have $B(x) = B(p)$. Thus any two elements of $B(p)$ cannot be separated by an open set, contradicting ($\mathsf{T0}$). A $\mathsf{T0}$-space therefore contains at least one point p for which $B(p) = \{p\}$.
If (V, τ) is ($\mathsf{T0}$) then each subset with the induced topology is also ($\mathsf{T0}$). Thus for each $A \subseteq V$ there is a $q \in A$ such that $B(q) \cap A = \{q\}$. The set A is dense-in-itself if and only if $B(p) \cap A \neq \{p\}$ for all $p \in A$. Thus no subset of a finite $\mathsf{T0}$-space is dense-in-itself, i.e., a finite $\mathsf{T0}$-space is scattered. □

The axiom ($\mathsf{T3}$) implies ($\mathsf{T5}$) in finite (and in fact already in second countable) spaces, while ($\mathsf{T5}$) always implies ($\mathsf{T4}$). As a consequence ($\mathsf{T3}$) and ($\mathsf{T3}_{1/2}$) are equivalent if V is finite, since ($\mathsf{T4}$) and ($\mathsf{T3}$) together always imply ($\mathsf{T3}_{1/2}$). Interestingly, the ($\mathsf{T4}$) axiom is closely related to connectedness:

Lemma 6. *A finite topological space satisfies* ($\mathsf{T4}$) *if and only if every connected component is ultra-connected.*

Proof. First note that (V, τ) satisfies ($\mathsf{T4}$) if and only if each of its connected components satisfies ($\mathsf{T4}$). Thus we may assume that (V, τ) is connected. The absence of disjoint closed sets trivially implies ($\mathsf{T4}$).
Suppose that (V, τ) is a connected $\mathsf{T4}$-space with two disjoint closed sets $A, A' \neq \emptyset$. Since V is finite we may construct the smallest open neighborhood $B(A)$ as the intersection of all open sets containing A. Since (V, τ) is connected A is not both open and closed, i.e., $B(A) \neq A$. Furthermore $C = X \setminus B(A)$ is a non-empty closed set (it must contain A'), and by ($\mathsf{T4}$) it must have an open neighborhood W disjoint from $B(A)$. Hence $W \subseteq X \setminus B(A) = C \subseteq W$, i.e., $W = C$. Thus C is both open and closed, contradicting the connectedness of (V, τ). □

A.4. Hierarchies.

Definition 3. *A system of* $\mathcal{B} \subseteq \mathcal{P}(V)$ *is a* (strong) hierarchy *on* V *if (i) the union of all elements of* $\mathcal{B}$ *equals* V *and (ii) for all* $P, Q \in \mathcal{B}$ *the intersection* $P \cap Q$ *equals* P, Q, *or* $\emptyset$, *see e.g.* [19]. *Let us call a topological space* (V, τ) *is* hierarchical *if it has a basis that is a hierarchy on* V.

It is clear that a hierarchy can be interpreted as the basis of a topological space: since the intersections of elements of $\mathcal{B}$ are again elements of $\mathcal{B}$, the open sets are exactly the unions of $\mathcal{B}$-elements.

Lemma 7. *A finite hierarchical space is (topologically) connected if and only if there exists $x \in V$ such that $B(x) = V$.*

Proof. Suppose V is hierarchical and connected. For each $x \in V$ we define $H(x)$ as the union of all basis elements $B(y)$ containing x. Obviously $H(x) \in \mathcal{B}$ as a consequence of the intersection property of hierarchies. By the same token, all basis elements that contain a point $z \in H(x)$ must be contained in $H(x)$. Thus $V \setminus H(x)$ is the union of all basis elements that do not contain a point in $H(x)$; it is therefore and open set, i.e., $H(x)$ and $V \setminus H(x)$ form a separation of V. Hence $V \setminus H(x) = \emptyset$ and $V = H(x) = B(y)$. The converse is obvious. □

Lemma 8. *A finite hierarchical space satisfies* (T5).

Proof. Let (V, τ) be finite, connected, and hierarchical. By lemma 7 there is a basis element satisfying $B(x) = V$. Thus (V, τ) is ultra-connected by lemma 3. Lemma 6 now implies that a finite hierarchical space satisfies (T4). It is obvious that every subspace of a finite hierarchical space is again hierarchical and therefore also satisfies (T4). Gaal's [17] result that a topological space is satisfies (T5) if and only if each subspace is a T4-space completes the proof. □

On the other hand it is obvious that a partition topology, an hence every finite T3-space, is hierarchical since all basis elements are disjoint. Starting from the definition of the T5 property, the following result is quite surprising:

Lemma 9. *A finite topological space satisfies* (T5) *if and only if it is hierarchical.*

Proof. We only have to show that (T5) implies that $\mathcal{B}$ is a hierarchy, the converse is lemma 8. Let V_1 through V_m be the components of V. Since each of the components is ultra-connected by lemma 6, there is a $z_j \in V_j$, $j = 1, \ldots, m$ such that $B(z_j) = V_j$. Next we determine the sets $X_j = \{y \in V_j | B(y) = V_j\}$. We add the sets V_j to a collection $\tilde{\mathcal{B}}$ and form the subspace $V' = \bigcup_j (V_j \setminus X_j)$. It is clear that the basis of the topological subspace V' of V consists of the basis elements $B(x) \cap V' = B(x)$ if $x \in V'$, i.e., the basis is $\mathcal{B} \setminus \tilde{\mathcal{B}}$. By Gaal's theorem every subspace of a T5-space is again T5 because every one of its subspaces satisfies (T4), i.e., we may repeat this procedure with V'. From each component of V' be obtain basis elements which we add to $\tilde{\mathcal{B}}$ and remove from $\mathcal{B}$. By construction they are pairwise disjoint because they form components of V' and each of them was contained entirely in a component at the previous step. Thus $\tilde{\mathcal{B}}$ is a hierarchy, and after at most $|V|$ steps we have moved all basis elements from $\mathcal{B}$ to $\tilde{\mathcal{B}}$. The basis of a finite T5-space is therefore a hierarchy. □

We are now in the position to express the separation axioms in a way that lends itself to computational surveys:

Proposition 1. *Let (V, τ) be a finite topological space. Then the separation axioms can be expressed as follows.*

(T1) $B(x) = \{x\}$.
(T0) $B(x) \neq B(y)$ *for all* $x \neq y \in V$.
(T3) $B(x) \cap B(y) = \emptyset$ *or* $B(x) = B(y)$.
(T5) $B(x) \cap B(y) = \emptyset$, $B(x)$, *or* $B(y)$.
(T4) *For each connected component V' of V there is $x \in V'$ such that $B(x) = V'$.*

Furthermore, none of the implications T1⟹T0 *and* T3⟹T5⟹T4 *can be reversed.*

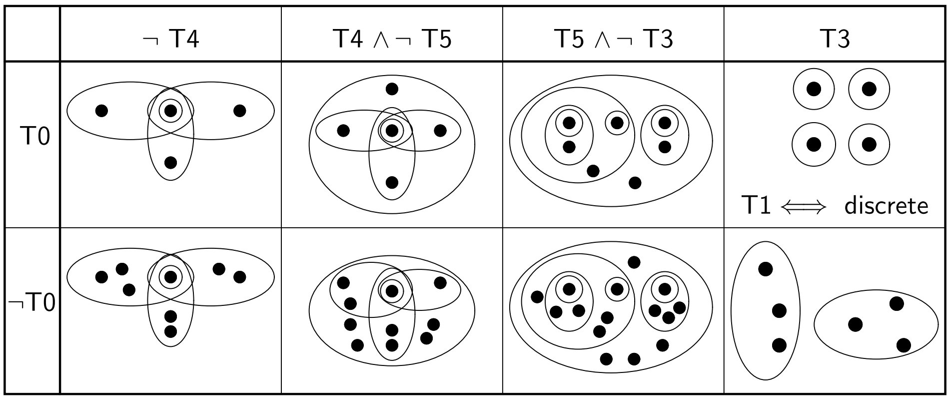

Figure 7. The implications T1⟹T0 and T3⟹T5⟹T4 between the separation axioms are not reversible. All examples shown here with the exception of the T3-spaces are connected. Elements of V are displayed as $\bullet$, ellipses denote the elements of the basis $\mathcal{B}$.

Proof. Since (T1) implies the discrete topology, the condition is trivial. Lemma 9 is equivalent to the condition for (T5), (T3) means that $\mathcal{B}$ is a partition topology, see lemma 4. Condition (T4) states that a V is T4-space iff each connected component is ultra-connected, lemmas 6 and 3. It remains to verify (T0): If V is a T0-space, then for all x, y holds $y \notin B(x)$ or $x \notin B(y)$, and hence $B(x) \neq B(y)$. Conversely, suppose $B(x) \neq B(y)$ but $x, y \in B(x) \cap B(y)$ (contradicting the separation axiom T0). In this case the axioms of topology require $B(x) \subseteq B(x) \cap B(y)$ and $B(y) \subseteq B(x) \cap B(y)$, i.e., $B(x) = B(y)$. Thus $B(x) \neq B(y)$ implies that $x \notin B(y)$ or $y \notin B(x)$, and hence (T0). The implications, which have already been discussed in the text, are obvious. Examples showing that the implications cannot be reversed are given in Figure 9. □

References

[1] P. Alberch. Morphological variation in the neotropical salamander genus bolitoglossa. *Evolution*, 37:906–919, 1983.

[2] A. Babajide, R. Farber, I. L. Hofacker, J. Inman, A. S. Lapedes, and P. F. Stadler. Exploring protein sequence space using knowledge based potentials. *J. Comp. Biol.*, 1999. submitted, Santa Fe Institute preprint 98-11-103.

[3] A. Babajide, I. L. Hofacker, M. J. Sippl, and P. F. Stadler. Neutral networks in protein space: A computational study based on knowledge-based potentials of mean force. *Folding & Design*, 2:261–269, 1997.

[4] C. Berge. *Hypergraphs*. Elsevier, Amsterdam NL, 1989.

[5] E. G. Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophys. J.*, 73:2393–2403, 1997.

[6] G. Choquet. Convergences. *Ann. Univ. Grenoble*, 23:55–112, 1947.

[7] J. C. Culberson. Mutation-crossover isomorphism and the construction of discriminating functions. *Evol. Comp.*, 2:279–311, 1995.

[8] B. A. Devay and H. A. Priestley. *Introduction to Lattices and Order*. Cambridge Univ. Press, Cambridge, UK, 1990.

[9] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. Principle of protein folding: A perspective from simple exact models. *Prot. Sci.*, 4:562–602, 1995.

[10] H. R. Fischer. Limesräume. *Math. Ann.*, 137:269–303, 1959.

[11] W. Fontana, T. Griesmacher, W. Schnabl, P. F. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degradation rate constants of RNA secondary structures. *Monatsh. Chem.*, 122:795–819, 1991.

[12] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[13] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaption. *Phys. Rev. A*, 40:3301–3321, 1989.

[14] W. Fontana and P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280:1451–1455, 1998.

[15] W. Fontana and P. Schuster. Shaping space. The possible and the attainable in RNA genotype-phenotype mapping. *J. Theor. Biol.*, 194:491–515, 1998.

[16] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for prediction of RNA duplex stability. *Proc. Natl. Acad. Sci. USA*, 83:9373–9377, 1986.

[17] S. A. Gaal. *Point Set Topology.* Academic Press, New York, 1964.

[18] P. Gitchoff and G. P. Wagner. Recombination induced hypergraphs: A new approach to mutation-recombination isomorphism. *Complexity*, 2:47–43, 1996.

[19] A. D. Gordon. A review of hierarchical classification. *J. R. Statist. Soc.*, A150:119–137, 1987.

[20] W. Grüner, R. Giegerich, D. Strothmann, C. M. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks. *Monath. Chem.*, 127:355–374, 1996.

[21] W. Grüner, R. Giegerich, D. Strothmann, C. M. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neutral networks and shape space covering. *Monath. Chem.*, 127:375–389, 1996.

[22] F. Hausdorff. Gestufte Räume. *Fund. Math.*, 25:486–502, 1935.

[23] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chemie*, 125(2):167–188, 1994.

[24] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. USA*, 93:397–401, 1996.

[25] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci. USA*, 86:7706–7710, 1989.

[26] D. C. Kent. Convergence functions and their related topologies. *Fund. Math.*, 54:125–133, 1964.

[27] J. Maynard-Smith, R. Burian, S. Kauffman, P. Alberch, J. Campell, B. Goodwin, R. Lande, D. Raup, and L. Wolpert. Developmental constraints and evolution. *Quart. Rev. Biol.*, 60:265–287, 1985.

[28] C. Reidys, C. Forst, and P. Schuster. Replication and mutation on neutral networks. *Bull. Math. Biol.*, 1998. in press, Santa Fe Institute Preprint 98-04-036.

[29] C. Reidys and P. F. Stadler. Bio-molecular shapes and algebraic structures. *Computers & Chem.*, 20:85–94, 1996.

[30] C. M. Reidys. Random induced subgraphs of generalized n-cubes. *Adv. Appl. Math.*, 19:360–377, 1997.

[31] C. M. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[32] J. Rienesl and G. P. Wagner. Constancy and change of basipodial variation patterns: a comparative study of crested and marbled newts – *triturus cristatus, triturus marmoratus* – and their natural hybrids. *J. Evol. Biol.*, 5:307–324, 1992.

[33] P. Schuster. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. *J. Biotechnology*, 41:239–257, 1995.

[34] P. Schuster and W. Fontana. Chance and necessity in evolution: lessons from RNA. *Physica D*, 1999. in press, Santa Fe Institute preprint 98-11-107.

[35] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. Lond. B*, 255:279–284, 1994.

[36] P. Schuster, P. F. Stadler, and A. Renner. RNA structures and folding: From conventional to new issues in structure predictions. *Curr. Opinions Structural Biol.*, 7:229–235, 1997.

[37] K. Schwenk. A utilitarian approach to evolutionary constraint. *Zoology*, 98:251–262, 1995.

[38] P. F. Stadler, R. Seitz, and G. P. Wagner. Evolvability of complex characters: Population dependent fourier decomposition of fitness landscapes over recombination spaces. *Bull. Math. Biol.*, 1999. in press, Santa Fe Institute Preprint 99-01-001.

[39] P. F. Stadler and G. P. Wagner. The algebraic theory of recombination spaces. *Evol. Comp.*, 5:241–275, 1998.

[40] L. A. Steen and J. A. Seebach, Jr. *Counterexamples in Topology.* Holt, Rinehart & Winston, New York, 1970.

[41] G. P. Wagner and L. Altenberg. Complex adaptations and the evolution of evolvability. *Evolution*, 50:967–976, 1996.

[42] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of RNA folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

[43] J. Weber. *Dynamics of Neutral Evolution – A case study on RNA secondary structures.* PhD thesis, Friedrich Schiller Universität, Jena, Germany, 1997.

[44] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

[45] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46(4):591–621, 1984.