# Genotype-Phenotype Maps

Peter F. Stadler

*Institut für Theoretische Chemie und Molekulare Strukturbiologie
Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

*Address for correspondence

**Abstract.** The current implementation of the Neo-Darwinian model of evolution typically assumes that the set of possible phenotypes is organized into a highly symmetric and regular space equipped at least with a notion of distance, for example, a Euclidean vector space. Recent computational work on the biophysical genotype-phenotype model defined by the folding of RNA sequences into secondary structures suggests a rather different picture. If phenotypes are organized according to genetic accessibility, the resulting space lacks a metric and is formalized by an unfamiliar structure, known as a pretopology. If recombination is taken into account, an even weaker structure, known as neighborhood space, must be used.

Patterns of phenotypic evolution — such as punctuation, irreversibility, and modularity — result naturally from the properties of the genotype-phenotype map, which, given the genetic accessibility structure, defines accessibility in the phenotype space. The classical framework, however, addresses these patterns by exclusively invoking natural selection on suitably imposed fitness landscapes. We extend the explanatory level for phenotypic evolution from fitness considerations alone to include the topological structure of phenotype space as induced by the genotype-phenotype map. The topological framework allows us to consider e.g. the continuity of an evolutionary trajectory in an unambiguous way.

Lewontin's notion of "quasi-independence" of characters can be formalized as the assumption that a region of the phenotype space is represented by a product space of orthogonal factors. In this picture each character corresponds to a factor of a region of the phenotype space. We consider any region of the phenotype space that has a given factorization as a "type", i.e., as a set of phenotypes that share the same set of phenotypic characters. Using the notion of local factorizations a theory of character identity can be developed that is based the correspondence of local factors in different regions of the phenotype space.

## 1. Introduction

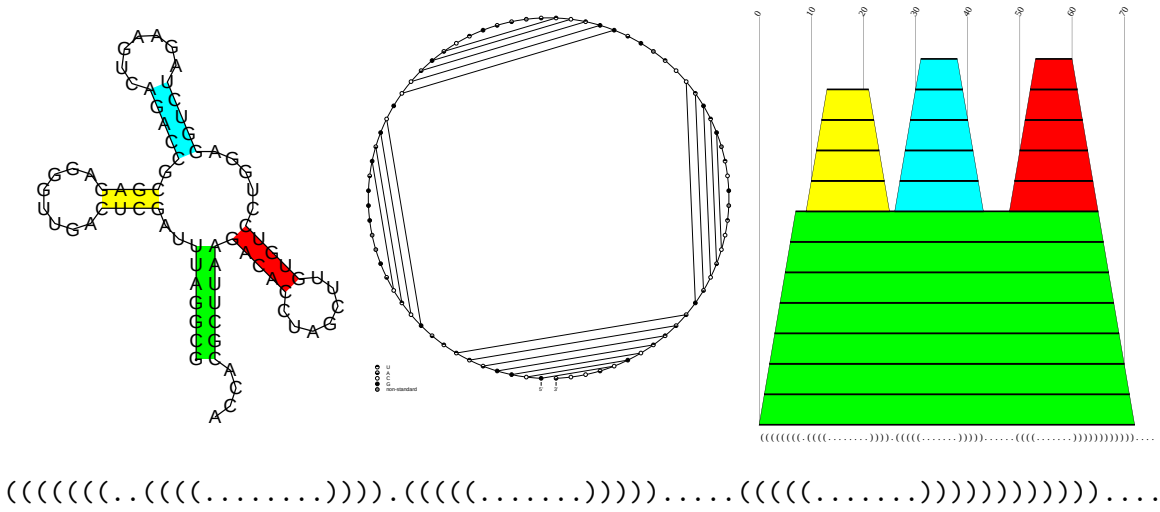
*Fitness landscapes* were introduced in the 1930s by Sewall Wright [1, 2] as a means of visualizing evolutionary adaptation. In this picture a population moves uphill on a kind of "potential function" due to the combined effects of mutation and selection. Modern genetics teaches us that mutation (and other genetic operators) act at the level of genotypes, while fitness-based selection is determined by the phenotypes. As a consequence, the relationships between genotype and phenotype, mathematically expressed as the *Genotype-Phenotype-map* (GP-map), play a crucial role in the understanding of evolution. Indeed, phenotypic innovation is the result of genetic modification mediated by the GP-map. Fontana & Schuster [3] therefore emphasize that the notion of phenotypic neighborhood is induced by the GP-map and that it may differ fundamentally from any notion of "nearness" among phenotypes based solely on the comparison of their morphological features.

From a mathematical point of view, the description of the structure of genotype space, phenotype space, and maps associated with them poses inherently topological questions. It turns out, however, that textbook point-set topology cannot be used without modifications. The reason for this complication is that the natural notions of neighborhood arising from the dominant modes of genetic variation are weaker than those usually employed in topology. A generalized topological theory, however, is well-suited for our purpose.

This contributions is organized as follows: In section 2 the best studied model of a GP-map, namely the folding of RNA sequences into their secondary structures, is outlined and the notion of accessibility and its impact on evolutionary processes is discussed. Then we introduce the language of *generalized topologies* that is necessary to describe the GP-map in a natural way. We will give precise definitions and state some of the most important results; for proofs and additional details the reader is referred to [4, 5, 6] and the references therein. The structure of the phenotype space is described in some detail in section 4. The notion of continuity of GP-maps and evolutionary trajectories is the topic of section 5. As a second example for the applicability of the topological language we briefly review concept of characters and its relation to product spaces in section 6.

## 2. The RNA Model and Neutral Networks

Qualitatively, there is ample evidence for neutrality both in the molecular record of natural evolution and in laboratory experiments under controlled conditions. On the other hand, very little detail is known about regularities in genotype-phenotype relations that cause the observed neutrality. In this section we briefly review the best studied *model*: RNA secondary structures. We consider RNA sequences as genotypes; the role of the phenotype is then played by the structure of the molecule, see e.g. [7]. This simplifying assumption is met indeed by RNA evolution experiments *in vitro* [8] as well as by the design of RNA molecules through artificial selection [9].



(((((((..((((........))))..(((((.......)))))).....(((((.......)))))))))))....

**Figure 1.** Four equivalent representations of the clover leaf shaped secondary structure of $\mathrm{tRNA}^{\mathrm{phe}}$. The conventional drawing emphasizes the helical regions shown in color, the circular drawing emphasizes the outer-planar nature of the graph, the "mountain representation" is obtained by using the sequence positions as $x$-axis. Below we show the dot-parenthesis string.

RNA secondary structures provide a discrete, coarse grained concept of structure that is 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. In the literature, secondary structures are routinely used to display, organize, and interpret experimental findings. They are oftentimes conserved over evolutionary times scales. 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 so-called pseudo-knots. By construction, secondary structures are 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 [10] that makes use of a extensive set of experimental determined energy parameters [11]. Data reported here have been obtained using the `Vienna RNA Package` [12].

For a given chain length $n$, there are $\alpha^n$ 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 [13]. In an exhaustive computational study all GC sequences with a chain length up to $n = 30$ were folded [14, 15]. Exhaustive studies

are obviously limited to short RNAs, since the number of objects like sequences and structures that can be handled on conventional present day computers is limited to some $10^9$; in many cases properties of longer sequences can be estimated from random samples.

From a more formal point of view we are interested in the structure of the neutral set $f^{-1}(\psi)$ of a phenotype $\phi$, i.e., the set of all genotypes that fold into the structure $\psi$. The second important topic is the mutual location of the neutral sets of different phenotypes. More precisely, we consider the induced subgraph $\mathcal{G}(\psi) = \mathcal{Q}_n[f^{-1}(\psi)]$ that is embedded in the sequence space (hypercube) $\mathcal{Q}_n$. The graph $G(\psi)$ is known as the *neutral network* of the phenotype $\psi$. The most basic quantities to characterize $G(\psi)$ are its number of vertices $|f^{-1}(\psi)|$ and edges $E_\psi$. The parameter $\bar{\lambda}(\psi) = |f^{-1}(\psi)|/(2E_\psi)$, i.e., the average degree of $G(\psi)$ measures the average number of *neutral neighbors*. It may serve as a convenient global characterization of neutral networks.

Four properties of the RNA GP-map were derived from computational studies [16]:

(i) **More sequences than structures**. For sequence spaces of chain lengths $n \geq 10$ there are orders of magnitude more sequences than structures and hence, the map is many-to-one. This fact can be proved combinatorially. Computational data indicate that only about $1.65^n$ secondary structure graphs are actually realized as the minimum energy structure of some RNA sequence. The GP-map of RNA is therefore highly redundant.

(ii) **Few common and many rare structures**. Relatively few common structures are opposed by a relatively large number of rare structures, some of which are formed by a single sequence only ("relatively" refers to the fact that the numbers of both common and rare structures increase exponentially with $n$, but the exponent for the common structures is smaller than that for the rare ones).

(iii) **Connectivity of neutral networks**. Neutral networks of common structures are connected unless specific and readily recognizable special features of RNA structures require specific non-random distribution in sequence space. These may cause $G(\psi)$ to decompose into a small number (typically 4 or fewer) connected components that are separated only be a small distance. This effect is much more pronounced in the restricted GC alphabet than in the natural GCAU alphabet. A random graph approach [17, 18] predicts that the neutral networks are connected provided the average neutrality $\bar{\lambda}$ exceeds a threshold that only depends on the nucleic acid alphabet. For the biophysical GCAU-alphabet the threshold value is $\lambda^* \approx 0.3700$. This value is significantly smaller than the average neutrality $\bar{\lambda}(\psi)$ for the common RNA structures.

(iv) **Shape space covering**. The neutral network $\mathcal{G}(\psi)$ is embedded in a compatible set $C(\psi)$ which includes all sequences that can form the structure $\psi$ at least as a suboptimal conformation. Computational studies show that $G(S)$ is well described in many respects by a random induced subgraph of the compatible part of sequence space $\mathcal{Q}[C(\psi)]$. On the other hand, one can prove rigorously that the compatible sets of any two phenotypes $C(\psi)$ and $C(\varphi)$ have a non-empty intersection [17]. As a result it is possible to define

a spherical ball with a diameter $R_{\text{cov}} \ll n$ which contains on the average for every common structure at least on sequence that folds into it.

Computational studies based on inverse folding by means of knowledge-based potentials [19] strongly indicate that all four features hold for protein spaces as well [20, 21]. Proteins, in contrast to RNA molecules, do not always form stable structures but may aggregate in aqueous solution when their constituents are too hydrophobic. This means that no useful structures will be available in certain parts of sequence space and the protein landscape is therefore *"holey"*. The concept of holey landscapes has been transferred also to the much more sophisticated problem of evolution of higher organisms and speciation [22].

The predictions (i) through (iv) are in agreement with experimental findings. For instance, Schultes & Bartel [23] describe an RNA that simultaneously carries the properties of two different RNA folds and which is connected by a neutral path with the two sequences from which it was originally designed. Other reports that are of interest in this context include [24, 25] for the protein case. Empirical evidence for a large degree of *functional* neutrality in protein space is described e.g. by Wain-Hobson and co-workers [26].

Neutrality has a number of important impacts on the dynamical behavior of a population of replicating individuals. Most importantly, there is a diffusive motion of the population's "center of gravity" through sequence space [27]. The diffusion constant is related to population size $N$, per digit mutation rate $p$ and the fraction $\lambda$ of neutral neighbors [28]. A constant "rate of innovation" is reported in [29] for the landscapes in which all neutral networks come close together, as in the case of RNA. A population therefore evolves by producing mutants in the boundary of the neutral network of the currently dominating species. Short, intermittent episodes of selection interrupt the diffusive behavior when a fitter mutant that invades the population and spreads through it.

The *accessibility* of fitter phenotypes therefore is the determining factor at large timescales, as pointed out by Walter Fontana and Peter Schuster [3, 30]. Phenotypes that are easily obtained from the present one would correspond to "continuous" steps along an evolutionary trajectory, while rare, unlikely mutations constitute "discontinuities". Continuity is an intrinsically topological term; it may not be surprising therefore, that the work of Fontana & Schuster can be reformulated in more rigorous mathematical terms [31, 4] that revealed the necessity for using a generalized version of point set topology.

## 3. Generalized Topologies

In this section we briefly review the mathematical foundations of a general theory of accessibility. The basic assumption is that, given a set of genotypes $A$ there is a well-defined set $A'$ of genotypes that can be reached ("accessed") in the next time-step. In the case of discrete synchronized generations the notion of a time-step is well defined. In general it is useful to consider the effective generation time as time unit.

Point set topology, as we shall see, provides a meaningful framework in which we can discuss the consequences of accessibility differences.

Topology textbooks, e.g. [32, 33], usually start out by defining a topology on a set $X$ by means of a collection $\mathcal{O}$ of open sets, or, equivalently, by means of a collection $\mathcal{C}$ of closed sets. By definition, $A \in \mathcal{C}$ if and only if $X - A \in \mathcal{O}$, i.e., $A$ is an open set if its complement is a closed set and *vice versa*. In a topological space the following three axioms hold:

(I1) $X \in \mathcal{C}$.
(I2) For every index set $I$ holds: If $A_i \in \mathcal{C}$ for all $i \in I$, then $\bigcap\{A_i | i \in I\} \in \mathcal{C}$.
(I3) If $A, B \in \mathcal{C}$ then $A \cup B \in \mathcal{C}$.

The corresponding axioms for open sets are obtained by exchanging unions and intersections. Setting $I = \emptyset$ in (I2) we see $\emptyset \in \mathcal{C}$. In lattice theory more general, so-called *intersection structures*, are considered that fulfill only (I2), see e.g., [34]. An intersection function is *topped* if (I1) holds, too.

Given an intersection structure $(X, \mathcal{C})$, a *closure function* can be defined that associates with each set $A \subseteq X$ the smallest closed set containing $A$

$$\mathsf{cl}(A) = \bigcap\{B \in \mathcal{C} | A \subseteq B\} \,. \tag{1}$$

The function $\mathsf{cl}$ has two important properties: (i) it is *isotone*, i.e., $A' \subseteq A$ implies $\mathsf{cl}(A') \subseteq \mathsf{cl}(A)$ and (ii) it is *idempotent*, i.e., $\mathsf{cl}(\mathsf{cl}(A)) = \mathsf{cl}(A)$. If (I1) also holds, the closure function is *expanding*, i.e., $A \subseteq \mathsf{cl}(A)$. In a topology, i.e., in the presence of (I3), finally, the closure function is additive, i.e., $\mathsf{cl}(A \cup B) = \mathsf{cl}(A) \cup \mathsf{cl}(B)$.

The "conjugate" or "dual" of the closure function is the *interior* defined by $\mathsf{int}(A) = X \setminus \mathsf{cl}(X \setminus A)$, i.e., $\mathsf{cl}(A) = X \setminus \mathsf{int}(X \setminus A)$. Closely related to the abstract concept of closure and interior of a set is the notion of *neighborhood*: $N$ is a neighborhood of $x$ if and only if $x$ lies in the interior of $N$. The neighborhood functions is thus defined by

$$\mathcal{N}(x) = \left\{ N \subseteq X \,\middle|\, x \in \mathsf{int}(N) \right\} \tag{2}$$

Conversely, given the neighborhoods of each point $x \in X$ it is possible to obtain the associated closure and interior functions [35]: $x \in \mathsf{cl}(A)$ iff $(X \setminus A) \notin \mathcal{N}(x)$ and $x \in \mathsf{int}(A)$ iff $A \in \mathcal{N}(x)$. In other words, closures, interiors, and neighborhoods are equivalent constructions on a set $X$. It is therefore possible to translate properties of the closure function $\mathsf{cl}$ into properties of the neighborhood function, and *vice versa*. Table 1 summarizes the basic axioms in all three languages.

**Table 1.** The basic axioms for extended topological spaces.

| | closure | interior | neighborhood |
|---|---|---|---|
| (K0) | $\mathsf{cl}(\emptyset) = \emptyset$ | $\mathsf{int}(X) = X$ | $\mathcal{N}(x) \neq \varnothing$ |
| (K1) | $A \subseteq B \implies \mathsf{cl}(A) \subseteq \mathsf{cl}(B)$<br>$\mathsf{cl}(A) \cup \mathsf{cl}(B) \subseteq \mathsf{cl}(A \cup B)$<br>$\mathsf{cl}(A \cap B) \subseteq \mathsf{cl}(A) \cap \mathsf{cl}(B)$ | $A \subseteq B \implies \mathsf{int}(A) \subseteq \mathsf{int}(B)$<br>$\mathsf{int}(A) \cup \mathsf{int}(B) \subseteq \mathsf{int}(A \cup B)$<br>$\mathsf{int}(A \cap B) \subseteq \mathsf{int}(A) \cap \mathsf{int}(B)$ | $N \in \mathcal{N}(x),\, N \subseteq N' \implies N' \in \mathcal{N}(x)$ |
| (K2) | $A \subseteq \mathsf{cl}(A)$ | $\mathsf{int}(A) \subseteq A$ | $N \in \mathcal{N}(x) \implies x \in N$ |
| (K3) | $\mathsf{cl}(A \cup B) \subseteq \mathsf{cl}(A) \cup \mathsf{cl}(B)$ | $\mathsf{int}(A) \cap \mathsf{int}(B) \subseteq \mathsf{int}(A \cap B)$ | $N', N'' \in \mathcal{N}(x) \implies N' \cap N'' \in \mathcal{N}(x)$ |
| (K4) | $\mathsf{cl}(\mathsf{cl}(A)) = \mathsf{cl}(A)$ | $\mathsf{int}(\mathsf{int}(A)) = \mathsf{int}(A)$ | $N \in \mathcal{N}(x) \iff \mathsf{int}(N) \in \mathcal{N}(x)$ |

Instead of starting with systems of open or closed sets one can also build the mathematical framework directly on the closure functions; for instance Čech's book [36] shows that many of the classical results of topology hold already for pretopologies that satisfy only (K0), (K1), (K2), and (K3) but lack idempotency (K4) of the closure function. This approach was pioneered by Day [35], Hammer [37, 38] and Gniłka [39]. Surprisingly, meaningful topological concepts can already be defined on a set $X$ endowed with an arbitrary set-valued set-function $\mathsf{cl}$. Almost all approaches to extending the framework of topology at least assume isotony (K1), *cf.* [40, 41, 37, 35, 42, 39] and many others. The importance of isotony is emphasized by numerous equivalent conditions, some of which are listed in Tab. 1. If (K1) holds we recover also more familiar relationship between closure and neighborhood:

$$\mathsf{cl}(A) = \left\{ x \in X \middle| \exists N \in \mathcal{N}(x) : \, N \cap A \neq \emptyset \right\} \tag{3}$$

In this case we may also express (K4) in a much more familiar form: *Every neighborhood $N$ of a point $x$ contains an open neighborhood.*

The concept of neighborhoods makes sense also for sets: $N$ is a neighborhood of $A$ if and only if $N$ is a neighborhood of each point of $A$, i.e.,

$$\mathcal{N}(A) = \bigcap_{x \in A} \mathcal{N}(x) \tag{4}$$

The notion of *boundary* can be derived from the closure function:

$$\partial A = \mathsf{cl}(A) \cap \mathsf{cl}(X \setminus A) = \mathsf{cl}(A) \setminus \mathsf{int}(A) \tag{5}$$

A topological theory based on boundaries was developed by [43]. From the boundary $\partial A$ we can recover closure and interior functions as $\mathsf{cl}(A) = A \cup \partial A$ and $\mathsf{int}(A) = A \setminus \partial A$. The biological interpretation of the boundary approach is discussed in some detail in [4].

Let us now return to the question of accessibility in genetic systems. In mathematical terms, we have to ask *What are the properties of $A'$ given $A$.* Clearly, an empty population cannot produce offsprings, hence $\emptyset' = \emptyset$. Furthermore, $A \subseteq B$ must imply $A' \subseteq B'$ since all the offsprings $A'$ of the sub-population $A$ of $B$ are still accessible even if $A$ is embedded in a larger population $B$. Finally, it appears safe to assume that we can at least in principle have the same genotypes as in the parent generation, i.e., $A \subseteq A'$. It follows that any "reasonable" genetic system will provide us with a closure function satisfying (K0), (K1), and (K2), i.e., that defines a neighborhood space on the set of genotypes [44].

Finite pretopological spaces correspond to (directed) graphs: $y$ is an out-neighbor of $x$ iff $y \neq x$ and $y \in \mathsf{cl}(x)$. Subspaces correspond to induced subgraphs. Connectedness in the topological and in the graph-theoretical sense coincide. In particular, the graph representing the pretopological space defined by point mutations of sequences with fixed length is of course the Hamming graph (generalized hypercube) that naturally represents the sequence spaces also without recourse to the topological formalism.

As argued in [45] and [44], however, even pretopological spaces are not general enough to deal with recombination. The abstract description of recombination spaces is pioneered in [45, 46, 47]. It is based on the notion of the *recombination function*

$\mathcal{R} : X \times X \to \mathcal{P}(X)$ assigning to each pair of parents $x$ and $y$ the *recombination set* $\mathcal{R}(x, y)$ introduced by [45] as the set of all their potential offsprings. Recombination in general satisfies two axioms:

(X1) $\{x, y\} \in \mathcal{R}(x, y)$,
(X2) $\mathcal{R}(x, y) = \mathcal{R}(y, x)$.

Condition (X1) states that replication may occur without recombination, and (X2) means that the role of the parents is exchangeable. Often a third condition (X3) $\mathcal{R}(x, x) = \{x\}$ is assumed which is not satisfied by models of unequal crossover [48, 5]. The *closure operator* associated with a recombination function was introduced by [45] as

$$\mathsf{cl}(A) = \bigcup_{x,y \in A} \mathcal{R}(x, y) \tag{6}$$

It is not hard to see that $\mathsf{cl}$ as defined in equ.(6) defines a neighborhood space [44].

## 4. Phenotype Space

Let us first consider the case of large (or even infinite) populations. In this case the entire neutral set $f^{-1}(\phi)$ of a phenotype $\phi$ can be covered by the population. Hence the phenotypes accessible from a set $\Phi$ of phenotypes consists of all mutants arising from $f^{-1}(\Phi)$. Hence we define the *accessibility closure* on phenotype space by

$$C_\infty(\Phi) = f\left(\mathsf{cl}\big(f^{-1}(\Phi)\big)\right) \tag{7}$$

In general, however, a population $\mathfrak{P}$ of genotypes will be much too small to completely cover the neutral sets. In this case we might want to use a more restrictive closure on the phenotype space

$$C_{\mathfrak{P}}(\Phi) = f\left(\mathsf{cl}\big(f^{-1}(\Phi) \cap \mathfrak{P}\big)\right) \tag{8}$$

Note that $C_{\mathfrak{P}}(\Phi) \subseteq C_{\mathfrak{Q}}(\Phi)$ whenever $\mathfrak{P} \subseteq \mathfrak{Q}$, i.e., smaller populations produce an effectively finer closure function on phenotype space. It seems to be more convenient, however, to base the phenotypic closure function on a notion of "sufficiently frequent" accessibility. Let $n^*(\psi, \phi)$ denote the average number of mutants of $x \in f^{-1}(\phi)$ that has phenotype $\psi$. We may say that $\psi$ is easily accessible from $\phi$ if $n^*(\psi, \phi)$ exceeds some threshold value. Similar notions of easy-accessibility are discussed in [3, 30, 31]. Fig. 2 shows an example of the resulting pretopological space for short RNA sequences. Note that the definition of easy-accessibility need not be symmetric.

In the RNA case there appears to be a natural definition of "frequent neighbors". As described in [30] there is a well-defined and relatively small subset of accessible secondary structures that appear as mutants of most of the sequences that fold into a particular common structure. This "shadow" may be regarded as the easily accessible structure. The important property of this set of neighbors is that, with few exceptions, they can be characterized in terms of just a few allowed structural changes: elongation and contraction of an existing stacked region or the complete removal of a stack, Fig. 3. These rules define the Fontana-Schuster pretopology for RNA secondary structures.

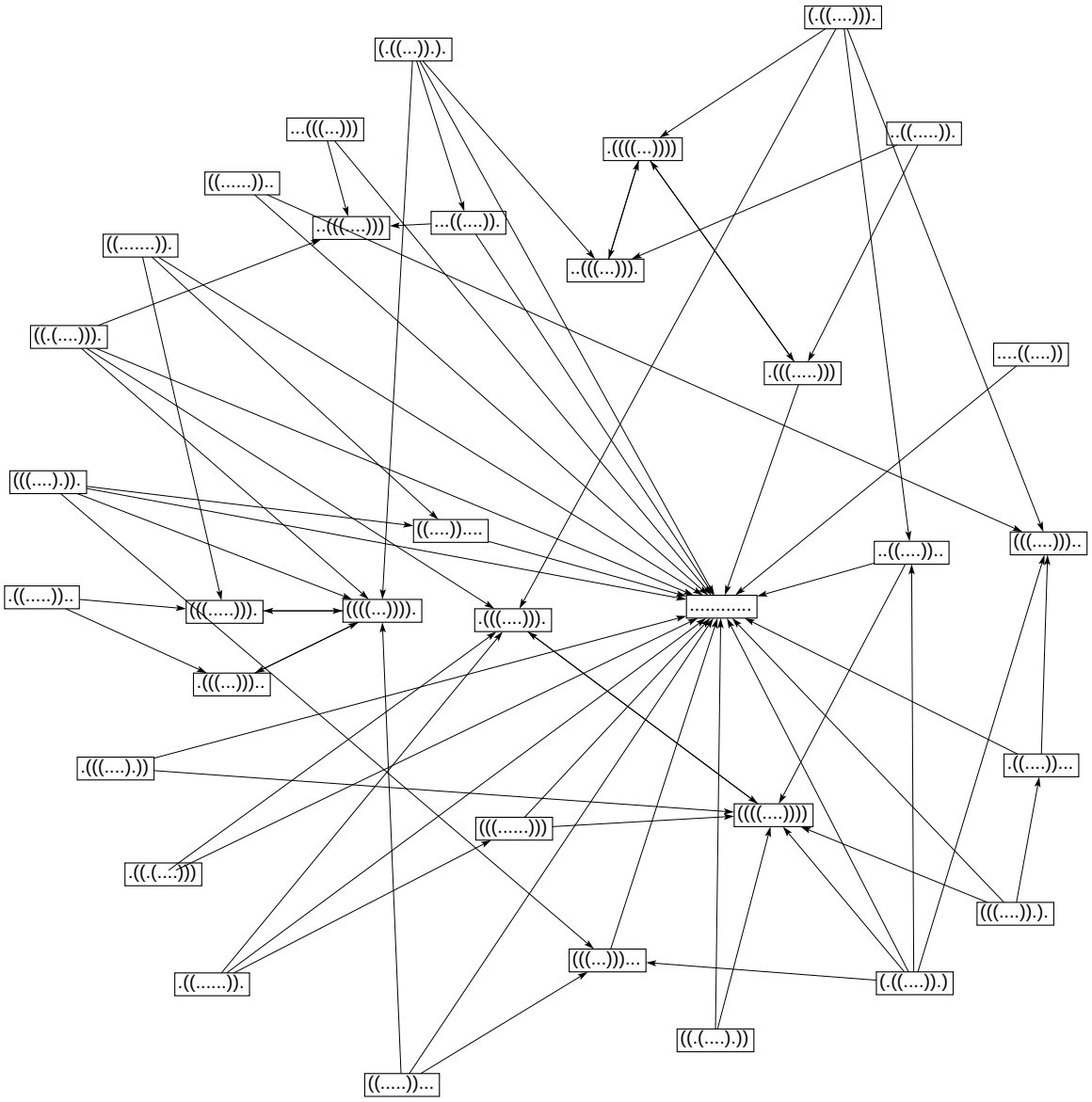



**Figure 2.** Pretopology of the GC12 shape space with accessibility defined as $n^*(\psi, \phi) > 1$. Note that some of the rare structures are not easily accessible from any other structure under this criterion.

For short sequences one finds significant deviations between the idealized FS-pretopology and frequency-based rules of easy-accessibility, as Fig. 2 shows. Numerical studies indicate, however, that the FS-pretopology is a good approximation for larger molecules [3, 30].

The notion of easy-accessibility is inherently stochastic since we ask "what is the *probability* to obtain a phenotype $\phi$ from $\psi$" Menger [49] suggested six decades ago that, due to inherent uncertainties of measurements, the distance $d(x, y)$ should be replaced by a probability distribution $P(x, y; d)$ describing the probability that the distance between $x$ and $y$ is at least $d$; see e.g. [50]. More recently, *probabilistic convergence spaces* have been introduced [51] where $(\mathcal{F}, x) \in q_\lambda$ means that the filter $\mathcal{F}$

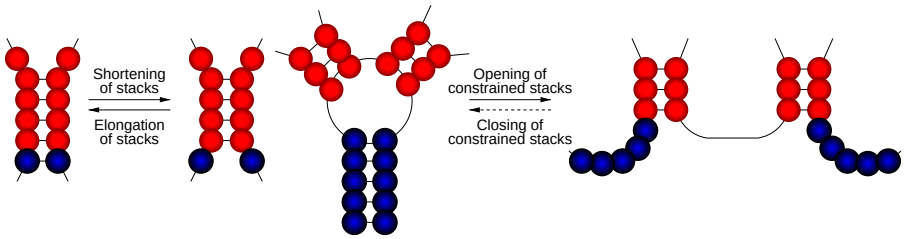



**Figure 3.** Fontana-Schuster pretopology. Contraction and elongation of stacks as well as the opening of constrained stacks in general lead to easily accessible structures. Closing a constrained stack, on the other hand, leads to structures that are inaccessible by point mutations in most sequence context.

converges to $x$ with probability at least $\lambda$. Obviously, $(\mathcal{F}, x) \in q_\lambda$ implies $(\mathcal{F}, x) \in q_{\lambda'}$ for all $\lambda' \leq \lambda$. The neighborhood filter of a convergence space is defined as

$$\mathcal{N}(x) = \bigcap \{\mathcal{F} \big| \mathcal{F} \to x\} \tag{9}$$

It follows that

(N*) $\mathcal{N}_{\lambda'}(x) \subseteq \mathcal{N}_\lambda(x)$ for all $\lambda' \leq \lambda$.

which we can use as an axiom for defining probabilistic neighborhood functions, i.e., $N \in \mathcal{N}_\lambda(x)$ means $N$ *is a neighborhood of* $x$ *with probability at least* $\lambda$. The associated probabilistic closure function is $C_\lambda : \mathcal{P}(X) \to \mathcal{P}(X)$ given by

$$C_\lambda(A) = \{y \in X \big| A \cap N \neq \emptyset \text{ for all } N \in \mathcal{N}[y, \lambda]\} \tag{10}$$

satisfying $C_\lambda(A) \subseteq C_{\lambda'}(A)$ for all $\lambda' \leq \lambda$.

The closure functions $C_\lambda$ now depend on the probability parameter $\lambda$. Instead, we might want to consider the probability $\xi(x, A)$ that a point $x$ is contained in the closure of $A$:

$$\xi(x; A) = \sup_\lambda \{x \in C_\lambda(A)\} \tag{11}$$

The map $x \mapsto \xi(x; A)$ can be interpreted as a *fuzzy closure* of $A$ since $\xi(x; A) \in [0, 1]$. By construction, $c(x, A) = 1$ for all $x \in A$. Note, however, that $C_\lambda$ does not fully specify a fuzzy closure function since $c(x, A)$ is determined only for ordinary sets $A$. An extension of the present framework to a probabilistic setting or to fuzzy set topology (see e.g. [52]) may prove useful in the future.

## 5. Continuity

Continuity is a property of a map $f : (X, \mathsf{cl}) \to (Y, C)$ between two (generalized) topological spaces $(X, \mathsf{cl})$ and $(Y, C)$. Intuitively, it implies that "small" changes in the argument $x$ of $f$ may cause only limited changes in the images $f(x)$. One can show that following four conditions are equivalent provided $\mathsf{cl}$ and $C$ are at least isotone. If they are satisfied, one says that $f$ is continuous.

(i) $\mathsf{cl}(f^{-1}(B)) \subseteq f^{-1}(C(B))$ for all $B \in \mathcal{P}(Y)$.
(ii) $f^{-1}(\mathrm{I}(B)) \subseteq \mathsf{int}(f^{-1}(B))$ for all $B \in \mathcal{P}(Y)$.

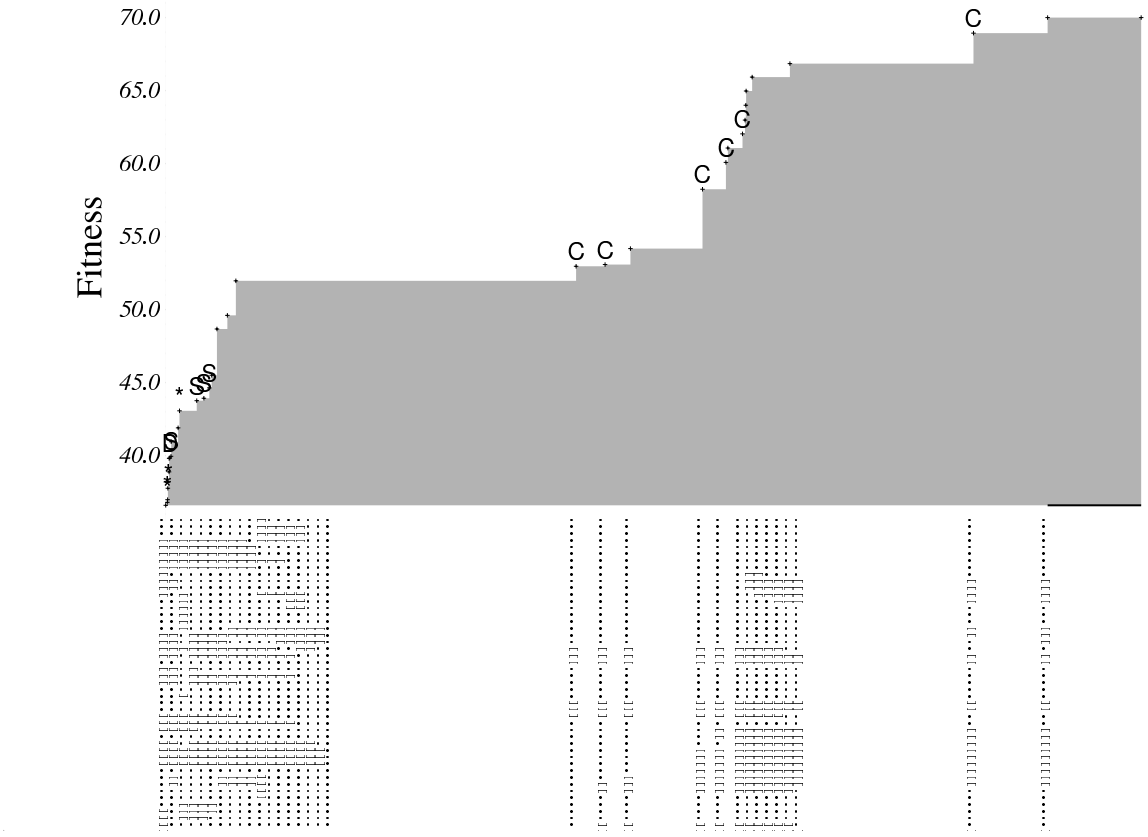



**Figure 4.** A simple evolutionary trajectory obtained from an adaptive walk. Fitness depends $f(\psi) = n - d(\psi, \tau)$ where $n = 70$ is the chain length, $\tau$ is the target structure which is attained at the end of the run and $d$ is a structure distance obtained by adding weights $1 + 0.1\epsilon$ for each base pairs that the two structures do not have in common; $\epsilon$ is uniformly distributed in $[-1, +1]$. Letters indicate the type of transitions (S for shift moves in which one side of a stack changes its position, C is the closing of a (constrained) stack, and $*$ indicated major re-foldings; these types are non-continuous transitions in the Fontana-Schuster pretopology, continuous transitions are not marked). The horizontal axis measures time in terms of attempted steps.

(iii) $B \in \mathcal{M}(f(x))$ implies $f^{-1}(B) \in \mathcal{N}(x)$ for all $x \in X$.
(iv) $f(\mathsf{cl}(A)) \subseteq \mathsf{cl}(f(A))$ for all $A \in \mathcal{P}(X)$.

It is not hard to verify that the GP-map $f : (X, \mathsf{cl}) \to (Y, C_\infty)$, with the phenotypic closure $C_\infty$ defined in equ.(7), is continuous. In fact, $C_\infty$ is the finest closure on $Y$ for which $f$ is everywhere continuous. More restrictive definitions of accessibility thus imply the existence of discontinuities.

An evolutionary trajectory is a map from the time axis into phenotype space. Thus an evolutionary trajectories is simply (complete) fossil record. An example is given in Fig. 4. More precisely, however, we have a time series $x(t)$ of genotypes and a resulting time series $f(x(t))$ of phenotypes. Since accessibility on genotype space is by construction defined by the genetic operators we expect $x(t)$ to be continuous. If the GP-map $f$ is continuous, then $f(x(t))$ is also continuous because the concatenation of

continuous functions is again a continuous function. If the GP-map $f$ is not continuous everywhere, however, we still might observe continuous trajectories.

Recall that in the case of mutation we have to consider finite pretopological spaces, i.e., graphs. It is shown in [4] that a transition between subsequent phenotypes is continuous iff it follows an arrow in the graph representation of phenotype space. The continuous transitions in the Fontana-Schuster pretopology for RNA secondary structures are given in Fig. 3.

As the example in Fig. 4 shows, simulated trajectories of RNA evolution exhibit both continuous and discontinuous transitions (see also the much more elaborate simulations in [3, 30]. The discontinuous transitions correspond to larger structural changes which can be interpreted as one kind of innovation. Note that even with the simple dynamics of the adaptive walk, in Fig. 4 we obtain the familiar picture of *punctuated equilibria* as a result of rapid, fitness-driven adaptation that interrupts periods of phenotypic stasis during which the genotype performs a random walk on the neutral network of the dominating phenotype. This reproduces the behavior of much more sophisticated population-based simulations, see e.g. [53].

## 6. Product Spaces and Characters

The basis for the development of a character concept based on the GP-map in [4, 6] is Lewontin's idea of quasi-independence [54]. *Variational characters* [6] are identified with factors or dimensions of a region of the phenotype space. Before discussion some of the implications of this approach we briefly outline its mathematical foundations. Details can be found in [4] for the special case of pretopological spaces and in [6] in the more general setting.

Products of generalized topological spaces play the crucial role in our discussion. Consider two isotonic closure space $(X_1, c_1)$ and $(X_2, c_2)$ with the associated neighborhood function $\mathcal{N}_1$ and $\mathcal{N}_2$. Their *product* $(X_1 \times X_2, c_1 \times c_2)$ of these to spaces consists of the set of all pairs $(x_1, x_2)$, $x_i \in X_i$, with the neighborhood of $(x_1, x_2)$ defined in the following way:

$$N \in \mathcal{N}(x_1, x_2) \iff \exists N_1 \in \mathcal{N}_1(x_1) \text{ and } N_2 \in \mathcal{N}_2(x_2) \text{ such that } N_1 \times N_2 \subseteq N \quad (12)$$

As a simple example of the properties of such a product space we remark that there is a simple formula for the closure of sets of the form $A_1 \times A_2$ [39, Thm.8.1]

$$\mathsf{cl}(A_1 \times A_2) = c_1(A_1) \times c_2(A_2) \quad (13)$$

Furthermore, if both $(X_1, c_1)$ and $(X_2, c_2)$ satisfy the axioms (K2), (K3), or (K4), respectively, then so does their product.

In the case of finite pretopologies, i.e., directed graphs, the topological product defined above coincides with the *strong graph product*. For more information of graph products we refer to the book [55]; an example is shown in Figure 5.

The factors in a product space are a generalization of coordinate axes. In order to define characters in a given phenotype space $(X, \mathsf{cl})$ we have to ask if we can represent the points by means of a set of "coordinates". The values of these "coordinates" then could be interpreted as the *states* of different characters which are defined by the

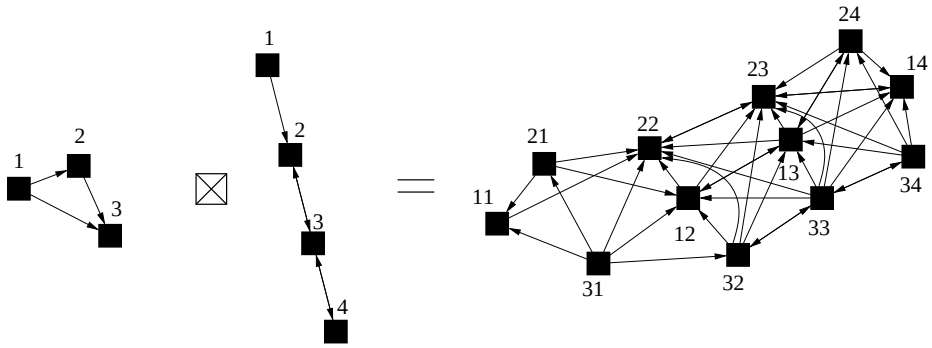



**Figure 5.** Graph Factorization

"coordinate axes". (One should keep in mind here that our coordinates, axes, and values by no means have to be numerical.)

In mathematical term, we ask whether $(X, \mathsf{cl})$ can be written as a non-trivial product of two or more generalized topological spaces. In [4, 6] a criterion for factorizability in terms of the so-called *rectangle condition* is derived.

It was argued already by [4] that it is unlikely that the space of all possible phenotypes will be factorizable as a whole. A local theory of factorization was therefore developed [6].

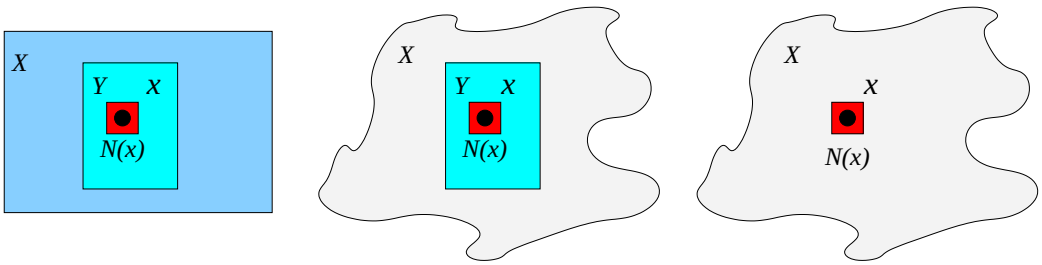



**Figure 6.** Global, regional, and local factorizations.

The crucial observation is that a factorization $X = X_1 \times X_2$ can projected down to any rectangular subspace $Y_1 \times Y_2$ with $Y_i \subseteq X_i$, $i = 1, 2$. In particular, we know from the rectangle condition that the neighborhood system of each point has a basis of rectangular neighborhoods. Generalizing this result we *define* a regional factorization on a set $Y \subseteq X$ as a factorization of the subset $Y$. In particular, the closure space is *locally factorizable* at a point $x$ if each neighborhood $N \in \mathcal{N}(x)$ contains a neighborhood $N' \subseteq N$ that is factorizable.

The existence of a regional factorization allows us to identify which local coordinates (characters) at different points correspond to each other. More formally, two points $x, y \in X$ have *consistent local factorizations* if there is a subset $Y \subseteq X$ such that $x, y \in \mathsf{int}(Y)$ and the subspace $Y \subset X$ has a factorization $Y \simeq \prod_j Y_k$. The restrictions of the factors $Y_k$ to a neighborhood of $x$ and $y$, respectively, defines corresponding local coordinates.

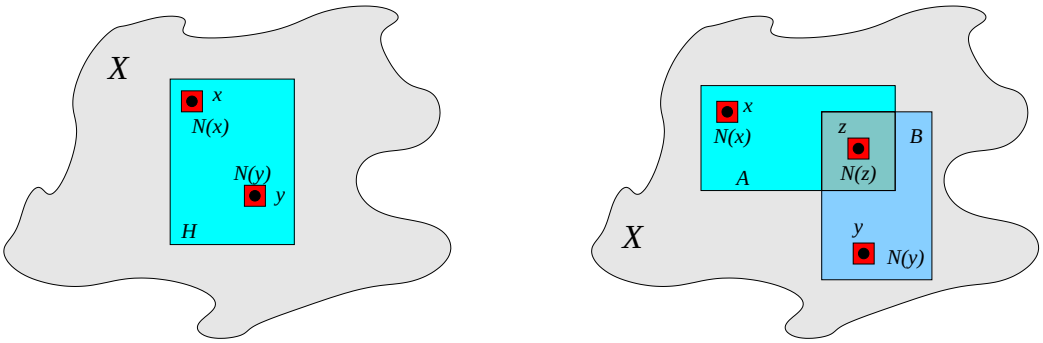



**Figure 7.** L.h.s.: A regional factorization defines local factorizations at its interior points $x$ and $y$.
R.h.s.: Two overlapping regional factorization define a correspondence of the local factors between $x$ and $z$, and $y$ and $z$, respectively, which allows us to identify corresponding factors between $x$ and $y$.

Regional factorizations of two connected sets $Y$ and $Z$ that overlap in a subset $U = Y \cap Z$ with non-empty interior $\mathsf{int}(U)$ define a regional factorization on $U$ that is consistent with both the factorizations of $Y$ and $Z$, Fig. 7. Hence a correspondence between factors of different regional factorizations can be "mediated" via a common region in phenotype space. That allows us to speak of the identity of characters (correspondence of factors) at two points $x$ and $y$ even if there is no regional factorization of set $Y$ that contains both $x$ and $y$ in its interior. However, we need a sequence $Y_1, \ldots, Y_n$ of factorizable regions with non-empty $\mathsf{int}(Y_{k-1} \cap Y_k)$ and factors that correspond with each other across these overlaps.

As argued in [6] this implies a topological notion of *homology* in terms of corresponding local factors. Since the factorizable regions must be connected, it follows that (topologically) homologous characters appear in connected subspaces of phenotype space. This is at least consistent with the *historical homology concept* which requires homologous characters to be related by a common ancestor. In [6] we argue that evolution will prefer trajectories that preserve a given factorization (decomposition into characters). Characters, i.e., local factors should promote evolvability because independent variability also allows independent adaptation locally. Areas in phenotype space that cannot be decomposed in characters should therefore exhibit reduced evolvability, and hence stir evolutionary trajectories back into the factorizable parts. Therefore, we should expect at least an approximate correspondence of the historical and the topological character concepts.

The topological character concepts suggests a notion of innovation at the character level, namely the introduction of new characters by means of splitting and possibly "recombining" existing local factors into new ones. This notion has to be distinguished from the notion of innovation as a large change of the phenotype.

## 7. Concluding Remarks

The topological theory of GP-map that we have reviewed here is based on very few assumptions:

(i) Genetic operators define a generalized topological structure on the space of genotypes.
(ii) The unfolding of the phenotype can be captured by a single genotype-phenotype map, given a constant environment.
(iii) We have some means of determining whether two phenotypes are the same or different.
(iv) The topological structure of the phenotype structure is determined by the accessibility relation which in turn is determined exclusively by the topological structure of genotype space and the GP-map

In particular, the theory does not require us to specify a particular *representation* of the phenotype. In practice, however, one will have to work with such a representation a given level of resolution; the RNA secondary structures used throughout this paper are just one example, anatomical descriptions of animals may be another one. The use of the topological language for an analysis of actual data, observed or simulated, therefore will require the solution of (at least) one additional theoretical problem which we have not touched so far: *when and in what sense is a representation of phenotypes "consistent" with the topological structure of phenotype space?*.

**Acknowledgments.** The research reviewed in this contribution is joint work with Walter Fontana, Bärbel M.R. Stadler, and Günter P. Wagner which was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* Proj.Nos. P-13887-MOB and P-13565.

## References

[1] Sewall Wright. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.
[2] Sewall Wright. "surfaces" of selective value. *Proc. Nat. Acad. Sci. USA*, 58:165–172, 1967.
[3] W. Fontana and P. Schuster. Continuity in evolution: On the nature of transitions. *Science*, 280:1451–1455, 1998.
[4] Bärbel M. R. Stadler, Peter F. Stadler, Günter Wagner, and Walter Fontana. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *J. Theor. Biol.*, 213:241–274, 2001.
[5] Bärbel M. R. Stadler, Peter F. Stadler, Max Shpak, and Günter P. Wagner. Recombination spaces, metrics, and pretopologies. *Z. Phys. Chem.*, 216:217–234, 2002.
[6] Günter Wagner and Peter F. Stadler. Quasi-independence, homology and the unity of type: A topological theory of characters. *J. Theor. Biol.*, 2002. submitted.
[7] Peter Schuster. Evolution *in silico* and *in vitro*: The RNA model. *Biol. Chem.*, 382:1301–1314, 2001.
[8] C. K. Biebricher and W. C. Gardiner. Molecular evolution of RNA *in vitro*. *Biophys. Chem.*, 66:179–192, 1997.
[9] David S. Wilson and Jack W. Szostak. *In Vitro* selection of fuctional nucleic acids. *Annu. Rev. Biochem.*, 68:611–647, 1999.
[10] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46(4):591–621, 1984.
[11] D.H. Mathews, J. Sabina, M. Zucker, and H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[12] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[13] Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 89:177–207, 1998.

[14] Walter Gruener, Robert Giegerich, Dirk Strothmann, Christian Reidys, Jacqueline Weber, Ivo L. Hofacker, Peter F. Stadler, and Peter Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monath. Chem.*, 127:355–374, 1996.

[15] Walter Gruener, Robert Giegerich, Dirk Strothmann, Christian Reidys, Jacqueline Weber, Ivo L. Hofacker, Peter F. Stadler, and Peter 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.

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

[17] Christian Reidys, Peter F. Stadler, and Peter Schuster. Generic properties of combinatory maps. Neutral networks of RNA secondary structure. *Bull. Math. Biol.*, 59:339–397, 1997.

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

[19] Manfred J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Computer-Aided Molec. Design*, 7:473–501, 1993.

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

[21] Aderonke Babajide, Robert Farber, Ivo L. Hofacker, Jeff Inman, Alan S. Lapedes, and Peter F. Stadler. Exploring protein sequence space using knowledge based potentials. *J. Theor. Biol.*, 212:35–46, 2001.

[22] Sergej Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends in Ecology and Evolution*, 12:307–312, 1997.

[23] Erik A. Schultes and David P. Bartel. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science*, 289:448–452, 2000.

[24] Seema Dalal, Suganthi Balasubramanian, and Lynne Regan. Protein alchemy: Changing $\beta$-sheet into $\alpha$-helix. *Nat. Struct. Biol.*, 4(7):548–552, 1997.

[25] A. D. Keefe and J. W. Szostak. Functional proteins from a random-sequence library. *Nature*, 410:715–718, 2001.

[26] Miguel Angel Martinez, Valérie Pezo, Philippe Marlière, and Simon Wain-Hobson. Exploring the functional robustness of an enzyme by *in vitro* evolution. *EMBO J.*, 15:1203–1210, 1996.

[27] B. Derrida and L. Peliti. Evolution in a flat fitness landscape. *Bull. Math. Biol.*, 53:355–382, 1991.

[28] Martijn A. Huynen, Peter F. Stadler, and Walter Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.

[29] Martijn A. Huynen. Exploring phenotype space through neutral evolution. *J. Mol. Evol.*, 43:165–169, 1996.

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

[31] J. Cupal, S. Kopp, and P. F. Stadler. Rna shape space topology. *Artif. Life*, 6:3–23, 2000.

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

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

[34] B. A. Davey and H. A. Priestley. *Introduction to Lattice and Order.* Cambridge Univ. Press, Cambridge UK, 1990.

[35] Mahlon M. Day. Convergence, closure, and neighborhoods. *Duke Math. J.*, 11:181–199, 1944.

[36] E. Čech. *Topological Spaces.* Wiley, London, 1966.

[37] P. C. Hammer. Extended topology: Set-valued set functions. *Nieuw Arch. Wisk. III*, 10:55–77, 1962.
[38] George C. Gastl and Preston C. Hammer. Extended topology. Neighborhoods and convergents. In N.N., editor, *Proceedings of the Colloquium on Convexity 1965*, pages 104–116, Copenhagen, DK, 1967. Københavns Univ. Matematiske Inst.
[39] Stanisław Gniłka. On extended topologies. I: Closure operators. *Ann. Soc. Math. Pol., Ser. I, Commentat. Math.*, 34:81–94, 1994.
[40] F. Hausdorff. Gestufte Räume. *Fund. Math.*, 25:486–502, 1935.
[41] Preton C. Hammer. General topology, symmetry, and convexity. *Trans. Wisconsin Acad. Sci., Arts, Letters*, 44:221–255, 1955.
[42] M. M. Brissaud. Les espaces prétopologiques. *C. R. Acad. Sc. Paris Ser. A*, 280:705–708, 1975.
[43] J. Albuquerque. La notion de "frontiere" en topologie. *Portug. Math.*, 2:280–289, 1941.
[44] Bärbel M. R. Stadler and Peter F. Stadler. Generalized topological spaces in evolutionary theory and combinatorial chemistry. *J. Chem. Inf. Comput. Sci.*, 42:577–585, 2002.
[45] P. Gitchoff and G. P. Wagner. Recombination induced hypergraphs: a new approach to mutation-recombination isomorphism. *Complexity*, 2:37–43, 1996.
[46] P. F. Stadler and G. P. Wagner. The algebraic theory of recombination spaces. *Evol. Comp.*, 5:241–275, 1998.
[47] 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.*, 62:399–428, 2000.
[48] M. Shpak and G. P. Wagner. Asymmetry of configuration space induced by unequal crossover: implications for a mathematical theory of evolutionary innovation. *Artificial Life*, 6:25–43, 2000.
[49] K. Menger. Statistical metrics. *Proc. Natl. Acad. Sci. (USA)*, 28:535–537, 1942.
[50] B. Schweizer and A. Sklar. *Probabilistic Metric Spaces*. North Holland, New York, 1983.
[51] G. Richardson and D. Kent. Probabilistic convergence spaces. *J. Austral. Math. Soc. (Series A)*, 61:1–21, 1996.
[52] Ying-Ming Liu and Mao-Kang Luo. *Fuzzy Topology*, volume 9 of *Advances in Fuzzy Systems: Applications and Theory*. World Scientific, Singapore, 1998.
[53] Peter Schuster. A testable genotype-phenotype map: Modeling evolution of RNA molecules. In Michael Lässig and Angelo Valleriani, editors, *Biological Evolution and Statistical Physics*, volume 585 of *Lecture Notes in Physics*, pages 56–83, Berlin, 2002. Springer-Verla.
[54] R. C. Lewontin. Adaptation. *Sci. Am.*, 239:156–169, 1978.
[55] W. Imrich and S. Klavžar. *Product Graphs: Structure and Recognition*. Wiley, New York, 2000.