R. Giegerich et al. (eds.)
Computer Science and Biology. GCB'99 Hannover, Germany. Univ. Bielefeld 1999, pp. 9-15

TOPOLOGY IN PHENOTYPE SPACE

JAN CUPAL[a,*], PETER SCHUSTER[a,b], PETER F. STADLER[a,b]

[a]Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 4277 52737, fax: 52793, E-Mail:
{jan,pks,studla}@tbi.univie.ac.at
*Address for correspondence
[b]The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA

1. Introduction

The course of biological evolution cannot be understood without the dichotomy of *genotype* and *phenotype*. While variation is introduced by mutation or recombination at the genotypic level, it is the phenotype that is subject to selection. The mapping from genotypes to phenotypes, hence, lies at the heart of any meaningful theory of evolution.

From the mathematical point of view, the genotype-phenotype map Φ is a function from a space of genotypes into a space phenotypes. Understanding the action of Φ hence requires a detailed knowledge about the structure of both spaces involved.

The questions that we have in mind here, concern, for example, the sensitivity of structures against mutations, as addressed by the inverse folding problem: given a structure x, what is the *neutral set* $S(x) = \{s|x = \Phi(s)\}$ of sequences that – under specified conditions – fold into this particular structure? Naturally, one may ask how the neutral set $S(x)$ is embedded in sequence space, and how the neutral sets $S(x)$ and $S(y)$ of two different structures x and y are located relative to each other. A series of publications, starting with [4], answers these questions in great detail for the example of RNA secondary structures [5, 8, 9, 19].

The genotypic part of the sequence-structure map Φ is reasonably well understood. In a molecular biology context, at least, it consists of a set $\mathcal{Q}$ of sequences attributed with a metric distance function defined as the minimal number of "elementary mutations" that are necessary to move from one sequence to the other. The resulting notion of the *sequence space* has become common knowledge since its first appearance in the thirties [25]. Recombination can be taken into account without changing the picture significantly [21].

The topological (and possibly metric) properties of phenotype spaces, on the other hand, are still largely uncharted territory. Previous explorations of the sequence structure map, see below, focus on exactly those features that can be described without reference to shape space beyond a definition of equality of structures. 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, biophysical, or biological similarities [6, 7]. Below we shall use RNA secondary structures as a example to elucidate some features of what we believe is a typical *phenotype space*.

2. RNA Secondary Structures

Biopolymers are a natural starting point for studying genotype-phenotype maps, since genotype and phenotype are here two aspects of the same molecule.

RNA secondary structures provide a discrete 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. Often they are conserved over evolutionary times scales, and *in vitro* selection experiments with RNA do indeed yield families of selected sequences that share distinctive features of secondary structure. RNA secondary structures can be predicted from the sequence information by means of a dynamic programming algorithm [11, 26] that makes use of a set of experimental determined energy parameters [24].

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 [12]. Typically, a secondary structure x is therefore obtained from many different sequences, i.e., the neutral sets $S(x)$ are usually very large. The size-distribution of neutral sets of different structures, however, is far from uniform [19].

The most important feature of the neutral sets of RNA secondary 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 [19, 20] and by means of a random graph model [17, 18]. These features seem to be generic for the sequence structure map, since they have been observed in computational studies not only of nucleic acids but also in polypeptide models, see e.g. [1].

3. Accessibility

The peculiar features of the sequence-structure map cause an evolving population to diffuse on the neutral networks of a particular secondary structure x until a superior phenotype x' is encountered. At this point the population jumps from $S(x)$ to the neutral network to $S(x')$, where the diffusion process resumes [6, 13, 16]. The evolutionary trajectories observed in computer simulations can be regarded as a sequence $(x_0, x_1, \dots)$ of those phenotypes on whose neutral networks the population is concentrated during subsequent diffusion phases. The question hence becomes whether there is a meaningful way of distinguishing between continuous (smooth, expectable), and discontinuous (surprising) evolutionary transitions.

The notion of (evolutionary) accessibility appears to be a suitable starting point for a general mathematical treatment of phenotypic evolution [6]. 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, or, in a more general context, by the genetic operator under consideration and (ii) $\Phi(s) = x$ and $\Phi(s') = y$.

Since sequence space is so large that not all possible sequences are ever realized in the course of simulation run (or during the history of evolution), Fontana & Schuster [6] argue that one should therefore consider more restrictive condition for accessibility. Following their argument, we will say that y is accessible from x, only if the probability is sufficiently high 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

$$D(y \leftarrow x) = S(y) \cap \partial S(x) \tag{1}$$

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. Figure 1 shows that the properties of phenotype space may depend strongly on the choice of the definition of "accessibility".

4. Finite Topologies

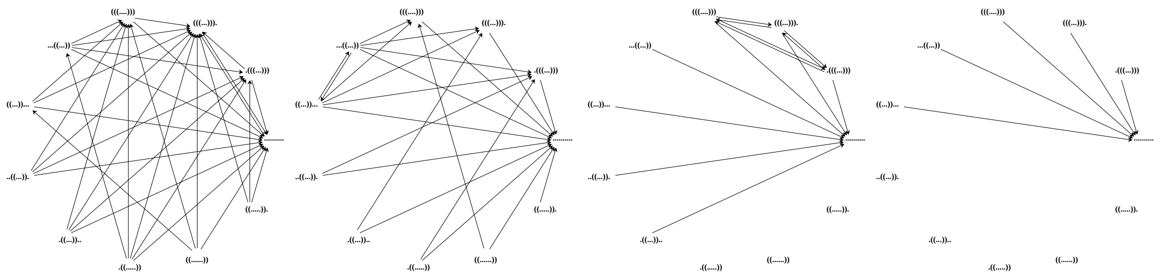
Let $N(x) = \{y \in V | y \leftarrow x\}$ be the set of the structures that are accessible from x (including x itself). Any particular definition of accessibility hence translates into a finite collection $\mathcal{N} = \{N(x)|x \in V\}$ of subsets of the shape space V, such that $x \in N(x)$ for all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

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, and (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 $\mathcal{B}$.

Since in our case V is a finite set, we encounter an extreme simplifications which explains why finite topologies have received very little attention apart from serving as simple counterexamples [22]. 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.

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 $x \rightarrow y$ (from x to y) if and only if $y \in B(x)$. An example is given if Figure 1.

Many of the basic topological properties such as connectedness and the separation axioms can be translated to properties of the basis graph $\vec{\Upsilon}$. Just to give some



$|D(y \leftarrow x)| \geq 5$ $|D(y \leftarrow x)| \geq 20$ $|D(y \leftarrow x)| \geq 50$ $|D(y \leftarrow x)| \geq 100$

Figure 1. Basis $\mathcal{B}$ of the accessibility topologies for RNA sequences of length $n = 10$ as a function of a threshold for the number $|D(y \leftarrow x)|$ that is used to define accessibility. Secondary structures are represented by strings; matching pairs of parenthesis denote base pairs, unpaired nucleotides are denoted by dots.

examples: (V, τ) is connected iff $\vec{\Upsilon}$ is (weakly) connected; the separation axioms are expressed as: $\mathsf{T1}$: $\vec{\Upsilon}$ contains no edges, $\mathsf{T0}$: $\vec{\Upsilon}$ contains no bidirectional edge, i.e., there is no pair of vertices x and y such that $x \to y$ and $y \to x$, $\mathsf{T3}$: $\vec{\Upsilon}$ is a union of disjoint cliques. For more details and proofs see [3].

5. RNA Secondary Structure Topologies

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. 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 $|D(y \leftarrow x)|$, it is straightforward to extract first $N(x)$ for each structure and then the non-redundant basis $\mathcal{B}$.

The most important question, of course, is whether the topologies of larger sequence spaces are 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) can be summarized as follows:
(i) For $n \geq 12$, the separation axiom $\mathsf{T0}$ is satiesfied.
(ii) There is a single giant connected component and a increasing number of very small (less then 10 sequences) components and isolated points.
The isolated points and small components are composed predominately of rare structures. It can be expected therefore that the connected component already encapsulates the relevant features of RNA shape space.

6. Continuity

The distinction between continuous and discontinuous transitions is a long-standing problem in the theory of evolution. Recently, transitions between phenotypes have been investigated in simulations of evolving RNA molecules. It turned out that evolutionary trajectories are usually "continuous" in the sense that they hardly ever involve structural changes that do not lead from one phenotype to an "easily accessible" one. This notion corresponds to the simple structural edit operations depicted in Figure 2. Discontinuous transitions are important, however, for the creation of structural novelty [7, 6].

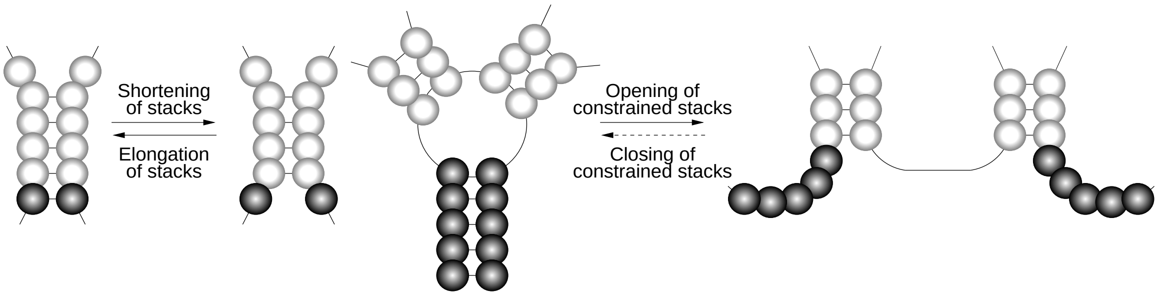


Figure 2. Structural changes corresponding to continuous evolutionary transitions. Shorting and elongation of stacks as well as opening of constrained stacks in general leads to easily accessible structures. Closing a constrained stack, on the other hand, leads to inaccessible structures and hence corresponds to discontinuous transitions.

From a more abstract point of view, *continuity* is a topological property of a map from one topological space into another one. Hence we define an evolutionary trajectory as a function from the time axis $\mathbb{R}$ into the phenotype space (V, τ), where τ is the "natural" topology of shape space that is implied by the accessibility relationships outlined in the previous section. In [3] we show that an evolutionary trajectory is continuous if and only if it follows the directed edges of the basis graph $\vec{\Upsilon}$.

Having defined the topology, it becomes a matter of observation or computer simulation whether "real" evolutionary trajectories are in fact continuous. Conforming previous work [6, 7], we find that most evolutionary transitions are indeed continuous in simulations of RNA evolution based on fitness functions that evaluate RNA structure.

7. Discussion

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 [23].

The genetic representation of a character determines the variability of the phenotype, and hence the *accessibility* of a particular phenotypic variant. The concept of developmental constraints can be understood as the limits of variability of traits caused by their representation or coding in the genome. In other words, developmental constraints are determined by (the lack of) accessibility, and hence by the intrinsic topology of phenotype space. 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. 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.

We note, finally, that the topological notion of continuity, and hence topology, might sometimes be too restrictive. Weaker structures, such as filter spaces or convergence spaces, as introduced for instance in [10, 2, 15, 14], appear to be promising starting points for a generalization of our approach.

Acknowledgements. This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* Proj. P13093-GEN, and by the *European Commission* within the framework of the Biotechnology Program (BIO-4-98-0189).

References

[1] 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.

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

[3] J. Cupal, S. Kopp, and P. F. Stadler. RNA shape space topology. *Alife*, 1999. submitted, SFI Preprint 99-03-022.

[4] 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.

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

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] I. L. Hofacker, P. Schuster, and P. F. Stadler. Combinatorics of RNA secondary structures. *Discr. Appl. Math.*, 88:207–237, 1998. SFI preprint 94-04-026.

[13] 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.

[14] M. Katětov. On continuity structures and spaces of mappings. *Comm. Math. Univ. Carolinae*, 6:257–278, 1965.

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

[16] C. Reidys, C. Forst, and P. Schuster. Replication and mutation on neutral networks. *Bull. Math. Biol.*, 1999. in press, Santa Fe Institute Preprint 98-04-036.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

[25] S. Wright. The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In D. F. Jones, editor, *International Proceedings of the Sixth International Congress on Genetics*, volume 1, pages 356–366, 1932.

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