Geometry and coarse-grained representations of landscapes

Konstantin Klemm, Jing Qin and Peter F. Stadler

1 Introduction

Combinatorial landscape theory provides a framework for the description of the thermodynamics and kinetics of a large class of complex systems. It has been proven to be a valuable concept in evolutionary biology, combinatorial optimization and the physics of disordered systems.

The notion of a "fitness landscape" originated in theoretical biology as a techinque to visualize evolutionary adaption in 1932 [25]. The basic ingredients are a set of discrete genetic structures, a fitness function using to evaluate every possible structure and a "mutation" function measuring the feasibility of transitions between pair of different structures. Due to the combined effects of mutation and selection, a population moves uphill/downhill on the landscape which provides evolutionary information in the form of accessibility or reachability. The intuitive of this theory gives rise to the method "evolutionary algorithm" in computer science for global search or solving combinatorial optimization problems such as the travelling sales-

Konstantin Klemm
Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, 04107 Leipzig, Germany, e-mail: klemm@bioinf.uni-leipzig.de

Jing Qin
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany, e-mail: jingqin@mis.mpg.de

Peter F. Stadler
Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, 04107 Leipzig, Germany.
Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany.
Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1,04103 Leipzig, Germany.
Institute for Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria.
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM87501, USA., e-mail: studla@bioinf.uni-leipzig.de

man problem. The equivalent notion of "energy landscapes" arose in physics as a natural description of disordered systems. In spin glasses, for instance, each spin configuration is assigned an energy describing its Hamiltonian which specifies the model [1]. In theoretical chemistry, energy landscapes are viewed as discretized models to approximate the smooth potential energy surfaces [16]. In structural biology, energy landscapes are used to understand the folding of biopolymers such as RNAs and proteins into their three-dimensional structures [6].

In formal terms, a *(combinatorial) landscape* consists of a *search space* or *configuration space* $\mathbb{X} = (V, \mathscr{T})$ and a fitness or energy function $f : V \to \mathbb{R}$ that evaluates each configuration. In general, $\mathscr{T}$ denotes a (generalized) topological structure on V. In this contribution we will restrict ourselves to the simplest case, namely undirected finite graphs $G = (V,E)$ as search spaces. Similarly, we will assume that the values of f are real numbers. We refer to [8] for some insights into landscapes over recombination spaces and to [20] for landscapes whose values are elements of a partially ordered set. For the sake of clarity we adopt the picture of physics and interpret f and an energy function so that we are interested in particular in configurations with low energy and dynamics that tend to minimize f.

In this contribution we focus on geometric and topological features of landscapes, i.e., on properties that arise from the interplay of the structure of G with the function f. These are of particular interest for an understanding of processes on combinatorial landscapes that are governed by local transitions, including in particular a wide variety of heuristic optimization algorithms from simulated annealing to genetic algorithms. Although the relationship between dynamical processes on combinatorial landscapes and geometric properties of the landscape itself has been a long-standing research problem we still lack a satisfactory theory [18]. Some progress has been made, however, in the analysis of the landscape structure itself. The hierarchical structure of local minina and the barriers between their basins of attraction plays a crucial role in this context.

2 Two examples

Before we proceed, let us briefly introduce two famous examples of combinatorial landscapes:

(A) TSP-landscapes. The travelling salesman problem (TSP) is probably the most frequently studied combinatorial optimization problem. Each potential solution to the TSP is a cyclic permutation among n cities, each city occurs once. The *configuration space* of the TSP landscape consists of all potential solutions of TSP. Each configuration is evaluated by the value of the distance of the total route. Two potentail solutions are *adjacent* in the underlying graph of the TSP landscape if their corresponding permutations can be transformed from each other by exchanging the positions of two cities.

(B) RNA-landscapes. The *RNA landscape* may serve as a prototype for biophysically interesting landscapes. An RNA sequence can be viewed as a string over the

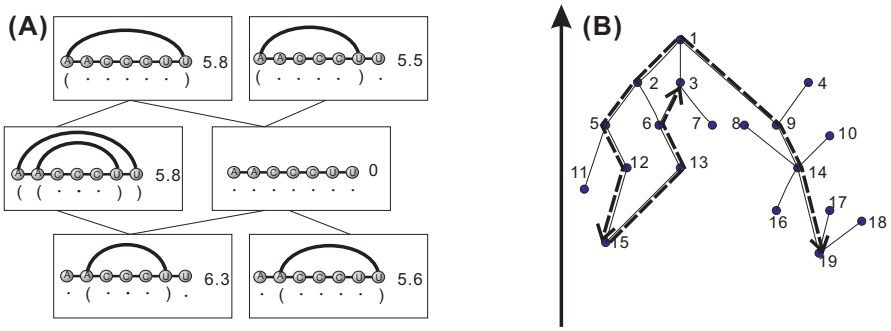

Fig. 1 Energy landscapes. **(A)** The configuration space $G = (V,E)$ of the RNA sequence AACCCUU. consists of six secondary structures. Two structures $x, y \in V$ are adjacent, $\{x,y\} \in E$, if x can be obtained from y by adding or removing a single arc. The folding energy $f(x)$ of a configuration x is displayed in the box next to the structure. **(B)** Three types of walks that can be defined on the landscape: a gradient walk $1 \to 9 \to 14 \to 19$ in which 9 is the unique gradient neighbor of 1, an adaptive walk $1 \to 2 \to 5 \to 12 \to 15$ and a hill climbing walk $15 \to 13 \to 6 \to 3$.

alphabet over four *bases* $\{\mathbf{A},\mathbf{U},\mathbf{G},\mathbf{C}\}$ of length n. Given an RNA sequence s, an RNA secondary structure is identified as a simple graph with vertex set $\{1,\dots,n\}$, whose edge set consists of the edges $\{\{i,i+1\}|1 \le i \le n-1\}$, together with a further collection of edges B_s such that if $\{i,j\}$, $\{k,\ell\} \in P_s$ with $i < j$ and $k < \ell$ then (i) the particular base combinations at pairing position i and j (k and ℓ) must be **AU**, **GU**, or **GC**; (ii) $i = k$ if and only if $j = \ell$; (3) $k \le j$ implies that $i < k < \ell < j$. An edge $\{i,j\}$ contained in P_s is called a *base pairs*. Those vertices not contained in a base pair are called unpaired. Condition (i) implies that each vertex is allowed to belong to at most one base pair. Condition (ii) excludes the formation of crossing base pairs, i.e. *pseudoknots*. For a given RNA sequence, the number of all valid secondary structures grows exponentially with the sequence length n. Its configuration space of the RNA folding landscape consists of all the valid secondary structures. Each secondary structure is called a configuration in the landscape. The energy of a secondary structure is calculated by `RNAeval` in `Vienna Package`[11] based on data from wet-lab experiments. Two configurations x, y are adjacent in the underlying graph, if y can be derived from x by adding or removing a base pair in x, see Fig. 1(B).

3 Local Minima, Walks, and Degeneracy

In this section we introduce the basic notations and concepts. Throughout, we consider a landscape (G,f) on finite undirected graph $G = (V,E)$ with a real-valued energy function $f : V \to \mathbb{R}$. We reserve calligraphic letters for systems of subsets of V and mappings between such systems.

3.1 Neighbours and minima

We write $N(x) = \{y|\{x,y\} \in E\}$ for the graph-theoretical neighborhood of a vertex x in G. A vertex $x \in V$ is a *strict local minimum* of (G,f) if $f(x) < f(y)$ for all $y \in N(x)$. If $f(x) \leq f(y)$, we call x a *weak local minimum*. A vertex x is a *global minimum*, also called *ground state* if $f(x) \leq f(y)$ for all $y \in V$. Since V is finite, a global minimum exists for all landscapes. It is not necessarily unique, however. The set of weak local minima of (G,f) will be denoted by $M(G,f)$.

A vertex $y \in N(x)$ with $f(x) = f(y)$ is a *neutral neighbor* of x. We say that $y \in N(x)$ is a *gradient neighbor* of x if $f(y) = \inf_{z \in N(x)} f(z)$ and $f(z) < f(x)$. Hence x has a gradient neighbor if and only if it is not a weak local minimum. In general, a configuration can have more than one gradient neighbor.

For computational purposes it is often desirable to *define* a unique gradient neighbor for each non-minimal vertex. An *energy sorted list* is a bijective mapping $L : \{1,2,\ldots,|V|\} \to V$ with $f(L_i) \leq f(L_j)$ for all indices i, j with $i < j \leq |V|$. Given L, a configuration $x \in V \setminus M(G,f)$ is assigned the unique gradient neighbor L_i with $i = \min\{j : L_j \in N(x)\}$, being the neighbor of x appearing earliest in the list.

3.2 Walks

An *adaptive* walk in (G,f) is a sequence of configurations $w_1, w_2, \ldots, w_\ell$ such that $\{w_{i-1}, w_i\} \in E$ and $f(w_{i-1}) \geq f(w_i)$ for all $1 < i \leq \ell$. Adaptive walks are often called "hill-climbing walks" in the context of maximization. A *gradient walk* is an adaptive walk $w_1, w_2, \ldots, w_\ell$ such that w_i is a gradient neighbor of w_{i-1} for $1 < i \leq \ell$. A *neutral walk* in (G,f) is an adaptive walk such that w_i is a neutral neighbor of w_{i-1} for $1 < i \leq \ell$.

A path is a walk in which no two vertices are visited twice. In particular, every gradient walk is a path. Furthermore, we note that every walk contains a path that is obtained by removing every part of a walk that leads from a vertex back to itself. Since G is finite, every path is necessarily finite as well.

3.3 Degeneracy

Major technical complications in the analysis of discrete landscapes arise from degeneracy, i.e., the presence of distinct vertices with the same value of f. A landscape (G,f) is *non-degenerate* if $f(x) = f(y)$ implies $x = y$. This condition is too strong for most practical applications since many landscape models have symmetries that lead to degeneracies. For instance, the tours in a TSP can start and end in any city along the way without changing the travel cost.

Denote by $G^f(x)$ the connected component of the induced subgraph $G[\{z \in V|f(z) = f(x)\}]$ that contains x. In the local search literature, $G^f(x)$ is often called

a *plateau* or neutral network [22]. Every neutral walk with starting configuration x is by construction confined to $G^f(x)$. The relation $x \sim_f y$ defined in V by $x \sim_f y \Leftrightarrow y \in G^f(x)$ is an equivalence relation and thus $\Pi = \{G^f(x)|x \in V\}$ is the set of all the equivalence classes in V. Therefore, it forms a partition of V.

A landscape is *locally non-degenerate* or *invertible on edges* if the following three equivalent conditions are satisfied: (i) $G^f(x)$ consists of a single vertex for all $x \in V$; (ii) there are no neutral walks on (G, f); (iii) $f(x) = f(y)$ implies $y \notin N(x)$. Clearly, if (G, f) is non-degenerate, then it is locally non-degenerate also. But the inverse statement is not true.

We note that strict local minima need not exist unless the landscape is locally non-degenerate. In the general case, therefore, we have to work with weak local minima and to accommodate neutral walks.

3.4 Reachability

The concept of adaptive walks implies a simple concept of reachability among the vertices of G: y is reachable from x, $x \rightsquigarrow y$, if there is an adaptive walk (and hence an adaptive path) starting at x that contains y. Naturally, one considers the system of sets

$$\mathscr{C}(x) = \{y \in V | x \rightsquigarrow y\} \tag{1}$$

on (G, f). Transitivity of $\rightsquigarrow$ immediately implies that $\mathscr{C}(y) \subseteq \mathscr{C}(x)$ whenever $x \rightsquigarrow y$, and by construction $x \in \mathscr{C}(x)$. Furthermore, let us consider set-wise reachability

$$\mathscr{C}(W) = \bigcup_{x \in W} \mathscr{C}(x) \tag{2}$$

so that $y \in \mathscr{C}(W)$ if there is a point $z \in \mathscr{C}(W)$ from which y can be reached. As shown in [19], the function $\mathscr{C} : 2^V \to 2^V$ satisfies Kuratowski's closure axioms and hence defines the "reachability topology" on V.

4 Basins and saddles

For each weak local minimum $x \in M(G, f)$ we define the gradient basin $\mathscr{G}(x)$ as the set of configurations $z \in V$ so that the unique gradient walk with starting point in z ends in x. We note for later reference that $\{\mathscr{G}(x)|x \in M(G, f)\}$ forms a partition of V. Analogously, we define the adaptive basin $\mathscr{A}(x) = \{z \in V | z \rightsquigarrow x\}$ for all local minima $x \in M(G, f)$. In contrast to the gradient basins, the adaptive basins form a covering of V that in general will not be a partition. By construction we have $x \in \mathscr{G}(x) \subseteq \mathscr{A}(x)$.

For $x, y \in M(G, f)$ and $s \in V$, we say that s is a *gradient saddle* between x and y if (i) there are neighbours $s' \in N(s) \cap \mathscr{G}(x)$ and $s'' \in N(s) \cap \mathscr{G}(y)$ with $f(s'), f(s'') \leq$

$f(s)$; and (ii) s is a configuration with minimal energy fulfilling property (i). In this case, we set the gradient saddle height $\mathsf{GS}(x,y) = f(s)$. We set $\mathsf{GS}(x,y) = \infty$ if x and y are not connected by a gradient saddle. This somewhat complicated definition of gradient saddle is the one we gave in the funnels paper so I decided to reproduce it here for consistency.

A *direct saddle* is defined analogously involving adaptive basins. We say that $s \in V$ is a direct saddle point between x and y if s is an element of $\mathscr{A}(x) \cap \mathscr{A}(y)$ with minimal energy. For any two local minima we define the *direct saddle height* $\mathsf{DS}(x,y) = f(s)$ if a direct saddle s between x and y exists. Otherwise we set $\mathsf{DS}(x,y) = \infty$.

If s is a direct saddle point between two local minima, it is also a gradient saddle for some, but not necessarily the same two local minima. In general we have $\mathsf{DS}(x,y) \leq \mathsf{GS}(x,y)$.

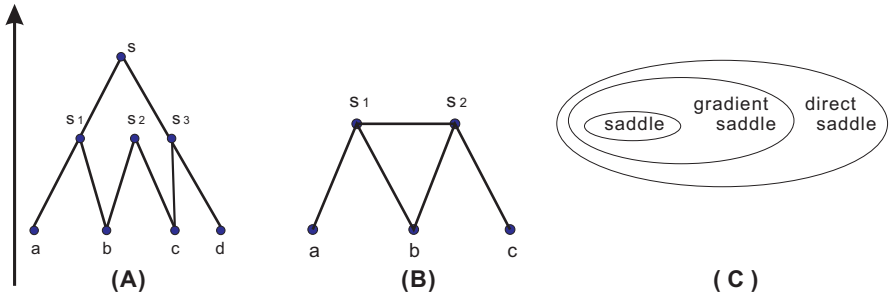

Fig. 2 **(A)** Saddles and direct saddles. Given a landscape in which the configuration space consists of $\{A,B,C,D,S_1,S_2,S_3,S\}$, we have $\mathsf{DS}[A,D] = f(S) > \mathsf{S}[A,D] = f(S_1)$. Therefore, S and $\{S_1,S_2,S_3\}$ is the direct saddle and the equivalent class of saddles between A and D, respectively. Furthermore, there exists relation between saddle height and direct saddle heights given by $\mathsf{S}[A,D] = \mathbf{min}\{\mathbf{max}\{\mathsf{DS}[A,B],\mathsf{DS}[B,C],\mathsf{DS}[C,D]\},\mathsf{DS}[A,D]\}$. **(B)** Direct saddles and gradient saddles. The configurations S_1 and S_2 are the direct saddles between A and C, but there does not exist any gradient saddle between A and C. **(C)** A set diagram of the sets of saddles, direct saddles and gradient saddles of a given landscape. Jing, please use the same symbols in the figure and the caption. I suggest to lower-case letters for all configurations.

Would it not be easier to first define the $\leftrightarrow\!\!\!\!\!\!\rightarrow$ relation and then use it for introducing saddle height?

The existence of a direct saddle point s between two local minima x and y implies that there is a path $\wp$ in G from x to y so that $f(v) \leq f(s)$ for all $v \in \wp$. This not necessarily the smallest bound on the "peak" of the path, however. Denote by $P_{x,y}$ the set of all possible walks between x and y in G. The *saddle height* between any two vertices is

$$\mathsf{S}(x,y) = \min_{\wp \in P_{x,y}} \max_{z \in \wp} f(z) \tag{3}$$

A configuration $s \in V$ is a *saddle point* between two distinct local minima $x,y \in M$ if (i) $f(s) = \mathsf{S}(x,y)$ and (ii) there is a path $\wp \in P_{x,y}$ so that $f(s) \geq f(z)$ for all $z \in \wp$. In contrast to gradient saddle points, thus, one can always find a saddle point since G is

assumed to be connected. In the degenerate case, it is common that the saddle point for two given local minima x, y is not unique. In this case, there exists an equivalence relation $\sim_{\mathsf{S}}$ between saddle points defined by Where is this used?

$$s_1 \sim_{\mathsf{S}} s_2 \Leftrightarrow \exists x, y \in G,\ s_1 \text{ and } s_2 \text{ are saddles between } x \text{ and } y. \tag{4}$$

It is well known that S is an ultrametric distance measure, i.e., $\mathsf{S}(x,y) \leq \max\{\mathsf{S}(x,z), \mathsf{S}(y,z)\}$ for all z [17]. Obviously, we have $\mathsf{S}(x,y) \leq \mathsf{DS}(x,y)$. We illustrate the differences between direct saddles, saddles, and gradient saddles in Fig. (2). We remark that computing saddle heights and saddle points is a difficult task in general. For landscapes of RNA secondary structure, for instance, the problem is NP-hard [15].

We say that x and y are *mutually accessible at level* h, in symbols $x \leftrightarrow\!\!\!\!\rho\,\underline{\ }^{h}\,\!\!\!\!\rightarrow y$, if there is walk $\wp \in P_{x,y}$ such that $f(z) \leq h$ for all $z \in \wp$. By construction, we have $x \leftrightarrow\!\!\!\!\rho\,\underline{\ }^{h}\,\!\!\!\!\rightarrow y$ iff and only if $h \geq \mathsf{S}(x,y)$. It is convenient to define the path connected sets

$$\mathscr{B}_h(x) = \{y \in V | x \leftrightarrow\!\!\!\!\rho\,\underline{\ }^{h}\,\!\!\!\!\rightarrow y\}\,. \tag{5}$$

for $h \geq f(x)$. Trivially, $x \rightsquigarrow y$ implies $x \leftrightarrow\!\!\!\!\rho\,\underline{\ }^{h}\,\!\!\!\!\rightarrow y$ for all $h \geq f(x)$. Thus we have $C(y) \subseteq \mathscr{B}_h(x)$ for all $x \in V$, $y \in \mathscr{B}_h(x)$, and $h \geq f(x)$. Thus $\mathscr{B}_h(x)$ is a closed set in the reachability topology.

The connection between direct saddles and saddles is elucidated in more detail by the following result. Given a path $P = (v_0, v_1, \ldots, v_\ell, v_{\ell+1}) \in G$, if $v_k > v_{k+1} = \cdots = v_{l-1} < v_l$, then all the configurations $w_j \in L$ for $k+1 \leq j \leq l-1$ are called *valley points*. Analogously, *peak points* are the configurations w_j with $k+1 \leq j \leq l-1$ if $w_k < w_{k+1} = \cdots = w_{l-1} > w_l$. A path $\wp = (x = w_0, w_1, \ldots, w_\ell, w_{\ell+1} = y) \in P(x,y)$ is a zig-zag path on (G, f) if

1. $\max_i f(w_i) = \mathsf{S}(x,y)$
2. If $w_k > w_{k+1} = \cdots = w_{l-1} < w_l$ then there is a minimal shelf L such that $w_j \in L$ for $k+1 \leq j \leq l-1$.
3. If $w_k < w_{k+1} = \cdots = w_{l-1} > w_l$ then each w_j with $k+1 \leq j \leq l-1$ is a direct saddle separating the nearest valley points that the path $\wp$ passed before and after w_j .

(Please reformulate without using shelves or move after the definition of shelves in the section on cvfs).

Theorem 1. *If $x, y \in V$ are two configurations so that neither $x \rightsquigarrow y$ nor $y \rightsquigarrow x$ then there is a zig-zag path connecting x and y.*

Proof. By construction, $x \leftrightarrow\!\!\!\!\rho\,\underline{\ }^{\mathsf{S}(x,y)}\,\!\!\!\!\rightarrow y$, hence there is a path $\wp$ from x to y whose height does not exceed $\mathsf{S}(x,y)$. Consider the graph $G^* = G/\sim_f$ derived from G by contracting any $G^f(x) \in \Pi$ into a vertex in G^*. In the meanwhile, we obtain a path $\wp^*$ in G^* from $\wp$ accordingly.

To prove the theorem, all we need is to first construct a "zig-zag" path $P^* \in G^*$ from $\wp^*$ and then prove the existence of a "zig-zag" path $P \in G$ such that P^* is the resulted graph of P after the contraction. The latter is trivial since by construction, $G^f(x)$ is connected for any $x \in G$. Therefore the proof reduces to the construction of

$P^* \in G^*$ from $\wp^*$. This construction is desribed as follows and illustrated in Fig. 3. Let $\{v_i\}_{i=1}^t$ denote the valley points in $\wp^*$. From each valley point v_i, a gradient walk is simulated to reach some local minimum ℓ_i. With out loss of generality, we set $v_0 = \ell_0 = x$, $v_{t+1} = \ell_{t+1} = y$ and assume that all ℓ_i are different configurations. In this context, we observe that there exists a pair of hill-climbing walks from "adjacent" local minima ℓ_i and ℓ_{i+1} to some peak point of $\wp^*$, denoted by p_i. By definition, $f(p_i) \geq \mathsf{DS}[\ell_i, \ell_{i+1}]$. Depend on whether they are equivalent or not, there are two cases. In case of $f(p_i) = \mathsf{DS}[\ell_i, \ell_{i+1}]$, then we just substitute the pair of sections $([v_i, p_i], [p_i, v_{i+1}])$ in $\wp^*$ into the pair of hill-climbing walks from ℓ_i and ℓ_{i+1} to p_i, respectively. Otherwise, by definition, there must exist a configuration d_i such that $f(d_i) = \mathsf{DS}[\ell_i, \ell_{i+1}] < f(p_i)$. In this case, we substitute the pair of sections $([v_i, p_i], [p_i, v_{i+1}])$ in $\wp^*$ into the pair of hill-climbing walks from ℓ_i and ℓ_{i+1} to d_i, respectively.

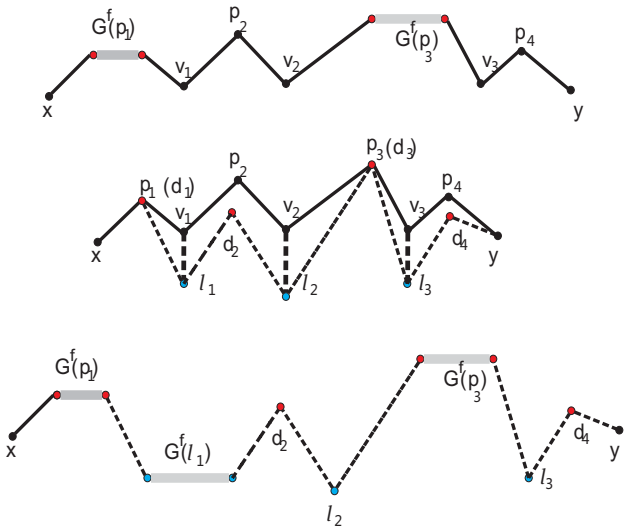

Fig. 3 An example to illustrate the construction $(\wp \to \wp^* \to P^* \to P)$ in the proof of Theorem 1. In which, bold lines in grey denote the path in $G^f(z)$, $z \in \{p_1, \ell_1, p_3\}$.

For each saddle point s, the *basin* below s [7] is the set $\mathscr{B}(s) := \mathscr{B}_{f(s)}(s)$ of configurations that can be reached from s by a path along which the energy of the configurations on the path never exceeds $f(s)$. An obvious connection between basins below saddle points and adaptive basins is the following:

$$\mathscr{B}(s) \subseteq \bigcup_{x \in \mathscr{B}(s) \cap M(G,f)} \mathscr{A}(x) \tag{6}$$

The analogous result for gradient walks holds only in non-degenerate landscapes.

It is not hard to verify that for any two saddles s' and s'' either $\mathscr{B}(s') \subseteq \mathscr{B}(s'')$, $\mathscr{B}(s'') \subseteq \mathscr{B}(s')$, or $\mathscr{B}(s'') \cap \mathscr{B}(s') = \emptyset$ is satisfied, i.e., the basins of a landscape give rise to a hierarchical structure [7]. This hierarchical structure is naturally represented by the *barrier tree* [24] whose leafs are the local minima and whose interior vertices are the saddles.

5 Barrier Trees

Methods to elucidate the basin structure of landscapes by means of barrier trees have been developed independently in different contexts. For instance, potential energy surfaces for protein folding [23, 10], molecular clusters [3] and the kinetics of RNA landscapes [7].

The barrier tree can by computed by the program *barriers* [7] via a flooding algorithm more refs on flooding algorithms. The program *barriers* takes an energy sorted list of the K configurations as input. This list may contain either all configurations or only the configurations below some threshold energy. The only part of *barriers* that relys on the geometric properties of the configuration space is the routine that generates all neighbors of each configuration in the list. Therefore, `barriers` with a time complexity of $O(\Delta \times K)$, where Δ denotes the maximum number of neighbors for a configuration in the landscape. To be precise, the program *barriers* proceeds each configuration on the list consecutively. First identifies local minima if it reads one. Each configuration x is then labeled by the lowest minimum L such that $x \in \mathscr{B}^a(L)$. In the meanwhile, any configurations that with the same energy is decomposed into connected components. The program then checks for each component whether it is a component of saddle points. For each adaptive basin $\mathscr{B}^a(L)$ appears in the barrier tree, the program also records the number of configurations in this basin and the partition function over them etc. .

6 Funnels

The presence of a large number of non-global local minima poses a difficulty for optimization, i.e., identifying movement towards global optima based on purely local information about the landscape. Several measures quantify this difficulty [12] also termed the *ruggedness* of a landscape. Despite being rugged, natural folding landscapes of biopolymeres (cf. the example in Section 2) allow for fast folding, i.e. a Markov chain quickly hits the global minimum after a relatively short time. The picture of a *funnel* [2] has been used for an arrangement of local minima and saddles that guide dynamics towards the optimum. In the following we present a rigorous definition of a funnel as the set of configurations that reach the global minimum by iterating exits from gradient basins over the lowest gradient saddle.

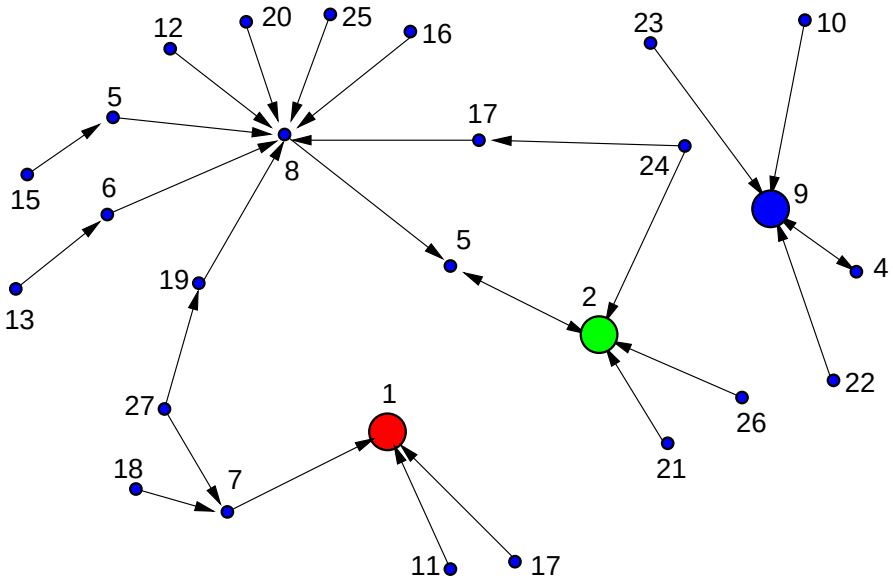

Fig. 4 Funnel digraph for the folding landscape of the RNA sequence xbix (CUGCGGCUUUGGCUCUAGCC). The funnel of the landscape contains the local minima $F \cap M = \{1,7,11,17,18,27\}$, node 1 being the unique ground state. In the funnel partition of the landscape, the set containing node 2 is the largest. This is consistent with the observation that a large part of the folding trajectories reach the node 2 whose energy lies 0.8 kcal above the energy of the ground state [24]. Figure reproduced from ref. [13].

First let us define the *gradient saddle network* of a landscape as a graph $(M(G,f),A)$ having as vertices the weak local minima of the landscape (G,f). Two local minima $x,y \in M$, $x \neq y$ form an edge $\{x,y\} \in A$ if there is a gradient saddle separating x and y. The gradient saddle network is also called *inherent structure network* [4].

The *funnel digraph* (M,B) includes arcs as transitions over the lowest gradient saddles. A pair of local minima $(x,y) \in M^2$ is contained in the arc set B if and only if $\{x,y\} \in A$ and

$$\mathsf{DS}(x,y) = \min\{\mathsf{DS}(x,z) : z \in M, \{x,z\} \in A\} \; . \tag{7}$$

The *funnel* of the landscape is a set $F \subseteq V$ containing those weak local minima, from which there is a directed path to a global minimum. Equivalently, the funnel may be defined recursively [13] as follows.

(i) F contains all global minima.
(ii)A local minimum $x \in M$ belongs to the funnel F if the funnel digraph contains an edge $\{x,y\} \in B$ with $y \in F$.

Thus $F \cap M$ is the maximal subset of M containing all global minima and fulfilling for all $x,y \in M$:

$$x \notin F \wedge (x,y) \in B \Rightarrow y \notin F \tag{8}$$

This determines which local minima belong to F. The funnel is completed by including in F all nodes in the gradient basins of these minima.

In practice, the funnel digraph may be obtained by Algorithm 1. Analogous to `barriers`, it scans the landscape from low to high energy. Each node x is assigned the local minimum $c(x)$ reached from x by a gradient walk, where $c(x) = x$ in the case that x is a local minimum itself. The set C contains all local minima reachable by gradient walks from the neighbours N^- of x that have occurred previously in the loop. If C contains more than one element, x may be a saddle point between pairs of local minima. It is a lowest saddle point (lowest energy exit) for a local minimum $r \in C$, if the current energy $f(x)$ is the lowest energy $h(r)$ at which r appears in C together with other minima. In this case, each pair (r,q) with $q \in C \setminus \{r\}$ is an arc of the funnel digraph.

Algorithm 1 computes the arc set B of the funnel digraph

Require: A landscape (V,E,f) with neighbourhood function N.
Require: An energy sorted list $L : \{1,2,\ldots,|V|\} \to V$.
  $B \leftarrow \emptyset$
  for all $i \in \{1,\ldots,|V|\}$ **do**
    $x \leftarrow L_i$
    $N^- = N(x) \cap \{L_i | i < j\}$
    if $N^- = \emptyset$ **then**
      $c(x) \leftarrow x$ // x is a local minimum
      $h(x) \leftarrow +\infty$
    else
      $j^* = \min\{j < i : L_j \in N^-\}$ // index of gradient neighbour of x
      $c(x) \leftarrow c(L_{j^*})$
      $C \leftarrow \{c(y) : y \in N^-\}$
      if $|C| > 1$ **then**
        for all $r \in C$ **do**
          if $e(r) = +\infty$ **then**
            $h(r) \leftarrow f(x)$ // first appearance of exit from r
          end if
          if $h(r) = f(x)$ **then**
            for all $q \in C \setminus C \setminus \{r\}$ **do**
              $B \leftarrow B \cup \{(r,q)\}$
            end for
          end if
        end for
      end if
    end if
  end for

After finding the funnel F of the landscape, one may be interested in the landscape outside the funnel. Thus the funnel may be removed and the residual landscape analyzed the same way. Iterating this procedure leads to the the *funnel partitioning* of a landscape, being a family $F_1, F_2, \ldots, F_k$. Here $F_1 = F$ is the funnel of the

landscape itself and, for all $2 \leq i \leq k$ F_i is the funnel of the landscape restricted to the subgraph induced by $V \setminus \bigcup_{j=1}^{i-1} F_j$.

The identification of funnels relies on knowledge of the gradient saddle network. For applied studies of real landscape instances, exact computation requires enumeration of all configurations or at least detection of all direct saddles. It is thus restricted to small instances [21]. In larger landscapes, the saddle network may be obtained by efficient sampling methods [14].

7 Combinatorial vector fields on graphs

7.1 basics

Given (G,f) we write $N^>(x) = \{y \in N(x) | f(x) > f(y)\}$ and $N^>[x] = N^>(x) \cup \{x\}$ and call x a *drainage point* if $N^>(x) \neq \emptyset$. Furthermore, we set $N^>(W) = \bigcup_{z \in W} N^>(z)$ for any subset $W \subseteq V$.

For a detailed investigation into the structure of adaptive walks on (G,f) we will need the neutral components together with their downhill neighbors. To this end we define for every subgraph H of G the subgraph $\overrightarrow{H}$ with vertex set $V(\overrightarrow{H}) = \bigcup_{x \in V(H)} N^>[x]$ and edge set $E(\overrightarrow{H}) = E(H) \cup \big\{\{x,y\} \in E | x \in V(H), y \in N^>(x)\big\}$. A particular role is played by the neutral components G^f. We call a graph $\overrightarrow{G^f(x)}$ a *shelf* of (G,f). For every shelf $A = \overrightarrow{G^f(x)}$ of (G,f) we distinguish between the "flat surface" $A^{\text{flat}} = V(\overrightarrow{G^f(x)})$ of the shelf, i.e., the vertices of of $G^f(x)$, and its exit points $A^> = \{y \in N^>(x') | x' \in V(\overrightarrow{G^f(x)})\}$.

Shelves are constructed such that their flat surfaces form a partition of the vertex set of G while their edge sets form a partition of the edge set of G [19]. In locally non-degenerate landscapes, the flat surfaces consist of single points so that each shelf consists of a vertex and its downhill neighbors.

A shelf is *locally minimal* if $A^> = \emptyset$. In this case $A^> \subseteq M(G,f)$, i.e., all vertices of locally minimal shelves are local minima. The converse is not true: shelves with exit points may also contain weak local minima. All strict local minima, are of course, correspond to locally minimal shelves that consist of a single vertex only.

Here we consider only the special case of combinatorial vector fields (cvf) on simple undirected graphs. For the general case we refer to [9].

Definition 1. A *combinatorial vector field* (cvf) on G is a map $\eta : V \to E \cup \{\varnothing\}$ such that, for all $e = \{x,y\} \in E$, $\eta^{-1}(e) \in \{\varnothing, \{x\}, \{y\}\}$.

It is easy to show that CVFs on G are in one-to-one correspondence with the relations $P \subset V \times V$ that satisfies

1. $(x,y) \in P$ implies $\{x,y\} \in E$ (consistency with G)
2. $(x,y) \in P$ and $(x,z) \in P$ implies $y = z$ (uniqueness)

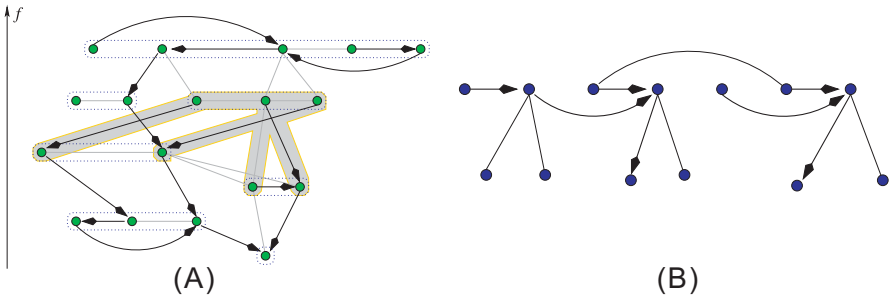

Fig. 5 **(A)** Example of a small highly degenerate landscape. Vertices of G are arranged according to the fitness values $f(x)$. Connected componentes $G^f(x)$ are indicated by dotted boxes. For one of them, the corresponding shelf-graph $\overrightarrow{G^f(x)}$ is highlighted in gray. A combinatorial vector field η (consistent with f) can be visualized as arrows correponding to the set P_η of oriented edges. **(B)** Combinatorial vector fields on a single shelf satisfied condition (A1).

3. $(x,y) \in P$ implies $(y,x) \notin P$ (antisymmetry)

The correspondence is established by $(x,y) \in P$ if and only if $\eta(x) = \{x,y\}$ [19]. We can therefore interpret a cvf as a subset P of directed edges so that each vertex has at most one successor. Note that in contrast to the outgoing arc, the number of incoming arcs is not restricted.

A vertex $x \in V$ is a *rest point* of η if $\eta(x) = \varnothing$, i.e., if x has not successor. The η-trajectory of x is the sequence of v_i, $i \geq 0$, of vertices such $x = v_0$ and $(v_i, v_{i+1}) \in P$. Thus a trajectory either ends in a rest point or it has infinite lengths. In the letter case it contains a finite directed cycle (limit cycle) that is visited infinitely often. The *ω-limit* $\omega_\eta(x)$ of a vertex x is either the (unique) rest point y at which the trajectory starting at x comes to an end, or the limit cycle in which it becomes trapped. Clearly for all $x \in V$, $\omega_\eta(x) \neq \emptyset$ and a vertex y is a rest point if and only if $\omega_\eta(y) = \{y\}$.

The *chain recurrent set* $\mathscr{R}_\eta$ of a combinatorial vector field η on G is defined as

$$\mathscr{R}_\eta = \bigcup_{x \in V} \omega_\eta(x) \tag{9}$$

i.e., it consists of the rest points and limit cycles.

Definition 2. Let η be a combinatorial vector field on G. A function $f : V \cup E \to \mathbb{R}$ is a *Lyapunov function* for η if

1. $f(v) \geq f(e) > f(v')$ if $\eta(v) = e$ and $e = \{v, v'\}$ and $v \notin \mathscr{R}_\eta$.
2. $f(v) = f(\eta(v)) = f(v')$ if $v \neq v'$ and v' and v are contained in a cycle.

The basic idea of [19] is now to study adaptive walks in terms of combinatorial vector fields on G for which the prescribed energy function f is a Lyapunov function. For convenience, we also consider the weaker condition

- (A1) $\eta(x) = \{x,y\}$ implies $f(x) \geq f(y)$.

and call a function of satisfying (A1) a weak Lyapunov function for η.

The crucial observation is that the ensembl of cvfs that have f as weak Lyaponov function together describe in a meaningful way all possible adaptive paths on (G,f): every adaptive path alternates between strictly downhill steps that take it from one shelf to the next and neutral paths along which it traverses a shelf from its entry point to an exit point.

From a technical point of view, the crucial result is that the set of all combinatorial vector fields on G has a product structure: it can be written as the set product of the sets of combinatorial vector fields on the individual shelves. Since consistency of a cvf with a given (weak) Lyapunov function also boils down to conditions that only refer to the individual shelves separately (see [19] for the technical details). It follows that the set combinatorial vector fields consistent with f is the direct product of the sets of combinatorial vector fields consistent with f on the individual shelves.

The importance of this result is the observation that it is sufficient to understand the admissible combinatorial vector fields on the shelves. In particular, it implies that combinatorial vector fields on locally non-degenerate landscapes are entirely characterized by their behavior on the trivial shelves $\overrightarrow{N^>(x)}$. If there are large shelves, on the other hand, quite complex vector field structures can be consistent with condition (A1) because degenerate fitness functions impose fewer constraints on the combinatorial vector fields. In particular they admit complex recurrent sets within individual sets.

7.2 Partition functions, path probabilities, reachability

Now consider a weight function $\omega : E \to \mathbb{R}$ defined on the edges of G. Since we are considering landscapes, we derive the weight function ω from the landscape (G,f). Interpreting f as a potential energy function, the most natural choice are Boltzmann weights of the form

$$\omega(\{x,y\}) = \exp(\beta|f(x)-f(y)|) \tag{10}$$

These weights increase with the steepness of the landscape along the edge. The "inverse temperature" β tunes our emphasis on steepness: For $\beta = 0$, all valid transitions receive the same weight 1. On the other hand, the steepest edges dominate in each set $N^>(x)$ for $\beta \to \infty$.

A natural choice for the weight of a combinatorial vector field on G is then

$$\omega(\eta) = \prod_{(x,y)\in\eta} \omega(\{x,y\}) \tag{11}$$

In which, $\omega(\{x,y\})$ is defined as in eqn. 10. The *partition function* of all combinatorial vector fields on (G,f)

$$Z = \sum_{\eta} \omega(\eta) \tag{12}$$

and its restriction to combinatorial vector fields that contain a particular transition $(u,w) \in P_\eta$.

$$Z_{(u,w)} = \sum_{\eta:(u,w)\in\eta} \omega(\eta). \tag{13}$$

With Boltzmann weights, equ.(10), Z simply counts the number of distinct combinatorial vector fields in the limit $\beta = 0$. On the other hand, $\omega(\eta)/Z \to 0$ unless η consists of edges of steepest descent only in the limit $\beta \to \infty$.

The weights $\omega(\eta)$ can be written as a product of the weights of restrictions of η to the shelves of G,

$$\omega(\eta) = \prod_{A\in\overrightarrow{\Pi}} \omega(\eta_A) \tag{14}$$

Theorem **??** therefore implies immediately that the partition functions are also products of partition functions restricted to the individual shelves:

$$Z = \sum_\eta \omega(\eta) = \sum_\eta \prod_{A\in\overrightarrow{\Pi}} \omega(\eta_A) = \prod_{A\in\overrightarrow{\Pi}} \sum_{\eta_A} \omega(\eta_A) = \prod_{A\in\overrightarrow{\Pi}} Z_A \tag{15}$$

Similarly, we can evaluate restricted partition functions such as

$$Z_{(u,w)} = \prod_{\substack{A\in\overrightarrow{\Pi}\\ \{u,w\}\notin E(A)}} \sum_{\eta_A} \omega(\eta_A) \times Z'_{(u,w)} = Z'_{(u,w)} \times \prod_{\substack{A\in\overrightarrow{\Pi}\\ \{u,w\}\notin E(A)}} Z_A \tag{16}$$

where $Z'_{(u,w)}$ is evaluated just like equ.(13) restricted to the shelf that contains the edge $\{u,w\}$.

For locally non-degenerate landscapes, these expressions are simplified greatly because each shelf contains only one "top point", say x, and edges of the form $\{x,y\}$ with $y \in N^>(x)$. Thus

$$Z = \prod_{x\in V} Z_x \qquad \text{and} \qquad Z_{(u,w)} = \prod_{x\in V\setminus\{u\}} Z_x \times Z'_{(u,w)} \tag{17}$$

with

$$Z_x = \sum_{y\in N^>(x)} \omega(x,y) \qquad \text{and} \qquad Z'_{(u,w)} = \omega(u,w). \tag{18}$$

In the case of locally degenerate landscapes, on the other hand, the computation of the partition functions for the individual shelves can be quite tedious and complex.

In the spirit of statistical mechanics, we endow the set of all combinatorial vector fields on (G,f) with the discrete probability measure

$$p(\eta) := \omega(\eta)/Z \tag{19}$$

In particular, then, the probability of picking a combinatorial vector field that contains the arc $(u,w) \in P_\eta$ is given by

$$p_{(uw)} = Z_{(u,w)}/Z = \frac{1}{Z_A} \sum_{\substack{\eta \in \mathrm{CVF}_{\mathfrak{X}}(A) \\ (u,w)\in\eta}} \omega(\eta) = Z'_{(u,w)}/Z_A\,, \tag{20}$$

where $A \in \overrightarrow{\Pi}$ is the (unique) shelf that contains the edge $\{u,w\} \in E(A)$. In other words, $p_{(uw)}$ is determined only by the combinatorial vector field on the shelf in which the restriction is defined.

Let us now consider trajectories connecting two vertices x and y. More precisely, we are interested in the probability to draw a combinatorial vector field that contains *an arbitrary* trajectory from x to y. We write $x \rightsquigarrow y$ for the set of all such trajectories in (G,f). Let $x \in A_x \in \overrightarrow{\Pi}$. Then

$$\mathbb{P}\{x \rightsquigarrow y\} := \frac{1}{Z} \sum_{\substack{\eta \\ x\rightsquigarrow y\in\eta}} \omega(\eta) = \frac{1}{Z} \sum_{z\in A_x^>} \sum_{\substack{\eta_{\overrightarrow{\Pi}\setminus A_x} \\ z\rightsquigarrow y}} \sum_{\substack{\eta_{A_x} \\ x\rightsquigarrow z}} \omega(\eta_{\overrightarrow{\Pi}\setminus A_x})\omega(\eta_{A_x}) \tag{21}$$

The partition function Z, on the other hand, can be decomposed in the following way:

$$Z = \sum_{\eta} \omega(\eta) = \sum_{\eta_{\overrightarrow{\Pi}\setminus A_x}} \omega(\eta_{\overrightarrow{\Pi}\setminus A_x}) \sum_{\eta_{A_x}} \omega(\eta_{A_x}) = Z_{\overrightarrow{\Pi}\setminus A_x} Z_{A_x} \tag{22}$$

Substituting this decomposition into equ.(21) yields

$$\mathbb{P}\{x \rightsquigarrow y\} = \sum_{z\in A_x^>} \sum_{\substack{\eta_{\overrightarrow{\Pi}\setminus A_x} \\ z\rightsquigarrow y}} \frac{1}{Z_{\overrightarrow{\Pi}\setminus A_x}} \omega(\eta_{\overrightarrow{\Pi}\setminus A_x}) \times \frac{1}{Z_{A_x}} \sum_{\substack{\eta_{A_x} \\ x\rightsquigarrow z}} \omega(\eta_{A_x}) \tag{23}$$

In order to compute this transition probability explicitly, we first consider paths within a shelf. Let us introduce the notation

$$T_{x\rightsquigarrow z} = \frac{1}{Z_{A_x}} \sum_{\substack{\eta_{A_x} \\ x\rightsquigarrow z}} \omega(\eta_{A_x}) \tag{24}$$

for the probability of a path within the shelf A_x from $x \in A_x^{\mathrm{flat}}$ to $z \in V(A_x)^>$. In other words, we consider paths that start in the "flat" part of the shelf, maybe stay on the flat for a while, and then end with a single downward step.

Before we proceed, we remark that $T_{x\rightsquigarrow z}$ can be computed trivially if the landscape is locally non-degenerate. Indeed, in this case $x \rightsquigarrow z$ can be realized exclusively by the arc $(x,z) \in P_\eta$, and hence

$$T_{x\rightsquigarrow z} = \omega(x,z)/Z_{\{x\}} \qquad\qquad Z_{\{x\}} = \sum_{y\in N^>(x)} \omega(x,y)\,. \tag{25}$$

In general, the situation is more complicated since we may have a nontrivial path in A_x^{flat} to some drainage point, say w, before taking the arc (w,z) to the exit point z. In the following let $D_x = \{u \in V(A_x^{\mathrm{flat}}) | N^>(u) \neq \emptyset\}$ denote the drainage points in A_x^{flat}. We have

$$T_{x\rightsquigarrow z} = \frac{1}{Z_{A_x}} \sum_{u\in D_x} \sum_{\substack{\eta_{A_x^{\text{flat}}} \\ x\rightsquigarrow u}} \sum_{\substack{\eta' \\ \text{on } N^>(u) \\ z\in N^>(u)}} \omega(\eta)\omega(\eta') \tag{26}$$

Introducing

$$T^>_{w\to z} := \frac{1}{Z_{\{w\}}} \sum_{\substack{\eta' \\ \text{on } N^>(w) \\ z\in N^>(w)}} \omega(\eta') = \frac{1}{Z_{\{w\}}}\omega(w,z) \tag{27}$$

we can rewrite equ.(26) in the form

$$T_{x\rightsquigarrow z} = \sum_{w\in D_x} \frac{Z_{\{w\}}}{Z_{A_x}} \sum_{\substack{\eta \\ \text{on } A_x^{\text{flat}} \\ x\rightsquigarrow w}} \omega(\eta) T^>_{w\to z} \tag{28}$$

We finally define the abbreviation

$$T^{\text{flat}}_{x\rightsquigarrow w} := \frac{Z_{\{w\}}}{Z_{A_x}} \sum_{\substack{\eta \\ \text{on } A_x^{\text{flat}} \\ x\rightsquigarrow w}} \omega(\eta) \tag{29}$$

and obtain the $T_{x\rightsquigarrow y}$ with $x \in A_x^{\text{flat}}$ and $y \in A_x^>$ as

$$T_{x\rightsquigarrow y} = \sum_{w\in D_x} T^{\text{flat}}_{x\rightsquigarrow w} T^>_{w\to y} \tag{30}$$

The probability $\tilde{P}(x \rightsquigarrow y)$ of a path that starts in x and terminates in y *such that* the final step is a downward step can be computed recursively because any path of this type consists of disjoint subpaths of the type described by equ.(24). The first subpath runs from the startand point x to some exit point $u \in N^>(V(G^f(x)))$, and continues from there to y.

$$\tilde{P}(x \rightsquigarrow y) = \sum_{u\in N^>(V(G^f(x)))} T_{x\rightsquigarrow u}\tilde{P}(u \rightsquigarrow y) \tag{31}$$

For fixed y, eq.(31) can be evaluated iteratively for all x with increasing fitness values $f(x) > f(y)$ and the following initializations: If $f(x) < f(y)$ then $\tilde{P}(x \rightsquigarrow y) = 0$ because of condition (A1). If $f(x) = f(y)$ then $\tilde{P}(x \rightsquigarrow y) = T^{\text{flat}}_{x\rightsquigarrow w}$ if $G^f(x) = G^f(y)$, and $\tilde{P}(x \rightsquigarrow y) = 0$ otherwise.

An arbitrary path from x to y, finally, is either of the type described by equ.(31), or it enters $G^f(y)$ at a vertex $z \in V(G^f(y))$ and continues within this set until it reaches y. Thus, the probability to reach y from x is

$$\mathbb{P}(x \rightsquigarrow y) = \sum_{z\in V(G^f(y))} \tilde{P}(x \rightsquigarrow z) T^{\text{flat}}_{z\rightsquigarrow y} \tag{32}$$

For completeness, finally, we set $\mathbb{P}(x \rightsquigarrow x) = 1$.

Definition 3. A vertex y is unreachable from x on (G,f) if there is no combinatorial vector field η that contains a trajectory from x to y.

In other words, y is unreachable from x if and only if $\mathbb{P}(x \rightsquigarrow y) = 0$. Note that this notion of "unreachable" is a slightly more precise way of saying "there is no adaptive walk from x to y".

A vertex set W is *mutually reachable* if for all $x,y \in W$ we have $\mathbb{P}(x \rightsquigarrow y) > 0$ and $\mathbb{P}(y \rightsquigarrow x) > 0$. Note that if the landscape is invertible on edges then all sets of mutually reachable points are trivial, consisting of a single vertex.

For each $x \in V$ we define the set

$$C(x) = \{y \in V | \mathbb{P}(x \rightsquigarrow y) > 0\} \tag{33}$$

of vertices reachable from x. By construction, $x \in C(x)$. Furthermore, $y \in C(x)$ implies $C(y) \subseteq C(x)$ because reachability is a transitive relation. It will be convenient to define $C(W) = \bigcup_{x \in W} C(x)$.

These are Kuratowski's axioms for a closure function of V, see e.g. [8]. Thus, C is a closure function that defines a (finite) topology τ_C on V. Clearly, a set W is closed in (V,τ_C) if it consists exactly of all vertices reachable from W. We call τ_C the *reachability topology* of the landscape. We note in passing that it may also be of interest to study in more detail the generalized, not idempotent, closure function defined by reachability on a single shelf.

In the following, we will need a characterization of connected sets.

Lemma 1. *A set W is* connected *in the topological space (V,τ_C) if and only if there is a (finite) sequence $x = x_0, x_1, \ldots, x_l = y$ such that $x_i \in C(x_{i-1})$ or $x_{i-1} \in C(x_i)$, i.e., if and only if $\mathbb{P}(x_{i-1} \rightsquigarrow x_i) > 0$ or $\mathbb{P}(x_i \rightsquigarrow x_{i-1}) > 0$.*

Proof. Recall that, in any topological space, the set $C(\{x\})$ is connected and the union of two intersecting connected sets is also connected. The condition above amounts to the existence of a (finite) chain of connected sets connecting any two points in W. Hence W is connected whenever the condition is satisfied. Conversely, suppose there is no such chain between x and y. Then there is a maximal set $U \subset W$ of points that are connected to x, while $y \notin U$. For every $z \in W \setminus U$, $C(z) \cap U = \emptyset$ and $z \notin C(U)$. Thus $C(W \setminus U) \cap U = \emptyset$ and $C(U) \cap (W \setminus U) = \emptyset$, i.e, W violates the Hausdorff-Lennes condition for connectedness.

7.3 (cvf)-Valleys, basins and direct saddles

In the following we will also need a slightly modified notion of maximality w.r.t. set inclusion. Usually, a set A is maximal for a property $\mathcal{Q}$ if A has propery $\mathcal{Q}$ but $A \cup \{x\}$ does not have property $\mathcal{Q}$ for all $x \notin A$. Here we need to modify this to "$A \cup R_x$ does not have property $\mathcal{Q}$ for all $x \notin A$" where R_x is the set of mutually reachable points.

The topology τ_C provides a useful device to describe the structure of the landscape. A natural notion is that of a "(cvf)-valley":

Definition 4. A *(cvf)-valley* is a *maximal* connected subset $W \subseteq V(G)$ such that all vertices $y \notin W$ are unreachable from every $x \in W$ and W is connected.

If W is a (cvf)-valley, then $x \in W$ implies $C(x) \subseteq W$ since by definition no vertices outside of W are reachable from within W. Therefore, $W = \bigcup_{x \in W} C(x) = C(W)$. The (cvf)-valleys are therefore the maximal closed connected sets w.r.t. the topology defined by C.

Consider a locally minimal shelf $G^f(x)$ and set $L := V(G^f(x))$. Local minimality means $N^>(L) = \emptyset$. Thus L together with the set of all points z for which *every* adaptive walk ends in L forms a (cvf)-valley W_L.

More generally, minimal closed sets correspond to the vertices that are restpoints in all admissible combinatorial vector fields, $C(x) = \{x\}$, or to unions (again over all admissible combinatorial vector fields) of $\sim_\eta$ equivalence classes. By transitivity of reachability, these sets are of the form $\{y \in V | \mathbb{P}(x \rightsquigarrow y > 0)$ and $\mathbb{P}(y \rightsquigarrow x > 0)\} \neq \{x\}$. Every adaptive walk in the landscape necessarily ends in one of these minimal closed sets. We can therefore label every $x \in V$ by the collection $\Xi(x)$ of minimal closed sets that are reachable from x, Fig. 6 **(A)**.

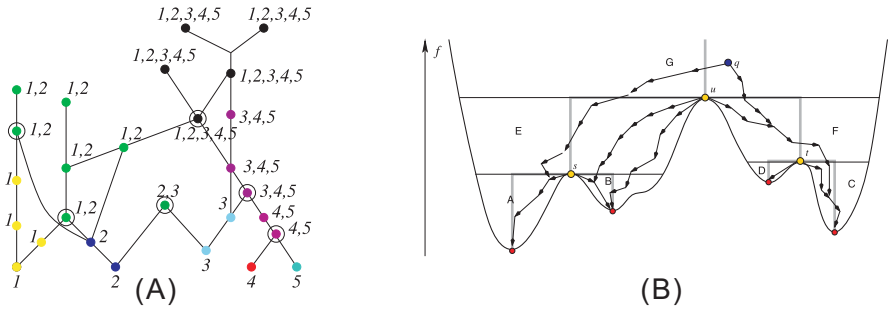

Fig. 6 (cvf)-Valley: **(A)** Each vertex is annotated with the list of reachable local minima. Each (cvf)-valley is characterized by such a list Υ and contains all vertices labeled by a subset if Υ. The minimal closed subsets, here $\{1\}$ through $\{5\}$ are always (cvf)-valleys. In addition, this landscape has the (cvf)-valleys $\{1,2\}$, $\{2,3\}$, $\{4,5\}$, $\{3,4,5\}$, and $\{1,2,3,4,5\}$. (cvf)-Valley connecting points are indicated by circles. **(B)** here are 7 (cvf)-valleys denoted A to G. The minimal (cvf)-valleys are A through D corresponding to the four local minima of the landscape. The (cvf)-valleys A and B are connected by the direct saddle point s. The direct saddle q between A and C has a strictly higher energy than saddle u between the same two (cvf)-valleys (assuming that there is no adaptive walk connecting u with a point in A). It is a saddle between A and D because it is the direct saddle between the (cvf)-valleys E and D and $A \subseteq E$. The barrier tree of the landscape reflects the inclusion relations of the (cvf)-valleys: $A, B \subseteq E$, $C, D \subseteq F$, $E, F \subseteq G$. The (cvf)-valley G corresponds to the entire landscape.

A (cvf)-valley can be identified by the set Υ of minimal closed sets that it contains.

Lemma 2. *A subset $W \subseteq V(G)$ is the (cvf)-valley labeled by Υ if and only if (i) W is connected, (ii) $x \in W$ implies $\Xi(x) \subseteq \Upsilon$, and (iii) $\Xi(x) \subseteq \Upsilon$ implies $x \in W$.*

Proof. Suppose W satisfies (i), (ii), and (iii). We first observe that (ii) implies that W is closed because every vertex y reachable from $x \in W$ satisfies $\Xi(y) \subseteq \Xi(x)$, and hence $y \in W$. To see that W is maximal, we argue as follows: Consider a vertex $z \in V \setminus W$. By (iii), $\Xi(z) \not\subseteq \Upsilon$. If z is contained in a minimal closed set $C \notin \Upsilon$, i.e., $\Xi(z) = \{C\}$, then $W \cup \{z\}$ is not connected because by construction z is not reachable from within W and no vertex in W can be reached from within C. On the other hand, if z is not contained in a minimal closed set C, then there is a minimal closed set $C' \in \Xi(z) \setminus \Upsilon$, and in particular a vertex $z' \in C'$ that is reachable from z. Since $z' \notin W \cup \{z\}$ while $z' \in C(W \cup \{z\})$, we we conclude that $W \cup \{z\}$ is not a closed set. Thus W is a maximal connected closed set.
Now suppose that W is a maximal closed connected set, and set $\Upsilon = \bigcup_{x \in W} \Xi(x)$. Then (ii) is trivially true and W contains in particular all minimal closed sets $C \in \Upsilon$. Now suppose that there is a vertex $z \notin W$ with $\Xi(z) \subseteq \Upsilon$. All adaptive walks emanating from z thus eventually reach W and all vertices y along such a walk satisfy $\Xi(y) \subseteq \Xi(z) \subseteq \Upsilon$. Hence we can expand W by the last mutually reachable subset R_y outside of W, contradicting maximality. Hence $\Xi(z) \subseteq \Upsilon$ implies $z \notin W$.

The (cvf)-valleys of the landscape (G, f) do not form a hierarchical structure. In Fig. 6 **(A)**, the valleys $\{1,2\}$ and $\{2,3\}$ are a counterexample. Nevertheless, the (cvf)-valleys are closely related to the barrier trees of the landscape. In particular, we can identify the (lowest) points that connect (cvf)-valleys with each other.

Definition 5. A vertex $u \in V$ is a *(cvf)-valley connecting vertex* if $\Xi(u)\Xi(v) \neq \emptyset$ for every $v \in C(u) \setminus W_u$, where W_u is the set of vertices that are mutually reachable from u.

In general, there can be multiple, disconnected, (cvf)-valley connecting vertices linking the same two (cvf)-valleys. In Fig. 6 **(A)**, there are two vertices connecting the (cvf)-valleys $\{1\}$ and $\{2\}$, which have different fitness values.

In order to connect our present discussion with earlier work, in particular [5, 7, 24], we briefly discuss the notation of saddle points in the context of our present formalism.

Definition 6. A vertex s is a direct saddle point separating two minimal closed sets W_1 and W_2 if (i) there are points $y_1 \in W_1$ and $y_2 \in W_2$ with $P(s \rightsquigarrow y_1) \neq 0$ and $P(s \rightsquigarrow y_2) \neq 0$, and (ii) there is no vertex s' with $f(s') < f(s)$ that also has property (i).

A direct saddle point is therefore a (cvf)-valley connecting point with minimal fitness connecting two (cvf)-valleys. In [8], basins of a landscape are discussed that are defined in terms of the connected components of $\{x \in V | f(x) < \eta\}$ where η is the fitness of a saddle point. This connects well to our present discussion. The subsets of (cvf)-valleys below a certain fitness threshold are always connected sets. Thus, basins are connected sets of the form

$$\bigcup_{W\in\mathscr{V}} \{x \in W | f(x) < \eta\} \tag{34}$$

constructed from maximal collections of (cvf)-valleys $W \in \mathscr{V}$. Saddle points, i.e., vertices of minimal fitness that connect distinct basins are therefore necessarily (cvf)-valley connecting points between (cvf)-valleys associated with the distinct basins that they merge. Given an arbitrary pair of disjoint (cvf)-valleys, their direct saddle can have a strictly larger value of f than the saddle point connecting the associated basins, Fig. 6 **(B)**.

We remark, finally, that the flooding algorithm implemented in the `barriers` program [5, 7] identifies saddle points as the lowest energy points that have neighbors with lower energy that are connected by means of gradient descent walks to local minima in two distinct valleys. This is equivalent to the existence of two adaptive walks starting at the saddle points that terminate in the same local minima. This flooding algorithm can easily be modified to keep track of the labelling $\Xi(x)$. In the non-degenerate case, $\Xi(x)$ is simply the union of the set $\Xi(y)$ of all neighbors of x that are reachable. In the degenerate case one has to keep track of all neighbors of the set W_x that is mutually reachable from x as described in [7]. This gives rise to the recursion

$$\Xi(x) = \bigcup_{x'\in W_x} \bigcup_{y\in N(x')\cap C(x)} \Xi(y) \tag{35}$$

Valley-connecting points are therefore recognizable in the course of the flooding algorithm as those vertices x for which the union $\Xi(x)$ does not coincide with the label set $\Xi(y)$ of at least one of the downward neighbors y.

References

1. Binder, K., Young, A.P.: Spin glasses: Experimental facts, theoretical concepts, and open questions. Rev. Mod. Phys. **58**, 801–976 (1986)
2. Dill, K.A., Chan, H.S.: From levinthal to pathways to funnels. Nat. Struct. Biol. **4**, 10–19 (1997)
3. Doye, J.P., Miller, M.A., Welsh, D.J.: Evolution of the potential energy surface with size for Lennard-Jones clusters. J. Chem. Phys. **111**, 8417–8429 (1999)
4. Doye, J.P.K.: Network topology of a potential energy landscape: A static scale-free network. Phys. Rev. Lett. **88**, 238,701 (2002)
5. Flamm, C., Fontana, W., Hofacker, I., Schuster, P.: RNA folding kinetics at elementary step resolution. RNA **6**, 325–338 (2000)
6. Flamm, C., Hofacker, I.L.: Beyond energy minimization: Approaches to the kinetic folding of RNA. Chemical Monthly **139**, 447–457 (2008)
7. Flamm, C., Hofacker, I.L., Stadler, P.F., Wolfinger, M.T.: Barrier trees of degenerate landscapes. Z. Phys. Chem. **216**, 155–173 (2002)
8. Flamm, C., Stadler, B.M.R., Stadler, P.F.: Saddles and barrier in landscapes of generalized search operators. In: C.R. Stephens, M. Toussaint, D. Whitley, P.F. Stadler (eds.) Foundations of Genetic Algortithms IX, *Lecture Notes Comp. Sci.*, vol. 4436, pp. 194–212. Springer, Berlin, Heidelberg (2007). 9th International Workshop, FOGA 2007, Mexico City, Mexico, January 8-11, 2007

9. Forman, R.: Combinatorial vector fields and dynamical systems. Math. Z. **228**, 629–681 (1998)
10. Garstecki, P., Hoang, T.X., Cieplak, M.: Energy landscapes, supergraphs, and "folding funnels" in spin systems. Phys. Rev. E **60**, 3219–3226 (1999)
11. Hofacker, I.L.: Vienna rna secondary structure server. Nucl. Acids Res. **31**(13), 3429–3431 (2003)
12. Kallel, L., Naudts, B., Reeves, C.R.: Properties of fitness functions and search landscapes. In: L. Kallel, B. Naudts, A. Rogers (eds.) Theoretical aspects of evolutionary computing, pp. 175–206. Springer-Verlag (2001)
13. Klemm, K., Flamm, C., Stadler, P.F.: Funnels in energy landscapes. Eur. Phys. J. B **63**, 387–391 (2008). ECCS 2007 contribution
14. Mann, M., Klemm, K.: Efficient exploration of discrete energy landscapes. Phys. Rev. E **83**(1), 011,113 (2011)
15. Maňuch, J., Thachuk, C., Stacho, L., Condon, A.: Np-completeness of the energy barrier problem without pseudoknots and temporary arcs. Natural Computing **10**(1), 391–405 (2011)
16. Mezey, P.G.: Potential Energy Hypersurfaces. Elsevier, Amsterdam (1987)
17. Rammal, R., Toulouse, G., Virasoro, M.A.: Ultrametricity for physicists. Rev. Mod. Phys. **58**, 765–788 (1986)
18. Reidys, C.M., Stadler, P.F.: Combinatorial landscapes. SIAM Review **44**, 3–54 (2002). SFI preprint 01-03-14
19. Stadler, B.M.R., Stadler, P.F.: Combinatorial vector fields and the valley structure of fitness landscapes. J. Math. Biol. **61**(6), 877–898 (2010). DOI 10.1007/s00285-010-0326-z. URL http://dx.doi.org/10.1007/s00285-010-0326-z
20. Stadler, P.F., Flamm, C.: Barrier trees on poset-valued landscapes. Genetic Prog. Evolv. Mach. **7-20**, 4 (2003)
21. Tomassini, M., Vérel, S., Ochoa, G.: Complex-network analysis of combinatorial spaces: The NK landscape case. Phys. Rev. E **78**, 066,114 (2008)
22. Van Nimwegen, E., Crutchfield, J.P.: Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? Bull. Math. Biol. **62**, 799–848 (2000)
23. Wales, D.J.: Decoding the energy landscape: extracting structure, dynamics and thermodynamics. Phil. Trans. R. Soc. A **370**(1969), 2877–2899 (2011)
24. Wolfinger, M.T., Svrcek-Seiler, W.A., Flamm, C., Hofacker, I.L., Stadler, P.F.: Exact folding dynamics of RNA secondary structures. J. Phys. A: Math. Gen. **37**, 4731–4741 (2004)
25. Wright, S.: The roles of mutation, inbreeding, crossbreeeding and selection in evolution. In: D.F. Jones (ed.) Proceedings of the Sixth International Congress on Genetics, vol. 1, pp. 356–366. Brooklyn Botanic Gardens, New York (1932)