Combinatorial Vector Fields and the Valley Structure of Fitness Landscapes

Bärbel M. R. Stadler · **Peter F. Stadler**

Abstract Adaptive (downhill) walks are a computationally convenient way of analyzing the geometric structure of fitness landscapes. Their inherently stochastic nature has limited their mathematical analysis, however. Here we develop a framework that interprets adaptive walks as deterministic trajectories in combinatorial vector fields and in return associate these combinatorial vector fields with weights that measure their steepness across the landscape. We show that the combinatorial vector fields and their weights have a product structure that is governed by the neutrality of the landscape. This product structure makes practical computations feasible. The framework presented here also provides an alternative, and mathematically more convenient, way of defining notions of valleys, saddle points, and barriers in landscapes. As an application, we propose a refined approximation for transition rates between macrostates that are associated with the valleys of the landscape.

1 Introduction

Fitness landscapes have played an important role in mathematical biology ever since Sewall Wright's seminal paper [30] as a means of conceptualizing evolutionary adaptation. As a mathematical structure — a finite but typically very large "search space" endowed with a cost or fitness function — landscapes appear naturally in many different contexts and fields of science. They can be seen as a generic formulation of combinatorial optimization problems in computer science and operations research [11], they describe the Hamiltonians of disordered systems in physics [2,19], and they are (discretized versions of) potential

B.M.R. Stadler

Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany. Tel.: ++49 341 9959 532, Fax: ++49 341 9959 658, E-mail: baer@bioinf.uni-leipzig.de

P.F. Stadler

Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany · Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany · Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany · Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria · Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
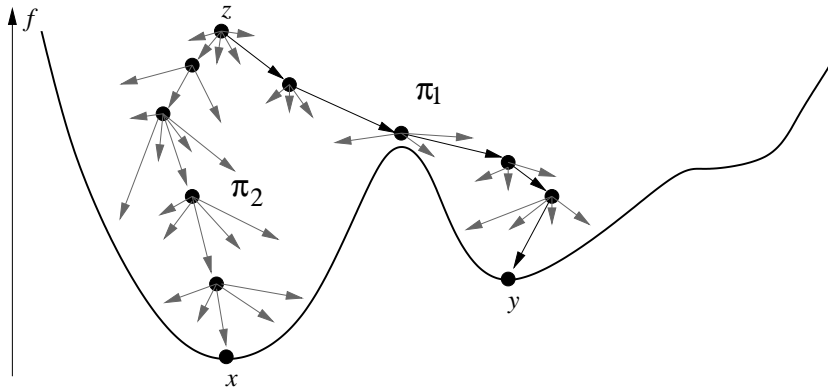
Fig. 1 Basins and membership in basins. A gradient descent walk starting from a vertex z may end up in a local minimum (here y) that is neither the deepest nor the closest accessible one. Different adaptive walks may connect z to different local minima, here x and y.

energy surfaces in theoretical chemistry [20]. In structural biology, energy landscapes are used to understand the folding of biopolymers into their three-dimensional structures, and discrete versions are used to study folding processes at the level of lattice protein models [21] and RNA secondary structures [5].

The relationship between dynamical processes on landscapes and geometric properties of the landscape itself is a long-standing research problem that is still far from a satisfactory solution [25]. In particular, *basins* and the barriers between them have attracted attention from several points of view. The replica approach to disordered spin systems for instance, emphasized the break-up of the state space into many valleys separated by saddle points [19]. Direct approaches to elucidate the basin structure naturally leads to the concept of *barrier trees*. In this picture, local minima are identified with the leaves of the tree, while the interior nodes represent (equivalence classes of) saddle points that separate local minima. This construction was introduced more or less independently in several different application domains, see e.g. [16, 1, 12, 27, 3, 4, 24].

A simple, but much less informative, measure for the size of a basin is the length distribution of the steepest descent (gradient) paths that terminate in a given local minimum [28, 9]. In contrast to gradient paths, an *adaptive walk* accepts steps whenever they improve fitness, instead of insisting on steepest descent. Gillespie [13] suggested to use adaptive walks as models of evolutionary adaptation. They have been studied extensively in NK models [28, 15, 8], in particular in the context of the maturation of the immune response [18, 17, 23], in RNA folding landscapes [9], and in a model of early vascular land plants [22]. The length distribution of adaptive walks appears to be linked to the size distribution of the basins; the details of this connection, however, remain to be elucidated.

In [29], a coarse-grained representation of energy landscapes is introduced in which the barrier trees represent a partition into "basins of attraction" of the local minima. This is achieved by assigning each point to (the basin of) the local minimum that is reached by gradient descent. Given this partition of the search space into "macro states" one can then derive an approximation for the transition rates between macro-states α and γ in the form

$$r_{\alpha,\gamma} = \sum_{x \in \alpha} \sum_{y \in \gamma} r_{xy} \mathsf{Prob}[y|\gamma] \tag{1}$$

where r_{xy} are the transition rates between micro-states and $\text{Prob}[y|\gamma]$ is the conditional probability for the system to be in microstate y given that it is in macrostate γ. In practise, the last term is approximated by the equilibrium distribution within γ [29]. The usefulness of this approximation, however, depends on the "correctness" of the assignment of microstates to macrostates.

As shown in Fig. 1 the use of adaptive or gradient walks may lead to unintuitive or undesired assignments. Adaptive walks, for instance, may connect points to several distinct local minima. Gradient walks, on the other hand, may well connect a point to a local minimum that is neither the closest (in terms of the number of steps) nor the deepest (in terms of the value of f) among the minima that can be reached by adaptive walks. In the case of degenerate landscapes [6], furthermore, gradient descent is not uniquely defined. Hence, a more rational and mathematically sound assignment of states to basins/valleys is highly desirable and may lead to an improved approximation of folding dynamics. Here, we propose to use combinatorial vector fields as a suitable framework for the study of adaptive walks and their interrelationships with the basin structure of a landscape.

Combinatorial vector fields were introduced by Robin Forman [10] to describe dynamical systems on simplicial complexes. A central concept in this theory is that of a *Lyapunov function F*. In analogy with a Lyapunov function in a vector space setting, the values of F decrease along trajectories. The basic idea of this contribution is to interpret adaptive walks as trajectories generated by combinatorial vectors for which the given fitness landscape behaves like a Lyapunov function. Adaptive walks are inherently stochastic objects. Thus we associate a weight with each combinatorial vector field. This weight can then be translated into a probability for the combinatorial vector field. The weights can naturally be chosen such that paths of gradient descent are at least locally more likely than all other trajectories. In this setting it becomes natural to model the probability that vertex z is connected to local minimum x as the probability of a combinatorial vector field containing a trajectory from z to x.

This contribution is organized as follows: In the next section we introduce the necessary notation, provide a convenient representation of combinatorial vector fields on undirected graphs and make precise in what sense the given fitness landscape is a Lyapunov function. We then proceed to show that combinatorial vector fields on landscapes have a natural product structure that can be exploited to derive a simple Markov model of adaptive walks. This in turn implies a canonical way in which microstates are partially assigned to alternative local minima that can be reached from them, and finally leads to a refinement of equ.(1).

2 Definitions and Preliminaries

Let $G(V,E)$ be a connected undirected simple finite graph with vertex set V and edge set E. We write $G[W]$, $W \subseteq V$, for the subgraph of G induced by V, i.e., $\{x,y\} \in E(G[W])$ if and only if $\{x,y\} \in E$ and $x,y \in W$. We write $N(x) = \{y|\{x,y\} \in E\}$ for the neighborhood of $x \in V$ and $N[x] = N(x) \cup \{x\}$ for its closed neighborhood.

Let $f : V \to \mathbb{R}$ be an arbitrary function defined on the vertex set of G. We call $(G,f) = (V,E,f)$ a *fitness landscape* and refer to f as the "fitness function". A vertex x is a strict local minimum of 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. Note that the notion of local optima explicitly depends on the graph structure, i.e., on the edge set E.

2.1 Combinatorial Vector Fields on Graphs

We will not need the complete formalism of combinatorial vector fields on simplicial complexes [10], hence we only provide definitions restricted to simple undirected graphs. For details we refer to the Appendix.

Definition 1 A *combinatorial vector field* (cvf) on G is a map $\eta : V \to E \cup \{\varnothing\}$ such that, for all $e \in E$, $\eta^{-1}(e)$ is either empty or consists of one of the two vertices incident with e.

A vertex $x \in V$ is a *rest point* of η if $\eta(x) = \varnothing$.

An η-*path* (of length r) is a sequence $\gamma = (v_0, e_0, v_1, e_1, ..., e_{r-1}, v_r)$ such that $e_i = \eta(v_i)$ and $\{v_i, v_{i+1}\} = e_i$. Since the edges $e_i \in \gamma$ are uniquely determined by the sequence $(v_0, ..., v_r)$, we may equivalently interpret γ as its sequence of vertices. An η-path is non-trivial if it contains at least two vertices. If $v_0 = v_r$, we say that γ is a closed path. Two vertices or edges $x, y \in V \cup E$ are η-equivalent, $x \sim_\eta y$, if there is a closed η-path γ with $x, y \in \gamma$. Again, it will be sufficient for our purposes to restrict ourselves to vertices.

An η-path $\psi = (x_0 = x, e_1, x_1, e_2, x_2, ...)$, $e_i = \eta(x_{i-1})$, $e_i = \{x_{i-1}, x_i\}$ is the η-*trajectory* starting at x. The associated vertex set $\hat{\psi}(x) = \{x_0, x_1, ...\}$ is the *orbit* of x. Note that $\hat{\psi}(x)$ is always finite.

Definition 2 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 set of vertices that appear infinitely often in the trajectory ψ starting at x.

Definition 3 The *chain recurrent set* $\mathscr{R}_\eta$ of a combinatorial vector field η on G is the set of vertices that are either rest points of η or that are contained in some non-trivial closed η-path.

Clearly, $\omega_\eta(x) \neq \emptyset$ and $\omega_\eta(x) \subseteq \mathscr{R}_\eta$ for all $x \in V$. A vertex is a rest point if and only if $\omega_\eta(x) = \{x\}$. Hence

$$\mathscr{R}_\eta = \bigcup_{x \in V} \omega_\eta(x) \tag{2}$$

Definition 4 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 $v \sim_\eta v'$.

Before we proceed it will be convenient to introduce an alternative representation of combinatorial vector fields on G as a relation on the vertex set $V(G)$.

Definition 5 Let $P \subset V \times V$ be a relation 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)
3. $(x, y) \in P$ implies $(y, x) \notin P$ (antisymmetry)

Lemma 1 *There is a one-to-one correspondence between combinatorial vector fields on G and relations (arc sets) satisfying the three properties in definition 5 such that $(x, y) \in P$ if and only if $\eta(x) = \{x, y\}$.*

Proof ($\Leftarrow$) Suppose P satisfies (1)-(3). $\eta(x) = e = \{x,y\}$ is a uniquely defined edge $e \in E$ according to properties (1) and (2). Property (3) implies that $y \notin \eta^{-1}(e)$, and hence $\eta^{-1}(e)$ is at most a single vertex for any $e \in E$.

($\Rightarrow$) Suppose η is a combinatorial vector field on G. Since $\eta : V \to E$ is a map, properties (1) and (2) follow immediately. From $\eta(x) = \{x,y\}$ we obtain $(x,y) \in P$. Now suppose $(y,x) \in P$, i.e., $\eta(y) = \{y,x\}$. This would imply $\eta^{-1}(\{x,y\}) = \{x\}$, a contradiction to the definition of η. Hence P must be antisymmetric.

In the following it will sometimes be convenient to identify a combinatorial vector field η with its equivalent orientation $P_\eta \subset V \times V$.

3 Combinatorial Vector Fields and Landscapes

Given a landscape (G, f) on G and let us define the following subsets and subgraphs: for all $x \in V$ let $N(x) = \{y \in V | \{x,y\} \in E\}$, $N[x] = N(x) \cup \{x\}$, and $N^>(x) = \{y \in N(x) | f(y) < f(x)\}$. Furthermore, let $N^>(W) = \bigcup_{z \in W} N^>(z)$. We call x a *drainage point* if $N^>(x) \neq \emptyset$.

Denote by $G^f(x)$ the connected component of $G[\{z \in V | f(z) = f(x)\}]$ that contains x. A landscape is *locally non-degenerate* or *invertible on edges* if $G^f(x)$ consists of a single vertex for all $x \in V$.

For any subgraph H of G we define $\overrightarrow{H}$ by the following sets of vertices and edges:

$$V(\overrightarrow{H}) = \bigcup_{x \in V(H)} N^>[x]$$

$$E(\overrightarrow{H}) = E(H) \cup \left\{ \{x,y\} \in E | x \in V(H), y \in N^>(x) \right\}$$

(3)

Let $\Pi = \{G^f(x) | x \in V\}$ be the set of maximal connected components of G on which the landscape is constant.

Definition 6 We will call the subgraphs $\overrightarrow{\Pi} = \{\overrightarrow{G^f(x)} | x \in V\}$ the shelves of the landscape.

Note that $\overrightarrow{\Pi}$ is uniquely determined by the landscape (G, f). It will be convenient, furthermore, to distinguish between the "flat surface" of the shelf A given by $A^{\text{flat}} = \{y \in A | A \in \Pi\}$ and its "exit points" $A^> = \{y \in N^>(x) | x \in A \text{ for } A \in \Pi\}$. A shelf is locally minimal if $A^> = \emptyset$.

Lemma 2 *The sets $V(A)$, $A \in \Pi$, form a partition of V. The sets $E(A)$, $A \in \overrightarrow{\Pi}$, form a partition of E.*

Proof If x is a vertex, it is trivially contained in some $V(G^f(x))$. Let x, y be two vertices with $f(x) = f(y)$ and $V(G^f(x)) \cap V(G^f(y)) \neq 0$. Then there exists a $z \in V(G^f(x)) \cap V(G^f(y))$ such that $f(x) = f(y) = f(z)$. Thus, every maximally connected component that contains x and z also has to contain y, and therefore $V(G^f(x)) = V(G^f(y))$. Let e be an edge that is contained in $E(\overrightarrow{G^f(x)}) \cap E(\overrightarrow{G^f(y)})$. Then there are two possibilities: (1)Both endpoints of e are in $V(G^f(x)) \cap V(G^f(y))$. Then, since $V(G^f(x)) = V(G^f(y))$, $e \in E(G^f(x))$ holds. (2)$e = \{x, N^>(x)\}$, where $x \in V(G^f(x)) \cap V(G^f(y))$. Since $V(G^f(x)) = V(G^f(y))$, it follows that $E(\overrightarrow{G^f(x)})$ form a partition of E.
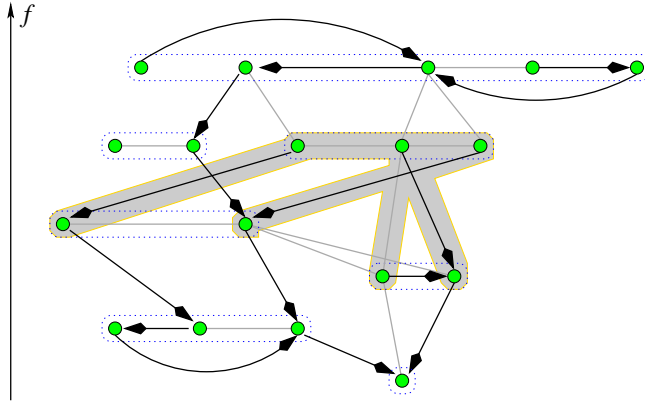
Fig. 2 Example of a small highly degenerate landscape. Vertices of G are arranged according to the fitness values $f(x)$. Connected componentes $\overline{G^f(x)}$ are indicated by dotted boxes. For one of them, the corresponding shelf-graph $\overrightarrow{G^f(x)}$ is highlighted in gray. A combinatorial vector field η (consistent with f) can be visualized as arrows correponding to the set P_η of oriented edges.

We can therefore write every combinatorial vector field η on G as a "vector" $(\eta_{A_1}, \ldots, \eta_{A_k})$ where η_{A_k} is the restriction of η to a shelf $A_k \in \overrightarrow{\Pi}$. In terms of the arcs we have

$$P_\eta = \bigcup_{A \in \overrightarrow{\Pi}} P_{\eta_A} \qquad \text{and} \qquad P_{\eta_A} \cap P_{\eta_B} \neq \emptyset \Rightarrow A = B \tag{4}$$

Definition 7 Given a simple undirected graph $G(V, E)$, a landscape (G, f) and a combinatorial vector field η on G we now consider the following properties

(A1) $(x, y) \in P_\eta$ implies $f(x) \geq f(y)$.

(A2) $(x, y) \in P_\eta$ and $x \notin \mathscr{R}_\eta$ implies $f(x) > f(y)$

(A3) $(x, y) \in P_\eta$ and $x \sim_\eta y$ implies $f(x) = f(y)$.

(B1) $N^>(x) \neq \emptyset$ implies that there is $(x, y) \in P_\eta$ with $y \in N^>(x)$.

(B2) $N^>(V(G^f(x))) \neq \emptyset$ implies that there is an η-path $\gamma = (x = z_1, \ldots, z_k = x')$ in $G^f(x)$ from x to an x' with $N^>(x') \neq \emptyset$.

(B3) $N^>(V(G^f(x))) \neq \emptyset$ implies that there is an η-path $\gamma = (x = z_1, \ldots, z_k = x')$ in $G^f(x)$ from x to an x' with $N^>(x') \neq \emptyset$ such that $N^>(z_j) = \emptyset$ for $2 \leq j \leq k - 1$.

(C1) $\mathscr{R}$ consists only of isolated points.

Conditions (A1) through (A3) specify in what sense the combinatorial vector field η "points downwards" in the landscape (G, f). We note that (A1) implies (A3). On the other hand, (A2) and (A3) together imply (A1). Suppose (A1) holds. Then, if $y \sim_\eta x$, we have $y \in V(G^f(x))$, i.e., the connected components of the chain recurrent set $\mathscr{R}_\eta$ are confined to the subgraphs $G^f(x)$ on which f is flat.

We observe that f can be extended to a Ljapunov function of η provided (A2) and (A3) are satisfied. More precisely, we have:

Lemma 3 *Suppose (A2) and (A3) are satisfied. Then there is a Ljapunov function $F : V \cup E$ for η satisfying $F(x) = f(x)$.*

Proof We define F on the edges as follows: if $\{x,y\} \in \eta(V)$ then $F(\{x,y\}) = (f(x) + f(y))/2$. This implies $f(x) > f(e) > f(y)$ if $(x,y) \in P_\eta$ and $x \notin \mathscr{R}_\eta$ and $f(x) = f(\{x,y\}) = f(y)$ if $x \sim_\eta y$. On the other hand, if $\{x,y\}$ is not oriented in η, i.e., $\{x,y\} \in E \setminus \eta(V)$, we set $F(\{x,y\}) = \max(f(x),f(y)) + 1$, so that $f(\{x,y\}) > f(x), f(y)$.

If (A1) but not necessarily (A2) and (A3) are satisfied, we can think of F as constructed in the proof above as a *weak Ljapunov function*.

Conditions (B1), (B2), and (B3) require that η points downwards in a certain sense whereever this is possible. Clearly (B3) implies (B2). Condition (B1) implies that the only rest points of η are the weak local minima of f. Conditions (B2) and (B3) imply that flat subgraphs $G^f(x)$ do not contain rest points unless all neighbors of $G^f(x)$ have a higher fitness, i.e., $G^f(x)$ form an extended strict local minimum of f. Otherwise, any $z \in G^f(x)$ is connected by an η-path to vertex y with $f(y) < f(z)$, i.e., $V(G^f(x)) \cap \mathscr{R} = \emptyset$. The distinction between (B2) and (B3) is that (B3) requires in addition that η "points downwards" on every drainage point.

Much of our subsequent discussion hinges on the observation that combinatorial vector fields satisfying (A1) as well as any combination $\mathfrak{X}$ of the properties defined above, can be composed of combinatorial vector fields on the shelves of (G,f) satisfying (A1) and $\mathfrak{X}$.

Lemma 4 *Suppose (A1) and any collection $\mathfrak{X}$ of additional properties (A2) through (C1) are satisfied. Then the restriction $\eta_{\overrightarrow{G^f(x)}}$ of η to the shelf $\overrightarrow{G^f(x)}$ is a combinatorial vector field that satisfies the same properties for the restricted landscape $(\overrightarrow{G^f(x)}, f_{|V(\overrightarrow{G^f(x)})})$.*

Proof We first observe that (A1) is satisfied globally if and only if it is true on every shelf because for each $x \in V$, $V(\overrightarrow{G^f(x)})$ contains all neighbors y of x for which $f(y) \leq f(x)$ holds. Since the $E(\overrightarrow{G^f(x)})$ form a partition of E, every possible arc of η is contained in exactly one shelf. The discussion above already showed that the connected components of the chain recurrent set $\mathscr{R}_\eta$ are contained in the flat parts $G^f(x)$ of the shelves, hence the assertion is true for (A2) and (A3). Conditions (B1) through (B3), and (C1), finally are already formulated restricted to shelves since $N^>(x) \subseteq V(\overrightarrow{G^f(x)})$.

Let $\mathrm{CVF}_{\mathfrak{X}}(A)$ denote the set of cvfs on the graph A that satisfy the conditions $\mathfrak{X}$ which in particular include 1a.

Theorem 1

$$\mathrm{CVF}_{\mathfrak{X}}(G) = \prod_{A \in \overrightarrow{\Pi}} \mathrm{CVF}_{\mathfrak{X}}(A) \tag{5}$$

Proof Using lemma 4 and equ.(4) we see that the union of combinatorial vector fields satisfying (A1) on the individual shelves is indeed a vector field satisfying (A1) on the entire graph G. The proof of lemma 4 shows in particular that each of the properties (A2) through (C1) holds globally if and only if it holds on each shelf provided (A1) is satisfied.

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 (A1) because degenerate fitness functions, therefore, impose fewer constraints on the combinatorial vector fields.
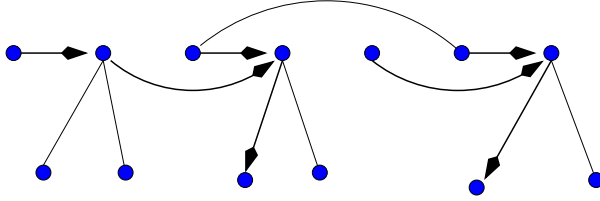
Fig. 3 Combinatorial vector fields on a single shelf. This example satisfies (A1) and (B2) but not (A2) [there are arcs connecting points with the same fitness that are not on a cylce], not (B1) [the trajectory does not point to one of the downward neighbors of the leftmost drainage point], and not (B3) [the trajectory passes through a drainage point without going downwards].

From here on, we will assume that η satisfies (A1) and any fixed collection $\mathfrak{X}$ of additional properties. Condition (A1) in particular implies that there is a close relationship between the local minima of the landcape (G, f) and the chain-recurrent set $\mathscr{R}_\eta$. The details, however, depend on the additional conditions. In any case, if x is a *strict* local minimum, then $x \in \mathscr{R}_\eta$.

Lemma 5 *Suppose (A1-A3), (B1) and (C1) are satisfied. Then $\mathscr{R}$ is the set of all weak local minima of (G, f).*

Proof ($\Rightarrow$) Let us assume that $x \in \mathscr{R}$ and (A1-A3), (B1) and (C1) are satisfied. (B1) implies that whenever there is a $y \in N^>(x)$, then there is an arc (x, y) and therefore $x \notin \mathscr{R}$. This means that for all $x \in \mathscr{R}$, $N^>(x) = \emptyset$ must hold. Thus, for all $y \in N(x)$, $f(y) \geq f(x)$ holds, i.e. x is a weak local minimum of f.

($\Leftarrow$) Now, let us assume that x is a weak local minimum. (A2) and (C1) imply that for all $y \in N(x)$, $f(y) > f(x)$ must hold. Thus, $\eta(x) = \varnothing$ and therefore $x \in \mathscr{R}$.

If we require (A1-A3),and (B2) or (B3), we see that all trajectories end in a shelf U with an empty set $N^>(U)$, i.e., all non-trivial recurrent cycles are confined to these locally strictly minimal shelves.

4 Weights and Partition Functions

Now consider a weight function $\omega : E \to \mathbb{R}$ defined on the edges of G. A natural choice for the weight of a combinatorial vector field on G is then

$$\omega(\eta) = \prod_{(x,y) \in \eta} \omega(\{x, y\}) \tag{6}$$

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

$$\omega(\{x, y\}) = \exp(\beta |f(x) - f(y)|) \tag{7}$$

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 transitions $(x, y) \in P_\eta$ receive the same weight 1. On the other hand, the steepest edges dominate in each set $N^>(x)$ for $\beta \to \infty$.

It will be useful to introduce the *partition function* of all combinatorial vector fields on (G, f)

$$Z = \sum_{\eta} \omega(\eta) \tag{8}$$

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

$$Z_{(u,w)} = \sum_{\eta : (u,w) \in \eta} \omega(\eta). \tag{9}$$

With Boltzmann weights, equ.(7), 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,

$$\omega(\eta) = \prod_{A \in \vec{\Pi}} \omega(\eta_A) \tag{10}$$

Theorem 1 therefore implies immediately that the partition functions are also products of partition functions restricted to the individual shelves:

$$Z = \sum_{\eta} \omega(\eta) = \sum_{\eta} \prod_{A \in \vec{\Pi}} \omega(\eta_A) = \prod_{A \in \vec{\Pi}} \sum_{\eta_A} \omega(\eta_A) = \prod_{A \in \vec{\Pi}} Z_A \tag{11}$$

Similarly, we can evaluate restricted partition functions such as

$$Z_{(u,w)} = \prod_{\substack{A \in \vec{\Pi} \\ \{u,w\} \notin E(A)}} \sum_{\eta_A} \omega(\eta_A) \times Z'_{(u,w)} = Z'_{(u,w)} \times \prod_{\substack{A \in \vec{\Pi} \\ \{u,w\} \notin E(A)}} Z_A \tag{12}$$

where $Z'_{(u,w)}$ is evaluated just like equ.(9) 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

$$Z = \prod_{x \in V} Z_x \qquad \text{and} \qquad Z_{(u,w)} = \prod_{x \in V \setminus \{u\}} Z_x \times Z'_{(u,w)} \tag{13}$$

with

$$Z_x = \sum_{y \in N^>(x)} \omega(x, y) \qquad \text{and} \qquad Z'_{(u,w)} = \omega(u, w). \tag{14}$$

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. The details depend, furthermore, on the exact combinations of properties $\mathfrak{X}$ in definition 7 that one cares to enforce.

5 Path Probabilities

In [29] an approximation for the folding kinetics of biopolymers is proposed that depends on a partition of V into basins. These basins are defined by assigning to each local minimum x the subset of vertices $B(x) \subset V$ that are transported to x by means of a gradient walk. For a fixed combinatorial vector field η, the natural analog is to define

$$B(x) = \{y \in V \,|\, x \in \psi_y\} \tag{15}$$

where ψ_y is the η-trajectory starting in y. In practise, of course this makes sense only for local minima of the landscape (G, f).

In the spirit of statistical mechanics, we endow the set of all combinatorial vector fields on (G, f) satisfying (A1) and a combination of properties $\mathfrak{X}$ with the discrete probability measure

$$p(\eta) := \omega(\eta)/Z \tag{16}$$

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

$$p_{(uw)} = Z_{(u,w)}/Z = \frac{1}{Z_A} \sum_{\substack{\eta \in \mathrm{CVF}_\mathfrak{X}(A) \\ (u,w) \in \eta}} \omega(\eta) = Z'_{(u,w)}/Z_A\,, \tag{17}$$

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

$$\mathbb{P}\{x \rightsquigarrow y\} := \frac{1}{Z} \sum_{\substack{\eta \\ x \rightsquigarrow y \in \eta}} \omega(\eta) = \frac{1}{Z} \sum_{z \in A_x^>} \sum_{\substack{\eta_{\overrightarrow{\Pi}\backslash A_x} \\ z \rightsquigarrow y}} \sum_{\substack{\eta_{A_x} \\ x \rightsquigarrow z}} \omega(\eta_{\overrightarrow{\Pi}\backslash A_x})\omega(\eta_{A_x}) \tag{18}$$

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

$$Z = \sum_\eta \omega(\eta) = \sum_{\eta_{\overrightarrow{\Pi}\backslash A_x}} \omega(\eta_{\overrightarrow{\Pi}\backslash A_x}) \sum_{\eta_{A_x}} \omega(\eta_{A_x}) = Z_{\overrightarrow{\Pi}\backslash A_x} Z_{A_x} \tag{19}$$

Substituting this decomposition into equ.(18) yields

$$\mathbb{P}\{x \rightsquigarrow y\} = \sum_{z \in A_x^>} \sum_{\substack{\eta_{\overrightarrow{\Pi}\backslash A_x} \\ z \rightsquigarrow y}} \frac{1}{Z_{\overrightarrow{\Pi}\backslash A_x}} \omega(\eta_{\overrightarrow{\Pi}\backslash A_x}) \times \frac{1}{Z_{A_x}} \sum_{\substack{\eta_{A_x} \\ x \rightsquigarrow z}} \omega(\eta_{A_x}) \tag{20}$$

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

$$T_{x \rightsquigarrow z} = \frac{1}{Z_{A_x}} \sum_{\substack{\eta_{A_x} \\ x \rightsquigarrow z}} \omega(\eta_{A_x}) \tag{21}$$

for the probability of a path within the shelf A_x from $x \in A_x^{\text{flat}}$ 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

$$T_{x \rightsquigarrow z} = \omega(x, z)/Z_{\{x\}} \qquad\qquad Z_{\{x\}} = \sum_{y \in N^>(x)} \omega(x, y). \tag{22}$$

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 $D_x = \{u \in V(A_x^{\text{flat}}) | N^>(u) \neq \emptyset\}$ denote the drainage points in A_x^{flat}. We have

$$T_{x \rightsquigarrow z} = \frac{1}{Z_{A_x}} \sum_{u \in D_x} \sum_{\substack{\eta_{A_x^{\text{flat}}} \\ x \rightsquigarrow u}} \sum_{\substack{\eta' \\ \text{on } N^>(u) \\ z \in N^>(u)}} \omega(\eta)\omega(\eta') \tag{23}$$

Introducing

$$T^>_{w \to z} := \frac{1}{Z_{\{w\}}} \sum_{\substack{\eta' \\ \text{on } N^>(w) \\ z \in N^>(w)}} \omega(\eta') = \frac{1}{Z_{\{w\}}} \omega(w, z) \tag{24}$$

we can rewrite equ.(23) in the form

$$T_{x \rightsquigarrow z} = \sum_{w \in D_x} \frac{Z_{\{w\}}}{Z_{A_x}} \sum_{\substack{\eta \\ \text{on } A_x^{\text{flat}} \\ x \rightsquigarrow w}} \omega(\eta) T^>_{w \to z} \tag{25}$$

We finally define the abbreviation

$$T^{\text{flat}}_{x \rightsquigarrow w} := \frac{Z_{\{w\}}}{Z_{A_x}} \sum_{\substack{\eta \\ \text{on } A_x^{\text{flat}} \\ x \rightsquigarrow w}} \omega(\eta) \tag{26}$$

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

$$T_{x \rightsquigarrow y} = \sum_{w \in D_x} T^{\text{flat}}_{x \rightsquigarrow w} T^>_{w \to y} \tag{27}$$

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.(21). The first subpath runs from the start point x to some exit point $u \in N^>(V(G^f(x)))$, and continues from there to y.

$$\tilde{P}(x \rightsquigarrow y) = \sum_{u \in N^>(V(G^f(x)))} T_{x \rightsquigarrow u} \tilde{P}(u \rightsquigarrow y) \tag{28}$$

For fixed y, eq.(28) 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.(28), 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

$$\mathbb{P}(x \rightsquigarrow y) = \sum_{z \in V(G^f(y))} \tilde{P}(x \rightsquigarrow z) T^{\text{flat}}_{z \rightsquigarrow y} \tag{29}$$

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

6 Reachability as a Topology

Definition 8 A vertex y is unreachable from x on (G, f) if there is no combinatorial vector field η (satisfying (A1) and a desired collection of properties $\mathfrak{X}$) 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". In non-degenerate landscapes, these statements are equivalent. Here we may also allow for certain walks that traverse "flat areas" of the landscape depending on which combinations of (A1-3), (B1-3) and (C1) we wish to enforce.

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

$$C(x) = \{y \in V | \mathbb{P}(x \rightsquigarrow y) > 0\} \tag{30}$$

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 $C(W) = \bigcup_{x \in W} C(x)$. One easily checks that the set-valued set-function $C : \mathfrak{P}(V) \to \mathfrak{P}(V)$ has the following five properties for all $W, W', W'' \in \mathfrak{P}(V)$.

(K1) $C(\emptyset) = \emptyset$
(K2) $W' \subset W'' \Longrightarrow C(W') \subset C(W'')$
(K3) $W \subseteq C(W)$
(K4) $C(W' \cup W'') = C(W') \cup C(W'')$
(K5) $C(C(W)) = C(W)$

These are Kuratowski's axioms for a closure function of V, see e.g. [7]. 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 6 *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.

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 $\mathscr{Q}$ if A has propery $\mathscr{Q}$ but $A \cup \{x\}$ does not have property $\mathscr{Q}$ for all $x \notin A$. Here we need to modify this to "$A \cup R_x$ does not have property $\mathscr{Q}$ for all $x \notin A$" where R_x is the set of mutually reachable points.

7 Valleys, Basins, and Barriers

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

Definition 9 A *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 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 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 valley W_L. Under conditions (A2), (A3), (B1), and (C1), L will contain sub-valleys W_z, each consisting of a single local minimum $z \in L$ and all those vertices z for which all adaptive walks lead to z. The importance of the valleys of type W_L lies in the fact that every trajectory of any combinatorial vector field statisfying (A1) and any of (B1-B3) ends in a uniquely defined 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 $\{y \in V | \mathbb{P}(x \rightsquigarrow y > 0) \text{ and } \mathbb{P}(y \rightsquigarrow x > 0)\} \neq \{x\}$. 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. 4.

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

Lemma 7 *A subset $W \subseteq V(G)$ is the 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

Fig. 4 Valley structure of a simple landscape. Each vertex is annotated with the list of reachable local minima. Each valley is characterized by such a list $\varUpsilon$ and contains all vertices labeled by a subset if $\varUpsilon$. The minimal closed subsets, here $\{1\}$ through $\{5\}$ are always valleys. In addition, this landscape has the valleys $\{1,2\}$, $\{2,3\}$, $\{4,5\}$, $\{3,4,5\}$, and $\{1,2,3,4,5\}$. Valley-connecting points are indicated by circles.

set. Thus W is a maximal connected closed set.

Now suppose that W is a maximal closed connected set, and set $\varUpsilon = \bigcup_{x \in W} \varXi(x)$. Then (ii) is trivially true and W contains in particular all minimal closed sets $C \in \varUpsilon$. Now suppose that there is a vertex $z \notin W$ with $\varXi(z) \subseteq \varUpsilon$. All adaptive walks emanating from z thus eventually reach W and all vertices y along such a walk satisfy $\varXi(y) \subseteq \varXi(z) \subseteq \varUpsilon$. Hence we can expand W by the last mutually reachable subset R_y outside of W, contradicting maximality. Hence $\varXi(z) \subseteq \varUpsilon$ implies $z \notin W$.

The valleys of the landscape (G, f) do not form a hierarchical structure. In Fig. 4, the valleys $\{1,2\}$ and $\{2,3\}$ are a counterexample. Nevertheless, the valleys are closely related to the barrier trees of the landscape. In particular, we can identify the (lowest) points that connect valleys with each other.

Definition 10 A vertex $u \in V$ is a *valley-connecting vertex* if $\varXi(u)\varXi(v) \neq \emptyset$ 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, valley-connecting vertices linking the same two valleys. In Fig. 4, there are two vertices connecting the valleys $\{1\}$ and $\{2\}$, which have different fitness values.

In order to connect our present discussion with earlier work, in particular [4,6,29], we briefly discuss the notation of saddle points in the context of our present formalism.
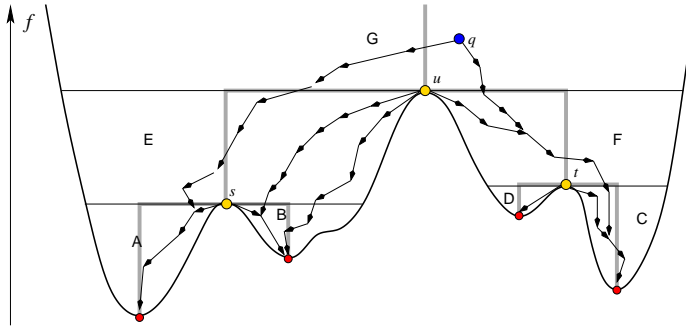
Fig. 5 Valleys, basins, and adaptive walks. In this example, there are 7 valleys denoted A to G. The minimal valleys are A through D corresponding to the four local minima of the landscape. The valleys A and B are connected by the direct saddle point s. The direct saddle q between A and C has a strictly higher energy than saddle u between the same two 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 valleys E and D and A $\subseteq$ E. The barrier tree of the landscape reflects the inclusion relations of the valleys: A, B $\subseteq$ E, C, D $\subseteq$ F, E, F $\subseteq$ G. The valley G corresponds to the entire landscape.

Definition 11 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 valley-connecting point with minimal fitness connecting two valleys. In [7], *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 valleys below a certain fitness threshold are always connected sets. Thus, basins are connected sets of the form

$$\bigcup_{W \in \mathscr{V}} \{x \in W | f(x) < \eta\} \tag{31}$$

constructed from maximal collections of valleys $W \in \mathscr{V}$. Saddle points, i.e., vertices of minimal fitness that connect distinct basins are therefore necessarily valley-connecting points between valleys associated with the distinct basins that they merge. Given an arbitrary pair of disjoint valleys, their direct saddle can have a strictly larger value of f than the saddle point connecting the associated basins, Fig. 5.

We remark, finally, that the flooding algorithm implemented in the `barriers` program [4,6] 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 [6]. This gives rise to the recursion

$$\Xi(x) = \bigcup_{x' \in W_x} \bigcup_{y \in N(x') \cap C(x)} \Xi(y) \tag{32}$$

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.

A detailed investigation of the relationships between the valley structure of the landscape and the basins and barriers of the landscape discussed in earlier work [4, 6, 7, 29] goes beyond the scope of this contribution. These connections will be explored in a forthcoming contribution.

8 Macrostates

In [29], macrostates were identified with "basins", i.e., set of vertices that are connected to the same local minimum (or equivalently, the same minimal valley) by means of gradient descent walks. Since gradient walks are not unique in degenerate landscapes, this definition becomes problematic in many applications. Here, we propose to relax the requirement that macrostates must be classes of a partition of microstates and instead consider macrostates that are linear combinations of microstates. Of course, we want the definition of macrostates to reflect the geometric structure of the landscape as closely as possible. Thus we index macrostates by the collection of minimal valleys. This makes sense because all trajectories under all combinatorial vector fields on the landscape end in one of these valleys. Instead of assigning each vertex $x \in V$ to a unique macrostate, we now distribute "shares" of x among all those minimal valleys that can be reached from x. Naturally, the share of valley α is computed as the probability

$$p_\alpha(x) = \sum_{y \in \alpha} P(x \rightsquigarrow y) \tag{33}$$

that a trajectory starting at x ends up in α. The partition function associated with this macrostate is then simply given by

$$Q_\alpha = \sum_{x \in V} \sum_{y \in \alpha} P(x \rightsquigarrow y) \exp(-\beta f(x)) \tag{34}$$

and we easily convince ourselves that

$$\sum_\alpha Q_\alpha = \sum_{x \in V} \exp(-\beta f(x)) \tag{35}$$

is satisfied, i.e., this construction of the macrostates still properly reflects the thermodynamics of the microscopic system.

We can now use our "mixed" macrostates to refine equ.(1). Since the contribution of y to the partition function of the macrostate α is $p_\alpha(y) \exp(-\beta f(y))$, the approximation assumption that the folding trajectory equilibrates within the macrostate becomes

$$\mathrm{Prob}[y|\gamma] \approx p_\gamma(y) \exp(-\beta f(y)/Q_\gamma. \tag{36}$$

A transition to x, on the other hand takes us to macrostate α with probability $p_\alpha(x)$, so that

$$r_{\alpha,\gamma} \approx \sum_{x,y \in V} p_\alpha(x) r_{xy} p_\gamma(y) \exp(-\beta f(y))/Q_\gamma \tag{37}$$

For deep valleys equ.(1) will still be an excellent approximation: Below the saddle point that separates γ from its most easily accessible neighbor we have $p_\gamma(y) = 1$, while $\exp(-\beta f(y))/Q_\gamma$ is already small above this saddle height. Similarly, in this type of landscape, $p_\alpha(x)$ will be

1 or at least close to 1 for points not much above the saddle point. Hence we approximately recover equ.(1). On the other hand, if the valleys are shallow or even nearly flat, separated by low saddles, then the stastistics of pathways leading to different valleys will play a large role [26].

9 Concluding Remarks

We have developed here a framework to study adaptive walks and associated structures for fitness landscapes with complex degeneracies. Combinatorial vector fields play a central role in this endeavour. They imply a deterministic "down-hill" dynamics on the landscape that essentially considers adaptive walks as admissible trajectories. The given landscape can be understood as a Lyapunov function of the admissible combinatorial vector fields. In case of landscapes with neutrality, there is some freedom to choose which steps to neutral neighbors are acceptable, i.e., how the Lyapunov function must behave along neutral steps. The crucial technical result is that the set of admissible combinatorial vector fields has a simple product structure, that allows combinatorial vector fields on the entire landscape to be constructed as the unrestricted combinations of combinatorial vector fields on the shelves of (G, f). The shelves are subgraphs determined by the connected components of neutral networks, and specialize to individual vertices and their downward neighbors in the non-degenerate case. Introducing weights for the combinatorial vector fields that depend on their steepness allows us to bring back the stochastic aspect of adaptive walks in a controlled way that is amenable to a thermodynamics-like formulation in terms of partition functions.

The framework represented here connects in a natural way to previous work on the basin and barrier structure of landscapes. In particular, reachability w.r.t. admissible combinatorial vectorfields on (G, f) defines a finite topology on the vertex set V of the underlying graph. This topological structure gives rise to a natural notion of "valleys" that is closely related to the barriers and saddle points investigated in previous work. This valley structure, in particular, can be identified in the course of the same flooding algorithm that is used to determine basins and barriers in the `barriers` program.

The framework developed here can be of interest much beyond that however. It becomes straightforward, for example, to investigate dependencies between adaptive walks. It follows immediately from the discussion here that two paths, one from x to y and the other from p to q, are independent if they traverse disjoint sets of shelves. On the other hand, they become dependent if they pass through a common point: uniqueness of trajectories implies that they must be identical from the intersection point. Otherwise they can only be realized by disjoint sets of combinatorial vector fields.

References

1. Becker, O.M., Karplus, M.: The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J. Chem. Phys. **106**, 1495–1517 (1997)
2. Binder, K., Young, A.P.: Spin glasses: Experimental facts, theoretical concepts, and open questions. Rev. Mod. Phys. **58**, 801–976 (1986)
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. Flamm, C., Fontana, W., Hofacker, I., Schuster, P.: RNA folding kinetics at elementary step resolution. RNA **6**, 325–338 (2000)
5. Flamm, C., Hofacker, I.L.: Beyond energy minimization: Approaches to the kinetic folding of RNA. Chemical Monthly **139**, 447–457 (2008)
6. Flamm, C., Hofacker, I.L., Stadler, P.F., Wolfinger, M.T.: Barrier trees of degenerate landscapes. Z. Phys. Chem. **216**, 155–173 (2002)
7. 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
8. Flyvbjerg, H., Lautrup, B.: Evolution in a rugged fitness landscape. Phys. Rev. A **46**, 6714–6723 (1992)
9. Fontana, W., Stadler, P.F., Bornberg-Bauer, E.G., Griesmacher, T., Hofacker, I.L., Tacker, M., Tarazona, P., Weinberger, E.D., Schuster, P.: RNA folding landscapes and combinatory landscapes. Phys. Rev. E **47**, 2083–2099 (1993)
10. Forman, R.: Combinatorial vector fields and dynamical systems. Math. Z. **228**, 629–681 (1998)
11. Garey, M., Johnson, D.: Computers and Intractability. A Guide to the Theory of $\mathcal{NP}$ Completeness. Freeman, San Francisco (1979)
12. Garstecki, P., Hoang, T.X., Cieplak, M.: Energy landscapes, supergraphs, and "folding funnels" in spin systems. Phys. Rev. E **60**, 3219–3226 (1999)
13. Gillespie, J.H.: Molecular evolution over the mutational landscape. Evolution **38**, 1116–1129 (1984)
14. Jonsson, J.: Simplicial Complexes of Graphs. Springer, Berlin (2007)
15. Kauffman, S.A., Levin, S.: Towards a general theory of adaptive walks on rugged landscapes. J. Theor. Biol. **128**, 11–45 (1987)
16. Klotz, T., Kobe, S.: "Valley Structures" in the phase space of a finite 3D Ising spin glass with $\pm i$ interactions. J. Phys. A: Math. Gen **27**, L95–L100 (1994)
17. Macken, C.A., Hagan, P.S., Perelson, A.S.: Evolutionary walks on rugged landscapes. SIAM J. Appl. Math. **51**, 799–827 (1991)
18. Macken, C.A., Perelson, A.S.: Protein evolution on rugged landscapes. Proc. Natl. Acad. Sci. USA **86**, 6191–6195 (1989)
19. Mézard, M., Parisi, G., Virasoro, M.A.: Spin Glass Theory and Beyond. World Scientific, Singapore (1987)
20. Mezey, P.G.: Potential Energy Hypersurfaces. Elsevier, Amsterdam (1987)
21. Mirny, L., Shakhnovich, E.: Protein folding theory: from lattice to all-atom models. Annu Rev Biophys Biomol Struct. **30**, 361–396 (2001)
22. Niklas, K.J.: Adaptive walks through fitness landscapes for early vascular land plants. Amer. J. Botany **84**, 16–25 (1997)
23. Perelson, A.S., Macken, C.A.: Protein evolution on partially correlated landscapes. Proc. Natl. Acad. Sci. USA **92**, 9657–9661 (1995)
24. Prügel-Bennett, A., Hallam, J.: Barrier trees MAX-SAT combinatorial optimization cost landscape heuristic search. IEEE Trans. Evol. Comput. **9**, 385–397 (2005)
25. Reidys, C.M., Stadler, P.F.: Combinatorial landscapes. SIAM Review **44**, 3–54 (2002). SFI preprint 01-03-14
26. Van Nimwegen, E., Crutchfield, J.P.: Metastable evolutionary dynamics: Crossing fitness barriers or escaping via neutral paths? Bull. Math. Biol. **62**, 799–848 (2000)
27. Wales, D.J., Miller, M.A., Walsh, T.R.: Archetypal energy landscapes. Nature **394**, 758–760 (1998)
28. Weinberger, E.D.: Local properties of Kauffman's N-k model: A tunably rugged energy landscape. Phys. Rev. A **44**, 6399–6413 (1991)
29. 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)
30. 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)

Appendix: Combinatorial Vector Fields on Simplicial Complexes

Let M be a simplicial complex constructed over the set K of simplices, see e.g. [14]. If σ and τ are simplices of M, we write $\sigma < \tau$ if σ lies in the boundary of τ.

Definition 12 A *combinatorial vector field* on M [10] is a map $\eta : K \to K \cup \{\varnothing\}$ such that

1. If $\eta(\sigma) \neq \varnothing$ then $dim\eta(\sigma) = dim(\sigma) + 1$, and $\sigma < \eta(\sigma)$.
2. If $\eta(\sigma) = \tau \neq \varnothing$, then $\eta(\tau) = \varnothing$.
3. For all $\sigma \in K$, $|\sigma^{-1}| \leq 1$.

The simplicial complex $M(G)$ associated with a graph G consists of the vertices and edges of G, $K = V \cup E$, with $v < e$ if and only if $v \in V$ is a vertex incident with the edge $e \in E$. Condition (i) implies that $\eta(v)$ is either $\varnothing$ or an edge incident with v for any vertex $v \in V$, while $\eta(e) = \varnothing$ for all $e \in E$ because there is no higher-dimensional simplex for and edge e to be mapped to. Condition (ii) thus becomes void on graphs, and we can view η on G simply as a map from vertices to edges. Condition (iii) finally implies that an edge $e \in E$ is the image of at most one of its two incident vertices. It follows that a combinatorial vector field on G is characterized by the properties stated in Definition 1.

Definition 13 A simplex $\sigma \in K$ is a *rest point* of η if $\eta(\sigma) = \varnothing$ and $\eta^{-1}(\sigma) = \emptyset$

On graphs, thus, $\{x,y\}$ is a rest point if and only if $\eta(x) \neq \{x,y\}$ and $\eta(y) \neq \{x,y\}$.

Definition 14 An η-path is a finite sequence of simplices $\gamma = (\sigma_0, \tau_0, \sigma_1, \tau_1, \ldots \sigma_{n-1}, \tau_{n-1}, \sigma_n)$ such that $\eta(\sigma_i) = \tau_i$ for $0 \leq i < n$ and $\sigma_{i+1} < \tau_i$.

On a graph, γ necessarily alternates between vertices and edges. Since the edge e_i is determined as $\eta(v_i)$ it follows that v_{i+1} is the other end of e_i. It is therefore uniquely determined by $\eta(v_i)$. Hence γ is an ordinary path on G whose sequence of vertices is determined by η (except for the start and end point).

Definition 15 [10] A *Lyapunovfunction* of the combinatorial vector field η is a function $F : M \to \mathbb{R}$ such that

1. if $\sigma \notin \mathscr{R}$ and $\tau > \sigma$ then
   (a) $F(\sigma) < F(\tau)$ if $\tau \neq \eta(\sigma)$
   (b) $F(\sigma) \geq F(\tau)$ if $\tau = \eta(\sigma)$
2. if $\sigma \in \mathscr{R}$ and $\tau > \sigma$ then
   (a) $F(\sigma) = F(\tau)$ if $\sigma \sim \tau$
   (b) $F(\sigma) < F(\tau)$ if $\sigma \not\sim \tau$

To specialize this to graphs, we simply translate the conditions one-by-one

1. If $x \in V \setminus \mathscr{R}$ and $x \in e$ then $F(x) \geq F(e)$ if $e = \eta(x)$ and $F(x) < F(e)$ otherwise.
2. If $x \in V \cap \mathscr{R}$ and $x \in e$ then $F(x) = F(e)$ if $x \sim e$ and $F(x) < F(e)$ otherwise

Now consider explicitly the "other end" of the edge e:

1. If $x \in V \cap \mathscr{R}$ and $\eta(x) = \{x,y\}$ then $F(y) = F(x)$.
2. If $e = \{x,y\} \in E \cap \mathscr{R}$ then $F(e) > F(x)$ and $F(e) > F(y)$.
3. If $e \in E \setminus \mathscr{R}$ then, w.l.o.g., $e = \{x,y\}$ and $e = \eta(x)$. In this case $F(x) \geq F(e) > F(y)$. This implies
4. If $x \in V \setminus \mathscr{R}$ and $\eta(x) = \{x,y\}$ then $F(y) < F(x)$.

This in turn easily translates to definition 4 in the main text.