Barrier Trees

The (Bio)Informatics View

Sven Findeiß

Seminar 'parallel algorithms and complex systems'
University of Leipzig

Leipzig, April 17, 2006

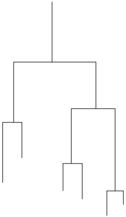
Outline

Definition of a Tree

the biological tree

A tree is a persevering (lasting over several years) plant, which has a distinct upright growing wooden trunk. The trunk is growing up from a root, and on it are branches and twigs situated.

the informatics tree

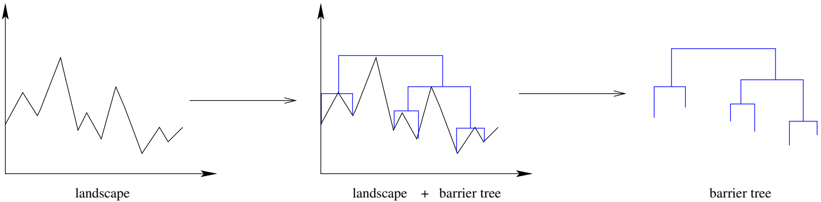


A tree is a special graph. We will consider G(V,E) to be a connected, undirected, acyclic, simple graph with vertex set V and edge set E.

Barrier Trees I

Definition

Barrier Trees are a method for representing the (fitness) landscape structure, of high-dimensional discrete spaces.



Barrier Trees II

Barrier Trees are a attractive technique to visualize some aspects of landscapes.

`examples:`

- physical processes e.g. disordered spin-systems
- chemical processes e.g. bio-polymer folding
- describe concept of evolutionary biology
- combinatorial optimization

Landscape I

- a landscape is a triplet $(\mathbb{X}, N, f)$
  - $\mathbb{X} \ldots$ set of configurations
  - $N \ldots$ topological structure
  - $f : \mathbb{X} \to \mathbb{R} \ldots$ cost or fitness function
- neighborhood function N is typically defined by a move set
  1. optimization algorithms chosen by the user
  2. biological aplications a mechanism of mutation or recombination

Landscape II

- configuration space $(\mathbb{X}, N)$ is a finite undirected graph

$$G(\mathbb{X}, E)$$
$$\mathbb{X} \ldots \text{vertex set}$$
$$E \ldots \text{edge set}$$

- edges connect configurations that can inter-converted by a single move

Landscape III

- Get a **landscape** by mapping configurations $x \in \mathbb{X}$, with a cost/fitness function f, from a decision space $\mathbb{X}$ into real numbers $\mathbb{R}$

$$f : \mathbb{X} \rightarrow \mathbb{R}$$

- decision space is a finite set $\mathbb{X}$ of configurations
- it is equipped with some notion of adjacency, nearness, distance or accessibility
- $f(x) \ldots$ values the fitness of the configuration x

The Example of a real Landscape

Neighborhood

- to describe a landscape we need a **neighborhood** function

$$N : \mathbb{X} \to P(\mathbb{X})$$

  $P(\mathbb{X}) \ldots$ the power set of $\mathbb{X}$

- `features of the neighborhood-function:`
  1. symmetric:
     $x, y \in \mathbb{X}$ and $x \in N(y) \to y \in N(x)$
  2. reflexive:
     $x \in N(x), \forall x$

Path in a Landscape

- A **path** in the landscape is a ordered list of configurations

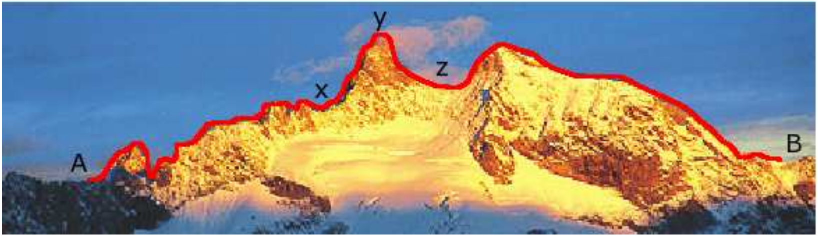
$$\pi = \{x_1, x_2, \ldots, x_n\}$$

such that

$$x_j \in \mathbb{X} \wedge x_{j+1} = N(x_j); \forall j$$

$\Rightarrow$ a path depends on the neighborhood function

Example I



1. $N(y) = \{x, y, z\}$
2. path from A to B

$$\Rightarrow \pi = \{A, \dots, x, y, z, \dots, B\}$$

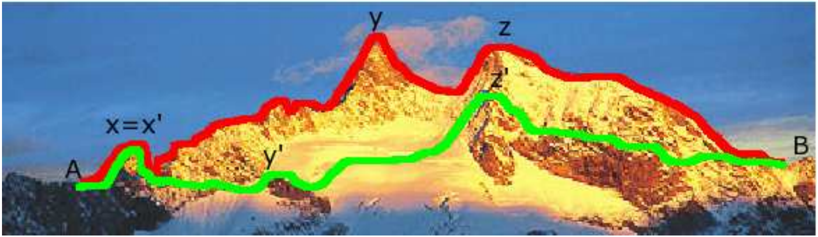
Accessibility

- a configuration y is accessible from x on level η if there is a path $\pi \in P_{xy}$ such that $f(z) \leq \eta; \forall z \in \pi$

$$x \overset{\eta}{\longleftrightarrow} y$$

- $P_{xy} \ldots$ set of all paths between x and y by a series of consecutive mutations
- `features of` $x \overset{\eta}{\longleftrightarrow} y$:
  1. symmetric:
     $x \overset{\eta}{\longleftrightarrow} y \Rightarrow y \overset{\eta}{\longleftrightarrow} x$
  2. transitive:
     $x \overset{\eta}{\longleftrightarrow} y \wedge y \overset{\eta}{\longleftrightarrow} z \Rightarrow x \overset{\eta}{\longleftrightarrow} z$
  3. reflexive:
     $\forall \eta \geq f(x)$

Example II



$$?A \xleftrightarrow{\eta} B?$$

$\pi = \{A, x, y, z, B\}$
$f(x) < \eta$
$f(y) > \eta$

$\Rightarrow$ B is not accessible from A

$\pi = \{A, x', y', z', B\}$
$f(x') < \eta$
$f(y') < \eta$
$f(z') \leq \eta$
$\Rightarrow$ B accessible from A

Minima

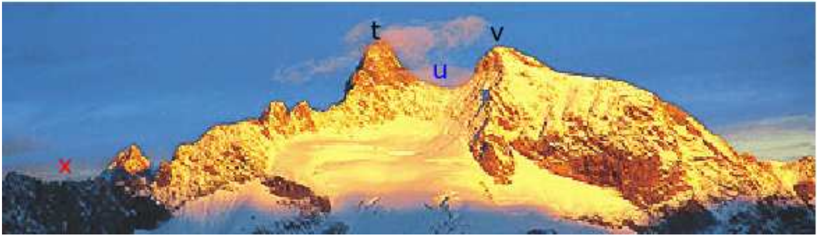
- x is a **local minimum** if

$$f(x) \leq f(y); \forall y \in N(x)$$

- x is a **global minimum** if

$$f(x) \leq f(y); \forall y \in \mathbb{X}$$

Example III



- $f(u) \leq f(t) \wedge f(u) \wedge f(v)$
  $\Rightarrow f(u) \leq f(n); \forall n \in N(u)$
  $\Rightarrow$ u is a local minimum
  but $f(x) < f(u)$
  $\Rightarrow$ not a global minimum
- $f(x) \leq f(n); \forall n \in \mathbb{X}$
  $\Rightarrow$ x is global minimum

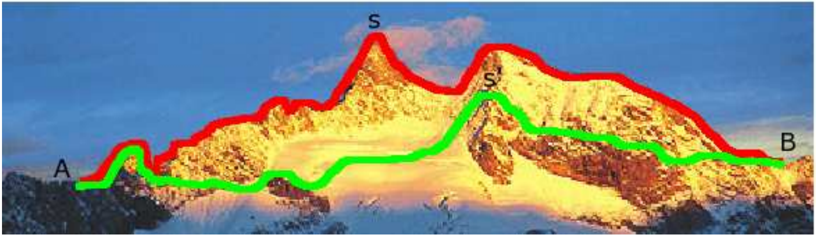
Saddle Point

- a **saddle point** between two minima is the highest cost configuration on the lowest cost path between the two minima
- s is a saddle point between x and y if:

$$\widehat{f}[x, y] = min_{\pi \in P_{xy}} max_{s \in \pi} f(s)$$

- $\widehat{f}[x, y] \ldots$ energy of the lowest saddle point s between the two minima x and y

Example IV



$$\widehat{f}[A,B] = min_{\pi \in P_{AB}} max_{s \in \pi} f(s)$$

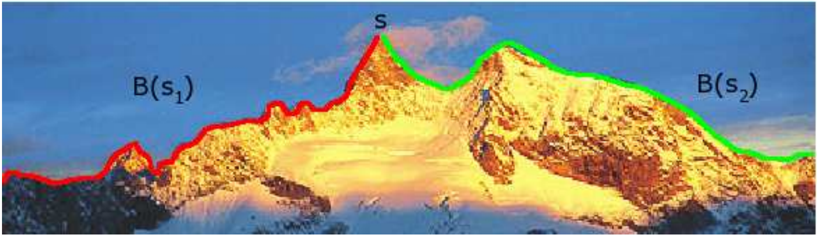
1. find the $max_{s \in \pi} f(s)$ of each path from A to B
   $\Rightarrow$ s and s'
2. take the min path (lowest saddle point) between A and B
   $\Rightarrow$ s' is the saddle point between A and B

Basin or Valley

- a saddle point connects a collection of configurations B(s)
- all configurations in B(s) can be reached by a path which never exceeds f(s)

We can say that B(s) is a **valley** or **basin** below the saddle s

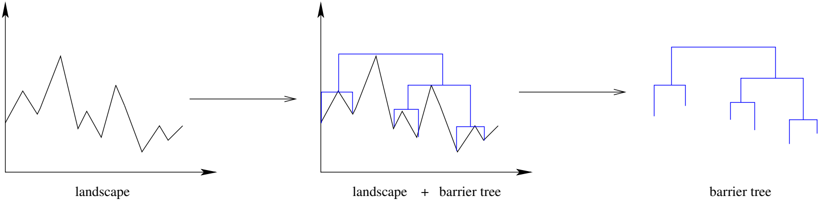
Example V



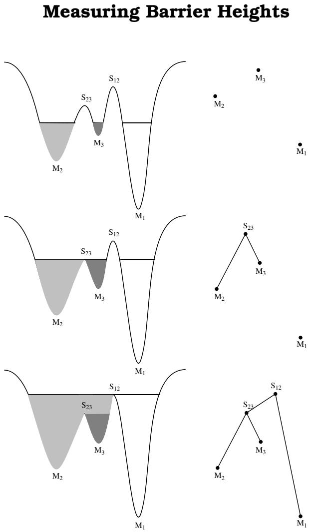
- $B(s_1)$ and $B(s_2)$ are the two colored basins below the saddle s

What we need to get the Barrier Tree

1. (fitness) landscape of the problem
2. local/global minima
3. saddle points
4. the connected basins under each saddle point



The Flooding Algorithm



The Algorithm:

Read conformations in energy sorted order.
For each confirmation x we have three cases:

(a) x is a local minimum if it has no neighbors we've already seen

(b) x belongs to basin $B(s)$, if all known neighbors belong to $B(s)$

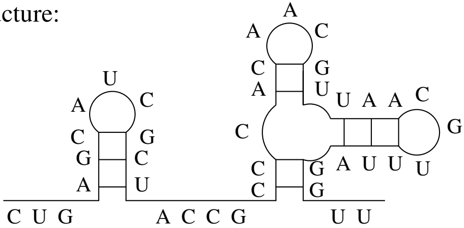
(c) if x has neighbors in several basins $B(s_1) \dots B(s_k)$ then it's a saddle point that *merges* these basins.
Basins $B(s_1), \dots, B(s_k)$ are then united and are assigned to the deepest of local minimum.

BioInformatics Background I

primary structure

C U G A G C A U C G C U A C C G C C C A C A A C G U U A A C G U U U A G G U U

secondary structure:



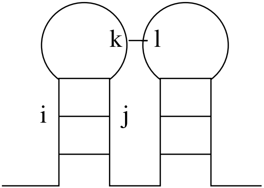
bracket–dot–notation:

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

BioInformatics Background II

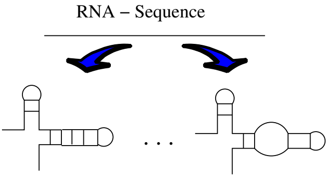
RNA secondary structure:

- list of base pairs (i,j)
- any base i `may` pair with at most one other base j
- six allowed pairs are {AU,UA,CG,GC,GU,UG}
- no pseudo-knots
  $\Rightarrow$ (i,j) and (k,l) are base pairs
  $\Rightarrow$ $i < k < j < l$ is not allowed



Folding Landscape of an RNA Molecule I

- different secondary structures could be shaped from a molecule sequence

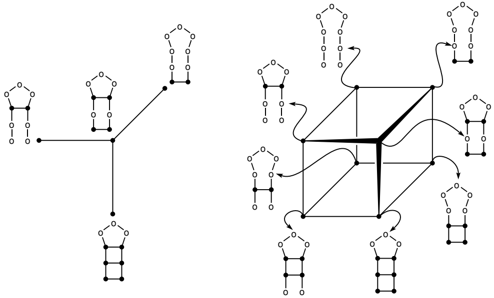


- secondary structures could be evaluate by a well established energy model
  $\Rightarrow$ each structure has its specific free energy

Folding Landscape of an RNA Molecule II

move set:

- typically addition or removal of a single base pair from the structure



$\Rightarrow$ Neighborhood Function

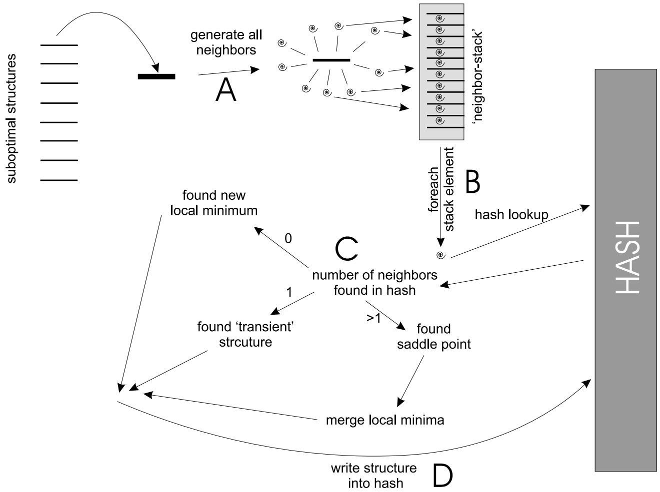
Energy Landscape

The energy landscape of a given RNA sequence is determined by:

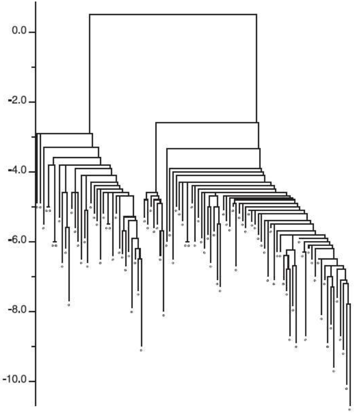
1. all legal secondary structures, the molecule can fold
2. the energies of all these secondary structures
3. the move set

The folding landscape is interesting, because the energy barriers could affect the lifetime or the effect of an RNA molecule.

Sketch of the Algorithm



Barrier Tree of SL RNA



- the graph is restricted to the first 100 local minima
- two alternative conformations are seperated by a high barrier
- one global minimum on the right side

$\Rightarrow$ SL-RNA is able to build two competing secondary structures with near equal free energy