The Structure of Fitness Landscapes

Peter F. STADLER

Lehrstuhl Bioinformatik, Institut für Informatik, Universität Leipzig

Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien

Santa Fe Institute, New Mexico

`http://www.tbi.univie.ac.at/~studla`

Agent Based Modelling and Simulation

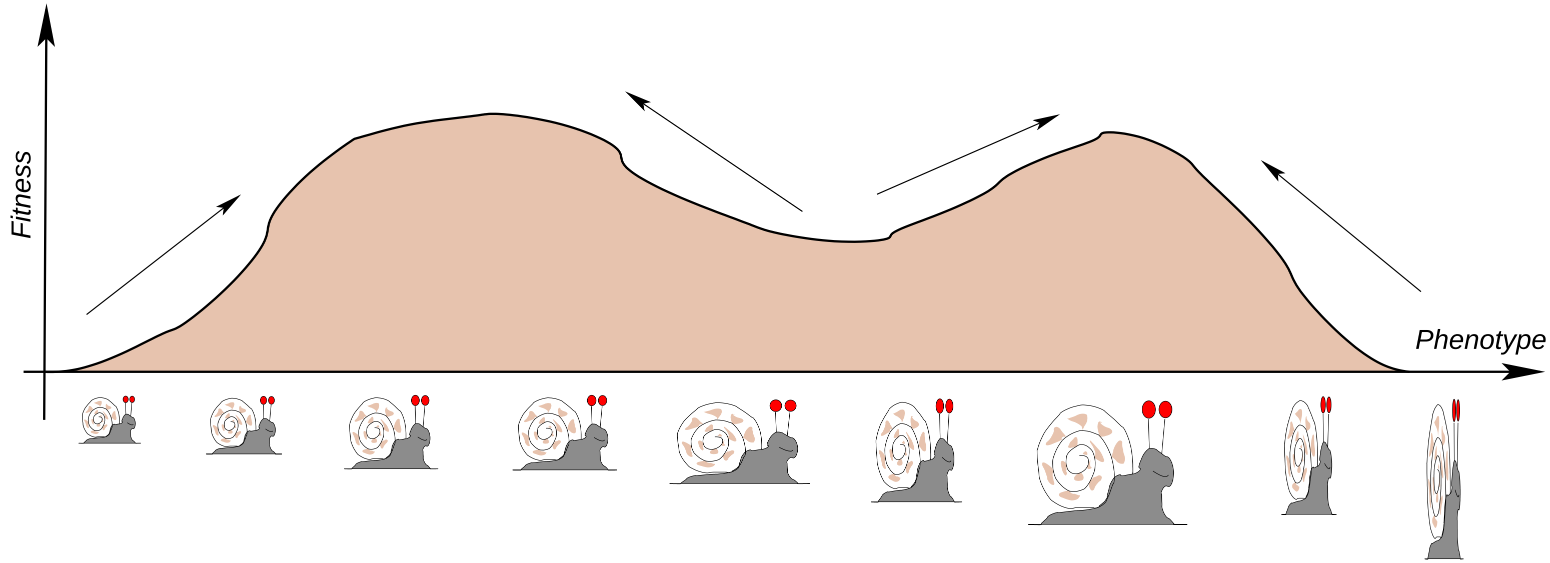
Minneapolis, Nov 3 2003

Landscapes

Great Basin NV

$$\text{height}(\text{longitude}, \text{latitude})$$

Sewall Wright's Fitness Landscape



More abstractly: A **landscape** is a triple $(V, \mathcal{X}, f)$ where

V is a set of *configurations*.

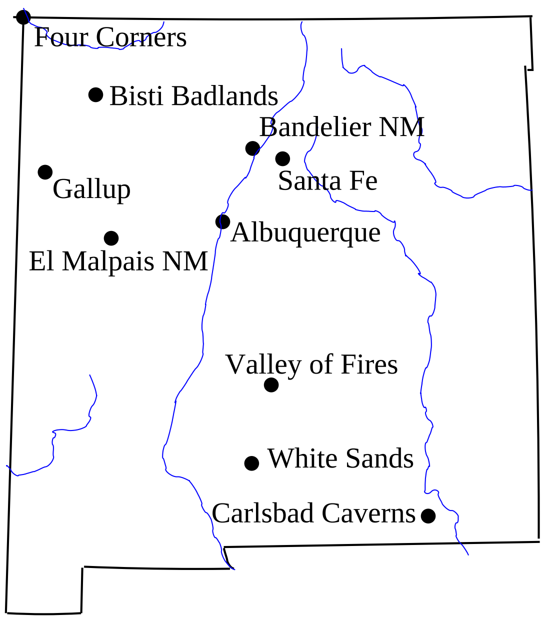
E.g.: gene sequences, tours of travelling salesman, spin configurations;

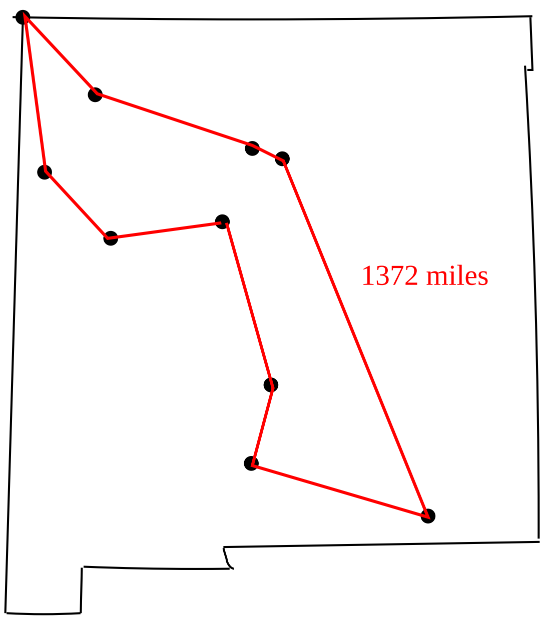
f is a *cost* or *fitness function* $f : V \to \mathbb{R}$;

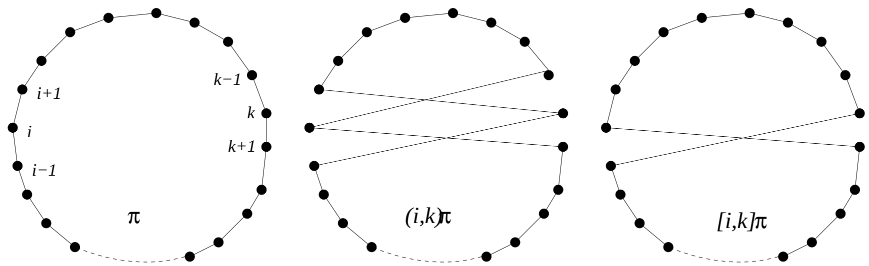
$\mathcal{X}$ is a way of defining "nearness", "closeness", "dissimilarity", or "accessibility" among the configurations.

E.g. an *adjacency relation* ($\Longrightarrow$graph), *transition matrix* ($\Longrightarrow$Markov Chain), or a *(pre)topology* on V.

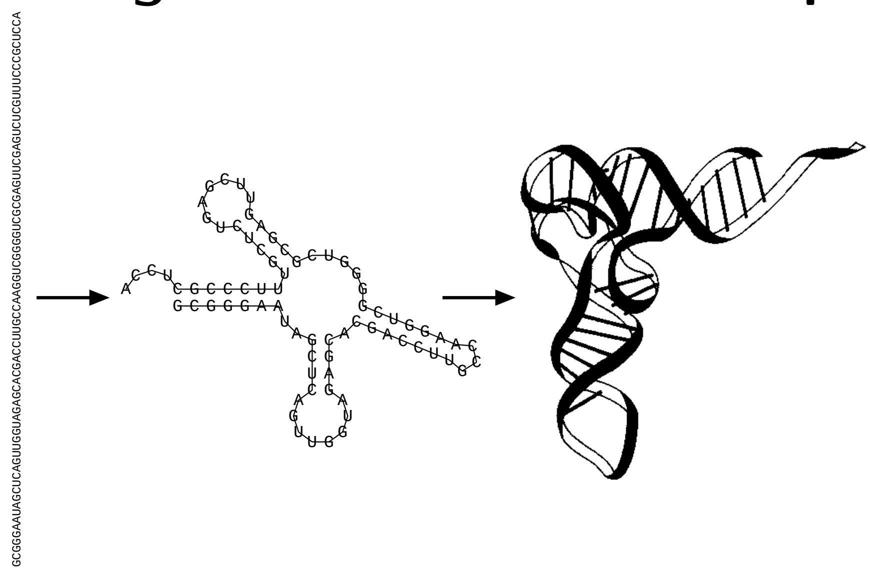
Combinatorial Landscapes







Biological Realistic Landscapes



Sequence ⟶ Shape/Structure ⟶ Fitness

Fitness landscapes *inherit* most their properties largely from the

sequence→shape or **genotype-phenotype map**

Ruggedness

Rugged: Bryce Canyon UT

Smooth: Capulin Volcano NM

Measures of Ruggedness:

- Number of Local Minima and Maxima
- Length of Adaptive Walks

Correlation Functions

Correlation Functions:
Random walk on V as defined through $\mathcal{X}$: $x_0, x_1, x_2, \ldots$ (Markov process on V).
$\Longrightarrow$ "Time Series" $f(x_0), f(x_1), f(x_2), \ldots$
$\Longrightarrow$ Autocorrelation Function

$$r(s) = \frac{\left\langle \big(f(x_{t+s}) - \langle f \rangle\big)\big(f(x_t) - \langle f \rangle\big) \right\rangle_t}{\left\langle \big(f(x_t) - \langle f \rangle\big)^2 \right\rangle_t}$$

For simplicity:

$$f(x) \to f(x) - \overline{f} \qquad \overline{f} = \frac{1}{|V|} \sum_{x \in V} f(x)$$

(Subtract the landscape average)

"Time series" = Markov process with transition matrix $\mathbf{T}$.

Theorem. $r(s) = \langle f, \mathbf{T}^s f\rangle \big/ \langle f, f\rangle$.

Theorem. $r(s)$ is an exponential function iff f is an eigenfunction of $\mathbf{T}$. We say $(V, f, \mathbf{T})$ is an *elementary* landscape.

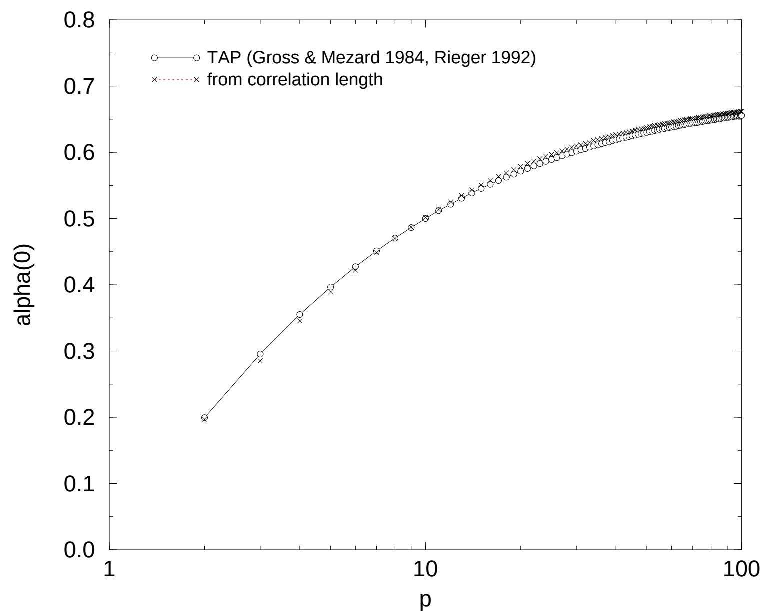
Correlation Length: $\ell = \sum_{s=0}^{\infty} r(s)$.

Associated *Laplacian matrix*: $-\mathbf{\Delta} = \mathbf{A} - \mathbf{D} = (\mathbf{T} - \mathbf{I})\mathbf{D}$

where $\mathbf{D}$ is the diagonal matrix of vertex degrees.

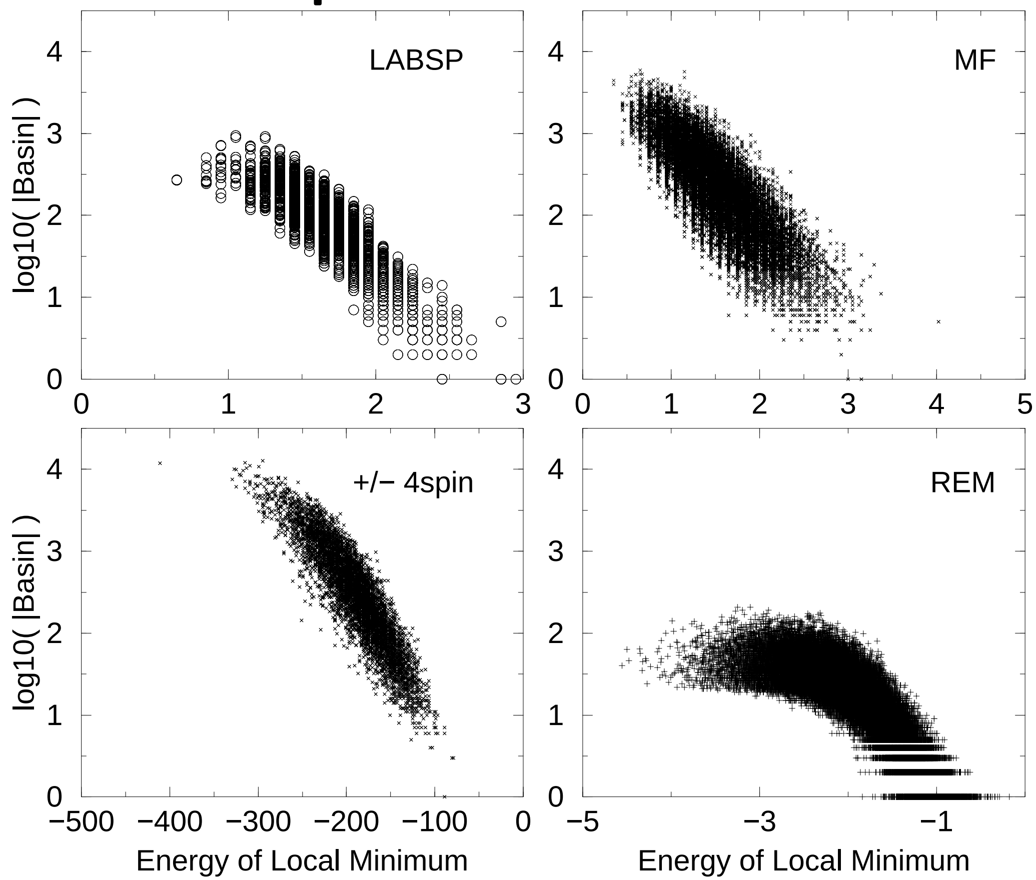
Ruggedness and Local Optima

Heuristic approximation: There are only a few local optima within a ball whose radius R is determined by the distance $d(x_0, x_\ell)$ that is reached by a random walk of length ℓ (correlation length).



$$\alpha(0) = \lim_{n\to\infty} \frac{1}{n} \ln \mathbb{E}[\mathcal{N}]$$

Local Optima and Basin Sizes



Nodal Domains

Let $f : V \to \mathbb{R}$ be an arbitrary function on a graph $G(V, E)$.

A *strong nodal domain* is a maximal connected induced subgraph $G[W]$ of G such that either $f(x) > 0$ for all $x \in W$ or $f(x) < 0$ for all $x \in W$.

A *weak nodal domain* is a maximal connected induced subgraph $G[W]$ of G such that either $f(x) \geq 0$ for all $x \in W$ or $f(x) \leq 0$ for all $x \in W$.

Each strong nodal domain is contained in a weak nodal domain.

Eigenfunction of Laplacians

$-\Delta f = 0$ on V implies that f is constant.

Theorem. (Grover)

If x_{min} and x_{max} is a local minimum and a local maximum, resp., then $f(x_{\mathrm{min}}) \leq 0 \leq f(x_{\mathrm{max}})$.

Suppose the eigenvalue of $-\Delta$ are ordered in the form $0 = \Lambda_1 \leq \Lambda_2 \leq \Lambda_3 \cdots \leq \Lambda_N$.

Theorem. (Courant for manifolds, Davies *et al.* 2001 for graphs)

If $-\Delta\psi = \Lambda_k \psi$ then ψ has

at most k **weak nodal domains** and

at most $k + \mathrm{mult}(\Lambda_k) - 1$ **strong nodal domains**.

Some Elementary Landscapes $-\Delta f = \Lambda(f - \overline{f})$

Problem	Move Set	D	Λ	ℓ/n	order
p-spin	Hamming	n	$2p$	$1/(2p)$	p
additive	Hamming	n	2	1/2	1
NAES	Hamming	n	4	1/4	2
WP	Hamming	n	4	1/4	2
GC	Hamming	$(\alpha-1)n$	2α	$(1-1/\alpha)/2$	2
XY-Hamiltonian	Hamming	$(\alpha-1)n$	2α	$(1-1/\alpha)/2$	2
	cyclic	$2n$	$8\sin^2(\pi/\alpha)$	$1/[4\sin^2(\pi/\alpha)]$	2
GBP	Exchange	$n^2/4$	$2(n-1)$	$1/8 \cdot n/(n-1)$	2
	Hamming	n	4	1/4	2
Linear Assignment	Transposition	$n(n-1)/2$	n	1	1
symmetric TSP	Transposition	$n(n-1)/2$	$2(n-1)$	1/4	2
	Inversions	$n(n-1)/2$	n	$(1-1/n)/4$	2
GMP	Transposition	$n(n-1)/2$	$2(n-1)$	$n/4$	2

NAES = Non-All-Equal-Satisfiability, WP = Weight Partitioning,
GC = Graph Coloring with α colors, GBP = Graph Bipartitioning,
TSP = Traveling Salesman Problem, GMP = Graph Matching Problem
XY-Hamiltonian: $\sum_{i<j} J_{ij} \cos(\frac{2\pi}{\alpha}(x_i - x_j))$

Elementary Landscape on the Boolean Hypercube

Spectral properties of $-\Delta$ well known:

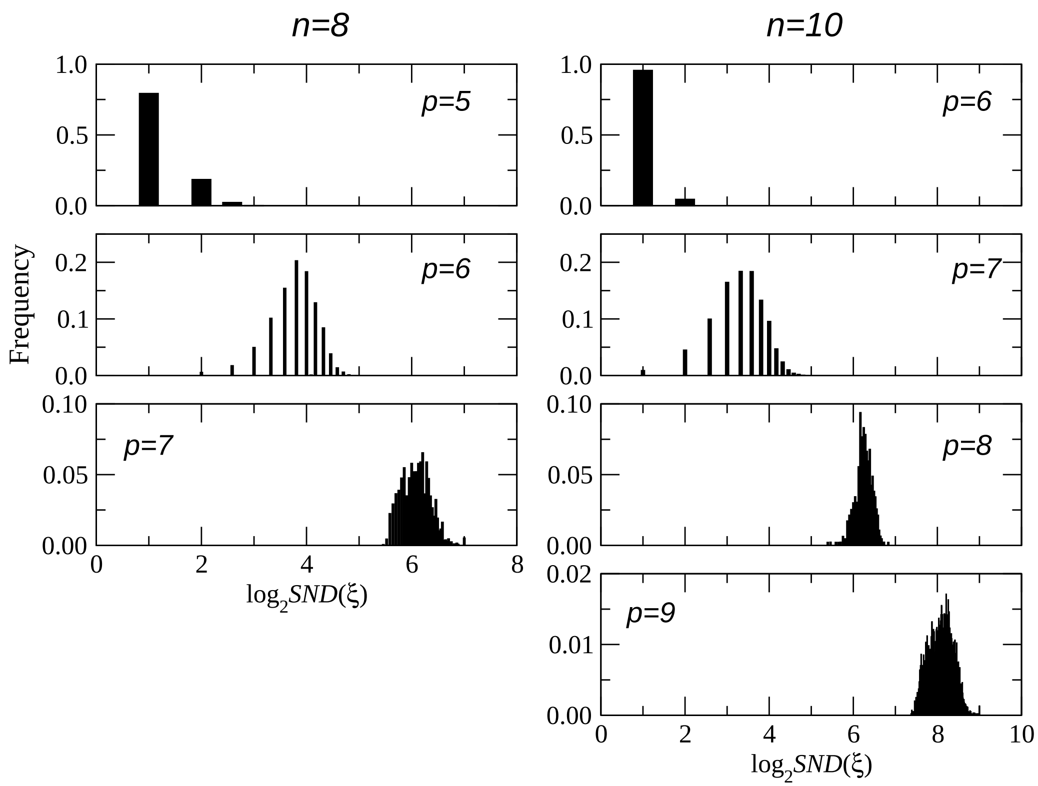
Eigenfunctions are the Walsh or p-spin functions

$\varepsilon_I(x) = \prod_{k\in I} x_k$ with eigenvalues eigenvalue $\Lambda_I = 2|I|$ and multiplicity $\mathrm{mult}(\Lambda_I) = \binom{n}{|I|}$.

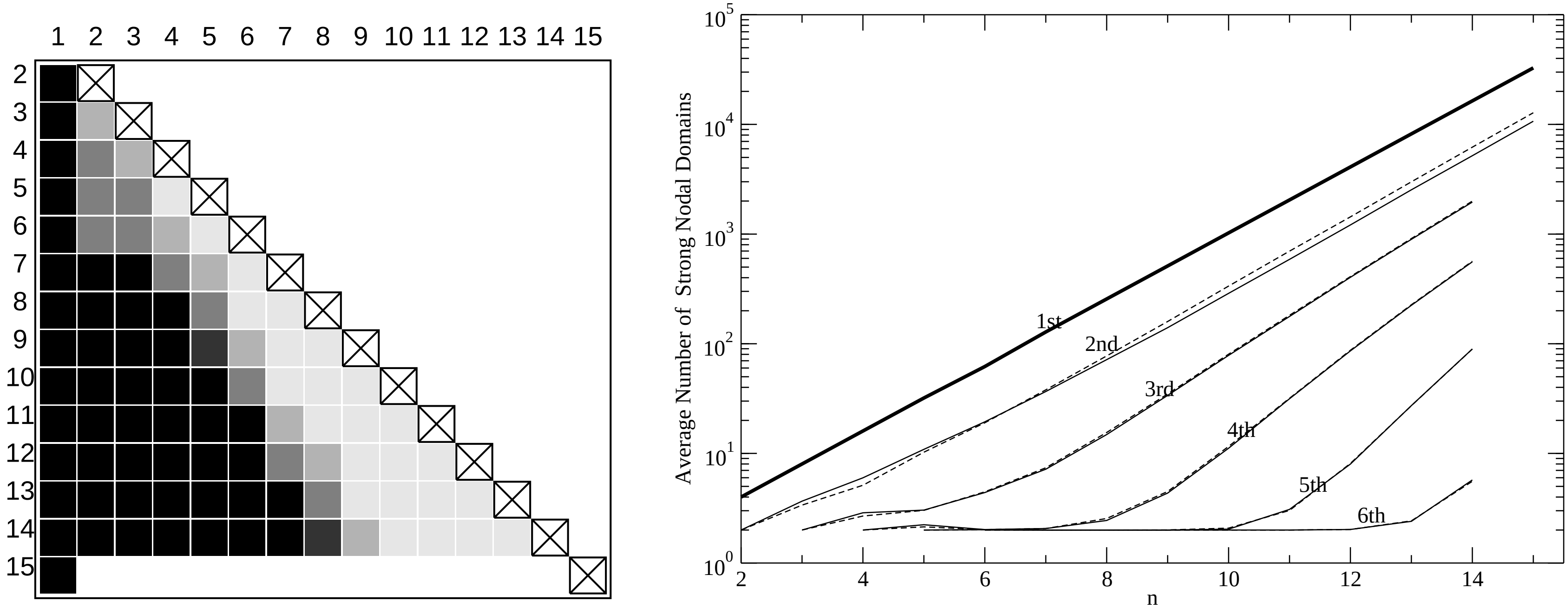
Here $I \subseteq \{1,\dots,n\}$ and $p = |I|$ is the *interaction order*.

Question: What do we know about an eigenfunction f with $-\Delta f = 2pf$.

p-spin functions $\qquad f(x) = \sum_{J:\,|J|=p} a_J \prod_{k\in J} x_k$



Expected number of strong nodal domains for i.i.d. Gaussian coefficients a_J.



Average number of nodal domains for the eigenvectors of the hypercubes with $n = 2$ to 15 as a function of p. 2 domains ■. 2^n nodal domains ⊠. Grey: $2-3$, $3-10$, and more than 10 nodal domains on average.

Fourier Transform on Graphs and Groups

Quasiabelian Cayley graphs: Let G be a group with identity $\imath$ and $S \subset$ G such that $x \in S \implies x^{-1} \in S$, $\imath \notin S$. The Cayley graph $\Gamma(\mathsf{G}, S)$ has vertex set G and edges $\{x, y\}$ iff $xy^{-1} \in S$. One says that $\Gamma(\mathsf{G}, S)$ is *quasi-abelian* if S is a conjugacy class.

Theorem. Let $\Gamma(\mathsf{G}, S)$ be a quasi-abelian Cayley graph with a finite group G.

(i) The function $\varepsilon_{ij}^k : \mathsf{G} \to \mathbb{C}$ defined as

$$\varepsilon_{ij}^k(u) = \frac{1}{\sqrt{|\mathsf{G}|}} \tilde{\rho}_{ij}^k(u) = \sqrt{\frac{d_k}{|\mathsf{G}|}} \boldsymbol{\rho}_{ij}^k(u^{-1}) \tag{1}$$

is an eigenvector of $\mathbf{A}(\Gamma)$ with eigenvalue $\Lambda_k = \frac{1}{d_k} \sum_{s \in S} \chi_k(s)$ where $\chi_k(s) = \mathrm{Tr}\boldsymbol{\rho}^k(s)$ is the character of ϱ^k at s.

(ii) All quasi-abelian Cayley graphs on G have a common basis of eigenvectors and hence their adjacency matrices commute.

(iii) A function $f : \mathsf{G} \to \mathbb{R}$ can be expanded in the form

$$f(s) = \sum_{ijk} a_{ij}^k \varepsilon_{ij}^k(s) \qquad \text{with} \qquad a_{ij}^k = \sqrt{\frac{d_k}{|\mathsf{G}|}} \widehat{f}_{ji}(\rho^k). \tag{2}$$

Amplitude Spectra

Let $\{\varphi_k\}$ be an orthonormal basis of eigenvectors of the Laplacian $-\Delta$. We will be interested in the "Graph Fourier Transform" $\{a_k\}$ where

$$f(x) = \sum_k a_k \varphi_k(x)$$

Collect terms that belong to the *same eigenspace*:

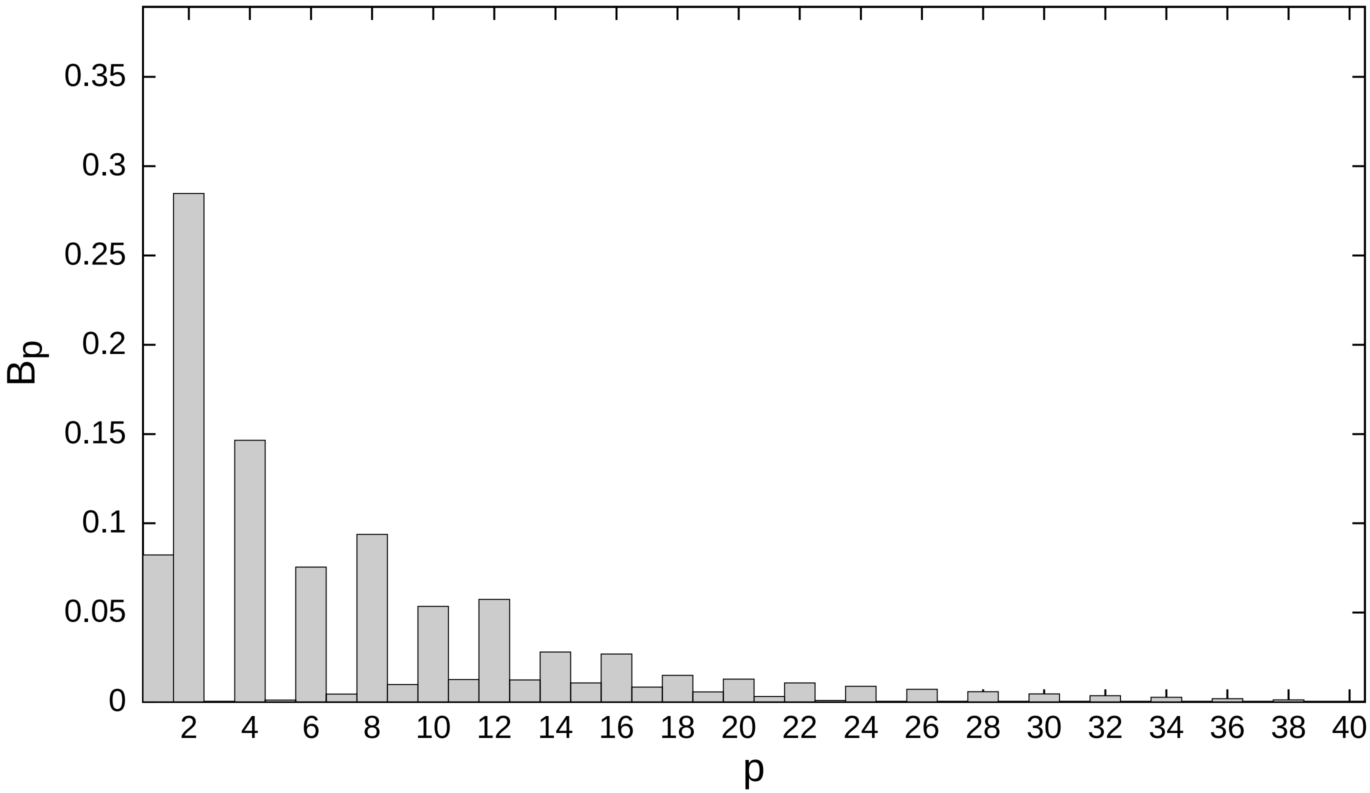
$$B_p = \frac{1}{\mathrm{var}[f]} \sum_{k: -\Delta\varphi_k = \Lambda_p \varphi_k} |a_k|^2$$

$\Longrightarrow$ *"Amplitude Spectrum of f"*.

Autocorrelation: $r(s) = \sum_{p \neq 0} B_p (1 - \Lambda_p / D)^s$

Correlation length $\ell := \sum_{s=0}^{\infty} r(s) = D \sum_{p \neq 0} B_p / \Lambda_p$

An RNA Example



GC alphabet, sequence length $n = 100$

Energy barriers

$$E[s,w] = \text{min}\left\{\text{max}\left[f(z)\middle|z \in \mathbf{p}\right] \middle| \mathbf{p} : \text{path from } s \text{ to } w\right\},$$

$$B(s) = \text{min}\left\{E[s,w] - f(s)\middle|w : f(w) < f(s)\right\}$$

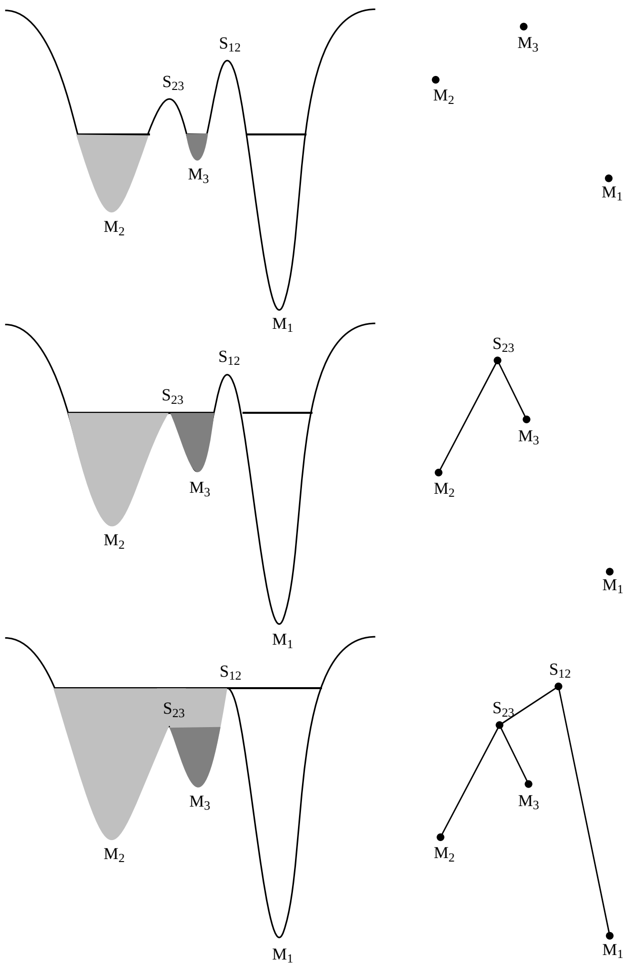
Depth and Difficulty

(borrowed from simulated annealing theory)

$$\mathsf{D} = \text{max}\left\{B(s)\middle|s \text{ is not a global minimum }\right\}$$

$$\psi = \text{max}\left\{\frac{B(s)}{f(s) - f(\text{min})}\middle|s \text{ is not a global minimum}\right\}$$

Calculating barrier trees



The flooding algorithm:

Read conformations in energy sorted order.

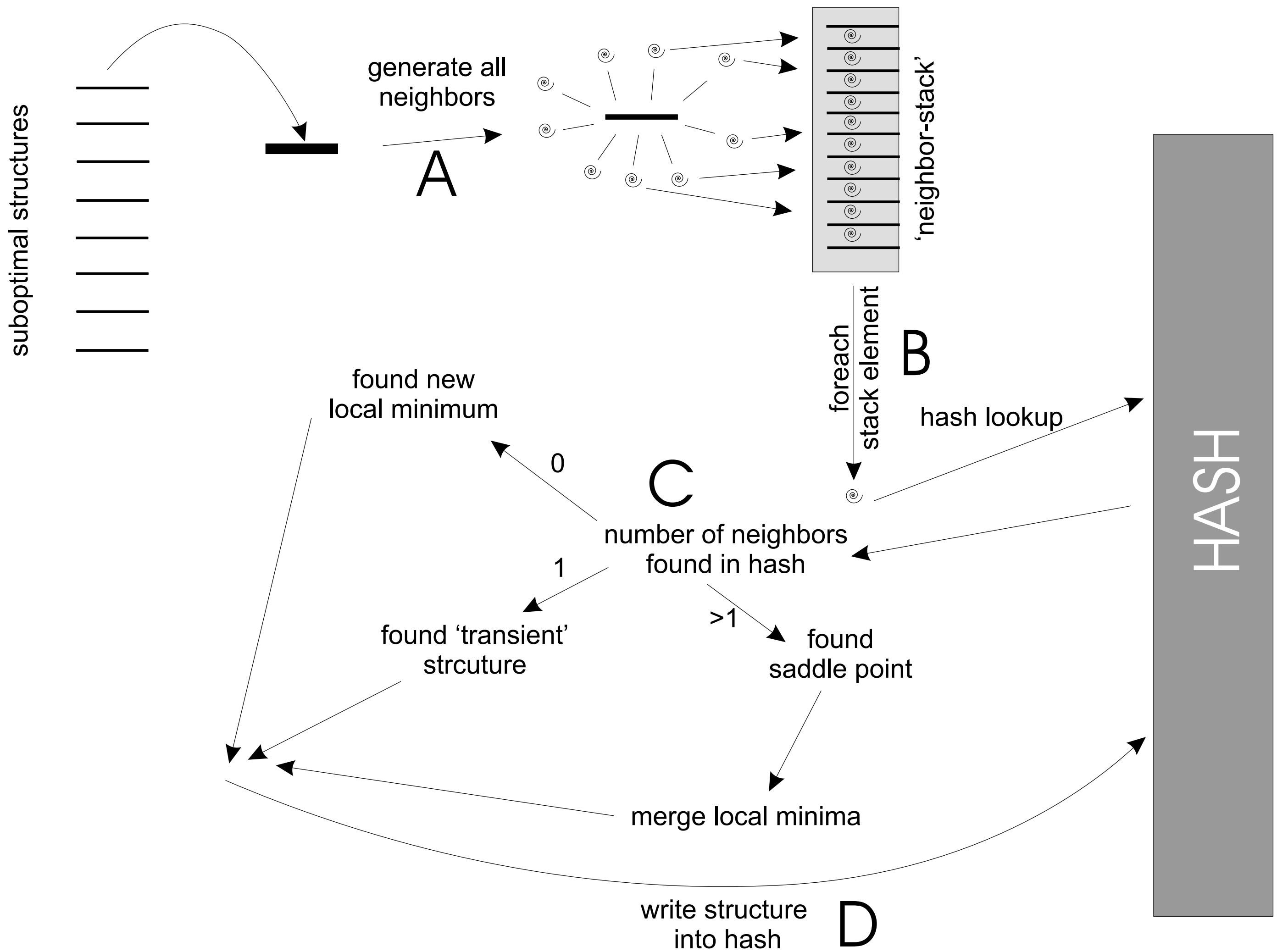
For each confirmation x we have three cases:

(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 $B(s_1) \ldots B(s_k)$ then it's a saddle point that *merges* these basins.
Basins $B(s_1), \ldots, B(s_k)$ are then united and are assigned to the deepest of local minimum.

The algorithm of `barriers`

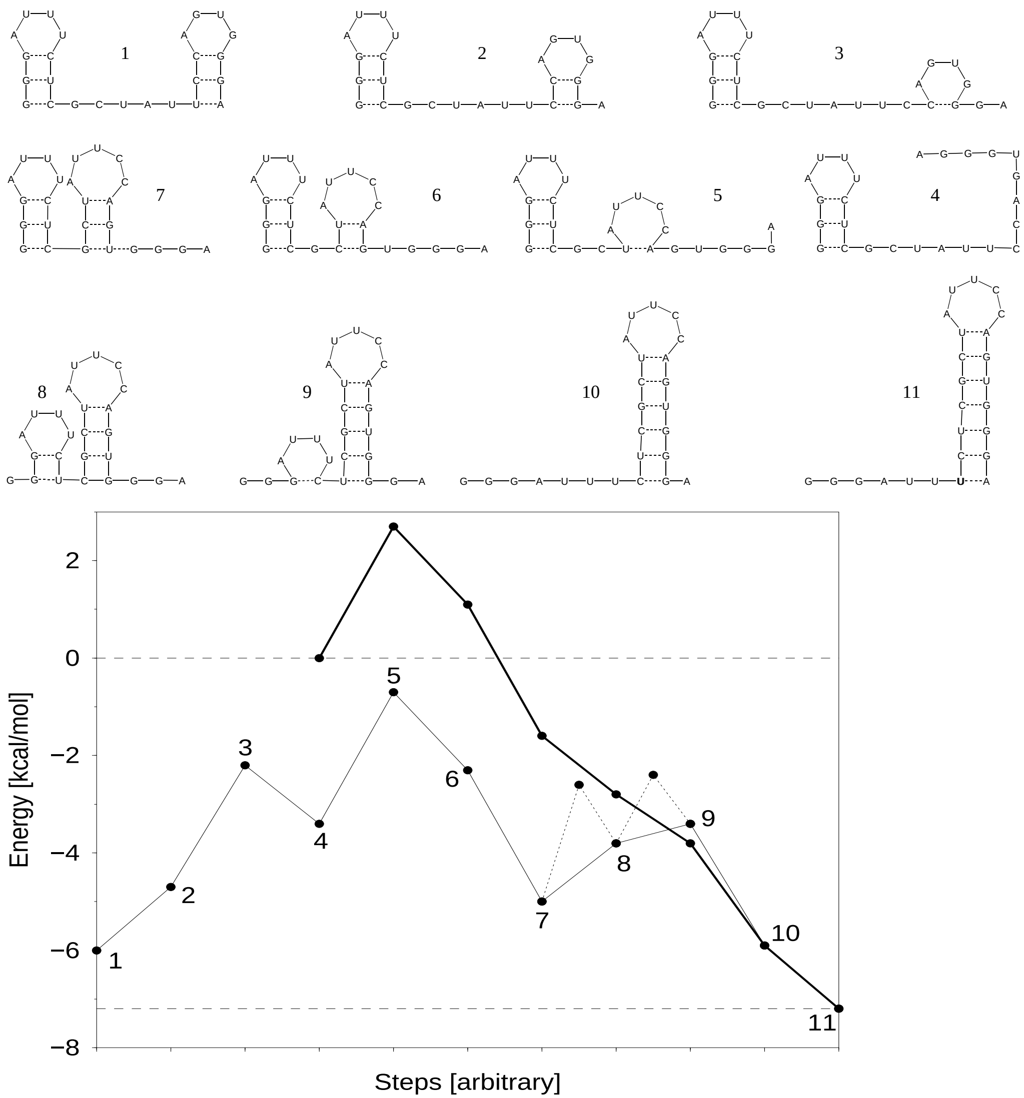


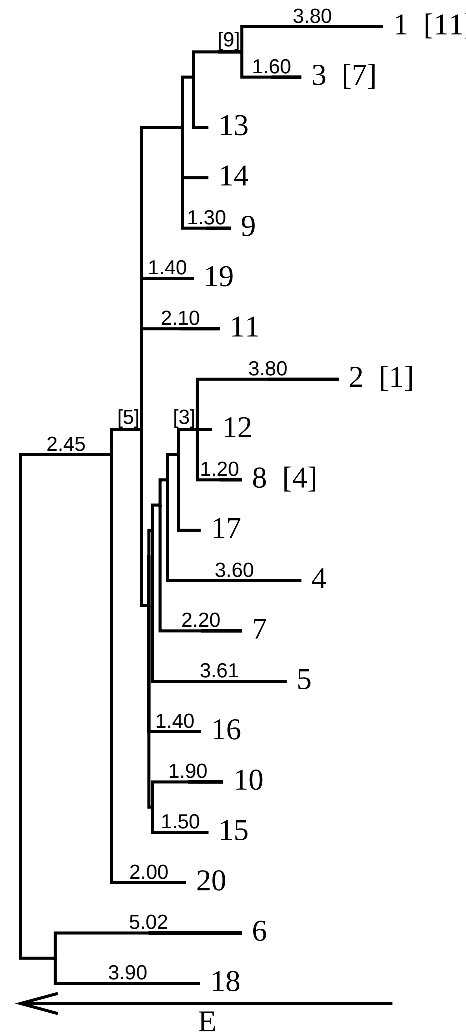
Information from the Barrier Trees

- Local minima
- Saddle points
- Barrier heights
- Gradient basins
- Partition functions and free energies of (gradient) basins
- Depth and Difficulty of the landscape

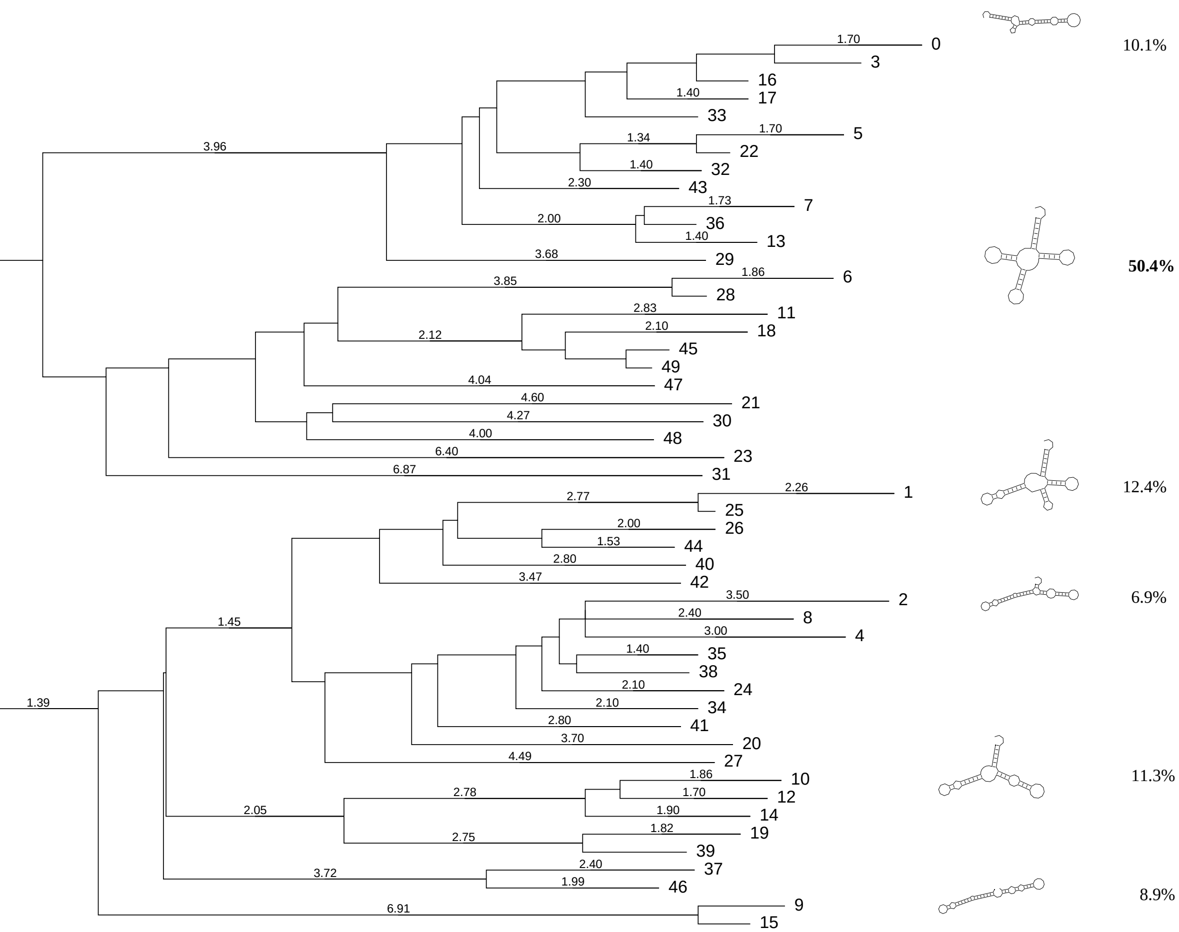
N.B.: A *gradient basin* is the set of all initial points from which a gradient walk (steepest descent) ends in the same local minimum.

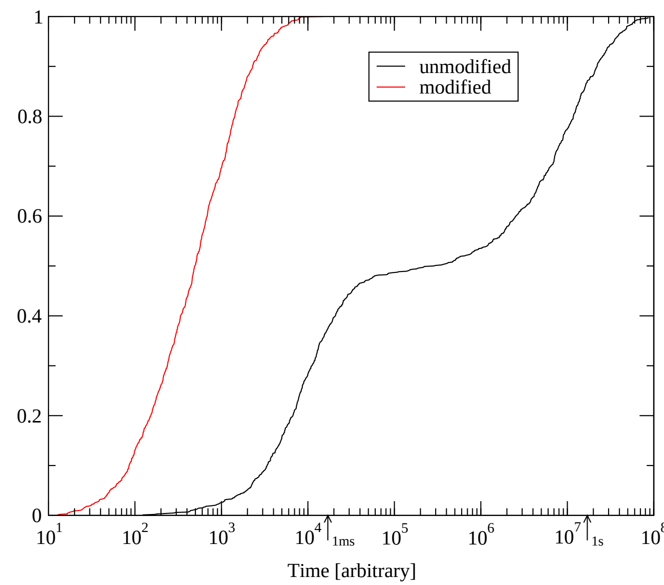
Energy Landscape of a Toy Sequence





Barrier Tree for $tRNA^{phe}$





Folding Kinetics

Transition rates from x to y:

$$r_{yx} = r_0 e^{-\frac{E^{\neq}_{yx} - E(x)}{RT}} \quad \text{for } x \neq y$$
$$r_{xx} = -\sum_{y \neq x} r_{yx}$$

Kinetics as a Markov process:

$$\frac{\mathrm{d}p_x}{\mathrm{d}t} = \sum_{y \in X} r_{xy} p_y(t)\,.$$

Transition states:

$$E^{\neq}_{yx} = \max\{E(x), E(y)\}$$

or more complex models (Tacker et al 1994, Schmitz et al 1996)

Reduced Description of the Folding Dynamics

Macrostates = Classes of a partition of the state space.

Partition function for a macro state:

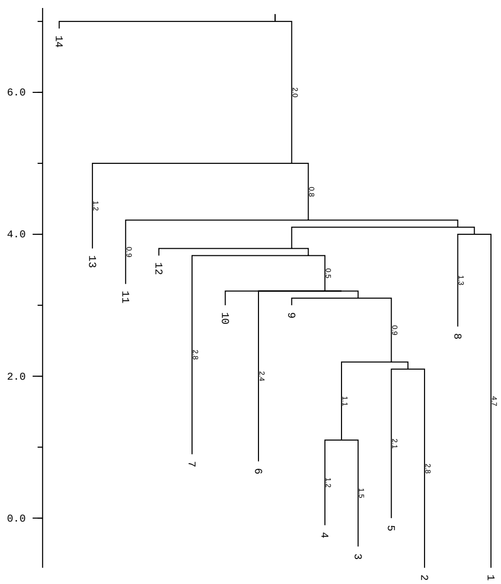
$$Z_\alpha = \sum_{x\in\alpha} \exp(-E(x)/RT)$$

Free energy of a macro state: $G(\alpha) = -RT \ln Z_\alpha$

$$r_{\beta\alpha} = \sum_{y\in\beta}\sum_{x\in\alpha} r_{yx}\,\mathrm{Prob}[x|\alpha] = \frac{1}{Z_\alpha}\sum_{y\in\beta}\sum_{x\in\alpha} r_{yx} e^{-E(x)/RT} \quad \text{for } \alpha \neq \beta$$

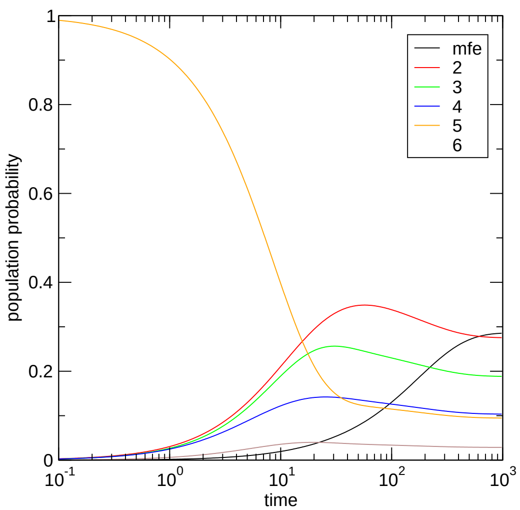
compute $r_{\beta\alpha}$ "on flight" while executing the `barriers` program.

Transition state free energy: $G^{\neq}_{\beta\alpha} = -RT \ln \sum_{y\in\beta}\sum_{x\in\alpha} e^{-\frac{E^{\neq}_{xy}}{RT}}$

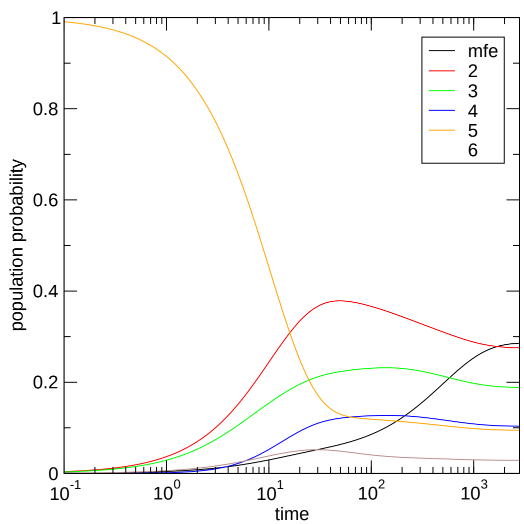


(a)

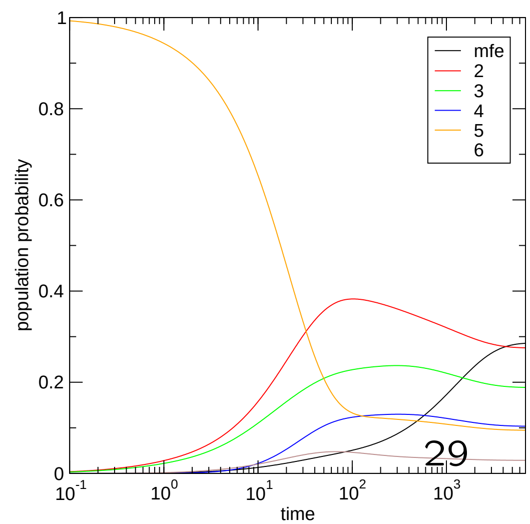
lilly: (a) barrier tree, (b) tree model, (c) transition rates of macro-states, (d) explicit computation with all states



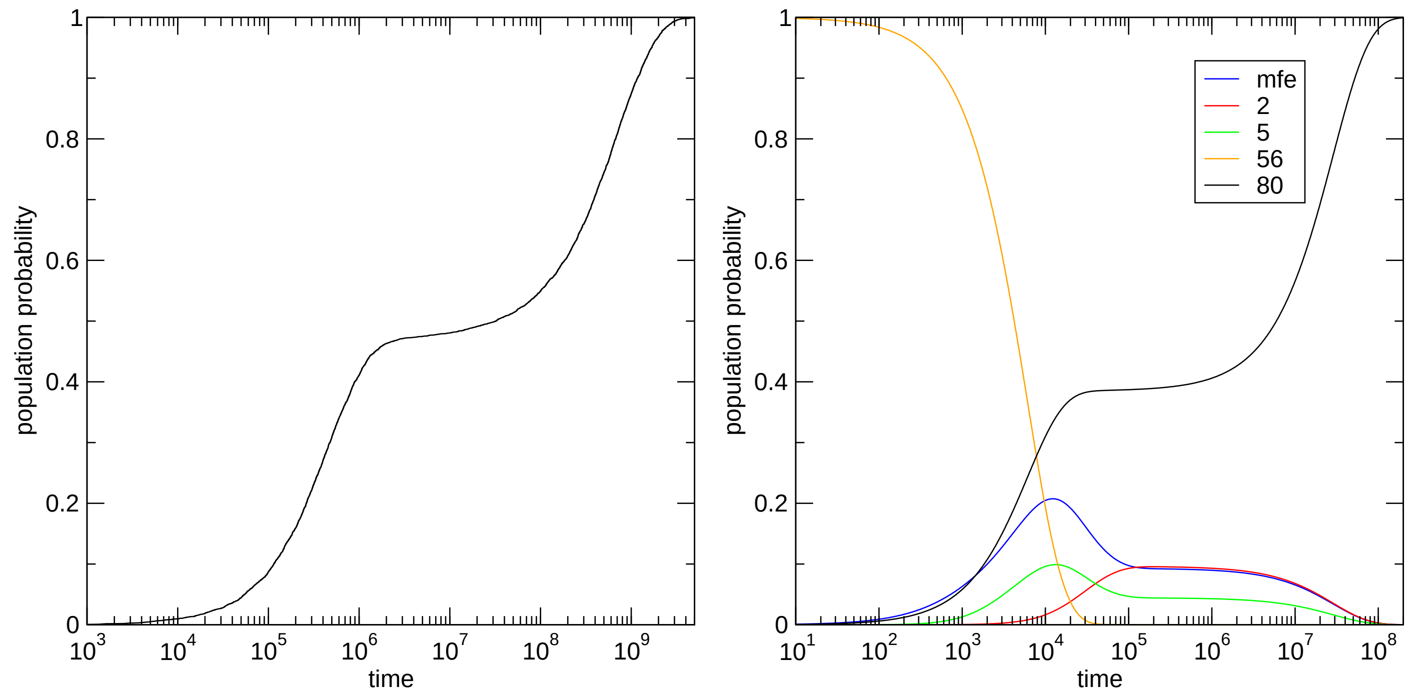
(b)



(c)

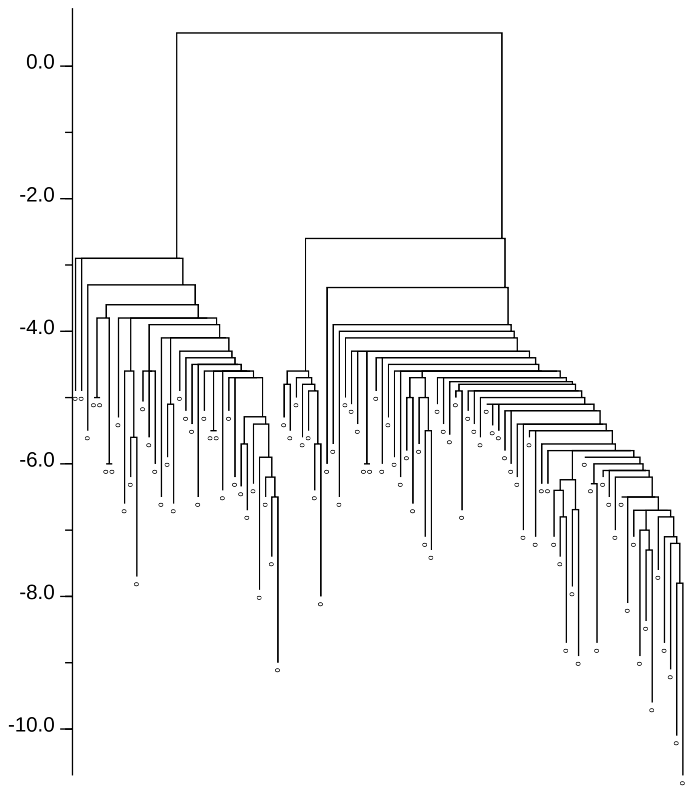


(d)

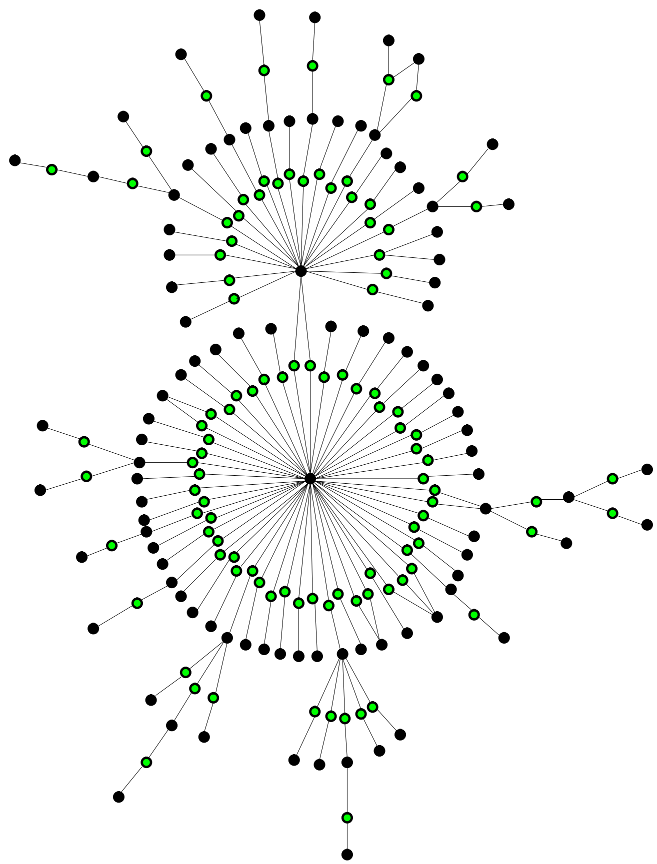


Refolding of a tRNA molecule.

More Topological Information: Merging Graphs



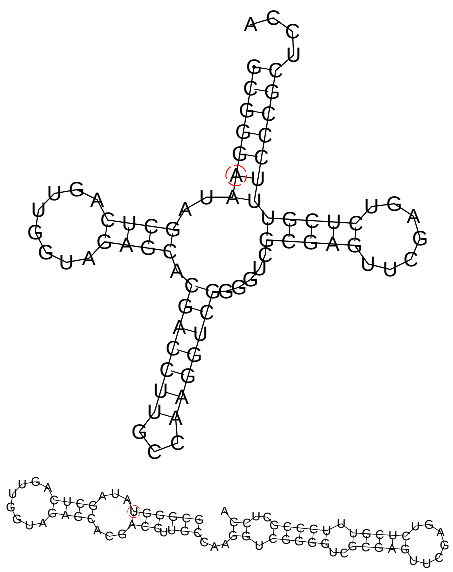
barrier tree



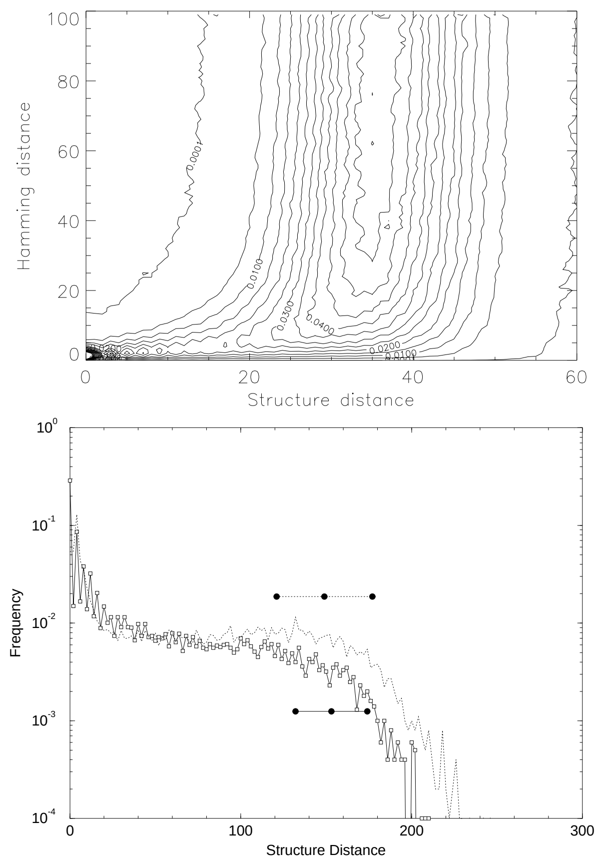
merging graph

tiny SL RNA

Sensitivity and Neutrality



Effect of a single point mutation



Distribution of structure distances

Neutrality

Neutrality = a substantial fraction of neighboring configurations with the *same* fitness.

Important feature of biopolymer landscapes.

Neutral neighbors:

$$\nu_x = \Big|\{y \in \partial x : f(x) = f(y)\}\Big|$$

Simple model: *additive landscapes*

$$f(x) = \sum_i c_i \vartheta_i(x)$$

with i.i.d. random coefficients c_i satisfying Prob[$c_j = 0$] = μ_0 and otherwise smooth distribution of the coefficients.

Examples: **short range spin glasses**.

Degree of neutrality

$\mathbb{E}[\nu_x]$ = expected number of neutral neighbors.

Suppose $\mu_0 > 0$. Then

$$\mathbb{E}[\nu_x] = \sum_{y:y\in\partial x} \mu_0^{c_x(y)}$$

where

$$c_x(y) = \Big|\{j|\vartheta_j(x) \neq \vartheta_j(y)\}\Big| \qquad y \in \partial x$$

Choose $\vartheta_j(y)$ as elementary landscapes with **given** correlation. Then μ_0 can be varied independent of the correlation.

$\Longrightarrow$ Correlation length ℓ and degree of neutrality ν_x can be tuned independently of each other in short range p-spin models.

Summary

Many combinatorial optimization problems are eigenfunctions of graph Laplacians with small eigenvalues

Ruggedness is closely linked to the Laplacian spectrum

Barrier Trees allow a direct link between landscape structure and dynamical properties

Neutrality and Ruggedness are independent properties

Thanks!

- Peter Schuster, Bärbel Stadler, Ivo Hofacker, Christoph Flamm, Walter Fontana, and many others @ U. Vienna and SFI
- Josef Leydold, Türker Bıyıkoğlu, Wim Hordijk (WU Wien)
- Günter Wagner (Yale), Rudi Seitz (MIT)
- Dan Rockmore, Peter Kestelec (Dartmouth),
- Christian Reidys (Los Alamos) Neutrality
- Ricardo García-Pelayo (Cd. Mexico), Viviane Oliveira, J. Fernando Fontanari (USP São Carlos)
- Tomaž Pisanski (Ljubljana)
- Gottfried Tinhofer, Oliver Bastert (TU München).

Some References

Combinatorial Landscapes, CM Reidys, PF Stadler. *SIAM Review* 44: 3-54 (2002).

Barrier Trees of Degenerate Landscapes, C Flamm, IL Hofacker, PF Stadler, MT Wolfinger. *Z.Phys.Chem.* 216: 155-173 (2002)

Discrete Nodal Domain Theorems EB Davies, GML Gladwell, J Leydold, PF Stadler *Lin.Alg.Appl.* 336: 51-60 (2001)

Neutrality in Fitness Landscapes CM Reidys, PF Stadler *Appl. Math. & Comput.* 117: 321-350

Random Field Models for Fitness Landscapes, PF Stadler, R Happel. *J.Math.Biol.* 38: 435-478 (1999) *RNA in silico: The Computational Biology of RNA Secondary Structures*, C Flamm, IL Hofacker, PF Stadler. *Adv. Complex Syst. 2: 65-90 (1999)*

See also `http://www.tbi.univie.ac.at/` for the preprint versions of these papers and further work in this direction