Chance and Randomess in Evolutionary Processes

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria and The Santa Fe Institute, Santa Fe, New Mexico, USA



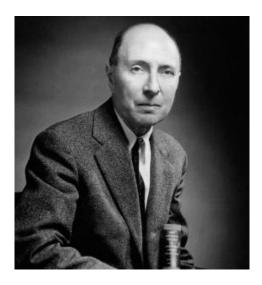
Concept of Probability in the Sciences ESI Wien, 29.– 30.10.2018 Web-Page for further information:

http://www.tbi.univie.ac.at/~pks

- 1. Is evolution possible?
- 2. "Non-probabilities"?
- 3. Protein folding a(n almost) solved example
- 4. Evolution The survival of the fittest?
- 5. Genotype-phenotype mapping and evolution
- 6. The interplay of adaptation and random drift
- 7. Natural selection and evolution

1. Is evolution possible ?

- 2. "Non-probabilities"?
- 3. Protein folding a(n almost) solved example
- 4. Evolution The survival of the fittest?
- 5. Genotype-phenotype mapping and evolution
- 6. The interplay of adaptation and random drift
- 7. Natural selection and evolution



Eugene P. Wigner 1902-1995

"Assembling elaborate structures with specific functions through **random** events is impossible."

> Statement used as argument against Darwinian evolution and in the context of a terrestrial origin of life



Fred Hoyle, 1915-2001

The argument is neither correct nor incorrect as long as it is not clearly said what is meant by **random**?

Three well-known different degrees of randomness are used, e.g., (i) in random numbers, (ii) in random walks, and (iii) in targeted random paths.

Eugene Wigner's or Fred Hoyle's argument applied to a bacterium:

All genomes have equal probability and all except one have no survival value or are lethal.

5'-end GCGGATTTAGCTCAGTTGGGAGAGCGCCAGACTGAAGATCTGGAGGTCCTGTGTTCGAUCCACAGAATTC......GCACCA 3'-end

Alphabet size: 4 Chain length: $\approx 1\ 000\ 000\ nucleotides$ Number of possible genomes: $4\ ^{1\ 000\ 000}$ Probability to find a given bacterial genome: $4^{-1\ 000\ 000} \approx 10^{-\ 600\ 000} = 0.000.....001$

E. Wigner. The probability of the existence of a self-reproducing unit. In: E.Shils, ed. The logic of personal knowledge. Routledge & Kegan Paul, London 1961, pp.231-238

F. Hoyle. The intelligent universe. A new view of creation and evolution. Holt, Rinehart and Winston. New York 1983

Eugene Wigner's and Fred Hoyle's arguments revisited:

Every single point mutation leads to an improvement and is therefore selected

5'-end GCGGATTTAGCTCAGTTGGGAGAGCGCCAGACTGAAGATCTGGAGGTCCTGTGTTCGAUCCACAGAATTC......GCACCA 3'-end

Alphabet size: 4

Chain length: \approx 1 000 000 nucleotides

Length of longest path to the optimum: 3×100000

Probability to find the optimal bacterial genome:

 $0.333.. \times 10^{-6} = 0.000000333..$

Myoglobin: 153 amino acid residues, MW 17.0 kDalton

amino acid sequence:

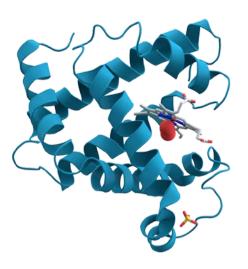
GLSDGEWQLV-LNVWGKVEAD-LAGHGQDVLI-RLFKGHPETL-EKFEKFKHLK-TEADMKASED-LKKHGNTVLT-ALGAILKKKG--HHDAELKPLA-ESHATKHKIP-IKYLEFISEA-IIHVLHSRHP-AEFGADAEGA-MDKALELFRK-DIAAKYKDLG-FHG

A	ala	alanine	I
С	cys	cysteine	к
D	asp	aspartic acid	L
Е	glu	glutamic acid	м
F	phe	phenylalanine	N
G	gly	glycine	P
н	his	histidine	Q

ile	isoleucine
lys	lysine
leu	leucine
met	methionine
asn	asparagine
pro	proline
glu	glutamine

R	arg	arginine
S	ser	serine
т	thr	threonine
v	val	valine
W	trp	tryptophan
Y	tyr	tyrosine

3**D** molecular structure:



Alphabet size: 20 Chain length: 153 amino acid residues Number of possible sequences: $20^{153} \approx 0.11 \times 10^{200}$ Probability to find the native sequence:

 $20^{-153} \approx 8.8 \times 10^{-200}$

Myoglobin – a small protein

1. Is evolution possible ?

2. "Non-probabilities" ?

- 3. Protein folding a(n almost) solved example
- 4. Evolution The survival of the fittest?
- 5. Genotype-phenotype mapping and evolution
- 6. The interplay of adaptation and random drift
- 7. Natural selection and evolution

Probability:
$$P_{(6)} = \frac{6}{45} \times \frac{5}{44} \times \frac{4}{43} \times \frac{3}{42} \times \frac{2}{41} \times \frac{1}{40} \cong 1.23 \times 10^{-7}$$

 $P_{(6)}^{-1} = 8145060$

Maximum number of tips: 52.5×10^6 at January 21, 1991

the Austrian lottery "6 out of 45"

- 1. Is evolution possible?
- 2. "Non-probabilities"?

3. Protein folding - a(n almost) solved example

- 4. Evolution The survival of the fittest?
- 5. Genotype-phenotype mapping and evolution
- 6. The interplay of adaptation and random drift
- 7. Natural selection and evolution

Lysozyme: 129 amino acid residues, MW: 14.4 kDalton

Ι

ĸ

L

М

Ν

Ρ

0

amino acid sequence:

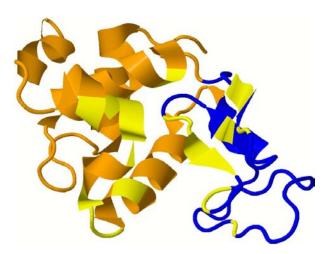
KVFGRCELAA-AMKRHGLDNT-RGYSLGNWVC-AAKFESNFNT-QAYNRNTDGS-TDYGILEINS-RWWCNDGWTP--GSRNLCNIPC-SALLSSDITA-SVNCAKKIVS-DGDGMNAYVA-YRNRCKGTDV-QAWIRGCRL

A	ala	alanine
С	cys	cysteine
D	asp	aspartic acid
Е	glu	glutamic acid
F	phe	phenylalanine
G	gly	glycine
H	his	histidine

ile isoleucine lys lysine leu leucine met methionine asn asparagine pro proline glu glutamine

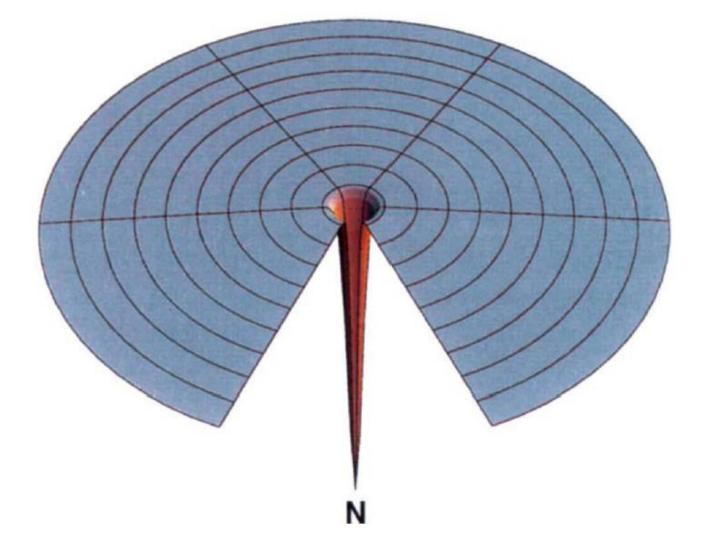
R	arg	arginine
S	ser	serine
т	thr	threonine
v	val	valine
W	trp	tryptophan
Y	tyr	tyrosine

3**D** molecular structure:



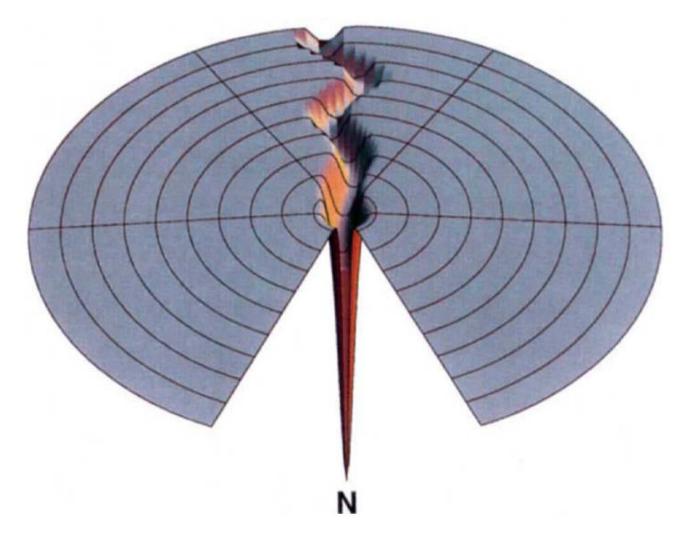
Conformations per amino acid residue: 3 Chain length: 129 amino acid residues Number of possible conformations: $8^{128} \approx 0.39 \times 10^{116}$ Probability to find the native conformation: $8^{-128} \approx 2.5 \times 10^{-116}$

Testing 10^{13} conformation per second it requires 1.3×10^{95} years to complete the search, but proteins of this chain lenghth fold in about a second.



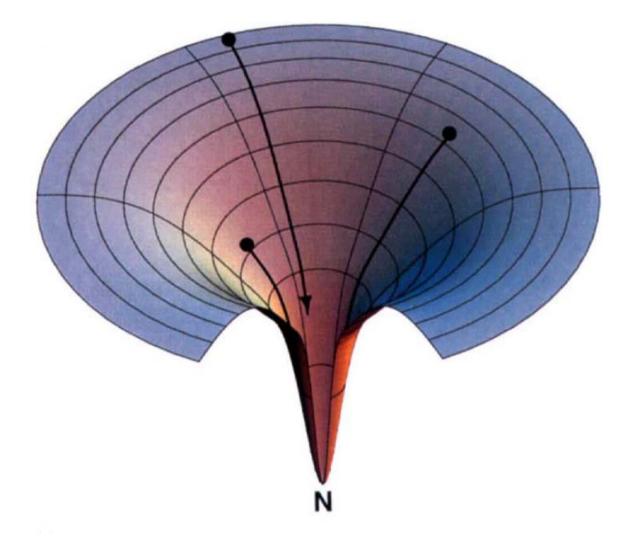
the golf-course landscape

Levinthal's paradox



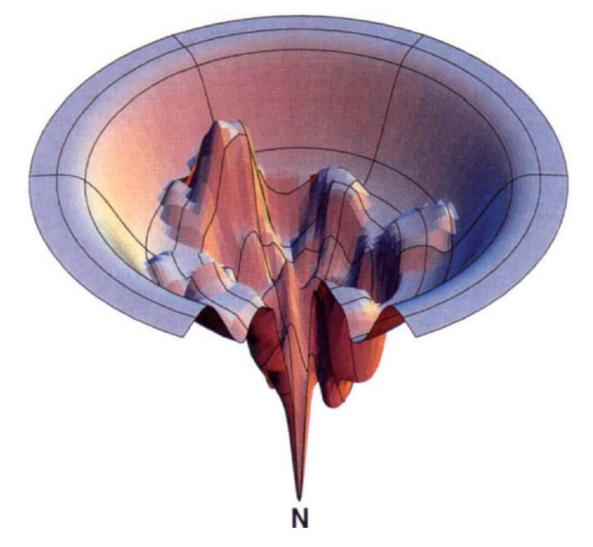
the "pathway" solution

Levinthal's paradox



the funnel landscape

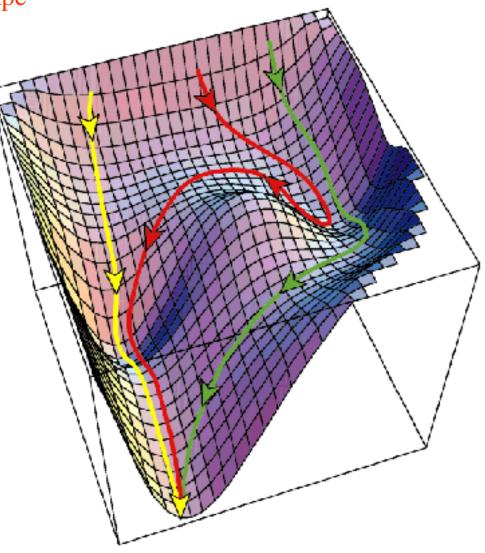
a solution to Levinthal's paradox



the structured funnel landscape

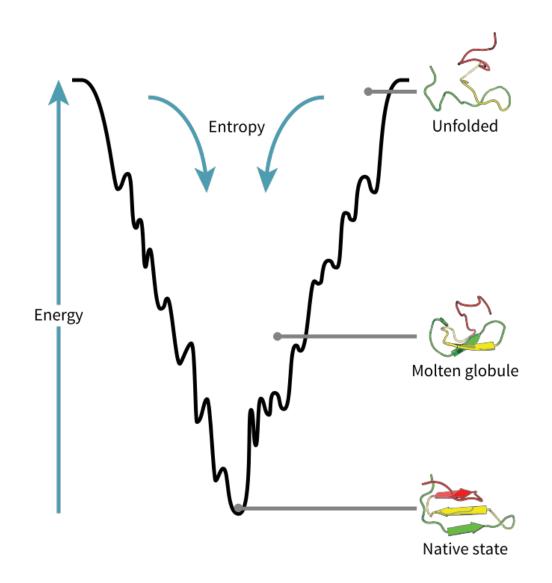
a realistic solution of Levinthal's paradox

An "all-roads-lead-to-Rome" landscape



The reconstructed folding landscape of a real biomolecule: "lysozyme"

Picture: C.M. Dobson, A. Šali, and M. Karplus, Angew.Chem.Internat.Ed. 37: 868-893, 1988



Statistical mechanics of protein folding

J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes. Proteins 21:167-195, 1995

But biological landscapes for biopolymer folding or evolution are high dimensional and much more complex than the toy examples shown here. However, protein and nucleic acid

folding landscapes can be investigated by experiment and evolution under controlled laboratory conditions provides insights into the mechanism of biological evolution.

$$U(\vec{R}) = \sum_{bonds} K_b (b - b_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{Urey-Bradley} K_{UB} (S - S_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{dihedrals} K_\phi (1 + \cos(n\phi - \delta)) + \sum_{impropers} K_\omega (\omega - \omega_0)^2 + \sum_{impropers} K_\omega ($$

Empirical force field for calculations of protein dynamics CHARMM: B.R. Brooks, ..., M. Karplus. J.Comp.Chem. 30:1545-1614, 2009

The origin of energy landscapes in chemistry is the Born-Oppenheimer approximation of quantum mechanics.

Newtonian dynamics on a molecular energy landscape

- 1. Is evolution possible?
- 2. "Non-probabilities"?
- 3. Protein folding a(n almost) solved example

4. Evolution - The survival of the fittest?

- 5. Genotype-phenotype mapping and evolution
- 6. The interplay of adaptation and random drift
- 7. Natural selection and evolution



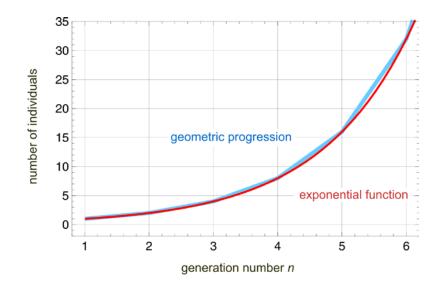


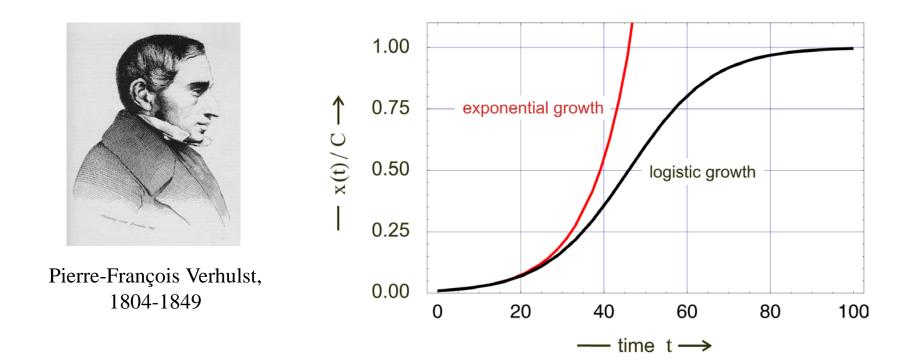
Thomas Robert Malthus, 1766 – 1834

Leonhard Euler, 1717 – 1783

geometric progression

exponential function





population: $\Pi = \{X\}$

the consequence of finite resources

$$\frac{dx}{dt} = f x \left(1 - \frac{x}{C} \right) \implies x(t) = \frac{C x_0}{x_0 + (C - x_0) \exp(-ft)}$$

the logistic equation: Verhulst 1838

Verhulst or logistic equation:

$$\frac{\mathrm{d}x}{\mathrm{dt}} = f\left(1 - \frac{x}{C}\right)x \text{ with } x(0) = x_0$$
$$x(t) = \frac{C x_0}{x_0 + (C - x_0)e^{-rt}}$$

basic structure of the equation:

$$\frac{\mathrm{d}x}{\mathrm{dt}} = \gamma_1 x - \gamma_2 x^2$$

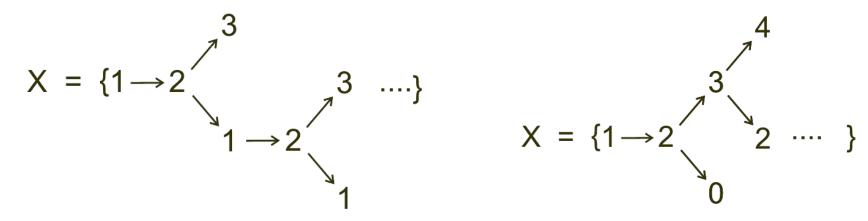
chemical models:

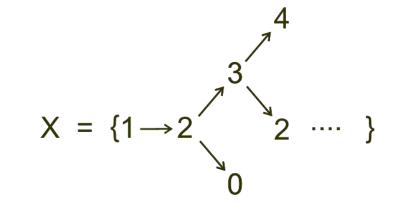
reversible autocatalytic reaction

annihilation reaction

absorbing barrier: $X = 0 \rightarrow dx/dt = 0$

 $A + X \rightarrow 2X$ and $2X \rightarrow A + X$ $A + X \rightarrow 2X$ and $2X \rightarrow \emptyset$

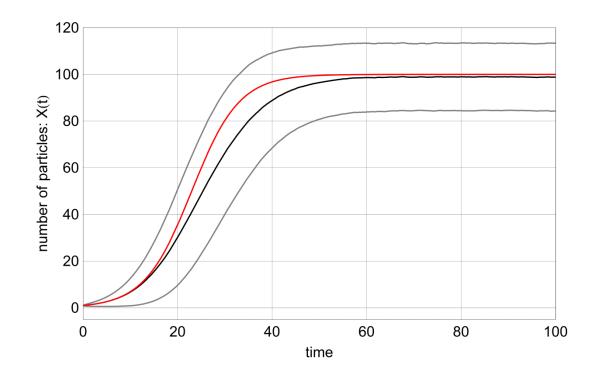




reversible autocatalytic reaction

reflecting barrier

annihilation reaction



logistic growth: $A + X \rightarrow 2X, 2X \rightarrow \emptyset$, expectation value and deterministic solution

	1		2		3		4		5	
	Х	extinct	Х	extinct	Х	extinct	Х	extinct	Х	extinct
numbers	99443	557	99478	522	99988	12	99984	16	99999	1
E±σ	9944.3 ± 7.5	55.7 ± 7.5	9947.8±9.3	52.2 ± 9.3	9998.8±1.0	1.2 ± 1.0	9998.4 \pm 0.8	1.6 ± 0.8	9999.9 ± 0.3	0. 1± 0.3

extinction in the logistic equation: N = 100; f = 0.2; h = 0.001, sample size: 10×10000 .

state of reproduction, S_1 and state of extinction S_0 X: $\lim_{t\to\infty} E(X(t)) = C$ and extinct: $\lim_{t\to\infty} X(t) = 0$

bistability in the logistic equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f x \left(1 - \frac{x}{C} \right) \implies \frac{\mathrm{d}x}{\mathrm{d}t} = f x - \frac{x}{C} f x$$
$$f x \equiv \Phi(t), C = 1: \quad \frac{\mathrm{d}x}{\mathrm{d}t} = x \left(f - \Phi \right)$$

$$X_1, X_2, ..., X_n: [X_i] = x_i; \sum_{i=1}^n x_i = C = 1$$

$$\frac{\mathrm{d}x_{j}}{\mathrm{d}t} = x_{j} \left(f_{j} - \sum_{i=1}^{n} f_{i} x_{i} \right) = x_{j} \left(f_{j} - \Phi \right) \quad ; \quad \Phi = \sum_{i=1}^{n} f_{i} x_{i}$$

Darwin's natural selection

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t} = 2\left(\langle f^2 \rangle - \langle f \rangle^2\right) = 2\operatorname{var}\{f\} \ge 0$$
 survival of the fittest

Generalization of the logistic equation to n variables yields selection.

$$\Pi = \{X_1, \dots, X_n\}$$

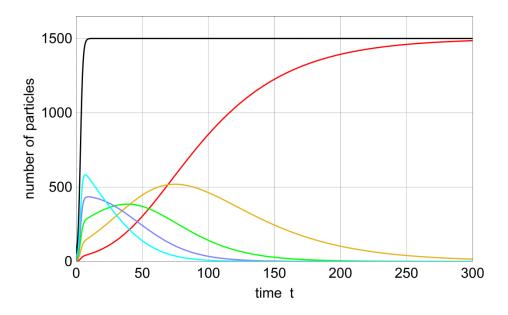
$$N(t) = (N_1(t), \dots, N_n(t)); \quad C(t) = \sum_{j=1}^n N_j(t)$$

$$C(t) = \frac{C(0)K}{C(0) + (K - C(0))e^{-\Phi}}$$
with $\Phi = \int_0^t \phi(\tau) d\tau$ and $\phi(t) = \frac{1}{C(t)} \sum_{i=1}^n f_i N_i(t)$

$$I_{\text{time t}}$$

$$x_{j}(t) = \frac{N_{j}(t)}{C(t)} = \frac{x_{j}(0)e^{f_{j}t}}{\sum_{i=1}^{n} x_{i}(0)e^{f_{i}t}}$$

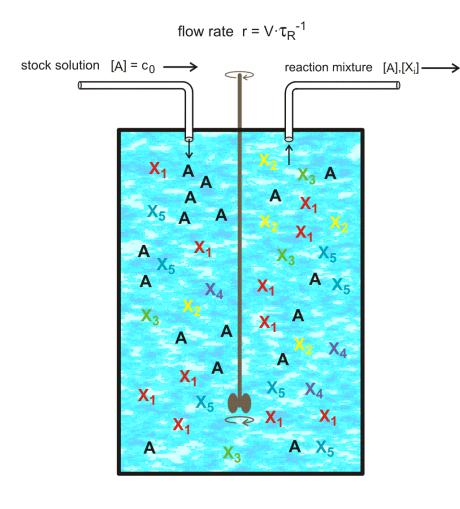
N(0) = (1,4,9,16,25)f = (1.10,1.08,1.06,1.04,1.02)



15

20

population:



$$\Pi = \{X_1, X_2, X_3, \dots, X_n\}$$

$$* \xrightarrow{r \cdot c_0} A$$

$$A + X_j \xrightarrow{f_j} 2X_j, j = 1, 2, \dots, n$$

$$A \xrightarrow{r} \varnothing$$

$$X_j \xrightarrow{r} \varnothing, j = 1, 2, \dots, n$$

$$\frac{da}{dt} = c_0 r - a \left(\sum_{j=1}^n f_j x_j + r\right)$$

$$\frac{dx}{dt} = x (f_j a - r), j = 1, 2, \dots, n$$

selection in the flow reactor

$$P_m(t) = \operatorname{Prob}\left(A(t) = m\right) \text{ and } P_{s_i}(t) = \operatorname{Prob}\left(X_i(t) = s_i\right)$$

$$\frac{\mathrm{d}P_{\mathbf{m}}}{\mathrm{dt}} = c_0 r \Big(P_{(\mathbf{m};m-1)} - P_{\mathbf{m}} \Big) + r \Big((m+1) P_{(\mathbf{m};m+1)} - m P_{\mathbf{m}} \Big) + r \sum_{i=1}^n \Big((s_i + 1) P_{(\mathbf{m};s_i+1)} - s_i P_{\mathbf{m}} \Big) + \sum_{i=1}^n f_i \Big((m+1) (s_i - 1) P_{(\mathbf{m};m+1,s_i-1)} - m s_i P_{\mathbf{m}} \Big)$$

$$\mathbf{m} = (m, s_1, \dots, s_n);$$
 $\mathbf{m}' = (m \pm 1, s_1, \dots, s_n) \equiv (\mathbf{m}; m \pm 1)$ or
 $\mathbf{m}' = (m, s_1, \dots, s_k \pm 1, \dots, s_n) \equiv (\mathbf{m}; s_k \pm 1)$

master equation for reproduction and selection in the flow reactor

Analysis of the solutions of chemical master equations through sampling of trajectories. The pioneering work has been done by Andrej Kolmogorov, Willi Feller, Joe Doob, David Kendall, and Maurice Bartlett.

The American physicist Daniel Gillespie revived the Kolmogoriv-Feller formalism and introduced a popular and highly efficient simulation tool for stochastic chemical reactions.



Daniel T. Gillespie, 1938 -

D.T. Gillespie, Annu.Rev.Phys.Chem. 58:35-55, 2007

In the limit of an infinite number of trajectories the distribution of the trajectory bundle converges to the probability distribution of the corresponding solution of the master equation.

Gillespie simulation of individual trajectories

Springer Series in Synergetics

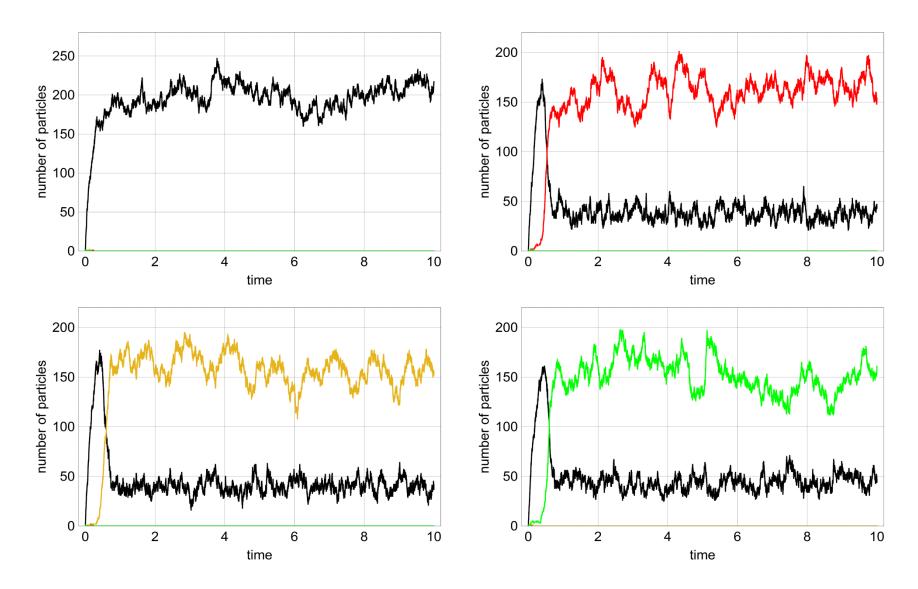
Peter Schuster

Stochasticity in Processes

Fundamentals and Applications to Chemistry and Biology



COMPLEXITY



color code: A , X_1 , X_2 , X_3

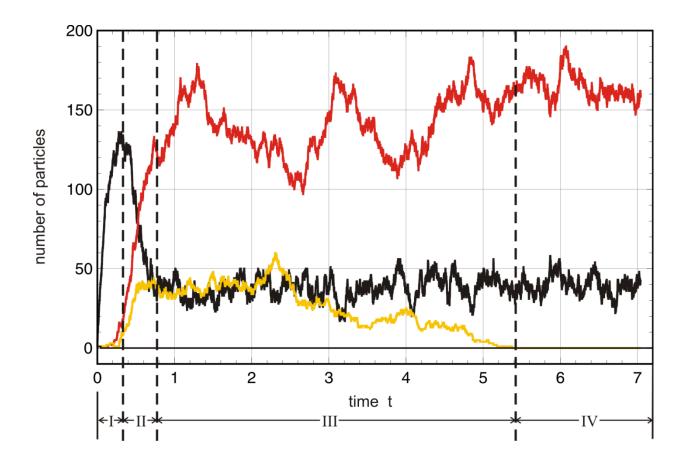
assorted sample of trajectories

		Population size N = 100				Population size N = 200			
$\Delta f/f$	t _e	A(t _e)	X ₁ (t _e)	X ₂ (t _e)	X ₃ (t _e)	$A(t_e)$	X ₁ (t _e)	X ₂ (t _e)	$X_3(t_e)$
0.0	600	1.5 ± 1.3	30.5 ± 3.9	34.2 ± 4.6	33.4 ± 4.1	0.5 ± 0.9	30.6 ± 4.6	<mark>30.9</mark> ±5.0	32.0 ± 4.7
0.02	600	1.8 ± 1.4	41.8 ± 4.8	<mark>32.9</mark> ±3.8	23.4 ± 4.0	0.6 ± 0.8	50.4 ± 5.7	27.7 ± 4.9	17.3 ± 2.6
0.04	400	2.4 ± 2.1	45.4 ± 5.0	31.3 ± 4.5	19.9 ± 2.5	0.7 ± 0.8	58.3 ± 4.6	<mark>25.6</mark> ± 4.5	11.0 ± 2.9
0.1	400	2.1 ± 1.7	59.8 ±5.5	<mark>28.0</mark> ± 4.1	10.0 ± 2.9	0.4 ± 0.5	73.9 ± 4.1	<mark>20.6</mark> ± 3.5	4.8 ± 1.9
0.2	400	1.9 ± 1.1	68.3 ± 4.5	<mark>23.1</mark> ± 3.7	6.7 ± 2.8	0.5 ± 0.7	76.6 ± 4.1	<mark>19.3</mark> ± 2.8	3.6 ± 1.7
0.4	400	2.3 ± 1.8	71.7 ± 6.0	<mark>20.8</mark> ±5.2	5.2 ± 2.4	0.9 ± 0.6	82.0 ± 4.2	13.8 ± 3.8	3.3 ± 1.7
1.0	200	2.7 ± 2.4	78.4 ± 4.7	15.8 ± 3.3	3.1 ± 1.5	0.9 ± 0.9	83.6 ± 4.0	12.6 ± 3.2	2.9 ± 1.5
1.8	200	4.3 ± 1.1	80.8 ± 2.9	13.6 ± 3.1	1.3 ± 1.2	1.5 ± 1.3	83.8 ± 3.3	12.7 ± 2.5	2.0 ± 1.7

n = 3: $X_1, f_1 = f + \Delta f / 2f$; $X_2, f_2 = f$; $X_3, f_3 = f - \Delta f / 2f$; f = 0.1

initial particle numbers: $X_1(0) = X_2(0) = X_3(0) = 1$

probability of selection



phase I: raise of [A];

phase II: random choice of convergence to a quasi-stationary state;

phase III: convergence to the quasi-stationary state;

color code: A, X_1 , X_2 , X_3

phase IV: fluctuations around the values of the quasi-stationary state

phases of the aproach towards steady states by individual trajectories

1 E \blacklozenge V E

neutral evolution in the Moran model

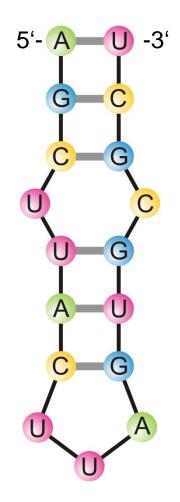
- 1. Is evolution possible?
- 2. "Non-probabilities"?
- 3. Protein folding a(n almost) solved example
- 4. Evolution The survival of the fittest?

5. Genotype-phenotype mapping and evolution

- 6. The interplay of adaptation and random drift
- 7. Natural selection and evolution

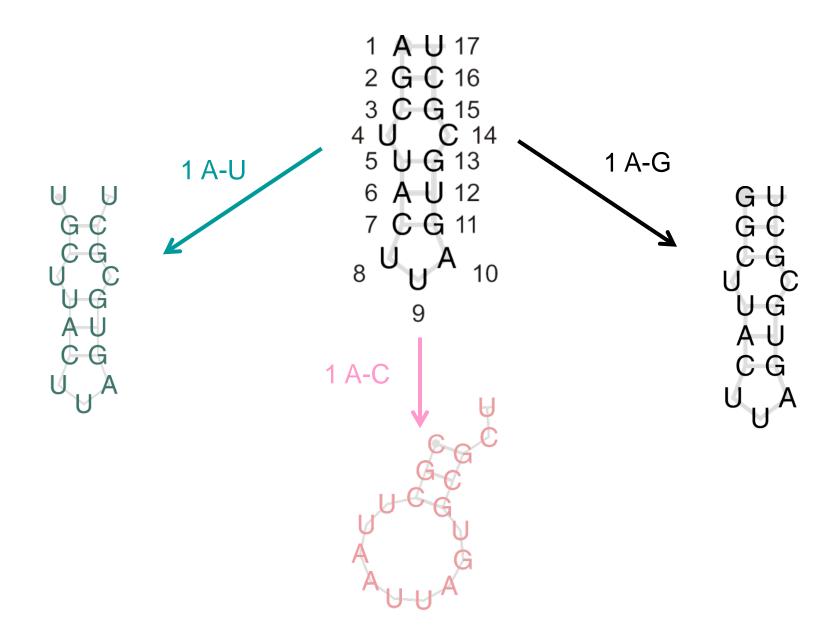
5'- AGCUUACUUAGUGCGCU-3'

5'-((((((((()))))))))-3'



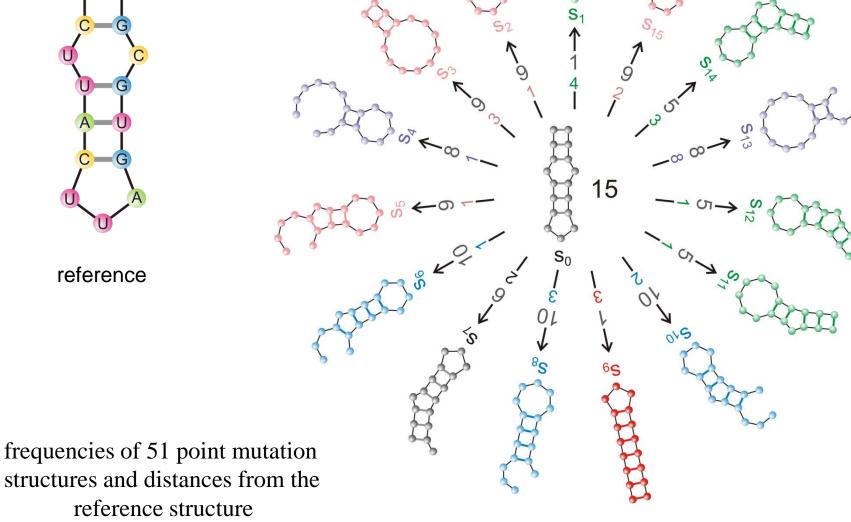
the minimum free energy structure of a small RNA molecule

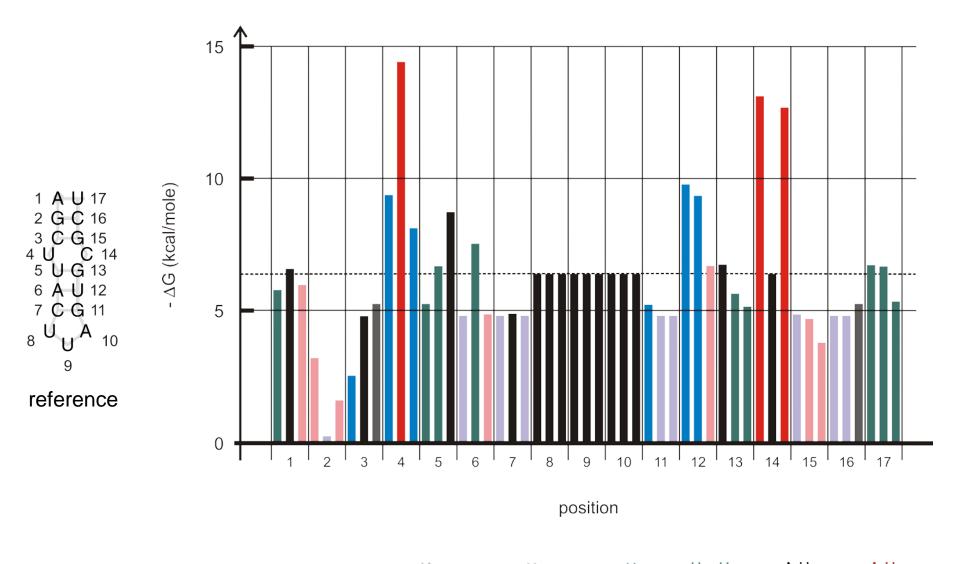
AGCUUAACUUAGUCGCU



AGCUUACUUAGUGCGCU $(((\cdot (((\cdot (((\cdot \cdot \cdot))))))))))$

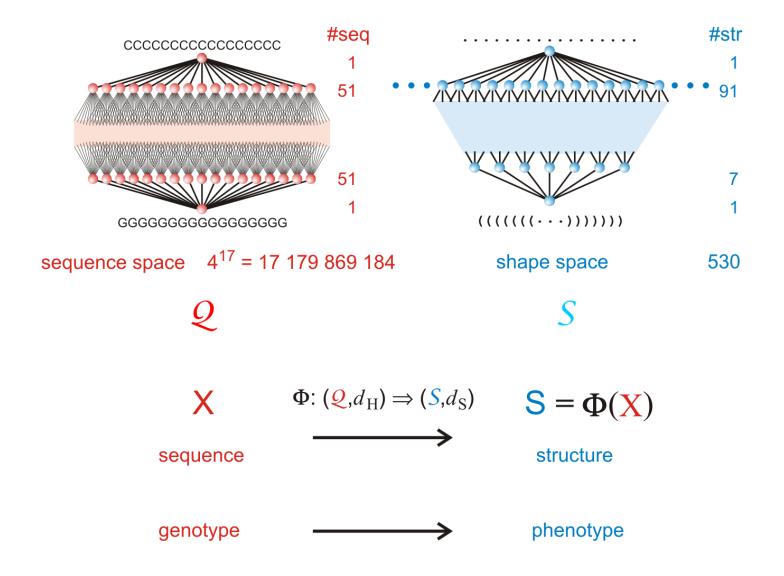




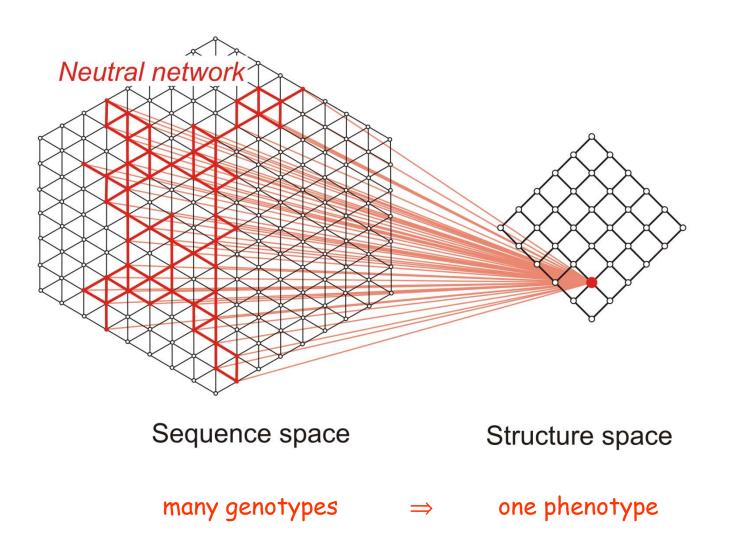


free energy of formation (ΔG_0) of 51 point mutants Of the reference sequence





formation of RNA secondary structures as genotype-phenotype mapping

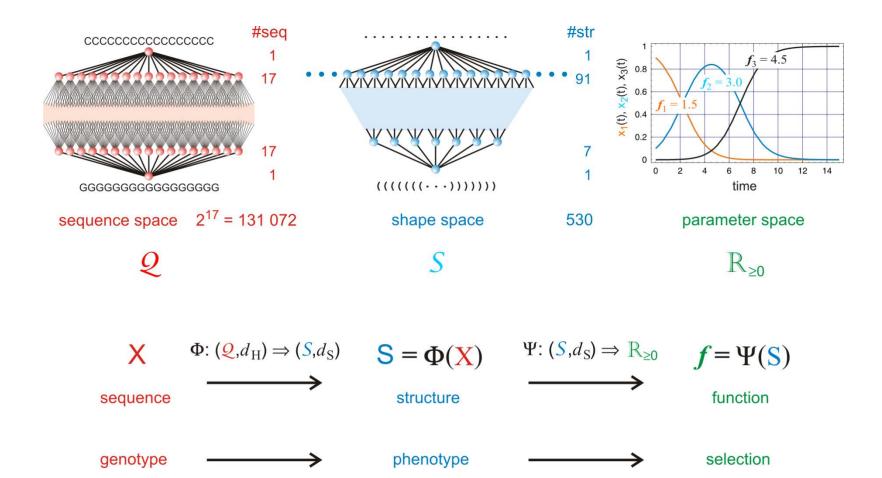


RNA sequence - structure mappings

- 1. ruggedness and neutrality
- 2. existence of extended neutral networks
- 3. shape space covering

The results 1. and 2. are certainly true also for other biopolymers, for example for proteins.

Evidence for ruggedness, neutrality and the existence of neutral networks was obtained also from virus evolution and *in vitro* experiments with bacteria.

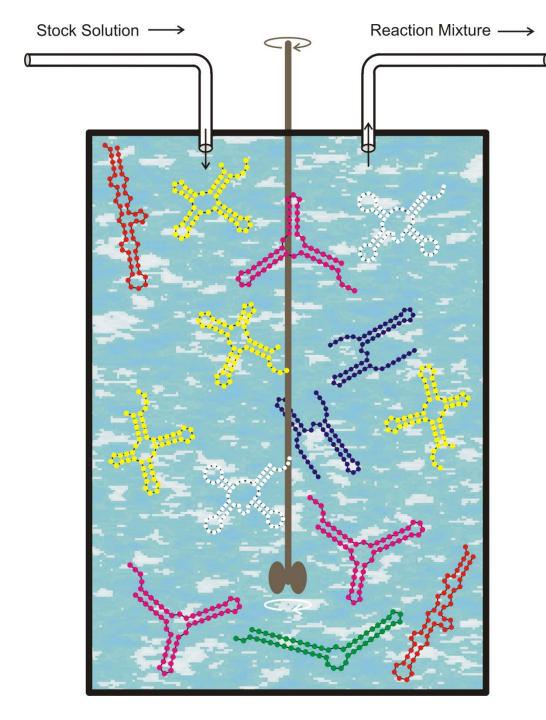


fitness of RNA secondary structures through evaluation of phenotypes

- 1. Is evolution possible?
- 2. "Non-probabilities"?
- 3. Protein folding a(n almost) solved example
- 4. Evolution The survival of the fittest?
- 5. Genotype-phenotype mapping and evolution
- 6. The interplay of adaptation and random drift
- 7. Natural selection and evolution

Evolution under controlled and analyzable conditions:

- (i) evolution *in silico*,
- (ii) evolution *in vitro*,
- (iii) virus evolution, and
- (iv) bacterial evolution.



replication rate constant or fitness:

 $f_k \!=\! \gamma \, / \, [\alpha + \Delta d_S^{(k)}] \text{ ; } \Delta d_S^{(k)} \!= d_H^{}(S_k^{},\!S_\tau^{})$

selection pressure:

The population size, N = # RNA, molecules, is determined by the flow:

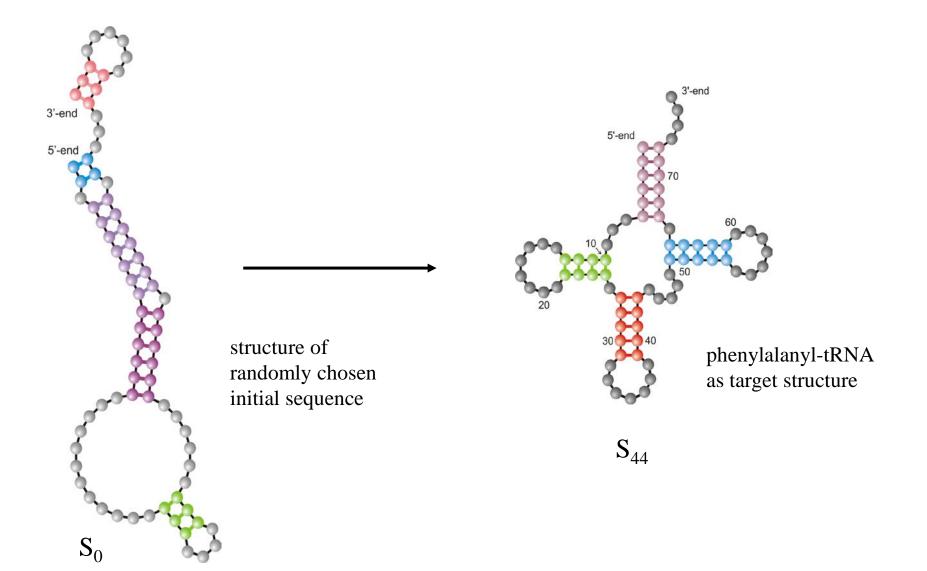
 $N(t)\approx\overline{N}\pm\sqrt{\overline{N}}$

mutation rate: p = 0.001 / nucleotide × replication

the flow reactor as a device for studying the evolution of molecules *in vitro* and *in silico*.

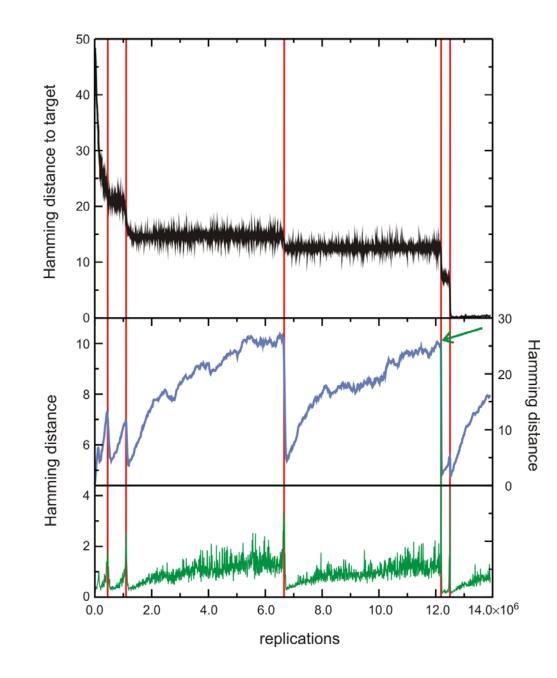
evolution in silico

W. Fontana, P. Schuster, *Science* **280** (1998), 1451-1455



evolution in silico.

W. Fontana, P. Schuster, Science 280 (1998), 1451-1455

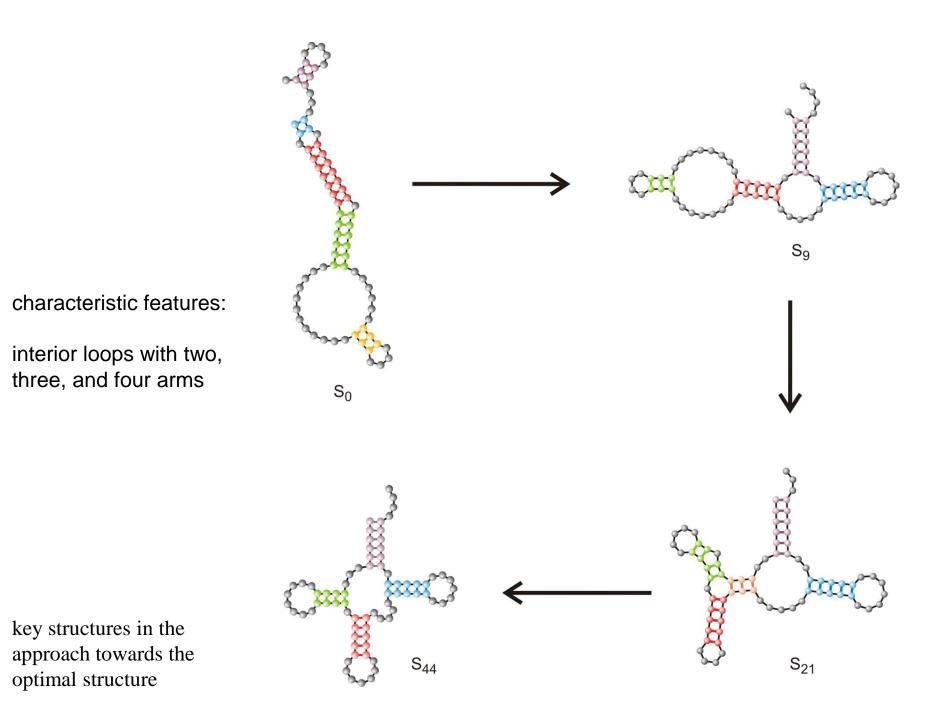


evolutionary trajectory

spreading of the population on neutral networks

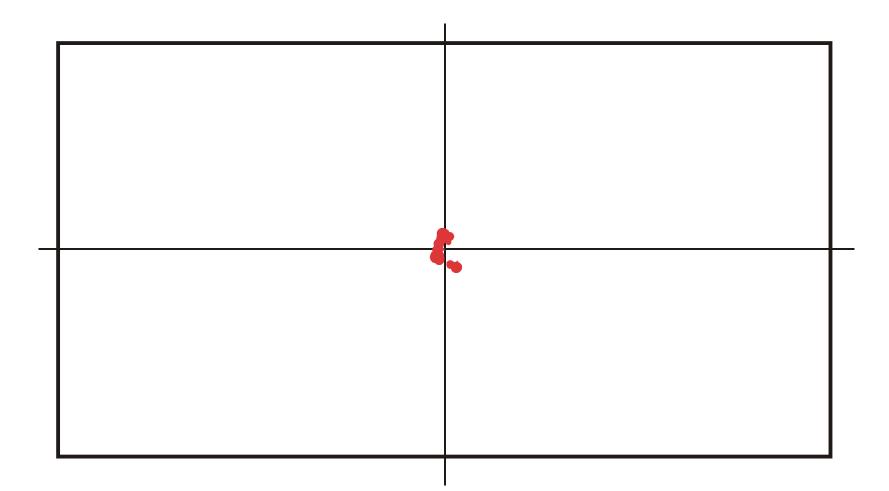
drift of the population center in sequence space

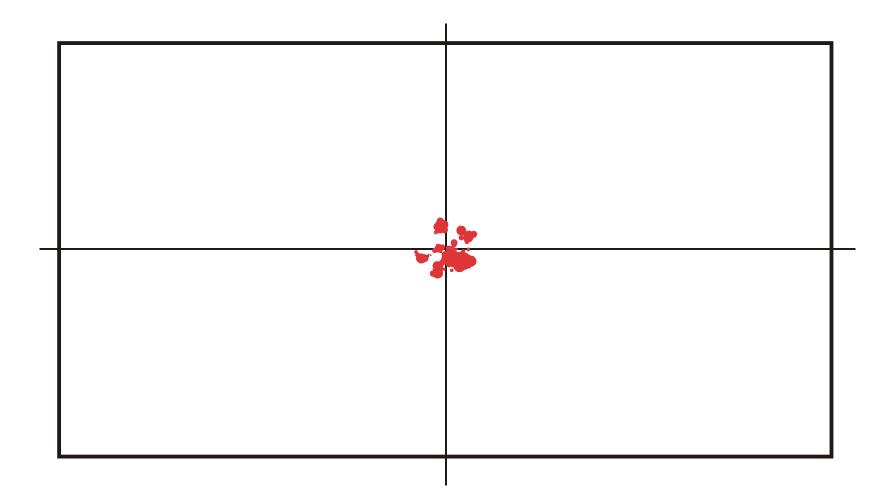
a targeted random walk to a predefind target structure

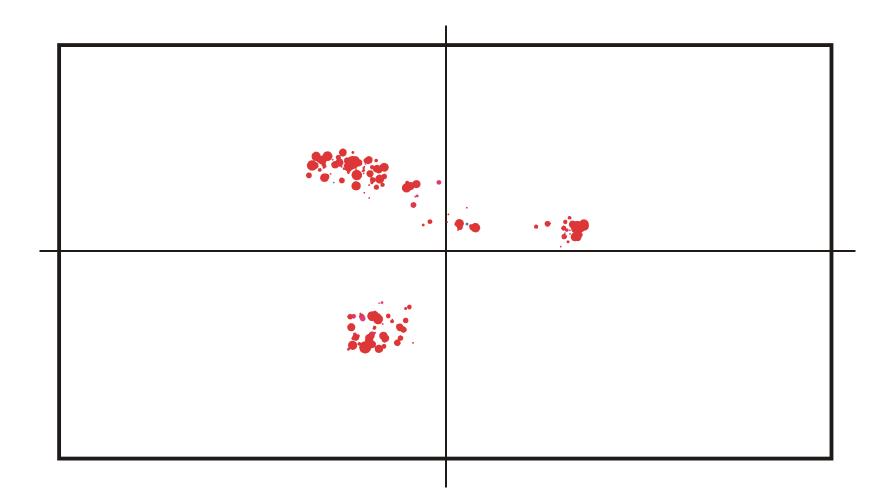


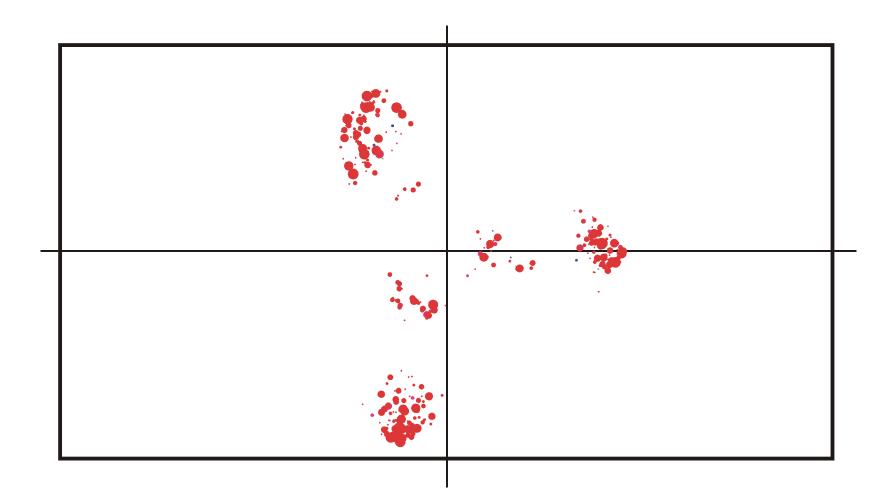
 •

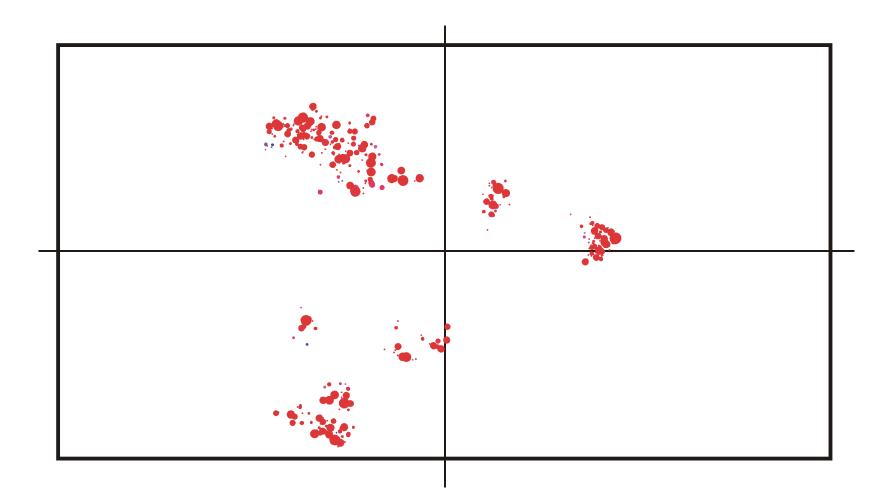
•

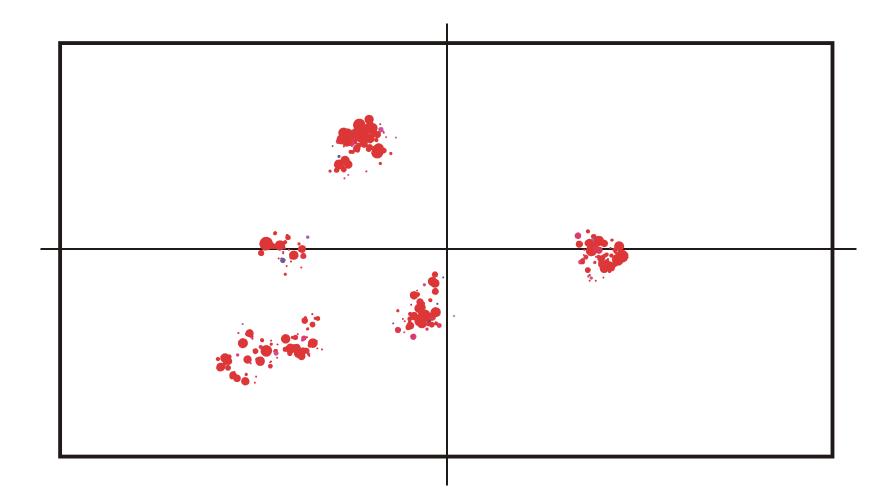


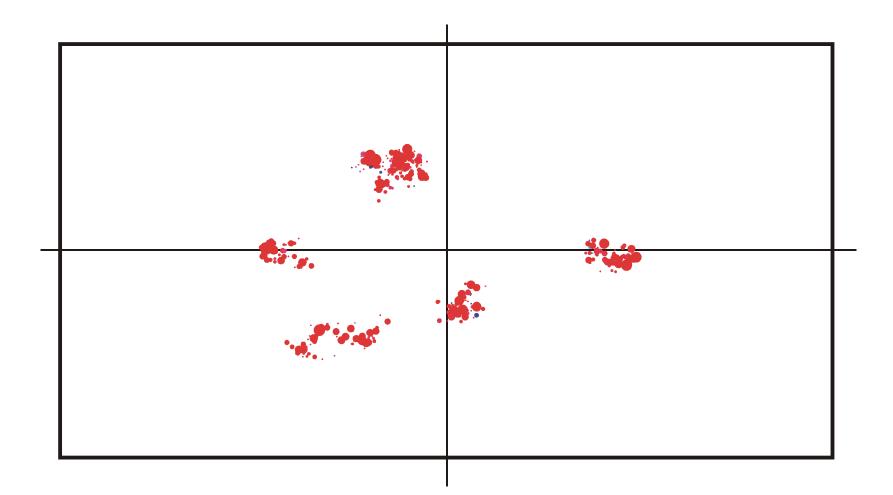


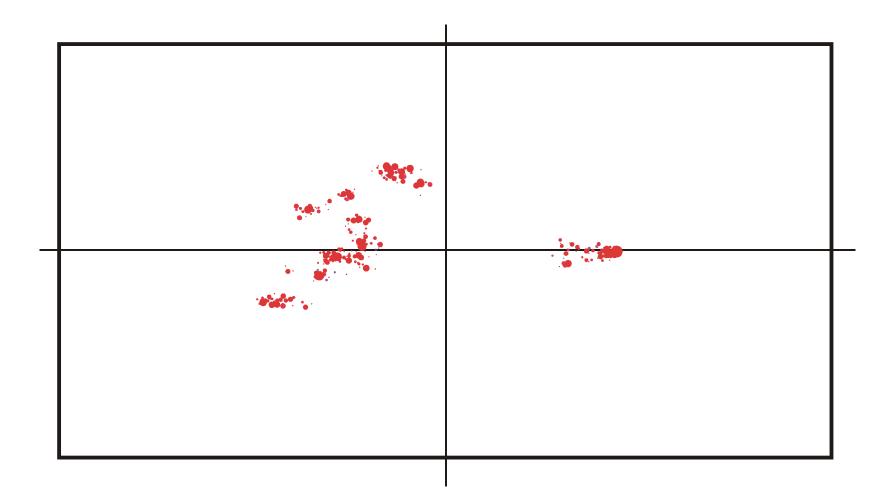


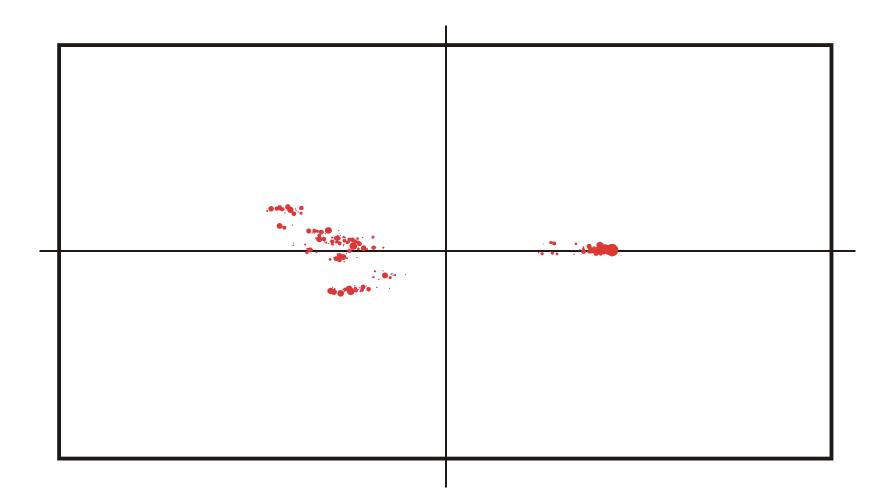


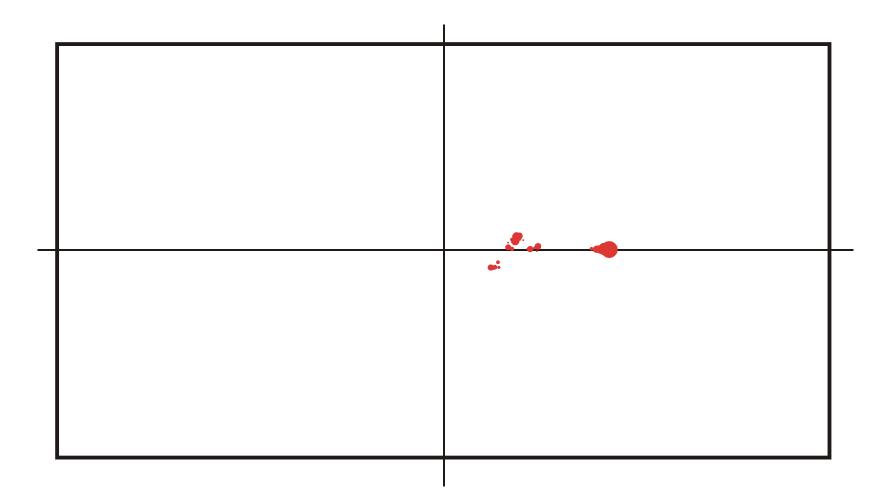


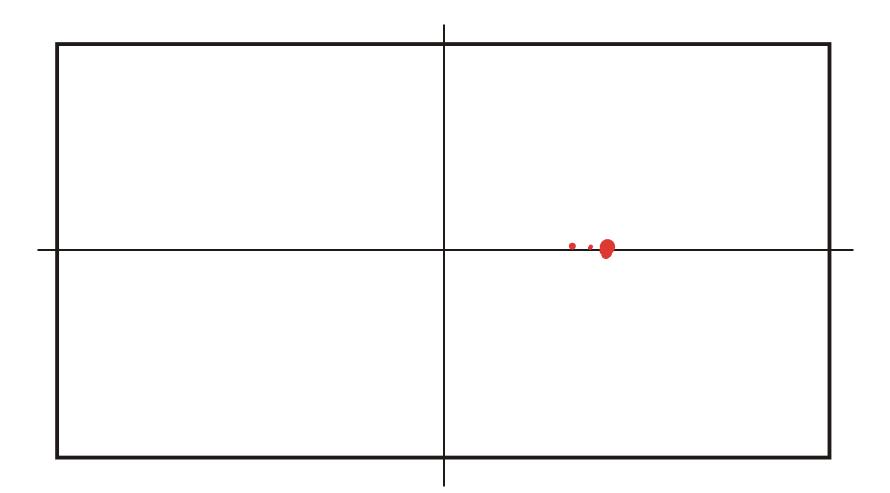


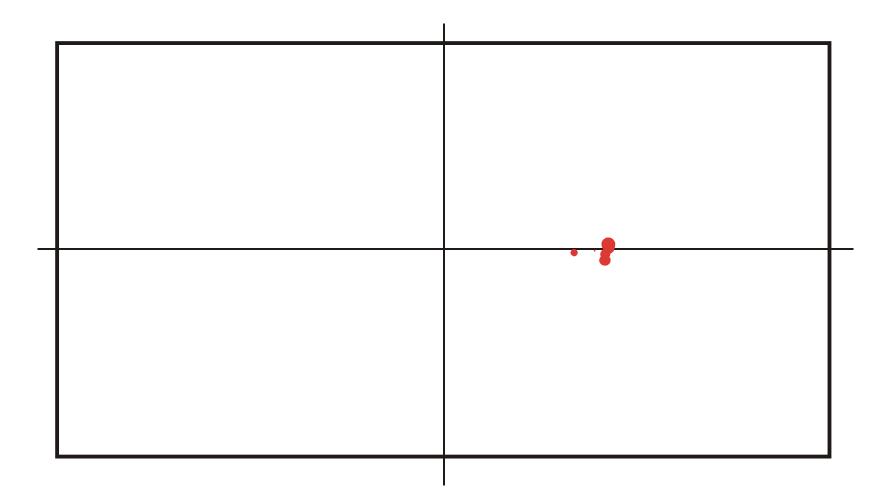


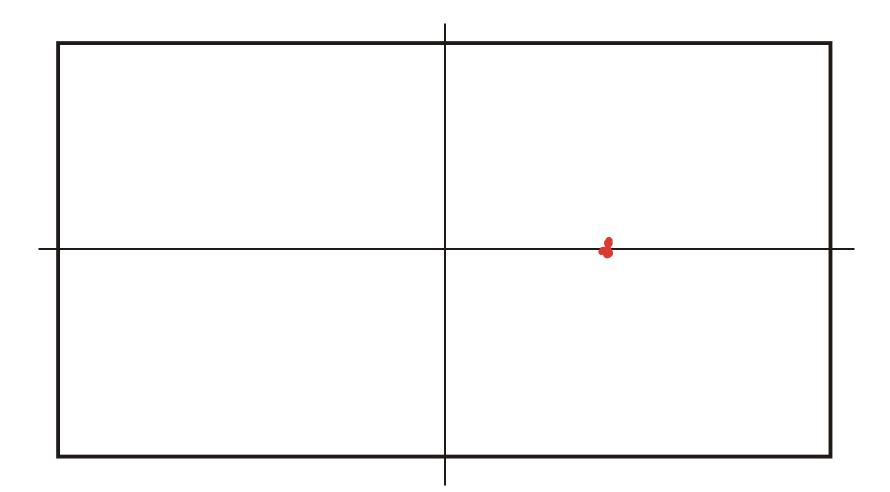


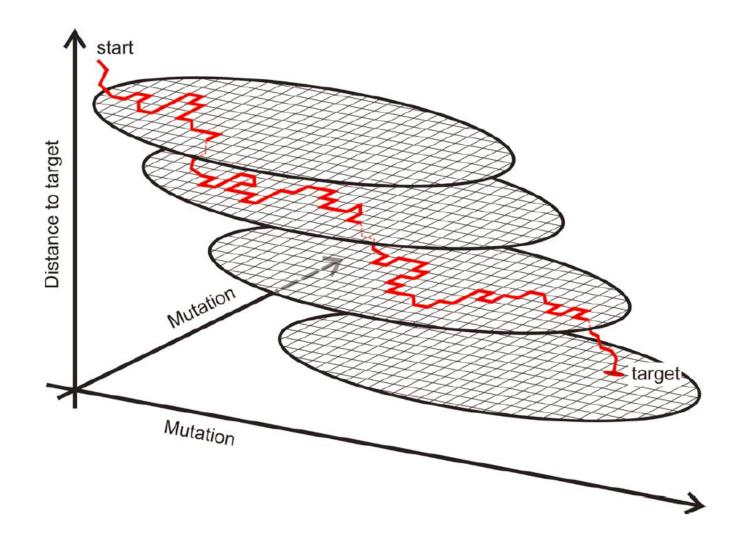












a sketch of optimization on neutral networks

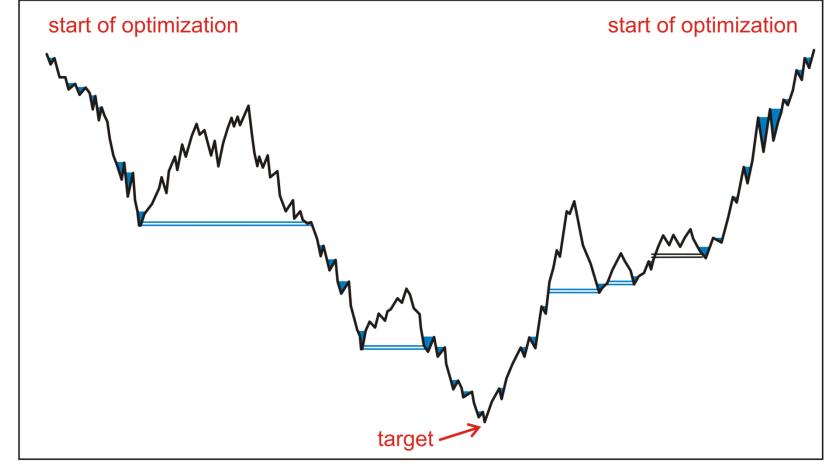
end of optimization end of optimization

start of optimization

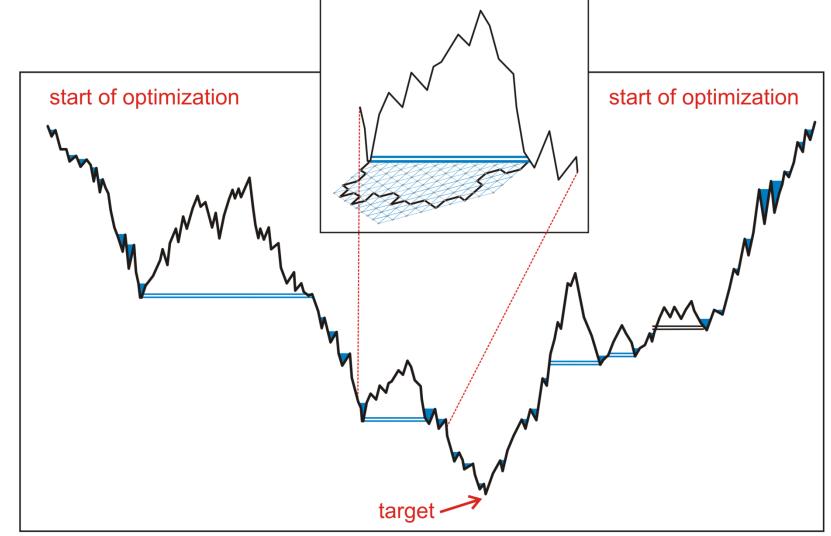
cost function

start of optimization

genotype space



genotype space



- 1. Is evolution possible?
- 2. "Non-probabilities"?
- 3. Protein folding a(n almost) solved example
- 4. Evolution The survival of the fittest?
- 5. Genotype-phenotype mapping and evolution
- 6. Natural selection and evolution

Reproduction leads to selection. In case of no effective fitness differences the selected variant is chosen at random.

Efficient evolution on natural fitness requires both adaptive periods of fitness increasing change and periods of phenotypic stasis with random drift in genotype space.

Thank you for your attention!

Web-Page for further information:

http://www.tbi.univie.ac.at/~pks