How Nature Circumvents Low Probabilities:

The Molecular Basis of Information and Complexity

Peter Schuster

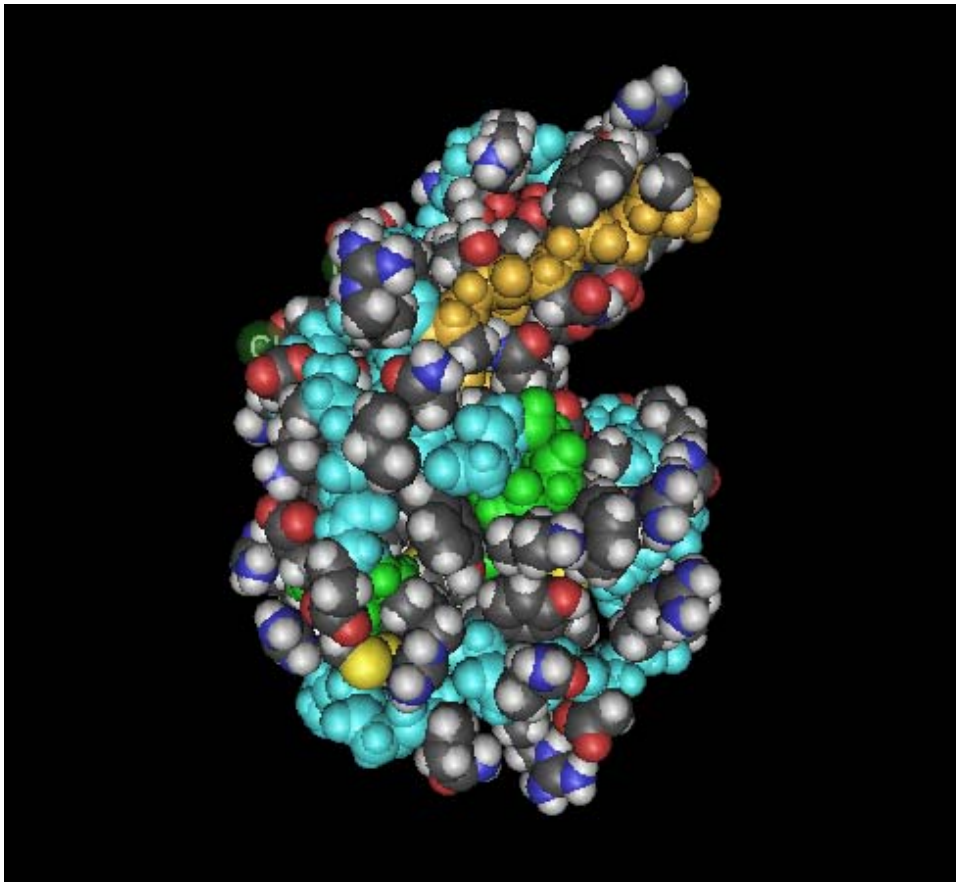
Institut für Theoretische Chemie
Universität Wien, Austria

Nonlinearity, Fluctuations, and Complexity

Brussels, 16.– 19.03.2005

Web-Page for further information:

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



Lysozyme – A small protein molecule

n = 130 amino acid residues

$6^{130} = 1.44 \times 10^{101}$ conformations

$20^{130} = 1.36 \times 10^{169}$ sequences

Protein folding: Levinthal's paradox

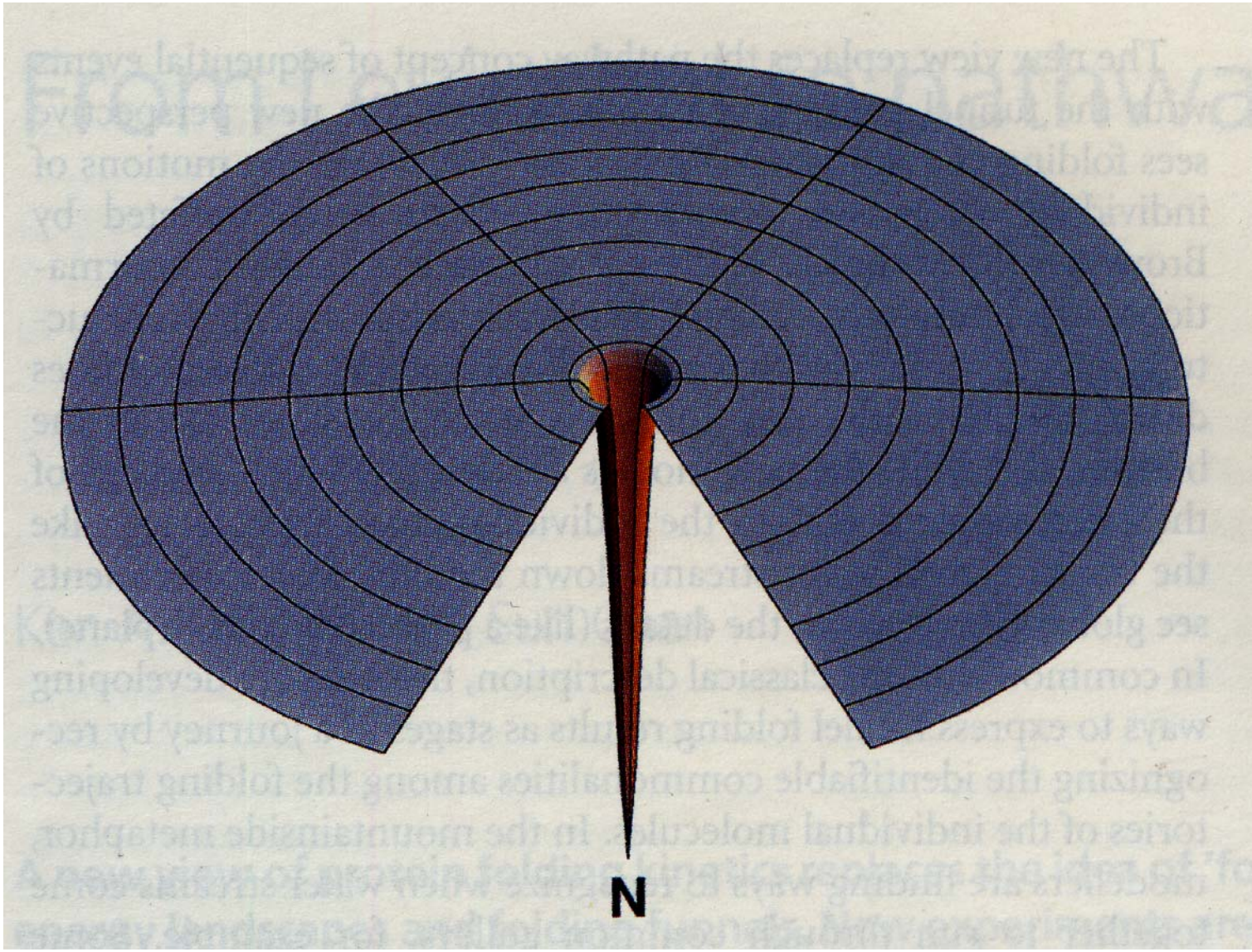
How can Nature find the native conformation in the folding process?

Evolution: Wigner's paradox

How can Nature find the optimal sequence of a protein in the evolutionary optimization process?

1. Solutions to Levinthal‘s paradox
2. Selection as solution to low probabilities in evolution
3. Origin of information by mutation and selection
4. Evolution of RNA phenotypes
5. The role of neutrality in molecular evolution
6. An experiment with RNA molecules
7. Summary

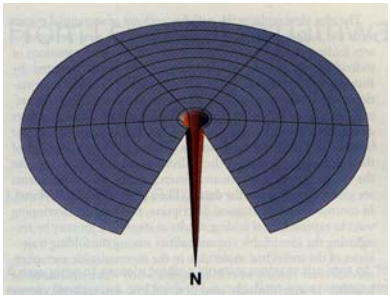
1. **Solutions to Levinthal‘s paradox**
2. Selection as solution to low probabilities in evolution
3. Origin of information by mutation and selection
4. Evolution of RNA phenotypes
5. The role of neutrality in molecular evolution
6. An experiment with RNA molecules
7. Summary

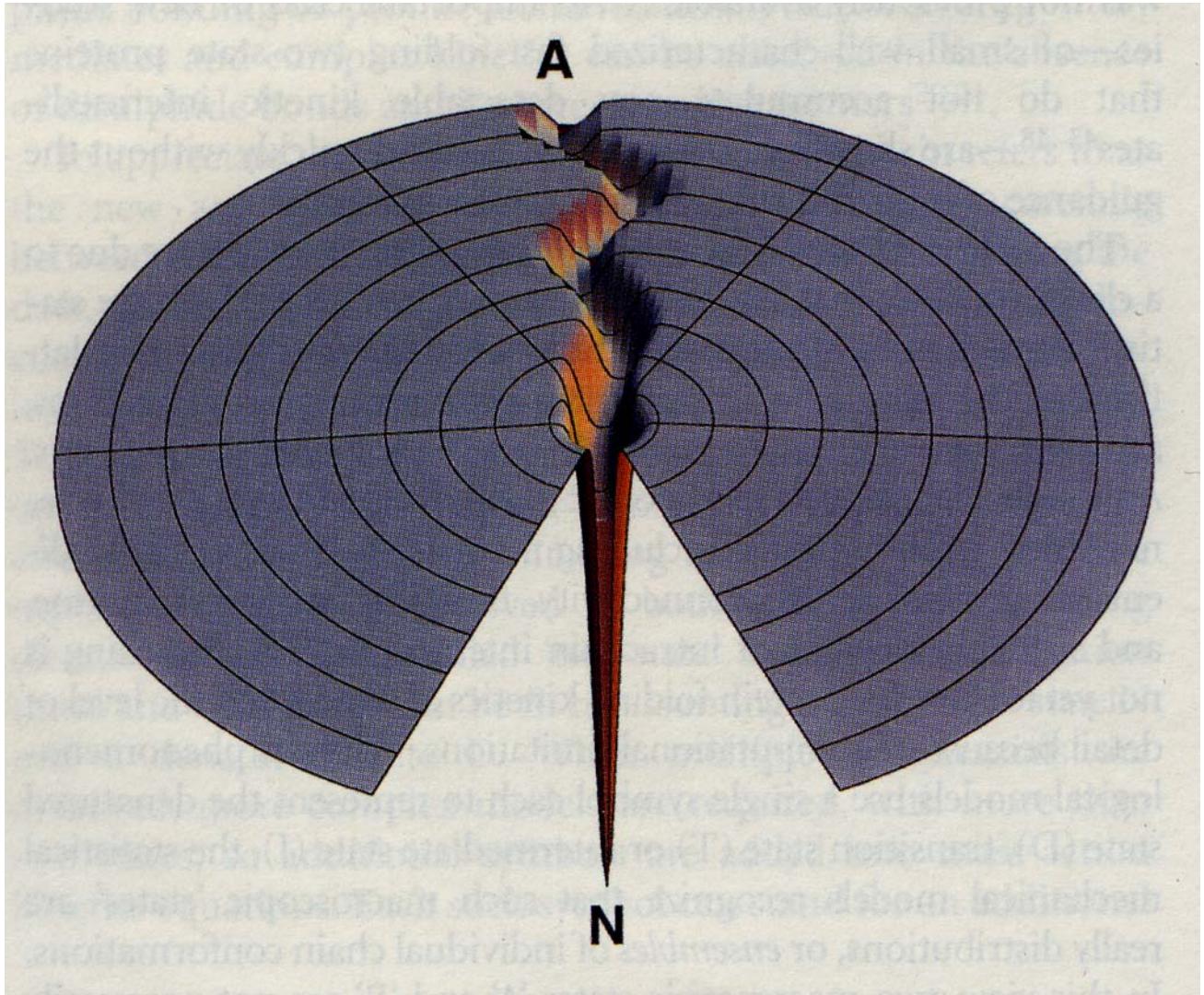


The golf course landscape

Levinthal’s paradox

K.A. Dill, H.S. Chan, Nature Struct. Biol. 4:10-19

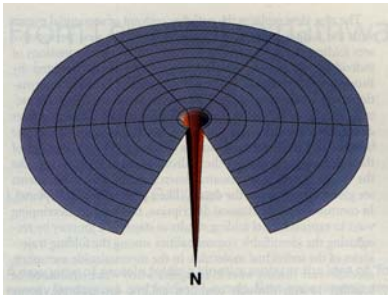


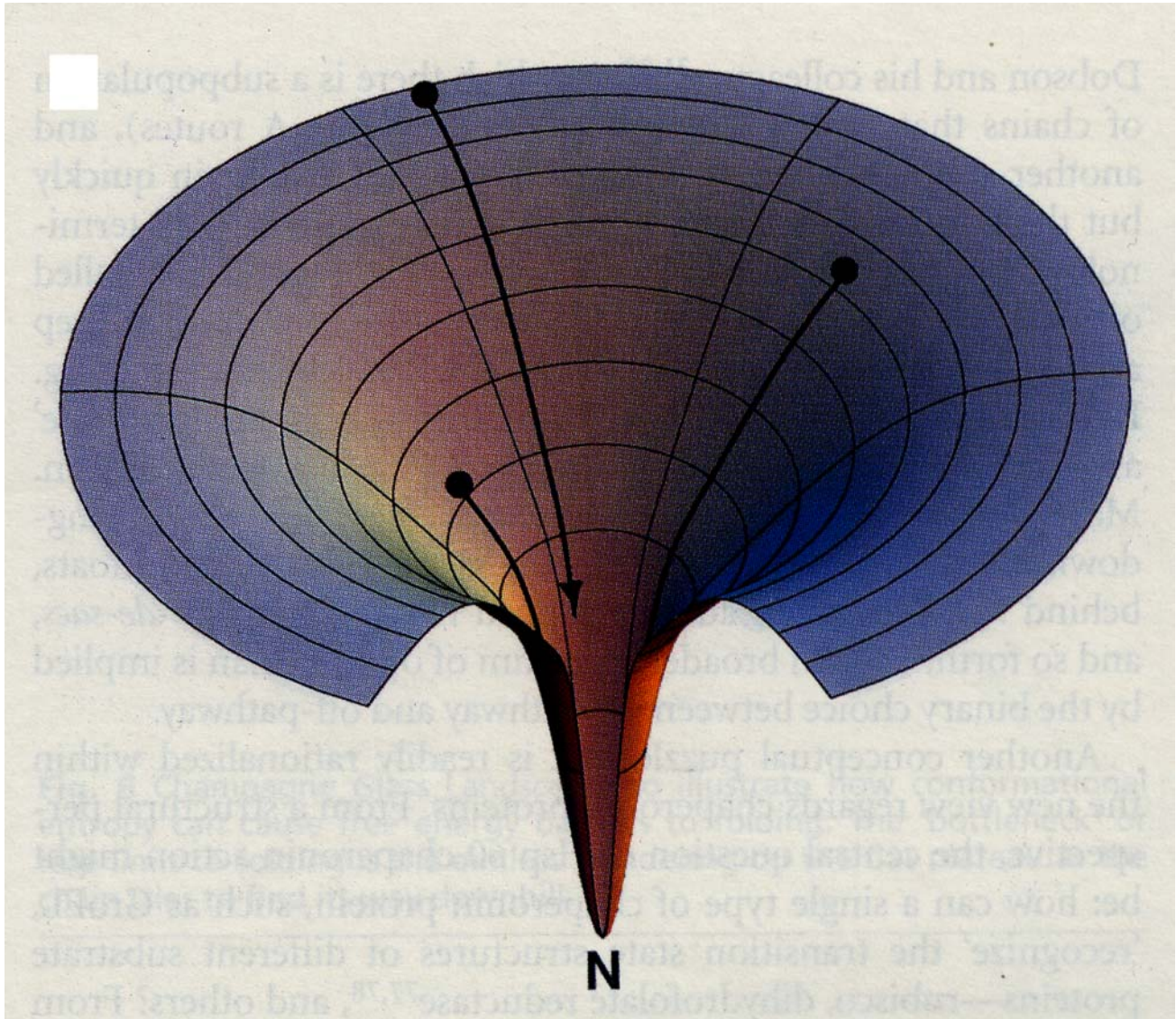


The pathway landscape

The pathway solution to Levinthal's paradox

K.A. Dill, H.S. Chan, Nature Struct. Biol. 4:10-19

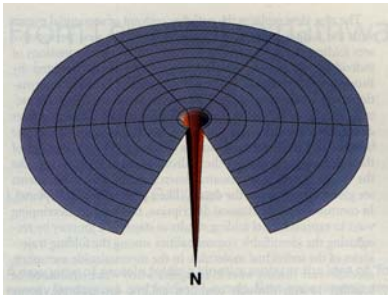


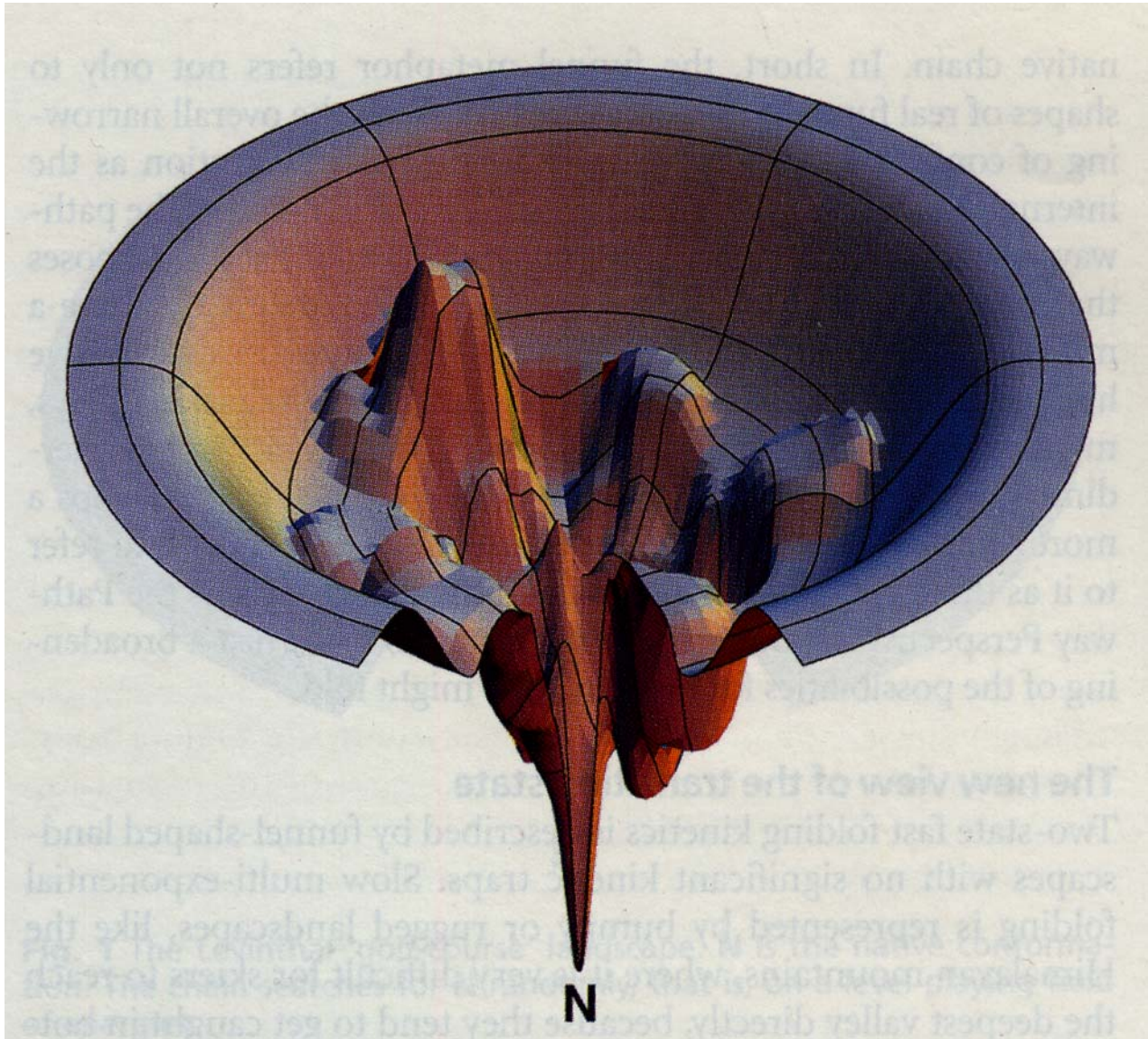


The folding funnel

The answer to Levinthal’s paradox

K.A. Dill, H.S. Chan, Nature Struct. Biol. 4:10-19



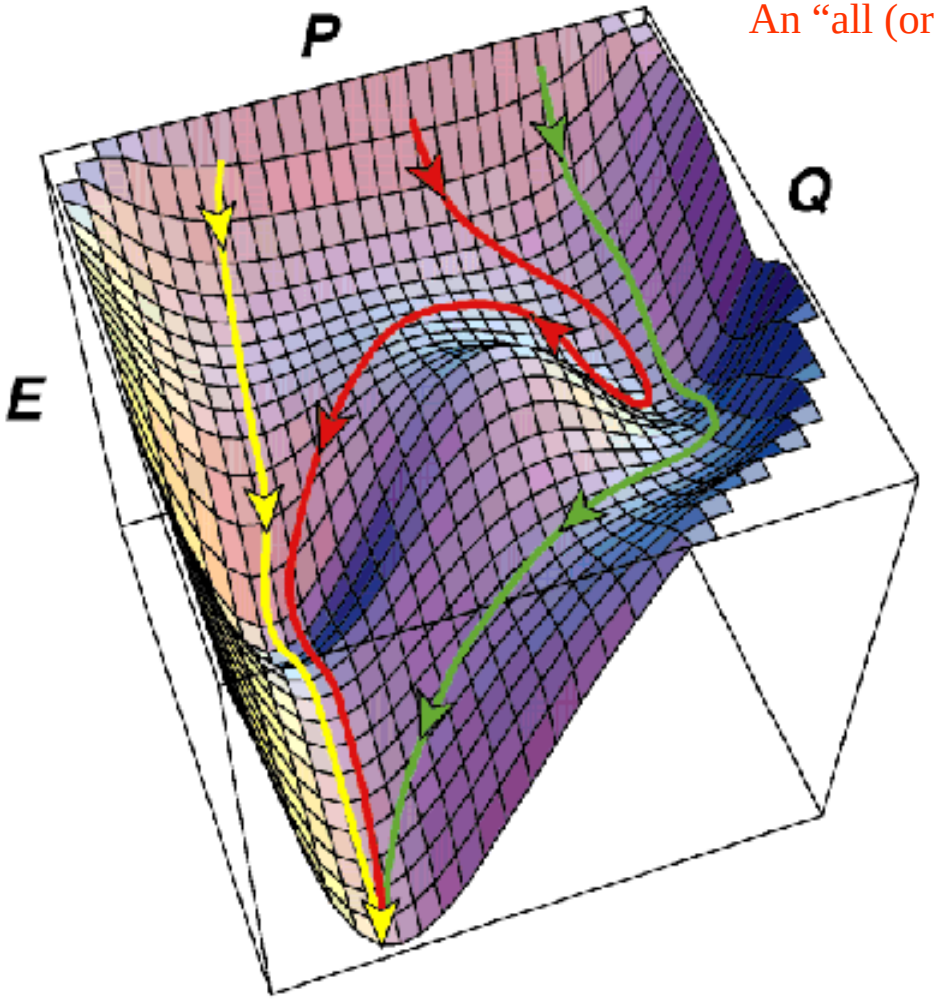


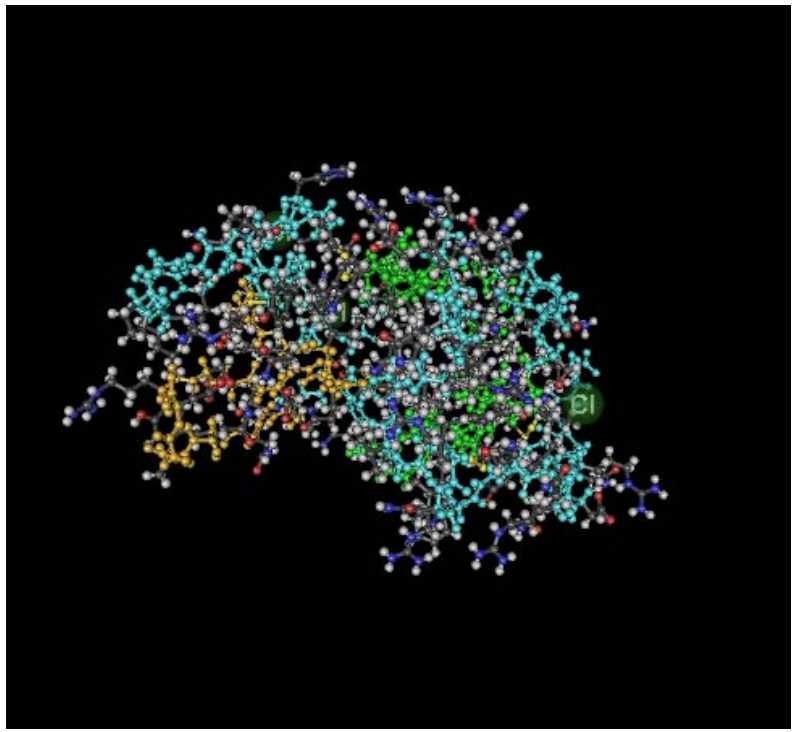
A more realistic folding funnel

The answer to Levinthal’s paradox

K.A. Dill, H.S. Chan, Nature Struct. Biol. 4:10-19

An “all (or many) paths lead to Rome” situation

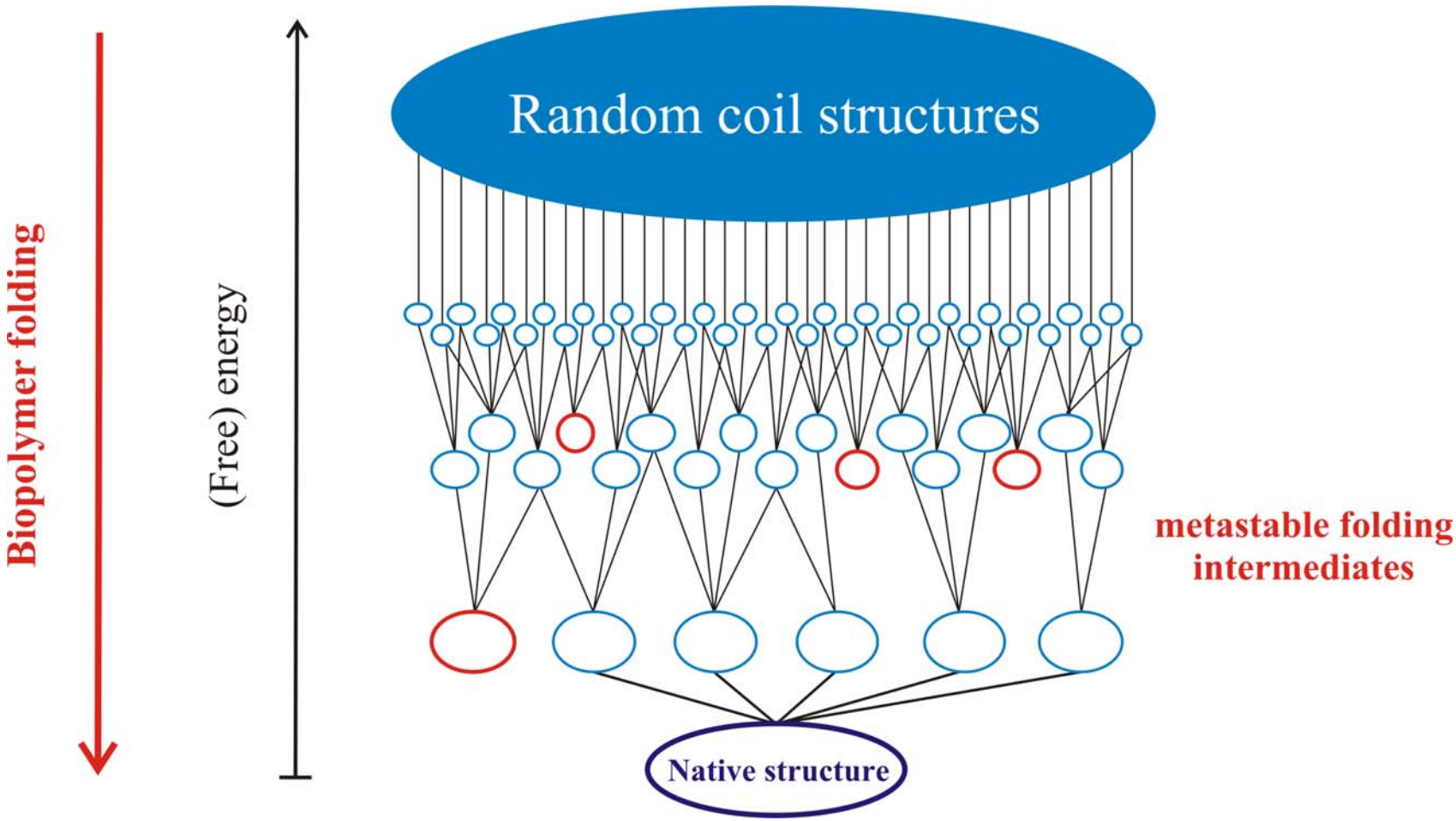




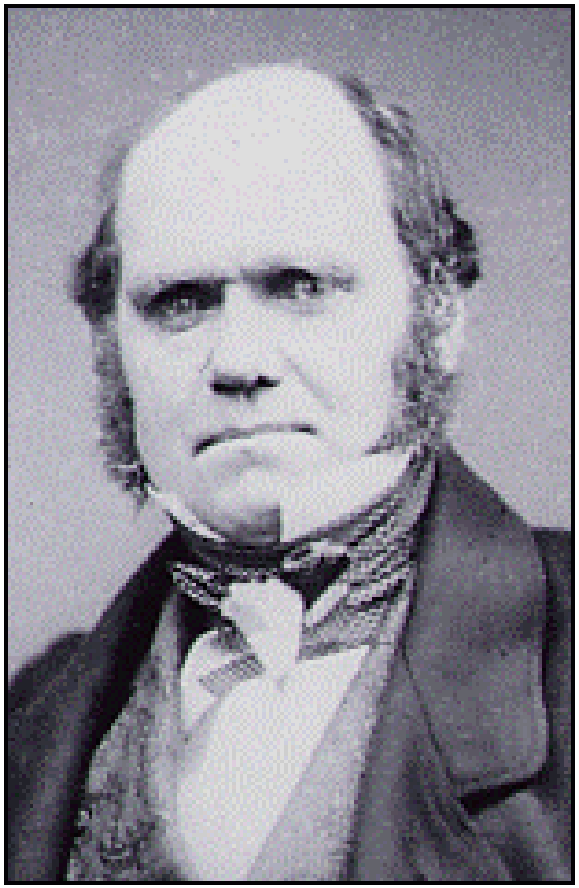
N **…** native conformation

A reconstructed free energy surface for lysozyme folding:

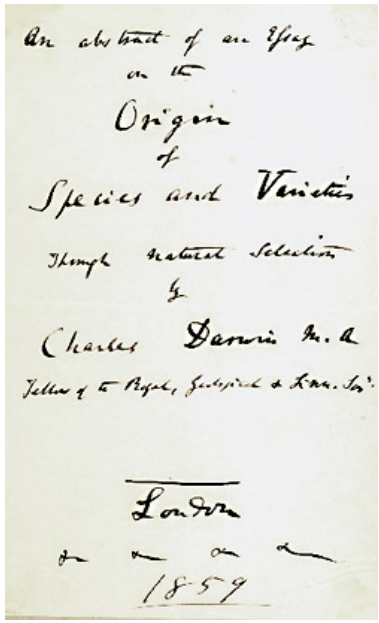
C.M. Dobson, A. Šali, and M. Karplus, Angew.Chem.Internat.Ed. 37: 868-893, 1988


Biopolymer folding
(Free) energy
Random coil structures
metastable folding intermediates
Native structure

1. Solutions to Levinthal‘s paradox
2. **Selection as solution to low probabilities in evolution**
3. Origin of information by mutation and selection
4. Evolution of RNA phenotypes
5. The role of neutrality in molecular evolution
6. An experiment with RNA molecules
7. Summary



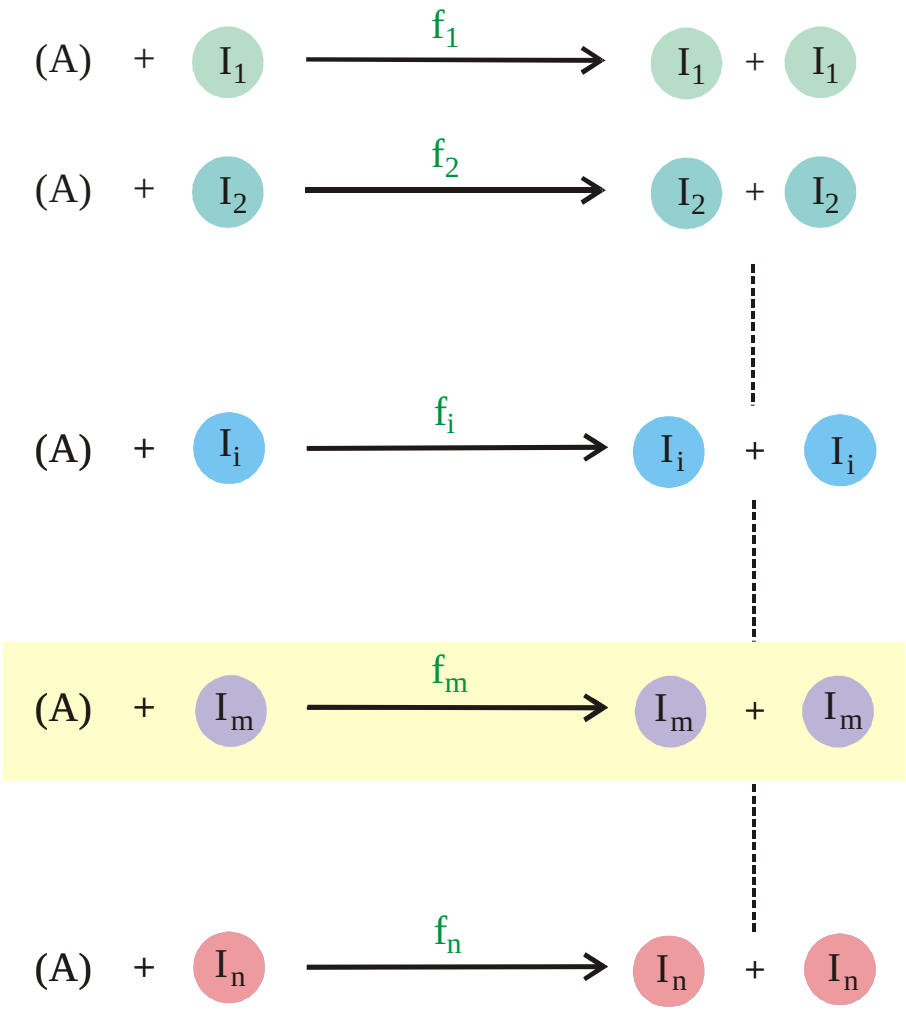
Charles Robert Darwin, 1809-1882



Earlier abstract of the
‚Origin of Species‘

Alfred Russell Wallace, 1823-1913

The two competitors in the formulation of evolution by natural selection



$$dx_i / dt = f_i x_i - x_i \Phi = x_i (f_i - \Phi)$$

$$\Phi = \Sigma_j f_j x_j ; \quad \Sigma_j x_j = 1 ; \quad i,j = 1,2,...,n$$

$$[I_i] = x_i \geq 0 ; \quad i = 1,2,...,n ;$$

$$[A] = a = \text{constant}$$

$$f_m = \max \{f_j; j=1,2,...,n\}$$

$$x_m(t) \rightarrow 1 \quad \text{for} \quad t \rightarrow \infty$$

Reproduction of organisms **or replication** of molecules as the basis of selection

Selection equation: $[I_i] = x_i \geq 0\ ,\ f_i > 0$

$$\frac{dx_i}{dt} = x_i\left(f_i - \phi\right), \quad i=1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j x_j = \overline{f}$$

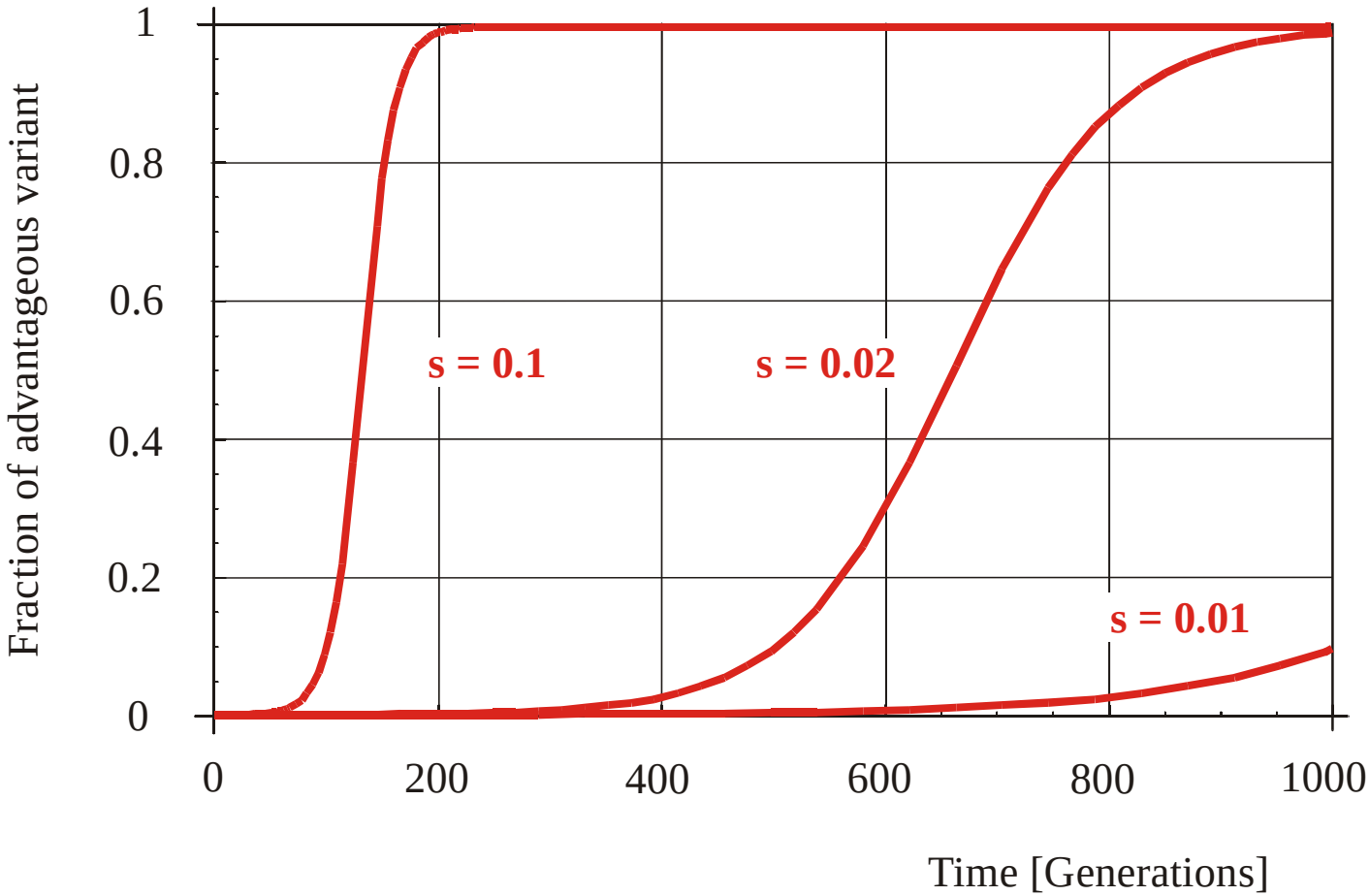
Mean fitness or dilution flux, ϕ(t), is a **non-decreasing function** of time,

$$\frac{d\phi}{dt} = \sum_{i=1}^{n} f_i \frac{dx_i}{dt} = \overline{f^2} - \left(\overline{f}\right)^2 = \mathrm{var}\{f\} \geq 0$$

Solutions are obtained by integrating factor transformation

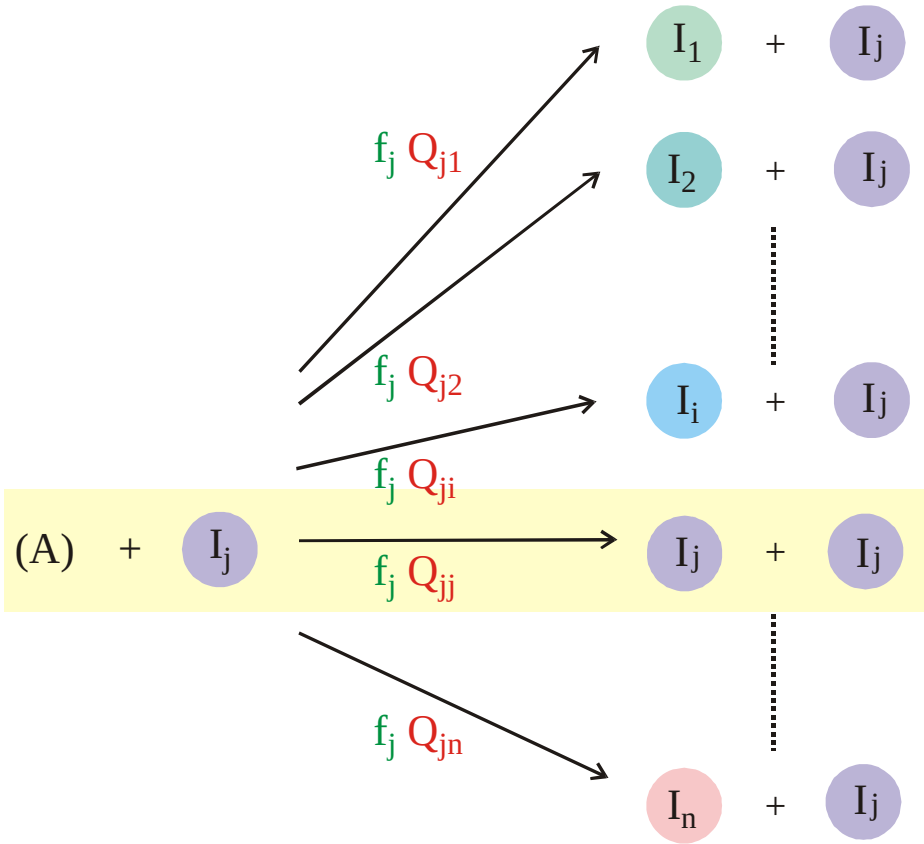
$$x_i(t) = \frac{x_i(0)\cdot\exp(f_i t)}{\sum_{j=1}^{n} x_j(0)\cdot\exp(f_j t)}; \quad i = 1,2,\cdots,n$$

$\mathbf{s} = (f_2 - f_1) / f_1;\ f_2 > f_1\ ;\ x_1(0) = 1 - 1/N\ ;\ x_2(0) = 1/N$



Selection of advantageous mutants in populations of N = 10 000 individuals

1. Solutions to Levinthal‘s paradox
2. Selection as solution to low probabilities in evolution
3. **Origin of information by mutation and selection**
4. Evolution of RNA phenotypes
5. The role of neutrality in molecular evolution
6. An experiment with RNA molecules
7. Summary



$$dx_i / dt = \Sigma_j f_j Q_{ji} x_j - x_i \Phi$$

$$\Phi = \Sigma_j f_j x_i ; \quad \Sigma_j x_j = 1 ; \quad \Sigma_i Q_{ij} = 1$$

$$[I_i] = x_i \geq 0 ; \quad i = 1,2,...,n ;$$

$$[A] = a = \text{constant}$$

$$Q_{ij} = (1-p)^{\ell - d(i,j)} \, p^{d(i,j)}$$

p ..... Error rate per digit

ℓ ..... Chain length of the polynucleotide

d(i,j) .... Hamming distance between I_i and I_j

Chemical kinetics of replication and mutation as parallel reactions

Mutation-selection equation: $[\mathrm{I}_i] = x_i \geq 0,\ f_i > 0,\ Q_{ij} \geq 0$

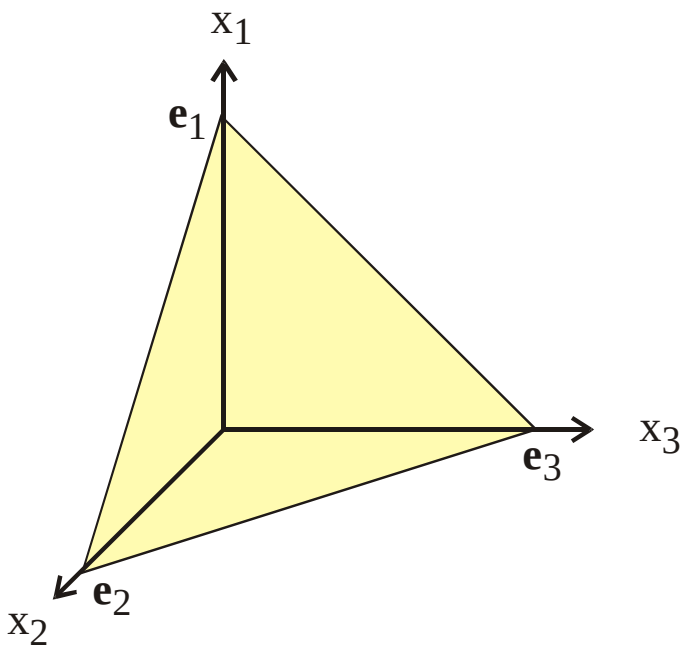
$$\frac{dx_i}{dt} = \sum_{j=1}^{n} f_j Q_{ji}\, x_j - x_i\, \phi, \quad i = 1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j\, x_j = \overline{f}$$

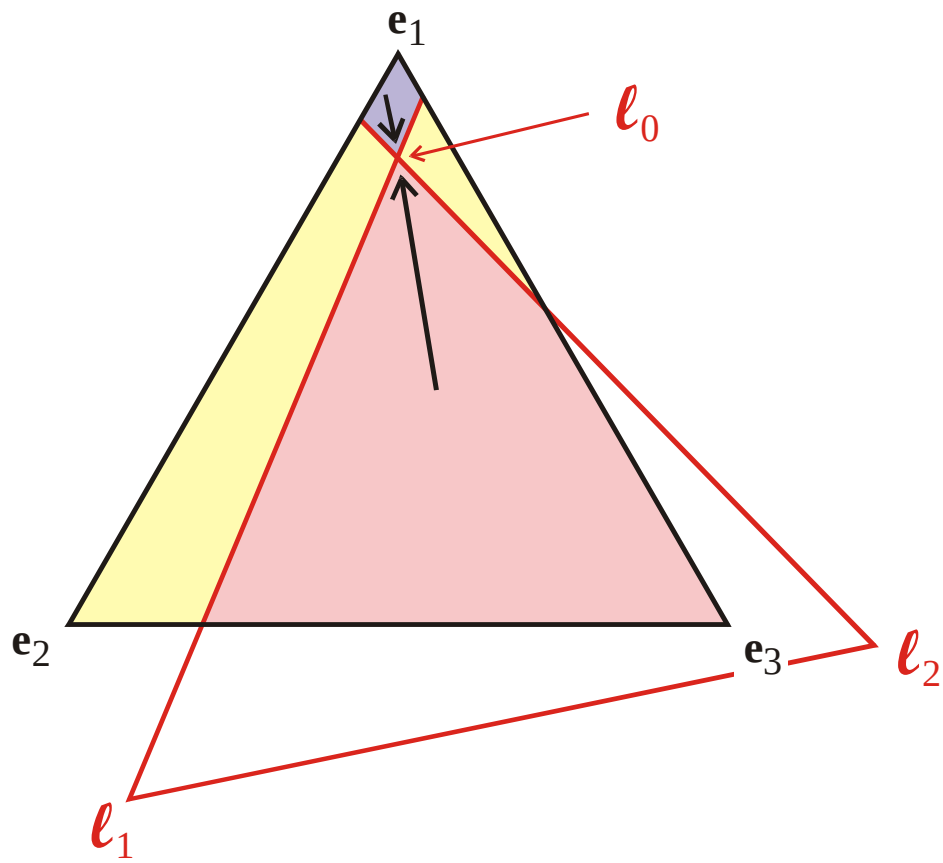
Solutions are obtained after integrating factor transformation by means of an eigenvalue problem

$$x_i(t) = \frac{\sum_{k=0}^{n-1} \ell_{ik} \cdot c_k(0) \cdot \exp(\lambda_k t)}{\sum_{j=1}^{n} \sum_{k=0}^{n-1} \ell_{jk} \cdot c_k(0) \cdot \exp(\lambda_k t)}; \quad i = 1,2,\cdots,n; \quad c_k(0) = \sum_{i=1}^{n} h_{ki}\, x_i(0)$$

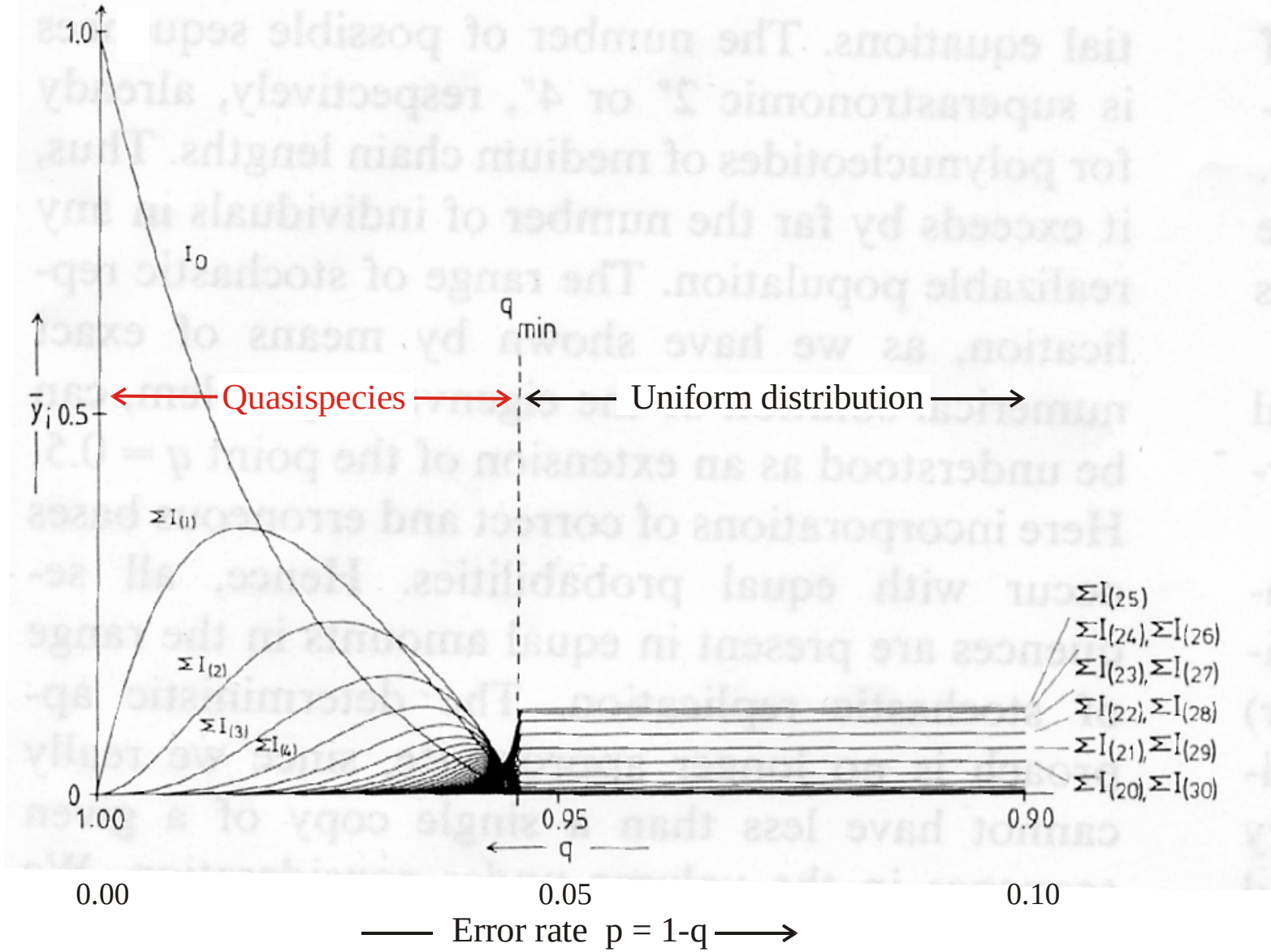
$$W \div \left\{ f_i Q_{ij};\ i, j = 1,2,\cdots,n \right\};\ L = \left\{ \ell_{ij};\ i, j = 1,2,\cdots,n \right\};\ L^{-1} = H = \left\{ h_{ij};\ i, j = 1,2,\cdots,n \right\}$$

$$L^{-1} \cdot W \cdot L \quad = \quad \Lambda \quad = \quad \left\{ \lambda_k;\ k = 0,1,\cdots,n-1 \right\}$$

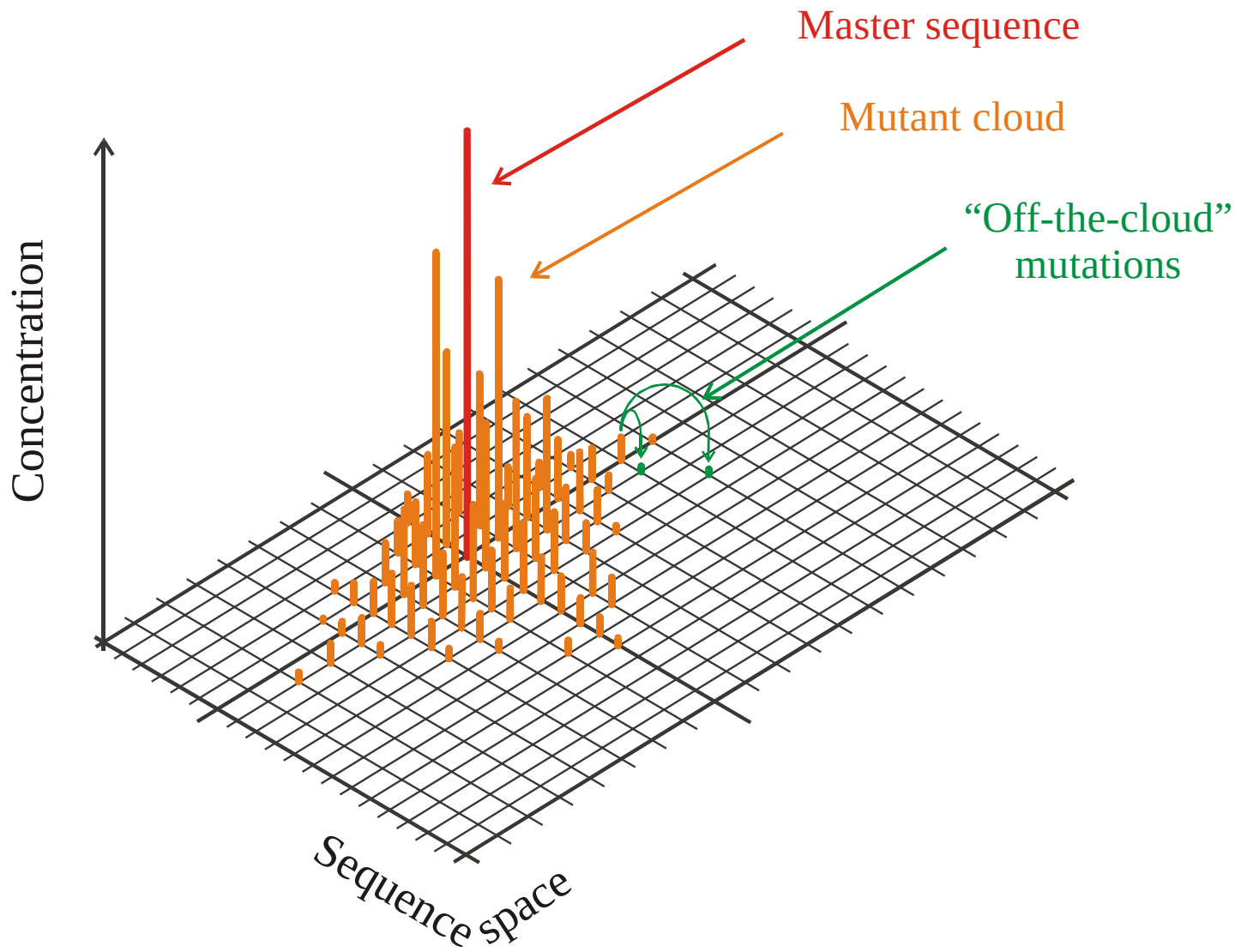




The quasispecies on the concentration simplex $\mathbb{S}_3 = \left\{ x_i \geq 0,\ i = 1,2,3;\ \sum_{i=1}^{3} x_i = 1 \right\}$

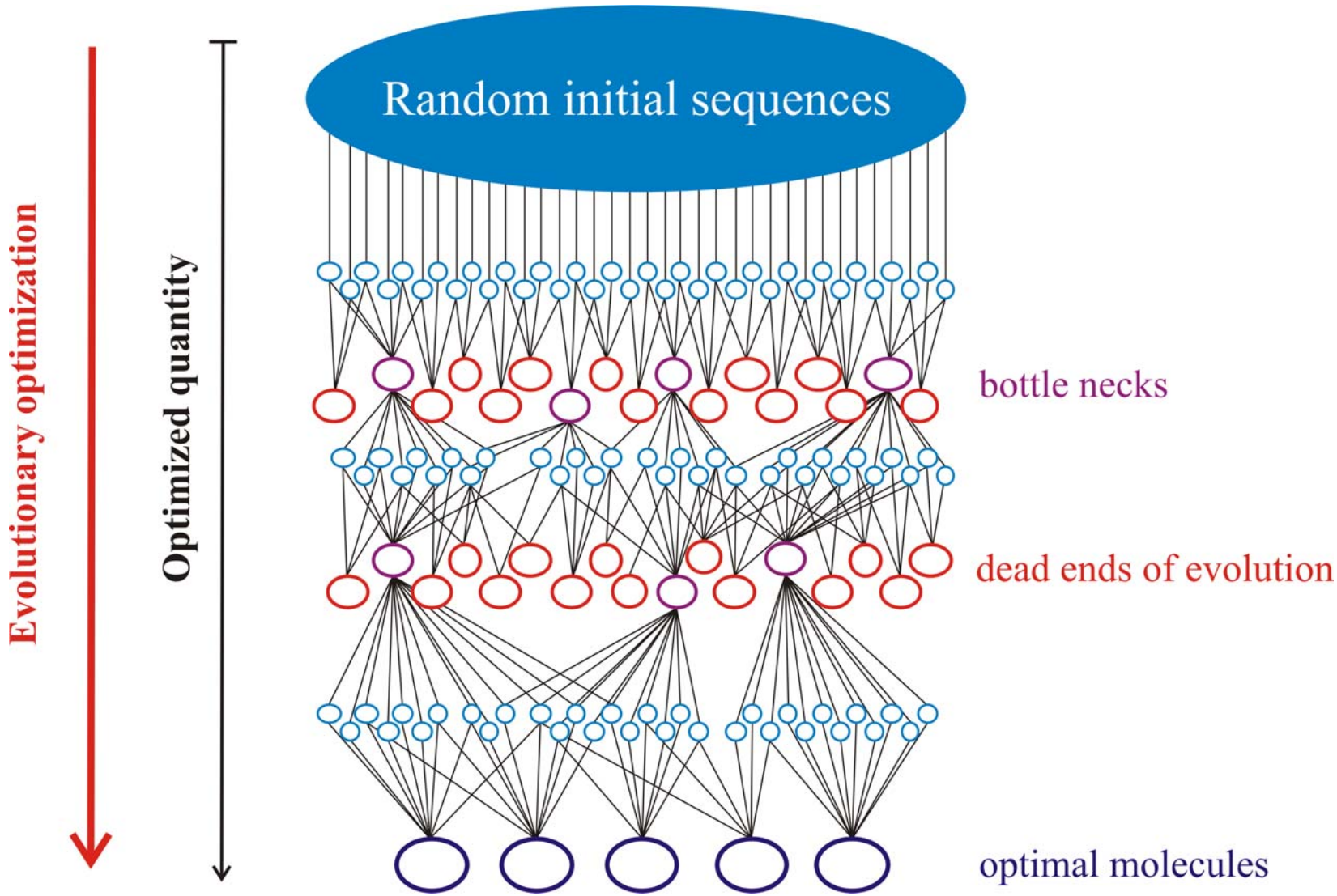


Quasispecies as a function of the replication accuracy q



The molecular quasispecies
in sequence space

1. Solutions to Levinthal‘s paradox
2. Selection as solution to low probabilities in evolution
3. Origin of information by mutation and selection
4. **Evolution of RNA phenotypes**
5. The role of neutrality in molecular evolution
6. An experiment with RNA molecules
7. Summary


Evolutionary optimization
Optimized quantity
Random initial sequences
bottle necks
dead ends of evolution
optimal molecules

Computer simulation of RNA optimization

Walter Fontana and Peter Schuster,
Biophysical Chemistry 26:123-147, 1987

Walter Fontana, Wolfgang Schnabl, and
Peter Schuster, Phys.Rev.A 40:3301-3321, 1989

PHYSICAL REVIEW A VOLUME 40, NUMBER 6 SEPTEMBER 15, 1989

Physical aspects of evolutionary optimization and adaptation

Walter Fontana, Wolfgang Schnabl, and Peter Schuster*
Institut für Theoretische Chemie der Universität Wien, Währingerstrasse 17, A 1090 Wien, Austria
(Received 2 February 1989; revised manuscript received 5 May 1989)

A model of an objective function based on polynucleotide folding is used to investigate the dynamics of evolutionary adaptation in finite populations. Binary sequences are optimized with respect to their kinetic properties through a stochastic process involving mutation and selection. The objective function consists in a mapping from the set of all binary strings with given length into a set of two-dimensional structures. These structures then encode the kinetic properties, expressed in terms of parameters of reaction probability distributions. The objective function obtained thereby represents a realistic example of a highly "rugged landscape." Ensembles of molecular strings adapting to this landscape are studied by tracing their escape path from local optima and by applying multivariate analysis. Effects of small population numbers in the tail of the sequence distribution are discussed quantitatively. Close upper bounds to the number of distinct values produced by our objective function are given. The distribution of values is explored by means of simulated annealing and reveals a random scatter in the locations of optima in the space of all sequences. The genetic optimization protocol is applied to the "traveling salesman" problem.

Biophysical Chemistry 26 (1987) 123–147
Elsevier

BPC 01133

A computer model of evolutionary optimization

Walter Fontana and Peter Schuster
Institut für theoretische Chemie und Strahlenchemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

Accepted 27 February 1987

Molecular evolution; Optimization; Polyribonucleotide folding; Quasi-species; Selective value; Stochastic reaction kinetics

Molecular evolution is viewed as a typical combinatorial optimization problem. We analyse a chemical reaction model which considers RNA replication including correct copying and point mutations together with hydrolytic degradation and the dilution flux of a flow reactor. The corresponding stochastic reaction network is implemented on a computer in order to investigate some basic features of evolutionary optimization dynamics. Characteristic features of real molecular systems are mimicked by folding binary sequences into unknotted two-dimensional structures. Selective values are derived from these molecular 'phenotypes' by an evaluation procedure which assigns numerical values to different elements of the secondary structure. The fitness function obtained thereby contains nontrivial long-range interactions which are typical for real systems. The fitness landscape also reveals quite involved and bizarre local topologies which we consider also representative of polynucleotide replication in actually occurring systems. Optimization operates on an ensemble of sequences via mutation and natural selection. The strategy observed in the simulation experiments is fairly general and resembles closely a heuristic widely applied in operations research areas. Despite the relative smallness of the system – we study 2000 molecules of chain length $\nu = 70$ in a typical simulation experiment – features typical for the evolution of real populations are observed as there are error thresholds for replication, evolutionary steps and quasistationary sequence distributions. The relative importance of selectively neutral or almost neutral variants is discussed quantitatively. Four characteristic ensemble properties, entropy of the distribution, ensemble correlation, mean Hamming distance and diversity of the population, are computed and checked for their sensitivity in recording major optimization events during the simulation.

1. Molecular evolution and optimization

Conventional population genetics treats mutation as an external stochastic source. Moreover, mutations are considered as very rare events. In the absence of genetic recombination populations of haploid organisms are expected to be usually homogeneous. Experimental evidence on viral and bacterial populations is available now and it contradicts these expectations. Mutations appear much more frequently than was originally assumed.

Dedicated to Professor Manfred Eigen on the occasion of his 60th birthday.
Correspondence address: P. Schuster, Institut für theoretische Chemie und Strahlenchemie der Universität Wien, Währingerstraße 17, A-1090 Wien, Austria.

The molecular approach considers error-free replication and mutation as parallel reactions within the same mechanism. Detailed information on the molecular mechanisms of polynucleotide replication provides direct insight into the nature of mutations and their role in evolution. Several classes of mutations are properly distinguished: point mutations, deletions and insertions. Point mutations are of special importance: they represent the most frequent mutations and are easily incorporated into theoretical models of molecular evolution. This does not mean, however, that the other classes of mutations are not important in evolution. To give an example: there is a general belief that insertions leading to gene duplication played a major role in the development of present day enzyme families.

The first theoretical model of molecular evolu-

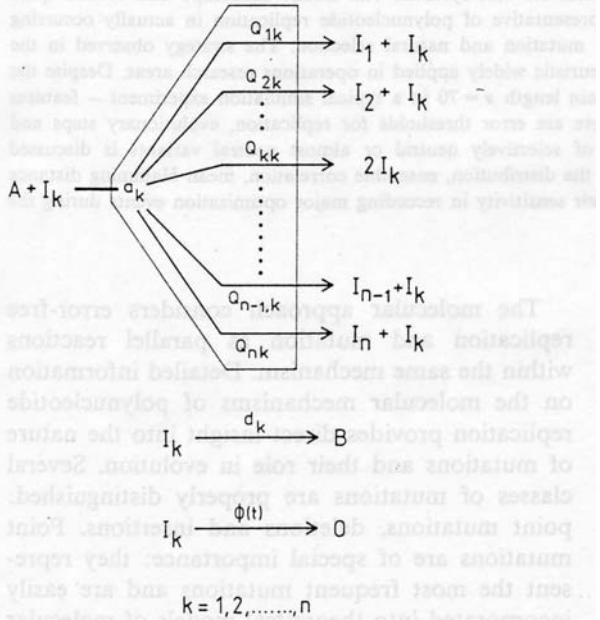


Fig. 1. The reaction network. Synthesis on template I_k proceeds with the rate constant a_k and leads with frequency Q_{ik} to a new template I_i preserving the old copy. Materials A needed for polymerization are assumed to be buffered. Degradation to waste products B occurs with rate d_k and a controlled unspecific flux $\Phi(t)$ removes templates from the system.

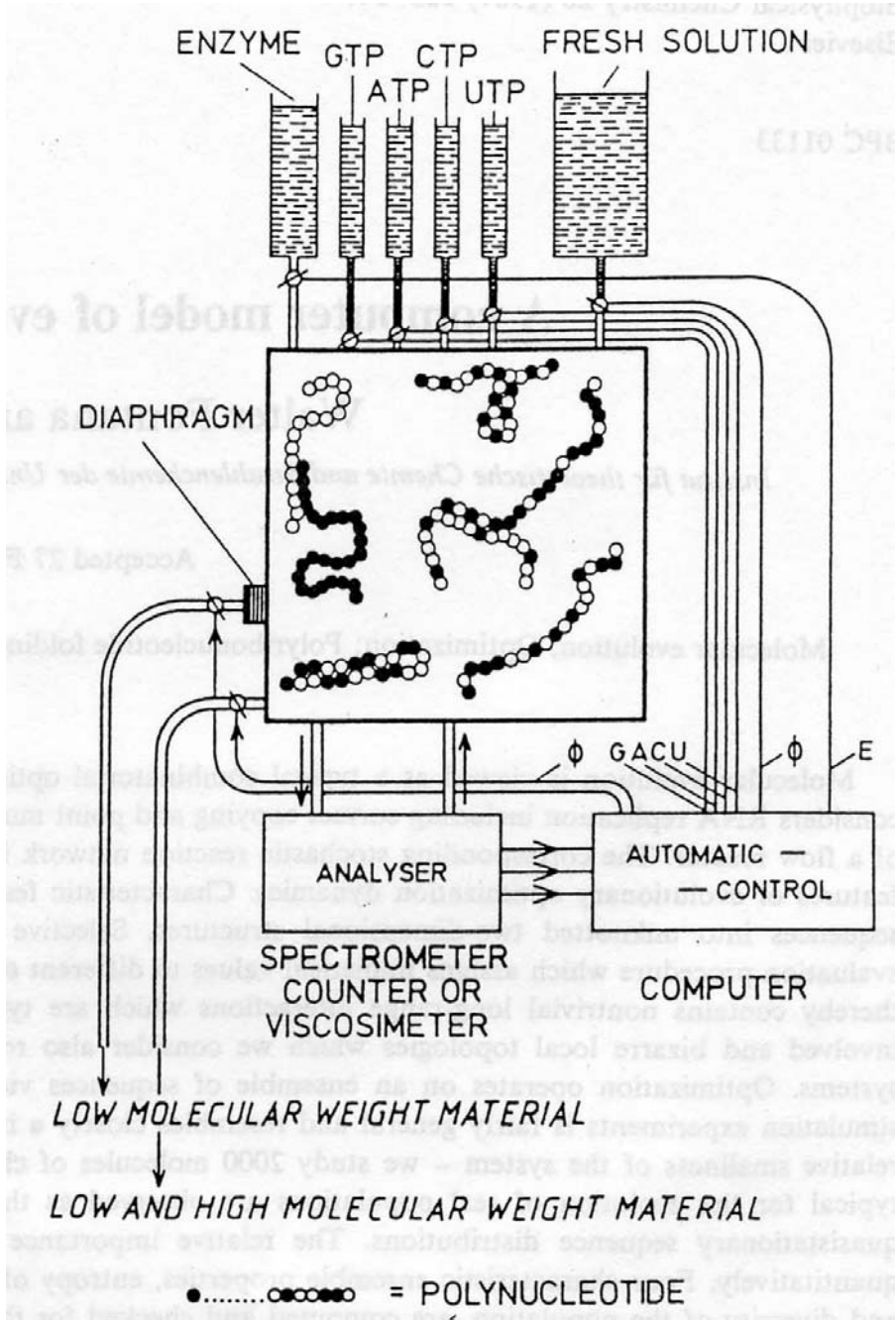
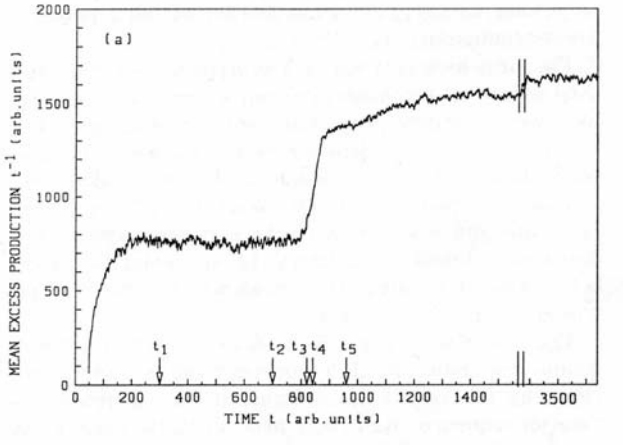
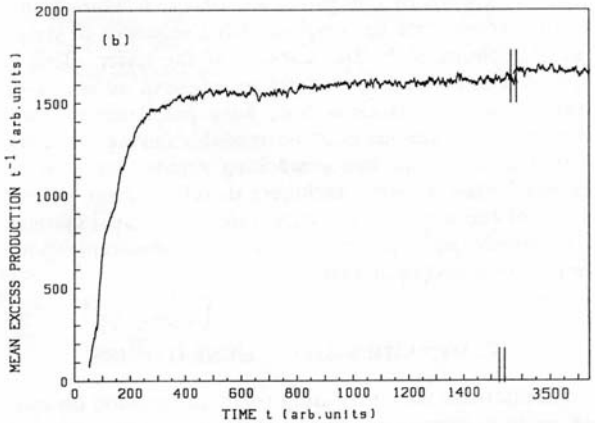
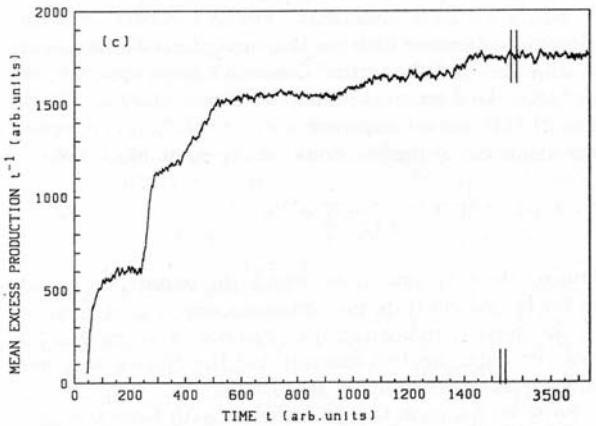
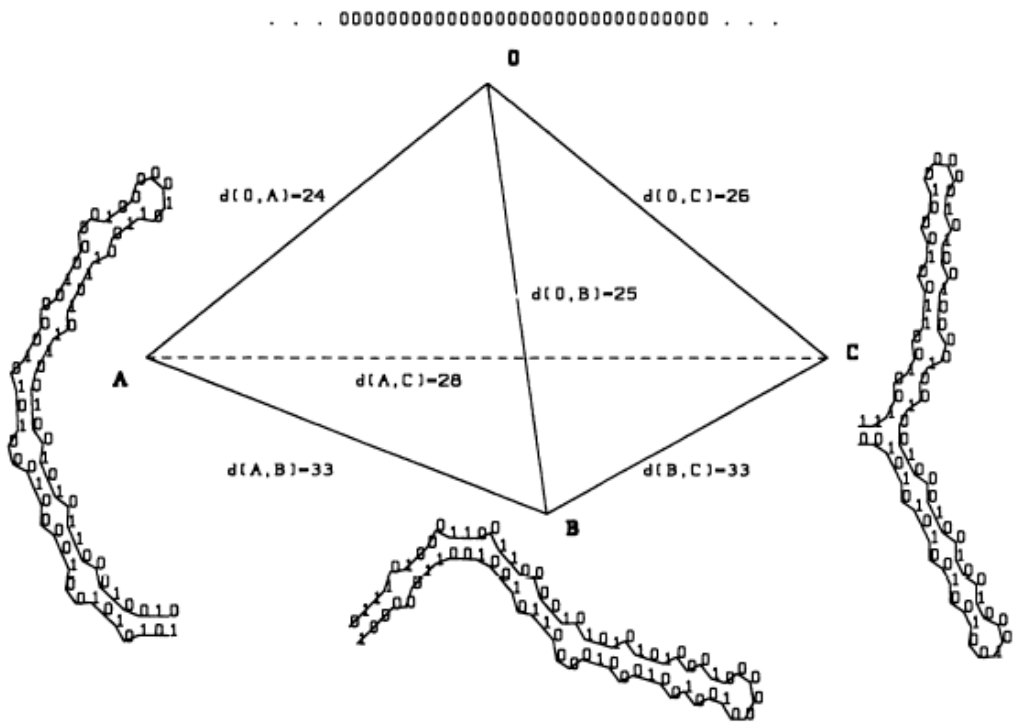


Fig. 2. The evolution reactor. This kind of flow reactor consists of a reaction vessel which allows for temperature and pressure control. Its walls are impermeable to polynucleotides. Energy-rich material is poured from the environment into the reactor. The degradation products are removed steadily. Material transport is adjusted in such a way that the concentration of monomers is constant in the reactor. A dilution flux Φ is installed in order to remove excess of polynucleotides produced by replication. Thus, the sum of the numbers of individual particles $\sum_i X_i(t) = N(t)$ may be controlled by the flux Φ. Under 'constant organization' Φ is adjusted such that $N(t) = \Theta$ is essentially constant. By this we indicate that fluctuations with standard deviation $\sigma = \sqrt{N}$ occur regularly. The regulation of Φ requires internal control, which can be achieved by logistic coupling.









Walter Fontana, Wolfgang Schnabl, and
Peter Schuster, Phys.Rev.A 40:3301-3321, 1989

Evolution *in silico*

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

random individuals. The primer pair used for genomic DNA amplification is 5′-TCTCCCTGGATTCT-CATTTA-3′ (forward) and 5′-TCTTTGTCTTCTGT-TCCACC-3′ (reverse). Reactions were performed in 25 μl using 1 unit of Taq DNA polymerase with each primer at 0.4 μM; 200 μM each dATP, dTTP, dGTP, and dCTP; and PCR buffer [10 mM tris-HCl (pH 8.3), 50 mM KCl_2,1.5 mM $MgCl_2$] in a cycle condition of 94°C for 1 min and then 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s followed by 72°C for 6 min. PCR products were purified (Qiagen), digested with Xmn I, and separated in a 2% agarose gel.

32. A nonsense mutation may affect mRNA stability and result in degradation of the transcript [L. Maquat, *Am. J. Hum. Genet.* **59**, 279 (1996)].
33. Data not shown; a dot blot with poly (A)$^+$ RNA from 50 human tissues (The Human RNA Master Blot, 7770-1, Clontech Laboratories) was hybridized with a probe from exons 29 to 47 of *MYO15* using the same condition as Northern blot analysis (*13*).
34. Smith–Magenis syndrome (SMS) is due to deletions of 17p11.2 of various sizes, the smallest of which includes *MYO15* and perhaps 20 other genes [(*6*); K-S Chen, L. Potocki, J. R. Lupski, *MRDD Res. Rev.* **2**, 122 (1996)]. *MYO15* expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for *MYO15* may explain a portion of the SMS phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in *MYO15* in trans to the SMS 17p11.2 deletion.
35. R. A. Fridell, data not shown.
36. K. B. Avraham *et al.*, *Nature Genet.* **11**, 369 (1995); X-Z. Liu *et al.*, *ibid.* **17**, 268 (1997); F. Gibson *et al.*, *Nature* **374**, 62 (1995); D. Weil *et al.*, *ibid.*, p. 60.
37. RNA was extracted from cochlea (membranous labyrinths) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly (A)$^+$ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human *MYO15*-specific oligonucleotide primers (forward, 5′-GCATGACCTGCCGGCTAAT-GGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGT-GCTCGGCTGGC-3′). Cycling conditions were 40 s at 94°C; 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR product is expected from amplification of the human *MYO15* cDNA. Amplification of human genomic DNA with this primer pair would result in a 2903-bp fragment.

38. We are grateful to the people of Bengkala, Bali, and the two families from India. We thank J. R. Lupski and K.-S. Chen for providing the human chromosome 17 cosmid library. For technical and computational assistance, we thank N. Dietrich, M. Fergusson, A. Gupta, E. Sorbello, R. Torkzadeh, C. Varner, M. Walker, G. Bouffard, and S. Beckstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Hinnant, I. N. Arhya, and S. Winata for assistance in Bali, and T. Barber, S. Sullivan, E. Green, D. Drayna, and J. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (Z01 DC 00035-01 and Z01 DC 00038-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.C.M.), the National Institute of Child Health and Human Development (R01 HD30428 to S.A.C.) and a National Science Foundation Graduate Research Fellowship to F.J.P. This paper is dedicated to J. B. Snow Jr. on his retirement as the Director of the NIDCD.

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

To distinguish continuous from discontinuous evolutionary change, a relation of nearness between phenotypes is needed. Such a relation is based on the probability of one phenotype being accessible from another through changes in the genotype. This nearness relation is exemplified by calculating the shape neighborhood of a transfer RNA secondary structure and provides a characterization of discontinuous shape transformations in RNA. The simulation of replicating and mutating RNA populations under selection shows that sudden adaptive progress coincides mostly, but not always, with discontinuous shape transformations. The nature of these transformations illuminates the key role of neutral genetic drift in their realization.

A much-debated issue in evolutionary biology concerns the extent to which the history of life has proceeded gradually or has been punctuated by discontinuous transitions at the level of phenotypes (*1*). Our goal is to make the notion of a discontinuous transition more precise and to understand how it arises in a model of evolutionary adaptation.

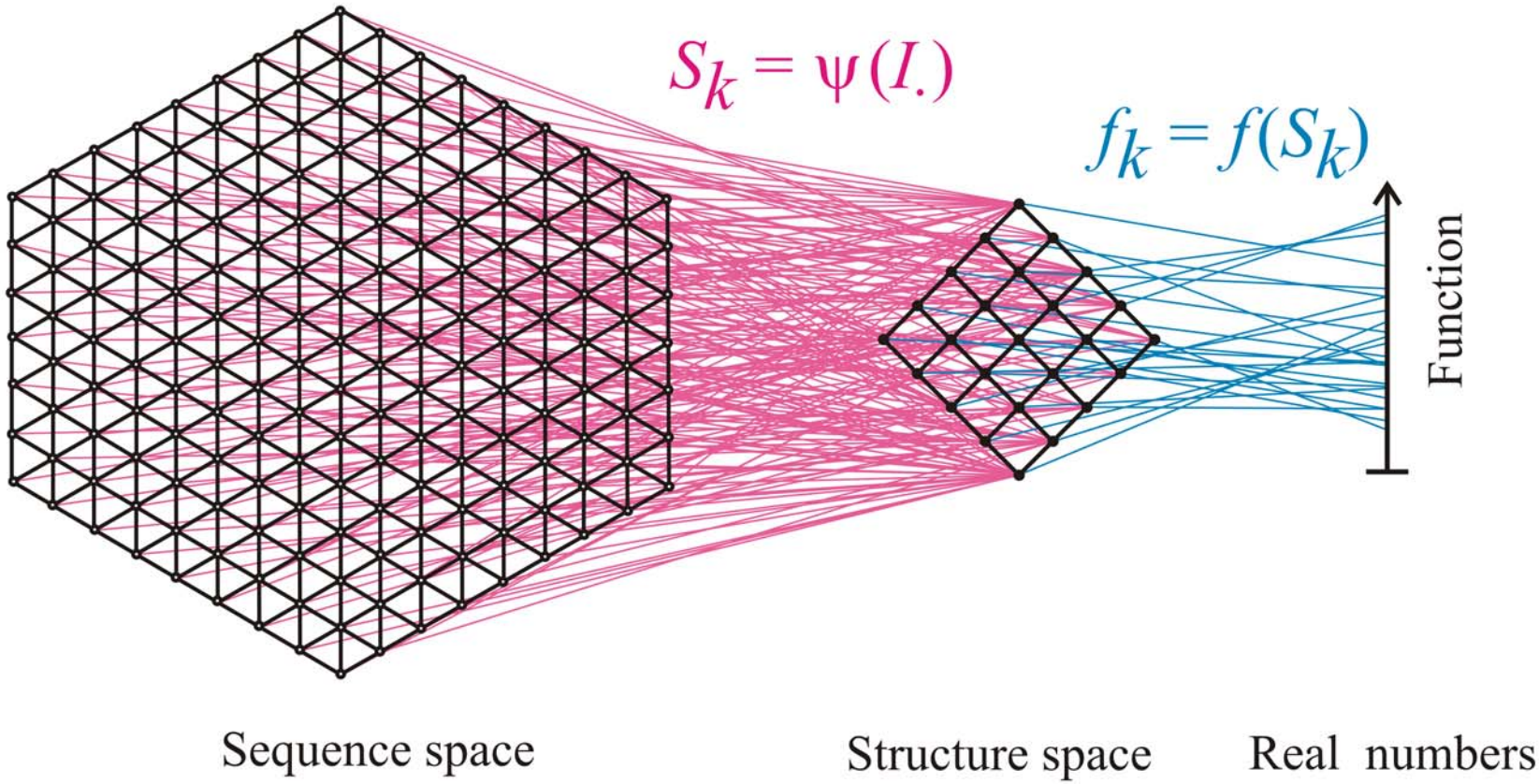
We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empirically well defined and obtain their biophysical and biochemical importance from being a scaffold for the tertiary structure. For the sake of brevity, we shall refer to secondary structures as "shapes." RNA combines in a single molecule both genotype (replicatable sequence) and phenotype (selectable shape), making it ideally suited for in vitro evolution experiments (*3*, *4*).

To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (*5*, *6*). In the laboratory, a goal might be to find an RNA aptamer binding specifically to a molecule (*4*). Although in the experiment the evolutionary end product was unknown, we thought of its shape as being specified implicitly by the imposed selection criterion. Because our intent is to study evolutionary histories rather than end products, we defined a target shape in advance and assumed the replication rate of a sequence to be a function of the similarity between its shape and the target. An actual situation may involve more than one best shape, but this does not affect our conclusions.

An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces the average distance to the target (inversely related to fitness) in the population against time. Aside from a short initial phase, the entire history is dominated by steps, that is, flat periods of no apparent adaptive progress, interrupted by sudden approaches toward the target structure (*7*). However, the dominant shapes in the population not only change at these marked events but undergo several fitness-neutral transformations during the periods of no apparent progress. Although discontinuities in the fitness trace are evident, it is entirely unclear when and on the basis of what the series of successive phenotypes itself can be called continuous or discontinuous.

A set of entities is organized into a (topological) space by assigning to each entity a system of neighborhoods. In the present case, there are two kinds of entities: sequences and shapes, which are related by a thermodynamic folding procedure. The set of possible sequences (of fixed length) is naturally organized into a space because point mutations induce a canonical neighborhood. The neighborhood of a sequence consists of all its one-error mutants. The problem is how to organize the set of possible shapes into a space. The issue arises because, in contrast to sequences, there are

Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.



Mapping from sequence space into structure space and into function

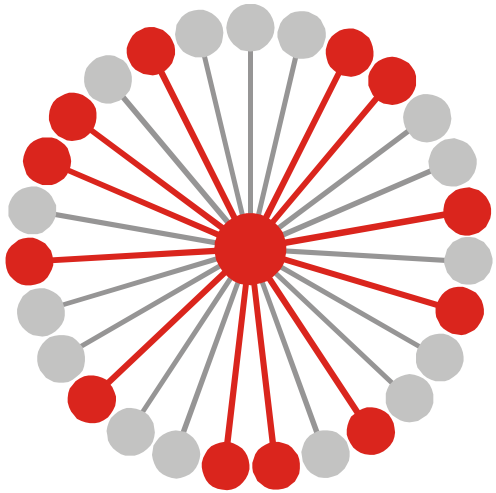
Neutral networks are sets of sequences forming the same object in a phenotype space. The neutral network $\mathbf{G_k}$ is, for example, the pre-image of the structure S_k in sequence space:

$$\mathbf{G_k} = \psi^{-1}(S_k) \pi \{\psi_j \mid \psi(I_j) = S_k\}$$

The set is converted into a graph by connecting all sequences of Hamming distance one.

Neutral networks of small biomolecules can be computed by exhaustive folding of complete sequence spaces, i.e. all RNA sequences of a given chain length. This number, $\mathbf{N=4^n}$, becomes very large with increasing length, and is prohibitive for numerical computations.

Neutral networks can be modelled by **random graphs** in sequence space. In this approach, nodes are inserted randomly into sequence space until the size of the pre-image, i.e. the number of neutral sequences, matches the neutral network to be studied.



$$G_k = \psi^{-1}(S_k) \cup \{ I_j \mid \psi(I_j) = S_k \}$$

$$\lambda_j = 12 / 27 = 0.444 , \quad \bar{\lambda}_k = \frac{\sum_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

Connectivity threshold: $\lambda_{cr} = 1 - \kappa^{-1/(\kappa-1)}$

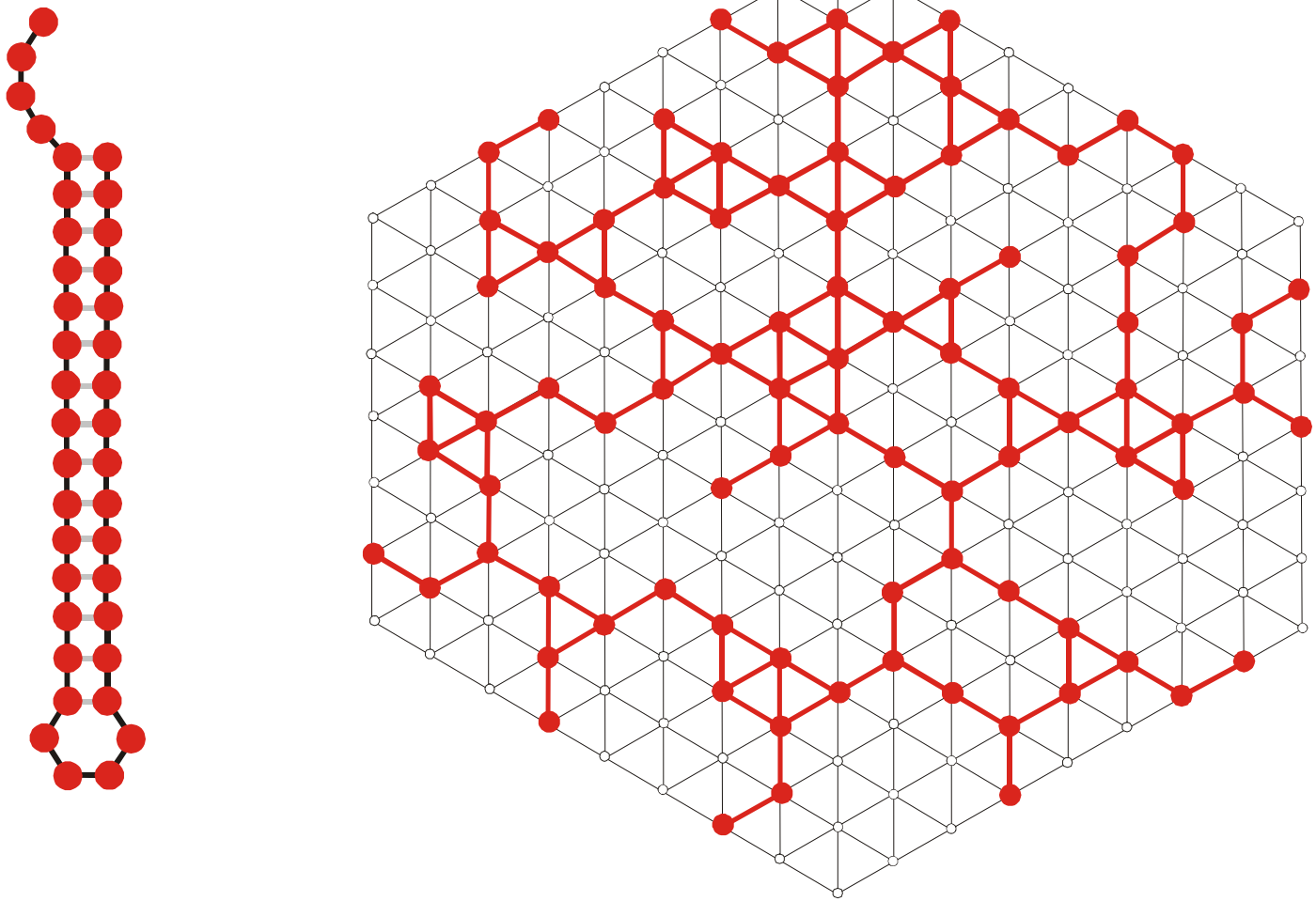
Alphabet size κ : **AUGC** $\Rightarrow \kappa = 4$

$\bar{\lambda}_k > \lambda_{cr}$ .... network G_k is connected

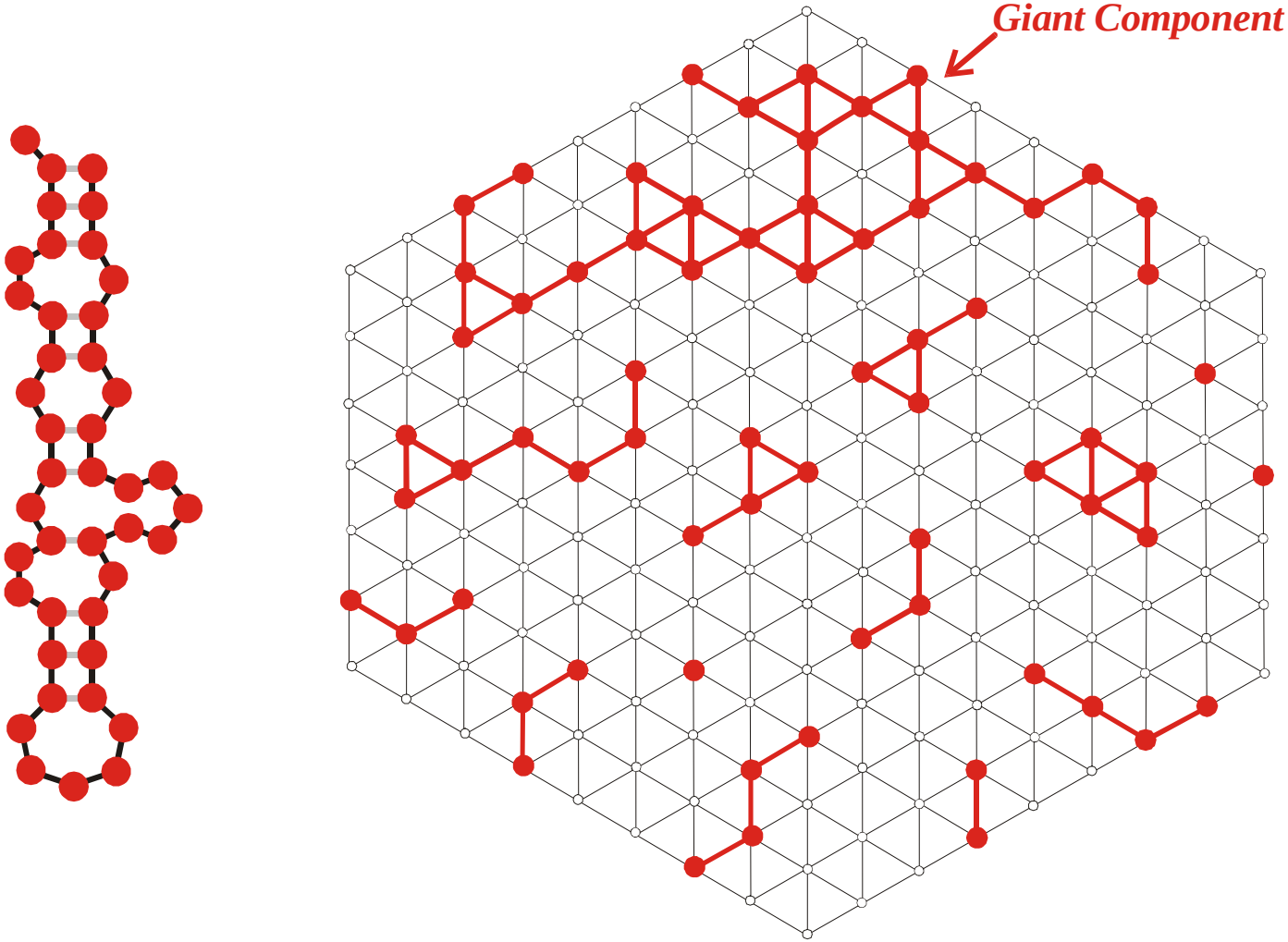
$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

κ	λ_{cr}	
2	0.5	**GC,AU**
3	0.423	**GUC,AUG**
4	0.370	**AUGC**

Mean degree of neutrality and connectivity of neutral networks



A connected neutral network formed by a common structure



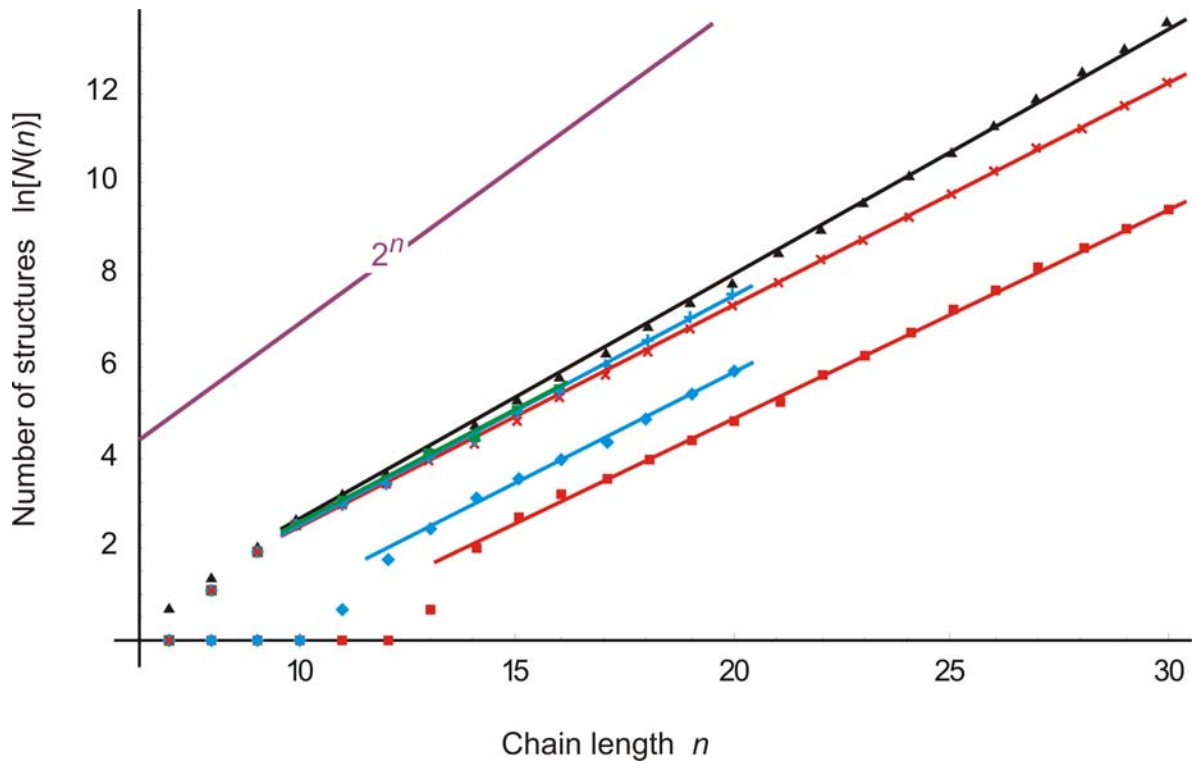
A multi-component neutral network formed by a rare structure

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures

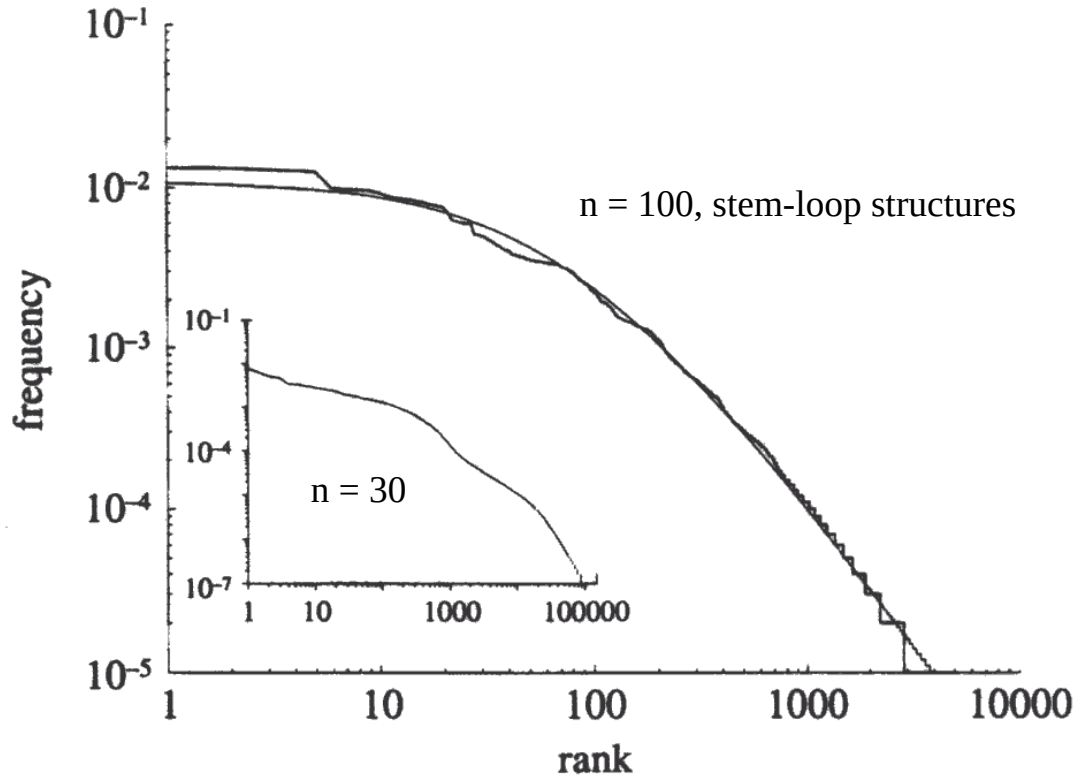


Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures



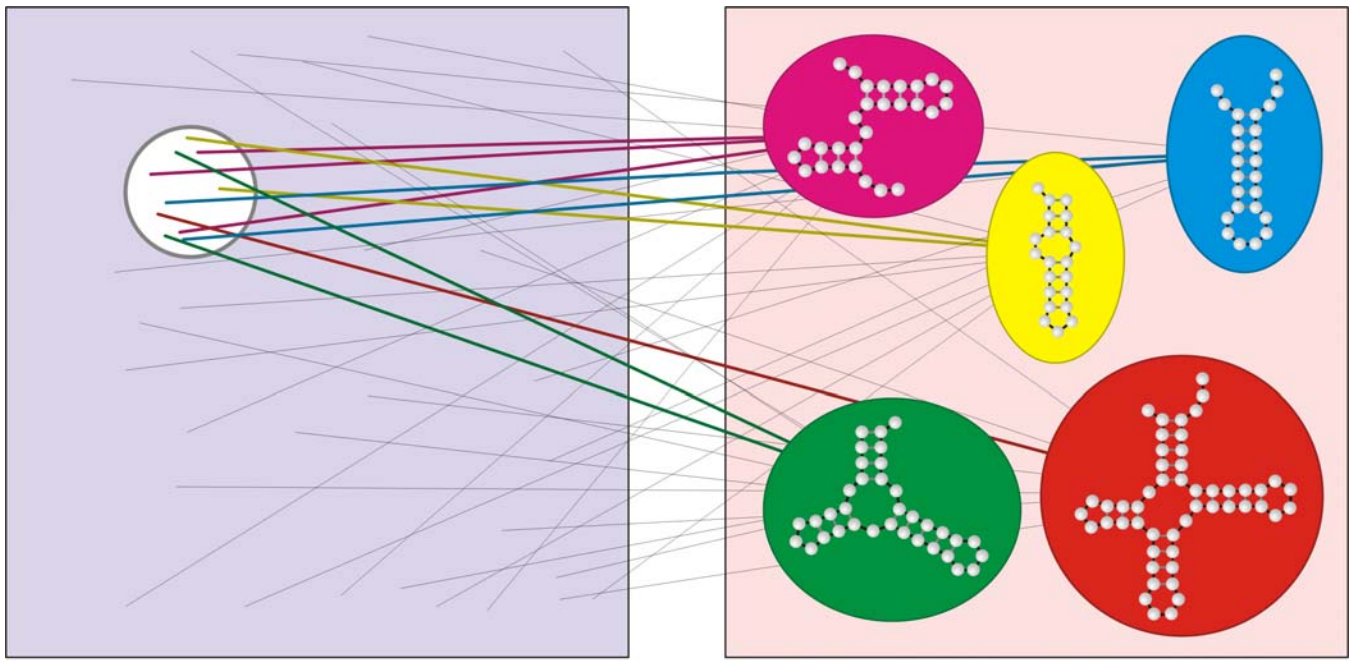
RNA secondary structures and Zipf's law

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures

Properties of RNA sequence to secondary structure mapping

1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures

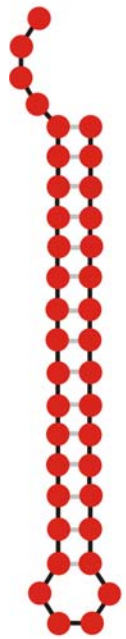


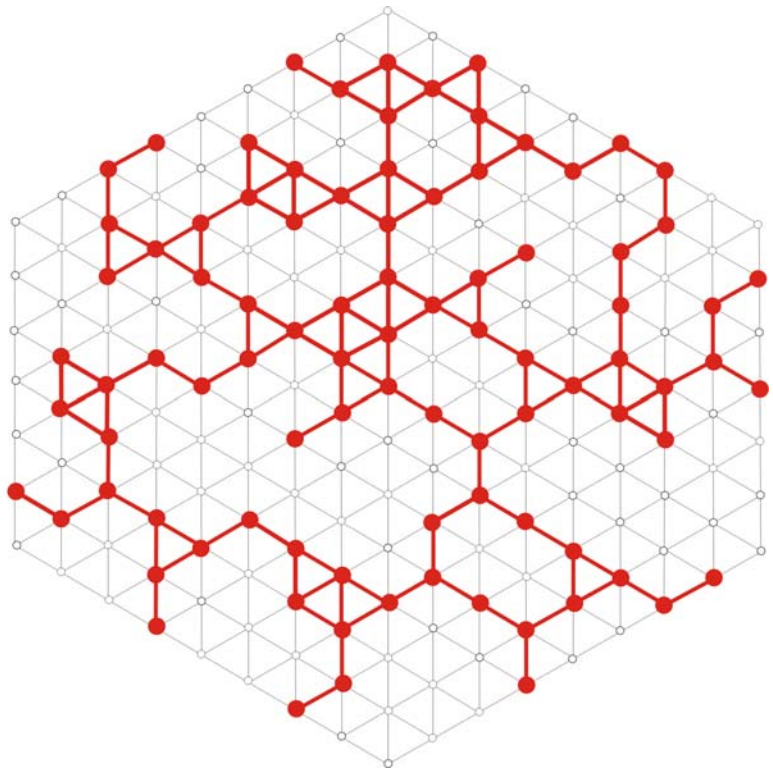
Properties of RNA sequence to secondary structure mapping

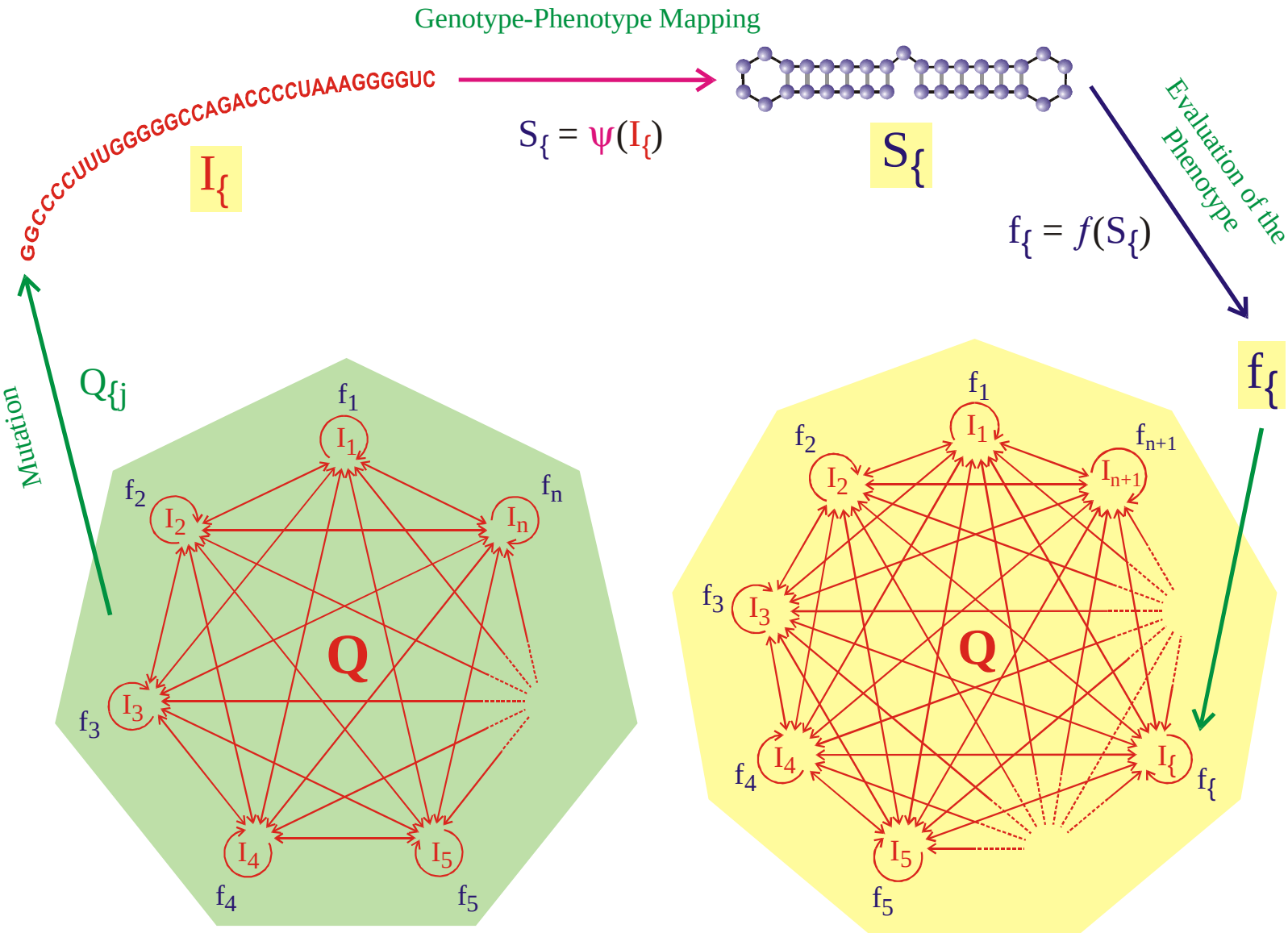
1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures
4. Neutral networks of common structures are connected

Properties of RNA sequence to secondary structure mapping

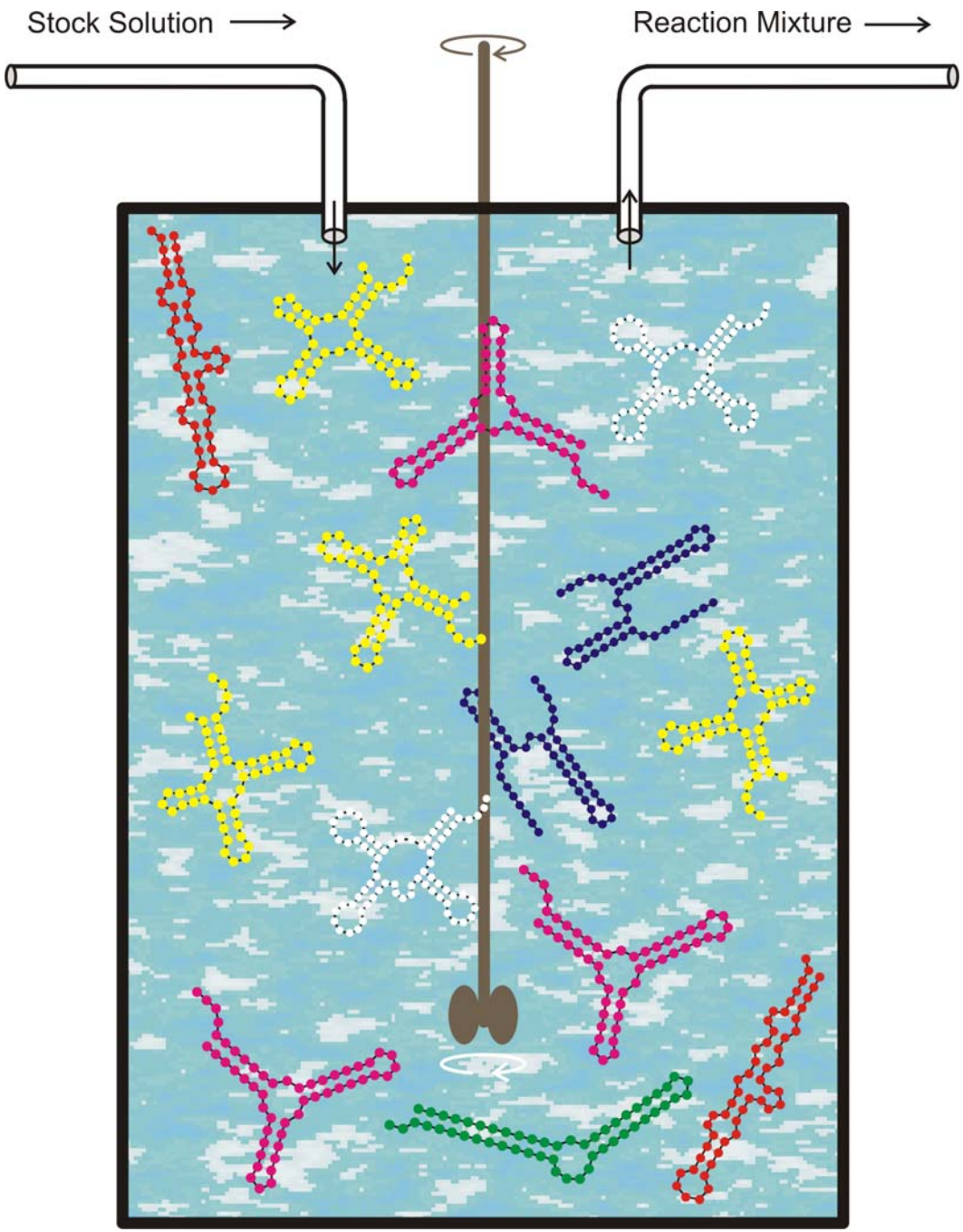
1. More sequences than structures
2. Few common versus many rare structures
3. Shape space covering of common structures
4. Neutral networks of common structures are connected







Evolutionary dynamics
including molecular phenotypes



Replication rate constant:

$$f_k = \gamma / [\alpha + \Delta d_S^{(k)}]$$

$$\Delta d_S^{(k)} = d_H(S_k, S_\tau)$$

Selection constraint:

Population size, N = # RNA molecules, is controlled by the flow

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

Mutation rate:

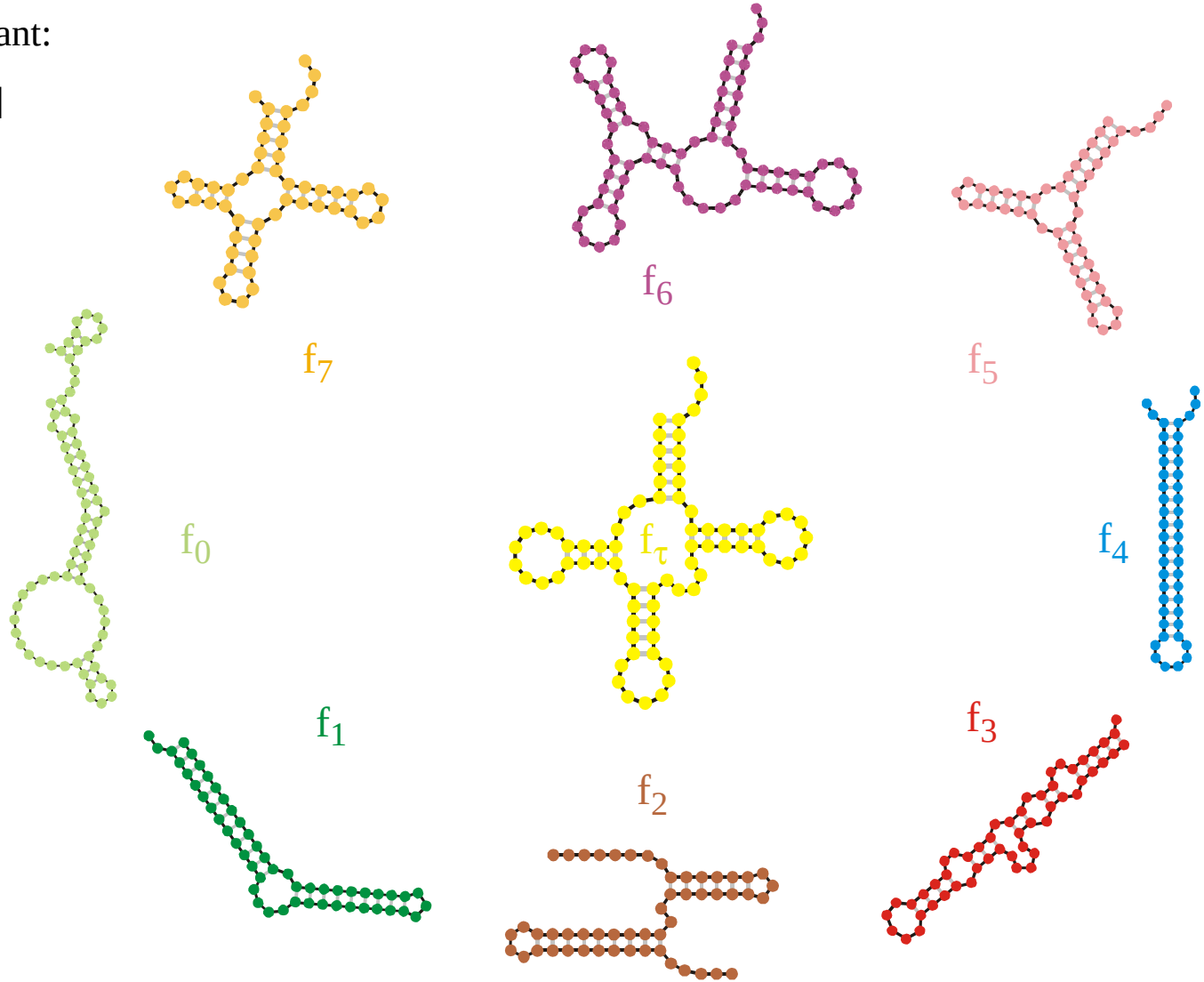
p = 0.001 / site × replication

The flowreactor as a device for **studies** of evolution ***in vitro*** and ***in silico***

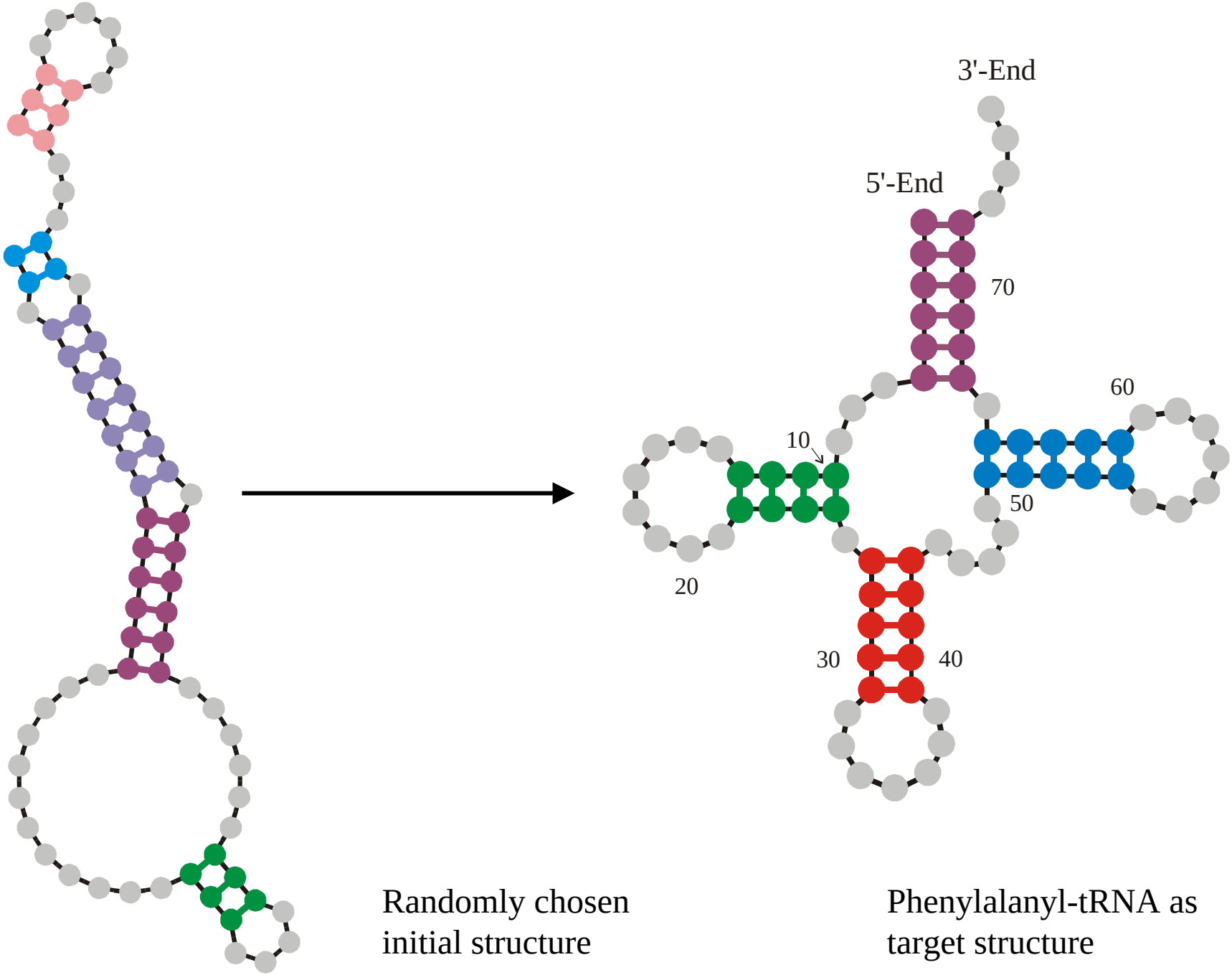
Replication rate constant:

$$f_k = \gamma / [\alpha + \Delta d_S^{(k)}]$$

$$\Delta d_S^{(k)} = d_H(S_k, S_\tau)$$

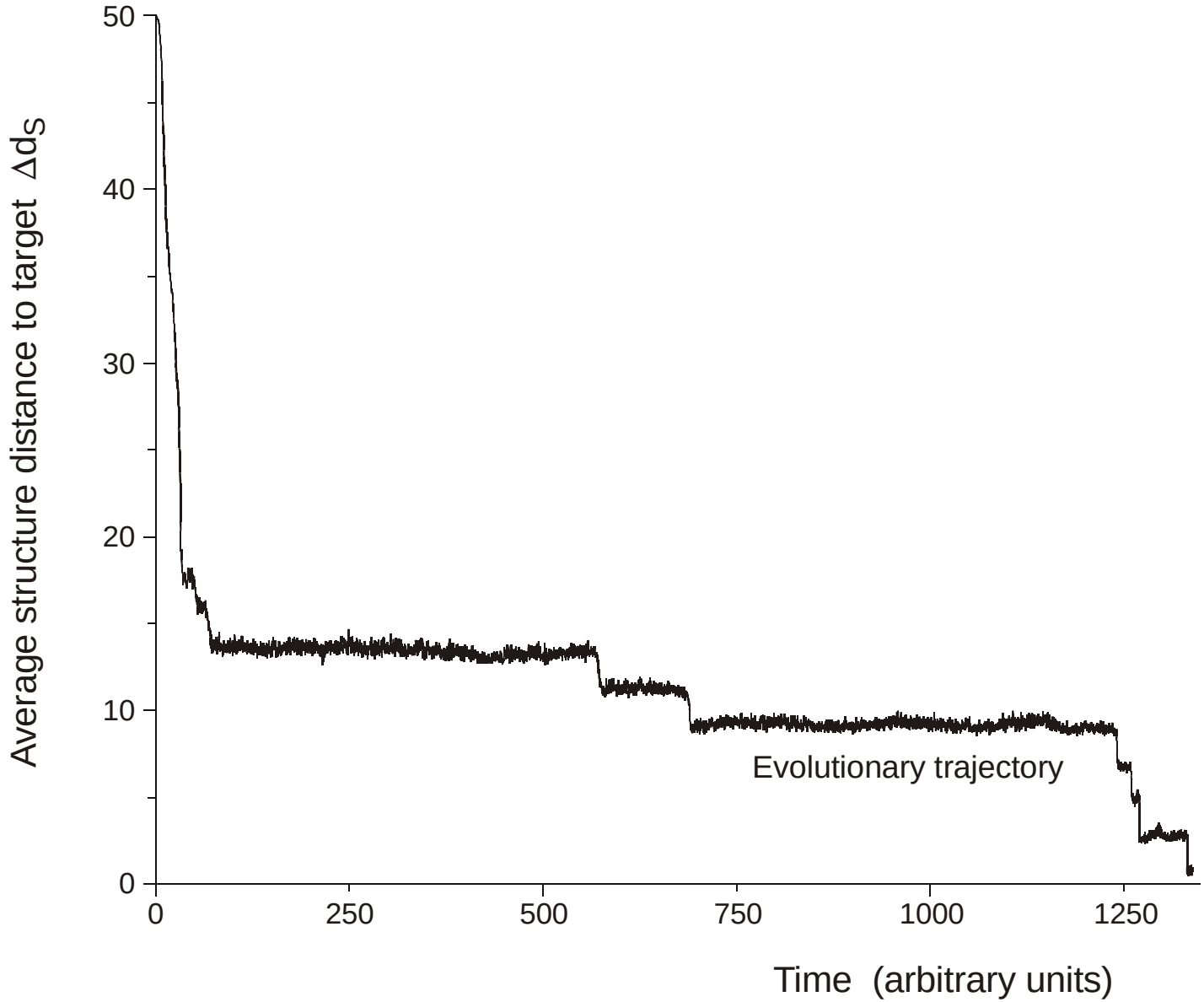


Evaluation of RNA secondary structures yields replication rate constants



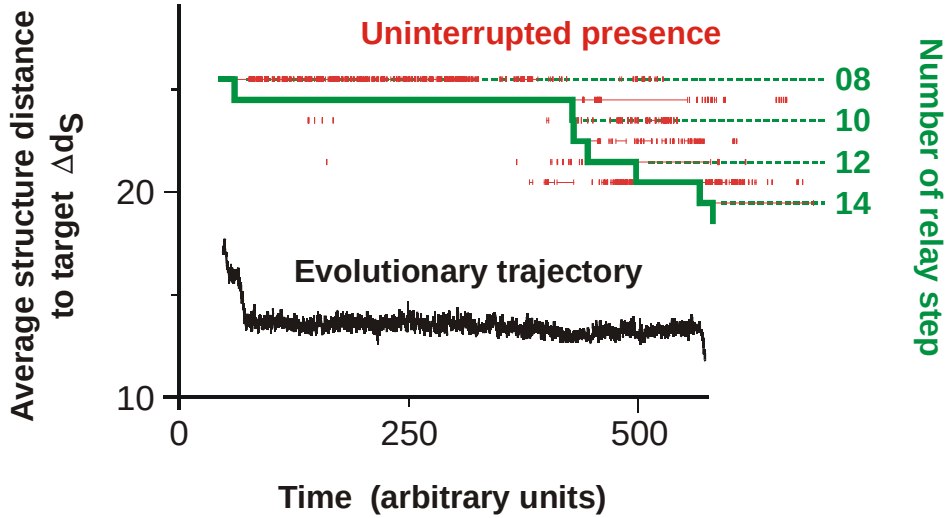
Randomly chosen initial structure

Phenylalanyl-tRNA as target structure



In silico optimization in the flow reactor: Evolutionary trajectory

28 neutral point mutations during a long quasi-stationary epoch




```
entry  GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8      .((((((((((((.........(((....)))........)))))).....((((.........))))))))))....
exit   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry  GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA
9      .((((((.((((.........(((....)))........)))))).....((((.........))))).))))))....
exit   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG
entry  UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
10     .(((((..((((.........(((....)))........)))))).....((((.........))))).)))))....
exit   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
```

Transition inducing point mutations **Neutral point mutations**

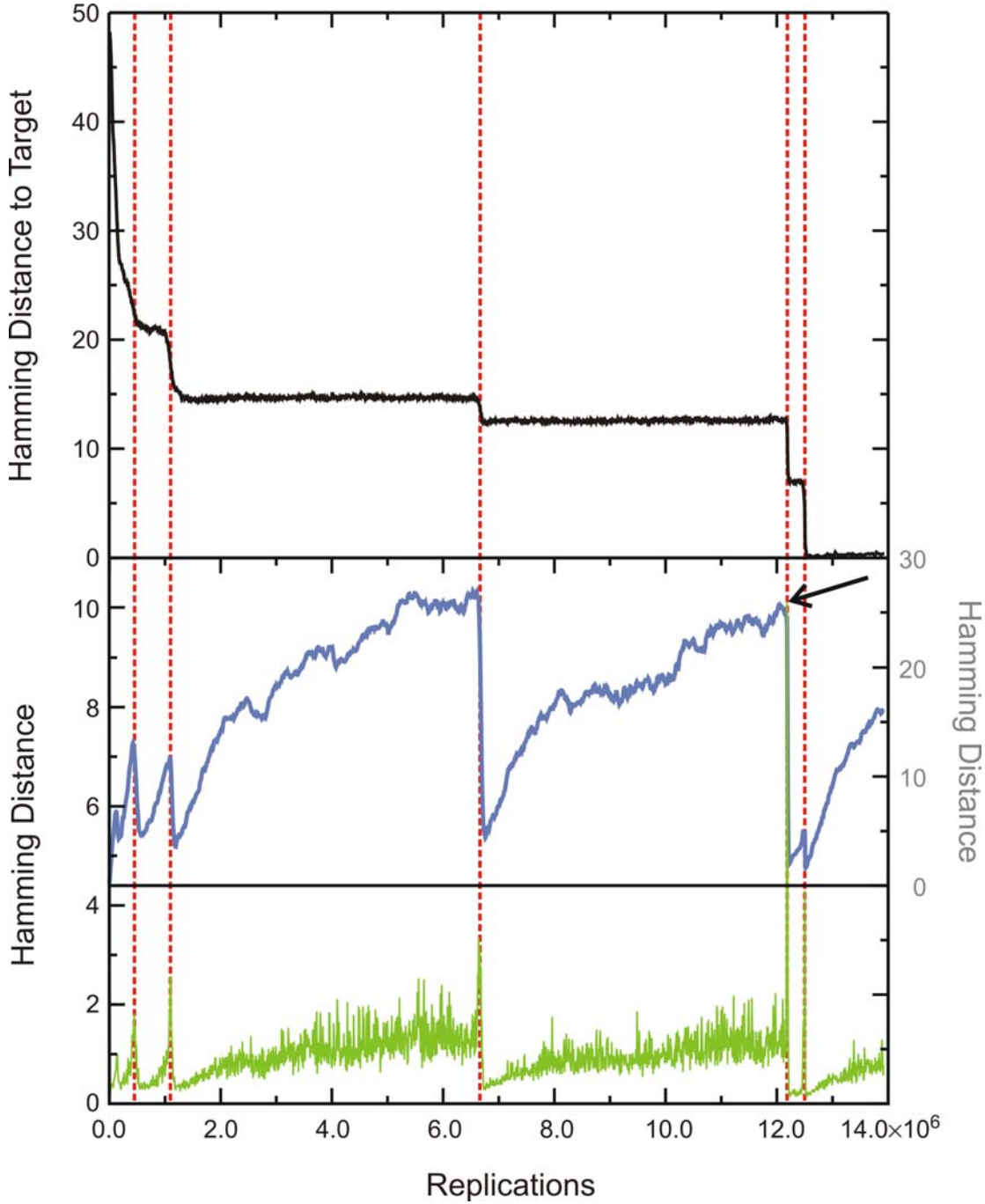
Neutral genotype evolution during phenotypic stasis

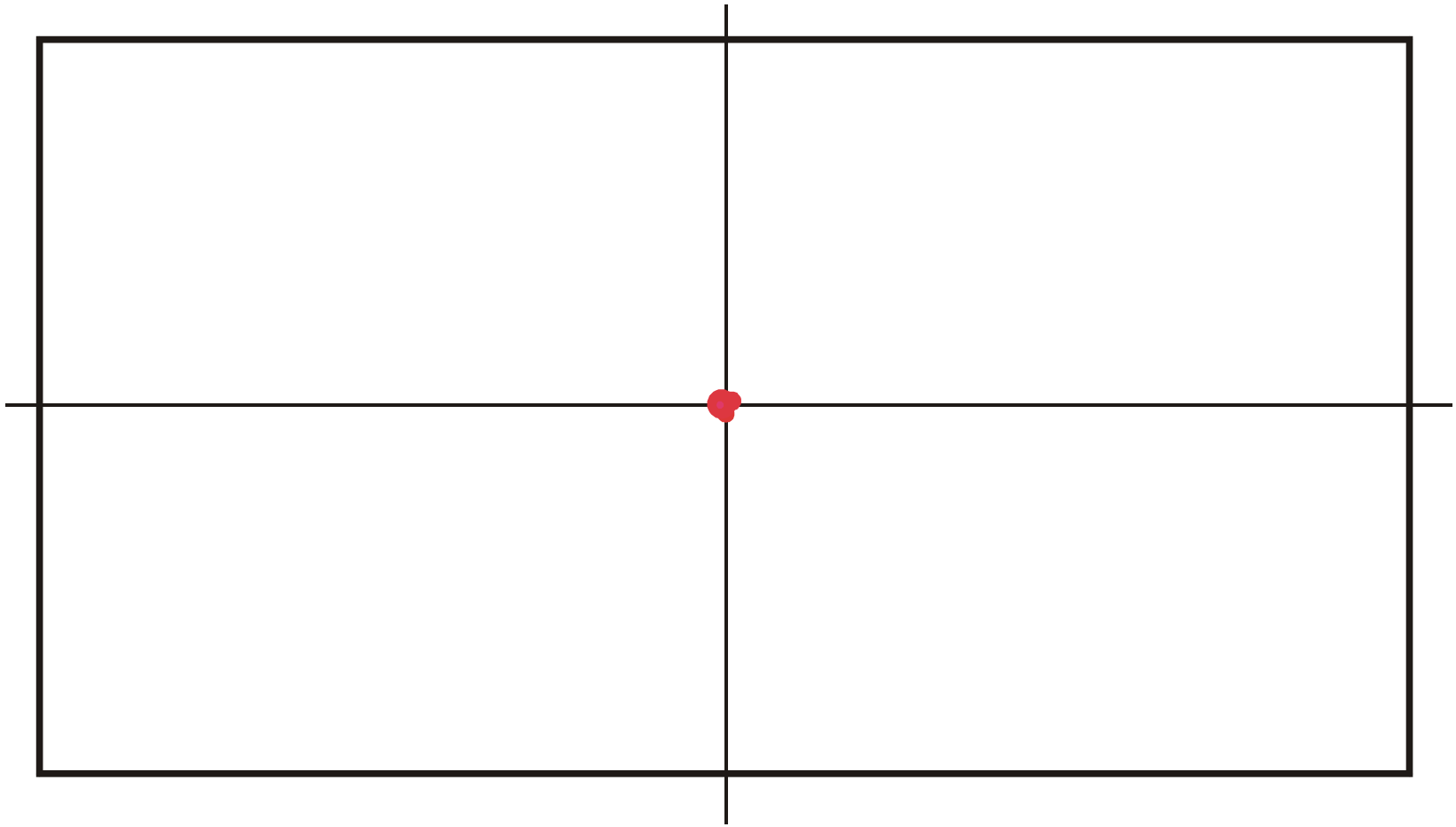
1. Solutions to Levinthal‘s paradox
2. Selection as solution to low probabilities in evolution
3. Origin of information by mutation and selection
4. Evolution of RNA phenotypes
5. **The role of neutrality in molecular evolution**
6. An experiment with RNA molecules
7. Summary

Evolutionary trajectory

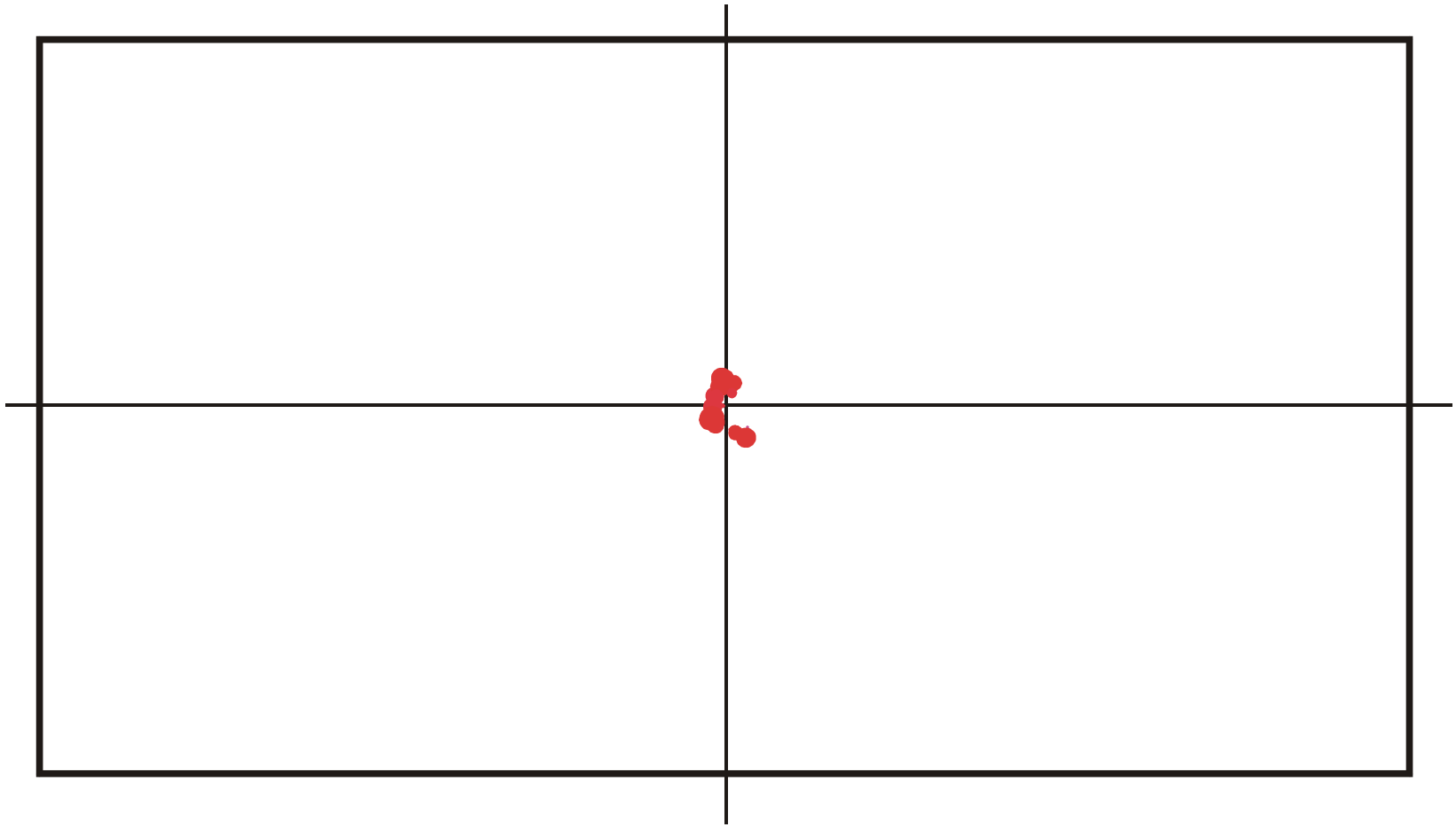
Spreading of the population through diffusion on the neutral network

Velocity of the population center in sequence space

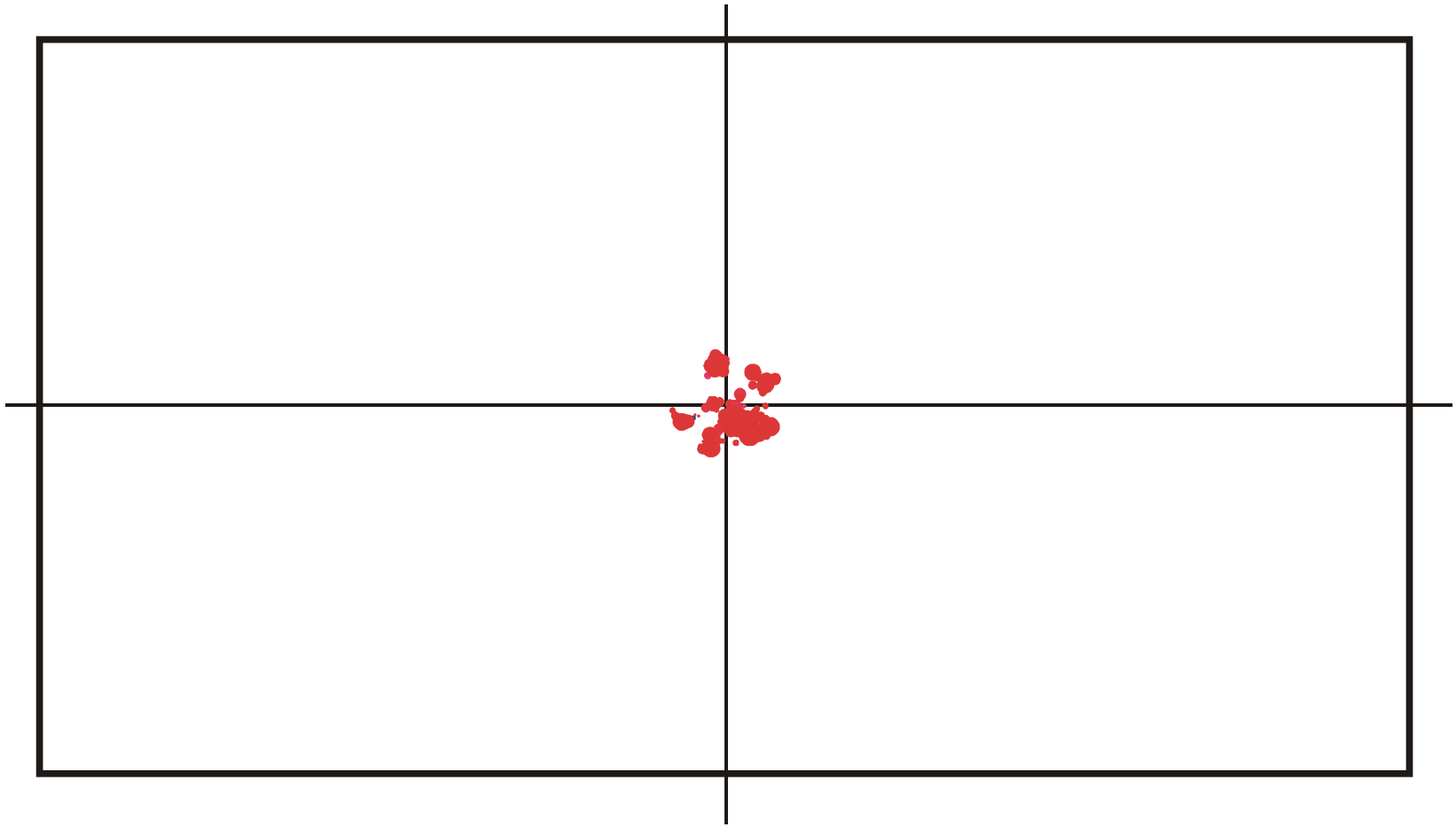




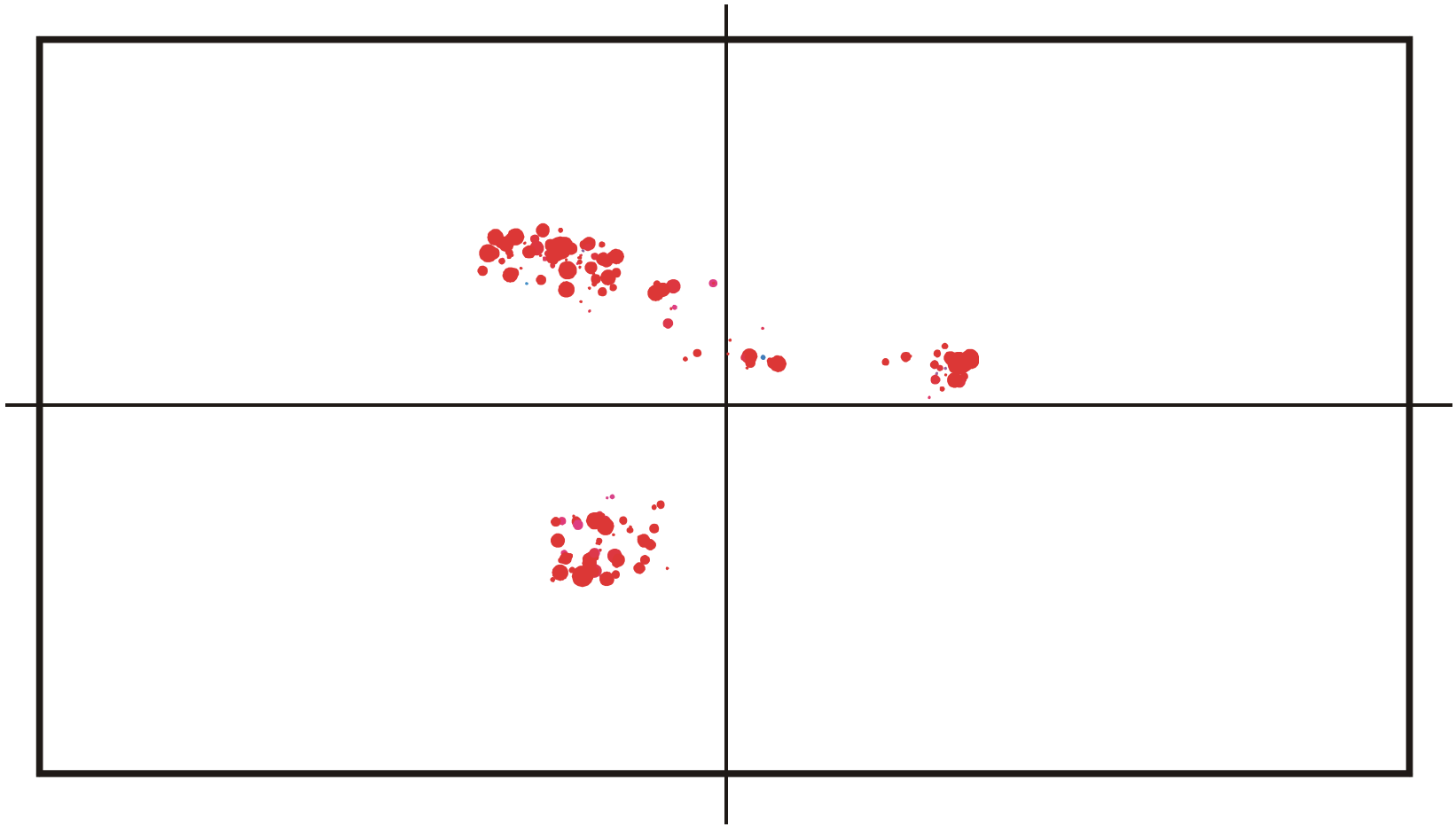
Spread of population in sequence space during a quasistationary epoch: t = 150



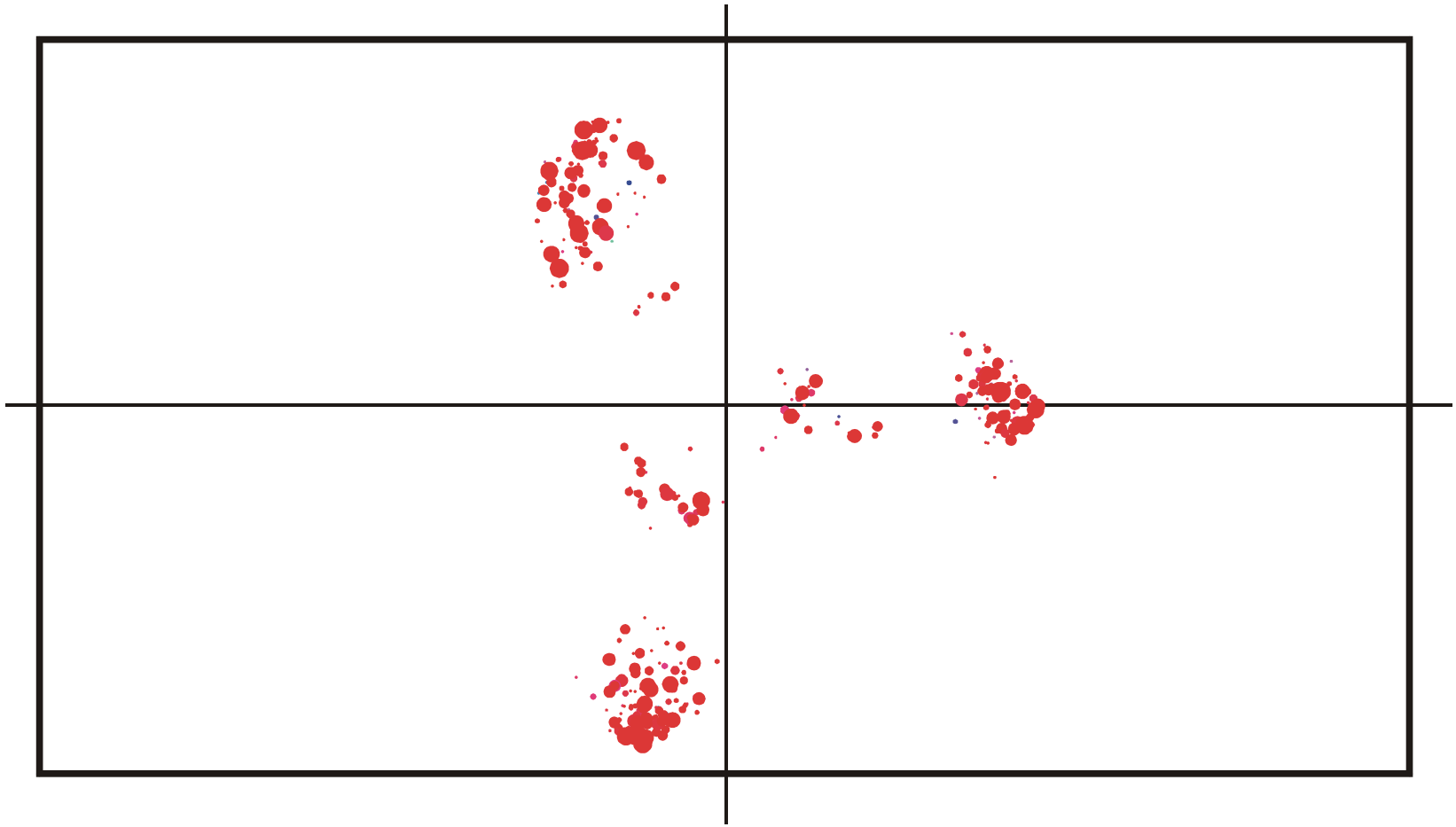
Spread of population in sequence space during a quasistationary epoch: t = 170



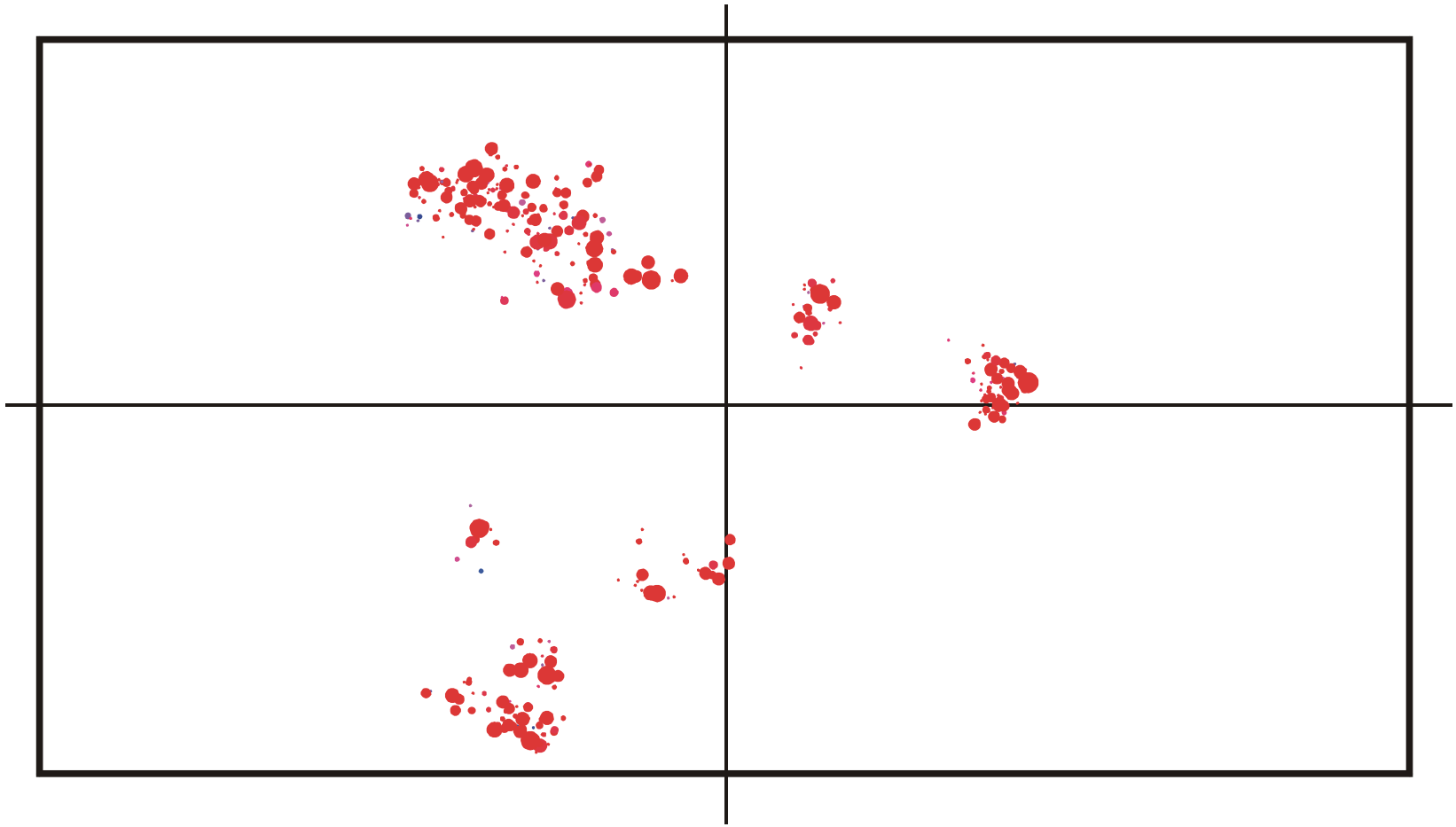
Spread of population in sequence space during a quasistationary epoch: t = 200



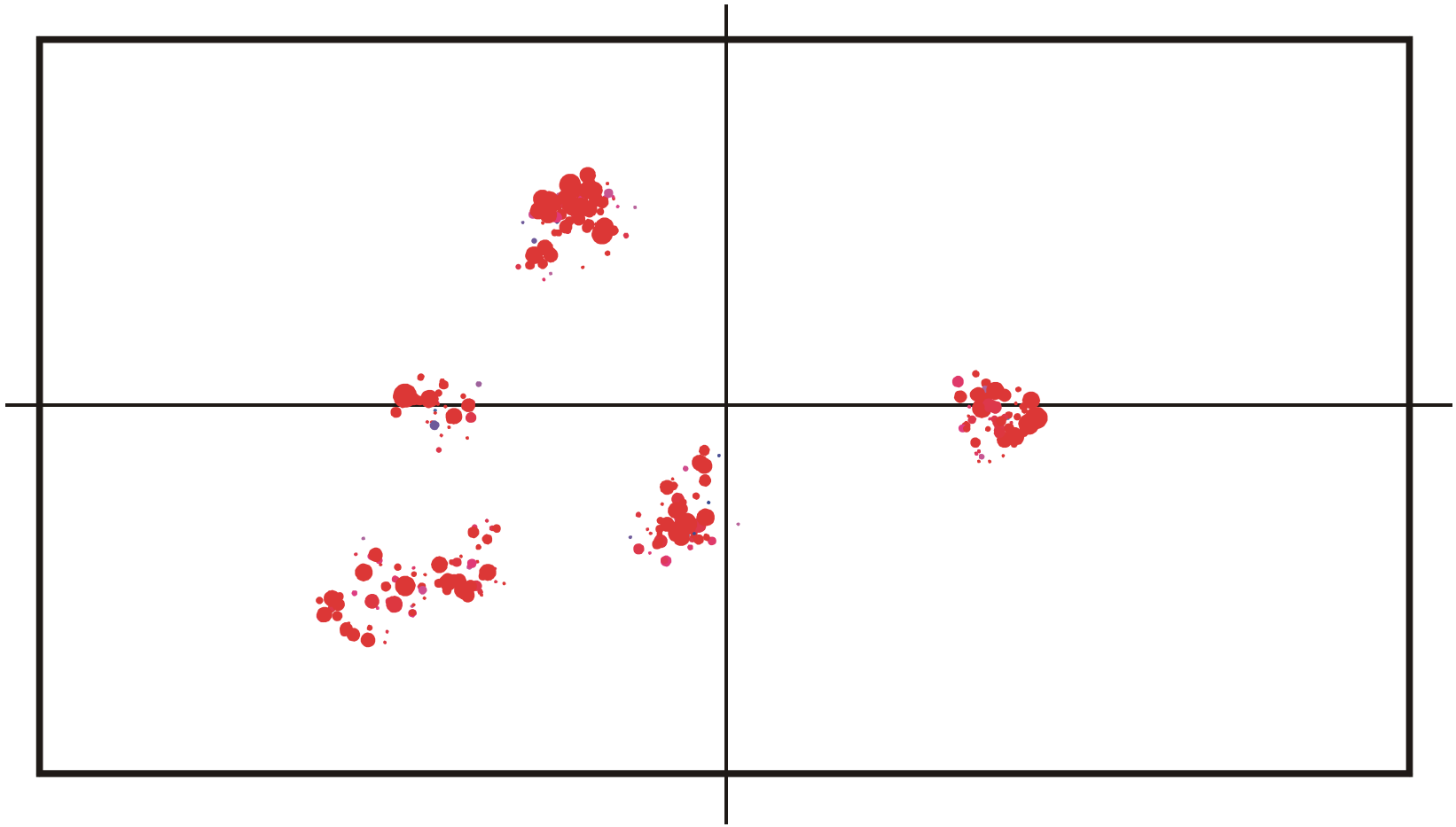
Spread of population in sequence space during a quasistationary epoch: t = 350



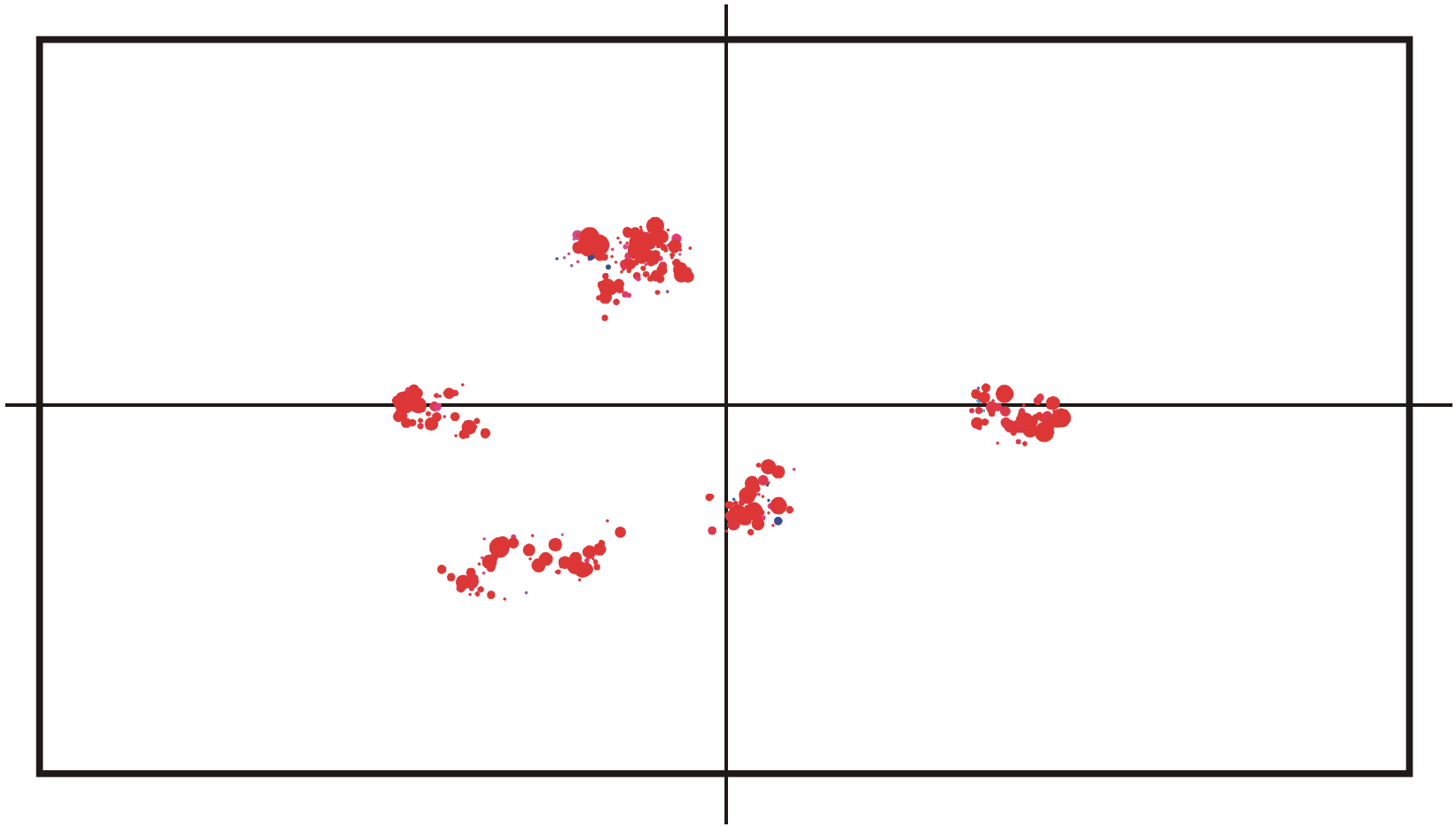
Spread of population in sequence space during a quasistationary epoch: t = 500



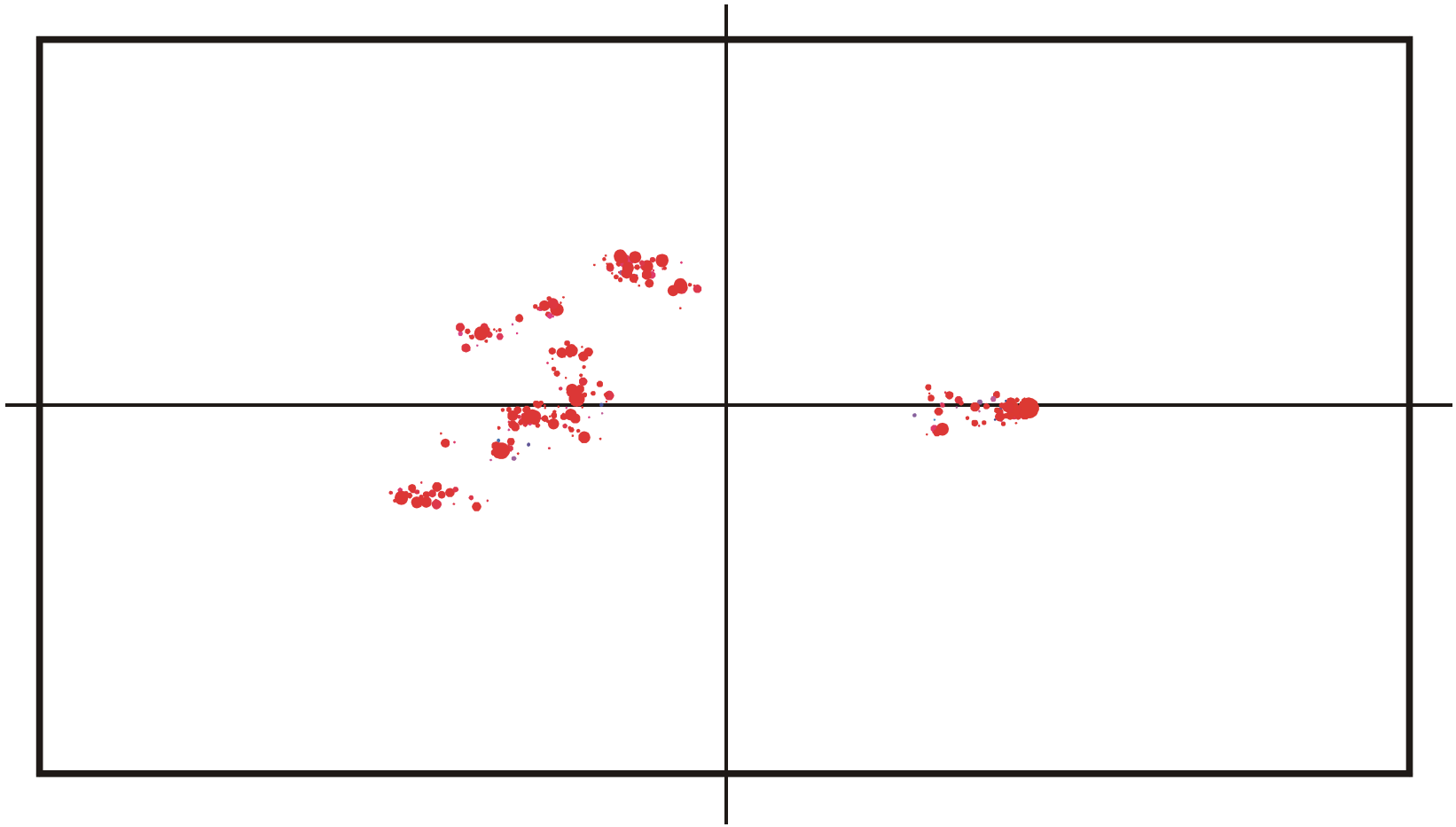
Spread of population in sequence space during a quasistationary epoch: t = 650



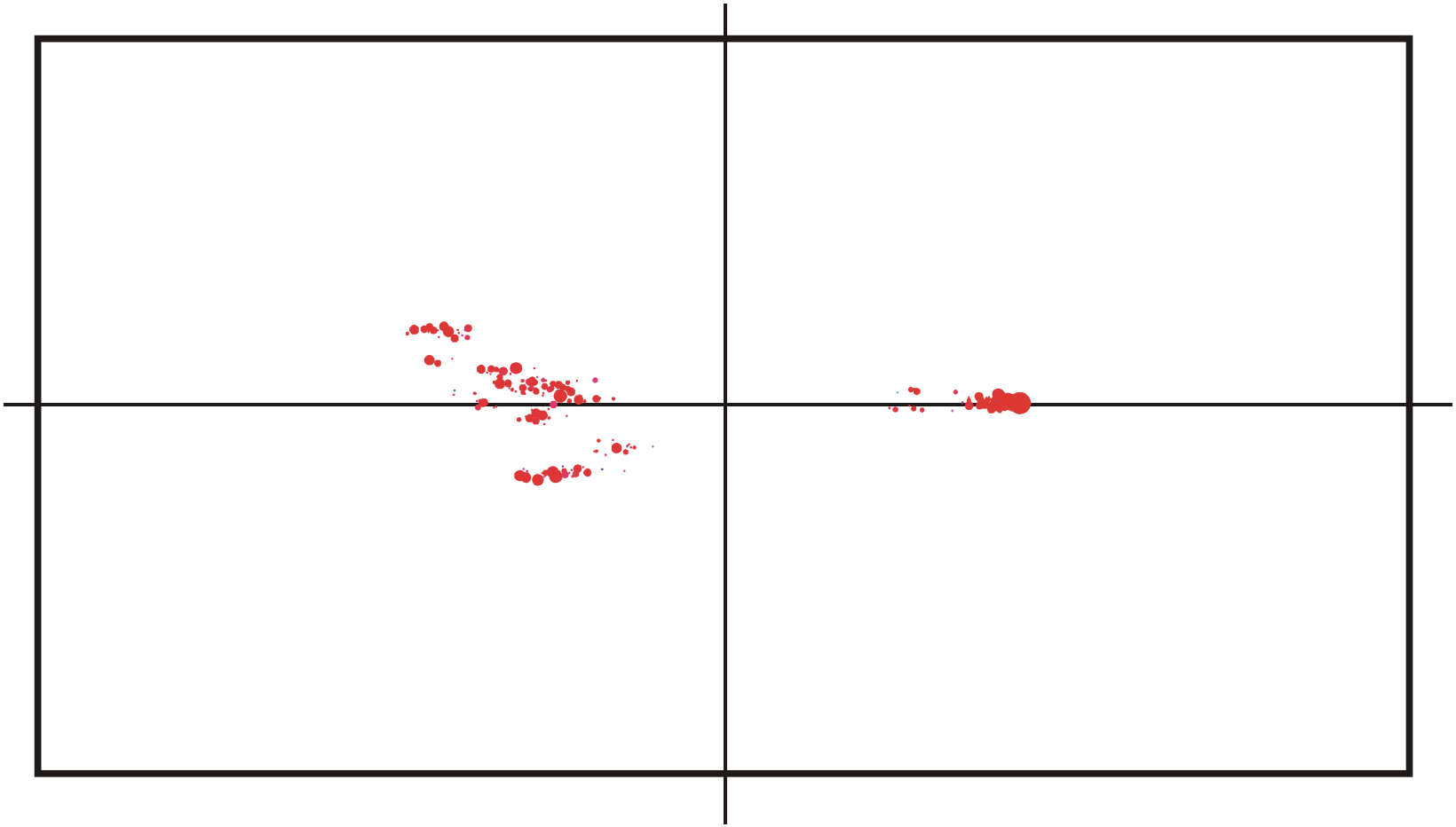
Spread of population in sequence space during a quasistationary epoch: t = 820



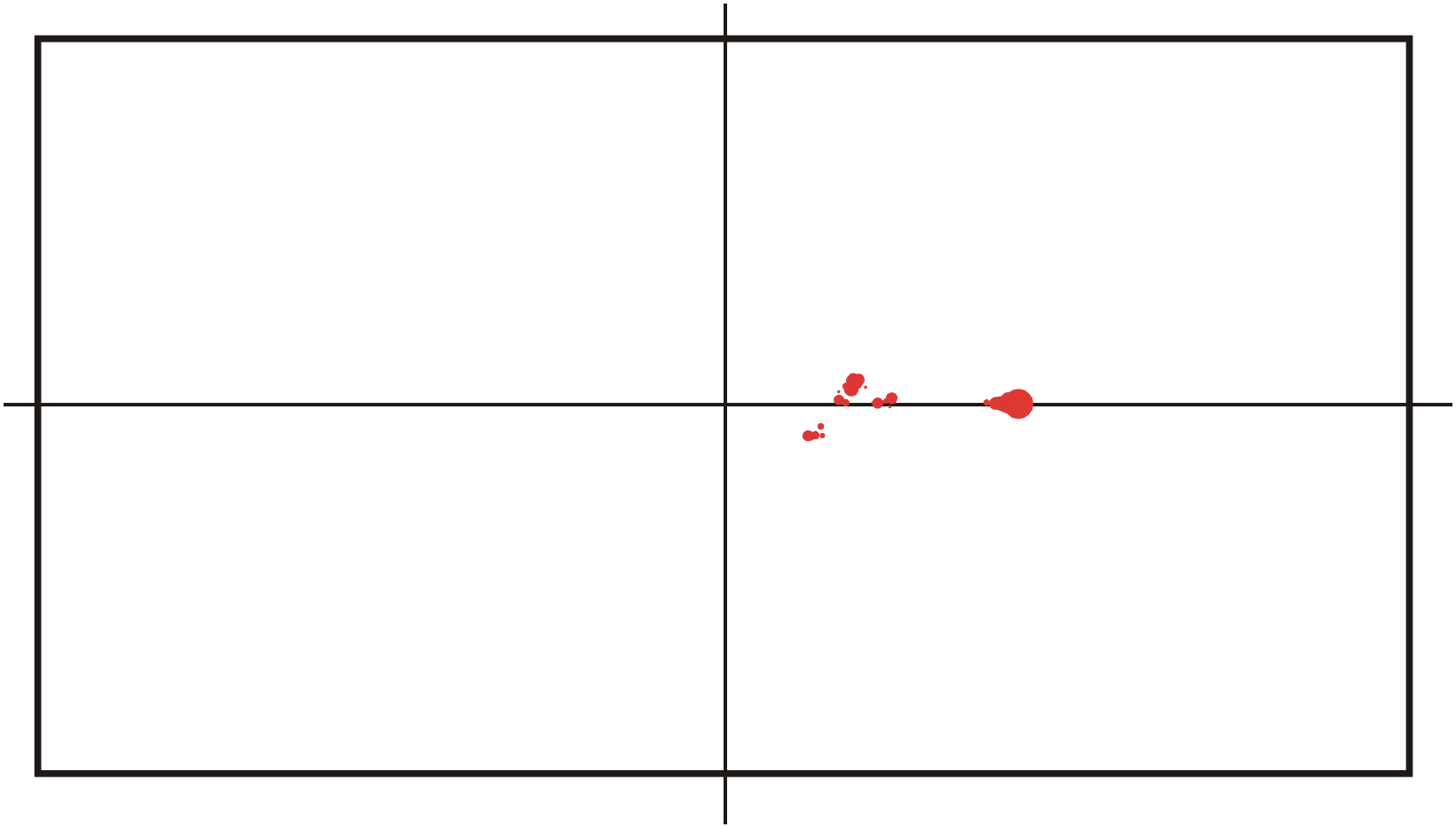
Spread of population in sequence space during a quasistationary epoch: t = 825



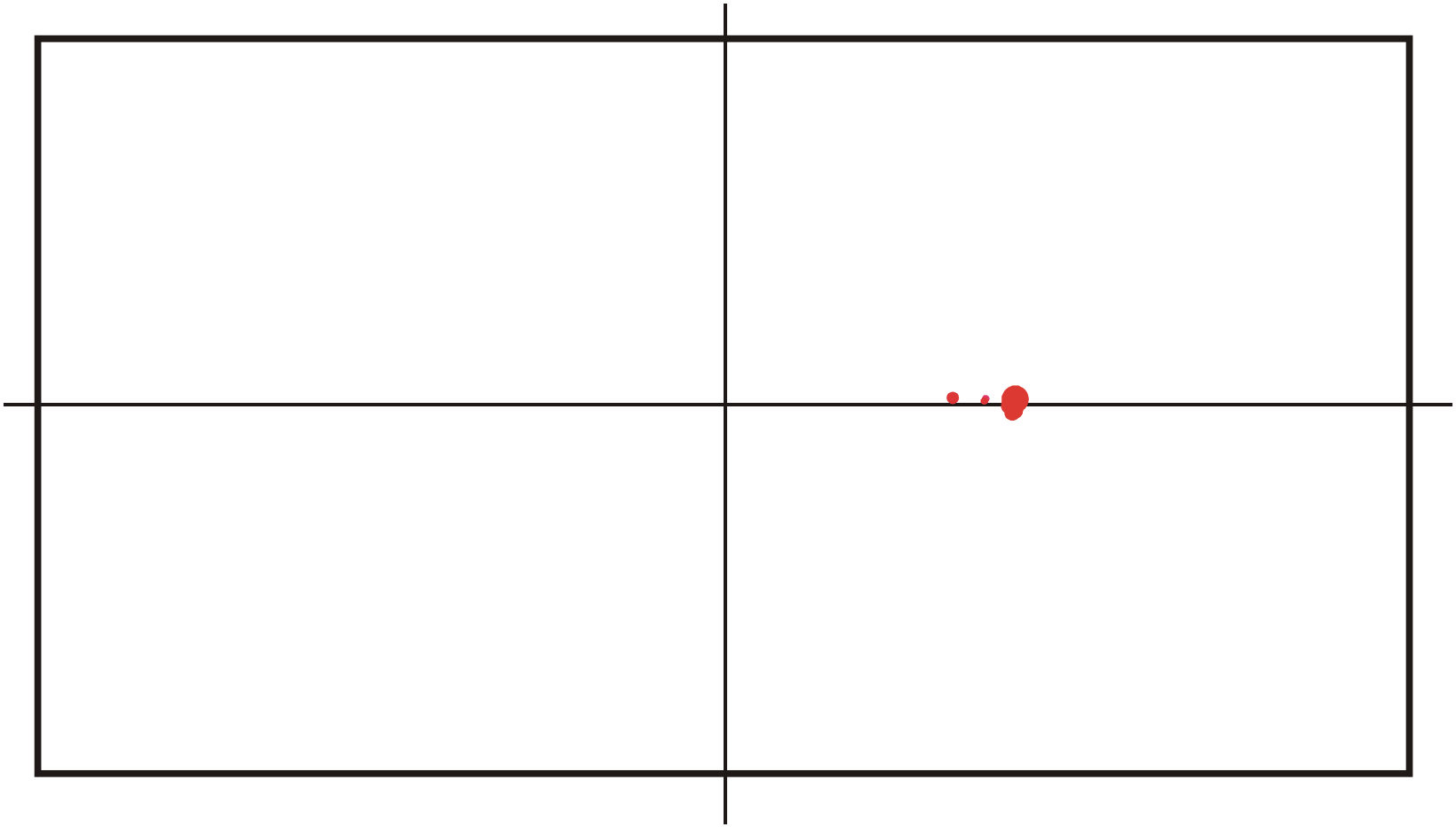
Spread of population in sequence space during a quasistationary epoch: t = 830



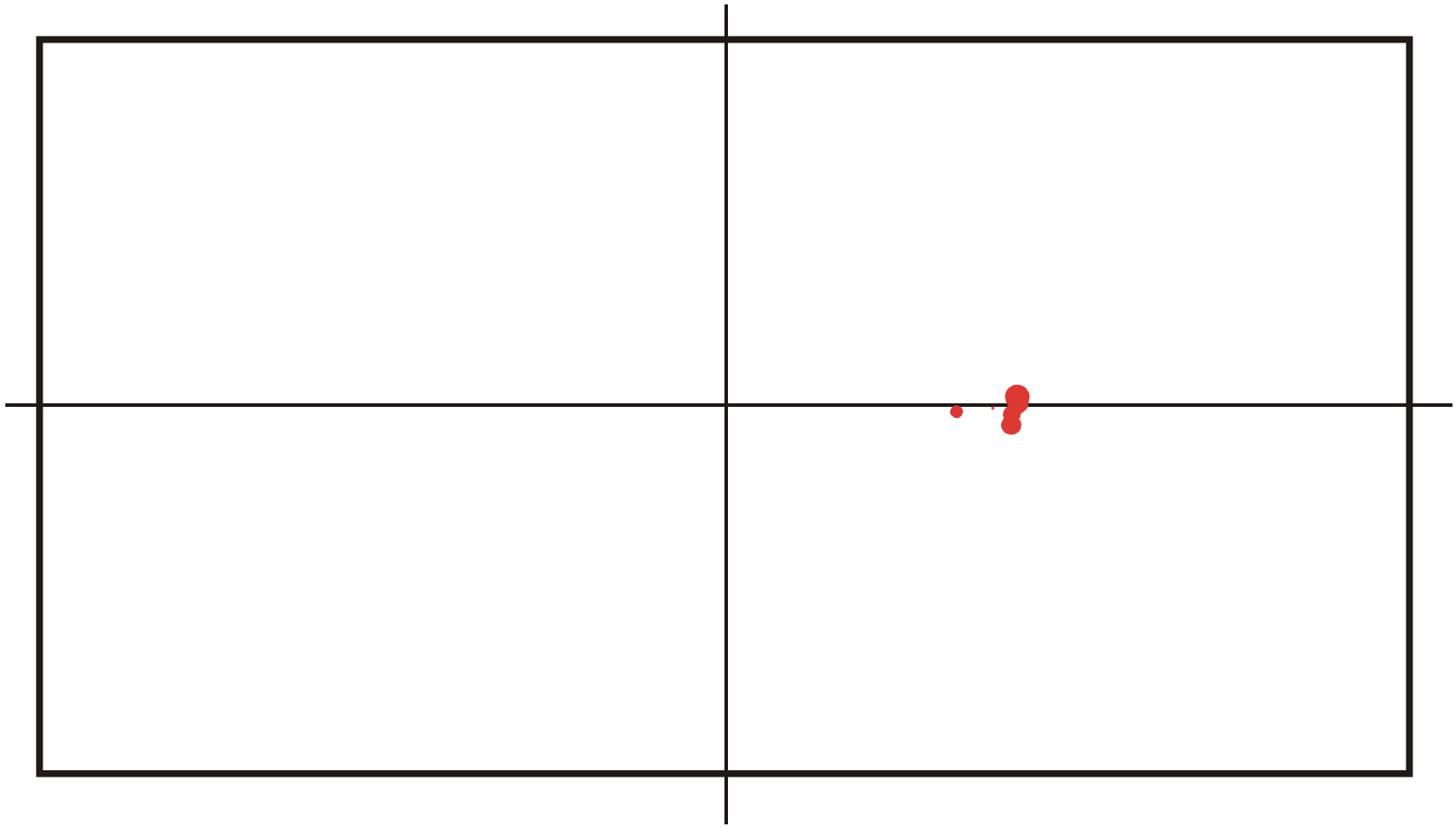
Spread of population in sequence space during a quasistationary epoch: t = 835



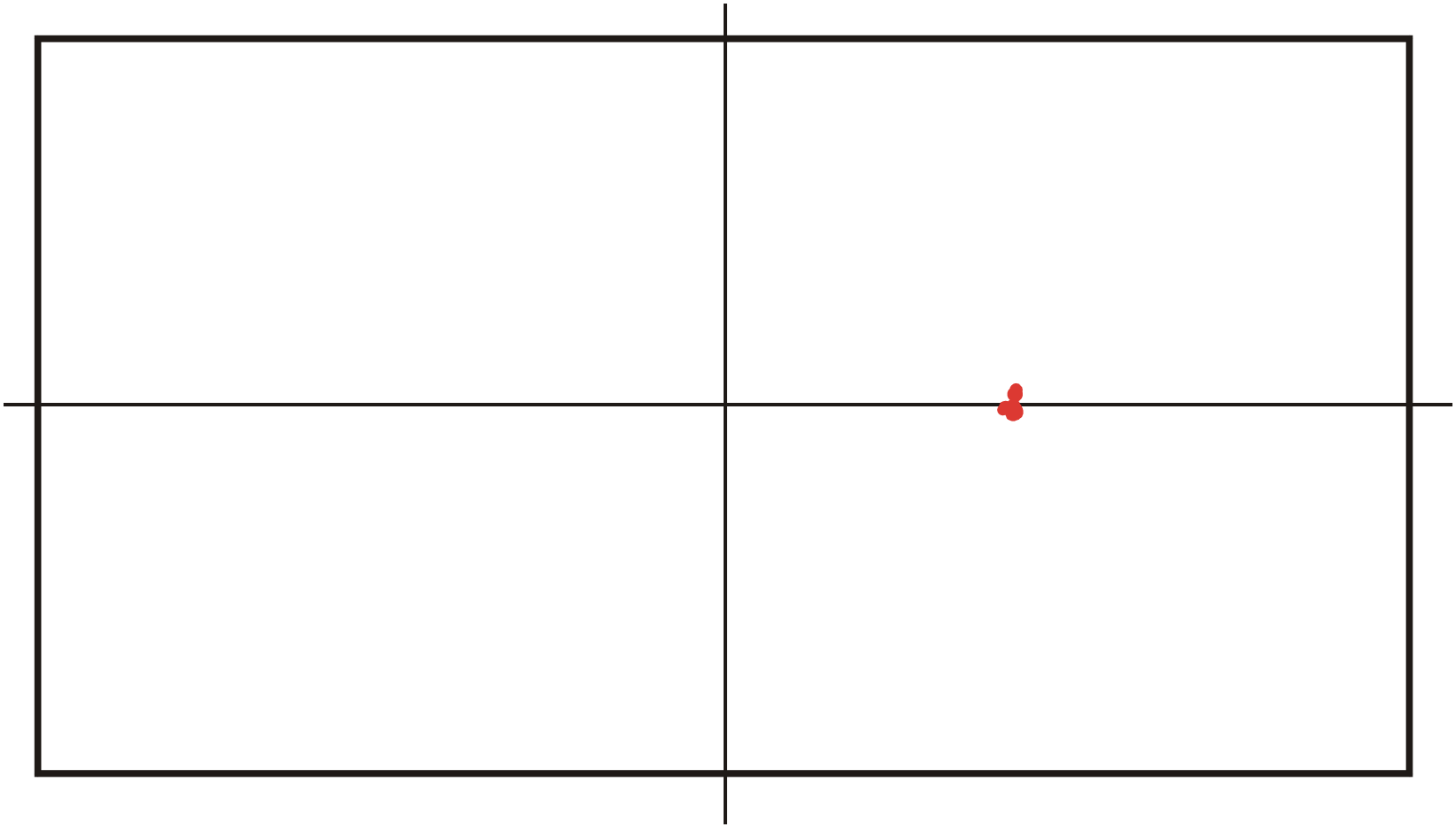
Spread of population in sequence space during a quasistationary epoch: t = 840



Spread of population in sequence space during a quasistationary epoch: t = 845



Spread of population in sequence space during a quasistationary epoch: t = 850



Spread of population in sequence space during a quasistationary epoch: t = 855

1. Solutions to Levinthal‘s paradox
2. Selection as solution to low probabilities in evolution
3. Origin of information by mutation and selection
4. Evolution of RNA phenotypes
5. The role of neutrality in molecular evolution
6. **An experiment with RNA molecules**
7. Summary

A ribozyme switch

E.A.Schultes, D.B.Bartel, Science **289** (2000), 448-452

minus the background levels observed in the HSP in the control (Sar1-GDP–containing) incubation that prevents COPII vesicle formation. In the microsome control, the level of p115-SNARE associations was less than 0.1%.
46. C. M. Carr, E. Grote, M. Munson, F. M. Hughson, P. J. Novick, *J. Cell Biol.* **146**, 333 (1999).
47. C. Ungermann, B. J. Nichols, H. R. Pelham, W. Wickner, *J. Cell Biol.* **140**, 61 (1998).
48. E. Grote and P. J. Novick, *Mol. Biol. Cell* **10**, 4149 (1999).
49. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
50. GST-SNARE proteins were expressed in bacteria and purified on glutathione-Sepharose beads using standard methods. Immobilized GST-SNARE protein (0.5 μM) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μM) in 1 ml of NS buffer containing 1% BSA for 2 hours at 4°C with rotation. Beads were briefly spun (3000 rpm for 10 s) and sequentially washed three times with NS buffer and three times with NS buffer supplemented with 150 mM NaCl. Bound proteins were eluted three times in 50 μl of 50 mM tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton X-100 for 15 min at 4°C with intermittent mixing, and elutes were pooled. Proteins were precipitated by MeOH/CH_3Cl and separated by SDS–polyacrylamide gel electrophoresis (PAGE) followed by immunoblotting using p115 mAb 13F12.
51. V. Rybin *et al.*, *Nature* **383**, 266 (1996).
52. K. G. Hardwick and H. R. Pelham, *J. Cell Biol.* **119**, 513 (1992).
53. A. P. Newman, M. E. Groesch, S. Ferro-Novick, *EMBO J.* **11**, 3609 (1992).
54. A. Spang and R. Schekman, *J. Cell Biol.* **143**, 589 (1998).
55. M. F. Rexach, M. Latterich, R. W. Schekman, *J. Cell Biol.* **126**, 1133 (1994).
56. A. Mayer and W. Wickner, *J. Cell Biol.* **136**, 307 (1997).
57. M. D. Turner, H. Plutner, W. E. Balch, *J. Biol. Chem.* **272**, 13479 (1997).
58. A. Price, D. Seals, W. Wickner, C. Ungermann, *J. Cell Biol.* **148**, 1231 (2000).
59. X. Cao and C. Barlowe, *J. Cell Biol.* **149**, 55 (2000).
60. G. G. Tall, H. Hama, D. B. DeWald, B. F. Horazdovsky, *Mol. Biol. Cell* **10**, 1873 (1999).
61. C. G. Burd, M. Peterson, C. R. Cowles, S. D. Emr, *Mol. Biol. Cell* **8**, 1089 (1997).
62. M. R. Peterson, C. G. Burd, S. D. Emr, *Curr. Biol.* **9**, 159 (1999).
63. M. G. Waters, D. O. Clary, J. E. Rothman, *J. Cell Biol.* **118**, 1015 (1992).
64. D. M. Walter, K. S. Paul, M. G. Waters, *J. Biol. Chem.* **273**, 29565 (1998).
65. N. Hui *et al.*, *Mol. Biol. Cell* **8**, 1777 (1997).
66. T. E. Kreis, *EMBO J.* **5**, 931 (1986).
67. H. Plutner, H. W. Davidson, J. Saraste, W. E. Balch, *J. Cell Biol.* **119**, 1097 (1992).
68. D. S. Nelson *et al.*, *J. Cell Biol.* **143**, 319 (1998).
69. We thank G. Waters for p115 cDNA and p115 mAbs; G. Warren for p97 and p47 antibodies; R. Scheller for rbet1, membrin, and sec22 cDNAs; H. Plutner for excellent technical assistance; and P. Tan for help during the initial phase of this work. Supported by NIH grants GM 33301 and GM42336 and National Cancer Institute grant CA58689 (W.E.B.), a NIH National Research Service Award (B.D.M.), and a Wellcome Trust International Traveling Fellowship (B.B.A.).

20 March 2000; accepted 22 May 2000

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes and David P. Bartel*

We describe a single RNA sequence that can assume either of two ribozyme folds and catalyze the two respective reactions. The two ribozyme folds share no evolutionary history and are completely different, with no base pairs (and probably no hydrogen bonds) in common. Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototype ribozymes through a series of neutral mutations. Thus, in the course of evolution, new RNA folds could arise from preexisting folds, without the need to carry inactive intermediate sequences. This raises the possibility that biological RNAs having no structural or functional similarity might share a common ancestry. Furthermore, functional and structural divergence might, in some cases, precede rather than follow gene duplication.

Related protein or RNA sequences with the same folded conformation can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to sequences with new macromolecular folds? When considering the origin of new folds, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can range very far in sequence space (*1*). For example, only seven nucleotides are strictly conserved among the group I self-splicing introns, yet secondary (and presumably tertiary) structure within the core of the ribozyme is preserved (*2*). Because these disparate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of mutational variants that were each functional. Hence, sequence heterogeneity among divergent isolates implies the existence of paths through sequence space that have allowed neutral drift from the ancestral sequence to each isolate. The set of all possible neutral paths composes a "neutral network," connecting in sequence space those widely dispersed sequences sharing a particular fold and activity, such that any sequence on the network can potentially access very distant sequences by neutral mutations (*3–5*).

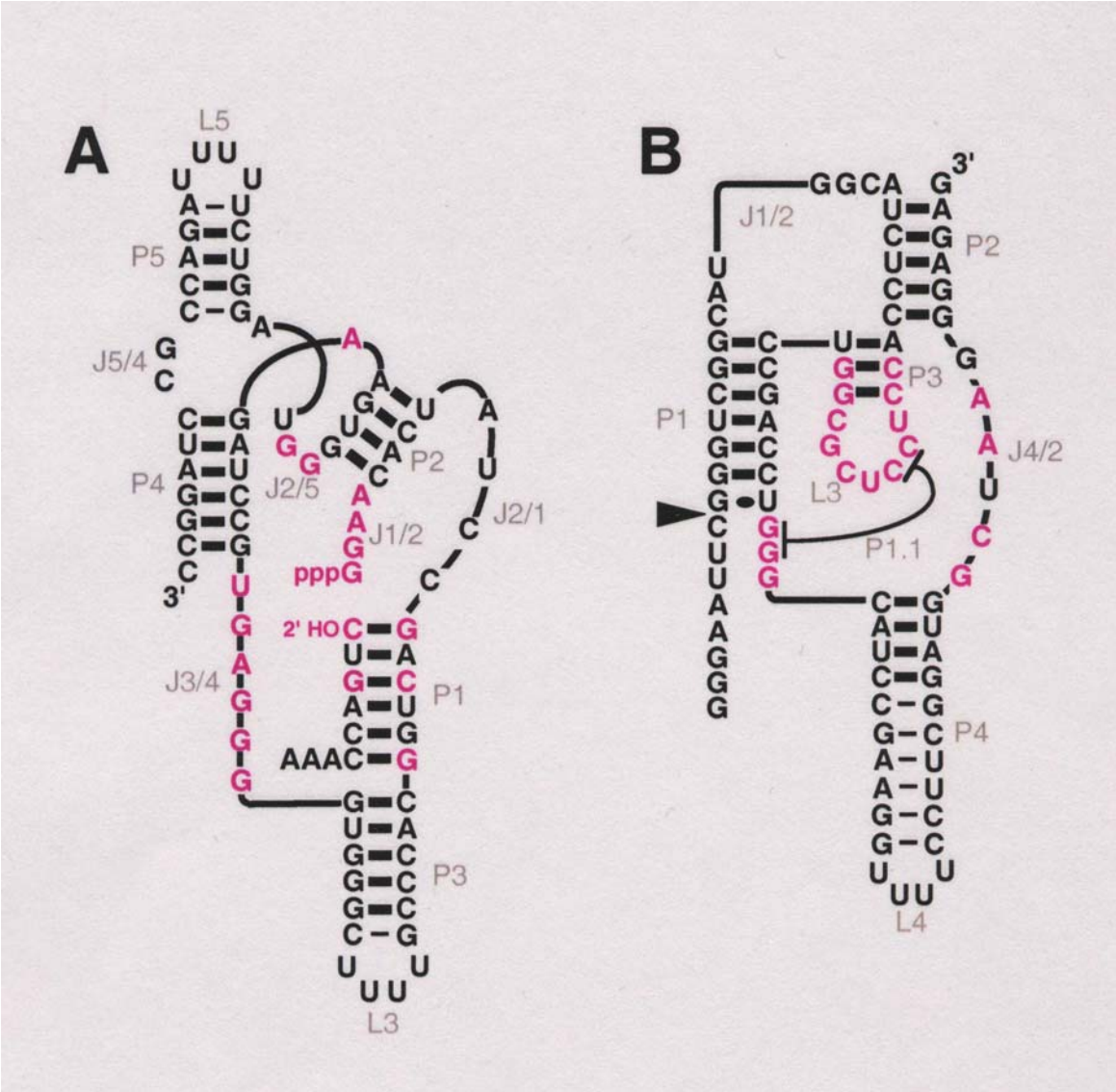
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. If so, a single sequence would be capable of folding into two different conformations, would have two different catalytic activities, and could access by neutral drift every sequence on both networks. With intersecting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry nonfunctional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

In choosing the two ribozymes for this investigation, an important criterion was that they share no evolutionary history that might confound the evolutionary interpretations of our results. Choosing at least one artificial ribozyme ensured independent evolutionary histories. The class III ligase is a synthetic ribozyme isolated previously from a pool of random RNA sequences (*9*). It joins an oligonucleotide substrate to its 5′ terminus. The prototype ligase sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 10 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isolate (*10*). The HDV ribozyme carries out the site-specific self-cleavage reactions needed during the life cycle of HDV, a satellite virus of hepatitis B with a circular, single-stranded RNA genome (*11*). The prototype HDV construct for our study (Fig. 1B) is a shortened version of the antigenomic HDV ribozyme (*12*), which undergoes self-cleavage at a rate similar to that reported for other antigenomic constructs (*13*, *14*).

The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

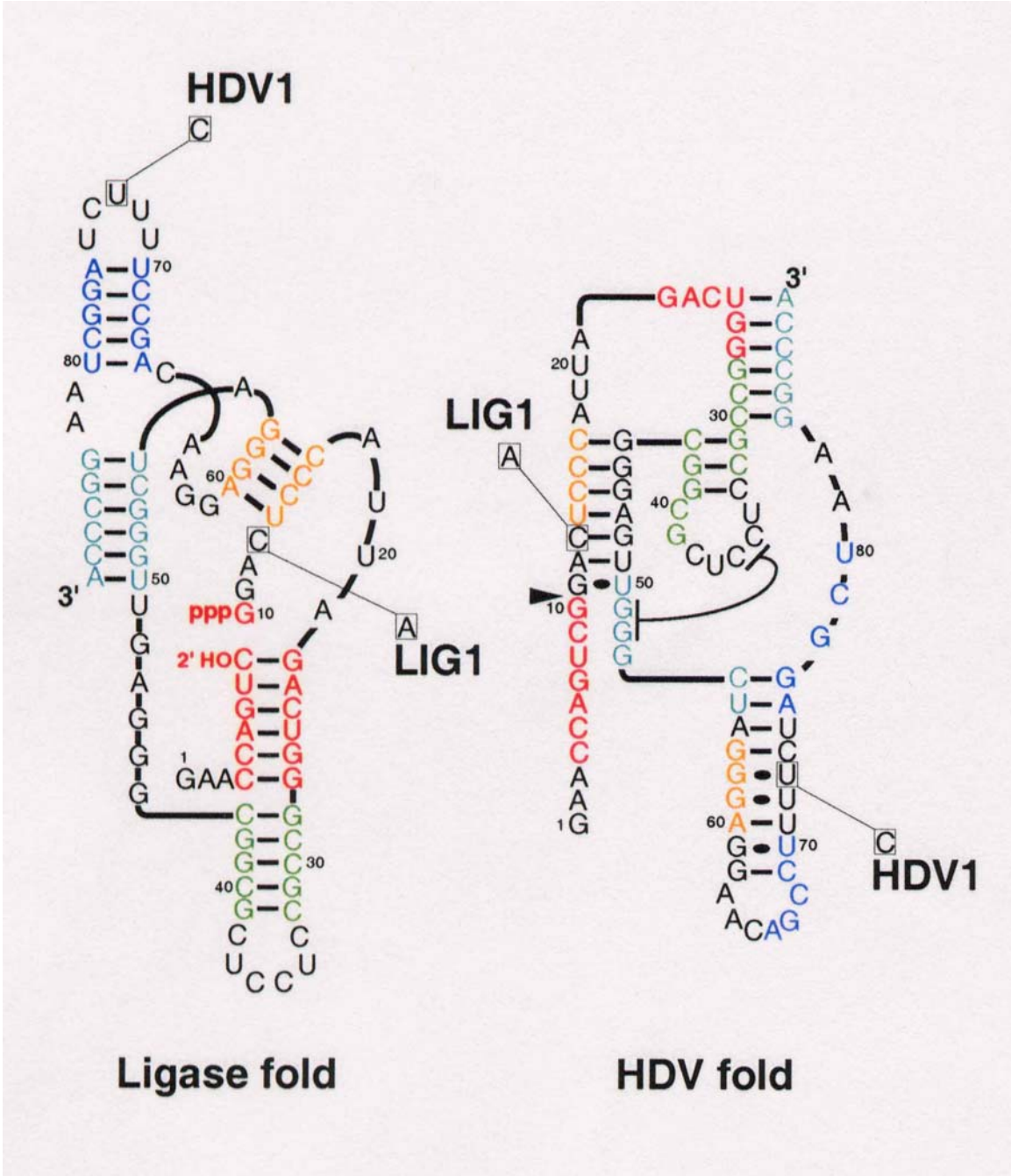
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

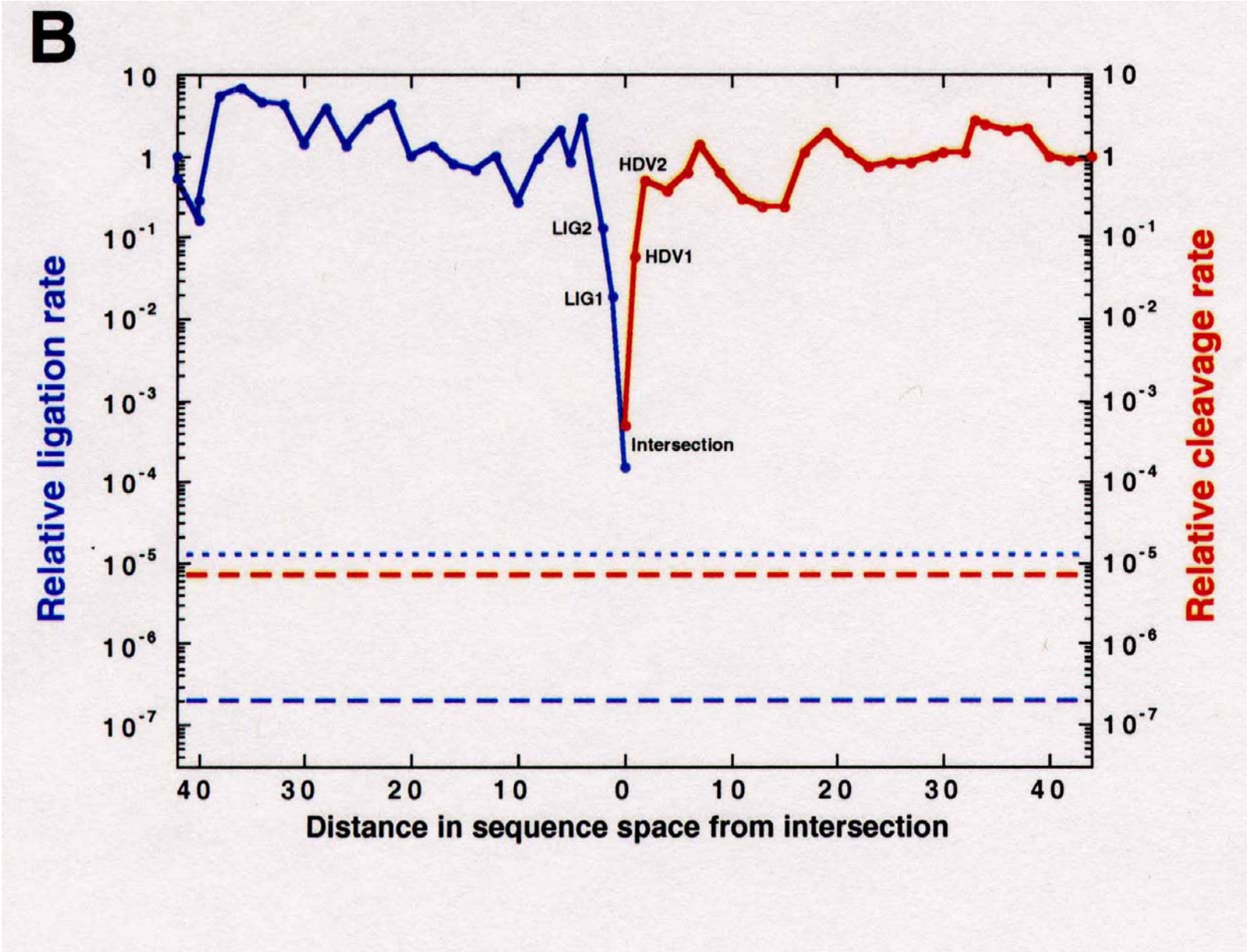


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

The sequence at the ***intersection***:

An RNA molecules which is 88 nucleotides long and can form both structures





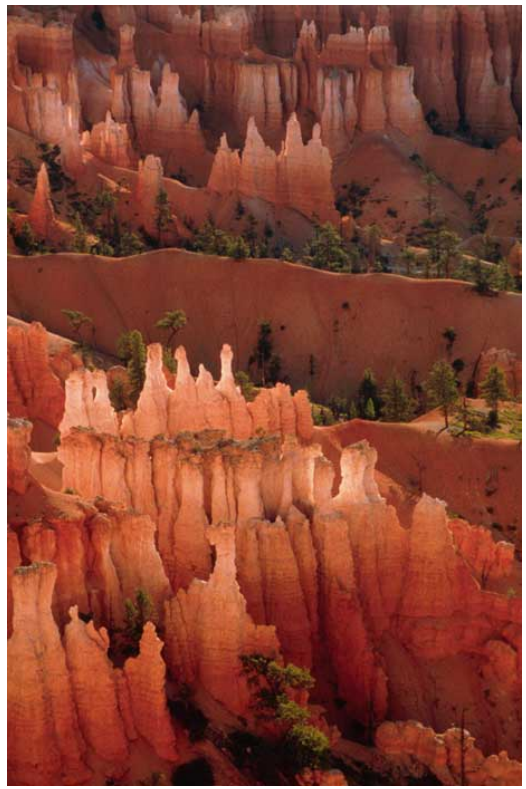
Two neutral walks through sequence space with conservation of structure and catalytic activity

1. Solutions to Levinthal‘s paradox
2. Selection as solution to low probabilities in evolution
3. Origin of information by mutation and selection
4. Evolution of RNA phenotypes
5. The role of neutrality in molecular evolution
6. An experiment with RNA molecules
7. **Summary**

Mount Fuji

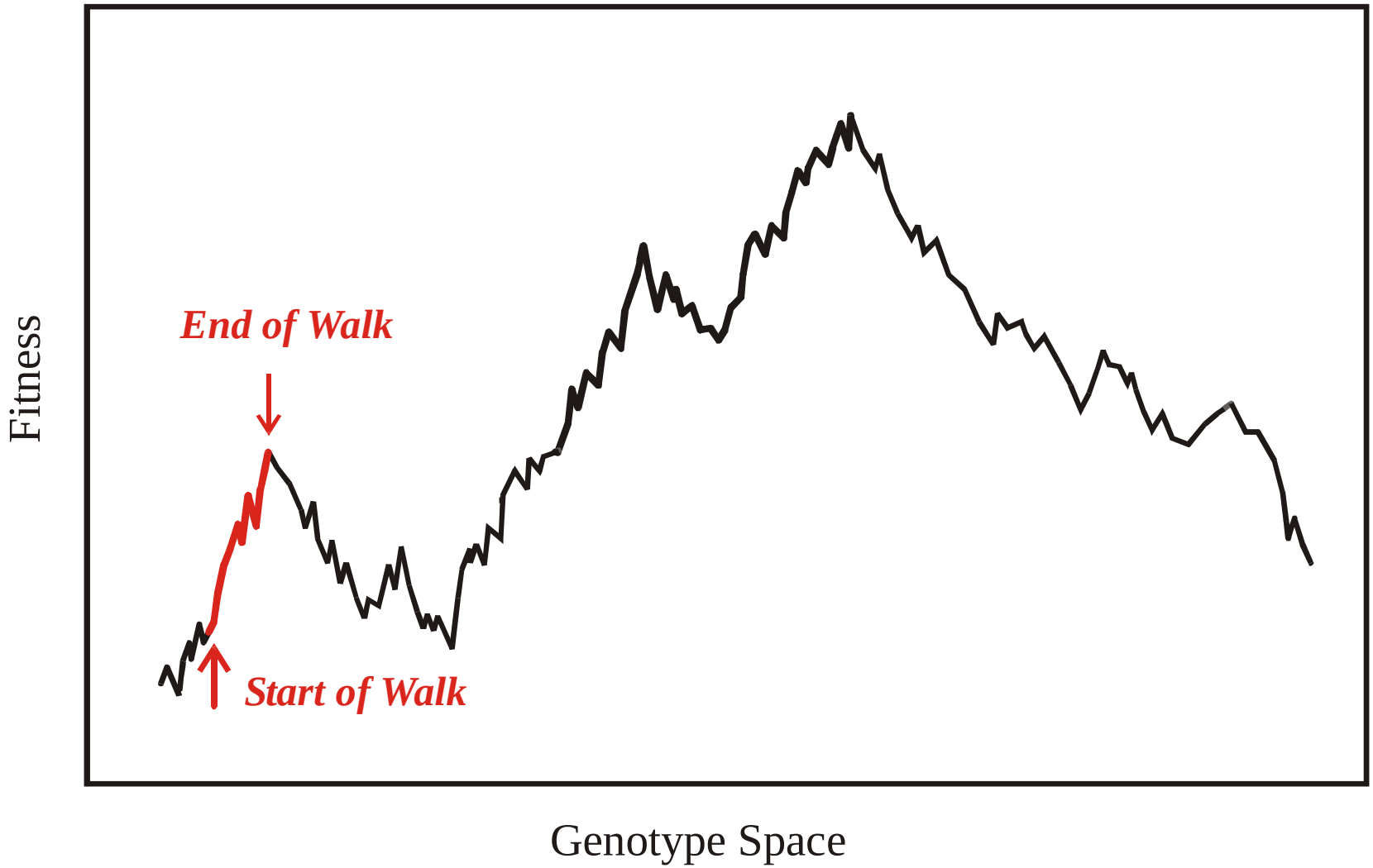
Example of a smooth landscape on Earth

Dolomites

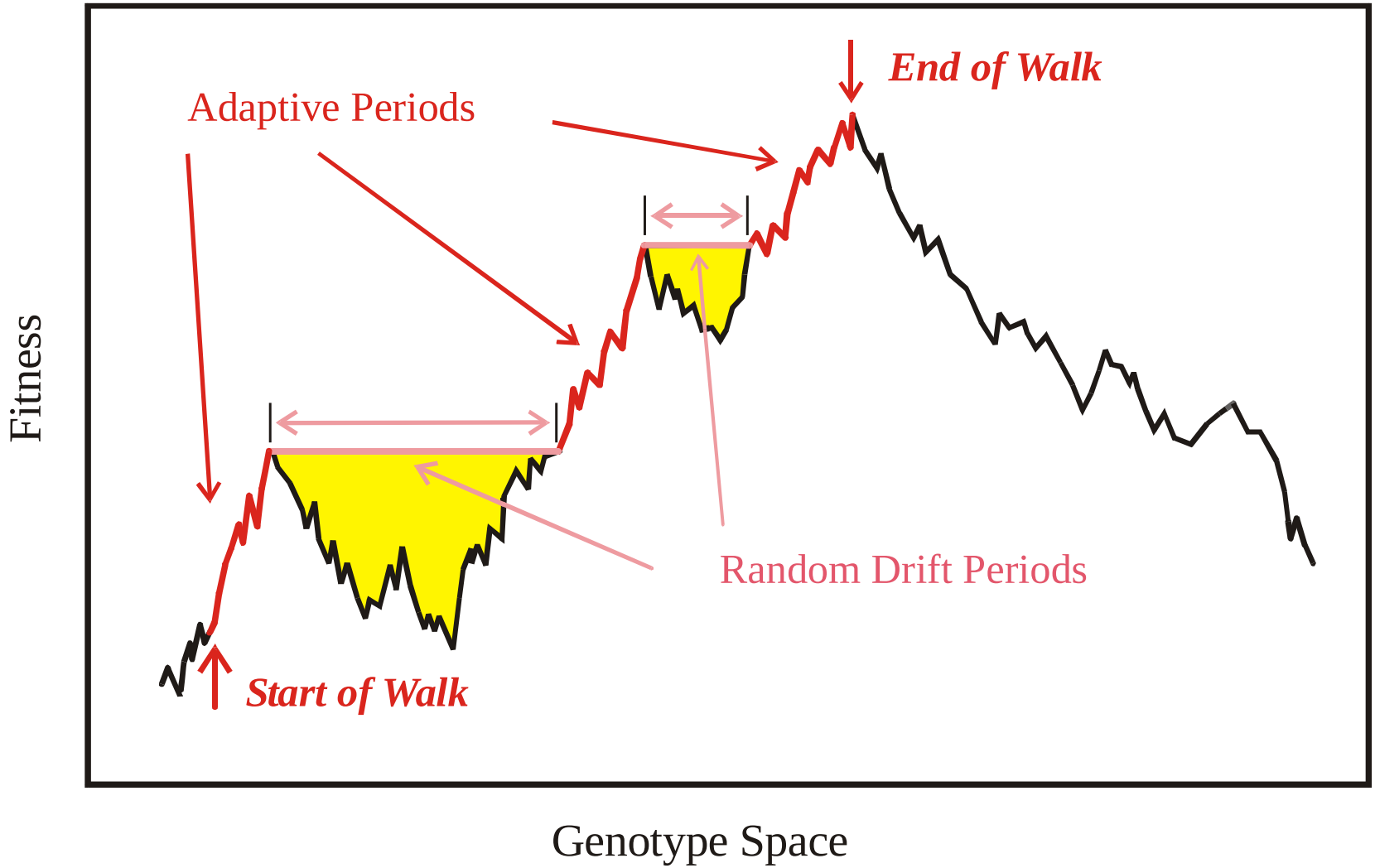


Bryce Canyon

Examples of rugged landscapes on Earth



Evolutionary optimization in absence of neutral paths in sequence space

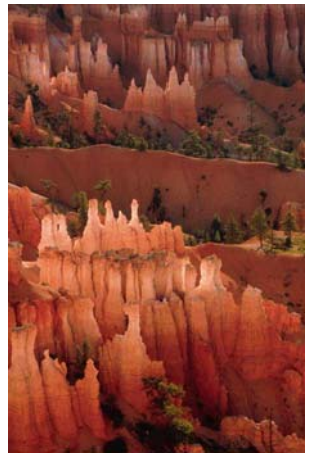


Evolutionary optimization including neutral paths in sequence space

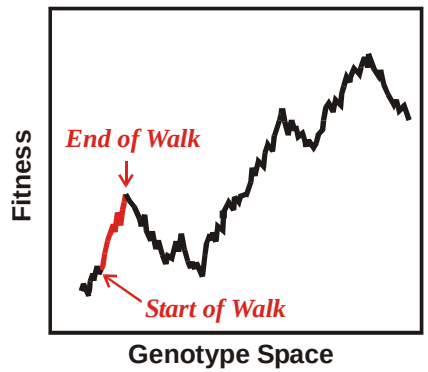
Grand Canyon

Example of a landscape on Earth with ‘neutral’ ridges and plateaus

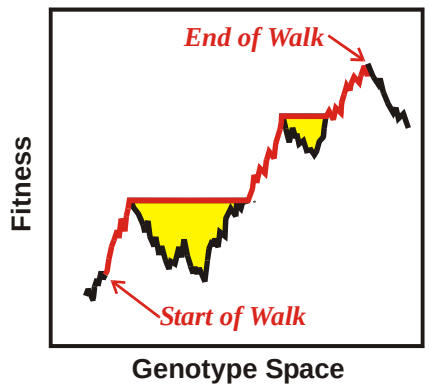
- Conformational and mutational landscapes of biomolecules as well as fitness landscapes of evolutionary biology are **rugged**.



- **Adaptive** or **non-descending walks** on rugged landscapes end commonly at one of the low lying local maxima.



- Selective neutrality in the form of **neutral networks** plays an active role in evolutionary optimization and enables populations to reach high local maxima or even the global optimum.



Acknowledgement of support

Universität Wien

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)

Projects No. 09942, 10578, 11065, 13093
13887, and 14898

Jubiläumsfonds der Österreichischen Nationalbank

Project No. Nat-7813

European Commission: Project No. EU-980189

Austrian Genome Research Program – GEN-AU

Siemens AG, Austria

Universität Wien and the Santa Fe Institute

Coworkers

Universität Wien

Walter Fontana, Harvard Medical School, MA

Christian Reidys, **Christian Forst**, Los Alamos National Laboratory, NM

Peter Stadler, **Bärbel Stadler**, Universität Leipzig, GE

Jord Nagel, **Kees Pleij**, Universiteit Leiden, NL

Ivo L.Hofacker, **Christoph Flamm**, Universität Wien, AT

Andreas Wernitznig, **Michael Kospach**, Universität Wien, AT
Ulrike Langhammer, **Ulrike Mückstein**, **Stefanie Widder**
Jan Cupal, **Kurt Grünberger**, **Andreas Svrcek-Seiler**, **Stefan Wuchty**
Stefan Bernhart, **Lukas Endler**

Ulrike Göbel, Institut für Molekulare Biotechnologie, Jena, GE
Walter Grüner, **Stefan Kopp**, **Jaqueline Weber**

Web-Page for further information:

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