RNA – A Model for Molecular Evolution

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

GDCh-Jahrestagung 2003

Fachgruppe Biochemie

München, 09.10.2003

Web-Page for further information:

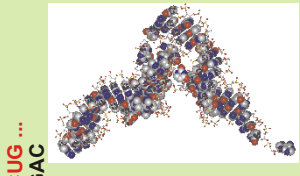
http://www.tbi.univie.ac.at/~pks

RNA *as transmitter of genetic information*

DNA
↓ transcription
...AGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUC...
messenger-**RNA**
↓ translation
protein

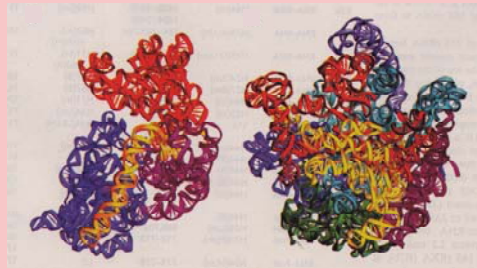
RNA as **working copy** of genetic information

RNA *as adapter molecule*



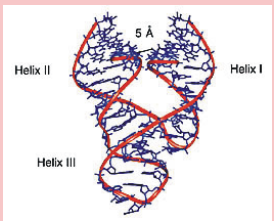
genetic code

RNA *is the catalytic subunit in supramolecular complexes*



The ribosome is a ribozyme !

RNA *as catalyst*



ribozyme

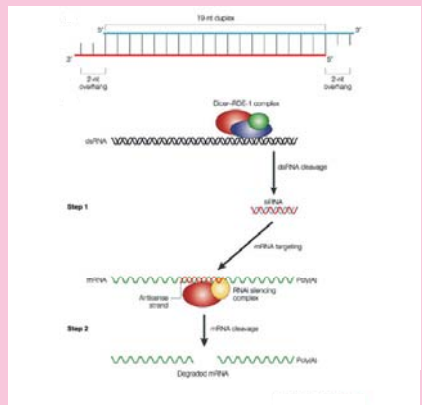
RNA

RNA *is modified by epigenetic control*

RNA editing

Alternative splicing of messenger **RNA**

RNA *as regulator of gene expression*



gene silencing by small interfering RNAs

The RNA *world as a precursor of the current* DNA + protein *biology*

RNA *as carrier of genetic information*

RNA viruses and retroviruses

RNA as information carrier in evolution *in vitro* and evolutionary biotechnology

Functions of RNA molecules

1. **Experiments on controlled evolution and RNA replication**

2. **Sequence-structure maps, neutral networks, and intersections**

3. **Optimization in the RNA model**

4. **What we can learn from molecules for evolution proper**

1. **Experiments on controlled evolution and RNA replication**

2. Sequence-structure maps, neutral networks, and intersections

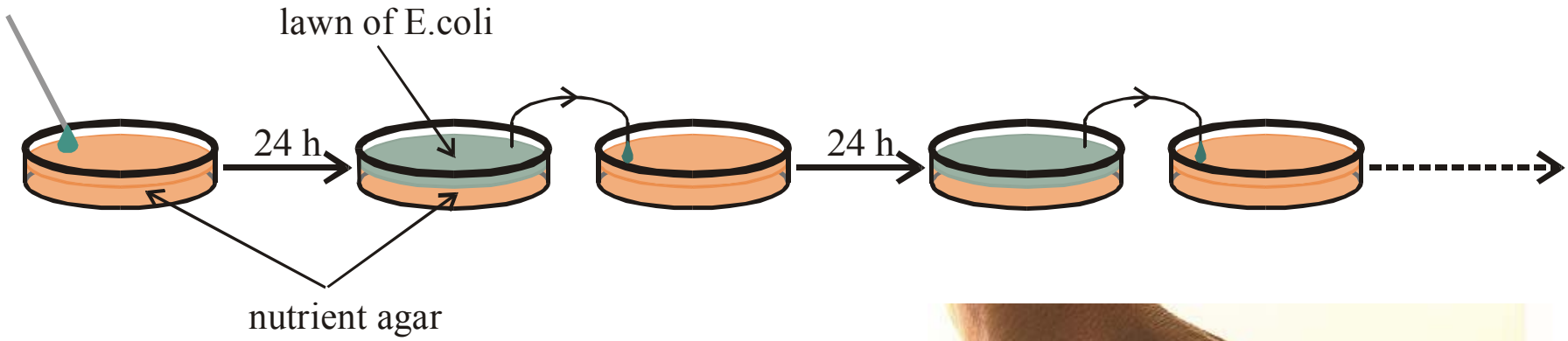
3. Optimization in the RNA model

4. What we can learn from molecules for evolution proper

Bacterial Evolution

S. F. Elena, V. S. Cooper, R. E. Lenski. ***Punctuated evolution caused by selection of rare beneficial mutants***. Science **272** (1996), 1802-1804

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. ***Genomic evolution during a 10,000-generation experiment with bacteria***. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812



Serial transfer of Escherichia coli cultures in Petri dishes

1 day Î 6.67 generations
1 month Î 200 generations
1 year Î 2400 generations

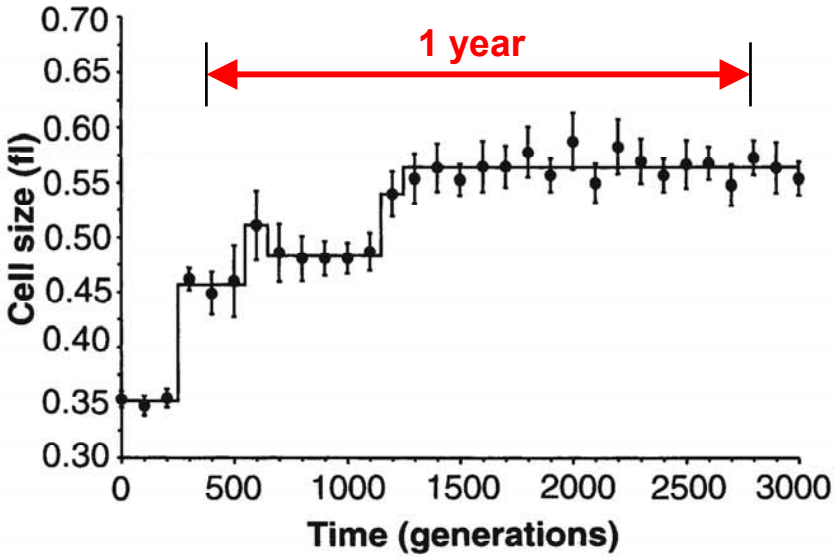


Fig. 1. Change in average cell size (1 fl = 10^{-15} L) in a population of *E. coli* during 3000 generations of experimental evolution. Each point is the mean of 10 replicate assays (*22*). Error bars indicate 95% confidence intervals. The solid line shows the best fit of a step-function model to these data (Table 1).

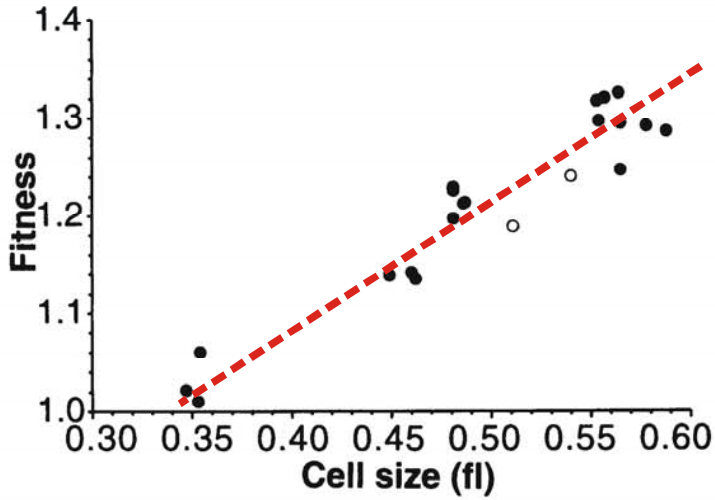
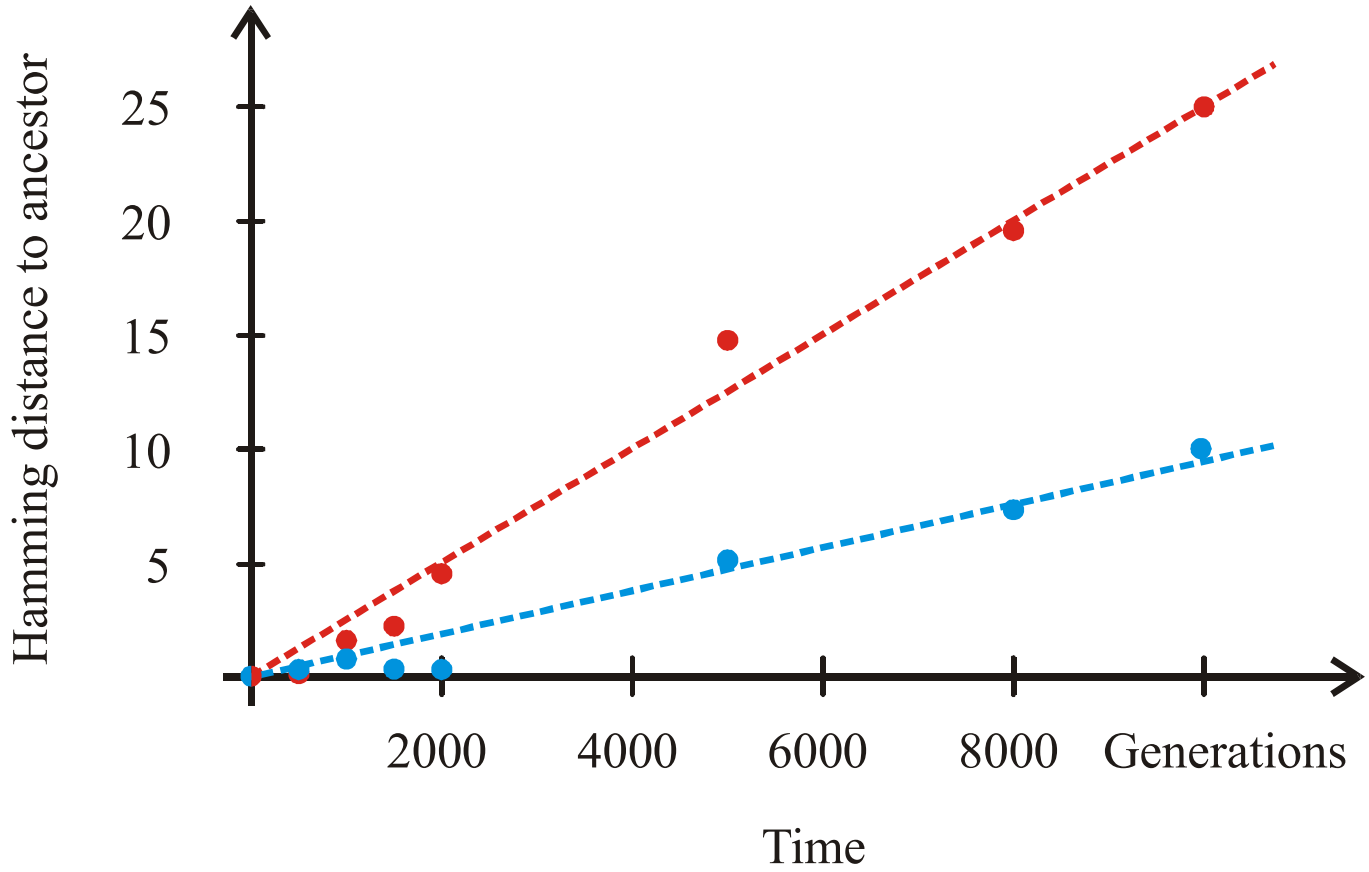


Fig. 2. Correlation between average cell size and mean fitness, each measured at 100-generation intervals for 2000 generations. Fitness is expressed relative to the ancestral genotype and was obtained from competition experiments between derived and ancestral cells (*6, 7*). The open symbols indicate the only two samples assigned to different steps by the cell size and fitness data.

Epochal evolution of bacteria in serial transfer experiments under constant conditions

S. F. Elena, V. S. Cooper, R. E. Lenski. *Punctuated evolution caused by selection of rare beneficial mutants*. Science **272** (1996), 1802-1804



Variation of genotypes in a bacterial serial transfer experiment

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. *Genomic evolution during a 10,000-generation experiment with bacteria*. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812

Evolution of RNA molecules based on Qβ phage

D.R.Mills, R.L.Peterson, S.Spiegelman, ***An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule***. Proc.Natl.Acad.Sci.USA **58** (1967), 217-224

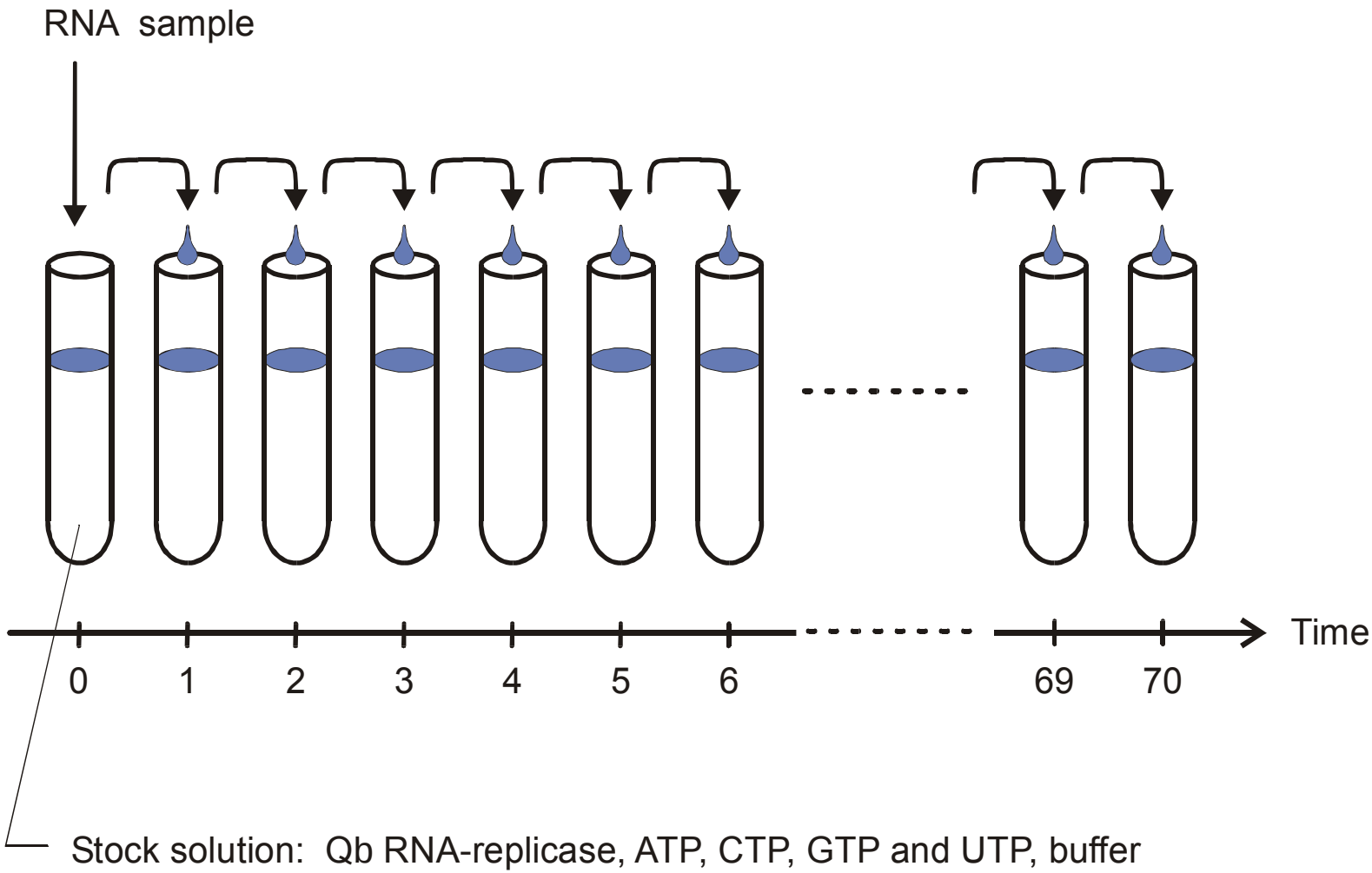
S.Spiegelman, ***An approach to the experimental analysis of precellular evolution***. Quart.Rev.Biophys. **4** (1971), 213-253

C.K.Biebricher, ***Darwinian selection of self-replicating RNA molecules***. Evolutionary Biology **16** (1983), 1-52

G.Bauer, H.Otten, J.S.McCaskill, ***Travelling waves of*** **in vitro** ***evolving RNA.*** *Proc.Natl.Acad.Sci.USA* **86** (1989), 7937-7941

C.K.Biebricher, W.C.Gardiner, ***Molecular evolution of RNA*** **in vitro**. Biophysical Chemistry **66** (1997), 179-192

G.Strunk, T.Ederhof, ***Machines for automated evolution experiments*** **in vitro** ***based on the serial transfer concept***. Biophysical Chemistry **66** (1997), 193-202



The serial transfer technique applied to RNA evolution ***in vitro***

Reproduction of the original figure of the serial transfer experiment with Qβ RNA

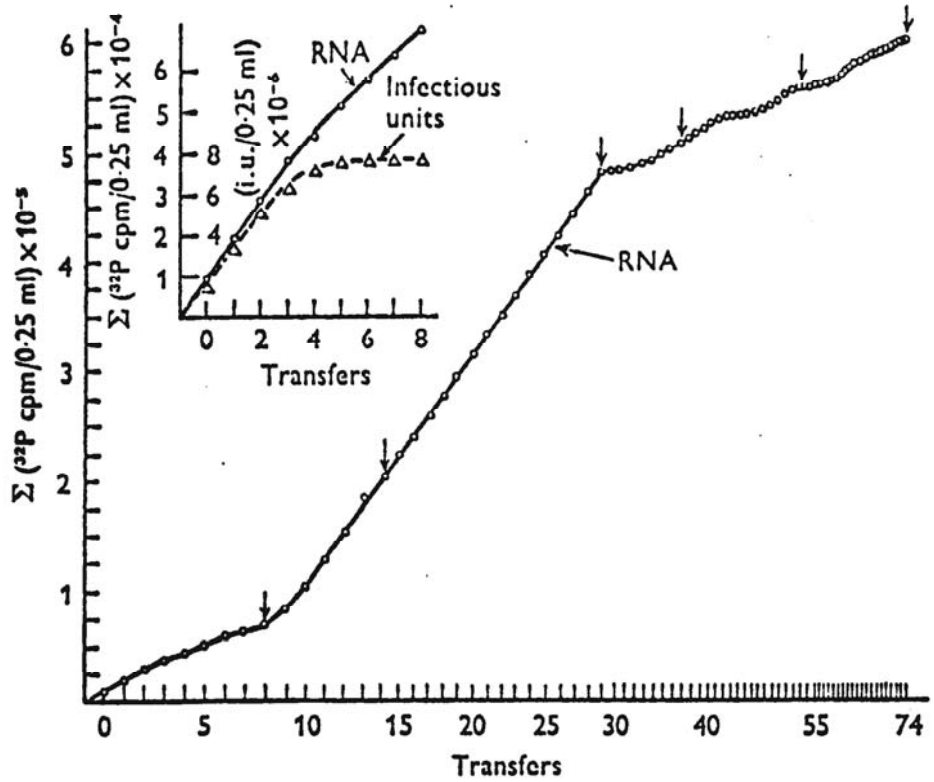
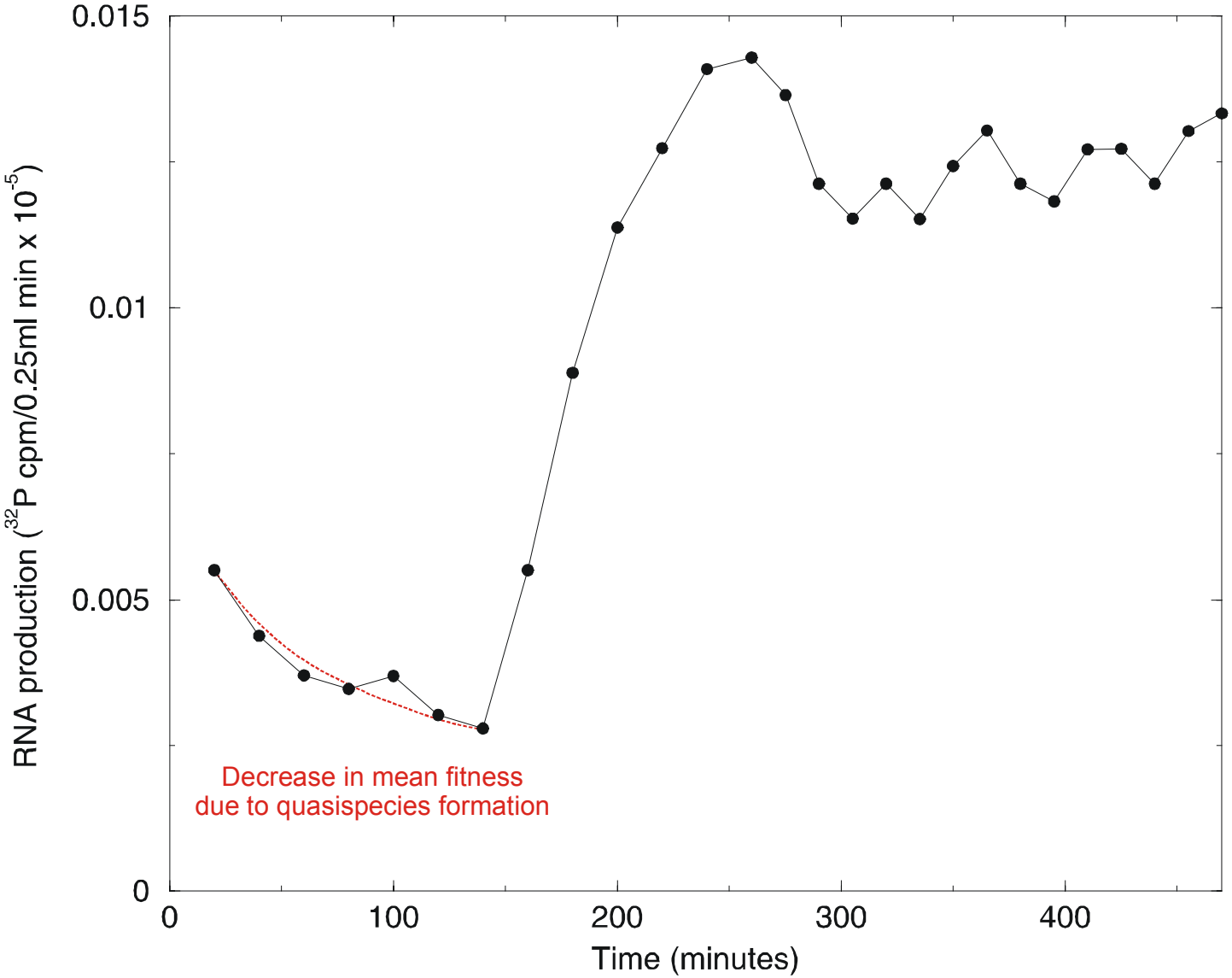


Fig. 9. Serial transfer experiment. Each 0·25 ml standard reaction mixture contained 40 μg of Qβ replicase and ^{32}P-UTP. The first reaction (0 transfer) was initiated by the addition of 0·2 μg ts-1 (temperature-sensitive RNA) and incubated at 35 °C for 20 min, whereupon 0·02 ml was drawn for counting and 0·02 ml was used to prime the second reaction (first transfer), and so on. After the first 13 reactions, the incubation periods were reduced to 15 min (transfers 14–29). Transfers 30–38 were incubated for 10 min. Transfers 39–52 were incubated for 7 min, and transfers 53–74 were incubated for 5 min. The arrows above certain transfers (0, 8, 14, 29, 37, 53, and 73) indicate where 0·001–0·1 ml of product was removed and used to prime reactions for sedimentation analysis on sucrose. The inset examines both infectious and total RNA. The results show that biologically competent RNA ceases to appear after the 4th transfer (Mills *et al.* 1967).

D.R.Mills, R,L,Peterson, S.Spiegelman, *An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule*. Proc.Natl.Acad.Sci.USA **58** (1967), 217-224



The increase in RNA production rate during a serial transfer experiment

No new principle will declare itself from below a heap of facts.

Sir Peter Medawar, 1985

Theory of molecular evolution

M.Eigen, ***Self-organization of matter and the evolution of biological macromolecules***. Naturwissenschaften **58** (1971), 465-526

C.J.Thompson, J.L.McBride, ***On Eigen's theory of the self-organization of matter and the evolution of biological macromolecules***. Math. Biosci. **21** (1974), 127-142

B.L.Jones, R.H.Enns, S.S.Rangnekar, ***On the theory of selection of coupled macromolecular systems.*** Bull.Math.Biol. **38** (1976), 15-28

M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle***. Naturwissenschaften **58** (1977), 465-526

M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part B: The abstract hypercycle***. Naturwissenschaften **65** (1978), 7-41

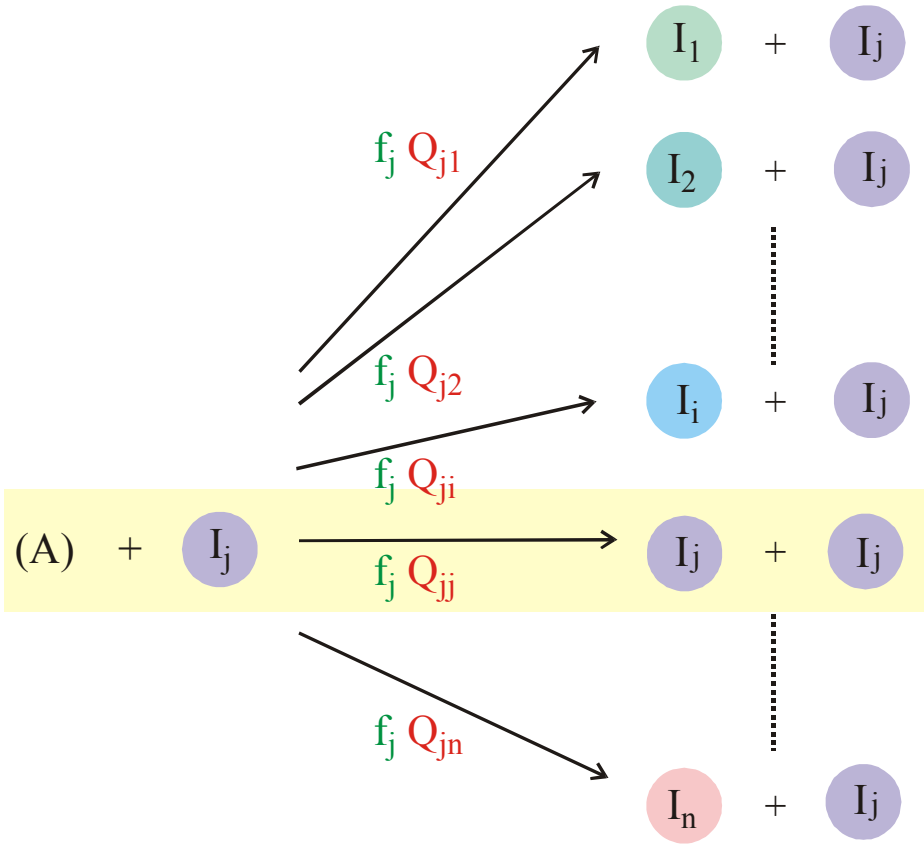
M.Eigen, P.Schuster, ***The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle***. Naturwissenschaften **65** (1978), 341-369

J.Swetina, P.Schuster, ***Self-replication with errors - A model for polynucleotide replication.*** Biophys.Chem. **16** (1982), 329-345

J.S.McCaskill, ***A localization threshold for macromolecular quasispecies from continuously distributed replication rates***. J.Chem.Phys. **80** (1984), 5194-5202

M.Eigen, J.McCaskill, P.Schuster, ***The molecular quasispecies***. Adv.Chem.Phys. **75** (1989), 149-263

C. Reidys, C.Forst, P.Schuster, ***Replication and mutation on neutral networks***. Bull.Math.Biol. **63** (2001), 57-94



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

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

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

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

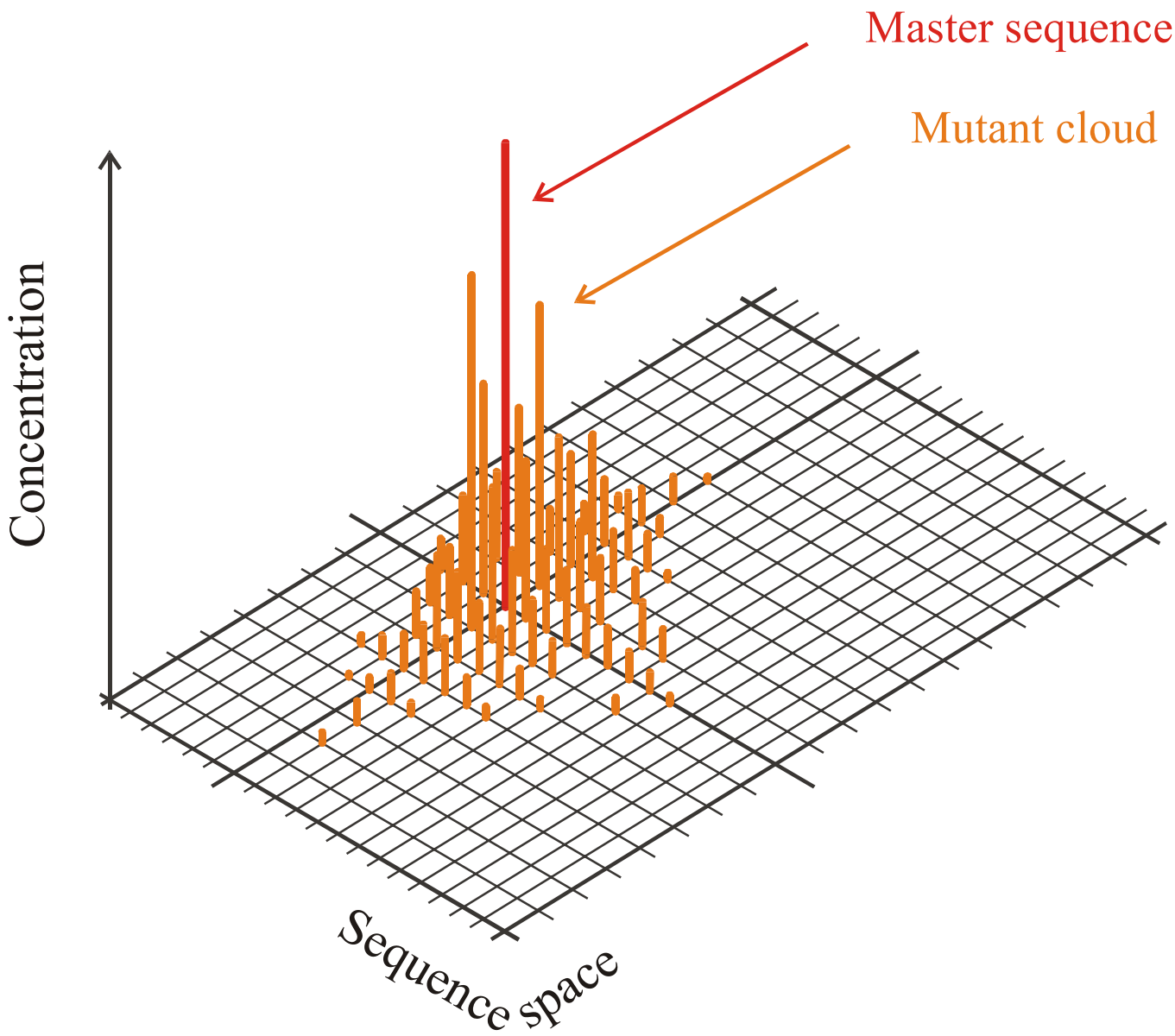
$$Q_{ij} = (1-p)^{\ell - d(i,j)} p^{d(i,j)}$$

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



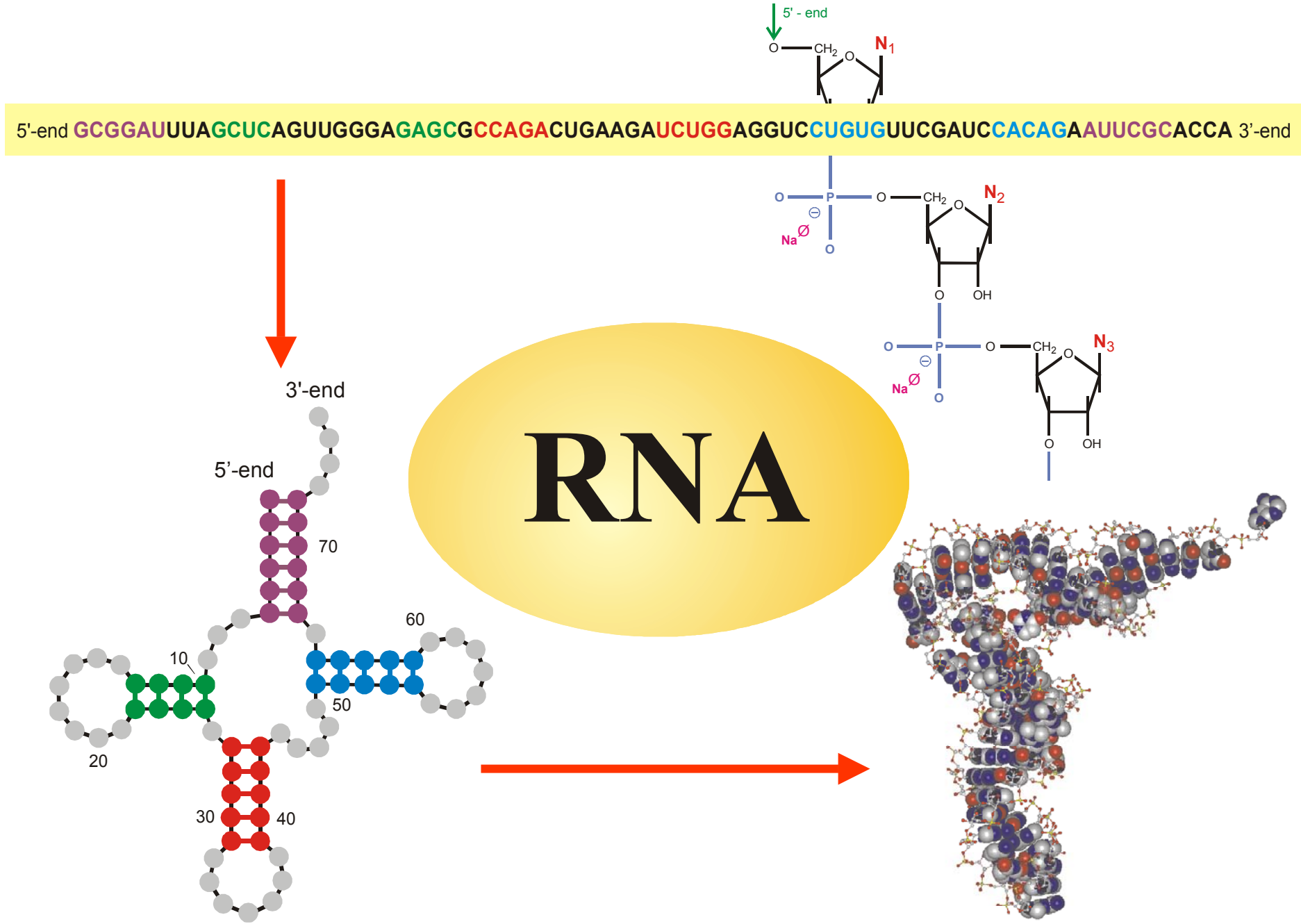
The molecular quasispecies in sequence space

1. Experiments on controlled evolution and RNA replication

2. Sequence-structure maps, neutral networks, and intersections

3. Optimization in the RNA model

4. What we can learn from molecules for evolution proper



Definition of RNA structure

How to compute RNA secondary structures

Efficient algorithms based on **dynamic programming** are available for computation of minimum free energy and many suboptimal secondary structures for given sequences.

M.Zuker and P.Stiegler. *Nucleic Acids Res*. **9**:133-148 (1981)

M.Zuker, *Science* **244**: 48-52 (1989)

Equilibrium partition function and base pairing probabilities in Boltzmann ensembles of suboptimal structures.

J.S.McCaskill. *Biopolymers* **29**:1105-1190 (1990)

The **Vienna RNA Package** provides in addition: inverse folding (computing sequences for given secondary structures), computation of melting profiles from partition functions, all suboptimal structures within a given energy interval, barrier tress of suboptimal structures, kinetic folding of RNA sequences, RNA-hybridization and RNA/DNA-hybridization through cofolding of sequences, alignment, etc..

I.L.Hofacker, W. Fontana, P.F.Stadler, L.S.Bonhoeffer, M.Tacker, and P. Schuster. *Mh.Chem*. **125**:167-188 (1994)

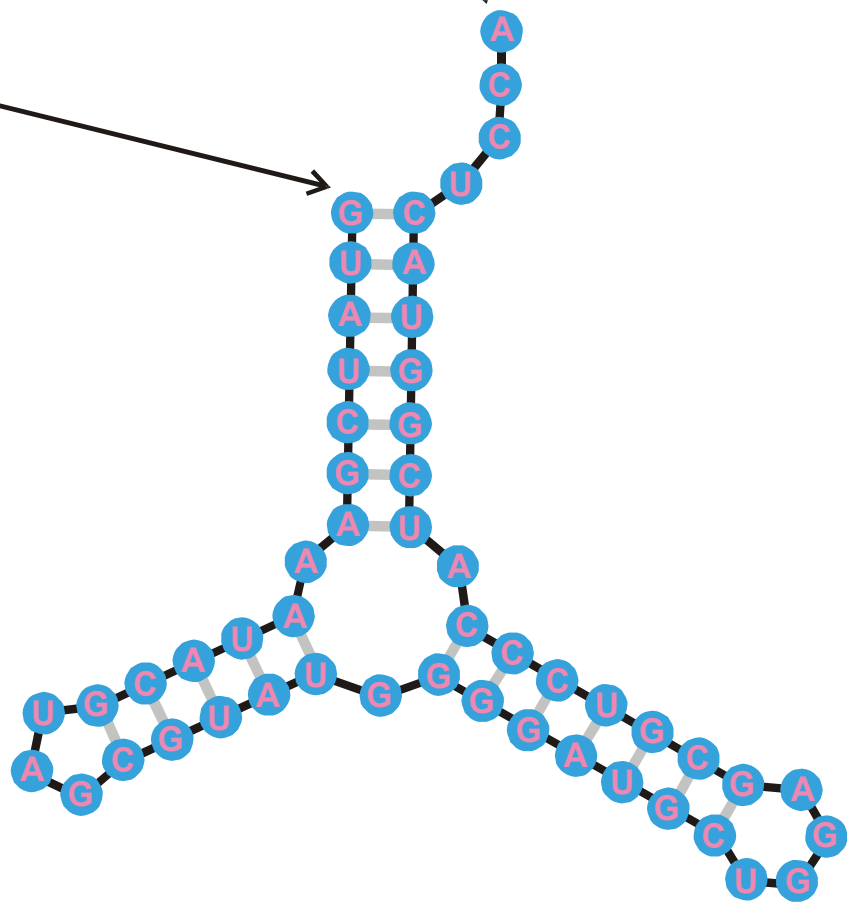
S.Wuchty, W.Fontana, I.L.Hofacker, and P.Schuster. *Biopolymers* **49**:145-165 (1999)

C.Flamm, W.Fontana, I.L.Hofacker, and P.Schuster. *RNA* **6**:325-338 (1999)

Vienna RNA Package: http://www.tbi.univie.ac.at

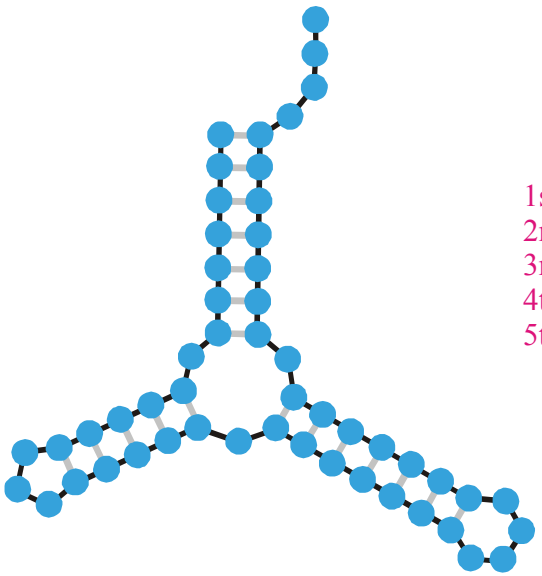
5’-end 3’-end

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



Folding of an RNA sequence into its secondary structure of minimum free energy

Base pair formation is the principle of folding RNA into secondary structures



Minimum free energy criterion

1st
2nd
3rd trial
4th
5th

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

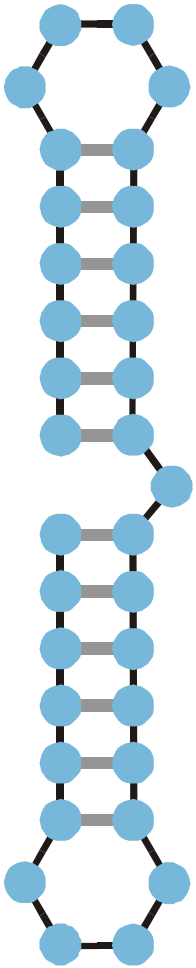
CUUCUUGAGCUAGUACCUAGUCGGAUAGGAUUUCCUAUCUCCAGGGAGGAUG

CUUUUCUUCACGUUAGAUGUGUAAUGGACAUGUGUUUAUUUAGGAAAGGCGC

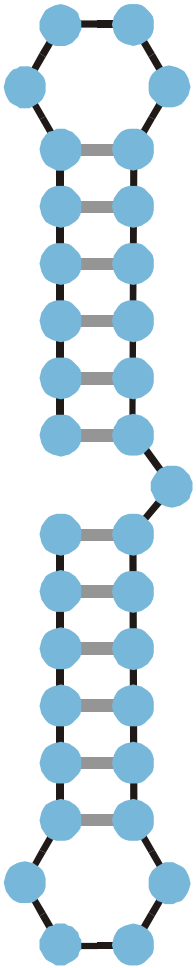
AUAACGUGAGUGUCUAAUACUGAUCGCUCCGGAGGGUGGUGGCGUUGUUAAU

Inverse folding of RNA secondary structures

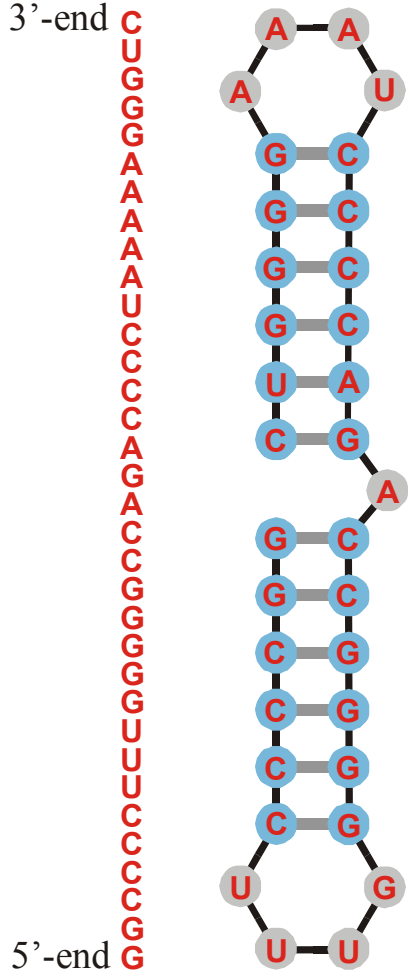
The inverse folding algorithm searches for sequences that form a given RNA secondary structure under the minimum free energy criterion.

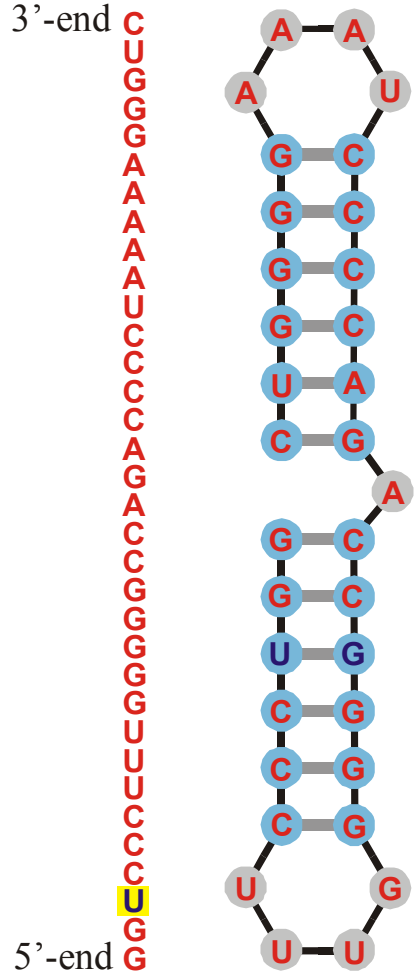


Structure

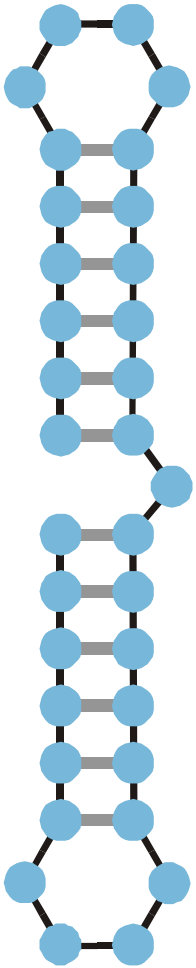


Structure

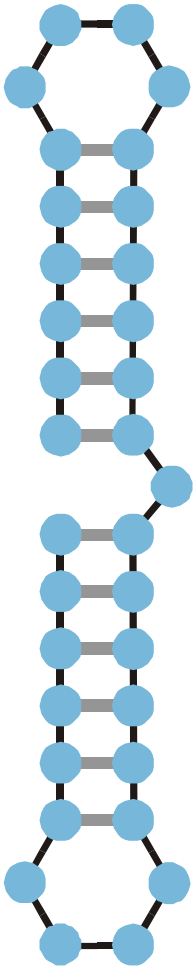




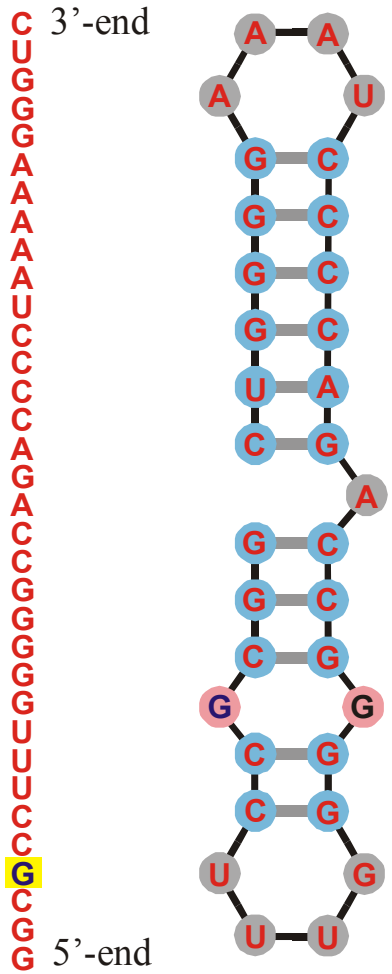
Compatible sequences



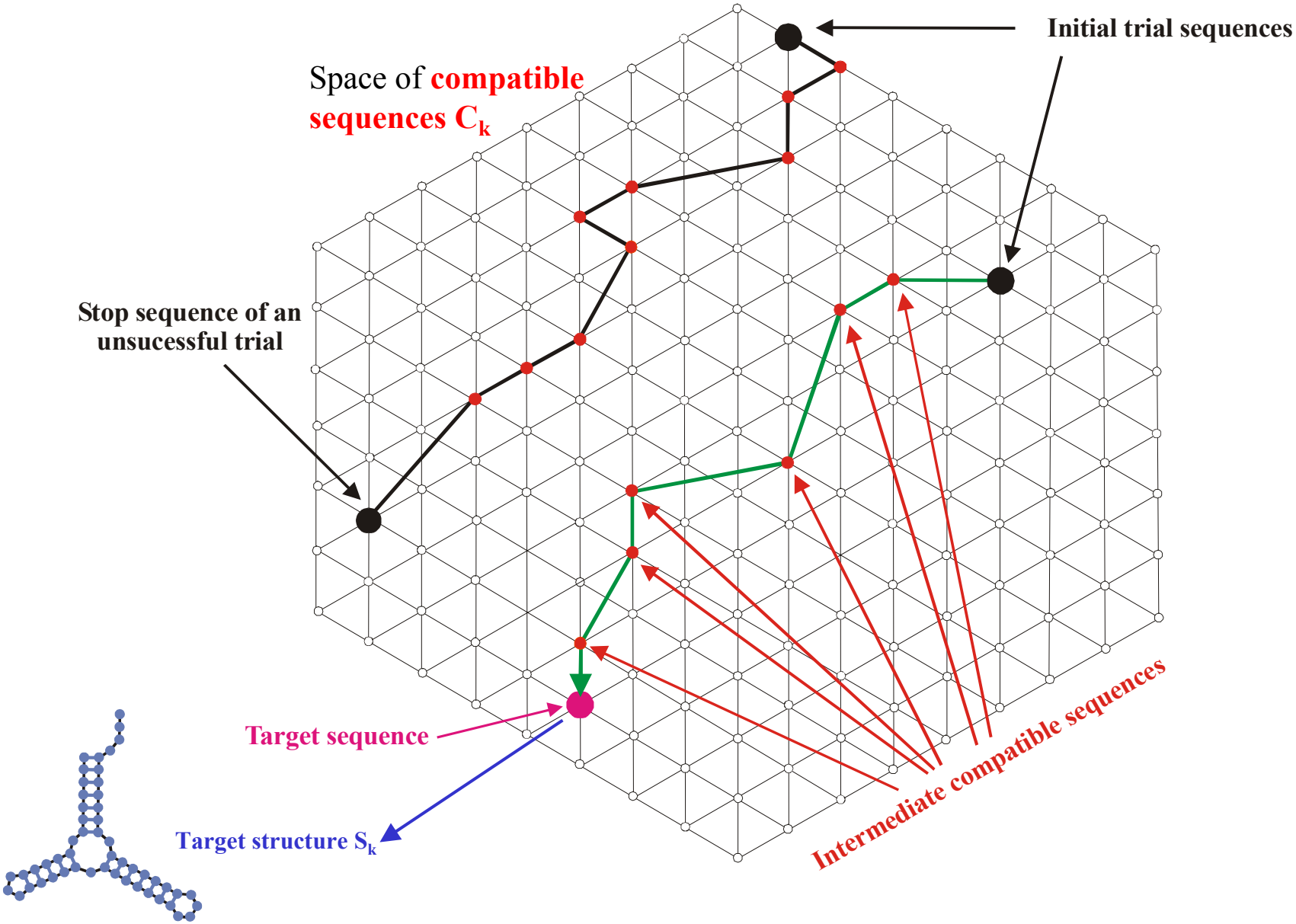
Structure



Structure



Incompatible sequence



Approach to the **target structure S_k** in the inverse folding algorithm

Theory of sequence – structure mappings

P. Schuster, W.Fontana, P.F.Stadler, I.L.Hofacker, ***From sequences to shapes and back: A case study in RNA secondary structures***. Proc.Roy.Soc.London **B 255** (1994), 279-284

W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks***. Mh.Chem. **127** (1996), 355-374

W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. II. Structure of neutral networks and shape space covering***. Mh.Chem. **127** (1996), 375-389

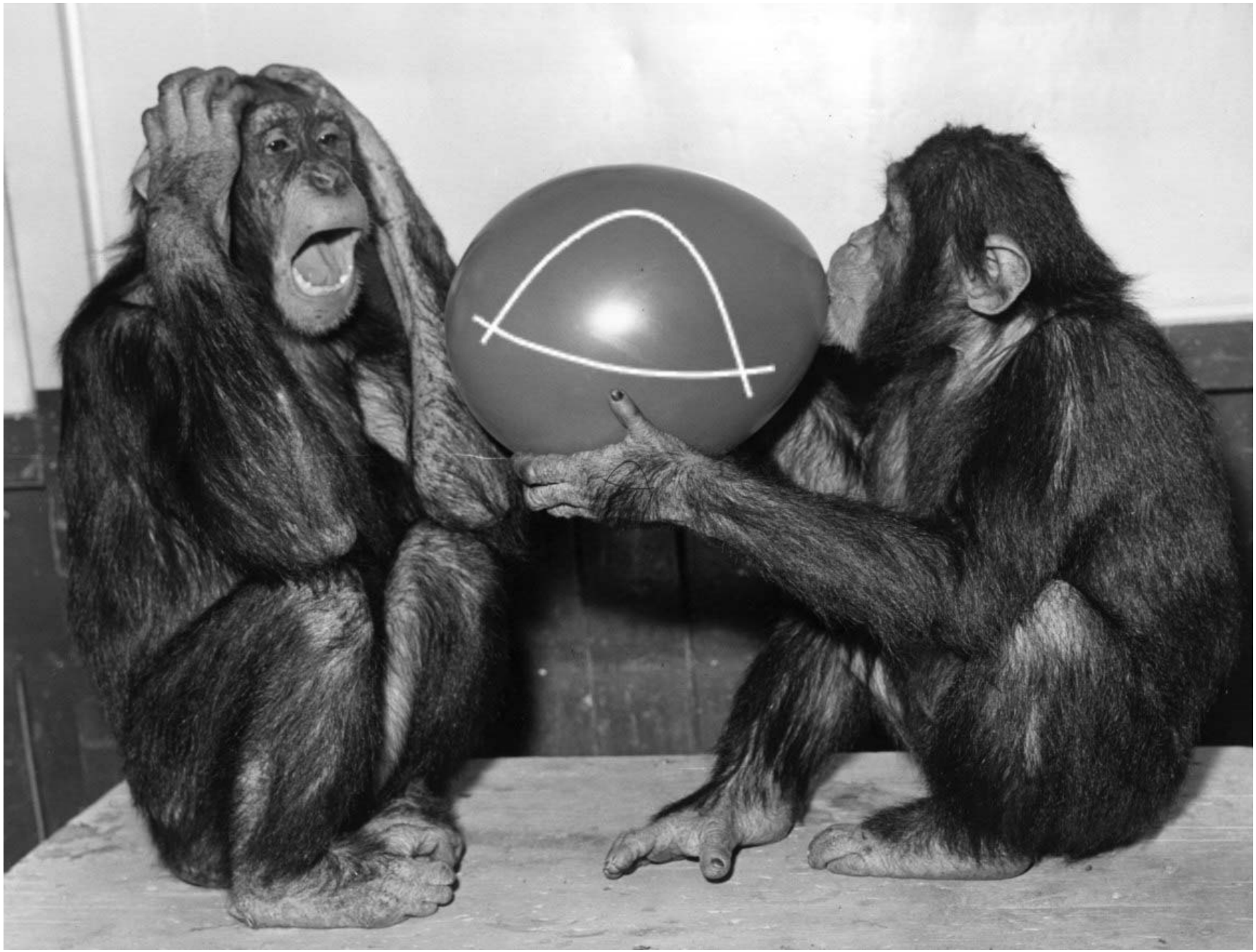
C.M.Reidys, P.F.Stadler, P.Schuster, ***Generic properties of combinatory maps***. Bull.Math.Biol. **59** (1997), 339-397

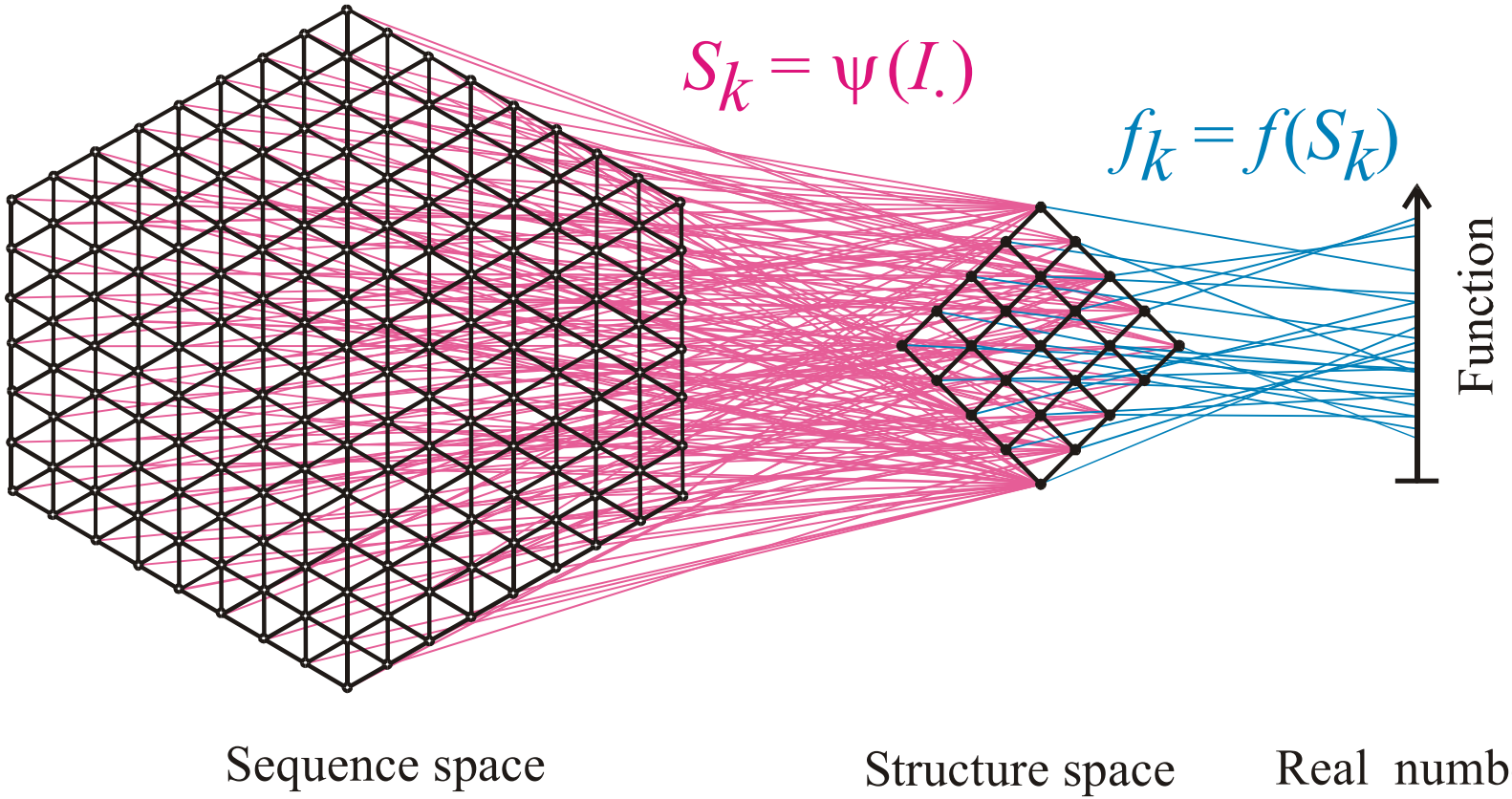
I.L.Hofacker, P. Schuster, P.F.Stadler, ***Combinatorics of RNA secondary structures***. Discr.Appl.Math. **89** (1998), 177-207

C.M.Reidys, P.F.Stadler, ***Combinatory landscapes***. SIAM Review **44** (2002), 3-54

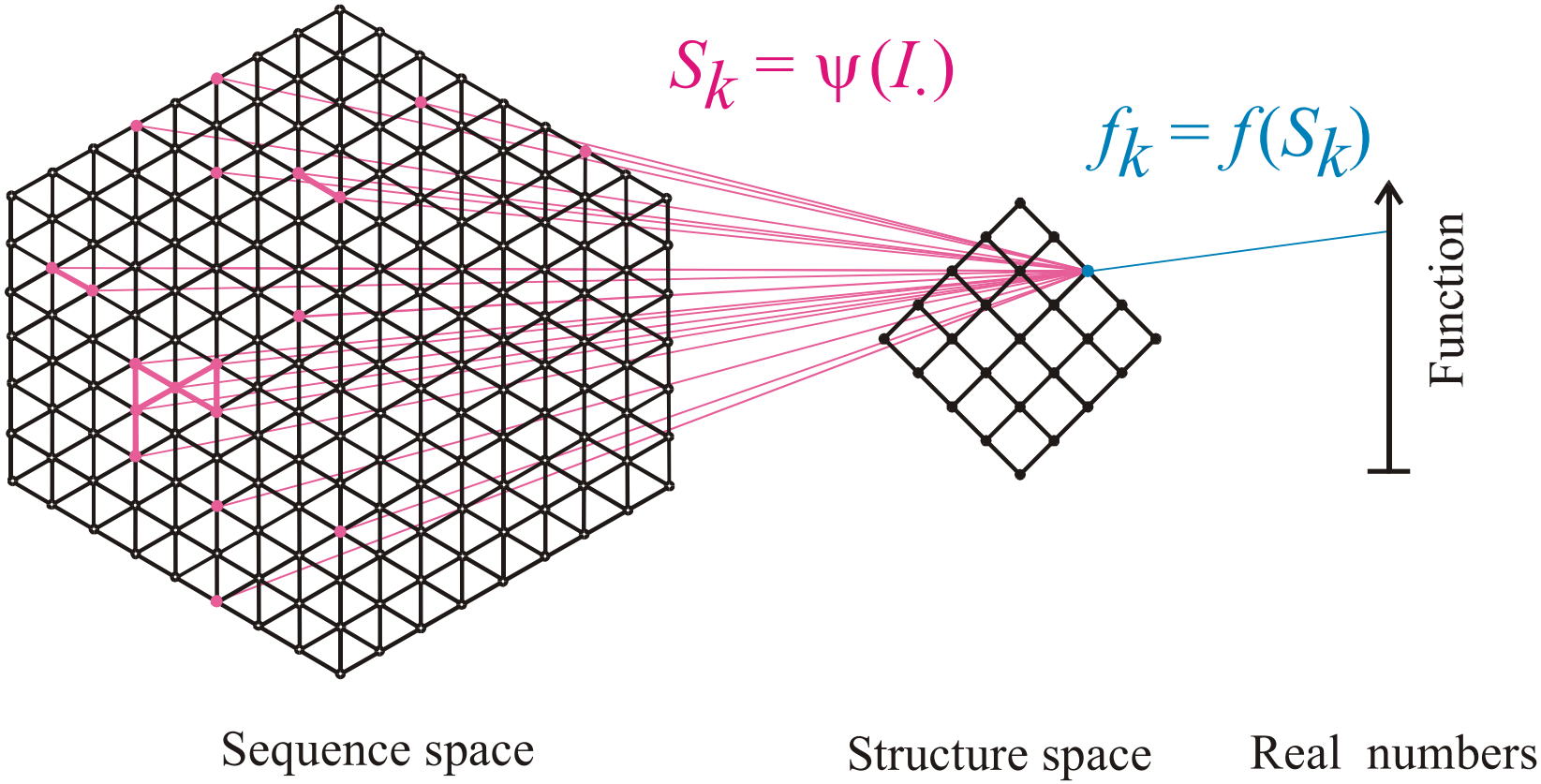
Sequence-structure relations are highly complex and only the simplest case can be studied. An example is the folding of RNA sequences into RNA structures represented in course-grained form as secondary structures.

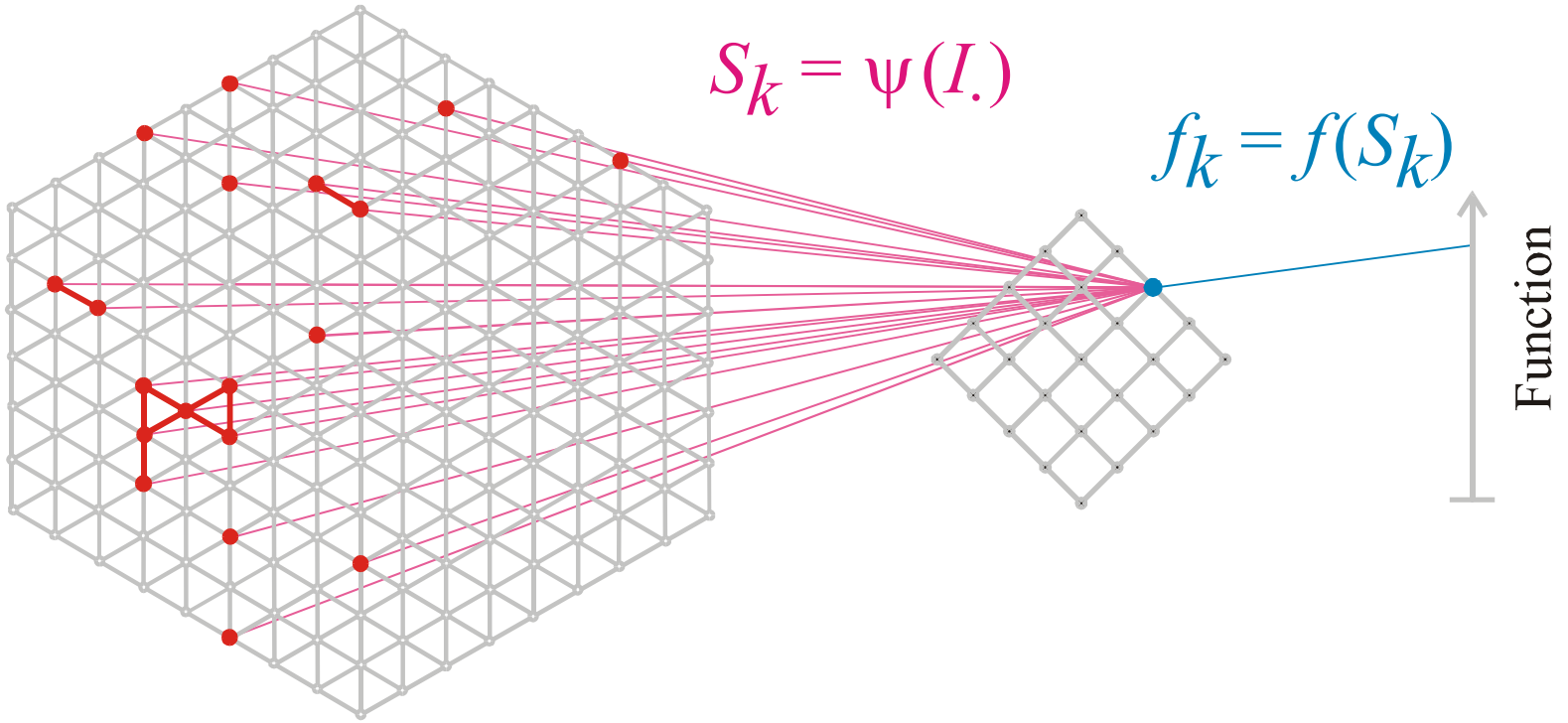
The RNA sequence-structure relation is understood as a mapping from the space of RNA sequences into a space of RNA structures.



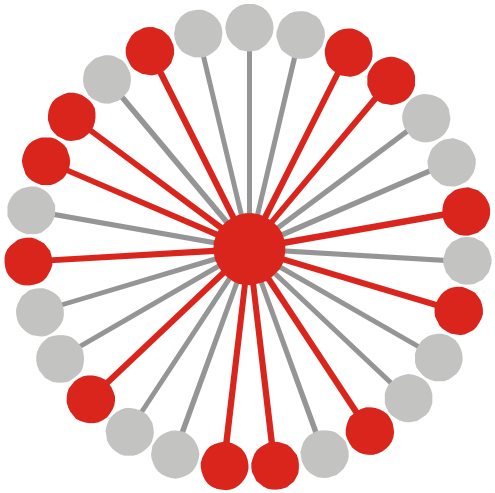


Mapping from sequence space into structure space and into function


$S_k = \psi(I_.)$
$f_k = f(S_k)$
Function
Sequence space
Structure space
Real numbers



The pre-image of the structure S_k in sequence space is the **neutral network G_k**



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

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

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

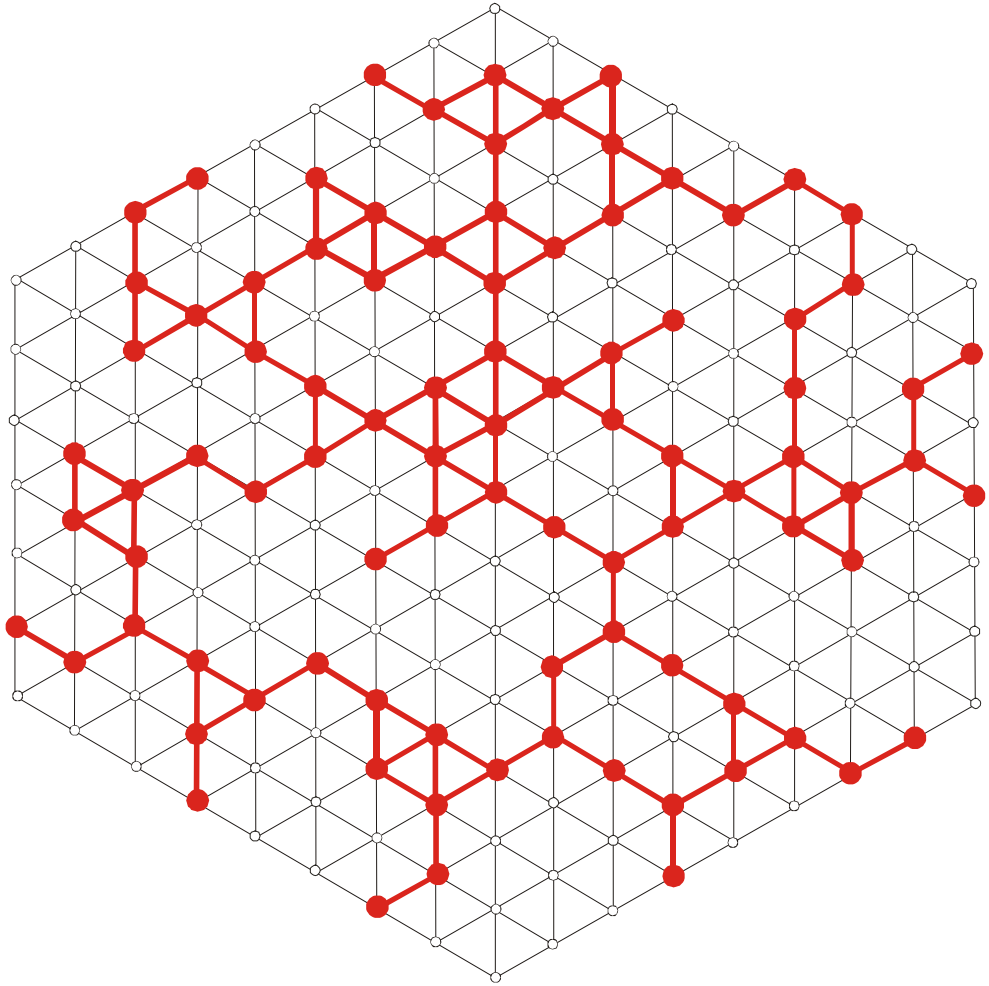
Alphabet size k : **AUGC** ñ k = 4

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

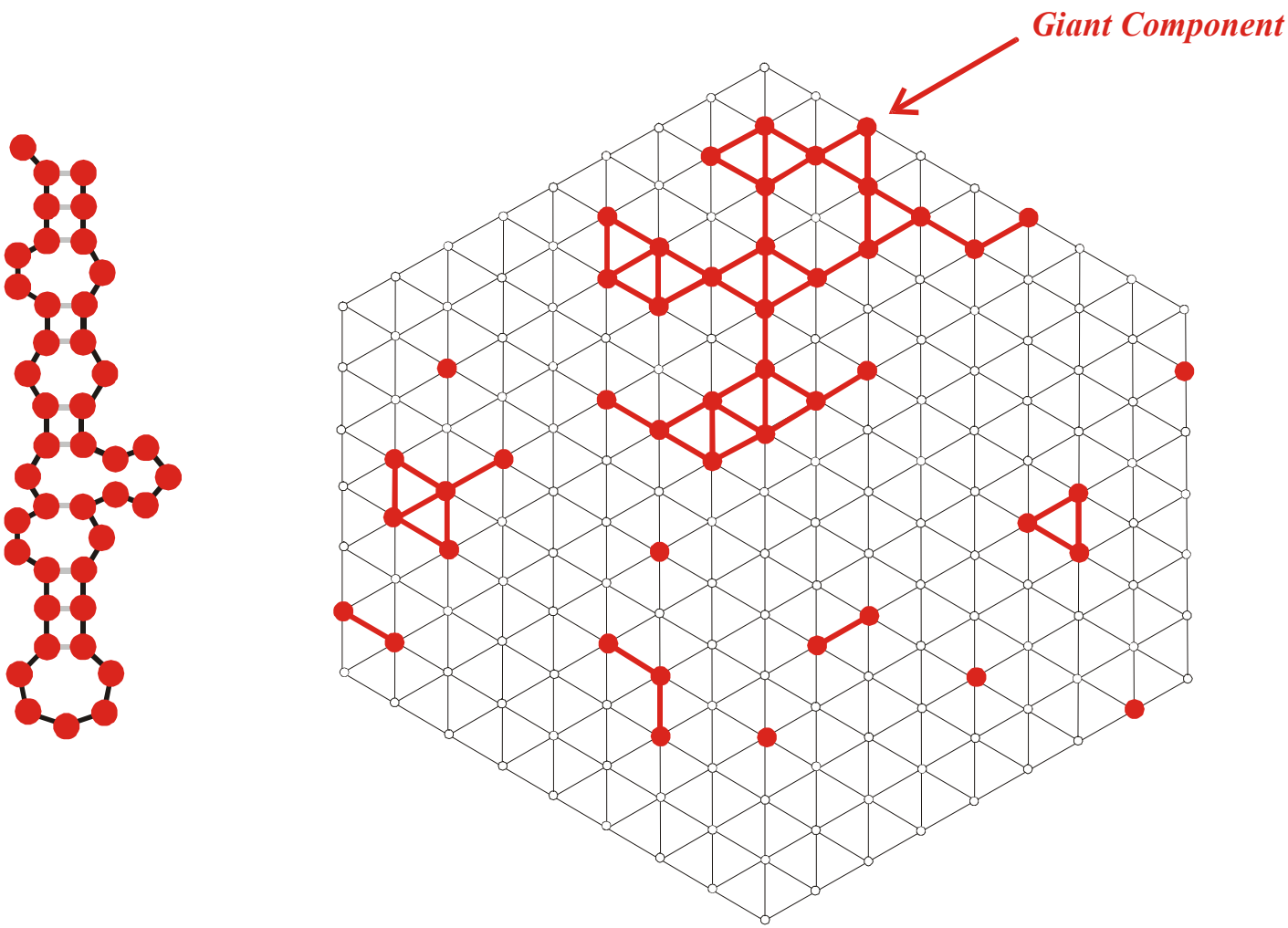
$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

k	I_{cr}	
2	0.5	**GC**
3	0.423	**GUC**
4	0.370	**AUGC**

Mean degree of neutrality and connectivity of **neutral networks**



A connected neutral network



A multi-component neutral network

From sequences to shapes and back: a case study in RNA secondary structures

PETER SCHUSTER[1,2,3], WALTER FONTANA[3], PETER F. STADLER[2,3] AND IVO L. HOFACKER[2]

[1] *Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany*
[2] *Institut für Theoretische Chemie, Universität Wien, Austria*
[3] *Santa Fe Institute, Santa Fe, U.S.A.*

SUMMARY

RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law: we find relatively few common and many rare ones. By using an algorithm for inverse folding, we show that sequences sharing the same structure are distributed randomly over sequence space. All common structures can be accessed from an arbitrary sequence by a number of mutations much smaller than the chain length. The sequence space is percolated by extensive neutral networks connecting nearest neighbours folding into identical structures. Implications for evolutionary adaptation and for applied molecular evolution are evident: finding a particular structure by mutation and selection is much simpler than expected and, even if catalytic activity should turn out to be sparse in the space of RNA structures, it can hardly be missed by evolutionary processes.

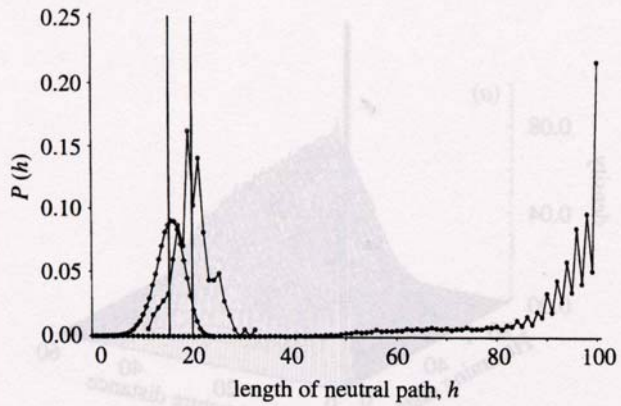
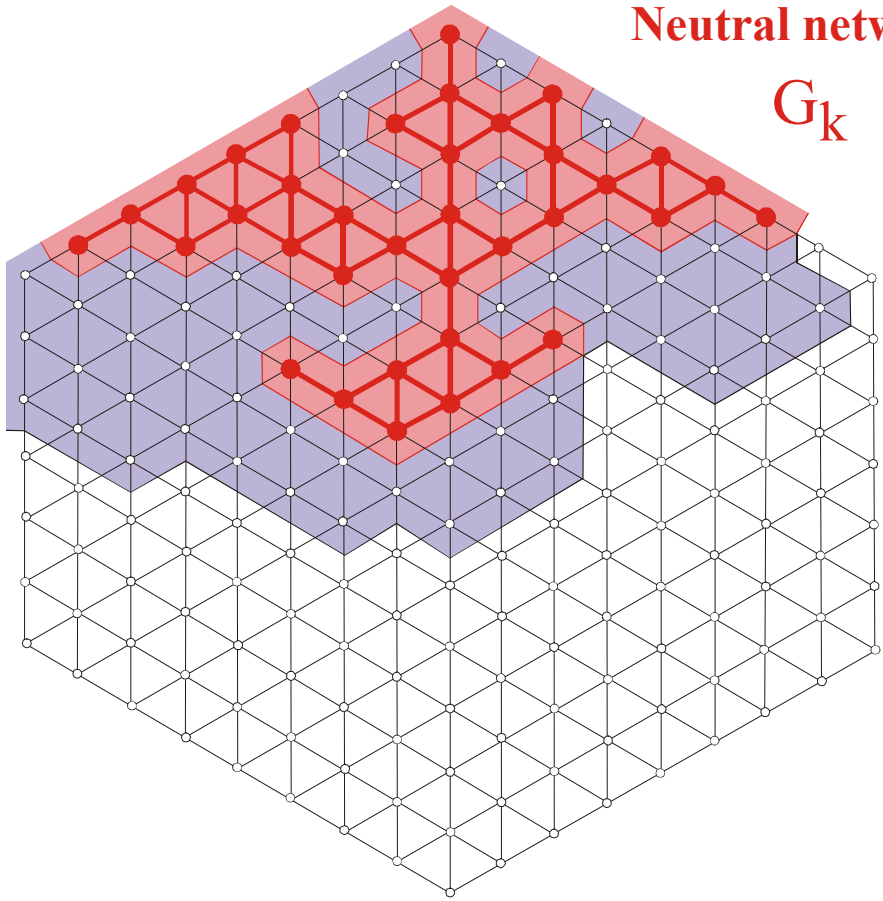


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and neighbours of Hamming distance 2, resulting from base pair exchanges in stacks. Two probability densities of Hamming distances are shown that were obtained by searching for neutral paths in sequence space: (i) an upper bound for the closest approach of trial and target sequences (open circles) obtained as endpoints of neutral paths approaching the target from a random trial sequence (185 targets and 100 trials for each were used); (ii) a lower bound for the closest approach of trial and target sequences (open diamonds) derived from secondary structure statistics (Fontana *et al.* 1993*a*; see this paper, §4); and (iii) longest distances between the reference and the endpoints of monotonously diverging neutral paths (filled circles) (500 reference sequences were used).

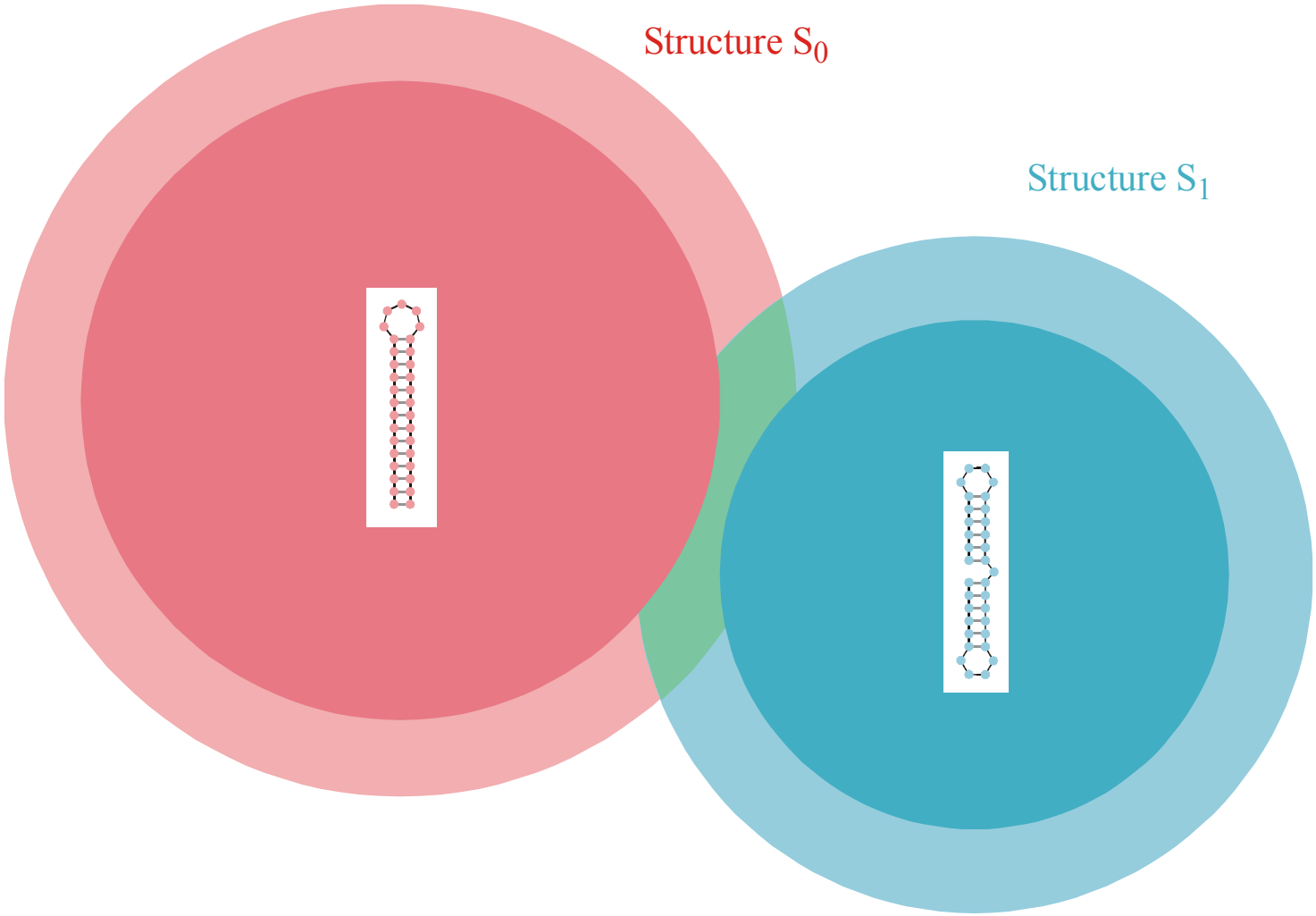
Proc. R. Soc. Lond. B (1994) **255**, 279–284
Printed in Great Britain

Reference for postulation and ***in silico*** verification of ***neutral networks***

G_k à C_k



The **compatible set C_k** of a structure $\mathbf{S_k}$ consists of all sequences which form $\mathbf{S_k}$ as its minimum free energy structure (**neutral network G_k**) or one of its suboptimal structures.



Intersection of two compatible sets: $C_0 \cap C_1$

The intersection of two compatible sets is always non empty: C_0 Ú C_1 â Ù

Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997
Elsevier Science Inc.

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES[1]

■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡ and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

†Los Alamos National Laboratory,
Los Alamos, NM 87545, U.S.A.

‡Institut für Theoretische Chemie der Universität Wien,
A-1090 Wien, Austria

§Institut für Molekulare Biotechnologie,
D-07708 Jena, Germany

(*E.mail: pks@tbi.univie.ac.at*)

Random graph theory is used to model and analyse the relationships between sequences and secondary structures of RNA molecules, which are understood as mappings from sequence space into shape space. These maps are non-invertible since there are always many orders of magnitude more sequences than structures. Sequences folding into identical structures form *neutral networks*. A neutral network is embedded in the set of sequences that are *compatible* with the given structure. Networks are modeled as graphs and constructed by random choice of vertices from the space of compatible sequences. The theory characterizes neutral networks by the mean fraction of neutral neighbors (λ). The networks are connected and percolate sequence space if the fraction of neutral nearest neighbors exceeds a threshold value ($\lambda > \lambda^*$). Below threshold ($\lambda < \lambda^*$), the networks are partitioned into a largest "giant" component and several smaller components. Structures are classified as "common" or "rare" according to the sizes of their pre-images, i.e. according to the fractions of sequences folding into them. The neutral networks of any pair of two different common structures almost touch each other, and, as expressed by the conjecture of *shape space covering* sequences folding into almost all common structures, can be found in a small ball of an arbitrary location in sequence space. The results from random graph theory are compared to data obtained by folding large samples of RNA sequences. Differences are explained in terms of specific features of RNA molecular structures. © *1997 Society for Mathematical Biology*

THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s′ *be arbitrary secondary structures and* **C**[s], **C**[s′] *their corresponding compatible sequences. Then,*

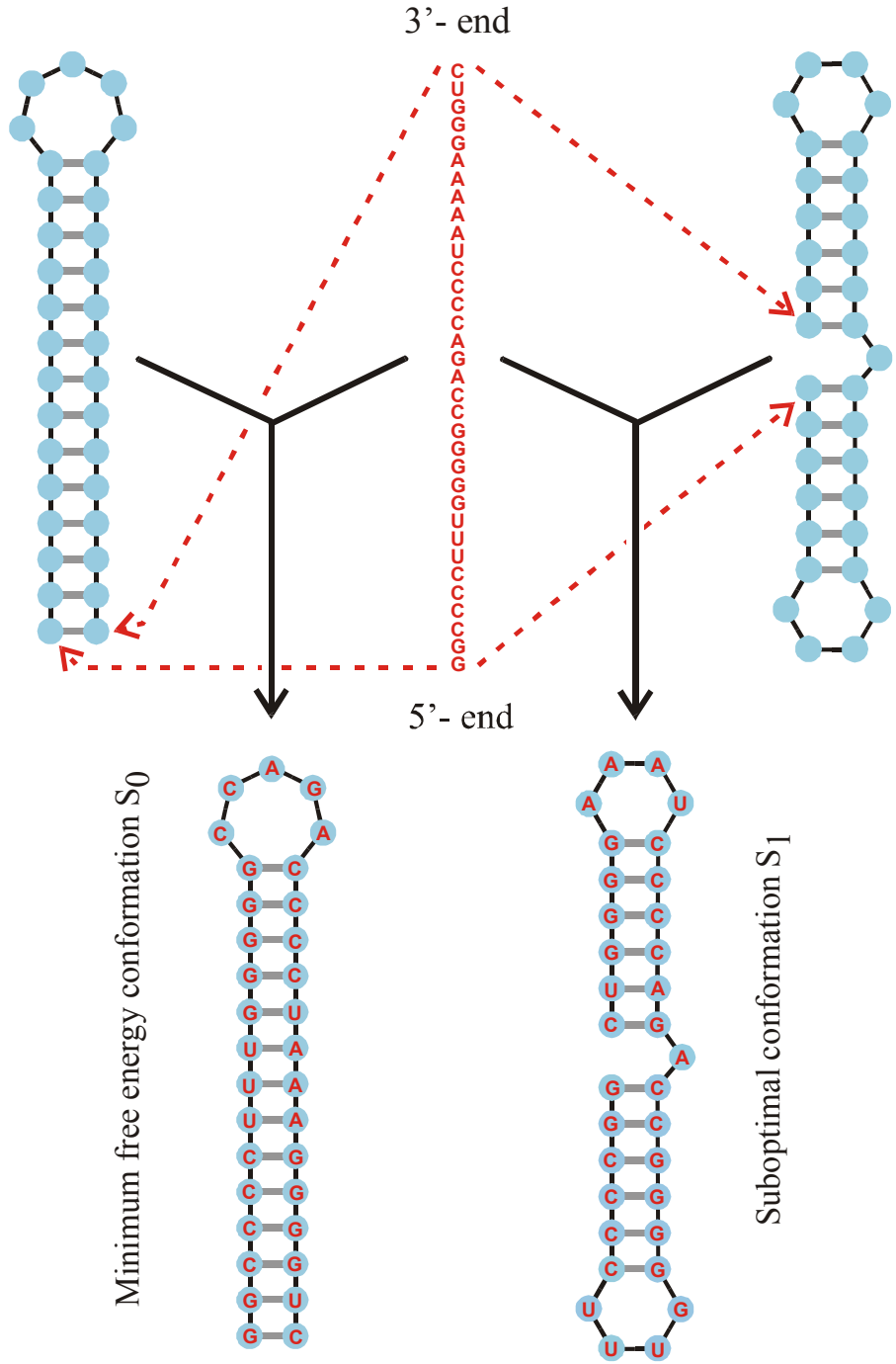
$$\mathbf{C}[s] \cap \mathbf{C}[s'] \neq \varnothing.$$

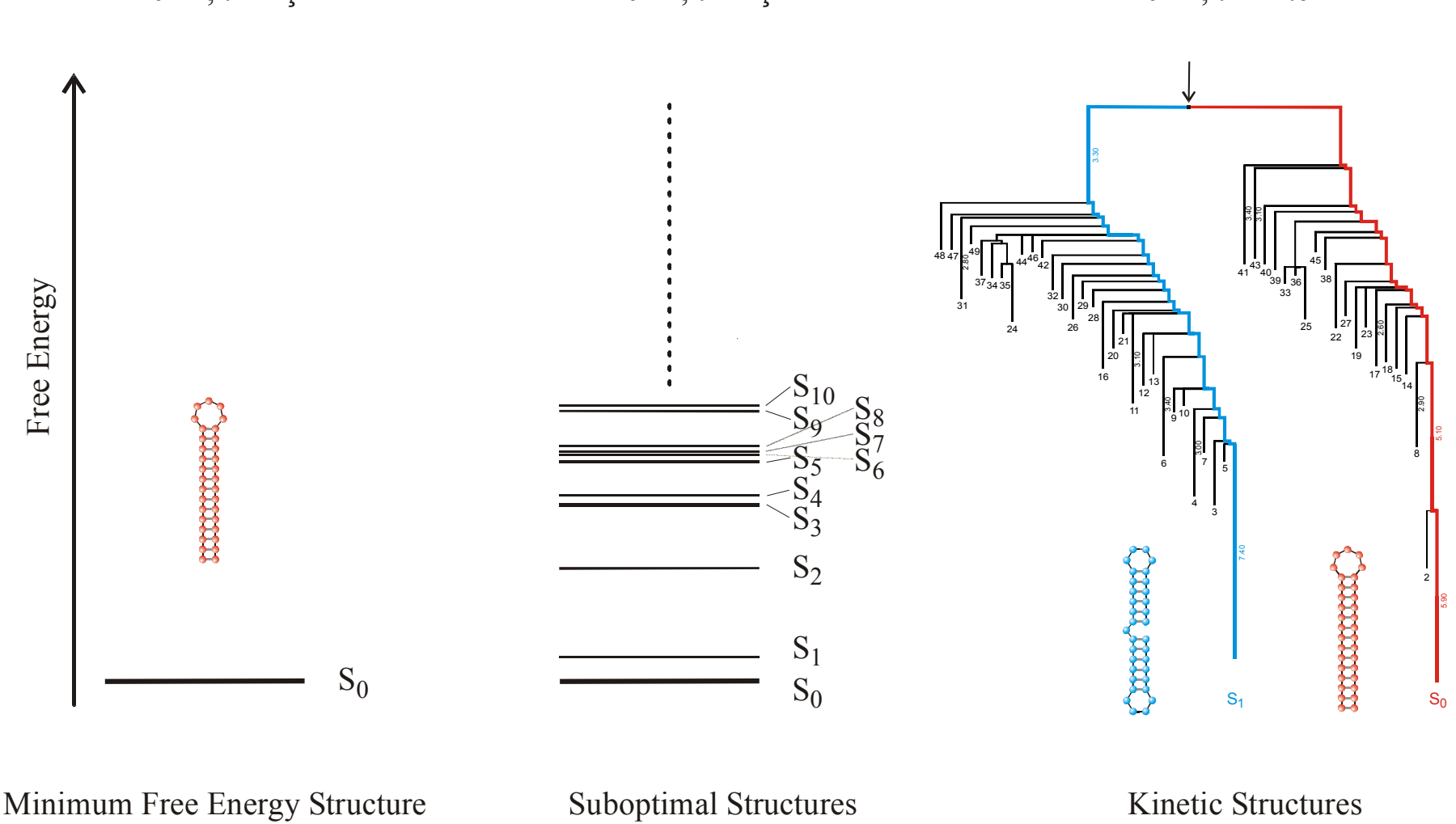
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\gamma(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either cycles or chains and the cycles have even order. A constraint for the sequence compatible to both structures appears only in the cycles where the choice of bases is not independent. It remains to be shown that there is a valid choice of bases for each cycle, which is obvious since these have even order. Therefore, it suffices to choose an alternating sequence of the pairing partners X and Y. Thus, there are at least two different choices for the first base in the orbit. ■

Remark. A generalization of the statement of theorem 5 to three different structures is false.

Reference for the definition of the intersection and the proof of the **intersection theorem**

A sequence at the **intersection** of two neutral networks is compatible with both structures





Minimum Free Energy Structure

Suboptimal Structures

Kinetic Structures

Different notions of RNA structure including suboptimal conformations and folding kinetics

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*).

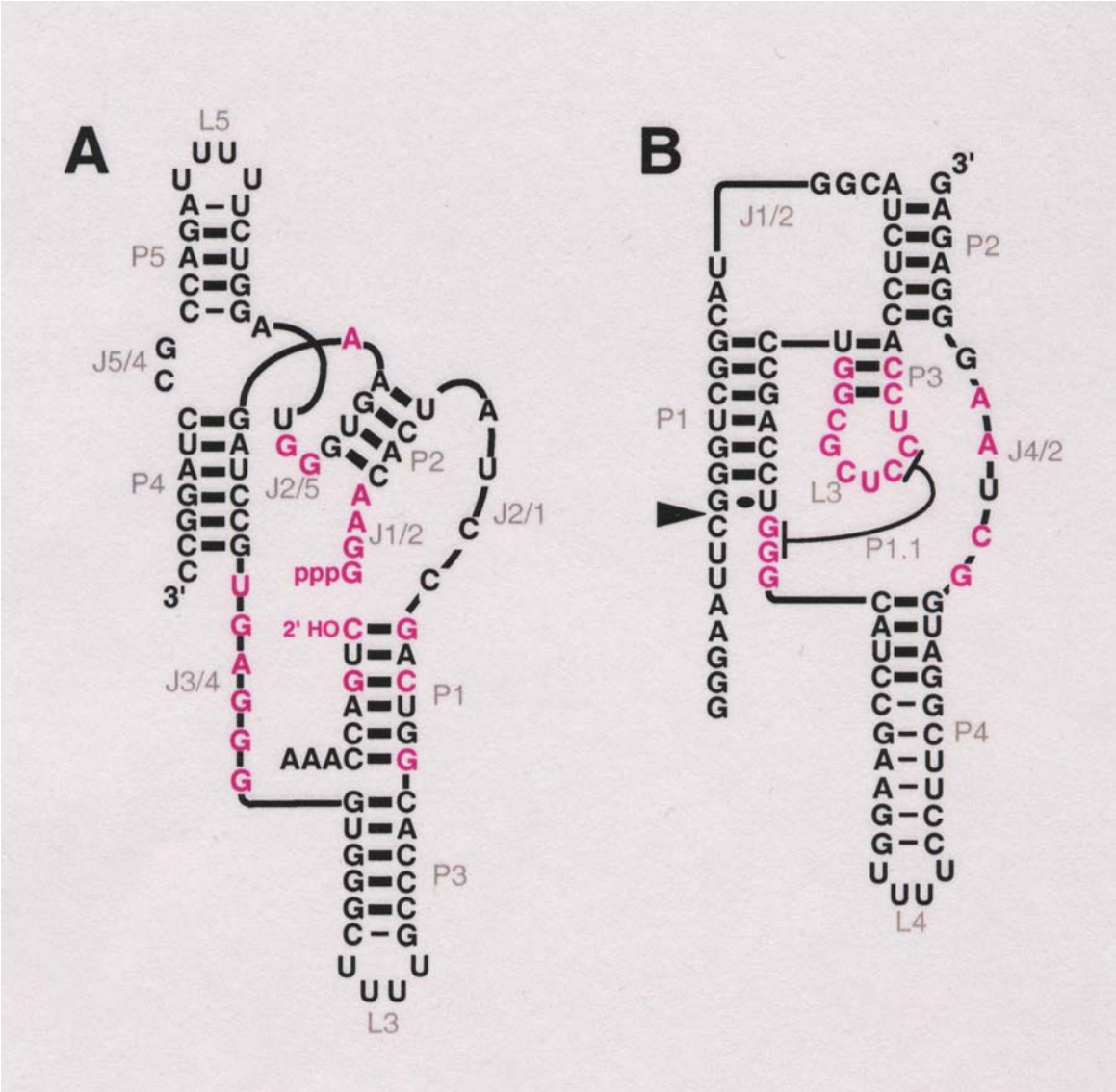
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. 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.

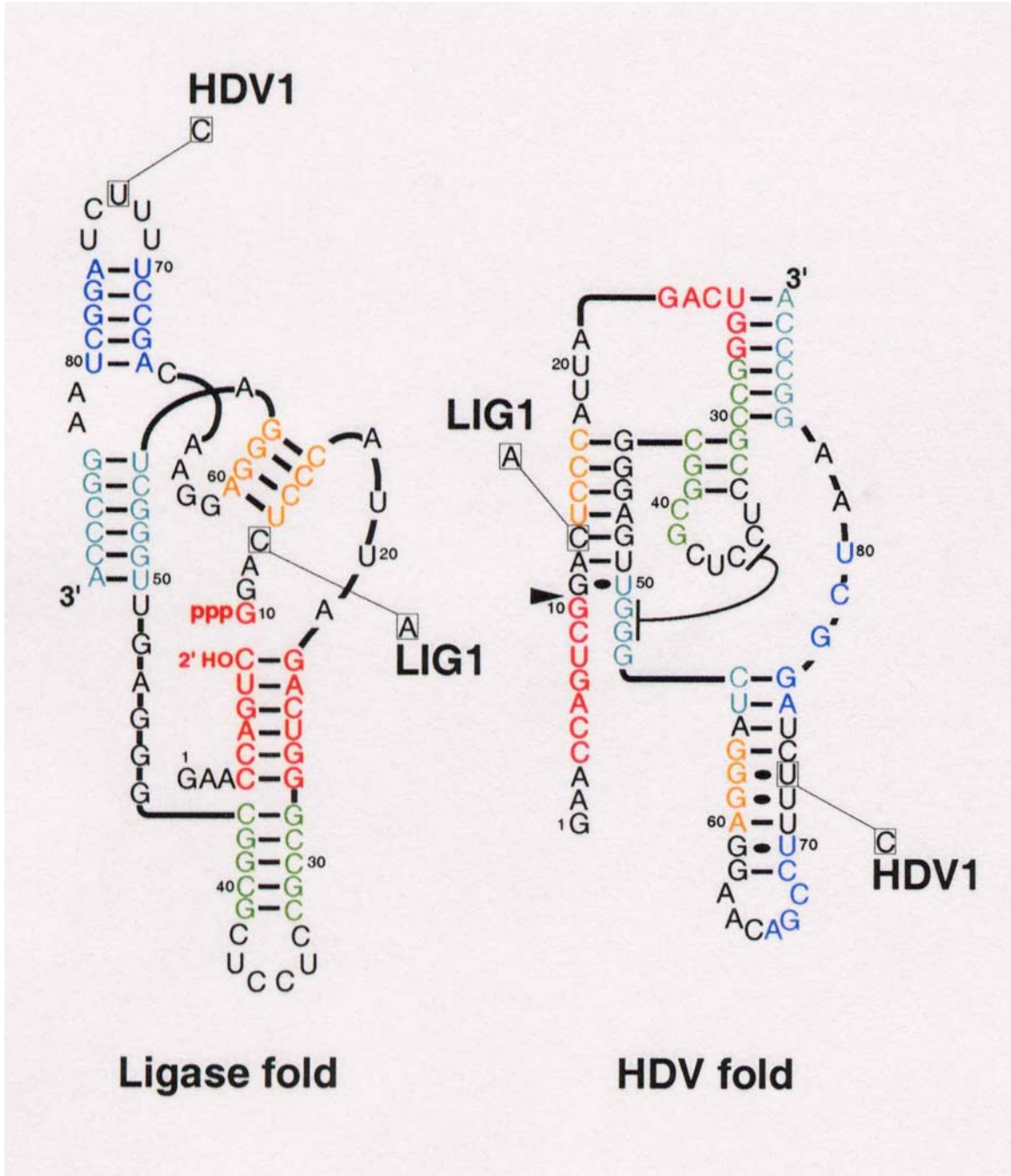
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

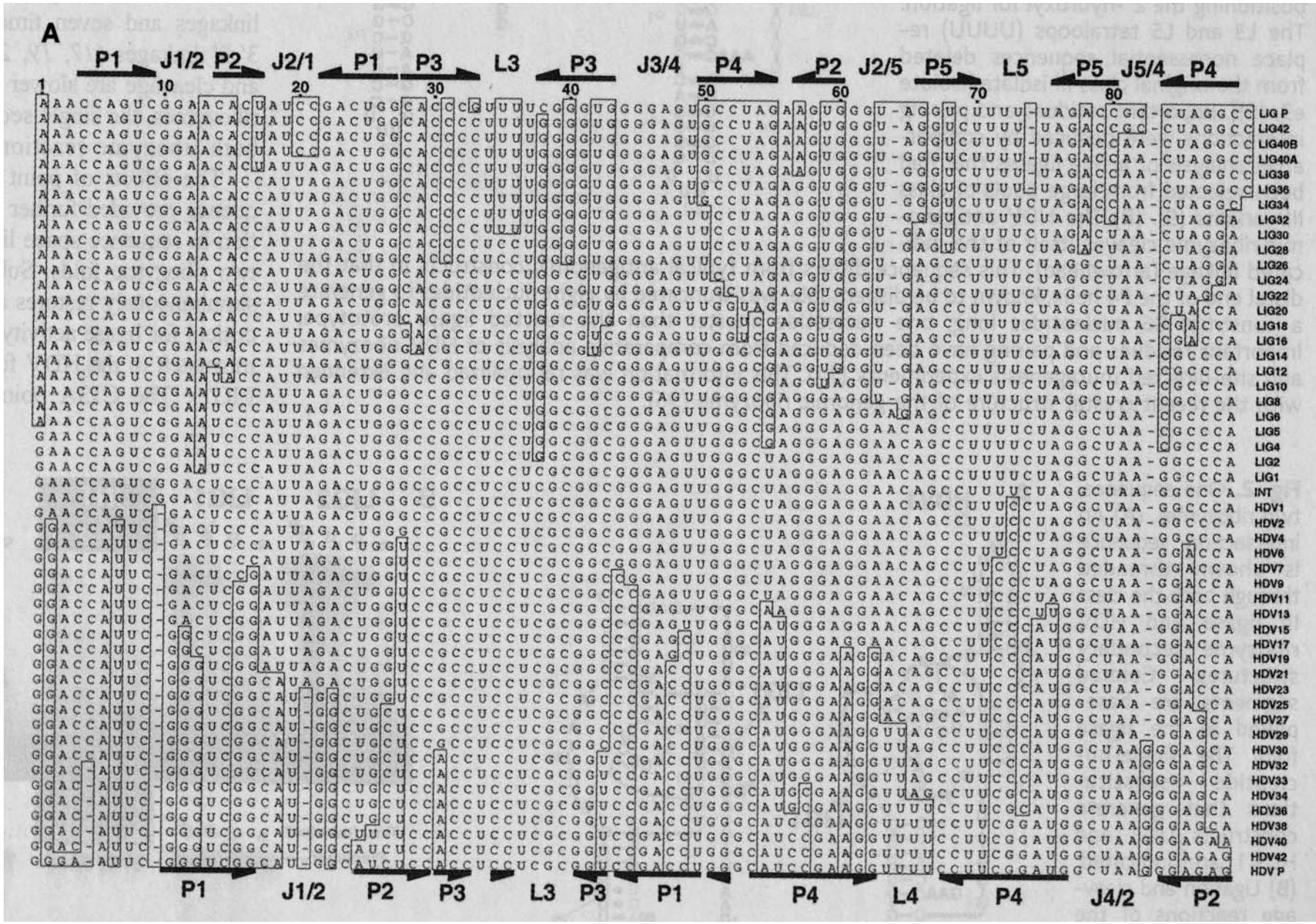


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-d-virus (**B**)

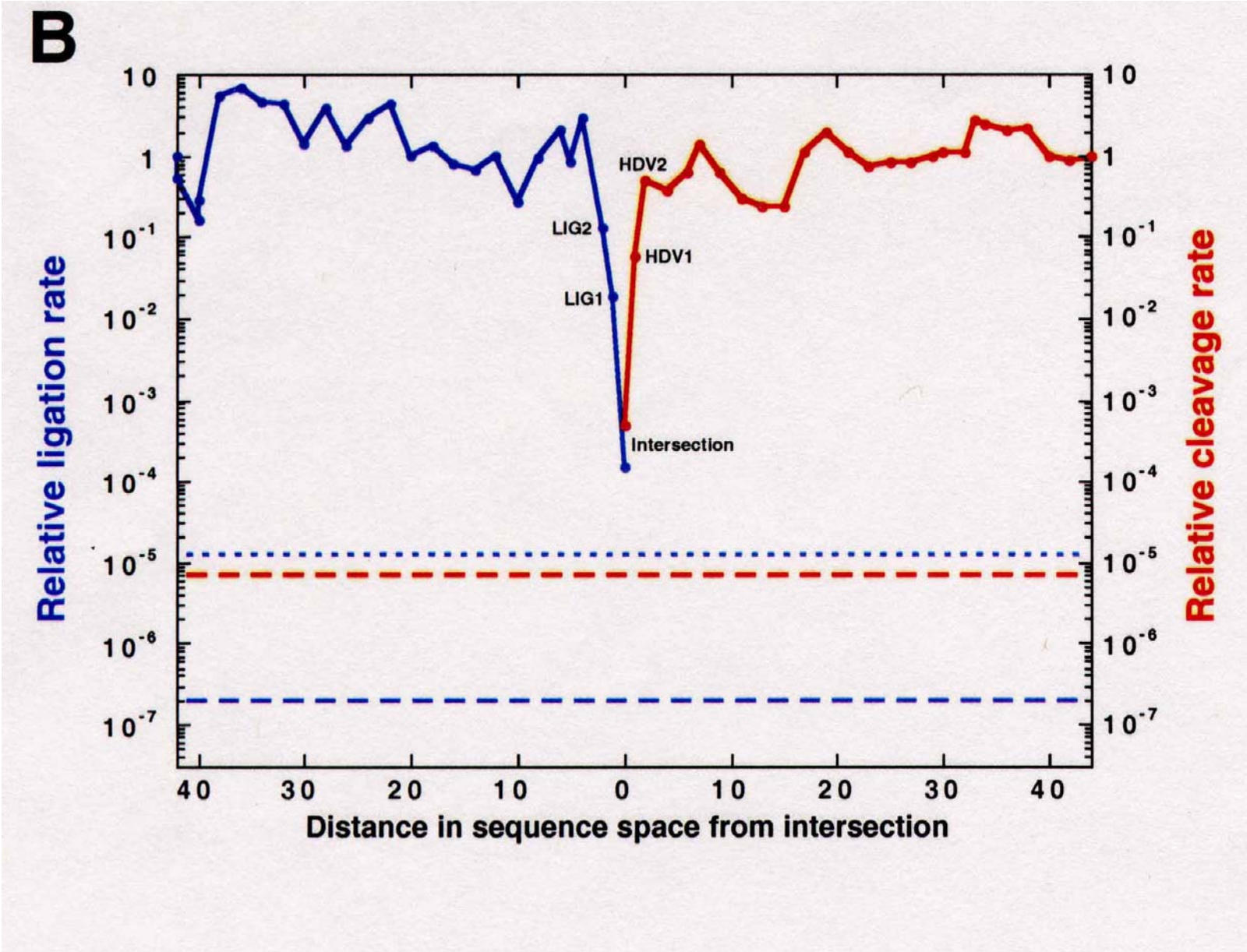
The sequence at the ***intersection***:

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





Sequence of mutants from the intersection to both reference ribozymes



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

1. Experiments on controlled evolution and RNA replication

2. Sequence-structure maps, neutral networks, and intersections

3. Optimization in the RNA model

4. What we can learn from molecules for evolution proper

Optimization of RNA molecules *in silico*

W.Fontana, P.Schuster, ***A computer model of evolutionary optimization***. Biophysical Chemistry **26** (1987), 123-147

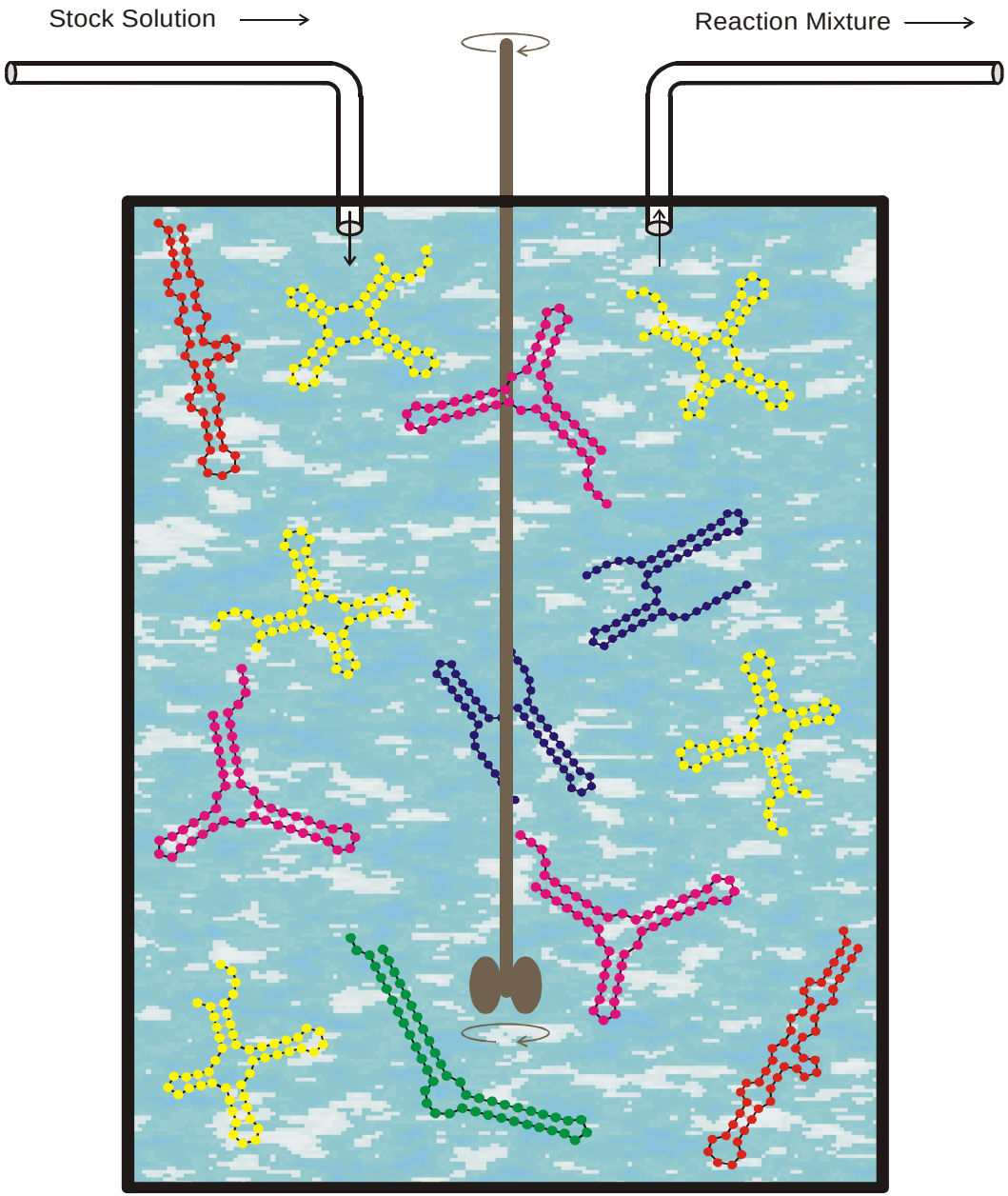
W.Fontana, W.Schnabl, P.Schuster, ***Physical aspects of evolutionary optimization and adaptation***. Phys.Rev.A **40** (1989), 3301-3321

M.A.Huynen, W.Fontana, P.F.Stadler, ***Smoothness within ruggedness. The role of neutrality in adaptation***. Proc.Natl.Acad.Sci.USA **93** (1996), 397-401

W.Fontana, P.Schuster, ***Continuity in evolution. On the nature of transitions***. Science **280** (1998), 1451-1455

W.Fontana, P.Schuster, ***Shaping space. The possible and the attainable in RNA genotype-phenotype mapping***. J.Theor.Biol. **194** (1998), 491-515

B.M.R. Stadler, P.F. Stadler, G.P. Wagner, W. Fontana, ***The topology of the possible: Formal spaces underlying patterns of evolutionary change.*** J.Theor.Biol. **213** (2001), 241-274

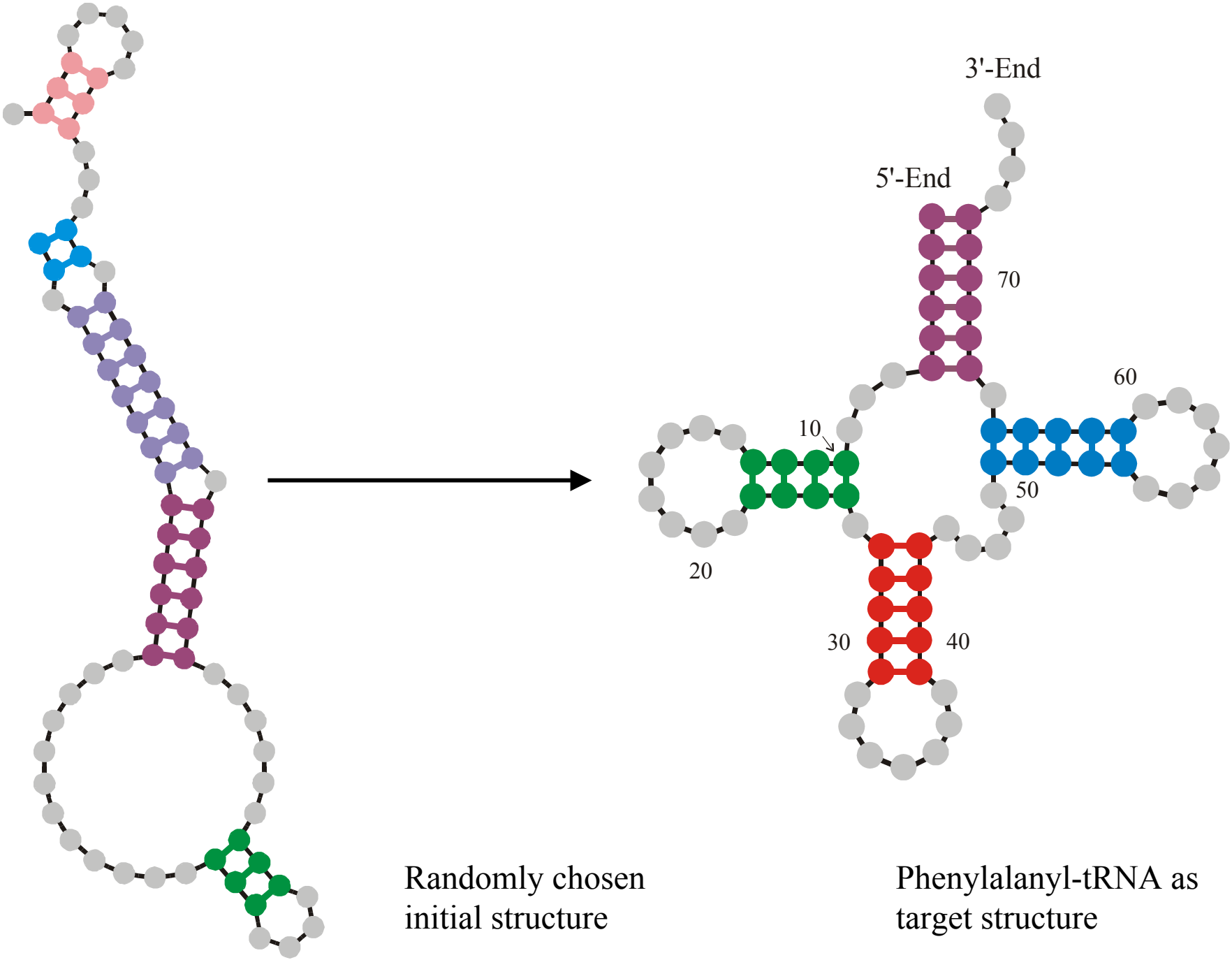


Fitness function:

$f_k = g / [a + Dd_S^{(k)}]$

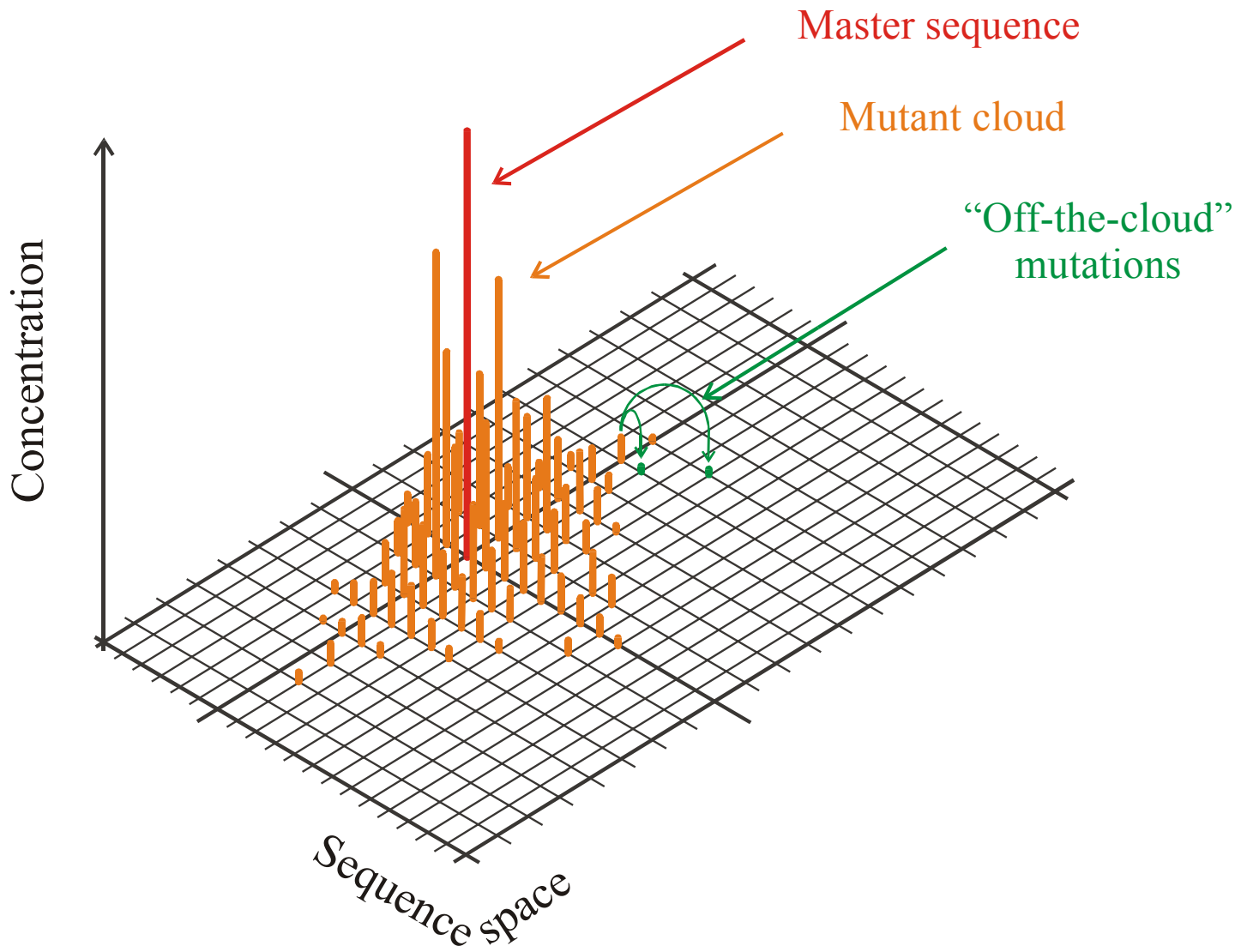
$Dd_S^{(k)} = d^s(I_k, I_t)$

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

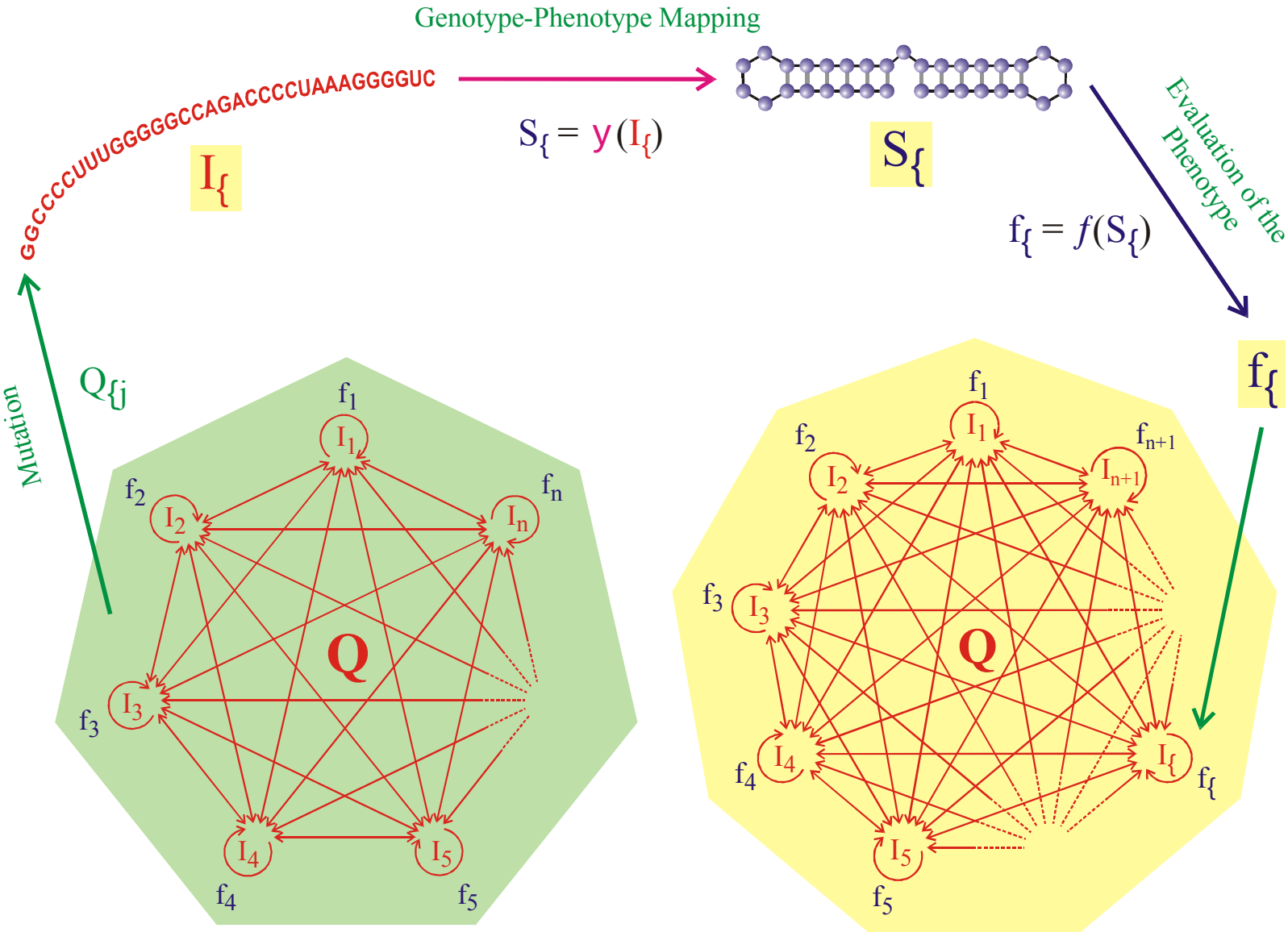


Randomly chosen
initial structure

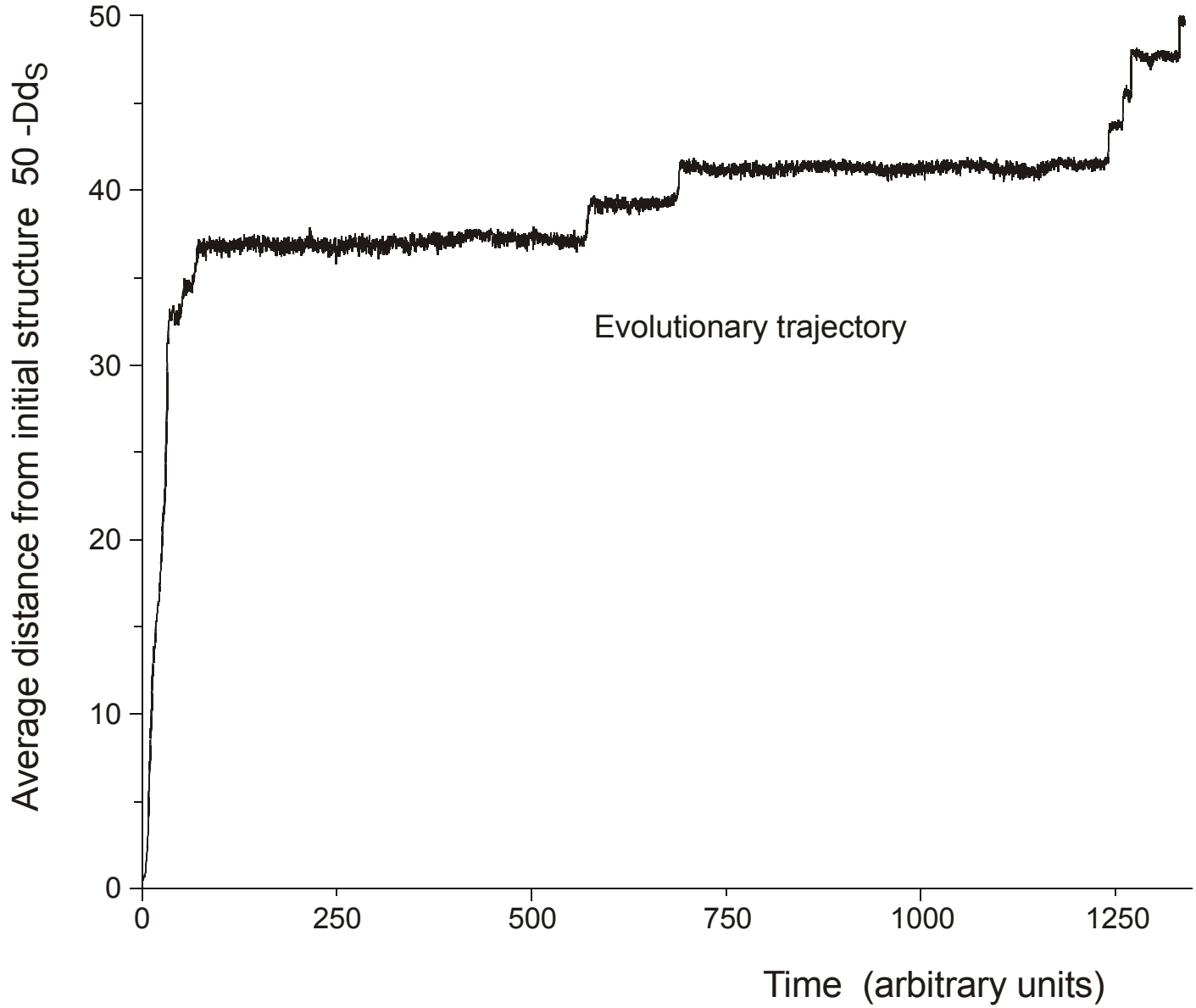
Phenylalanyl-tRNA as
target structure



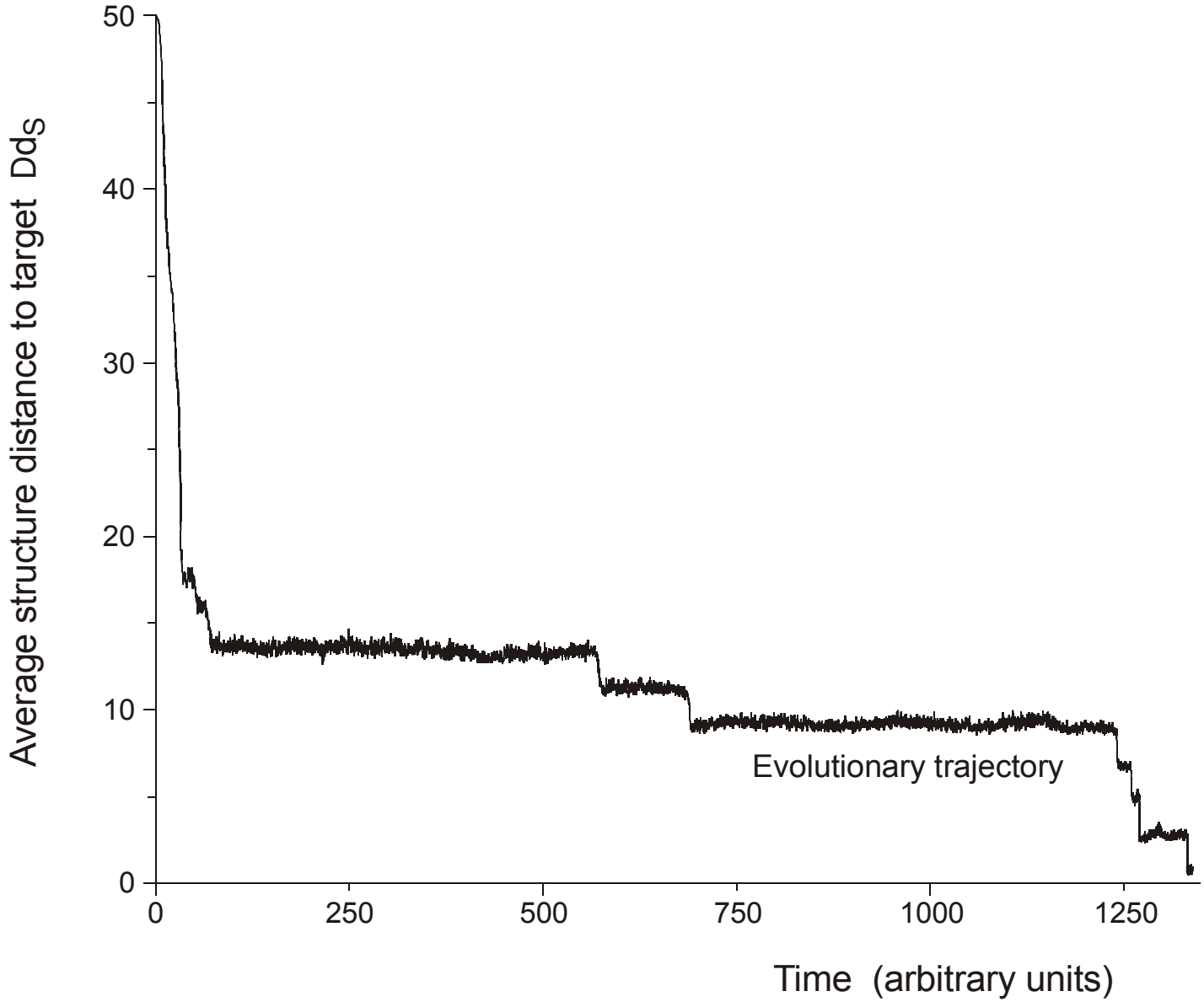
The molecular quasispecies in sequence space



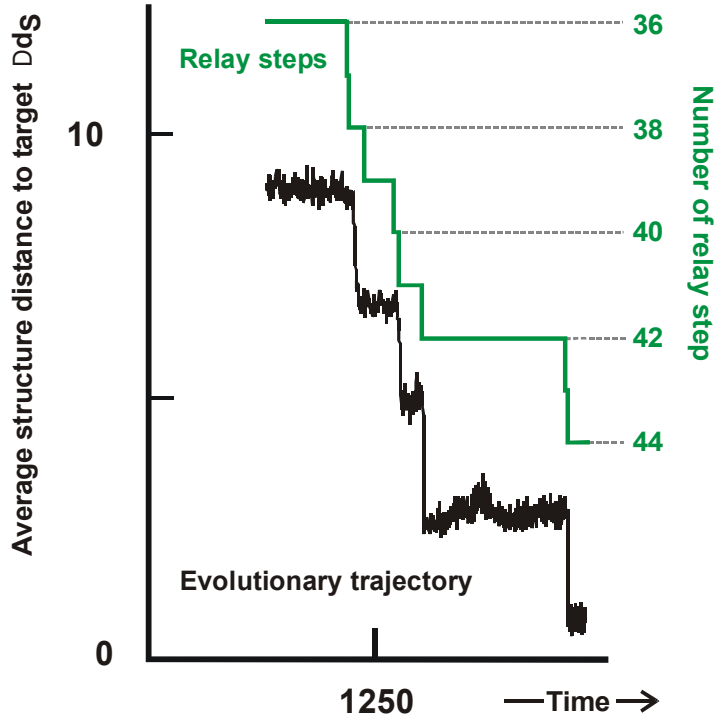
Evolutionary dynamics
including molecular phenotypes

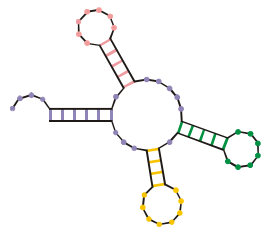


In silico optimization in the flow reactor: Trajectory (**biologists‘ view**)



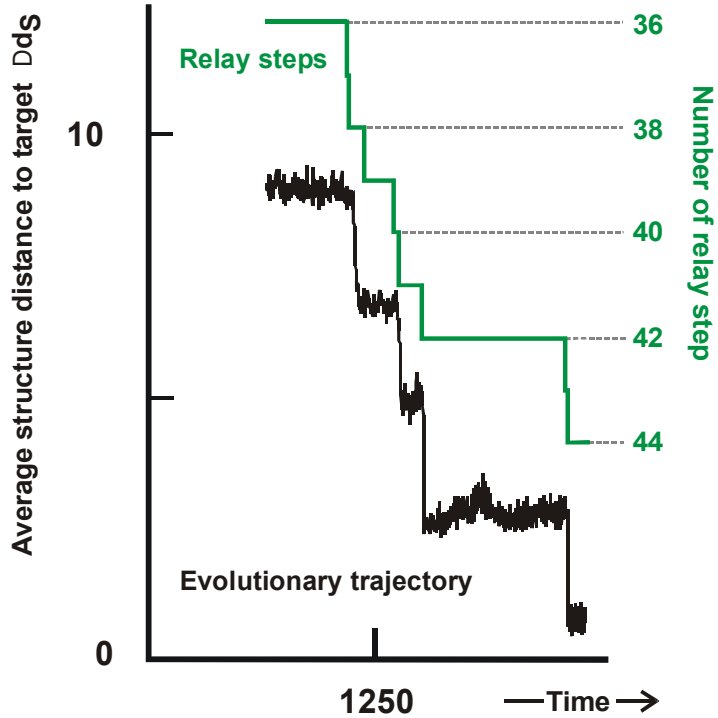
In silico optimization in the flow reactor: Trajectory (**physicists‘ view**)

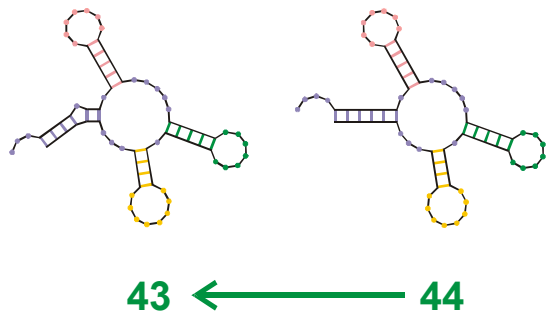




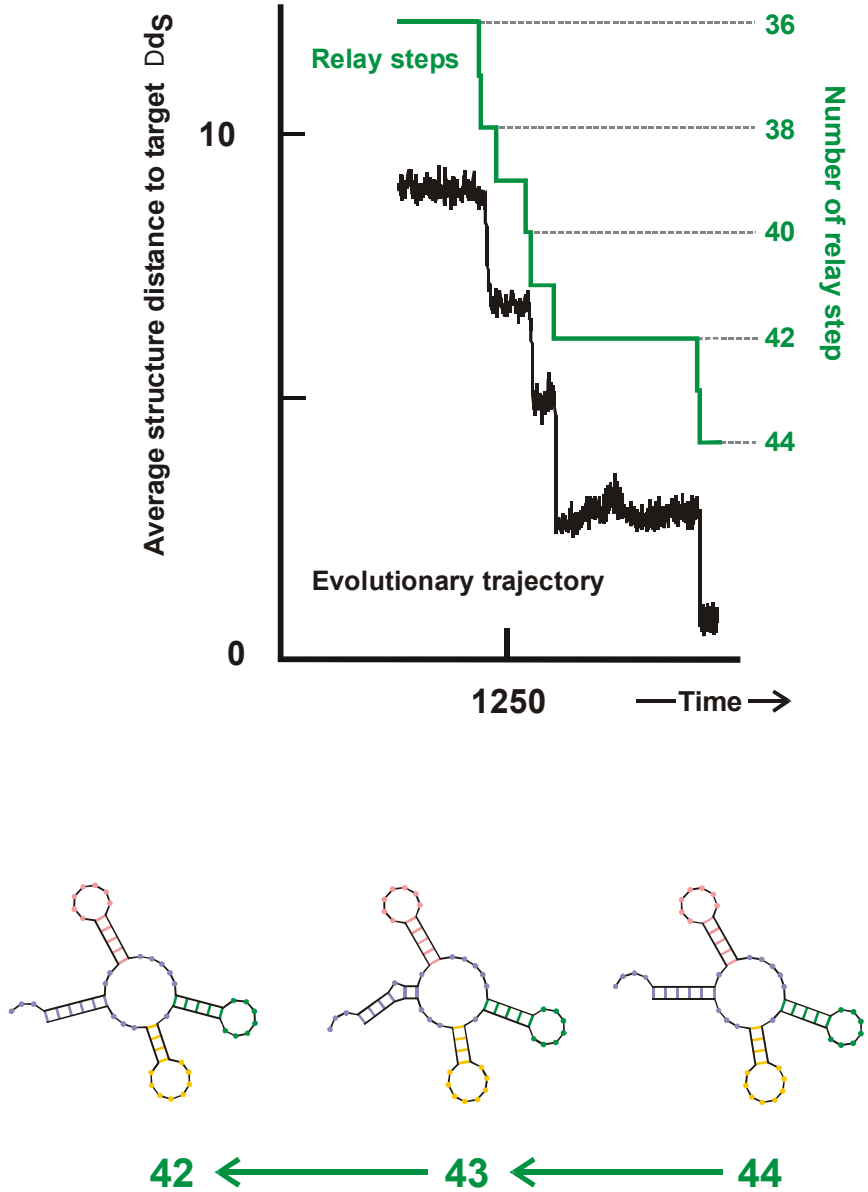
44

Endconformation of optimization

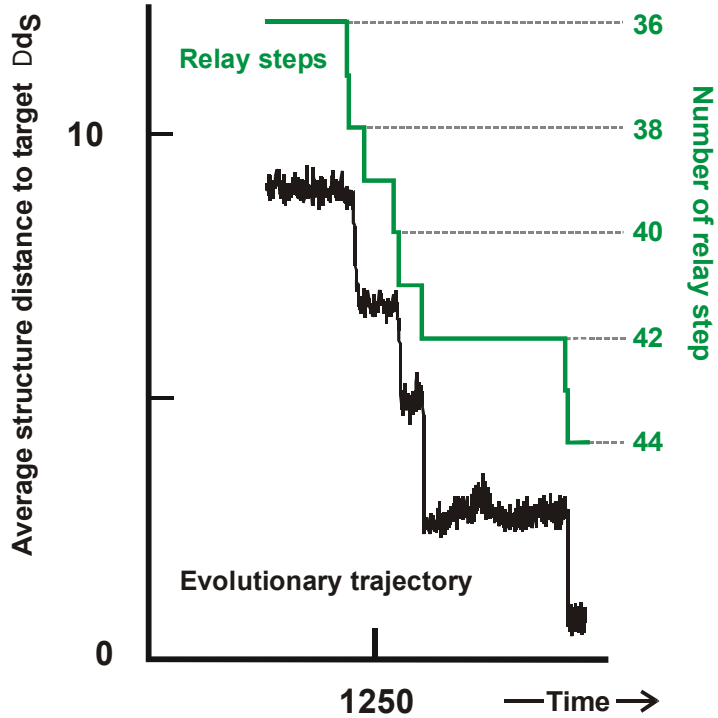


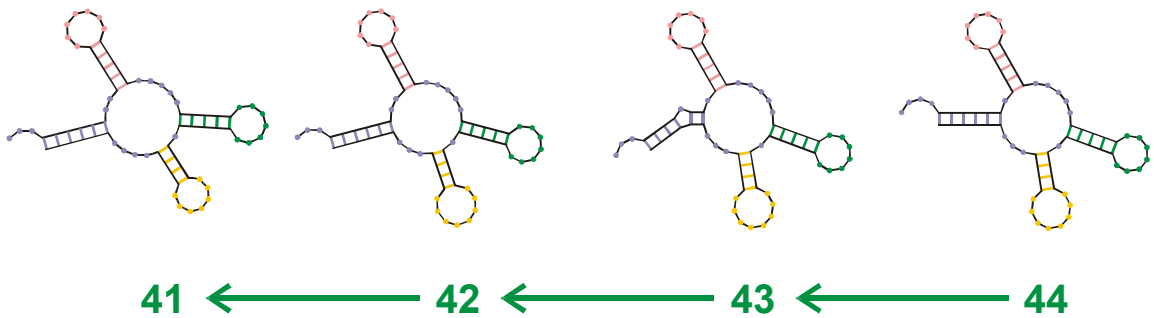


Reconstruction of the last step 43 Á 44

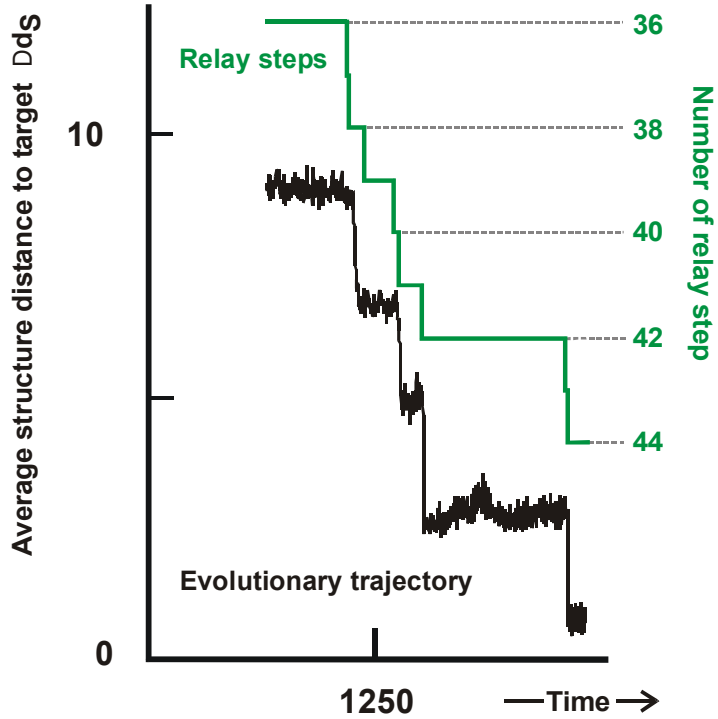


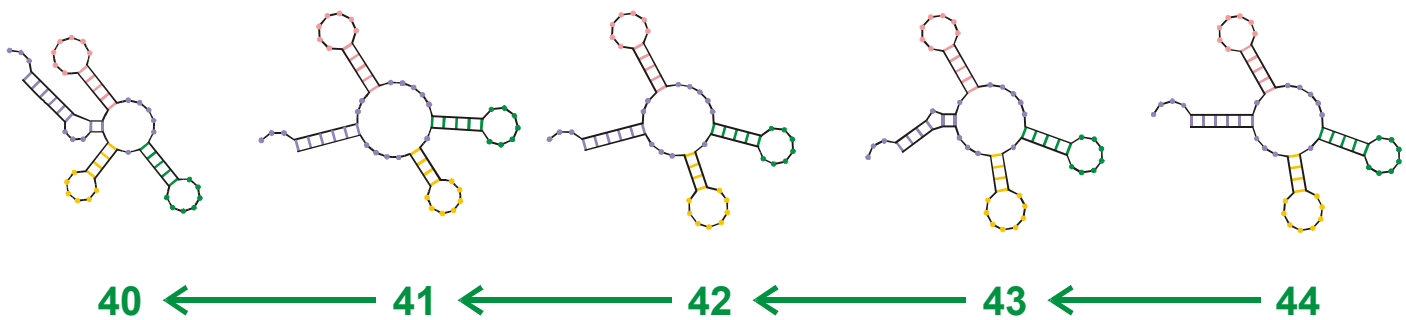
Reconstruction of last-but-one step 42 Á 43 (Á 44)



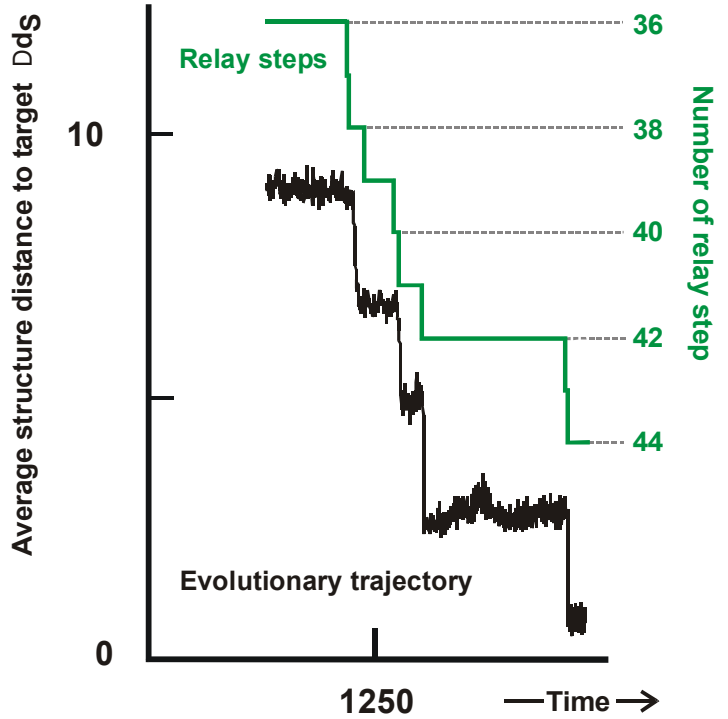


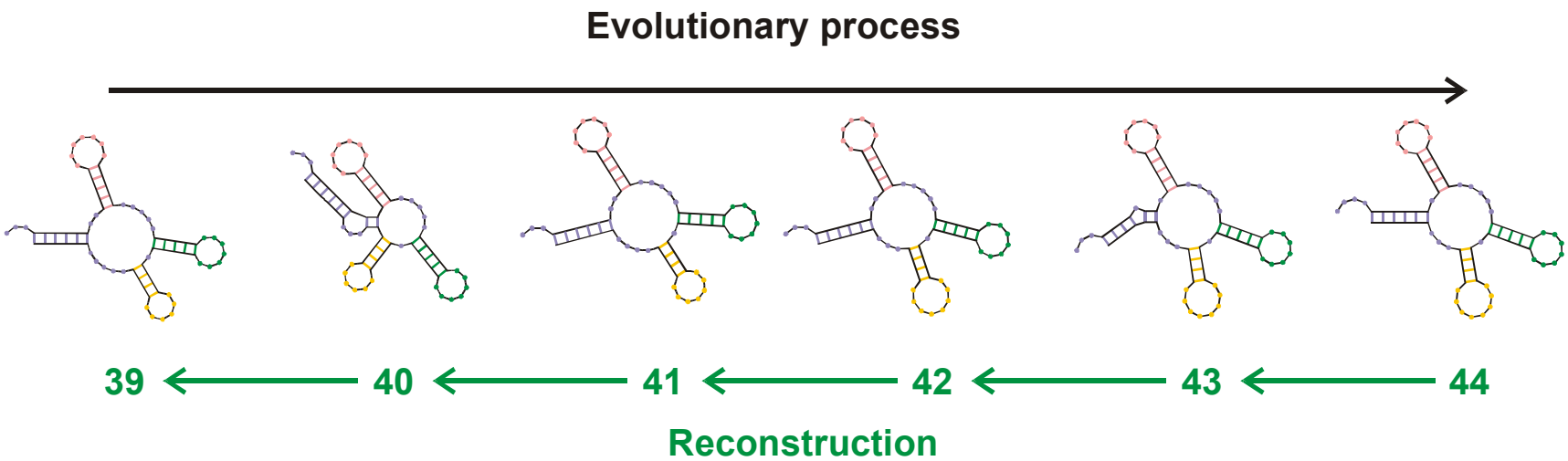
Reconstruction of step 41 Á 42 (Á 43 Á 44)



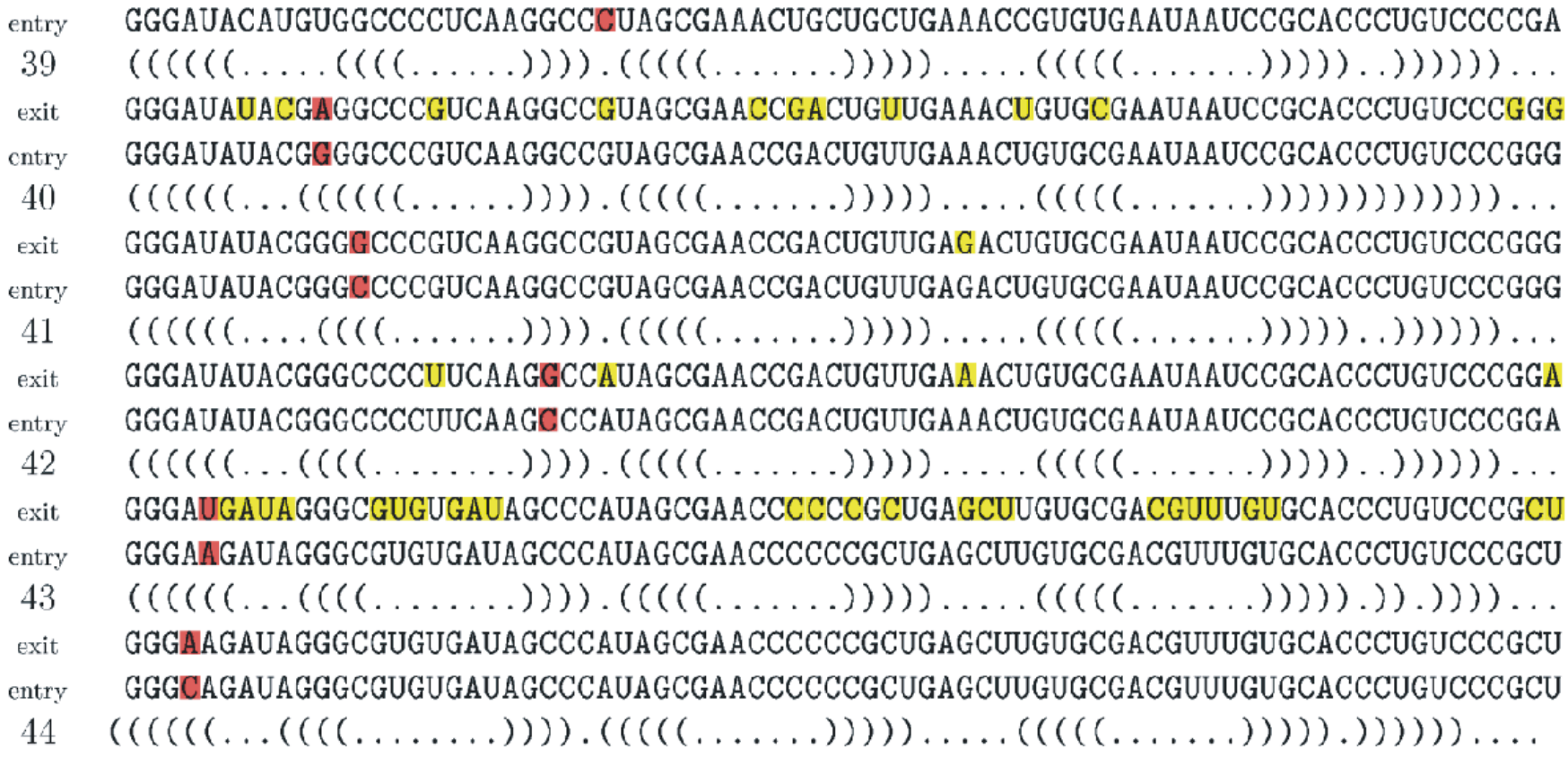


Reconstruction of step 40 Á 41 (Á 42 Á 43 Á 44)



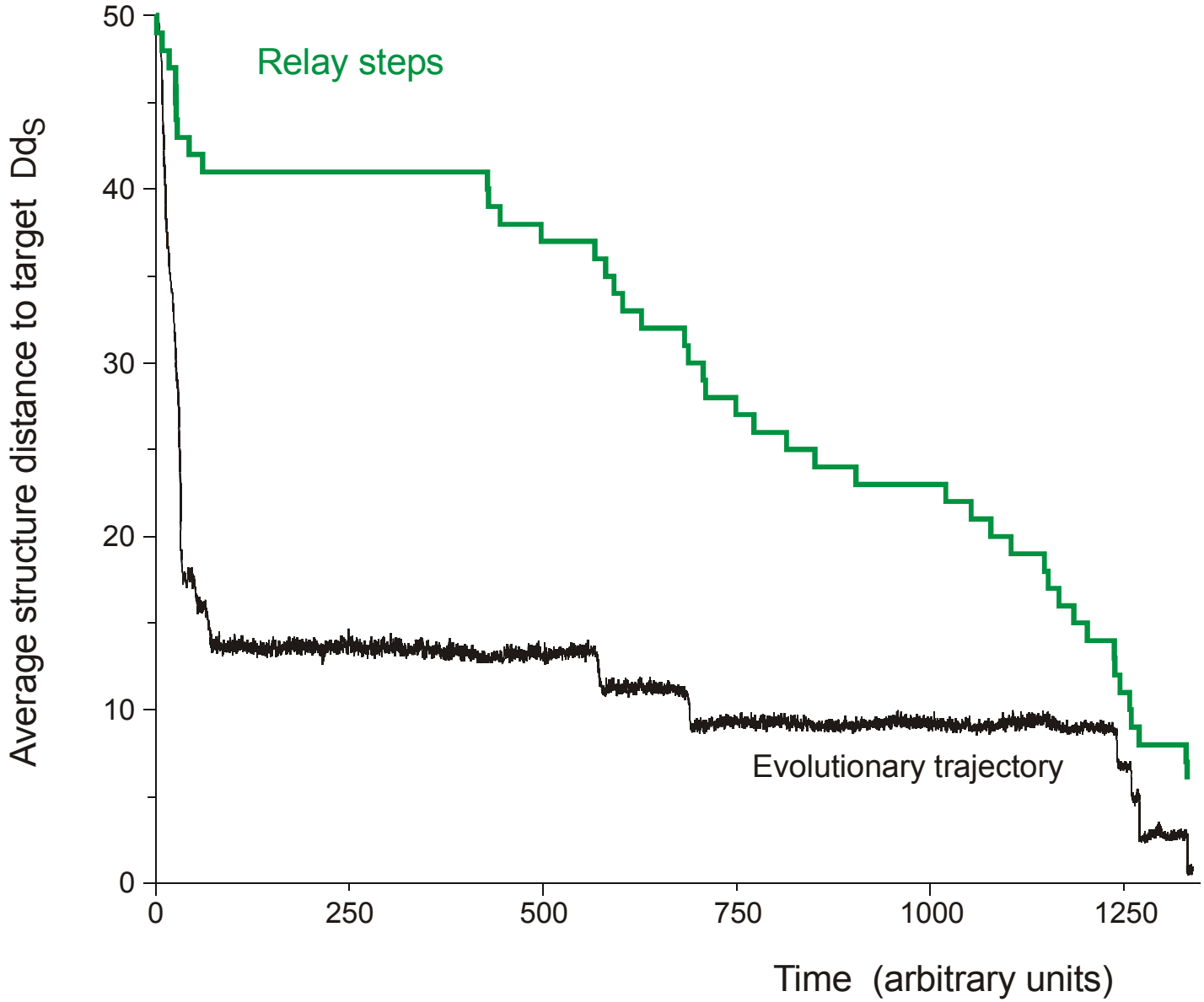


Reconstruction of the relay series



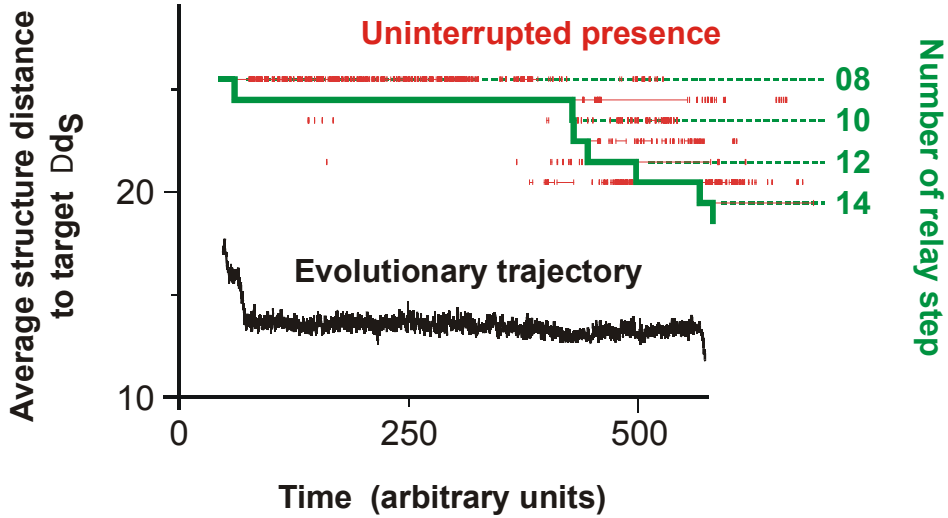
Transition inducing point mutations **Neutral point mutations**

Change in RNA sequences during the final five relay steps 39 Á 44



In silico optimization in the flow reactor: Trajectory and relay steps

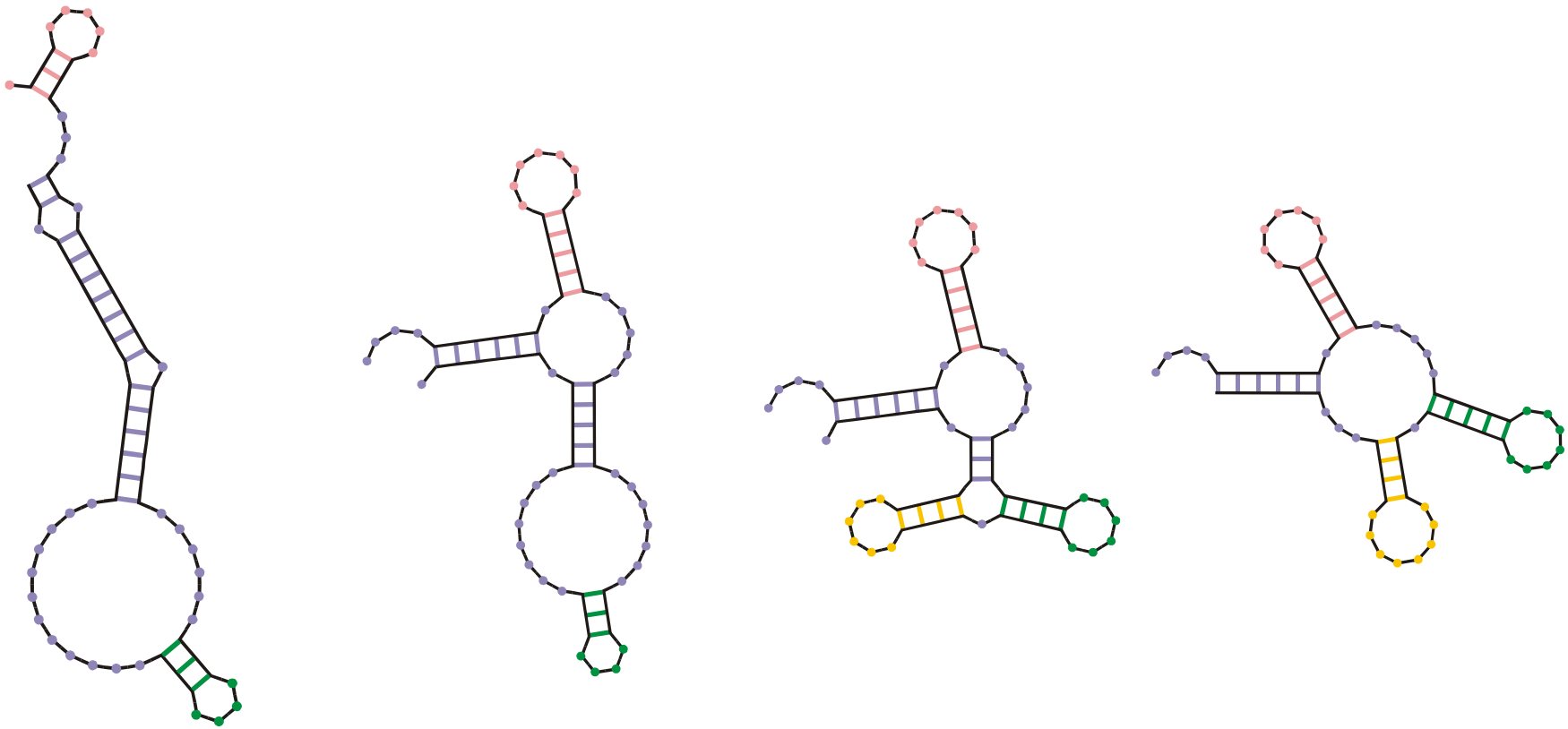
28 neutral point mutations during a long quasi-stationary epoch



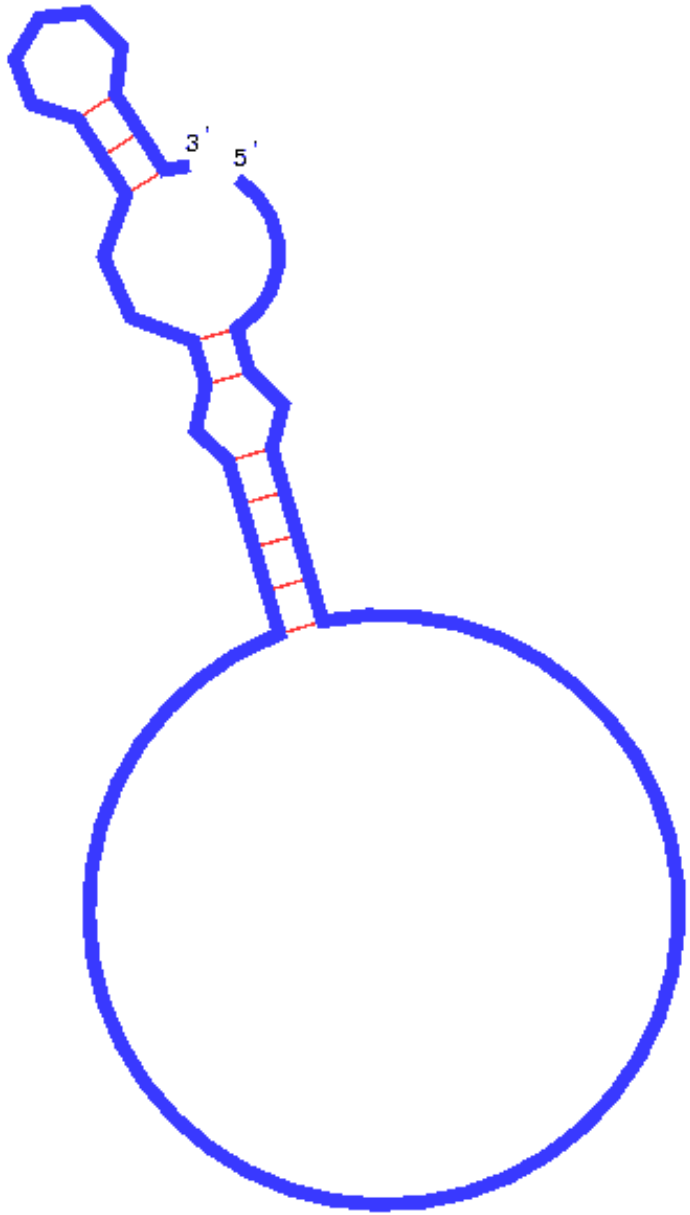

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))........)))))).....((((.........))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations — **Neutral point mutations**

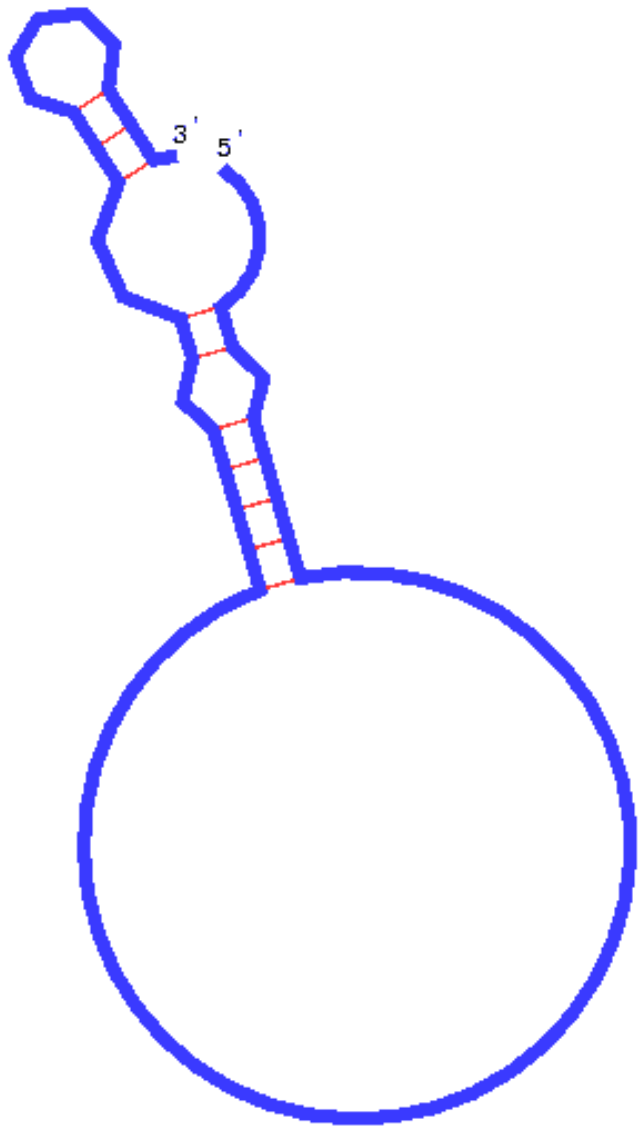
Neutral genotype evolution during phenotypic stasis



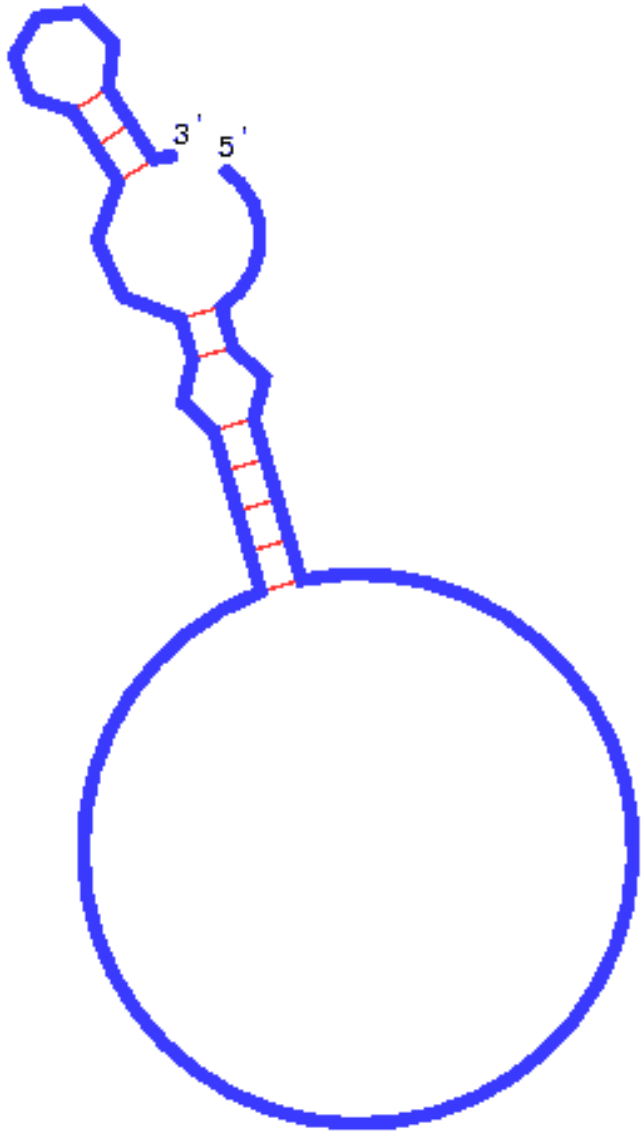
Three important steps in the formation of the tRNA clover leaf from a randomly chosen initial structure corresponding to three **main transitions**.



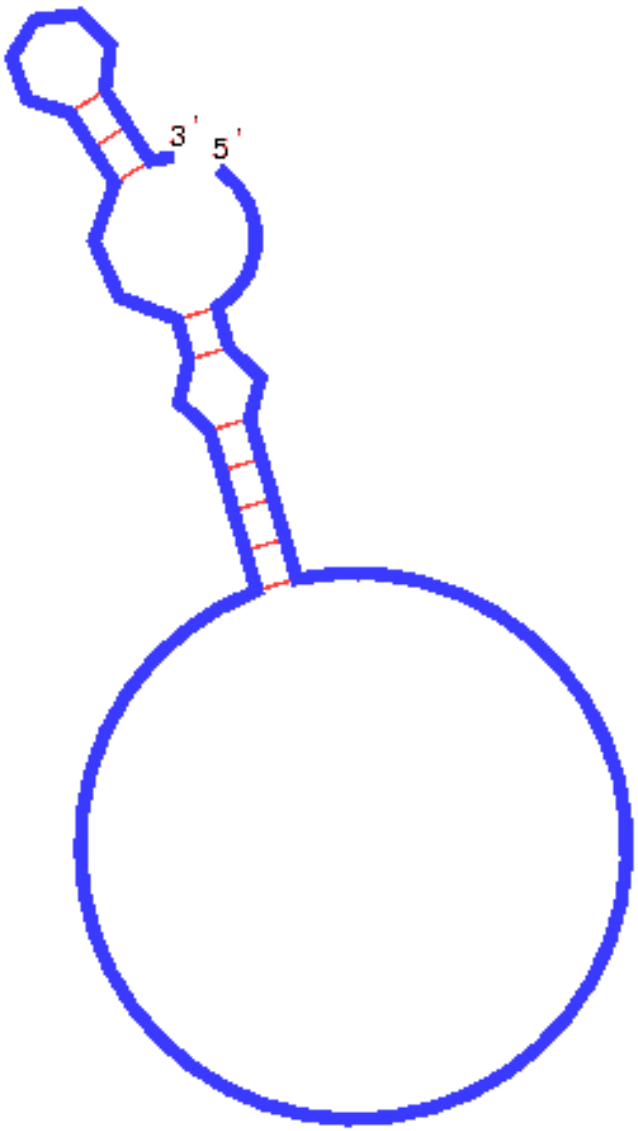
Movie of a short optimization trajectory over the **AUGC** alphabet.



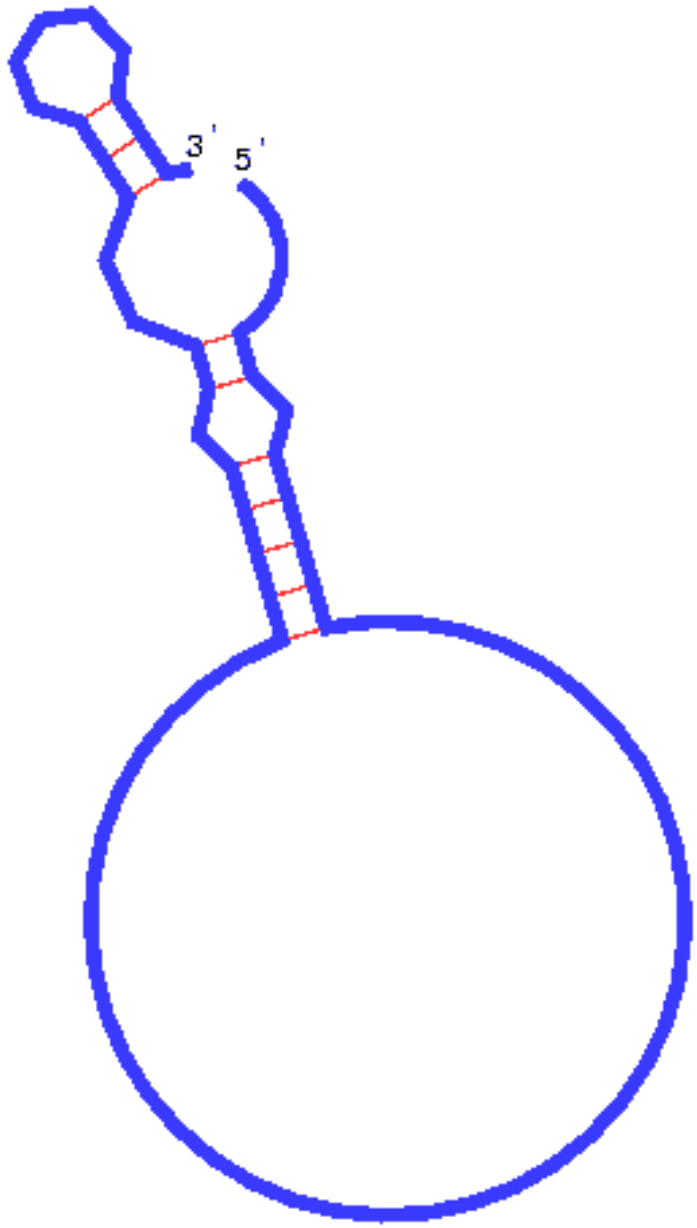
Movie of a long optimization trajectory over the **AUGC** alphabet.



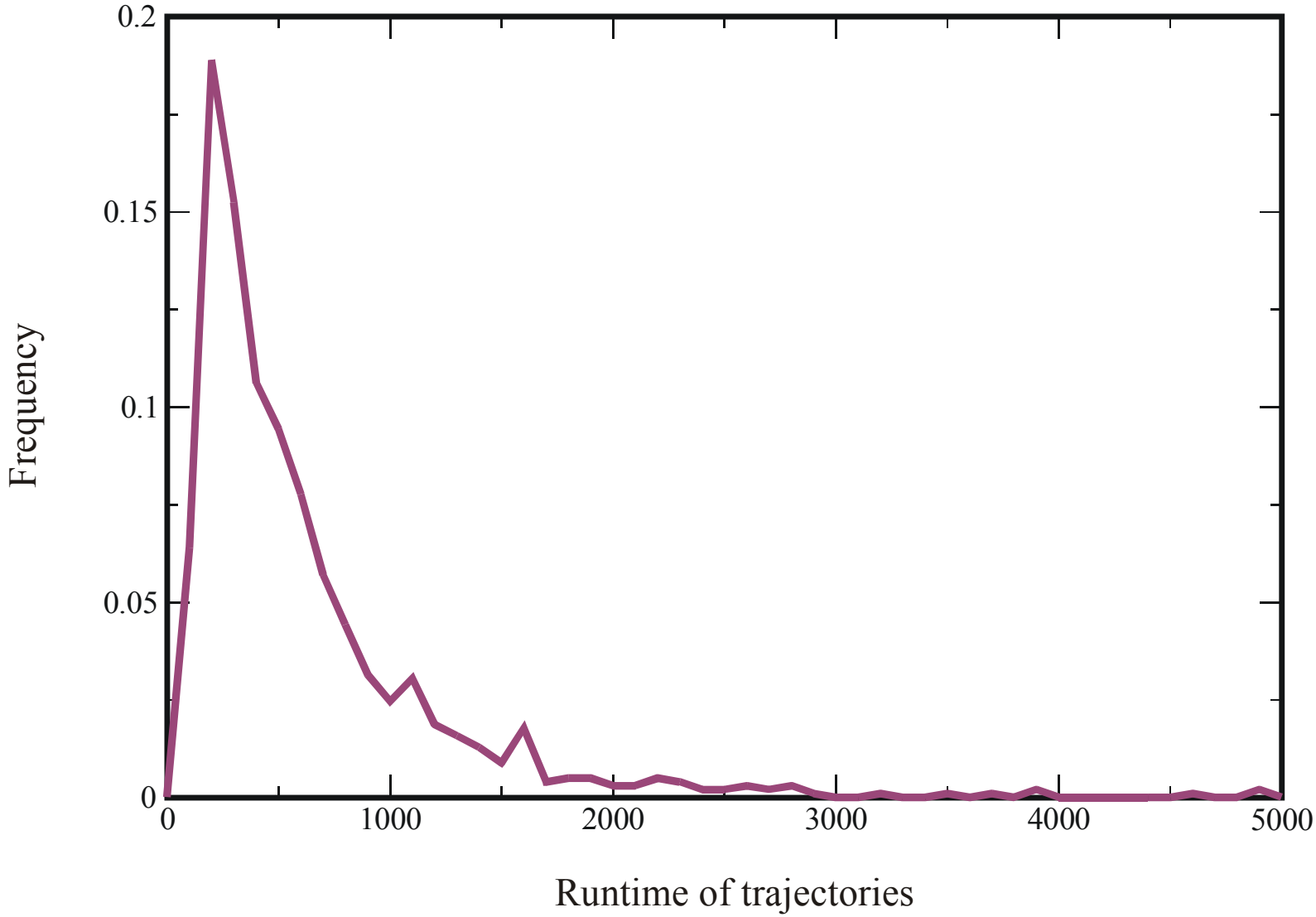
Movie of a short optimization trajectory over the **GUC** alphabet.



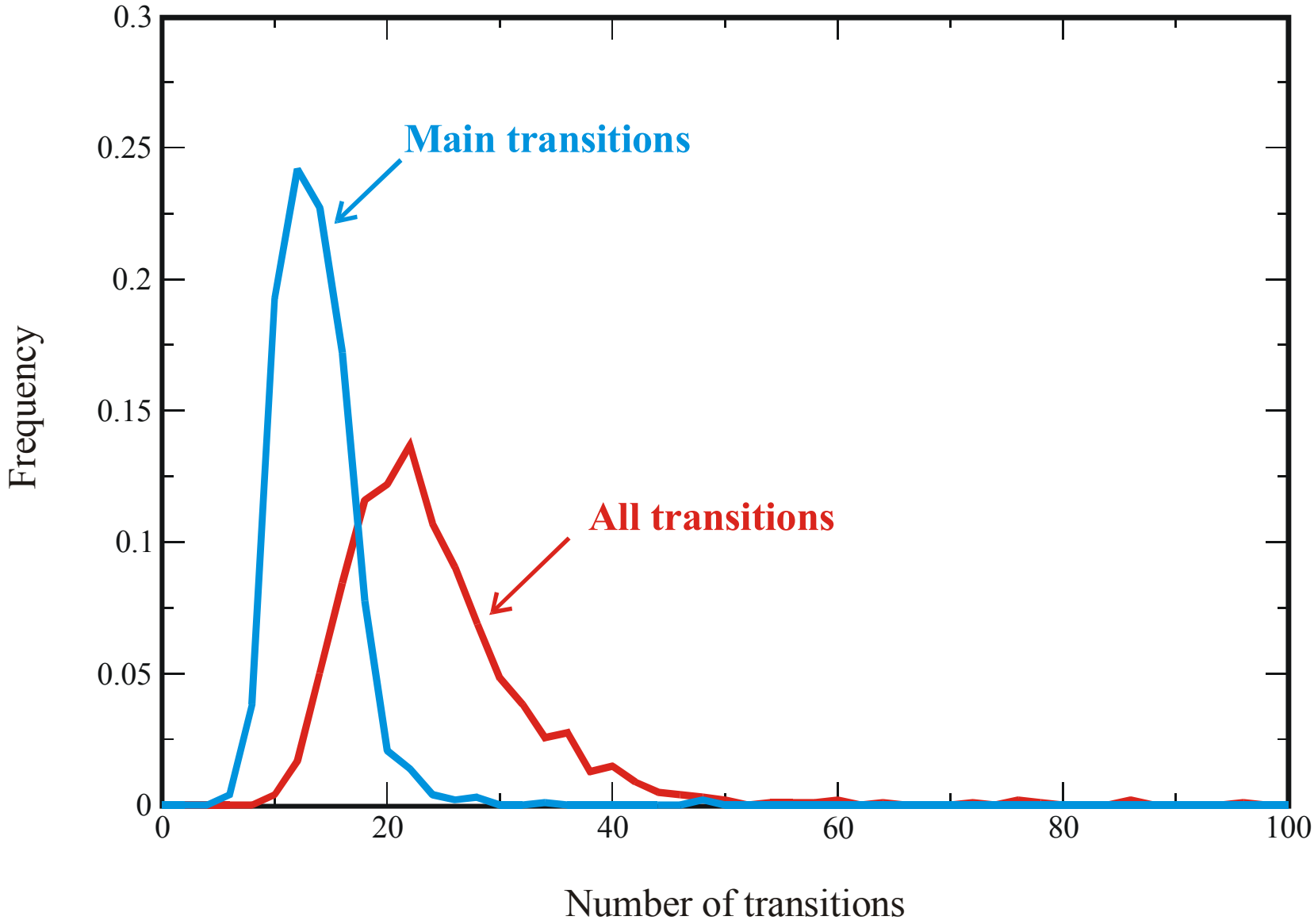
Movie of a short optimization trajectory over the **GC** alphabet.



Movie of a long optimization trajectory over the **GC** alphabet.



Statistics of the lengths of trajectories from initial structure to target (**AUGC**-sequences)

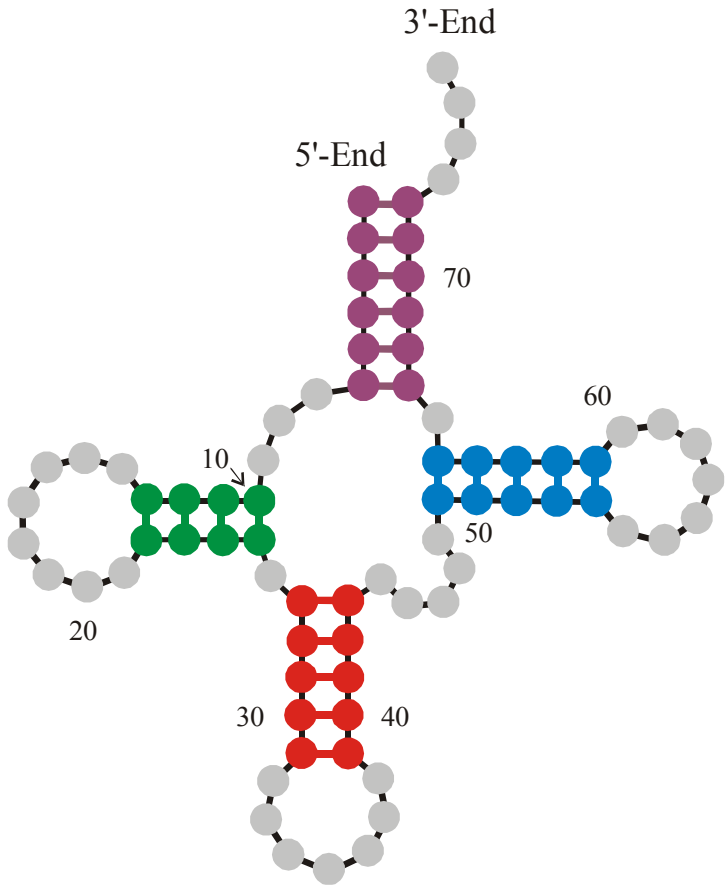


Statistics of the numbers of transitions from initial structure to target (**AUGC**-sequences)

Alphabet	Runtime	Transitions	Main transitions	No. of runs
AUGC	385.6	22.5	12.6	1017
GUC	448.9	30.5	16.5	611
GC	2188.3	40.0	20.6	107

Statistics of trajectories and relay series (mean values of log-normal distributions)

Stable tRNA clover leaf structures built from binary, **GC**-only, sequences exist. The corresponding sequences are found through inverse folding. Optimization by mutation and selection in the flow reactor turned out to be a hard problem.



The neutral network of the tRNA clover leaf in **GC** sequence space is not connected, whereas to the corresponding neutral network in **AUGC** sequence space is close to the connectivity threshold, $\mathbf{l}_{\mathbf{cr}}$. Here, both inverse folding and optimization in the flow reactor are more effective than with **GC** sequences.

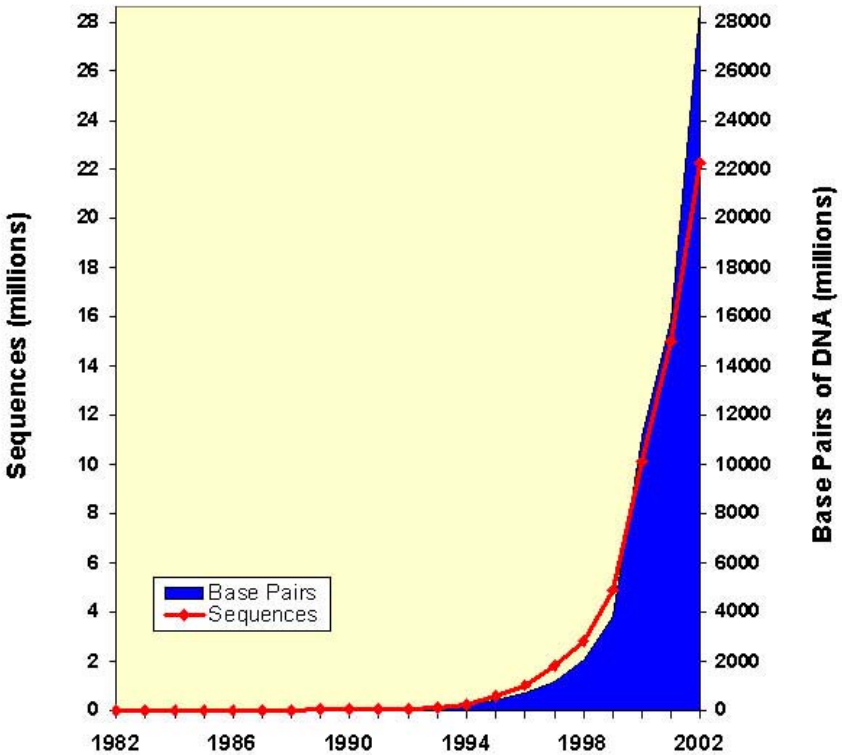
The hardness of optimization depends on the connectivity of neutral networks.

1. Experiments on controlled evolution and RNA replication

2. Sequence-structure maps, neutral networks, and intersections

3. Optimization in the RNA model

4. What we can learn from molecules for evolution proper



Source: NCBI

Fully sequenced genomes

- *Organisms* 751 projects

153 complete (16 A, 118 B, 19 E)

(*Eukarya* examples: mosquito (pest, malaria), sea squirt, mouse, yeast, homo sapiens, arabidopsis, fly, worm, …)

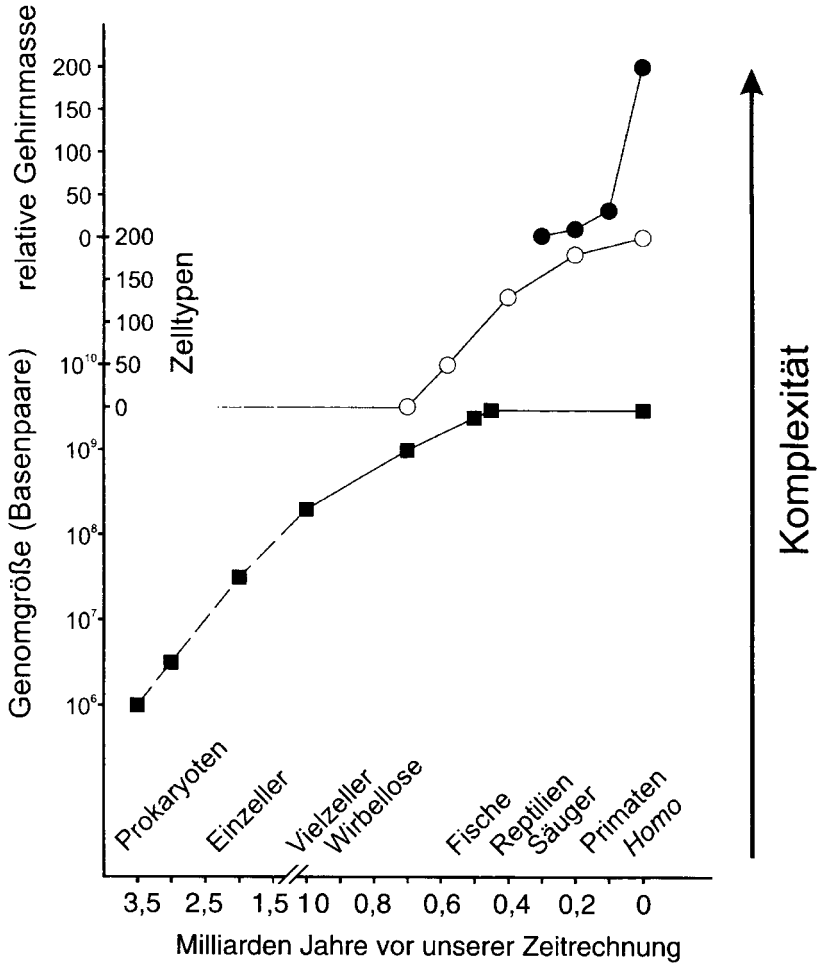
598 ongoing (23 A, 332 B, 243 E)

(*Eukarya* examples: chimpanzee, turkey, chicken, ape, corn, potato, rice, banana, tomato, cotton, coffee, soybean, pig, rat, cat, sheep, horse, kangaroo, dog, cow, bee, salmon, fugu, frog, …)

- *Other structures with genetic information*

68 phages
1328 viruses
35 viroids
472 organelles (423 mitochondria, 32 plastids, 14 plasmids, 3 nucleomorphs)

Source: Integrated Genomics, Inc.
August 12th, 2003



4.10 Die Zunahme der Komplexität ist ein wesentlicher Aspekt der biologischen Evolution, wobei höhere Komplexität sowohl durch Vergrößerung der Zahl von miteinander in Wechselwirkung stehenden Elementen als auch durch Differenzierung der Funktionen dieser Elemente entstehen kann. In dieser Abbildung wird zwischen drei Phasen oder Strategien der Evolution von Komplexität unterschieden. *Untere Kurve*: Zunahme der Genomgröße; logarithmische Auftragung der Zahl der Basenpaare im Genom von Zellen seit Beginn der biologischen Evolution (Daten aus Abbildung 2.3). *Mittlere Kurve*: Zunahme der Zahl der Zelltypen in der Evolution der Metazoa (Daten aus Abbildung 4.8). *Obere Kurve*: Zunahme des relativen Gehirngewichts (bezogen auf die Körperoberfläche) bei Säugetieren (Daten aus Wilson 1985). Für die Abszisse wurden zwei Skaleneinteilungen verwendet, eine für den Zeitraum $>10^9$ Jahre, eine andere für den Zeitraum $<10^9$ Jahre vor der Gegenwart. Oberhalb der Abszisse sind die Namen einiger wichtiger taxonomischer Einheiten angeführt, deren Evolution in etwa beim jeweiligen Wortbeginn einsetzt.

Wolfgang Wieser. Die Erfindung der Individualität oder die zwei Gesichter der Evolution. Spektrum Akademischer Verlag, Heidelberg 1998.

A.C.Wilson. The Molecular Basis of Evolution. Scientific American, Oct.1985, 164-173.

Evolution (cartoon Î 1980)

Acknowledgement of support

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

The Santa Fe Institute and the Universität Wien

The software for producing RNA movies was developed by
Robert Giegerich and coworkers at the Universität Bielefeld

Universität Wien

Universität Wien

Coworkers

Walter Fontana, Santa Fe Institute, NM

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

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

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 Svrček-Seiler, Stefan Wuchty

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