Some Mathematical Challenges from Molecular Biology

Part II

Peter Schuster

Institut für Theoretische Chemie und Molekulare
Strukturbiologie der Universität Wien

Mathematisches Kolloquium

Zürich, 11.11.2003

Web-Page for further information:

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

1. Prolog – Mathematics and the life sciences in the 21st century

2. Replication kinetics of RNA molecules and evolution

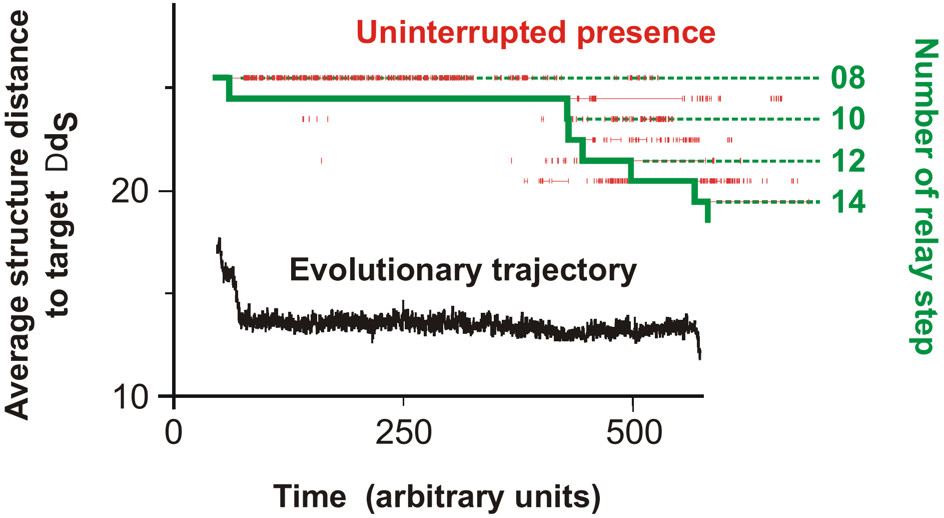
3. **RNA evolution *in silico***

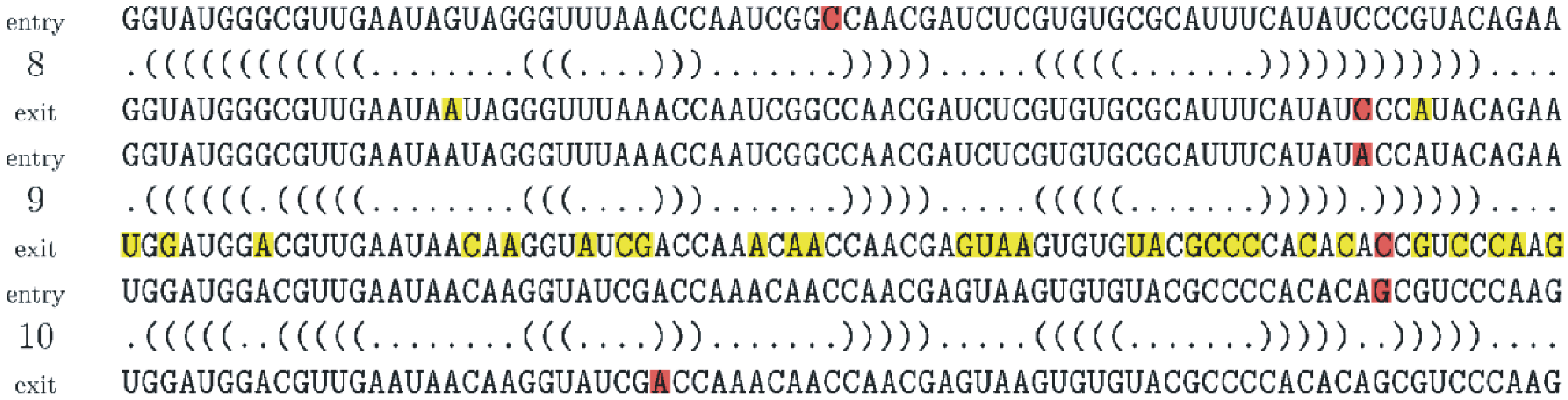
4. Sequence-structure maps, neutral networks, and intersections

5. Reference to experimental data

6. Summary

28 neutral point mutations during a long quasi-stationary epoch



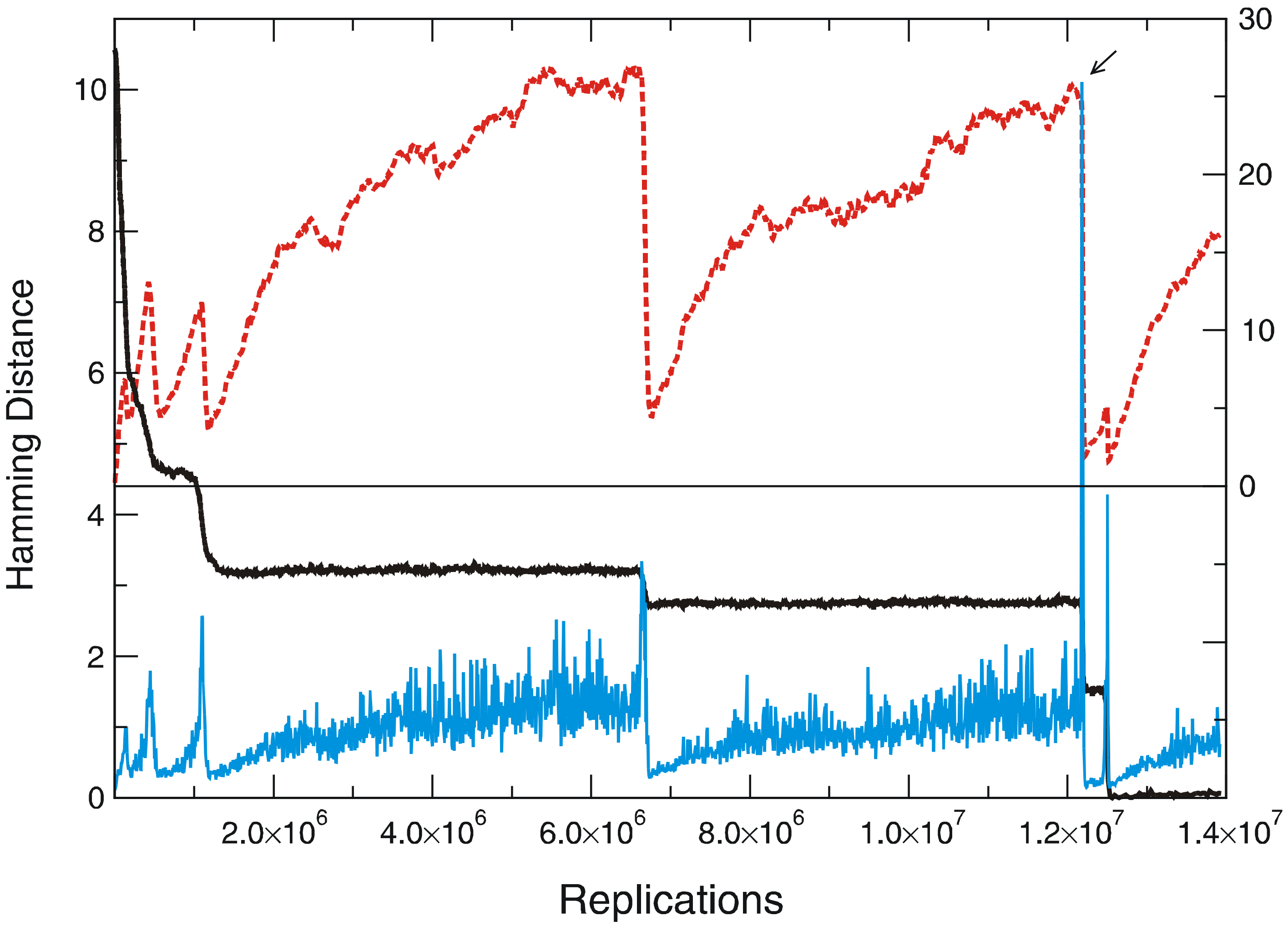


```
entry  GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8      .((((((((((((........(((....))).......))))).....(((((.......))))))))))))....
exit   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry  GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations

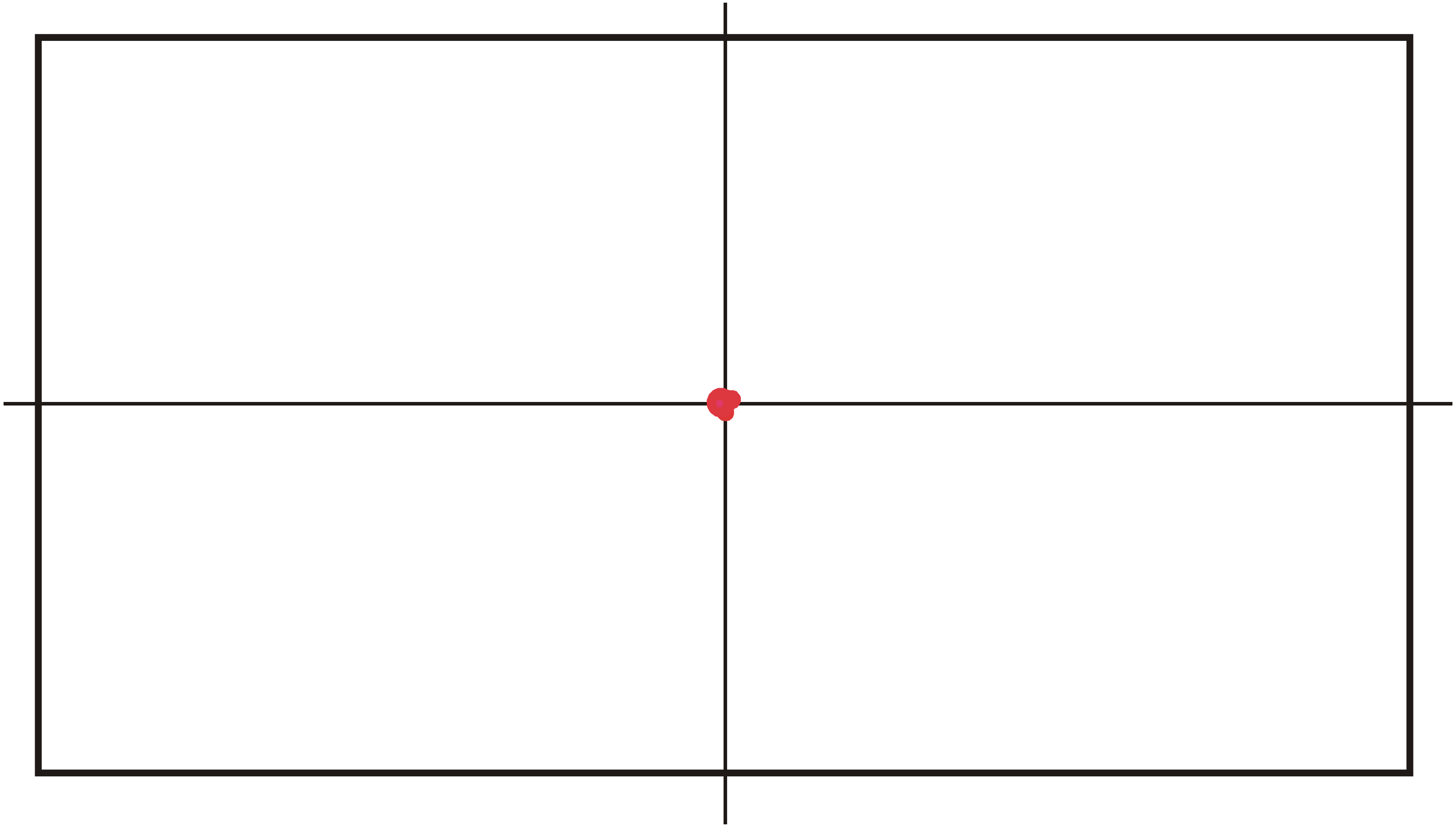
Neutral point mutations

Neutral genotype evolution during phenotypic stasis

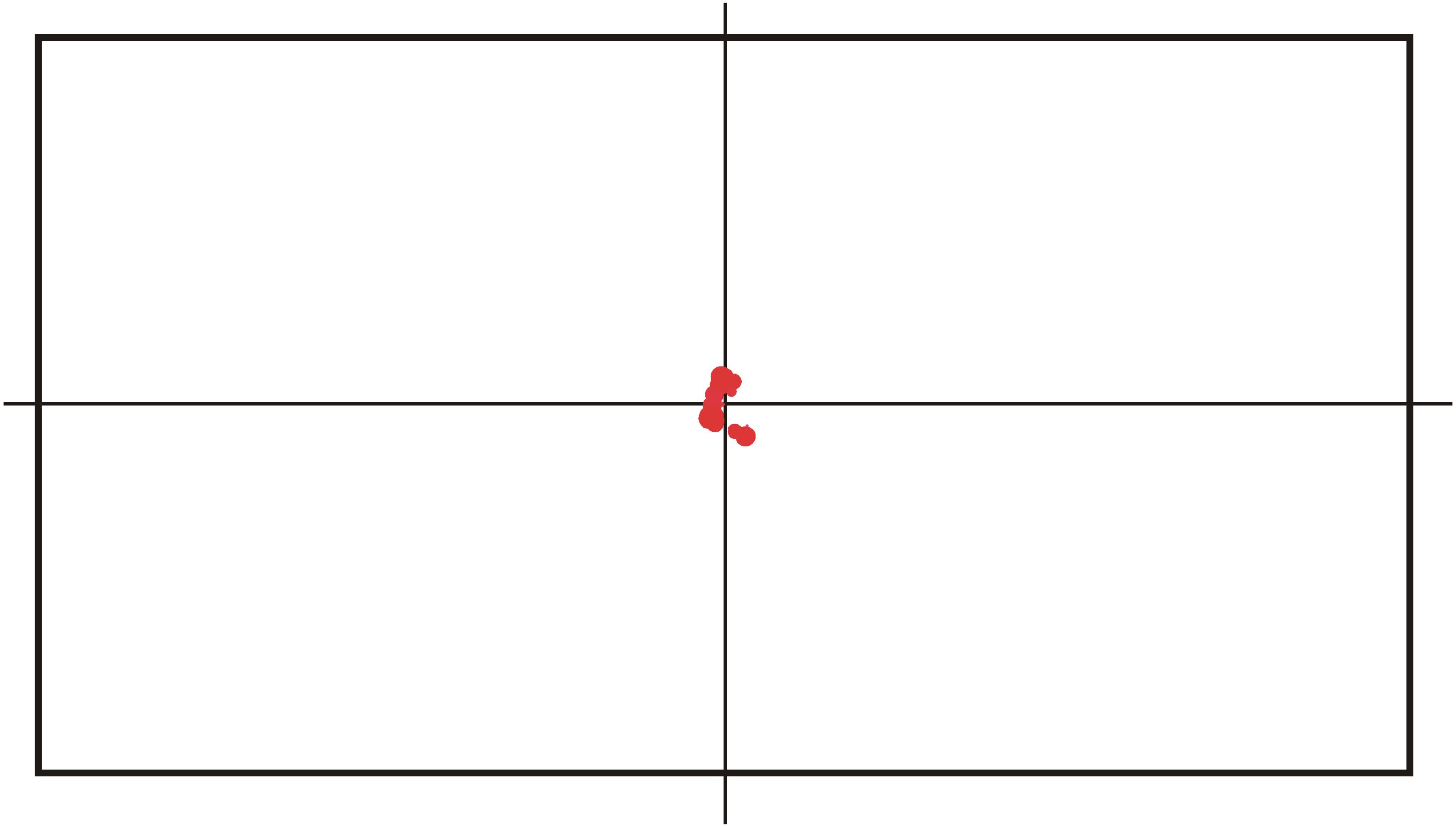


Variation in genotype space during optimization of phenotypes

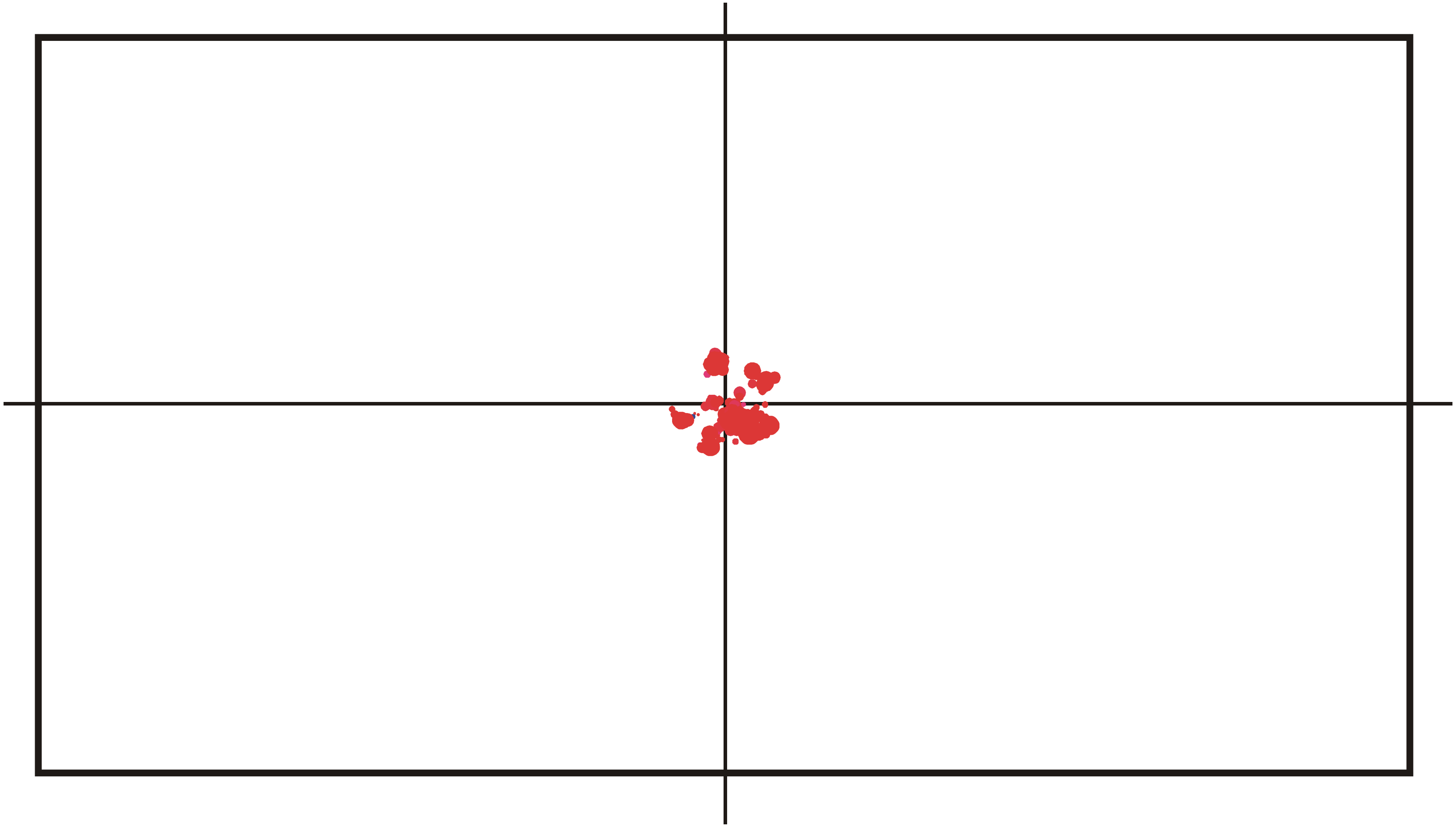
Mean Hamming distance within the population and **drift velocity of the population center** in sequence space.



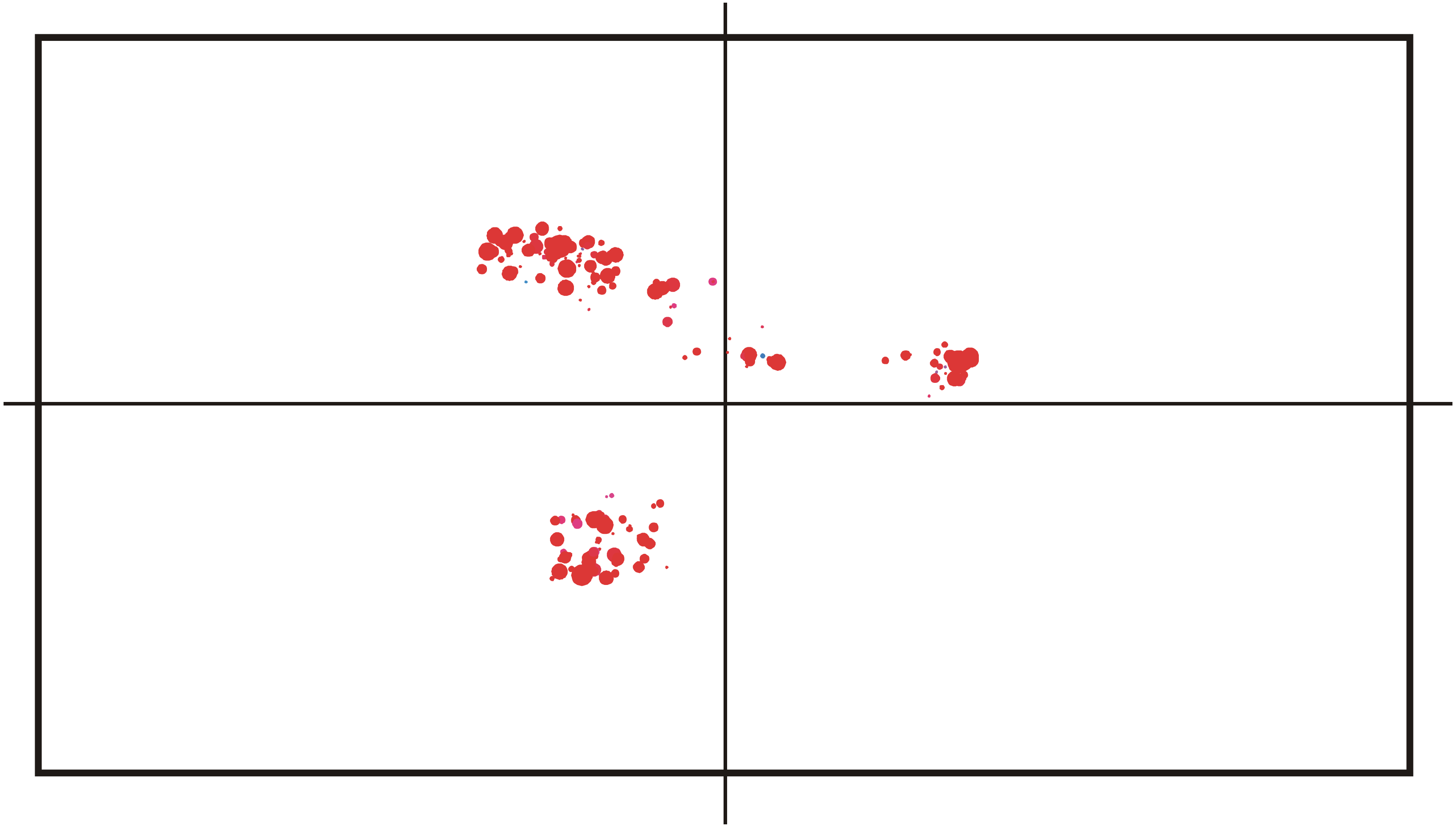
Spread of population in sequence space during a quasistationary epoch: t = 150



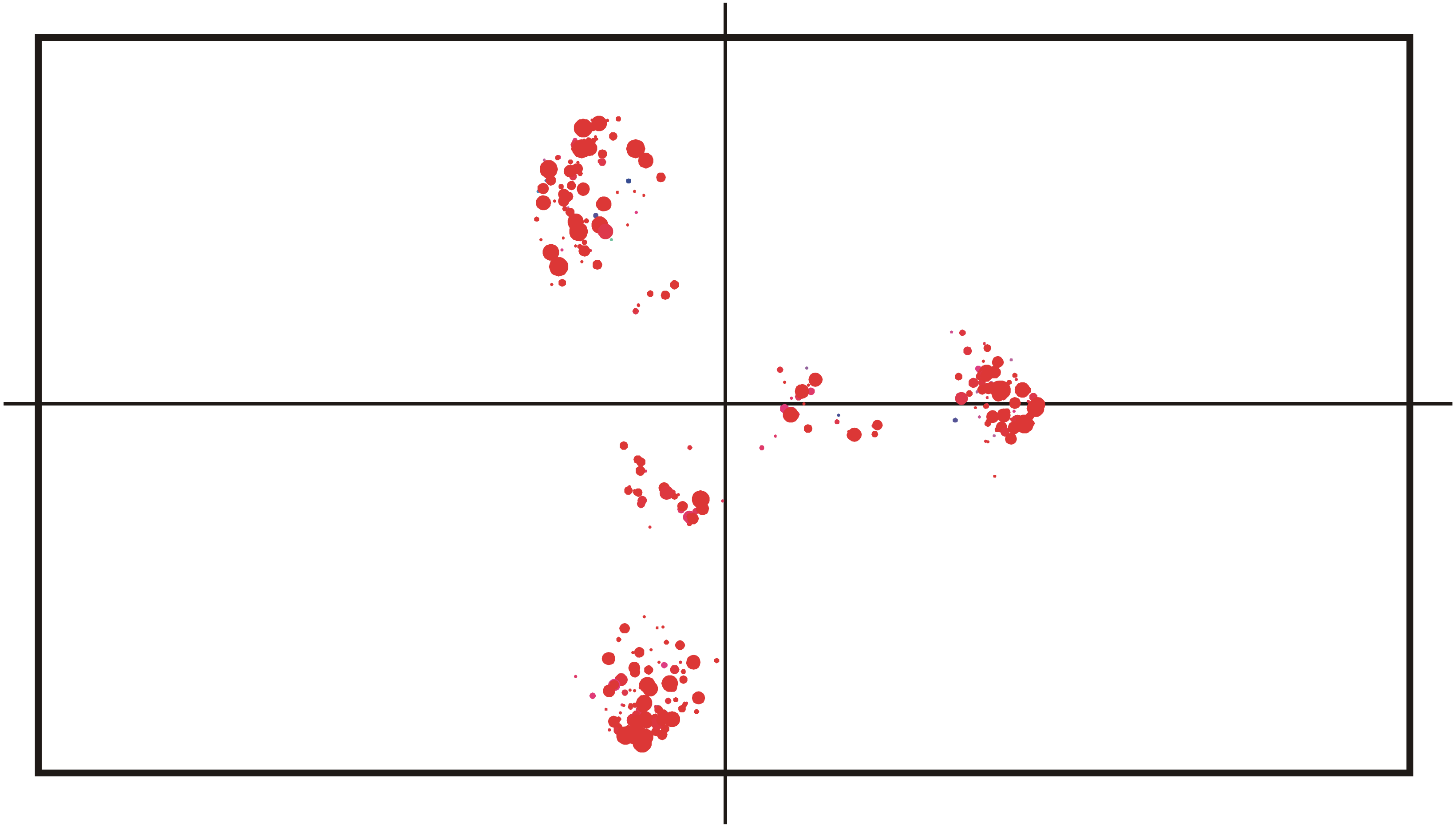
Spread of population in sequence space during a quasistationary epoch: t = 170



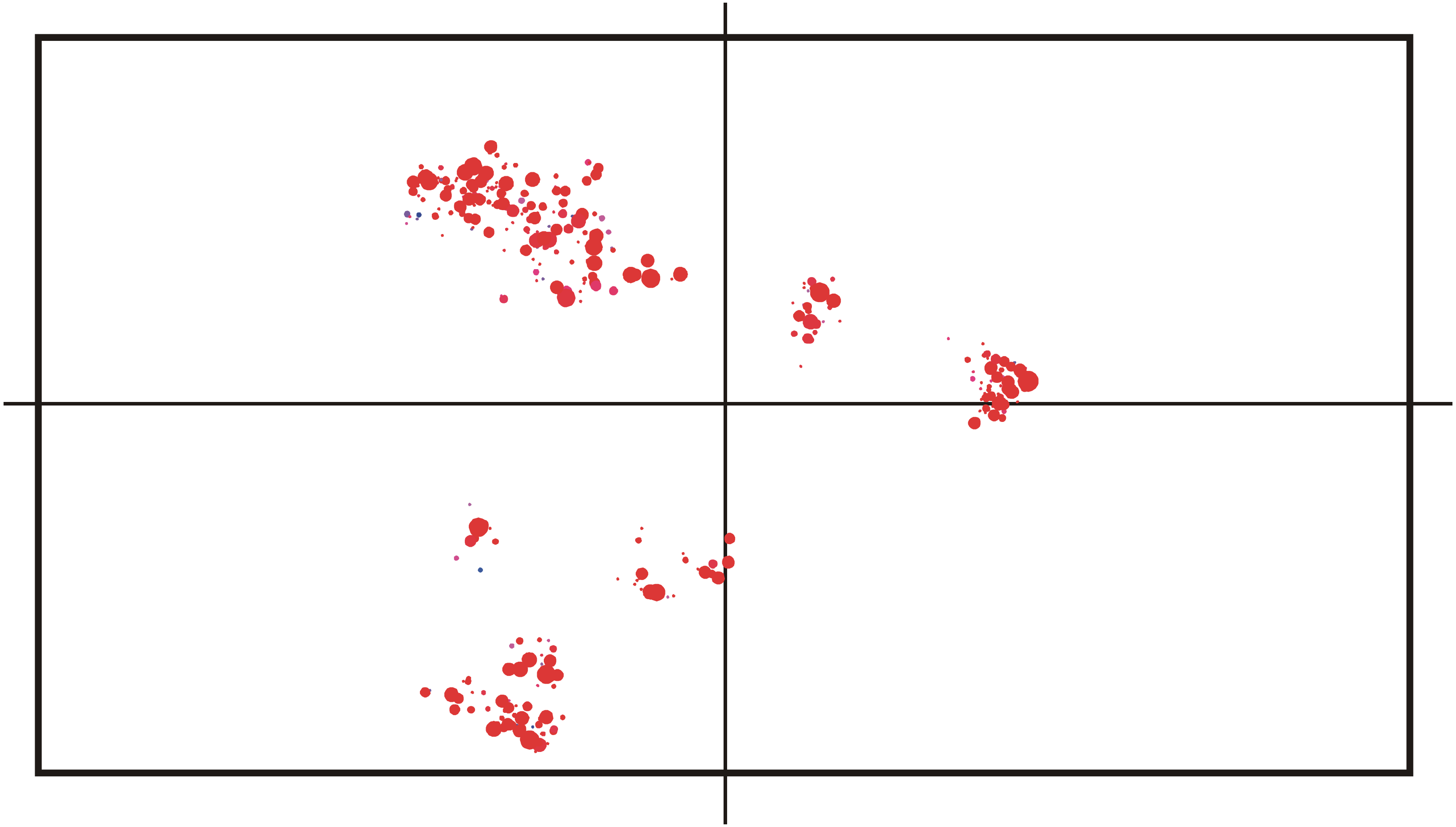
Spread of population in sequence space during a quasistationary epoch: t = 200



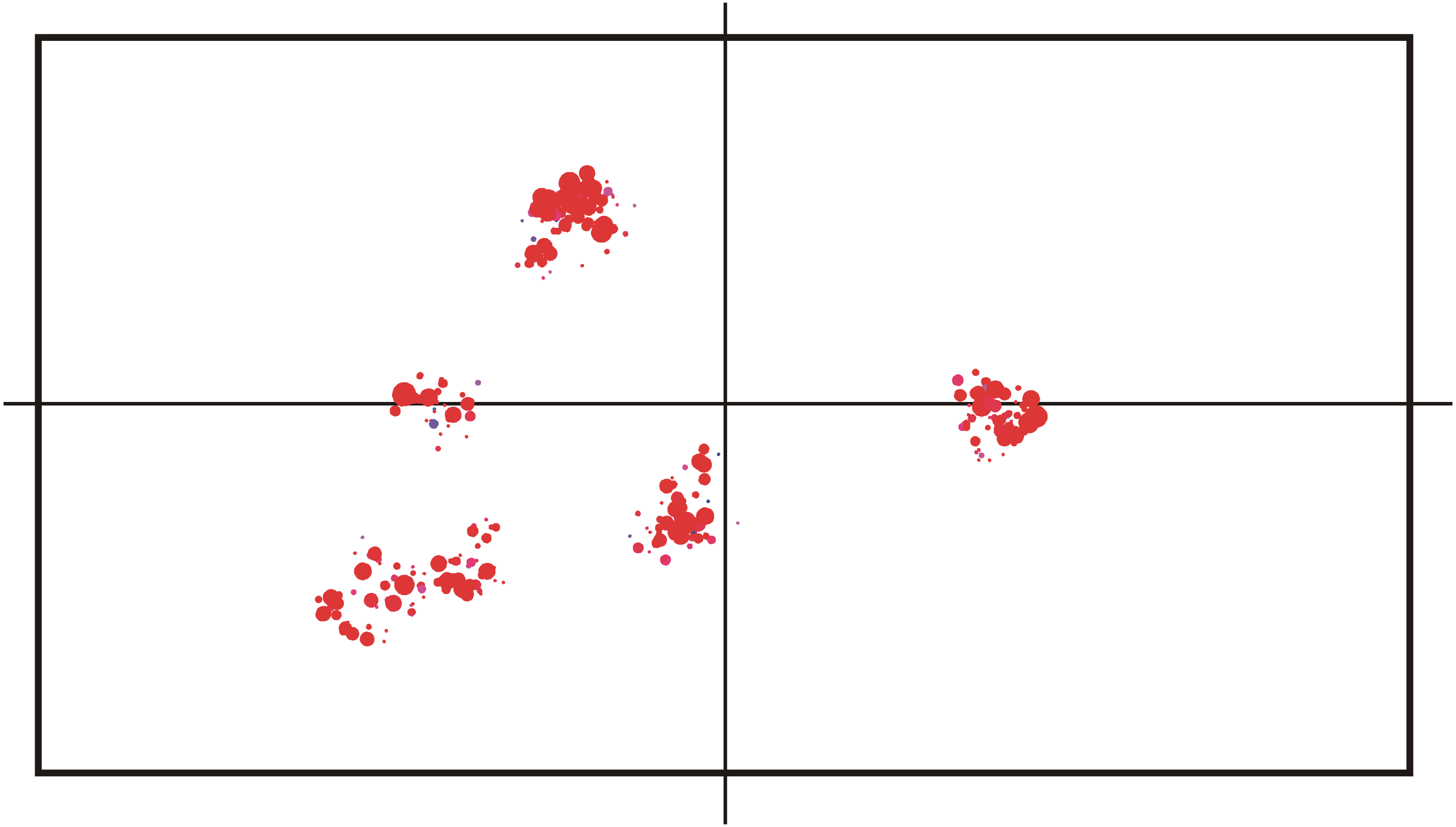
Spread of population in sequence space during a quasistationary epoch: t = 350



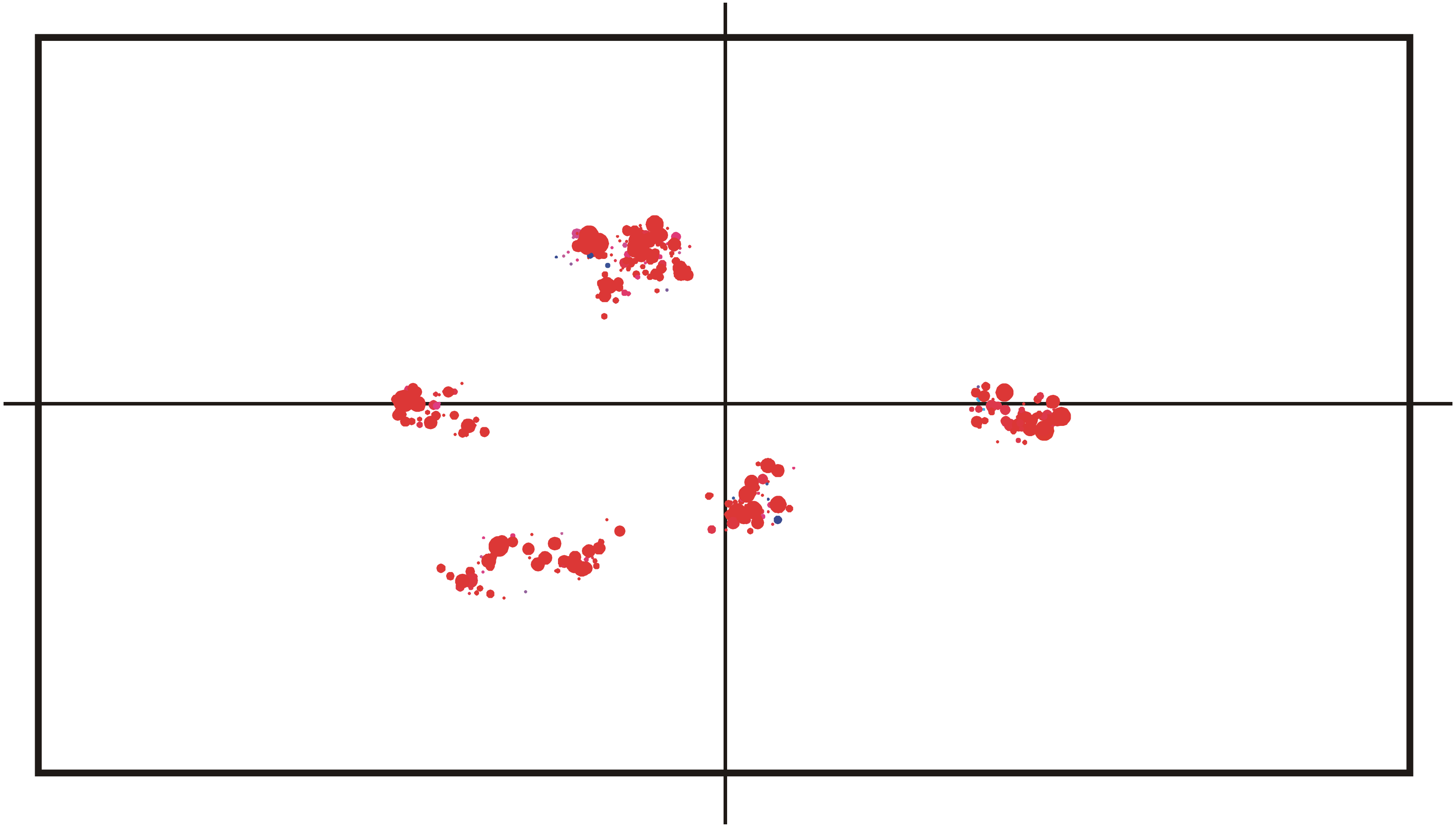
Spread of population in sequence space during a quasistationary epoch: t = 500



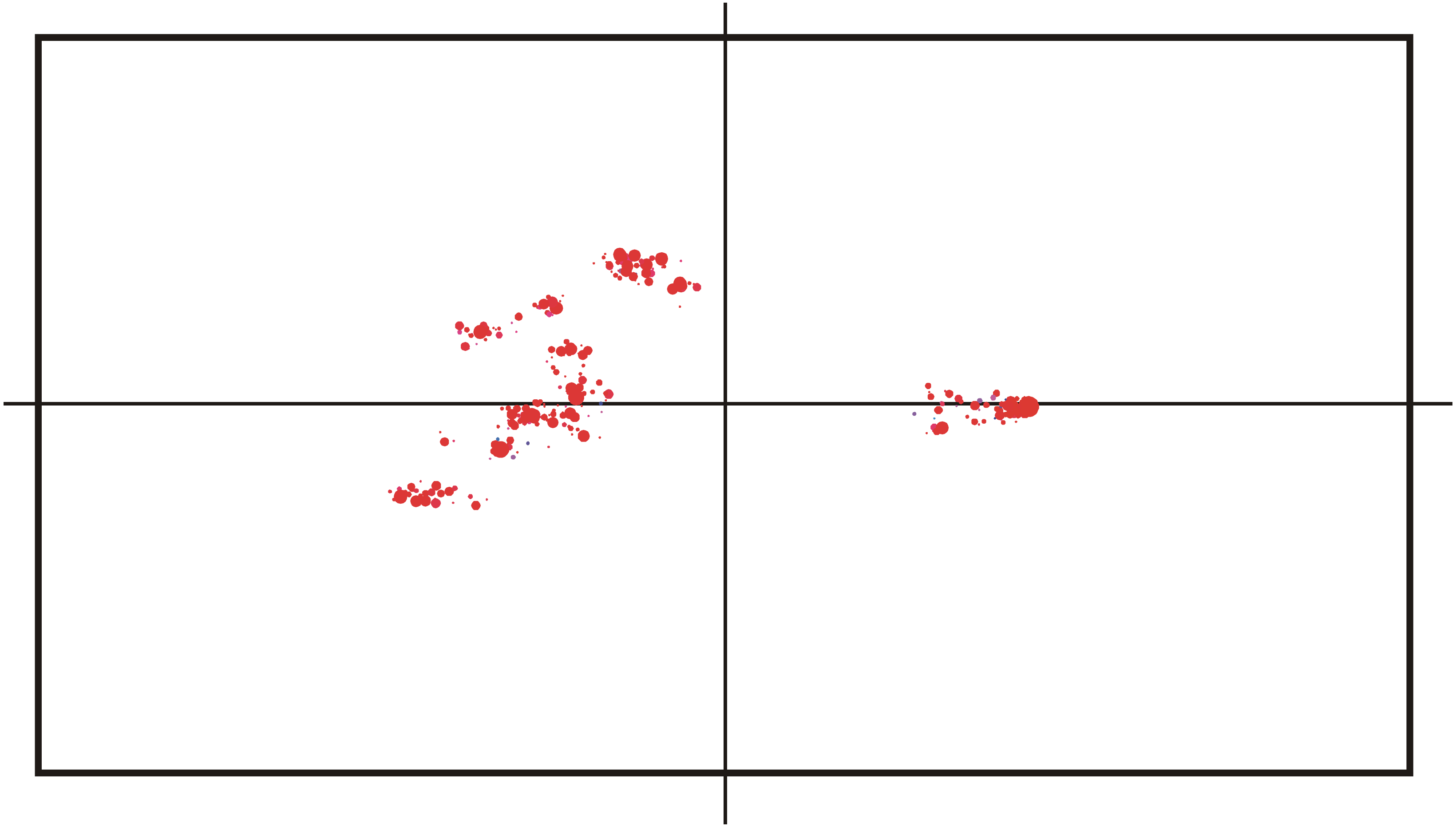
Spread of population in sequence space during a quasistationary epoch: t = 650



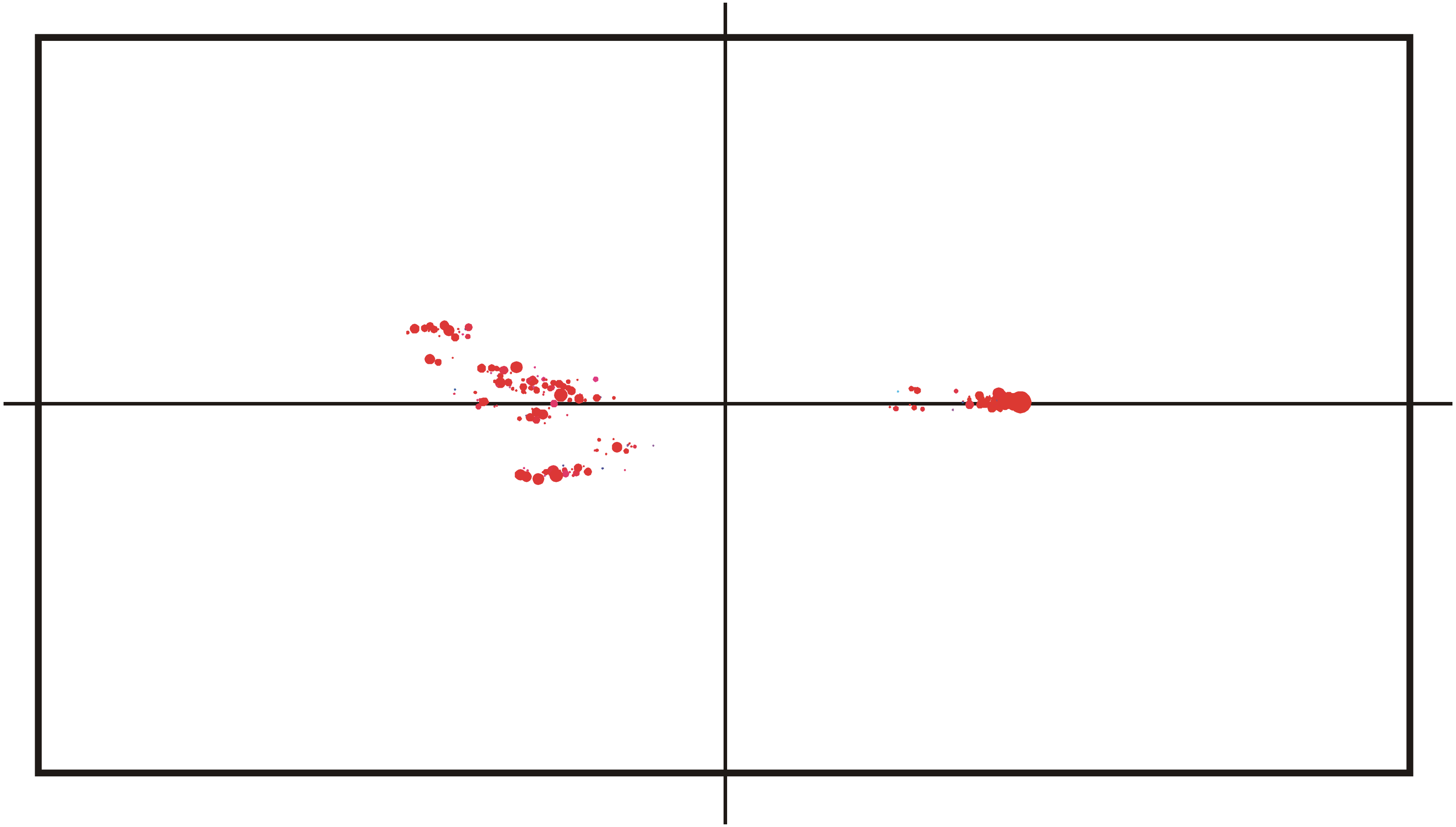
Spread of population in sequence space during a quasistationary epoch: t = 820



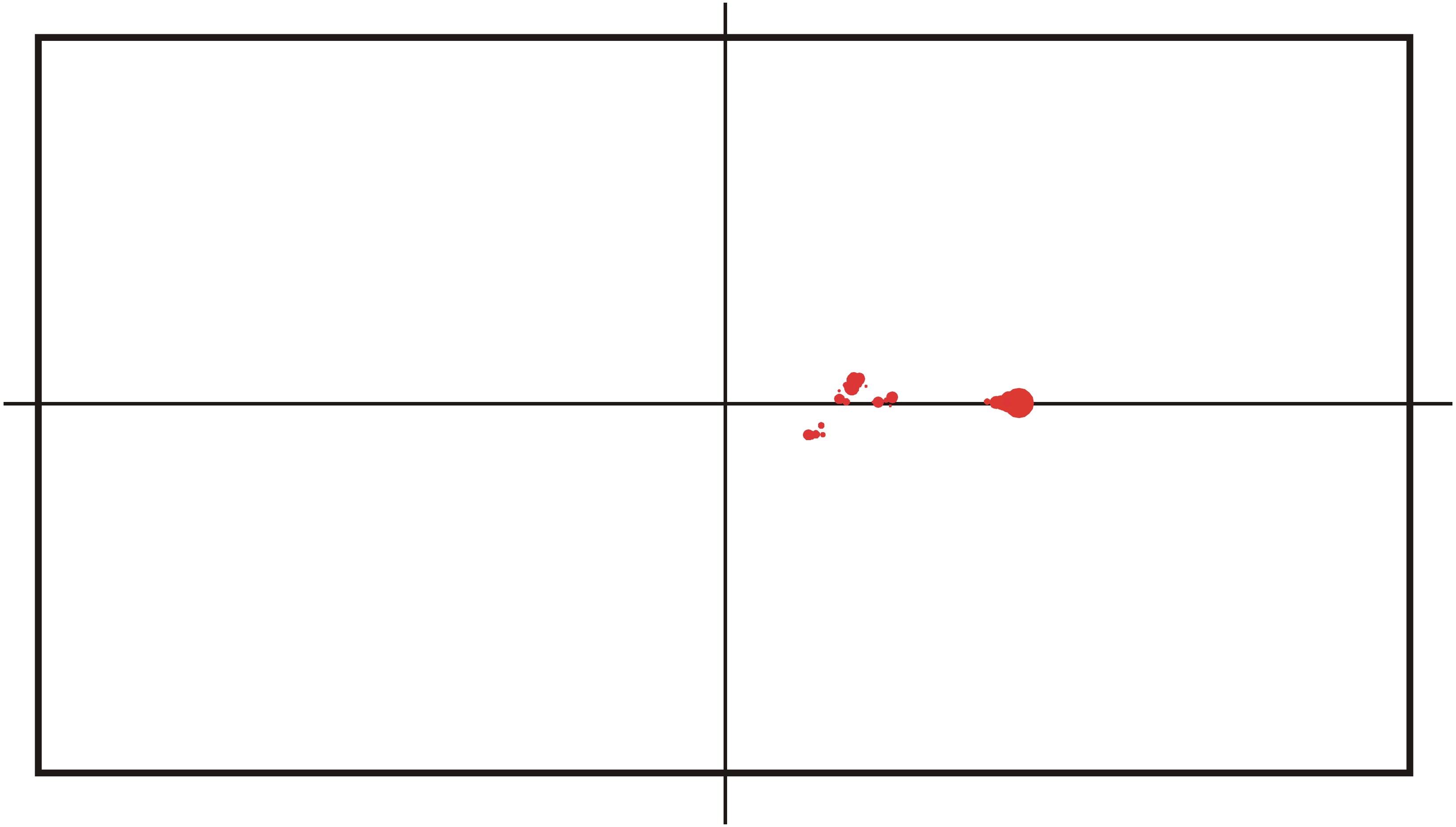
Spread of population in sequence space during a quasistationary epoch: t = 825



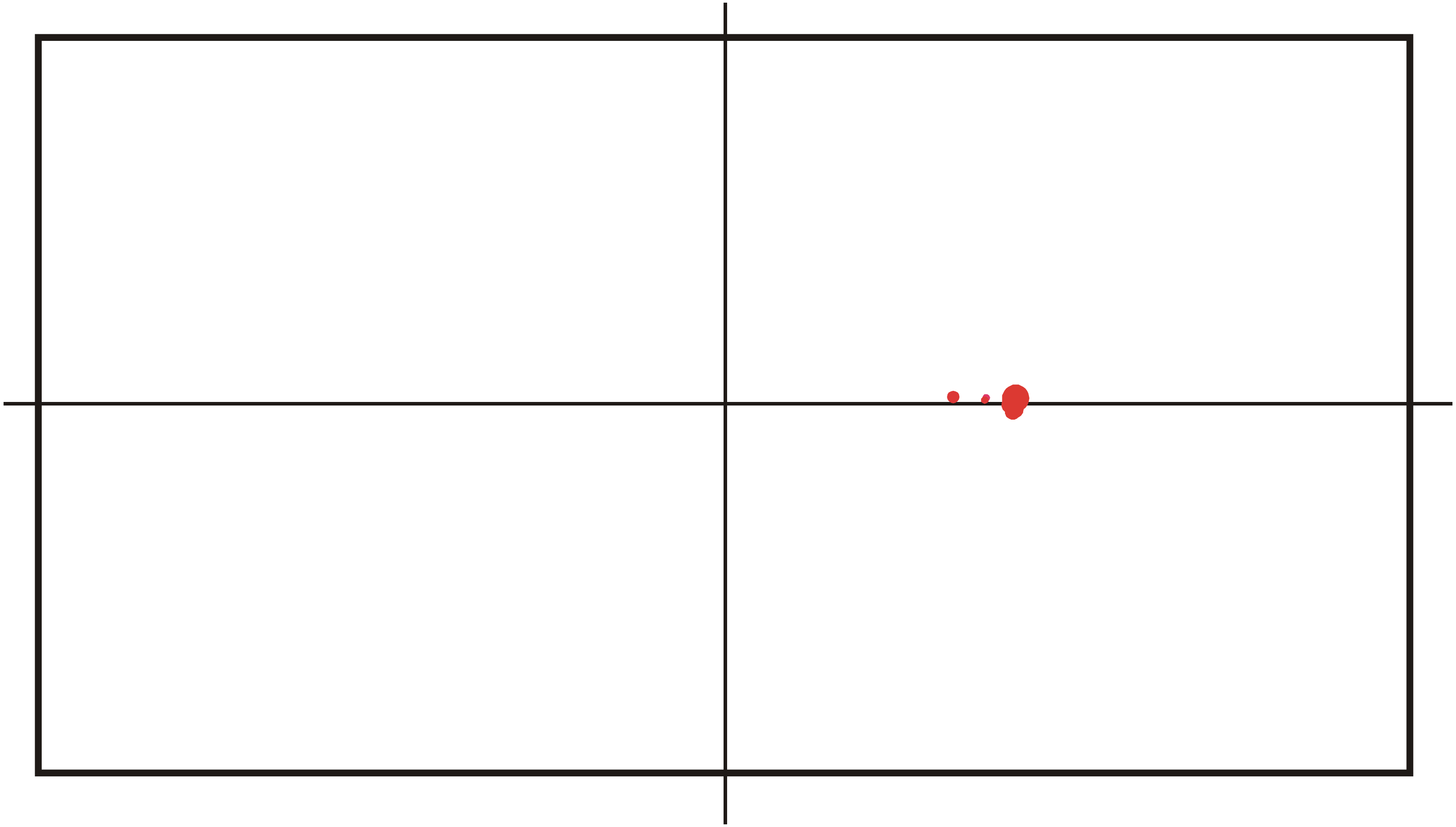
Spread of population in sequence space during a quasistationary epoch: t = 830



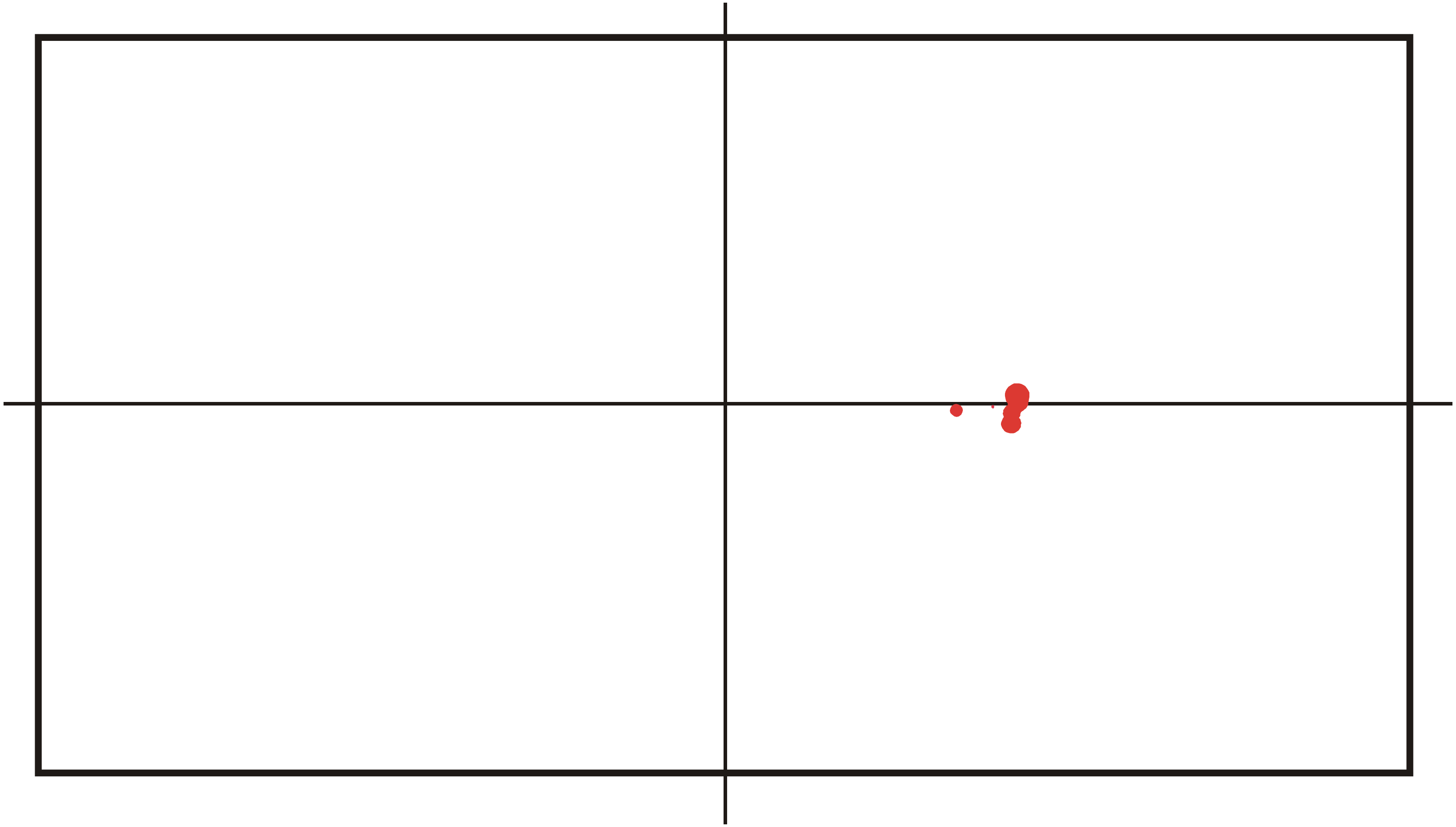
Spread of population in sequence space during a quasistationary epoch: t = 835



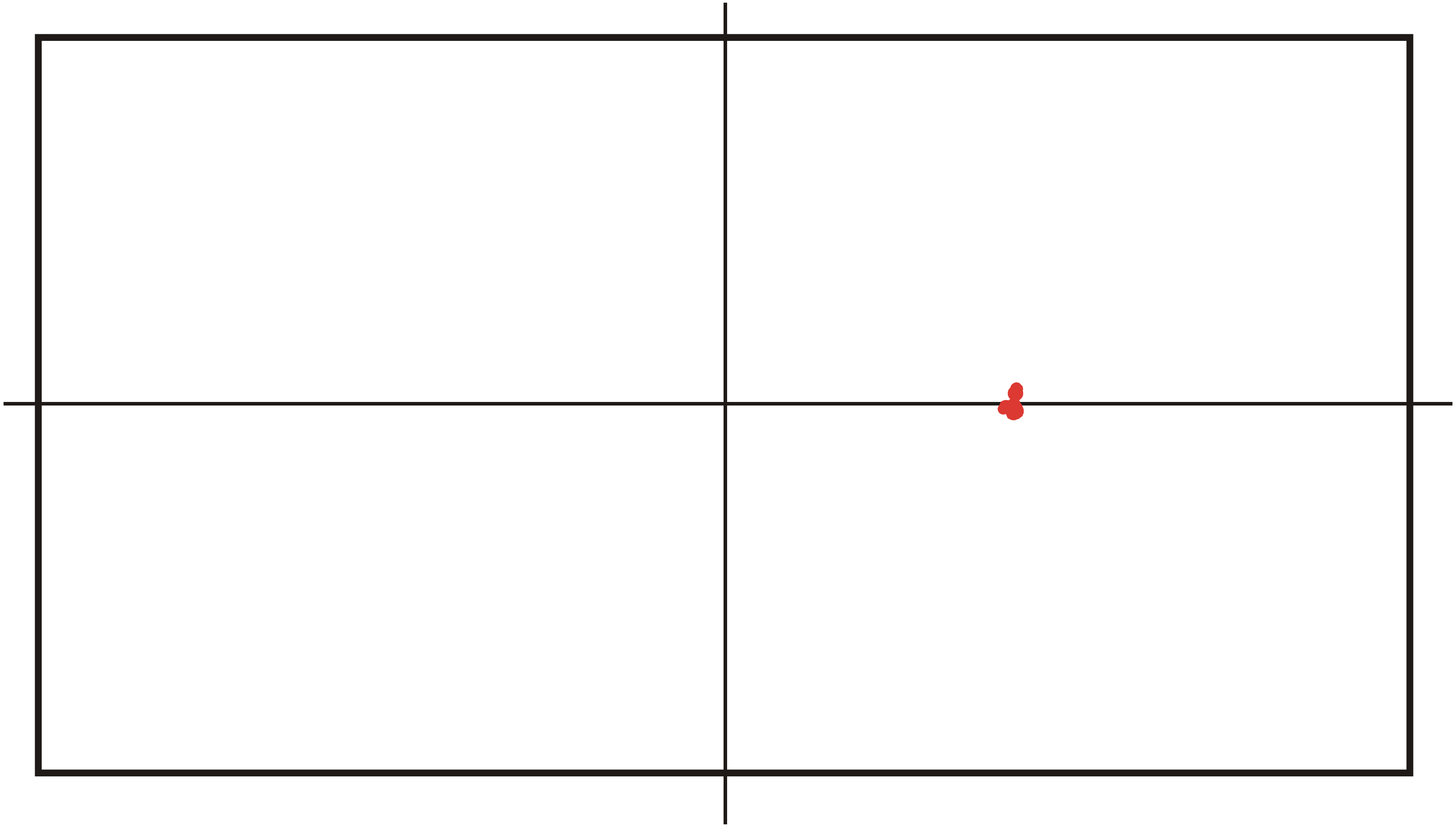
Spread of population in sequence space during a quasistationary epoch: t = 840



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



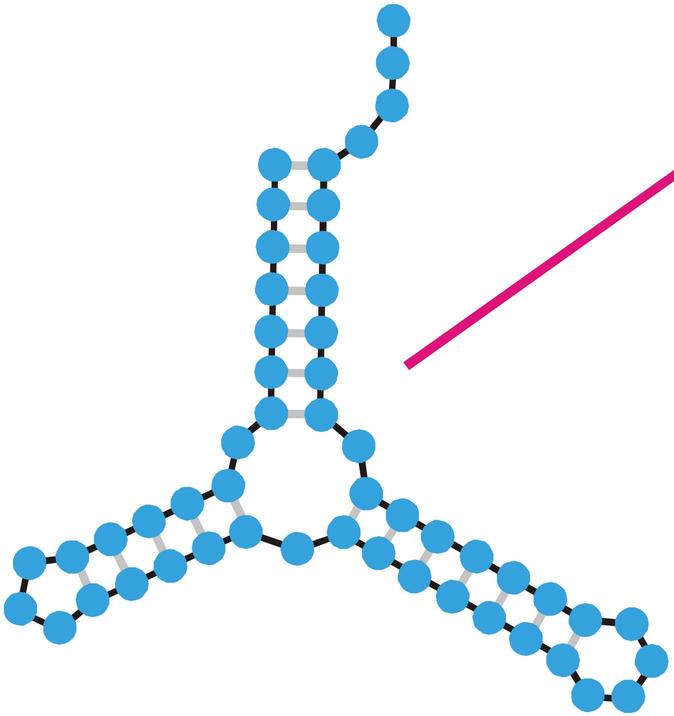
Spread of population in sequence space during a quasistationary epoch: t = 850



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

1. Prolog – Mathematics and the life sciences in the 21st century
2. Replication kinetics of RNA molecules and evolution
3. RNA evolution *in silico*
4. **Sequence-structure maps, neutral networks, and intersections**
5. Reference to experimental data
6. Summary

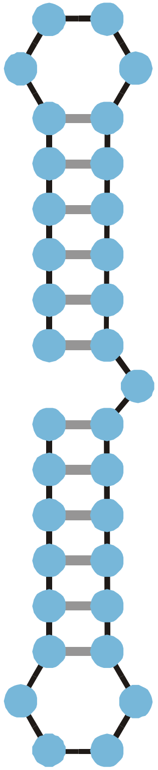
GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA



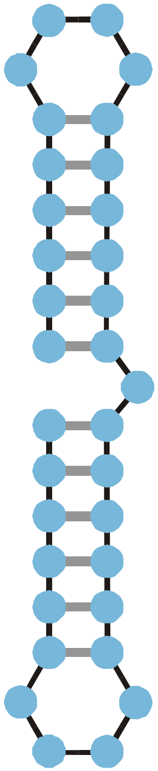
Minimum free energy
criterion

Inverse folding of RNA secondary structures

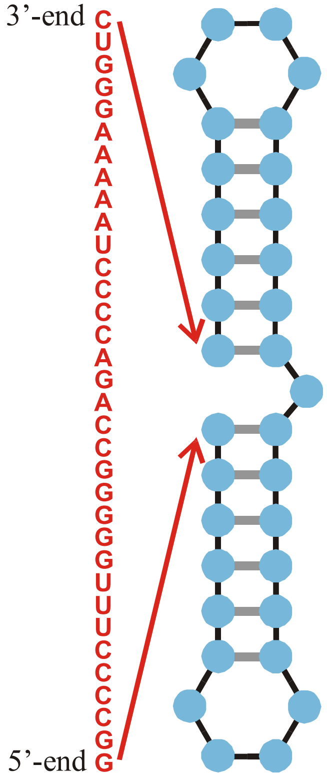
The idea of inverse folding algorithm is to search for sequences that form a given RNA secondary structure under the minimum free energy criterion.



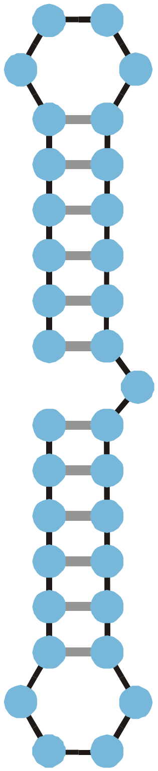
Structure



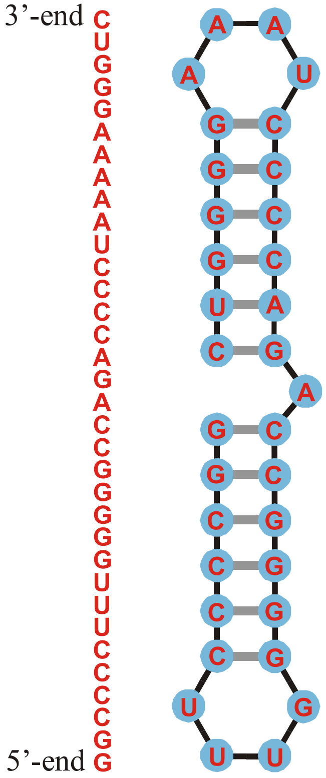
Structure



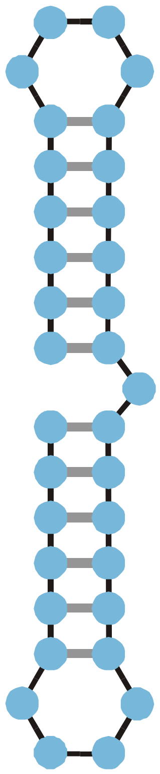
Compatible sequence



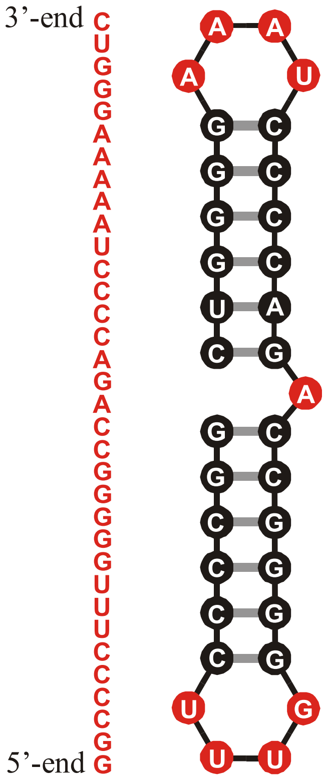
Structure



Compatible sequence



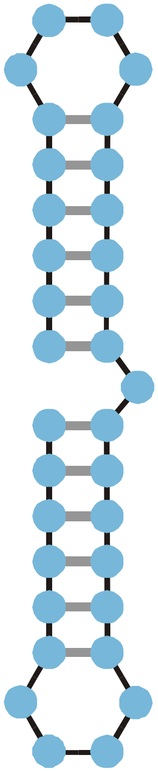
Structure



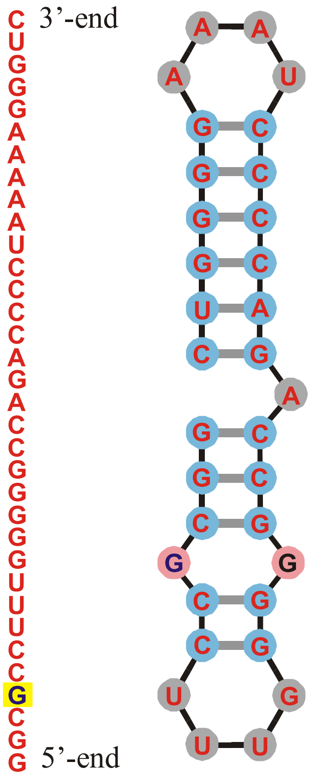
Compatible sequence

Single nucleotides: A,U,G,C

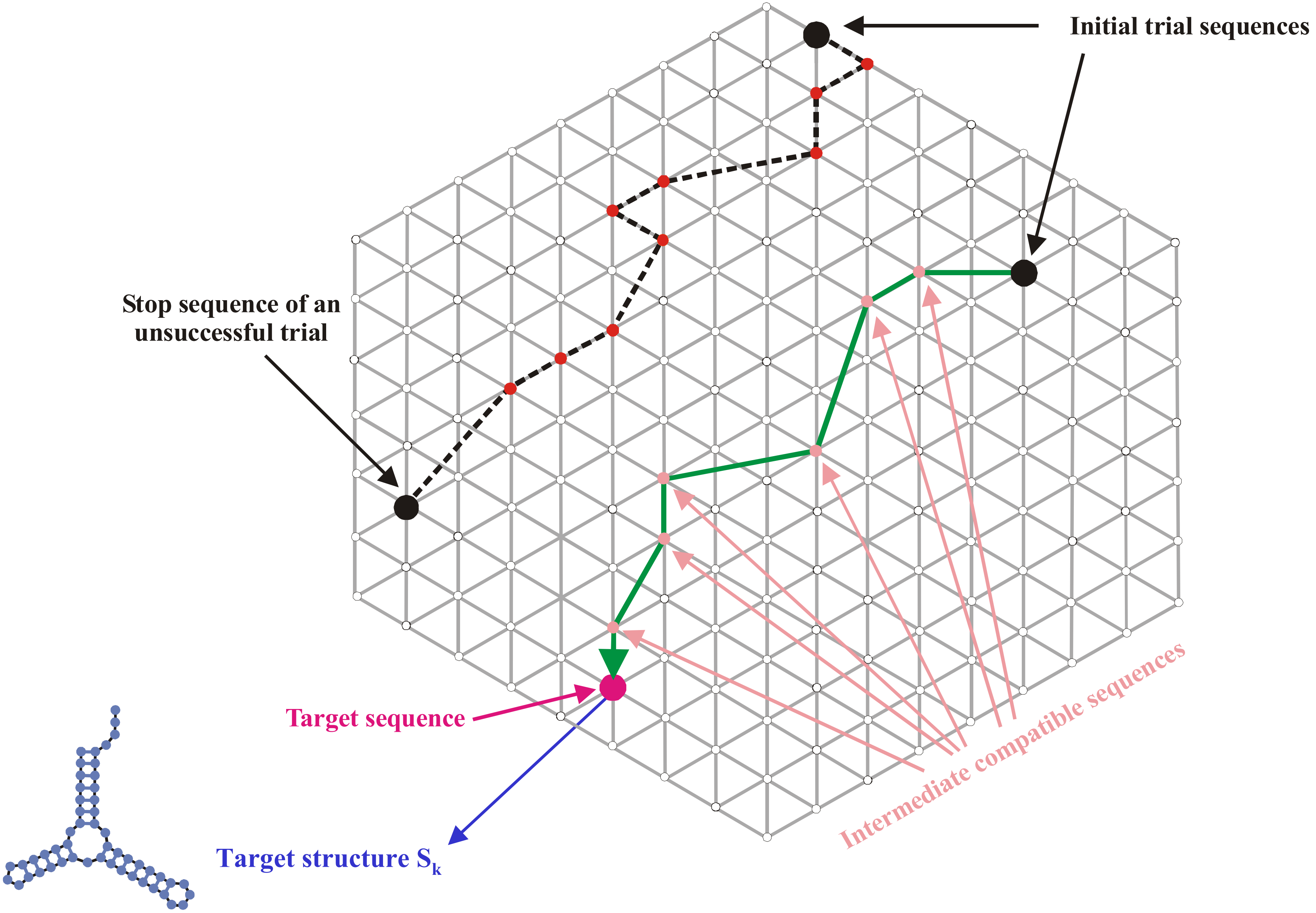
Base pairs: AU , UA
GC , CG
GU , UG



Structure



Incompatible sequence



Initial trial sequences

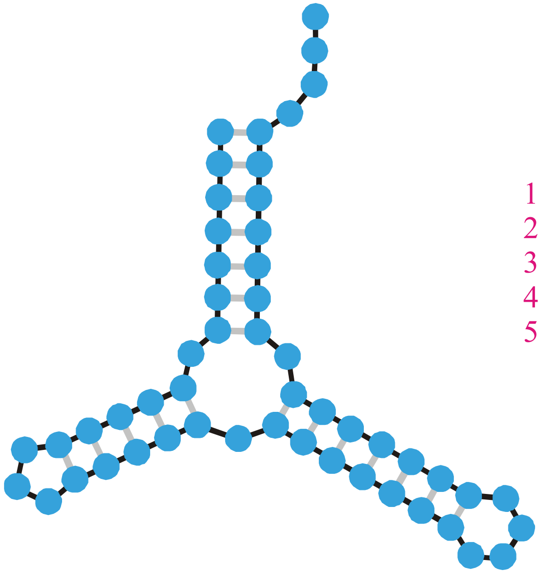
Stop sequence of an unsuccessful trial

Intermediate compatible sequences

Target sequence

Target structure S_k

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



Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th

GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

UGGUUACGCGUUGGGGUAACGAAGAUUCCGAGAGGAGUUUAGUGACUAGAGG

CUUCUUGAGCUAGUACCUAGUCGGAUAGGAUUUCCUAUCUCCAGGGAGGAUG

CUUUUCUUCACGUUAGAUGUGUAAUGGACAUGUGUUUAUUUAGGAAAGGCGC

AUAACGUGAGUGUCUAAUACUGAUCGCUCCGGAGGGUGGUGGCGUUGUUAAU

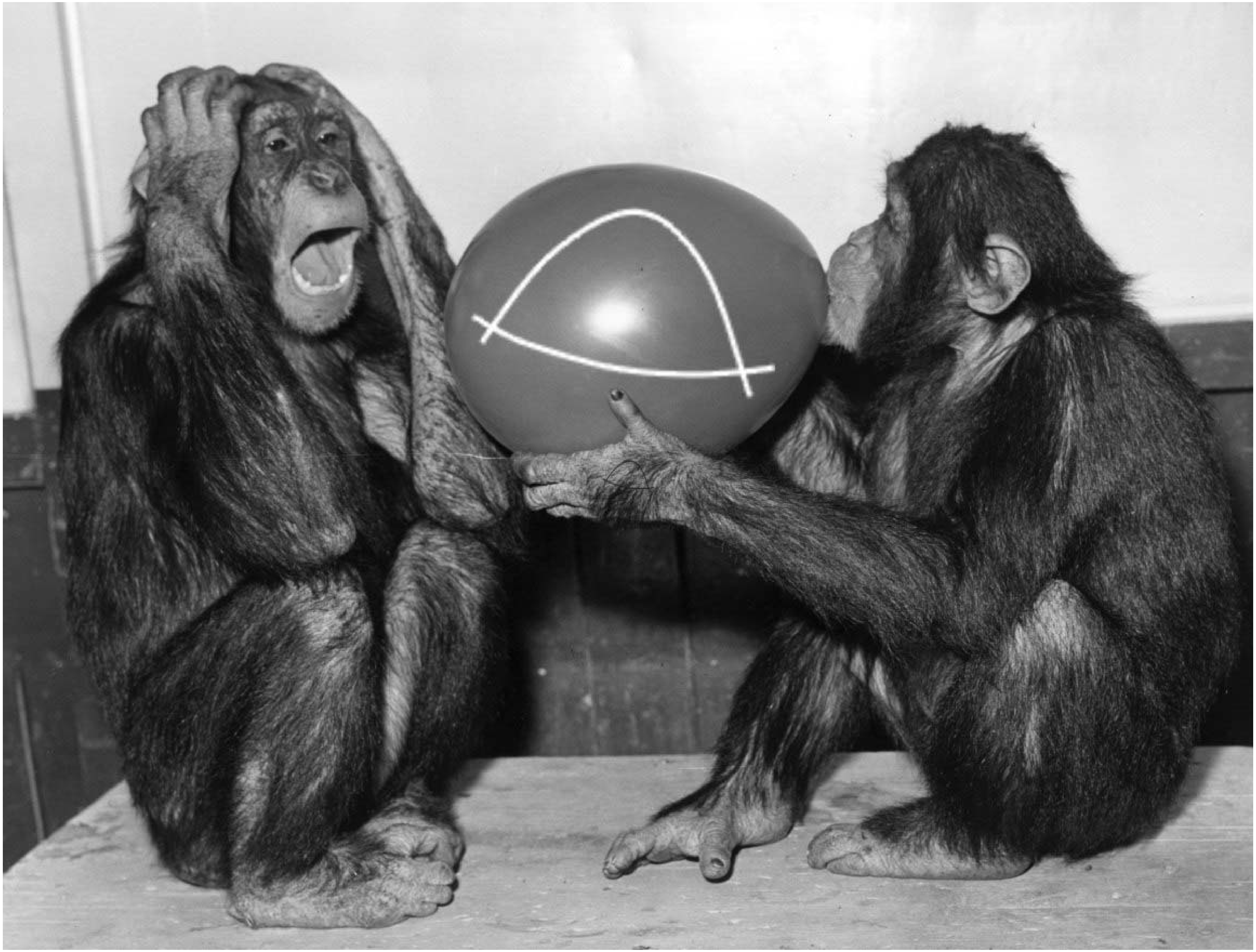
Inverse folding of RNA secondary structures

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

RNA **sequences** as well as RNA secondary **structures** can be visualized as objects in **metric spaces**. At constant chain length the sequence space is a (generalized) hypercube.

The **mapping** from RNA **sequences** into RNA secondary **structures** is many-to-one. Hence, it is redundant and not invertible.

RNA **sequences**, which are mapped onto the same RNA secondary **structure**, are **neutral** with respect to **structure**. The pre-images of structures in sequence space are **neutral networks**. They can be represented by graphs where the edges connect sequences of Hamming distance $d_H = 1$.



Theory of genotype – phenotype mapping

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

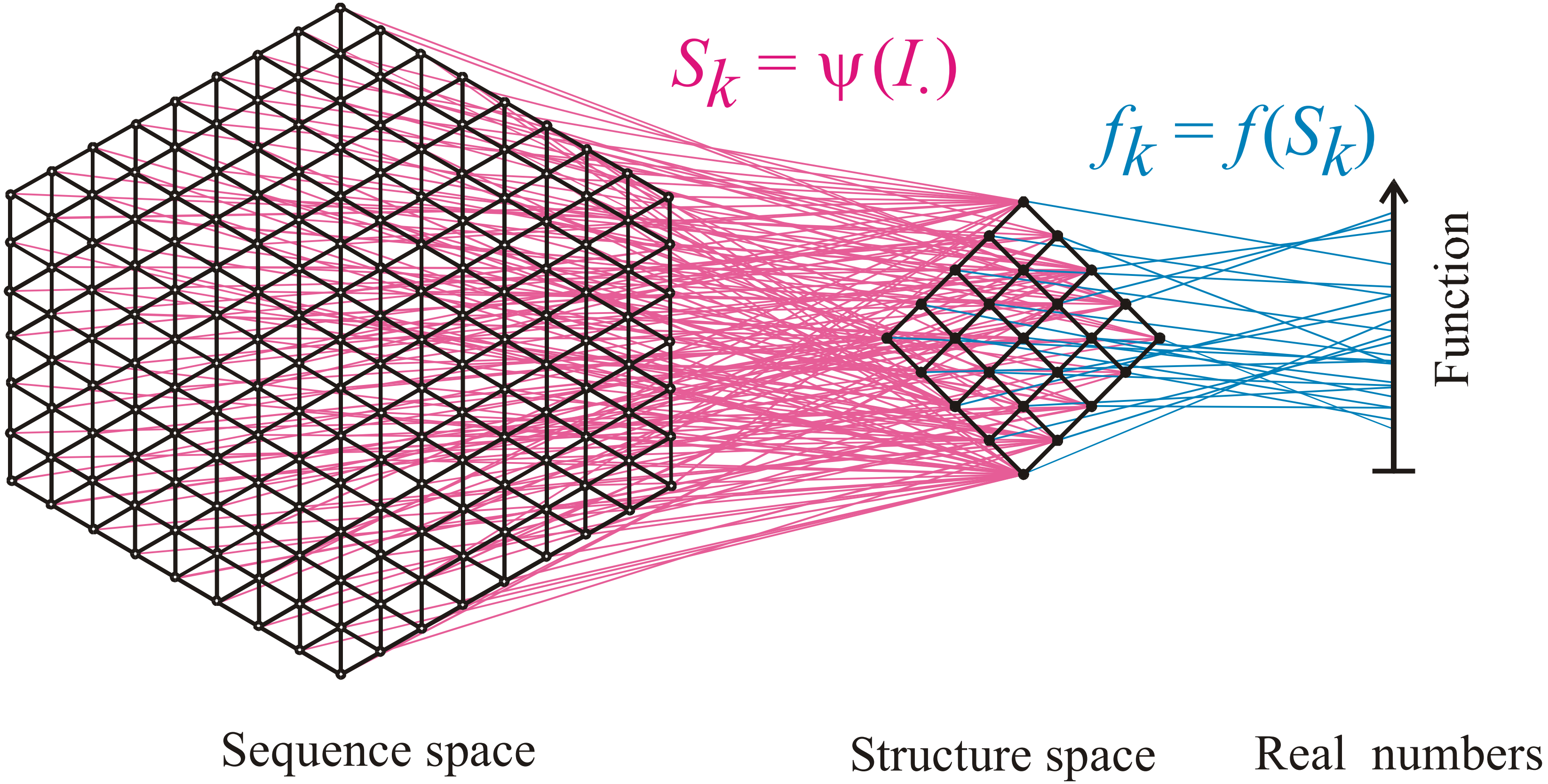
W.Grüner, R.Giegerich, D.Strothmann, C.Reidys, I.L.Hofacker, P.Schuster, ***Analysis of RNA sequence structure maps by exhaustive enumeration. I. Neutral networks***. Mh.Chem. **127** (1996), 355-374

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

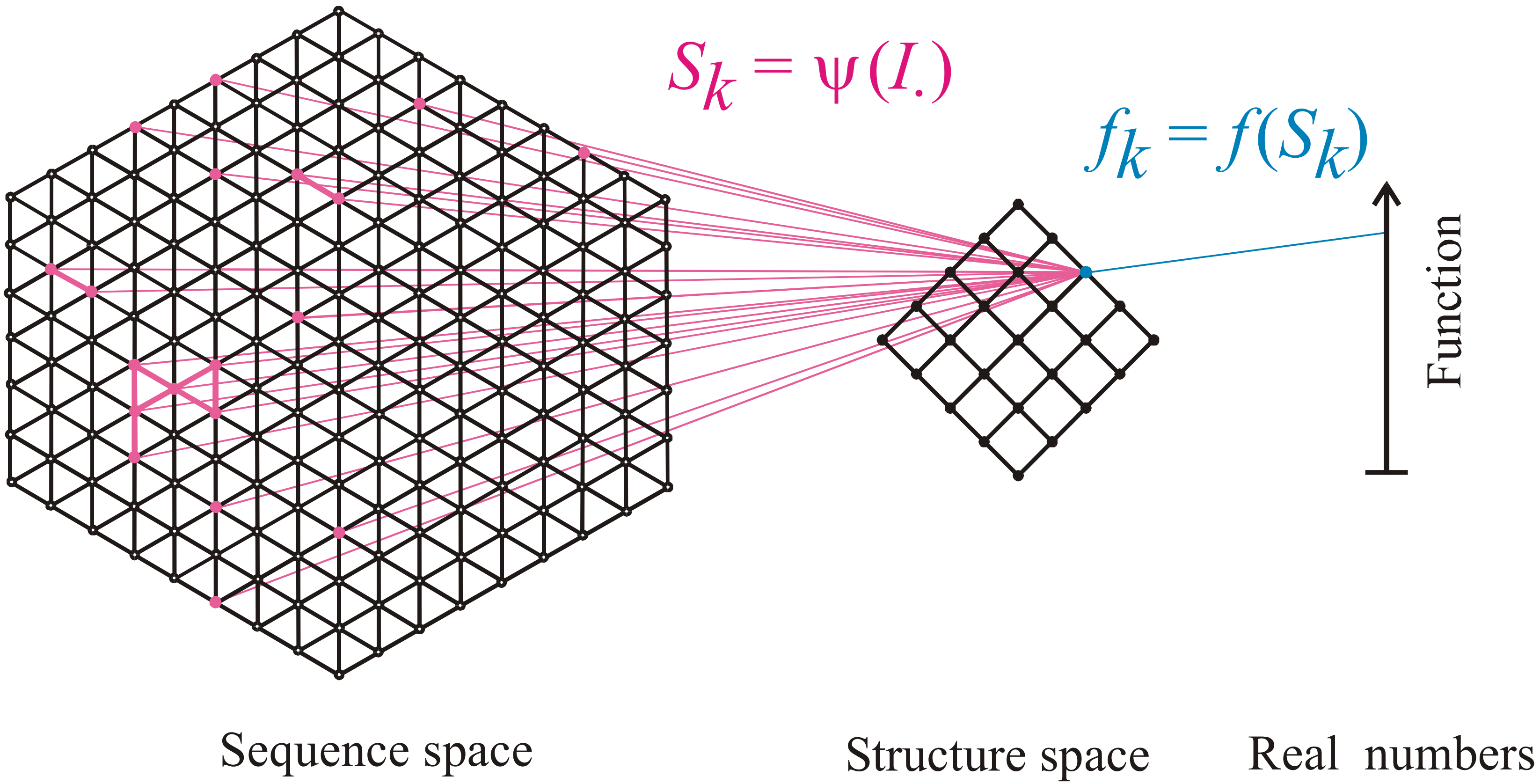
C.M.Reidys, P.F.Stadler, P.Schuster, ***Generic properties of combinatory maps***. Bull.Math.Biol. **59** (1997), 339-397

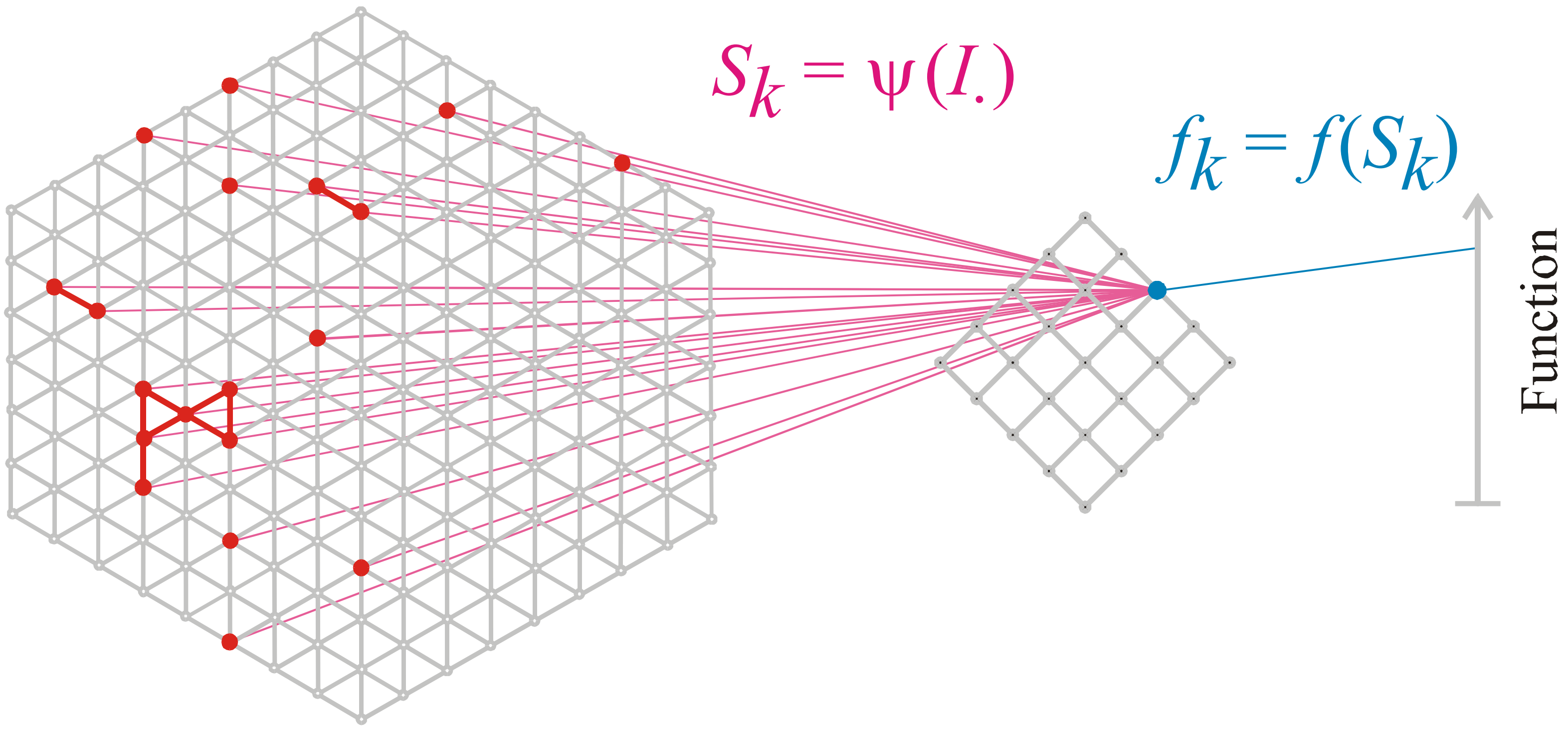
I.L.Hofacker, P. Schuster, P.F.Stadler, ***Combinatorics of RNA secondary structures***. Discr.Appl.Math. **89** (1998), 177-207

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



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



$S_k = \psi(I.)$

$f_k = f(S_k)$

Function

Sequence space

Structure space

Real numbers

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

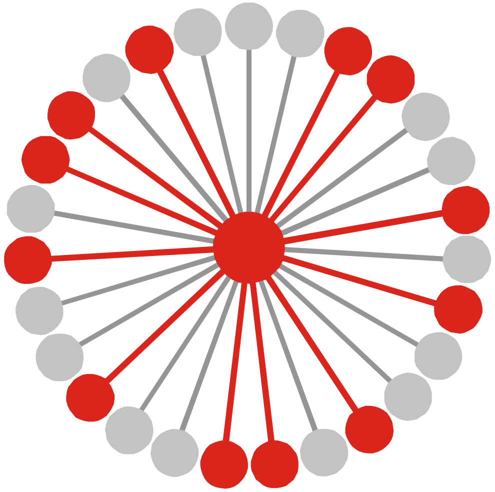
Neutral networks are sets of sequences forming the same structure. **G_k** is the pre-image of the structure S_k in sequence space:

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

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

Neutral networks of small RNA molecules can be computed by exhaustive folding of complete sequence spaces, i.e. all RNA sequences of a given chain length. This number, **$N=4^n$** , 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.



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

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

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

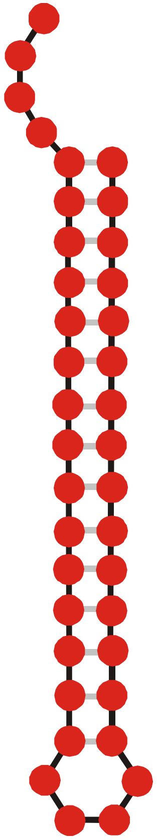
Alphabet size k : **AUGC** ñ k = 4

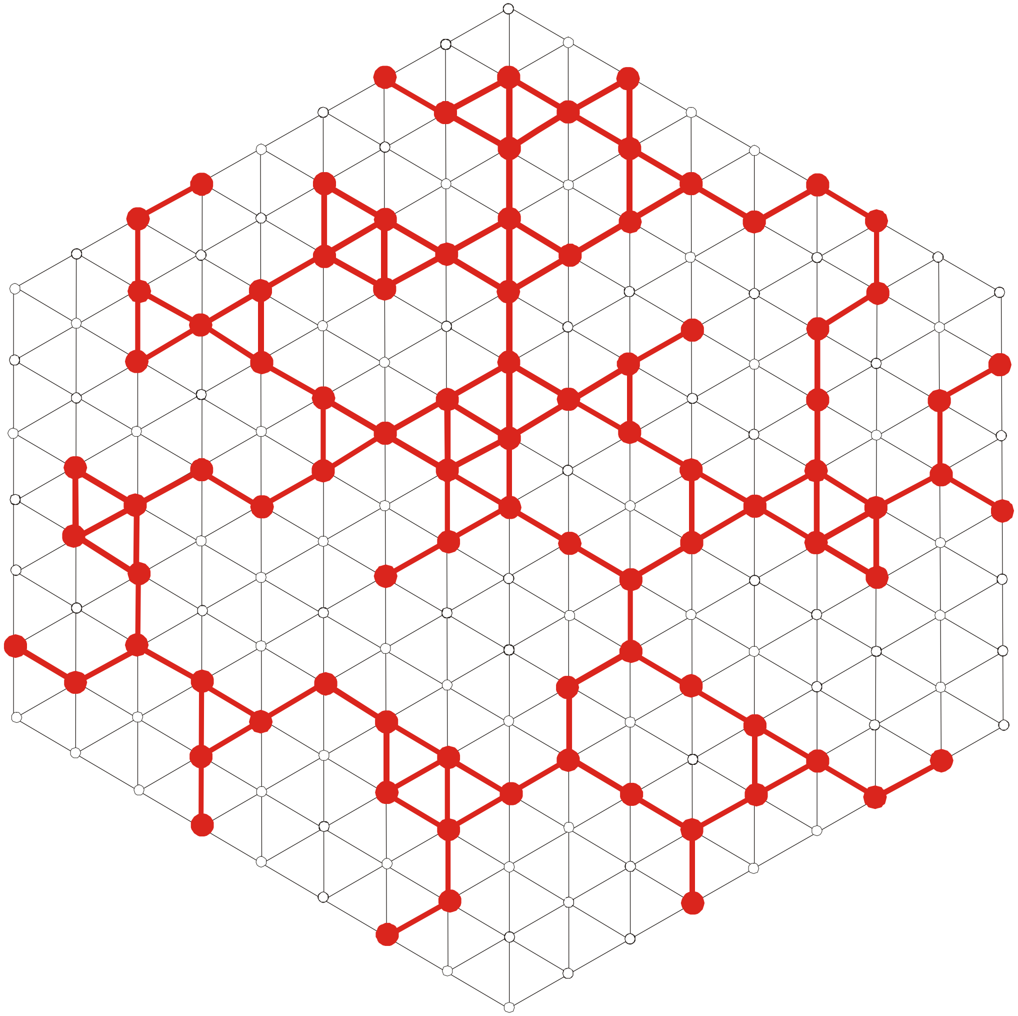
$\bar{\lambda}_k > \lambda_{cr}$ . . . . network **G_k** is connected

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

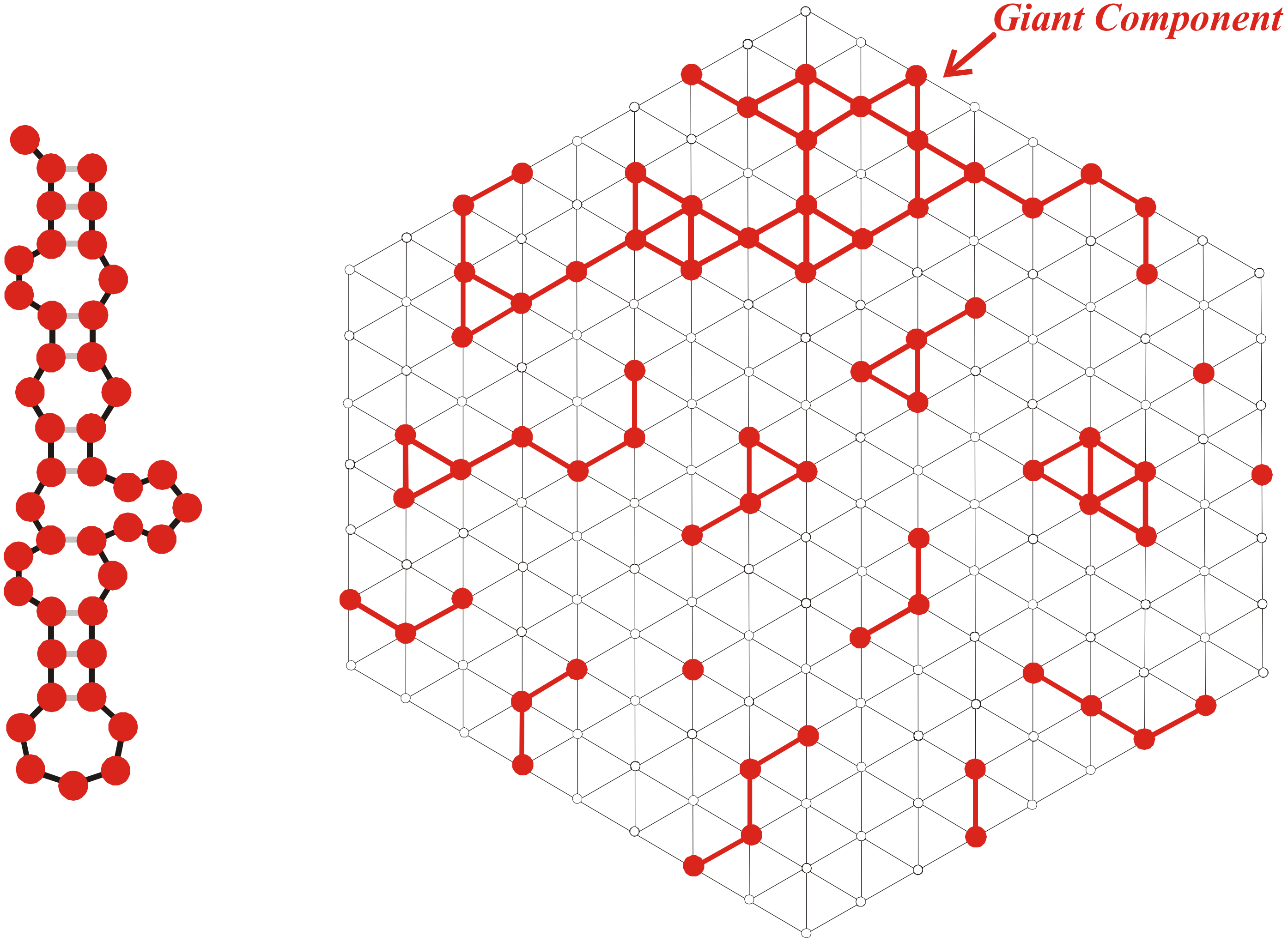
k	l_{cr}	
2	0.5	**GC,AU**
3	0.423	**GUC,AUG**
4	0.370	**AUGC**

Mean degree of neutrality and connectivity of neutral networks

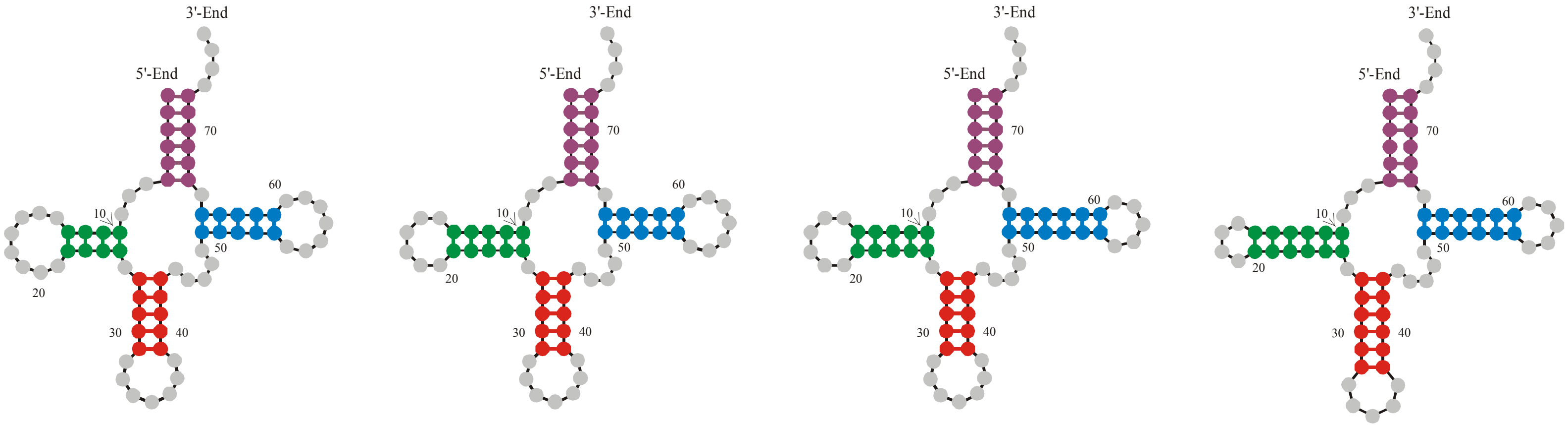




A connected neutral network

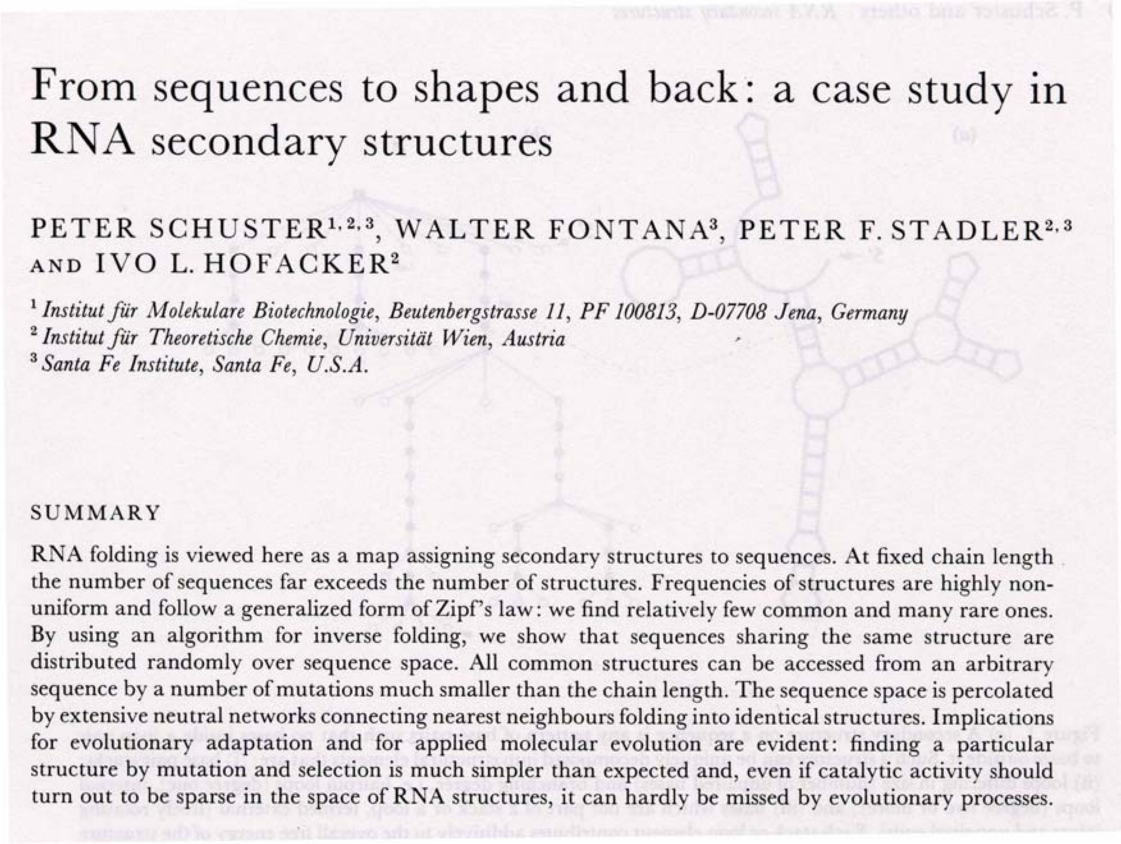


A multi-component neutral network



Alphabet	**Degree of neutrality `l**			
AU	- -	- -	- -	0.073 Ä 0.032
AUG	- -	0.217 Ä 0.051	0.207 ± 0.055	0.201 Ä 0.056
AUGC	0.275 Ä 0.064	0.279 Ä 0.063	0.289 ± 0.062	0.313 Ä 0.058
UGC	0.263 Ä 0.071	0.257 Ä 0.070	0.251 ± 0.068	0.250 Ä 0.064
GC	0.052 Ä 0.033	0.057 Ä 0.034	0.060 ± 0.033	0.068 Ä 0.034

Degree of neutrality of cloverleaf RNA secondary structures over different alphabets



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.

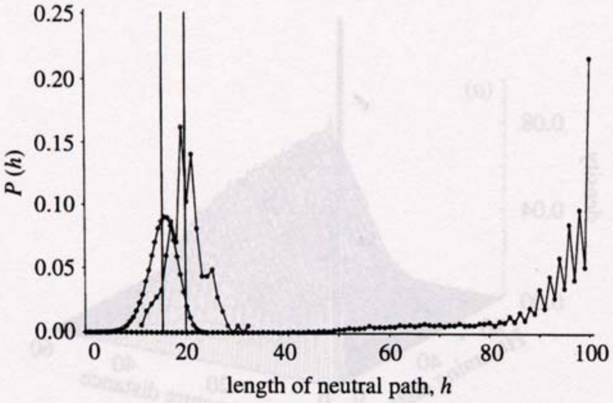
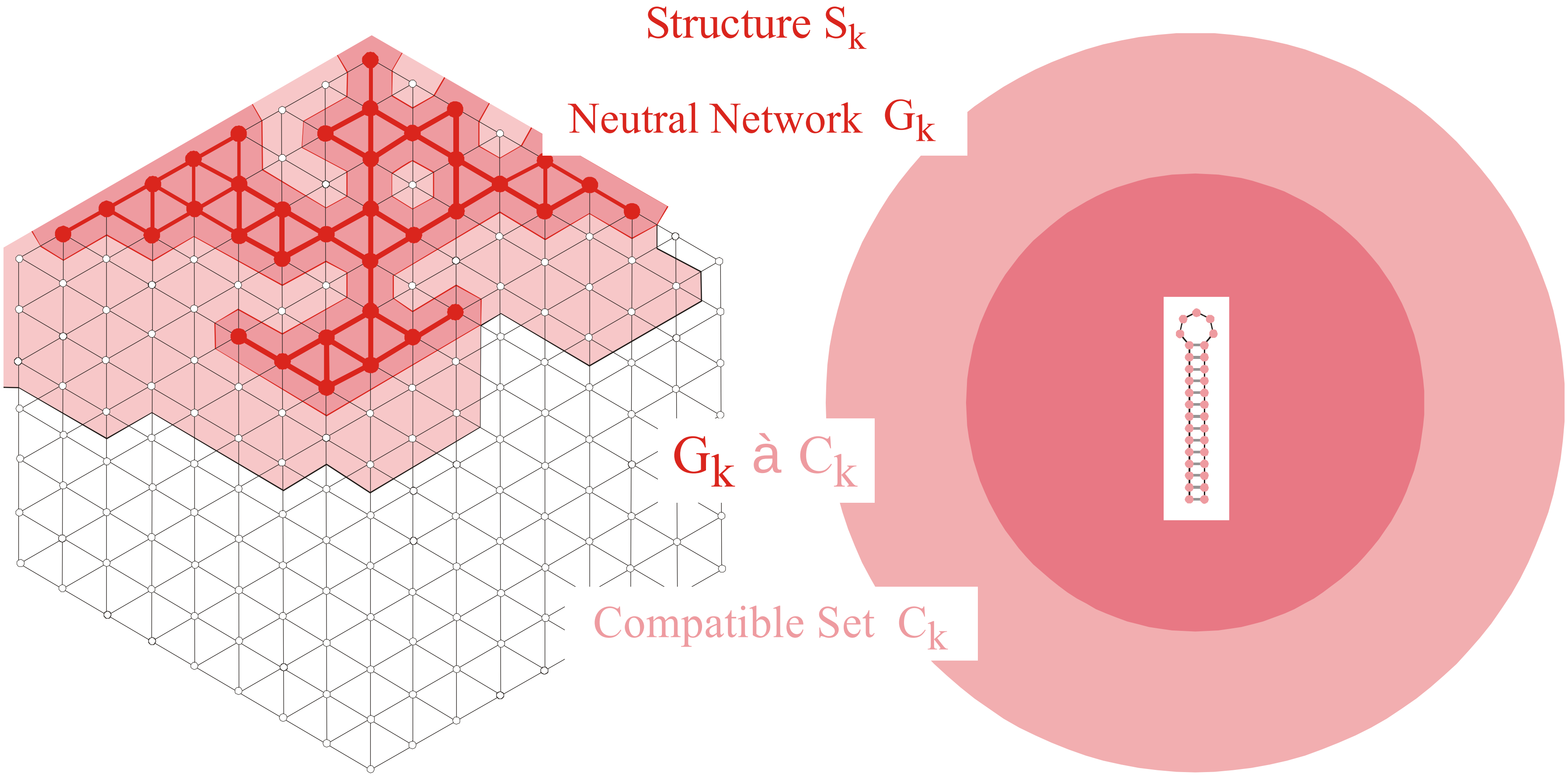


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and 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).

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



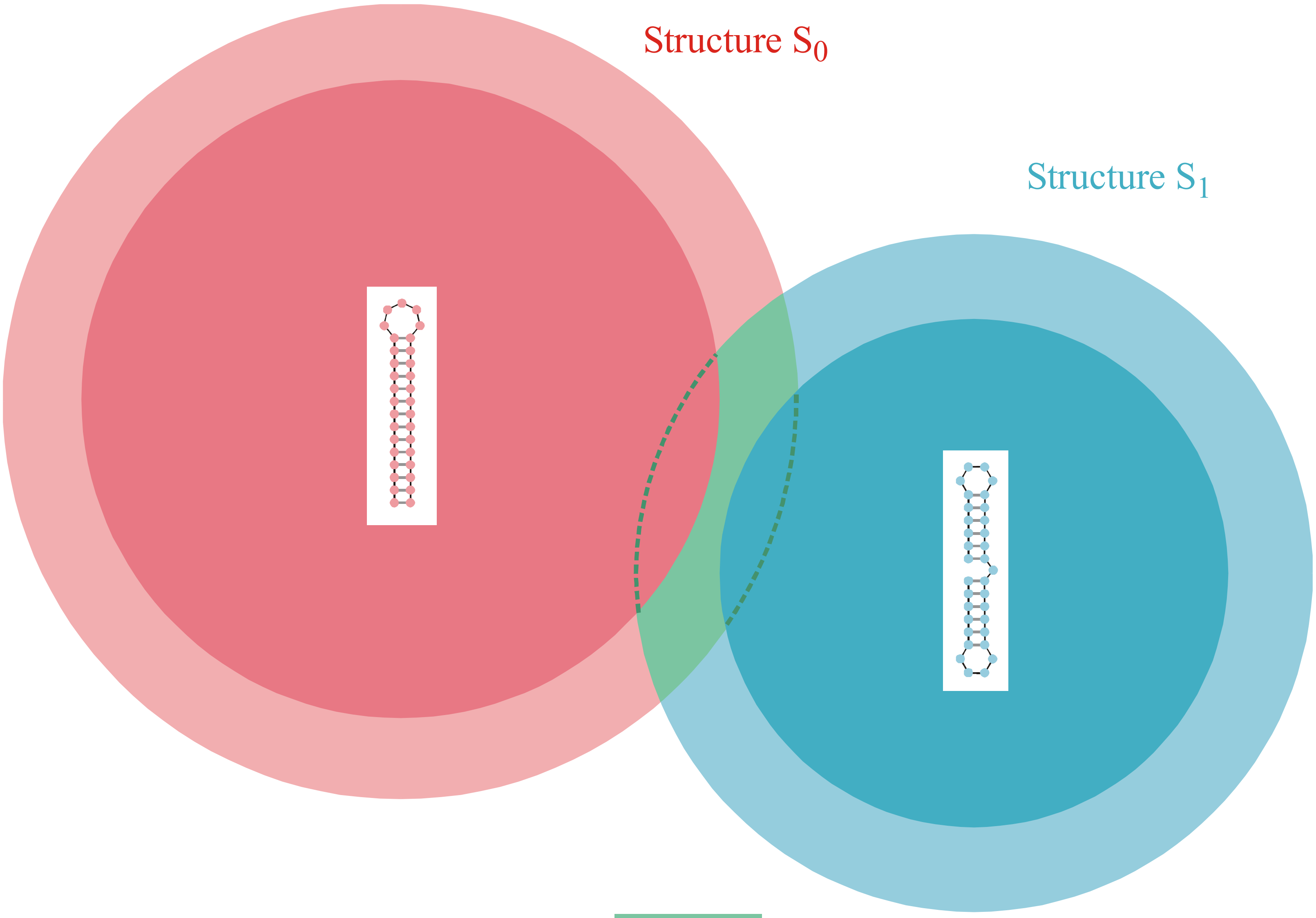
Structure S_k

Neutral Network G_k

G_k à C_k

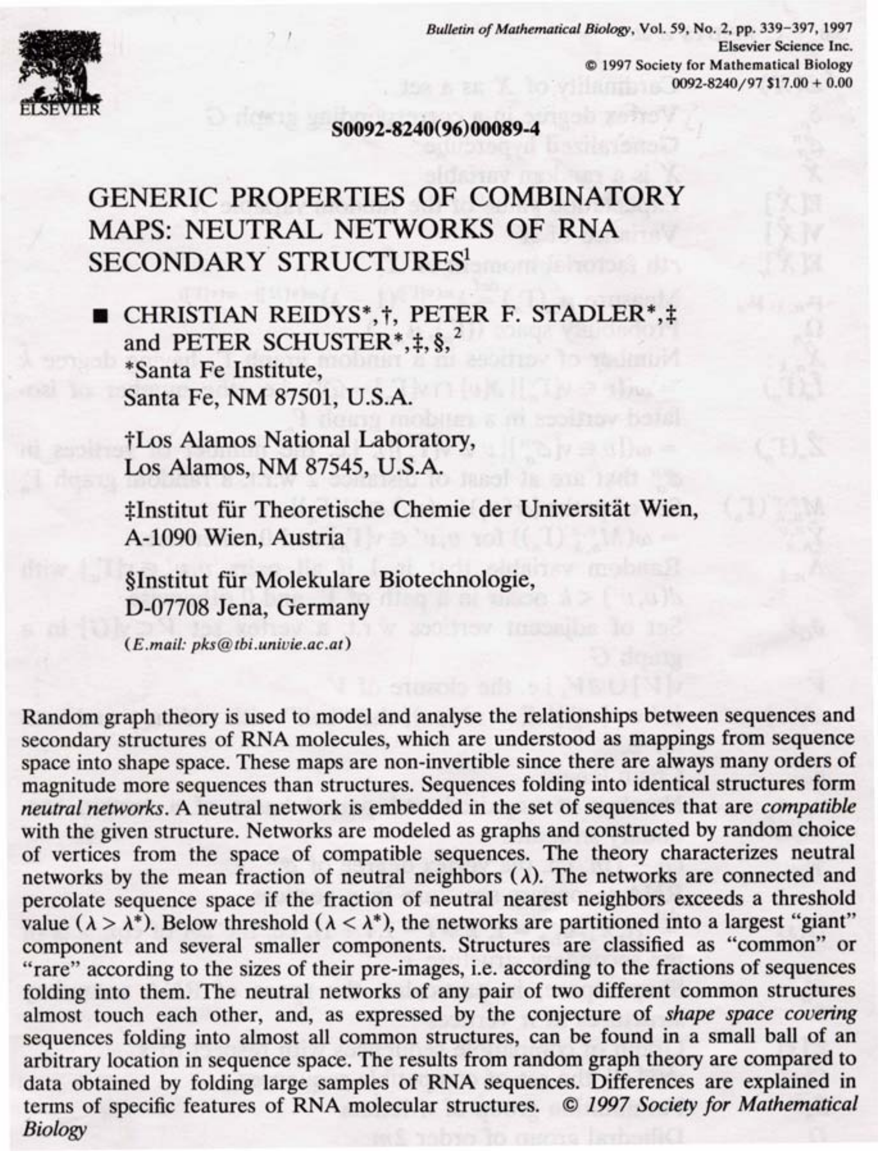
Compatible Set C_k

The **compatible set C_k** of a structure **S_k** consists of all sequences which form **S_k** as its minimum free energy structure (the 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.
© 1997 Society for Mathematical Biology
0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

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

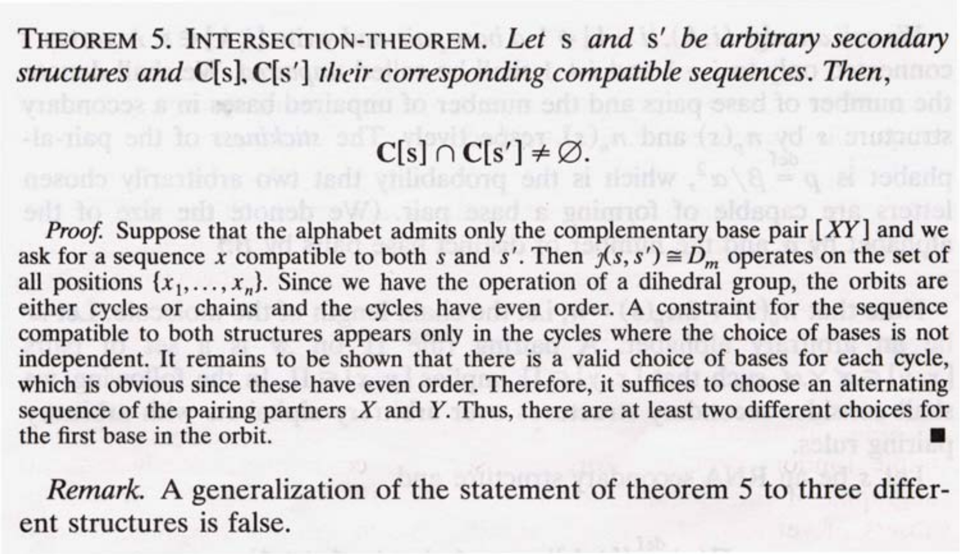
■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡ and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

†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*



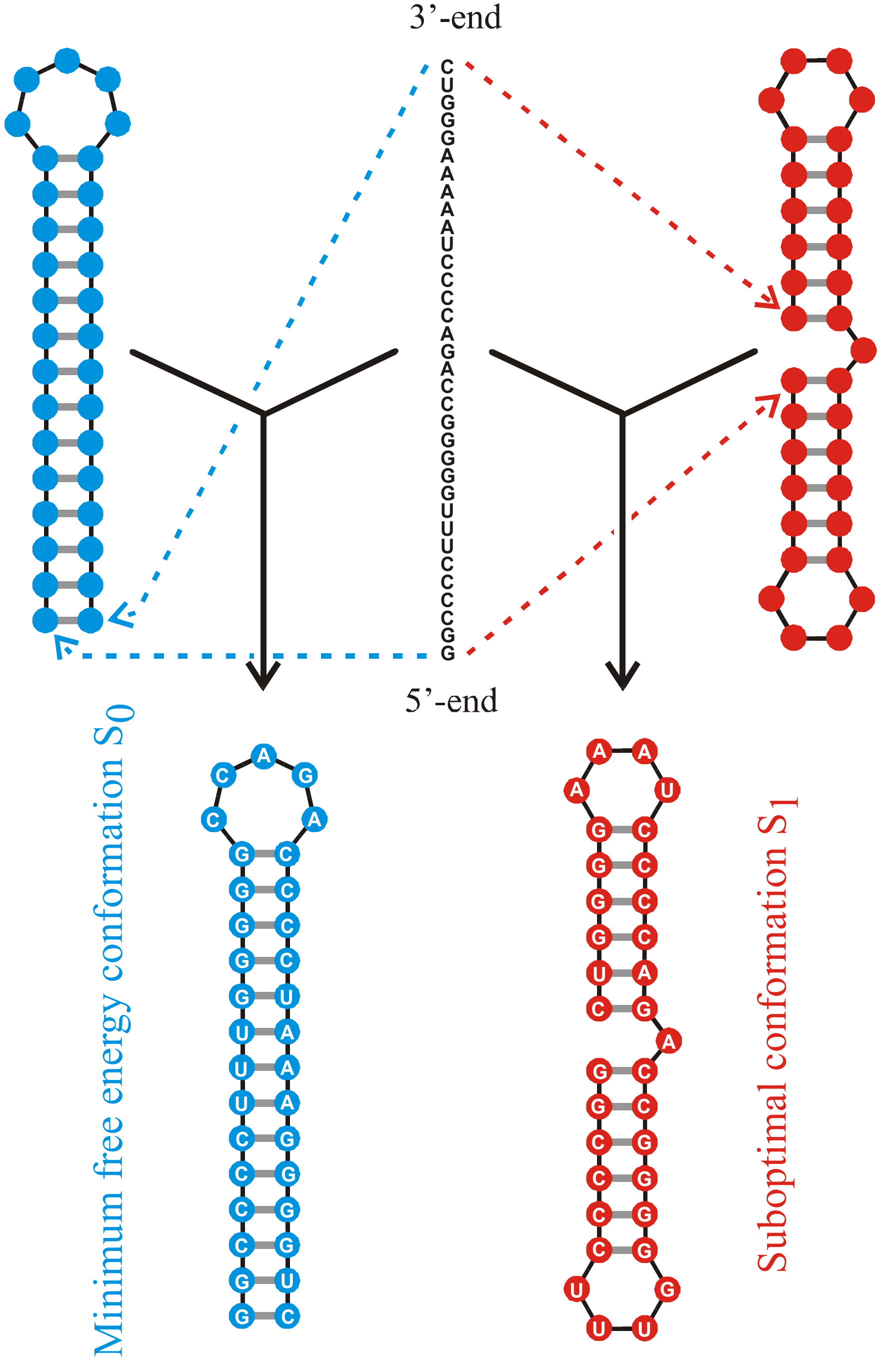
THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s′ *be arbitrary secondary structures and* C[s], C[s′] *their corresponding compatible sequences. Then,*

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

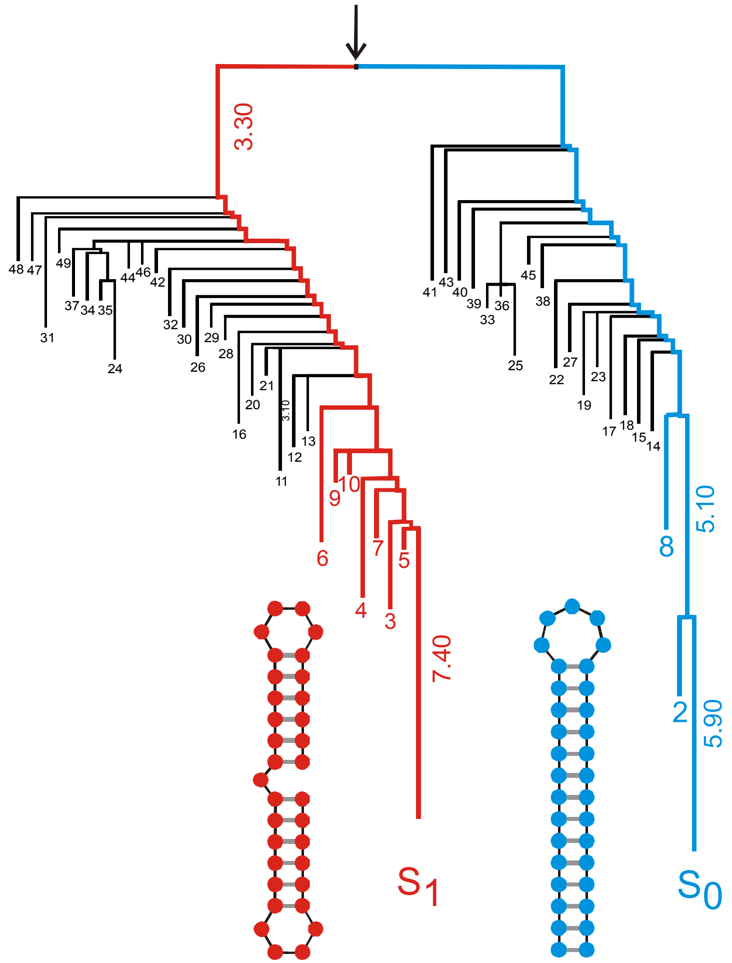
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\mathcal{J}(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either 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**


3'-end
CUGGGAAAAAUCCCCAGACCGGGGUUUCCCCGG
5'-end
Minimum free energy conformation S_0
Suboptimal conformation S_1

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



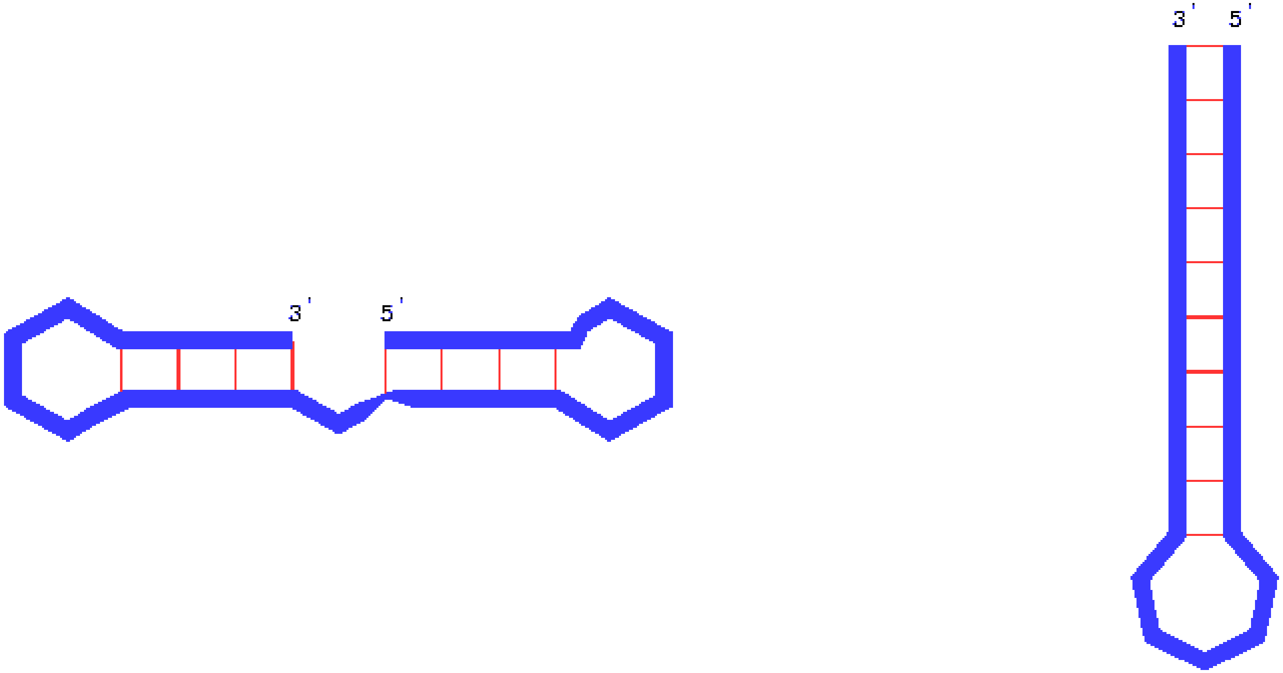
basin '1'

basin '0'

long living metastable structure

minimum free energy structure

Barrier tree for two long living structures



Kinetics of RNA refolding between a long living metastable conformation and the minmum free energy structure

1. Prolog – Mathematics and the life sciences in the 21st century

2. Replication kinetics of RNA molecules and evolution

3. RNA evolution *in silico*

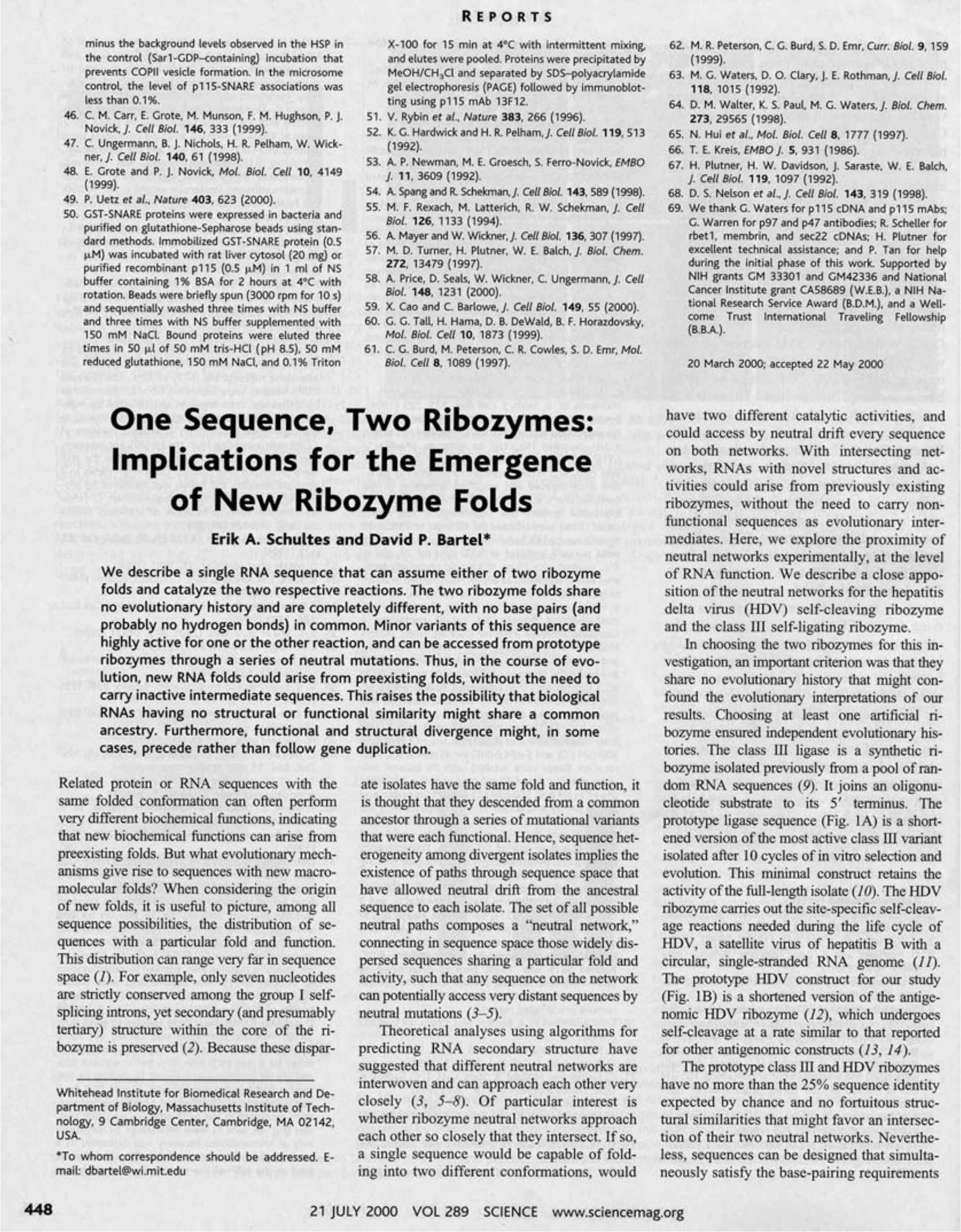
4. Sequence-structure maps, neutral networks, and intersections

5. **Reference to experimental data**

6. 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*).

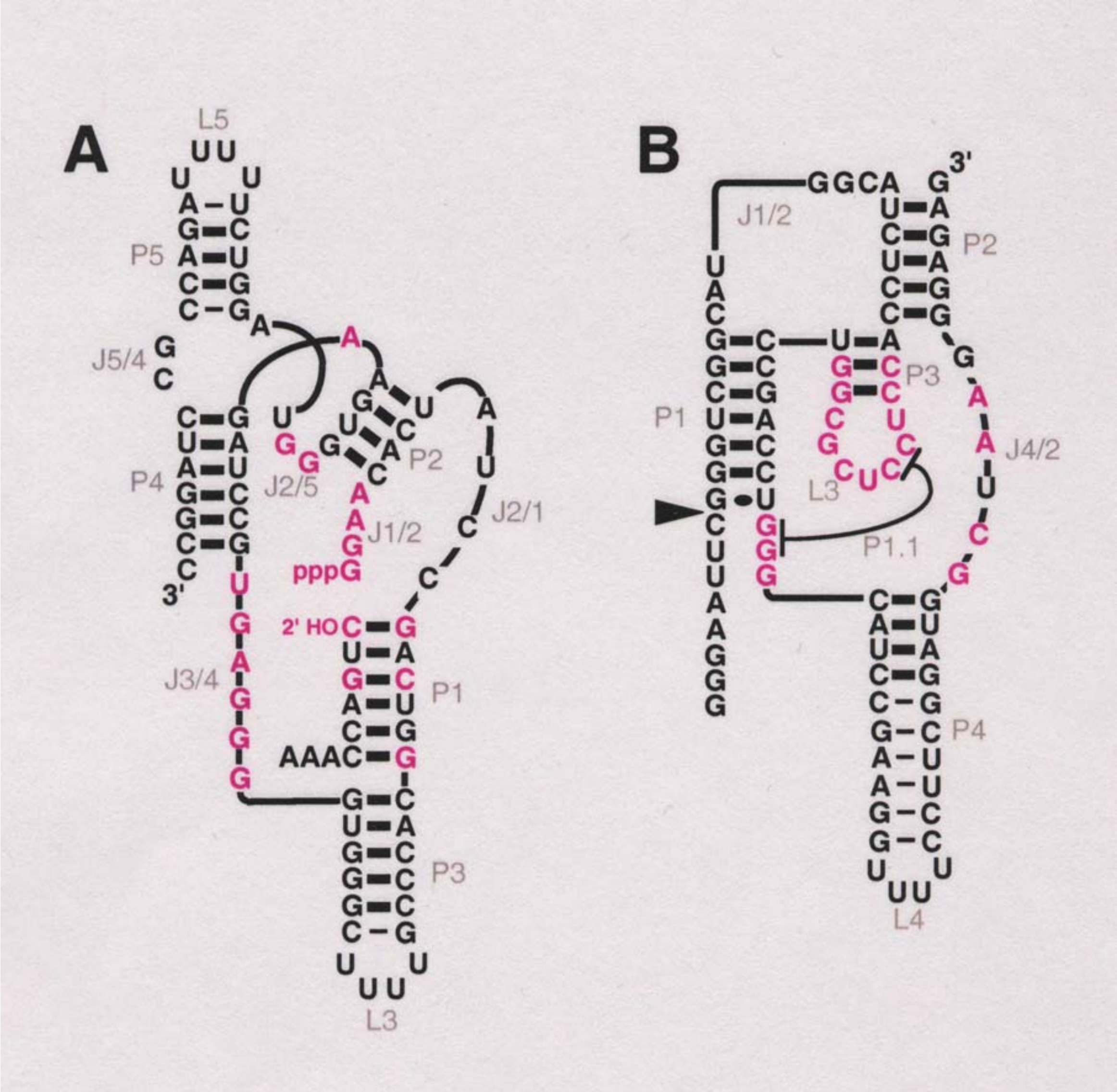
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect.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

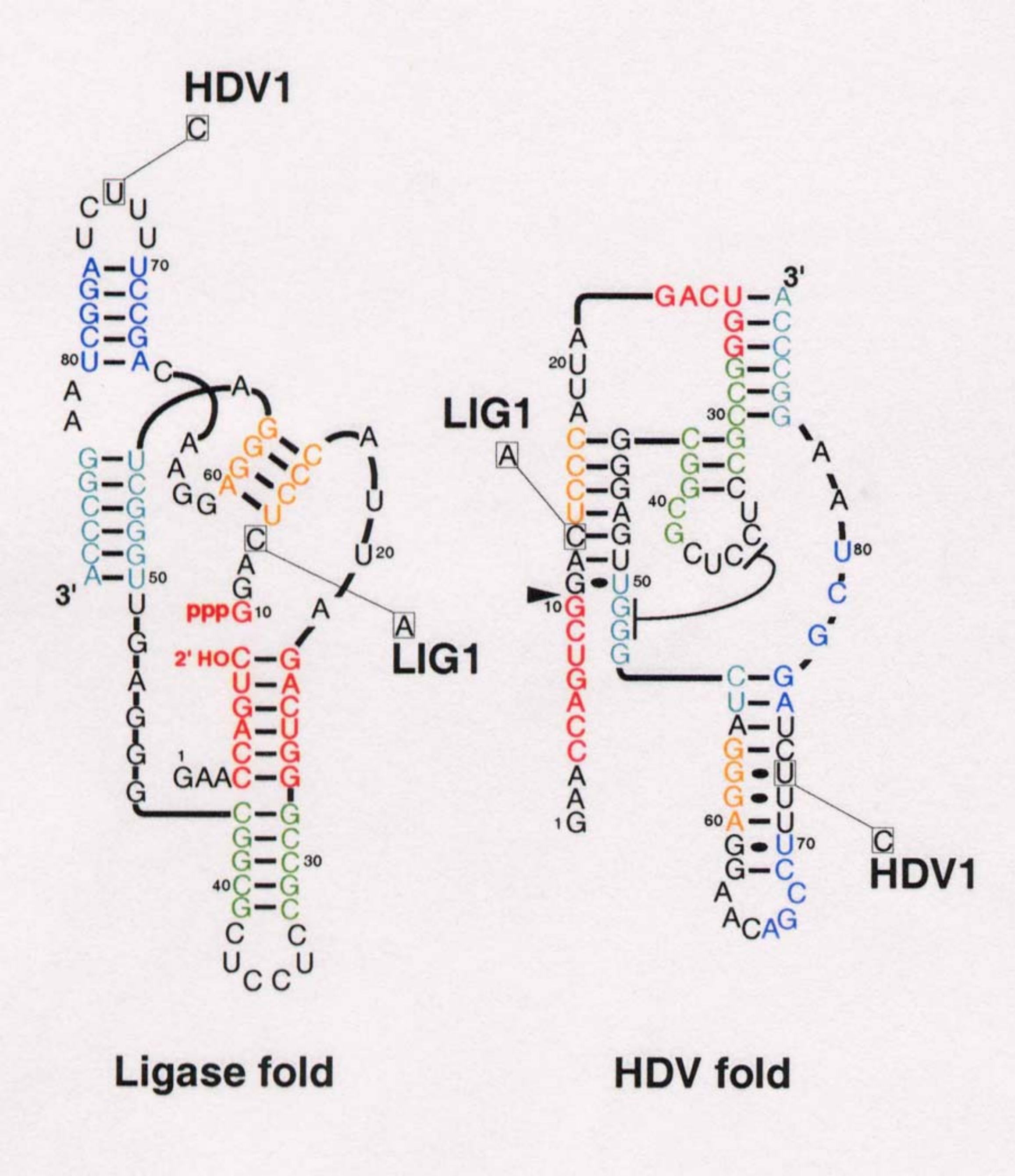
The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

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

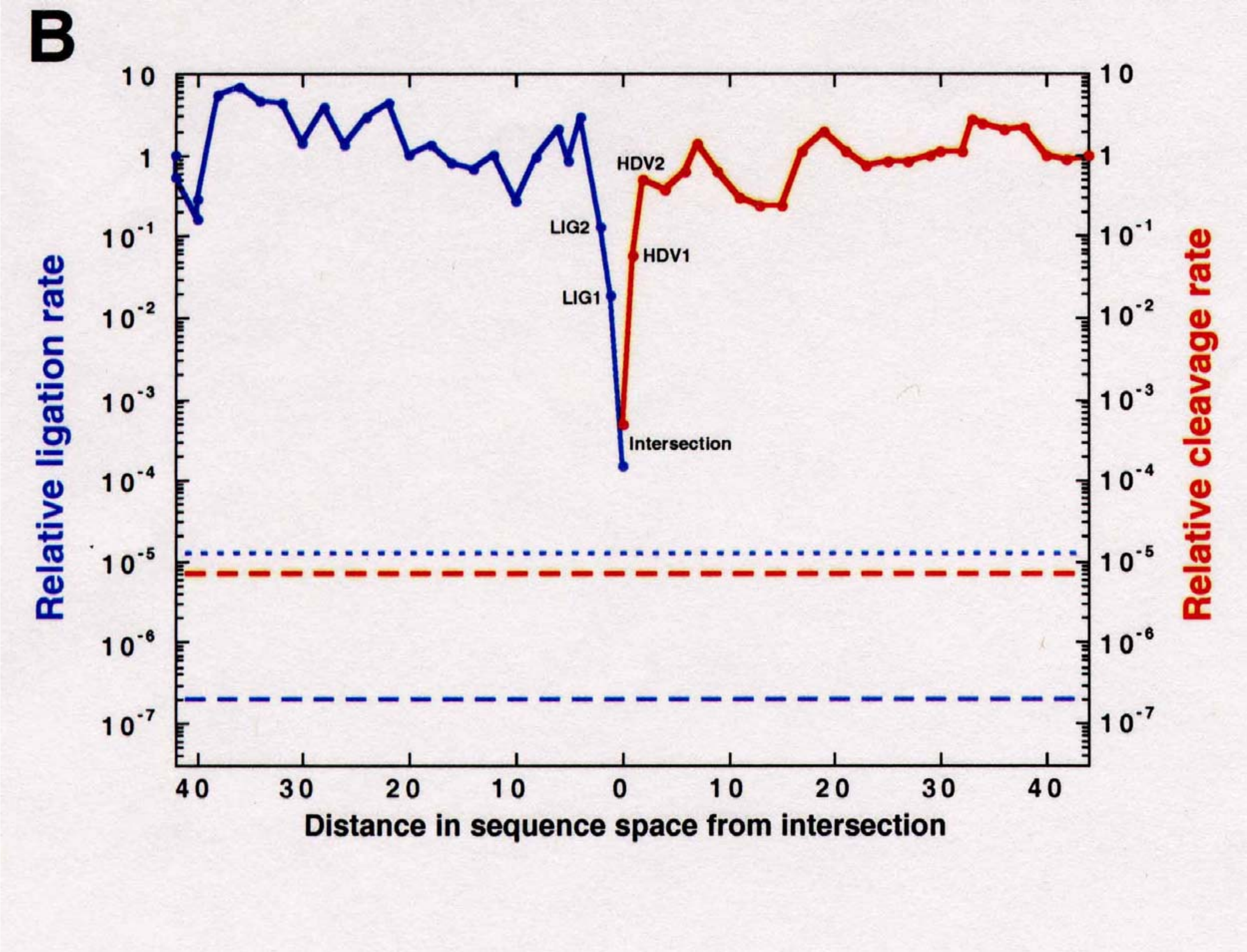


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-d-virus (**B**)

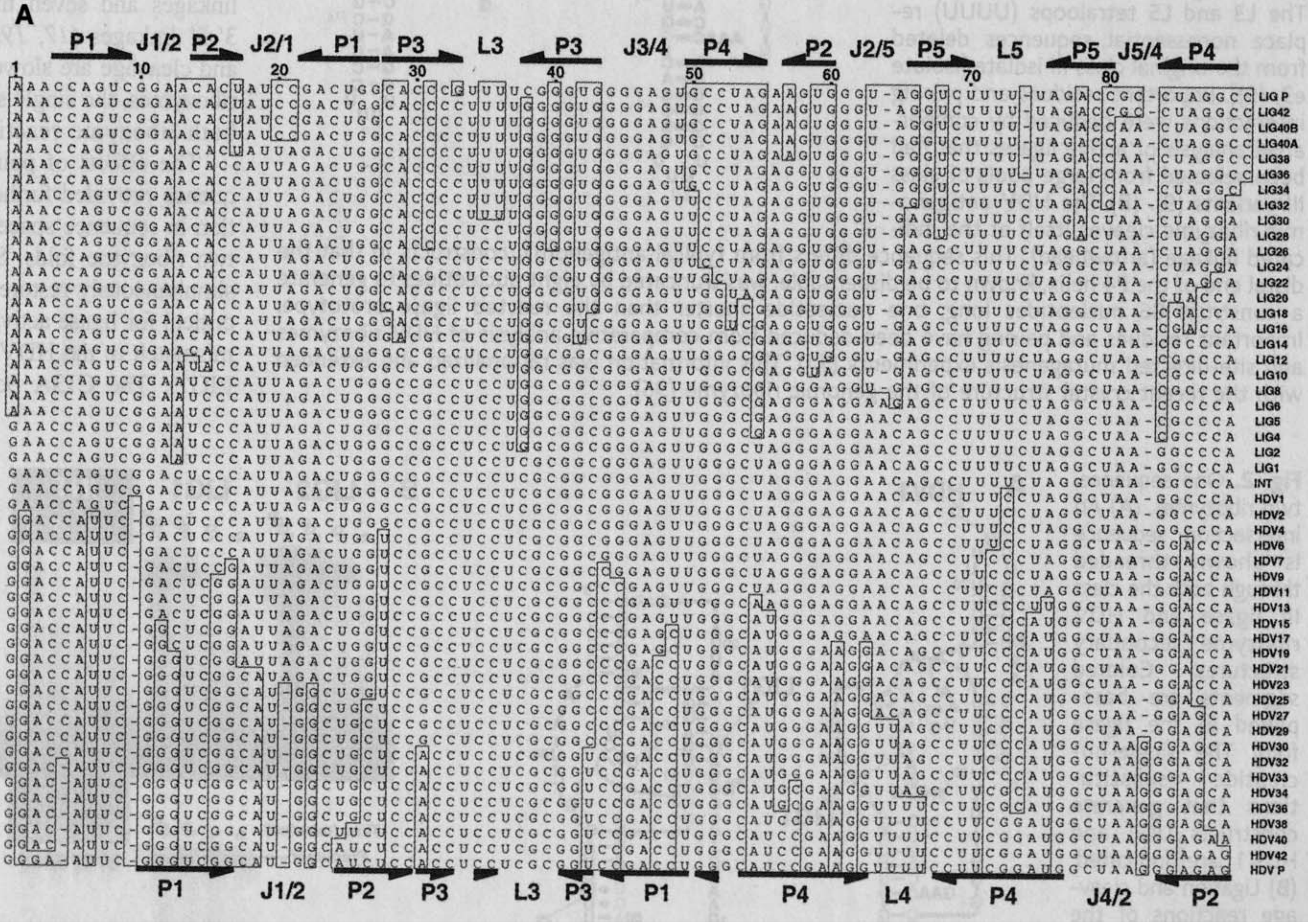


The sequence at the ***intersection***:

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



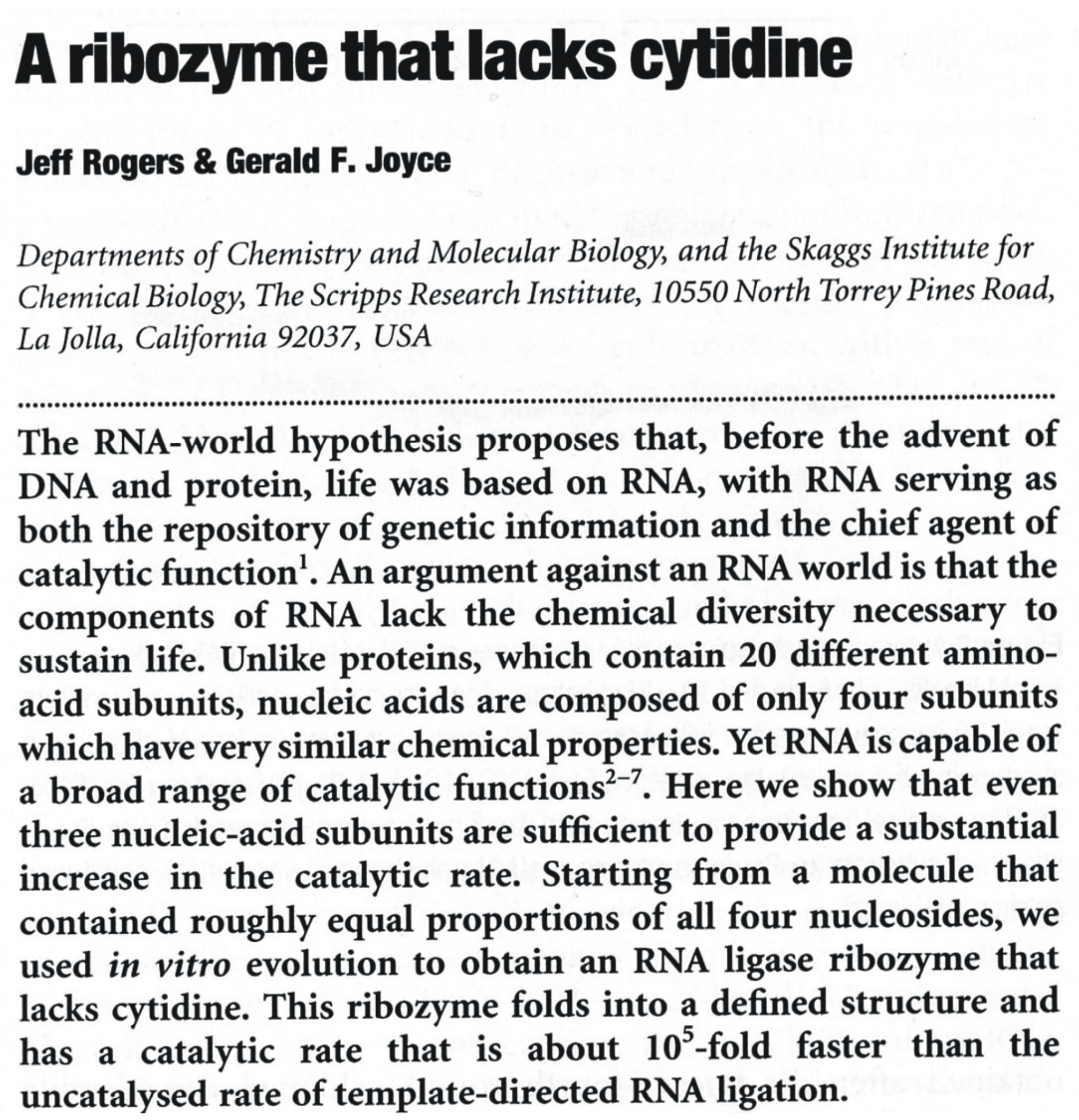
Two neutral walks through sequence space with conservation of structure and catalytic activity



```
AAACCAGUCGGAACACUAUCCGACUGGCACCCGUUUUCGGGUGGGGAGUGCCUAGAAGUGGGU-AGGUCUUUU-UAGACCGC-CUAGGCC  LIG P
AAACCAGUCGGAACACUAUCCGACUGGCACCCCUUUUGGGGUGGGGAGUGCCUAGAAGUGGGU-AGGUCUUUU-UAGACCGC-CUAGGCC  LIG42
AAACCAGUCGGAACACUAUCCGACUGGCACCCCUUUUGGGGUGGGGAGUGCCUAGAAGUGGGU-AGGUCUUUU-UAGACCAA-CUAGGCC  LIG40B
AAACCAGUCGGAACACUAUCCGACUGGCACCCCUUUUGGGGUGGGGAGUGCCUAGAAGUGGGU-GGGUCUUUU-UAGACCAA-CUAGGCC  LIG40A
AAACCAGUCGGAACACUAUUAGACUGGCACCCCUUUUGGGGUGGGGAGUGCCUAGAAGUGGGU-GGGUCUUUU-UAGACCAA-CUAGGCC  LIG38
AAACCAGUCGGAACACCAUUAGACUGGCACCCCUUUUGGGGUGGGGAGUGCCUAGAGGUGGGU-GGGUCUUUU-UAGACCAA-CUAGGCC  LIG36
AAACCAGUCGGAACACCAUUAGACUGGCACCCCUUUUGGGGUGGGGAGUGCCUAGAGGUGGGU-GGGUCUUUUCUAGACCAA-CUAGGC   LIG34
AAACCAGUCGGAACACCAUUAGACUGGCACCCCUUUUGGGGUGGGGAGUUCCUAGAGGUGGGU-GGGUCUUUUCUAGACCAA-CUAGGA   LIG32
AAACCAGUCGGAACACCAUUAGACUGGCACCCCUUUUGGGGUGGGGAGUUCCUAGAGGUGGGU-GAGUCUUUUCUAGACUAA-CUAGGA   LIG30
AAACCAGUCGGAACACCAUUAGACUGGCACCCCUCCUGGGGUGGGGAGUUCCUAGAGGUGGGU-GAGUCUUUUCUAGACUAA-CUAGGA   LIG28
AAACCAGUCGGAACACCAUUAGACUGGCACCCCUCCUGGGGUGGGGAGUUCCUAGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CUAGGA   LIG26
AAACCAGUCGGAACACCAUUAGACUGGCACGCCUCCUGGCGUGGGGAGUUCCUAGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CUAGGA   LIG24
AAACCAGUCGGAACACCAUUAGACUGGCACGCCUCCUGGCGUGGGGAGUUGCUAGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CUAGCA   LIG22
AAACCAGUCGGAACACCAUUAGACUGGCACGCCUCCUGGCGUGGGGAGUUGGUAGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CUACCA   LIG20
AAACCAGUCGGAACACCAUUAGACUGGCACGCCUCCUGGCGUGGGGAGUUGGUCGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CGACCA   LIG18
AAACCAGUCGGAACACCAUUAGACUGGGACGCCUCCUGGCGUCGGGAGUUGGUCGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CGACCA   LIG16
AAACCAGUCGGAACACCAUUAGACUGGGACGCCUCCUGGCGUCGGGAGUUGGGCGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CGCCCA   LIG14
AAACCAGUCGGAACACCAUUAGACUGGGCCGCCUCCUGGCGGCGGGAGUUGGGCGAGGUGGGU-GAGCCUUUUCUAGGCUAA-CGCCCA   LIG12
AAACCAGUCGGAAUACCAUUAGACUGGGCCGCCUCCUGGCGGCGGGAGUUGGGCGAGGUAGGU-GAGCCUUUUCUAGGCUAA-CGCCCA   LIG10
AAACCAGUCGGAAUCCCAUUAGACUGGGCCGCCUCCUGGCGGCGGGAGUUGGGCGAGGGAGGU-GAGCCUUUUCUAGGCUAA-CGCCCA   LIG8
AAACCAGUCGGAAUCCCAUUAGACUGGGCCGCCUCCUGGCGGCGGGAGUUGGGCGAGGGAGGAAGAGCCUUUUCUAGGCUAA-CGCCCA   LIG6
AAACCAGUCGGAAUCCCAUUAGACUGGGCCGCCUCCUGGCGGCGGGAGUUGGGCGAGGGAGGAACAGCCUUUUCUAGGCUAA-CGCCCA   LIG5
GAACCAGUCGGAAUCCCAUUAGACUGGGCCGCCUCCUGGCGGCGGGAGUUGGGCGAGGGAGGAACAGCCUUUUCUAGGCUAA-CGCCCA   LIG4
GAACCAGUCGGAAUCCCAUUAGACUGGGCCGCCUCCUGGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUUUCUAGGCUAA-GGCCCA   LIG2
GAACCAGUCGGAAUCCCAUUAGACUGGGCCGCCUCCUCGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUUUCUAGGCUAA-GGCCCA   LIG1
GAACCAGUCGGACUCCCAUUAGACUGGGCCGCCUCCUCGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUUUCUAGGCUAA-GGCCCA   INT
GAACCAGUCGGACUCCCAUUAGACUGGGCCGCCUCCUCGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUUCCUAGGCUAA-GGCCCA   HDV1
GAACCAGUC-GACUCCCAUUAGACUGGGCCGCCUCCUCGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUUCCUAGGCUAA-GGCCCA   HDV2
GGACCAUUC-GACUCCCAUUAGACUGGGCCGCCUCCUCGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUUCCUAGGCUAA-GGCCCA   HDV4
GGACCAUUC-GACUCCCAUUAGACUGGUCCGCCUCCUCGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUUCCUAGGCUAA-GGACCA   HDV6
GGACCAUUC-GACUCCCAUUAGACUGGUCCGCCUCCUCGCGGCGGGAGUUGGGCUAGGGAGGAACAGCCUUCCCUAGGCUAA-GGACCA   HDV7
GGACCAUUC-GACUCCGAUUAGACUGGUCCGCCUCCUCGCGGCCGGAGUUGGGCUAGGGAGGAACAGCCUUCCCUAGGCUAA-GGACCA   HDV9
GGACCAUUC-GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGAGUUGGGCUAGGGAGGAACAGCCUUCCCUAGGCUAA-GGACCA   HDV11
GGACCAUUC-GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGAGUUGGGCAAGGGAGGAACAGCCUUCCCUUGGCUAA-GGACCA   HDV13
GGACCAUUC-GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGAGUUGGGCAUGGGAGGAACAGCCUUCCCAUGGCUAA-GGACCA   HDV15
GGACCAUUC-GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGAGCUGGGCAUGGGAGGAACAGCCUUCCCAUGGCUAA-GGACCA   HDV17
GGACCAUUC-GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGAGCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA   HDV19
GGACCAUUC-GGGUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA   HDV21
GGACCAUUC-GGGUCGGCAUAGACUGGUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA   HDV23
GGACCAUUC-GGGUCGGCAU-GGCUGGUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA   HDV25
GGACCAUUC-GGGUCGGCAU-GGCUGCUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGAGCA   HDV27
GGACCAUUC-GGGUCGGCAU-GGCUGCUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAA-GGAGCA   HDV29
GGACCAUUC-GGGUCGGCAU-GGCUGCUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA   HDV30
GGACCAUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA   HDV32
GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA   HDV33
GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUAGCCUUCGCAUGGCUAAGGGAGCA   HDV34
GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGCAUGGCUAAGGGAGCA   HDV36
GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA   HDV38
GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA   HDV40
GGAC-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAG   HDV42
GGGA-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAG   HDV P
```

Sequence of mutants from the intersection to both reference ribozymes

Nature **402**, 323-325, 1999



A ribozyme that lacks cytidine

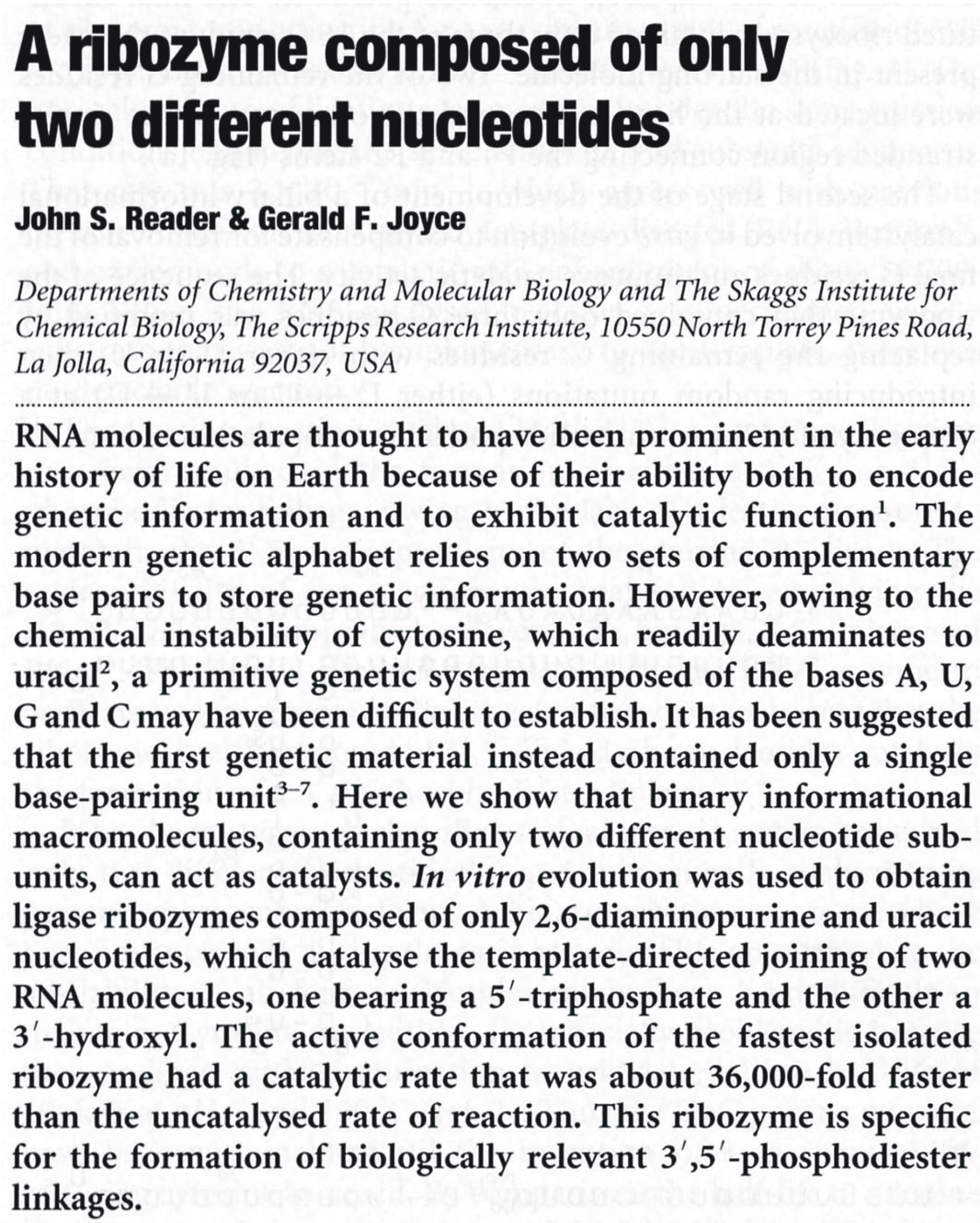
Jeff Rogers & Gerald F. Joyce

Departments of Chemistry and Molecular Biology, and the Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

The RNA-world hypothesis proposes that, before the advent of DNA and protein, life was based on RNA, with RNA serving as both the repository of genetic information and the chief agent of catalytic function[1]. An argument against an RNA world is that the components of RNA lack the chemical diversity necessary to sustain life. Unlike proteins, which contain 20 different amino-acid subunits, nucleic acids are composed of only four subunits which have very similar chemical properties. Yet RNA is capable of a broad range of catalytic functions[2–7]. Here we show that even three nucleic-acid subunits are sufficient to provide a substantial increase in the catalytic rate. Starting from a molecule that contained roughly equal proportions of all four nucleosides, we used *in vitro* evolution to obtain an RNA ligase ribozyme that lacks cytidine. This ribozyme folds into a defined structure and has a catalytic rate that is about 10^5-fold faster than the uncatalysed rate of template-directed RNA ligation.

Catalytic activity in the
AUG alphabet

Nature **420**, 841-844, 2002



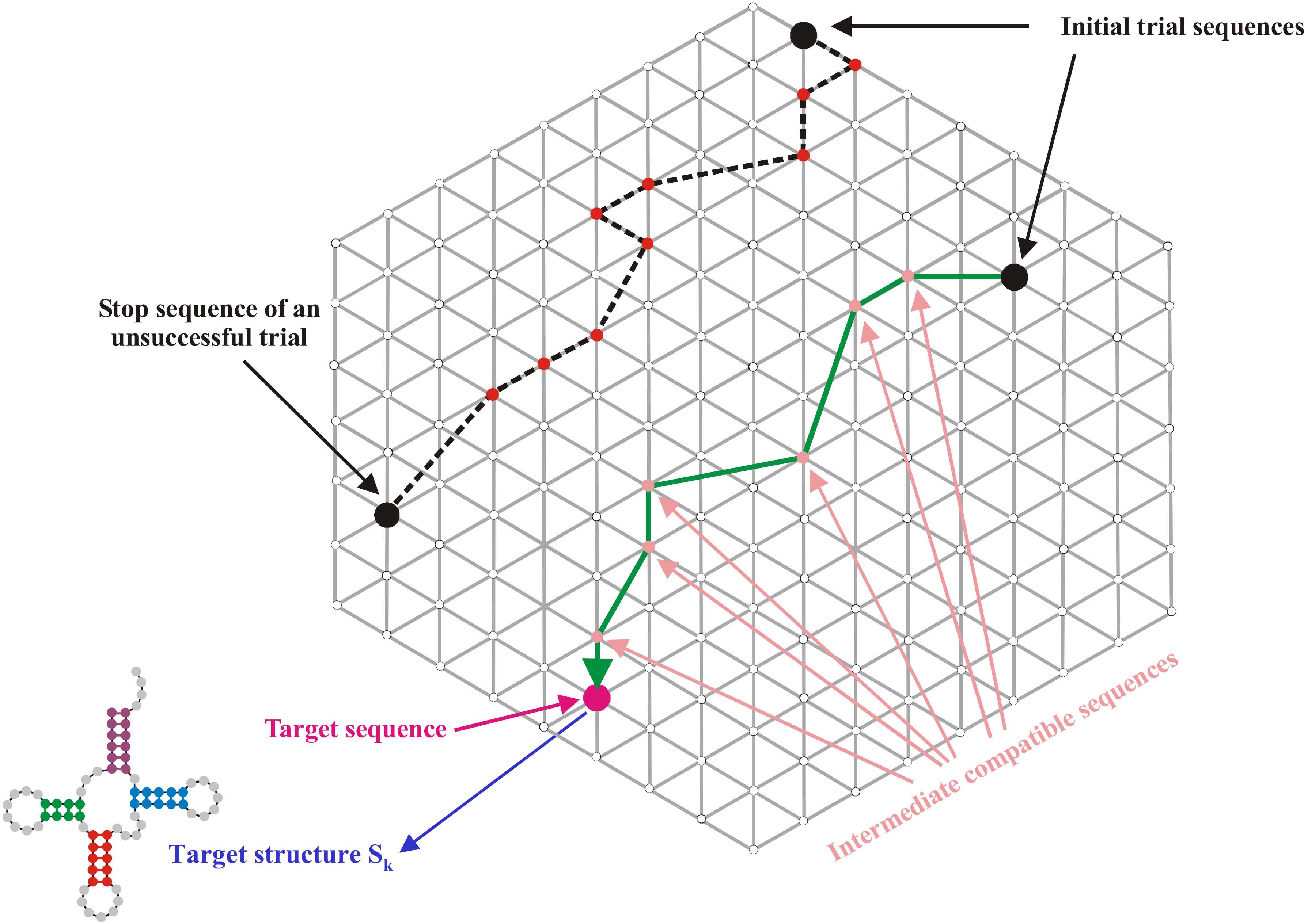
A ribozyme composed of only two different nucleotides

John S. Reader & Gerald F. Joyce

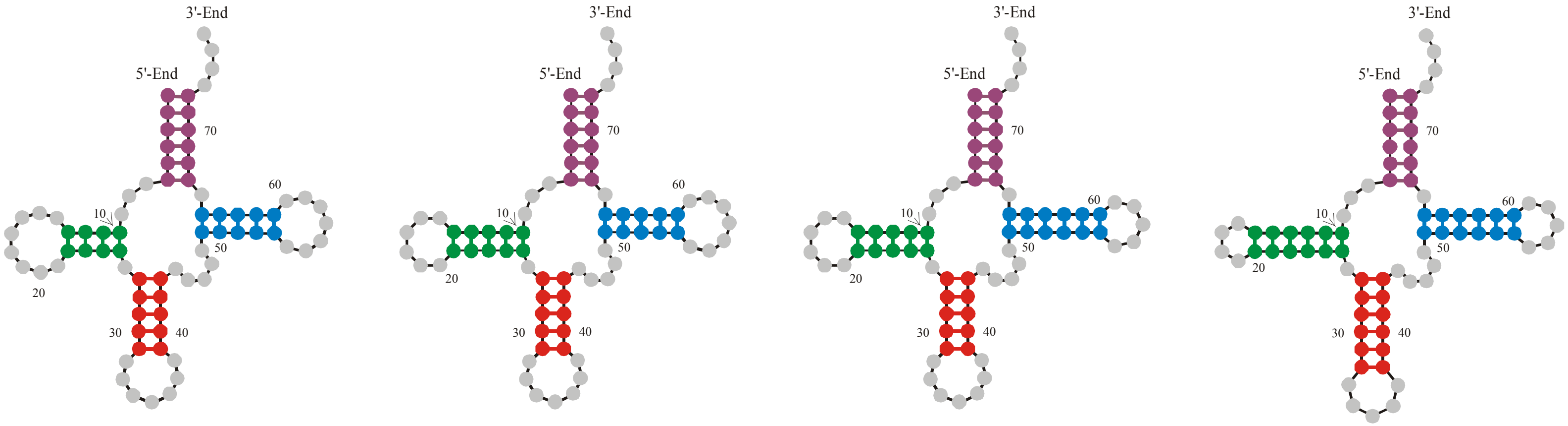
Departments of Chemistry and Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

RNA molecules are thought to have been prominent in the early history of life on Earth because of their ability both to encode genetic information and to exhibit catalytic function[1]. The modern genetic alphabet relies on two sets of complementary base pairs to store genetic information. However, owing to the chemical instability of cytosine, which readily deaminates to uracil[2], a primitive genetic system composed of the bases A, U, G and C may have been difficult to establish. It has been suggested that the first genetic material instead contained only a single base-pairing unit[3–7]. Here we show that binary informational macromolecules, containing only two different nucleotide subunits, can act as catalysts. *In vitro* evolution was used to obtain ligase ribozymes composed of only 2,6-diaminopurine and uracil nucleotides, which catalyse the template-directed joining of two RNA molecules, one bearing a 5′-triphosphate and the other a 3′-hydroxyl. The active conformation of the fastest isolated ribozyme had a catalytic rate that was about 36,000-fold faster than the uncatalysed rate of reaction. This ribozyme is specific for the formation of biologically relevant 3′,5′-phosphodiester linkages.

Catalytic activity in the
DU alphabet


Initial trial sequences
Stop sequence of an unsuccessful trial
Target sequence
Target structure S_k
Intermediate compatible sequences

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



Alphabet	Probability of successful trials in inverse folding			
AU	- -	- -	- -	0.051 Ä 0.006
AUG	- -	0.003 Ä 0.001	0.026 ± 0.006	0.374 Ä 0.016
AUGC	0.794 Ä 0.007	0.884 Ä 0.008	0.934 ± 0.009	0.982 Ä 0.004
UGC	0.548 Ä 0.011	0.628 Ä 0.012	0.697 ± 0.020	0.818 Ä 0.012
GC	0.067 Ä 0.007	0.086 Ä 0.008	0.087 ± 0.008	0.127 Ä 0.006

Accessibility of cloverleaf RNA secondary structures through inverse folding

Evolution of RNA molecules based on Qβ phage

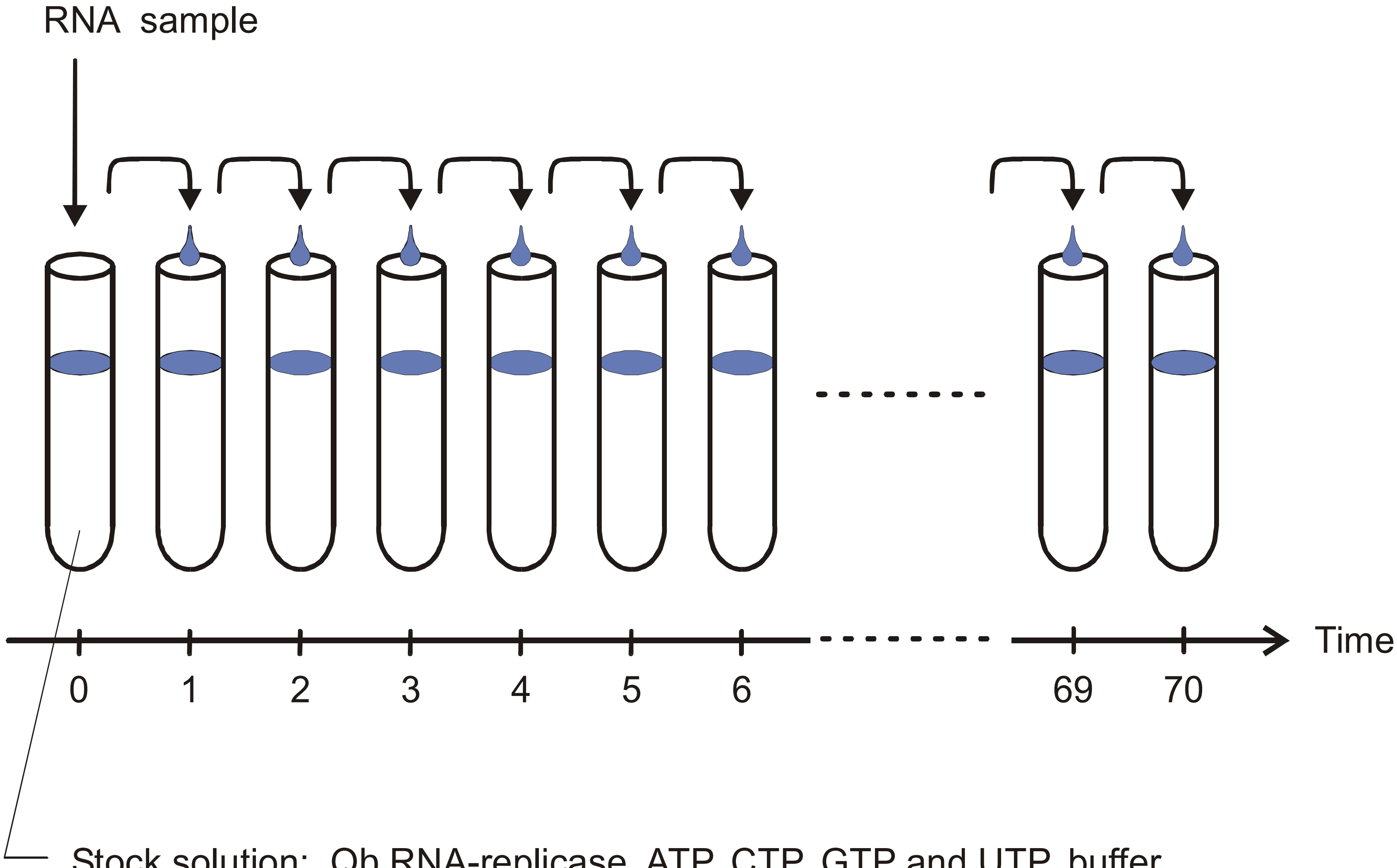
D.R.Mills, R,L,Peterson, S.Spiegelman, ***An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule***. Proc.Natl.Acad.Sci.USA **58** (1967), 217-224

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

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



Stock solution: Qb RNA-replicase, ATP, CTP, GTP and UTP, buffer

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

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

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

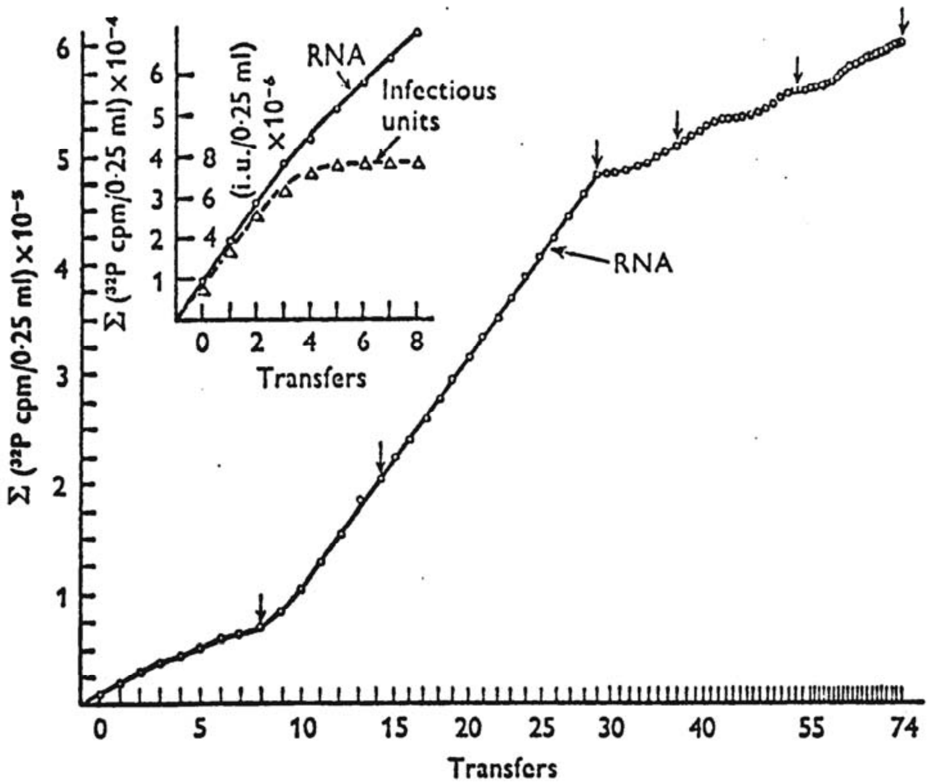
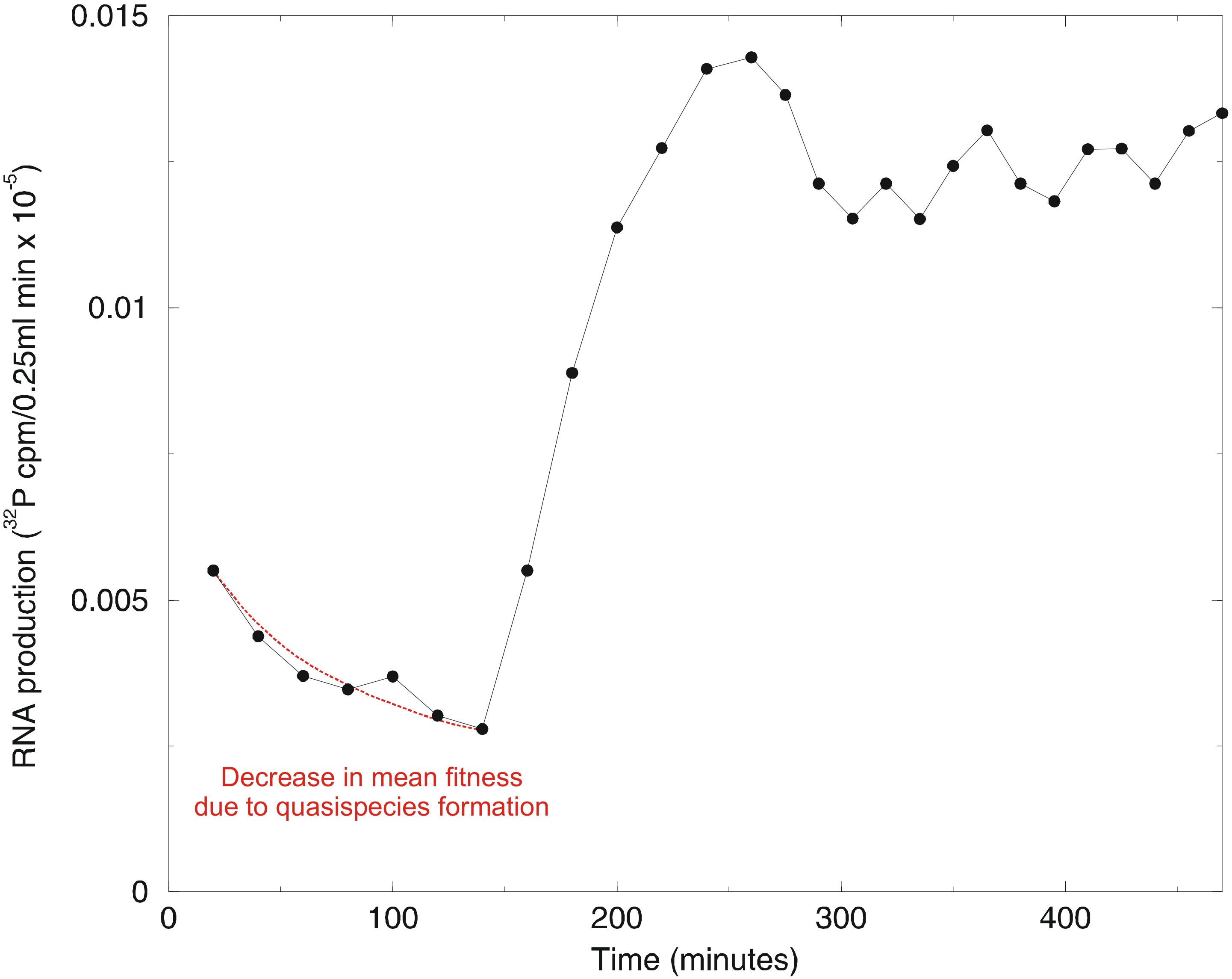


Fig. 9. Serial transfer experiment. Each 0·25 ml standard reaction mixture contained 40 μg of Qβ replicase and ^{32}P-UTP. The first reaction (0 transfer) was initiated by the addition of 0·2 μg ts-1 (temperature-sensitive RNA) and incubated at 35 °C for 20 min, whereupon 0·02 ml was drawn for counting 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).



The increase in RNA production rate during a serial transfer experiment

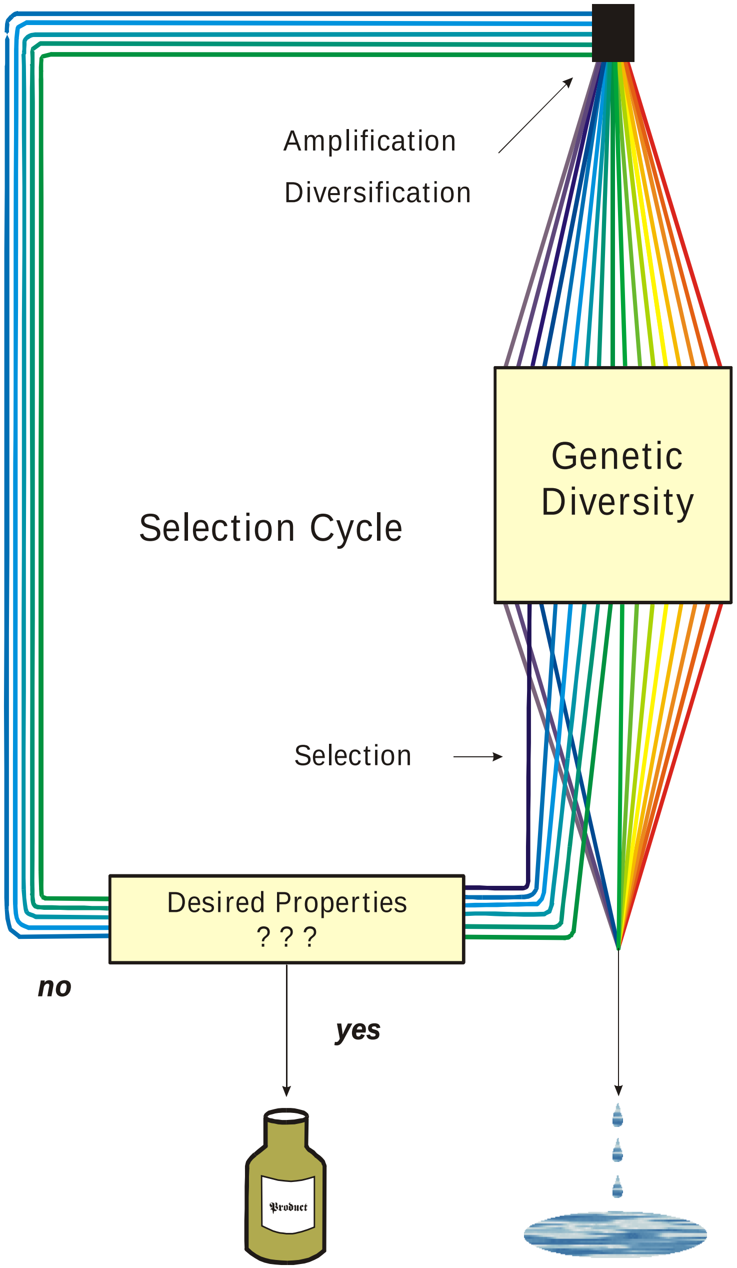
Evolutionary design of RNA molecules

D.B.Bartel, J.W.Szostak, **In vitro** ***selection of RNA molecules that bind specific ligands***. Nature **346** (1990), 818-822

C.Tuerk, L.Gold, **SELEX** - ***Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage*** **T4** ***DNA polymerase***. Science **249** (1990), 505-510

D.P.Bartel, J.W.Szostak, ***Isolation of new ribozymes from a large pool of random sequences***. Science **261** (1993), 1411-1418

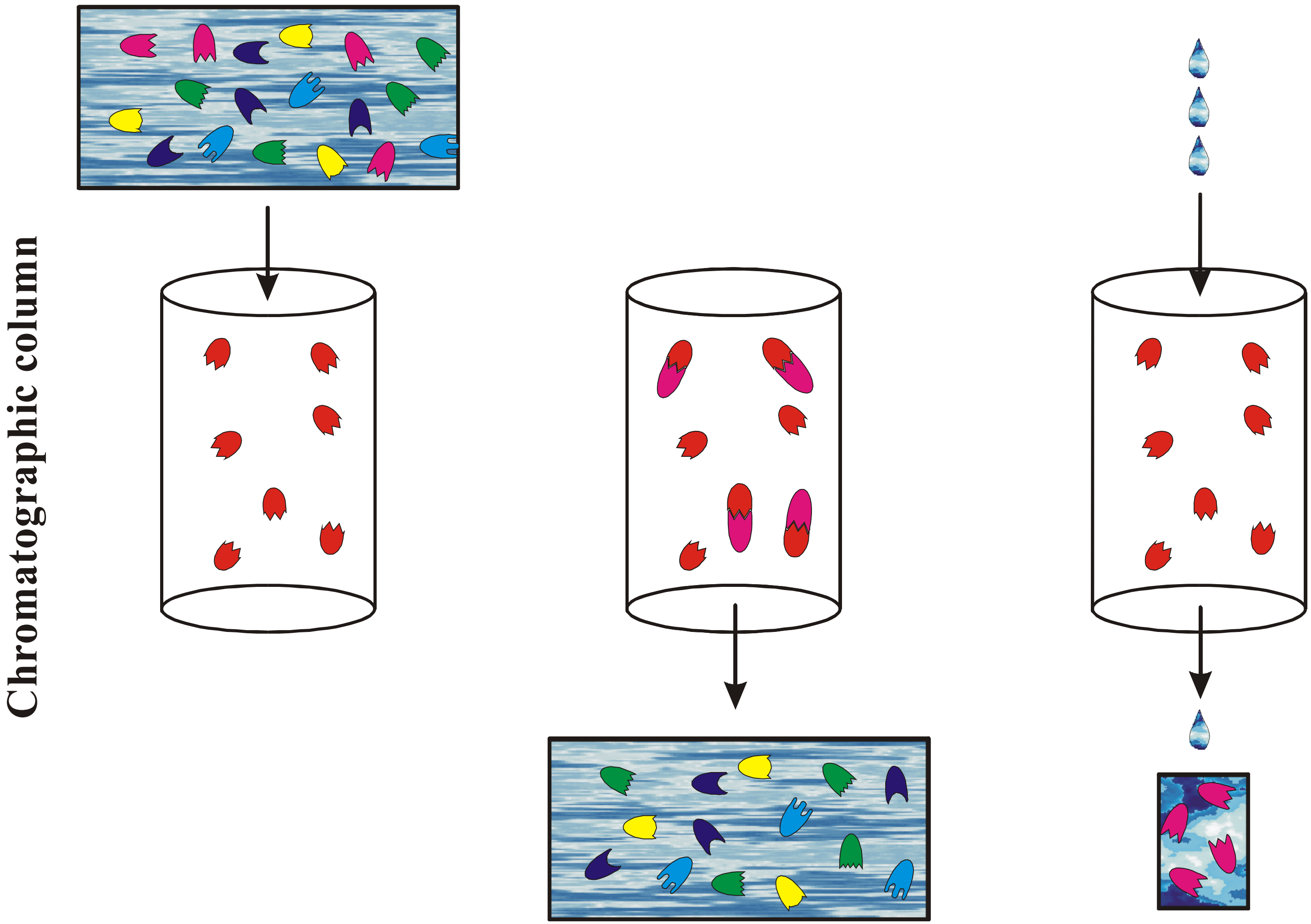
R.D.Jenison, S.C.Gill, A.Pardi, B.Poliski, ***High-resolution molecular discrimination by RNA***. Science **263** (1994), 1425-1429



Selection cycle used in applied molecular evolution to design molecules with predefined properties

Retention of binders

Elution of binders



Chromatographic column

The SELEX technique for the evolutionary design of ***aptamers***

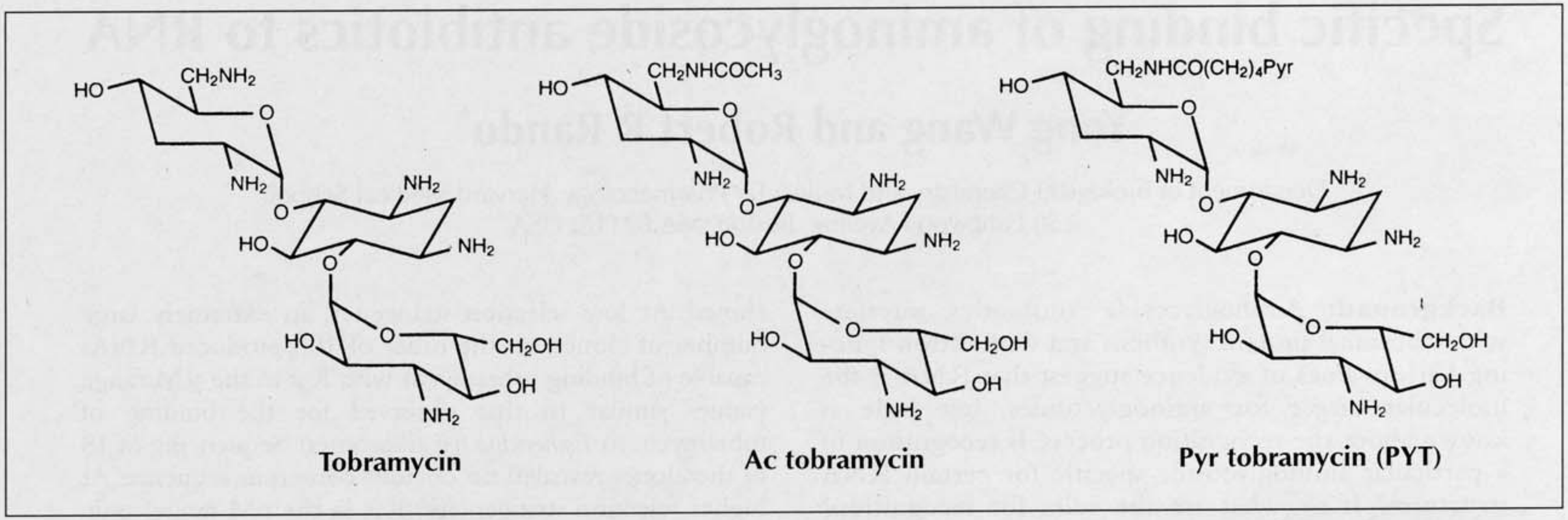
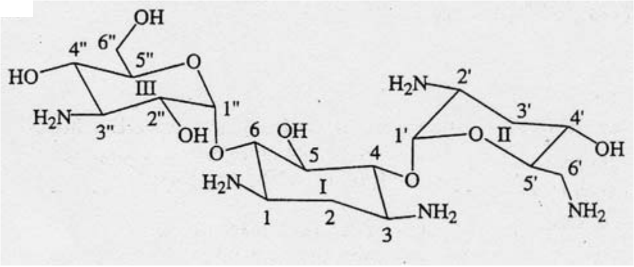


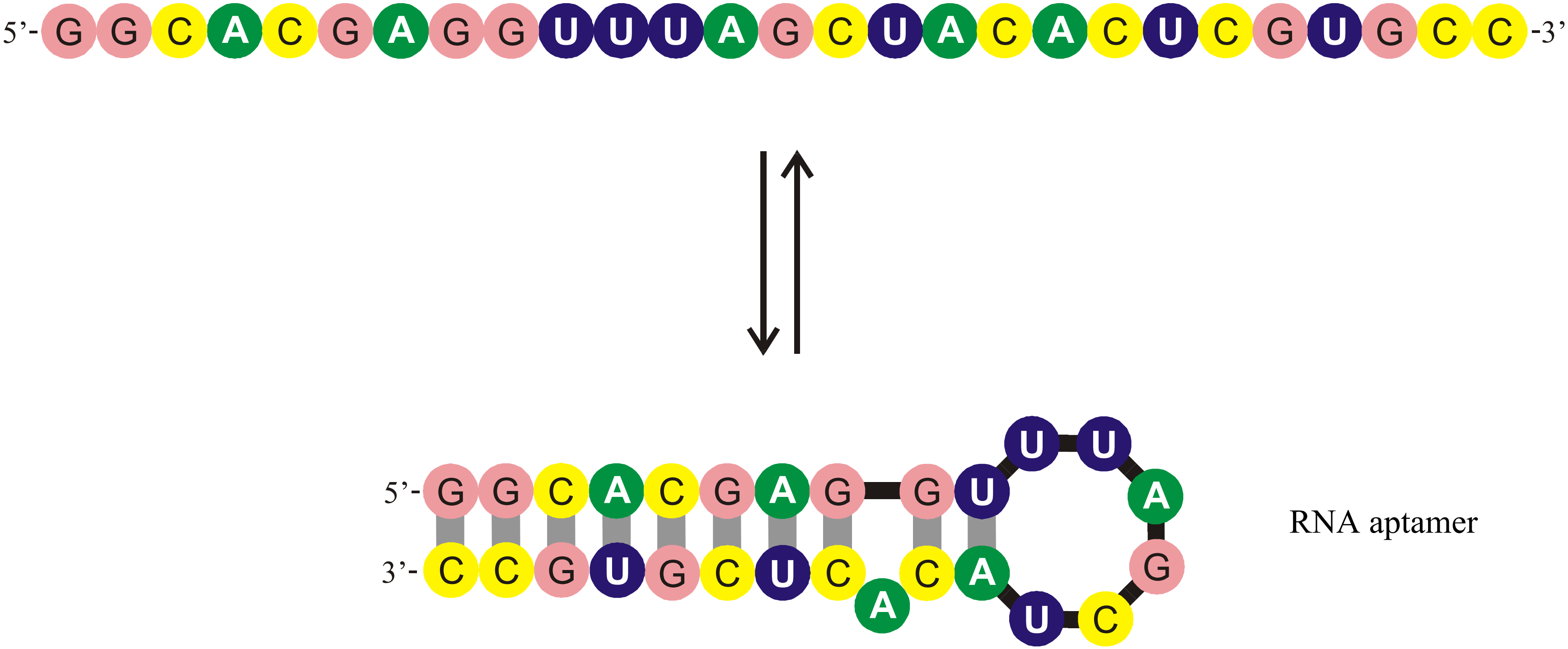
Fig. 1. Structures of tobramycin and analogs used in these studies.

Aptamer binding to aminoglycosid antibiotics: Structure of ligands

Y. Wang, R.R.Rando, ***Specific binding of aminoglycoside antibiotics to RNA***. Chemistry & Biology **2** (1995), 281-290



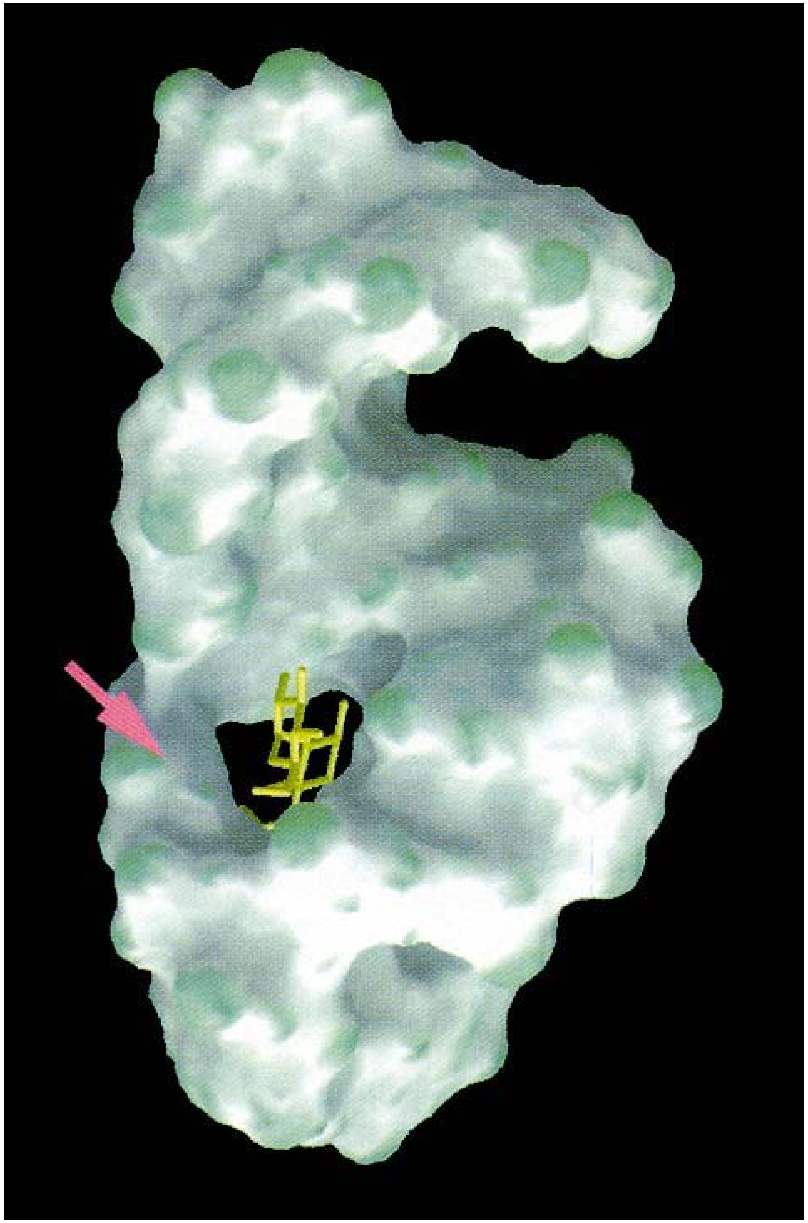
tobramycin



RNA aptamer

Formation of secondary structure of the tobramycin binding RNA aptamer

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, ***Saccharide-RNA recognition in an aminoglycoside antibiotic-RNA aptamer complex.*** Chemistry & Biology **4**:35-50 (1997)



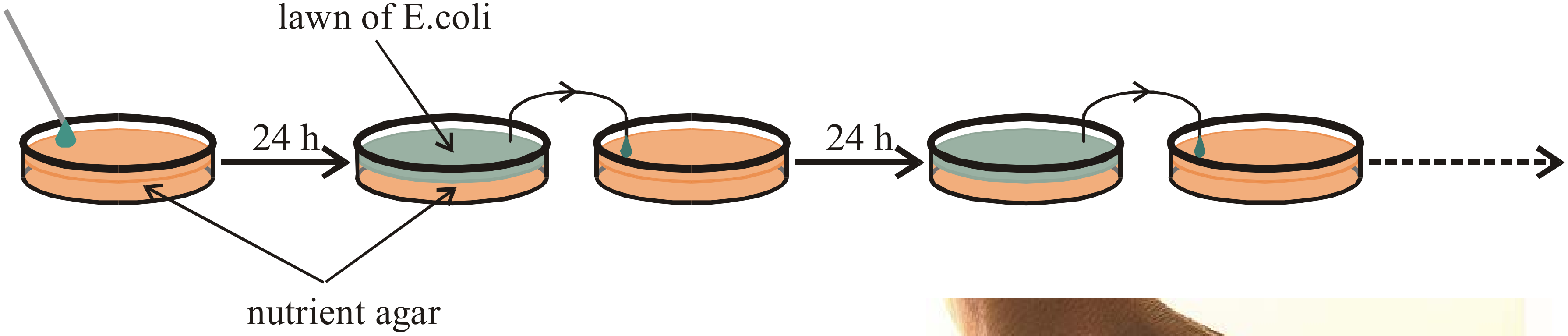
The three-dimensional structure of the tobramycin aptamer complex

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, Chemistry & Biology **4**:35-50 (1997)

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

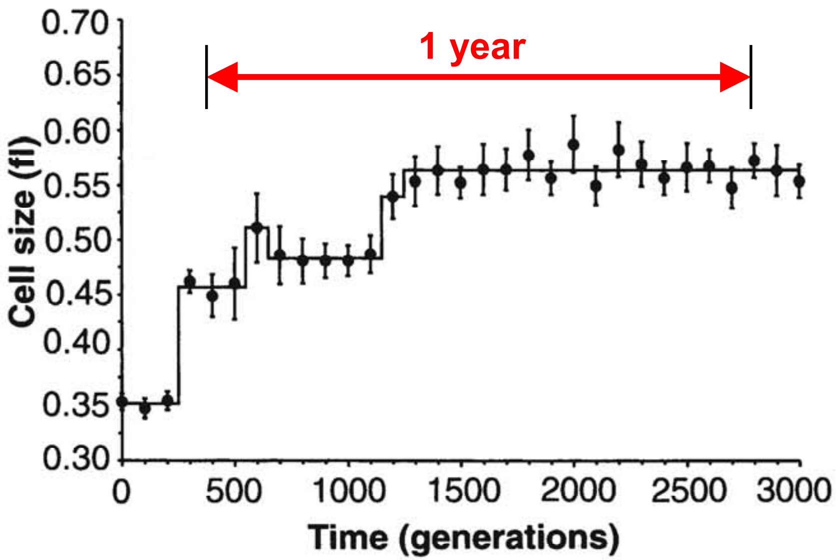


Fig. 1. Change in average cell size (1 fl = 10^{-15} L) in a population of *E. coli* during 3000 generations of experimental evolution. Each point is the mean of 10 replicate assays (22). Error bars indicate 95% confidence intervals. The solid line shows the best fit of a step-function model to these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

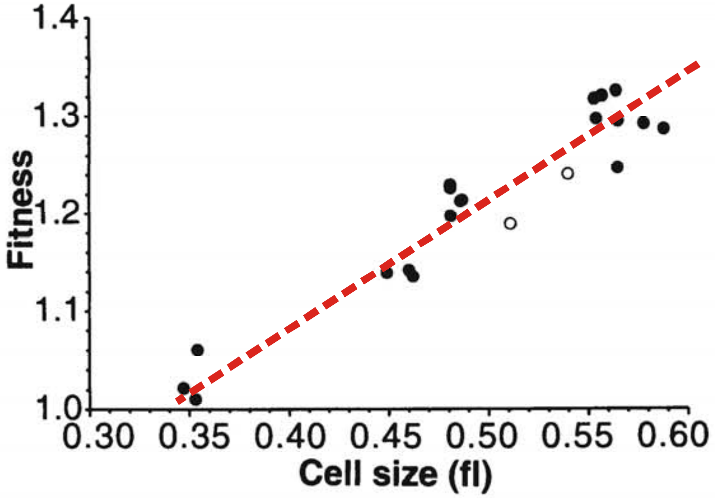
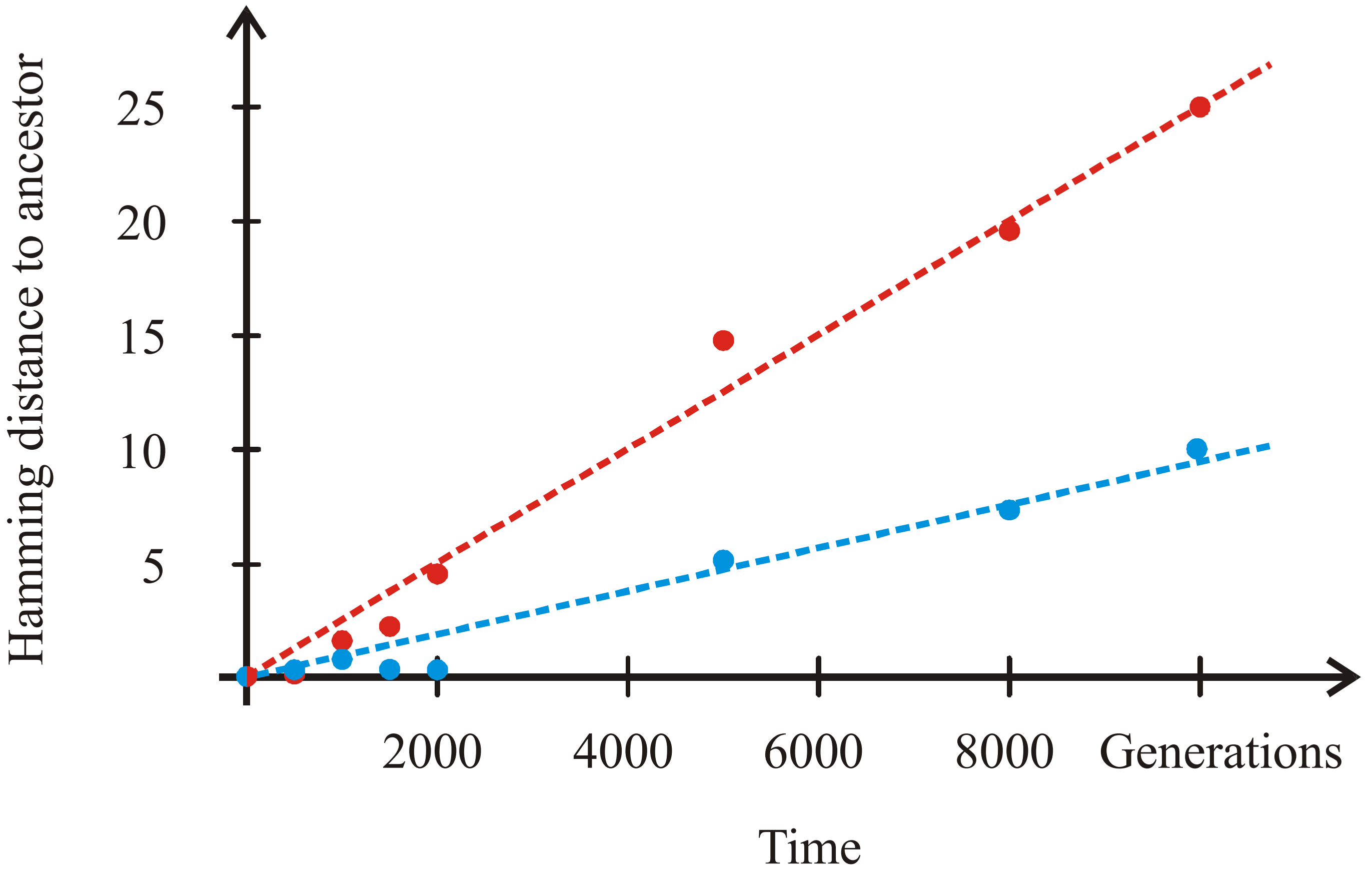


Fig. 2. Correlation between average cell size and mean fitness, each measured at 100-generation intervals for 2000 generations. Fitness is expressed relative to the ancestral genotype and was obtained from competition experiments between derived and ancestral cells (*6, 7*). The open symbols indicate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

1. Prolog – Mathematics and the life sciences in the 21st century

2. Replication kinetics of RNA molecules and evolution

3. RNA evolution *in silico*

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

5. Reference to experimental data

6. **Summary**

Concluding remarks

(i) The RNA model allows for detailed insights into evolutionary optimization and experimental tests of predictions. Evolution occurs in steps: short adaptive phases are interrupted by long quasi-stationary epochs of neutral evolution.

(ii) RNA molecules share features with much more complex elements when they are subsumed in populations. The elements of a population are related by a genetic mechanism.

(iii) Creation of information and learning by trial and error occur at the level of populations although the individual elements are subjected to random processes.

(iv) In this sense the population is more than the sum of its elements. It carries a temporary memory of its past in the form of molecular species that had been selected in previous adaptive phases.

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

Coworkers

Universität Wien

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