# **Evolution** *in vitro* and **Evolutionary Biotechnology**

Peter Schuster

# Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

RNA Secondary Structures in Dijon

Dijon, 24.– 26.06.2002

	Generation time	10 000 generations	10 <sup>6</sup> generations	10 <sup>7</sup> generations
RNA molecules	10 sec	27.8 h = 1.16 d	115.7 d	3.17 a
	1 min	6.94 d	1.90 a	19.01 a
Bacteria	20 min	138.9 d	38.03 a	380 a
	10 h	11.40 a	1 140 a	11 408 a
Higher multicelluar	10 d	274 a	27 380 a	273 800 a
organisms	20 a	20 000 a	2 × 10 <sup>7</sup> a	2 × 10 <sup>8</sup> a

Generation times and evolutionary timescales

## Evolution of RNA molecules based on $Q\beta$ 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

RNA sample



Stock solution: QV 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  $Q\beta$  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  $\mu$ g of Q $\beta$  replicase and <sup>33</sup>P-UTP. The first reaction (0 transfer) was initiated by the addition of 0.2  $\mu$ g 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



Complementary replication as the simplest copying mechanism of RNA



 $GAAUCCCGAA \rightarrow GAAUCCCGUCCCGAA$ Insertion  $GAAUCCCGAA \rightarrow GAAUCCA$ Deletion



Mutations represent the mechanism of variation in nucleic acids

$$dx_{j} / dt = \sum_{i} f_{i}Q_{ji} x_{i} - x_{j} \Phi$$

$$\Phi = \sum_{i} f_{i} x_{i}; \quad \sum_{i} x_{i} = 1; \quad \sum_{i} Q_{ij} = 1$$

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

$$p \dots Error rate per digit$$

$$d(i,j) \dots Hamming distance$$

$$between I_{i} and I_{j}$$

$$[A] = a = constant$$



Chemical kinetics of replication and mutation



The molecular quasispecies in sequence space



The increase in RNA production rate during a serial transfer experiment

Ronald Fisher's conjecture of optimization of mean fitness in populations does not hold in general for replication-mutation systems: In general evolutionary dynamics the mean fitness of populations may also decrease monotonously or even go through a maximum or minimum. It does also not hold in general for recombination of many alleles and general multi-locus systems in population genetics.

Optimization of fitness is, nevertheless, fulfilled in most cases, and can be understood as a useful heuristic.



Selection of **QV**-RNA through replication in a capillary

G.Bauer, H.Otten, J.S. McCaskill, *Proc.Natl.Acad.Sci.USA* **90**:4191, 1989 FIG. 3. Evolution of a new quasi-species along the capillary. (Upper) Front position measured using setup B. (Lower) Gel containing the fractions at 2.5-mm intervals. Regression lines are shown for the periods before and after 170 min. Aliquots  $(2 \ \mu)$  of the fractions were withdrawn after 240 min, mixed with  $2 \ \mu$  of loading buffer, boiled for 3 min to melt the double strands, immediately chilled on dry ice, and loaded into the gel slots. The polyacrylamide gel contained 13% (wt/vol) acrylamide and 0.26% N,N'-methylenebisacrylamide in running buffer (100 mM Tris borate, pH 8.3). Electrophoresis was for 6 hr at 5 V/cm at 4°C (16). Lane MNV<sub>11</sub> contains MNV<sub>11</sub> single strands (plus and minus strands) as reference. The concentration shift to new bands is centered at 12 mm where the velocity changes.

## **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



**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

# **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



The SELEX technique for the evolutionary design of *aptamers* 

# 5'-GGCACGAGGUUUAGCUACACUCGUGCC-3'

 $4^{27} = 1.801 \pm 10^{16}$  possible different sequences

Combinatorial diversity of sequences:  $N = 4^{\{}$ 

Combinatorial diversity of heteropolymers illustrated by means of an RNA aptamer that binds to the antibiotic tobramycin



### Formation of secondary structure of the tobramycin binding RNA aptamer

L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, 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) A ribozyme switch

E.A.Schultes, D.B.Bartel, *One sequence, two ribozymes: Implication for the emergence of new ribozyme folds*. Science **289** (2000), 448-452



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



The sequence at the *intersection*:

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



Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997 Elsevier Science Inc. © 1997 Society for Mathematical Biology 0092-8240/97 517.00 + 0.00

#### S0092-8240(96)00089-4

### GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES<sup>1</sup>

 CHRISTIAN REIDYS\*, †, PETER F. STADLER\*, ‡ and PETER SCHUSTER\*, ‡, §,<sup>2</sup>
 \*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 ( $\lambda$ ). 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,

### $C[s] \cap C[s'] \neq \emptyset.$

**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  $j(s, s') \cong D_m$  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*



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

Α															
P1	.11/2	P2	.12/1	D1	Da	13	D2	12/4	D4		10/E DE	uences del		-	
-				Concession and			FO	00/4	P4		J2/5 P	L5	P5 J5/4	P4	
1	10		20		30		40		50	60	)	70	80	C	
AAACCAG	UCGGA	ACACU	AUCCG	ACUGGIC	ACCCLGI	Junnc	GGGUG	GGGAG	UGCCUI	GAAGUG	GGU-AGG	UCUUUU-UA	GACCGC-C	UAGGC	LIGP
AAACCAG	UCGGA	ACACIU	AUCCG	ACUGGIC	ACCCCT	JUUUG	GGGUG	GGGAG	UGCCUI	GAAGUG	GGU-AGG	UCUUUU-UA	GACCGC-C	UAGGC	C LIG42
AAACCAG	UCGGA	ACACI	AUCCO	ACUGGIC	ACCCCC	JU UUG	GGGGUG	GGGAG	UGCCUA	GAAGUG	GGU-AGG	UCUUUU-UA	GACCAA-C	UAGGC	LIG40B
AAACCAG	UCGGA	ACACU	AUUAG	ACUGGIC	ACCCCC	TUNING	GCCTC	GGGAG	UGCCUA	GAAGUG	GGU-GGG	UCUUUU-UA	GACCAA-C	UAGGC	C LIG40A
AAACCAG	UCGGA	ACACC	AUUAG.	ACUGGIC	ACCCCI	JUUUG	GGGUG	GGGAG	UGCCUA	GAGGUG	GGU-GGGG		GACCAA-C	UAGGCO	LIG38
AAACCAG	UCGGA	ACACC	AUUAG.	ACUGGC	ACCCCI	JUUUUG	GGGUG	GGGAG	UGCCUA	GAGGUG	GGU - GGG	CUUUUUCUA	GACCAA-C	UAGGCC	LIG36
AAACCAG	UCGGA	ACACC	AUUAG.	ACUGGC	ACCCCU	JUUUG	GGGUG	GGGAG	UUCCUA	GAGGUG	GGU-GGG	CUUUUCUA	GACICAA-C	UAGGA	LIG32
AAACCAG	UCGGA	ACACC	AUUAG	ACUGGIC	ACCCCU	luuug	GGGUG	GGGAG	UUCCUA	GAGGUG	GGU-GAGI	CUUUUCUA	GACUAA-C	UAGGA	LIG30
AAACCAG	UCGGA	ACACC	AUUAG	ACUGGIC	ACCCCC	ICCUG	GGGUG	GGGAG	UUCCUA	GAGGUG	GGU-GAGU	JCUUUUCUA	GACUAA-C	UAGGA	LIG28
AAACCAG	UCGGA	ACACC	AUUAG	ACUGGIC	ACCCCC	IC C DIG	GGGGUG	GGGAG	UUCCUA	GAGGUG	GGU-GAGO	CCUUUUCUA	GGCUAA-C	UAGGA	LIG26
AAACCAG	UCGGA	ACACC	AUUAG	ACUGGIC	ACGCCI	ICCUG	GCGUG	GGGAG	UUCCUA	GAGGUG	GGU-GAGO	CUUUUCUA	GGCUAA-C	UAGGA	LIG24
AAACCAG	UCGGA	ACACC	AUUAG	ACUGGIC	ACGCCU	CCUG	GCGUG	GGGAG	UUGGUA	GAGGUG	GGU-GAGO	CUUUUUCUA	GGCUAA-C	UAGICA	LIG22
AAACCAG	UCGGA	ACACC	AUUAGI	ACUGGLC	ACGCCU	CCUG	GCGUG	GGGAG	UUGGUC	GAGGUG	GGU-GAGO	спнинсих	GGCUAA-C	GACCA	LIGIO
AAACCAG	UCGGA	ACACC	AUUAGI	ACUGGG	ACGCCU	ICCUG	GCGUC	GGGAG	UUGGUC	GAGGUG	GGU-GAGO	CUUUUCUA	GGCUAA-C	GACCA	LIGIS
AAACCAG	UCGGA	ACACC	AUUAGI	ACUGGGL	ACGCCU	ICCUG	GCGUC	GGGAG	UUGGGC	GAGGUG	GGU-GAGO	CUUUUCUA	GGCUAA-C	GCCCA	LIG14
AAACCAG	UCGGA	ACACC	AUUAGI	ACUGGG	CCGCCU	CCUG	GCGGC	GGGAG	UUGGGC	GAGGUG	GGU-GAGO	CUUUUCUA	GGCUAA-C	GCCCA	LIG12
AAACCAG	UCGGAL	AUCCC	AUUAGI	ACUGGG	CCGCCU	CCUG	GCGGC	GGGAG	UUGGGC	GAGGUA	GGU-GAGO	CUUUUCUA	GGCUAA-C	GCCCA	LIGIO
AAACCAG	UCGGA	AUCCC	AUUAGA	ACUGGG	CCGCCU	CCUG	GCGGC	GGGAG	UUGGGC	GAGGGA	GGU-GAGO	CUUUUCUA	GGCUAA-C	GCCCA	LIG8
LAAACCAG	UCGGA	AUCCC	AUUAGA	ACUGGG	CCGCCU	CCUG	GCGGC	GGGAGI	UNGGGC	GAGGGA	GGAALGAG	CUUUUCUA	GGCUAA-C	GCCCA	LIG6
GAACCAG	UCGGA	AUCCC	AUUAGI	ACUGGG	CCGCCU	CCUG	GCGGC	GGGAGI	UUGGGC	GAGGGA	GGAACAGO	CUUUUUCUA	GGCUAA-C	GCCCA	LIGS
GAACCAG	UCGGAL	AUCCC	AUUAGI	ACUGGG	CCGCCU	cculd	GCGGC	GGGAGI	UUGGGC	UAGGGA	GGAACAGO	CUUUUCUA	GGCUAA-G	GCCCA	1102
GAACCAG	UCGGAL	AUCCC	AUUAGI	ACUGGG	ccgccu	ccuc	GCGGC	GGGAGI	UUGGGC	UAGGGA	GGAACAGO	CUUUUCUA	GGCUAA-G	GCCCA	LIGI
GAACCAG	UCGGA	CUCCC	AUUAGA	CUGGG	CCGCCU	ccuc	GCGGC	GGGAGI	JUGGGC	UAGGGA	GGAACAGO	CUUUUCUA	GGCUAA-G	GCCCA	INT
GAACCAG	UCLGA	CUCCC	AUUAGA	CUGGG	CCGCCU	CCUC	GCGGC	GGGAGI	JUGGGC	UAGGGA	GGAACAGO	CUUUCCUA	GGCUAA-G	GCCCA	HDV1
GGACCAU	UC-GA	CUCCC	AUUAGA	CUGGG	CCGCCU	CCUC	GCGGC	GGGAG	JUGGGC	UAGGGA	GGAACAGO	CUUUCCUA	GGCUAA-G	GCCCA	HDV2
GGACCAU	UC-GAG	CUCCC.	AUUAGA	CUGGU	CCGCCU	CCUC	GCGGC	GGGAGI	INGGGC	UAGGGA	GGAACAGC	CUUUCCUA	GGCUAA-G	GCCCA	HDV4
GGACCAU	UC-GAO	UCCC.	AUUAGA	CUGGU	CCGCCU	CCUC	GCGGC	GGGAGI	JUGGGC	HAGGGA	GGAACAGC	CUUCCCUA	GGCUAA-G	GACCA	HDV6
GGACCAU	UC-GAC	cucco.	AUUAGA	CUGGU	ccgccu	ccuc	GCGGC	CGGAGI	JUGGGC	UAGGGA	GGAACAGO	CUUCCCUA	GGCUAA-G	GACCA	HOVO
GGACCAU	UC-GAC	CUCIGG	AUUAGA	CUGGU	CCGCCU	CCUC	GCGGC	CCGAGI	JUGGGC	UAGGGA	GGAACAGO	CUUCCCUA	GGCUAA-G	GACCA	HDV11
GIGACCAU	UCI-GAC	UCIGG	AUUAGA	CUGGU	cccccu	ccuc	GCGGC	CCGAGI	JUGGGC	AAGGGA	GGAACAGC	cuucccuu	GGCUAA-G	GACCA	HDV13
GGACCAID	I C - GA	TOCGG	AUUAGA	CUGGIU	CCGCCU	CCUC	GCGGC	CCGAGL	JUGGGC	AUGGGA	GGAACAGC	CUUCCCAU	GGCUAA-G	GACCA	HDV15
GGACCAU	UC-GGC	UCGG	AUUAGA	CUGGU	CGCCU	CCUC	GCGGC	CCGAGO	JUGGGC	AUGGGA	GGAACAGC	CUUCCCAU	GGCUAA-G	GACCA	HDV17
GGACCAUI	JC-GGC	UCGG	AUUAGA	CUGGU	CGCCU	CCUC	acaad	COGAGO	UGGGGC	AUGGGA	AGGACAGC	CUUCCCAU	GGCUAA-G	GACCA	HDV19
GGACCAUI	JC-GGC	UCGG	CAUAGA	CUGGU	CCGCCU	CCUCI	GCGGC	CCGACO	UGGGGC	AUGGGA	AGGACAGC	CUUCCCAU	GGCUAA-G	GACCA	HDV21
GGACCAUI	JC-GGC	UCGG	CAU-GG	CUGGUO	CCGCCU	CCUC	GCGGC	CCGACO	UGGGC	AUGGGAL	AGGACAGC	CUUCCCAU	GGCUAA-G	GACCA	HDV23
GGACCAUI	JC-GGG	UCGG	CAU-GG	CUGCUG	CCGCCU	ccuco	GCGGC	CCGACO	UGGGC	AUGGGA	AGGACAGC	CUUCCCAU	GGCUAA-G	GAGCA	HDV27
GGACCAUI		UCGGG	CAU-GG	CUGCUG	cccccu	CCUC	GCGGC	CCGACO	UGGGC	AUGGGA	AGGUUAGC	CUUCCCAU	GGCUAA-G	GAGCA	HDV29
GGACCAUL		UCGGG	CAU-GG	CUGCUC	cgccu	CCUCO	GCGGC	CCGACC	UGGGC	AUGGGA	AGGUUAGC	CUUCCCAU	GGCUAAGG	GAGCA	HDV30
GIGACTAUT	IC-GGG	UCGGG	ATT	CHGCHC	CACCU	CCUCC	GCGGU	COGAICO	UGGGC	AUGGGA	AGGUUAGC	CUUCCCAU	GGCUAAGG	GAGCA	HDV32
GGAC-AUL	JC-GGG	UCGGG	AU-GG	CUGCUC	CACCU	CCUC	CGGU	CCGACC	UGGGGC	AUGGGAL	AGGUUAGC	CUUCCICAU	GGCUAAGG	GAGCA	HDV33
GGAC-AUT	C-GGG	UCGGG	CAU-GG	CUGCUC	CACCU	ccuco	GCGGU	CCGACC	UGGGC	AUGCGAL	GGUUUUUC	CUUCGCAU	GGCUAAGG	GAGCA	HDV34
GGAC-AUT	JC-GGG	UCGGO	CAU-GG	CUGCUC	CACCU	ccuco	GCGGU	CCGACC	UGGGC	AUCCIGAL	GGUUUUUC	CUNCGGAU	GGCUAAGG	GAGCA	HOV36
GGAC-AUU	C-GGG	UCGGG	CAU-GG	CUUCUC	CACCU	ccuco	GCGGU	CCGACC	UGGGC	AUCCGAL	GGUUUUC	CUUCGGAU	GGCUAAGG	GAGAA	HDV40
GGAC - AUT	G-GGG	UCIGGO	AU-GG	CAUCUC	CACCU	ccuca	GCGGU	CCGACC	UGGGC	AUCCGAL	GGUUUUC	CUUCGGAU	GGCUAAGG	GAGAG	HDV42
GO CA - AU	CELOIG G	CIG G C	a ou de	CAUCUC	CLAICCU	ccuca	C G GIU	CCIGAICC	UGGGC	AUCCIGAZ	GGUUUUC	CUUCGGAU	GGCUAAGG	GAGAG	HDV P
	P	1	J1/2	P2	P3	L	3 P3	P1	A	P4	14	P4	.14/2	P2	
							15 15 M	107			and the second s			1 44	

Sequence of mutants from the intersection to both reference ribozymes

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

### PETER SCHUSTER<sup>1, 2, 3</sup>, WALTER FONTANA<sup>3</sup>, PETER F. STADLER<sup>2, 3</sup> and IVO L. HOFACKER<sup>2</sup>

<sup>1</sup> Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany
 <sup>2</sup> Institut für Theoretische Chemie, Universität Wien, Austria
 <sup>3</sup> 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 nonuniform 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. 1993a; 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 279

© 1994 The Royal Society

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

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

Sir Peter Medawar, 1985

# Coworkers

Walter Fontana, Santa Fe Institute, NM

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

Peter Stadler, Universität Leipzig, GE

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

Bärbel Stadler, Andreas Wernitznig, Universität Wien, AT Michael Kospach, 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