

RNA evolution *in vitro* and *in silico*

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Österreich
und
The Santa Fe Institute, Santa Fe, New Mexico, USA



Institute of Structural Molecular Biology, UCL
London, 11.11.2009

Web-Page for further information:

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

Diagram illustrating the structure of a DNA double helix, showing three helices (Helix I, Helix II, Helix III) and a scale bar indicating 5 Å.



Evolutionary biotechnology
RNA aptamers, artificial ribozymes,
allosteric ribozymes



RNA

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

RNA as **working copy** of genetic information

Gene regulation by small interfering RNAs

Riboswitches controlling transcription and translation through metabolites

1. RNA replication *in vitro* and *in vivo*
2. Evolution of RNA molecules
3. RNA sequences and structures
4. Evolutionary optimization of RNA structure

1. RNA replication *in vitro* and *in vivo*
2. Evolution of RNA molecules
3. RNA sequences and structures
4. Evolutionary optimization of RNA structure

Evolution of RNA molecules based on Q β 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. Weissmann, *The making of a phage*. FEBS Letters **40** (1974), S10-S18

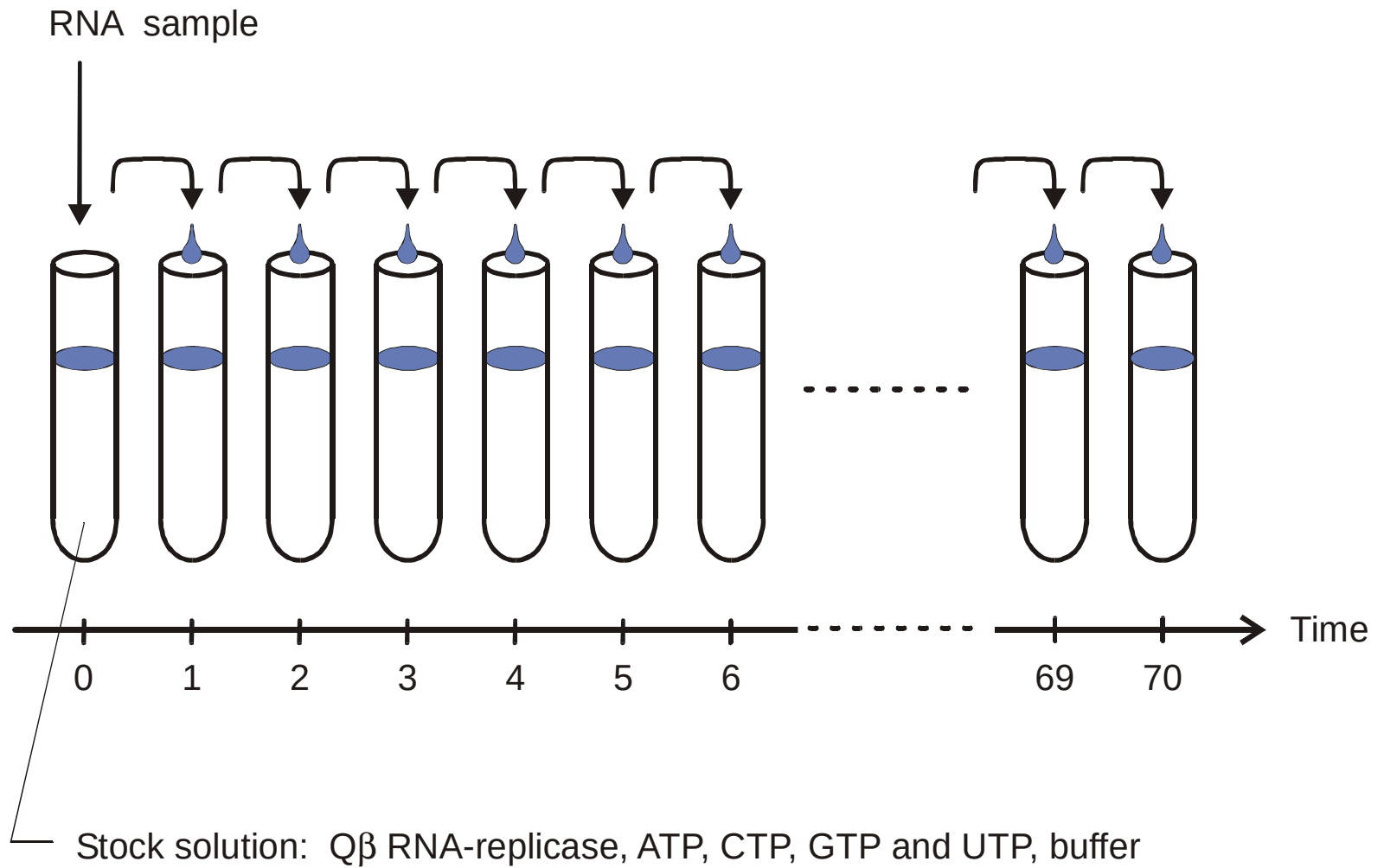
C.K.Biebricher, *Darwinian selection of self-replicating RNA molecules*. Evolutionary Biology **16** (1983), 1-52

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

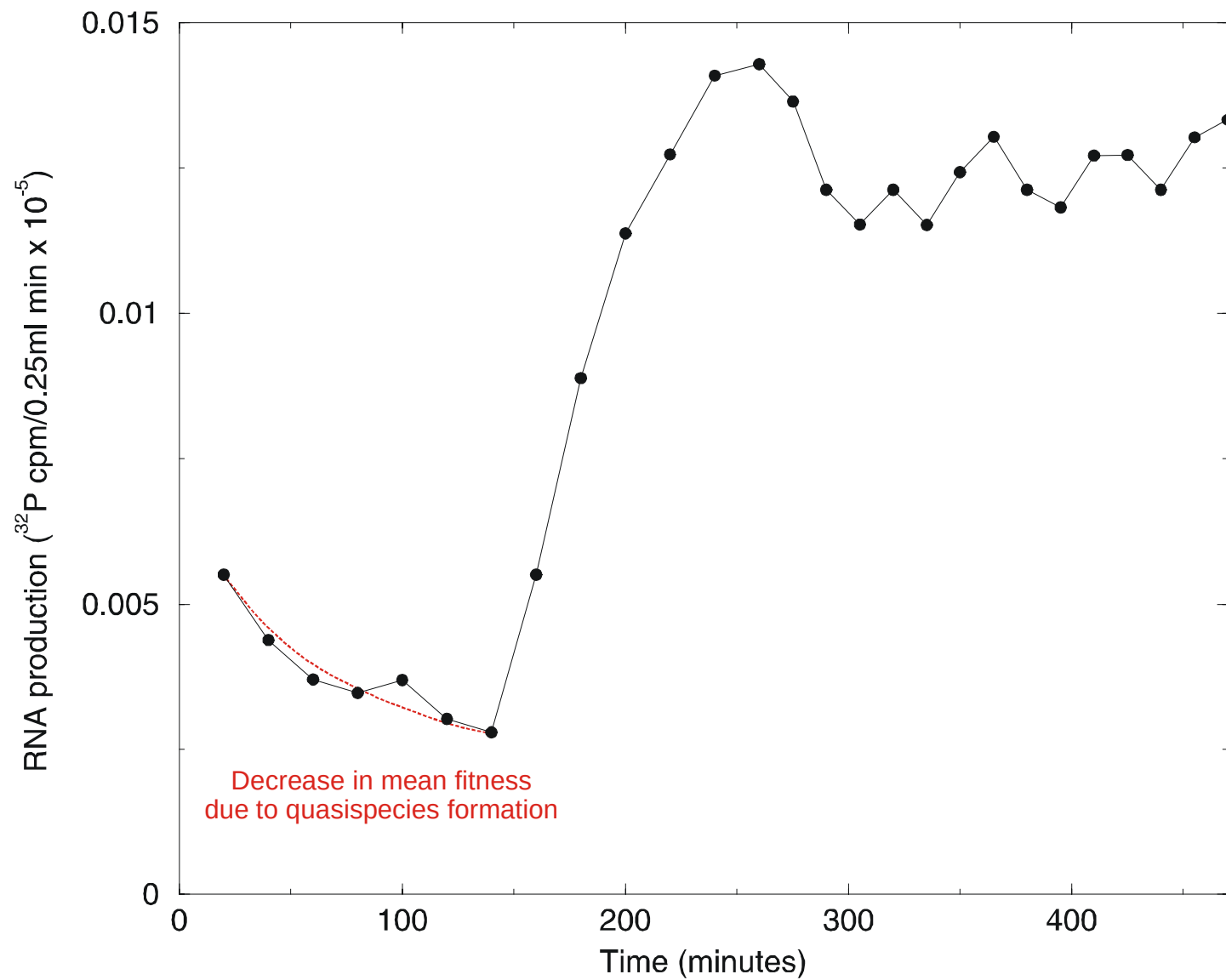
C.K.Biebricher, W.C.Gardiner, *Molecular evolution of RNA in vitro*. Biophysical Chemistry **66** (1997), 179-192

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

F.Öhlenschläger, M.Eigen, *30 years later – A new approach to Sol Spiegelman's and Leslie Orgel's in vitro evolutionary studies*. Orig.Life Evol.Biosph. **27** (1997), 437-457



Application of serial transfer to RNA evolution *in vitro*

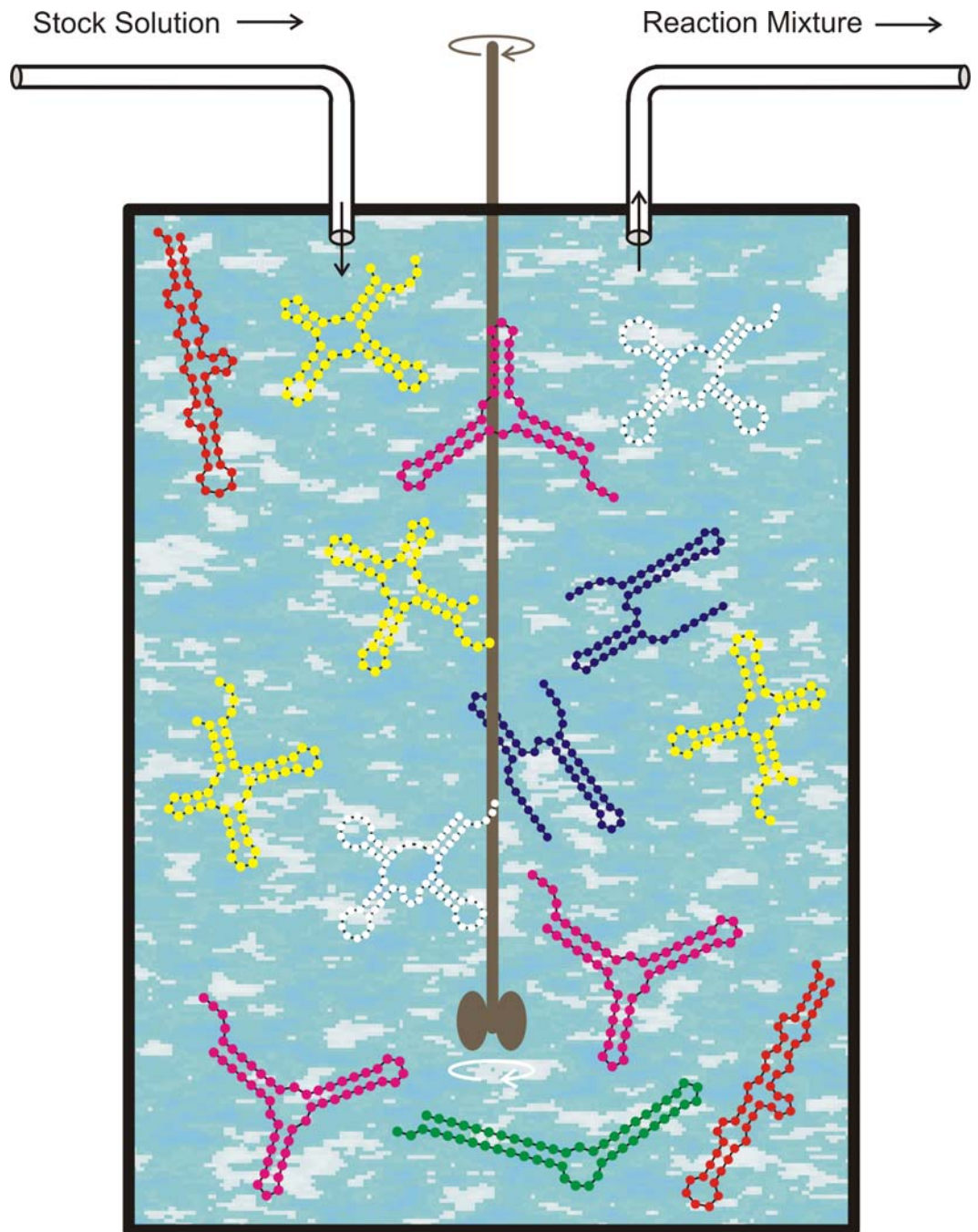


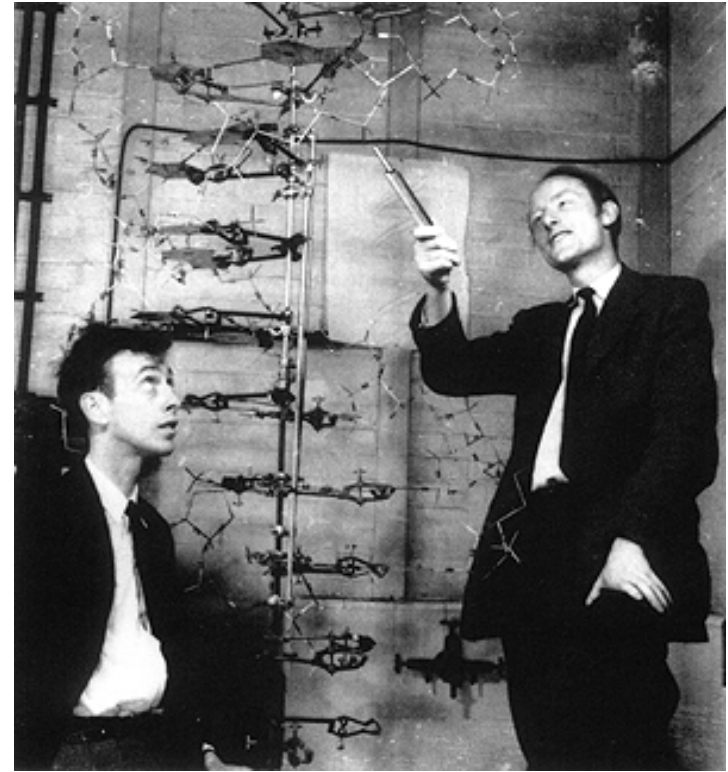
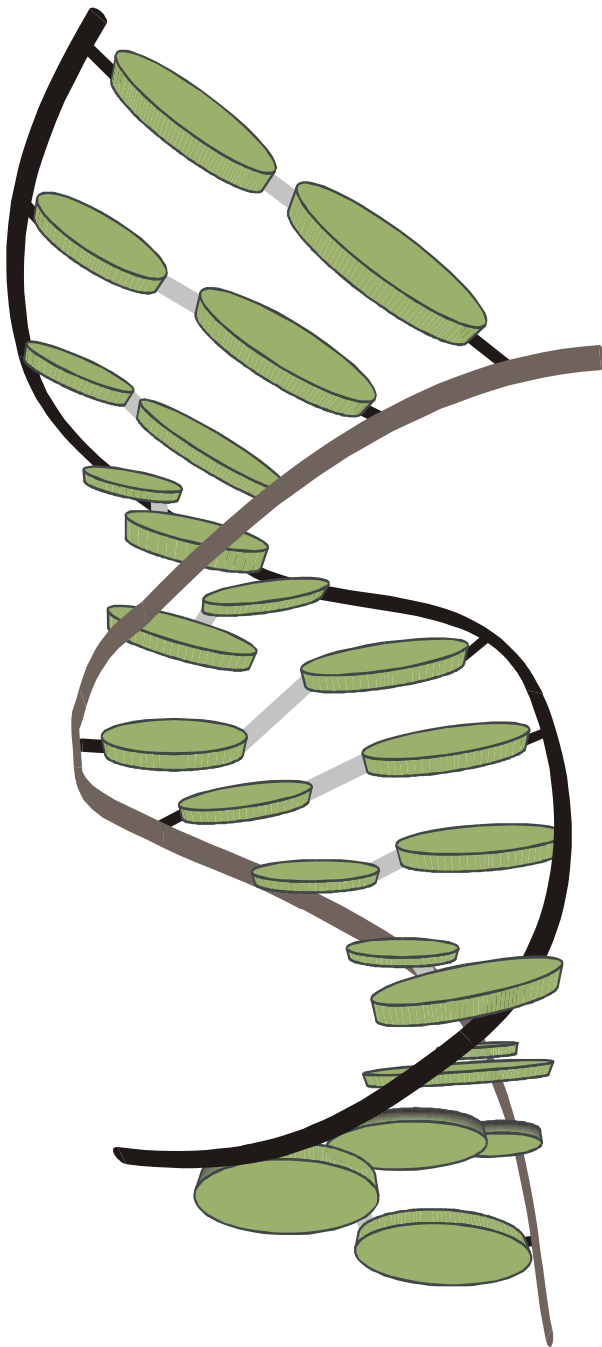
The increase in RNA production rate during a serial transfer experiment

Stock solution:

activated monomers, **ATP, CTP, GTP, UTP (TTP)**;
a replicase, an enzyme that performs
complementary replication;
buffer solution

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



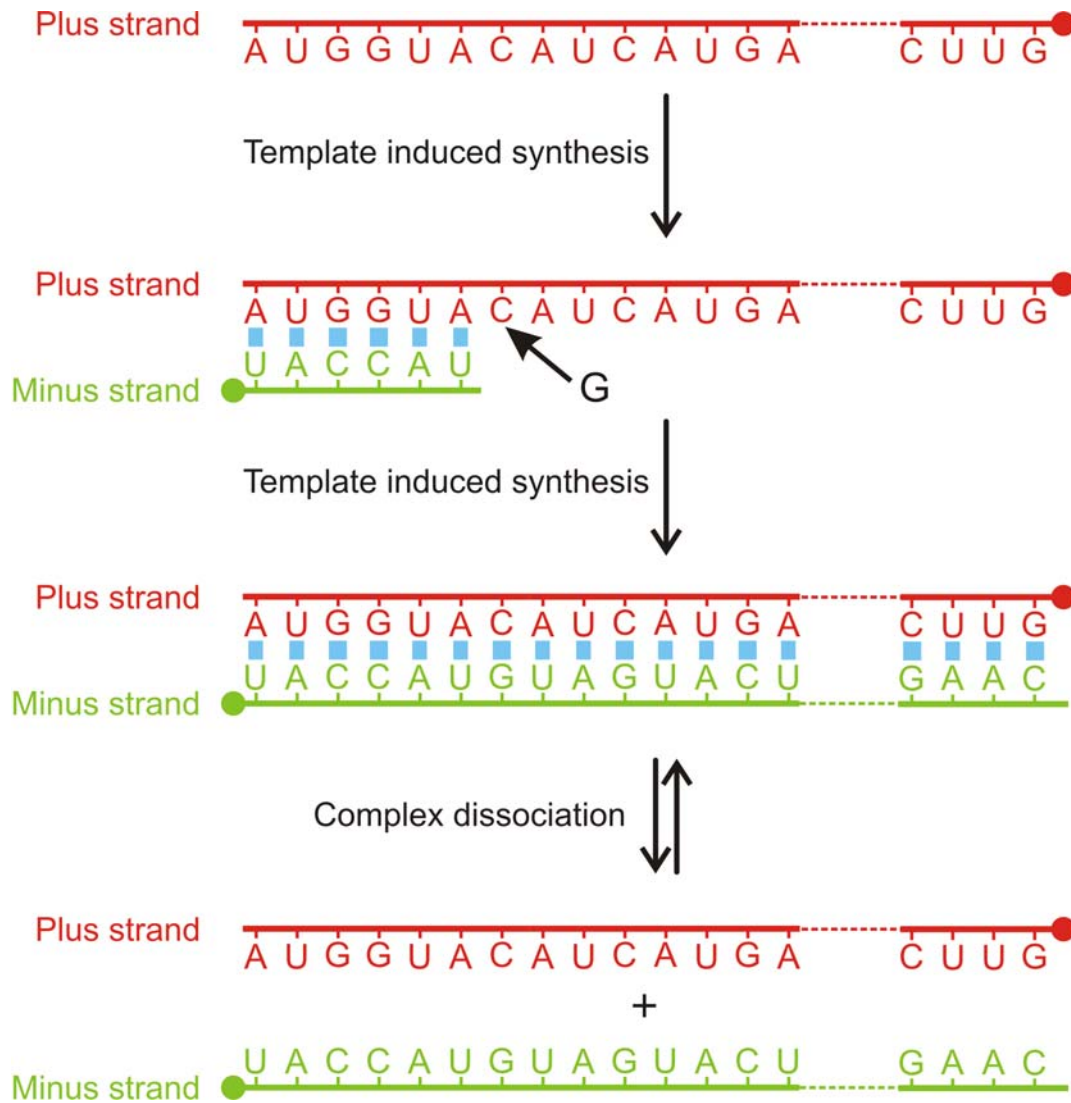


James D. Watson, 1928-, and Francis H.C. Crick, 1916-2004

Nobel prize 1962

1953 – 2003 fifty years double helix

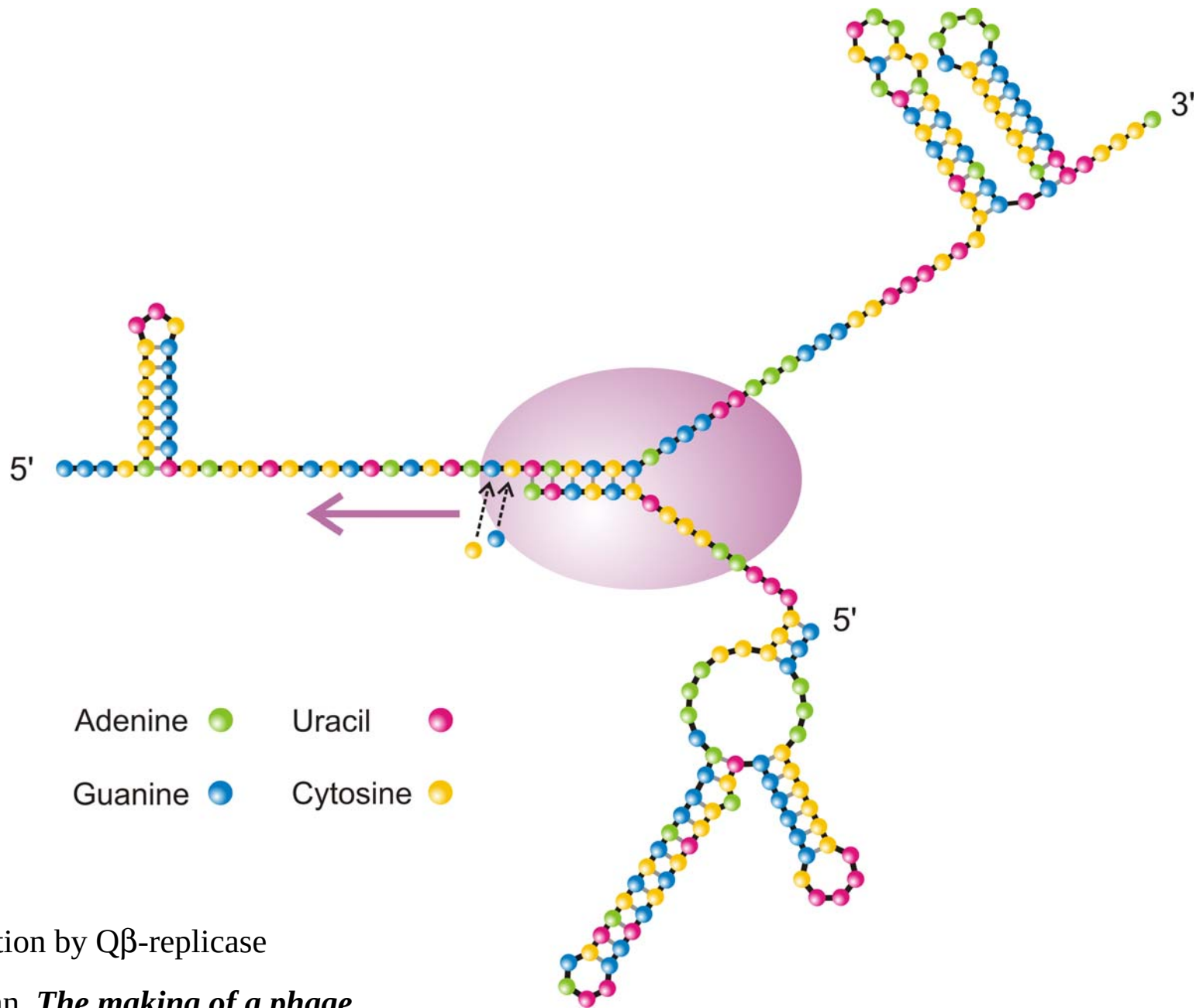
The three-dimensional structure of a short double helical stack of B-DNA



Complementary replication is the simplest copying mechanism of RNA.

Complementarity is determined by Watson-Crick base pairs:

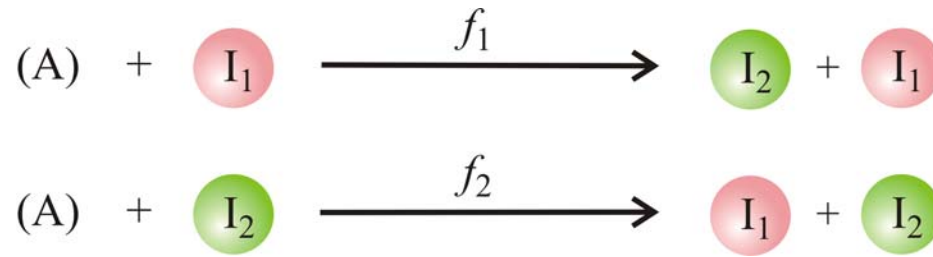
$G \equiv C$ and $A = U$



RNA replication by Q β -replicase

C. Weissmann, ***The making of a phage.***

FEBS Letters **40** (1974), S10-S18



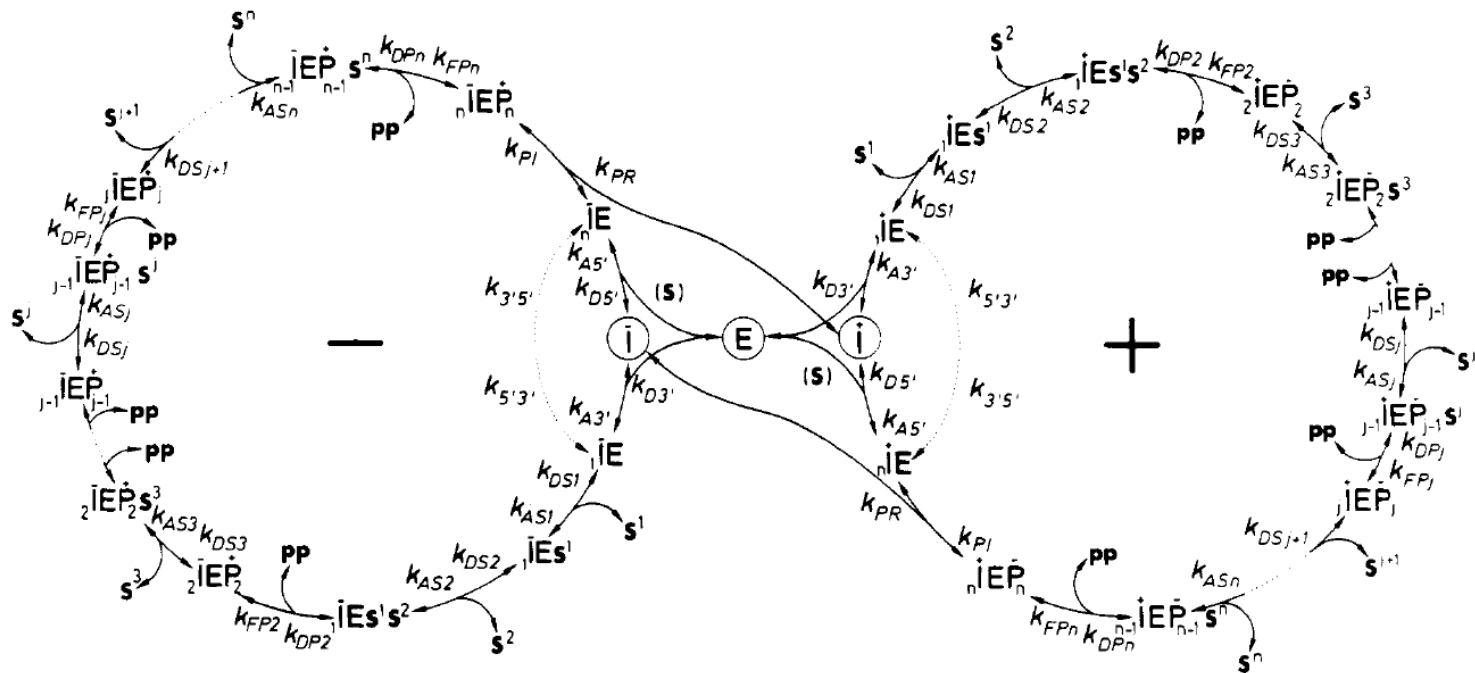
$$\frac{dx_1}{dt} = f_2 x_2 \quad \text{and} \quad \frac{dx_2}{dt} = f_1 x_1$$

$$x_1 = \sqrt{f_2} \xi_1, \quad x_2 = \sqrt{f_1} \xi_2, \quad \zeta = \xi_1 + \xi_2, \quad \eta = \xi_1 - \xi_2, \quad f = \sqrt{f_1 f_2}$$

$$\eta(t) = \eta(0) e^{-ft}$$

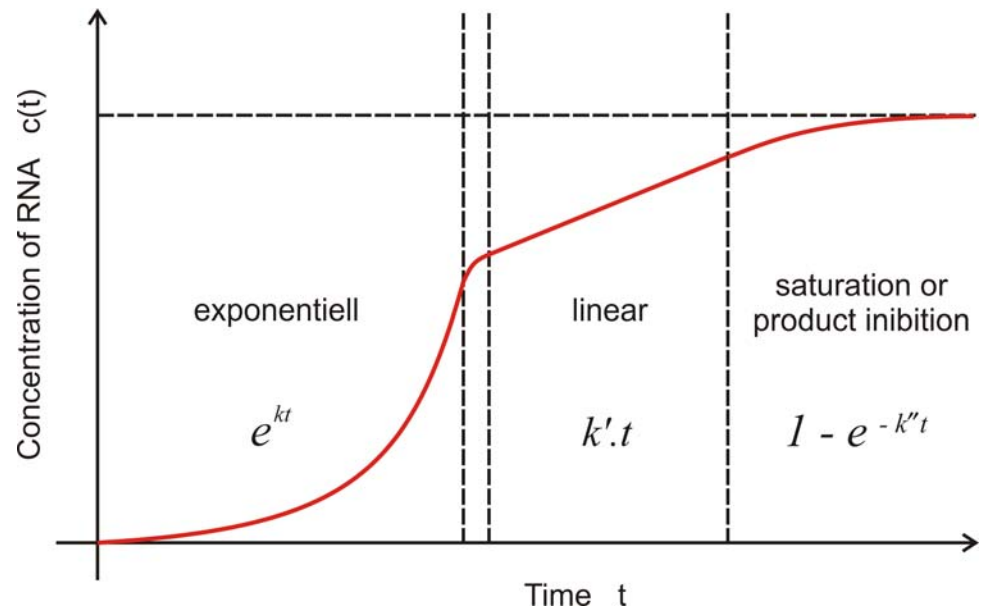
$$\zeta(t) = \zeta(0) e^{ft}$$

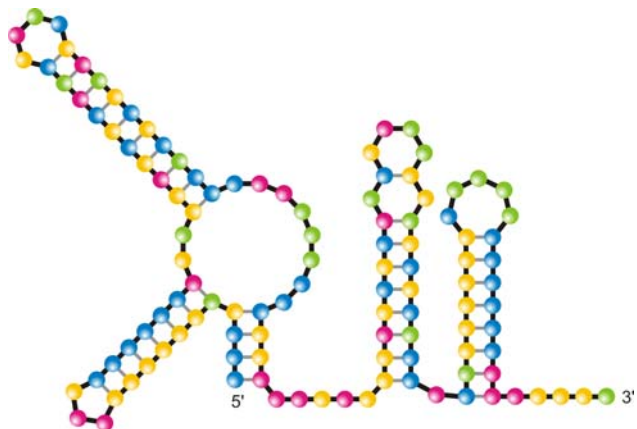
Complementary replication as the simplest molecular mechanism of reproduction



Kinetics of RNA replication

C.K. Biebricher, M. Eigen, W.C. Gardiner, Jr.
Biochemistry **22**:2544-2559, 1983





$$\Delta G_{\text{fold}} = - 68.5 \text{ kcal / mole}$$



$$\Delta G_{\text{fold}} = - 98.4 \text{ kcal / mole}$$



$$\Delta G_{\text{fold}} = - 277.4 \text{ kcal / mole}$$

$$\Delta G_{\text{bind}} = - 72.1 \text{ kcal / mole}$$

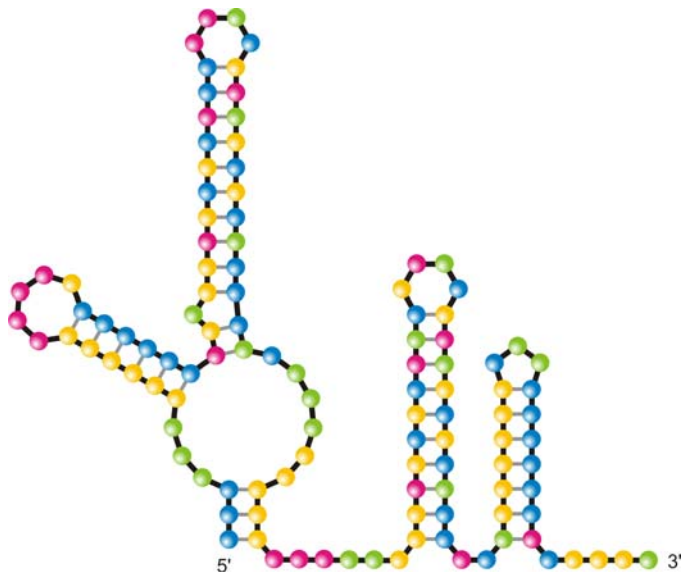
Adenine 

Uracil 

Guanine 

Cytosine 

SV11 plus strand



$$\Delta G_{\text{fold}} = - 71.1 \text{ kcal / mole}$$



$$\Delta G_{\text{fold}} = - 101.9 \text{ kcal / mole}$$



$$\Delta G_{\text{fold}} = - 277.4 \text{ kcal / mole}$$

$$\Delta G_{\text{bind}} = - 72.1 \text{ kcal / mole}$$

Adenine 

Uracil 

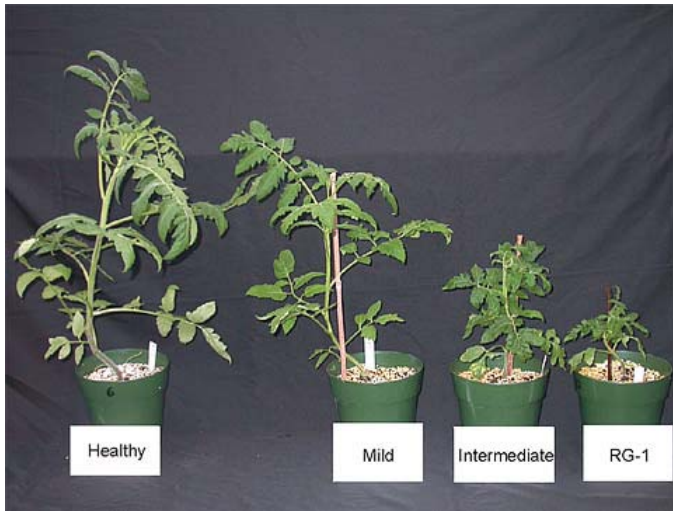
Guanine 

Cytosine 

SV11 minus strand

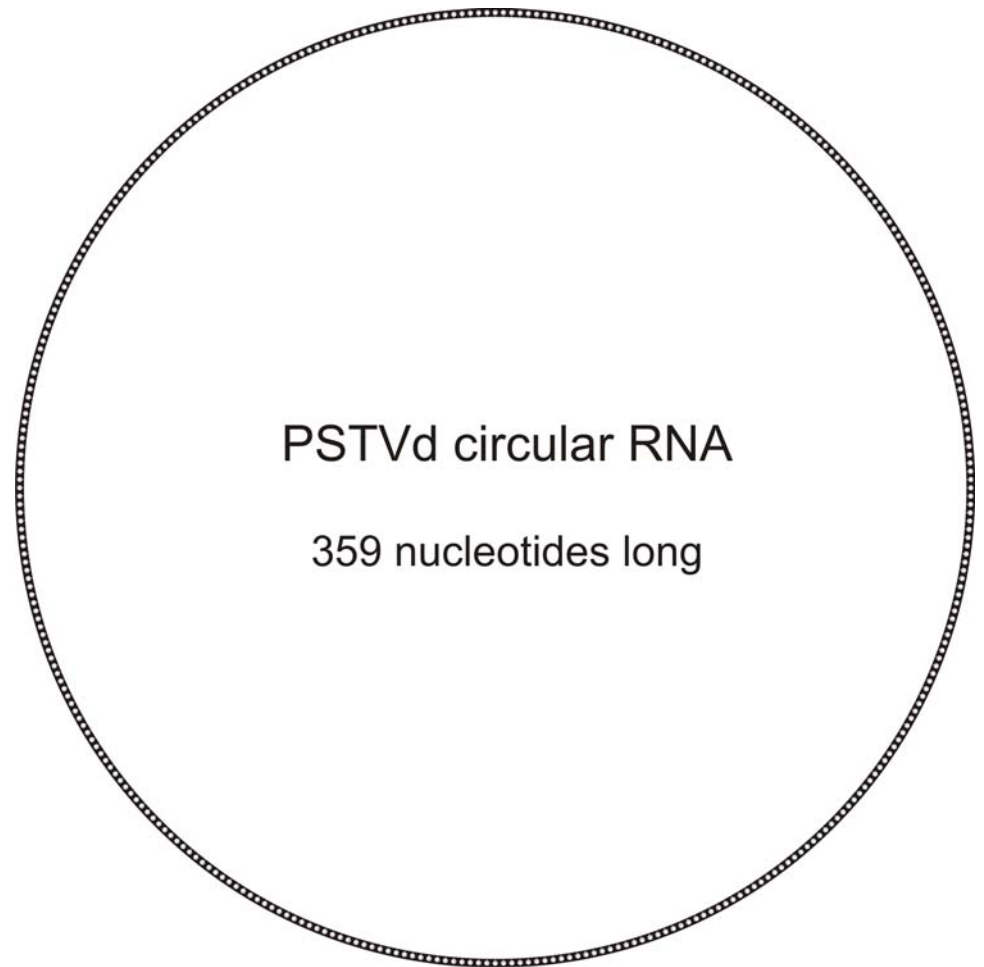
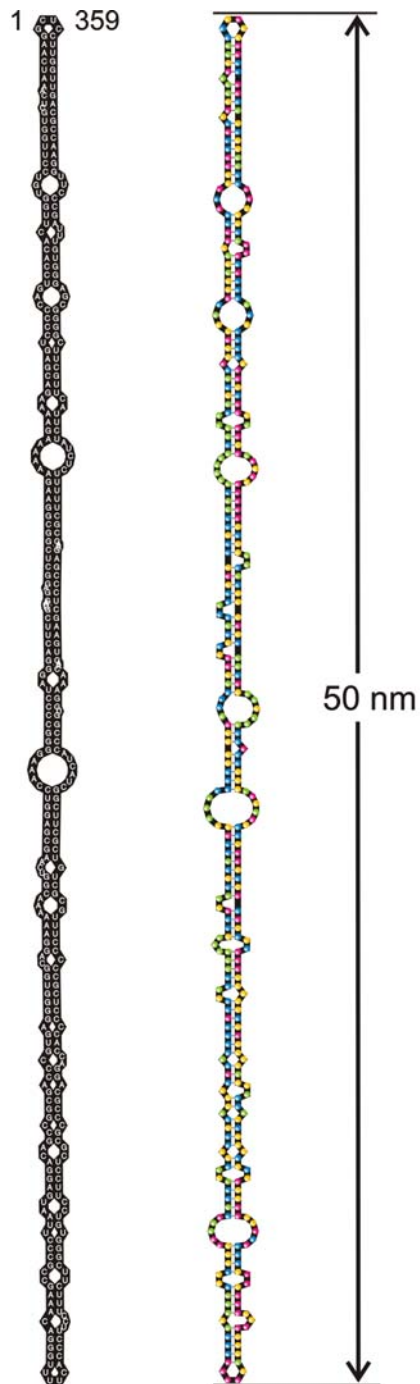


J. Demez. European and mediterranean plant protection organization archive. France



R.W. Hammond, R.A. Owens. Molecular Plant Pathology Laboratory, US Department of Agriculture

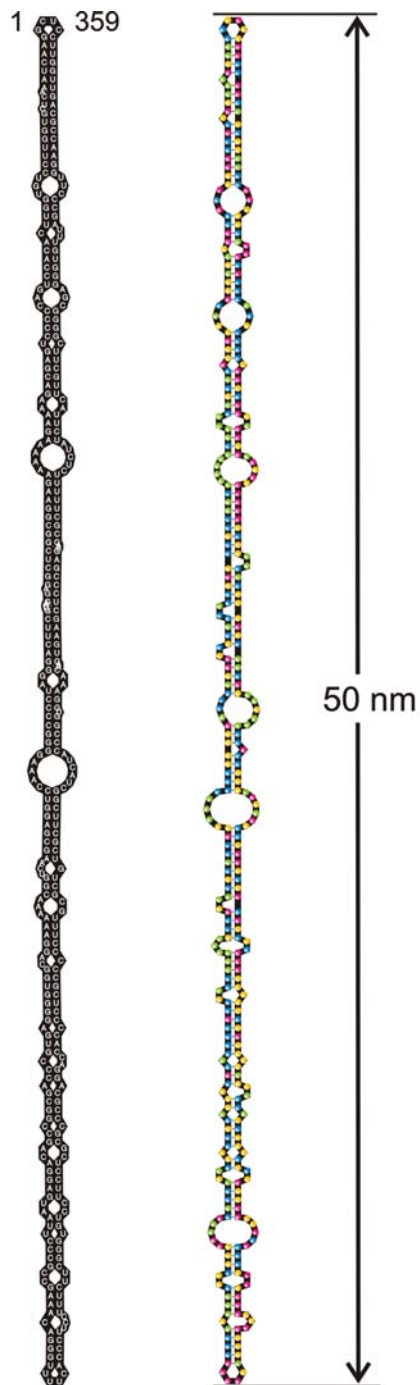
Plant damage by viroids



- Adenine
- Uracil
- Guanine
- Cytosine

Nucleotide sequence and secondary structure
of the potato spindle tuber viroid RNA

H.J.Gross, H. Domdey, C. Lossow, P Jank,
M. Raba, H. Alberty, and H.L. Sänger.
Nature **273**:203-208 (1978)



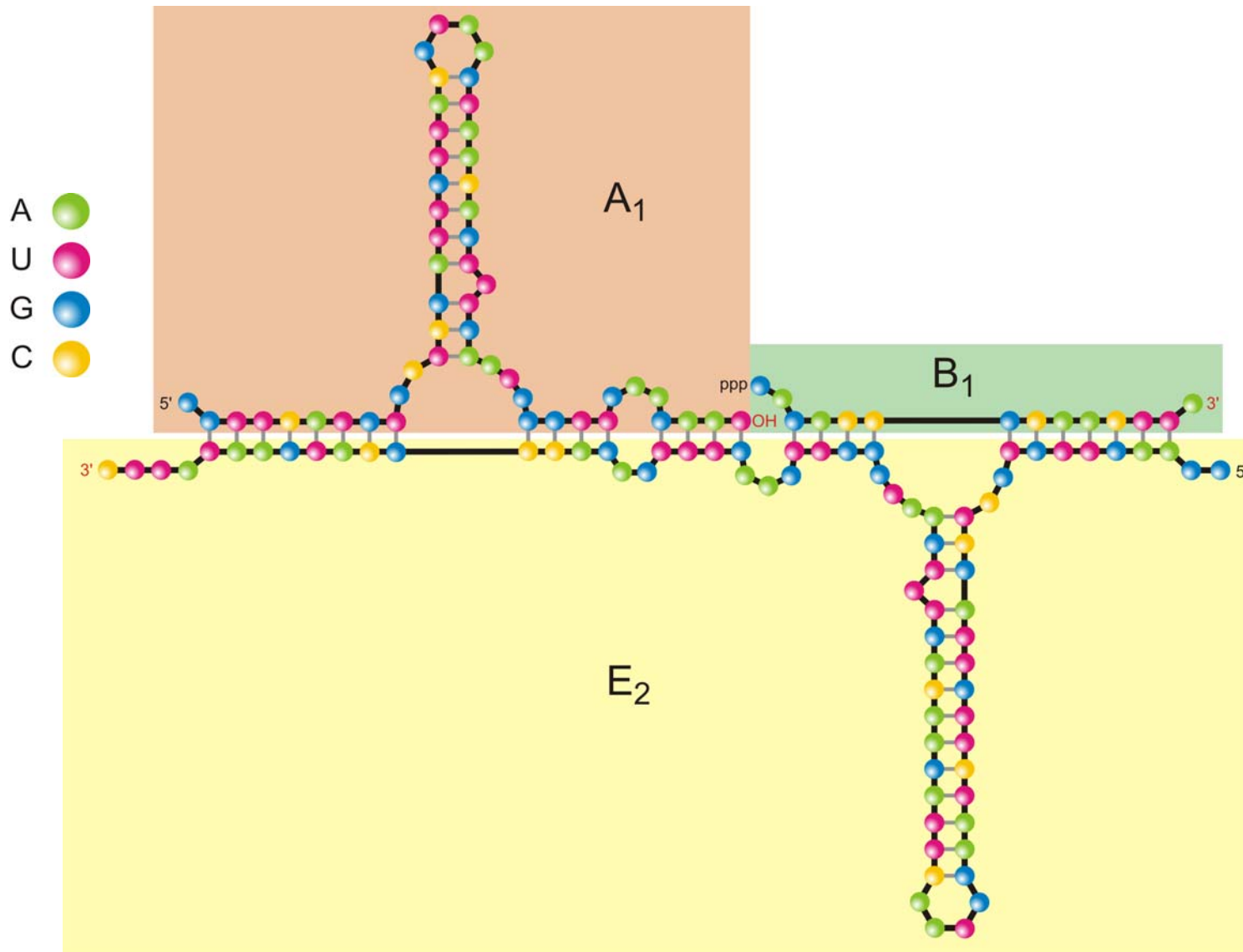
Vienna RNA Package 1.8.2

Biochemically supported structure

- Adenine
- Uracil
- Guanine
- Cytosine

Nucleotide sequence and secondary structure
of the potato spindle tuber viroid RNA

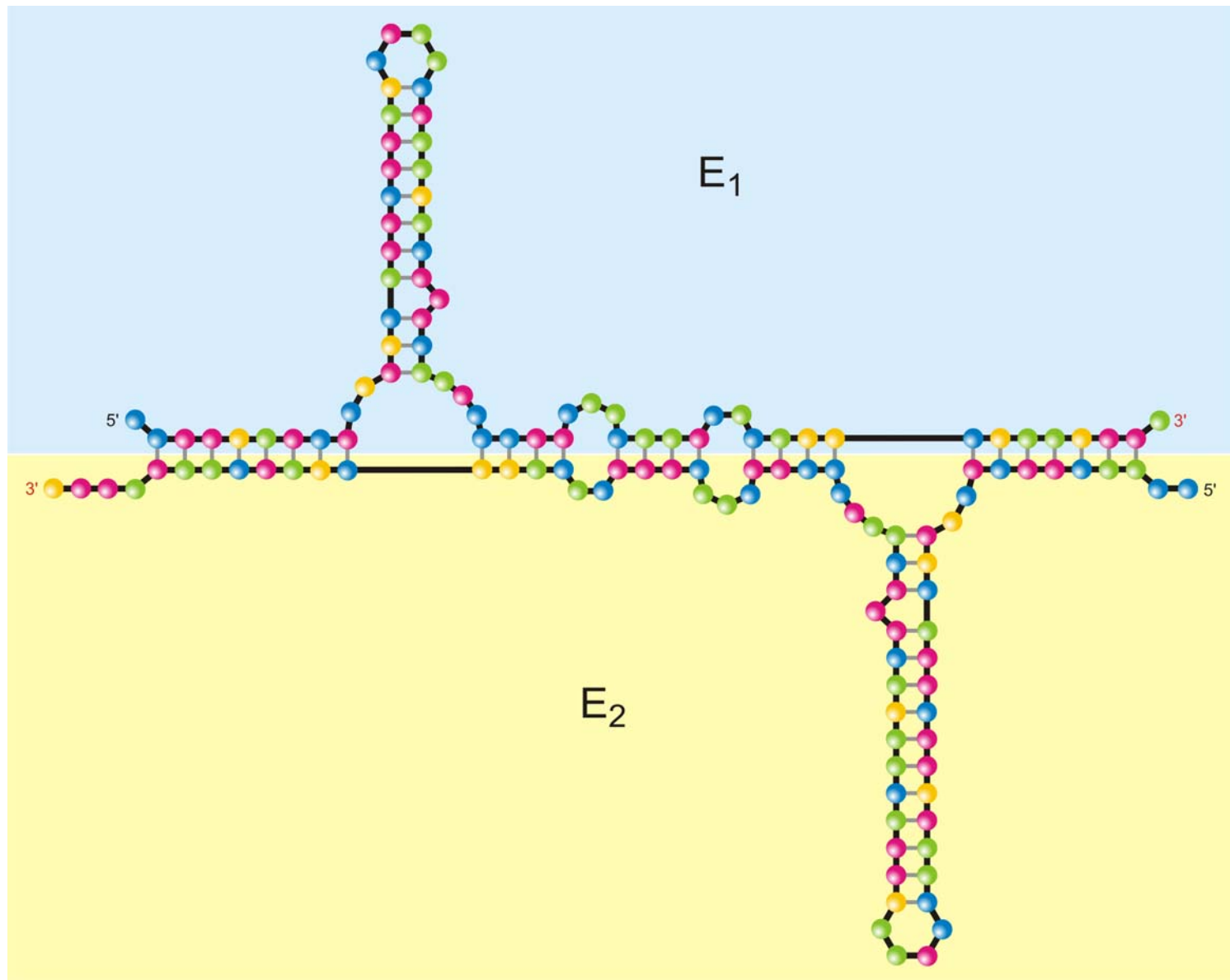
H.J.Gross, H. Domdey, C. Lossow, P Jank,
M. Raba, H. Alberty, and H.L. Sanger.
Nature **273**:203-208 (1978)



An example of two ribozymes growing exponentially by cross-catalysis.

T.A. Lincoln, G.F. Joyce. 2009. Self-sustained replication of an RNA enzyme. *Science* 323:1229-1232

A ●
U ●
G ●
C ●



An example of two ribozymes growing exponentially by cross-catalysis.

T.A. Lincoln, G.F. Joyce. 2009. Self-sustained replication of an RNA enzyme. *Science* 323:1229-1232

1. RNA replication *in vitro* and *in vivo*
2. **Evolution of RNA molecules**
3. RNA sequences and structures
4. Evolutionary optimization of RNA structure

Selforganization of Matter
and the Evolution of Biological Macromolecules

MANFRED EIGEN*

Max-Planck-Institut für Biophysikalische Chemie,
Karl-Friedrich-Bonhoeffer-Institut, Göttingen-Nikolausberg

I. Introduction	465	V. Selforganization via Cyclic Catalysis? Proteins	498
I.1. Cause and Effect	465	V.1. Recognition and Catalysis by Enzymes	498
I.2. Penetration of Selforganization	467	V.2. Selforganizing Enzyme Cycles (Theory)	499
I.2.1. Evolution Must Start from Random Events	467	V.2.1. Catalytic Networks	499
I.2.2. Information Requires Information	467	V.2.2. Selfreplicating Loop and Its Variants	499
I.2.3. Information Originates or Gains Value by Selection	469	V.2.3. Competition between Different Cycles	501
I.2.4. Selection Occurs with Special Substances under Special Conditions	470	V.3. Can Protein Replication Theories?	501
II. Phenomenological Theory of Selection	473	VI. Selforganization by Enzymal Catalytic Function	503
II.1. The Concept "Information"	473	VI.1. The Requirement of Cooperation between Nucleic Acids and Proteins	503
II.2. Phenomenological Equations	474	VI.2. A Selforganizing Hyper-Cycle	503
II.3. Selection Strains	476	VI.2.1. The Model	503
II.4. Selection Equilibrium	479	VI.2.2. Theoretical Treatment	505
II.5. Quality Factor and Error Distribution	480	VI.3. On the Origin of the Code	506
II.6. Kinetics of Selection	481	VII. Relative Experiments	511
III. Stochastic Approach to Selection	484	VII.1. The QP-Replicase System	511
III.1. Limitations of a Deterministic Theory of Selection	484	VII.2. Darwinian Evolution in the Test Tube	512
III.2. Fluctuations around Equilibrium States	484	VII.3. Quantitative Selection Studies	513
III.3. Fluctuations in the Steady State	485	VII.4. "Mines One" Experiments	514
III.4. Stochastic Models as Markov Chains	487	VIII. Conclusion	515
III.5. Quantitative Discussion of Three Prototypes of Selection	487	VIII.1. Limits of Theory	515
IV. Selforganization Based on Complementary Interactions: Nucleic Acids	490	VIII.2. The Concept "Value"	515
IV.1. True "Self-Organization"	490	VIII.3. "Diagnosis" and the "Origin of Information"	516
IV.2. Complementary Interaction and Selection (Theory)	492	VIII.4. The Phenology of Selection and Evolution	517
IV.3. Complementary Base Recognition (Experimental Data)	494	VIII.5. "Indeterminable" but "Inevitable"	518
IV.3.1. Single Pair Formation	494	VIII.6. Can the Phenomena of Life be Explained by Our Present Concepts of Physics?	520
IV.3.2. Cooperative Interactions in G-CpG- and Poly-nucleotide	495	IX. Deutsche Zusammenfassung	520
IV.3.3. Conclusions about Recognition	496	Acknowledgements	522
		Literature	522

I. Introduction

I.1. "Cause and Effect"

The question about the origin of life often appears as a question about "cause and effect". Physical theories of macroscopic processes usually involve answers to such questions, even if a statistical interpretation is given to the relation between "cause" and "effect". It is mainly due to the nature of this question that many scientists believe that our present physics does not offer any obvious explanation for the existence of life.

* Formerly, presented as the "Robbins Lectures" at Penna College, Chicago, in spring 1970.

234 Naturwissenschaften 1971

1971

The Hypercycle

A Principle of Natural Self-Organization

Part A: Emergence of the Hypercycle

Manfred Eigen

Max-Planck-Institut für biophysikalische Chemie, D-3400 Göttingen

Peter Schuster

Institut für theoretische Chemie und Strahlenchemie der Universität, A-1090 Wien

This paper is the first part of a trilogy, which comprises a detailed study of a special type of functional organization and demonstrates its relevance with respect to the origin and evolution of life. Self-replicating macromolecules, such as RNA or DNA in a suitable environment exhibit a behavior, which we may call Darwinian and which can be formally represented by the concept of the quasi-species. A quasi-species is defined as a given distribution of macromolecular species with closely interrelated sequences, dominated by one or several (hypothesis) master copies. External constraints enforce the selection of the best adapted distribution, continuously referred to as the wild-type. Most important for Darwinian behavior are the criteria for internal stability of the quasi-species. If these criteria are violated, the information stored in the nucleotide sequence of the master copy will disseminate irreversibly leading to an error catastrophe. As a consequence, selection and evolution of RNA or DNA molecules is limited with respect to the amount of information that can be stored in a single replicative unit. An analysis of experimental data regarding RNA and DNA replication at various levels of organization reveals that a sufficient amount of information for the build-up of a translation machinery can be gained only via integration of several different replicative units (replicative cycles) through reciprocal linkages. A stable functional integration then will raise the system to a new level of organization and thereby enlarge its information capacity considerably. The hypercycle appears to be such a form of organization.

Preview on Part B: The Abstract Hypercycle

The mathematical analysis of dynamical systems using methods of differential topology yields the result that there is only one type of mechanism which fulfills the following requirements. The information stored in each single replicative unit (or reproductive cycle) must be maintained, i.e., the respective master copies must compete favorably with their error distributions. Despite their competitive behavior these units must establish a cooperation which includes all functionally integrated species. On the other hand, the cycle as a whole must continue to compete strongly with any other single entity or isolated ensemble which does not contribute to its integrated function.

These requirements are crucial for a selection of the best adapted functionally linked ensemble and its evolutive optimization. Only

Naturwissenschaften 64, 541-565 (1977) © by Springer-Verlag 1977

1977

Molecular Quasi-Species¹

Manfred Eigen,* John McCaskill,

Max-Planck-Institut für biophysikalische Chemie, Am Fassberg, D-3400 Göttingen-Nikolausberg, BRD

and Peter Schuster*

Institut für theoretische Chemie und Strahlenchemie, der Universität Wien, Währinger Strasse 17, A-1090 Wien, Austria (Received: June 9, 1988)

The molecular quasi-species model describes the physicochemical organization of monomers into an ensemble of heteropolymers with combinatorial complexity by ongoing template polymerization. Polynucleotides belong to the simplest class of such molecules. The quasi-species itself represents the stationary distribution of macromolecular sequences maintained by chemical reactions effecting error-prone replication and by transport processes. It is obtained deterministically, by mass-action kinetics, as the dominant eigenvector of a value matrix, W , which is derived directly from chemical rate coefficients, but it also exhibits stochastic features, being composed of a significant fraction of unique individual macromolecular sequences. The quasi-species model demonstrates how macromolecular information originates through specific nonequilibrium autocatalytic reactions and thus forms a bridge between reaction kinetics and molecular evolution. Selection and evolutionary optimization appear as new features in physical chemistry. Concentration bias in the production of mutants is a new concept in population genetics, relevant to frequently mutating populations, which is shown to greatly enhance the optimization processes. The present theory relates to actually replicating ensembles, but this restriction is not essential. A sharp transition is exhibited between a drifting population of essentially random macromolecular sequences and a localized population of close relatives. This transition at a threshold error rate was found to depend on sequence lengths, distributions of selective values, and population sizes. It has been determined generally for complex landscapes and for special cases, and, it was shown to persist generally in the presence of nearly neutral mutants. Replication dynamics has much in common with the equilibrium statistics of complex spin systems: the error threshold is equivalent to a magnetic order-disorder transition. A rational function of the replication accuracy plays the role of temperature. Experimental data obtained from test-tube evolution of polynucleotides and from studies of natural virus populations support the quasi-species model. The error threshold seems to set a limit to the genome lengths of several classes of RNA viruses. In addition, the results are relevant even in eucaryotes where they contribute to the exon-intron debate.

Preview on Part C: The Abstract Hypercycle

A realistic model of a hypercycle relevant with respect to the origin of the genetic code and the translation machinery is presented. It includes the following features referring to natural systems:

- 1) The hypercycle has a sufficiently simple structure to admit an organization with finite probability under prebiotic conditions.
- 2) It permits a continuous emergence from slowly accumulated in RNA-like precursors, originally being members of a stable RNA quasi-species and having been amplified to a level of higher abundance.
- 3) The organizational structure and the properties of single functional units of this hypercycle are well reflected in the present genetic code in the translation apparatus of the prokaryotic cell, as well as in certain bacterial viruses.

I. The Paradigm of Unity and Diversity in Evolution

Why do millions of species, plants and animals, exist, while there is only one basic molecular machinery of the cell: one universal genetic code and unique chemicalities of the macromolecules?

The geneticists of our day would not hesitate to give an immediate answer to the first part of this question. Diversity of species is the outcome of the tremendous branching process of evolution with its myriads of single steps of reproduction and mutation. It in-

cludes very small numbers of different bases, even just two, readily provides numbers of different entities so enormous that neither consecutive nor parallel physical realization is possible. The physical chemistry of finite systems of such macromolecules must deal with both known regularities and the advent of unique copolymeric sequences. Normally this would present no difficulty in a statistical mechanical analysis of typical behavior, where rare events play no significant role, but with autocatalytic polymerization processes even unique single molecules may be amplified to determine the fate of the entire system. Potentially creative, self-organizing around unique events, the dynamics of the simplest living chemical system is invested with regularities that both allow and limit efficient adaptation. The quasi-species model is a study of these regularities.

The fundamental regularity in living organisms that has invited explanation is adaptation. Why are organisms so well fitted to their environments? At a more chemical level, why are enzymes

optimal catalysts? Darwin's theory of natural selection has provided biologists with a framework for the answer to this question. The present model is constructed along Darwinian lines but in terms of specific macromolecules, chemical reactions, and physical processes that make the notion of survival of the fittest precise. Not only does the model give an understanding of the physical limitations of adaptation, but also it provides new insight into the role of chance in the process. For an understanding of the structure of this minimal chemical model it is first necessary to recall the conceptual basis of Darwin's theory.

Darwin recognized that new inheritable adaptive properties were not induced by the environment but arose independently in the production of offspring. Lasting adaptive changes in a population could only come about by natural selection of the heritable traits or genotype based on the full characteristics or phenotype relevant for producing offspring. A process of chance, i.e., uncorrelated with the developed phenotype, control changes in the genotype from one generation to the next and generates the diversity necessary for selection. Three factors have probably prevented chemists from gaining a clear insight into these phenomena in the past, despite the discovery of the polymeric nature of the genotype (DNA): the complexity of a minimum replication phenotype, the problem of dealing with a huge number of variants, and the nonequilibrium nature of these ongoing processes.

The formulation of a tractable chemical model based on Darwin's principle may be understood in several steps:

1. The major constituents of the system have to be inherently self-reproductive. Only two classes of molecules are presently

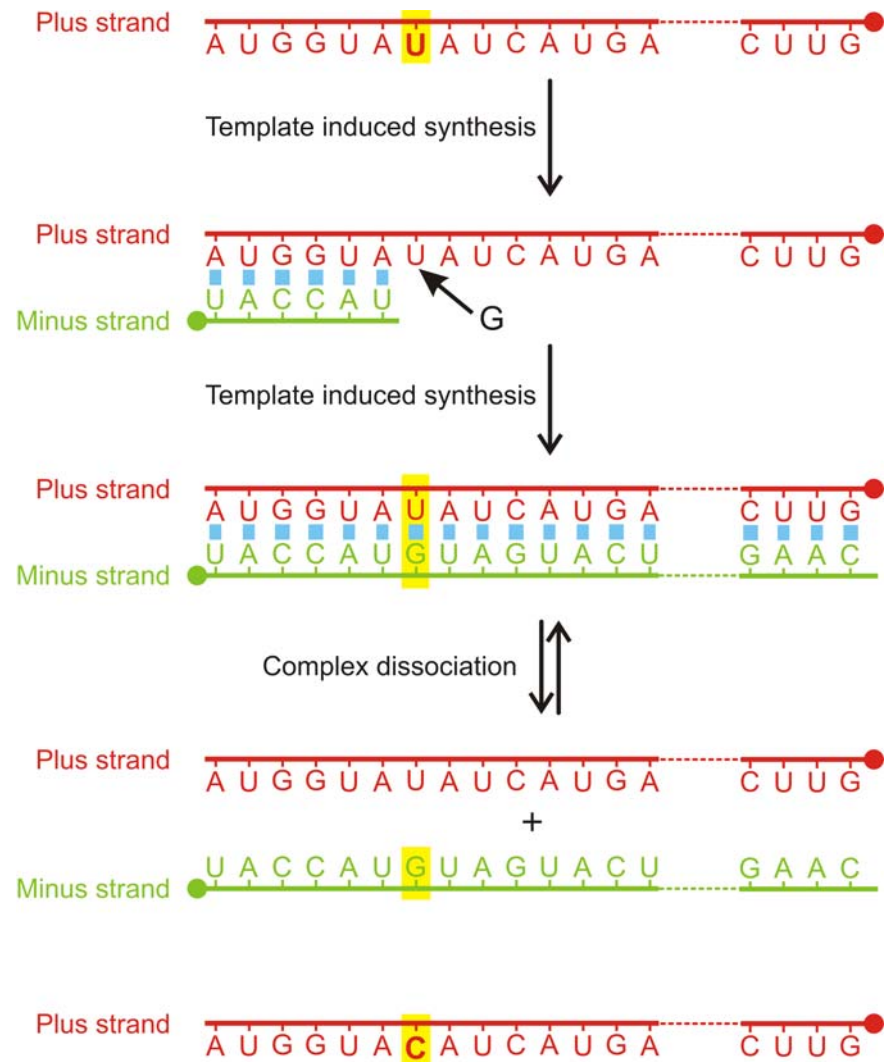
¹ This is an abridged account of the quasi-species theory that has been submitted in comprehensive form to *Advances in Chemical Physics*.

(*) Eigen, M.; McCaskill, J. S.; Schuster, P. *Adv. Chem. Phys.*, in press.

0022-3654/88/2022-6881\$01.50/0

© 1988 American Chemical Society

1988

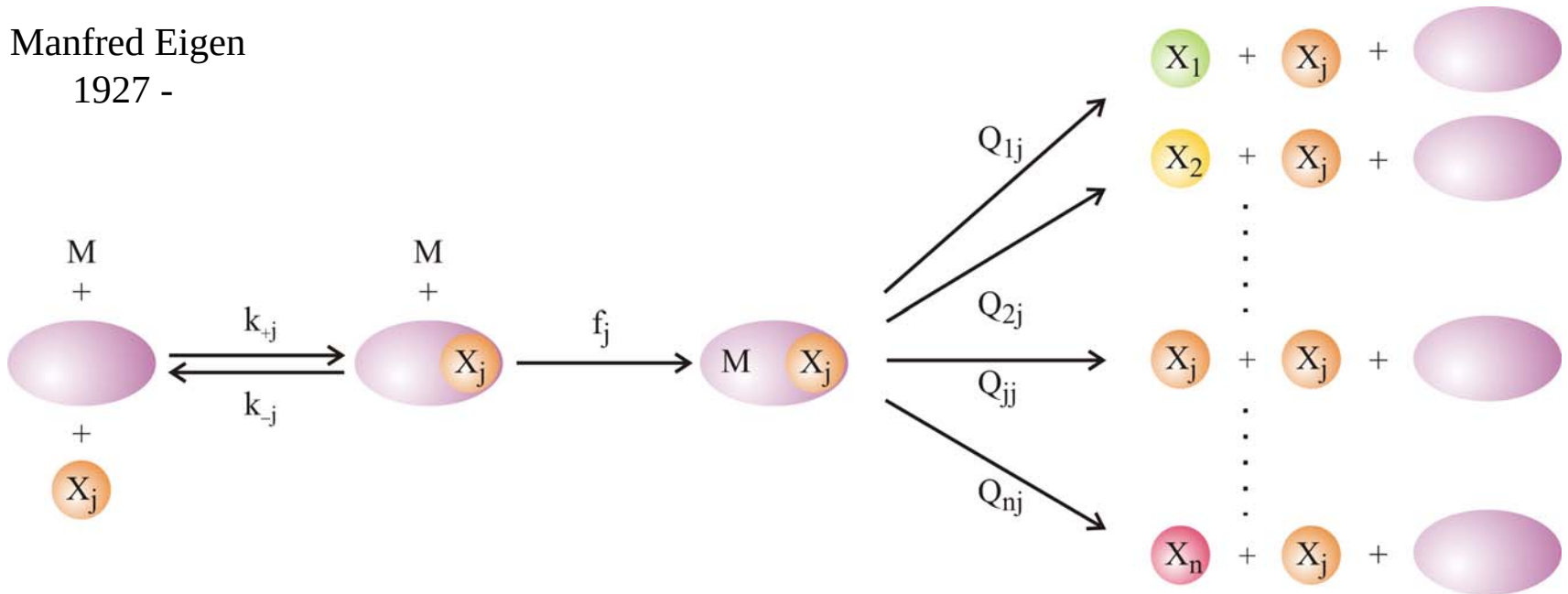


Replication and mutation are parallel chemical reactions.



Manfred Eigen
1927 -

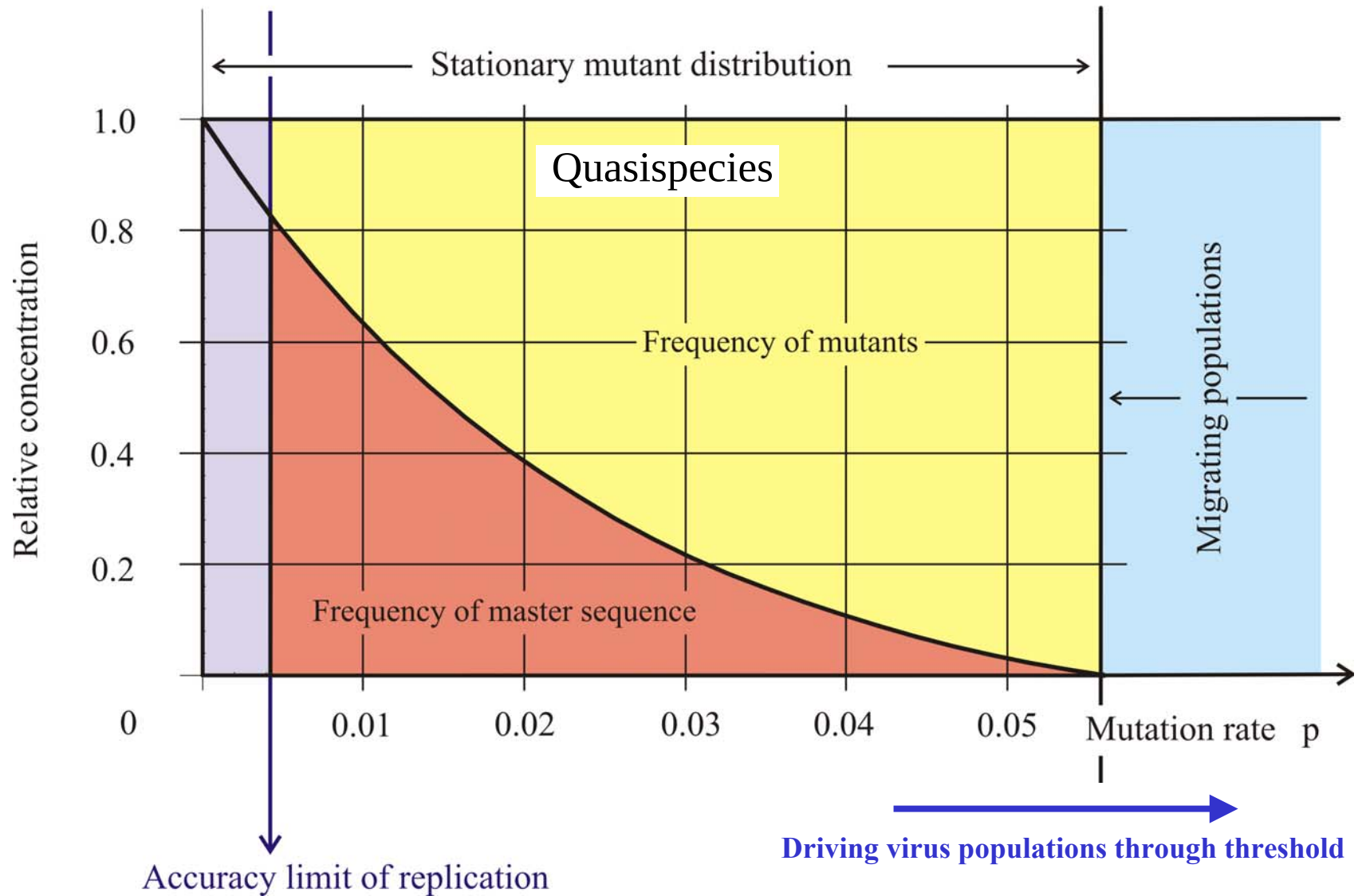
$$\frac{dx_j}{dt} = \sum_{i=1}^n Q_{ji} f_i x_i - x_j \Phi; \quad j=1,2,\dots,n$$



Mutation and (correct) replication as parallel chemical reactions

M. Eigen. 1971. *Naturwissenschaften* 58:465,

M. Eigen & P. Schuster. 1977. *Naturwissenschaften* 64:541, 65:7 und 65:341



The error threshold in replication

Chain length and error threshold

$$Q \cdot \sigma = (1-p)^n \cdot \sigma \geq 1 \Rightarrow n \cdot \ln(1-p) \geq -\ln \sigma$$

$$p \dots \text{constant} : n_{\max} \approx \frac{\ln \sigma}{p}$$

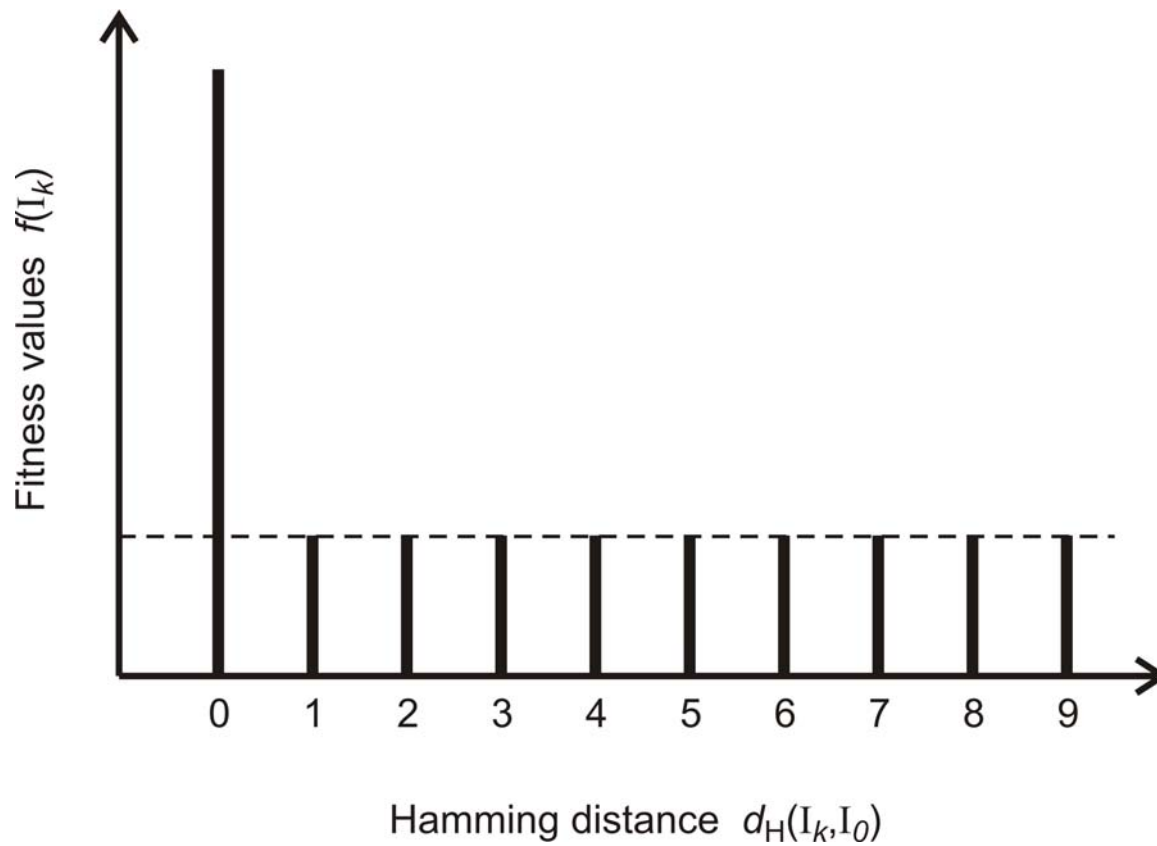
$$n \dots \text{constant} : p_{\max} \approx \frac{\ln \sigma}{n}$$

$Q = (1-p)^n$... replication accuracy

p ... error rate

n ... chain length

$\sigma = \frac{f_m}{\sum_{j \neq m} f_j}$... superiority of master sequence



A fitness landscape showing an error threshold

SELF-REPLICATION WITH ERRORS

A MODEL FOR POLYNUCLEOTIDE REPLICATION **

Jörg SWETINA and Peter SCHUSTER *

Institut für Theoretische Chemie und Strahlenchemie der Universität, Währingerstraße 17, A-1090 Wien, Austria

Received 4th June 1982

Revised manuscript received 23rd August 1982

Accepted 30th August 1982

Key words: Polynucleotide replication; Quasi-species; Point mutation; Mutant class; Stochastic replication

A model for polynucleotide replication is presented and analyzed by means of perturbation theory. Two basic assumptions allow handling of sequences up to a chain length of $n = 80$ explicitly: point mutations are restricted to a two-digit model and individual sequences are subsumed into mutant classes. Perturbation theory is in excellent agreement with the exact results for long enough sequences ($n > 20$).

1. Introduction

Eigen [8] proposed a formal kinetic equation (eq. 1) which describes self-replication under the constraint of constant total population size:

$$\frac{dx_i}{dt} = x_i \sum_j w_{ij} x_j - \frac{x_i}{\phi} \quad i = 1, \dots, n \quad (1)$$

By x_i we denote the population number or concentration of the self-replicating element I_i , i.e., $x_i = [I_i]$. The total population size or total concentration $c = \sum_i x_i$ is kept constant by proper adjustment of the constraint ϕ : $\phi = \sum_i \sum_j w_{ij} x_j$. Characteristically, this constraint has been called 'constant organization'. The relative values of diagonal

(w_{ii}) and off-diagonal (w_{ij} , $i \neq j$) rates, as we shall see in detail in section 2, are related to the accuracy of the replication process. The specific properties of eq. 1 are essentially based on the fact that it leads to exponential growth in the absence of constraints ($\phi = 0$) and competitors ($n = 1$).

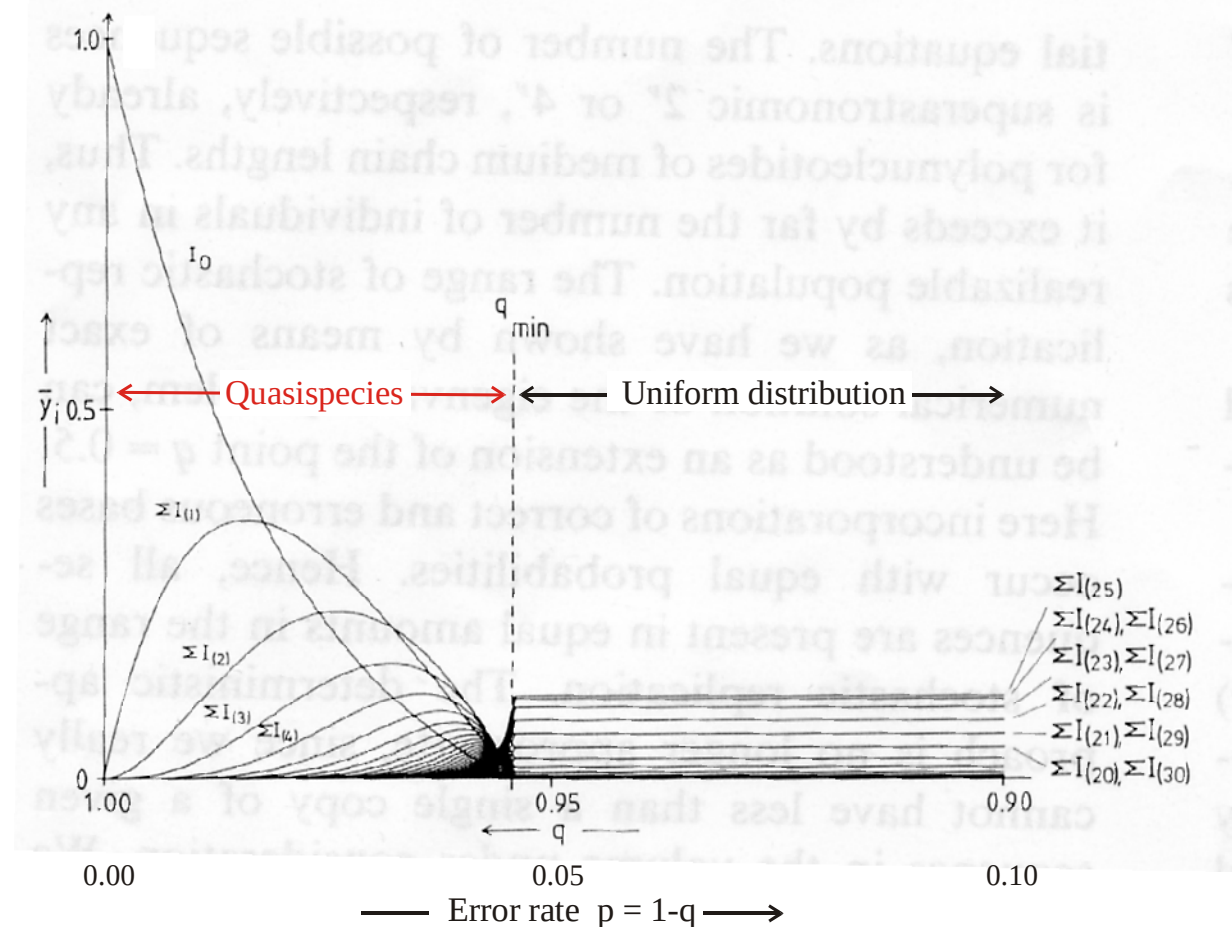
The non-linear differential equation, eq. 1 – the non-linearity is introduced by the definition of ϕ at constant organization – shows a remarkable feature: it leads to selection of a defined ensemble of self-replicating elements above a certain accuracy threshold. This ensemble of a master and its most frequent mutants is a so-called 'quasi-species' [9]. Below this threshold, however, no selection takes place and the frequencies of the individual elements are determined exclusively by their statistical weights.

Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional perturbation technique

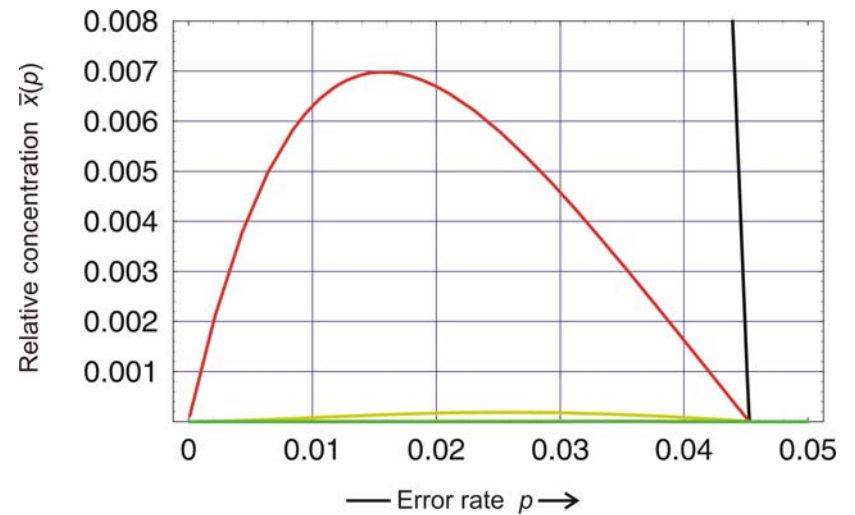
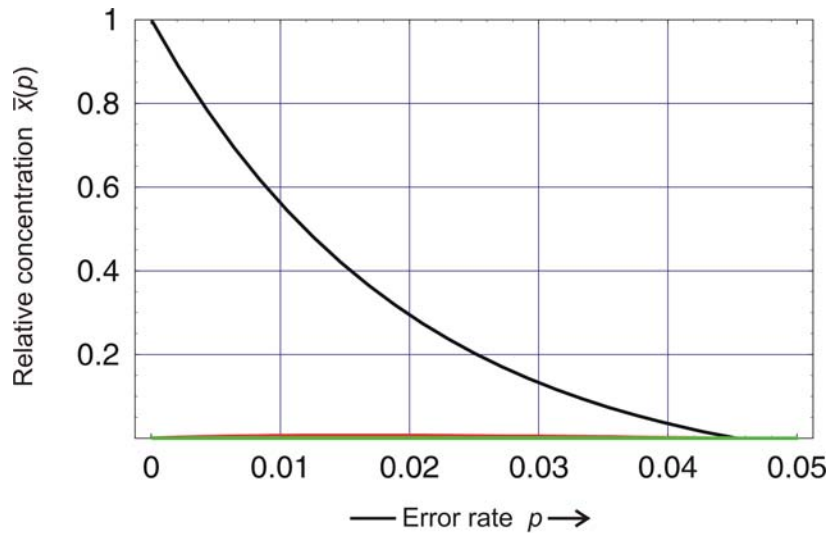
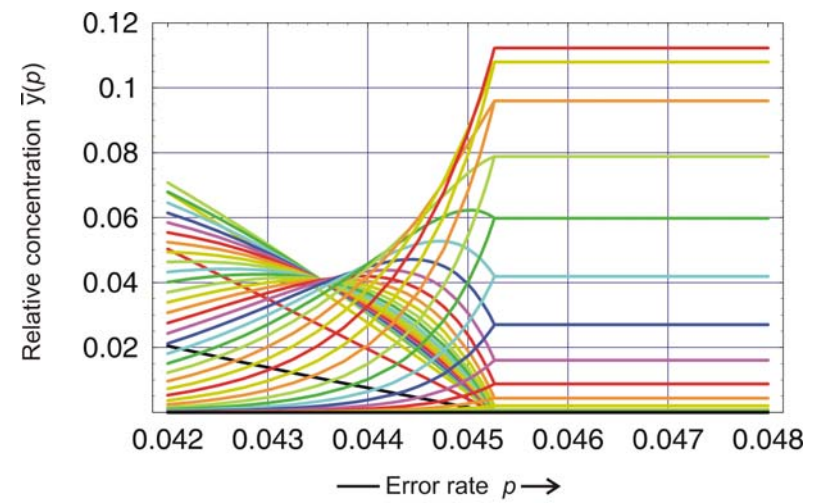
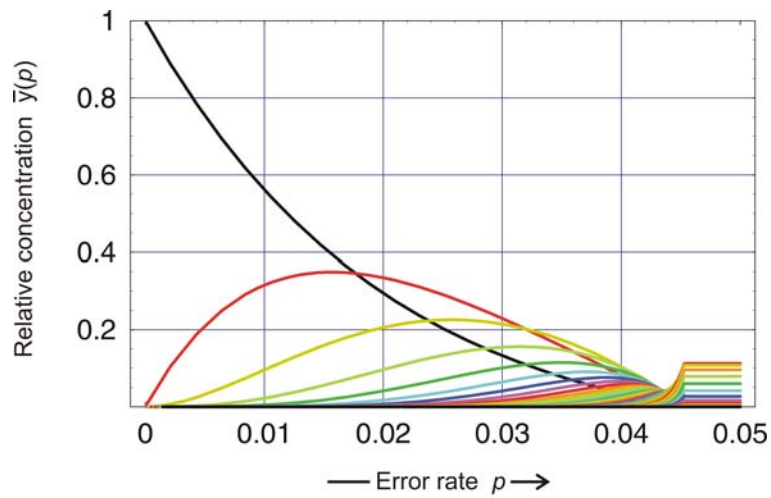
* Dedicated to the late Professor B.L. Jones who was among the first to do rigorous mathematical analysis on the problems described here.

** This paper is considered as part II of Model Studies on RNA replication. Part I is by Gansner and Schuster [14].

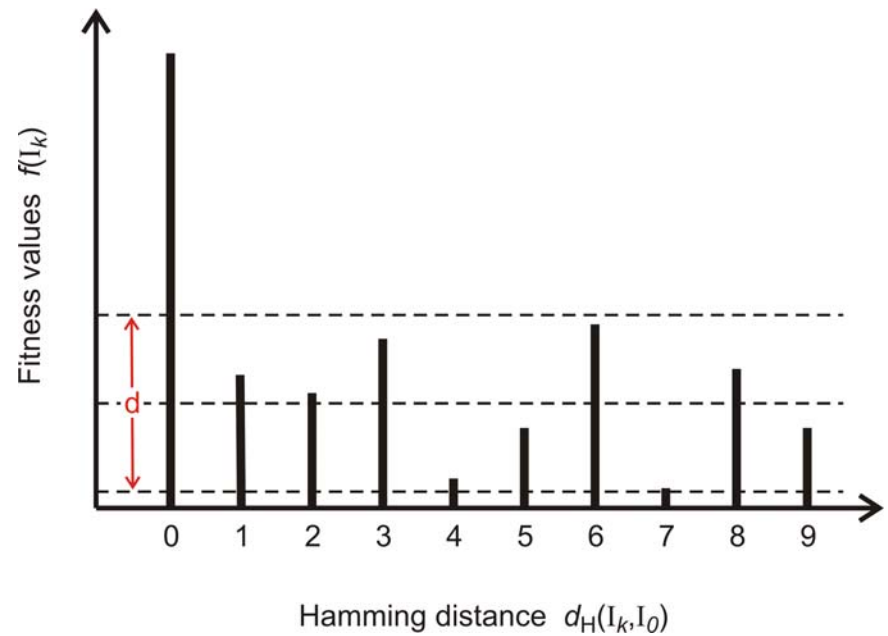
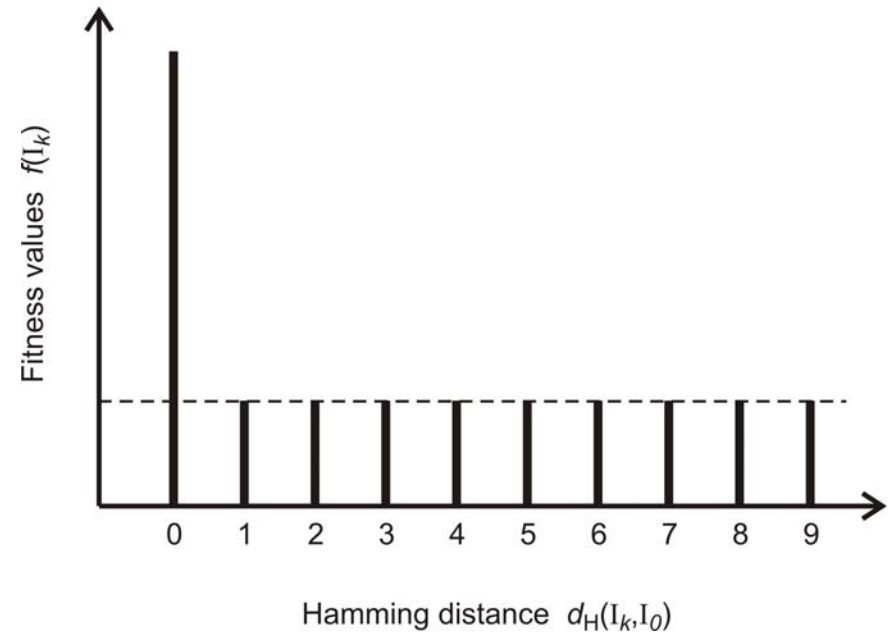
* All summations throughout this paper run from 1 to n unless specified differently: $\sum_i = \sum_{i=1}^n$ and $\sum_{i,j} = \sum_{i=1}^n \sum_{j=1}^n$, respectively.



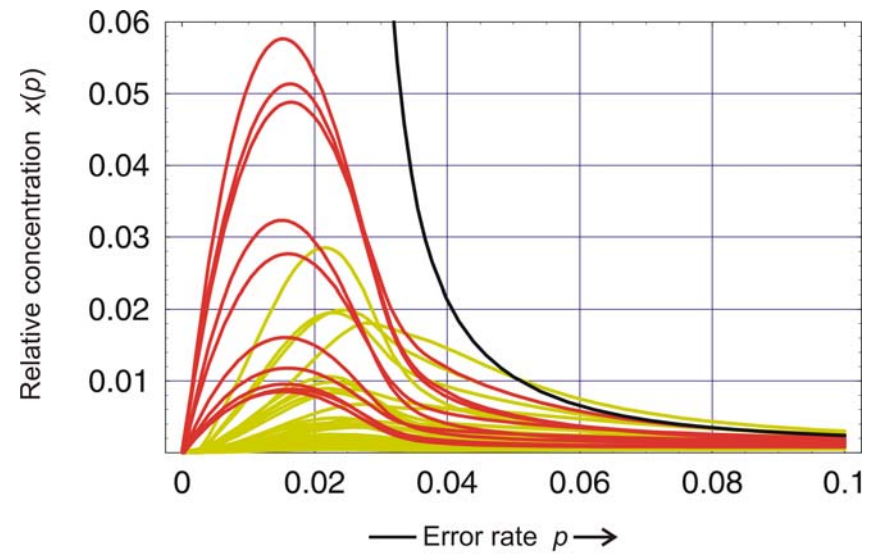
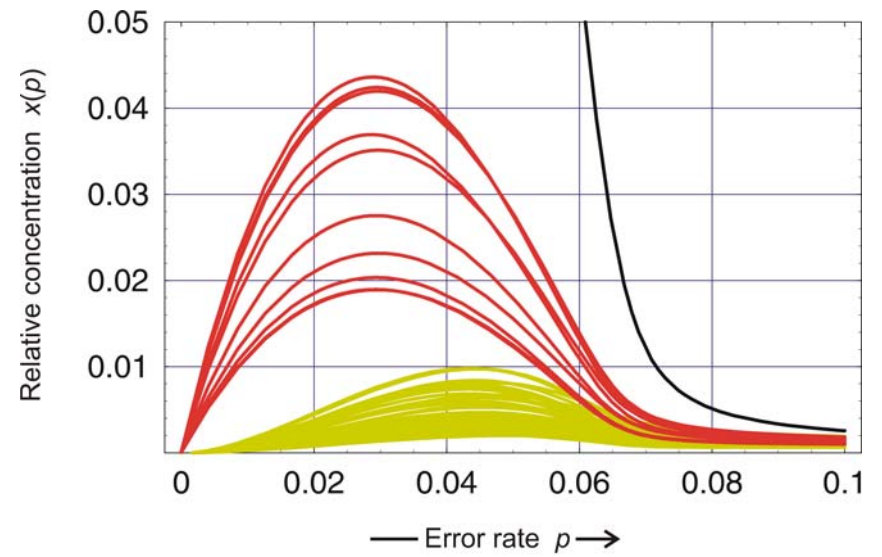
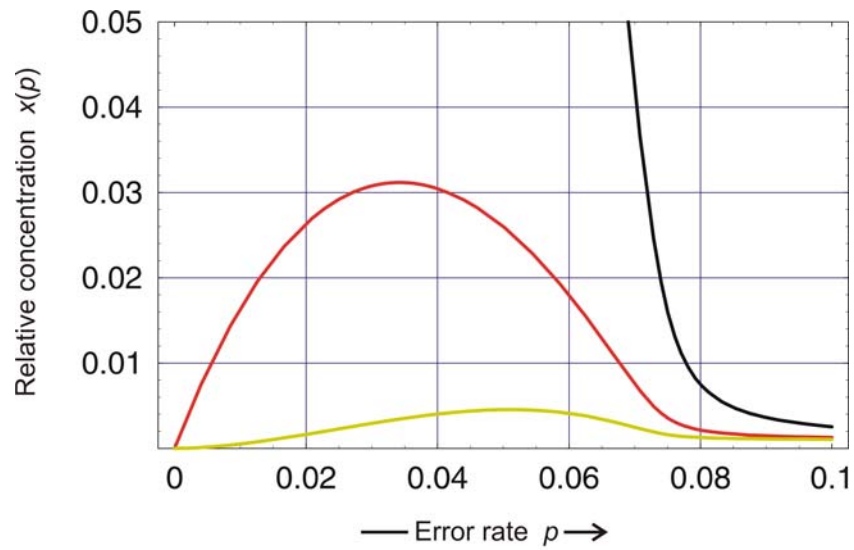
Stationary population or **quasispecies** as a function of the mutation or error rate p



Error threshold on a single peak fitness landscape with $n = 50$ and $\sigma = 10$



Fitness landscapes showing error thresholds



Error threshold: Individual sequences

$n = 10$, $\sigma = 2$ and $d = 0, 1.0, 1.85$

1. RNA replication *in vitro* and *in vivo*
2. Evolution of RNA molecules
- 3. RNA sequences and structures**
4. Evolutionary optimization of RNA structure

Prediction of RNA secondary structures: from theory to models and real molecules

Peter Schuster^{1,2}

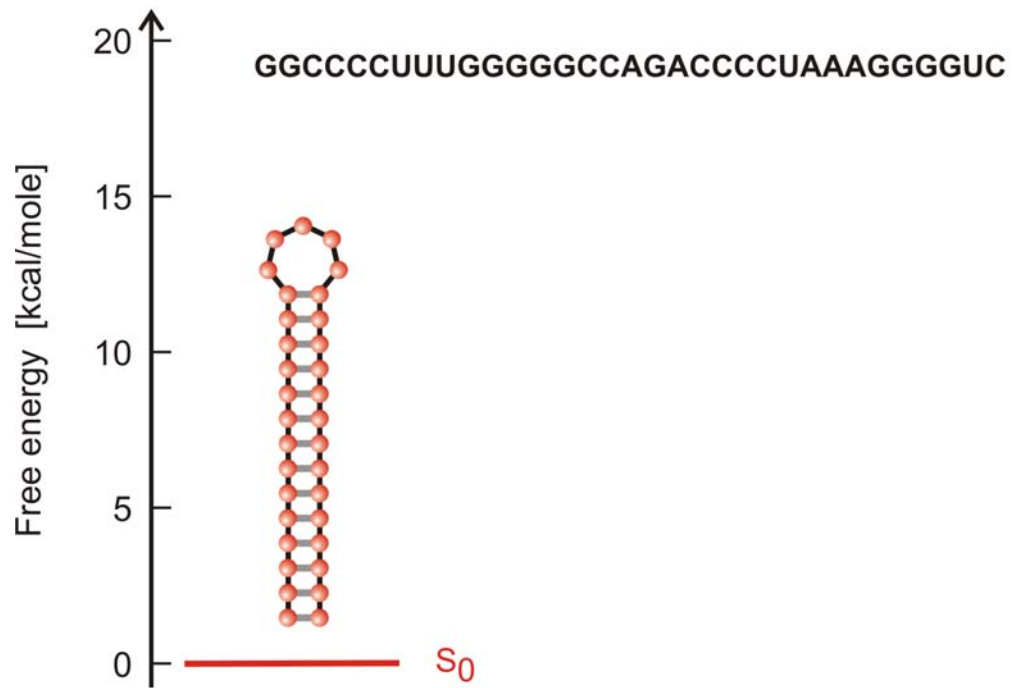
¹Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, A-1090 Vienna, Austria

²The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

E-mail: pks@tbi.univie.ac.at

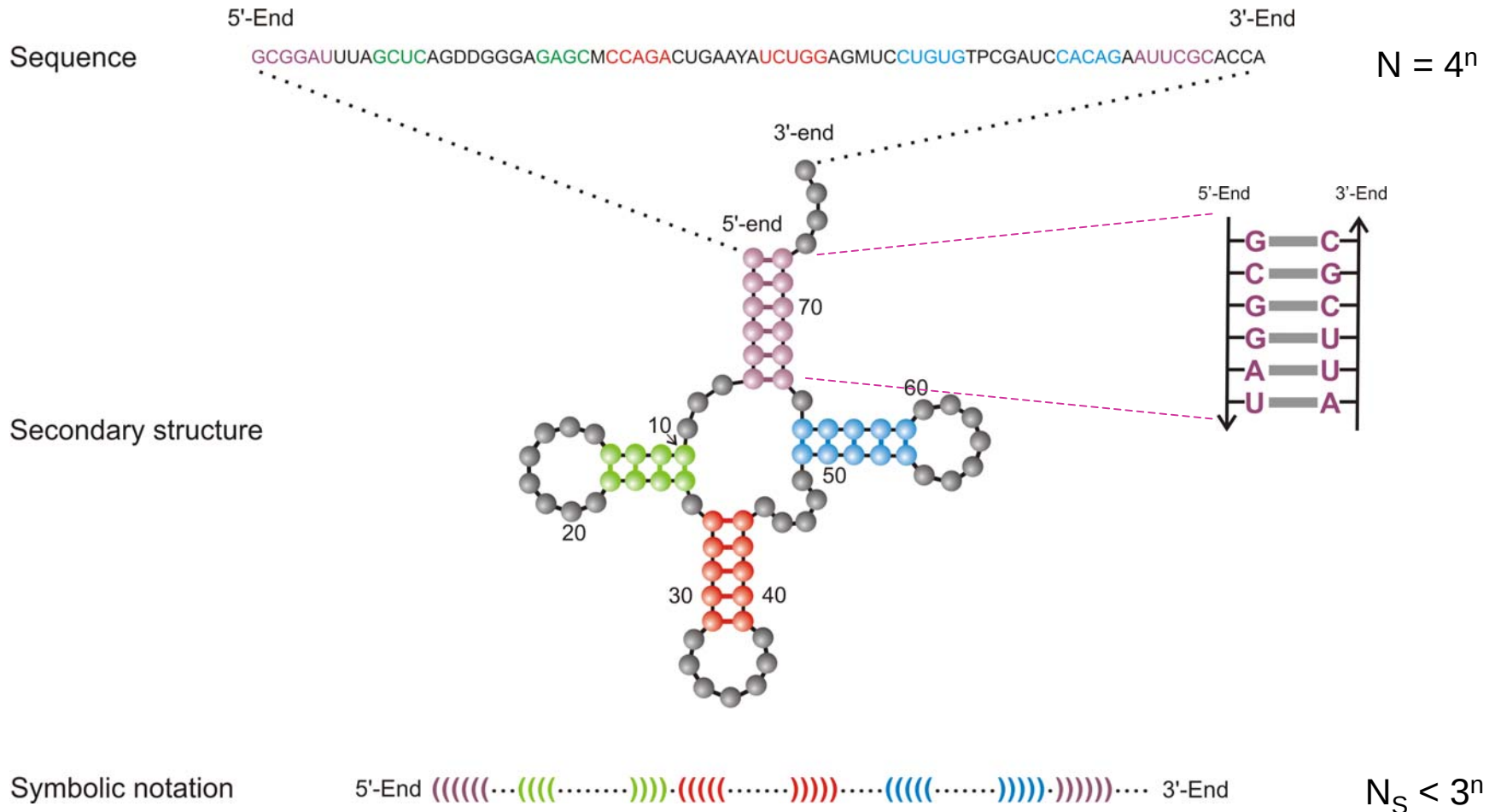
The notion of RNA (secondary) structure

- 1. Minimum free energy structure**
2. Many sequences one structure
3. Suboptimal structures
4. Kinetic structures



Minimum free energy structure

Extension of the notion of structure



Criterion: Minimum free energy (mfe)

Rules: $_ (_) _ \in \{AU, CG, GC, GU, UA, UG\}$

A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs

Fast Folding and Comparison of RNA Secondary Structures

**I. L. Hofacker^{1,*}, W. Fontana³, P. F. Stadler^{1,3}, L. S. Bonhoeffer⁴, M. Tacker¹
and P. Schuster^{1,2,3}**

¹ Institut für Theoretische Chemie, Universität Wien, A-1090 Wien, Austria

² Institut für Molekulare Biotechnologie, D-07745 Jena, Federal Republic of Germany

³ Santa Fe Institute, Santa Fe, NM 87501, U.S.A.

⁴ Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, U.K.

Summary. Computer codes for computation and comparison of RNA secondary structures, the Vienna RNA package, are presented, that are based on dynamic programming algorithms and aim at predictions of structures with minimum free energies as well as at computations of the equilibrium partition functions and base pairing probabilities.

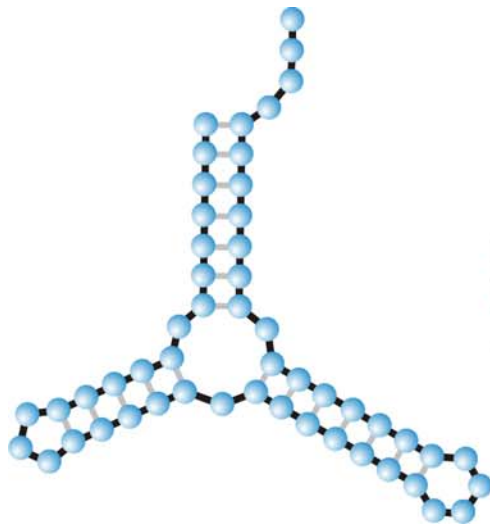
An efficient heuristic for the inverse folding problem of RNA is introduced. In addition we present compact and efficient programs for the comparison of RNA secondary structures based on tree editing and alignment.

All computer codes are written in ANSI C. They include implementations of modified algorithms on parallel computers with distributed memory. Performance analysis carried out on an Intel Hypercube shows that parallel computing becomes gradually more and more efficient the longer the sequences are.

Keywords. Inverse folding; parallel computing; public domain software; RNA folding; RNA secondary structures; tree editing.

The notion of RNA (secondary) structure

1. Minimum free energy structure
- 2. Many sequences one structure**
3. Suboptimal structures
4. Kinetic structures



Minimum free energy
criterion

1st
2nd
3rd trial
4th
5th



Inverse folding

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUUAUCUCUGUCGGGCUAGGGCGC
GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG
UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG
CAUUGGUGCUAAUGAUUUAGGGCUGUAUUCUGUAUAGCGAUCAGUGUCCG
GUAGGCCCUUGACAUAAAGAUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG

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

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

PETER SCHUSTER^{1,2,3}, WALTER FONTANA³, PETER F. STADLER^{2,3}
AND IVO L. HOFACKER²

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.

Space of genotypes: $I = \{I_1, I_2, I_3, I_4, \dots, I_N\}$; Hamming metric

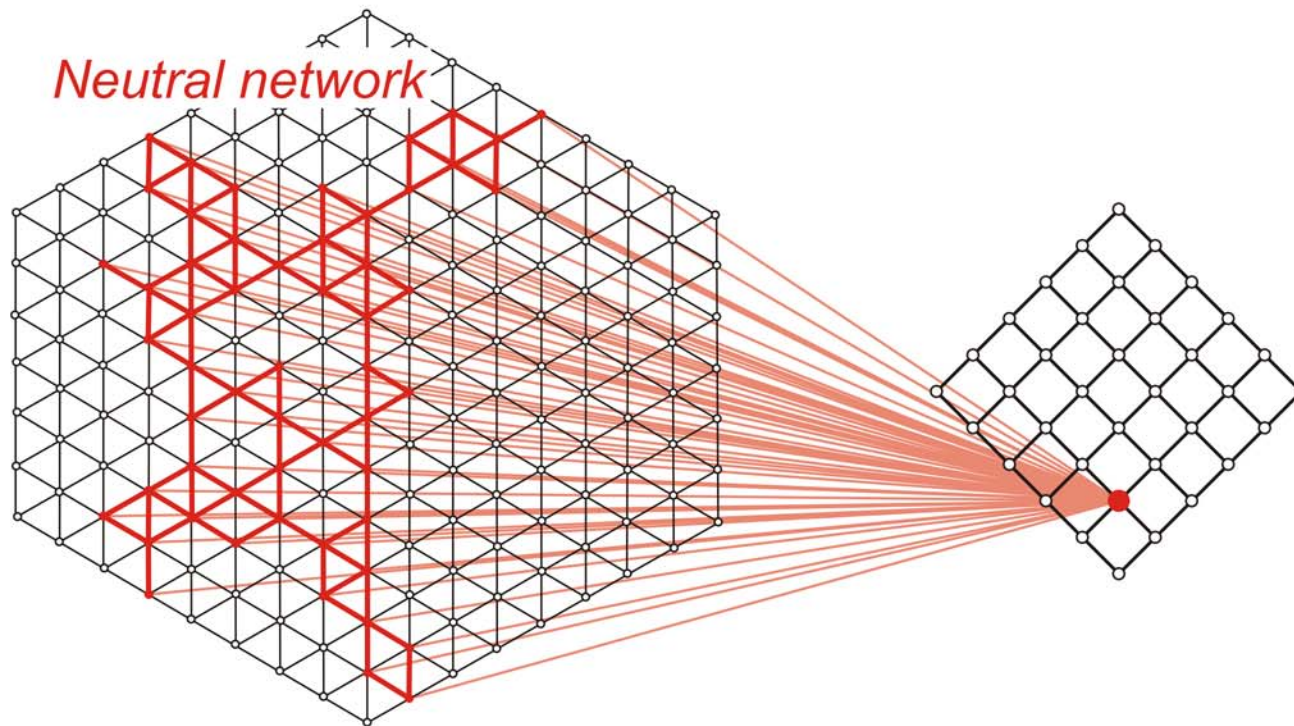
Space of phenotypes: $S = \{S_1, S_2, S_3, S_4, \dots, S_M\}$; metric (not required)

$$N \gg M$$

$$\psi(I_j) = S_k$$

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

A mapping ψ and its inversion



Sequence space

Structure space

many genotypes

$\Rightarrow$

one phenotype



S0092-8240(96)00089-4

GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES¹

■ 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.

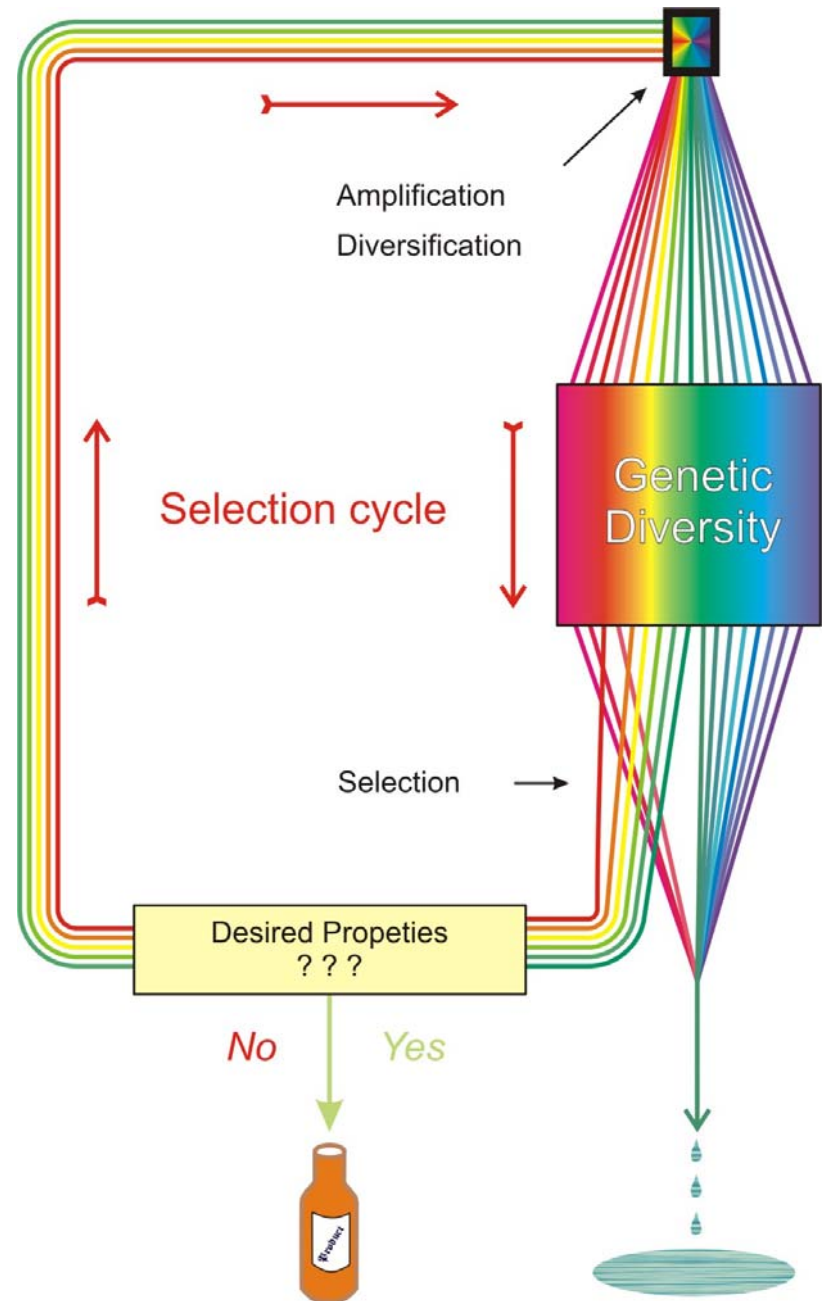
‡Institut für Theoretische Chemie der Universität Wien,
A-1090 Wien, Austria

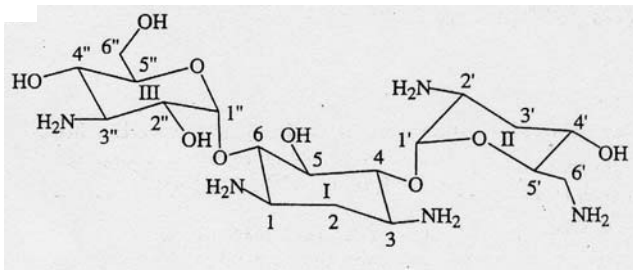
§Institut für Molekulare Biotechnologie,
D-07708 Jena, Germany

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

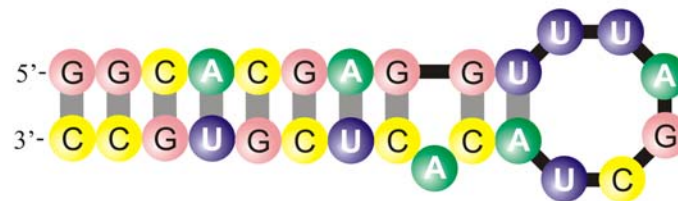
Random graph theory is used to model and analyse the relationships between sequences and secondary structures of RNA molecules, which are understood as mappings from sequence space into shape space. These maps are non-invertible since there are always many orders of magnitude more sequences than structures. 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

An example of ‘artificial selection’
with RNA molecules or ‘breeding’ of
biomolecules





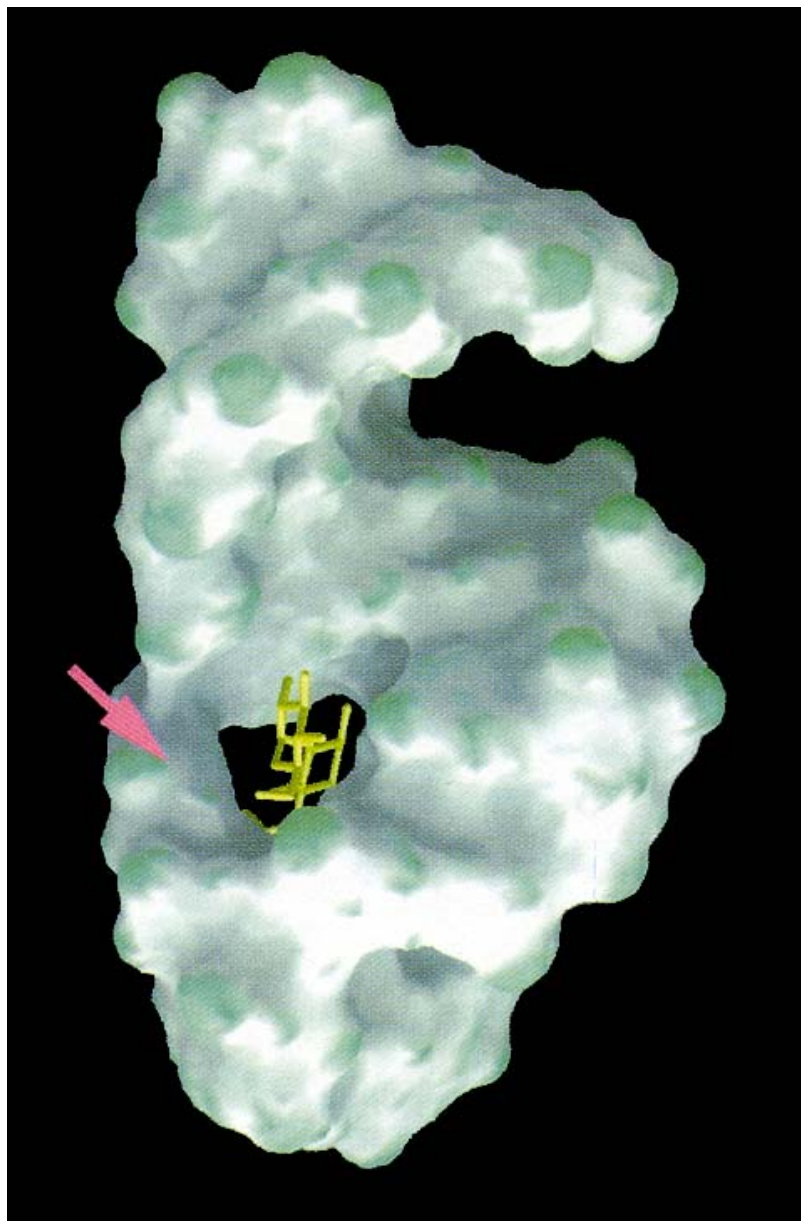
tobramycin



RNA aptamer, n = 27

Formation of secondary structure of the tobramycin binding RNA aptamer with $K_D = 9 \text{ nM}$

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)

Evolution of aptamers with a new specificity and new secondary structures from an ATP aptamer

ZHEN HUANG¹ and JACK W. SZOSTAK²

¹Department of Chemistry, Brooklyn College, Ph.D. Programs of Chemistry and Biochemistry, The Graduate School of CUNY, Brooklyn, New York 11210, USA

²Howard Hughes Medical Institute, Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

ABSTRACT

Small changes in target specificity can sometimes be achieved, without changing aptamer structure, through mutation of a few bases. Larger changes in target geometry or chemistry may require more radical changes in an aptamer. In the latter case, it is unknown whether structural and functional solutions can still be found in the region of sequence space close to the original aptamer. To investigate these questions, we designed an *in vitro* selection experiment aimed at evolving specificity of an ATP aptamer. The ATP aptamer makes contacts with both the nucleobase and the sugar. We used an affinity matrix in which GTP was immobilized through the sugar, thus requiring extensive changes in or loss of sugar contact, as well as changes in recognition of the nucleobase. After just five rounds of selection, the pool was dominated by new aptamers falling into three major classes, each with secondary structures distinct from that of the ATP aptamer. The average sequence identity between the original aptamer and new aptamers is 76%. Most of the mutations appear to play roles either in disrupting the original secondary structure or in forming the new secondary structure or the new recognition loops. Our results show that there are novel structures that recognize a significantly different ligand in the region of sequence space close to the ATP aptamer. These examples of the emergence of novel functions and structures from an RNA molecule with a defined specificity and fold provide a new perspective on the evolutionary flexibility and adaptability of RNA.

Keywords: Aptamer; specificity; fold; selection; RNA evolution

RNA 9:1456-1463, 2003

Evidence for neutral networks and shape space covering

Evolutionary Landscapes for the Acquisition of New Ligand Recognition by RNA Aptamers

Daniel M. Held, S. Travis Greathouse, Amit Agrawal, Donald H. Burke

Department of Chemistry, Indiana University, Bloomington, IN 47405-7102, USA

Received: 15 November 2002 / Accepted: 8 April 2003

Abstract. The evolution of ligand specificity underlies many important problems in biology, from the appearance of drug resistant pathogens to the re-engineering of substrate specificity in enzymes. In studying biomolecules, however, the contributions of macromolecular sequence to binding specificity can be obscured by other selection pressures critical to bioactivity. Evolution of ligand specificity *in vitro*—unconstrained by confounding biological factors—is addressed here using variants of three flavin-binding RNA aptamers. Mutagenized pools based on the three aptamers were combined and allowed to compete during *in vitro* selection for GMP-binding activity. The sequences of the resulting selection isolates were diverse, even though most were derived from the same flavin-binding parent. Individual GMP aptamers differed from the parental flavin aptamers by 7 to 26 mutations (20 to 57% overall change). Acquisition of GMP recognition coincided with the loss of FAD (flavin-adenine dinucleotide) recognition in all isolates, despite the absence of a counter-selection to remove FAD-binding RNAs. To examine more precisely the proximity of these two activities within a defined sequence space, the complete set of all intermediate sequences between an FAD-binding aptamer and a GMP-binding aptamer were synthesized and assayed for activity. For this set of sequences, we observe a portion of a neutral network for FAD-binding function separated from GMP-binding function by a distance of three muta-

tions. Furthermore, enzymatic probing of these aptamers revealed gross structural remodeling of the RNA coincident with the switch in ligand recognition. The capacity for neutral drift along an FAD-binding network in such close approach to RNAs with GMP-binding activity illustrates the degree of phenotypic buffering available to a set of closely related RNA sequences—defined as the set's functional tolerance for point mutations—and supports neutral evolutionary theory by demonstrating the facility with which a new phenotype becomes accessible as that buffering threshold is crossed.

Key words: Aptamers — RNA structure — Phenotypic buffering — Fitness landscapes — Neutral evolutionary theory — Flavin — GMP

Introduction

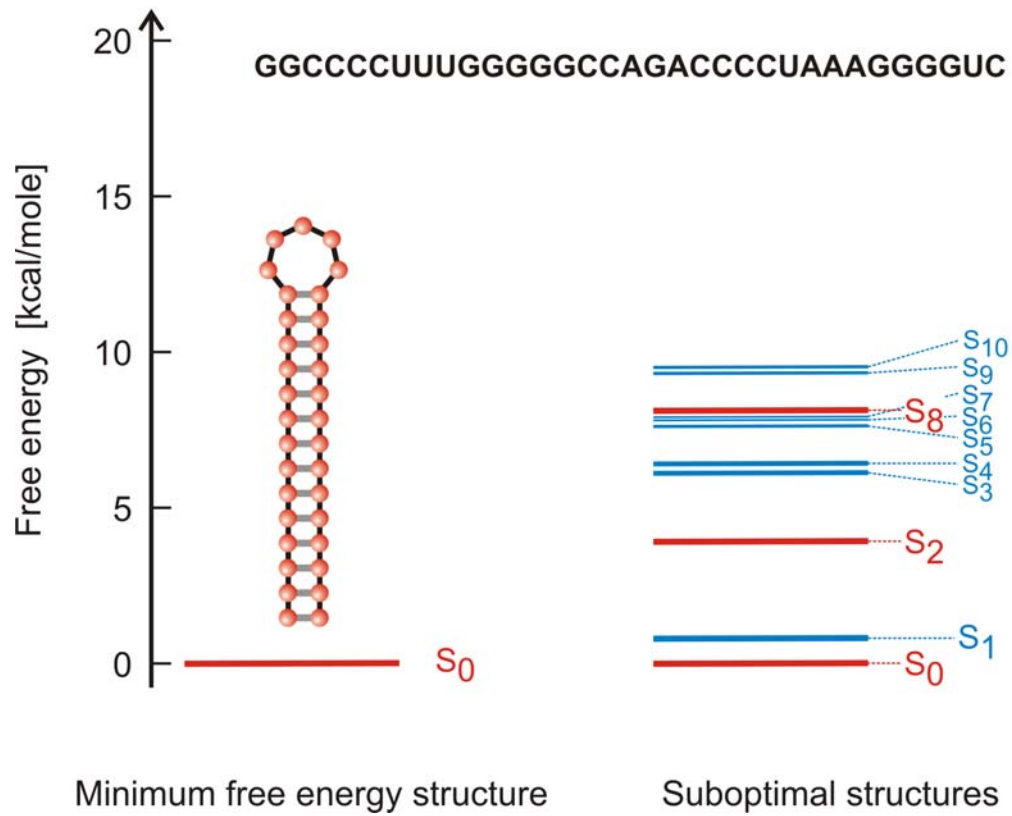
RNA aptamers targeting small molecules serve as useful model systems for the study of the evolution and biophysics of macromolecular binding interactions. Because of their small sizes, the structures of several such complexes have been determined to atomic resolution by NMR spectrometry or X-ray crystallography (reviewed by Herman and Patel 2000). Moreover, aptamers can be subjected to mutational and evolutionary pressures for which survival is based entirely on ligand binding, without the complicating effects of simultaneous selection pressures for bioactivity, thus allowing the relative contributions of each activity to be evaluated separately.

Evidence for neutral networks and intersection of aptamer functions

Correspondence to: Donald H. Burke; email: dhburke@indiana.edu

The notion of RNA (secondary) structure

1. Minimum free energy structure
2. Many sequences one structure
- 3. Suboptimal structures**
4. Kinetic structures



Extension of the notion of structure

Stefan Wuchty¹

Walter Fontana^{1,2}

Ivo L. Hofacker¹

Peter Schuster^{1,2}

¹ Institut für Theoretische
Chemie,
Universität Wien,
Währingerstrasse 17,
A-1090 Wien, Austria

² Santa Fe Institute,
1399 Hyde Park Road,
Santa Fe, NM 87501 USA

Received 13 May 1998;
accepted 6 August 1998

Complete Suboptimal Folding of RNA and the Stability of Secondary Structures

Abstract: An algorithm is presented for generating rigorously all suboptimal secondary structures between the minimum free energy and an arbitrary upper limit. The algorithm is particularly fast in the vicinity of the minimum free energy. This enables the efficient approximation of statistical quantities, such as the partition function or measures for structural diversity. The density of states at low energies and its associated structures are crucial in assessing from a thermodynamic point of view how well-defined the ground state is. We demonstrate this by exploring the role of base modification in tRNA secondary structures, both at the level of individual sequences from *Escherichia coli* and by comparing artificially generated ensembles of modified and unmodified sequences with the same tRNA structure. The two major conclusions are that (1) base modification considerably sharpens the definition of the ground state structure by constraining energetically adjacent structures to be similar to the ground state, and (2) sequences whose ground state structure is thermodynamically well defined show a significant tendency to buffer single point mutations. This can have evolutionary implications, since selection pressure to improve the definition of ground states with biological function may result in increased neutrality. © 1999 John Wiley & Sons, Inc. *Biopoly* 49: 145–165, 1999

Keywords: RNA secondary structure; suboptimal folding; density of states; tRNA; modified bases; thermodynamic stability of structure; mutational buffering; neutrality; dynamic programming

GGCUAUCGUACGUUUAC**C**AAAAGUCUACGUUGGACCCAGGCA**U**UGGACG

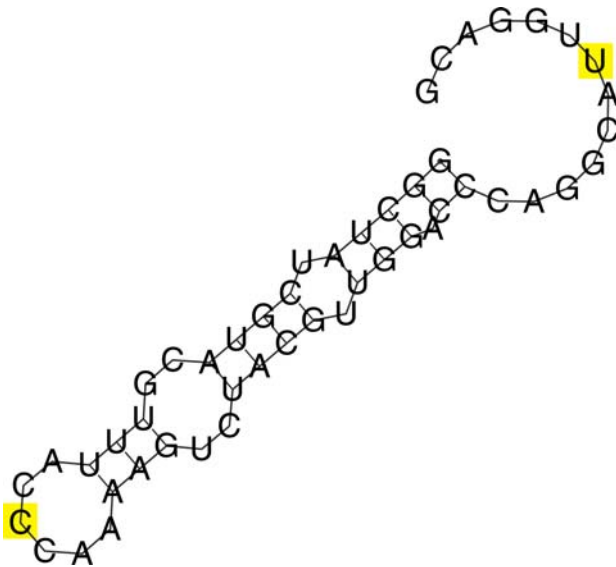
(((((.((((.((((.....)))..))))..)))..))..... -7.30
.....(((((((.(.....((((.....))))..))..)))))) -6.70
.....(((((((.(.....((((.....))))..))..)))))) -6.60
..(((((((.((((.....)))..))))..))..((((.....))).. -6.10
(((((((.((((.....)))..))))..))..).((.....)). -6.00
(((((((.((((.....)))..))))..))..).((.....)). -6.00
..(((((((.((((.....)))..))))..))..).((.....)). -6.00

GGCUAUCGUACGUUUAC**A**AAAAGUCUACGUUGGACCCAGGCA**U**UGGACG

(((((.((((.((((.....)))..))))..)))..))..... -7.30
..(((((((.((((.....)))..))))..))..).((.....)). -6.50
..(((((((.((((.....)))..))))..))..).((.....)). -6.30
..(((((((.((((.....)))..))))..))..).((.....)). -6.10
(((((((.((((.....)))..))))..))..).((.....)). -6.00
(((((((.((((.....)))..))))..))..).((.....)). -6.00
..(((((((.((((.....)))..))))..))..).((.....)). -6.00

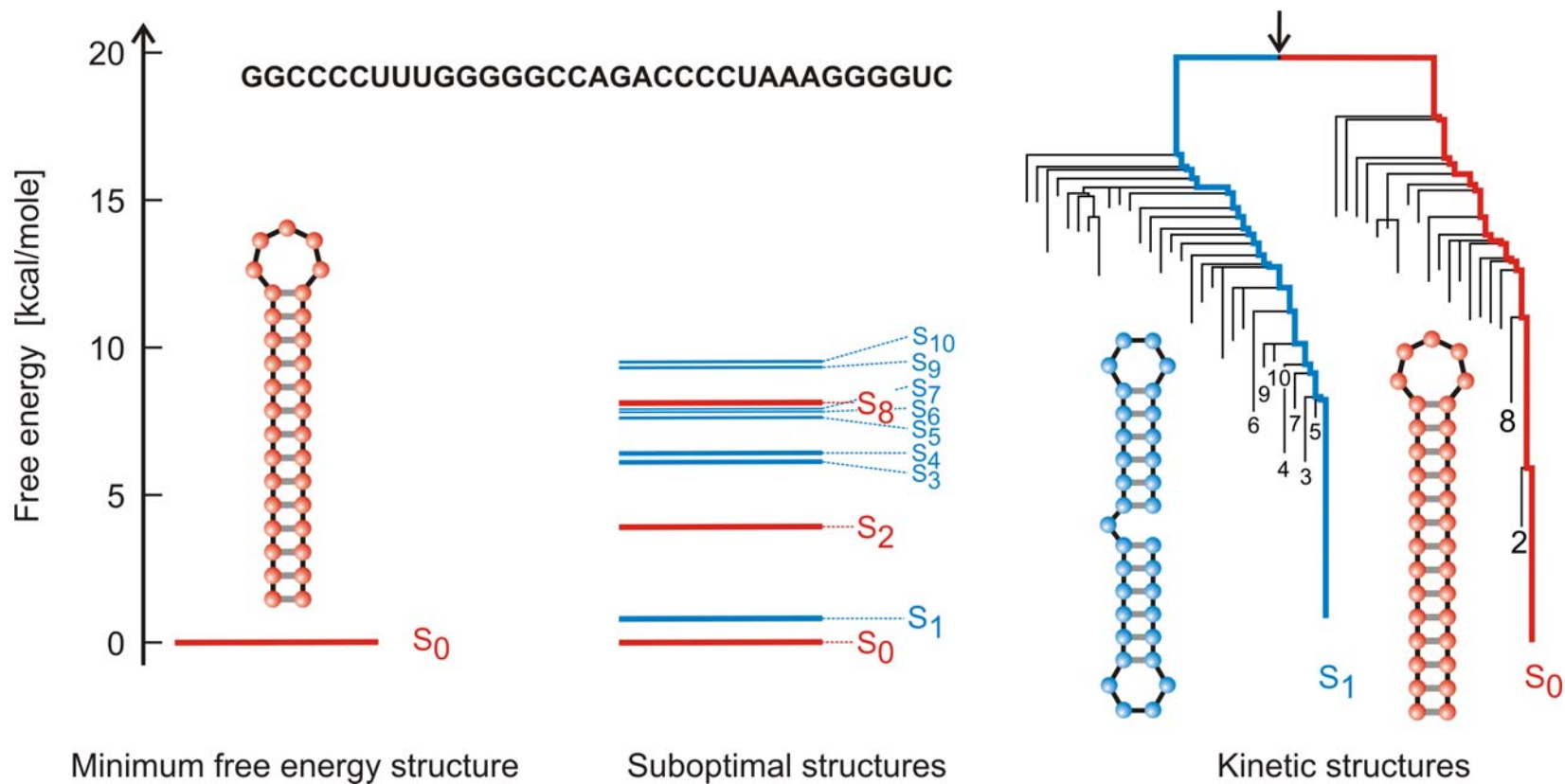
GGCUAUCGUACGUUUAC**C**AAAAGUCUACGUUGGACCCAGGCA**A**UGGACG

(((((.((((.((((.....)))..))))..)))..))..... -7.30
..(((((((.((((.....)))..))))..))..).((.....)). -7.20
.....(((((((.(.....((((.....))))..))..)))))) -6.70
.....(((((((.(.....((((.....))))..))..)))))) -6.60
(((((((.((((.....)))..))))..))..).((.....)). -6.50
..(((((((.((((.....)))..))))..))..).((.....)). -6.30
..(((((((.((((.....)))..))))..))..).((.....)). -6.30
.....(((((((.((((.....)))..))))..))..).((.....)). -6.30
..(((((((.((((.....)))..))))..))..).((.....)). -6.10
.....(((((((.((((.....)))..))))..))..).((.....)). -6.10
.....(((((((.((((.....)))..))))..))..).((.....)). -6.10
(((((((.((((.....)))..))))..))..).((.....)). -6.00
(((((((.((((.....)))..))))..))..).((.....)). -6.00
..(((((((.((((.....)))..))))..))..).((.....)). -6.00
.....(((((((.((((.....)))..))))..))..).((.....)). -6.00



The notion of RNA (secondary) structure

1. Minimum free energy structure
2. Many sequences one structure
3. Suboptimal structures
4. **Kinetic structures**



Extension of the notion of structure

RNA folding at elementary step resolution

CHRISTOPH FLAMM,¹ WALTER FONTANA,^{2,3} IVO L. HOFACKER,¹
and PETER SCHUSTER¹

¹Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, A-1090 Wien, Austria

²Santa Fe Institute, Santa Fe, New Mexico 87501 USA

ABSTRACT

We study the stochastic folding kinetics of RNA sequences into secondary structures with a new algorithm based on the formation, dissociation, and the shifting of individual base pairs. We discuss folding mechanisms and the correlation between the barrier structure of the conformational landscape and the folding kinetics for a number of examples based on artificial and natural sequences, including the influence of base modification in tRNAs.

Keywords: conformational spaces; foldability; RNA folding kinetics; RNA secondary structure

The Folding Algorithm

A sequence **I** specifies an energy ordered set of compatible structures $\mathfrak{S}(\mathbf{I})$:

$$\mathfrak{S}(\mathbf{I}) = \{S_0, S_1, \dots, S_m, \mathbf{O}\}$$

A trajectory $\mathfrak{T}_k(\mathbf{I})$ is a time ordered series of structures in $\mathfrak{S}(\mathbf{I})$. A folding trajectory is defined by starting with the open chain **O** and ending with the global minimum free energy structure S_0 or a metastable structure S_k which represents a local energy minimum:

$$\mathfrak{T}_0(\mathbf{I}) = \{\mathbf{O}, S(1), \dots, S(t-1), S(t), \\ S(t+1), \dots, S_0\}$$

$$\mathfrak{T}_k(\mathbf{I}) = \{\mathbf{O}, S(1), \dots, S(t-1), S(t), \\ S(t+1), \dots, S_k\}$$

Master equation

$$\frac{dP_k}{dt} = \sum_{i=0}^{m+1} (P_{ik}(t) - P_{ki}(t)) = \sum_{i=0}^{m+1} k_{ik} P_i - P_k \sum_{i=0}^{m+1} k_{ki} \\ k = 0, 1, \dots, m+1$$

Transition probabilities $P_{ij}(t) = \text{Prob}\{S_i \rightarrow S_j\}$ are defined by

$$P_{ij}(t) = P_i(t) k_{ij} = P_i(t) \exp(-\Delta G_{ij}/2RT) / \Sigma_i$$

$$P_{ji}(t) = P_j(t) k_{ji} = P_j(t) \exp(-\Delta G_{ji}/2RT) / \Sigma_j$$

$$\Sigma_k = \sum_{k=1, k \neq i}^{m+2} \exp(-\Delta G_{ki}/2RT)$$

The symmetric rule for transition rate parameters is due to Kawasaki (K. Kawasaki, *Diffusion constants near the critical point for time dependent Ising models*. Phys.Rev. **145**:224-230, 1966).

Formulation of kinetic RNA folding as a stochastic process

Efficient computation of RNA folding dynamics

Michael T Wolfinger¹, W Andreas Svrcek-Seiler¹, Christoph Flamm¹,
Ivo L Hofacker¹ and Peter F Stadler²

¹ Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien,
Währingerstraße 17, A-1090 Wien, Austria

² Bioinformatik, Institut für Informatik, Universität Leipzig, D-04103 Leipzig, Germany

E-mail: ivo@tbi.univie.ac.at

Received 22 December 2003, in final form 8 March 2004

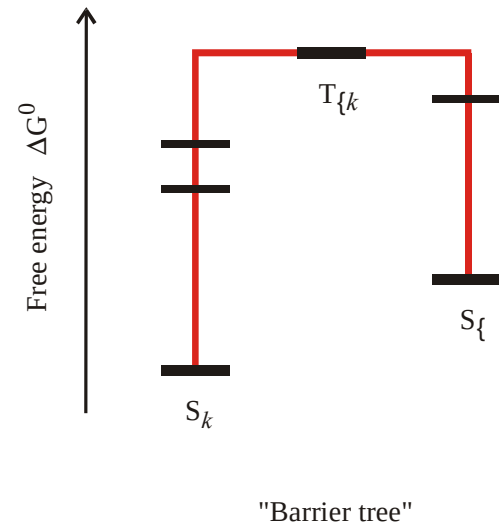
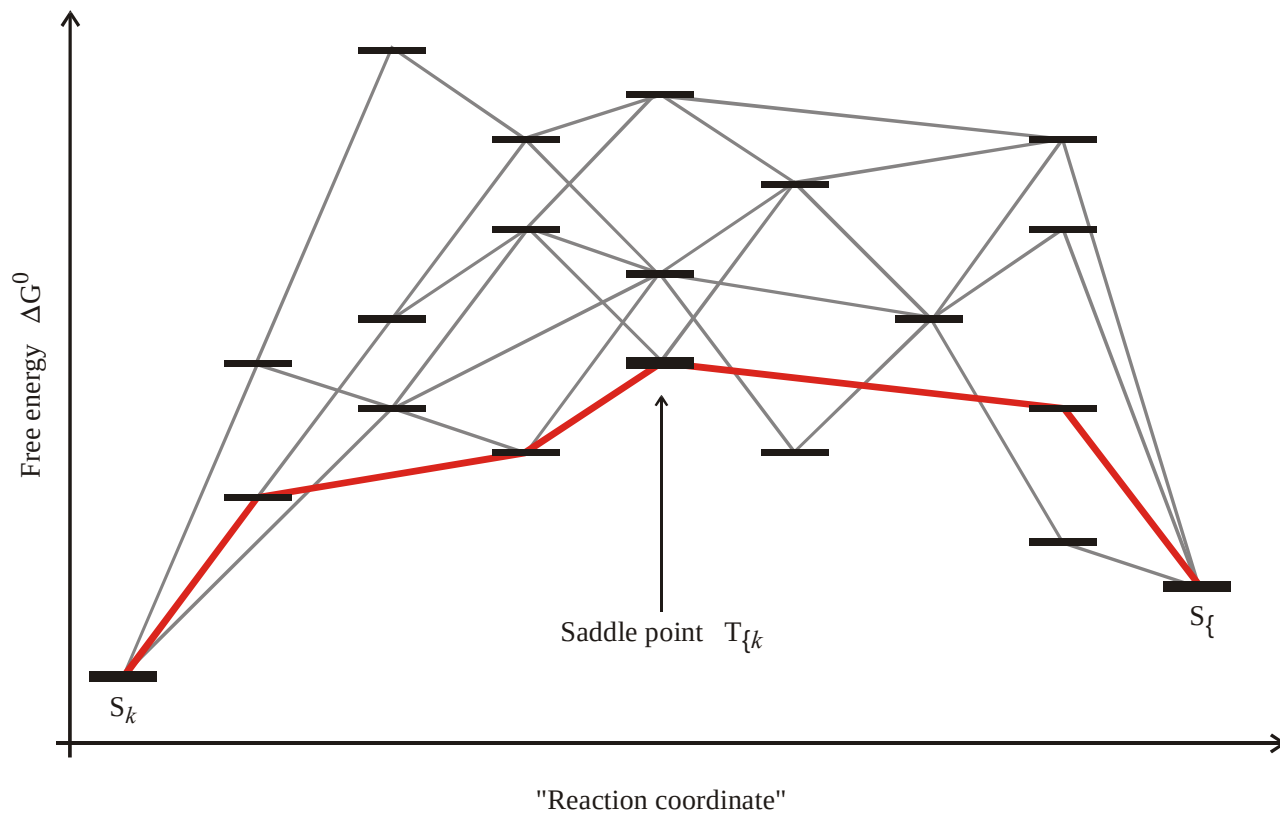
Published 14 April 2004

Online at stacks.iop.org/JPhysA/37/4731 (DOI: 10.1088/0305-4470/37/17/005)

Abstract

Barrier trees consisting of local minima and their connecting saddle points imply a natural coarse-graining for the description of the energy landscape of RNA secondary structures. Here we show that, based on this approach, it is possible to predict the folding behaviour of RNA molecules by numerical integration. Comparison with stochastic folding simulations shows reasonable agreement of the resulting folding dynamics and a drastic increase in computational efficiency that makes it possible to investigate the folding dynamics of RNA of at least tRNA size. Our approach is readily applicable to bistable RNA molecules and promises to facilitate studies on the dynamic behaviour of RNA switches.

PACS numbers: 87.14.Gg, 87.15.He, 87.15.Aa, 87.15.Cc



Definition of a ,barrier tree‘

Structural parameters affecting the kinetics of RNA hairpin formation

J. H. A. Nagel, C. Flamm¹, I. L. Hofacker¹, K. Franke², M. H. de Smit,
P. Schuster¹ and C. W. A. Pleij^{*}

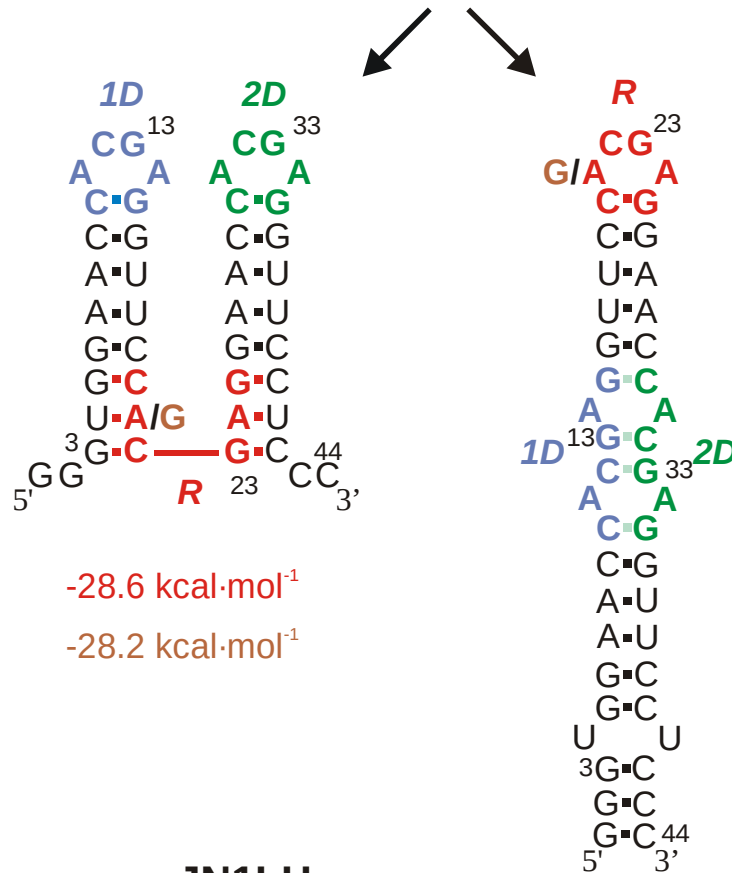
Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, 2300 RA Leiden, The Netherlands,
¹Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, A-1090 Vienna, Austria
and ²IBA NAPS GmbH Rudolf-Wissell-Strasse 28 D-37079 Göttingen, Germany

Received January 28, 2005; Revised and Accepted June 7, 2006

ABSTRACT

There is little experimental knowledge on the sequence dependent rate of hairpin formation in RNA. We have therefore designed RNA sequences that can fold into either of two mutually exclusive hairpins and have determined the ratio of folding of the two conformations, using structure probing. This folding ratio reflects their respective folding rates. Changing one of the two loop sequences from a purine- to a pyrimidine-rich loop did increase its folding rate, which corresponds well with similar observations in DNA hairpins. However, neither changing one of the loops from a regular non-GNRA tetra-loop into a stable GNRA tetra-loop, nor increasing the loop size from 4 to 6 nt did affect the folding rate. The folding kinetics of these RNAs have also been simulated with the program 'Kinfold'. These simulations were in agreement with the experimental results if the additional stabilization energies for stable tetra-loops were not taken into account. Despite the high stability of the stable tetra-loops, they apparently do not affect folding kinetics of these RNA hairpins. These results show that it is possible to experimentally determine relative folding rates of hairpins and to use these data to improve the computer-assisted simulation of the folding kinetics of stem-loop structures.

^{1D} ^R ^{2D}
 GGGUGGAAC**CACGAG**GUUUC**CACGAG**GAAC**CACGAG**GUUCCUCCC
 3 13 23 33 44
 G



-28.6 kcal·mol⁻¹

-28.2 kcal·mol⁻¹

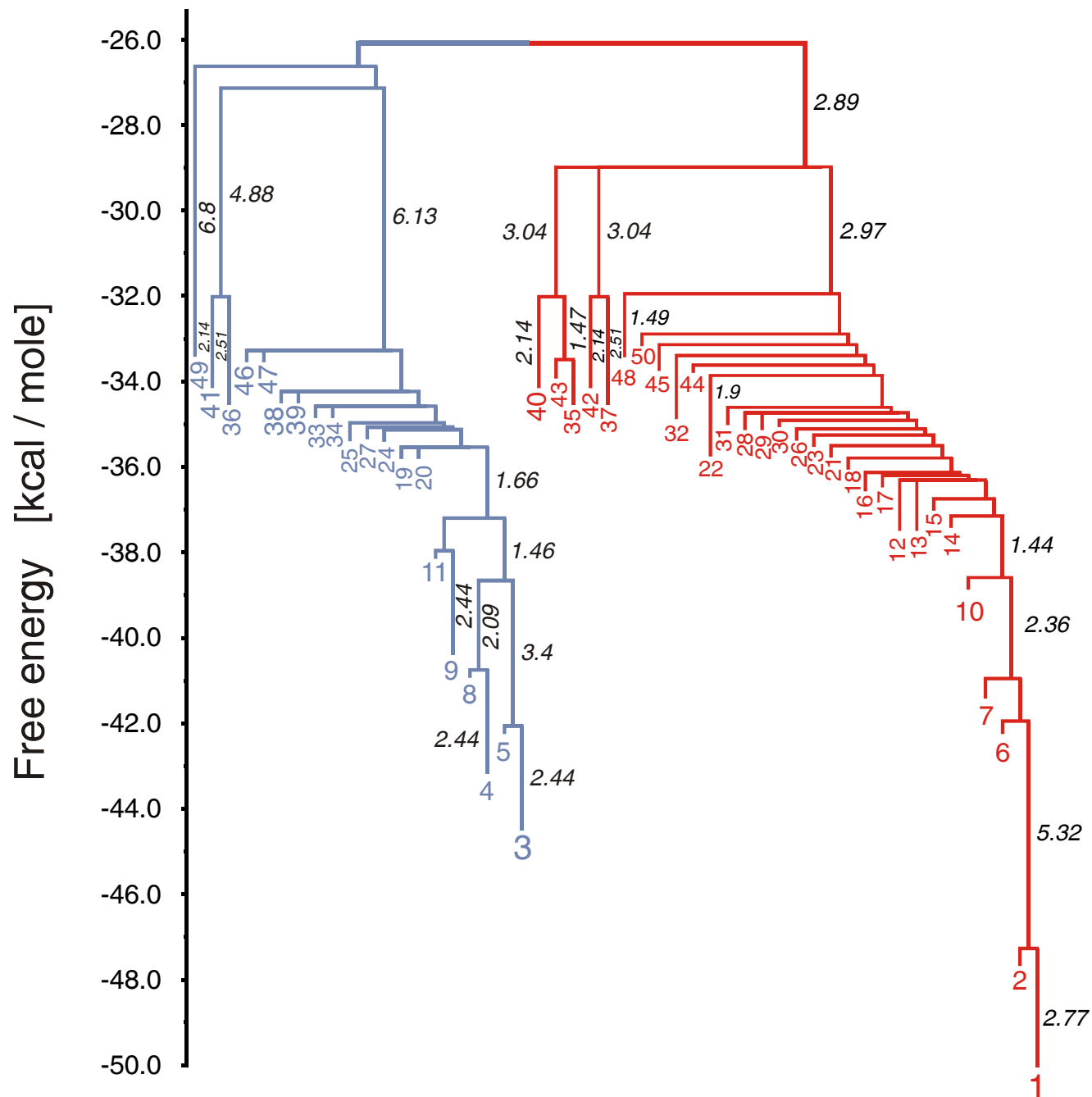
JN1LH

-28.6 kcal·mol⁻¹

-31.8 kcal·mol⁻¹

An experimental RNA switch

J1LH barrier tree



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 μ M) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μ M) 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 μ l of 50 mM Tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton

REPORTS

X-100 for 15 min at 4°C with intermittent mixing, and elutes were pooled. Proteins were precipitated by MeOH/CH₂Cl₂ 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. Horadzsky, *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 rbt1, 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 dis-

crete 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

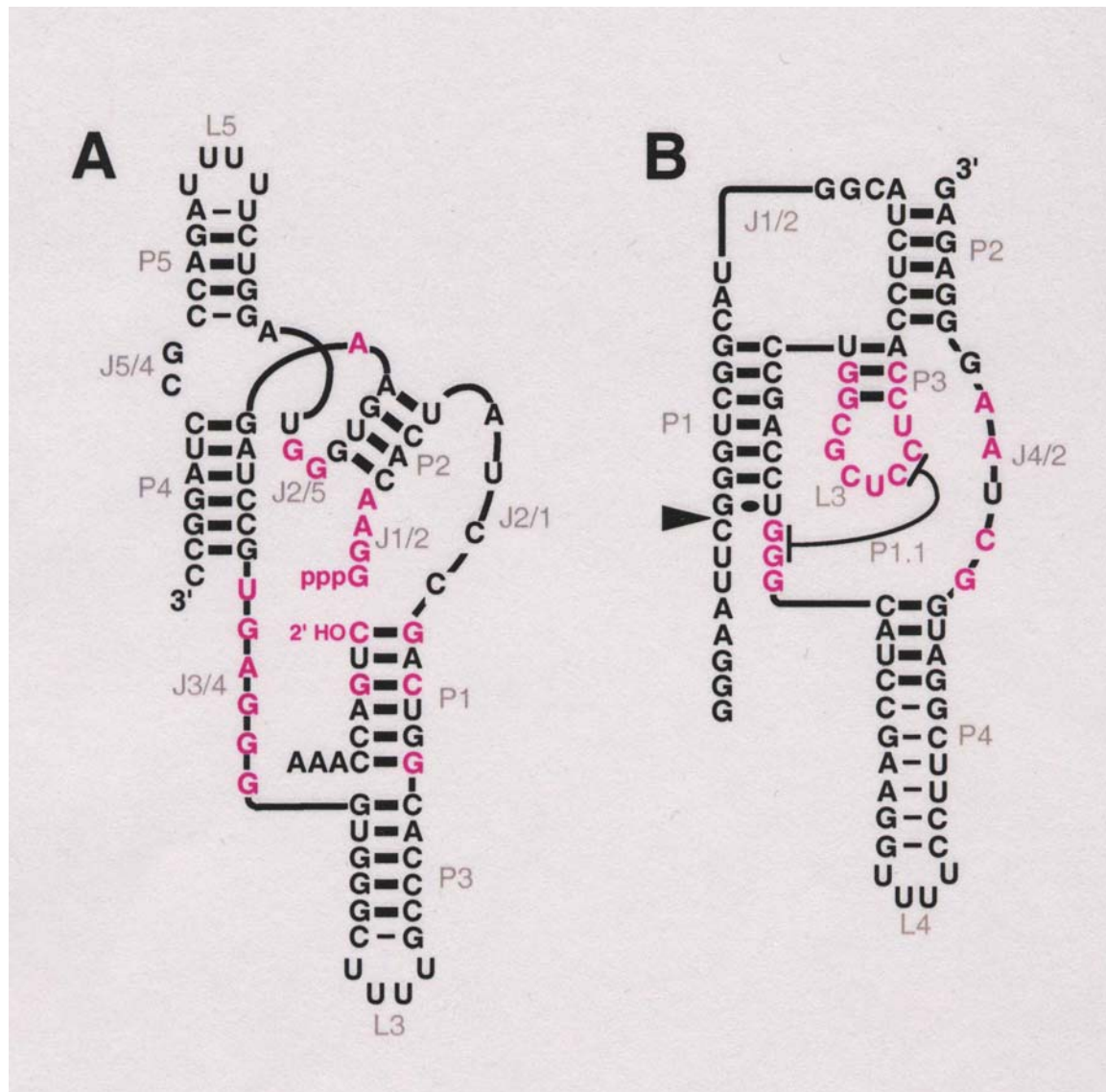
have two different catalytic activities, and could access by neutral drift every sequence on both networks. With intersecting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry non-functional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

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

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

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

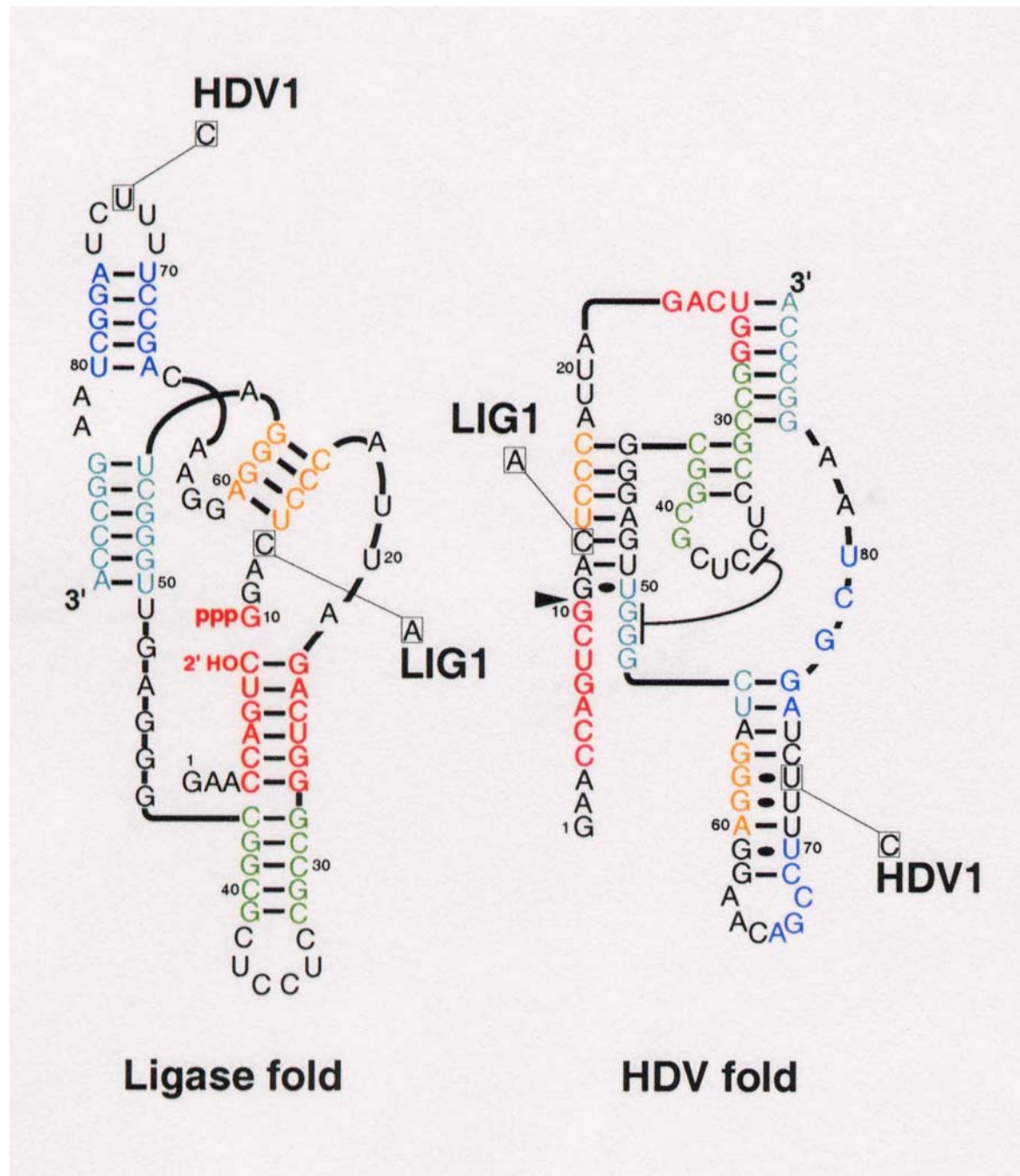
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

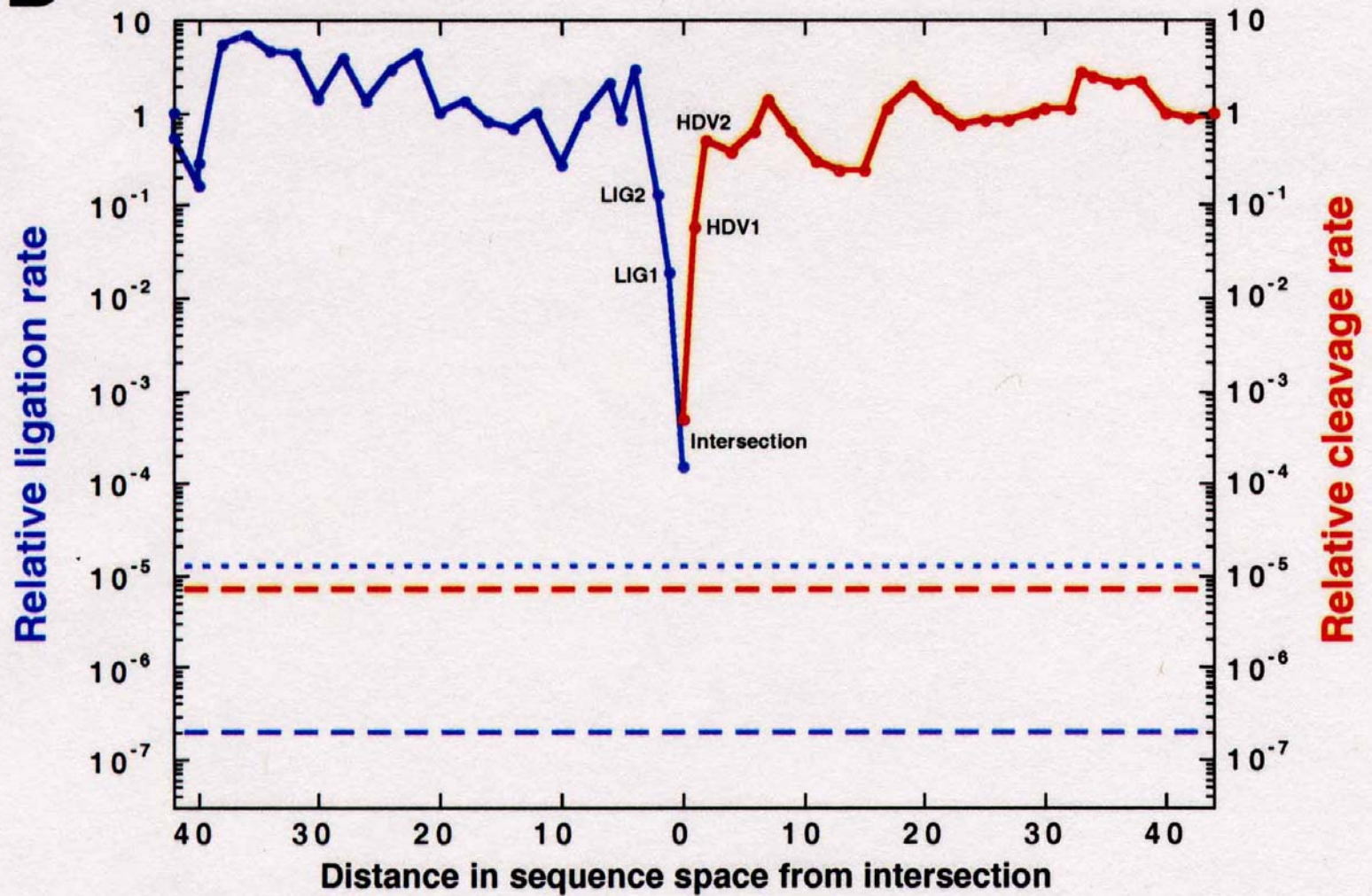


Two ribozymes of chain lengths $n = 88$ nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis- δ -virus (**B**)

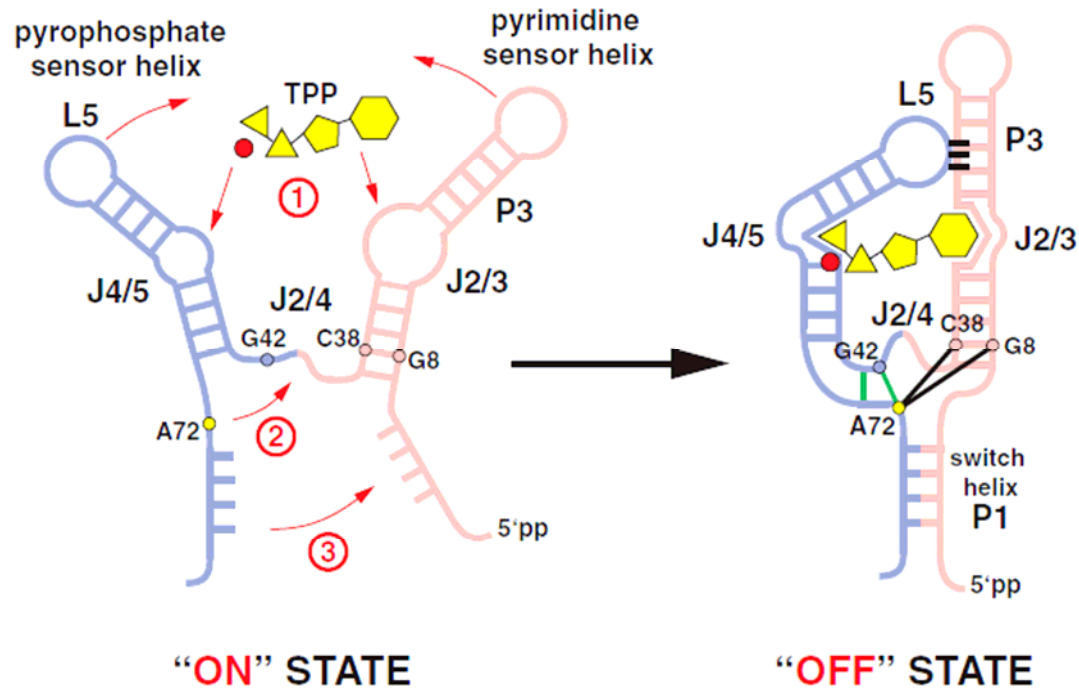
The sequence at the *intersection*:

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



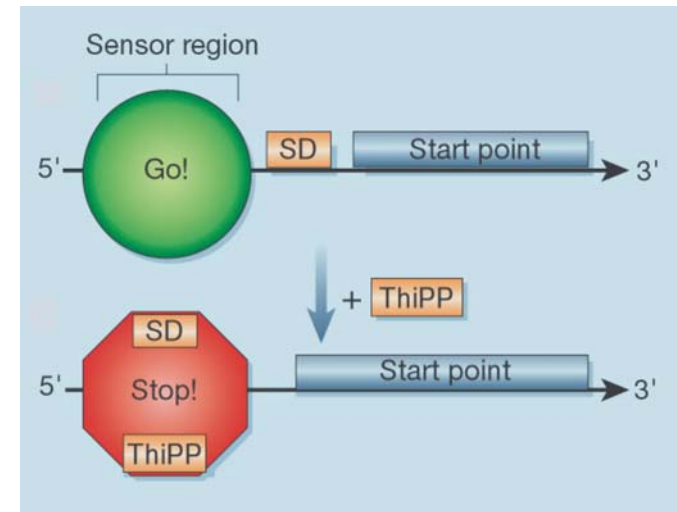
B

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

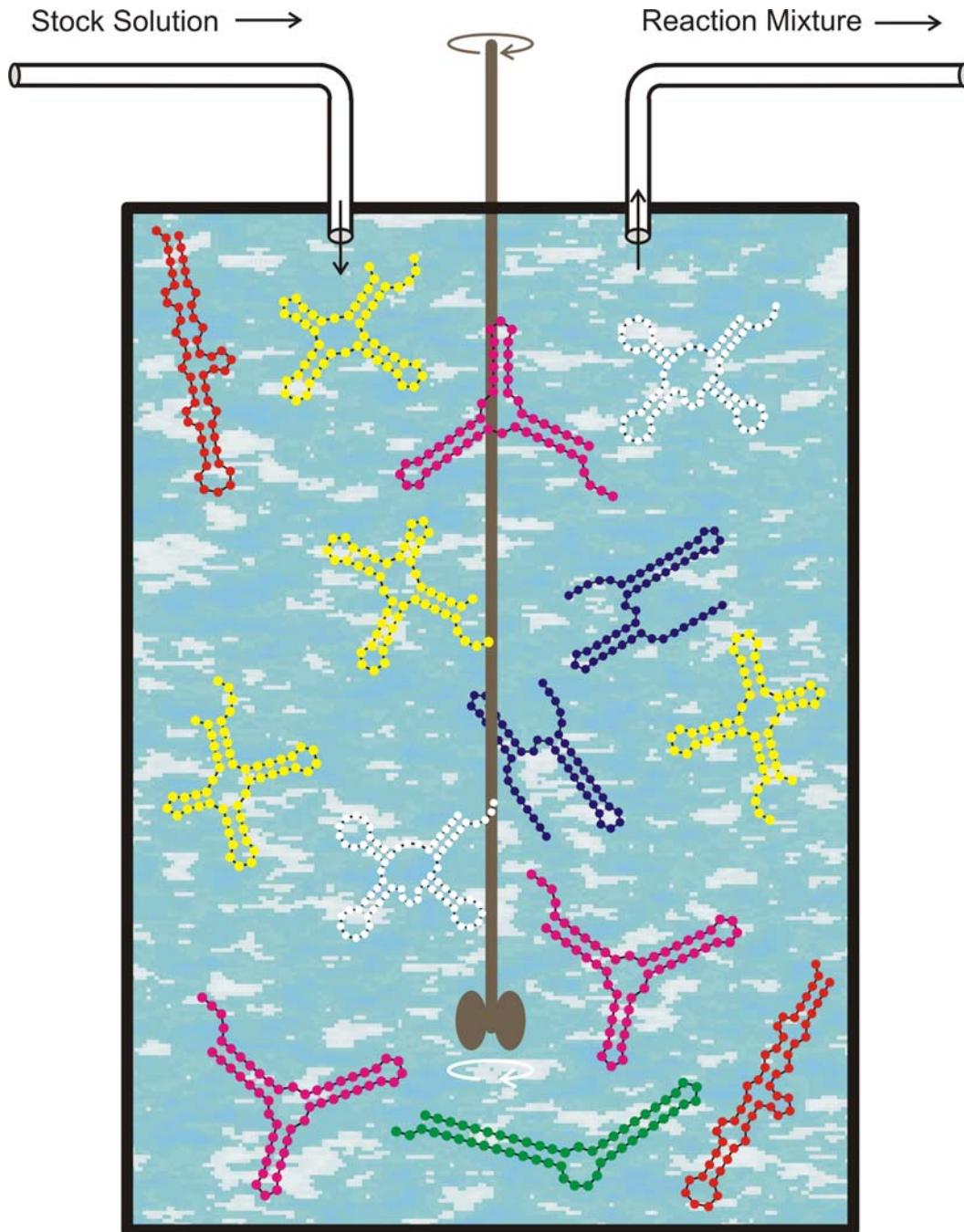


The thiamine-pyrophosphate riboswitch

S. Thore, M. Leibundgut, N. Ban.
Science **312**:1208-1211, 2006.



1. RNA replication *in vitro* and *in vivo*
2. Evolution of RNA molecules
3. RNA sequences and structures
4. **Evolutionary optimization of RNA structure**



Computer simulation using
Gillespie's algorithm:

Replication rate constant:

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

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

Selection constraint:

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

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

Mutation rate:

$$p = 0.001 / \text{site} \times \text{replication}$$

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

random individuals. The primer pair used for genomic DNA amplification is 5'-TCTCCTGGATTCT-CATTAA-3' (forward) and 5'-TCTTTGTCTCTGT-TGCAAC-3' (reverse). Reactions were performed in 25 µl using 1 unit of Taq DNA polymerase with each primer at 0.4 µM, 200 µM each dATP, dTTP, dCTP, and dGTP, and PCR buffer [10 mM Tris-HCl (pH 8.3), 50 mM KCl, 1.5 mM MgCl₂] in a cycle condition of 94°C for 1 min and then 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s followed by 72°C for 6 min. PCR products were purified (Qiagen), digested with Xmn I, and separated in a 2% agarose gel.

32. A nonsense mutation may affect mRNA stability and result in degradation of the transcript [L. Maquat, *Am. J. Hum. Genet.* **59**, 279 (1996)].

33. Data not shown; a dot blot with poly (A)⁺ RNA from 50 human tissues (The Human RNA Master Blot, 7770-1, Clontech Laboratories) was hybridized with a probe from exons 29 to 47 of MYO15 using the same condition as Northern blot analysis [13].

34. Smith-Magenis syndrome (SMS) is due to deletions of 17p11.2 of various sizes, the smallest of which includes MYO15 and perhaps 20 other genes [6]; K-S Chen, L. Potocki, J. R. Lupski, *MROD Res. Rev.* **2**, 122 (1996). MYO15 expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for MYO15 may explain a portion of the SMS

phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in MYO15 in trans to the SMS 17p11.2 deletion.

35. R. A. Fiedel, data not shown.

36. K. B. Avraham *et al.*, *Nature Genet.* **11**, 369 (1995); X-Z. Liu *et al.*, *ibid.* **17**, 268 (1997); F. Gibson *et al.*, *Nature* **374**, 62 (1995); D. Weil *et al.*, *ibid.*, p. 60.

37. RNA was extracted from cochlea (menstrual labyrinth) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly (A)⁺ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human MYO15-specific oligonucleotide primers (forward, 5'-GCATGACCTGCGGGTAAT-GGG-3'; reverse, 5'-CTCAGGGCTTCTGCGATGGT-GCTGGCTGGC-3'). Cycling conditions were 40 s at 94°C, 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR

product is expected from amplification of the human MYO15 cDNA. Amplification of human genomic DNA with this primer pair would result in a 2903-bp fragment.

38. We are grateful to the people of Bengala, Bali, and the two families from India. We thank J. R. Lupski and K-S. Chen for providing the human chromosome 17 cosmid library. For technical and computational assistance, we thank N. Dietrich, M. Ferguson, A. Gupta, E. Sorbello, R. Torzadeh, C. Varner, M. Walker, G. Bouffard, and S. Beckstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Hinnant, I. N. Arhya, and S. Winata for assistance in Bali, and T. Barber, S. Sullivan, E. Green, D. Drayna, and J. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (201 DC 00035-01 and 201 DC 00038-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.G.M.), the National Institute of Child Health and Human Development (R01 HD30428 to S.A.C.) and a National Science Foundation Graduate Research Fellowship to F.J.P. This paper is dedicated to J. B. Snow Jr. on his retirement as the Director of the NIDCD.

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

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

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

We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (2). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empiri-

cally well defined and obtain their biophysical and biochemical importance from being a scaffold for the tertiary structure. For the sake of brevity, we shall refer to secondary structures as "shapes." RNA combines in a single molecule both genotype (replicable sequence) and phenotype (selectable shape), making it ideally suited for in vitro evolution experiments (3, 4).

To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (5, 6). In the laboratory, a goal might be to find an RNA aptamer binding specifically to a molecule (4). Although in the experiment the evolutionary end product was unknown, we thought of its shape as being specified implicitly by the imposed selection criterion. Because our intent is to study evolutionary histories rather than end products, we defined a target shape in advance and assumed the replication rate of a sequence to be a function of

the similarity between its shape and the target. An actual situation may involve more than one best shape, but this does not affect our conclusions.

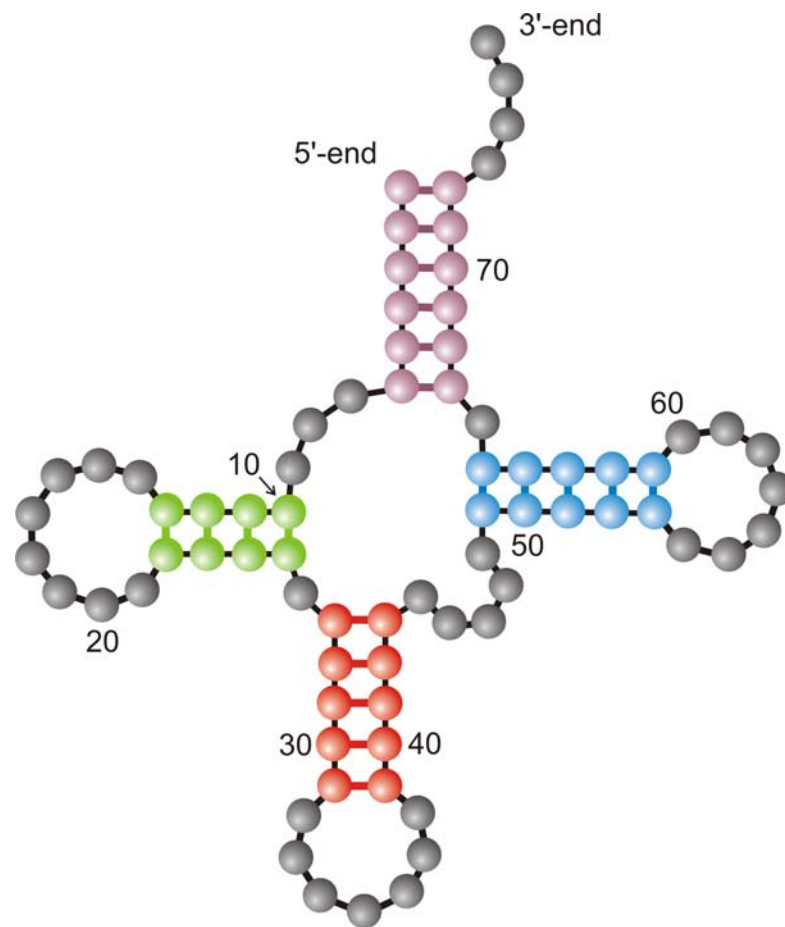
An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces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

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

Evolution *in silico*

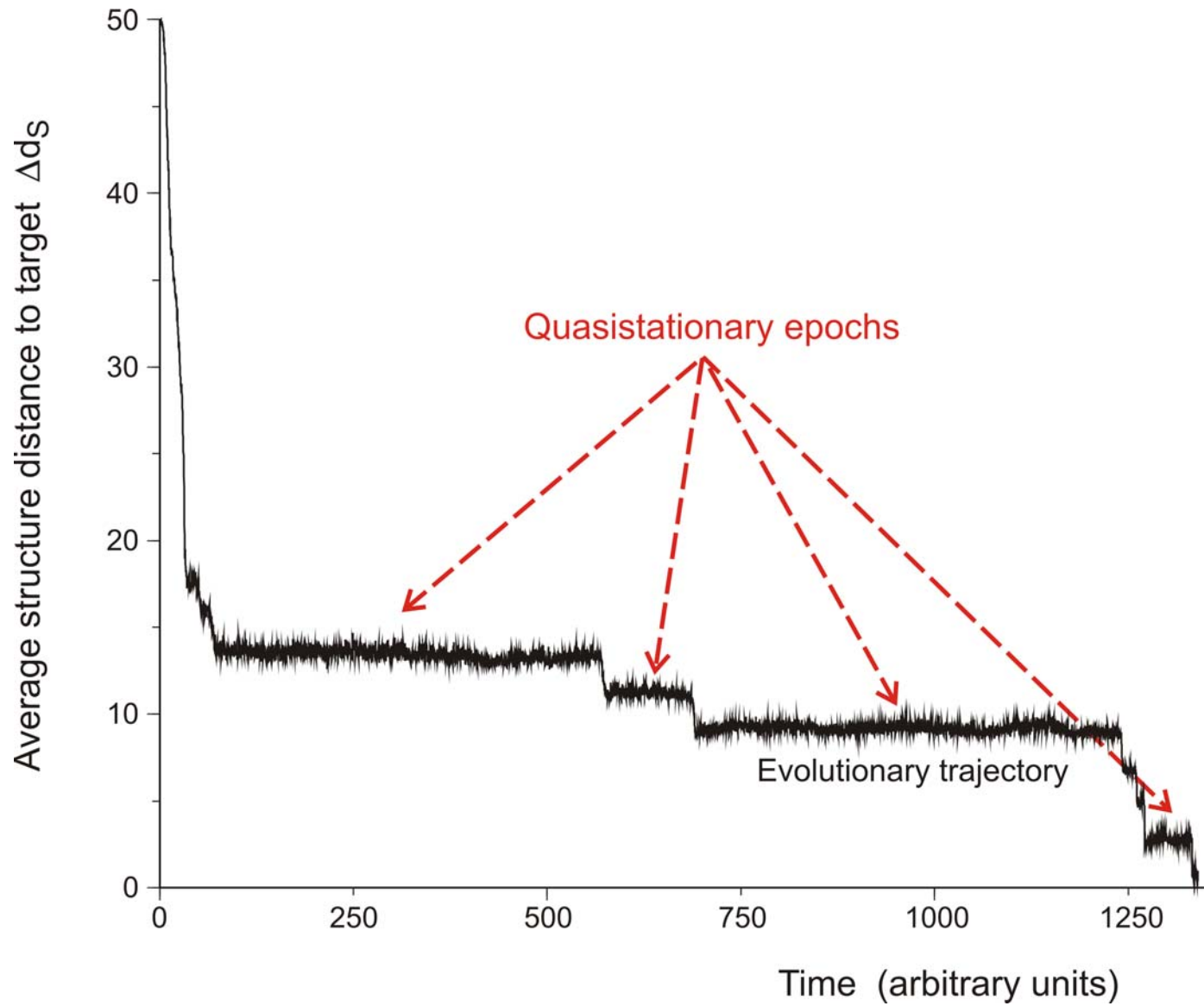
W. Fontana, P. Schuster,
Science **280** (1998), 1451-1455

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



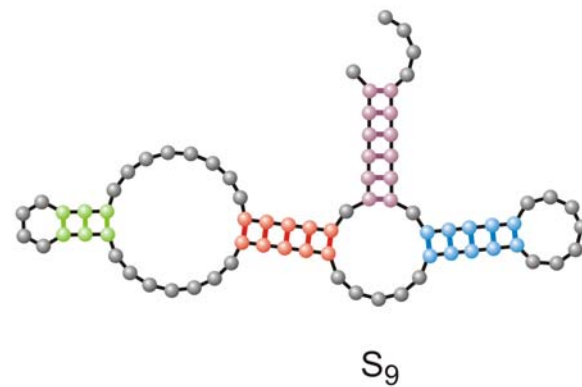
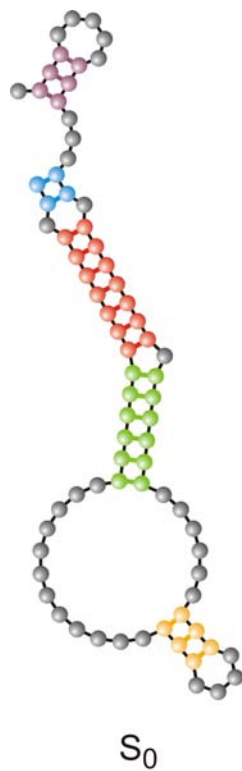
Structure of
randomly chosen
initial sequence

Phenylalanyl-tRNA as
target structure

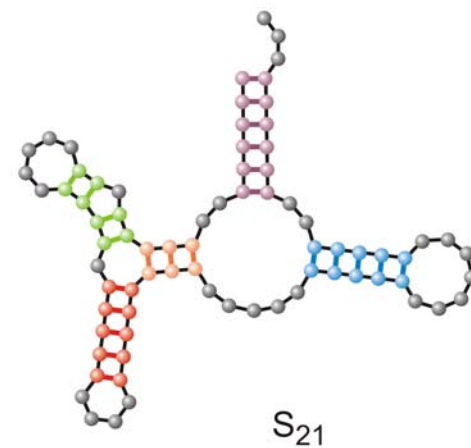
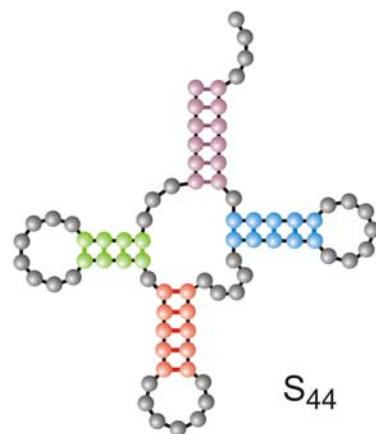


In silico optimization in the flow reactor: Evolutionary Trajectory

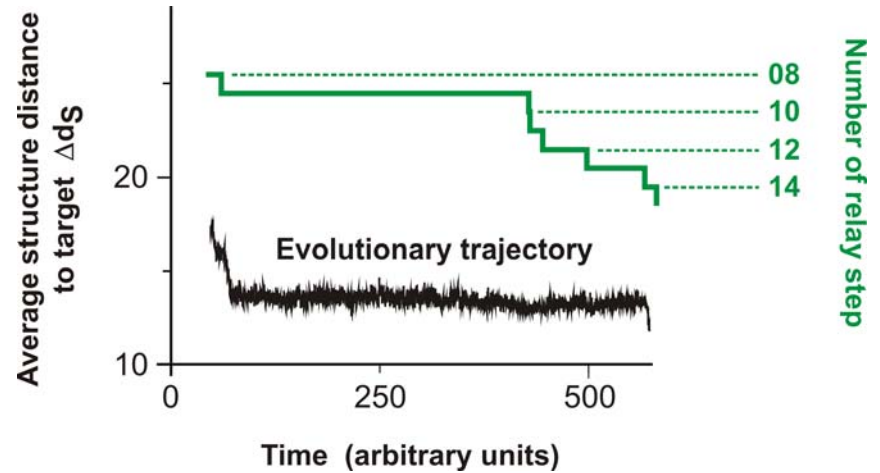
Randomly chosen
initial structure



Phenylalanyl-tRNA
as target structure



28 neutral point mutations during a long quasi-stationary epoch

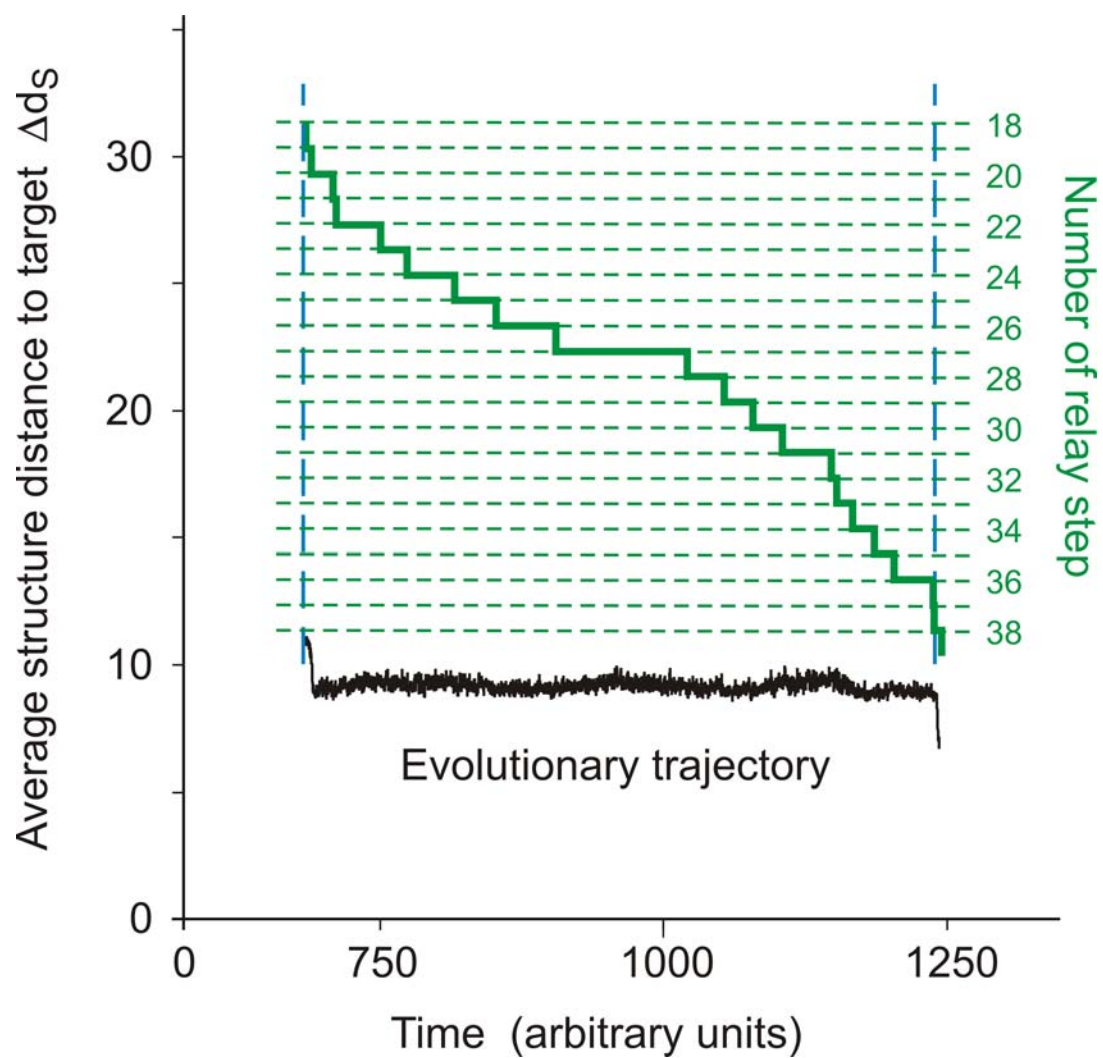


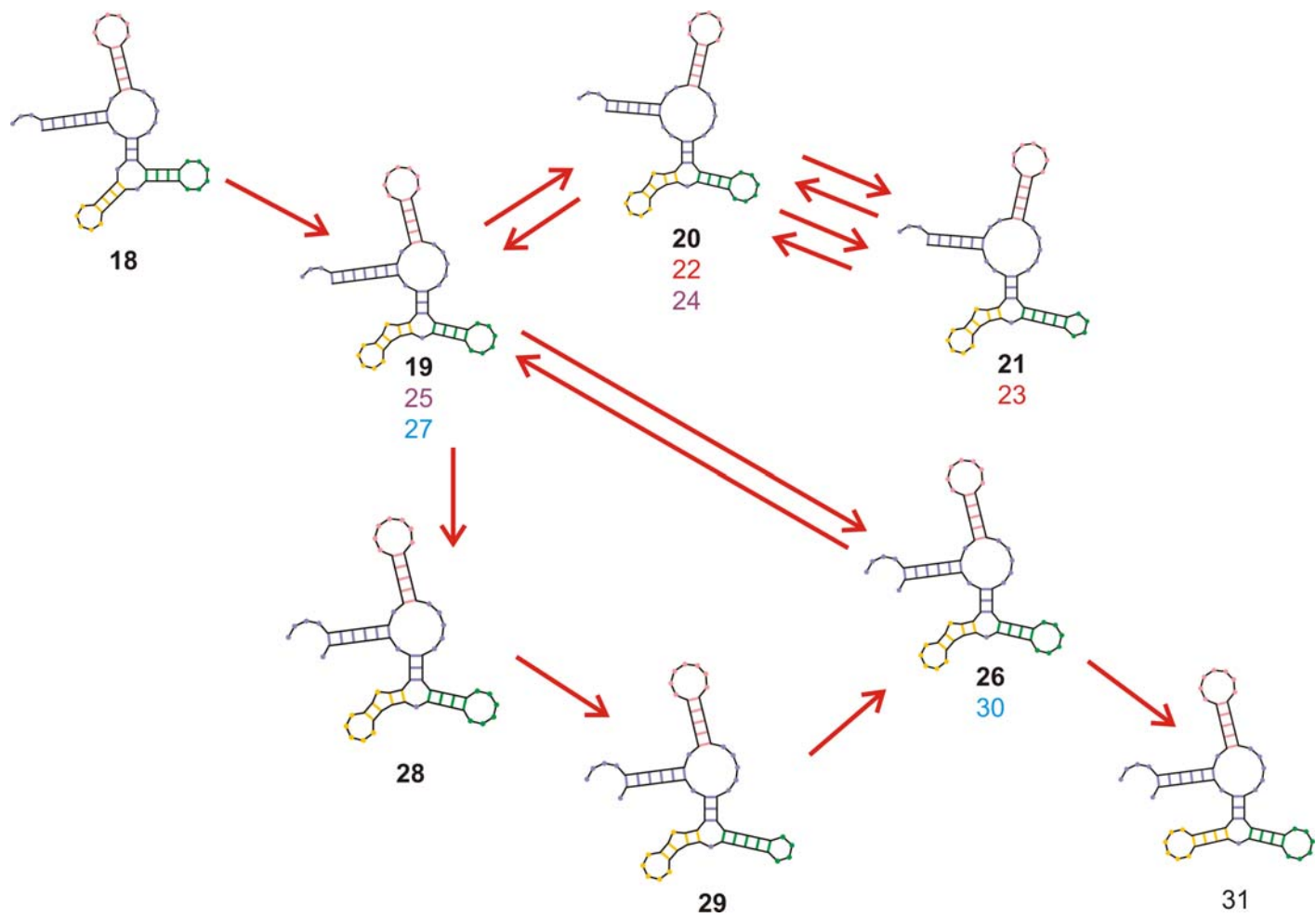
entry	GGUAUGGGCGUUGAAUAGG	GGUUAAACCAUCGG	CAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8	.((((((((((((((((.....((((.....)))).....)))))).....((((((((.....)))))))))).....		
exit	GGUAUGGGCGUUGAAUA	AUAGG	GGUUAAACCAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry	GGUAUGGGCGUUGAAUAAUAGG	GGUUAAACCAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUAACCAUACAGAA	
9	.((((((((((((((((.....((((.....)))).....)))))).....((((((((.....)))))))))).....		
exit	UGGAUGGACGUUGAAUAA	CAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG	
entry	UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG		
10	.((((((((((((((((.....((((.....)))).....)))))).....((((((((.....)))))))))).....		
exit	UGGAUGGACGUUGAAUAAACAAGGUAUCG	ACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG	

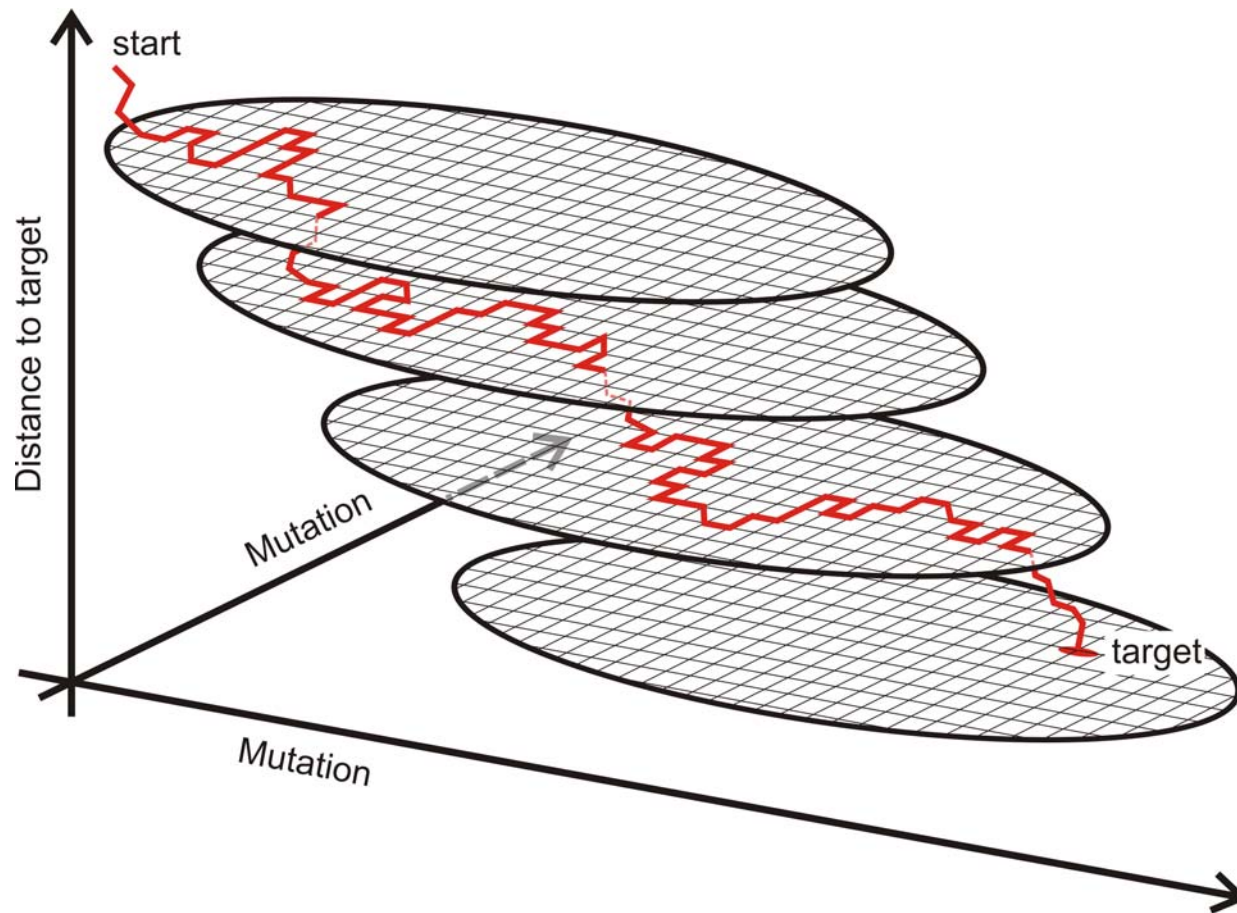
Transition inducing point mutations
change the molecular structure

Neutral point mutations leave the
molecular structure unchanged

Neutral genotype evolution during phenotypic stasis







A sketch of optimization on neutral networks

Coworkers

Peter Stadler, Bärbel M. Stadler, Universität Leipzig, GE

Paul E. Phillipson, University of Colorado at Boulder, CO

Heinz Engl, Philipp Kügler, James Lu, Stefan Müller, RICAM Linz, AT

Jord Nagel, Kees Pleij, Universiteit Leiden, NL

Walter Fontana, Harvard Medical School, MA

Martin Nowak, Harvard University, MA

Christian Reidys, Nankai University, Tien Tsin, China

Christian Forst, Los Alamos National Laboratory, NM

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

Ivo L.Hofacker, Christoph Flamm, Andreas Svrček-Seiler, Universität Wien, AT

Kurt Grünberger, Michael Kospach, Andreas Wernitznig, Stefanie Widder,
Stefan Wuchty, Jan Cupal, Stefan Bernhart, Lukas Endler, Ulrike Langhammer,
Rainer Machne, Ulrike Mückstein, Erich Bornberg-Bauer,
Universität Wien, AT



Universität Wien

Acknowledgement of support

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)
Projects No. 09942, 10578, 11065, 13093
13887, and 14898



Universität Wien

Wiener Wissenschafts-, Forschungs- und Technologiefonds (WWTF)
Project No. Mat05

Jubiläumsfonds der Österreichischen Nationalbank
Project No. Nat-7813

European Commission: Contracts No. 98-0189, 12835 (NEST)

Austrian Genome Research Program – GEN-AU: Bioinformatics
Network (BIN)

Österreichische Akademie der Wissenschaften

Siemens AG, Austria

Universität Wien and the Santa Fe Institute

Thank you for your attention!

Web-Page for further information:

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

