Complex behavior from simple molecular systems

Peter Schuster

Institut für Theoretische Chemie, Universität Wien, Austria
and
The Santa Fe Institute, Santa Fe, New Mexico, USA

Franqui Symposium 2008 in honor of Pierre Gaspard

Complex Systems: A fundamental Science Perspective

Brussels, 03.– 06.09.2008

Web-Page for further information:

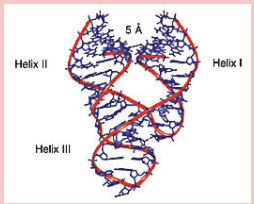
http://www.tbi.univie.ac.at/~pks

Review article:

Peter Schuster. Prediction of RNA molecules: From theory to models and real molecules. *Rep.Prog.Phys.* **69**:1419-1477, 2006.

RNA

RNA *as catalyst*



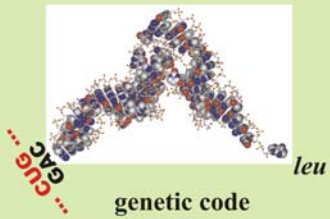
Ribozyme

RNA *as transmitter of genetic information*

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

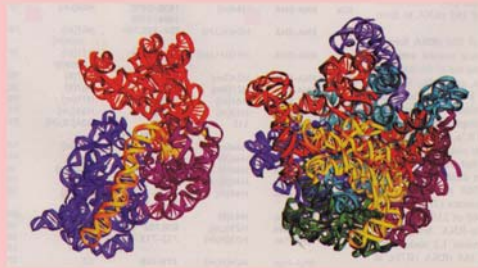
RNA as **working copy** of genetic information

RNA *as adapter molecule*



genetic code

RNA *is the catalytic subunit in supramolecular complexes*

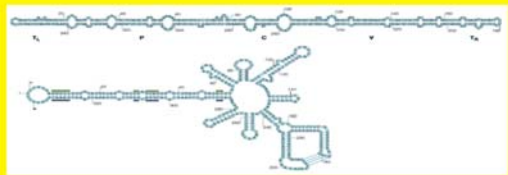


The **ribosome** is a **ribozyme !**

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

RNA *is modified by epigenetic control*

RNA editing, alternative splicing



Viroids

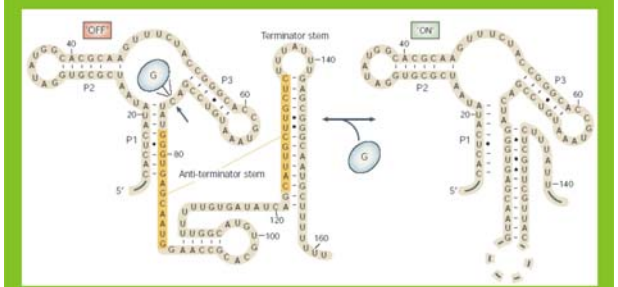
RNA *as carrier of genetic information*

RNA viruses and retroviruses

RNA evolution *in vitro*

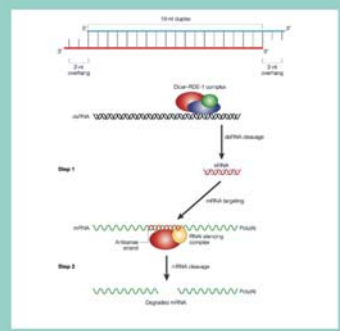
Evolutionary biotechnology
RNA aptamers, artificial ribozymes, allosteric ribozymes

***Allosteric control of transcribed* RNA**



Riboswitches controlling transcription and translation through **metabolites**

RNA *as regulator of gene expression*



Gene regulation by small interfering RNAs

RNA – The magic molecule

ENCODE stands for
ENCyclopedia Of DNA Elements.

ENCODE Project Consortium.
Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.
Nature **447**:799-816, 2007

118. Emanuelsson, O. *et al.* Assessing the performance of different high-density tiling microarray strategies for mapping transcribed regions of the human genome. *Genome Res.* advance online publication, doi: 10.1101/gr.5014606 (21 November 2006).
119. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. Science **296,** 916–919 (2002).
120. Bhinge, A. A., Kim, J., Euskirchen, G., Snyder, M. & Iyer, V. R. Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). *Genome Res.* **17,** 910–916 (2007).
121. Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature Methods* **2,** 105–111 (2005).
122. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17,** 877–885 (2006).
123. Rada-Iglesias, A. *et al.* Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Hum. Mol. Genet.* **14,** 3435–3447 (2005).
124. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436,** 876–880 (2005).
125. Halees, A. S. & Weng, Z. PromoSer: improvements to the algorithm, visualization and accessibility. *Nucleic Acids Res.* **32,** W191–W194 (2004).
126. Bajic, V. B. *et al.* Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol.* **7,** (Suppl 1; S3) 1–13 (2006).
127. Zheng, D. & Gerstein, M. B. A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol.* **7,** S13.1–S13.10 (2006).
128. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet* **1,** e78 (2005).
129. Turner, B. M. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nature Struct. Mol. Biol.* **12,** 110–112 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Leja for providing graphical expertise and support. Funding support is acknowledged from the following sources: National Institutes of Health, The European Union BioSapiens NoE, Affymetrix, Swiss National Science Foundation, the Spanish Ministerio de Educación y Ciencia, Spanish Ministry of Education and Science, CIBERESP, Genome Spain and Generalitat de Catalunya, Ministry of Education, Culture, Sports, Science and Technology of Japan, the NCCR Frontiers in Genetics, the Jérôme Lejeune Foundation, the Childcare Foundation, the Novartis Foundations, the Danish Research Council, the Swedish Research Council, the Knut and Alice Wallenberg Foundation, the Wellcome Trust, the Howard Hughes Medical Institute, the Bio-X Institute, the RIKEN Institute, the US Army, National Science Foundation, the Deutsche Forschungsgemeinschaft, the Austrian Gen-AU program, the BBSRC and The European Molecular Biology Laboratory. We thank the Barcelona SuperComputing Center and the NIH Biowulf cluster for computer facilities. The Consortium thanks the ENCODE Scientific Advisory Panel for their advice on the project: G. Weinstock, M. Cherry, G. Churchill, M. Eisen, S. Elgin, J. Lis, J. Rine, M. Vidal and P. Zamore.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. The list of individual authors is divided among the six main analysis groups and five organizational groups. Correspondence and requests for materials should be addressed to the co-chairs of the ENCODE analysis groups (listed in the Analysis Coordination group) E. Birney (birney@ebi.ac.uk); J. A. Stamatoyannopoulos (jstam@u.washington.edu); A. Dutta (ad8q@virginia.edu); R. Guigó (rguigo@imim.es); T. R. Gingeras (Tom_Gingeras@affymetrix.com); E. H. Margulies (elliott@nhgri.nih.gov); Z. Weng (zhiping@bu.edu); M. Snyder (michael.snyder@yale.edu); E. T. Dermitzakis (md4@sanger.ac.uk) or collectively (encode_chairs@ebi.ac.uk).

The ENCODE Project Consortium

Analysis Coordination Ewan Birney[1], John A. Stamatoyannopoulos[2], Anindya Dutta[3], Roderic Guigó[4,5], Thomas R. Gingeras[6], Elliott H. Margulies[7], Zhiping Weng[8,9], Michael Snyder[10,11] & Emmanouil T. Dermitzakis[12]

Chromatin and Replication John A. Stamatoyannopoulos[2], Robert E. Thurman[2,13], Michael S. Kuehn[2,13], Christopher M. Taylor[3], Shane Neph[2], Christoph M. Koch[12], Saurabh Asthana[14], Ankit Malhotra[3], Ivan Adzhubei[14], Jason A. Greenbaum[15], Robert M. Andrews[12], Paul Flicek[1], Patrick J. Boyle[3], Hua Cao[13], Nigel P. Carter[12], Gayle K. Clelland[12], Sean Davis[16], Nathan Day[2], Pawandeep Dhami[12], Shane C. Dillon[12], Michael O. Dorschner[2], Heike Fiegler[12], Paul G. Giresi[17], Jeff Goldy[2], Michael Hawrylycz[18], Andrew Haydock[2], Richard Humbert[2], Keith D. James[12], Brett E. Johnson[13], Ericka M. Johnson[13], Tristan T. Frum[13], Elizabeth R. Rosenzweig[13], Neerja Karnani[3], Kirsten Lee[2], Gregory C. Lefebvre[12], Patrick A. Navas[13], Fidencio Neri[2], Stephen C. J. Parker[15], Peter J. Sabo[2], Richard Sandstrom[2], Anthony Shafer[2], David Vetrie[12], Molly Weaver[2], Sarah Wilcox[12], Man Yu[13], Francis S. Collins[7], Job Dekker[19], Jason D. Lieb[17], Thomas D. Tullius[15], Gregory E. Crawford[20], Shamil Sunyaev[14], William S. Noble[2], Ian Dunham[12] & Anindya Dutta[3]

Genes and Transcripts Roderic Guigó[4, 5], France Denoeud[5], Alexandre Reymond[21,22], Philipp Kapranov[6], Joel Rozowsky[11], Deyou Zheng[11], Robert Castelo[5], Adam Frankish[12], Jennifer Harrow[12], Srinka Ghosh[6], Albin Sandelin[23], Ivo L. Hofacker[24], Robert Baertsch[25,26], Damian Keefe[1], Paul Flicek[1], Sujit Dike[6], Jill Cheng[6], Heather A. Hirsch[27], Edward A. Sekinger[27], Julien Lagarde[5], Josep F. Abril[5,28], Atif Shahab[29], Christoph Flamm[24,30], Claudia Fried[30], Jörg Hackermüller[32], Jana Hertel[30], Manja Lindemeyer[30], Kristin Missal[30,31], Andrea Tanzer[24,30], Stefan Washietl[24], Jan Korbel[11], Olof Emanuelsson[11], Jakob S. Pedersen[26], Nancy Holroyd[12], Ruth Taylor[12], David Swarbreck[12], Nicholas Matthews[12], Mark C. Dickson[33], Daryl J. Thomas[25,26], Matthew T. Weirauch[25], James Gilbert[12], Jorg Drenkow[6], Ian Bell[6], XiaoDong Zhao[34], K.G. Srinivasan[34], Wing-Kin Sung[34], Hong Sain Ooi[34], Kuo Ping Chiu[34], Sylvain Foissac[4], Tyler Alioto[4], Michael Brent[35], Lior Pachter[36], Michael L. Tress[37], Alfonso Valencia[37], Siew Woh Choo[34], Chiou Yu Choo[34], Catherine Ucla[22], Caroline Manzano[22], Carine Wyss[22], Evelyn Cheung[6], Taane G. Clark[38], James B. Brown[39], Madhavan Ganesh[6], Sandeep Patel[6], Hari Tammana[6], Jacqueline Chrast[21], Charlotte N. Henrichsen[21], Chikatoshi Kai[23], Jun Kawai[23,40], Ugrappa Nagalakshmi[10], Jiaqian Wu[10], Zheng Lian[41], Jin Lian[41], Peter Newburger[42], Xueqing Zhang[42], Peter Bickel[43], John S. Mattick[44], Piero Carninci[40],Yoshihide Hayashizaki[23,40], Sherman Weissman[41], Emmanouil T. Dermitzakis[12], Elliott H. Margulies[7], Tim Hubbard[12], Richard M. Myers[33], Jane Rogers[12], Peter F. Stadler[24,30,45], Todd M. Lowe[25], Chia-Lin Wei[34], Yijun Ruan[34], Michael Snyder[10,11], Ewan Birney[1], Kevin Struhl[27], Mark Gerstein[11,46,47], Stylianos E. Antonarakis[22] & Thomas R. Gingeras[6]

Integrated Analysis and Manuscript Preparation James B. Brown[39], Paul Flicek[1], Yutao Fu[8], Damian Keefe[1], Ewan Birney[1], France Denoeud[5], Mark Gerstein[11,46,47], Eric D. Green[7,48], Philipp Kapranov[6], Ulaş Karaöz[8], Richard M. Myers[33], William S. Noble[2], Alexandre Reymond[21,22], Joel Rozowsky[11], Kevin Struhl[27], Adam Siepel[25, 26]†, John A. Stamatoyannopoulos[2], Christopher M. Taylor[3], James Taylor[49,50], Robert E. Thurman[2,13], Thomas D. Tullius[15], Stefan Washietl[24] & Deyou Zheng[11]

Management Group Laura A. Liefer[51], Kris A. Wetterstrand[51], Peter J. Good[51], Elise A. Feingold[51], Mark S. Guyer[51] & Francis S. Collins[52]

Multi-species Sequence Analysis Elliott H. Margulies[7], Gregory M. Cooper[33]†, George Asimenos[53], Daryl J. Thomas[25,26], Colin N. Dewey[54], Adam Siepel[25,26]†, Ewan Birney[1], Damian Keefe[1], Minmei Hou[49,50], James Taylor[49,50], Sergey Nikolaev[22], Juan I. Montoya-Burgos[55], Ari Löytynoja[1], Simon Whelan[1]†, Fabio Pardi[1], Tim Massingham[1], James B. Brown[39], Haiyan Huang[43], Nancy R. Zhang[43,56], Peter Bickel[43], Ian Holmes[57], James C. Mullikin[7,48], Abel Ureta-Vidal[1], Benedict Paten[1], Michael Seringhaus[11], Deanna Church[58], Kate Rosenbloom[26], W. James Kent[25,26], Eric A. Stone[33], NISC Comparative Sequencing Program*, Baylor College of Medicine Human Genome Sequencing Center*, Washington University Genome Sequencing Center*, Broad Institute*, Children's Hospital Oakland Research Institute*, Mark Gerstein[11,46,47], Stylianos E. Antonarakis[22], Serafim Batzoglou[59], Nick Goldman[1], Ross C. Hardison[60,59], David Haussler[25,26,60], Webb Miller[49,50,61], Lior Pachter[36], Eric D. Green[7,48] & Arend Sidow[33,62]

Transcriptional Regulatory Elements Zhiping Weng[8,9], Nathan D. Trinklein[33]†, Yutao Fu[8], Zhengdong D. Zhang[11], Ulaş Karaöz[8], Leah Barrera[63], Rhona Stuart[64], Deyou Zheng[11], Srinka Ghosh[6], Paul Flicek[1], David C. King[60, 59], James Taylor[49, 50], Adam Ameur[69], Stefan Enroth[69], Mark C. Bieda[70], Christoph M. Koch[12], Heather A. Hirsch[27], Chia-Lin Wei[34], Jill Cheng[6], Jonghwan Kim[71], Akshay A. Bhinge[71], Paul G. Giresi[17], Nan Jiang[72], Jun Liu[34], Fei Yao[34], Wing-Kin Sung[34], Kuo Ping Chiu[34], Vinsensius B. Vega[34], Charlie W.H. Lee[34], Patrick Ng[34], Atif Shahab[29], Edward A. Sekinger[27], Annie Yang[27], Zarmik Moqtaderi[27], Zhou Zhu[27], Xiaoqin Xu[20], Sharon Squazzo[20], Matthew J. Oberley[73], David Inman[73], Michael A. Singer[72], Todd A. Richmond[72], Kyle J. Munn[72,74], Alvaro Rada-Iglesias[74], Ola Wallerman[74], Jan Komorowski[69], Gayle K. Clelland[12], Sarah Wilcox[12], Shane C. Dillon[12], Robert M. Andrews[12], Joanna C. Fowler[12], Phillippe Couttet[12], Keith D. James[12], Gregory C. Lefebvre[12], Alexander W. Bruce[12], Oliver M. Dovey[12], Peter D. Ellis[12], Pawandeep Dhami[12], Cordelia F. Langford[12], Nigel P. Carter[12], David Vetrie[12], Philipp Kapranov[6], David A. Nix[6], Ian Bell[6], Sandeep Patel[6], Joel Rozowsky[11], Ghia Euskirchen[10], Stephen Hartman[10], Jin Lian[41], Jiaqian Wu[10], Alexander E. Urban[10], Peter Kraus[10], Sara Van Calcar[68], Nate Heintzman[68], Tae Hoon Kim[68], Kun Wang[68], Chunxu Qu[68], Gary Hon[68], Rosa Luna[75], Christopher K. Glass[75], M. Geoff Rosenfeld[75], Shelley Force Aldred[33], Sara J. Cooper[33], Anason Halees[8], Jane M. Lin[9], Hennady P. Shulha[9], Xiaoling Zhang[8], Mousheng Xu[8], Jaafar N. S. Haidar[9], Yong Yu[9], Ewan Birney*,[1], Sherman Weissman[41], Yijun Ruan[34], Jason D. Lieb[17], Vishwanath R. Iyer[71], Roland D. Green[72], Thomas R. Gingeras[6], Claes Wadelius[74], Ian Dunham[12], Kevin Struhl[27], Ross C. Hardison[60,59], Mark Gerstein[11,46,47], Peggy J. Farnham[70], Richard M. Myers[33] Bing Ren[68] & Michael Snyder[10,11]

UCSC Genome Browser Daryl J. Thomas[25,26], Kate Rosenbloom[26], Rachel A. Harte[26], Angie S. Hinrichs[26], Heather Trumbower[26], Hiram Clawson[26], Jennifer Hillman-Jackson[26], Ann S. Zweig[26], Kayla Smith[26], Archana Thakkapallayil[26], Galt Barber[26], Robert M. Kuhn[26], Donna Karolchik[26], David Haussler[25,26,60] & W. James Kent[25,26]

Variation Emmanouil T. Dermitzakis[12], Lluis Armengol[76], Christine P. Bird[12], Taane G. Clark[38], Gregory M. Cooper[33]†, Paul I. W. de Bakker[77], Andrew D. Kern[20], Nuria Lopez-Bigas[5], Joel D. Martin[60,59], Barbara E. Stranger[12], Daryl J. Thomas[25,26], Abigail Woodroffe[78], Serafim Batzoglou[53], Eugene Davydov[53], Antigone Dimas[12], Eduardo Eyras[5], Ingileif B. Hallgrímsdóttir[79], Ross C. Hardison[60,59], Julian Huppert[12], Arend Sidow[33,62], James Taylor[49,50], Heather Trumbower[26], Michael C. Zody[77], Roderic Guigó[4,5], James C. Mullikin[7], Gonçalo R. Abecasis[78], Xavier Estivill[76,80] & Ewan Birney[1].

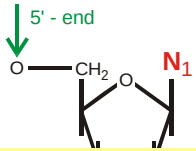
***NISC Comparative Sequencing Program** Gerard G. Bouffard[7,48], Xiaobin Guan[48], Nancy F. Hansen[48], Jacquelyn R. Idol[7], Valerie V.B. Maduro[7], Baishali Maskeri[48], Jennifer C. McDowell[48], Morgan Park[48], Pamela J. Thomas[48], Alice C. Young[48] & Robert W. Blakesley[7,48] **Baylor College of Medicine, Human Genome Sequencing Center** Donna M. Muzny[63], Erica Sodergren[63], David A. Wheeler[63], Kim C. Worley[63], Huaiyang Jiang[63], George M. Weinstock[63] & Richard A. Gibbs[63]; **Washington University Genome Sequencing Center** Tina Graves[64], Robert Fulton[64], Elaine R. Mardis[64] & Richard K. Wilson[64] **Broad Institute** Michele Clamp[65], James Cuff[65], Sante Gnerre[65], David B. Jaffe[65], Jean L. Chang[65], Kerstin Lindblad-Toh[65] & Eric S. Lander[65,66] **Children's Hospital Oakland Research Institute** Maxim Koriabine[67], Mikhail Nefedov[67], Kazutoyo Osoegawa[67], Yuko Yoshinaga[67], Baoli Zhu[67] & Pieter J. de Jong[67]

Affiliations for participants: [1]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK. [2]Department of Genome Sciences, 1705 NE Pacific Street, Box 357730, University of Washington, Seattle, Washington 98195, USA. [3]Department of Biochemistry and Molecular Genetics, Jordan 1240, Box 800733, 1300 Jefferson Park Ave, University of Virginia School of Medicine, Charlottesville, Virginia 22908, USA. [4]Genomic Bioinformatics Program, Center for Genomic Regulation, [5]Research Group in Biomedical Informatics, Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra, c/o Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. [6]Affymetrix, Inc., Santa Clara, California 95051, USA. [7]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [8]Bioinformatics Program, Boston University, 24 Cummington St., Boston, Massachusetts 02215, USA. [9]Biomedical Engineering Department, Boston University, 44 Cummington St., Boston, Massachusetts 02215, USA. [10]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. [11]Department of Molecular Biophysics and Biochemistry, Yale University, PO Box 208114, New Haven, Connecticut 06520, USA. [12]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [13]Division of Medical Genetics, 1705 NE Pacific Street, Box 357720, University of Washington, Seattle, Washington 98195, USA. [14]Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [15]Department of Chemistry and Program in Bioinformatics, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA. [16]Genetics Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [17]Department of Biology and Carolina Center for Genome Sciences, CB# 3280, 202 Fordham Hall, The University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [18]Allen Institute for Brain Sciences, 551 North 34th Street, Seattle, Washington 98103, USA. [19]Program in Gene Function and Expression and Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA. [20]Institute for Genome Sciences & Policy and Department of Pediatrics, 101 Science Drive, Duke University, Durham, North Carolina 27708, USA. [21]Center for Integrative Genomics, University of Lausanne, Genopode building, 1015 Lausanne, Switzerland. [22]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva, Switzerland. [23]Genome Exploration Research Group, RIKEN Genomic Sciences Center (GSC), RIKEN Yokohama Institute, 1-7-22, Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan. [24]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria. [25]Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA. [26]Center for Biomolecular Science and Engineering, Engineering 2, Suite 501, Mail Stop CBSE/ITI, University of California, Santa Cruz, California 95064, USA. [27]Department of Biological Chemistry & Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA. [28]Department of Genetics, Facultat de Biologia, Universitat de Barcelona, Av Diagonal, 645, 08028, Barcelona, Catalonia, Spain. [29]Bioinformatics Institute, 30 Biopolis Street, #07-01 Matrix, Singapore, 138671, Singapore. [30]Bioinformatics Group, Department of Computer Science, [31]Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. [32]Fraunhofer Institut für Zelltherapie und Immunologie - IZI, Deutscher Platz 5e, D-04103 Leipzig, Germany. [33]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. [34]Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Singapore. [35]Laboratory for Computational Genomics, Washington University, Campus Box 1045, Saint Louis, Missouri 63130, USA. [36]Department of Mathematics and Computer Science, University of California, Berkeley, California 94720, USA. [37]Spanish National Cancer Research Centre, CNIO, Madrid, E-28029, Spain. [38]Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, UK. [39]Department of Applied Science & Technology, University of California, Berkeley, California 94720, USA. [40]Genome Science Laboratory, Discovery and Research Institute, RIKEN Wako Institute, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan. [41]Department of Genetics, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA. [42]Department of Pediatrics, University of Massachusetts Medical School, 55 Lake Avenue, North Worcester, Massachusetts 01605, USA. [43]Department of Statistics, University of California, Berkeley, California 94720, USA. [44]Institute for Molecular Bioscience, University of Queensland, St. Lucia, QLD 4072, Australia. [45]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. [46]Department of Computer Science, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA. [47]Program in Computational Biology & Bioinformatics, Yale University, PO Box 208114, New Haven, Connecticut 06520-8114, USA. [48]NIH Intramural Sequencing Center, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [49]Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [50]Center for Comparative Genomics and Bioinformatics, Huck Institutes for Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [51]Division of Extramural Research, National Human Genome Research Institute, National Institute of Health, 5635 Fishers Lane, Suite 4076, Bethesda, Maryland 20892-9305, USA. [52]Office of the Director, National Human Genome Research Institute, National Institute of Health, 31 Center Drive, Suite 4B09, Bethesda, Maryland 20892-2152, USA. [53]Department of Computer Science, Stanford University, Stanford, California 94305, USA. [54]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 6720 MSC, 1300 University Ave, Madison, Wisconsin 53706, USA. [55]Department of Zoology and Animal Biology, Faculty of Sciences, University of Geneva, 1205 Geneva, Switzerland. [56]Department of Statistics, Stanford University, Stanford, California 94305, USA. [57]Department of Bioengineering, University of California, Berkeley, California 94720-1762, USA. [58]National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA. [59]Department of Biochemistry and Molecular Biology, Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [60]Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA. [61]Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. [62]Department of Pathology, Stanford University School of Medicine, Stanford, California 94305, USA. [63]Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA. [64]Genome Sequencing Center, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, Saint Louis, Missouri 63108, USA. [65]Broad Institute of Harvard University and Massachusetts Institute of Technology, 320 Charles Street, Cambridge, Massachusetts 02141, USA. [66]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. [67]Children's Hospital Oakland Research Institute, BACPAC Resources, 747 52nd Street, Oakland, California 94609, USA. [68]Ludwig Institute for Cancer Research, 9500 Gilman Drive, La Jolla, California 92093-0653, USA. [69]The Linnaeus Centre for Bioinformatics, Uppsala University, BMC, Box 598, SE-75124 Uppsala, Sweden. [70]Department of Pharmacology and the Genome Center, University of California, Davis, California 95616, USA. [71]Institute for Cellular & Molecular Biology, The University of Texas at Austin, 1 University Station A4800, Austin, Texas 78712, USA. [72]NimbleGen Systems, Inc., 1 Science Court, Madison, Wisconsin 53711, USA. [73]University of Wisconsin Medical School, Madison, Wisconsin 53706, USA. [74]Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, SE-75185 Uppsala, Sweden. [75]University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, California 92093, USA. [76]Genes and Disease Program, Center for Genomic Regulation, c/o Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. [77]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [78]Center for Statistical Genetics, Department of Biostatistics, SPH II, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, USA. [79]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK. [80]Universitat Pompeu Fabra, c/o Dr. Aiguader 88, Barcelona Biomedical Research Park Building, 08003 Barcelona, Catalonia, Spain. †Present addresses: Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA (G.M.C.); Department of Biological Statistics & Computational Biology, Cornell University, Ithaca, New York 14853, USA (A.S.); Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT, UK (S.W.); SwitchGear Genomics, 1455 Adams Drive, Suite 2015, Menlo Park, California 94025, USA (N.D.T.; S.F.A.).

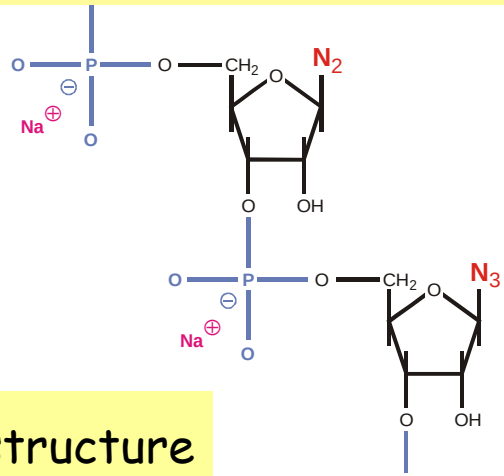
411 Autoren aus 80 Institutionen

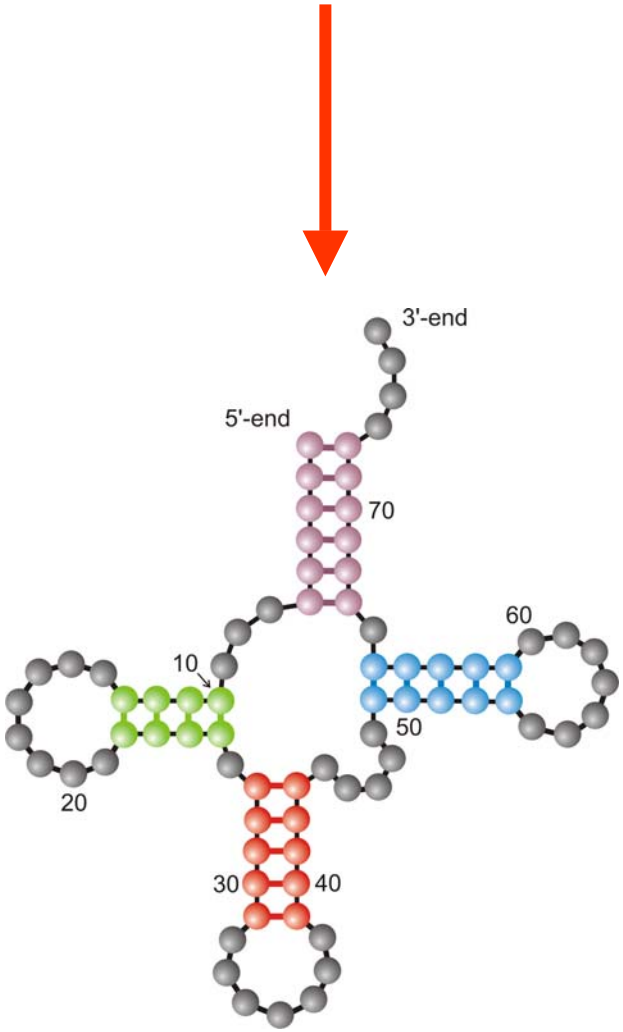
1. Minimum free energy structures of RNA
2. Suboptimal structures of RNA
3. Kinetic folding and RNA switches
4. Chemistry of Darwinian evolution
5. Consequences of neutrality
6. Evolutionary optimization of RNA structure

1. **Minimum free energy structures of RNA**
2. Suboptimal structures of RNA
3. Kinetic folding and RNA switches
4. Chemistry of Darwinian evolution
5. Consequences of neutrality
6. Evolutionary optimization of RNA structure

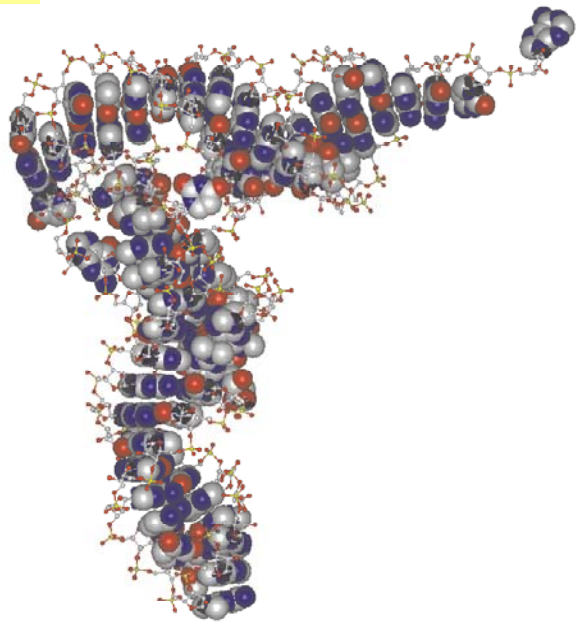


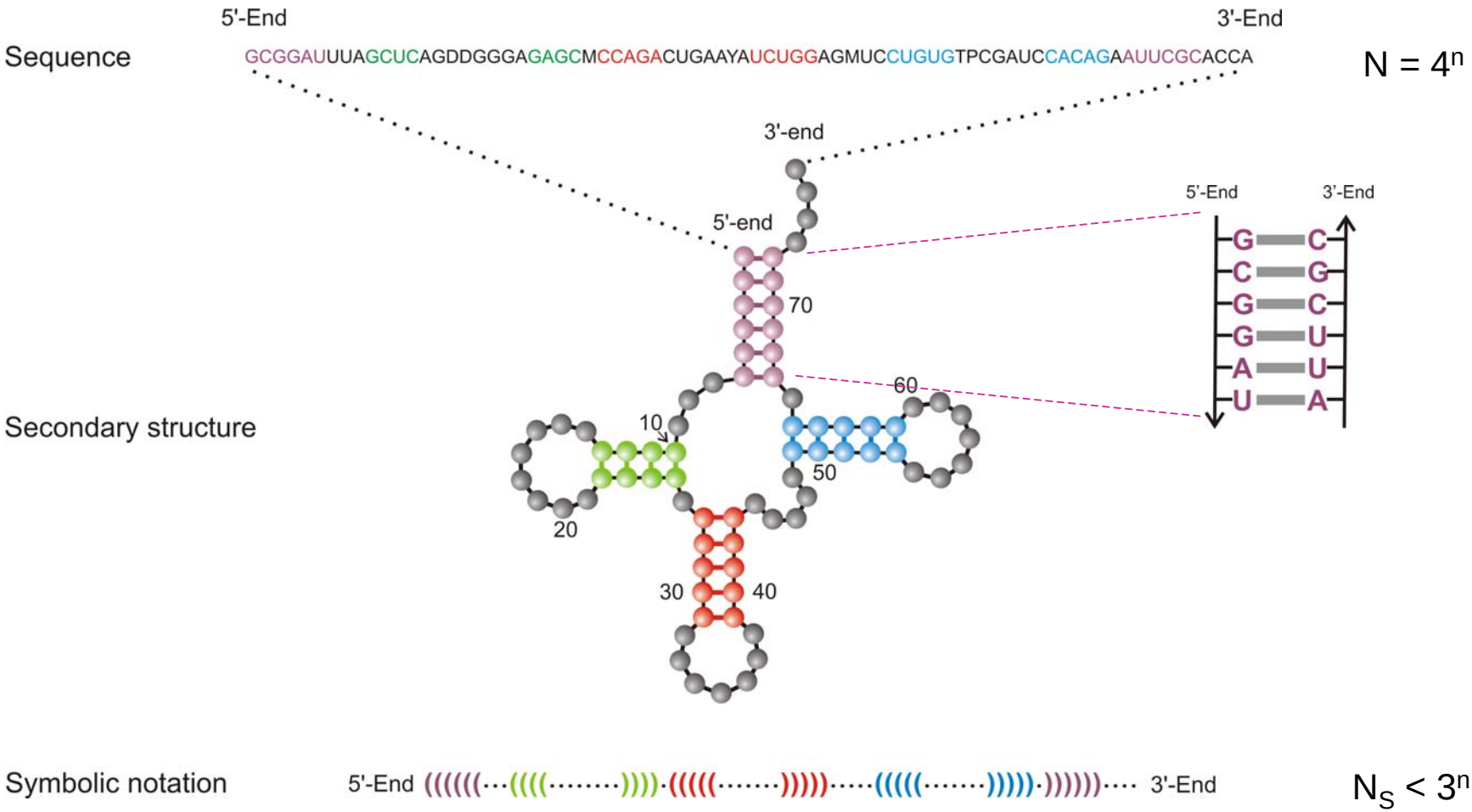
5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





Definition of RNA structure





Criterion: Minimum free energy (mfe)

Rules: _ (_) _ ∈ {**AU**,**CG**,**GC**,**GU**,**UA**,**UG**}

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

The paradigm of structural biology

What is neutrality ?

Selective neutrality =
= several genotypes having the same fitness.

Structural neutrality =
= several sequences forming molecules with the same structure.

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

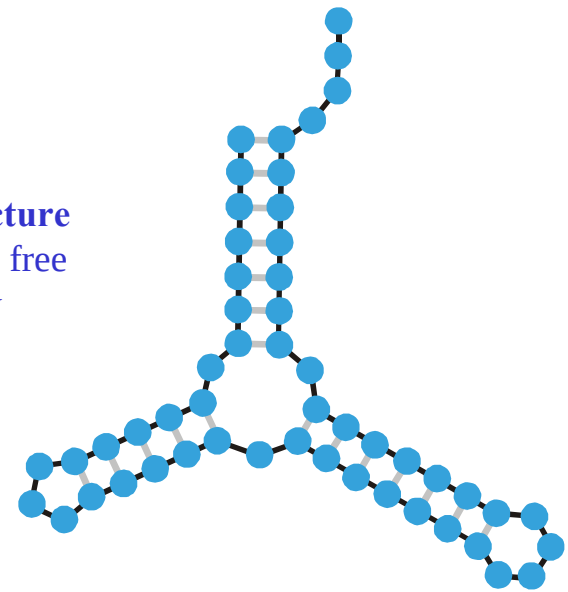
RNA folding:

Structural biology, spectroscopy of biomolecules, understanding **molecular function**

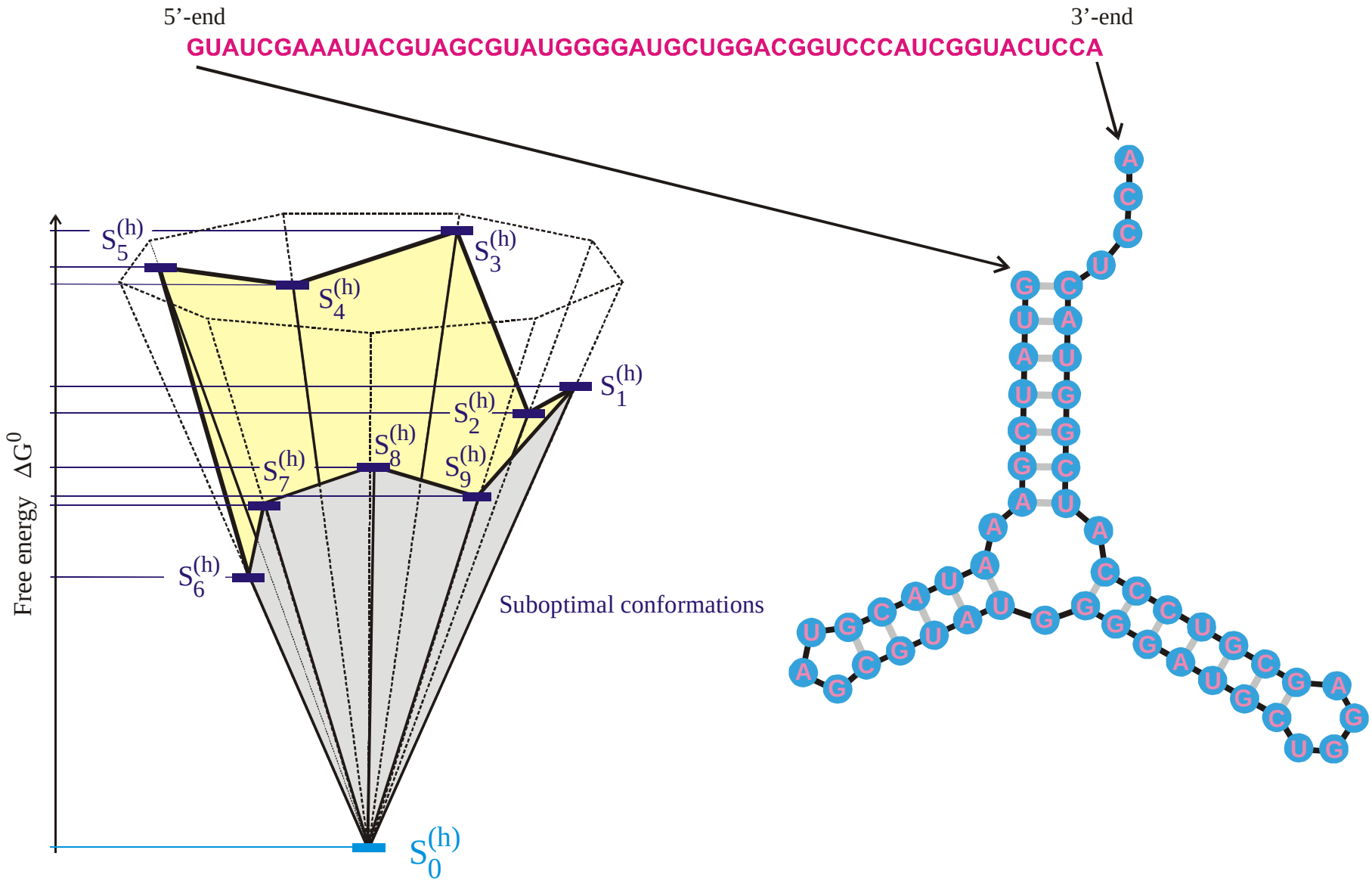
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

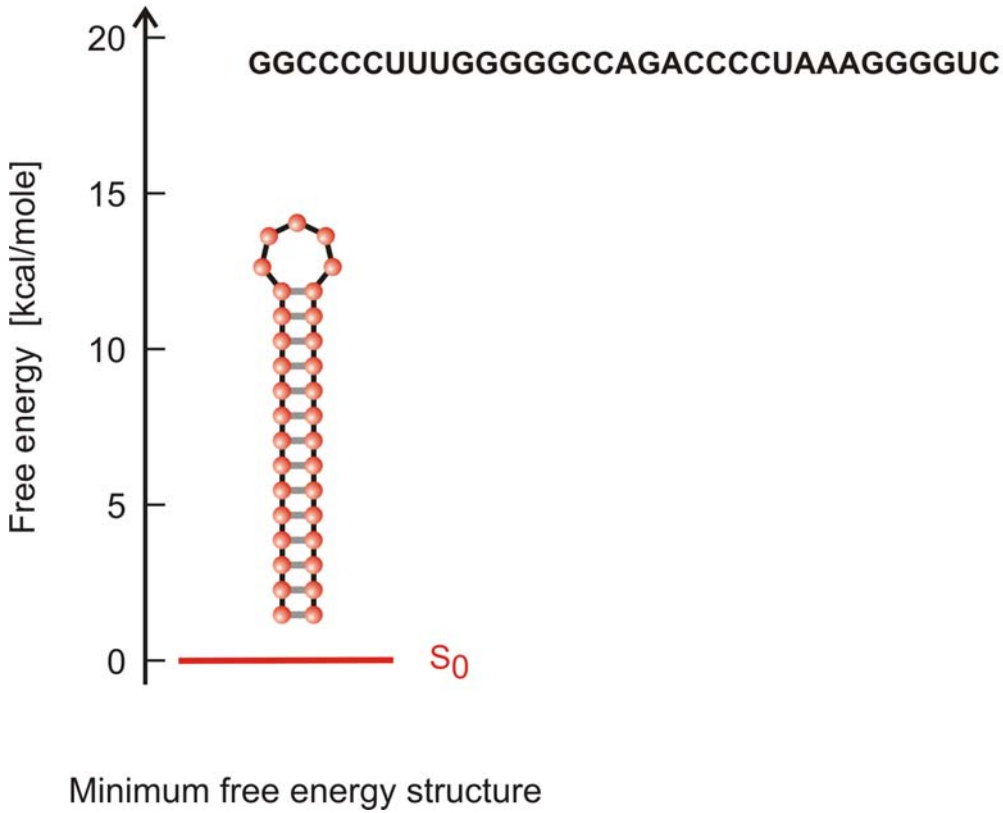
RNA structure of minimal free energy



Sequence, structure, and design



The minimum free energy structures on a discrete space of conformations



Extension of the notion of structure

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

Structural biology, spectroscopy of biomolecules, understanding **molecular function**

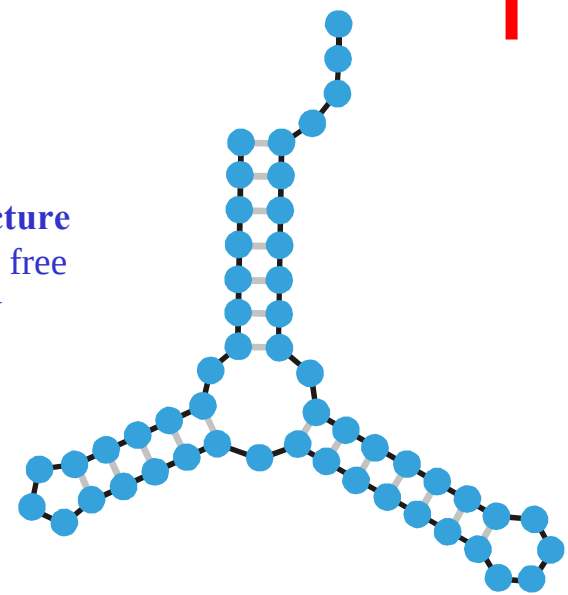
Iterative determination of a sequence for the given secondary structure

Inverse Folding Algorithm

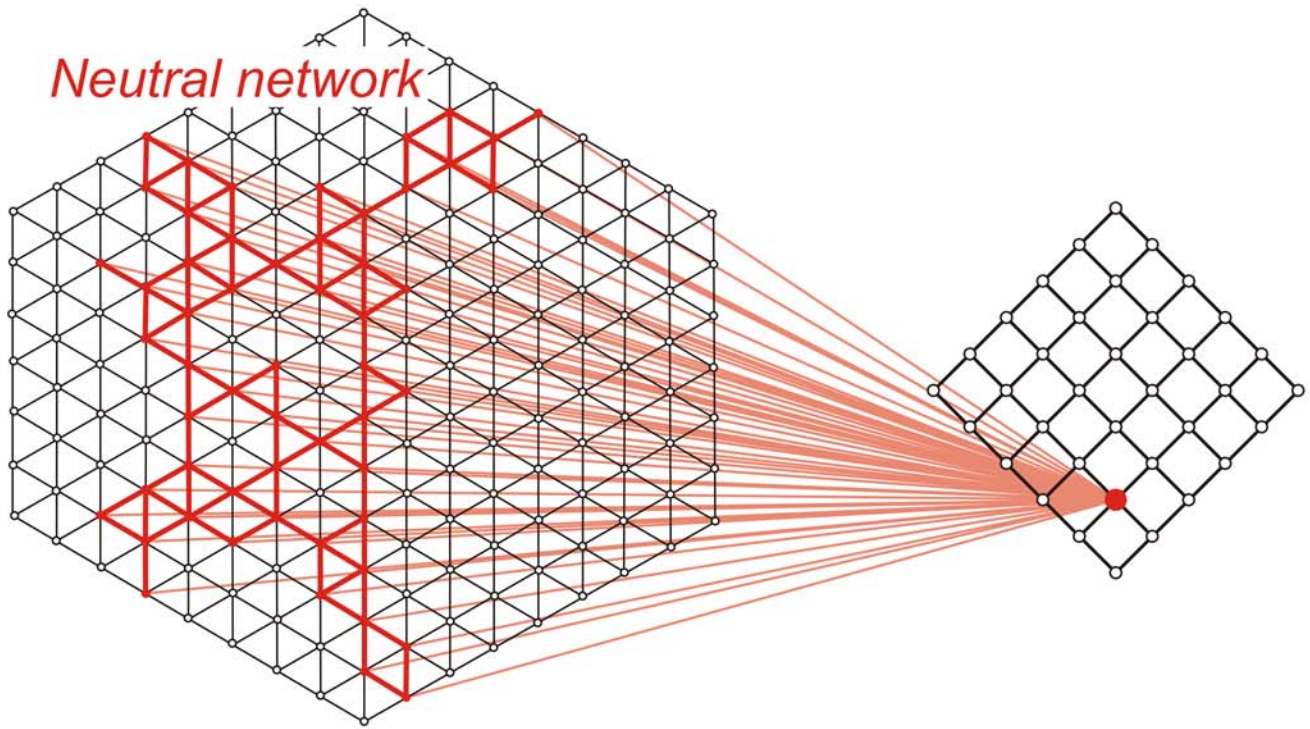
Inverse folding of RNA:

Biotechnology, **design of biomolecules** with predefined structures and functions

RNA structure of minimal free energy



Sequence, structure, and design



Sequence space

Structure space

many sequences ⇒ one structure

Space of genotypes: $Q = \{X_1, X_2, X_3, X_4, \ldots, X_N\}$; Hamming metric

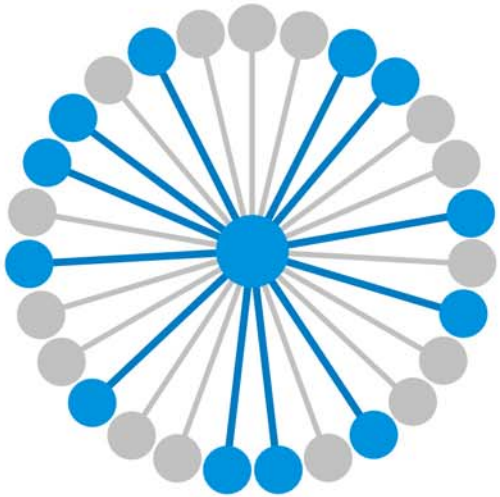
Space of phenotypes: $S = \{S_1, S_2, S_3, S_4, \ldots, S_M\}$; metric (not required)

$$N >> M$$

$$\psi(X_j) = S_k$$

$$\mathbf{G}_k = \psi^{-1}(S_k) \doteq \{ X_j \mid \psi(X_j) = S_k \}$$

A mapping ψ and its inversion



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

$$\bar{\lambda}_k = \frac{\sum_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

$\lambda_j = 12 / 27 = 0.444$

Alphabet size κ :

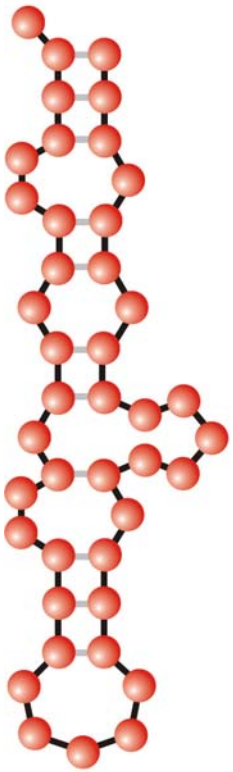
κ	λ_{cr}	
2	0.5	**AU,GC,DU**
3	0.423	**AUG , UGC**
4	0.370	**AUGC**

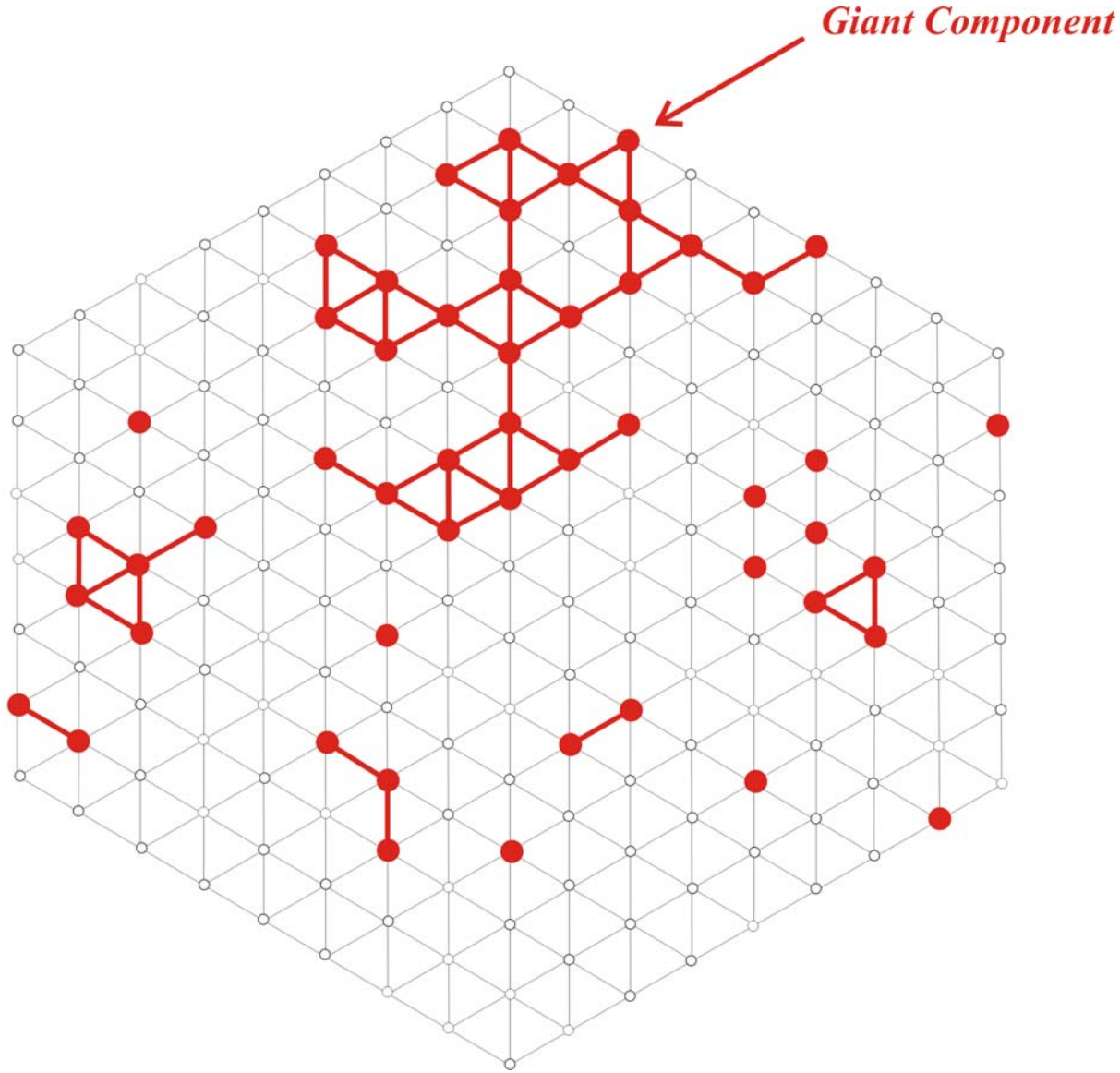
$\bar{\lambda}_k > \lambda_{cr}$ .... network G_k is connected

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

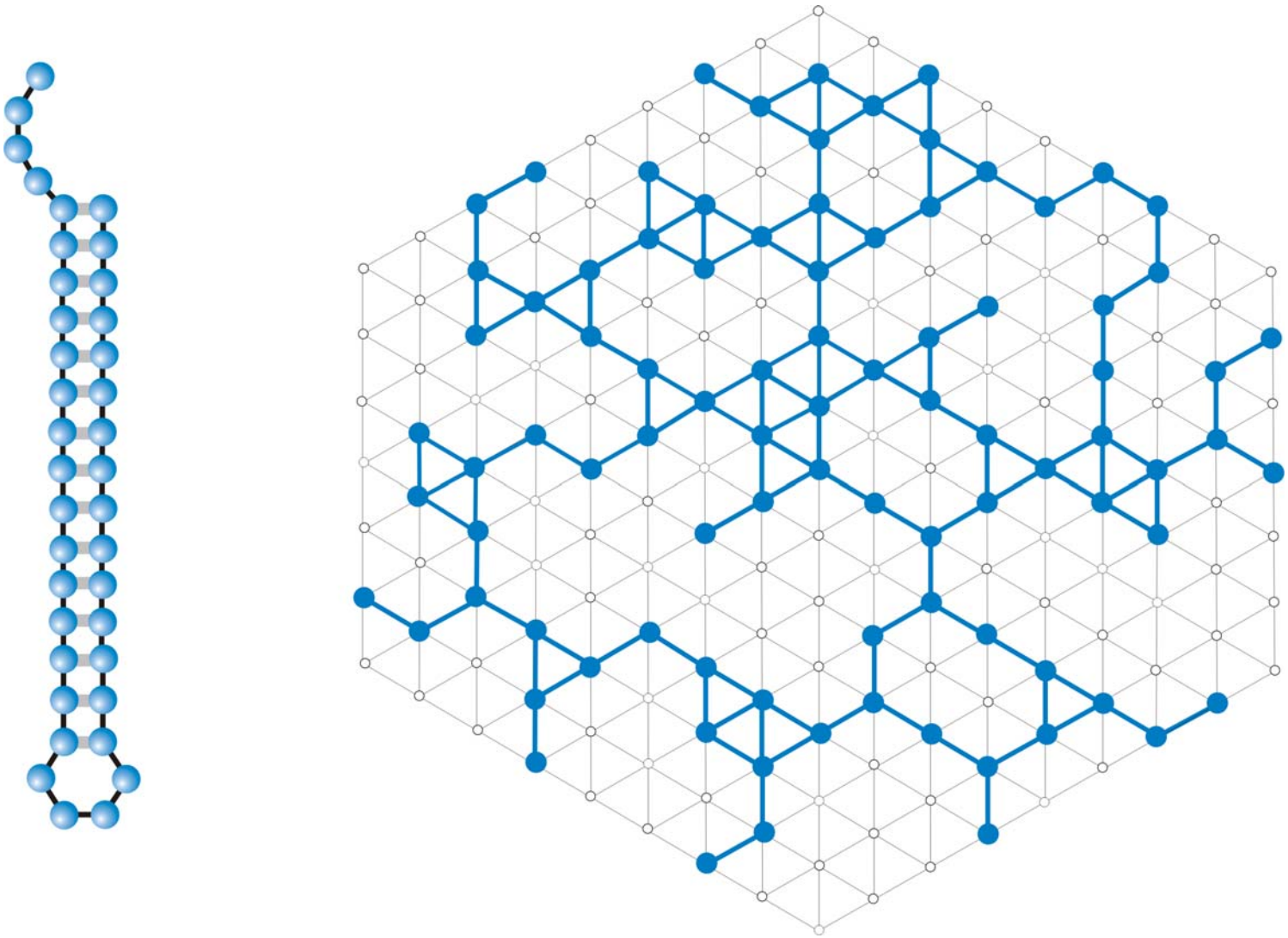
Connectivity threshold: $\lambda_{cr} = 1 - \kappa^{-1/(\kappa-1)}$

Degree of neutrality of neutral networks and the connectivity threshold

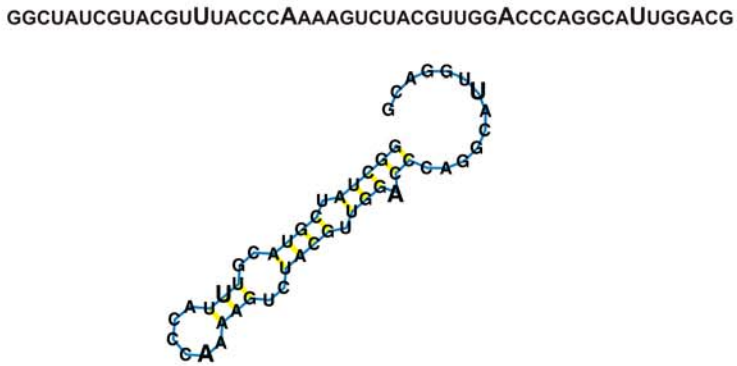




A multi-component neutral network formed by a rare structure: $\lambda < \lambda_{cr}$



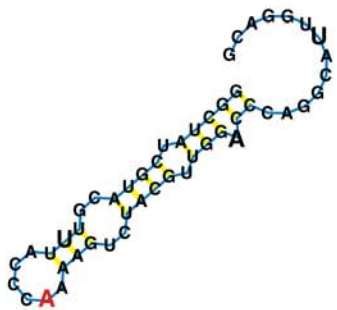
A connected neutral network formed by a common structure: $\lambda > \lambda_{cr}$



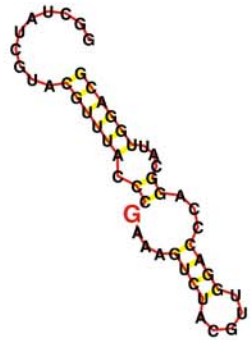
One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG

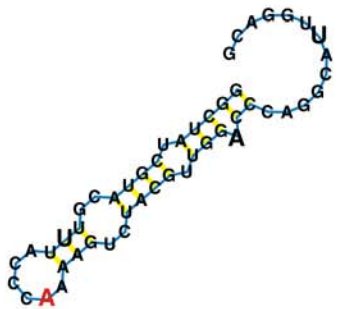


One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

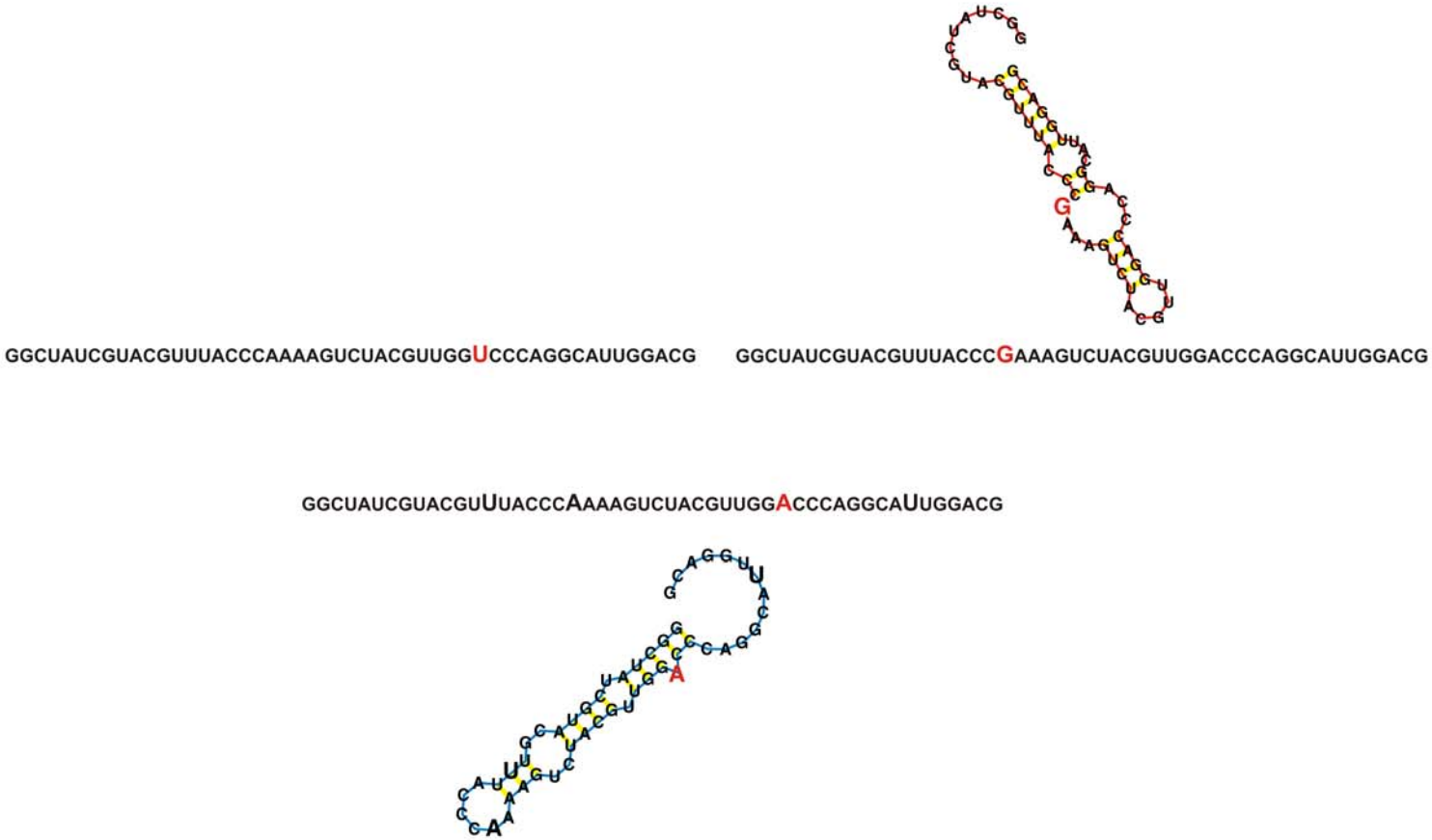


GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

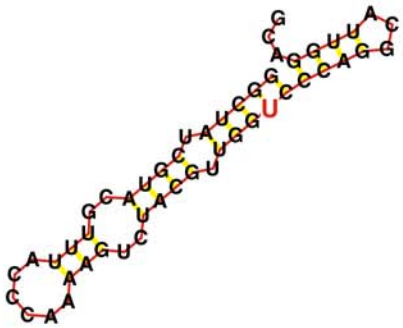
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



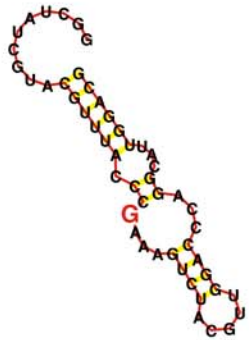
One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space



One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

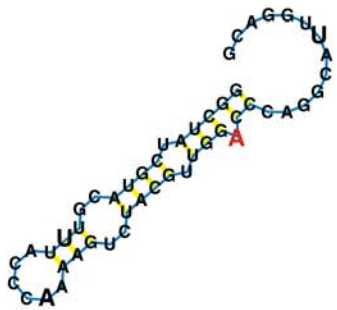


GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGUCCCAGGCAUUGGACG



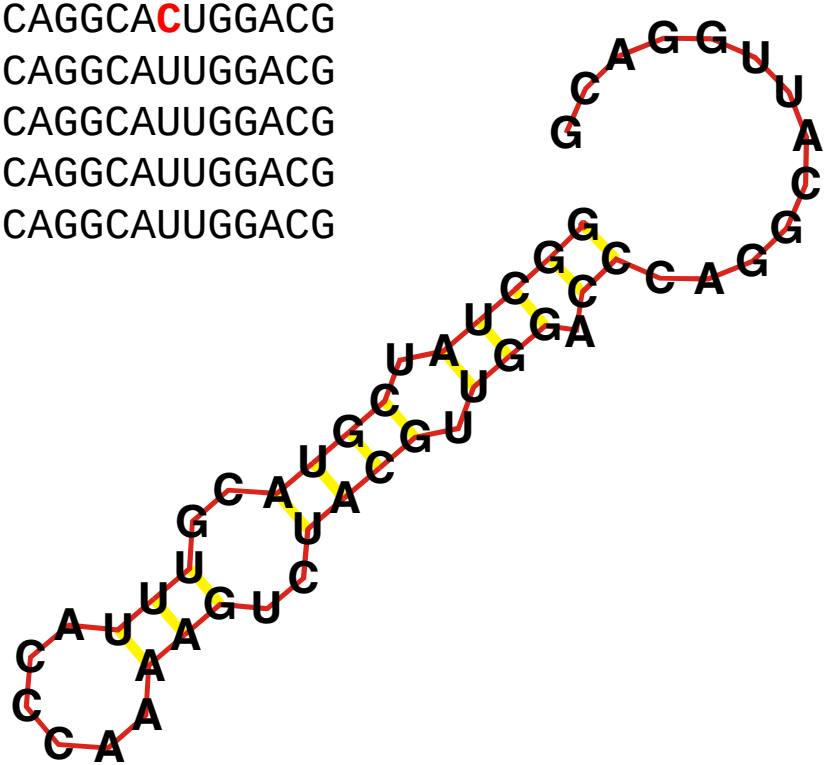
GGCUAUCGUACGUUUACCCGAAAGUCUACGUUGGACCCAGGCAUUGGACG

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG



One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

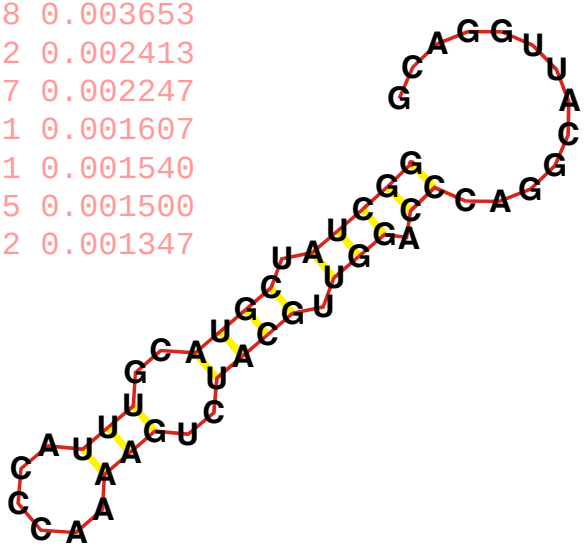
GGCUAUCGUA**U**GUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUU**A**GACG
GGCUAUCGUACGUUUAC**U**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACG**C**UUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGC**C**AUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGU**G**UACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUA**A**CGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCC**U**GGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**C**UGGACG
GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGG**U**CCCAGGCAUUGGACG
GGCUA**G**CGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCC**G**AAAGUCUACGUUGGACCCAGGCAUUGGACG
GGCUAUCGUACGUUUACCCAAAAG**C**CUACGUUGGACCCAGGCAUUGGACG



One error neighborhood – Surrounding of an RNA molecule of chain length n=50 in sequence and shape space

	Number	Mean Value	Variance	Std.Dev.
Total Hamming Distance:	150000	11.647973	23.140715	4.810480
Nonzero Hamming Distance:	99875	16.949991	30.757651	5.545958
Degree of Neutrality:	50125	**0.334167**	0.006961	**0.083434**
Number of Structures:	**1000**	**52.31**	85.30	**9.24**

1	(((((.((((..(((......)))..)))).))).))..............	50125	0.334167
2	..(((.((((..(((......)))..)))).)))..............	2856	0.019040
3	((((((((((..(((......)))..)))))))).))..............	2799	0.018660
4	(((((.((((..((((....))))..)))).))).))..............	2417	0.016113
5	(((((.((((.((((......)))).)))).))).))..............	2265	0.015100
6	(((((.(((((.(((......))).)))))).))).))..............	2233	0.014887
7	(((((..(((..(((......)))..)))..))).))..............	1442	0.009613
8	(((((.((((..((........)))..)))).))).))..............	1081	0.007207
9	((((..((((..(((......)))..)))).))).))..............	1025	0.006833
10	(((((.((((..(((......)))..)))).)))))..............	1003	0.006687
11	.((((.((((..(((......)))..)))).))))..............	963	0.006420
12	(((((.(((...(((......)))...))).))).))..............	860	0.005733
13	(((((.((((..(((......)))..)))).)).)))..............	800	0.005333
14	(((((.((((...(......)...)))).))).))..............	548	0.003653
15	(((((.((((................)))).))).))..............	362	0.002413
16	((.((.((((..(((......)))..)))).)).))..............	337	0.002247
17	(.(((.((((..(((......)))..)))).))).))..............	241	0.001607
18	(((((.(((((((((.....)))))))).))).))..............	231	0.001540
19	((((..((((..(((......)))..)))).))))..............	225	0.001500
20	((....((((..(((......)))..)))).....))..............	202	0.001347



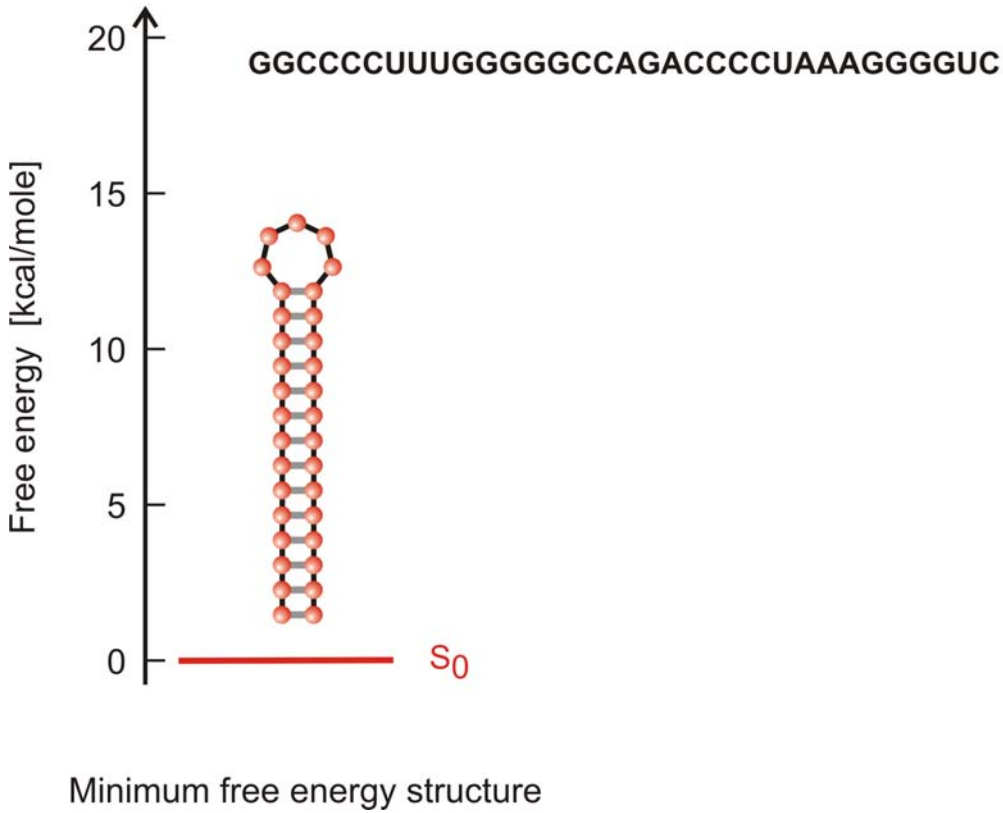
Shadow – Surrounding of an RNA structure in shape space:
AUGC alphabet, chain length n=50

Results from RNA minimun free energy structures:

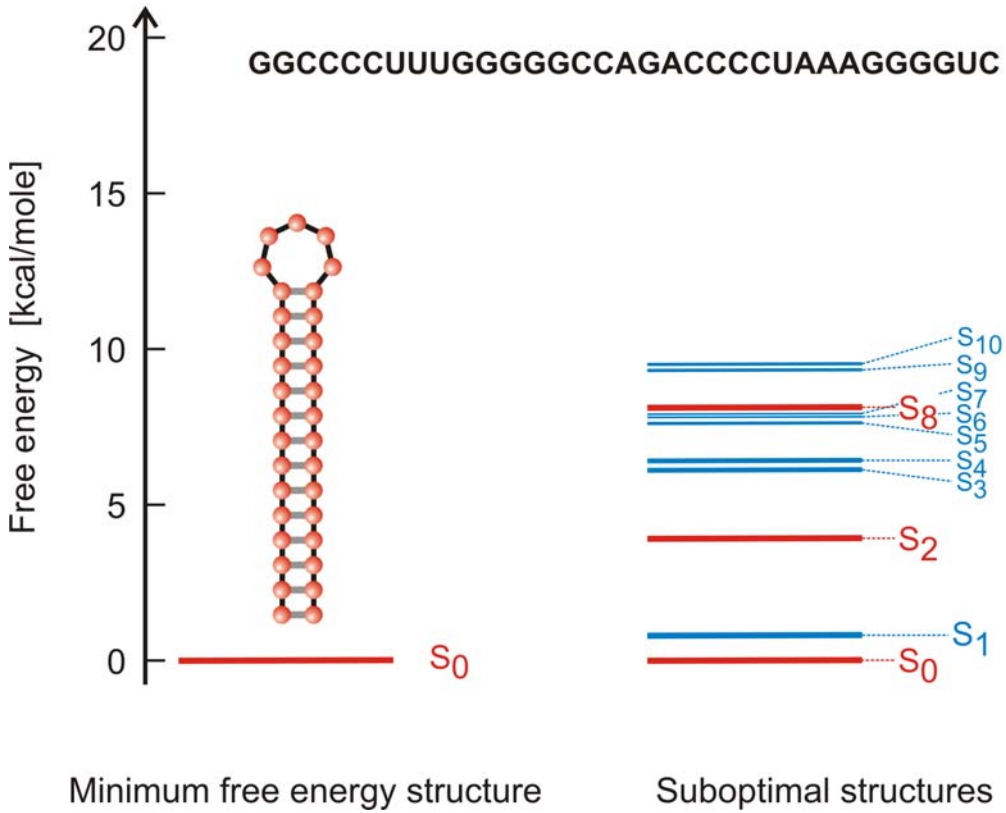
• RNA minimum free energy structures show neutrality: Many sequences fold into the same (secondary) structure.

• The single base mutation neighborhood contains structures from neutral sequences **and** a great variety of other structures: Biopolymer landscapes are **rugged**.

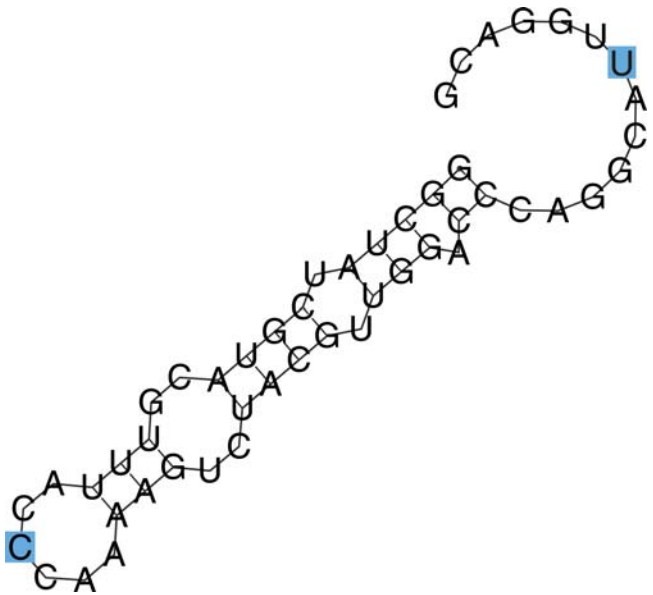
1. Minimum free energy structures of RNA
2. **Suboptimal structures of RNA**
3. Kinetic folding and RNA switches
4. Chemistry of Darwinian evolution
5. Consequences of neutrality
6. Evolutionary optimization of RNA structure



Extension of the notion of structure



Extension of the notion of structure



GGCUAUCGUACGUUUAC**C**CAAAAGUCUACGUUGGACCCAGGCA**U**UGGACG

```
(((((.((((..(((......)))..)))).))).)).............  -7.30

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

GGCUAUCGUACGUUUAC**A**CAAAAGUCUACGUUGGACCCAGGCAUUGGACG

```
(((((.((((..(((......)))..)))).))).)).............  -7.30

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

GGCUAUCGUACGUUUACCCAAAAGUCUACGUUGGACCCAGGCA**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
```

At equilibrium and temperature T the conformations form a Boltzmann ensemble that contains S_j with the Boltzmann weight $\gamma_j(T) = g_j \exp\left(-(\varepsilon_j - \varepsilon_0)/RT\right)/Q(T)$, where R is the Boltzmann constant for 1 mole, $R = N_L \cdot k_B$, and $Q(T)$ is the partition function

$$Q(T) = \sum_i g_i \exp\left(-(\varepsilon_i - \varepsilon_0)/RT\right).$$

$$\gamma_j(T) = g_j \exp\left(-(\varepsilon_j - \varepsilon_0)/RT\right)/Q(T)$$

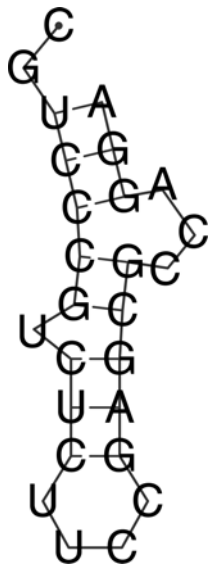
$$P(X, T) = \sum_k \gamma_k(T)\, A(S_k) \quad \text{or} \quad p_{ij}(X, T) = \sum_k \gamma_k(T)\, a_{ij}(S_k)$$

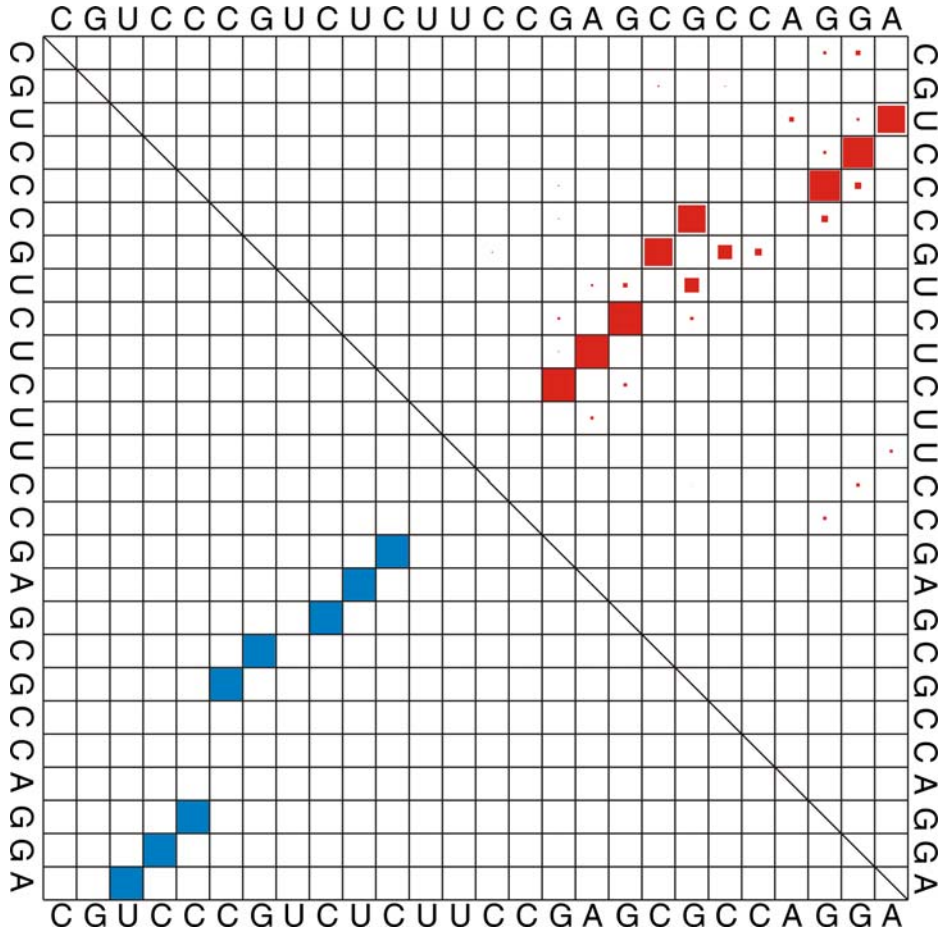
$A(S_k)$ … adjacency matrix of structure S_k

$p_{ij}(X, T)$ … base pairing probability

X … sequence

Usage of the partition function to anayze the spectrum of suboptimal states



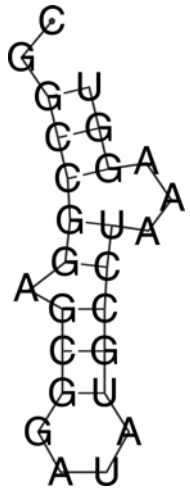


mfe-weight: 0.46336

CGUCCCGUCUCUUCCGAGCGCCAGGA

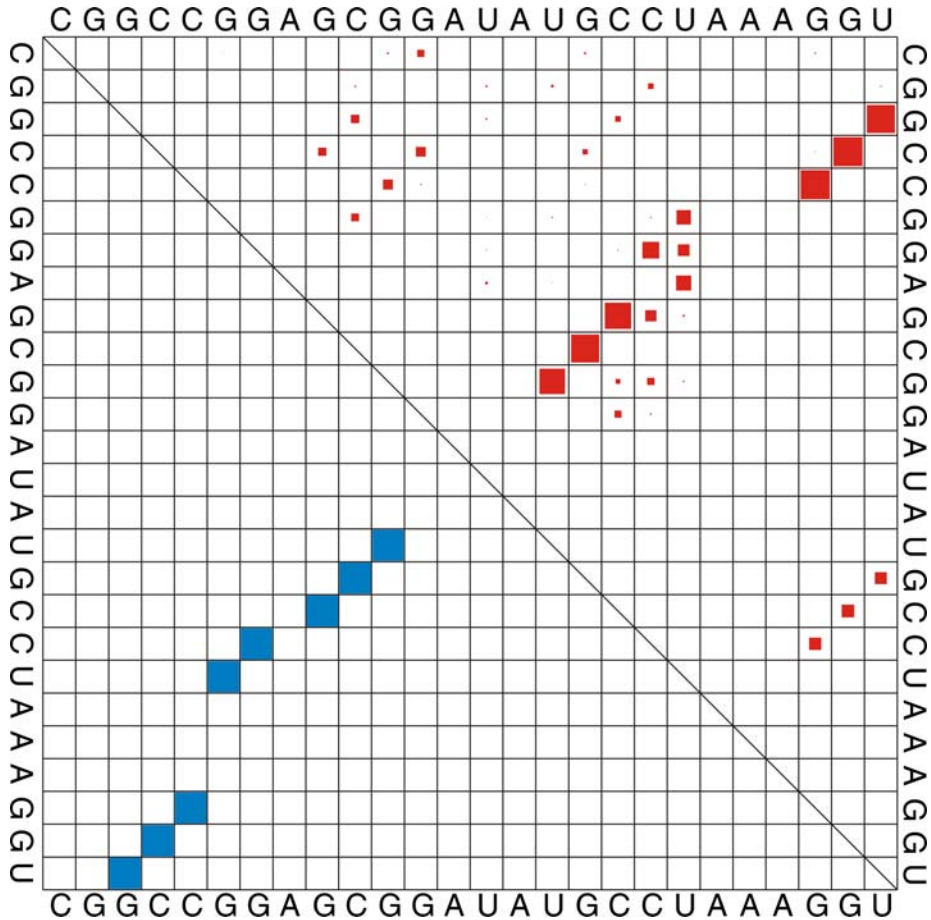
```
..(((((.(((....)))))...)))  -4.50

..(((.(((((....))).)).)))  -3.70
...((((.(((....)))))...)).  -3.60
.....((.(((....)))))......  -3.00
...((.(((((....))).)).)).  -2.80
..(((.(.(((....))).).).)))  -2.60
(.(..((.(((....)))))..).).  -2.50
```

Suboptimal structures and partition function of a small RNA molecule: n = 26



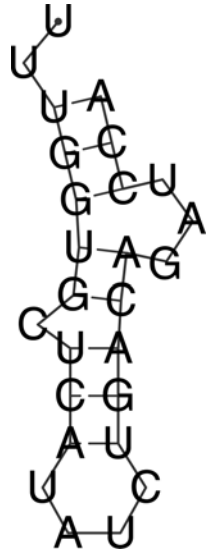
CGGCCGGAGCGGAUAUGCCUAAAGGU

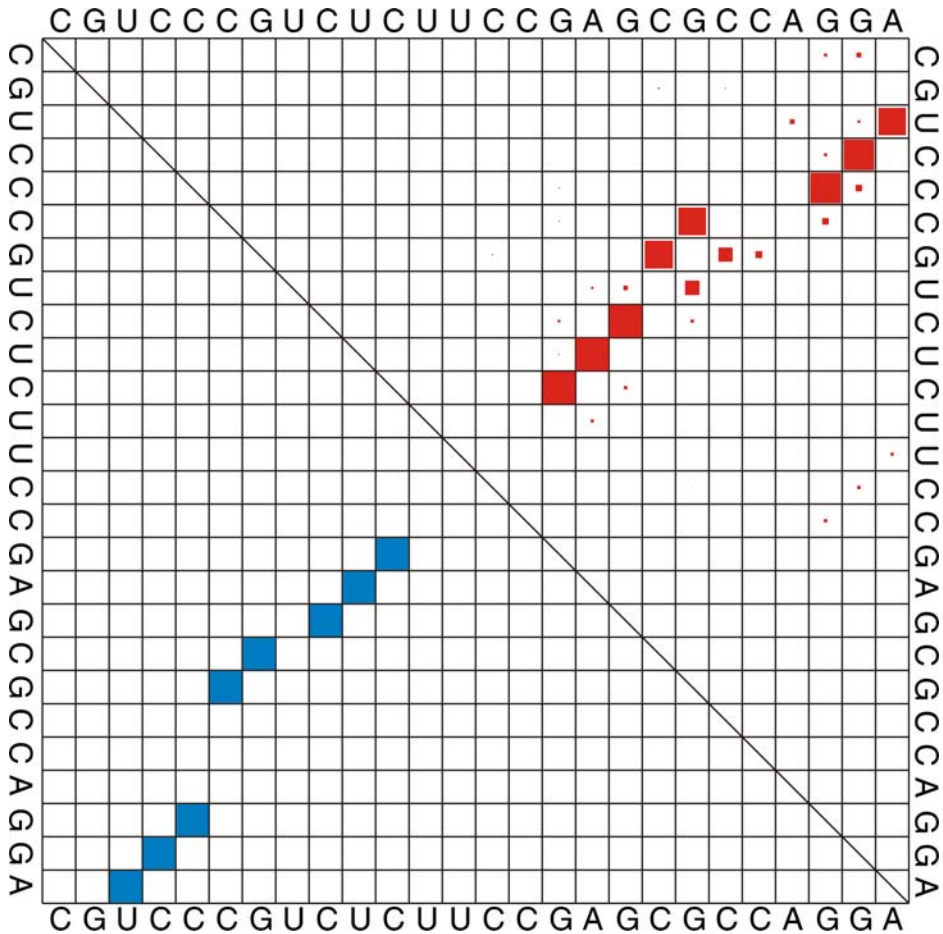
```
..(((((.(((....)))))...)))   -3.70

..(((...(((....)))).....)))  -3.60
..(((.(.(((....))).)...)))   -3.50
..(((..(((....))).)...)))    -3.30
..(((..(((....)).))...)))    -3.30
..(((.(.(((....))))....)))   -3.10
(.((....)).)....(((....)))   -2.90
..(((.....(.....)....)))     -2.90
...(((...)))....(((....)))   -2.90
..(((((.(......))))...)))    -2.70
..(((...(......).....)))     -2.60
...((.....)).....(((....)))  -2.60
..(((.(.(......)).)...)))    -2.50
..(((..((.(.....)))...)))    -2.50
.(((............)))......    -2.30
..(((..(((......)).)...)))   -2.30
..(((..(((......).))...)))   -2.30
.....((.(((....)))))......   -2.20
```



mfe-weight: 0.13642

Suboptimal structures and partition function of a small RNA molecule: n = 26



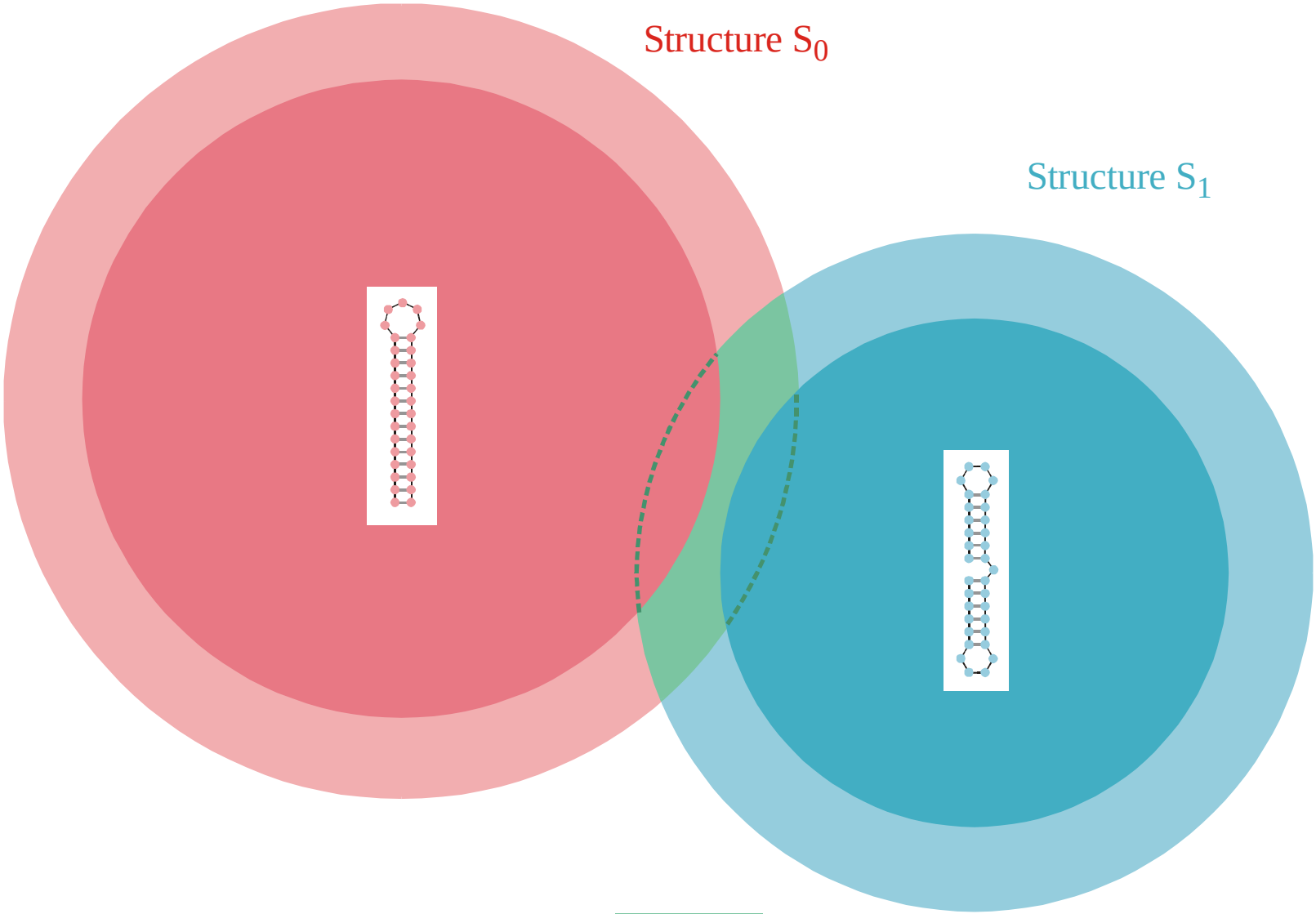


mfe-weight: 0.09514

UUUGGUGCUCAUAUCUGACAGAUCCA

```
..((((((.(((....)))))...)))  -1.10

..(((...(((....))).....)))  -1.00
...((((.(((....))))))...).  -1.00
...((...(((....))).....)).  -0.90
..((((((.(......))))...)))  -0.70
..(((...((......)).....)))  -0.60
...((((.((......))))...)).  -0.60
...((...((......)).....)).  -0.50
.....((.(((....))))).......  -0.20
..(((.(.(((....))))....)))  -0.10
..(((..((((....))))..)..)))  -0.10
((((...(........).))))....   0.00
...((.(.(((....))))....)).   0.00
...((..((((....))))..)..)).   0.00
.........................   0.00
```

Suboptimal structures and partition function of a small RNA molecule: n = 26



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

The intersection of two compatible sets is always non empty: $C_0 \cap C_1 \notin \varnothing$

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

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

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

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

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

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

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

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

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

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

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

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

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

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

Results from RNA suboptimal structures:

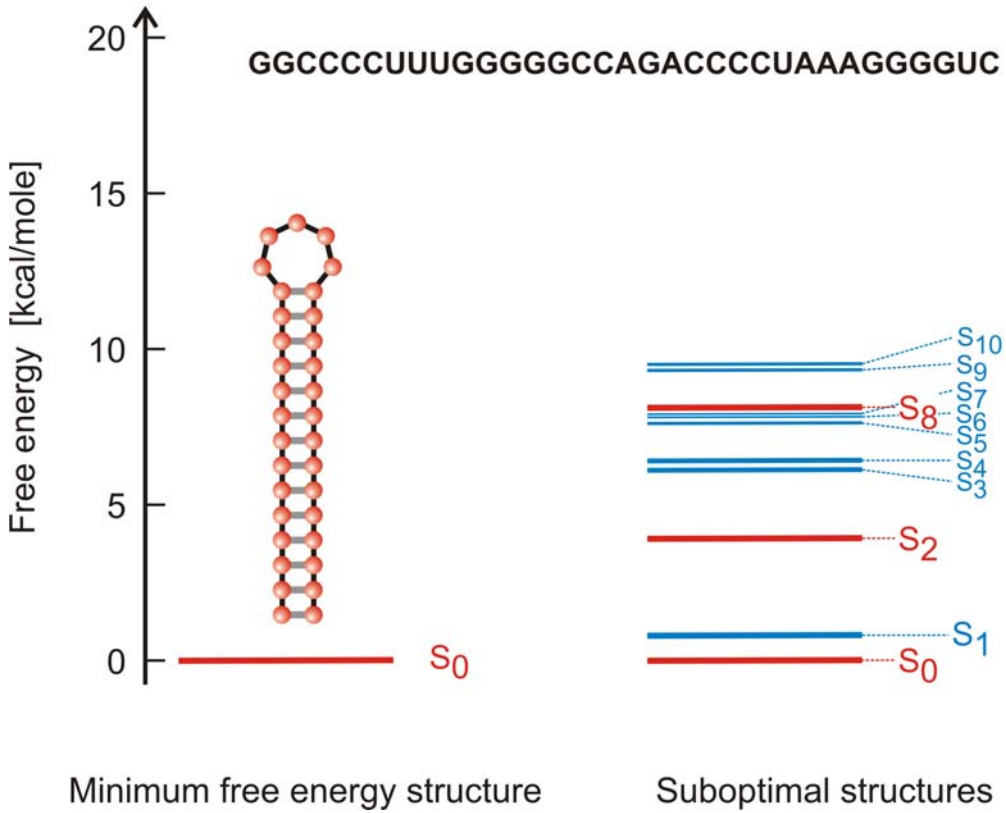
• Neutral RNA sequences differ with respect to their spectra of suboptimal structures.

• Suboptimal RNA structures with low free energies contribute substantially to the partition function.

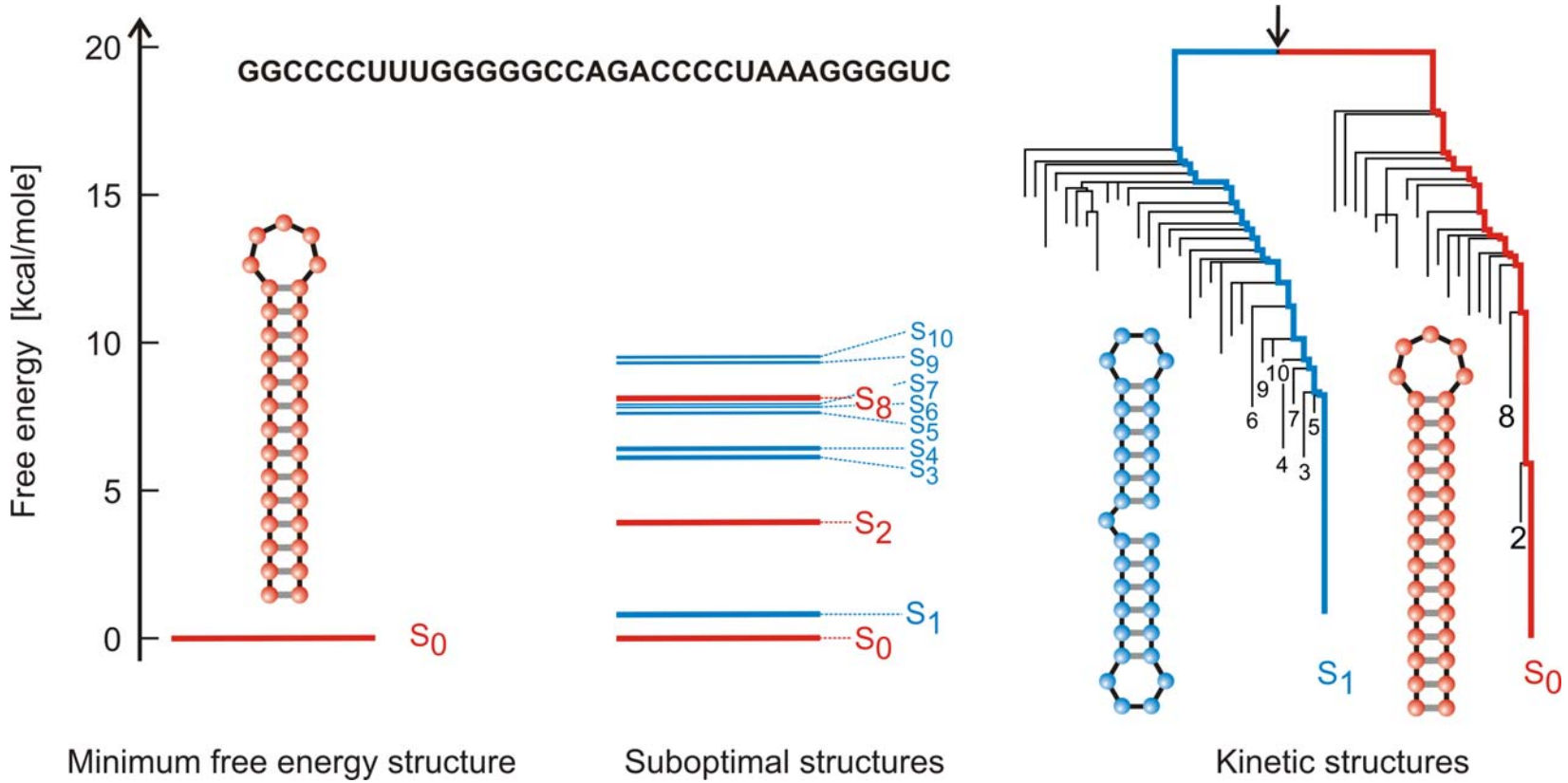
• Nature selects for stable structures in the sense that the contribution of the mfe structure to the partition function is large.

• For every pair of structures it is possible to find a sequence that can form both. This is not (always) true for three structures.

1. Minimum free energy structures of RNA
2. Suboptimal structures of RNA
3. **Kinetic folding and RNA switches**
4. Chemistry of Darwinian evolution
5. Consequences of neutrality
6. Evolutionary optimization of RNA structure



Extension of the notion of structure



Extension of the notion of structure

The Folding Algorithm

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

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

A trajectory $\mathfrak{T}_k(I)$ is a time ordered series of structures in $\mathfrak{S}(I)$. 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:

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

Master equation

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

$$k = 0,1,\ldots,m+1$$

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

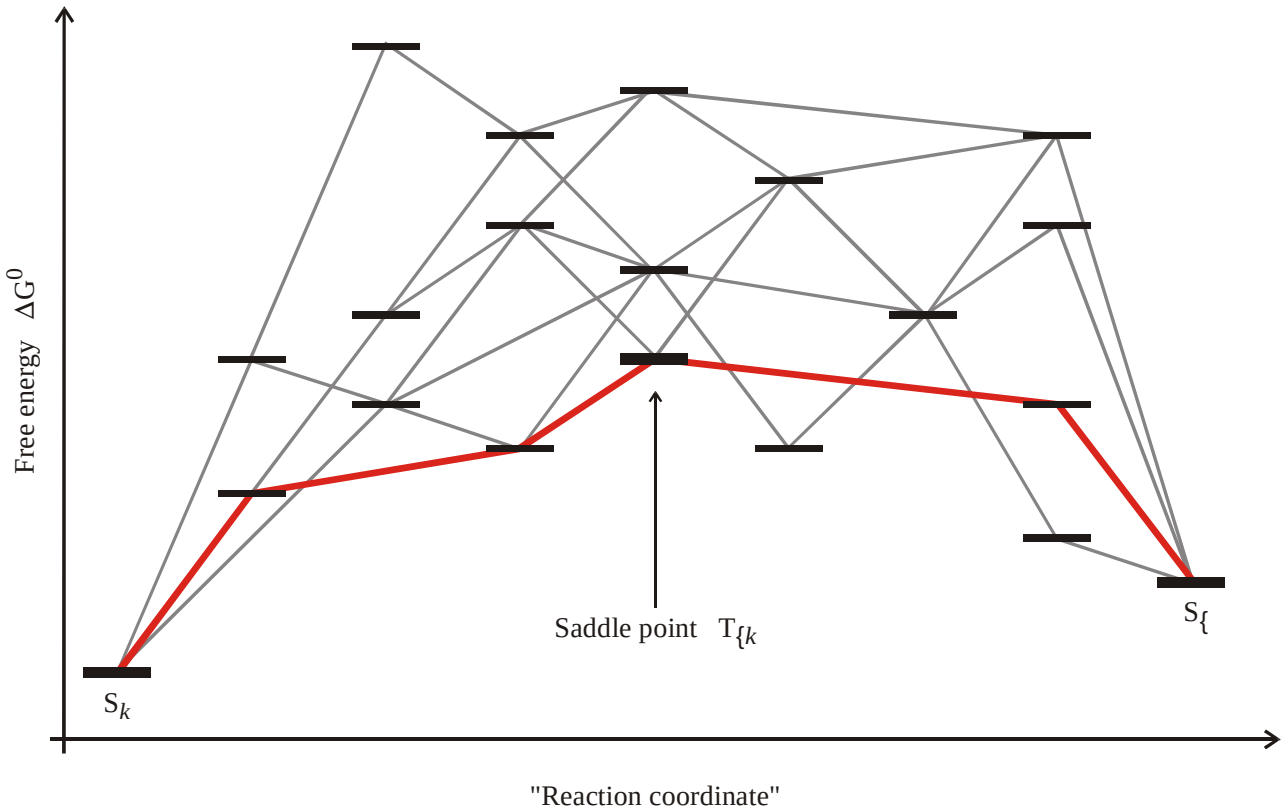
$$P_{ij}(t) = P_i(t)\, k_{ij} = P_i(t) \exp(-\Delta G_{ij}/2RT) / \Sigma_i$$

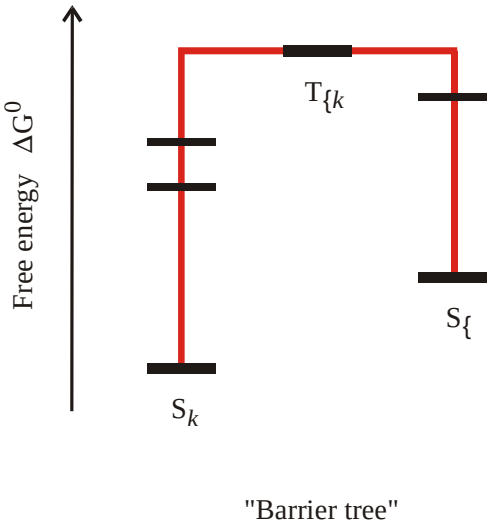
$$P_{ji}(t) = P_j(t)\, k_{ji} = P_j(t) \exp(-\Delta G_{ji}/2RT) / \Sigma_j$$

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

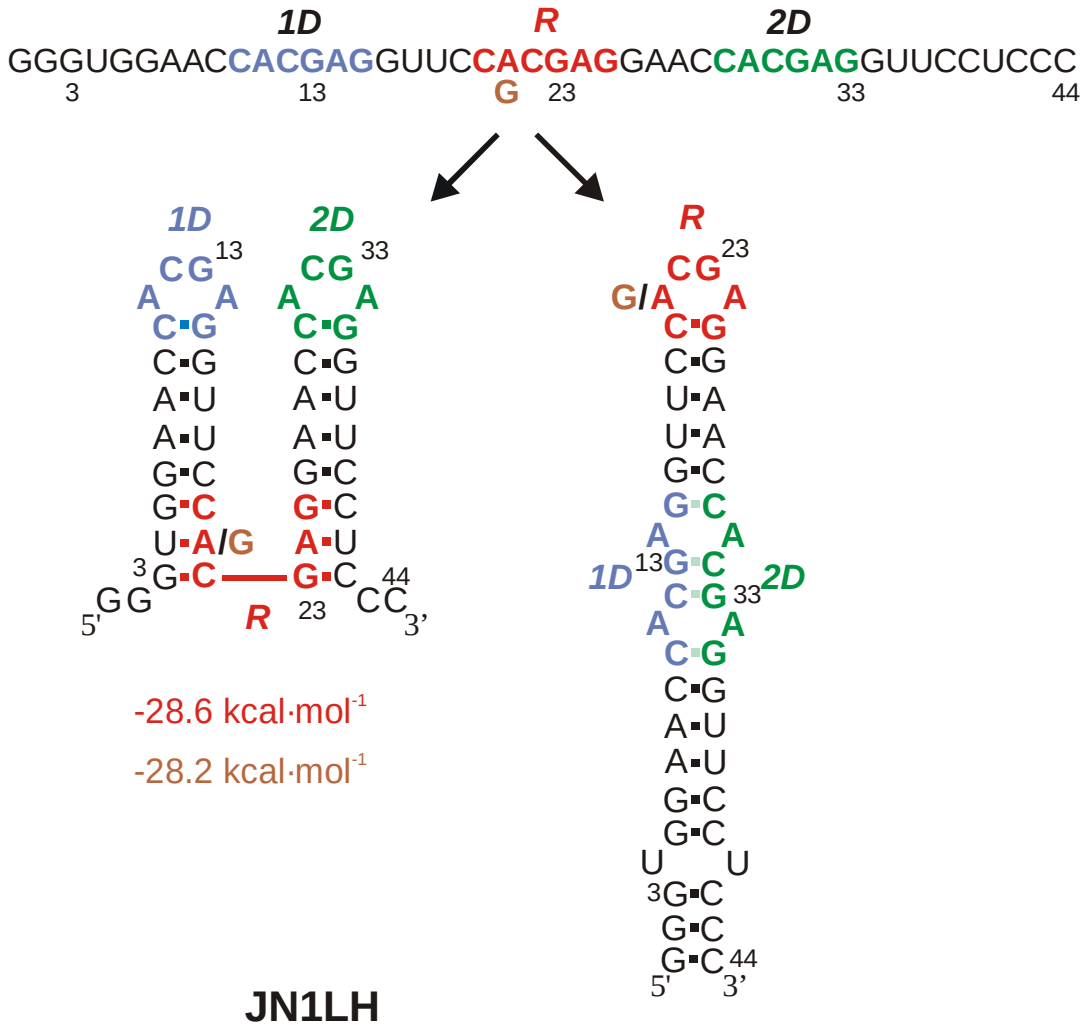
The symmetric rule for transition rate parameters is due to Kawasaki (K. Kawasaki, *Diffusion constants near the critical point for time dependent Ising models*. Phys.Rev. **145**:224-230, 1966).

Formulation of kinetic RNA folding as a stochastic process





Definition of a ‚barrier tree‘

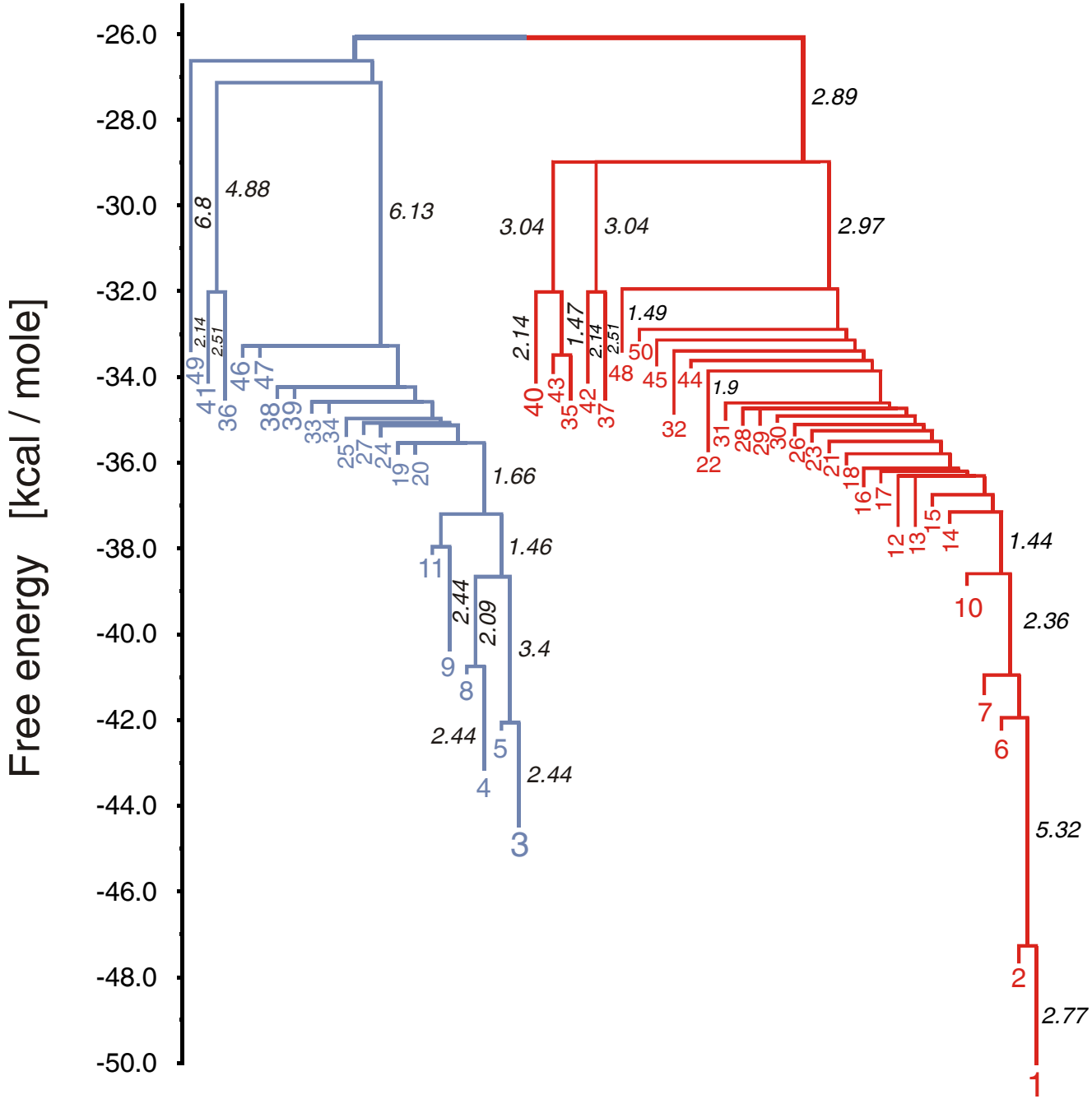


An experimental RNA switch

JN1LH

J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke, M.H. de Smit, P. Schuster, and C.W.A. Pleij.

Structural parameters affecting the kinetic competition of RNA hairpin formation. ***Nucleic Acids Res.*** **34**:3568-3576 (2006)



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

20 March 2000; accepted 22 May 2000

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

Erik A. Schultes and David P. Bartel*

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

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

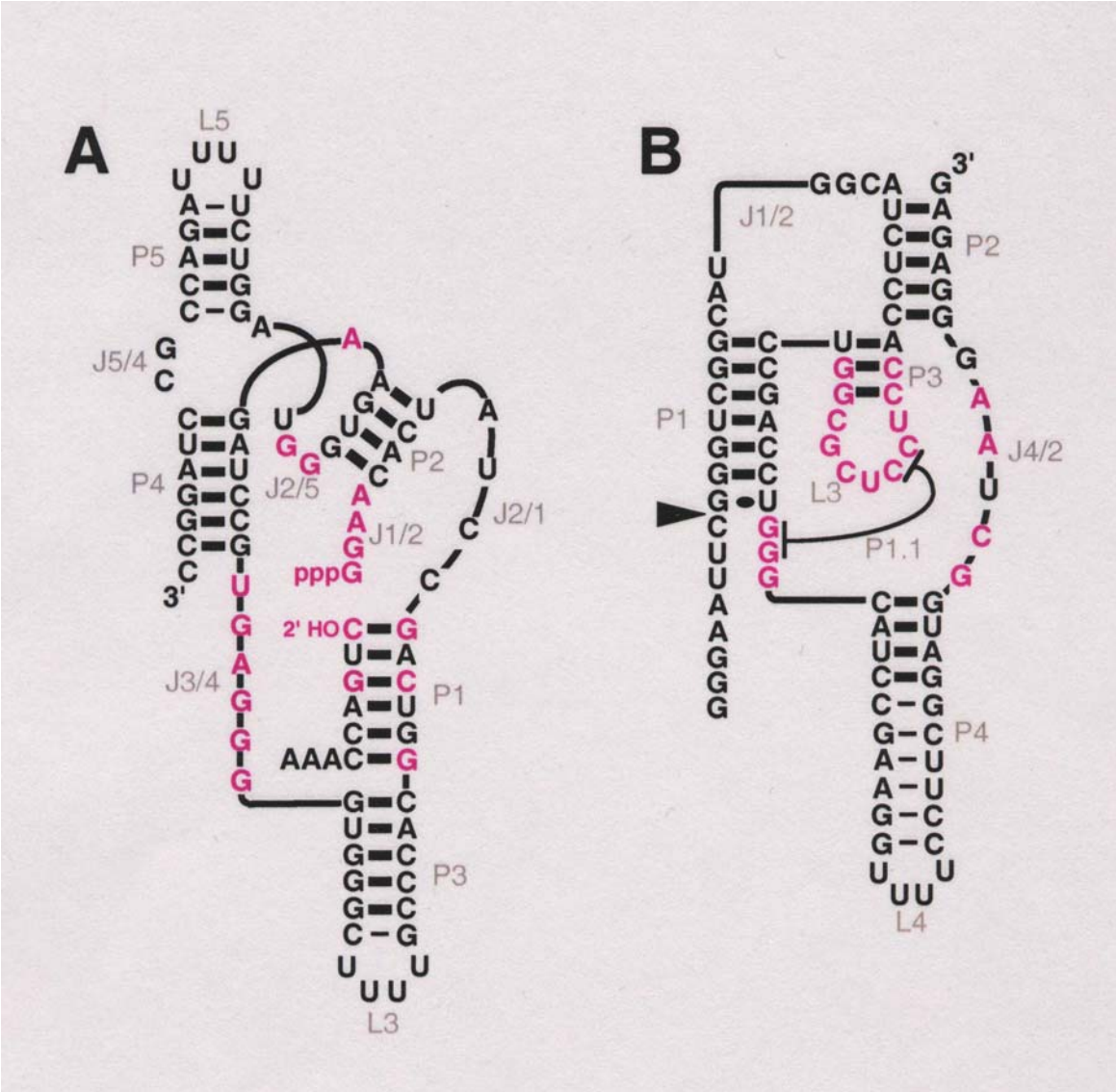
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect.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

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

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

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

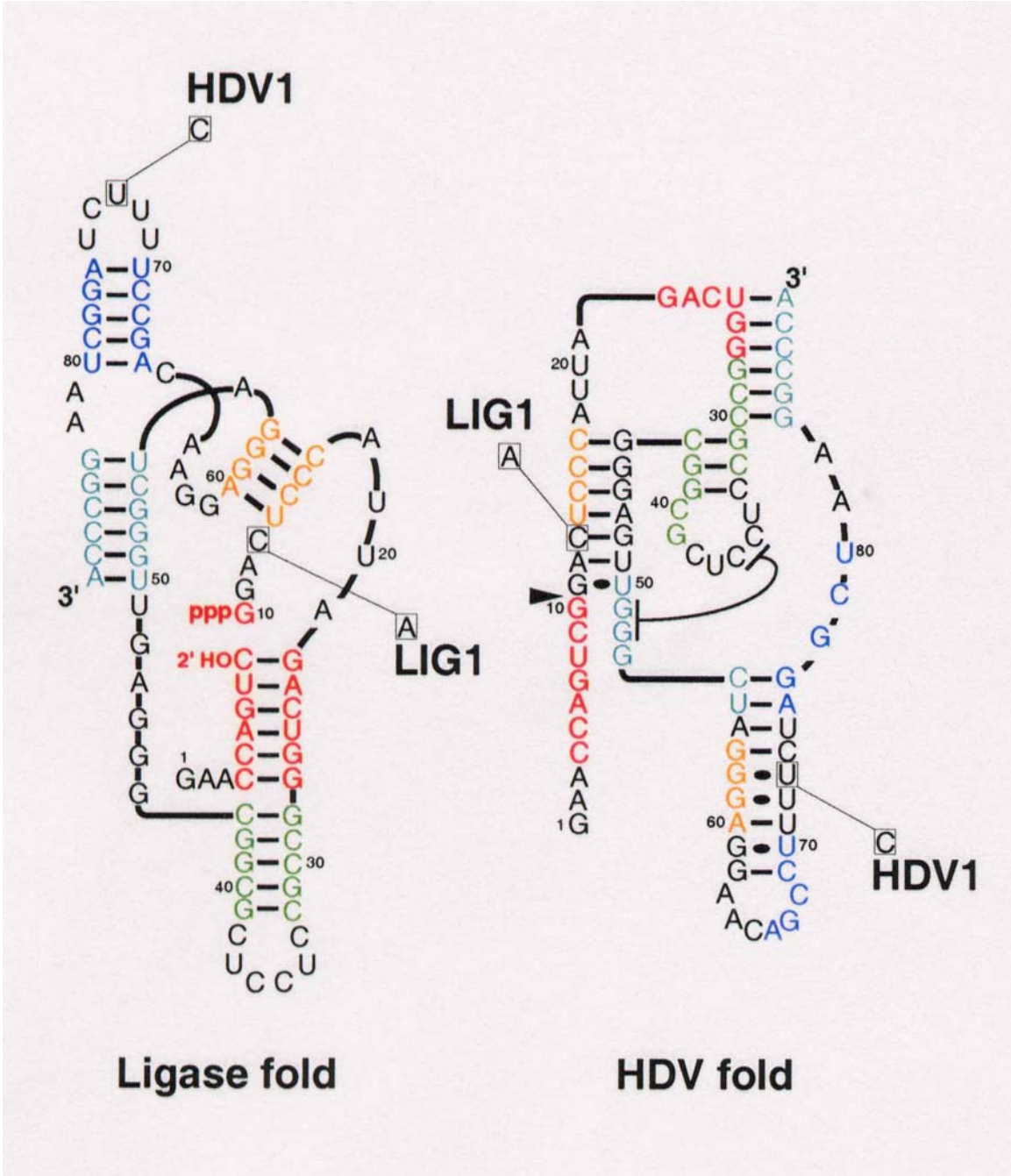
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

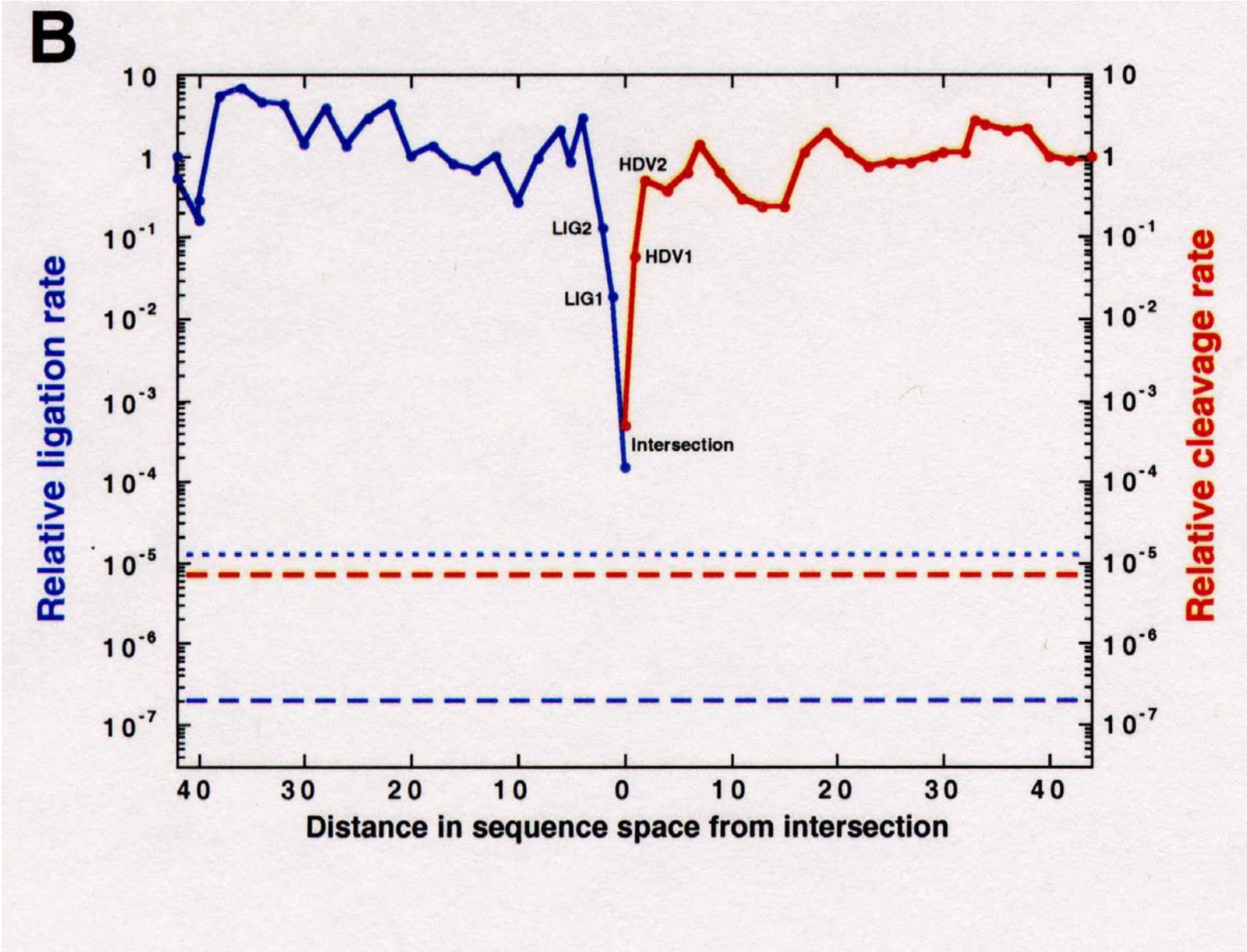


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

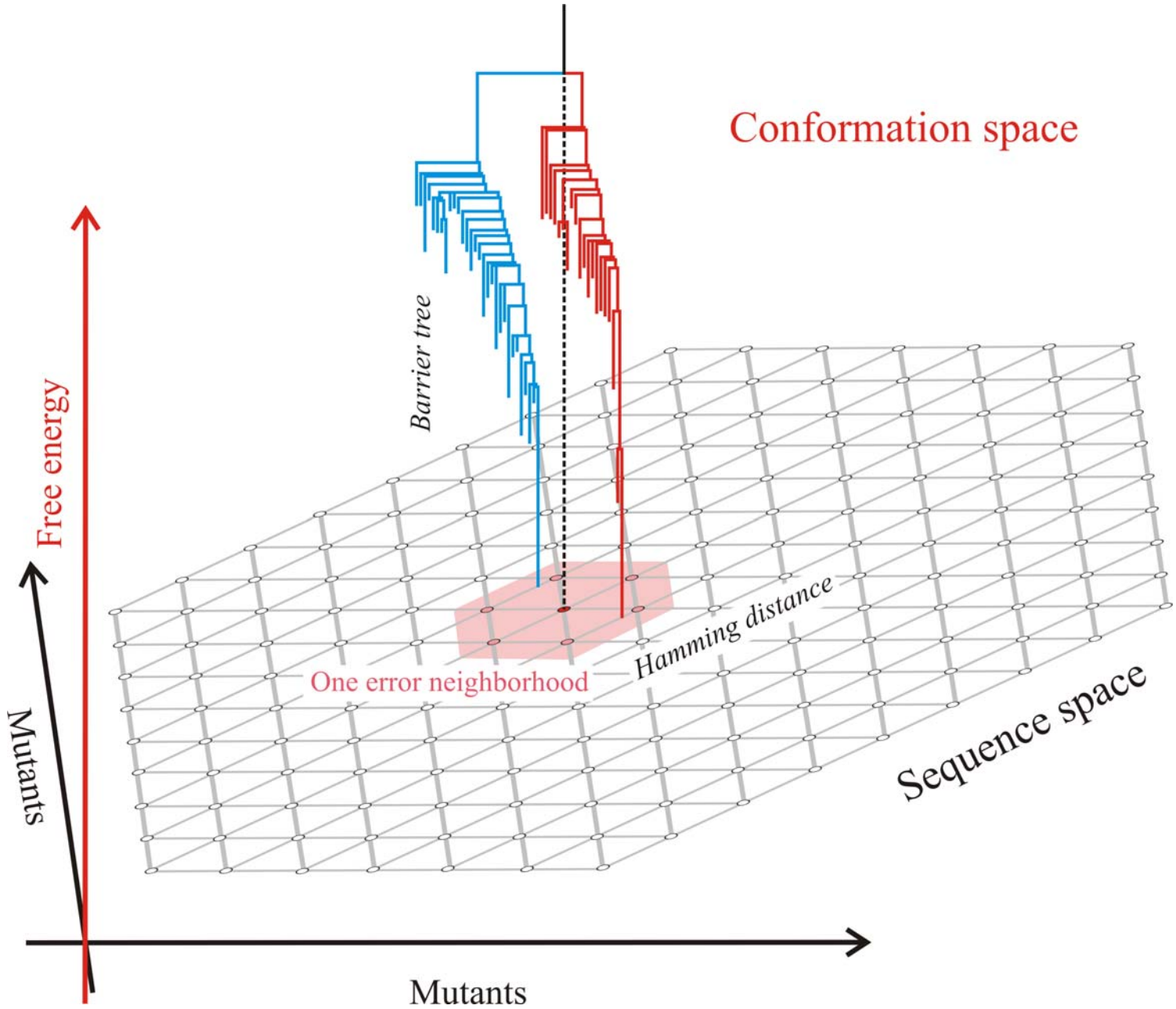
The sequence at the ***intersection***:

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

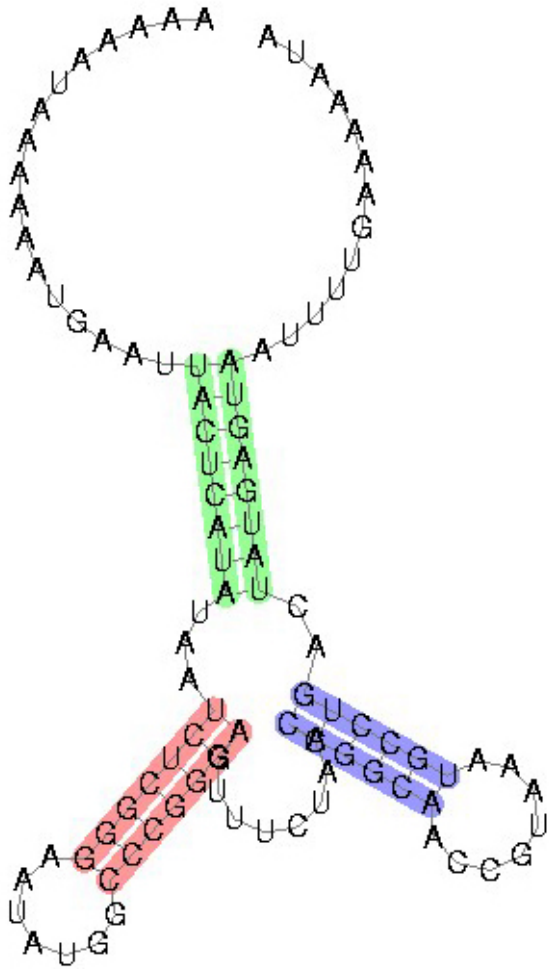




Two neutral walks through sequence space with conservation of structure and catalytic activity


Conformation space
Barrier tree
Free energy
Mutants
Mutants
One error neighborhood
Hamming distance
Sequence space

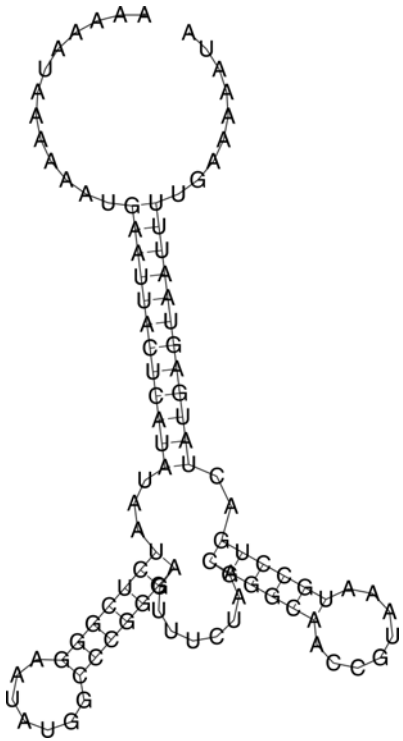
A natural metabolic riboswitch

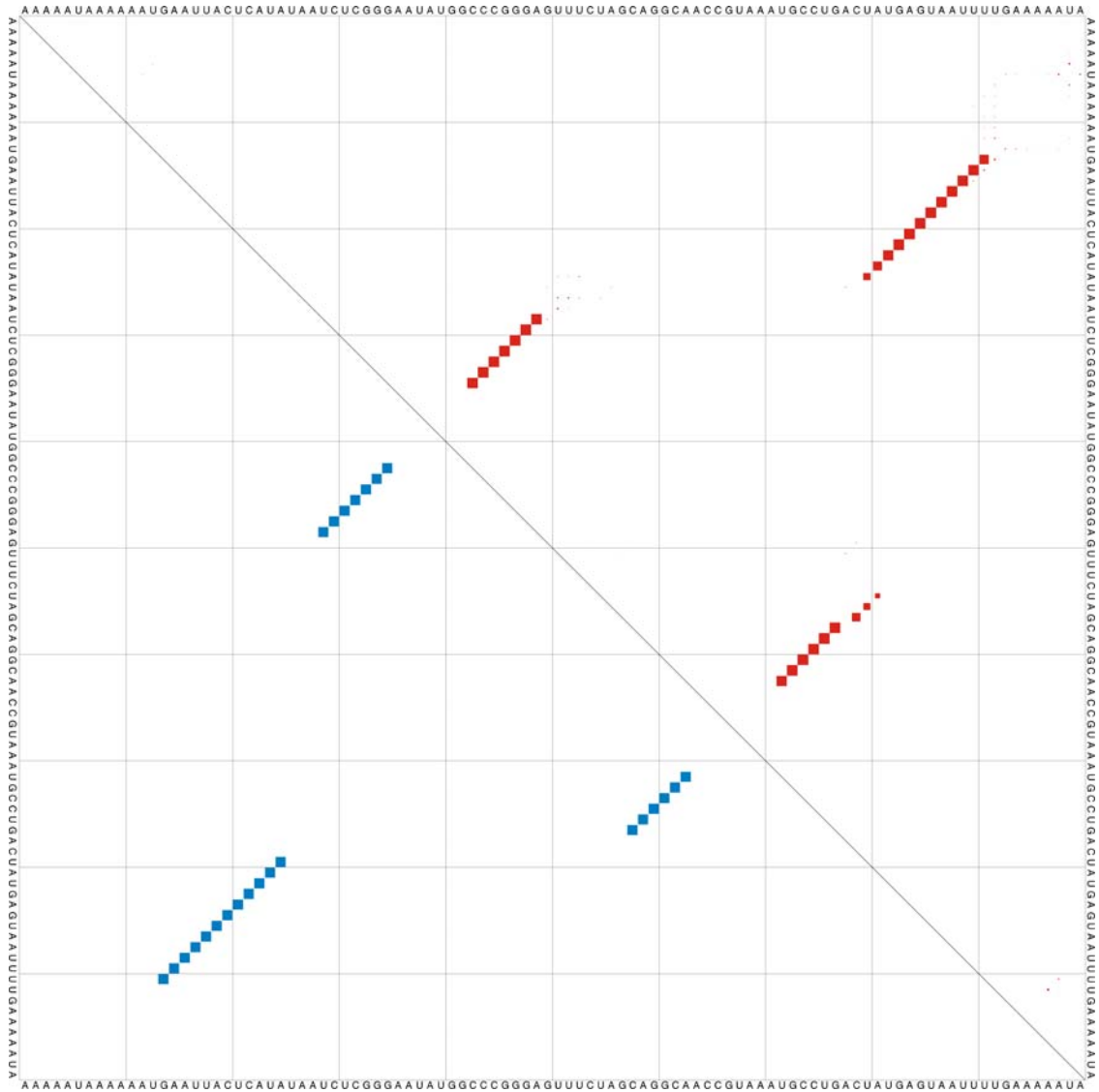


The purine riboswitch

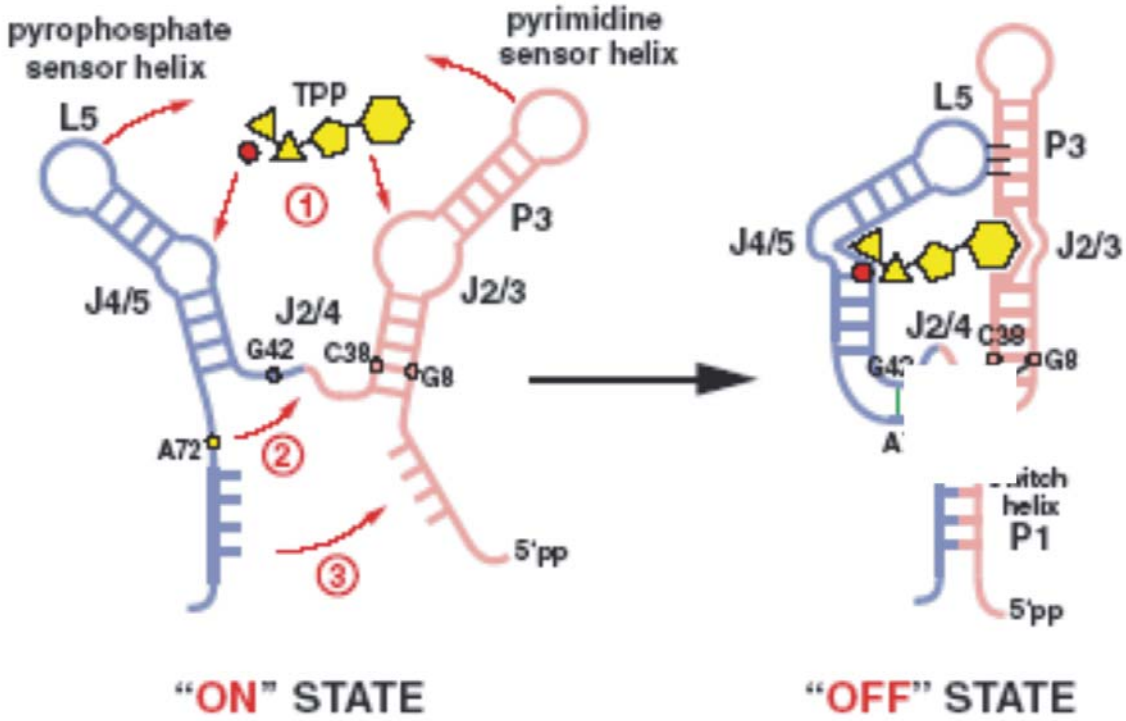
M. Mandal, B. Boese, J.E. Barrick, W.C. Winkler, and R.R. Breaker. 2003. *Molecular Cell*. **11**:1419-1420, *Cell* **113**:577-586.

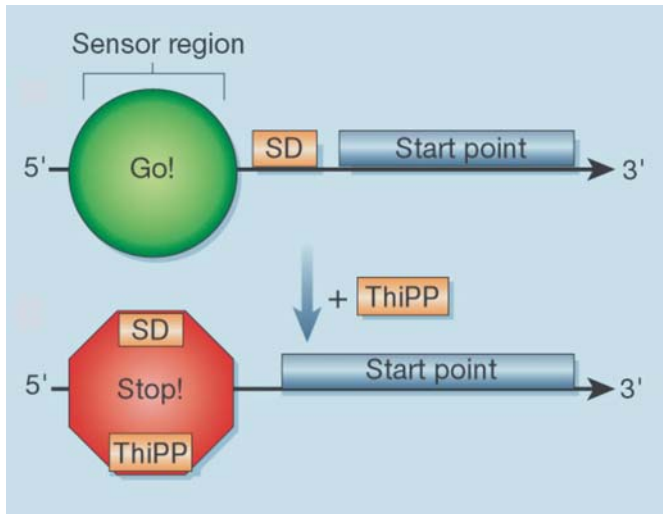
The purine riboswitch: *Molecular Cell*. 2003. **11**:1419-1420.





mfe-weight: 0.1459





The thiamine-pyrophosphate riboswitch

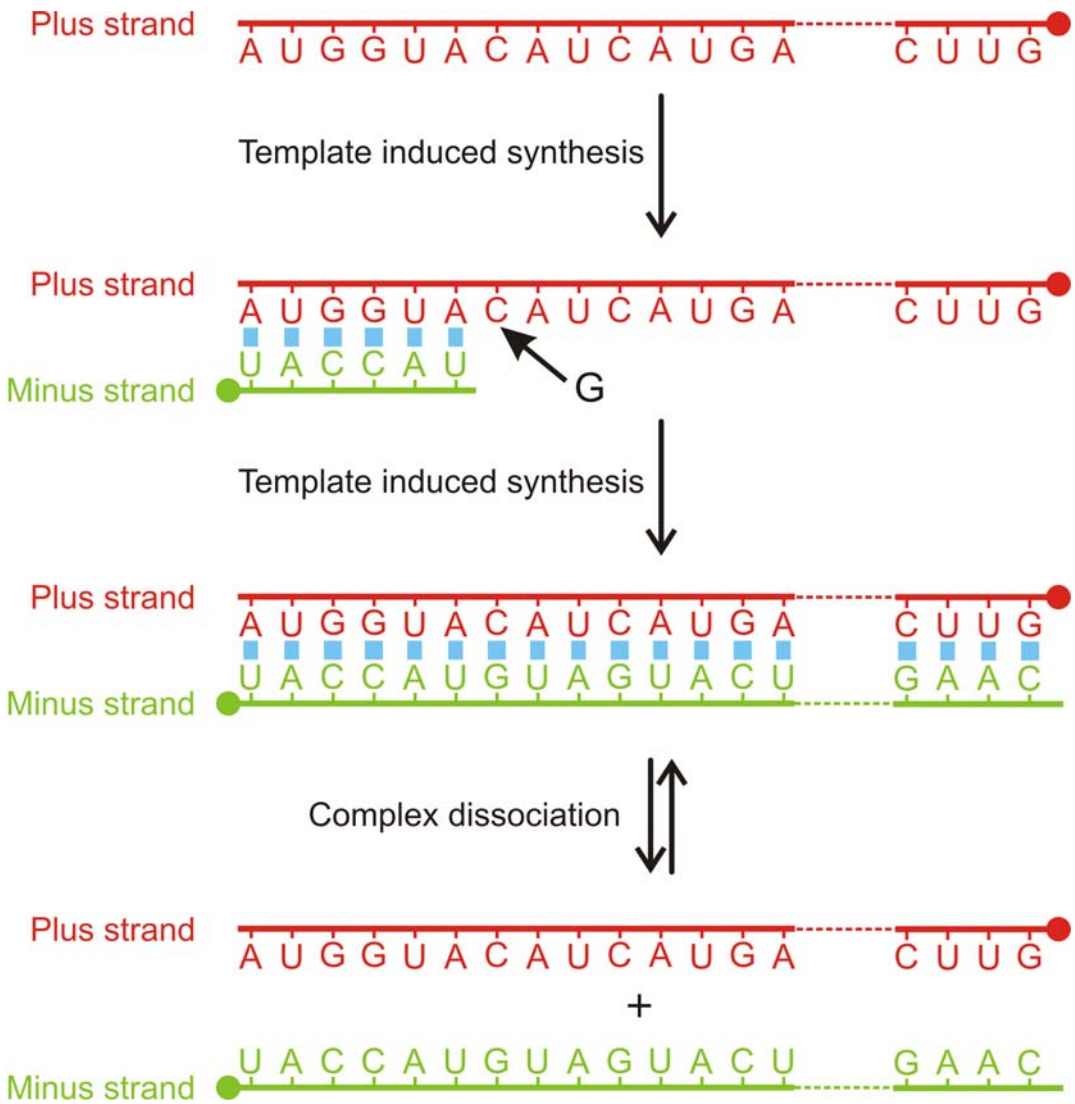
S. Thore, M. Leibundgut, N. Ban.
Science **312**:1208-1211, 2006.

Results from RNA folding kinetics:

• In addition to the minimum free energy structure RNA molecules can exist in one, two or more long-lived metastable structures.

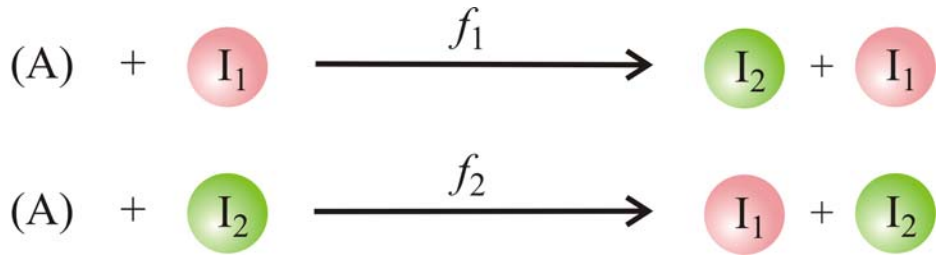
• RNA switches are molecules with two or more long-lived conformations that allow for metabolic control.

1. Minimum free energy structures of RNA
2. Suboptimal structures of RNA
3. Kinetic folding and RNA switches
4. **Chemistry of Darwinian evolution**
5. Consequences of neutrality
6. Evolutionary optimization of RNA structure



Complementary replication is the simplest copying mechanism of RNA.
Complementarity is determined by Watson-Crick base pairs:

G≡C and **A=U**



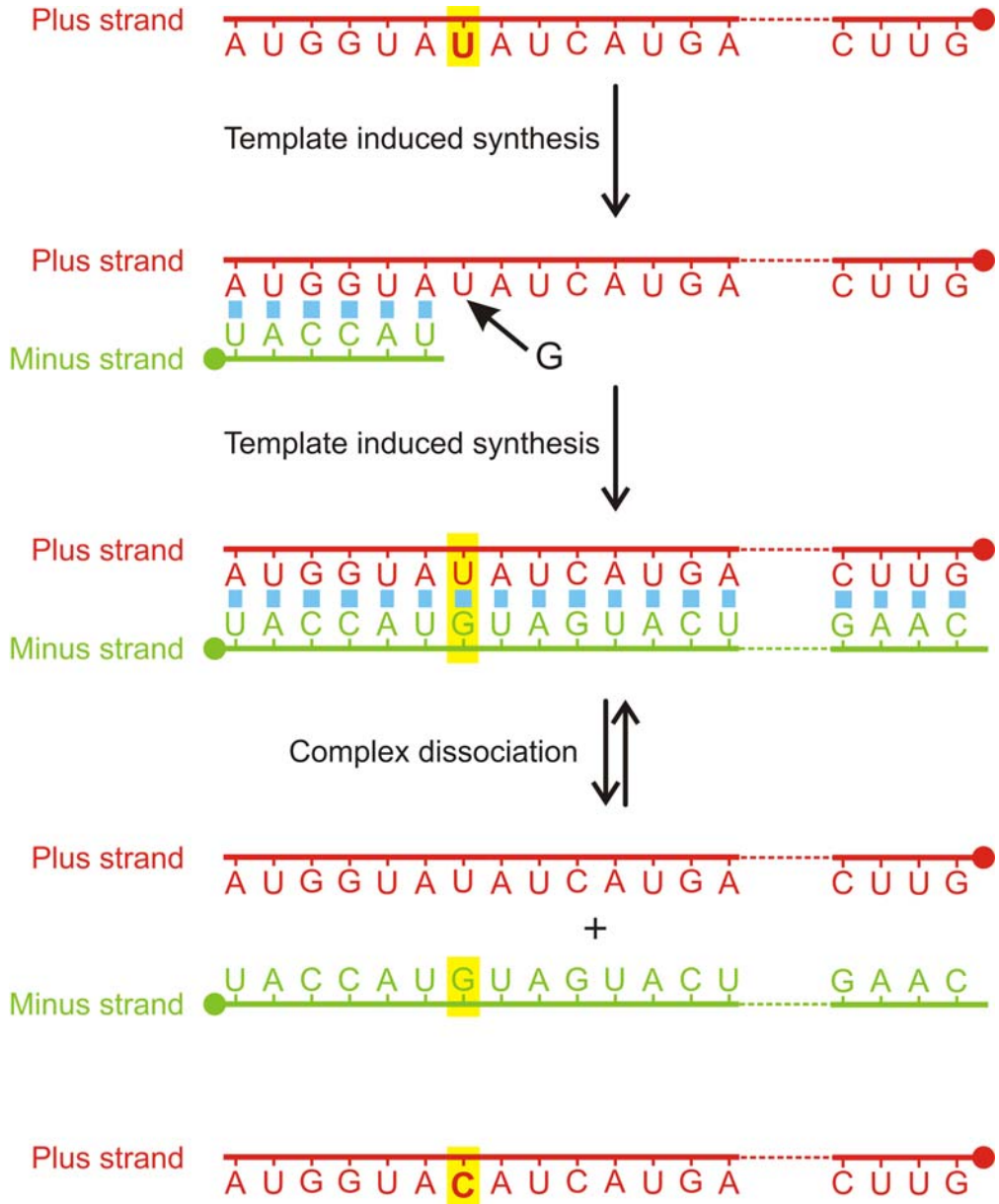
$$\frac{dx_1}{dt} = f_2\, x_2 \quad \text{and} \quad \frac{dx_2}{dt} = f_1\, x_1$$

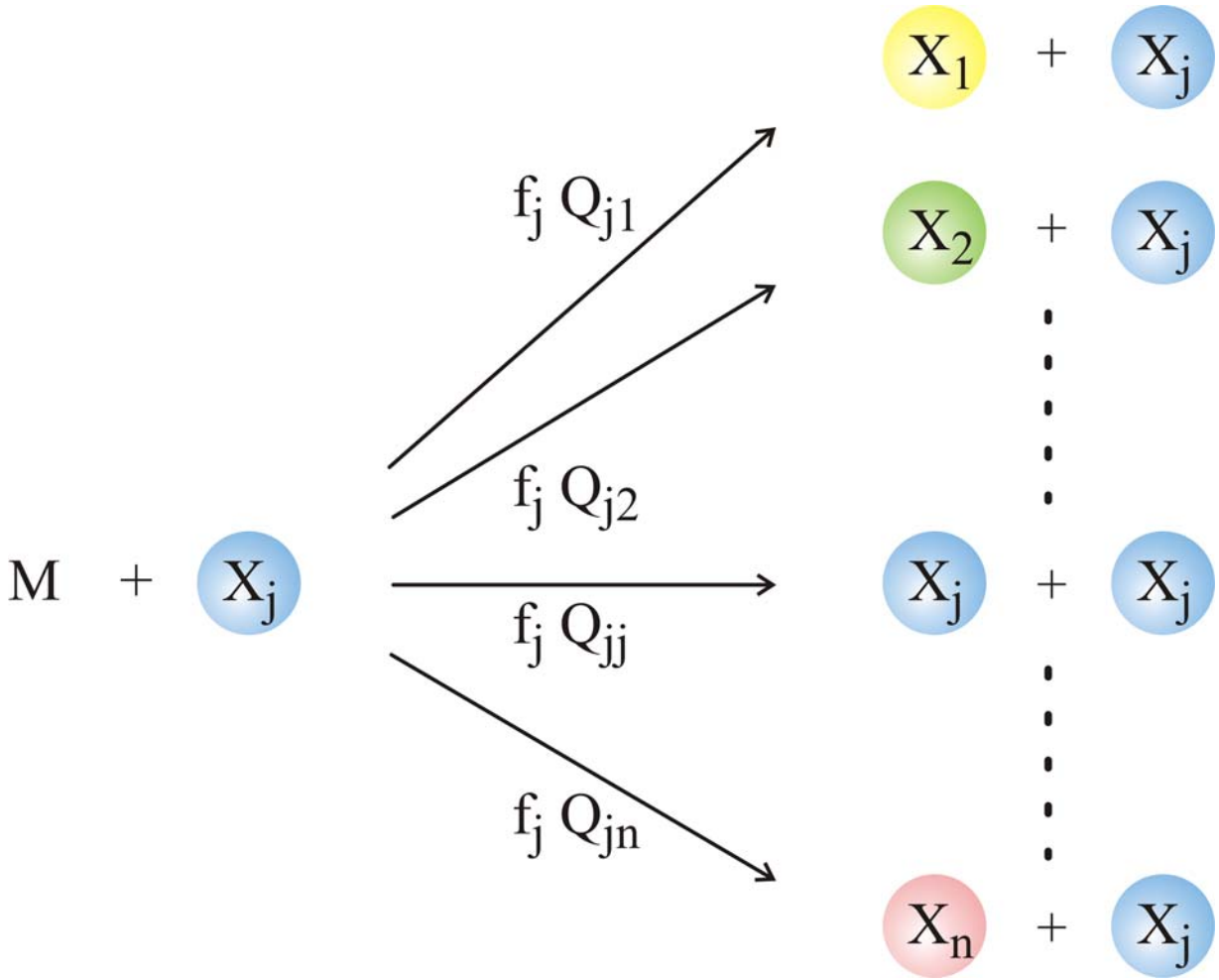
$$x_1 = \sqrt{f_2}\;\xi_1\,, \quad x_2 = \sqrt{f_1}\;\xi_2\,, \quad \zeta = \xi_1 + \xi_2\,, \quad \eta = \xi_1 - \xi_2\,, \quad f = \sqrt{f_1 f_2}$$

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

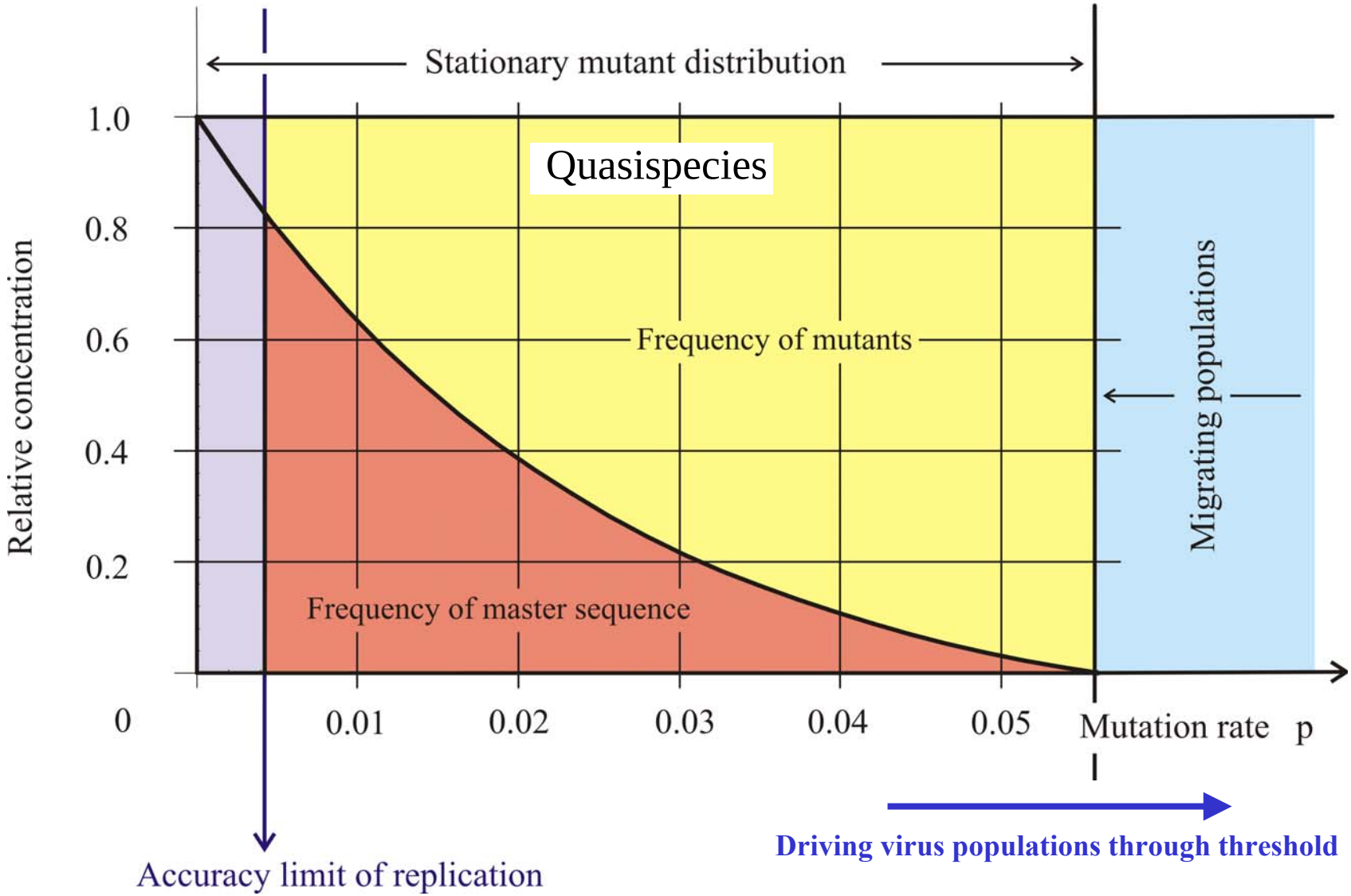
$$\zeta(t) = \zeta(0)\, e^{ft}$$

Complementary replication as the simplest molecular mechanism of reproduction

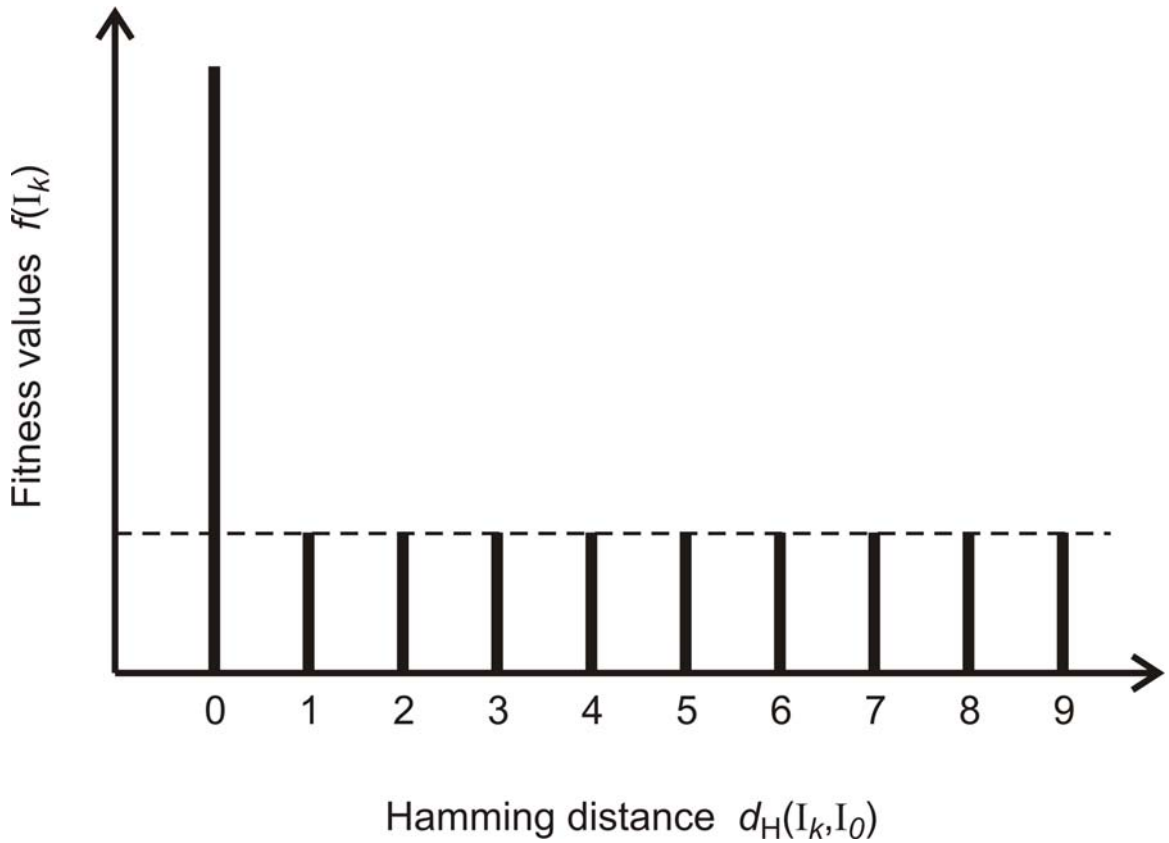

Plus strand
A U G G U A U A U C A U G A
C U U G
Template induced synthesis
Plus strand
A U G G U A U A U C A U G A
C U U G
U A C C A U
Minus strand
G
Template induced synthesis
Plus strand
A U G G U A U A U C A U G A
C U U G
U A C C A U G U A G U A C U
G A A C
Minus strand
Complex dissociation
Plus strand
A U G G U A U A U C A U G A
C U U G
+
U A C C A U G U A G U A C U
G A A C
Minus strand
Plus strand
A U G G U A C A U C A U G A
C U U G



Chemical kinetics of replication and mutation as parallel reactions



The error threshold in replication



A fitness landscape showing an error threshold

Biophysical Chemistry 16 (1982) 329–345
Elsevier Biomedical Press

329

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 $\nu \approx 80$ 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 ($\nu > 20$).

1. Introduction

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

$$\frac{dx_i}{dt} = \dot{x}_i = \sum_j w_{ij}x_j - \frac{x_i}{c}\phi; \; i = 1,\ldots,n \; ^{\dagger} \qquad (1)$$

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.

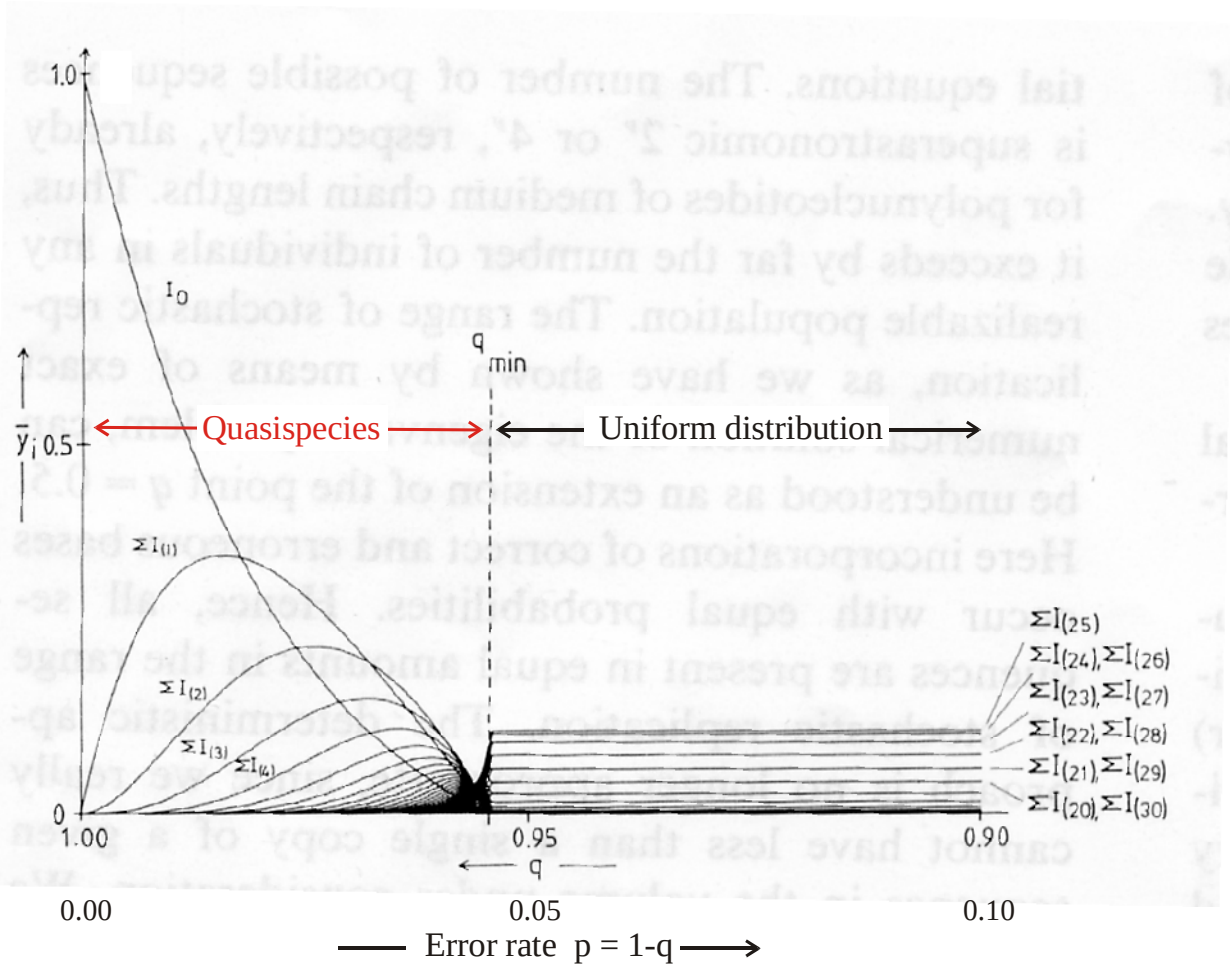
Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional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

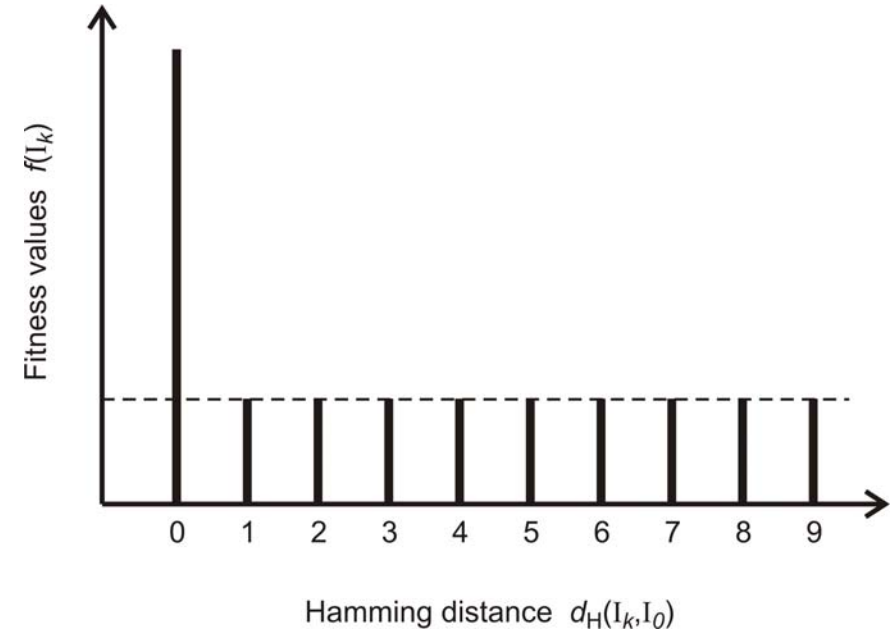
** This paper is considered as part II of Model Studies on RNA replication. Part I is by Gassner and Schuster [14].

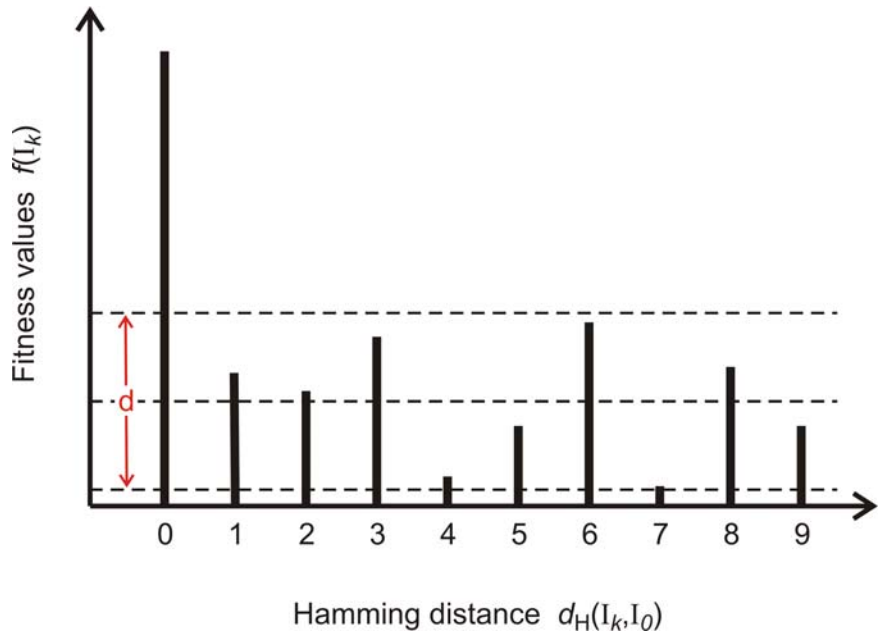
† All summations throughout this paper run from 1 to n unless specified differently: $\Sigma_i = \Sigma_{i=1}^{n}$ and $\Sigma_{i,i\neq j} = \Sigma_{i=1}^{j-1} + \Sigma_{i=j+1}^{n}$, respectively.

0301-4622/82/0000-0000/$02.75

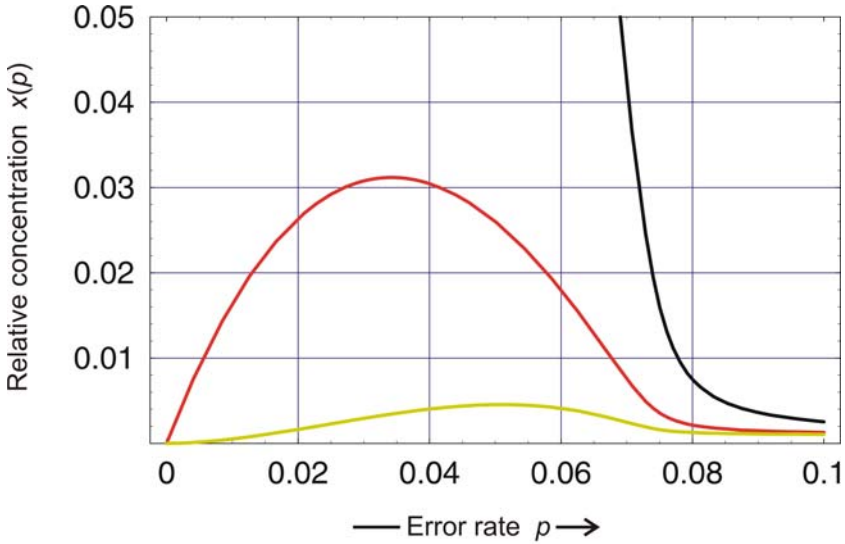


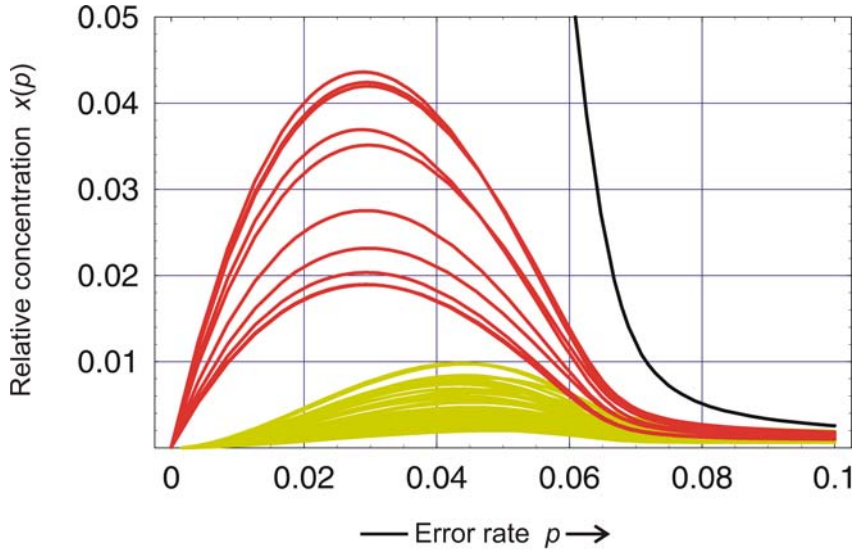
Stationary population or quasispecies as a function of the mutation or error rate *p*

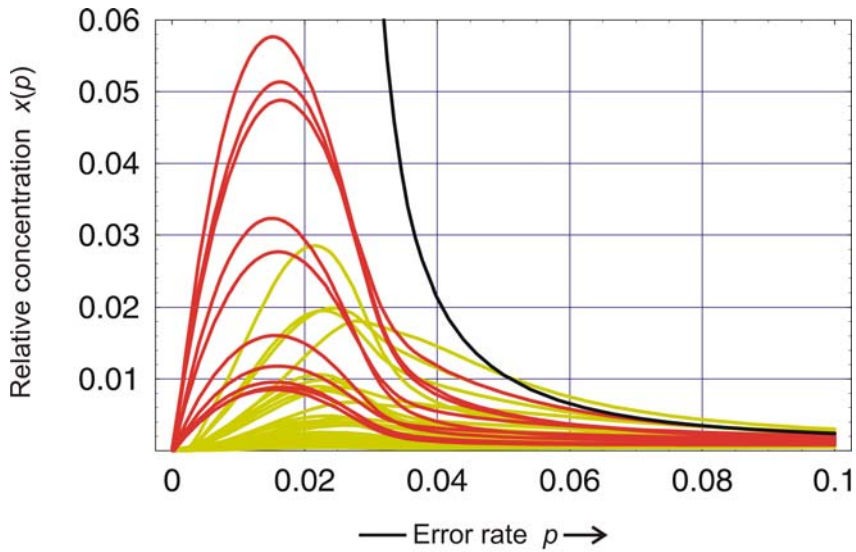




Fitness landscapes showing error thresholds







Error threshold: Individual sequences

n = 10, σ = 2 and d = 0, 1.0, 1.85

Evolution of RNA molecules based on Qβ phage

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

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

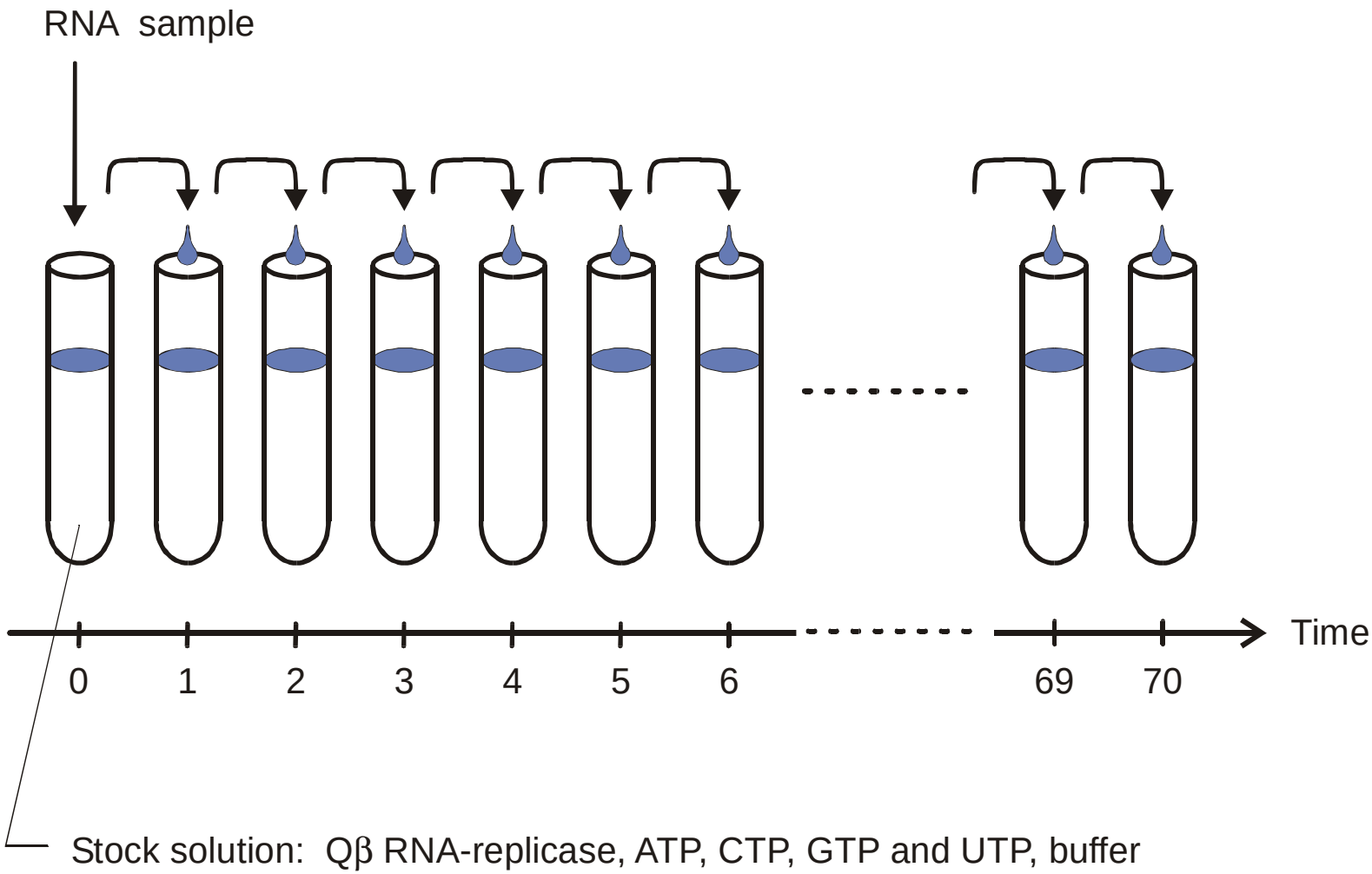
C.K.Biebricher, ***Darwinian selection of self-replicating RNA molecules***. Evolutionary Biology **16** (1983), 1-52

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

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

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

F.Öhlenschlager, 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



Anwendung der seriellen Überimpfungstechnik auf RNA-Evolution in Reagenzglas

Evolutionary design of RNA molecules

A.D. Ellington, 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

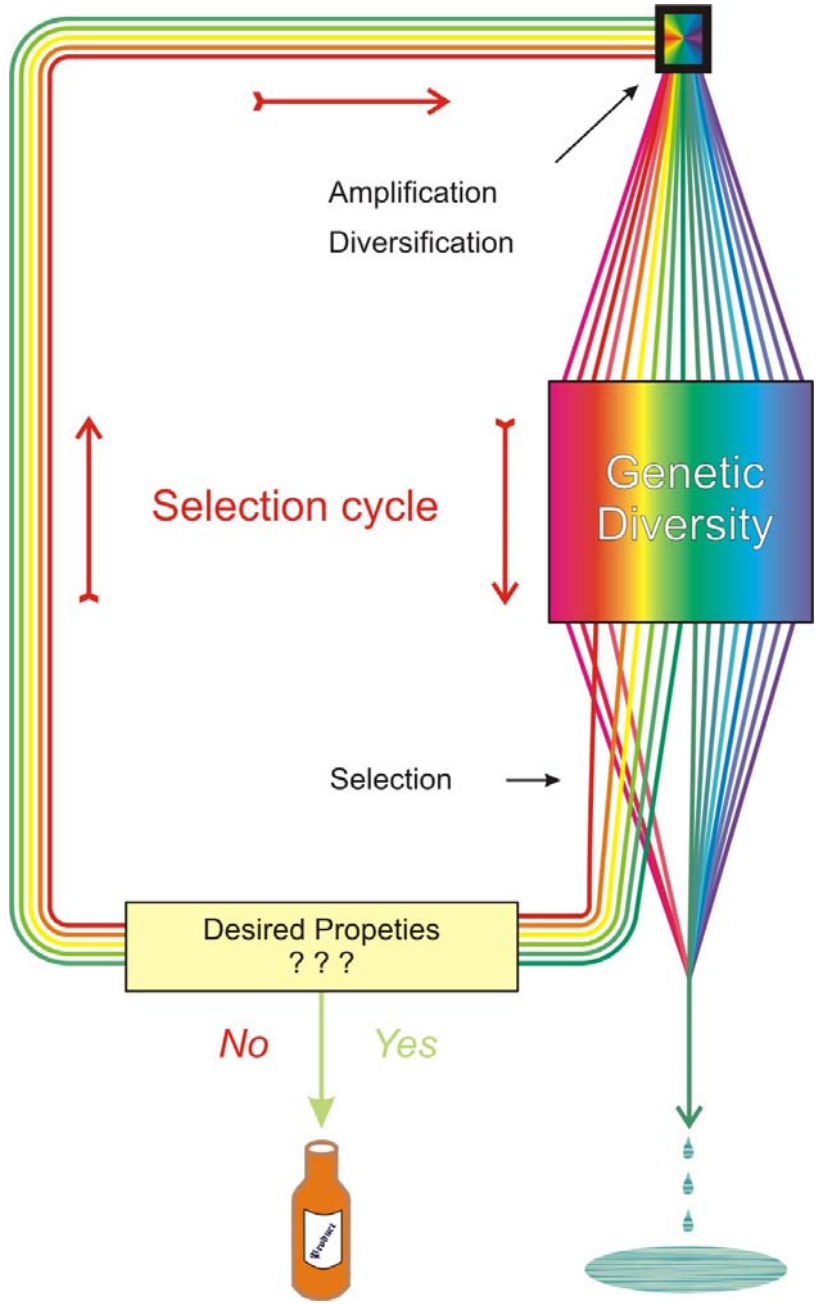
D.P. Bartel, J.W. Szostak, ***Isolation of new ribozymes from a large pool of random sequences***. Science **261** (1993), 1411-1418

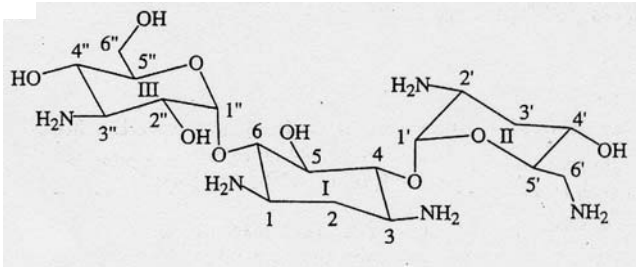
R.D. Jenison, S.C. Gill, A. Pardi, B. Poliski, ***High-resolution molecular discrimination by RNA***. Science **263** (1994), 1425-1429

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

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

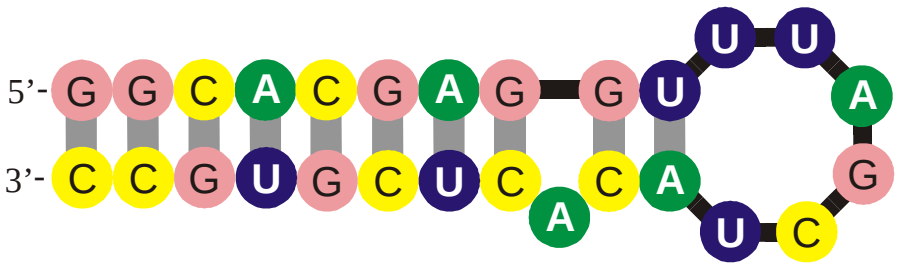
An example of ‘artificial selection’ with RNA molecules or ‘breeding’ of biomolecules





tobramycin

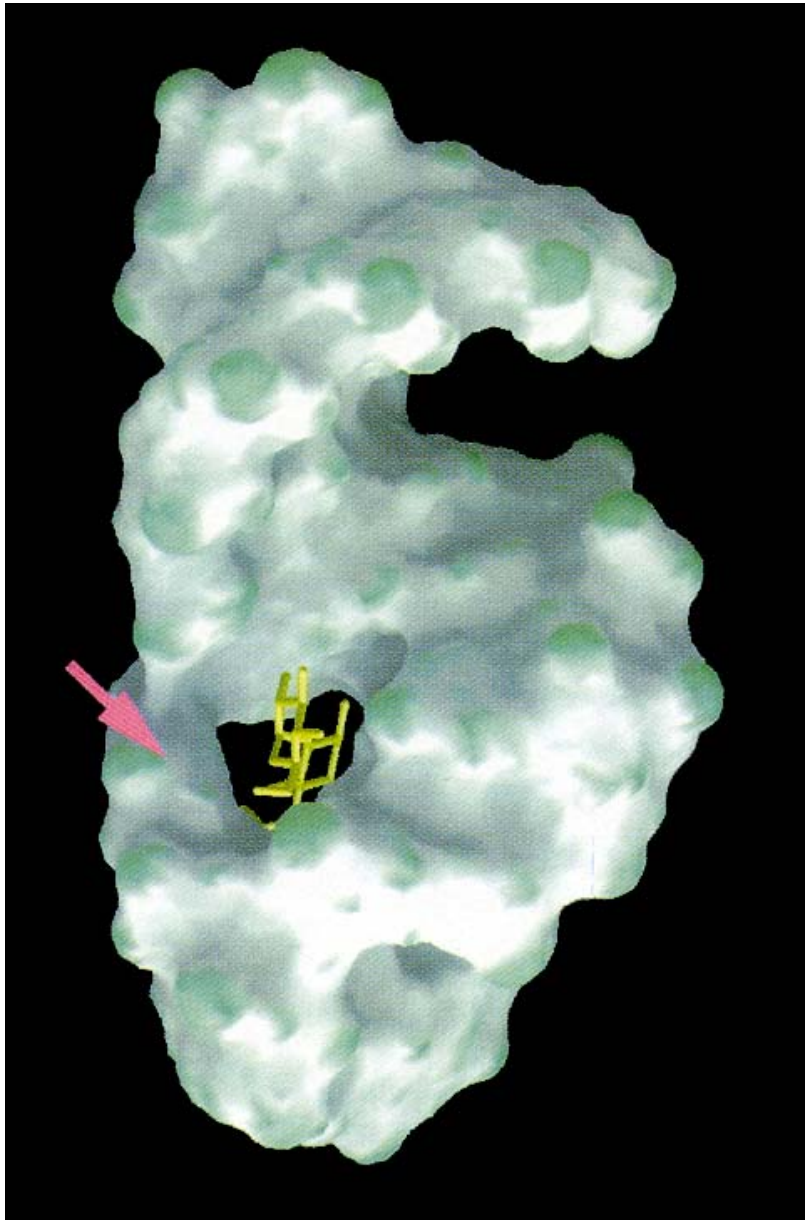
5’- G G C A C G A G G U U U A G C U A C A C U C G U G C C -3’



RNA aptamer

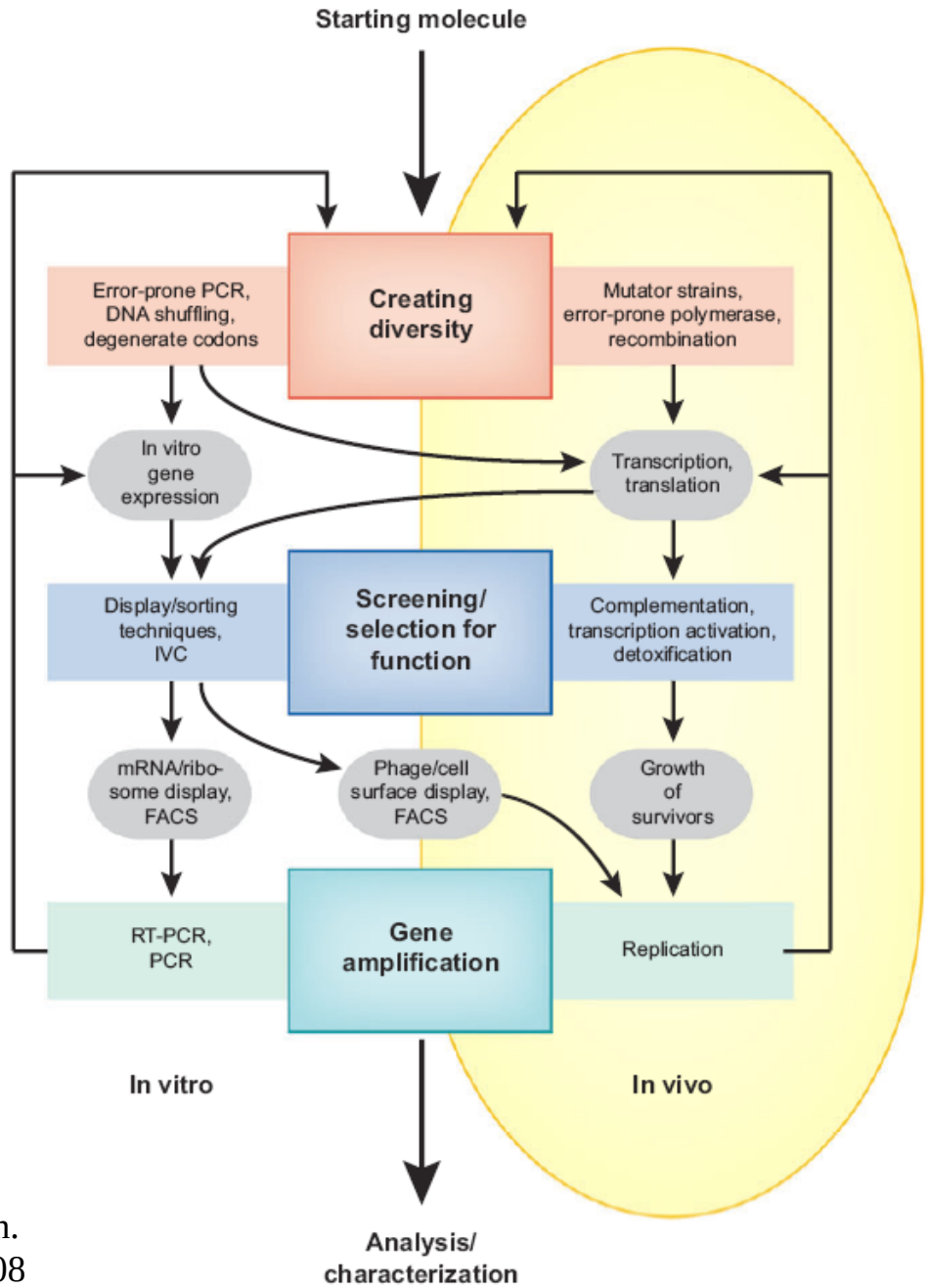
Formation of secondary structure of the tobramycin binding RNA aptamer with $K_D = 9$ nM

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)



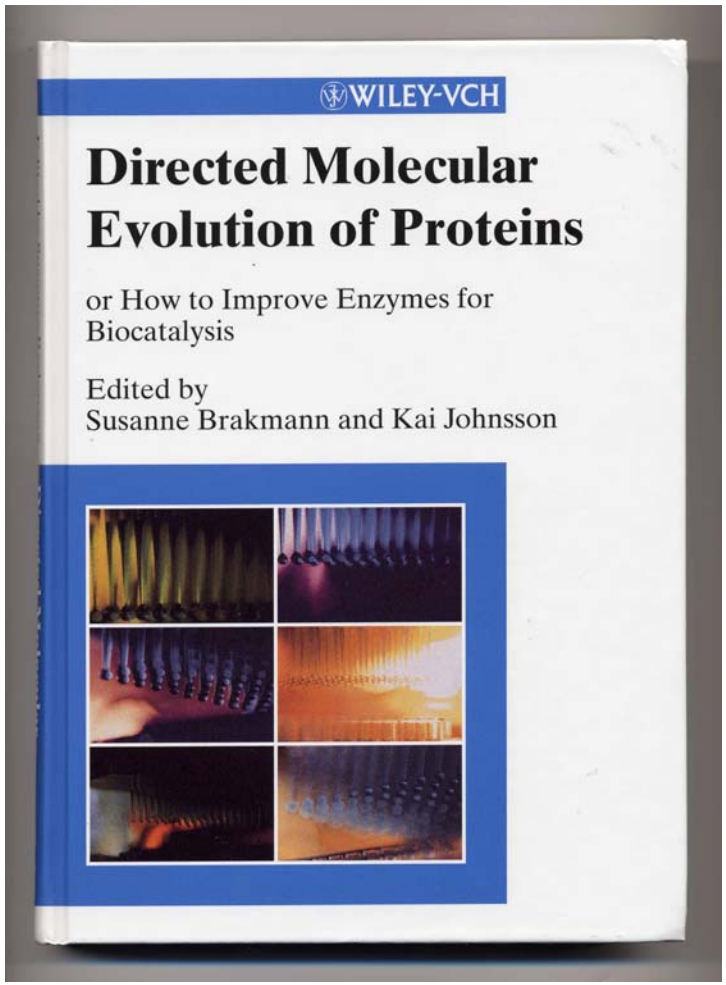
The three-dimensional structure of the tobramycin aptamer complex

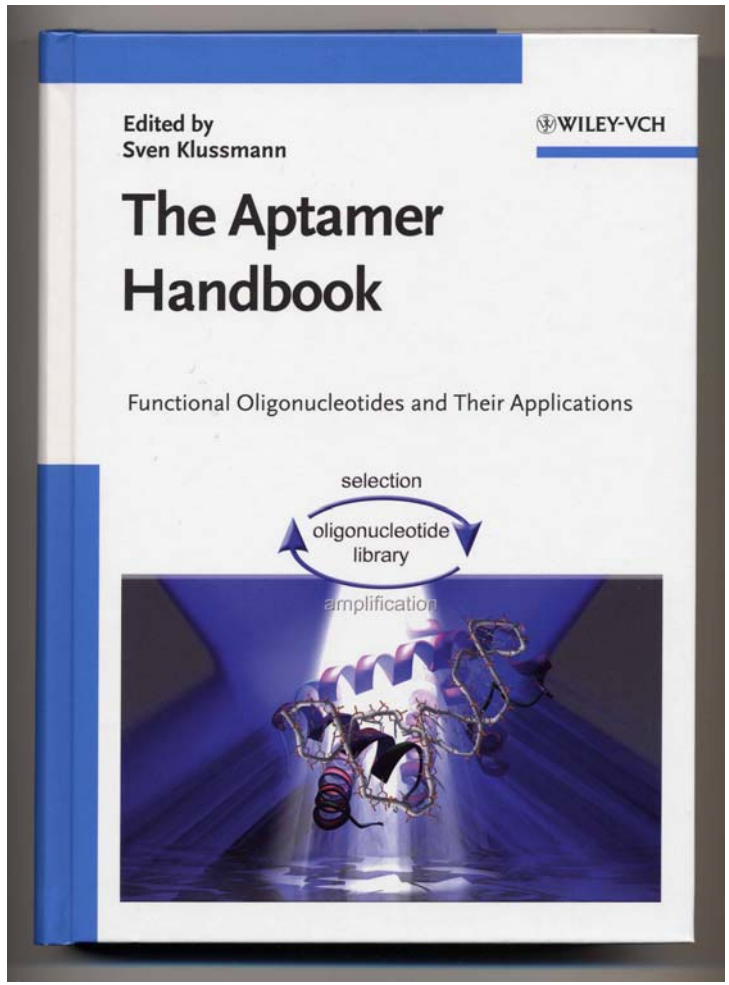
L. Jiang, A. K. Suri, R. Fiala, D. J. Patel, Chemistry & Biology **4**:35-50 (1997)



Schematic overview of the principal processes, strategies, and techniques of directed evolution. Today, numerous experimental methods are available to perform the fundamental processes of true Darwinian evolution (*central boxes*) in the laboratory, either in vivo within microorganisms or entirely in vitro in the test tube. Arrows indicate possible routes for connecting individual evolutionary steps. Abbreviations: PCR, polymerase chain reaction; RT-PCR, reverse transcription PCR; IVC, in vitro compartmentalization; FACS, fluorescence-activated cell sorting.

Christian Jäckel, Peter Kast, and Donald Hilvert.
Protein design by directed evolution.
Annu.Rev.Biophys. **37**:153-173, 2008





Application of molecular evolution to problems in biotechnology

Artificial evolution in biotechnology and pharmcology

G.F. Joyce. 2004. Directed evolution of nucleic acid enzymes. *Annu.Rev.Biochem*. **73**:791-836.

C. Jäckel, P. Kast, and D. Hilvert. 2008. Protein design by directed evolution. *Annu.Rev.Biophys*. **37**:153-173.

S.J. Wrenn and P.B. Harbury. 2007. Chemical evolution as a tool for molrcular discovery. *Annu.Rev.Biochem*. **76**:331-349.

Results from replication kinetics and molecular evolution in laboratory experiments:

• Evolutionary optimization does not require cells and occurs in molecular systems too.

• *In vitro* evolution allows for production of molecules for predefined purposes and gave rise to a branch of biotechnology.

• Novel antiviral strategies were developed from known molecular mechanisms of virus evolution.

1. Minimum free energy structures of RNA
2. Suboptimal structures of RNA
3. Kinetic folding and RNA switches
4. Chemistry of Darwinian evolution
5. **Consequences of neutrality**
6. Evolutionary optimization of RNA structure

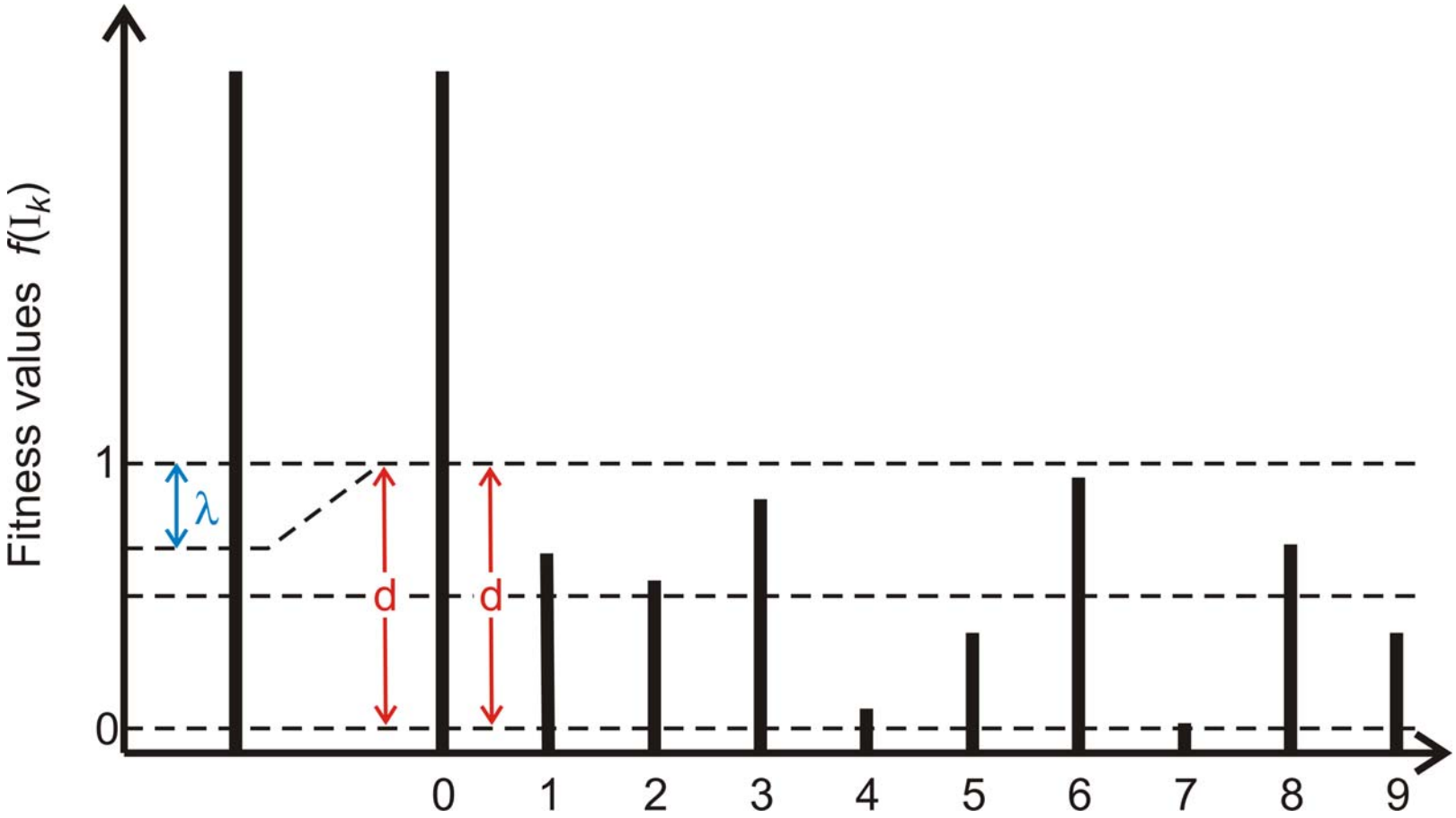
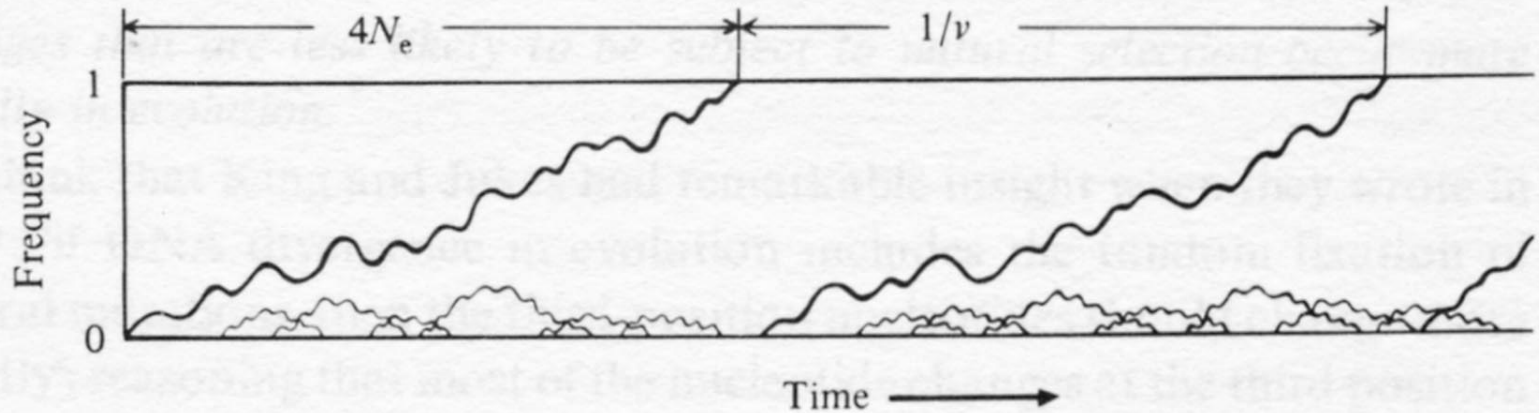

Fitness values $f(I_k)$
1
0
λ
d
d
0
1
2
3
4
5
6
7
8
9

Fig. 3.1. Behavior of mutant genes following their appearance in a finite population. Courses of change in the frequencies of mutants destined to fixation are depicted by thick paths. N_e stands for the effective population size and v is the mutation rate.



Motoo Kimura

Is the Kimura scenario correct for frequent mutations?

Bulletin of Mathematical Biology Vol. 50, No. 6, pp. 635–660, 1988.
Printed in Great Britain.

0092-8240/88$3.00 + 0.00
Pergamon Press plc
Society for Mathematical Biology

STATIONARY MUTANT DISTRIBUTIONS AND EVOLUTIONARY OPTIMIZATION

■ PETER SCHUSTER and JÖRG SWETINA
Institut für theoretische Chemie
und Strahlenchemie der Universität Wien,
Währingerstraße 17,
A 1090 Wien,
Austria

Molecular evolution is modelled by erroneous replication of binary sequences. We show how the selection of two species of equal or almost equal selective value is influenced by its nearest neighbours in sequence space. In the case of perfect neutrality and sufficiently small error rates we find that the Hamming distance between the species determines selection. As the error rate increases the fitness parameters of neighbouring species become more and more important. In the case of almost neutral sequences we observe a critical replication accuracy at which a drastic change in the "quasispecies", in the stationary mutant distribution occurs. Thus, in frequently mutating populations fitness turns out to be an ensemble property rather than an attribute of the individual.

In addition we investigate the time dependence of the mean excess production as a function of initial conditions. Although it is optimized under most conditions, cases can be found which are characterized by decrease or non-monotonous change in mean excess productions.

1. Introduction. Recent data from populations of RNA viruses provided direct evidence for vast sequence heterogeneity (Domingo *et al.*, 1987). The origin of this diversity is not yet completely known. It may be caused by the low replication accuracy of the polymerizing enzyme, commonly a virus specific, RNA dependent RNA synthetase, or it may be the result of a high degree of selective neutrality of polynucleotide sequences. Eventually, both factors contribute to the heterogeneity observed. Indeed, mutations occur much more frequently than previously assumed in microbiology. They are by no means rare events and hence, neither the methods of conventional population genetics (Ewens, 1979) nor the neutral theory (Kimura, 1983) can be applied to these virus populations. Selectively neutral variants may be close with respect to Hamming distance and then the commonly made assumption that the mutation backflow from the mutants to the wilde type is negligible does not apply.

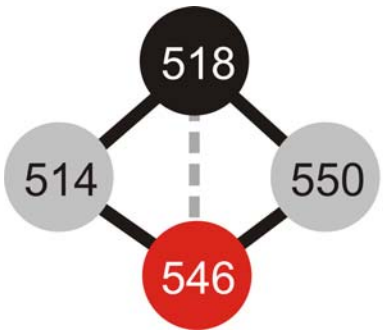
A kinetic theory of polynucleotide evolution which was developed during the past 15 years (Eigen, 1971; 1985; Eigen and Schuster, 1979; Eigen *et al.*, 1987; Schuster, 1986); Schuster and Sigmund, 1985) treats correct replication and mutation as parallel reactions within one and the same reaction network

Neutral network

$\lambda = 0.01,\ s = 367$

$\mathbf{d_H = 1}$

$$\lim_{p\to 0} x_1(p) = x_2(p) = 0.5$$



Neutral network

$\lambda = 0.01,\ s = 877$

$\mathbf{d_H = 2}$

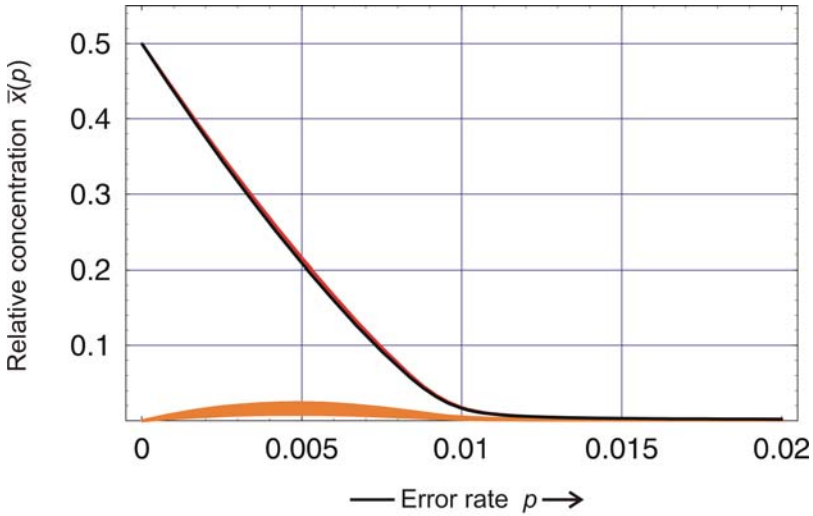
$$\lim_{p\to 0} x_1(p) = a$$

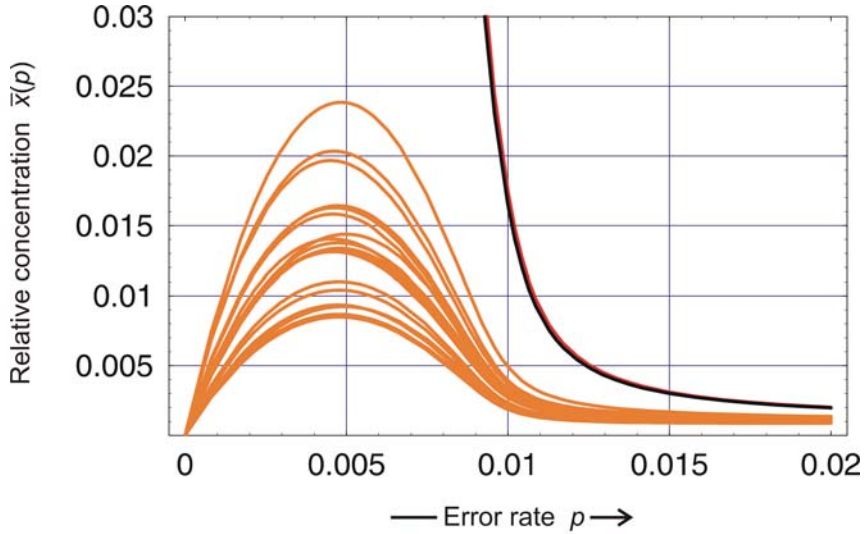
$$\lim_{p\to 0} x_2(p) = 1 - a$$

$\mathbf{d_H \geq 3}$

random fixation in the sense of Motoo Kimura

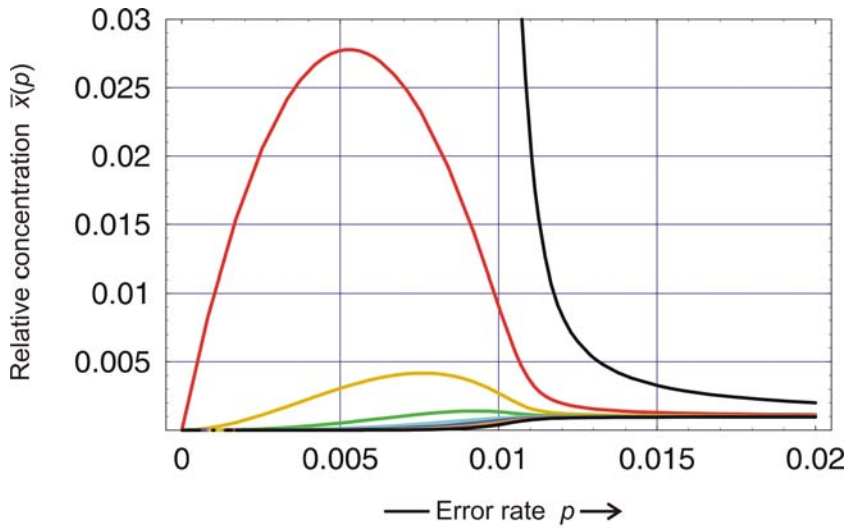
Pairs of genotypes in neutral replication networks





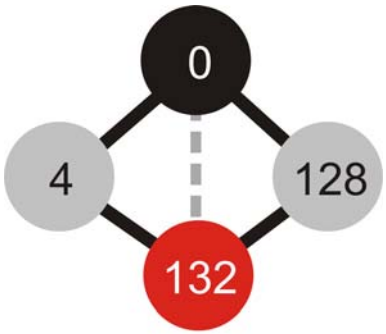
Neutral network

$\lambda = 0.01,\ s = 367$



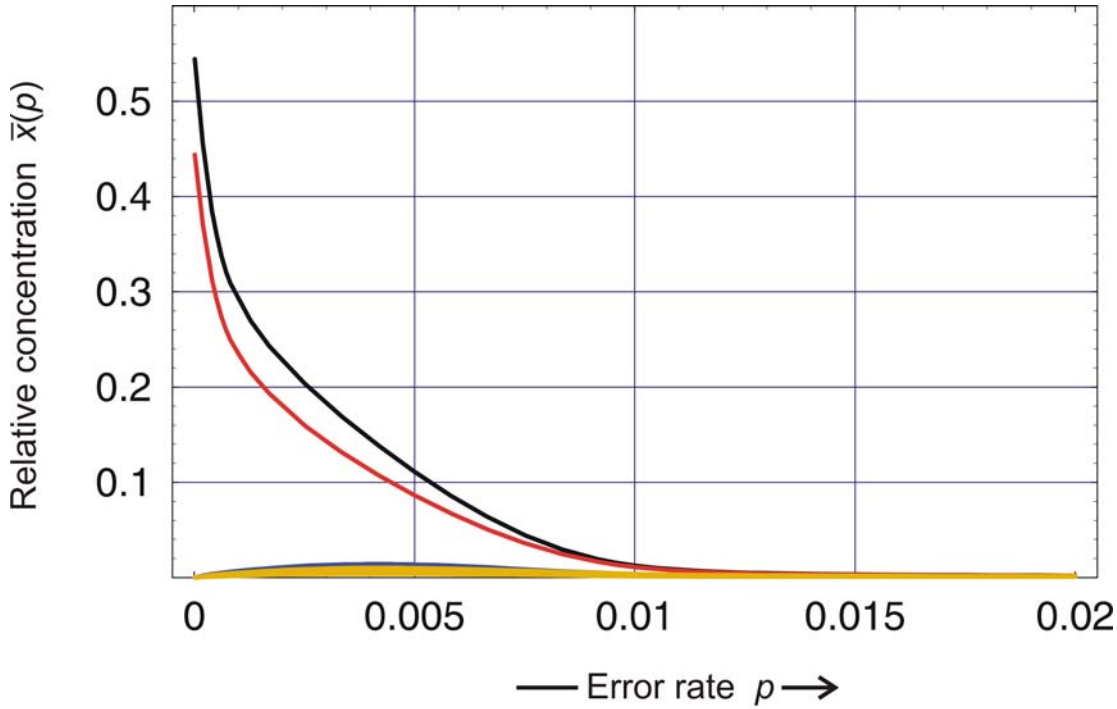
Neutral network: Individual sequences

$n = 10,\ \sigma = 1.1,\ d = 1.0$



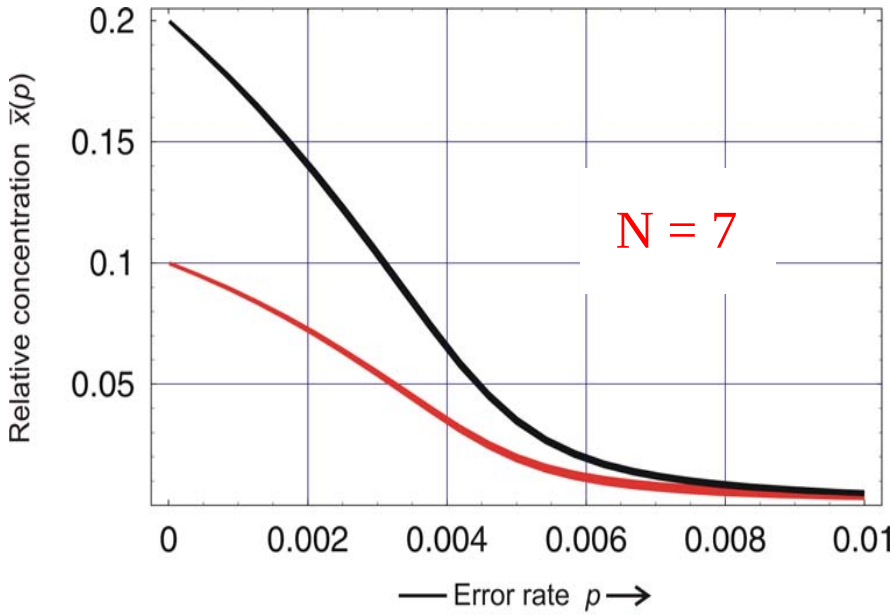
Neutral network

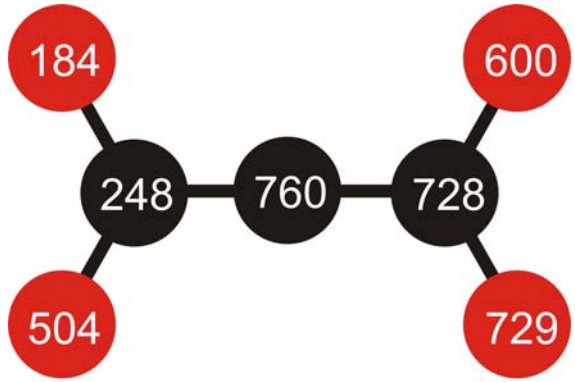
$\lambda = 0.01,\ s = 877$



Neutral network: Individual sequences

n = 10, σ = 1.1, d = 1.0

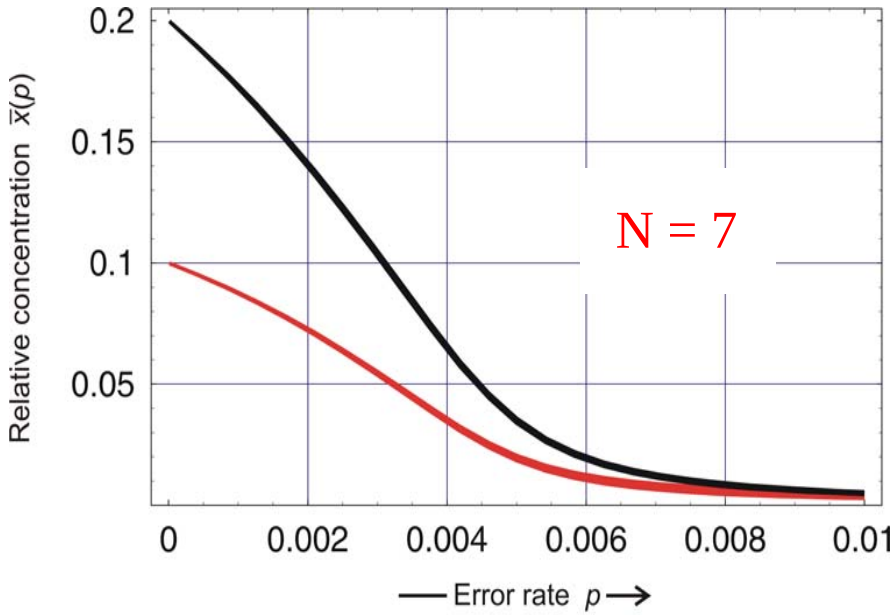


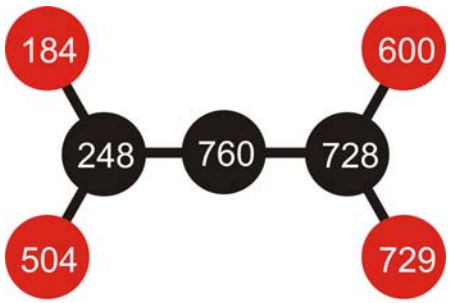


Neutral network

$\lambda = 0.10$, s = 229

Neutral networks with increasing λ: $\lambda = 0.10$, s = 229





Neutral network

$\lambda = 0.10,\ s = 229$

Perturbation matrix W

$$W = \begin{pmatrix} f & 0 & \varepsilon & 0 & 0 & 0 & 0 \\ 0 & f & \varepsilon & 0 & 0 & 0 & 0 \\ \varepsilon & \varepsilon & f & \varepsilon & 0 & 0 & 0 \\ 0 & 0 & \varepsilon & f & \varepsilon & 0 & 0 \\ 0 & 0 & 0 & \varepsilon & f & \varepsilon & \varepsilon \\ 0 & 0 & 0 & 0 & \varepsilon & f & 0 \\ 0 & 0 & 0 & 0 & \varepsilon & 0 & f \end{pmatrix}$$

Eigenvalues of W

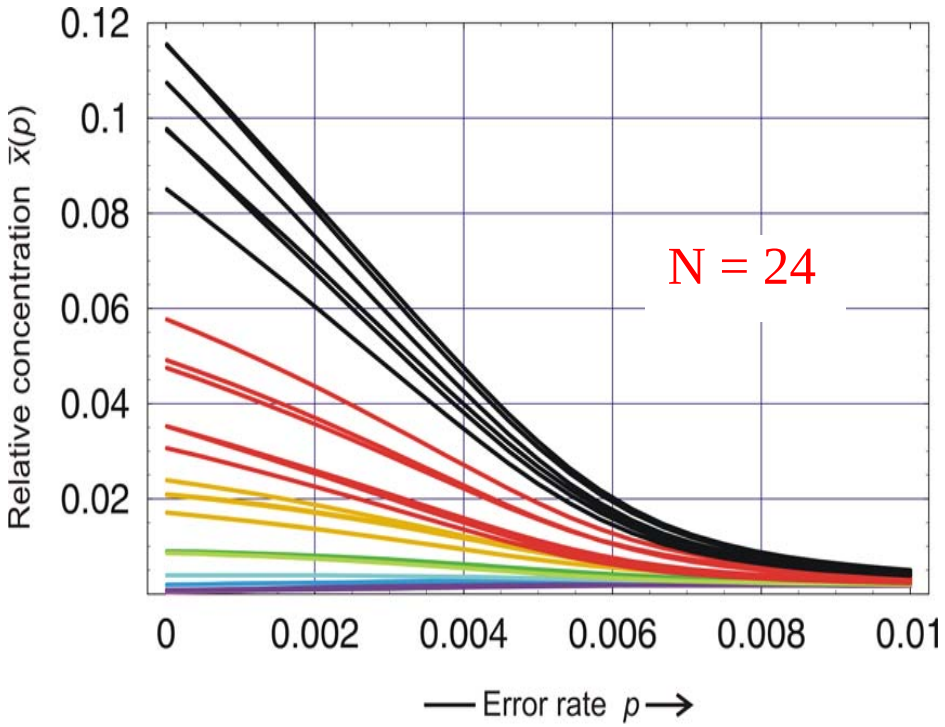
$$\lambda_0 = f + 2\varepsilon ,$$

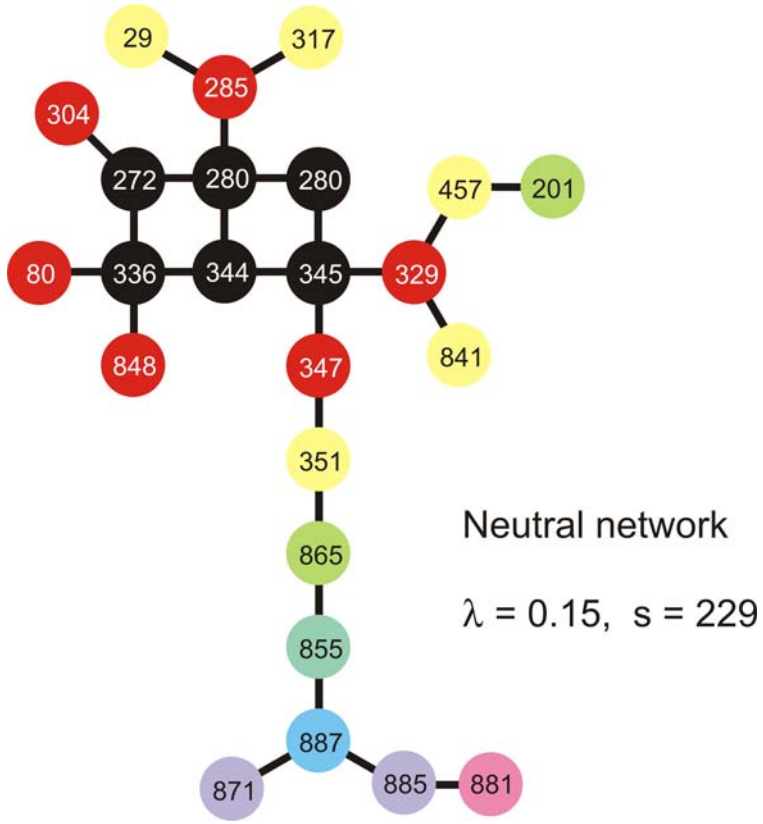
$$\lambda_1 = f + \sqrt{2}\varepsilon ,$$
$$\lambda_{2,3,4} = f ,$$
$$\lambda_5 = f - \sqrt{2}\varepsilon ,$$
$$\lambda_6 = f - 2\varepsilon .$$

Largest eigenvector of W

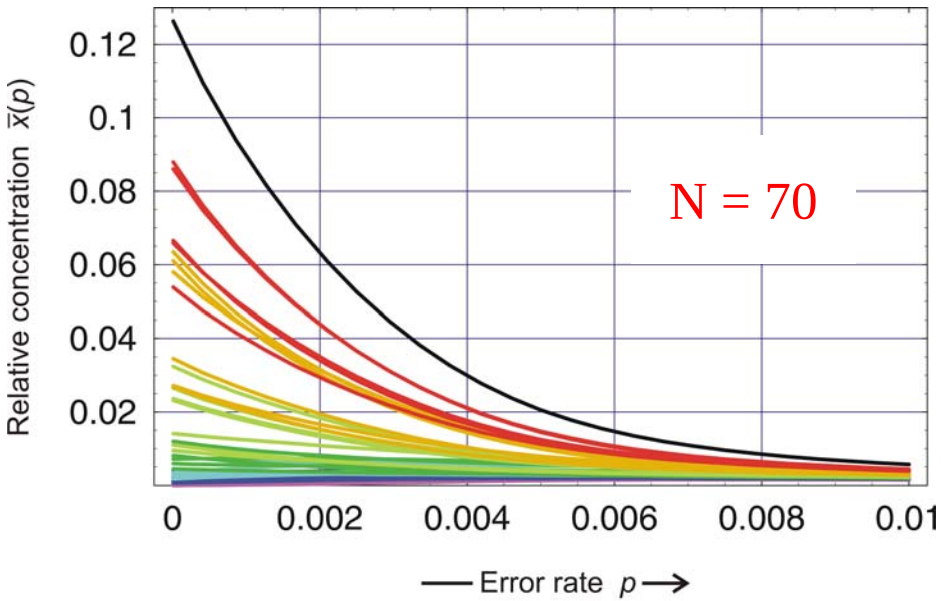
$$\xi_0 = (0.1, 0.1, 0.2, 0.2, 0.2, 0.1, 0.1) .$$

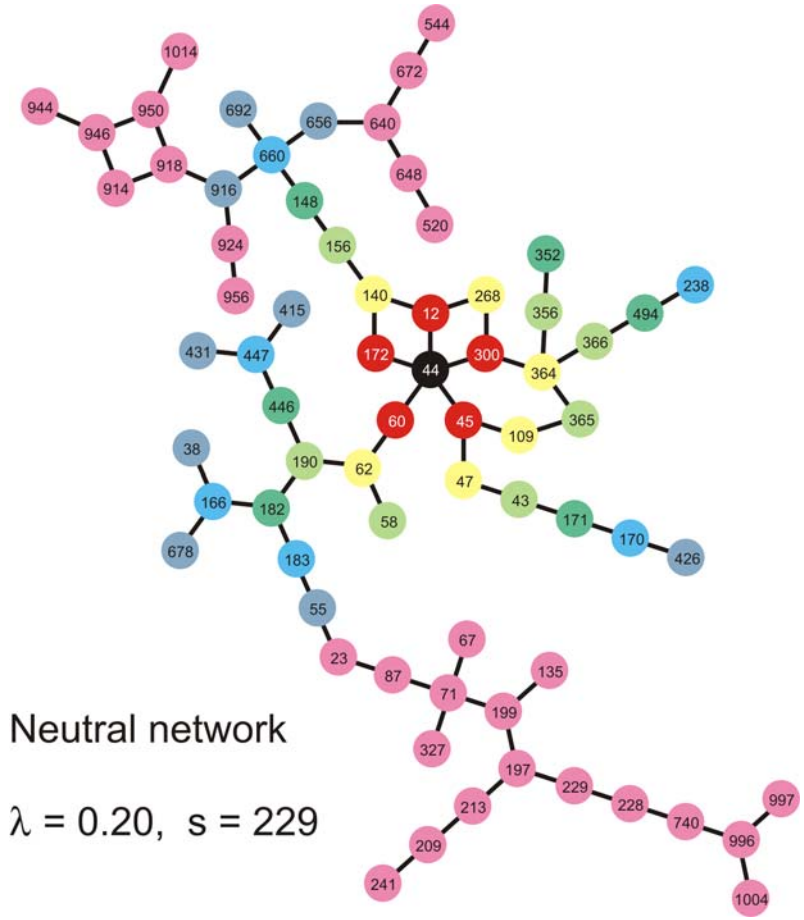
Neutral networks with increasing λ: λ = 0.10, s = 229





Neutral networks with increasing λ: λ = 0.15, s = 229





Neutral networks with increasing λ: λ = 0.20, s = 229

Results from replication kinetics and RNA neutral networks:

• RNA sequences with Hamming distance d = 1 and d = 2 form strongly coupled replication ensembles. For d > 2 random drift in the sense of Kimura's theory occurs.

• Direct evidence that neutrality is increasing the repertoire of structures and properties in populations.

• Implication for virus replication in infected hosts.

Neutrality in evolution

Charles Darwin: *„ ... neutrality might exist ..."*

Motoo Kimura: *„ ... neutrality is unaviodable and represents the main reason for changes in genotypes and leads to molecular phylogeny ..."*

Current view: *„ ... neutrality is essential for successful optimization on rugged landscapes ..."*

Proposed view: *„ ... neutrality provides the genetic reservoir in the rare and frequent mutation scenario ..."*

1. Minimum free energy structures of RNA
2. Suboptimal structures of RNA
3. Kinetic folding and RNA switches
4. Chemistry of Darwinian evolution
5. Consequences of neutrality
6. **Evolutionary optimization of RNA structure**

Evolution *in silico*

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

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

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

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

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

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

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

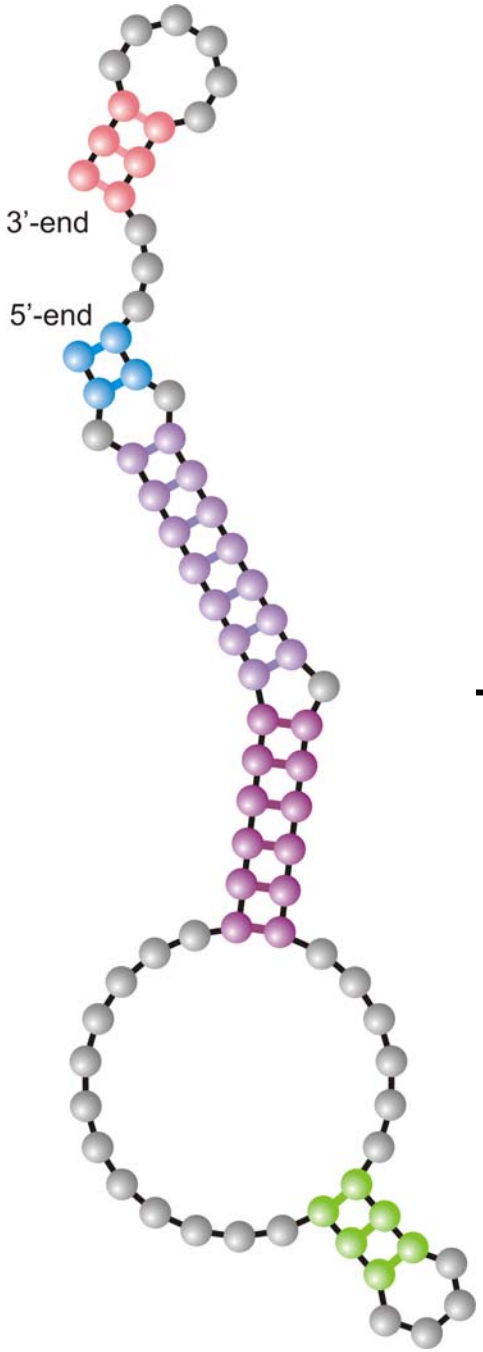
We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level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

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

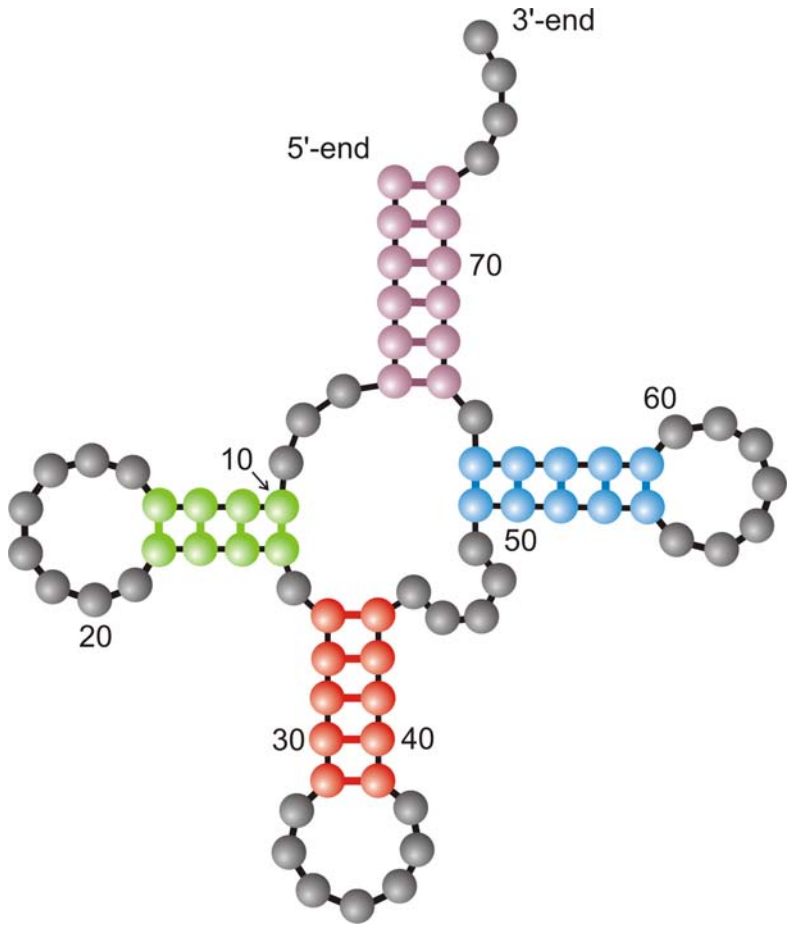
An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces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

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

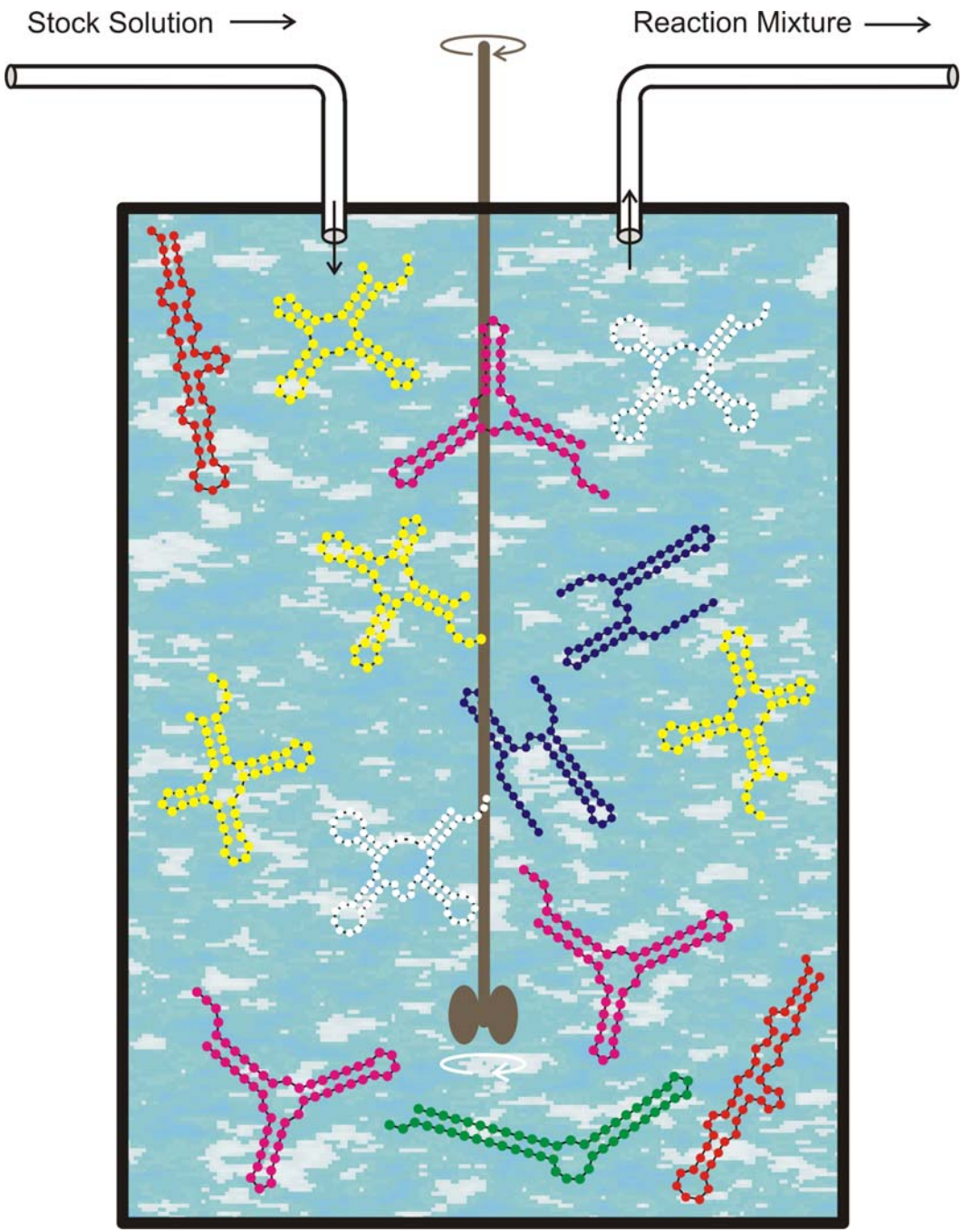
Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.



Structure of randomly chosen initial sequence



Phenylalanyl-tRNA as target structure



Replication rate constant

(Fitness):

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

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

Selection pressure:

The population size,

N = # RNA moleucles,

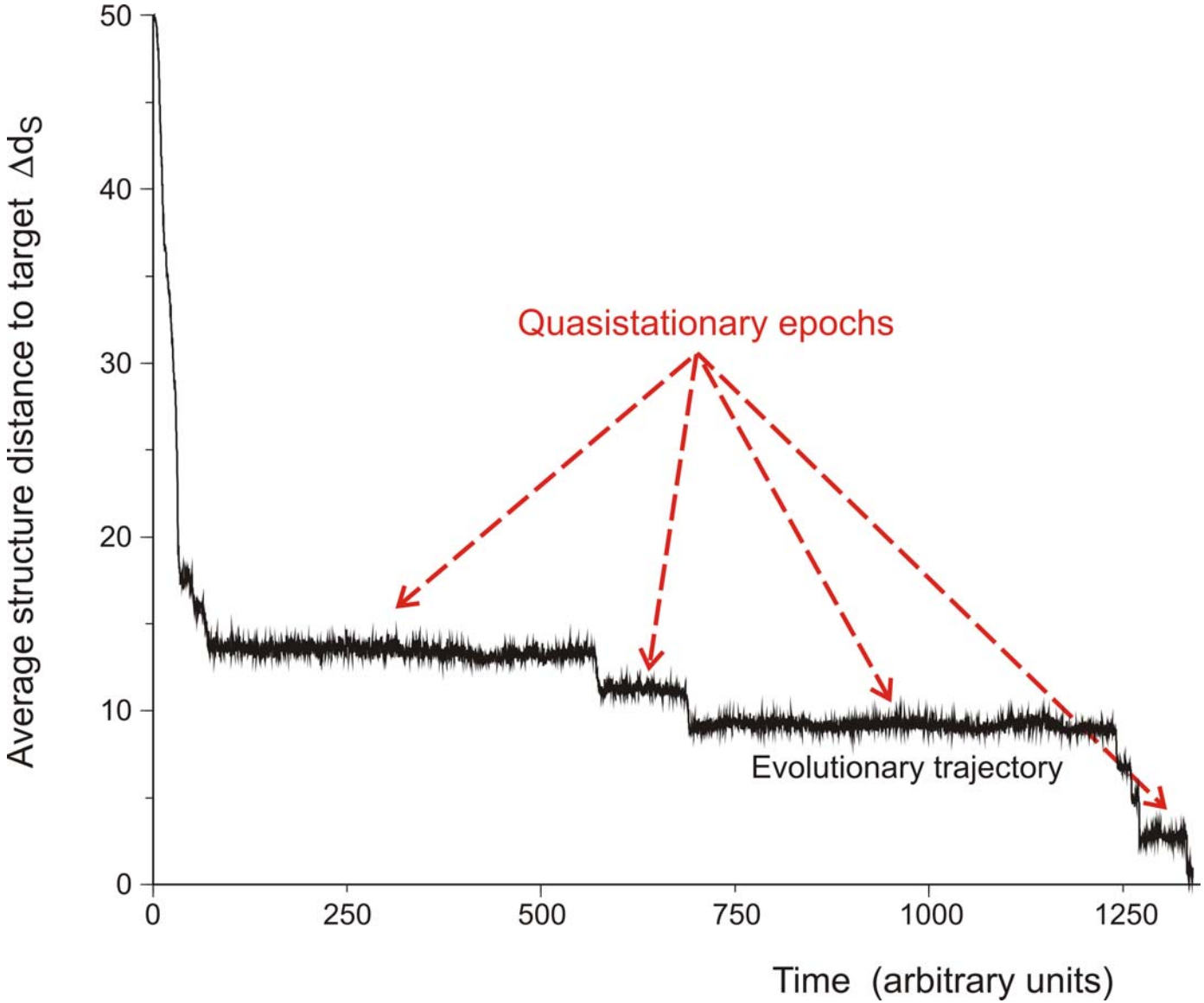
is determined by the flux:

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

Mutation rate:

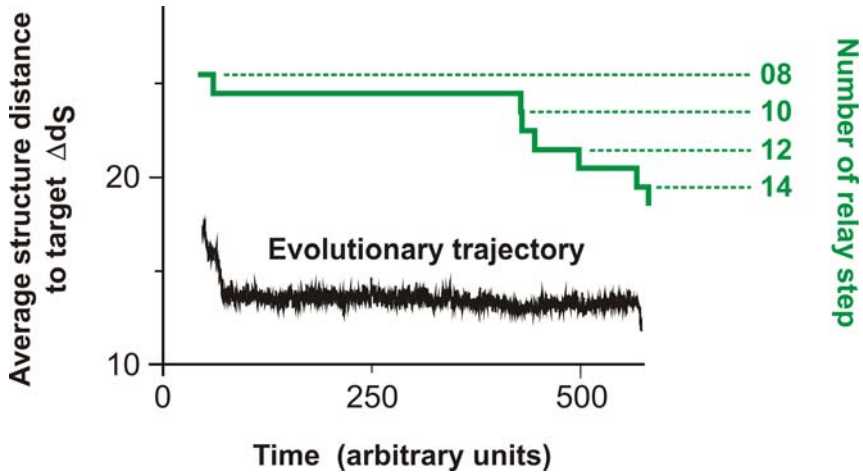
p = 0.001 / Nucleotide × Replication

The flow reactor as a device for studying the evolution of molecules ***in vitro*** and ***in silico***.



In silico optimization in the flow reactor: Evolutionary Trajectory

28 neutral point mutations during a long quasi-stationary epoch



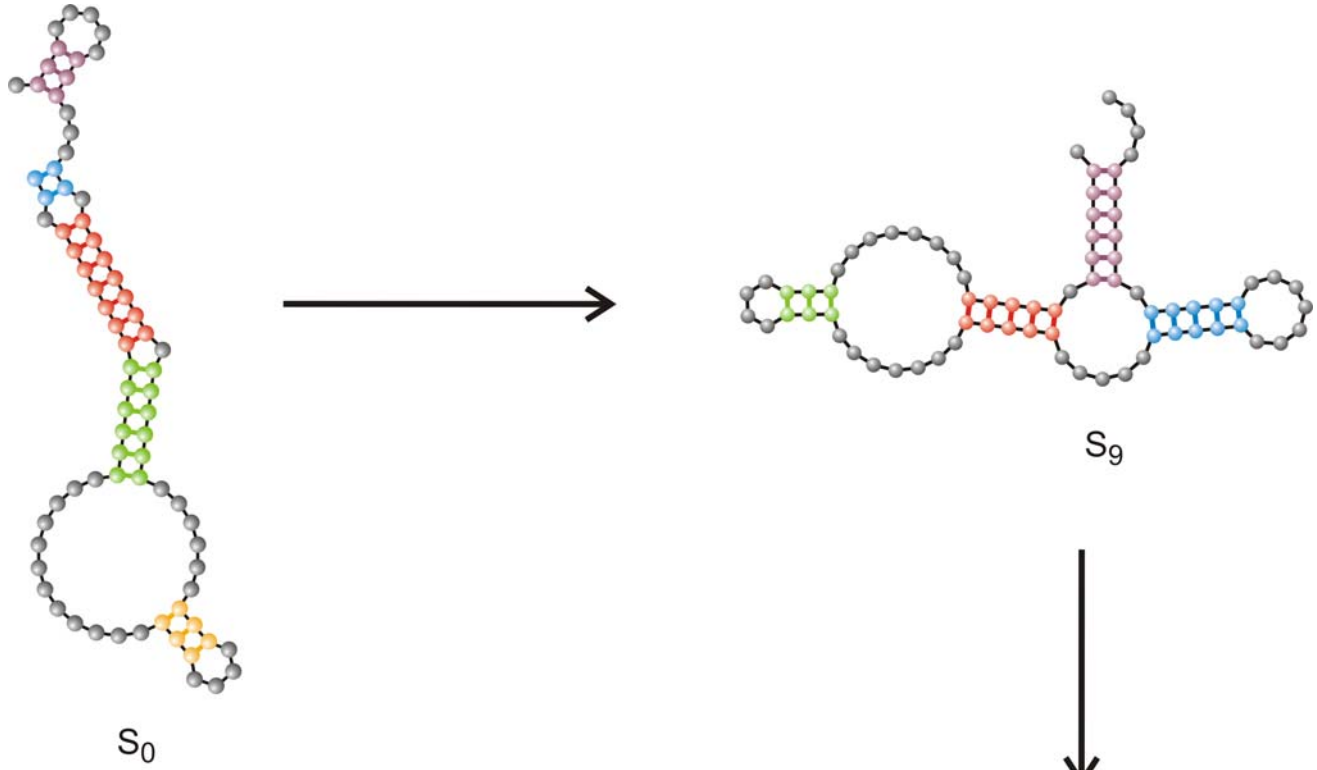

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))).......)))))).....((((.......))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA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

Transition inducing point mutations change the molecular structure

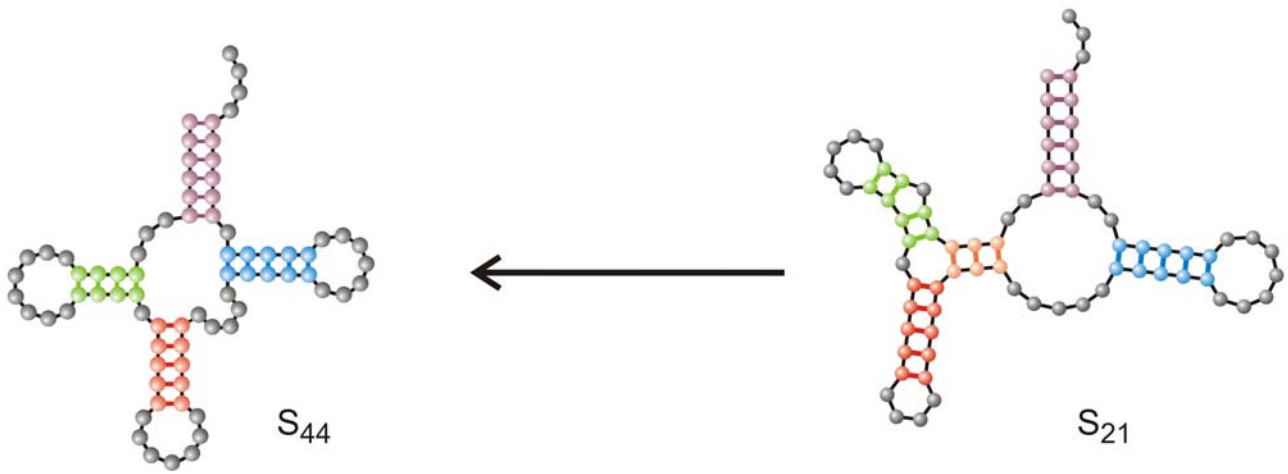
Neutral point mutations leave the molecular structure unchanged

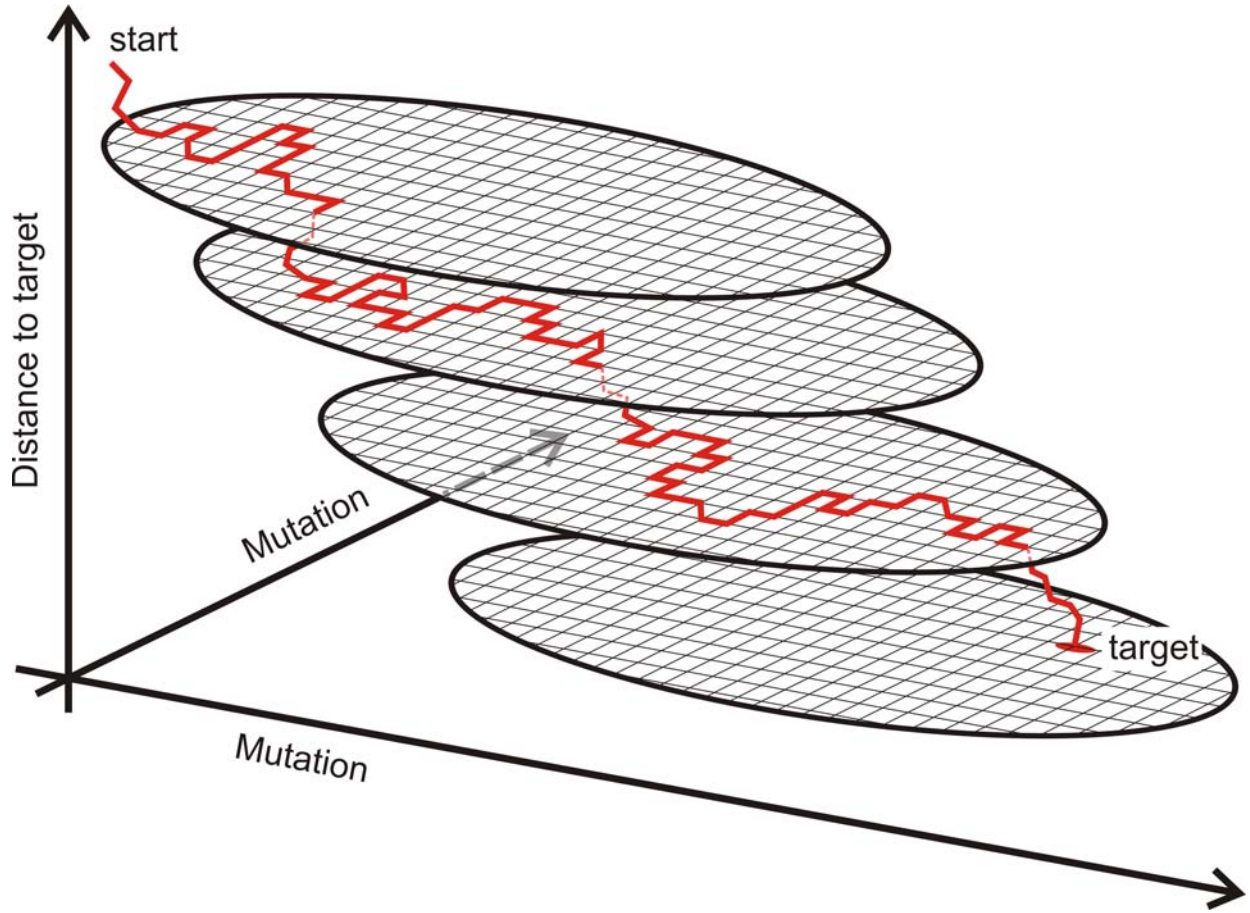
Neutral genotype evolution during phenotypic stasis

Randomly chosen
initial structure



Phenylalanyl-tRNA
as target structure





A sketch of optimization on neutral networks

Results from *in silico* simulation of RNA evolution:

• Evolutionary optimization occurs on two time scales: Fast adaptive phases and random walk on neutral networks.

• Neutral networks are essential for searching sequence space.

Acknowledgement of support

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

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

Universität Wien

Coworkers

Universität Wien

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

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

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**, Universität Wien, AT

Jan Cupal, **Stefan Bernhart**, **Lukas Endler, Ulrike Langhammer**, **Rainer Machne, Ulrike Mückstein**, **Hakim Tafer, Thomas Taylor,** Universität Wien, AT

INSTITUTE OF PHYSICS PUBLISHING REPORTS ON PROGRESS IN PHYSICS

Rep. Prog. Phys. **69** (2006) 1419–1477 doi:10.1088/0034-4885/69/5/R04

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

Peter Schuster[1,2]

[1]Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, A-1090 Vienna, Austria
[2]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

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

Web-Page for further information:

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