RNA Structures

Stability, Folding and the Role of Hydrogen Bonding and Protons

Peter Schuster

Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Schloß Ringberg, 05.03.2002



The chemical formula of RNA consisting of nucleobases, ribose rings, phosphate groups, and sodium counterions

Structural Constraints and Hydrogen Bonding in RNA

Single stranded RNA molecules form structures, which combine double-helical stacking (A-type) regions with loops and metal ion (Mg²) coordinated centers.



The three-dimensional structure of a short double helical stack



Canonical Watson-Crick base pairs:

cytosine – guanine uracil – adenine

W.Saenger, Principles of Nucleic Acid Structure, Springer, Berlin 1984



Canonical Watson-Crick base-pair

Wobble base-pairs

Wobble base pairs in RNA double-helical stacks



S.A. Benner *et al.*, Reading the palimpsest: Contemporary biochemical data and the RNA world. In: R.F.Gesteland and J.F.Atkins, eds. The RNA World, pp.27-70. CSHL Press, 1993



Classification of purinepyrimidine base pairs









A·G N7-N1, amino-carbonyl



amino-N1

A.G N7-amino, amino-N3



A-A N1-amino, symmetric

H



G-G N1-carbonyl, symmetric



H

н

нин

R

A·A N7-amino,

symmetric

R

N-H

н

H

A·A N1-amino,

N7-amino

R

G-G N3-amino, symmetric



G-G N7-N1,

carbonyl-amino



G•G N1-carbonyl, N7-amino

Classification of purine-purine base pairs



4-carbonyl-N3

Classification of pyrimidinepyrimidine base pairs



General classification of base pairs

N.B.Leontis and E. Westhof, RNA 7:499-512 (2001)



Stacking of heterocyclic aromatic molecules without sugar-phosphate backbone

Example: N6,N9-dimethyl adenine, D. Pörschke and F. Eggers, Eur.J.Biochem. 26:490-498 (1972)



Stacking of RNA single strands

Example: poly-**A**, D.Pörschke. Elementary steps of base recognition and helix-coil transitions in nucleic acids. In: I.Pecht and R.Rigler, eds. Chemical Relaxation in Molecular Biology, pp.191-218. Springer-Verlag, Berlin 1977.



Three-dimensional structure of phenylalanyl-transfer-RNA

RNA Secondary Structures and their Properties

RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots. Secondary structures are folding intermediates in the formation of full three-dimensional structures.

D.Thirumalai, N.Lee, S.A.Woodson, and D.K.Klimov. *Annu.Rev.Phys.Chem.* **52**:751-762 (2001)



Symbolic Notation

Definition of the secondary structure of phenylalanyl-tRNA



Minimum Free Energy Structure

Suboptimal Structures

Kinetic Structures

Different notions of RNA structure

RNA Minimum Free Energy Structures

Efficient algorithms based on dynamical programming are available for computation of secondary structures for given sequences. Inverse folding algorithms compute sequences for given secondary structures.

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

Vienna RNA Package: http://www.tbi.univie.ac.at (includes inverse folding, suboptimal structures, kinetic folding, etc.)

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

UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG







Mapping from sequence space into phenotype space and into free energies



Point mutations as moves in sequence space



$$\mathbf{G}_{\mathbf{k}} = \mathbf{m}^{-1}(\mathbf{S}_{\mathbf{k}}) \mid \mathbf{O}\mathbf{I}_{j} \mid \mathbf{m}(\mathbf{I}_{j}) = \mathbf{S}_{\mathbf{k}} \mathbf{Q}$$

$$\lambda_j = 12 / 27$$
, $\bar{\lambda}_k = \frac{\hat{O}_{j \in |G_k|} \hat{J}(k)}{|G_k|}$

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

Alphabet Size _: AUGC $i = 4$		cr
_	2	0.5
$\lambda_k > \lambda_{cr} \dots$ Network G_k is connected	3	0.4226
$\bar{\lambda}_k < \lambda_{cr} \dots$ Network G_k is not connected	4	0.3700

Mean degree of neutrality and connectivity of neutral networks



A connected neutral network



A multi-component neutral network

Kinetic Folding of RNA at Elementary Step Resolution

The RNA folding process is resolved to base pair **closure**, base pair **cleavage** and base pair **shift**. The kinetic folding behavior is determined by computation of a sufficiently large ensemble of individual folding trajectories and taking an average over them. The folding behavior is illustrated by barrier trees showing the path of lowest energy between two local minima of free energy.

C.Flamm, W.Fontana, I.L.Hofacker and P.Schuster. RNA, 6:325-338 (2000)





Move set for elementary steps in kinetic RNA folding



Folding dynamics of the sequence **GGCCCUUUGGGGGGCCAGACCCCUAAAAAGGGUC**









Search for local minima in conformation space

Local minimum





Barrier tree of a sequence with two conformations

A ribozyme switch

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



The "hammerhead" ribozyme

The smallest known catalytically active RNA molecule



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



The sequence at the *intersection*:

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



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

S0092-8240(96)00089-4

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

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

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

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

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

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

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

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

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

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

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

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



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

P1 _ J1/2 P2 _ J2/ P1 _ P3 _ L3 _ P3 _ J3/4 _ P4 _ P2 _ J2/5 _ P5 _ L5 _ P5 _ J5/4 _ P4 _ P3 _ P3 _ P3 _ P3 _ P3 _ P3 _ P	A																
P P		P1	.11/	2 P2	.12/1	D1	Da	13	D2	12/4	D4	00	10/5 0	leb appende		00000	
P1 J12 P2 P3 L3 P3 P3 P4 P5 P5 P1 P4 L4 P4 L4 P4 L4 P4		-				Research ser			10	00/4	P4	-2	J2/5 P	5 L5	P5 J5/4	P4	
P1 J12 P2 P3 L3 P3 P1 P4 L4 P4 L4 P4 L4 P4	1.	038.0	10	[[20	-	30		40	1	50	60	22.200	70	80	Children and	
	AAA	CCAG	UCGGA	ACACI	AUCCG	ACUGGIC	ACCCC	nnnn	GGGGUG	GGGAG	UGCCUAG	AAGUG	GGU-AGG	ปิดดากการ	AGACCGC-	CUAGGCO	LIGP
	AAA	CCAGI	CGGA	ACACI	AUCCG	ACUGGIC	ACCCC	00000	GGGGUU	GGGAG	UGCCUAG	AAGUG	GGU-AGG	UCUUUU-U	AGACCGC-	CUAGGCO	LIG42
P1 J1/2 P2 P3 L3 P3 P1 P4 L4 P4 L4 P4 L4/2 P2	AAA	CCAGI	UCGGA	ACACI	AUCCO	ACUGGIC	ACCCC		GGGGUG	GGGAG	UGCCUAG	AAGUG	GGU-AGG	ucuuuu-u	AGACCAA-	CUAGGCO	LIG40B
AAACCABUUCGAAACACCAUUAGACUGG C CCUUUUUG C GUUGG CAUUUGU AAG CUGGU CAUUUUU AAG CUGGU CAUUUUU AAG CUAG CU	AAA	CCAGI	JCGGA	ACACI	AUUAG	ACUGGIC	ACCCC		GGGGUG	GGGAG	UGCCUAG	AAGUG	GGU-GGG	UCUUUU-U	AGACCAA-	CUAGGCO	LIG40A
P1 J12 P2 P3 L3 P3 P1 P4 L4 P4	AAA	CCAGI	JCGGA	ACACO	AUUAG	ACUGGIC	ACCCC	UUUUU	GGGGUG	GGGAG	UGCCUAG	AGGUG	G GIU - G G G		AGACCAA-	CUAGGCO	LIG38
AAACCAAUUCGAAACCAUUAGACUUGGACGCCUUCUUGGGGGGGG	AAA	CCAGI	JCGGA	ACACO	AUUAG	ACUGGIC	ACCCC	UUUUU	GGGGUG	GGGAG	UGCCUAG	AGGUG	GGU-GGG	UCUUUUUCU	A GACCAA-	CUAGGCC	LIG36
P1 1/2 P2 P3 L3 P3 P1 P4 L4 P4	AAA	CCAGI	JCGGA	ACACO	AUUAG	ACUGGC	ACCCC	u u u u u	GGGGUG	GGGAG	UUCCUAG	AGGUG	GGU-GGG	UCUUUUCU.	AGACICAA -	CUAGGA	LIG32
AAACCASUUCGAALCACUUAGACUUGGACGCCUUCUUGGCGUUGGGUUGG	AAA	CCAGI	JCGGA	ACACO	AUUAG	ACUGGIC	ACCCC	UUUUU	GGGGUG	GGGAG	UUCCUAG	AGGUG	GGU-GAG	UCUUUUCU.	AGACUAA-	CUAGGA	LIG30
An ac charues an a charue a chur a ch	AAA	CCAGI	ICGGA	ACACC	AUUAG	ACUGGIC	ACCCC	uccul	GGGGUG	GGGAG	UUCCUAG	AGGUG	GGU-GAG	UCUUUUCU.	AGACUAA-	CUAGGA	LIG28
AAACCAAUUGGAALCACUUUGAACUUGGAUGGUUCUUUGGGUUGGUUAGGUUA	AAA	CCAGI	ICGGA	ACACC	AUTIAG	ACUGGIC	ACCCC	UCCUC	GGGGUG	GGGAG	UUCCUAG	AGGUG	GGU-GAG	ccuuuucu.	AGGCUAA-	CUAGGA	LIG26
A A A C A BUCC GA LA C A C LA ULA GA C UG GA C C UC C ULG C G LUG C ULG C G LA GUUG GU C A A GUUG C ULUU C ULA GC ULA A C C ALGA C LUG C A A C C AU ULA GA C ULA C C AU ULA GA C ULG C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C C ULUU C ULA GA C ULA C C ALGA C ALGA C ULU C ALGA C ULUU C ULA GA C ULA C C ALGA C ALGA C ULU C ULU C ALGA C ULUU C ULA GA C ULA C C ALGA C ALGA C ULU U C ULA GA C ULA C C ALGA C ALGA C ULU U C ULA GA C ULA C C ALGA C ALGA C ULU U C ULU C ALGA C ULUU U U ULU ALGA C ULU C ALGA C ULUU C ULU C ALGA C ULUU U ULU ALGA C ULU C ALGA C ULUU C ULU C	AAA	CCAGE	JCGGA	ACACO	AUUAG	ACUGGIC	ACGCC	recue	GCGUG	GGGAG	UULCCUAG	AGGUG	GGU-GAG	ccuuuucu.	AGGCUAA-	CUAGGA	LIG24
A A A C A BU C G A L C A C A U U A G C U U G C G C C U C U U G C G C G G A U U G G C A G U G G D G G U - G A G C U U U U C U A G C U A A - G A C A U U C A U U A G C U A A - G A C A U U G A A C A U C G A U U A A C U U U A G C U A A - G A C A U U A G C U A A - G A C A U U A G C U A A - G A C A U U G A U C A U U A A C U U U A G C U A A - G A C A U U G A U A A - G A C A U U G A U C A U U A A C U U U A G C U A A - G A C A U U G A U A A - G A C A U U G A U A A C U U U C A A C A U U G A U A A C U U U A G C U A A C A U U G A U A A C U U U A G C U A A C A U U G A U A A C U U U A G C U U A G C U A A C A U U G A U A A C U U U C A A C U U U A A C	AAA	CCAGU	JCGGA	ACACO	AUUAG	ACUGGIC	ACGCC	ICCUG	GCGUG	GGGAG	UUGCUAG	AGGUG	GGU-GAG	CCUUUUCU.	AGGCUAA-	CUAGICA	LIG22
A A A C A BU C G A LU LA A C LU LU G A LU G G A C GU C LU G C GU C G G A BU LU G G C A G GU G G LU - A C C LU LU LU LU A G C LU A G C LU A C C LU LU CU A G C LU A - G A C C A LU A A C A BU C A LU A G A LU G G C C G C G C LU C LU G C G G G G A BU LU G G C A G GU G G U - G G C C LU LU LU LU A G C LU A G C LU A - G A C C A LU A A C A BU C A LU A G A LU G G C C G C C C LU C LU G C G G G G G A DU LU G G C A G G A G G A G U G G C C LU LU LU LU A G C LU A G C LU A - G A C C A LU A A C C A BU C C A LU A G A LU G C C C C C C LU C LU G C G C G G G A DU LU G C G A G G A G G A G G A G G A C G C LU LU LU LU A G C LU A G C LU A - G C C C A LU A A C C A BU L A G C C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C C LU LU C LU A G C LU A - G C LU LU C LU A G C LU A - G C LU LU C LU A G C LU A - G A C A C C LU LU C LU A G C LU A - G A C A C C LU LU C LU A G C LU A - G A C A C A C C LU LU C LU A G C LU A - G A C A C A C C LU LU C LU A G C LU A - G A C A C A C C LU LU C LU A G C LU A - G A C A C A C C LU LU C LU A G C LU A - G A C A C A C C LU LU C LU A G C	AAA	CCAGI	JCGGA	ACACO	AUUAG	ACUGGIC	ACGCC	JCCUG	GCGUG	GGGAG	UUGGUCG	AGGUG	G GIU - GAG	CCUUUUCU.	AGGCUAA-	UACCA	LIG20
A A A C C A G U C G A AL C A C U U U A G A C U G G G G G C U C U U G C G C G G G A G U U G C G C G G G A G U U G C G C U U U U U U U A G C U A A - G G C C U U U U U U A G C U A A - G G C C C A U U A G C U A A - G G C C U U U U U U A G C U A A - G G C C C A U U A G C U A A - G G C C U U U U U U A G C U A A - G G C C C A U U A G C U A A - G G C C U U U U U U A G C U A A - G G C C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U U U U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G G C C C A U U A G C U A A - G G C C C A U A A - G G C C A U U A G C U A A - G G C C C A U A A - G C U C U C C A U A A C C A U U A G C U A A - G G C C C A U A A - G C C U C U C C A U A A C G C C U U U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U C C A U A A A C C A U U A G C U C C A U U A A C C U U U U A G C U A A - G G C C A U U A G C U A A - G G C C A U U A G C U A A - G C C C A U U A G C U C C A U A A A C C A U U A A C C A U U A A C C U U U A A C C U U A G C U A A - G G C C A U U U A G C U C C A U U A A C U U A G C U C C A U U A A C U U A G C U A A - G G C C A U U U A G C U A A - G G C C A U U U A G C U A A - G G C C A U U U A G C U A A - G G C C A U U U A G C U A A - G G C C A U U U A G C U A A - G G C C A U U U A G C U A A - G G C C A U U U A A C U U A G C U A A - G G C C A U U U A A C U U A G C U A A - G G C C A U U U A A C U U A G C U A A - G G C C A U U U A A C U U G G A U A A C A U U C A U U C A U G C U A A - G G A C A U U C A U G C U A A - G G A	AAA	CCAGU	ICGGA	ACACC	AUUAG	ACUGGG	ACGCCI	JCCUG	GCGUC	GGGAG	UUGGUCG	AGGUG	GGU - GAG	ссицинси	AGGCUAA-	GACCA	LIGIS
A A C C A G U C G A LUA C A LUA G A C U G G C C C C U C U G C G C C G C G G A G U U G G C G A G G LU G U U U U U LUA G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U U U C U A G C U A A - G C C C A U U U U C U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C U U U U C U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C C A U U A G C U A A - G C C U U U U C U A G C U A A - G C C C A U U A G C U A A - G C C U U U U C U A G C U A A - G C C C A U U A G C U A A - G C C U U U U C U A G C U A A - G C C C A U U A G C U A - G C C C A U U A G C U A - G C C C A U U A G C U A - G C C C A U U A G C U A - G C C C A U U A G C U A - G G C C A U U A G C U A - G G C C A U U A G C U A - G G C C A U U A G C U A - G G C C A U U A G C U A - G G C C A U U A G C U A - G G C C A U U A G C U A - G G C C A U U A G C U A - G G C C A U U U A G C U A - G G C C A U U A G C U A - G G C C A U U U A G C U A - G G C C A U U U A G C U A - G G C C A U U A - G G C C A U U A - G G C C A U U A - G G C C A U U A - G G C C A U U A - G G C C A U U A - G G C C A U U A - G G C C A U U U A G C U A - G G C C A U U A - G G C C A U U U A G C U A - G G C C A U U A - G G C C A U U A - G C U C C A U U A G C U A A - G C U U A - G A C C A U U G C A U G C A A C C C U U U U U A G C U A - G G C C A U U A - G G C C A U U U A G C U A - G G C C A U U U A - G C U A - G A C C A U U U A - G C U A - G A C C A U U U A - G C U A - G A U C C A U U A - G A U C C A U U C A U A C C U U C C A U G C U A A - G G C C A U U C A U C C A U U C C A U G C U A A - G A C C A U U C A U G C U U A - G A U C C A U U C A U G C U U A - G A U C C A U G C U	AAA	CCAGU	JCGGA	ACACC	AUUAG	ACUGGG	ACGCCI	JCCUG	GCGUC	GGGAG	UUGGGCG	AGGUG	GGU-GAG	CCUUUUCU	AGGCUAA-	GCCCA	LIGIA
AAACCAGUUGGAA UAGAAUUAGAAUUAGACUGGGCGCGCUUCUUGGCGGGGGAGUUGGGCGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAG	AAA	CCAGU	CGGA	ACACC	AUUAG.	ACUGGG	CCGCCI	JCCDG	GCGGC	GGGAG	UUGGGCG	AGGUG	GGU-GAG	CCUUUUCU	AGGCUAA -	GCCCA	LIG12
AAACCAGUUCGAAUUGGALUUGGAUUGGCCGCCUUCUUGGGGGGGGGAGUUGGGCGAGGGAGG	AAAC	CCAGL	CGGA	AULAICC	AUUAG	ACUGGG	CCGCCI	JCCUG	GCGGC	GGGAGI	JUGGGCG	AGGUA	GGU-GAG	ccuuuucu	AGGCUAA-	GCCCA	LIGIO
AAACCAGUUCGAAUUAAAAUUGGGCCGCUUCUUGGGGGGGG	AAA	CAGE	CGGA	AUCCC	AUUAG	ACUGGG	CCGCCI	JCCUG	GCGGC	GGGAGI	JUGGGCG	AGGGA	GGU-GAG	ccuuuucu	AGGCUAA-	GCCCA	LIG8
AAACCAAUUG GALUUCALUUAGALUUAGALUUG GCCUCCUUCUUG GGGGGGGGALUUG GCUAGGGAACAGCCUUUUCUAGGCUAA - GGCCCA Lude GAACCAAUUG GALUUCALUUAGALUUG GCCUCCUUCUG GGGGGGGALUUG GCUAGGGAACAGCCUUUUCUAGGCUAA - GGCCCA Lude GAACCAAUUG GALUUCCAUUAGALUUG GCCUCCUUCUG GGGGGGGGALUUG GCUAGGGAACAGCCUUUUCUAGGCUAA - GGCCCA LUde GAACCAAUUG GALUUCCAUUAGALUUG GCCUUCUUG GGGGGGGGALUUG GCUAGGGAACAGCUUUUCUAGGCUAA - GGCCCA HOV GAACCAAUUG GALUUCCAUUAGALUUG GCCUUCUUG GGGGGGGGALUUG GCUAGGGAACAGCUUUUCUAGGCUAA - GGCCCA HOV GAACCAUUG GALUUCCAUUAGALUUG GCCUUCUUG GGGGGGGGAAUUG GCUAGGGAACAGCUUUUCUAGGCUAA - GGCCCA HOV GAACCAUUG GALUUCGAUUAGALUUG GCCUUCUUG GGGGGGGGAAUUG GCUAGGGAACAGCUUUUCUAGGCUAA - GGCCCA HOV GAACCAUUG GALUUCGAUUAGALUUG GCCUUCUUG GGGGGGGGAAUUG GCUAGGGAACAGCUUUUCUAGGCUAA - GGCCCA HOV GAACAUUC GALUUCGAUUAGALUUG GCCUUCUUG GGGGGGGGAAUUG GCUAGGGAACAGCUUUCCUAGGCUAA - GGCCA HOV GAACAUUC GALUUCGAUUAGALUG GCCUUCUUCGCGGGGGGGAAUUG GCUAGGGAACAGCUUUCCUAGGCUAA - GGACCA HOV GAACAUUC GALUUCGAUUAGALUG GCCUUCUUCGCGGGGGGGAAUUG GCUAGGGAACAGCUUUCCUAGGCUAA - GGACCA HOV GAACAUUC GALUUCGAUUAGALUG GCCUUCUUCGCGGCGGGGGAAUUG GCUAGGGAACAGCUUUCCUAGGCUAA - GGACCA HOV GAACAUUC GALUUCGAUUAGALUG GCCUUCUUCGCGGCGGGGGGAAUUG GCUAGGGAACAGCUUUCCUAGGCUAA - GGACCA HOV GAACAUUC GALUUG GAUUAGALUG GCCUUCUUCGCGGCGGGGGGAAUUG GCUAGGGAACAGCUUUCCCUAGGCUAA - GGACCA HOV GAACAUUC GALUUAGALUG GUCGCUUCUUCGCGGCGGGGGAAUUG GCUAGGGAACAGCUUUCCCUAGGCUAA - GGACCA HOV GAACAUUC GALUUAGALUG GUCGCUUCUUCGCGGCGGGGGAAUUG GCUAGGAACAGCUUUCCCUAGGCUAA - GGACCA HOV GAACAUUC GALUUAGALUG GUCGCUUCUUCGCGGCGGGGAAUGGGAACAGCUUUCCCUUCCCAUGGCUAA - GGACCA HOV GAACAUUC GGUUG GAUUAGALUG GUCGCUUCUUCGCGGCGGGGAAUUG GCAUGGGAACAGCUUUCCCUUCCC	AAA	CAGI	CGGA	AUCCO	AUUAG	ACUGGG	CCGCCC	JCCUG	GCGGC	GGGAGI	JUGGGCG	AGGGA	GGAAGAG	ccuuuucu	AGGCUAA-	GCCCA	LIG6
AACCABUEGGALUCCCAUUAGGCUGGCUCUUGGCGGGGGAGUUGGGCUAGGGAGGAACAGCCUUUUCUAGGCUAA-GGCCCA LIGA GAACCABUEGGALUCCCAUUAGGCUGGCCUCUUGGCGGGGGGGGGGGGG	GAAC	CCAGU	CGGA	AUCCC	AUUAG	ACUGGG	CCGCCC	ICCUG	GCGGGC	GGGAGI	JUGGGCIG	AGGGA	GGAACAG	ccuuuucu	AGGCUAA-C	GCCCA	LIG5
GAACCAGUCGGALUUCAGUUAGACUGGGCCGCCUCUUGGGGGGGGGG	GAAC	CAGU	CGGA	AUCCC	AUUAG	ACUGGG	CCGCCI	ICCUG	GCGGC	GGGAGI	INGGGCC	AGGGA	GGAACAG	CCUUUUCUI	AGGCUAA-K	JGCCCA	LIG4
GAACCAGUCGGACUCCCAUUAGACUGGGCGGCUUCUCGGGGGGGG	GAAC	CCAGU	CGGA	AUCCC	AUUAG	ACUGGG	CCGCCI	JCCUC	GCGGC	GGGAGI	IUGGGCU	AGGGA	GGAACAG	CCUUUUCUI	AGGCUAA-C	GCCCA	LIG2
GAACCAGUC GGACUCCCAUUAGACUGGGCGCGCCUCCUGGGGGGGGGG	GAAC	CCAGU	CGGA	cuccc	AUUAG	ACUGGG	CCGCCI	JCCUC	GCGGC	GGGAGI	JUGGGCU	AGGGA	GAACAG	ссининси	GCCUAA-(GCCCA	LIGT
GAACCAGUUC - GACUCCCAUUAGACUUGGGCCGCCUUCUUCGCGGGGGGGAGUUUGGGCUAGGGAACAGGCUUUUCCUAGGCUAA - GGCCCA HUV2 GGACCAUUC - GACUCCCAUUAGACUUGGUCCGCCUUCUUCGCGGGGGGGAGUUUGGCUAGGGAACAGCCUUUCCUAGGCUAA - GGCCCA HUV4 GGACCAUUC - GACUCCCAUUAGACUUGUCCGCCUUCUUCGCGGGCGGGAGUUUGGCUAAGGAACAGCCUUUCCUAGGCUAA - GGCCCA HUV4 GGACCAUUC - GACUCCGAUUAGACUUGUCCGCCUUCUUCGCGGCGGGGGAGUUUGGCUAAGGGAACAGCCUUUCCUUAGGCUAA - GGACCA HUV7 GGACCAUUC - GACUCGGAUUAGACUUGUCCGCCUUCUUCGCGGCCGGAGUUUGGCUAAGGGAACAGCCUUUCCUUAGGCUAA - GGACCA HUV7 GGACCAUUC - GACUCGGAUUAGACUUGUCCGCCUUCUUCGCGGCCGGAGUUUGGGCAAGGGAACAGCCUUUCCUUAGGCUAA - GGACCA HUV1 GGACCAUUC - GACUCGGAUUAGACUUGUCCGCCUUCUUCGCGGCCGGAGUUUGGGCAAGGGAACAGCCUUUCCUUGGGUAA - GGACCA HUV1 GGACCAUUC - GACUCGGAUUAGACUUGUCCGCCUUCUUCGCGGCCGGAGUUUGGGCAAGGGAACAGCCUUUCCUUUCGAUGGCUAA - GGACCA HUV1 GGACCAUUC - GGCUCGGAUUAGACUUGUCCGCCUUCUUCGCGGCCGGAGUUUGGCAUGGGAAGGAA	GAAC	CCAGU	CGGA	CUCCC	AUUAGI	ACUGGG	CCGCCI	CCUC	GCGGC	GGGAGI	JUGGGCU	AGGGA	GGAACAG	ccuuuldcu	GGCUAA-C	GCCCA	HOVA
GGACCAUUC - GACUCCCAUUAGACUGGGCCGCGCGGGAGUUGGGCUAAGGGAGGAACAGCCUUUCCUAGGCUAA-GGACCA HOWA GGACCAUUC - GACUCCCAUUAGACUGGUCCGCCUCCUCGCGGCGGGAGUUGGGCUAAGGGAACAGCCUUUCCCUAGGCUAA-GGACCA HOWA GGACCAUUC - GACUCCCAUUAGACUGGUCCGCCUCCUCCUCGCGGCGGGAGUUGGGCUAAGGGAACAGCCUUUCCCUAGGCUAA-GGACCA HOWA GGACCAUUC - GACUCCCAUUAGACUGGUCCGCCUCCUCCUCGCGGCGGGGAGUUGGGCUAAGGAACAGCCUUUCCCUAGGCUAA-GGACCA HOWA GGACCAUUC - GACUCCGGAUUAGACUGGUCCGCCUCCUCCUCGCGGCCGGGAGUUGGGCUAAGGAACAGCCUUUCCCUAGGCUAA-GGACCA HOWA GGACCAUUC - GACUCGGAUUAGACUGGUCCGCCUCCUCCUCGCGGCCCGGAGUUGGGCUAAGGAACAGCCUUUCCCUAGGCUAA-GGACCA HOWA GGACCAUUC - GACUCGGAUUAGACUGGUCCGCCUCCUCCUCGCGGCCCGGAGUUGGGAUGGAACAGCCUUUCCCUUUCCCGAUGGCUAA-GGACCA HOWA GGACCAUUC - GGCUCGGAUUAGACUGGUCCGCCUCCUCCUCGCGGCCCGGACUUGGGAAGGAA	GAAC	CAGU	C-GA	cuccc	AUUAGI	ACUGGG	CCGCCI	ICCUC	GCGGC	GGGAGI	JUGGGCU	AGGGA	GAACAG	CCUUUCCU	GGCUAA-C	GCCCA	HOVO
GGACCAUUC GACUCCCAUUAGACUGGUCCGCCUCCUCGCGGCGGGAGGUUGGGCUAGGGAACAGCCUUUCCCUAGGCUAA-GGACCA HDVF GGACCAUUC GACUCCGAUUAGACUGGUCCGCCUCCUCGCGGCGGGAGUUGGGCUAAGGAGAACAGCCUUUCCCUAGGCUAA-GGACCA HDVF GGACCAUUC GACUCCGAUUAGACUGGUCCGCCUCCUCGCGGCGGAGUUGGGCUAAGGAGAACAGCCUUCCCUAGGCUAA-GGACCA HDVF GGACCAUUC GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGCGCGGAGUUGGGCUAAGGAGAACAGCCUUCCCUUCCCUAGGCUAA-GGACCA HDVF GGACCAUUC GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCGGAGUUGGGCAAGGAGGAACAGCCUUCCCUCGUGGCUAA-GGACCA HDVF GGACCAUUC GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCGGAGUUGGGCAUGGGAGGAACAGCCUUCCCUCGAUGGCUAA-GGACCA HDVF GGACCAUUC GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCGGAGUUGGGCAUGGGAAGGAA	GGAC	CAUU	C-GA	CUCCC	AUUAGI	ACUGGG	CCGCCI	ICCUC	GCGGC	GGGAGI	JUGGGCU	AGGGA	GAACAG	CCUUUCCUI	GGCUAA-G	GCCCA	HDV4
GGACCAUUC - GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGGCGGGGGGGG	GGAC	CAUD	GIGA	CUCCCC	AUUAGI	ACUGGU	CCGCCI	ICCUC	GCGGC	GGGAGI	JUGGGCU	AGGGA	GGAACAG	ccuuuccu	GGCUAA-C	GACCA	HDV6
GACCAUUC - GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGGCCGAGUUGGGCUAGGGAGGAACAGCCUUUCCCUAGGCUAA - GGACCA HOVI GGACCAUUC - GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGGCCGAGUUGGGCAUGGGAACAGCCUUUCCCUUGGGCUAA - GGACCA HOVI GGACCAUUC - GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCGAGUUGGGCAUGGGAAGGAA	GGAC	CAILED	CJGA	cucco	AUUAGA	ACUGGIO	CCGCCC	CCUC	GCGGC	GGGAGI	JUGGGCU	AGGGA	GGAACAG	CCUUCCCUA	GGCUAA-G	GACCA	HDV7
GACCAUUC - GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCGAGUUGGGCAAGGGAACAGCCUUCCCUAGGCUAA-GGACCA HOV13 GGACCAUUC - GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCGAGUUGGGCAUGGGAAGGAA	GGAC	CAUD	C-GA	CUCCCG	AUUAGA	ACUGGIU	CCGCCC	CCUC	GCGGC	CGGAGI	JUGGGCU	AGGGA	GAACAG	ccuncccur	GGCUAA-G	GACCA	HDV9
GACCAUUC - GACUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGAAGUUGGGCAUGGGAAGACAGCCUUUCCAUGGCUAA - GGACCA HDVIs GGACCAUUC - GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGACGUGGGCAUGGGAAGGAA	GGAC	CAUU	C-GA	CUCGG	AUUAGI	ACUGGI	CCGCCI	CCUC	GCGGGC	COGAGI	UGGGGCU	AGGGAG	GGAACAG	CCUUCCCU	GGCUAA-G	GACCA	HDV11
Gacc Auu c - GG c u c G auu A G A c u G G U c c G c c u c c u c c c c G A G C u G G G C A U G G G A A C A G C C U U c C C A U G G C U A A - G G A C C A HDVIS G A C C A U U c - GG G U C G G A U U A G A C U G G U C G C C U C C U C G C G G C C G A G C U G G G C A U G G G A C A G C C U U C C C A U G G C U A A - G G A C C A HDVIS G G A C C A U U C - G G G U C G G A U U A G A C U G G U C G C C U C C U C G C G G C C U G G G C A U G G G A A G G A C A G C C U U C C C A U G G C U A A - G G A C C A HDVIS G G A C C A U U C - G G G U C G G C U U A G A C U G G U C C G C C U C C U C G C G G C C U G G C C U G G C A U G G C A U G G C U A A - G G A C C A G G A C C A U U C - G G G U C G G C A U - G G C U G G C C U C C C C C C C C C C C	GGAC	CAUU	C-GA	CUCGG	AUUAGA	ACUGGIU	CCGCCI	CCUC	GCGGC	COGAGI	DGGGGCA	AGGGAG	GAACAG	ccuucccuu	GGCUAA-G	GACCA	HDV13
GACCAUUC - GGCUCGGAUUAGACUGGUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA HDV19 GACCAUUC - GGGUCGGCAU - GGCUGGUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA HDV23 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA HDV23 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACGCCUUCCCAUGGCUAA-GGACCA HDV23 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGACGCCUUCCCAUGGCUAA-GGACCA HDV23 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAA-GGACCA HDV23 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGCCGGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAA-GGACCA HDV30 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV33 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV33 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV33 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCCGCGGUCCGACCUGGGCAUGCGAAGGUUUAGCCUUCCGCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCC	GGAC	CAUU	C-GG	CUCGG	AUUAGI	ACUGGU	CCGCCU	CCUC	GCGGC	CCGAGO	UGGGCA	UGGGA	GAACAG	CUUCCCAU	GGCUAA-G	GACCA	HDV15
GGACCAUUC - GGGUUCGGCAU AGACUGGUCCGCCUUCUUCGCGGCCGGCCGAUGGGCAUGGGAAGGUUUUCCUUCGCAUGGCUAA - GGACCA HDV21 GGACCAUUC - GGGUUCGGCAU - GGCUGGUCCGCCUUCUUCGCGGCCGGCCGGACCUGGGCAUGGGAAGGACAGCCUUUCCCAUGGCUAA - GGACCA HDV23 GGACCAUUC - GGGUUCGGCAU - GGCUGCUCCGCCUUCUUCGCGGCCCGACCUGGGCAUGGGAAGGUUUGCCUUUCCCAUGGCUAA - GGACCA HDV27 GGACCAUUC - GGGUUCGGCAU - GGCUGCUCCGCCUUCUUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUUCCCAUGGCUAA - GGACCA HDV27 GGACCAUUC - GGGUUCGGCAU - GGCUGCUCCGCCUUCUUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUUCCCAUGGCUAA - GGAGCA HDV27 GGACCAUUC - GGGUUCGGCAU - GGCUGCUCCGCCUUCUUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUUCCCAUGGCUAA - GGAGCA HDV28 GGACCAUUC - GGGUUCGGCAU - GGCUGCUCCGCCUUCUUCGCGGCCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUUCCCAUGGCUAAGGGAGCA HDV29 GGACCAUUC - GGGUUCGGCAU - GGCUGCUCCACCUUCUCGCGGUUCGACCUGGGCAUGGGAAGGUUAGCCUUUCCCAUGGCUAAGGGAGCA HDV30 GGAC - AUUC - GGGUUCGGCAU - GGCUGCUCCACCUUCUCGCGGUUCGGACCUGGGCAUGGGAAGGUUAGCCUUUCCCAUGGCUAAGGGAGCA HDV33 GGAC - AUUC - GGGUUCGGCAU - GGCUGCUCCACCUUCUCGCGGUUCGGACGUGGGCAUGGGAAGGUUAGCCUUUCCCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUUCGGCAU - GGCUGCUCCACCUUCUCCCCGGUUCGGACGUGGGCAUGGGAAGGUUUUCCUUCGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUUCGGCAU - GGCUGCUCCACCUUCUCCCGCGUUCGGACCUGGGCAUGCGAAGGUUUUUCCUUCGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUUCGGCAU - GGCUUCUCCACCUUCUCCCGCGUUCGGCAUGGGCAUCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUUCGGCAU - GGCUUCUCCACCUUCUCCCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUUCGGCAU - GGCUUCUCCACCUUCCUCCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCC	GGAC	CAUU	C-GG	CUCGG	AUUAGI	ACUGGU	CCGCCU	CCUC	GCGGC	CCGAGO	UGGGCA	UGGGA	GGACAG	CUUCCCAL	GGCUAA-G	GACCA	HOVIO
GGACCAUUC-GGGUCGGCAU-GGCUCGCCCCCCCCCCCCGACCUCGCGGCCAUGGGCAUGGGAAGGACAGCCUUCCCAUGGCUAA-GGACCA GGACCAUUC-GGGUCGGCAU-GGCUGCCCCCCCCCCCCCC	GGAC	CAUU	C-GG	GUCGG	AUUAGI	ACUGGU	CCGCCL	CCUC	GCGGC	CCGACO	UGGGCA	UGGGAL	GGACAGO	CCUUCCCAU	GGCUAA-G	GACCA	HDV21
GGACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAA - GGACCA HDV27 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAA - GGAGCA HDV27 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGCCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAA - GGAGCA HDV29 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAA - GGAGCA HDV29 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV30 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUUAGCCUUCCGCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGCAAGGUUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGCAAGGUUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGCAAGGUUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUCGAAGGUUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGGAAG HDV40 HDV	GIGAC	CAUU	C-GG	guciee	CAUAGA	cuccu	CCGCCU	CCUC	GCGGC	CCGACO	UGGGCA	UGGGAI	GGACAGO	CUUCCCAU	GGCUAA - G	GACCA	HDV23
GACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGCCCGACQUGGGCAUGGGAAGGUUUAGCCUUCCCAUGGCUAA-GGAGCA HDV29 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCGCCUCCUCGCGGUCCGACQUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV29 GGACCAUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACQUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV39 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACQUGGGCAUGGGAAGGUUAGCCUUCCCAUGGCUAAGGGAGCA HDV39 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACQUGGGCAUGGGAAGGUUUAGCCUUCCCAUGGCUAAGGGAGCA HDV39 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUAGCCUUCCGCAUGGCUAAGGGAGCA HDV34 HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGGCAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40	COAC	CAUU	CI-GG	GUCIGG	CAU-GO	cuccio	CCGCCU	CCUC	GCGGC	CCGACC	UGGGCA	UGGGAI	GGACAGO	CCUUCCCAU	GGCUAA - G	GACCA	HDV25
Gaccaluuc - Gggucggcau - Ggcugcuccaccuccucgcgguccgaccugggcau - Ggaucgau - Ggaucgcuccaccuccucgcgguccgaccugggcau - Auuc - Gggucggcau - Ggcuggucggau - Ggcugcuccaccuccucgcgguccgaccuggggcau - Auuc - Gggucggcau - Ggcuggucggau - Ggcugcuccaccuccucgcgguccgaccuggggau - Auuc - Gggucggau - Ggcuggucggau - Ggcugguu - Ggcuu - Gguu - Ggcuu - Gguu - Guu	GGAC	CAUL	CIGG	GUCIGG	CAU-GO	ac u gic u	CCGCCU	CCUC	GCGGC	CCGACC	UGGGCA	UGGGAI	GGACAGO	CCUUCCCAU	GGCUAA-G	GAGCA	HDV27
GGACCAUUC-GGGUCGGCAU-GGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUUUCCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUCGGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA HDV30 GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAAGA HDV40 GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAAGA HDV40 HDV40 GGAC-AUUC-GGGUCGGCAU-GGCUCCGCUCCUCCCUCCCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAAGA HDV40 GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAAGA HDV40 H	GGAC	CAUU	C-GG	ducies	CAURGO	CUGCU	CCGCCC	CCUC	GCGGC	COGACO	UGGGCA	UGGGAI	GUUAGO	CCUUCCCAU	GGCUAA-G	GAGCA	HDV29
GGAC AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUUAGCCUUCCCAUGGCUAAGGGAGCA HDV32 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUUAGCCUUCGCAUGGCUAAGGGAGCA HDV34 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGGGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 GGAC - AUUC - GGGUCGGCAU - GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HDV36 HDV40 GGAC - AUUC - GGGUCGGCAU - GGCUCCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 GGAC - AUUC - GGGUCGGCAU - GGCUUCCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 GGAC - AUUC - GGGUCGGCAU - GGCUUCCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 GGAC - AUUC - GGGUCGGCAU - GGCUUCCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 GGAC - AUUC - GGGUCGGCAU - GGCAUCCCACCUCCUCCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 GGAC - AUUC - GGGUCGGCAU - GGCAUCCCACCUCCUCCCCCCCGGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 GGAC - AUUC - GGGUCGGCAU - GGCAUCCCACCUCCUCCCCCCCCCCCGGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 HDV40 CUUC - GGGUCGGCAU - GGCAUCCCACCUCCUCCCCCCCCCCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 HDV40 HDV40	GGAC	CAUU	C-GG	GUCIGG	CAU-GO	CUGCH	collectu	CCUC	GCGGU	COGACO	UGGGCA	UGGGAI	GGUUAGO	CUUCCCAU	GGCUAAGG	GAGCA	HDV30
GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HOV3 GGAC - AUUC - GGGUCGGCAU - GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUGCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HOV3 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HOV3 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HOV3 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA HOV3 GGAC - AUUC - GGGUCGGCAU - GGCUUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA GGAC - AUUC - GGGUCGGCAU - GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HOV40 HOV40 GGAC - AUUC - GGGUCGGCAU - GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HOV40 HOV42 HOV42 HOV40 HOV42 HOV40 HOV42 HOV40 HOV42 HOV40 HOV40 HOV40 HOV42 HOV40 H	GGAC	AUU	C-GG	ducias	CAU-GO	CUGCU	CCACCU	CCUC	GCGGII	CCGACC	UGGGGCA	GGGGAI	GGUUAGO	CUUCCICAU	GGCUAAGG	GAGCA	HDV32
GGAC-AUUC-GGGUCGGCAU-GGCUGCUCCACCUCCUCGCGGUCCGACCUGGGCAUCGGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGCA GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGAA GGAC-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUUCCUUCGGAUGGCUAAGGGAGAA GGAC-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 HDV42 GGGA-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 HDV42 HDV4 HDV42 HDV4 HDV42 HDV42 HDV42 HDV44 HDV4	GGAC	-AUU	C-GG	ducias	CAU-GO	CUGCU	CCACCU	CCUC	GCGGU	CCGACC	UGGGCA	UCCCAL	GGUUAGO	CUDCCCAU	GGCUAAGG	GAGCA	HDV33
GGAC-AUUC-GGGUCGGCAU-GGCUCGCCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGCA GGAC-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA GGAC-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA GGAC-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA GGAC-AUUC-GGGUCGGCAU-GGCAUCUCCACCUCCUCGCGGUCCGACCUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA HDV40 HDV42 P1 J1/2 P2 P3 L3 P3 P1 P4 L4 P4 J4/2 P2	GGAC	-AUU	C-GG	GUCGG	CAU-GG	CUGCU	CCACCU	CCUC	GCGGU	CCGACC	UGGGCA	UGICIGAZ	GIGUUUU	CUUCGCAU	GGCUAAGG	GAGCA	HDV34
GGAC-AUUC-GGGUCGGCAU-GGCUUCUCCACCUCCUCGCGGUCGGACGUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAA GGAC-AUUC-GGGUCGGCAU-GGCUCGGCCUCCUCGCGGUCGGACGUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAG GGAC-AUUC-GGGUCGGCAU-GGCUCGGCUCCGCGGUCGGACGUGGGCAUCCGAAGGUUUUCCUUCGGAUGGCUAAGGGAGAG HDV40 P1 J1/2 P2 P3 L3 P3 P1 P4 L4 P4 J4/2 P2	GGAC	-AUU	C-GG	GUCIGG	CAU-GG	CUGCU	CCACCU	CCUC	GCGGU	CCGACC	UGGGCA	UCCGAL	GGUUUUG	CUUCGGAU	GGCUAAGG	GAGCA	HDV38
P1 J1/2 P2 P3 L3 P3 P1 P4 L4 P4 J4/2 P2 P3 P1 P2 P3 P1 P4 L4 P4 J4/2 P2	GGAC	AUU	C - GG	gucies	CAU-GG	CUNCU	CCACCU	CCUC	GCGGU	CCGACC	UGGGCA	UCCGAF	GGUUUUU	CUUCGGAU	GGCUAAGG	GAGAA	HDV40
P1 J1/2 P2 P3 L3 P3 P1 P4 L4 P4 .14/2 P2	COCAC	AUU	GIGG	GUCIGG	CAU-GG	CAUCUC	CACCU	CCUC	GCGGU	CCGACC	UGGGCA	UCCGAR	GGUUUUC	CUUCGGAU	GGCUAAGG	GAGAG	HDV42
P1 J1/2 P2 P3 L3 P3 P1 P4 L4 P4 .14/2 P2	OLG G A	AUD	CELOIG I	G O CIG G	La VE GLG	CAUCUC	CLAICCU	ccuc	GCGGIU	CCIGAICC	UGGGCA	UCCIGAA	GGUUUUG	CUUCGGAU	GGCUAAGG	GAGAG	HDV P
			P	1	J1/2	P2	P3	L	3 P3	P1	Are	P4	L4	P4	J4/2	P2	

Sequence of mutants from the intersection to both reference ribozymes

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

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

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

SUMMARY

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



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

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

© 1994 The Royal Society

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

Coworkers

Walter Fontana, Santa Fe Institute, NM

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

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

Bärbel Stadler, Ulrike Mückstein, Andreas Wernitznig, Stefanie Widder, Stefan Wuchty, Universität Wien, AT

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