# Computational methods for RNA folding: Algorithms and Software

Michael Zuker

September 30, 2009.

– Typeset by  $\mbox{Foil}T_{\!E}\!X$  –

# What is RNA primary structure?

RNA primary structure is the sequence of ribonucleotides (bases), and can be written as  $5' - r_1 r_2 \dots r_n - 3'$ , where *n* is the number of bases and each  $r_i$  is either A, C, G or U. Of course, bases are sometimes modified, either naturally or *in vitro*.

Nucleobase	Adenine	Guanine	Cytosine	Uracil
Nucleoside	Adenosine	Guanosine	Cytidine	Uridine
(NB + D-ribose)				
Nucleotide	AMP, Ap	GMP, Gp	CMP, Cp	UMP, Up
(NS+phosphate)			•	

#### Ambiguous codes for RNA/DNA

A,G	C,U/T	A,U/T	C,G	A,C	G,U/T
R	Y	W	S	М	K
C,G,U/T	A,G,U/T	A,C,U/T	A,C,G	A,C,G,U/T	
В	D	Н	V	N	

# **Remembering the Ambiguous Base Letters**



#### What is RNA secondary structure?

RNA secondary structure is the collection (set) of base pairs that form in 3D. The hydrogen bonds of base pairs and the stacking of adjacent base pairs are responsible for most of the thermodynamic stability of an RNA. The most common base pairs are Watson-Crick (W-C): C·G, G·C, A·U and U·A. Base pairs between G and U, G·U and U·G, are called wobble pairs. Other pairs are called non-canonical.

A base pair between  $r_i$  and  $r_j$  is denoted as  $r_i \cdot r_j$  (i < j) or simply by  $i \cdot j$  when the context is clear. A secondary structure is a collection, *S*, of base pairs that satisfy:

- 1. If  $i \cdot j \in S$ , then j i > 3. The number 3 is called the minimum hairpin loop size. This "rule" is broken by the existence of tetra-loops.
- 2. If  $i \cdot j, i' \cdot j' \in S$ , then j = j' if i = i' and i = i' if j = j'. This rule excludes base triples, and is violated in some structures.
- 3. If  $i \cdot j, i' \cdot j' \in S$ , then either i < j < i' < j' ( $i \cdot j$  precedes  $i' \cdot j'$ ) or i < i' < j' < j ( $i \cdot j$  includes  $i' \cdot j'$ ). Violations of this rule also occur and create "pseudoknots".

# Anything can pair with anything else in twelve different ways (almost).

Recent work of Leontis and Westhof describe 12 distinct ways in which any two bases can pair.

- Some cannot occur.
- Some can be modeled but have not yet been observed in structures derived by X-ray crystalography or NMR methods.
- Many base pairs are structurally equivalent (isosteric).

http://www.bgsu.edu/departments/chem/RNA/pages/

PDB/index.html

#### **Non-canonical BPs Deduced by Comparative Sequence Analysis**

The figure on the right displays five consecutive non-canonical BPs in a portion of a largesubunit rRNA.

Statistics were gathered from 5S rRNA, grpl introns, RNAse P RNA, SSU and LSU rRNA.



# **Nomenclature and Definitions**

- Complete: Entire structures.
- Core: "Reliable" portions of rRNA structures deduced from an alignment with respect to E. coli rRNA.
- ext-: Base pairs as indicated.
- ext+: Base pairs added when found in small symmetric interior loops (up to  $5 \times 5$ ).
- "X<sub>p</sub>Y" refers to a di-nucleotide, not to a BP. Think 5' XY 3'.
- "XY" refers to a BP, canonical or not. Thus AU or GG are base pairs.
- "WX/YZ" refers to "tandem base pairs", canonical or not. Think  $\frac{5' WY 3'}{3' XZ 5'}$ .
- "WX/YZ" is clearly equivalent to "ZY/XW".
- F(WX/YZ) is the ratio of the observed number of base pair stacks of type WX/YZ divided by the expected number. The expected number is based on observed frequencies of di-nucleotides.
- F(WX/YZ) measures under or over representation of tandem BPs.

# **The Numbers**

Data	Sequences	Stacks	Base pairs	Nucleotides
Complete ext-	1900	337659	436327	1597471
Complete ext+	1900	345586	441117	1597471
core rRNA ext-	8058	1671686	2245790	10998639
core rRNA ext+	8058	1737666	2287177	10998639

The core data, though less reliable, especially for Eukaryote rRNA, was used because of the very large number of base pairs that could be analyzed.

# Statistics on tandem base pairs - I

		Com	plete		Core rRNA			
	Co	ount	F(WX/YZ)		Count		F(WX/YZ)	
WX/YZ	-ext	+ext	-ext	+ext	-ext	+ext	-ext	+ext
AA/AA	121	170	0.24	0.25	105	358	0.29	0.33
AA/AC	33	45	0.83	0.79	79	167	0.99	0.73
AA/AG	501	1283	0.64	0.86	412	13490	0.48	1.60
AA/AU	626	882	0.62	0.67	1746	7037	1.00	1.24
AA/CA	46	78	0.91	0.59	71	175	0.95	0.54
AA/CC	23	25	2.14	1.70	99	571	1.56	0.92
AA/CG	725	878	1.10	1.17	2883	3284	1.46	1.36
AA/CU	51	66	2.13	1.33	38	136	1.35	1.07
AA/GA	104	140	0.38	0.33	144	226	0.71	0.42
AA/GC	809	967	1.20	1.25	1627	5614	1.21	1.14
AA/GG	44	48	0.47	0.46	77	90	0.72	0.68
AA/GU	219	282	0.68	0.66	289	853	0.77	0.73
AA/UA	666	831	0.71	0.74	1316	2494	0.91	0.95
AA/UC	49	54	2.33	1.76	62	80	1.29	1.04
AA/UG	135	160	0.60	0.58	243	351	0.81	0.69
AA/UU	220	291	0.84	0.77	215	875	0.77	1.11
AC/AC	36	40	0.65	0.61	64	66	0.89	0.83

# Statistics on tandem base pairs - II

		Comp	olete		Core rRNA			
	Со	unt	F(WX/YZ)		Count		F(WX/YZ)	
WX/YZ	-ext	+ext	-ext	+ext	-ext	+ext	-ext	+ext
AC/AG	36	103	0.53	0.56	121	235	0.80	0.50
AC/AU	337	518	0.77	0.73	1404	1794	1.22	0.94
AC/CA	32	34	0.68	0.68	120	128	1.23	1.16
AC/CC	31	33	1.39	1.21	80	85	1.50	1.29
AC/CG	1517	1529	1.83	1.79	10168 10372		1.95	1.87
AC/CU	17	18	1.31	1.17	126	126 129		1.49
AC/GA	43	59	0.66	0.66	149	149 491		0.94
AC/GC	693	764	1.14	1.12	2435	2827	1.49	1.48
AC/GG	46	49	0.70	0.64	163	165	0.88	0.83
AC/GU	240	260	0.80	0.81	1311	1361	1.15	1.10
AC/UA	618	636	1.08	1.07	700	796	1.36	1.28
AC/UC	62	62	1.66	1.60	28	29	1.44	1.32
AC/UG	177	190	0.76	0.75	1121	1123	1.18	1.12
AC/UU	55	84	0.94	0.94	91	94	1.27	1.16
AG/AC	40	43	0.57	0.53	33	116	0.98	0.57
AG/AG	394	549	0.50	0.44	2146	2719	0.62	0.54
AG/AU	575	708	0.64	0.69	2343	3132	0.75	0.76

# Statistics on tandem base pairs - III

		Comp	olete		Core rRNA			
	Co	unt	F(WX/YZ)		Count		F(WX/YZ)	
WX/YZ	-ext	+ext	-ext	+ext	-ext	+ext	-ext	+ext
AG/CC	25	27	1.00	0.91	48	111	0.98	0.75
AG/CG	2803	3000	2.06	2.14	19452	26253	2.15	2.75
AG/CU	50	60	0.55	0.52	182	253	0.80	0.84
AG/GA	55	56	0.42	0.40	164	167	0.70	0.67
AG/GC	1804	2214	1.49	1.70	2188	8923	1.11	1.68
AG/GG	49	55	0.60	0.58	214	214 225		0.68
AG/GU	1351	1566	1.17	1.26	5687	8033	0.85	1.08
AG/UA	608	869	0.93	1.09	2566	6124	1.02	1.49
AG/UC	7	8	0.94	0.98	28	69	1.19	0.94
AG/UG	305	511	0.58	0.69	2303	6474	0.95	1.11
AG/UU	194	282	0.67	0.65	184	825	0.91	0.92
AU/AC	560	683	0.86	0.97	2440	2866	1.32	1.28
AU/AG	645	701	0.60	0.61	4551	10955	0.98	1.29
AU/CC	149	190	1.36	1.31	511	549	1.67	1.49
AU/CU	336	393	0.95	0.91	513	686	1.20	1.10
AU/GG	552	666	1.45	0.95	1444	1472	0.99	0.94

		Comp	olete			Core r	RNA									
	Co	unt	F(WX/YZ)		Count		F(WX/YZ)									
WX/YZ	-ext	+ext	-ext	+ext	-ext	+ext	-ext	+ext								
AU/UC	547	586	1.15	1.12	1791	1990	1.35	1.35								
AU/UU	1447	1589	1.37	1.43	8180	8483	2.29	2.25								
CA/AC	55	79	1.09	1.24	613	623	1.54	1.48								
CA/AG	39	52	0.88	0.88	291	324	1.25	1.15								
CA/AU	249	340	1.01	0.98	633	874	1.37	1.18								
CA/CC	10	12	1.34	1.34	47	49	1.28	1.14								
CA/CG	768	826	1.07	1.04	2293	2406	1.06	0.98								
CA/CU	15	15	0.76	0.75	54	60	1.13	1.04								
CA/GC	499	582	1.43	1.40	2216	2431	1.63	1.47								
CA/GG	67	73	0.62	0.56	773	778	1.02	0.98								
CA/GU	66	101	0.92	0.73	188	293	1.25	0.95								
CA/UC	16	27	1.81	1.50	53	66	1.42	1.28								
CA/UG	177	179	0.67	0.65	985	993	1.22	1.16								
CA/UU	37	41	0.85	0.80	68	81	1.31	1.12								
CC/AG	13	16	0.99	0.94	68	70	1.18	1.09								
CC/AU	87	101	1.69	1.52	360	368	1.63	1.54								

# Statistics on tandem base pairs - IV

# Statistics on tandem base pairs - V

		Comp	olete			Core rRNA				
	Co	unt	F(W)	<td>Co</td> <td>unt</td> <td colspan="2">F(WX/YZ)</td>	Co	unt	F(WX/YZ)			
WX/YZ	-ext	+ext	-ext	+ext	-ext	+ext	-ext	+ext		
CC/CC	9	11	0.70	0.30	46	47	0.65	0.56		
CC/CG	187	202	0.91	0.89	1180	1252	1.28	1.09		
CC/CU	12	14	2.69	1.64	44	45	1.56	1.39		
CC/GC	207	219	0.77	0.76	1144	1616	1.22	0.97		
CC/GG	35	36	0.44	0.42	680	687	0.97	0.92		
CC/GU	57	58	0.93	0.93	153	156	1.20	1.11		
CC/UC	13	15	1.52	1.41	30	31	1.06	0.99		
CC/UG	86	87	0.92	0.90	880	885	1.27	1.22		
CC/UU	48	95	0.86	0.96	170	174	1.25	1.17		
CG/AG	723	764	0.86	0.87	5998	6049	1.03	0.99		
CG/CU	428	644	0.91	1.15	1617	2154	1.14	1.27		
CG/GG	403	505	0.62	0.55	3761	3772	0.82	0.79		
CG/UC	390	633	1.23	1.28	1275	1848	1.31	1.23		
CG/UU	1260	1326	1.78	1.80	3107	3870	1.51	1.75		
CU/AG	35	37	1.84	1.76	108	168	1.09	0.97		
CU/AU	307	339	0.95	0.93	1023	1143	1.51	1.38		

# Statistics on tandem base pairs - VI

		Comp	olete		Core rRNA			
	Со	unt	F(WX/YZ)		Count		F(WX/YZ)	
WX/YZ	-ext	+ext	-ext	+ext	-ext	+ext	-ext	+ext
CU/CG	492	564	1.10	1.04	2887	3060	1.19	1.10
CU/CU	60	63	0.93	0.89	489	489	1.07	1.03
CU/GG	24	26	0.68	0.60	203	229	1.12	0.96
CU/GU	105	109	1.27	1.21	204	225	1.18	1.15
CU/UC	6	6	1.11	1.08	80	80 80		1.10
CU/UG	124	126	0.82	0.80	944	944 950		1.29
CU/UU	46	67	0.69	0.64	362	362 366		0.98
GA/AG	738	966	0.70	0.73	1788	4959	0.63	0.77
GA/AU	2326	2641	1.94	1.96	9546	13561	1.55	2.04
GA/CG	1366	1401	1.30	1.28	15634	15920	2.05	1.97
GA/CU	17	19	1.16	1.06	73	124	1.31	0.89
GA/GG	335	343	1.79	1.68	134	159	0.59	0.53
GA/GU	260	277	0.65	0.63	1444	1704	0.85	0.78
GA/UG	183	189	0.61	0.59	1043	1069	0.89	0.83
GA/UU	55	64	0.57	0.53	245	653	0.91	0.98
GC/CU	398	429	1.09	1.07	3931	3947	1.48	1.42
GC/GG	471	513	0.60	0.61	3171	3175	0.71	0.68

# Statistics on tandem base pairs - VII

		Comp	olete		Core rRNA			
	Co	unt	F(WX/YZ)		Count		F(WX/YZ)	
WX/YZ	-ext	+ext	-ext	+ext	-ext	+ext	-ext	+ext
GC/UU	1796	2179	2.42	2.69	15064	15462	3.31	3.15
GG/AU	354	377	0.72	0.73	2207	2211	1.06	1.01
GG/CU	21	24	1.08	0.87	94	116	0.85	0.69
GG/GG	10	14	0.29	0.29	91	93	0.41	0.38
GG/GU	101	115	0.44	0.42	255	386	0.58	0.50
GG/UG	68	76	0.47	0.44	380	380 385		0.65
GG/UU	46	49	0.52	0.50	119	119 194		0.60
GU/CU	63	64	0.68	0.66	429	431	1.13	1.08
GU/UU	238	256	0.60	0.61	1117	1168	0.92	0.87
UA/CU	320	358	1.45	1.34	2147	2166	2.26	2.14
UA/UU	969	1154	1.06	1.17	3784	3974	1.84	1.83
UC/CU	15	16	3.63	2.37	25	28	1.54	1.39
UC/GU	56	66	0.92	0.96	244	249	1.07	1.00
UC/UU	200	235	1.05	0.94	534	707	1.14	1.03
UG/UU	645	678	1.03	1.04	1830	1983	1.07	1.08
UU/UU	397	459	0.57	0.61	1418	1425	0.90	0.86

# **Non-canonical BP Count**

Note: The W-C base pair counts are restricted to those occurring next to a non-canonical base pair. The numbers in the table are the base pair counts for complete and core BPs, ext- and ext+

BP	Complete	BP	Complete+	BP	Core	BP	Core+
CC	992	CC	1141	CC	5540	CC	6696
GG	2626	GG	2969	AA	9406	GG	14137
CU	4282	CU	5143	GG	13766	CU	22054
AA	4372	GU	5350	CU	19618	GU	29072
GU	4656	AA	6200	GU	21050	AC	31997
AC	6617	AC	7555	AC	28918	AA	35801
UU	7653	UU	8849	UU	36488	UU	40334
AU	12278	AU	14562	AU	49205	AU	72675
AG	15679	AG	19003	AG	79287	CG	124235
CG	17739	CG	20139	CG	102031	AG	133581

The BP columns are ordered from least frequent to most.

# What is RNA tertiary structure?

Base triples and pseudoknots are best excluded from the definition of secondary structure. They comprise some of the interactions that characterize tertiary structure. Angles between pairs of adjacent helices are part of tertiary structure. Adding (sufficient) tertiary structure to secondary structure can lead to the modeling of atomic resolution structures.

A helix is a collection of two or more adjacent base pairs.

#### 3D structure for *S. cerevisiae* Phe tRNA.

An atomic resolution model is complicated and full of details. What can secondary structure show?



#### A 3D-like secondary structure for *S. cerevisiae*

This image preserves some features of the 3D structure.



1TRA yeast Phe tRNA (S. cerevisiae)

# The "cloverleaf" secondary structure for *S. cerevisiae*

This traditional secondary structure representation for tRNA does not convey 3D structure information.



1TRA yeast Phe tRNA (S. cerevisiae)

#### **Representations of RNA secondary structures.**



The traditional representation of secondary structure. Example is a group II intron from *Microscilla sp. PRE1* Source: Steve Zimmerly www.fp.ucalgary.ca/ group2introns/species.htm)

- E: exterior loop
- H: hairpin loop
- B: bulge loop
- I: interior loop
- M: multi-branch loop

Colors depend on the probability of base pairs or the probability that a single-stranded base is single-stranded.

# **Color annotation of secondary structure.**

The eight colors used to annotate probabilities. They emphasize the very high and very low probabilities. 0.9990< Prob. <=1.0000</td>0.9900< Prob. <=0.9990</td>0.9000< Prob. <=0.9900</td>0.6500< Prob. <=0.9000</td>0.3500< Prob. <=0.6500</td>0.1000< Prob. <=0.3500</td>0.0100< Prob. <=0.1000</td>0.0000< Prob. <=0.0100</td>

# **Circle Plot & Tree Diagram**



Two plots in one.

- Bases are drawn along the circumference of a circle
- Base pairs are circular arcs the intersect the circle at right angles.
- Black lines (edges) within the circle comprise a "tree representation" of the secondary structure.

Every base pair and multibranched loop is a node. Nodes connecting consecutive base pairs can be collapsed into a single "helix" node.

Colors depend on the probability of base pairs, as in the standard plot.

# Side by side comparison - 1



### **Pseudoknot example**

Two base pairs (BPs),  $r_i \cdot r_j$  and  $r_{i'} \cdot r_{j'}$ , can be called "incompatible" if i < i' < j < j'. That is, they violate the third rule for RNA secondary structure.

A pseudoknot is created by two incompatible helices (stems). That is, every BP in one is incompatible with every BP in the other.

Stem 1:  $C_1 \cdot G_{15}$ ,  $C_2 \cdot G_{14}$ ,  $C_3 \cdot G_{13}$ ,

is incompatible with

Stem 2:  $U_8 \cdot A_{23}$ ,  $C_9 \cdot G_{22}$ ,  $C_{10} \cdot G_{21}$ ,  $G_{11} \cdot C_{20}$ ,  $A_{12} \cdot U_{19}$ .

Is this possible? – Yes.

3D model of simple pseudoknot. Coordinates by F. Major



# Same model - orientation shows coaxial stacking



# In the circle plot, intersecting BP arcs indicate a pseudoknot



#### Can you find the pseudoknot?



#### Now you can. P3 and P7 create a pseudoknot.



#### Now you can see them in the original plot.



# Dot plot

Structure dot plot for M\_sp\_l1\_phylo



# Side by side comparison - 2



# **Prediction of RNA secondary structure**

What are the common methods?

- Comparative, or phylogenetic methods.
  - Considered the "gold standard".
  - Labor intensive
  - Requires numerous homologous sequences that can be well aligned.
- Free energy minimization methods
  - Works on single sequences.
  - Fast, cheap and easy to perform.
  - Unreliable in general.
  - Cannot predict pseudoknots. (Some time consuming exceptions exist.)

# **Comparative methodology**

- A "golden rule" in biology: Structure is conserved more than sequence.
- This principle can be used to predict RNA secondary structure.
- It is used together with site directed mutagenesis to confirm the existence of specific base pairs.
- It can be used, for example, to design non-virulent strains of an RNA virus by interrupting significant secondary structure.

# SSU rRNA: Escherichia coli versus Deferribacter thermophilus




#### Secondary structure comparison between two 16S rRNAs.



Escherichia coli

Deferribacter thermophilus

Compare a small domain in one with the corresponding domain in the other.

- BP is conserved Both bases unchanged.
- BP is conserved Both bases change (compensatory change)
- BP is conserved One base changes (W-C  $\leftrightarrow$  wobble pair)
- BP not conserved One base changes (W-C  $\leftrightarrow$  non-canonical pair)

#### Need an alignment of homologous sequences

Given *m* RNA sequences,  $R_1, R_2, \ldots, R_m$ . The *i*<sup>th</sup> sequence has length  $n_i$ . After alignment, they all have a common length, *n*. They can be written as

$$\begin{array}{rclcrcrcrcrcrc} R_1 &=& r_1(1), & r_1(2), & r_1(3), & \dots, & r_1(n), \\ R_2 &=& r_2(1), & r_2(2), & r_2(3), & \dots, & r_2(n), \\ R_3 &=& r_3(1), & r_3(2), & r_3(3), & \dots, & r_3(n), \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ R_m &=& r_m(1), & r_m(2), & r_m(3), & \dots, & r_m(n). \end{array}$$

 $R_k(i)$  is the *i*<sup>th</sup> "base" in the *k*<sup>th</sup> sequence. It is A, C, G, U or "-". The last symbol stands for an inserted gap.

Constructing a "correct" alignment is usually slow and difficult work. Many methods have been developed to automate this procedure.

"Correct" refers to computing an alignment that captures the evolution of these RNAs.

# Some History: Quick-n-Dirty

Copied from the Gutell Lab web site:

http://www.rna.ccbb.utexas.edu/CAR/1D/Paradigms/Visual

Red-dot Green-dot: 1978–1981; attributed to Carl Woese

- Strict covariation.
- Can detect W-C and wobble pairs only (consecutive, anti-parallel & nested)
- Provides a comparative proof for a helix. Two or more compensatory base changes within a helix.
- Less similar sequences provide more supporting changes.
- Aligns sequences to maximize primary structure similarity.
- Analyzes just two sequences at one time.
- Identifies secondary interactions only.

# Red-dot, green-dot: Gutell tRNA example



**Figure 1**. Reddot-greendot examples from tRNA. Symbols used: +: transition; -: transversion; |: deletion; \*: ambiguous nucleotide. Experimentally verified helices from the secondary structure are boxed and connected with black lines. Nucleotide position numbers refer to the S.cerevisiae Phe reference sequence. Sequence names are shown as *amino acid:organism*.



Figure 2. Results of the reddot-greendot analysis shown on tRNA secondary structure diagrams. Base pairs which are predicted with the method are shown with red tick marks. A. Sequences with 85.3% similarity. B. 72.9% similarity. C. 53.4% similarity.

#### September 30, 2009. - RNA Course

#### Green dot Red dot: E. coli vs. Z. mays 16S rRNA.

1	0	20	30 	40 	50	)	60 	70 		80 	90	) 1(	00	110 	12	0 	130 	14	0	150 	160 
AAAUUGAAG	AGUUUGAU		GAUUGAAC		GCAGGCCUA	ACACAUGC	AAGUCGA		AACAGGI	AAGÀAGCU	UGCUUC			GGAĊG  @    GAACG	GGUGAGUA	AUGUCUG  @ @@ @ ACGCGUA	GGAÀAC @   @  AGAACC	UGCCUGA		GAUAA  @ @   GAACAA	CAACUGGA
1		 20	 30	40	5(	 )	 60	 70				80	91		100	11	 .0	120	13	)	140
17	70 	180	190 	200 	210	) 2. I	20 	230 		240	250	) 20	50 	270 	28	D I	290 	30	0 	310 	320 
AACGGUAGC			GCAAGACC	AAAGAGGG   @   @@	GGGACCUU(	CGGGCCUCU		GGAUGU		GAUGGGAU @@ @@			GUAACGG	CUCÁC	CUAGGCGA			GUCUGAG		CCAGCC	ACACUGGĂ
150	16	 0 17	   0	 180	ANUCC	1	90	200	GCUCG	210	220	) 2:	 30	240	25	   0	260	27	   0	280	290
33	30	340	350	360	370	) 3	80 I	390 I		400	410	) 42	20	430 I	44	0	450	46	0	470 I	480
ACUGAGACA	CGGUCCA	GACUCCUACO	GGAGGCAG	cagugggg	GAAUAUUG(	CACAAUGGG  @	cgcaago    @	CUGAUC	GCAGCC/  @   @	augċcgcg 	UGUAUC	GAAGAAGGC(  @@		JGUÅA ⊉  @	AGUACUUU		GAGGAA	GGGAGUA	AAGUUAAI	JACCUU	JGCUCAUU  @  @
ACUGAGACA	ACGGCCCAO   )0	GACUCCUACC   310	GGAGGCAG   320	CAGUGGGG   330	3400000C0	CGCAAUGGG   ) 3	CGAAAGC   50	CUGACO   360	GGAGCA	AUGCCGCG   370	UGGAGO   380	GUGGAAGGCO	CUACGGGU   90	2GUCA   400	ACUUCUUU	JCUCGGA   0	GAAGAA   420				ACAAU
49	0	500	510	520	53(	) 5	40	550		560	570	) 58	30	590	60	0	610	62	0	630	640
GACGUUACC	 CCGCAGAA(  @ @@   (	 GAAGCACCGC @     @	 CUAACUCC      @	 GUGCCAGC	CAGCCGCGG	 GUAAUACGG. 	 AGGGUGC      @	LAAGCGI	UUAAUC(	 GGAAUUAC           @ @	UGGGCC	 GUAAAGCGC/                 @@	 ACGCAGGC( @@ @   @	 GUUU   @	GUUAAGUC	 AGAUGUG <mark>@   @@     @</mark>	AAAUCC	.CCGGGCU  @	 CAACCUG(         @@	 GGAACU(       @@@	 }CAUCUGA     @@@@
GACGGUAUC	UGAGGAAU	JAAGCAUCGO	CUAACUCU	GUGCCAGC	CAGCCGCGC   70	JUAAGACAG	AGGAUGC   490	AAGCGI	UUAUCCO   500	GGAAUGAU   510	UGGGCC	JUAAAGCGU	CUGUAGGU	GCUU 5	UUCAAGUC	CGCCGUC	AAAUCC	CAGGGCU	CAACCCU   570	GACAG	3CGGUGGA   0
65	50	660	670	680	69(	) 7	00	710	500	720	730	) 74	10	750	76	p p	770	78	0	790	800
UACUGGCAA @   @@	 \GCUUGAGI    @	 UCUCGUAGAC  @@@    @	  GGGGGGUAG  @@ @@@	 AAUUCCAC         @   @	GUGUAGCO	 GGUGAAAUG 	 CGUAGAG  @ @	 AUCUGO	GAGGAA1 @ @   (	 UACCGGUG @   @@@	 GCGAAG 	 GGCGGCCCC( @     @@@   @@	 CUGGACGA2         @   @@	 AGACU  @	GACGCUCA	 GGUGCGA  @@@	  AAGCGU     @@	GGGGAGC	 AAACAGG 	 AUUAGA  	 JACCCUGG @    @@
AACUACCAA	AGCUGGAGI	UACGGUAGGO	GCAGAGGG	AAUUUCCO	GUGGAGCO	GUGAAAUG	CAUUGAG	AUCGG	AAAGAA(   660	CACCAACG	GCGAAA	AGCACUCUGO		ACACU 7	GACACUGA	GAGACGA	AAGCUA	GGGGAGC	AAAUGGG   730	AUUAGA	JACCCCAG
81	10	820	830	840	85(	)	860	87(	0	880	89	90 9	900	910	9:	20	930	9	40	950	
UAGUCCACO	 GCCGUAAA(	 CGAUGUCGAC           @@@	 :UUGGAGGU  @  @ @	 UGUGCCCL @@@@@@	JUGAGGCGI  @  @@	 JG - GCUUCC     @@  @  @	 GGAGCUA  @	ACGCGI	 UUAAGU(               (	 CGACCGCC @@@	UGGGGZ	 AGUACGGCC(           @@	 GCAAGGUU/	 AAAAC ₽    @	UCAAAUGA	 AUUGACG 	 	CGCACAA	 GCGGUGGJ	AGCAUG	JGGUUU
		CGAUGGAUAC	ÚAĠĠUĠĊÚ	GUGCGACU				ÁĊĠĊĠĭ		AUCĊĊĠĊĊ	UGGGGZ			GÀÀÀĊ			ġġġġċċ	ĊĠĊĂĊĂĂ   0		AGCAUG	JĠĠŬŬŬ 
960	970	980	990	100	00 2	1010	1020	1(	030	1040	1	L050	1060	10	70	1080	109	0	1100	111	) D
AAUUCGAUG	 GCAACGCG2	 AAGAACCUU <i>I</i> 	 .CCUGGUCU   @  @	UGACAUCO	 CACGGAAGI  @@@   @@	 JUUUCAGAG. @@ @@@  @	 AUGAGAA   @@@@@@		 CUUCGG(  @	 GAACCGUG         @@@	AGACAC	 GGUGCUGCAU	 JGGCUGUC(	GUCAG	UCGUGUU	 GUGAAAU    @   @@	IGUUGGG	 UUAAGUC	 CCGCAAC( @	GAGCGC.	ACCCUU
AAUUCGAUG		AAGAACCUUZ	CCAGGGCU	UGACAUGO	CGCGAAU	CUCUUGAA	AGAGAGO	GGUGCO		GAACGCGG	ACACAC			GUCAG			iguuggg		UCGCAAC	SAGCGC.	ACCCUC
1120	1130	1140	94 1150	11	950 160	980 1170	1180	:	980 1190	1200	0	1210	1220	1	230	1240	12	50	1260	12	70
AUCCUUUGU	JUGCCAGC		GGAACUCA	AAGGAGAG		 GAUAAACUG	 GAGGAAG	GUGGGG	 GAUGACO	 GUCAAGUC	AUCAUC	 GGCCCUUACO			CACGUGCU	 ACAAUGG	CGCAUA	CAAAGAG	 AAGCGAC	CUCGCG.	 AGAGCAAG
GUGUUUAGU	JUGCCA (	CUAUGAGUUL	IGGAACCCU	GAACAGAC	CGCCGGU	GUUAAGCCG	GAGGAAG	GAGAG	GAUGAG	GCCAAGUC	AUCAUC	GCCCCUUAU	GCCCUGGG	CGACA	CACGUGCU	ACAAUGG	GCGGGA	CAAAGGG	UCGCGAU	CUCGCG.	AGGGUGAG
1070	108	1300	1310	1100	1110 320	1120	1 1340	.130	114 1350	40 1360	1150	1160 1370	11	/0	1180 1390	11 1400	.90	1200 410	121	1.0	1220 430
CGGACCUCA		CGUCGUAGUC	CGGAUUGG	AGUCUGCZ	 AACUCGACI	 JCCAUGAAG	UCGGAAU	ICGCUA	 GUAAUCO	 GUGGAUCA	G-AAUO	 GCCACGGUGA	AAUACGUU	CCCGG	GCCUUGUA		CCCGUC	 ACACCAU	 GGGAGUG	GUUGC.	
CUAACUCCA	AAAAACCCO	GUCCUCAGUU	ICGGAUUGC	AGGCUGC <i>I</i>	ACUCGCCI	JGCAUGAAG	CAGGAAU	ICGCUA(	GUAAUC	GCCGGUCA	GCCAU	ACGGCGGCGA	AUCCGUU		GCCUUGUA	CACACCG	CCCGUC	ACACUAU	AGGAGCU	GCCAG	JUUUGAAG
1230 1440	124	40 12 1460	14	1260 70	1270 1480	1280	1	.290	130	15	1310 20	1320	13:	30	1340	13	50	1360	13'	70	1380
UAGGUA - GC		UCGGGAGGGG	GCUUACCA	l CUUUGUG#		CUGGGGUGA	AGUCGUA	ACAAG	GUAACCO	GUAGGGGA	ACCUG	GGUUGGAU		JA							
UCAUUACCC	CUUAACCG	@@      @ UAAGGAGGGG 	GAUGCCUA	200000000 AGGCUAGO I	9@ @@@   GCUUGCGA( 	@     CUGGAGUGA 	AGUCGUA	ACAAGO	@   GUAGCC(	@@    GUACUGGA 	@@    AGGUG( 	@      CGGCUGGAU( 	CACCUCCU	<mark>@</mark> זע ו							
1390	140	00 14	10	1420	1430	1440	1	450	140	6 Ó	1470	1480	149	θÓ							

#### Annotated secondary structure plot for *E. coli* 16S rRNA.



# Modern approach. A no-brainer alignment.

# The twenty 5S rRNAs shown below all have the same length. Alignment is simple. No gaps are introduced.

GCUUACGGCCAUACCAGCCUGAAUACGCCCGAUCUCGUCCGAUCUCGGAAGCUAAGCAGGGUCGGGCCUGGUUAGUACUUGGAUGGGAGACCGCCUGGGAAUACCAGGUGCUGUAAGCUU GCUUACGGCCAUACCACCCUGAGCACGCCCGAUCUCGUCCGAUCUCGGAAGCUAAGCAGGGUCGGGCCUGGUUAGUACUUGGAUGGGAGACUGCCUGGGAAUACCAGGUGUUGUAAGCUU GCUUACGGCCACACCCAACCUGAGCAAGCCCGAUCUCGUCUGAUCUCGGAAGCCAAGCAGGGUUGGGCCUGGUUAGUACUUGGAUGGGAGACUGCCUGGGAAUACCAGGUGUUGUAAGCUU GCCUACGACCAUACCACCAUGAGUAUACCGGUUCUCGUCCGAUCACCGGAGUCAAGCAUGGUCGGGCCGGGUCAGUACCUGGAUGGGUGACCGCCUGGGAACACCUGGUGUUGUAGGCCU GCCUACGACCAUACCACCAUGAAUACACCGGUUCUCGUCCGAUCACCGAAGUUAAGCAUGGUCAGGCCGGGUCAGUACCUGGAGUGGUGACCGCCUGGGAACACCCGGUGUUGUAGGCCU GUCUACGACCAUACCACAAUGAACACACCGGUUCUCGUCCGAUCACCGAAGUUAAGCAUUGUCGGGCCAGGAUAGUACCUGGAUGGGGGACCGCCUGGGAACGCCUGGUGUCGUAGACUU GUCUACGGCCAUACCACCGGGAAAAAACCGGUUCUCGUCCGAUCACCGAAGUCAAGCCCGGUAGGGCCAGGUUAGUACUUGGAUGGGUGACCGCCUGGGAAUACCUGGUGCUGUAGACUU GCCUGCGGCCAUACCACGUUGAAUGCACCGGUUCCCAUCUGAACACCGAAGUUAAGCAACGUCGGGCCAGCUUAGUACCUGGAUGGGUGACCGCCUGGGAAUCGCUGCUGCAGGCUU GCCUACGGCCAUACCACGUUGAAAACACCGGUUCUCGUCUGAUCACCGAAGUUAAGCAACGUAGGGCCUGCCCAGUACUUGGAUGGGUGACCGCCUGGGAACAGCAGGUGUUGUAGGCUU GCCUAGGACCAUAUCACGUUGAAUGCACCGGUUCUCGUCCGAUCACCGAAGUUAAGCAACGUCGAGCCCGGUUAGUACUUGGAUGGGUGACCGCCUGGGAAUACCGGGUGUUCUAGGCCU GCCAACGUCCAUACCAUGUUGAAUACACCGGUUCUCGUCCGAUCACCGAAGUCAAGCAACAUCGGGCGUGGUCAGUACUUGGAUGGGUGACCGCCUGGGAACACCACGUGAUGUUGGCUU GAUAGCGUCCAUACCACACUGAAAACACCGGUUCUCGUCCGAUCACCGCAGUUAAGCAGUGUCGGGCCCAGUUAGUACUUGGAUGGGUGACCGCCUGGGAAUACUGGGUGUCGCUACCUU ACCAACGGCCAUACCACGUUGAAAGUACCCAGUCUCGUCAGAUCCUGGAAGUCACACAACGUCGGGCCCGGUCAGUACUUGGAUGGGUGACCGCCUGGGAACACCGGGUGCUGUUGGCAU GUCGACGCUCAUACUAGGUUGGGUCCACCCGAUCUCGUUCGAUCUCGGCAGUUAAACAACCUUAGGCCUCGUUAGUACUUGAAUGCGUGAGCGUCUGGGAAUACGAGGUGUGUCGACUU GUUGUCGGCCAUACUAAGGUGAAAACACCGGAUCCCAUUCGAACUCCGAAGUUAAGCGCCUUAAGGCUGGGUUAGUACUAAGGUGGGGGACCGCUUGGGAAGUCCCAGUGUCGACAACCU GUUGGUGGCCAUACUAAGCCUAAAGCACCGGAUCCCAUUCGAACUCCGAAGUUAAGCGGCUUAAGGCGAGGUUAGUACUAAGGUGGGGGACCGCUUGGGAAGUCCUCGUGUUGACAACC GUUAUCGGCCAUACUAAGCCAAAAGCACCGGAUCCCAUUCGAACUCCGAAGUUAAGCGGCUUAAGGCAUGGUUAGUACUAAGGUGGGGGACCGCUUGGGAAGCCCAUGUGCUGAUAGCUU GCUAUCGGCCAUACUAAGCCAAAUGCACCGGAUCCCAUCCGAACUCCGAAGUUAAGCGGUUUAAGGCCUGUUAAGUACUGAGGUGGGGGACCACUCGGGAACUUCAGGUGCUGAUAGCUU 

#### Then what? How to find BPs conserved by compensatory mutations?

#### **Mutual information content**

- M(i, j) = "mutual information" between columns *i* and *j*. It measures the "degree of correlation".
- A large M(i, j) suggests that  $r_k(i) \cdot r_k(j)$  exists for all (or most) k between 1 and m.
- Base pairs that are 100% conserved yield **no** mutual information.
- M(i, j) is the "relative entropy" between a pair of probability distributions. If  $f_{i,j}(B_1, B_2)$  is the observed frequency of the base pair,  $B_1 \cdot B_2$ , in columns *i* and *j*, and if  $f_i(B)$  is the observed frequency of *B* in column *i*, then

$$M(i,j) = \sum_{B_1,B_2 \in \{A,C,G,U\}} f_{i,j}(B_1,B_2) \log_2 \frac{f_{i,j}(B_1,B_2)}{f_i(B_1)f_j(B_2)}.$$

Comment: The sum of  $f_{i,j}(B_1, B_2)$  over all pairs and the sum of  $f_i(B)$  over all bases may be < 1, since gaps are ignored.

### Mutual information plot for 20 Eukaryote 5S rRNAs



#### **Twenty sequences perfectly aligned**

- Only (4+4+2+4=14) out of 39 (40 with non-canonical U · U) base pairs are identified (35%)
- There is a fair amount of "noise".
- 100% conserved base pairs not shown. They greatly add to the "noise", but can be useful to "fill in" or extend stems (helices). Total of 86 base pairs in plot.

#### Comparative model for Bombyx mori 5S rRNA



BPs with MI  $\geq 0.6$  are annotated in color. Its free energy can be used to compare with the minimum energy folding.

### Alignment of 302 5S rRNAs from Eukaryotes (M. Szymanski)

A_kurodai	6-9CU-AC93CCAUACCAC-6-U-U5-AACCUACC03UUCUC0UCC-6AUCACC0A-AGUCA3CAAC0UC03GC0U5-GU-UAGUACUAGU03GUGACC-9CUU03GAACACCAC0UGCU04A63CG	G_compres A-CAU-UCGGCCAUAUCAA-GUU-GA-CAAACGCCGUAUCCCAUCCC	U
A_rufus	IUU - ACGGCCMACCAC UU - AMAIIACCGGUUUUUUUUUUUUC- GAUCACCAA- AGUCAACCAA- CGUUCAGUUUGUUUUAAUGGUUACCA CGUUUGGUAACCA CGUUUUUUUAAUGACCA CGUUUUUUAAUGACCA CGUUUUUUAAUCA CGUUUUUUAAUCA CGUUUUUUAAUCA	C. crispata = -A-CNU=U0302CNUAC0A-G-AC0A-CNAUACCCCUCCCUCUGAAC0902C-301UAA010CCCCCG-CU-UNAUAC0A00400-AC1C090AN07C	J
I_illeceb	G-CUU-ACGGCCAUAUCAC-G-C-UG-AAUACACCGGUUCUCGUCC-GAUCACCGA-AGUUAAGCAACGUAGAGCCUA-GU-CAGUACUUGGAUGGGUGACC-GCCUGGGAAUACUAGGUGCUGUAAGCAU	H_sanguinG-UCU-ACGACCAUACCAC-G-U-UG-AAAACACCGGUUCUCGUCC-GAUCACCGA-AGUUAAGCAACGUAGGGCCUG-CC-CAGUACUUGGAUGGGUGAC-GCCUGGGAACAGCAGGUGUUGUAGACU	JU
<pre>0_vulgaris S_officina</pre>	C = 0 - A C = A C = A A A A C A C = C = 0 - A A A A C A C = C = 0 - A A A C A C = C = 0 - A A A C A C = 0 - A C = 0 - A C = 0 - A A A A C A C = 0 - A C = 0 - A C = 0 - A A A A C A C = 0 - A A A C A C = 0 - A C = 0 - A C = 0 - A A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - A A A C A C = 0 - 0 - A A A C A C = 0 - 0 - A A A C A C = 0 - A A A C A C = 0 - 0 - A A A C A C = 0 - 0 - 0 - A A A C A C = 0 - 0 - 0 - A A A C A C = 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0	$H_{a} mines = -6 - CC + COSCCMACCK - 3 - 0 = - A A A A C C COSUUC GUC = A A C A C C A A A G A C C A A A A A C C C A A A A$	JU UU
L_polyphe A diadema	G-UCA-UCGUUCAUACCAC-G-U-UG-ANACCACCGGUUCUCGUU-GAUCCCCGA-AGCUNAGCAACGUUCGGUCCG-UJ-UNGUACUAGGAGGUAAC-ACCUGGGAAACAC	$\texttt{H}_{\texttt{coulata}} = -6 - \text{CCU} - 6036 \text{Coularch} - 6 - 0 - 0 - 0 - \text{Augaccouggeutcouggeut} - 6 - \text{Augaccouggeut} - 6 - \text{Augaccouggeut} - 6 - \text{CCU} - 6036 - 6 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - $	30
E_califor	$ C C - k \cdot C G - C C - k \cdot C - k \cdot L - k \cdot M (C - K - C - G - K - K - C - C - K - K - K - K - K - K$	$D_{misskis} = -0 0 + KGAUCAUACUUS-0 - C-0 - UAACACCCGUUCUCAACC-AACCACAGAUAUAACGUULUAGCUUG-UGU-UAUACUUGUUGUCUGUUGUC-GACCUGGAAUCA C$	J
L_oceanica			d====
H_gammarus C_pagurus	0-102-hC00CCL0ACCC- $0-U-00$ -hAAACCCCCUUUUCUUCUC-C-0AUCACC0A-hAUUAACAAUU000UUU-0U-UHAUACUU00AU000U0ACC-0CCU000AACACCA0AU0CUU0U00CUU	$\begin{array}{llllllllllllllllllllllllllllllllllll$	J U
Spirobolus	-a-010a+k000C040k040a-0-U-100-kMackac0a00000000000-aM00k0000-b000000000000000000000000000000	= CUU - KGACCMUUCAC U - U - ANNCACCMUCUCCCGUCC-GUCUMCA-AGUUAACACGUUANNCN-GU- U IAGUACUGGAAUCG-GCUUGGAAUCC	J
D_melanog			00
A_domestic A magnolia	G-CCA-ACCUCCAUDCCAC-G-U-UG-ADACACCGUUCUCCUCC-GAUCACCGA-AGUDAGCAGCCUCGGGCGCG-GU-UAGUACUUCGGUCGGGGAC-GCUGGGAACCCC-=-GCGUGACGUUGGCAU-= G-G-G-A-ACCUCCAUDCCAC-G-U-UG-ADACACCGUUCUCCUCC-GAUCACCGA-AGUDAGCGGCGUCGGGCGGG-GU-UAGUACUUCGGAUGGGUCGGCGC-GCUGGGAACCCC-=-GCGUGACGUUGGCAU-=	$ \begin{array}{llllllllllllllllllllllllllllllllllll$	JU
P_cynthia	G-C2-kCUICCUIKCC-O-U-U-MANACKCOJUUUCCUIC-GAUCKCQA-kUUAACGU-020CCGG-U-CAUKUUU00000000000-C-ACU000AACACCGCUU0000U000000-C- 	R_pachypt	JN
B_moril B_moril			A
B_mori3 A pernyi	O-CCA-ACGUCCAUACCAU-O-U-U0-ANIACACCOGUUCUCCGUCC-GAUCACCGA-AGUCAACGACAUCOGGCGGA-GU-CAGUACUUGGAUGGGGGGACC-GCCUGGGAACACUACGUGAUGUUGGCCU	$ \begin{array}{llllllllllllllllllllllllllllllllllll$	JU (III
T_molitor	G-ACA-ACGOCCAUAGCAC-G-C-UG-UAAAUACCOGUUCUCGUCC-GAUCACCGA-AGUCAAGCGUUGGCGCG-GU-UAGUACUUGGAUGGUGACC-GCUUGGGAACACCGCGUGCCGUUGUCAU	japonic2G-UCG-ACGCUCAUACUAG-G-U-UG-GGUACACCCGAUCUCGUUC-GAUCUCGGAAUGUUAAGCAGCCUGAGGCCUC-GU-UAGUACUUGAAUGCGUGAGC-GUCUGGGAAUACGAGGUGGUGUGGACU	JU
H_rufipes A_vulgaris	0 + 0 + - A COCCALOUGHC + - 0 + - 0 + - 0 A A A A A C COSUUCE + A U + A C C A A U + A	P_TetTetuta =AUG-AUG-COURCEC-A-C-UG-AAACACCUUUUUUUUUCC-GAUCACUA-AUGAAACACUUUUUUUUUU	JU UU
A_pectini H_nulcher	0-UUU-ACGACCAUACUAC-0-U-U0-AAUACACCOGUUCUCCOUCC-GAUCACCGA-AGUUAACGACGOGOGUUGC-GU-UAGUACUUGGAUGGAGAGACC-GCCUGGGAAUACCXGGUGGUGGUGGGCUUGCCUUGCCUU- 	$\label{eq:linear} \begin{split} & auxital & -e-CCU-ACGACAUACCAC-C-A-UG-AMUCACCGGUUCUCGUCC-GAUCACCGA-AGUUAGCCGGCGG-GH-CAGUACGGUGGGUGGCGACCC-GCUGGGAGACACCOGUGUUGGAGCCG-GUCCGGUGGGGGCGG-GH-CAGUACGCCG-C$	2U
P_depress	G-CCU-ACGACCAUACCAU-G-C-UG-AAUAUACCOGUUCUCGUCC-GAUCACCGA-AGUCAAGCAGCAUA00GCCCG-GU-UAGUACUUGGAUGGGAGACACC-GCCUGGGAAUACC	L_japonicaG-UCU-ACGCCAUACCAC-C-G-GG-AAAAAACCGGUUCUCGUCC-GAUCACCGA-AGUCAAGCCCGGUAGGGCCAG-GU-UAGUACUUGGAUGGGGGAACC-GCCUGGGAAUACCUGGUGCUGUAGACU	JU
S_kowalev1 S_kowalev2	G-CCU-AC03CCAUGCAC-G-U-AG-AD05CACC0000CC00CC-GAUGACC0A-A0UAAGC00C00C000G-G0-UAUACU00GAU00AAACC-G0C00GAAUACCACG00C000A6C0- G-CCU-AC03CCAUACAC-G-U-AG-AD05CACC00000C000CC-GAUGACC0A-A0UAAGC00C003GC00G-G0-UAUACU00GAU03GAACC-G0C00GAUACCACG00C000A6C0	$\mathbb{C}_{quinque}$ ,G-CU-ACBACCBARCAC-A-A-U-GARDACCBUUCUUCU-GARCACCB-ABUAACABUGBCBUCCBU-GU-CABARCCBUBBUBAC-GCUBBACACC	20 UU
H_roretzi	UCU-ACG9CCAUBCCACU-UG-ACB6CACCCCAUCUCCUUC-CAUCUCCUA-ACUUBACAUGUAGGAUCA-CU-UBCULGGAUGGGUGAC-CCCUGGAUBLCU	Sealtatr = -0-UCI-ACGCCAUACCAC-0-N-UG-ANUACACCGUUCUCGUUCC-GAUCACCGA-AGUUAGCCUUGGUUCGGUCCAGA-UAGUUAGGUGGGCCAG-GAUGAGUUCUGGUUCGGUUC	300-
E_japonic2	- 02 - CCL ACARCANUCKC- C- C- 09 - AND CCCCCOMUCCANCE AND COMA ACCUMACADOUCAACCUM- 09 - UNDIACUUCANORANCE - ACCUMUNANCE - A - ODU UNDIACUUCANOCUM A - ODU UNDIACUUCANOCUM A - ODU UNDIACUMACUMACUM A - ODU UNDIACUMACUMACUM A - ODU UNDIACUMACUMACUMACUMACUMACUMACUMACUMACUMACUM	L_NLIMI = -DUU ACGCCMUACKU-U-SAAGCCCCGCCCUCUCUCUC-AUCACCA-AGUAACAACGUCCGGCCCC-GU UAGACUDGAUGGGCCCC-CCCUGGAUACACCGGGGGCGGUGAACAC	JU
<pre>s_japonici S_canicul</pre>			00
M_fossilS	6-CUU-ACG9CCAUACAC-C-C-UG-AGCACGCCGAUCUCGUCC-GAUCUCGUA-AGCUAGGGGGGGGGG	=-6-UCI-ACGACCAUAUCAC-6-U-UG-AAAACACGGUUCUGGUCC-GAUCACCGA-AGUUAACGA-AGUUAACGCAGCCGA-GGUUAGGAUGGAUGGAUGACC-GCCUGGAAUACCGGGUGCUGUAGACU	JU
T_tincaS		MedulisG-UCU-ACGACCAUAUCAC-G-U-UG-AAAACACCCGUUCUCGUCC-GAUCACCGA-AGUUAAGCAACGUCGAGCCCG-GU-UAGUACUUGGAUGGGGGGACC-GCCUGGGAAUACCGGGUGUUGUAGACA	A
T_tincaO X_boreal2	-d-CU0-ACOCCALACCAC-C-U-0-AAACOCCCAAUCCCAAUCCUCU-AUCUCUUA-AUCUAGAGGUUCGGCCUG-GU-INGUACUUGGUUGGAGUAGC-GCCUGGGAAUACCAGUUGCUUUAACUU	$\mathbf{R}_{\text{polymor}} = -G_{\text{abu}} - Gradue Gradue Cardee G_{\text{constructed}} = -G_{\text{constructed}} = -G_{cons$	A
S_oshimae X laevisl	0-UUU-ACGACCAUDICAC-0-UI-U0-ANDAUCCOSIUCUC-GUICACCGA-AGUCAACCGA-AGUCAACCGACCGACCGACCGACCGACCGACCGACCGACC	$ L_{\rm beterop} = -6 - 6 {\rm AU} - 6 {\rm CO3G}({\rm CU}) {\rm ACC} - 6 {\rm AC} - {\rm CU} + {\rm CU} {\rm ACC} {\rm CO3G} - 6 {\rm AC} - {\rm AG} {\rm AC} {\rm CU} {\rm AC} {\rm CC} - {\rm AC} - {\rm AG} {\rm AC} {\rm CU} {\rm AC} {\rm CC} - {\rm AC} - {\rm AG} {\rm AC} {\rm AC} {\rm AC} {\rm AC} - {\rm AC} - {\rm AC} {\rm AC} {\rm AC} {\rm AC} - {\rm AC} - {\rm AC} - {\rm AC} {\rm AC} - {\rm A$	1
X_laevis2		P_nudum =-d030U-0C030UAUACCAC-C-G-UU-AAU0CACC03AUCCC0UCG-GAACUCC0U-AGUUAA0C0CC0UUG03CC0G-AA-UAGUACU003AU03GUGACC-UCCC003AAGUCCCG3UGCC02CCCA	AC
X_1aevisi X_tropic3	-d - CC + ACOCCALCCACC+ C - C - 0 + ALACCOCCALUCCCALUCCCUA + ALCALUCACUA + ALCALUCACUA + 0 - ULAUACCUA ALUGAALUACC - C - CUGGAALUACC 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0	$D_a cum na \\ - 0 0 0 0 - 0 0 0 0 0 0 0 0 0 0 0 0 0$	JU UU
B_taurus X tropic2	UU-ACGCCAUACCA-C-C-CU-AACCGCCCAUUUCGUU-AUUUCGGA-ACCUAGGAGGUUGGCCUA-GU-IUAUAUUUGAUGGAGACC-C-CUGGAAUACC	$\mathbb{E}_{avense} = -0000U - 0000U - 000U - 0000U - 0000U - 0000U - 0000U - 0000U - 0000U - 0000$	2 0C
X_tropic1	G-CU-ACG9CNUACCC-C-C-UG-ANACCCCCQUCU-GUCU-GUCUCQA-AGUNACG003CACG9CUGCU-GU-UAGUACCUGAUG9A0ACC-GCCUG9ANALACC	G_bilbaG_0GU-GGAU-GGAU-GGAU-GGAU-GCAULACGACCCAUCA-GACUCCCA-GUUAGCACGUUGGCUUG-AG-UAGUAGUAGCAC-UCCUGGGAUCCCAGUGUUGCACCCU	JC
T_carolina			00
G_gallus1 G_gallus2	G-CCU-ACGGCUNICCCAC-C-CU-G-GUNACGCCCGNUCUCGUCU-AUCUCGA-AGCUNAGCMGGUCGGGCCUG-GU-UNGULCUGGUGGGACCC-UCCUGGANAIACCGGGUGCUGUAGCUUU G-CCU-ACGGCUNICCCAC-C-C-U-G-GUNACGCCCGNUCUGU-AUCUCGA-AGCUNAGCMGGUCGGGCCUG-GU-UNGULCUGGAUGGACCU-UCCUGGANAIACC	E_bolanica =-0-08U-0C0AUCAUACU6C-0-UU-ANUGACCGAUCCAUCA-AAACUCCGC-AGUUAAGCGCCUUGGACUA-AG-UAGUCUGAGUGACCCUCCCGGAUAGUCCUAGUUGUCACCUU	JC CU
P_walt101	G-CCU-ACGGCCNUACCAC-C-C-UG-ANUGGCCUGAUCUCGUCI-GAUCUCGGA-AGCUAAGCAGGUCGGGCCUG-GU-UAGUACUUGGAUGGGAGACU-GCCUGGGAMUACC	S_olariesaG-001-0000U0NUNCCAC-C-A-CT-ANUGCACCOGNUCCCAUCA-GAACUCCOC-AGUUAACCOGUUGUGGGAGA-AG-UAGUACUAGAUGGGGUCC-UCCUGGAAAUGC	2U
P_waitiS P_walt102	-d - CC + ACOCCALACCAC + C - C + 0 + ADVCCCCALAUCCCALUCCCUA + AUCUACUA + AUCUACUACUACUCCUCC- d + IAGDACUUGACUUGACUUGACUUCCUACUCCUACUCCUACUCCUACUCCUCACUCCUCC	$\texttt{M}_{\texttt{J}\texttt{J}\texttt{P}\texttt{F}\texttt{C}\texttt{S}} = -0 - 300 - 6030 - C340 - C - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - $	JU== UUU=
A_castell1	GAU-ACG9CCAUACUGC-GC-AG-AAAGCACCGCUUCCCAUCC-GAACAGCGA-AGUUAAGCUGCGCCAGGCGGU-GU-LIAGUACUGGGGUGGGCCACC-ACCCGGGAAUCCA	J_media =-0-G0I-0C0AUCAUACCAG-C-0-UU-A000CACCA0UCCAUCA-GAACUCOCC-A00UAACGCCCCCGG-AG-UAGUACUCOUGOGGAAGUCACC-UCCCGGAAGUCCCGGUGUUGACCCUU B	JU
Ac_crinita	A-GGA-ACGGCCAUACCAC-G-U-CG-AUCGCACCACAUCCCGUCC-GCUCUGUGA-AGUUAAGCGACGUCGGGGCCAG-GC-UAGUACUACGGUGGGGGGACC-ACGUGGGAAGCCCUGUUGCUGUUUCCU	P_silvestG-00U-0C0AUCAUACCAG-C-G-UU-AAUGCACCCGAUCCCAUCA-GAACUCCGC-AGUUAAACCGCGCUUGGGCUAG-AG-UAGUACUGGGAUGGGGGGGGGG	2
B_japonicu B_vorax	0 - 0 - 0 - 0 - 0 - 0 - 0 0 A A A C A C C A A A C C C - A A C C C C	$\texttt{B}_{\texttt{D}}\texttt{B}_{D$	U
C_campylu C_colnoda	0-CUG-UC00CCUACUA-0-0-U0-ACACACCOAUCC-0AUCCC0A-AUUAA0CCUCA-AUUAA0CUC0ACUA0U0000000000	H_annus =G-00U-GOAUCAUACCA-C-A-CU-AUGCACCGGAUCCCAUCA-GAACUCCGC-AGUAGCGUUGGCGAG-AG-UAGUAGGAUGGGUGACC-CCCUGGGAUGCCUCGIGUUGCAACC	2C
C_fascic	G-AU-ACCACCIACUUG-A-G-UG-AAAACACCAUAUCCCOUCC-GAUUIGUGA-AGUUAACCCCACAGGCULA-GU-UAGUACUGAGGUCAGGUGAUG-ACUCAGGAACCCGAUUGCCULACUCC	L_minor =-0-001-02010/0104-0	J
C_paradox D_discoid	UUD-ACUDEVUALCUC-C	$L_1 \text{Line is} = -A - \text{Div} - \text{div}$	CU
E_gracill Eu gracil2	0-C0U-AC00CAURCUAC-C-0-0G-ANIACACCUONACCCOUIC-GAUUICAGA-AGUIDAGCOGGUCGAGCCCA-GI-IUGUACUGAGUGGOCGACC-ACUUGGGAACACU0GGGUGCUGUACCCUU	Magnolia        0-02U-02URACCAG-C-A-CU-ANUCACCCGARCCCGARCCCA-ARUCAACCUUGACCUUGACGAA-AC-UUGIAUGAGUCACC-UCCUGGAAAUCC	2 C
E_woodruf	CUS-UCGGCCURCURA-C-C-CA-ANUCACCOGAUCCC-URCUCCGA-AUURAGOGUUUAAGCCUG-UI-AUUACUGGGUGGGGGGGCC-AUUGGGAUCC-URCUGGGAUCC-URCUGGGAUCC-URCUGGGAUCC-URCUGGGAUCC-URCUGGAUGAUGGCUUC		2
M_inverta			á
P_tetraure P_polyceph	0-0103-01002-01002-01002-0100000000000	V_faba =A-G0U=G0AUCAUACCA-C-A-CU-AUGCACCGGAUCCCAUCA-GAACUCCGC-ACGUAGCGCUGGGCGAG-AG-UAGUAGGAUGGCGACC-UCCUGGGAAGUCC-U-U-UGGGUGCGACCCU	20 CII
P_berghei	U-0AC-UCQUICAUKCUAC-A-O-U0-OCCCCACACADUCCCUQA-AGUUAC-GAUGUAGUGAGUGAGUGAGUGAGUGAGUGAGUGAGUCC-GCUQQQAAACQ	\$_alba =-0_00U-002UUAC0G-C-A-CU-ANICOCOGNCCCANCA-GACUCCOC-AGUUAC000000A-AG-UAGUAC0A00000000A-C-UCCC0GAA0UCC0GIGUU2CACCC-	
T_paravor			CU
7_thermoph 7_cruzi	G - NUN - UCGGCCAUACUIA - G - G - UG - AAACACCCGAUCCCAUC - GACUICCGA - AGUUAACGCCCUIAAGGCUGG - GU - UAGUACUAAGGUGGGGGAACC - GCUUGGGAAAGUC AGUGUCGAUANCC AGUGUCGAUANCC AGUGUCGAUANCC AGUGUCGAUANCC AGUGUCGAUANCC AGUGUCGAUANCC	S_cereale = -0-GAU=GCAUCAUACCAG-C→-CT-AAAGCACCGGAUCCCAUC→GAACUCCGA-GAUGUAGCGGCUUGGGCGAG-AG-UAGUACUAGAUGGGUGGCC-UCCUGGAAGUCCU-C-CGGGUGCUGGCAUGCAC- S_cereale = -0-GAU=GCAUCAUACCAG-C→-CT-AAAGCACCGGAUCCCAUC→GAACUCCGA-AGUUAGCGGCUUGGGCGAG-AG-UAGUACUAGAUGGGUGGGCGUGGGGAGUCCU-C-U-CGGGUGUGGCAUCCAUCCAU- = -0-GAU=GCAUCAUACUACCA-CT-A-CT-AAAGCACCGGAUCCCAUCCAA-AGUUAGCGGCUUGGGCGGUGGGGGUGGGGU	2C C
S_carestia	$-x_{-} UUU - 6069 \\ C 2 UU \\ C - 2 U \\ C - 2$	B_gisplsx	
T_mesent1	- A - U U - SCIOSCANSING - U - C - U - ANALVACCOGAUCCOGUC - AUCUGACA- AGUADACCOGUCA - U - U AUVACCACOUGOGARCC- ACUGANUCCU	C_1116gtus = 0 -00U-ecocochulation - 0 -AbadeAccouldececoule - Baltean generalization - 0 - Couldean and - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 - 0 -	J
<pre>T_mesent2 G_primul</pre>	0 to $ 0$ to $$	$C_a xyma = -6 - CCU - ACCGO CAUJUCA - U - 6 - CU - AUXAIXCAU UUCCCUUCC - GAUCALCAU DAGUCAU CUU UAACCGU AACCGU ACCGU ACCGU ACCGU ACCGU AACCGU AACCGU AACCGU AACCGU AACCGU AACCGU$	U
T_papilion T_violea	-x-luce-luceaccalacatu-e-x-ce-tataccaccaccaccaccac-aucuscac-aucuscac-aucuscaccaccaturatu-e-e-aucuscaccaturatu-e-au-aucuscaccaturatu-e-e-aucuscaccaturatu-e-e-aucuscaccaturatu-e-e-aucuscaccaturatu-e-e-aucuscaccaturatu-e-e-aucuscaccaturatu-e-eaucuscaccaturatu-e-eaucuscaccaturatu-e-e	$ \begin{array}{llllllllllllllllllllllllllllllllllll$	
U_fusispor	UCC-NCGCCMURGAC-U-U-U-C-NANCACCOCMUCCCGUICC-GAUCIOCCC-NUUMALCOMOUCCCCUL-UI-UIUMUACCACGUIGGGGGACC-NCGCGGGAUCCU	C_mellibing	J
A_eduiis B_adusta	R-UC-ACOCCALIAGAC-UC-UO-ANACCACCCAUCCCOUC-ACUCICC-AUUACCAAUUACCAAUUACCCAUUA-GI-INGUACCACCAUGGODGAAUCCU	$c\_paraph1 = -d_000 - d_00cchalled - dc_a - A A A A C C C 000 C C C A A C A A C A C$	
C_pallida S commune	- h - UIC2 - hCOSC2LUAGRAC - U - C - US - AAACCACCGAUCCCGUICC - GAUCUGCGC - AGUIAACCAGAGUGCCGCCUL - GU - UAGUACCACGUIGGGGAACC - ACGCGGAAUCCU h AGUICCAIGUIGGGGACC - U - C - US - AAACCACCGAUCCGUICC - GAUCUGCGC - AGUIAACCAGAUGCCGCCUL - GU - UAGUACCAGGUIGGGGAACC - ACGCGGAAUCCU h AGUICCAUGUGGGAACC - U - C - US - AAACCACCGAUCCGUICC - GAUCUGGCG - AGUIAACCAGAUGCGCCUL - GU - UAGUACCAGGUIGGGGAACC - ACGCGGAAUCCU h	$C_{2} rego_{2} =G_{2} U - GOU - GOOGCCAUACCA-G-U - A_{2} - AAAGCACCAUUCCCOUCC-GAUCAAUGACGAUAACUACUAAGGUCUG-AUGAGIAGUGAGGAGACC-AUGAGAAUUCAGAUGCUGCAAUCUCCGUCAAUGUCUGCAAUCCCGA-ACCAGGAACCAUACUCUCAUGAGAAUCCUGOU-GOUGCUGCAAUCUCUGUGAGAACC-AUGAGGAGACC-AUGAGGAGACC-AUGAGGAGACC-AUGAGGAGACC-AUGAGGAGACCAUGACUGUGAGAAUCA-UGAUGCUGCAAUCUCUGUGAGAAUCGAUGCUGCAAUCUCUGUGAGAUC$	J
C_radiatus	&-UCC-ACGOCCAUAGAC-U-C-UG-AAAGCACCGCAUCCCGUCC-GAUCUGCGC-AGUUAACCAGAGUGCCGCCUA-GU-UAGUACCACGGUGGGGGACC-ACGCGGGAAUCCUGGGUGCUGUGGUU	Z_helleni =-A-OCU-ACGOCCAUAUCCA-G-A-UG-AAAAUAAGACUUCCCGUCC-GAUCAGUCCUAUUCAAGCAUCUGAGAOCCUU-ACUAAGUACUAUAGUGGGUGACC-AUAUGGGAACUGAAGGUGCUGUAGUU-	
<pre>% semperv</pre> <pre>% d_phoenic</pre>		$\mathbb{R}_{\text{creering}} = -A \cup \mathbb{C}_{\text{constraint}} = A \cup \mathbb{C}_{\text{constraint}}$	
G_applana P ostreat	- h - UIC2 - hCOSC2LUAGRAC - U - C - C - AAACCACCGAUCCCGUICC - GAUCUGCGA - AGUUAACCAGAUGUGCGCCCCA - GU - UAGUACCACGUUGGGGAACC - ACGOGGAAAUCCU G GGUGCCUUGGGUU	$R_{\text{himm}1is} = -\lambda - UCU - 60363 CCUACCC - \lambda - C - U - AAUCACCCCAUCCCUCU - CAUCUCCCA AUUACCACUUGACCC- ACUCUCCCUC- CAUCUCCCACUCUCCCUCUC- CAUCUCCCCACUCUCCCUUC- C - AUUACCACUUCCCCUC- C - AUUACCACUUCCCUUC- C - AUUACCACUUCCCCUC- C - AUUACCACUUCCCCUC- C - AUUACCACUUCCCCUC- C - AUUACCACUUCCCCUC- C - AUUACCACUUCCCUUC- C - AUUACCACUUCCUUCCUUCCUUCCUUCCUUCCUUCCUUC$	
R_cyanoxa	UCC-ACGGCCAUAGAC-UC-UG-AAAGCACCGCAUCCCGUUC-GAUCUGCGC-AGUUAACGAGUGCGCCCCCUGU-UAGUACCACGGUGGGGGACC-ACGGCGGAAUCCU	K_microstA-CNU-ACGGCCACADUCA-G-C-UG-AAAACUGGGCAUCCCGUCCC-GUCUGCCAUACAUAACUAACCAAUGGCAA-AU-UAGUACUGAGGGGGGGGGCAACC-ACGUGGGGAAUCCC	3
S_paradoxa			
C_tussilag C apobasid	- h - 10 C - h COOCCALIAGGAC - C - U - 0 - h A h A CACCOCAU CCCOII C - G A UCIGACO - h GUIDACCAGGUIDAGGUIA - GU - I M GU h A CCACGUIDAGGUAL - C - L A GUIDACCAGGUIDAGGUAL - C - L A GUIDACCAGGUIDAGGUAL - C - L A GUIDACCAGGUIDAGGUIDA - GUIDACCAGGUIDAGGUAL - C - L A GUIDACCAGGUIDAGGUIDA - GUIDACCAGGUIDAGGUAL - GUIDACCAGGUIDAGGUAL - C - L A GUIDACCAGGUIDAGGUIDA - GUIDACCAGGUIDAGGUAL - GUIDACCAGGUIDAGGUAL - C - L A GUIDACCAGGUIDAGGUIDAC - A CUIDACCAGGUIDAGGUIDAC - A CUIDACCAGGUIDAGGUAL - C - L A GUIDACCAGGUIDAGGUIDAC - A CUIDACCAGGUIDAGGUIDAC - C - L A GUIDACCAGGUIDAGGUIDAC - L A GUIDACCAGGUIDAGGUIDAC - C - L A GUIDACCAGGUIDAGGUIDAC - L A GUIDACCAGGUIDAGGUIDAC - C A GUIDACCAGGUIDAGUIDAC - C A GUIDACCAGGUIDAGGUIDAC - C A	$C_{\rm humicol} = -A - UCC - ACOGCCLUAGGAC - U - C AG - AAAADACCCCGAUCCCOUCC - GAUCUGCGA - AGUCUACUACUAGUACCCCUA GU UAGUACCAUGUGGGGGGAC - ACADGGGAAUCCU GGUGCGUGGGUU G GGUGCGUGGGUU GGUGCGUGGGUU GGUGCGUGGGUU GGUGCGUGGGUU GGUGCGUGGGUU GGUGCGUGGGUU GGUGCGUGGGUU GGUGCGUGGGUU GUGCGUGGGUU GUGCCUAGUU GUGCGUGGGUU GUGCGUGGGUU GUGCGUGGGUU GUGCGUGGGUU GUGCGUGGGUU GUGCGUGGUU GUGCGUGGUU$	
G_clavaria R malvinel	$- x_{-} U c C_{-} A c O C C C C C C C C C C C C C C C C C C$	$P\_\text{partim} = -A-UCC-ACGGCUAGGAC-C-U-UG-AAAACACCGCAUCCCGUCC-GAUCUGCCC-AGUUACCAGGGUGCCCCCUA-GU-UAGUACCAGGUGGGGGAUCC-ACGCOGGAAUCCJ=-A=GGIGCUGIGGUU-D-AGGIGCUGIGUU-D-AGGIGCUGIGUU-D-AGGIGCUGIGUU-D-AGGIGCUGIGGUU-D-AGGIGCUGIGGUU-D-AGGIGCUGIGGUU-D-AGGIGCUGIGGUU-D-AGGIGCUGIGGUU-D-AGGIGCUGIGGUU-D-A$	 (l
G_sabinae	A-UUU-GGGGCCACACCAC-A-G-UG-AACUCACCGCAUCCCGUCU-GAUCUGCGC-AGUCAAUCACUGUGGGGCCUC-GU-AAGUAGUAGUGGGGGGACC-AACUGCGAACACGAU-GGUGCCCCAGGU	P_rhodory A-UCC-ACGCCAUAGAN-A-U-CC-AAAGCACCGCAUCCCGUCC-AAUUACCG-AGUUAACCGAUAUGCCGCCUA-GU-UAGUACCACGGUGGGGGACC-ACGCGGGAAUCCUGGGUGCUGUGGGUUG	
P_suaveoi S_salmonic	R-1010-6000CCLACCCC-A-0-0-AACCC0CCCCC0UUCGUC-AUC00CC-AUC0ACUC-AUC0AUCA-00-CAULADUCG000CCA-AU-CAULADUCG00000AUC-AUC0CAUAAUCUU-0000CCCCCAUAAU	S magning = -AACCA-UCROCCAUACAUA-C-C-AG-AAAAUACCCUAUCCCUCC-GAUCOCCUA-BUCAACGUAUGAUGCCUA-GU-CAUACGUBUGUGUCC-ACGUAUGAUCCAC-GUGCUAUGUUGUUGUUGUUGUUGUUGUUGUUGUUGUUGUUGUUG	J U
U_scabios S polygoni	$- x_{-} UU - c c c c c c - x_{-} U - x_{-} U U - x_{-} U - x_{-}$	C_glegangAMUU-ACOGCACAUCAU-C-0-UG-AMAGACCAAUCCCAUCC-GAUCUUGC-AGUCAUCCAGUCGAGUCG	J
R_toruloid		P_blaks1ENNVU-KO3CCNIACMA-U-A-GI-AKCKCCGGANCCCCUCU-GAICUCCCCCGGAIAGC-CU-CAGINCCUNGGOGGAGAC-C-CAUGGANACGCUUGUGGGAAACGCUUGUGGGAAACGC	J
S_vulgare	- A-GUD-UIGACCACACULA-G-C-GA-CALACACUGGGGCCCGUCC-GUICCGCA-ADUCALUCCCULADGUCAC-UIG-AGUICCGALADUCGGALAGUICGGGAL-AUCCGALADUCGGALAUCCCULADGUCAC-UIG-AGUICCGALADUCGGALAUCGGCALADUCGGALAUCCGALAUCCGALAUCCGALAUCCGALAUCCGALAUCGGALAUCCGAUCCG	C_SEGIAT GUCUU-ACQCCAUUCACA-C-C-AG-AAAGCACCAAUCCCOUCC-GUCUUCAA-AUUUAACCUUGAUGACUUC-UU-L-CUGLCUUGUUGUUGUUGUUGUUCACA-CACACCUUAAUUCCCUAAUCCCUAAUCCCUAAUUCCCUAAUCCUAAUUCCCUAAUCCCUAAUCCCUAAUUCCCUAAUCCUAAUUCCCUAAUCCUAAUUCUUC	
L_pyritorm P_poarum			U
G_phoenic T_controv		$D_{accuration} = -\lambda - GUIV - ACOACCUMICAA - C-U-UG - AMAGCACCAACCCCCURCA - ADIUGUIGAA - AGUUMAGCAGUUG-GUIU-GUI-UUGUIGUIGUIGUIGUIGUIGUIGUIGUIGUIGUIGUIGU$	J U
T_anomala		A paragitiG-GUU-ACGGCCAUACGUA-G-C-AG-ALAGGACCCCAUCUCGUCC-GAUCUGGGA-AGUUAACGUCUAGGGGGGGGGG	J
o_nordei G_primul		imitate ===================================	du
U_maydis F thuemen		A_chrysogsA-CAU-ACGACCAUACCUA-G-U-GG-AAAAUACGGGAUCCCCUC-CCUCUCCCAUAGUCACCUCAAGGGCUC-AU-UAGUAGUGGGUGGGGACG-ACCAGGAAUACCGCCUCUUGUAUGUGGUGGGCGACG-ACCAGGAAUACCGCCUCUUGUAUGUGGUGGGCGACG-ACCAGGAAUACCGCCUCUUGUAUGUGGUGGGGCGACG-ACCAGGAAUACC	
G_primul			
M_acetabut			
r_oedoceph S_aggrega	-*- UCC-uCC-uCutractackanda-u-u-u-uu-aaasta-uco-aaasta-uco-aabtulaccaa-aduulaaccaa-aduulaaccaa-aduulaacaa-au-au-aduulaacaa-au-aduulaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-aduulaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-aduulaacaa-au-au-au-au-au-au-au-au-au-au-au-au-a	A_nigulaiA-Cub-ActuAccubacgu-g-u-g0-AdAACAggetUteContecCetCACCOACCUBACUBACCAACACGGUAG-GI-UbgUAGGUGGGGGGACC-AAUGCAAUCCCUACUUUGUAGG	
T_visurgen		$\label{eq:rescaled} \begin{array}{l} n_1(a_1) = -aCall - ACOACCOMBOGUE-G_0-U_0GACOAACCOMBCUTUCCOCCUCC-C-CACCOACCOACCUCCACUCCACUCCACU$	
C_reinha2		H_jadiniiG-GUU-GCGGCCAUAUCUA-C-C-AG-AAAGCACCUUCUCCUCCCCCAUCAUCUAUGUAACCUCCUAGAGCCUG-UCGAUGAUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGU	J
Chiorella Chlamydom	-AUGU-ACUUCAACUUCAACUACCA-C-A-CU-AAAGCACCOAUC-CAUCA-AAACUCAAACGUGAUGAGCUCC-AC-UAGUACOAGUUUAAGCAUU-ACCUGGAUAACCC	$x_{\text{microst}} = -a - \text{cm} - a $	J U
C_pyrenoi		L_11pofer>-00U-0030C01JaC0A-0>A-AANOCACHUUCCC00UCC-0AUCAACAUGUUUU0000CC0A-UU-UUU0000000000ACC>COCU00AUC07	
K_flaccidu		R_EPUELICo A-CAU-ACAACCAUAACU-A-A-NO-ACAAUU020CAUCCCOUCC-GCUCUCCCAUACACAAUU02000CAUCCCCU-U-U-UU01AUU020000000000000000000000000000000	
<pre>n_riexiis C_monilig</pre>			·
P_minor P_subcord	DIQUIGCUUCULTUCULC-U-U-COAAACACCCOQUICCCAUCA-CAACUUCQGA-AGUUAAGCGGAUUQCCCGA-UCUGUUACUGAUUQGUUACUGGAUCA-C-UCOGGAAUCAC	$\label{eq:product} p_{max} = -0-GUU-OCOCCALAUCUA-O-C-AAAACACCOUVICCCOUCC-AAUCACURAGUAACUCUAAAACUCCUAAAACCUC-ACCAGUAGUAAAACACCUC-AUAACACACUCAUAACAACUCAAUCUAAGUUCCUAACACAAUCUAAGUUCCUAACACAAUCUAAGUUCCUAACUCAAUCUAAAACAAUCUAAUAAAACUUCUAAAAAACUUCUAAAAAAA$	
Prasinocla			3
S_quadric			
Spirogyra U_pertusa	NUGCI-ACGGICAUACCAC-C-A-CG-ANACCACCGAUACCCAUAC-GAUCUGAC-AGUUAACGIGGUUGGGICCAG-AU-UAGUUCUGGGUUGAGGAUC-ACUUGGGAUCCAC-U	T_abundans        >-0:U-AC007CNLA0AAC-A-C-U-AAAACAC030UUCCC0UCC-04CUCAGCCUA0UUACAAACANN0UUAUCCQAG-GU-UAGUACUM0000300UACC-ACAU030AUUCCUCULAC004AUUCC	
S_grevill		Y_lipolyt	3
C_CONNII D_tenue			
<pre>n_foetidus C_parameci</pre>		$\label{eq:linear} \begin{tabular}{lllllllllllllllllllllllllllllllllll$	
A_lubricum		3. paleber 2>-NUU-MORCCCMACMA-C-C-C-C-ALMACKUCUUUCCCUUCCAUUUCCAUUUCCAUUUCCAUUUCCAUUUCCAUUUCCAUUUCCAUUUUCCAUUUUCCAUUUUCCAUUUUUCCAUUUUUU	 (l
E_bicyclis	- A - GGA - ACGGCCAUACCAC - G - U - CG - AUCCCACCACAUCCCCUUCC - GCUUGUGA - AGUUAAGCGCCUUCGGCCAC - GC- UAGUACUACGGUGGGGGACC - G- U - GGUGCUGUUCCU	B_alba =-A-UCC-ACGGCAUMAGAC-U-C-MG-AMAGUACCCCAUCCCGGUCC-AUCCGCGCA-AGUCAGCGCGUAGUACCGCCUA-GU-UMAGUACCACGGGGGGGAGAUCCUQGGGGCGGGGGG	
s_fulvell C_caldari	A-CUB-ACERCCAURCCAC-9-U-CU-AUCEACCACAUCCCOUCC-ACUUUGAA-AGUUAACOSCCUCOGOGCC03-CC-UAGUACACOGOGOGCC3-CC-UAGUACACOCCGOUGCUGUCCU	U_rusisporA-UCC-ACUUCLAUAUAUA-U-U-U-G-AAAGACOCCAUCCONUCC-AAUCUGCC-AGUUAACCGAUGUGCOCCU-GI-UAGUACACGUUGGGGGGAC-ACGCGGAAUCUGGGUGCUUGGGUU D_stilatA-UCC-ACUUCLAUAUAUA-C-CUG-AAAGACOCCAUCCGOUCC-AAUCUGCON-AGUUAACCGAUGUGGGGCG-G-U-UAGUACACGGUGGGGGGAC-ACGCGGAAUCUGGUGCUUGGGUU-	J
P_palmata P_umbilics	-sh-coll-collecculation-ch-ah-chlococcculation-chlococcality (coll-advited) = dotted and (coll-advited) = dotted	8.albest>-btc-Cx03cCAUA00Ac-U-C-U-C-W-ANACAC020AUCC00UcC-AUUC03CA-AUUAAC0AAUUC00CCC0U-UAUAAC0AC0U000000AUCC>AOC000AUUCU	
P_yezoenzi	- A-C00-AC00CCNINICC0AG-A-CA-COCUNACCCANUCCOANUCCAAUUCCAAUUCCAAUUCCOCAGUUG GI-UAGUAUCUGUGAMAAUC-ACAOOCAACCC	E autorioutak-PEC-ACOSCIUMORAC U C US-IMACACOSCIUCCOUCC-GAUCISCAA AUTORACIACIONCEA-GU-UMURACACIGUIGODODACC-ACOCOGUNUCUCUDODUCUCUUGUIG	
<pre>P_cenera B_ectocarp</pre>		a_valiates ====================================	
G_amansi2 G_amansiS	A-CAU-UURECAUACUU-GUA-GUC-AU-CAURUCUCCOULCCOALCUARCUARGA-AUCUNAUCUCCCUCAGOCCA-GU-AGUIAUUCOSUCAGOZAUG-ACCUGOZAULUC	P_tagines =-A-UUU-0020CUMACUM-CA-UU0-ANAUKUUGAUCCOUUU-0MUCUGAC-AUGUAUAGCACUMAGOACCA-00-AGUAUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGU	

# "Low noise" MI plot



- Most BPs are detected.
- The irregularities in stems is a consequence of embedded gaps in the alignment.
- Alignment BPs must be converted when a particular sequence is extracted and "degapped".

#### "Low noise" MI plot - Bombix mori numbering



Helices appear properly when numbering with respect to a single sequence.

#### Comparative model for *Bombyx mori* 5S rRNA - Update



BPs with MI  $\geq 0.5$  are annotated in color. All base pairs are supported by comparative data.

# **Transfer RNA – tRNA**

- A huge number are known.
- Secondary structure deduced from perhaps 12 sequences in 1969 (Michael Levitt)
- For this presentation, 654 aligned tRNAs were selected from Sprinzl's database

#### Sample entry:





#### MI plot for 654 aligned tRNAs (Sprintzl, 1993)



With 654 sequences, Secondary structure is very well determined using MI. The quality of the alignment is critical.

MI plot + conserved BPs for 654 aligned tRNAs (Sprintzl, 1993)



#### MI plot + conserved BPs for 654 aligned tRNAs (Sprintzl, 1993)



Plotting conserved BPs adds noise. Only one extra BP is discovered. Is it worth it?

### **MI plot & tRNA-TGC Halobacterium cutirubrum**



Halobacterium cutirubrum tRNA-TGC

## **Energy Minimization**

- How is energy assigned? Answer: "nearest-neighbor" energy rules are used.
- A stem with *n* BPs is broken into n-1 "BP stacks". Energy  $\Delta G$  is assigned to the "BP stacks", but takes into account hydrogen bonds and stacking. These energies are negative "favorable".
- Mismatched BPs at the ends of stems also contribute to stability.
- The loops are destabilizing.

#### $\delta G$ for BP stacks

NN (nearest neighbor) free energies for RNA at 37°. Doug Turner's group at the University of Rochester.

$$\begin{split} \delta G \left( \begin{array}{c} 5' - \mathrm{CGAGTATTCGG} - 3' \\ 3' - \mathrm{GCTCATAAGCC} - 5' \end{array} \right) &= \delta G \left( \begin{array}{c} 5' - \mathrm{CG} - 3' \\ 3' - \mathrm{GC} - 5' \end{array} \right) + \\ \delta G \left( \begin{array}{c} 5' - \mathrm{GA} - 3' \\ 3' - \mathrm{CT} - 5' \end{array} \right) &+ \delta G \left( \begin{array}{c} 5' - \mathrm{AG} - 3' \\ 3' - \mathrm{TC} - 5' \end{array} \right) + \\ \delta G \left( \begin{array}{c} 5' - \mathrm{TA} - 3' \\ 3' - \mathrm{AT} - 5' \end{array} \right) &+ \\ \delta G \left( \begin{array}{c} 5' - \mathrm{TA} - 3' \\ 3' - \mathrm{AT} - 5' \end{array} \right) + \\ \delta G \left( \begin{array}{c} 5' - \mathrm{CG} - 3' \\ 3' - \mathrm{AT} - 5' \end{array} \right) + \\ \delta G \left( \begin{array}{c} 5' - \mathrm{CG} - 3' \\ 3' - \mathrm{CG} - 3' \\ 3' - \mathrm{GC} - 5' \end{array} \right) + \\ \delta G \left( \begin{array}{c} 5' - \mathrm{CG} - 3' \\ 3' - \mathrm{CG} - 3' \\ 3' - \mathrm{CC} - 5' \end{array} \right) + \\ \end{split}$$

- Don't sum "scores" as in sequence alignment.
- Consider two BPs at a time.
- Consecutive  $\delta G$ 's are not independent.

# **Other stacking free energies**

 $\delta G$  for mismatched pairs and dangling



In the example structure on the left:

- Stacking of the  $C_{13}\cdot A_{19}$  mismatch stabilizes the H-loop.
- Stacking of the  $C_4 \cdot C_{27}$  mismatch and the  $U_8 \cdot U_{24}$  mismatch stabilizes the I-loop.
- These negative (favorable) energies are added to the unfavorable (positive) energies of the H-loop or the I-loop. They are really associated with the adjacent stem. The energy assignment to the loop is done for algorithmic reasons.
- Stacking of single bases at the end of stems is also considered (not shown).

# Entropic terms for loops: $\delta G$ is unfavorable



In the same example structure on the left:

- Both the H-loop and the I-loop have penalty energies that grow logarithmically with loop size (number of single-stranded bases).
- $\delta G \approx 1.75 RT \ln(l)$ , for loop size *l*.
- In addition, there is an I-loop asymmetry penalty. The asymmetry of the I-loop in the example is 1 = |5 4|. This is the difference between the number of single-stranded bases on each side of the loop.

#### **Energy methods for single sequences - Multi-structure dot plots**

The energy dot plot (EDP)

- The minimum free energy (mfe) of a folding,  $\Delta G_{\rm mfe}$ , can be computed.
- Let  $\delta G \geq 0$  be a free energy increment.
- All secondary structures with free energies between  $\Delta G_{mfe}$  and  $\Delta G_{mfe}$  +  $\delta G$  are superimposed in a single plot.
- Different colors are used for different values of  $\delta G$ .
- For  $\delta G_1 < \delta G_2$ , with colors  $c_1$  and  $c_2$ , respectively,  $c_1$  is "on top" and obscures  $c_2$ .

The probability dot plot (PDP)

- All base pairs with probabilities above some cutoff are plotted as "dots".
- The area of the dots is (usually) proportional to the probability.
- The dots are colored to indicate a probability range.

#### **Algorithms - Free energy minimization**

- Simplified model that assigns energies to base pairs and ignores loop instabilities.
- For RNA sequence,  $R = r_1 r_2 ... r_n$ , let e(i, j) be the free energy of pairing  $r_i$  with  $r_j$ . This can be  $+\infty$  when a base pair is impossible.
- If S is a secondary structure on R, then the free energy of S is denoted by ΔG(S) and is defined as

$$\Delta G(\mathbf{S}) = \sum_{r_i \cdot r_j \in \mathbf{S}} e(i, j).$$

- Reasonable values of *e* at 37 °C are -3, -2 and -1 kcal/mol for GC, AU and GU base pairs, respectively.
- The goal is to compute  $\Delta G_{mfe}(R) = \min_{S \in S} \Delta G(S)$ , where "mfe" refers to minimum free energy and S is the collection of all secondary structures.
- If  $1 \le i \le j \le n$ , let  $E(i, j) = \Delta G_{mfe}(R_{i,j})$ , where  $R_{i,j} = r_i \dots r_j$ .
- E(i, j) = 0 if j i < 4, since no structure can form in such a case.
- Otherwise,

$$E(i,j) = \min\left\{\frac{E(i+1,j), E(i,j-1), e(i,j) + E(i+1,j-1), \min_{k=i+1}^{j-1} (E(i,k) + E(k+1,j))\right\}$$

September 30, 2009. - RNA Course

#### Algorithms - Free energy minimization - II

In (color coded) words:

- Fragments of length  $\leq$  4 have 0 folding energy, since they cannot fold. Otherwise,
- $r_i$  is unpaired, or
- $r_j$  is unpaired, or
- $r_i$  and  $r_j$  pair with each other, or
- $r_i$  and  $r_j$  both pair, but **not** with each other. In this case,  $r_i$  pairs with  $r_{k_1}$  and  $r_j$  pairs with  $r_{k_2}$ , where  $i < k_1 < k_2 < j$ . The k in the recursion can be any integer satisfying  $k_1 \le k < k_2$ .

The above equations describe the "fill algorithm".

- E(i, j) is computed recursively for ever larger sequence fragments.
- $\Delta G(R) = E(1, n)$ , but what structure(s) has (have) this free energy?
- The "traceback algorithm" computes a structure.

# Algorithms - Free energy minimization - Traceback - I



An illustrated traceback chart for RNA folding using base pair dependent energy rules.

#### Algorithms - Free energy minimization - Traceback - II

- Start: Set i = 1 and j = n. Put *i* and *j* on to the "traceback stack".
- Recursion:
  - 1. If the traceback stack is empty, the traceback terminates. Otherwise, take *i* and *j* from the traceback stack.
  - 2. If  $E_{i+1,j} = E_{i,j}$ , then *i* is not paired.
    - (a) If j i > 3, set i = i + 1 and continue with 2.
    - (b) If  $j i \le 3$ , continue with 1.

Otherwise, continue with 3.

- 3. If  $E_{i,j-1} = E_{i,j}$ , then *j* is not paired.
  - (a) If j i > 3, set j = j 1 and continue with 3.
  - (b) If  $j i \le 3$ , continue with 1.

Otherwise, continue with 4.

- 4. If  $E_{i,j} = e(i,j) + E_{i+1,j-1}$ , then  $r_i$  pairs with  $r_j$ . Add i.j to the list of base pairs, set i = i + 1 and j = j 1 and continue with 2. Otherwise, continue with 5.
- 5. If  $E_{i,j} = E_{i,k} + E_{k+1,j}$ , for some  $k \in (i, j)$ , put the fragment  $k + 1 \dots j$  on to the traceback stack (*i* is k + 1 and *j* is the current *j*) and deal with  $i \dots k$  by setting j = k and continue with 2. (Note that some *k* must exist if this point is reached.)

#### Fill and traceback example.

Given the sequence R = GCAGCACCCAAAGGGAAUAUGGGAUACGCGUA. The base pair folding energies are -3, -2 and -1 for GC (CG), AU (UA) and GU (UG) base pairs, respectively. E(i, j) appears in row *i* and column *j* of the triangular array below.

1 2 3 4 5	5 7 8 9 10 11 12 13 14	15 16 17 18 19 20 21 22	2 23 24 25 26 27 28 29 30 31 32	
G C A G C	CCCCAAAGG	G A A U A U G C	G G A U A C G C G U A	
	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$

# Fill & traceback example: A traceback route

5 6 7 8 9 10 11 12 13 14	15 16 17 18 19 20 21	22 23 24 25 26	27 28 29 30 31 32	
САСССАААGG	G A A U A U G	i G G A U A	CGCGUA	
-3 -3 -3 -3 -6 -6 -6 -6 -6 -9	-12 -12 -12 -14 -14 -16	6 - 17 - 17 - 17 - 18 - 18	-20 -21 -24 -24 -25 -25	G 1
0 0 0 -3 -3 -3 -3 -3 -6 -9	-9 -9 -9 -11 -11 -13 -16	6 -17 -17 -17 -17 -17	-20 -21 -22 <mark>-24</mark> -24 -24	C 2
0 0 0 -3 -3 -3 -3 -3 -6 -6	-9 -9 -9 -11 -11 -13 -14	-14 -14 -14 -16 -16	-18 -18 <mark>-21</mark> -21 -23 -23	A 3
0 0 0 -3 -3 -3 -3 -3 -6 -6	-9 -9 -9 -11 -11 -12 -14	-14 -14 -14 -15 -15	-18 -18 <mark>-21</mark> -21 -22 -22	G 4
000000000-3-6	-9 -9 -9 -11 -11 -14	-14 -14 -14 -15 -15	-17 <mark>-18</mark> -20 -21 -21 -21	C 5
0000000-3-6	-9 -9 -9 -11 -11 -11 -11	-11 -13 -13 -15 -15	- <mark>15</mark> -18 -18 -18 -20 -20	A 6
000000-3-6	-9 -9 -9 -9 -9 -11 -11	-11 -13 -13 -13 -15	- <mark>15</mark> -18 -18 -18 -18 -20	C 7
00000-3-6	-6 -6 -6 -6 -8 -8	6 -10 -10 -10 -10 -12	-15 -15 -15 -18 -18 -18	C 8
0 0 0 0 -3 -3	-3   -3 -3 -4 -4 -5 -7	' -7 -7 -7 -8 -9	-12 -12 -15 -15 -15 -15	C 9
0 0 0 0 0	0 0 0 -2 -2 -4 -4	-4 -4 -4 -6 -6	-9 -9 -12 -12 -14 -14	A 10
0 0 0 0	0 0 0 -2 -2 -4 -4	-4 -4 -4 -6 -6	-9 -9 -12 -12 -14 -14	A 11
0 0 0	0 0 0 -2 -2 -3 -3	8 - 3 - 3 - 4 - 6 - 6	-9 -9 -12 -12 -14 -14	A 12
0 0	0 0 0 -1 -1 -2 -2	2 - 2 - 2 - 3 - 5 - 6	-9 -9 -12 -12 -13 -13	G 13
0	0 0 0 -1 -1 -2 -2	2 - 2 - 2 - 3 - 5 - 6	-9 -9 -12 -12 -12 -12	G 14
	0 0 0 0 0 -2 -2	2 -2 -2 -2 -4 -6		G 15
	0 0 0 0 -2 -2	2 - 2 - 2 - 2 - 4 - 6		A 16
		0 -1 -1 -2 -4 -6		A 17
		0 -1 -1 -2 -4 -6		
	0 (			
	ť			
				G 22
				A 24
				11 25
		0 0		Δ 26
		U		C 27
				G 28
				C 29
			0 0 0	G 30
				U 31
			0	A 32
			e e	1

# **Resulting Structure(s)**



The structure on the left corresponds to the traceback route depicted in red on the previous page.

This alternative structure is also mfe. It would be found by following the "-15"s to position (6,25).

# **Instability of Predictions**

Problems in folding prediction became evident as soon a minimum free energy folding predictions became available. How was this detected?

- 1. Folding a slightly different homologous sequence could result in a totally different predicted secondary structure.
- 2. Small changes in the free energy parameters could dramatically alter folding prediction.
- 3. Adding folding constraints would change predicted sturctures, but in many cases the mfe increased very little.
- 4. Attempts to model equilibrium folding of elongating RNA molecules failed to give satisfactory results.

On the next twelve pages, mfe foldings are predicted for the same 5S rRNA on the segments  $5' - r_1 \dots r_N - 3'$ , for  $N = 10, 20, \dots 110, 123$ . Base pairs in previous foldings are not conserved in general.

Thus, it became important to predict sub-optimal foldings as well as mfe foldings. September 30, 2009. - RNA Course

### Folding Halobacterium saccharovorum 5S rRNA: 1–10

dG = -4.2 Halosacc

### Folding Halobacterium saccharovorum 5S rRNA: 1–20

dG = -4.2 Halosacc








R

120

3' - C













dG = -34.8 Halosacc



dG = -47.6 Halosacc

#### The suboptimal algorithm: Circularize the RNA

**Apparent Lack of Symmetry:** 



For any  $i_{\lambda}j$  pair,  $V(i_{\lambda}j)$  gives the best folding energy for the included region only.

Solution: "Circularize" the sequence.



Define V(i, j) = e(i, j) + E(i + 1, j - 1). That is, V(i, j) is the mfe of all foldings on  $R_{i,j}$  where  $r_i$  and  $r_j$  pair. Call  $R_{i,j}$  an "included" fragment of R. Circularize the RNA (some RNAs. such as viroids, are naturally circular.) For i < j, let E(j,i) and V(j,i) be defined for the "excluded" fragment, Then  $R_{j,i} = r_j r_{j+1} \dots r_n r_1 \dots r_i.$ V(i, j) + V(j, i) - e(i, j) is the mfe over all foldings containing the base pair,  $r_i \cdot r_j$ . Q: How to compute E and V for excluded fragments?

#### **The suboptimal algorithm: Folding** mod *n*



Apply the regular fill algorithm to the doubled sequence.

#### The suboptimal algorithm: Practical details

- I' region,  $n < i \le j \le 2n$ , need not be defined or filled, since E(i, j) = E(i n, j n) and V(i, j) = V(i n, j n).
- The region defined by  $j i \ge n$  is not required (unused).
- Only 1/2 of the doubled array need be defined in computer memory.
- The folding rules are slightly altered for fragments containing the origin. For example, the base pair,  $r_1 \cdot r_n$  can exist, implying that  $r_n \cdot r_{n+1}$  exists, even though *n* and n+1 are adjacent.



#### The suboptimal algorithm: Results

• For base pair energy rules, the mfe on any folding containing the base pair,  $r_i \cdot r_j$ , is

$$\delta G_{i,j} = V(i,j) + V(j,i+n) - e(i,j).$$

For loop energy rules,  $\delta G_{i,j} = V(i,j) + V(j,i+n)$ .

A structure containing r<sub>i</sub>·r<sub>j</sub> with free energy δG<sub>i,j</sub> may be computed by computing and combining two tracebacks; one for R<sub>i,j</sub> beginning with r<sub>i</sub>·r<sub>j</sub> and one for R<sub>j,i+n</sub>, starting with r<sub>j</sub>·r<sub>i+n</sub>.

#### **The Equilibrium Partition Function: Definition**

The definition of the partition function for all secondary structures on R is

$$Z(R) = \sum_{\mathbf{S}\in\mathcal{S}} e^{-\frac{\Delta G(\mathbf{S})}{RT}}$$

This is a weighted counting of all structures. Note that the lower the free energy, the higher the weighting. According to statistical mechanical theory, this Boltzmann weighting gives the probability density for every folding. That is, the probability of any particular folding, **S**, is given by  $\exp(-\Delta G(\mathbf{S})/RT)/Z$ .

The number of secondary structures grows roughly as  $1.8^n$ , but the computation can be performed in reasonable time using a recursion similar to the original dynamic programming algorithm for computing an optimal folding. We need to introduce new terms.

#### Algorithm for base pair dependent rules.

AUXILIARY PARTITION FUNCTIONS:

 $Q(i, j) = Z(R_{i,j})$ Q'(i, j): *restricted*, must contain  $r_i \cdot r_j$  base pair

That is, we consider partition functions for every fragment of the original sequence. The Q's are required for loop dependent energy rules, but not for the simpler base pair energy rules. Simple recursions exist for base pair dependent energy rules.

where Q(j+1,j) is defined to be 1 for convenience.

#### Algorithm in picture form.

Partition function – base pair dependent rules



The figure on the left illustrates the recursion step for computing the partition function for base pair energy rules. For each summand, the contribution is broken into 3 components, as shown by the colors.

#### Cases:

- 1. i is unpaired
- 2. i pairs with a base k, where  $Q_{j+1,j} = 1$

#### **Partition functions: Computing probabilities**

When the recursion is finished, we know the probability, P(S), of any structure S. That is:

$$P(\mathbf{S}) = \frac{e^{-\frac{\Delta G(\mathbf{S})}{RT}}}{Q(1,n)}$$

As with the suboptimal algorithm, the partition function algorithm may also be run on the **doubled sequence**.

For  $1 \le i \le j \le n$ , Q(i, j) and Q'(i, j) are partition functions for the included fragment,  $R_{i,j}$  and Q(j, i+n) & Q'(j, i+n) are partition functions for the excluded fragment,  $R_{j,i+n}$ .

The probability of the base pair  $r_i \cdot r_j$  is given by

$$P(r_i \cdot r_j) = \frac{Q'(i,j)Q'(j,i+n)}{e^{-\frac{e(i,j)}{RT}}Q(1,n)}.$$

The exponential term in the denominator corrects for the fact that both Q'(i, j) and Q'(j, i + n) count the energy contribution of the base pair  $r_i \cdot r_j$ . This term is absent when base pair energy rules are used.

September 30, 2009. – RNA Course

#### **Partition functions: Stochastic traceback**

#### Ratios of partition functions are conditional probabilities.

For any fragment,  $R_{i,j}$  of R,

• 
$$P_{i,j}(i) = \frac{Q(i+1,j)}{Q(i,j)}$$
 is the probability that  $r_i$  is single-stranded.

•  $P_{i,j}(k) = \frac{Q'(i,k)Q(k+1,j)}{Q(i,j)}$  is the probability that  $r_i$  pairs with  $r_k$ , for  $i < k \le j$ .

- These probabilities are **conditional**.
- That is, they are conditioned on the premise that no base pair of the form r<sub>h</sub>·r<sub>l</sub> exists for h < i and i ≤ l ≤ j or for i ≤ h ≤ j and l > j.

#### Partition functions: Stochastic traceback algorithm.

Algorithm flowchart:

- Initial conditions: The list of base pairs and the "traceback stack" are both empty.
- Start: Set i = 1 and j = n. Put *i* and *j* on to the "traceback stack".
- Recursion:
  - 1. If the traceback stack is empty, the traceback terminates. Otherwise, take i and j from the traceback stack.
  - 2. Choose a random number, *x*, between 0 and 1 (uniform distribution). If  $x \le P_{i,j}(i)$ , continue with 3; otherwise, continue with 4
  - 3.  $r_i$  will be unpaired.
    - (a) If j i > 3, set i = i + 1 and continue with 2.
    - (b) If  $j i \le 3$ , continue with 1.
  - 4. Find the unique k,  $i < k \le j$  such that  $\sum_{h=i}^{k-1} P_{i,j}(h) < x$  and  $\sum_{h=i}^{k} P_{i,j}(h) \ge x$ .
    - (a) Add  $r_i \cdot r_k$  to the list of base pairs.
    - (b) If  $j k \ge 3$ , place k + 1 and j on to the traceback stack.
    - (c) Set i = i + 1, j = k 1. Continue with 2 if j i > 3; otherwise, continue with 1

#### The folding problem is ill-conditioned in general

As bad as it gets.

Three *conjugate*<sup>1</sup> foldings of a short sequence. From left to right,  $\Delta G = -4.9$  kcal/mol (mfe folding),  $\Delta G = -4.7$  and  $\Delta G = -4.6$ !



<sup>&</sup>lt;sup>1</sup>Two foldings are called conjugate if they share no base pairs in common. September 30, 2009. – RNA Course

#### Other views of the "three conjugate" foldings example





Clockwise from above left:

- 1. Structure dot plot. Superposition of three conjugate foldings.
- EDP (energy dot plot). Superposition of all possible foldings within 0.3 kcal/mol from mfe.
- 3. PDP (probability dot plot) A plot of all base pairs with probabilities  $\geq 0.01$ . The maximum probability is 0.44 (poor).



#### Heresy: Statistics on one Datum

For a single secondary structure, predicted or otherwise:

- Not all base pairs are created equal.
- The EDP and PDP can be used to distinguish.
- In a secondary structure plot:
  - 1. color base pairs according to the probability that they form.
  - 2. color single-stranded bases according to the probability that they are **not** paired.
- Higher probability base pairs are more likely to be correct.

#### The *consensus* structure.

The consensus structure contains all base pairs whose probabilities are > 50%, and no others. (Ye Ding & Chip Lawrence call this the "centroid".)

Quiz:

- Prove that the consensus structure is valid (no base triples). [20 points]
- Prove that the consensus structure contains no pseudoknots. [40 points]

#### Good versus Bad RNAs

- A good RNA has an "uncluttered" EDP or PDP. Base pairs in a mfe folding are more likely to be correct. High probability base pairs are "usually" correct, but a good RNA has a lot more high probability base pairs than a bad RNA.
- A bad RNA has a "cluttered" EDP or PDP. A mfe folding has fewer correct base pairs. Even though high probability base pairs are likely to be correct, there are relatively few of them.

# Example: A good RNA versus a bad RNA.E. californium 5S rRNAD. discoidium 5S rRNA

Output of boxplot\_ng (®) by D. Stewart and M. Zuker



0.0099 <= Probability <= 1.0000

Output of boxplot\_ng (®) by D. Stewart and M. Zuker

Probability Dotplot for D\_discoid

0.0099 <= Probability <= 0.9995





"Uncluttered" PDP versus a "cluttered" one.

#### **Consensus structures for the same pair of 5S rRNAs**



Not much predicted with confidence *versus* a lot predicted with high confidence. In this case, the second folding is a very good prediction.

#### The numbers

Output of ct\_boxplot (®) by D. Stewart and M. Zuker Output of ct\_boxplot (®) by D. Stewart and M. Zuker





\*Counts for each structure include overlap dots.

#### \*Counts for each structure include overlap dots.

#### Entropy: Quantifying "good" and "bad"

If  $P = \{p_i\}_{i=1}^{\infty}$  is a countable probability distribution (may be finite), then

$$\sum_{i=1}^{\infty} p_i = 1.$$

The (Shannon) entropy of P is defined by

$$\mathfrak{S}(P) = -\sum_{i=1}^{\infty} p_i \log(p_i).$$

(Treat  $0\ln(0)$  as 0.) For infinite distributions, the entropy might be  $+\infty$ . We do not consider infinite or continuous distributions here. The base for the logarithm is not specified and may be chosen arbitrarily to scale the entropy. For example, base 2 gives the entropy in "bits".

When there are *n* "outcomes",  $1 \le i \le n$  and the maximum entropy distribution sets  $p_i = 1/n$ , so that  $\mathfrak{S}(P) = -\sum_{i=1}^{n} p_i \log(p_i) = \log(n)$ .

If S(R) is the set of all foldings of  $r_1r_2...r_n$ , then the maximum entropy loss to select a single folding is given by making all foldings equally likely, and so:

 $\mathfrak{S}(\mathcal{S}(R)) \sim Cn - 1.5 \ln(n)$  as  $n \to \infty$ , for some constant *C*. What is *C*?

#### September 30, 2009. - RNA Course

#### The expected number of foldings

Suppose that an RNA sequence is totally random, with equal probabilities for A, C, G and U. Then the probability that two different bases can pair is  $\frac{3}{8}$ . This represents 1/16 for A·U, U·A, C·G, G·C, G·U and U·G. In general, call this probability *p*.

Many years ago, I showed that the expected number of foldings,  $T_n$ , on a random RNA sequence of length *n* is given by

$$T_n \sim \frac{(1+4\sqrt{p})^{\frac{1}{4}}}{2\sqrt{\pi}p^{\frac{3}{4}}} n^{-\frac{3}{2}} \left(\frac{1+\sqrt{1+4\sqrt{p}}}{2}\right)^{2n+2},$$

as  $n \to \infty$ .

This implies a maximum entropy of  $2\log\left(\frac{1+\sqrt{1+4\sqrt{p}}}{2}\right)$  per base. This reduces to  $2\log(1+\sqrt{1+4\sqrt{p}})-2$  bits/base.

For p = 3/8, the maximum entropy is 1.029 per base. For p = 1/4, it is 0.900 bits per base.

#### The entropy of the Boltzmann distribution of all foldings.

The partition function for S(R) is defined by

$$Z = \sum_{\mathbf{S} \in \mathcal{S}(R)} \exp\left(-\frac{\Delta G(\mathbf{S})}{RT}\right)$$
  
The probability of structure **S** is  $\frac{e^{-\frac{\Delta G(\mathbf{S})}{RT}}}{Z}$   
The log probability is  $-\frac{\Delta G(\mathbf{S})}{RT} - \ln Z$ , so  
 $\mathfrak{S}(\mathcal{S}(R)) = \frac{\overline{\Delta G(\mathbf{S})} - \Delta G_{\text{ens}}}{RT}.$ 

 $\Delta G_{\text{ens}}$  is computed directly by standard algorithms and  $\overline{\Delta G(\mathbf{S})}$  can be estimated by averaging the free energies of a stochastic sample of foldings. Replacing *RT* with *RT* ln(2) expresses the entropy in bits.

I have chosen entropy per base to be the overall measure of "well-definedness" for RNA folding.

#### **Entropy of random RNA**

For each of 302 Eukaryote 5S rRNAs, 100 random sequences were generated with the same length and dinucleotide distribution. For each of these, the entropy of the Boltzmann distribution was computed with a sample of 100 foldings. GC content in the random sequences was  $55.57 \pm 3.79\%$ .

The entropy was  $0.226 \pm 0.007$  bits/base.

The entropies of 100 randomized versions of the RNase P RNA from *Desulfovibrio desulfuricans* (length 360) gave similar results, as did foldings of the first 150 bases of these sequences.

#### Good and bad revisited

- Good refers to low entropy RNAs, for which many base pairs are predicted with high probability.
- Bad refers to high entropy RNAs, for which few base pairs are predicted with high probability.
- Problems: As is usual in biology, there are exceptions to each "rule".
  - 1. In "good" or "bad" RNAs (especially for "good"), one expects that base pairs with high probability should be correct. Some high probability base pairs are not correct and some correct base pairs have very low probabilities.
  - 2. In homologous RNAs that are similar enough to be readily aligned, some high probability base pairs in one may be low probability base pairs in the other, and *vice versa*.

#### **Table of computations**

Туре	Number	min S	$\overline{\mathfrak{S}}$	max S	r GC content	$r  \delta G_{\mathrm{mfe}}$
5S rRNA	439	0.123	0.223	0.311	-0.308	0.500
Eubacteria		Thermus sp1		Planctomyces		
				limnophilus		
5S rRNA	57	0.063	0.165	0.246	-0.386	0.768
Archaea		Sulfolobus		Halobacterium		
		faci2		saccharovorum		
5S rRNA	302	0.106	0.182	0.264	-0.119	0.543
Eukaryota		Lineus		Spirogyra		
		geniculatus		sp		
group I	81	0.161	0.212	0.307	-0.431	0.362
introns		Staurastrum sp. Chl		<i>Zea mays</i> Chl		
		(M753) SSU rRNA		tRNA Leu		
group II	51	0.177	0.217	0.258	-0.770	0.597
introns		<i>Trichodesmium</i> Chl		Sinorhizobium		
		erythraeum (sp I1)		<i>meliloti</i> (I1)		
RNase P	179	0.161	0.205	0.263	-0.573	0.762
Eubacteria		Desulfovibrio		Campylobacter		
		desulfuricans		jejuni		
RNase P	36	0.162	0.213	0.278	-0.745	0.693
Archaea		Aeropyrum		Methanococcus		
		pernix		maripaludus		
tRNA	803	0.090	0.201	0.308	-0.207	0.443
		Bacillus		(Chl) <i>Euglena</i>		
		subtillus GAT		<i>gracilis</i> CAT		

Minimum, average and maximum entropies, in bits/base, for different types of RNA. *r* is the correlation of entropy with GC content or minimum free energy.

#### Best versus worst. 5S rRNA from Archaea. I Halobacterium saccharovorum

Sulfolobus faci2

Output of boxplot\_ng (®) mfold\_util 4.0

#### Probability Dotplot for Sulfaci2

0.0010 <= Probability <= 1.0000

Output of boxplot\_ng (®) mfold\_util 4.0

Probability Dotplot for Halosacc

0.0010 <= Probability <= 0.9532





#### Best versus worst. 5S rRNA from Archaea. II



September 30, 2009. - RNA Course
#### Best versus worst. 5S rRNA from Archaea. III Halobacterium saccharovorum

Sulfolobus faci2

Output of ct\_boxplot (®) mfold\_util 4.0



Output of ct\_boxplot (®) mfold\_util 4.0

Structure dot plot for Halosacc phylo/consen



\*Counts for each structure include overlap dots.

\*Counts for each structure include overlap dots.

#### Best versus worst. 5S rRNA from Eubacteria. I Planctomyces limnophilus

#### Thermus sp1

Output of boxplot\_ng (®) mfold util 4.0

#### i lancioniyces in

Output of boxplot\_ng (®) mfold\_util 4.0

Probability Dotplot for Pmyc\_limno

0.0010 <= Probability <= 0.9996



Probability Dotplot for Thermu\_sp1



### Best versus worst. 5S rRNA from Eubacteria. II



#### Best versus worst. 5S rRNA from Eubacteria. III Planctomyces limnophilus

#### Thermus sp1

Output of ct\_boxplot (®) mfold\_util 4.0



Output of ct\_boxplot (®) mfold\_util 4.0

Structure dot plot for Pmyc\_limno phylo/consen



\*Counts for each structure include overlap dots.

\*Counts for each structure include overlap dots.

# **Problems I**

The tandem base pairs in black have probabilities of roughly 1/2 of 1 percent. They occur in the phylogenetic models for 5S rRNA and seem correct in terms of alignment and covariation.

This motif strongly contradicts the energy rules. What is happening here?

The table below shows good conservation evidence for the base pair  $A_{27} \cdot U_{52}$ . MI = 0.85 bits.

Number	-	A	С	G	U	Total
A	0	0	0	0	246	246
С	0	0	0	0	0	0
G	0	0	40	0	0	40
U	0	16	0	0	0	16
	0	0	0	0	0	0
Total	0	16	40	0	246	302



## Trypanosomes



#### Copied from the Matthews Lab web site: www.biology.ed.ac.uk/research/groups/kmatthews/

African trypanosomes are parasites, spread between mammals by tsetseflies. They are responsible for epidemics of sleeping sickness in sub-Saharan Africa. This disease is always fatal unless treated and is the second biggest killer behind HIV in parts of Africa.

Trypanosomes are very ancient and interesting organisms: they are among the earliest branching eukaryotes. Fundamental discoveries have been made in these parasites which have had major impact on what we know about all cells. They also are unique in many respects, an example of the "differently evolved".

Life cycle: Different forms in insect and mammalian hosts.

Tyrpanosomes are ugly (nasty) protists. Their biology is very odd (RNA editing). It is not surprising that they are "ugly" in terms of RNA folding.

## **Problems II - The Ugly**

The U3 RNAs from *L. collosoma* and *T. brucei* are problematic. Their entropy is high, but the problem is that they contradict each other.



## **Amastin mRNA 3'-UTRs**

Amastin surface proteins belong to a large family of developmentally regulated proteins comprising up to 45 members that have recently been discovered in the genus *Leishmania* and are highly similar to the amastin proteins in *Trypanosoma cruzi*.

Pfam Family: Amastin (PF07344) Summary: Amastin surface glycoprotein

This family contains the eukaryotic surface glycoprotein amastin (approximately 180 residues long). In Trypanosoma cruzi, amastin is particularly abundant during the amastigote stage.

Reference: Teixeira SM, Russell DG, Kirchhoff LV, Donelson JE; , *J Biol Chem* 1994;**269**:20509-20516.: A differentially expressed gene family encoding "amastin," a surface protein of *Trypanosoma cruzi* amastigotes. PUBMED:8051148

### **Amastin species tree**

```
Species tree for PF07344
# Generated from Pfam version 22.0
   +--Eukarvota (58)
      +--Euglenozoa (58)
         +--Kinetoplastida (58)
            +--Trypanosomatidae (58)
               +--Trypanosoma brucei (2)
               +--Leishmania mexicana (1)
               +--Leishmania amazonensis (1)
               +--Leishmania (1)
                  +--Leishmania braziliensis (1)
               +--Leishmania major (39)
               +--Trypanosoma (13)
                  +--Trypanosoma cruzi (13)
               +--Leishmania donovani infantum (1)
```

Goal: Look for a common RNA structure in the 3'-UTRs of 60 amastin mRNAs that would be involved in expression control.

#

# **Very ugly – Amastin mRNA 3'-UTRs**

Facte	•
i acis	•

- The 60 UTRs cannot be aligned in any meaningful way.
- As with the Trypanosome U3 RNAs, the PDPs are very dissimilar.
- Worse, some of the UTRs have very low entropy, and still contradict one another.

Right: Alignment of the two lowest entropy amastin UTRs.

CLUSTAL W	W (1.83)	multiple	sequence	alignment
-----------	----------	----------	----------	-----------

AL356456_111721-112125 AL365476_10543-10944	CG-AUCCGCUCAGACGCCCAUGCGCAUACACACCAUGCCACGGCAUACCG CGUGUGUGCGCAGAGCAUGCGGCGUCUCCUGCGUCCGACCCGGUGGGCUG ** * ** *** ** * * * * * * * * * * * *
AL356456_111721-112125 AL365476_10543-10944	CGGCCUGAAGCGACUGAGAGCGAAGCGUGGCAUGCGAGCAGAGCGCAG AUAUCAUGAGGGCACGGGUGCCGGAAGCAUGGCAUACGGCCGGGUUGCAC * ** * * * * ** ***** ***** ** * * * *
AL356456_111721-112125 AL365476_10543-10944	GGUCUCGGAG-GGCCCUGGCCUGCGGCUGGCCAACUGCCUCCCUGAGCUC GGGCUCAGAAAGGCUCUCACCGGAGGCUGGCGGACUUGUUCCUCGAGCUC ** *** ** *** *** ** ** ******* *** **
AL356456_111721-112125 AL365476_10543-10944	AUGUGGCACGCACCUGGGUCACCCCCUGCUGCCC-GCGCGACAUCCGGCACGCGCCUGGACCACCACCACCACCUCUGUUACCCCGCAGGGC***
AL356456_111721-112125 AL365476_10543-10944	AGACGUCACCGCCACCGCACCGGAGCAGGGCUCUG-GGCGCUGCCCGCCU GGACGCCACCGCCACCGCCUCGAGCAGAGCCCUGUGACGCCGACCGC **** *********** * **** * *** * ****
AL356456_111721-112125 AL365476_10543-10944	GGAUCAGGGCGUGCAGUCCACCCAGGCCCGCGUGGUGCCGGCGGGGCAGC GCAGCACGUACCGGAGGAAGGGAUGG-GCAGGUGCUGUUGG *** *** * *** * *** * *** ** *
AL356456_111721-112125 AL365476_10543-10944	ACACAUGGGAAGCGGGGACGGCGGAAGGGCAGGACGGGCUCGUCCUAGGC GCGCGUGGCGCGGUUCGGCGGCAGGGCGGCUCGUUCUCGGC * * *** ** ** ***** *****
AL356456_111721-112125 AL365476_10543-10944	GGCCACGAUGCAGCCGUGGUUCAACAGAAGUAGAUGAGCGGUGGUU GACCAGGAGAUGUCGCCGCCGUGCGGCCAGAGGGUGGGGGCGCUGUGGUU * *** ** ** ** *** ** * * * * * * * *
AL356456_111721-112125 AL365476_10543-10944	GU

### **Complot: Dot plot superposition using alignment**

As ugly as it can be. **NO** common base pairs with probabilities  $\geq 0.001!$ , with the exception of four nonsense pairs (in black). This occurs despite the fact that both *Leishmania major* 3'-UTRs have significantly low entropy.



### Rfam - Database of aligned RNAs with consensus structure

The Rfam database, at

www.sanger.ac.uk/Software/Rfam/, and at rfam.janelia.org/ is a significant and growing web resource.

- Currently<sup>2</sup>, 1371 families. Each family has a *seed* file as well as a full alignment file.
- The purpose of the database is to use families to search genomic data for similar sequences.
- Many entries contain little or no mutual information to support consensus structures annotated in the alignment files.

A total of 14654 sequences from the 607 Rfam seed files (Version 8.1, October 2007) were processed.

### Some remarks on Rfam

The use of single sequence entropy calculations on Rfam RNAs might identify families for which the alignment needs to be improved, or families that might better be split into two (or more) groups.

Low entropy foldings might be useful in generating an initial folding for a family that could then be refined as needed using the other members together with a multiple sequence alignment.

This method fails badly with the amastin 3'-UTRs. It works in other cases.

## **Acknowledgments**

- NIGMS: Algorithm and software development support.
- Doug Turner, Dave Mathews et al. RNA energy parameters and algorithm development
- HHMI, Sean Eddy: Sabbatical support.
- Darin Stewart, Alex Yu, Nick Markham: mfold\_util display software
- Nick Markham: UNAFold software
- Rfam consortium (Sean Eddy, Alex Bateman *et al.*): Rfam database and amastins
- Jim Brown: RNAse P RNA database
- Robin Gutell: group I intron database
- Sprinzl, Szymanski, Zimmerly already mentioned
- many others