

WEB SUPPLEMENT: Exploring the energy landscape of k -point mutagens of RNA

P. Clote^{1,2}, J. Waldispühl^{1,3,4,*}, B. Behzadi³, J.-M. Steyaert³,

¹ Department of Biology, Higgins 355, Boston College, Chestnut Hill, MA 02467, USA,

² Department of Computer Science (courtesy appt.), Boston College, Chestnut Hill, MA 02467,

USA, ³ LIX, Ecole Polytechnique, 91128 Palaiseau Cedex, FRANCE, ⁴ LIAFA, Université Denis

Diderot, 2 place Jussieu, 75251 Paris Cedex, FRANCE.

MINIMUM ENERGY FOR K -POINT MUTAGENS

Assuming that $E_k^T(r, s)$ has been defined for all values $0 \leq k \leq n$ and all $1 \leq r < s \leq n$ such that $|s - r| < |j + 1 - i|$, we define $E_k^T(i, j + 1)$ for $1 \leq k \leq n$ by the recursive definition given in Figure 1.

In Tables 5 and 6, we give for each distance, the value ρ_{opt} of the number of mutations $k \leq 10$, such that the k -superoptimal secondary structure has best score (note that ρ_{opt} depends on the metric chosen). Although finding some regularity among the computed values of ρ_{opt} appears to be difficult, the occurrence of non-zero values for ρ_{opt} appears to provide evidence for the important role of admitting point mutations in structure determination. Some prediction improvements for sequences from Green Non-Sulfur Bacteria appear to be substantially better than that for Beta Purple Bacteria. At this point, we cannot explain the differences between these two RNA classes; however, further studies should elucidate this question.

In Tables 5 and 6, we give for each distance, the value ρ_{opt} of the number of mutations $k \leq 10$, such that the k -superoptimal secondary structure has best score (note that ρ_{opt} depends on the metric chosen). Although finding some regularity among the computed values of ρ_{opt} appears to be difficult, the occurrence of non-zero values for ρ_{opt} appears to provide evidence for the important role of admitting point mutations in structure determination. Some prediction improvements for sequences from Green Non-Sulfur Bacteria appear to be substantially better than that for Beta Purple Bacteria. At this point, we cannot explain the differences between these two RNA classes; however, further studies should elucidate this question.

In some cases (see for example *B-bronchiseptica* RNA in the Beta Purple Bacteria class, or *SMB-B3* RNA in the Green Non Sulfur Bacteria class), the k -superoptimal structure is quite close to the real (i.e. phylogenetic) secondary structure. Moreover, the upper-bound of $k \leq 10$ is most often not required to obtain the best results. This indicates that competing secondary structures generally appear when one admits only a small number of mutations for structural RNA. It would be interesting to contrast this situation with random RNA

of the same dinucleotide frequency (see Clote (2005); Clote et al. (2005)).

Note that some large values of base pair distance are surely due to the fact that phylogenetic structures are determined by covariation of base pairs in high quality multiple sequence alignments, and hence include only a proper subset of the set of experimentally determined base pairs. In particular, phylogenetic structures are generally *not* saturated in the sense of (Zuker and Sankoff, 1984) or locally optimal in the sense of (Clote, 2005).

Since k -superoptimal structures are mfe structures for sequences distinct from the originally given RNA sequence, we additionally post-process the k -superoptimal structures by replacing mutated bases by the original ones and by removing unauthorized base pairs (e.g. if a base pair (i, j) occurs where mutagen nucleotides $s'_i = G$, $s'_j = C$, and if $s_i = C = s_j$, then base pair (i, j) is removed from the k -superoptimal structure). Using this process, we report for each sequence and for each distance measure, the score (denoted *best score*) of that k -superoptimal structure, for $k \leq 10$, which is closest to the phylogenetic structure as measured by the distance measure considered. In Tables 5 and 6, we show the results for Beta purple Bacteria respectively Green Nonsulfur Bacteria.

Figure 4 presents an example of how a superoptimal structure of tRNA can better predict the anticodon region.

0.1 Experiments using AMSAG

Here, we present results from computer experiments using AMSAG with functional RNA sequences. Specifically, we compute k -superoptimal secondary structures, for $k \leq 10$, for a diverse collection of RNA sequences of length at most 345 nt. (only two small classes are shown, although we have performed computations on a number of RNA classes – see the web supplement link at <http://clavius.bc.edu/~clotelab/>). We describe two types of result. The first concerns the relation between k , the number of mutations, and the energy of the k -superoptimal secondary structure. Figure 6 shows an approximately linear dependence for *small* values of k , while data on the web supplement shows a non-linear dependence when all values $0 \leq k \leq n$ are considered. Since RNA sequence evolution depends on rare mutations tolerated by the organism, we

*Corresponding author.

$$E_k^T(i, j + 1) = \min\{A, B, C, D, E, F\} \quad (1)$$

where

- A is $E_k^T(i, j)$,
- B is $E_{k-1}^T(i, j)$,
- C is $\min_{r=i}^{j-\theta} a_{s_r, s_{j+1}} + (\min_{c=0}^k E_c^T(i, r-1) + E_{k-c}^T(r+1, j))$,
- D is $\min_{x \in \{A, C, G, U\} - \{a_{j+1}\}} \min_{r=i}^{j-\theta} a_{s_r, x} + (\min_{c=0}^{k-1} E_c^T(i, r-1) + E_{k-1-c}^T(r+1, j))$,
- E is $\min_{r=i}^{j-\theta} \min_{x \in \{A, C, G, U\} - \{a_r\}} a_{x, s_{j+1}} + (\min_{c=0}^{k-1} E_c^T(i, r-1) + E_{k-1-c}^T(r+1, j))$,
- F is $\min_{r=i}^{j-\theta} \min_{x \in \{A, C, G, U\} - \{a_r\}} \min_{y \in \{A, C, G, U\} - \{a_{j+1}\}} a_{x, y} + (\min_{c=0}^{k-2} E_c^T(i, r-1) + E_{k-1-c}^T(r+1, j))$.

Fig. 1. Recursive computation of the minimum energy of all k -point mutagens, according to the Nussinov-Jacobson model. By use of tracebacks, as explained in Clote and Backofen (2000), the corresponding minimum free energy structures can be produced.

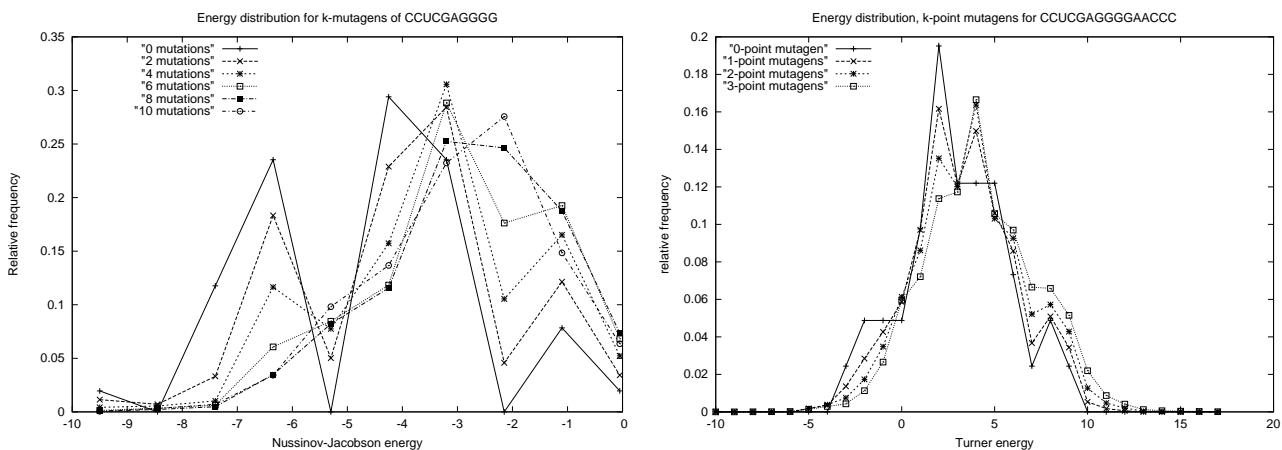


Fig. 2. Nussinov-Jacobson energy distribution for all k -point mutagens, for even $0 \leq k \leq 10$, of the 10-mer prefix CCUCGAGGGG from SECIS element fruA. Nussinov-Jacobson energies are integers from -9 to 0 . Statistics on secondary structure energies for these and other k -point mutagens can be found in the web supplement. Distributions obtained by creating histograms of energies produced by brute force enumeration.

focus on values of k which are small when compared to sequence length n .

At the extremity, when $k = n$, the situation is particularly simple. For any RNA sequence s of even length $n = 2m$, the n -superoptimal secondary structure is a hairpin loop with $m - 2$ base pairs and 4 centrally located unpaired bases (a tetraloop), since this is the minimum free energy structure of the sequence s' consisting of m occurrences of G followed by m occurrences of C. In the case that $n = 2m + 1$ is odd, because of energy bonus of a tetraloop, the n -superoptimal secondary structure is a hairpin loop with $m - 2$ base pairs and 4 centrally located unpaired bases along with an unpaired rightmost base (i.e. a 3' dangle). See Figure 5 for an example structure for $n = 41$.

We performed benchmark tests for two classes of RNA from the RNaseP database (Brown, 1999) – namely *Beta Purple Bacteria* and *Green Non Sulfur Bacteria*. For each RNA sequence s and each $k \leq 10$, we computed the k -superoptimal secondary structure and its energy (0-superoptimal structure

is just the mfe structure, as determined by Zuker's method). For these experiments, 6 Gb of memory and 2 hours of cpu time on a 666Mhz DEC-alpha were required to compute the k -superoptimal secondary structures and their energies, for $k \leq 10$, for a given RNA sequence of 300 nucleotides.

In Figure 6, for small values of $k \ll n$, we plot the minimum free energy over all k -mutagens (y -axis) of RNA from the Beta Purple Bacteria family, as a function of the number k of allowed mutations (x -axis).

For the first k -mutagens ($0 \leq k \leq 10$), Figure 6 clearly shows a *linear* dependance of mfe as a function of the number of mutations. The slope is similar for both RNA families, and approximately equal to -4.93 ± 0.65 .

REFERENCES

- Brown, J. W. (1999). The ribonuclease p database. *Nucleic Acids Res.*, 27(314).
- Clote, P. (2005). An efficient algorithm to compute the landscape of locally optimal RNA secondary structures with respect to

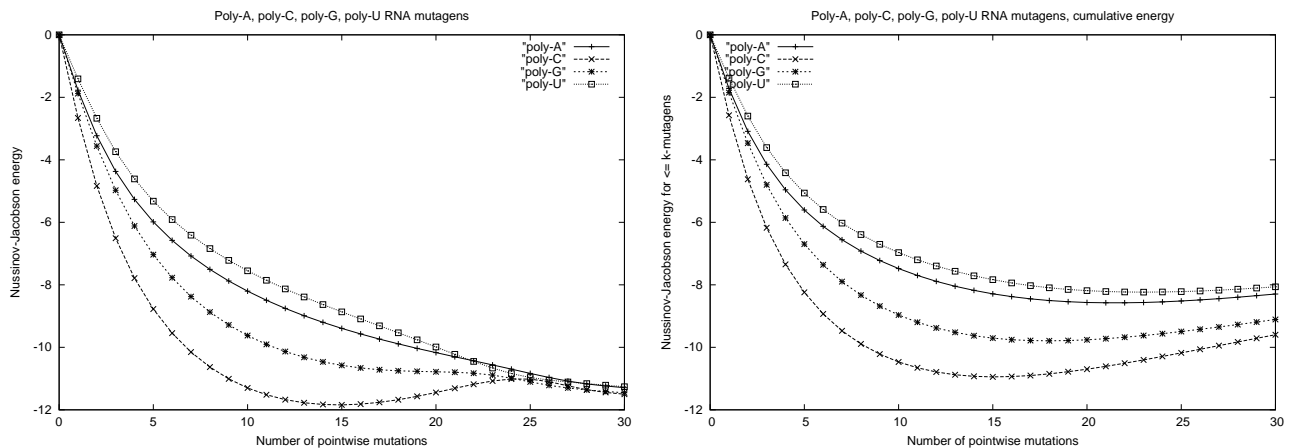


Fig. 3. This figure depicts the relation between number k of mutations (x -axis) and average energy of all k -point mutagens (resp. $\leq k$ -point mutagens) of toy 30 nt. sequences consisting only of a single type of nucleotide (i.e. poly-A, poly-C, poly-G and poly-U sequences of length 30). (i) The left panel considers k -point mutagens on poly-A, poly-C, poly-G and poly-U sequences. The order of the curves, from top to bottom, is: poly-U, poly-A, poly-G and poly-C. (ii) The right panel considers $\leq k$ -point mutagens on poly-A, poly-C, poly-G and poly-U sequences; i.e. at position k on the x -axis, the average energy over all $\leq k$ -point mutants of the respective sequence is given. Note the approximate linear comportment of the curves, which maintain the same order from top to bottom as in (i): poly-U, poly-A, poly-G and poly-C. See web supplement for a similar curve obtained for hammerhead type III ribozyme.

rna	length	#BP	coarse				w. coarse				BP dist.			
			or.	s.o.	rat.	ρ_{opt}	or.	s.o.	rat.	ρ_{opt}	or.	s.o.	rat.	ρ_{opt}
A-eutrophus	337	105	47	31	0.34	10	279	255	0.09	2	72	68	0.06	2
B-bronchiseptica	240	68	12	3	0.75	7	104	95	0.09	1	23	15	0.35	6
C-testosteroni	271	73	41	38	0.07	4	218	194	0.11	3	78	70	0.10	3
ESH167F	274	73	36	25	0.31	4	255	185	0.27	3	83	80	0.04	3
ESH183D	278	76	13	11	0.15	5	113	100	0.12	1	33	31	0.06	1
ESH26-4	278	73	27	22	0.19	7	179	-	0.00	0	73	66	0.10	4
Mxa1	308	87	30	27	0.10	3	182	-	0.00	0	93	92	0.01	3
N-europaea	281	78	40	28	0.30	10	339	337	0.01	9	123	117	0.05	7
N-gonorrhoeae	354	105	66	60	0.09	5	441	386	0.12	10	119	-	0.00	0
T-thioparus	235	54	33	29	0.12	9	261	235	0.10	10	89	66	0.26	2

Table 1. Evaluation of superoptimal structure prediction improvement for Beta Purple Bacteria. For each distance measure (coarse, weighted coarse and base pair metric), we report the distance between the phylogenetic structure and (i) the mfe structure of the original sequence (column denoted “or.”), (ii) that k -superoptimal structure, for $k \leq 10$ which is closest to the phylogenetic structure as measured by the distance measure under consideration (column denoted “s.o.”). We designate this k -superoptimal structure as the *best scoring structure*. The columns “rat.” compute the improvement ratio, defined by $(d(S_0, S_p) - d(S_{\rho_{opt}}, S_p)) / d(S_0, S_p)$, where S_p is the phylogenetic secondary structure, S_0 the mfe structure of the original (wild type) sequence, and $S_{\rho_{opt}}$ the superoptimal structure having best score. The columns ρ_{opt} list the number of mutations which occur for the best scoring superoptimal structure.

the Nussinov-Jacobson energy model. *J. Computational Biology*, 12(1):83–101.

Clote, P. and Backofen, R. (2000). *Computational Molecular Biology: An Introduction*. John Wiley & Sons. 279 pages.

Clote, P., Ferrè, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random rna of the same

dinucleotide frequency. *RNA*, 11(5):578–591.

Zuker, M. and Sankoff, D. (1984). RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621.

rna	length	#BP	coarse				w. coarse				BP dist.			
			or.	s.o.	rat.	ρ_{opt}	or.	s.o.	rat.	ρ_{opt}	or.	s.o.	rat.	ρ_{opt}
C-aurantiacus	324	89	51	31	0.39	4	323	231	0.28	4	108	90	0.17	3
H-aurantiacus	323	89	44	39	0.11	4	320	279	0.13	5	128	113	0.12	4
SMB-B3	283	71	28	13	0.54	10	319	222	0.30	10	130	93	0.28	3
T-album	328	91	48	37	0.23	7	335	315	0.06	4	123	93	0.24	3
T-roseum	345	95	37	34	0.08	8	275	-	0.00	0	127	97	0.24	5

Table 2. Evaluation of superoptimal structure prediction improvement for Green Non Sulfur Bacteria. For each distance measure (coarse, weighted coarse and base pair metric), we report the distance between the phylogenetic structure and (i) the mfe structure of the original sequence (column denoted “or.”), (ii) that k -superoptimal structure, for $k \leq 10$ which is closest to the phylogenetic structure as measured by the distance measure under consideration (column denoted “s.o.”). We designate this k -superoptimal structure as the *best scoring structure*. The columns “rat.” compute the improvement ratio, defined by $(d(S_0, S_p) - d(S_{\rho_{opt}}, S_p))/d(S_0, S_p)$, where S_p is the phylogenetic secondary structure, S_0 the mfe structure of the original (wild type) sequence, and $S_{\rho_{opt}}$ the superoptimal structure having best score. The columns ρ_{opt} list the number of mutations which occur for the best scoring superoptimal structure.

RNA id.	score	0	1	2	3	4	5	6	7	8	9	10	best
A-eutrophus	sens.	70.48	71.43	72.38	72.38	66.67	67.62	69.52	69.52	68.57	68.57	69.52	72.38
	spec.	64.35	64.66	64.96	63.87	57.85	57.72	59.35	59.84	58.54	58.06	58.40	64.96
B-bronchiseptica	sens.	86.76	91.18	91.18	67.65	67.65	69.12	94.12	92.65	94.12	94.12	94.12	94.12
	spec.	80.82	83.78	81.58	55.42	54.76	55.29	79.01	76.83	77.11	75.29	74.42	79.01
C-testosteroni	sens.	50.68	52.05	52.05	54.79	52.05	52.05	54.79	39.73	39.73	42.47	42.47	54.79
	spec.	46.84	47.50	46.91	50.00	45.78	44.71	47.06	34.12	33.33	35.63	34.83	50.00
ESH167F	sens.	49.32	50.68	47.95	47.95	38.36	38.36	38.36	38.36	38.36	38.36	38.36	50.68
	spec.	43.90	44.58	43.21	43.75	31.46	30.77	30.11	31.82	31.11	30.43	30.11	44.58
ESH183D	sens.	78.95	80.26	80.26	80.26	72.37	78.95	80.26	64.47	64.47	73.68	73.68	80.26
	spec.	77.92	78.21	74.39	72.62	63.22	69.77	70.11	52.13	52.13	62.22	60.87	78.21
ESH26-4	sens.	50.68	50.68	50.68	50.68	61.64	61.64	61.64	60.27	54.79	54.79	54.79	61.64
	spec.	50.00	46.84	46.25	44.58	51.72	51.14	49.45	48.89	44.44	43.01	41.67	51.72
Mxa1	sens.	49.43	49.43	49.43	49.43	49.43	49.43	49.43	49.43	44.83	44.83	44.83	49.43
	spec.	46.74	46.24	45.74	45.74	44.33	43.88	43.88	43.88	38.61	38.24	37.86	46.74
N-europaea	sens.	25.68	12.16	12.16	12.16	12.16	22.97	12.16	31.08	31.08	31.08	31.08	31.08
	spec.	21.84	10.71	10.59	10.47	10.23	19.77	10.23	23.96	23.71	23.47	23.00	23.96
N-gonorrhoeae	sens.	43.81	41.90	41.90	41.90	38.10	38.10	28.57	28.57	28.57	40.95	40.95	43.81
	spec.	43.40	40.37	39.64	38.94	35.09	34.78	25.00	24.79	24.59	33.59	33.33	43.40
T-thioparus	sens.	31.48	31.48	55.56	31.48	31.48	31.48	31.48	31.48	31.48	31.48	31.48	55.56
	spec.	24.64	23.29	40.54	22.67	22.37	22.08	22.08	21.79	21.52	20.73	20.48	40.54
Total	sens.	54.31	53.68	55.20	51.78	49.62	51.52	52.03	50.63	49.62	52.41	52.54	55.20
	spec.	50.12	48.68	49.38	45.48	42.22	43.47	43.48	41.82	40.48	42.10	41.57	49.38

Table 3. Evaluation of superoptimal structure prediction improvement for Beta Purple Bacteria. For each number k of mutation, we report the sensitivity ((# basepair correctly predicted) / (# observed basepairs in phylogenetic structure)) and specificity ((# basepair correctly pasepair) / (# observed basepairs in k -supersecondary structure)) evaluated between the phylogenetic structure and the corresponding k -superoptimal structure. The last column recall the score obtained at the level where the best sensitivity is observed.

RNA id.	score	0	1	2	3	4	5	6	7	8	9	10	best
C-aurantiacus	sens.	48.31	49.44	49.44	56.18	56.18	56.18	56.18	56.18	56.18	56.18	44.94	56.18
	spec.	40.95	44.90	43.14	48.08	47.62	45.87	45.45	44.64	43.86	44.25	35.09	48.08
H-aurantiacus	sens.	34.83	39.33	39.33	39.33	39.33	39.33	38.20	40.45	40.45	40.45	40.45	40.45
	spec.	30.69	34.65	33.98	33.65	35.71	35.00	31.48	32.73	32.43	32.14	31.58	32.73
SMB-B3	sens.	18.31	18.31	18.31	47.89	47.89	47.89	47.89	30.99	29.58	29.58	29.58	47.89
	spec.	15.29	15.12	14.94	36.56	36.17	35.42	35.42	22.45	21.00	20.79	21.43	36.56
T-album	sens.	40.66	40.66	40.66	58.24	58.24	29.67	29.67	29.67	29.67	29.67	29.67	58.24
	spec.	34.91	34.58	33.94	47.75	47.32	23.68	23.48	23.28	23.08	23.08	22.88	47.75
T-roseum	sens.	38.95	49.47	49.47	45.26	45.26	56.84	44.21	55.79	55.79	38.95	38.95	56.84
	spec.	34.91	44.76	44.34	37.39	37.07	46.96	35.59	45.30	44.92	30.58	30.58	46.96
Total	sens.	37.01	40.46	40.46	49.43	49.43	45.98	42.99	43.22	42.99	39.31	37.01	49.43
	spec.	32.01	35.41	34.71	40.80	40.95	37.45	34.19	34.00	33.39	30.32	28.50	40.95

Table 4. Evaluation of superoptimal structure prediction improvement for Green Non Sulfur Bacteria. For each number k of mutation, we report the sensitivity ((# basepair correctly predicted) / (# observed basepairs in phylogenetic structure)) and specificity ((# basepair correctly pasepair) / (# observed basepairs in k -supersecondary structure)) evaluated between the phylogenetic structure and the corresponding k -superoptimal structure. The last column recall the score obtained at the level where the best sensitivity is observed.

RNA id.	score	0	1	2	3	4	5	6	7	8	9	10	best
A-eutrophus	sens.	70.48	71.43	72.38	72.38	66.67	67.62	69.52	69.52	68.57	68.57	69.52	72.38
	spec.	64.35	64.66	64.96	63.87	57.85	57.72	59.35	59.84	58.54	58.06	58.40	64.96
B-bronchiseptica	sens.	86.76	91.18	91.18	67.65	67.65	69.12	94.12	92.65	94.12	94.12	94.12	94.12
	spec.	80.82	83.78	81.58	55.42	54.76	55.29	79.01	76.83	77.11	75.29	74.42	79.01
C-testosteroni	sens.	50.68	52.05	52.05	54.79	52.05	52.05	54.79	39.73	39.73	42.47	42.47	54.79
	spec.	46.84	47.50	46.91	50.00	45.78	44.71	47.06	34.12	33.33	35.63	34.83	50.00
ESH167F	sens.	49.32	50.68	47.95	47.95	38.36	38.36	38.36	38.36	38.36	38.36	38.36	50.68
	spec.	43.90	44.58	43.21	43.75	31.46	30.77	30.11	31.82	31.11	30.43	30.11	44.58
ESH183D	sens.	78.95	80.26	80.26	80.26	72.37	78.95	80.26	64.47	64.47	73.68	73.68	80.26
	spec.	77.92	78.21	74.39	72.62	63.22	69.77	70.11	52.13	52.13	62.22	60.87	78.21
ESH26-4	sens.	50.68	50.68	50.68	50.68	61.64	61.64	61.64	60.27	54.79	54.79	54.79	61.64
	spec.	50.00	46.84	46.25	44.58	51.72	51.14	49.45	48.89	44.44	43.01	41.67	51.72
Mxa1	sens.	49.43	49.43	49.43	49.43	49.43	49.43	49.43	49.43	44.83	44.83	44.83	49.43
	spec.	46.74	46.24	45.74	45.74	44.33	43.88	43.88	43.88	38.61	38.24	37.86	46.74
N-europaea	sens.	25.68	12.16	12.16	12.16	12.16	22.97	12.16	31.08	31.08	31.08	31.08	31.08
	spec.	21.84	10.71	10.59	10.47	10.23	19.77	10.23	23.96	23.71	23.47	23.00	23.96
N-gonorrhoeae	sens.	43.81	41.90	41.90	41.90	38.10	38.10	28.57	28.57	28.57	40.95	40.95	43.81
	spec.	43.40	40.37	39.64	38.94	35.09	34.78	25.00	24.79	24.59	33.59	33.33	43.40
T-thioparus	sens.	31.48	31.48	55.56	31.48	31.48	31.48	31.48	31.48	31.48	31.48	31.48	55.56
	spec.	24.64	23.29	40.54	22.67	22.37	22.08	22.08	21.79	21.52	20.73	20.48	40.54
Total	sens.	54.31	53.68	55.20	51.78	49.62	51.52	52.03	50.63	49.62	52.41	52.54	55.20
	spec.	50.12	48.68	49.38	45.48	42.22	43.47	43.48	41.82	40.48	42.10	41.57	49.38

Table 5. This table evaluates the performance of AMSAG with respect to base pair distance for RNAs of Beta Purple Bacteria, where the phylogenetic structure is known. For each value of $0 \leq k \leq 10$, we report the *sensitivity*, i.e. the number of correctly predicted base pairs in the k -superoptimal structure divided by the number of base pairs in the phylogenetic structure. Additionally, we report the *specificity*; i.e. the number of correctly predicted base pairs in the k -superoptimal structure divided by the number of base pairs in the k -superoptimal structure. The last column gives the score obtained at the level where highest sensitivity is observed.

RNA id.	score	0	1	2	3	4	5	6	7	8	9	10	best
C-aurantiacus	sens.	48.31	49.44	49.44	56.18	56.18	56.18	56.18	56.18	56.18	56.18	44.94	56.18
	spec.	40.95	44.90	43.14	48.08	47.62	45.87	45.45	44.64	43.86	44.25	35.09	48.08
H-aurantiacus	sens.	34.83	39.33	39.33	39.33	39.33	39.33	38.20	40.45	40.45	40.45	40.45	40.45
	spec.	30.69	34.65	33.98	33.65	35.71	35.00	31.48	32.73	32.43	32.14	31.58	32.73
SMB-B3	sens.	18.31	18.31	18.31	47.89	47.89	47.89	47.89	30.99	29.58	29.58	29.58	47.89
	spec.	15.29	15.12	14.94	36.56	36.17	35.42	35.42	22.45	21.00	20.79	21.43	36.56
T-album	sens.	40.66	40.66	40.66	58.24	58.24	29.67	29.67	29.67	29.67	29.67	29.67	58.24
	spec.	34.91	34.58	33.94	47.75	47.32	23.68	23.48	23.28	23.08	23.08	22.88	47.75
T-roseum	sens.	38.95	49.47	49.47	45.26	45.26	56.84	44.21	55.79	55.79	38.95	38.95	56.84
	spec.	34.91	44.76	44.34	37.39	37.07	46.96	35.59	45.30	44.92	30.58	30.58	46.96
Total	sens.	37.01	40.46	40.46	49.43	49.43	45.98	42.99	43.22	42.99	39.31	37.01	49.43
	spec.	32.01	35.41	34.71	40.80	40.95	37.45	34.19	34.00	33.39	30.32	28.50	40.95

Table 6. Similar evaluation, as explained in Table 5, for Green Non Sulfur Bacteria.

RNA id.	score	0	1	2	3	4	5	6	7	8	9	10	best
C-opuntiae	sens.	87.80	97.56	70.73	80.49	92.68	75.61	75.61	51.22	51.22	51.22	51.22	97.56
	spec.	58.06	63.49	41.43	46.48	58.46	43.66	42.47	28.77	28.38	28.00	28.00	63.49
P-mississippiensis	sens.	94.44	85.19	94.44	94.44	77.78	87.04	87.04	87.04	77.78	77.78	77.78	94.44
	spec.	77.27	70.77	73.91	72.86	59.15	62.67	61.84	61.04	53.85	53.16	52.50	77.27
R-norvegicus	sens.	25.00	25.00	25.00	25.00	25.00	25.00	25.00	10.00	10.00	10.00	10.00	25.00
	spec.	11.63	11.76	11.24	11.11	10.99	10.87	10.64	4.26	4.21	4.12	4.08	11.76
S-japonicus	sens.	68.25	68.25	69.84	69.84	69.84	69.84	69.84	69.84	69.84	69.84	69.84	69.84
	spec.	44.33	43.43	45.36	44.90	44.90	44.44	44.00	43.56	43.14	42.72	42.31	45.36
S-octosporus	sens.	55.56	52.38	52.38	53.97	53.97	46.03	46.03	46.03	46.03	19.05	19.05	55.56
	spec.	37.63	35.48	35.48	35.79	34.69	29.90	29.00	29.00	28.16	11.32	11.11	37.63
S-pombe	sens.	50.00	50.00	52.94	52.94	52.94	52.94	52.94	54.41	54.41	54.41	54.41	54.41
	spec.	35.42	34.69	37.50	37.11	37.11	36.73	36.36	37.00	36.63	36.27	35.92	37.00
S-versatilis	sens.	68.25	68.25	69.84	69.84	69.84	69.84	69.84	69.84	69.84	69.84	69.84	69.84
	spec.	44.33	43.43	45.36	44.90	44.90	44.44	44.00	43.56	43.14	42.72	42.31	45.36
T-syrichta	sens.	14.29	14.29	14.29	19.05	19.05	16.67	21.43	14.29	19.05	19.05	19.05	21.43
	spec.	6.74	6.59	6.52	8.60	8.89	8.14	10.34	6.25	8.25	8.16	8.08	10.34
W-fluorescens	sens.	98.00	98.00	98.00	98.00	96.00	96.00	96.00	96.00	40.00	40.00	40.00	98.00
	spec.	63.64	62.82	62.03	61.25	59.26	58.54	57.83	57.14	21.51	21.28	21.28	63.64
Z-bailii	sens.	66.67	66.67	66.67	74.36	66.67	66.67	51.28	51.28	66.67	51.28	53.85	74.36
	spec.	44.07	43.33	42.62	42.03	41.94	40.62	31.25	29.41	36.62	27.78	28.77	42.03
Z-rouxii	sens.	56.45	56.45	53.23	53.23	53.23	53.23	53.23	53.23	54.84	53.23	53.23	56.45
	spec.	35.00	34.65	32.35	31.43	31.43	31.13	30.84	29.46	30.36	29.20	28.95	34.65
Total	sens.	63.43	62.96	62.50	64.04	62.81	61.57	60.96	58.18	54.48	50.77	50.93	64.04
	spec.	40.33	39.57	38.87	39.00	38.61	37.36	36.47	34.06	31.24	28.73	28.55	39.00

Table 7. Evaluation of superoptimal structure prediction improvement for Nuclear. For each number k of mutation, we report the sensitivity ((# basepair correctly predicted) / (# observed basepairs in phylogenetic structure)) and specificity ((# basepair correctly predicted) / (# observed basepairs in k -supersecondary structure)) evaluated between the phylogenetic structure and the corresponding k -superoptimal structure. The last column recall the score obtained at the level where the best sensitivity is observed.

```
#folding energy = -24.15
#max. number of mutations = 0
GGGGCTATAGCTCAGCTGGGAGAGCGCCTGCTTTGCACGCAGGAGTCTGCGGTTCGATCCCGCATAGCTCCACCA
((((((((((...))))(((((((...(((...))))(((((((...))))))))))...))))))....

#folding energy = -28.54
#max. number of mutations = 1
GGGGCTATAGCTCAGCTGGGAGAGCGCCTGCTTTGCACGCAGGAGTCTGCGGTTCGATCCCGCATAGCTCCACCA
((((((((((...))))(((((((...(((...))))(((((((...))))))))))...))))))....

#folding energy = -33.76
#max. number of mutations = 2
GGGGCTATAGCTCAGCTGGGAGAGCGCCTGCTTTGCACGCAGGAGTCTGCGGTTCGATCCCGCATAGCTCCACCA
((((((((((...))))...((((((((((.....))))))))))(((((((.....))))))))))....
```

Fig. 4. The set $\mathcal{M}_{\leq 2}$ of k -superoptimal structures for $k \leq 2$ for a tRNA from *E. coli* with anticodon TGC. Note that the anticodon and the correct structure of the tRNA can be retrieve with 2 mutations.

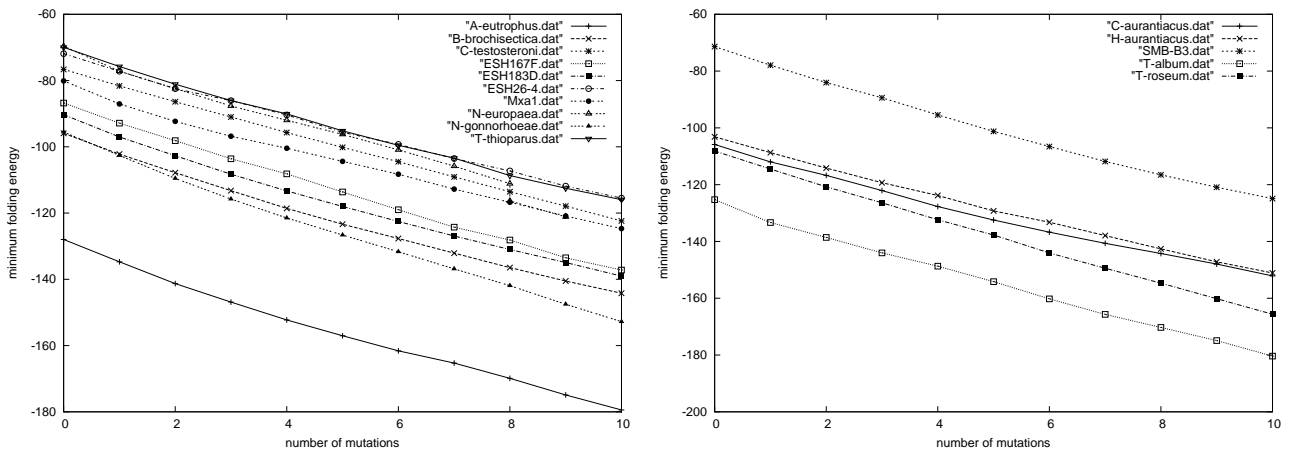


Fig. 6. Minimum folding energy values as a function of the maximum number of mutations allowed in a given RNA sequence. Left diagram displays graphs for the sequences from the Beta Purple Bacteria family, while the right diagram displays graphs for the sequences from the Green Non-Sulfur Bacteria family. In data on the web supplement, we display similar graphs for all values of $0 \leq k \leq 50$, for 36 SECIS elements, each of length 50. The SECIS data shows that the displayed graphs in this figure are only initially linear portions of a graph which levels out to a horizontal asymptote. The asymptotic value is the energy value, corresponding to the AMSAG-computed minimum free energy of the 50-mer consisting of 25 guanines followed by 25 cytosines.

