Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 26, Issue Number 1, (2008)

Reconstruction of Natural RNA Sequences from RNA Shape, Thermodynamic Stability, Mutational Robustness, and Linguistic Complexity by Evolutionary Computation

http://www.jbsdonline.com

N. Dromi[1,2]
A. Avihoo[1]
D. Barash[1,3,*]

[1]Department of Computer Science
Ben-Gurion University
Beer-Sheva 84105, Israel
[2]Rosetta Genomics
Weizmann Science Park
Rehovot 76706, Israel
[3]Institute of Evolution
University of Haifa
Haifa 31905, Israel

Abstract

The process of designing novel RNA sequences by inverse RNA folding, available in tools such as RNAinverse and InfoRNA, can be thought of as a reconstruction of RNAs from secondary structure. In this reconstruction problem, no physical measures are considered as additional constraints that are independent of structure, aside of the goal to reach the same secondary structure as the input using energy minimization methods. An extension of the reconstruction problem can be formulated since in many cases of natural RNAs, it is desired to analyze the sequence and structure of RNA molecules using various physical quantifiable measures. In prior works that used secondary structure predictions, it has been shown that natural RNAs differ significantly from random RNAs in some of these measures. Thus, we relax the problem of reconstructing RNAs from secondary structure into reconstructing RNAs from shapes, and in turn incorporate physical quantities as constraints. This allows for the design of novel RNA sequences by inverse folding while considering various physical quantities of interest such as thermodynamic stability, mutational robustness, and linguistic complexity. At the expense of altering the number of nucleotides in stems and loops, for example, physical measures can be taken into account. We use evolutionary computation for the new reconstruction problem and illustrate the procedure on various natural RNAs.

Introduction

The problem of computationally predicting an RNA secondary structure given a sequence has been advanced extensively over the past three decades. Moreover, both RNA sequence and structure have been subjected to computational and theoretical studies covering a variety of different aspects. One interesting direction involves investigating the physical properties of an RNA secondary structure, addressing questions such as how the secondary structures of natural and random RNA sequences differ and how they evolve, a review of which is available in (1). It is reasonable to assume that physical properties of an RNA secondary structure can be put to use in the process of computationally designing novel RNA sequences with favorable characteristics, given an RNA secondary structure for which several RNA sequences fold into by energy minimization. This problem is known as the RNA inverse folding problem.

The problem of inverse folding of RNAs was discussed at length in the seminal paper on the Vienna RNA package (2), where an algorithm called RNAinverse was derived and has subsequently been used as the gold standard in a variety of computational simulations since then. Recently, improvements to RNAinverse have been suggested in terms of computational efficiency and different methodologies, among which are the RNA-SSD (3) and the INFO-RNA (4). In RNAinverse, the strategy of adaptive walk was used and local optima were found according to two different

[*]Email: dbarash@cs.bgu.ac.il

criteria, namely a structural distance between the minimum free energy structure of the designed sequence and the target structure (mfe-mode) and the probability of folding into the target structure (p-mode). The RNA-SSD (RNA Secondary Structure Designer) is based on a recursive stochastic local search that also tries to minimize a structural distance. The INFO-RNA (Inverse Folding of RNA) is similar to RNAinverse with the following contributions: a dynamic programming algorithm chooses a sequence that adopts the lowest energy a sequence can have when folding into a target structure, and the mutating step of the sub-sequences is performed by a stochastic local search that can avoid local minima. More details can be found in the references above for these three main RNA inverse folding algorithms.

A natural RNA (*e.g.*, an RNA virus or a microRNA) differs from an inversely folded RNA not only by sequence composition, but also by physical properties. When designing RNA molecules, those differences may potentially be quite important, and therefore there is a clear motivation to insert physical considerations into the design problem. Here, it is proposed to include physical properties in the following way. When simulating RNA evolution, several quantifiable measures have been suggested to analyze the sequence and structure of an RNA molecule. Among others [*e.g.*, (5)], three basic measures that have been discussed in prior works and can be calculated by computational means are thermodynamics stability, mutational robustness, and linguistic complexity. These measures can be added to the RNA inverse folding problem as constraints, and they are introduced in the next paragraph while their exact calculation is described in the ***Methods*** section.

The question of how natural RNA sequences differ from random ones has been investigated previously in several works. For example, it was found that tRNAs are unusually stable thermodynamically compared with random RNA sequences of the same length and base compositions (6). A comprehensive review on the physical and computational aspects of RNA secondary structure is available in (1). In a recent work of Borenstein and Ruppin (7) it was shown that the structure of miRNA precursor stem-loops are significantly more mutationally robust in comparison with random RNA sequences with similar stem-loop structures. Aside of thermodynamic stability and mutational robustness, an additional physical measure that was investigated in (8, 9) is linguistic complexity. The linguistic complexity of natural RNA sequences is expected to be lower than that of random sequences. For example, repetitive sequences in the eukaryotic genome, such as tandem repeats or CpG islands in genes promoters, reduce the total complexity of natural sequences as compared to random ones. This can also be attributed to the fact that millions of years of random mutations would have reduced the high complexity of ancient RNA sequences, unless the complex sequences are maintained as such by specific selection pressure. Complexity is expected to be reduced in a system that is not completely deterministic, which is definitely the case in the presence of random mutations, because an increase in entropy due to the second law of thermodynamics translates to a decrease in complexity. This also makes the higher sequence complexity an indication of some functional load carried by the sequence. Thus, thermodynamic stability, mutational robustness, and linguistic complexity are three examples of physical measures that may contribute to RNA inverse folding by providing robust constraints.

Here, the design problem of RNA sequences from secondary structure, known as the RNA inverse folding problem, is revisited with the goal of incorporating physical measures and thereby allowing a useful extension for designing RNA sequences that mimic the behavior of natural RNAs. In the traditional formulation of the problem, the inverse folding is performed from RNA secondary structure to sequences, without taking into account physical constraints (although, other constraints are allowed, such as sequence constraints). Thus, as a first step, the idea of incorporating physical measures as constraints into the RNA inverse folding can be suggested. However, when attempting to include such constraints while restricting both secondary structure and physical measures, no solutions are ob-

tained in the majority of cases. To resolve this issue, it is argued that at least in a considerable number of cases, physical measures are more important than strictly obeying the secondary structure instead of allowing the inclusion of an extra base pair in a stem, for example. Thus, instead of performing the inverse folding from secondary structure to sequences, one can relax the problem to perform the inverse folding from an RNA "shape" to sequences, and thereby incorporate physical measures as constraints. An RNA shape is a family of structures, sharing a common pattern of nested and adjacent helices. For example, as demonstrated in (10-12), an RNA shape can be taken as a coarse-grained representation of the secondary structure, such as a tree-graph representation that captures the secondary structure motifs but not the complete information of all the base pairings that form the secondary structure. The idea in the relaxation to shapes is to provide as input an RNA coarse-grained representation of the secondary structure, derived from a given natural RNA sequence, instead of its secondary structure (for an illustration of the difference between RNA secondary structures and RNA shapes, see Figure 1). This permits taking into account the various physical measures calculated from the given RNA sequences as constraints to the RNA inverse folding problem. The desired outputs are RNA sequences that possess the same shape as the initially given RNA sequence, with the additional trait that the resultant RNA sequences are also similar in their physical properties to the initial one. In at least some design problems it is anticipated that this procedure will offer an improvement over RNA inverse. If one of the designed sequences and the input natural RNA sequence are found to have highly distinct secondary structures then this designed sequence can be discarded in favor of another, but there should be many cases of two secondary

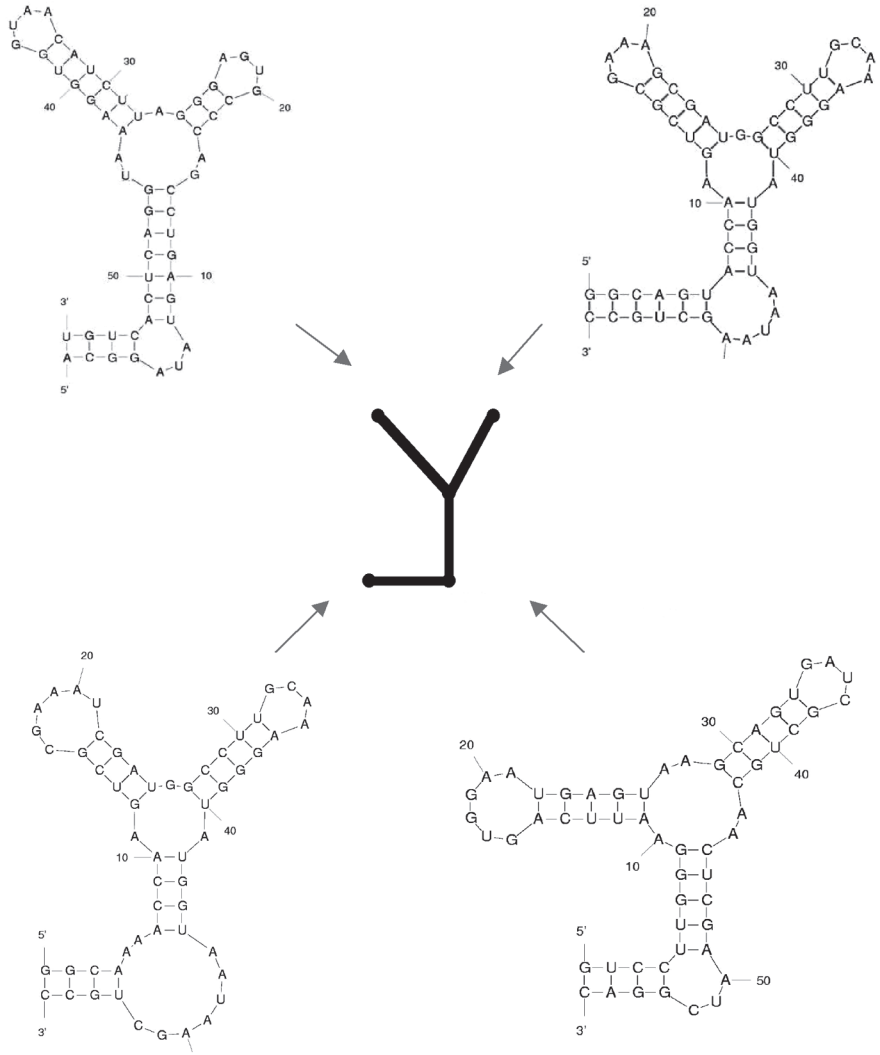


Figure 1: An illustration of RNA secondary structures vs. RNA shapes.

structures that are slightly different but relax into the same shape (see illustration in Figure 1), and therefore the relaxation from secondary structure to shape in order to accommodate for the additional constraints is justified. Without this relaxation there will be no solution to the extended design problem in the majority of the cases, while with this relaxation many interesting candidate solutions that mimic the favorable properties of natural RNAs can be found, as will be illustrated in the ***Results and Discussion*** section. It should be noted that such a relaxation may also produce different three-dimensional structures that do not necessarily have the same properties as the original naturally folded sequences, an issue that should be considered, but these three-dimensional structures may also occur in locations that are not biologically important or relevant to the design problem.

Throughout our work, we use an RNA optimization-led simulator where the simulations are carried out by a simple evolutionary computation strategy as described in the ***Methods*** section. It should be noted, however, that an evolutionary scheme like the simple genetic algorithm (GA) employed in this work is not necessarily a good model for RNA evolutionary dynamics. The justification for the usage of a simple GA in this problem is in terms of its ability to solve multi-objective optimization problems and its convenience in the implementation. Computational simulations and analyses are important for understanding theoretical notions in RNA evolution such as neutral networks (13), continuity in evolution (14), topology and fitness landscapes (15), distribution of beneficial fitness effects (16), and mutational robustness (7, 17, 18). Evolutionary computation (EC) may well assist in performing additional simulations in that context to model RNA evolution by computational means. Although we chose a rather ordinary evolutionary computation strategy in this work, employing a simple GA on a parallel platform in a master-slave fashion as described in (19), more advanced strategies as in (20, 21) may prove beneficial in a variety of problems associated with modeling RNA evolution.

In the next section, we describe the computational methods used in this work, starting from the procedures that are used to calculate the three physical measures. Finally, in the last section, we show results obtained by our inverse RNA folding method that incorporates the physical measures and discuss their potential advantage in comparison to the standard inverse RNA folding.

Methods

First, the procedures to calculate the three physical measures discussed in the previous section are provided. Second, the extraction of the RNA secondary structure to a simplified coarse-grained representation, or equivalently its shape, is described. Third, details of the evolutionary computation strategy used in this work for performing the extended RNA inverse folding from shape and physical measures to sequences is given.

Thermodynamic Stability

The first measure used for incorporation into the RNA inverse folding problem is thermodynamic stability. In accordance with previous works described in the review paper on RNA secondary structure that considers physical aspects (1), we will characterize an RNA molecule as thermodynamically stable if: (i) its ground state free energy is low; (ii) the RNA has few alternative secondary structures, if at all, at energies close to the ground state energy (6). In the example test cases reported (see below), without loss of generality, we measured thermodynamic stability in terms of the ground state free energy as predicted by energy minimization, in units of kcals/mole. It is possible to add the difference between the ground state free energy and that of the alternative structure with the lowest free energy to the estimation of the thermodynamics stability, depending on the problem at hand and what type of sequences are desired as output in the design

problem. Adding this additional constraint to the RNA inverse folding problem can significantly reduce the number of solutions obtainable and may prove important in some design problems. Thus, although we used property (i) above for the purpose of illustration, it is possible to combine properties (i) and (ii) for the estimation of thermodynamic stability in future work.

Mutational Robustness

The second measure is the robustness of the RNA molecule to remain with the same secondary structure as a response to single point mutations. We measured the mutational robustness in accordance with the neutrality calculation described in (7). That is, the neutrality of an RNA sequence of length L is calculated by $\eta=\langle(L-d)/L\rangle$, where d is the distance between the secondary structure of the original sequence and the secondary structure of the mutant, averaged over all $3L$ one-mutant neighbors. We used Vienna's RNAdistance routine to evaluate d using a tree-edit distance. Although in this work we evaluated d according to a tree-edit distance between secondary structures, for the purpose of a better comparison with RNAinverse, it should be noted that when relaxing the RNA inverse folding problem to finding similar shapes then the evaluation of d can be performed using RNAdistance with Bruce Shapiro's coarse grained representation option (essentially calculating a tree-edit distance between shapes). In any case, η is a number between zero and one, representing the average fraction of the structure that remains unchanged after a mutation occurs, thereby measuring mutational robustness.

Linguistic Complexity

For the third measure, we refer to the notion and measure of linguistic complexity introduced in (8). In accordance with the formalism of linguistic complexity, every sequence can be characterized by its vocabulary. One can measure the complexity by the extent to which the maximally non-repetitive possible vocabulary is used. For any sequence of length L its complexity C is defined as the product of Ui's, where i ranges from 1 to $L-1$, and Ui corresponds to the ratio of the actual to maximal vocabulary sizes for word length i. Originally, the linguistic complexity was used for the analysis of DNA and protein sequences, as well as for texts of human languages (9). Here, the linguistic complexity is being used for the analysis of an RNA molecule in measuring sequence complexity for the 4-letter alphabet 'A', 'U', 'C', and 'G' (another way is to measure the structure complexity by using the 3-letter alphabet '(', '.', ')', generated by Vienna's dot-bracket representation). The linguistic complexity is a number between zero and one and the calculation is relatively fast since no secondary structure prediction is necessary.

Relaxation of the RNA Inverse Folding Problem into Shapes

For incorporating the physical measures listed above as constraints in the inverse RNA folding problem, one needs to relax the problem since such constraints can yield no solution as was noticed in our simulations. Thus, instead of the RNA secondary structure, we use the simplified coarse-grained representation (10) as its shape. This is a convenient choice provided by the routine 'b2shapiro' in the Vienna RNA package (2). There are other ways to represent shapes, such as in the recent work on RNA abstract shapes (12). As a consequence, the inverse RNA folding problem becomes a reconstruction problem given an RNA shape and physical measures as constraints, yielding desired solutions in each case as will be demonstrated in the next section. The justification for relaxing the RNA inverse folding problem from that of an RNA secondary structure to an RNA shape emanates from the fact that if in the designed sequences there are a few more nucleotides in stems/loops but nevertheless these motif elements remain the same as in the initial input, such sequences are interesting to examine as candidate solutions to the design problem.

Number of Constraints

It is worthwhile noting that flexibility can be kept when considering how many physical measure constraints to impose and the accuracy of obeying each constraint, which is dependent on the design problem at hand. For example, one can require to obey all three constraints but allow a certain deviation for each one, or one can require to obey only one constraint with almost no deviation from it. For the purpose of illustration in most of the example test cases of our manuscript except the last one, the latter requirement is chosen, without loss of generality. This enables to find a solution from the one hand, because of avoiding the complication of trying to meet several constraints, but it is not easy to find a solution from the other hand, because of demanding only a slight deviation (our choice was a deviation of 0.0001 in all measures) from the wildtype value. In multiple constraints problems, it is possible that there will be no solutions at all or a variety of solutions in case the constraints are easy to meet, depending on the specific problem and the specific constraints that are dealt with. For example, in the final test case of Figure 9, we test our method on a multiple constraint problem that can be of interest and indeed find a solution.

Proposed Methodology

Neglecting efficiency considerations at present, we suggest a simple evolutionary computation strategy as means to solve the reconstruction problem. Basically, given a natural sequence, one can extract its shape and physical properties, followed by constructing other sequences that possess the same shape and physical properties. In order to construct such sequences, one can try to simulate numerous sequences as solution candidates, fold them by energy minimization prediction methods, extract their shape and physical properties, and check if these match the initial shape and physical measures. Despite the fact that for a reliable use of energy minimization prediction one would favor a relatively short RNA sequence (<100 nts), the sequence space is still considerably large. In order to perform a guided search for solution candidates, instead of by trial and error, it is convenient to employ a simple genetic algorithm in the case of RNA sequences. Instead of the standard binary string representation used in the traditional genetic algorithm (19), one can use a quad-string representation in which each allele contains 'A' , 'U', 'C', and 'G'.

As a consequence, we have developed an optimization-based RNA sequence and structure evolution simulator. The simulator implements a simple genetic algorithm strategy, whereby an individual in the population is an RNA sequence. The population of the first generation is a collection of random sequences. In order to determine the population in the next generation, we use an objective ("fitness") function that yields advantage to sequences for which their reduction to the Shapiro coarse-grained representation (2, 10) is close in distance to the initial Shapiro coarse-grained representation of target RNA sequence. Implicitly, thermodynamic stability is contained in this fitness function since the generated sequences are folded using a thermodynamically based algorithm and their produced shape checked, although this does not interfere with imposing an additional thermodynamic stability constraint, in case it is desired to choose specifically those sequences that are as thermodynamically stable as the wildtype sequence. Distances between Shapiro coarse-grain representations are calculated in a straight-forward way using the RNAdistance routine available in the Vienna RNA package, which computes a tree edit distance (2). Thus, we would like to optimize the fitness function that essentially employs a tree edit distance desired to be minimized. In the passage from generation i to generation i+1, we perform several steps that are common in a simple GA. First, a selection operator is used, in order to give preference to "better" sequences by allowing them to pass to the next generation. The fitness of each sequence is determined by the fitness function mentioned above. An "elite" group of the "best" selected sequences (the group's size is determined by a predefined parameter that is usually taken as 10% of the general population size) is transferred from generation *i* to generation *i+1* with-

out a change in sequence composition. Second, a single point crossover operator is used, by taking two sequences that were chosen in the previous step along with a random crossover point, and creating two sequences from the constituents of those sequences surrounding the crossover point. The occurrence of a crossover event between two selected sequences is determined by a predefined probability value (the value of 0.9 is used in our implementation). Third, mutations on the new sequence created in the previous step are performed according to a small predefined probability value (the value of 0.1 is used in our implementation). When the GA arrives to the final generation according to some threshold parameter, the process stops. The solution is recorded and a new procedure starts again, since this is a stochastic method and similar to the approach used in RNAinverse there is no guarantee that all solutions will be found. The solutions that are found can be far away in terms of sequence composition from the wildtype sequence, unless a sequence conservation is desired to some extent, in which case the method can be modified accordingly, for example by starting with an initial population of sequences containing the same base composition as the wildtype. The size of the population (number of sequences in each generation) is also determined by a predefined parameter value that will be stated and explained below in the next subsection on parallelization.

Parallelization

To reduce the amount of computation to a reasonable level that enabled us to generate the results appearing in the next section within several hours, we used a parallel platform with a simple "master-slave" strategy that operates in the following way (see Figure 2 for illustration). The procedure is divided into two stages. The first stage, being the more computationally intensive, is performed by some "slave" processors that communicate their results to a "master" processor, which in turn is responsible for the second stage. In the first stage, we only use the distance between two Shapiro coarse-grained representations as the fitness function. Each time we find a sequence in which there is an exact match with the target sequence as far as their respective coarse-grained Shapiro representations (optimal fitness), we communicate this sequence to the "master" processor and continue in order to find more such sequences in parallel with the "slave" processors. In the second stage, the population of the first generation contains the sequence with the desired shape obtained from a "slave" processor during the first stage while including some additional random sequences. The fitness function is formulated as a combination of how close the sequences are to the target sequence in terms of one of the physi-

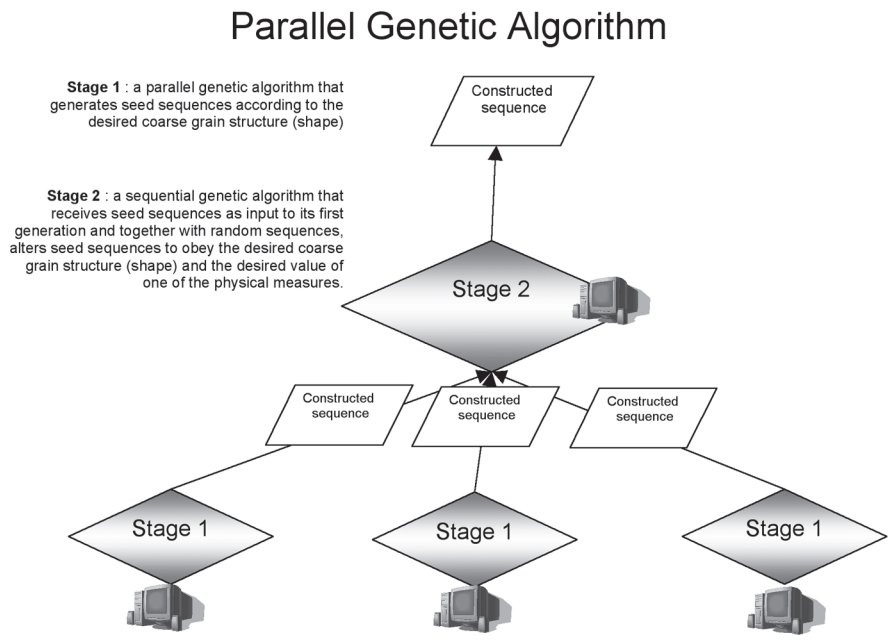


Figure 2: The "master-slave" strategy with the two stages used for the parallel genetic algorithm.

cal measures chosen, *e.g.*, mutational robustness, and the distance of the Shapiro coarse-grained representation of the sequences to that of the target sequence as in the first stage. The GA in the second stage allows favorable variations in the sequence that was transmitted from the first stage to accommodate the physical measure constraint, while retaining its desired shape. After the algorithm finds the final sequence composition as required (optimal fitness) when using the newly formulated fitness function, being a combination of shape and physical measure proximities, the sequence is written to a 'Results' file, and the GA of the second stage starts all over again with the first generation containing random sequences and a new sequence transmitted from the first stage. The first stage is performed with a large population size (150-1000 individuals) and relatively few generations (50-200), whereas the second stage is executed with a small population size (~10 individuals) and a large number of generations (200-2000). The combination of the two stages described above and illustrated in Figure 2 was found to yield in practice an efficient procedure for finding the results reported in the next section, compared to other parallel strategies that have been tried with less success.

Results and Discussion

First, we demonstrate that the numerical values of the three measures described in the previous section (thermodynamic stability, mutational robustness, and linguistic complexity) are significantly different in our test case examples of natural RNAs vs. random RNA sequences of the same length. Thus, the motivation to include these measures as constraints when solving the standard RNA folding problem for generating designed sequences with favorable properties is clear. Second, we show concrete examples of solutions to the RNA inverse folding problem by our method that take into account each one of the measures separately as a constraint, while preserving the shape and deviating only slightly with respect to the secondary structure of the input RNA sequence. These examples serve as illustration for our method and its potential success in other cases of interest.

Validation for Random vs. Natural RNAs

Tables I-III show the results of the comparison between the thermodynamic stability, mutational robustness, and sequence linguistic complexity values, respectively, of random RNA sequences vs. either natural microRNA sequences or natural RNA sequence pieces cut from a model ribosome. The dataset of the microRNA sequences was constructed from the miRNA Registry (25) and included 1732 sequences of sizes 58-100 nts (most of the sequences were 70-100 nts long). The ribosomal RNA dataset consisted of 20 short pieces (ranging from 39-93 nts in length) that were computationally cut from a well-known ribosome structure for which an experimentally derived secondary structure is available (26). On each piece, the secondary structure prediction by energy minimization conveys high

Table I

Thermodynamic stability characteristics of random sequences compared to natural RNA sequences. The natural sequences are significantly more stable than the random sequences. Suboptimal solutions in this case were calculated using mfold (22) with the energy rules of (23) and that mimic the favorable properties of natural ones with default percentage of suboptimality by the method of (24). Energy units in the table are calculated in Kcals/mole.

RNA type	Distance from the first sub-optimal solution Kcals/mole		Minimum free energy Kcals/mole		Number of sub-optimal solutions near the ground state	
	Average	Standard-deviation	Average	Standard-deviation	Average	Standard-deviation
Pre-miRNA	4.48	2.256	-36.36	6.8	1.336	1.34
Ribosomal RNA	5.7349	2.42	-30.7	12.47	0.6	0.8
Random sequences	1.0377	0.9559	-17.85	3.29	11.35	5.44

secondary structure agreement with the experimental result. The random RNA dataset included 10,000 random sequences with uniform base-composition, and same length distribution as the dataset of the microRNA sequences.

The results of the comparison between the specific natural RNA sequences we chose for illustration vs. random sequences, in all of the three measures (Tables I-III), show as expected that natural RNA molecules have significantly higher thermodynamic stability and mutational robustness, and a lower sequence linguistic complexity compared to random RNA sequences. In addition, since one may claim that a large share of the anticipated increase or decrease levels of these measures can be attributed to an intrinsic outcome of the secondary structure, *e.g.*, any RNA sequence that folds into a stem-loop structure will be more thermodynamically stable and mutationally robust than a random sequence, we included Table IV. In Table IV, a comparison between natural sequences and their inversely folded sequences as a reference is performed, to show that the measures behave as expected not only in reference to completely random sequences. We note that in some cases, such as the mir-30 in Table IV and the mir-146 in Figure 3, the sequence linguistic complexity of the inversely folded sequence may sometime be lower than the natural sequence but there is no attribution to an intrinsic outcome of the secondary structure in this measure that is sequence based. This measure is perhaps not sensitive enough to detect a clear trend unless natural vs. completely random sequences are being compared, as in Table III. However, with the thermodynamic stability and mutational robustness, there is clearly a decrease for the inversely folded sequences with respect to their natural input sequences. To conclude, inclusion of all these physical measures in the inverse folding problem is potentially desired for designing sequences with favorable properties.

Table II
The average and standard deviation values of mutational robustness of natural RNAs compared to random RNA sequences. The natural RNAs are significantly more robust than the random sequences (the robustness is a numerical value between zero and unity, as explained in the ***Methods*** section).

RNA type	Average mutational robustness	Standard deviation of mutational robustness
Pre-miRNA	0.932	0.0257
Ribosomal RNA	0.9088	0.0442
Random sequences	0.7449	0.106

Table III
The sequence linguistic complexity of natural RNAs compared to random RNA sequences.

RNA type	Average sequence complexity	Standard deviation of sequence complexity
Pre-miRNA	0.507	0.0748
Ribosomal RNA	0.4395	0.101
Random sequences	0.555	0.0776

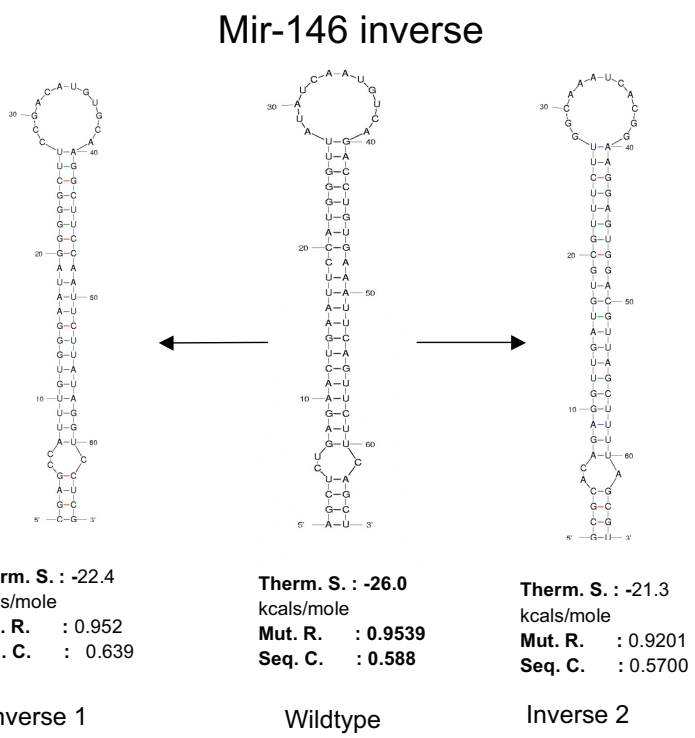
Table IV
The average values of the measures of 500 sequences constructed using RNA-inverse for three natural sequences (compared with the values of the measures of the natural sequences from which they were constructed). The table shows that the sequences constructed by RNA-inverse have lower thermodynamic stability and mutational robustness, and in most cases higher linguistic complexity than the natural sequences (the same phenomena that was observed for random sequences).

Molecule (the target sequence)	Thermodynamic stability Kcals/mole		Mutation robustness		Sequence linguistic complexity	
	average	Standard deviation	average	Standard deviation	average	Standard deviation
P5abc sub-domain **inverse**	**-15.21**	**4.01**	**0.766**	**0.071**	**0.564**	**0.067**
P5abc sub-domain	**-25.6**	**-**	**0.9458**	**-**	**0.557**	**-**
Ribosomal-RNA 53 **inverse**	**-14.21**	**3.36**	**0.7978**	**0.071**	**0.5507**	**0.0771**
Ribosomal-RNA 53	**-30.5**	**-**	**0.9563**	**-**	**0.38**	**-**
Pre-mir30 **inverse**	**-35.01**	**5.625**	**0.950**	**0.0199**	**0.532**	**0.0725**
Pre-mir30	**-40.6**	**-**	**0.9694**	**-**	**0.6521**	**-**

Comparison of Suggested Method with RNAinverse

Next, we compared our suggested extension to the RNA inverse folding problem that includes the physical measures against RNAinverse (2), without loss of generality as the same conclusions hold in case we compare our method with InfoRNA (4) or RNA-SSD (3). The comparison was performed for three test cases, in which their structure is well predicted by energy minimization: mir-146 (27), P5abc subdomain of the *tetrahymena thermophila* group I intron ribozyme (28), and a piece

Figure 3: mir-146 solution of RNAinverse. In the middle, the secondary structure drawing of the input sequence is given. On the sides, two secondary structure drawings of RNAinverse representative solutions are given. Values for the physical measures appear below the drawings. Values for the input reference sequence are in bold.



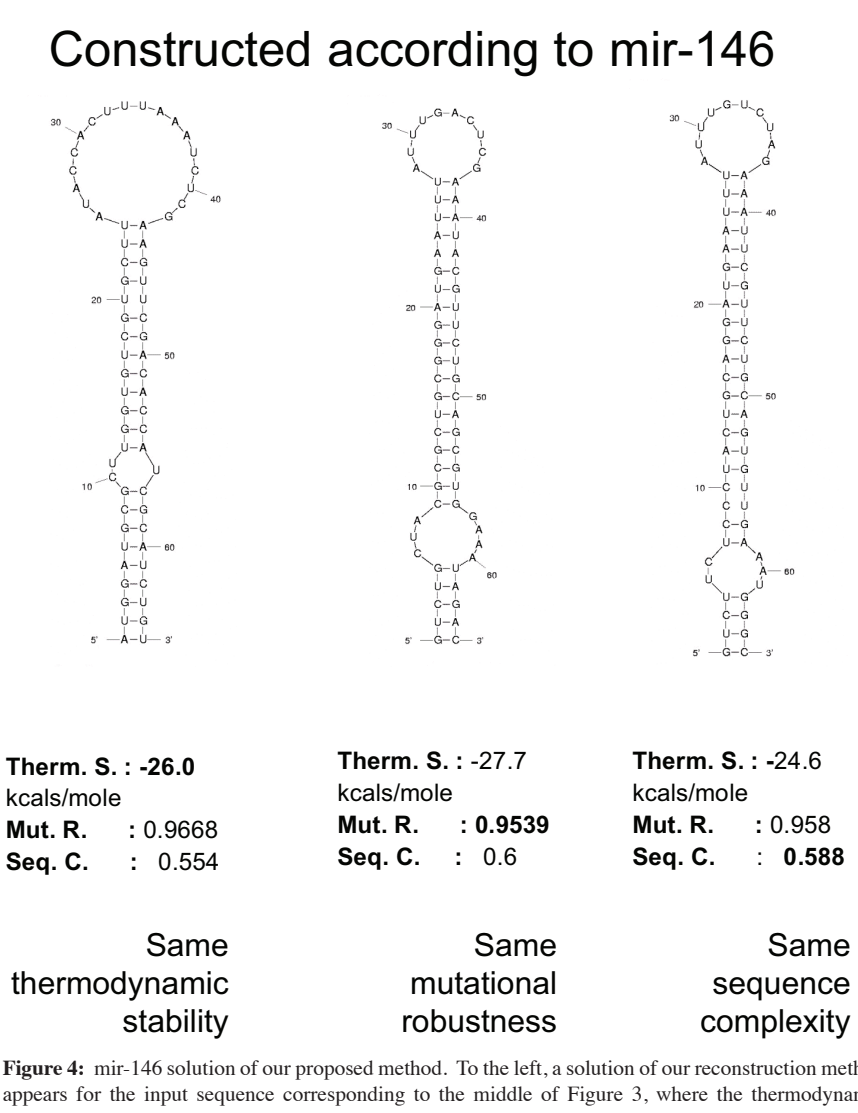


Figure 4: mir-146 solution of our proposed method. To the left, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 3, where the thermodynamic stability is the same as in the input sequence and appears in bold. In the middle, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 3, where the mutational robustness is the same as in the input sequence and appears in bold. To the right, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 3, where the sequence linguistic complexity is the same as in the input sequence and appears in bold.

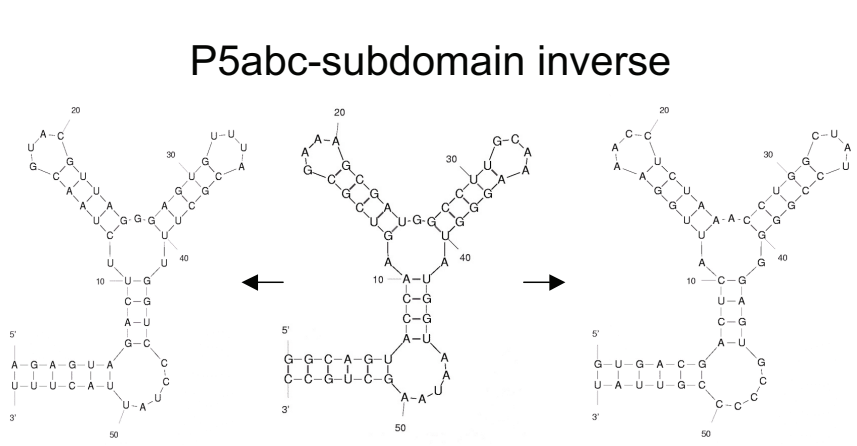


Figure 5: P5abc subdomain solution of RNAinverse. In the middle, the secondary structure drawing of the input sequence is given. On the sides, two secondary structure drawings of RNAinverse representative solutions are given. Values for the physical measures appear below the drawings. Values for the input reference sequence are in bold.

Therm. S. : -11.7
kcals/mole
Mut. R. : 0.7447
Seq. C. : 0.5267

Inverse 1

Therm. S. : -25.6
kcals/mole
Mut. R. : 0.9458
Seq. C. : 0.557

Wildtype

Therm. S. : -15
kcals/mole
Mut. R. : 0.6711
Seq. C. : 0.4921

Inverse 2

from the *thermus thermophilus* ribosome (26). In Figure 3, the wildtype structure of mir-146 is depicted in the middle, same result obtained by using either mfold (22) or RNAfold (29) for the prediction of the secondary structure or the reported experimental secondary structure (27). On the left and on the right, indicated by two outgoing arrows from the middle, the folded structure of two representative results obtained by RNAinverse (2) are given. The two results of RNAinverse deviate in their physical measures from the initial structure in the middle (labeled as 'wildtype'), as can be viewed by examining the numerical values below the secondary structure drawings. The numerical values for the physical measures of the wildtype initial structure are in bold. In Figure 4, the results of our reconstruction method from shape using physical parameters as constraints is given, for the wildtype initial structure of mir-146 that is available in the middle of Figure 3. Note that the loop sizes have slightly been modified with respect to the secondary structure drawings of Figure 3, at the expense of preserving the values of the physical measures: same thermodynamic stability with respect to Figure 3 in the middle can be found in Figure 4 to the left (the numerical value is indicated in bold), same mutational robustness with respect to Figure 3 in the middle can be found in Figure 4 in the middle (the numerical value is indicated in bold), and same sequence linguistic complexity

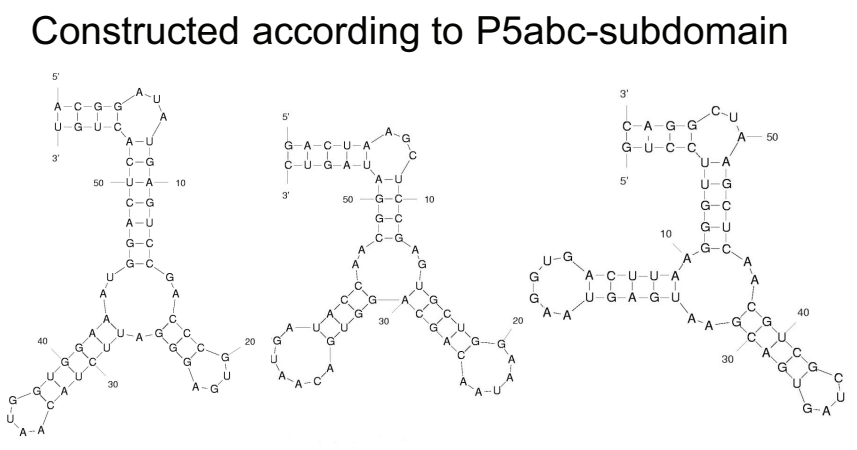


Figure 6: P5abc subdomain solution of our proposed method. To the left, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 5, where the thermodynamic stability is the same as in the input sequence and appears in bold. In the middle, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 5, where the mutational robustness is the same as in the input sequence and appears in bold. To the right, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 5, where the sequence linguistic complexity is the same as in the input reference sequence and appears in bold.

Therm. S. : -25.6
kcals/mole
Mut. R. : 0.9479
Seq. C. : 0.5793

Same thermodynamic stability

Therm. S. : -17.4
kcals/mole
Mut. R. : 0.9458
Seq. C. : 0.6158

Same mutational robustness

Therm. S. : -17.5
kcals/mole
Mut. R. : 0.9008
Seq. C. : 0.557

Same sequence complexity

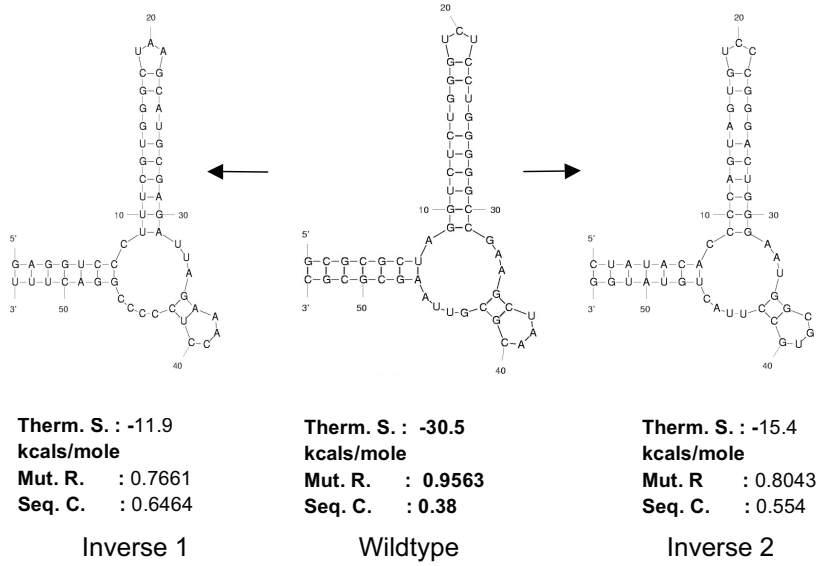


Figure 7: Ribosomal piece solution of RNAinverse. In the middle, the secondary structure drawing of the input sequence is given. On the sides, two secondary structure drawings of RNAinverse representative solutions are given. Values for the physical measures appear below the drawings. Values for the input sequence are in bold.

with respect to Figure 3 in the middle can be found in Figure 4 to the right (the numerical value is indicated in bold). The same illustration as in Figures 3 and 4 for mir-146 are given in Figures 5 and 6 for the P5abc subdomain of the *tetrahymena thermophila* group I intron ribozyme and in Figures 7 and 8 for the piece that was taken from the *thermus thermophilus* ribosome. Note that in all designed structures generated by RNAinverse or our approach, the free energy is not as low as in the wildtype structure, in good agreement with the fact that natural sequences tend to be more thermodynamically stable than other sequences. Finally, in Figure 9, an additional experiment was performed with the mir-146 that was presented in Figure 3. Instead of obtaining an RNA designed sequence with the same shape and an exact same physical measure value as the natural RNA, as in Figure 4, the objective is to

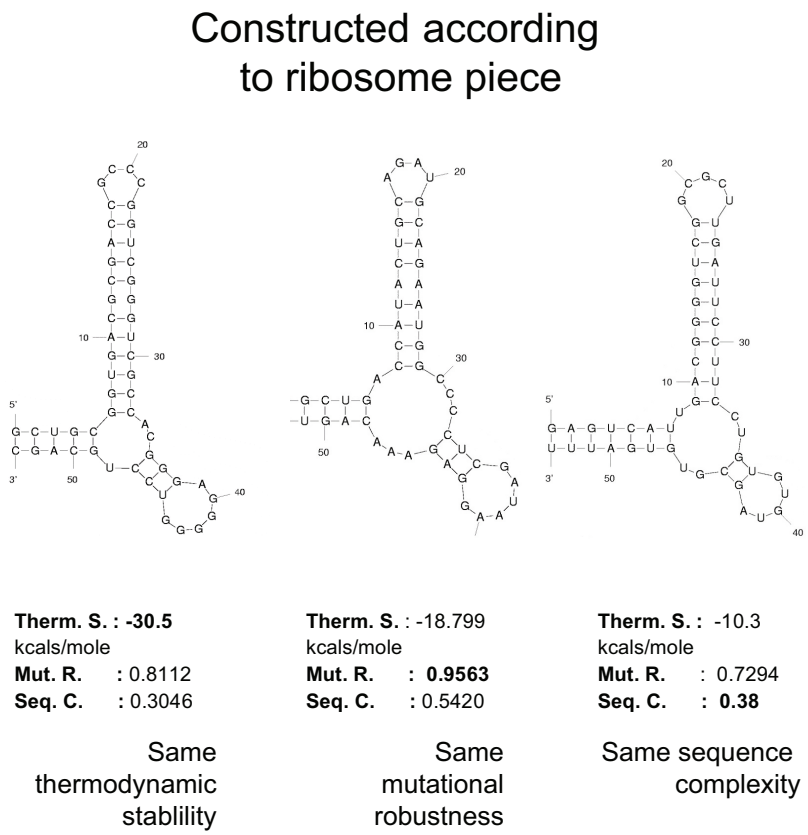
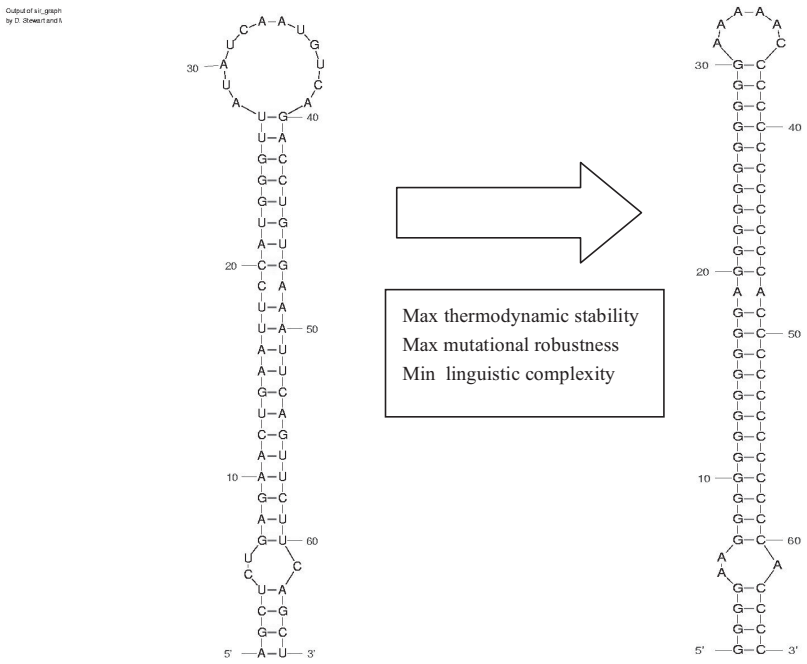


Figure 8: The ribosomal piece solution of our proposed method. To the left, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 7, where the thermodynamic stability is the same as in the input sequence and appears in bold. In the middle, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 7, where the mutational robustness is the same as in the input sequence and appears in bold. To the right, a solution of our reconstruction method appears for the input sequence corresponding to the middle of Figure 7, where the sequence linguistic complexity is the same as in the input sequence and appears in bold.

maintain the same shape as the natural sequence while maximizing thermodynamic stability, maximizing mutational robustness, and minimizing linguistic complexity. The resulting designed sequence is observed in Figure 9. This also shows that the concept of maintaining the exact shape while obeying certain physical constraints is quite general in scope, and can be applied to a variety of problems. Here, we only picked very few representative problems, and carefully selected them such that we start from an initial input that is well-predicted with energy minimization compared to the experimental result. These specific problems represent a wide class of other problems for which interesting solutions can be found. In some of the cases there may be no solutions, depending on the problem at hand. For example, if we impose strict constraints on extreme values of mutational robustness and linguistic complexity that are not usually found in random sequences and demand that the designed sequences deviate very slightly from these extreme values, obviously there can be no solution to the problem. On the other hand, if the numerical values of the constraints are in the range that is typically found for random sequences and we permit the designed sequences to deviate from the these values less strictly, there will be a variety of solutions found for the problem. As for efficiency comparison with RNAinverse, a fair comparison would demand an RNAinverse-like implementation for the extended problem. Here, for the extended inverse problem presented, we chose an approach that is easy to parallelize and convenient, namely parallel GAs, since the goal of the present work was not to devise an efficient algorithm. It is quite likely that at least for a single processor, an RNAinverse-like approach might be found more efficient, which can be dealt with in future work on this problem.



The structure of mir-146. Calculated measures are: thermodynamic stablity of -22.4 kcals/mole, mutational robustness of 0.952, and linguistic complexity of 0.639.

Sequence designed by our method with the same "shape" as mir-146 and max possible thermodynamic-stability (-72.6 kcals/mole), max possibly mutational robustness (0.971), min possible linquistic complexity (0.006).

Figure 9: Starting from the sequence of mir-146 (and its secondary structure prediction) on the left, a solution of the suggested reconstruction method appears on the right that is characterized by maximum thermodynamic stability (a decrease from -22.4 to -72.6 kcals/mole) and maximum mutational robustness (an increase from 0.952 to 0.971) and minimum linguistic complexity (a decrease from 0.639 to 0.006).

Computational Expenses and Cost Reduction

The three measures are not equal with respect to their computational expense. It takes $O(n^2)$, where n is the length of the sequence, to compute the linguistic complexity of the sequence. On the other hand, it takes $O(n^3)$ to compute the thermodynamic stability of the sequence, and $O(n^4)$ to compute the mutational robustness of an RNA struc-

ture since it entails the calculation of 3n single point mutations and each takes $O(n^3)$ to calculate the minimum free energy structure. It should be noted that the two stage approach (Figure 2) is considerably more efficient than a single stage GA, because on a single processor we noticed that in most cases of a convergence to a desired solution, the method first finds a sequence with the desired coarse-grain structure (or shape) and only then improves the structure towards a sequence with a desired physical measure. Thus, if we break the process into two stages and parallelize the first stage (as illustrated in Figure 2), we achieve a significant reduction in computation.

Physical Properties in an Evolutionary Process

An underlying assumption in this work is that the evolutionary process that natural RNAs undergo brings about certain physical properties. An interesting experiment would be to check whether a simple evolutionary process, either being carried out by a computer simulation or by observing how natural sequences evolve, can serve as further guidance into which physical properties are being generated and preserved in various RNA classes of interest (*e.g.*, miRNAs, RNA viruses) without imposing explicit selection of these properties.

Implications for RNA Design and Future Work

To the best of our knowledge, this study brings for the first time the awareness of physical quantifiable measures to the RNA inverse folding problem. The implications of this approach to RNA design are in potentially improving previous algorithms that so far only took into consideration the sequence and structure of RNAs, which is indicative of their topology. By bringing physical measures to the design of novel RNA sequences, one can take advantage of previous studies that highlighted the differences between random and natural RNAs in terms of the physical measures. Here, we demonstrated the approach by using thermodynamic stability, mutational robustness, and linguistic complexity. In future work, other physical measures of interest or a combination of several ones can be incorporated.

Acknowledgements

We thank Edward N. Trifonov and the Genome Diversity Center at the Institute of Evolution, University of Haifa. We also thank Dror Berman and Eyal Schwartz for their help in the initial stage. The research was supported by the Lynn and William Frankel Center for Computer Sciences at Ben-Gurion University and a grant from the Israel USA binational Foundation BSF 2003291.

References and Footnotes

1. Higgs, P. G. *Quarterly Review of Biophysics 33*, 199-253 (2000).
2. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., Schuster, P. *Monatsh Chem 125*, 167-188 (1994).
3. Aguirre-Hernandez, R., Hoos, H. H., Condon, A. *BMC Bioinformatics 8*, 34 (2007).
4. Busch, A., Backofen, R. *Nucleic Acids Res 35*, W310-W313 (2007).
5. Le, S. Y., Chen, J. H., Konings, D., Maizel, J. V. *Bioinformatics 19*, 354-361 (2003).
6. Higgs, P. G. *J Phys I (France) 3*, 43-59 (1993).
7. Borenstein, E., Ruppin, E. *Proc Natl Acad Sci USA 103*, 6593-6598 (2006).
8. Trifonov, E. N. Making sense of the human genome. In *Structure and Methods 1*, pp. 69-77. Eds., Sarma, R., Sarma, M. Adenine Press, New York (1990).
9. Popov, O., Segal, D. M., Trifonov, E. N. *BioSystems 38*, 65-74 (1996).
10. Shapiro, B. A. *Comput Appl Biosci 4*, 387-393 (1988).
11. Le, S. Y., Nussinov, R., Maizel, J. V. *Comput Biomed Res 22*, 461-473 (1989).
12. Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., Giegerich, R. *Bioinformatics 22*, 500-503 (2006).
13. Schuster, P., Fontana, W., Stadler, P. F., Hofacker, I. L. *Proc R Soc Lond Ser B Biol Sci 255*, 279-284 (1994).
14. Fontana, W., Schuster, P. *Science 280*, 1451-1455 (1998).
15. Stadler, B. M. R., Stadler, P. F. The topology of evolutionary biology. In *Modeling in Molecular Biology*, pp. 267-286. Eds., G. Ciobanu, G. Rozenberg. Natural Computing Series, Springer Verlag (2004).

16. Cowperthwaite, M. C., Bull, J. J., Meyers, L. A. *Genetics 170*, 1449-1457 (2005).
17. Wagner, A., Stadler, P. F. *J Exper Zool (Molec Devel Evol) 285*, 119-127 (1999).
18. Meyers, L. A., Lee, J. F., Cowperthwaite, M. C., Ellington, A. D. *J Mol Evol 58*, 618-625 (2004).
19. Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley (1989).
20. Shapiro, B. A., Wu, J. C., Bengali, D., Potts, M. J. *Bioinformatics 17*, 137-148 (2001).
21. Fogel, G. B., Porto, V. W., Weekes, D. G., Fogel, D. B., Griffey, R. H., McNeil, J. A., Lesnik, E., Ecker, D. J., Sampath, R. *Nucleic Acids Res 30*, 5310-5317 (2002).
22. Zuker, M. *Nucleic Acids Res 31*, 3406-3415 (2003).
23. Mathews, D. H., Sabina, J., Zuker, M., Turner, D. *J Mol Biol 288*, 911-940 (1999).
24. Zuker, M. *Science 244*, 48-52 (1989).
25. Griffiths-Jones, S. *Nucleic Acids Res 32*, D109-D111 (2004).
26. Yusupov, N. M., Yusupova, G. Z., Baucom, A., Lieberman, K., Earnest, T. N., Cate, J. H., Noller, H. F. *Science 292*, 883-896 (2001).
27. Krol, J., Sobczak, K., Wilczynska, U., Drath, M., Jasinska, A., Kaczynska, D., Krzyzosiak, W. J. *J Biol Chem 279*, 42230-42239 (2004).
28. Wu, M., Tinoco, I., Jr. *Proc Natl Acad Sci USA 95*, 11555-11560 (1998).
29. Hofacker, I. L. *Nucleic Acids Res 31*, 3429-3431 (2003).

Date Received: January 7, 2008

Communicated by the Editor Ramaswamy H. Sarma