In Silico Biology 4 (2004) 593–604
IOS Press

STR^2: A Structure to String Approach for Locating G-Box Riboswitch Shapes in Pre-Selected Genes

Oriel Bergig[a], Danny Barash[a], Evgeny Nudler[b] and Klara Kedem[a,*]
[a]*Department of Computer Science, Ben-Gurion University, Beer Sheva 84105, Israel*
E-mail: {bergig, dbarash, klara}@cs.bgu.ac.il
[b]*Department of Biochemistry, New York University Medical Center, NY 10016, USA*
E-mail: evgeny.nudler@med.nyu.edu

Edited by E. Wingender; received 5 August 2004; revised 29 October 2004; accepted 31 October 2004; published 16 November 2004

ABSTRACT: Traditional sequence-based search methods such as BLAST and FASTA can be used to identify sequence similarities. Recently, there is a growing interest in performing RNA shape similarity searches inside selected genes to locate RNA structure motifs that are known to possess functionally important roles. For example, in the newly discovered RNA genetic control elements called "riboswitches", the box domain is known to be highly conserved among various bacterial species in both its nucleotide composition and shape. However, in non-bacterial species, shape conservation is likely to become more important than sequence conservation when searching for riboswitch patterns. For this purpose, we present an approach tailored for detecting RNA shape similarities. We extend the Structure to String (STR^2) method that was initially proposed to locate shape similarities in proteins to identify predicted secondary structures of RNAs. The STR^2 for RNAs is a translation of a secondary structure to a string of characters, after which known sequence-based search algorithms with an efficient implementation are being used. We validate that the STR^2 succeeds to locate G-box riboswitches in prokaryotes, as expected. Subsequently we show running examples when attempting to detect G-box riboswitch candidates in eukaryotes.

Availability: The binaries and source code are available upon request.

KEYWORDS: STR^2, string inexact matching, RNA folding prediction, dynamic programming, suffix tree, RNA shapes, riboswitches

INTRODUCTION

The collection of complete genomes from a variety of model organisms has prompted new challenges in searching and analyzing the gathered data. Traditional sequence-based search methods such as BLAST [Altschul *et al.*, 1990] and FASTA [Pearson and Lipman, 1988] can be used to scan entire genomes for sequence similarities. However, as more detailed knowledge is accumulated on the relationship between structure and function, there is an increasing demand for methods that take structure similarities into account. These methods can either work in conjunction or as alternatives to the traditional sequence

*Corresponding author.

Electronic publication can be found in *In Silico Biol.* **4**, 0049 <http://www.bioinfo.de/isb/2004/04/0049/>, 16 November 2004.

based ones. Taken together, they can lead to important discoveries, by locating functionally meaningful structural elements.

Recently, unique RNA genetic control elements were discovered in bacteria that regulate gene expression without the participation of proteins [Winkler *et al.*, 2002; Mironov *et al.*, 2002]. These RNA genetic control elements called "riboswitches" bind small molecules with high affinity and as a consequence they respond with conformational switching [Winkler and Breaker, 2003; Nudler and Mironov, 2004; Vitreschak *et al.*, 2004]. Their secondary structure is indicative of their function. The various natural riboswitches that were found to control vitamin, amino acid, and purine metabolism in bacteria consist of two neighboring domains. The domain that binds small molecules and is highly conserved among many species of bacteria, called aptamer or "box", and the expression platform domain that undergoes conformational change. The consensus secondary structure of the box domain has been constructed in several works [Grundy and Henkin, 1998; Miranda-Rios *et al.*, 2001; Ro dal *et al.*, 2003]. It possesses a unique shape and thus can be used as a fingerprint to search for riboswitch pattern. Because riboswitches are believed to be the derivatives of an ancient genetic control system, it is logical to assume that they have undergone compensatory neutral mutations in evolution as simulated for RNAs in several models [Higgs, 1998]. Thus, on an evolutionary timescale as was shown for tRNAs, ancient riboswitches that are more widely distributed across the phylogenetic landscape relative to newer ones should possess a considerable amount of structure similarity whereas their base sequence is likely to be much less conserved.

We introduce the Structure to String (STR^2) method that was initially proposed for searching similarities in the tertiary structure of proteins (see web server at http://www.cs.bgu.ac.il/catalina/STR2). Unlike search programs that mostly rely on sequence similarity such as the SequenceSniffer used in [Sudarsan *et al.*, 2003; Barrick *et al.*, 2004], or programs that incorporate information about numbers/lengths of stems/loops such as the RNA-Pattern used in [Rodionov *et al.*, 2002] and more sophisticated packages that use motif descriptors such as the RNAMotif [Macke *et al.*, 2001], Structure to String (STR^2) is conceptually unique being a purely structural based approach. It relies on the geometry of the drawn secondary structure and further represents the RNA shape by letters, also taking into account the transitions between stems and loops as a unique feature. We emphasize that the STR^2 can be used in conjunction with the aforementioned methods in searching for with the added contribution that RNA shapes are of particular importance in ancient riboswitches that have apparently accumulated numerous compensatory neutral mutations and are therefore difficult to detect by relying mostly on sequence conservation. Therefore, the STR^2 package can be used on top of recently introduced web servers that attempt to identify riboswitch motifs using sequence considerations [Bengert and Dandekar, 2004], or as a post-processing step after running the RNAMotif [Macke *et al.*, 2001] or RNAProfile [Pavesi *et al.*, 2004], FastR [Bafna and Zhang, 2004], HomoStRscan [Le *et al.*, 2004] on large data sets.

We apply the STR^2 to search for G-box shapes in selected genes that are participating in purine metabolism. First, we validate our method in prokaryotes, locating known box shapes [Mandal *et al.*, 2003] with notably few false positives. Second, we illustrate the initiation of a search for G-box shapes in eukaryotic genes, with the aim of collecting potentially new riboswitch candidates that have not been discovered yet in higher organisms.

Methods

The search method we propose is the Structure to String (STR^2). Given a *query* sequence to search among a set of target sequences, STR^2 will find predicted sub-structures of the *query* similar to predicted

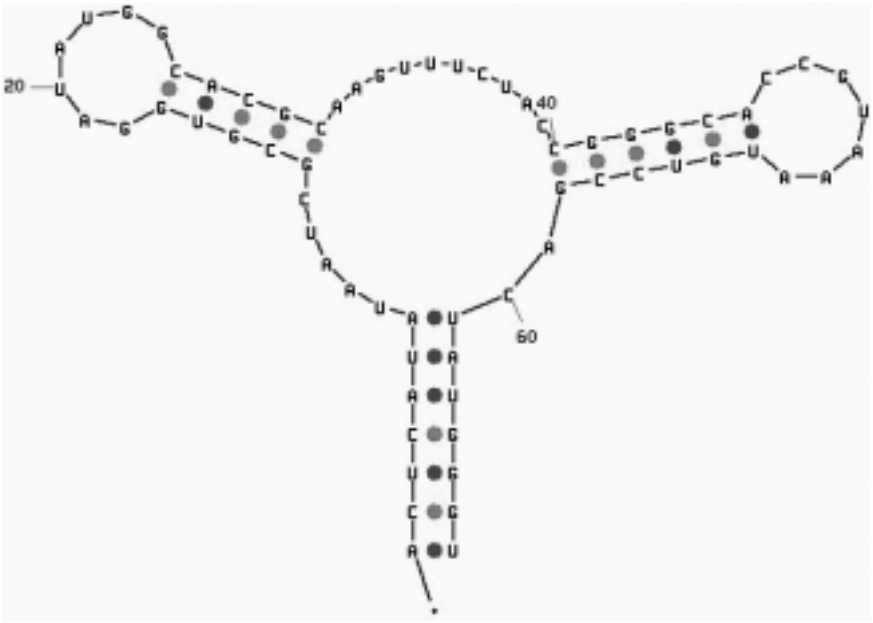

Fig. 1. G-box structure query. The predicted secondary structure of G-box in *Bacillus subtilis-xpt* [Mandal *et al.*, 2003] using Mfold [Zuker, 2003]. Other predicted G-box secondary structures have the same structural shape as the one above.

sub-structures of the target. To illustrate the approach, the *query* we use is the G-box depicted in Fig. 1, which is 68 nt long in size. Target sequences were extracted from selected genes and were then used to predict a set of secondary structures, namely the target structures, as follows: first, by cropping from the sequence segments of 68 nt, stepping 4 nt between overlapping windows, then by predicting the secondary structure in each window.

Translation of a structure to a string of letters

Instead of comparing the query structure to target structures, the STR^2 transforms the problem of structure similarity to inexact string matching. It then applies fast string algorithms to solve the latter. The translation is performed using a fragment library, which consists of a small number of short structure fragments, each associated with a unique letter. In this work we found that 5 letters, consisting of 3 nucleotides each, represent well the variety of short fragments in the G-box (see Fig. 2). To translate a secondary structure to a string of letters, we decompose the secondary structure to overlapping small structure fragments and normalize the distances between fragments. Then, we translate each fragment to a letter by superimposing the fragment on all fragments in the library and picking the one with the smallest minimum RMS distance. The minimum RMS distance between two fragments is calculated in a standard way (see for example [Kabsch, 1978], minimizing the root-mean-square). We represent this secondary structure fragment with the letter in the library associated with the nearest (smallest minimum RMS distance) library fragment. Repeating this process consecutively on the secondary structure from beginning to end converts it into a shape-representing character string. We note that by definition, a local representation which uses consecutive fragments involves a loss of information. Thus, when comparing two translated mRNA structures, a local inconsistency between two fragments may cause a global dissimilarity although the resulting character string will be affected only locally. Therefore, we

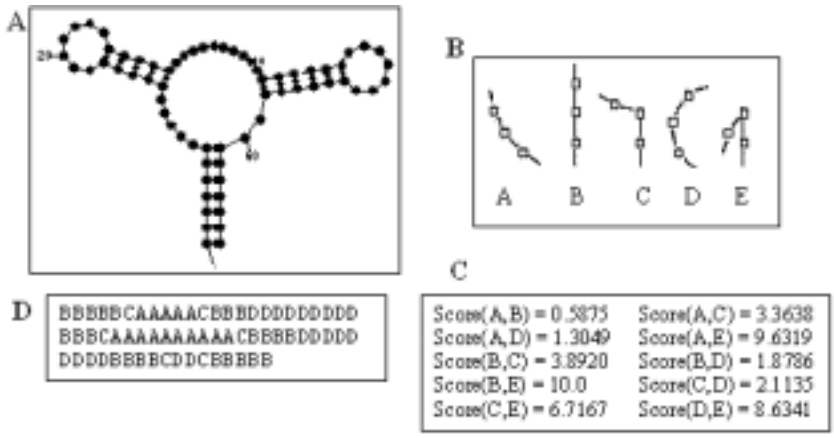

Fig. 2. Shape representing string of letters. The shape representing string of the G-box secondary structure. (A) The predicted secondary structure of the G-box. (B) The fragment library with 5 characters, each representing 3 consecutive nucleotides. (C) The similarity score between the various characters within the fragment library. (D) G-box structure translated to a string of letters based on the fragment library, starting from the 5′-end.

choose to have an overlap between fragments, avoiding the limitation of this approach in many of the cases. An additional loss of information occurs because we represent a fragment structure as one of only five library fragments. However, because we use minimum RMS distance to decide which fragment to choose rather than an arbitrary division of the space of the possible fragments shapes [see Park and Levitt, 1995] for a further discussion on possible representations) and since the variety of fragments possible in a folded mRNA structure is small, the approach is feasible. With queries similar in motif complexity to the G-box, the representation scheme suggested is robust.

We define a distance score between the characters. The score is decided according to the minimum RMS distance between the fragments (see Fig. 2) associated with the characters, normalized to the range from 0 to 10. A score of ‘0’ is assigned in the case of same characters and ‘10’ in the case of the most distant two characters (or fragments). Introducing scores between characters allows calculating a similarity between two character strings, defined as the sum of the scores between any two aligned characters. A similarity between two shape-representing strings corresponds to a similarity between their correlated structures.

For search of the G-box we constructed a fragment library that consists of five different characters, each representing three consecutive nucleotides. The concept is analogous to the case of protein tertiary structure, in which a fragment library can be constructed [Kolodny *et al.*, 2002] for various applications, but here it is implemented on RNA secondary structure. Figure 2(B) illustrates the selected fragments and their assigned characters. The letter characteristics are as follows: the character ‘A’ represents three consecutive nucleotides within a large loop as in the multibranch loop of Fig. 2(A), ‘B’ represents a stem segment, ‘C’ represents a turn as in Fig. 2(A), and ‘D’ represents three consecutive nucleotides within a short loop as in the hairpins of Fig. 2(A). It also represents the bottom right-most part of the multibranch loop (around location 60) since there are less nucleotides in this part compared to the bottom left-most part. Therefore, its turn is sharper resembling the situation in the hairpins at the top, after the distances between nucleotides are normalized. Character ‘E’ represents three consecutive nucleotides in a sharp turn, it does not exist in the G-Box, but we noticed it exists in many of the secondary structures predicted by MFOLD. ‘E’ is a character that is most distant from all the others, see Fig. 2(C) where the scores between pairs of characters based on their minimum RMS distance are presented. Combinations with the letter ‘E’ receive the highest scores compared to all the others.

Structure searching with string algorithms

A search for the query means searching the character string representing the query within those representing the targets. Thus, it is an *inexact string matching* problem. Dynamic programming is a common way to approach this type of problems, and in particular the local alignment problem [Gusfield, 1997] that we are trying to solve. However, its complexity on two strings of size m and size n is $O(nm)$. When searching segments in genes that are considerably long, this procedure may become impractical. Faster results are achieved with exclusion methods [Gusfield, 1997], as was also implemented in BLAST [Altschul *et al.*, 1990]. In such a case, the following assumption should hold: two character strings are similar if they consist of a shorter common portion. This assumption is correct in the case of mRNA secondary structures, in particular with conserved structures like the G-box and with a small fragment library. Therefore, to gain efficiency we implement the STR^2 method by encoding try and target strings in a fast search data structure, a generalized suffix tree [Gusfield, 1997]. We perform an *exact string matching* of the query in the target by applying an algorithm that finds all exact matches in the query that exist in the target and are longer than a "minimal length" parameter. This parameter receives a value between 5 and 15 (see Supplementary Section for more details concerning parameter sensitivity). The algorithm to find all exact matches performs only one pass on the query, therefore its online runtime is linear in the size of the query and its offline runtime is linear in the size of the target. The method benefits from the minimal length parameter introduced here since it allows a compromise between speed and accuracy.

STR^2 expands the exact matches to longer inexact matches in a local alignment manner. Local alignment allows extracting two regions that exhibit high similarity from two character strings using a dynamic programming (DP) algorithm. Because we acquire the exact matches from the previous stage, we extend these matches to the left and to the right in an iterative manner. We calculate the score of the growing inexact matches in each iteration as was done in the local alignment DP table. However, there is no need to fill all the cells of the DP table, only a part of one of its diagonals for each exact match. We stop extending the matching region when its score reaches zero, then we shorten the region to the length where its score was a local minimum. The score between letters in a local alignment algorithm are being shifted with different values to set a compromise between the similarity and length of a matching region. Here, because we shift the scores introduced previously with values between 1 and 10, we call this parameter the "shift parameter". The results are ordered starting from the most similar structure (receiving best score) to the lowest. We choose to stop after a predefined number of acquired results because we do not calculate the statistical significance of the scores. Implementing statistical significance is an extension suggested for future work. The search method presented here is novel in RNAs for several reasons. First, it is a purely structure based method, measuring shape similarity as one would measure sequence similarity. Second, by using the fragment library we modulate not only the stem and the loop, but also the loop sizes and transitions between stems and loops. Loop sizes are also influential in representing the transitions, since they affect the geometric curvature associated with the transition. Representing the transitions between secondary structure elements in an adequate manner is important for assessing shape similarities.

RESULTS AND DISCUSSION

We use the G-box (guanine binding) riboswitch domain [Mandal *et al.*, 2003] as the query sequence for illustrating our proposed method. This conserved box resides in the $5'$-UTR of bacterial genes that are

largely involved in purine metabolism. Its secondary structure is well predicted by Mfold [Zuker, 2003] when applied on the bacterial sequences listed in [Mandal *et al.,* 2003], conforming to the consensus model found in that reference. Thus, we can compare the G-box shape with secondary structures from target gene sequences by the folding prediction of these sequences.

The G-Box Secondary Structure Query

The secondary structure of the G-box domain (68 nt) is composed of a three stem junction with a multibranch loop connecting two hairpins and the 5′-3′ end. In [Mandal *et al.*, 2003; Ji *et al.*, 2004], the G-box consensus and the conserved region common to the list of bacterial species are given. Using Mfold, we predict the secondary structure of the G-box in the aforementioned bacterial sequences. Although some alternations exist among the conserved sequences of the G-box domain, all of these sequences result in a similar predicted shape, which is the one we use for the search query (Fig. 1).

Validation of STR^2 on Known G-Box Instances in Prokaryotes

In Mandal *et al.*, 2003, the existence of the G-box was reported in the 5′-UTR of several bacterial mRNAs. In order to test the STR^2, we performed two example experiments reported below. In the first experiment we extracted a sequence of 15000 nt from the complete genome of the Bacillus halodurans from location 645000 to location 660000 to test the method and its signal to noise level on a large amount of data. Two genes encoding to mRNAs with a G-box domain are present in locations 648460 and 650328. We folded the extracted sequence in windows of 68 nt each and with a step of 4 nt between each window. The first fold is from location 645000 to 645068, the second is from location 645004 to 645072, and so on. Stepping only four nucleotides between prediction windows might cause the same predicted shape to appear twice or more, but assures we do not miss a region that could match the query. In each fold we included the optimal as well as suboptimal solutions since we assume that some suboptimal predictions may correspond to correct structures. This increases the pool of candidates and may also introduce some incorrect structures. Since we considered the suboptimal solutions as well, a sequence of 15000 nt yielded a total of 7542 optimal and suboptimal folds. This target set was processed using the STR^2 to locate the G-box.

In the second experiment we performed a search for two known G-box instances in two selected genes from *Bacillus subtilis* since we designate the method for searching in selected genes. The first gene is the *pbuG* (locus tag BSU06370) and the second is the *purE* (locus tag BSU06420). Since riboswitches may occur around 200–300 nt [Nudler and Mironov, 2004] upstream to the gene, we extracted a total of 500 nt upstream to the gene. In order to verify that no false positives occur, we extracted a region up to the end of the gene. Therefore, we extracted the regions 693500–695333 and 697500–698442 for the *pbuG* and *purE*, respectively. To perform this search and all others presented in this paper we used a value of 10 for the "minimal length" parameter and a value of 5 for the "shift" parameter after experimenting with several possibilities, some are described in the Supplementary section. The known locations presented in Mandal *et al.*, 2003, were all found as depicted in Fig. 3.

Figures 3(A) and 3(B) are the predicted secondary structures of the *Bacillus halodurans* at positions 648460 and 650328. Figures 3(C) and 3(D) are the predicted secondary structures of the *Bacillus subtilis* at positions 693795 and 697731. Predicted structures were found using the STR2 where the query was the G-box sequence found upstream to the gene of *Bacillus subtilis-xpt*. The target for the structures depicted in 3A and 3B was a segment of 15000 nt extracted from the Bacillus halodurans genome. The target for structure depicted in 3C was the pbuG gene of the Bacillus subtilis starting 500 nt upstream to

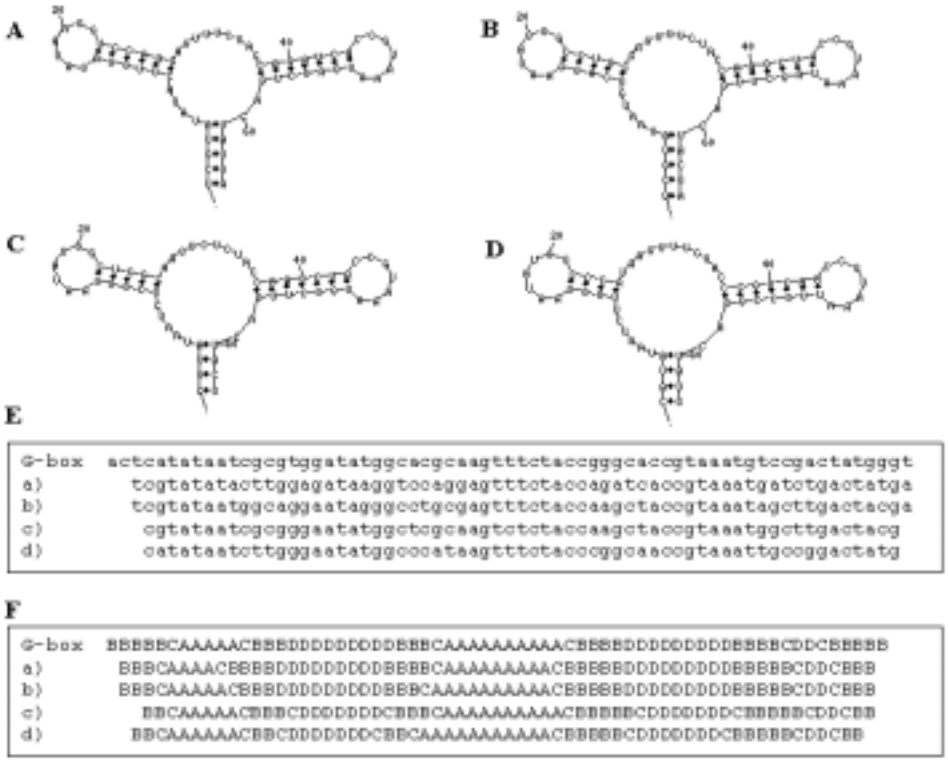

Fig. 3. G-box structure targets. The folded secondary structures of four G-box domains: two from *Bacillus halodurans* and two from *Bacillus subtilis*. (A) First located G-box starting at position 648460. (B) Second located G-box starting at position 650328. (C) Located G-box upstream to *pbuG*. (D) Located G-box upstream to *purE*. (E) The G-box query sequence, aligned to the sequences corresponding to the located G-box. (F) The G-box query STR^2 character string, aligned to those corresponding to the located G-box character strings.

the gene, and the target for the structure depicted in Fig. 3D was the *purE* gene of the *Bacillus subtilis* starting 500 nt upstream to the gene. The fragment library was constructed as described previously.

The STR^2 was able to find all four G-box shapes without the appearance of false positives as will be analyzed when examining the parameter sets. This is because their shape representing character string is similar to that of the query, as depicted in Fig. 3(F). In Fig. 3(E), the sequences corresponding to the G-box shapes found are presented along with the query sequence. A similarity in sequences exists although an even more pronounced similarity appears between the STR^2 shape representing character strings (Fig. 3(F)) and the query.

We checked whether a standard geometrical similarity measure such as the minimum RMS distance would have found those two predicted structures. The RMS distance was calculated between the two located structures and the G-box query structure, resulting in large distances. Obviously, for RNA secondary structure similarity, the minimum RMS distance fails to identify two similar shapes. This can be easily explained by suggesting the following example. Suppose we examine a loop with three stems (as in the G-box) compared to another similar shape, but with a difference in the angle between a stem and a loop because of a variation in the size of the loop. This difference will be responsible for a high minimum RMS value although the shapes are quite similar. In contrast, our proposed geometrical similarity measure is only affected from those differences in a local manner.

Table 1
Search results

Gene	Accession number	Extracted sequence	Similar G-Box structure start positions
ADE12	NC_001146	234001 to 236101	234048, 235433, 234219, 234209
ADE1	NC_001133	169021 to 170521	169985, 169352, 169733
ADE13	NC_001144	844621 to 845083	845015

STR^2 results when searching for G-Box like structures in three different genes participating in purine biosynthesis. The last column contains the positions in the extracted sequence where a G-Box like structure was found using STR^2.

STR^2 Search for G-Box Like Secondary Structure in Eukaryotes

Understanding the principles behind the STR^2 method and their careful implementation ensures that if the query structure has a similar structure in the target, the structure will be found. Encouraged by the validation in prokaryotes, we extended our search to several eukaryotic genes. As an example for illustrating the method, we concentrate on the ADE12 gene in *Saccharomyces cerevisiae* that is participating in purine biosynthesis in yeast (see Fig. 4). Trial results in two other participating genes are reported in Table 1. Notice that the G-box query length and the length of the regions found to be similar to the G-box vary. This was possible because in some cases we used results of STR2 as a seed for extending the folding prediction and because we are searching for a similar region in a local alignment manner as described previously. Using local alignment for our search possesses the disadvantages that were mentioned in the Methods section, but has the advantage that it allows us to find partial matches.

A Secondary Structure Similar to G-Box is Found by STR^2 in the Saccharomyces Cerevisiae Chromosome XIV

One of the genes that participate in the biosynthesis of purine in Saccharomyces cerevisiae is ADE12 [Marks *et al.*, 2003]. This gene is located between positions 234412 and 235713 in the sequence (locus tag YNL220W, accession number NC_001146 in GenBank). We extract the sequence of the *Saccharomyces cerevisiae* starting from location 234001 to location 236101. We fold the sequence in windows of 68 nt with a step of 4 nt in each fold, resulting with 1168 predicted secondary structures. In those target structures, we search for the query structure of the G-box. Our findings are presented in Fig. 4. All located structures have a noticeable structure similarity with the G-box structure extracted from *Bacillus subtilis-xpt* and used as the query for this search. Notice that the four predicted similar structures do not possess any sequence similarity, as can be observed in Fig. 4(E). However, some character string similarity exists among the translated secondary structures to strings, as presented in Fig. 4(F). These are the strings that have enough inexact similarity for the STR^2 to designate them among the total of all 1168 predicted secondary structures.

CONCLUSIONS

Traditionally, the search for functionally important elements in various genes has been performed using sequence similarity methods. We introduce a purely structural-based approach that relies on a representation of an RNA secondary structure as a string of letters, consequently applying string algorithms. We apply the Structure to String (STR^2) method to search for the G-box (guanine binding) domain [Mandal *et al.*, 2003] of purine riboswitch in eukaryotic genes. This particular task on the purine

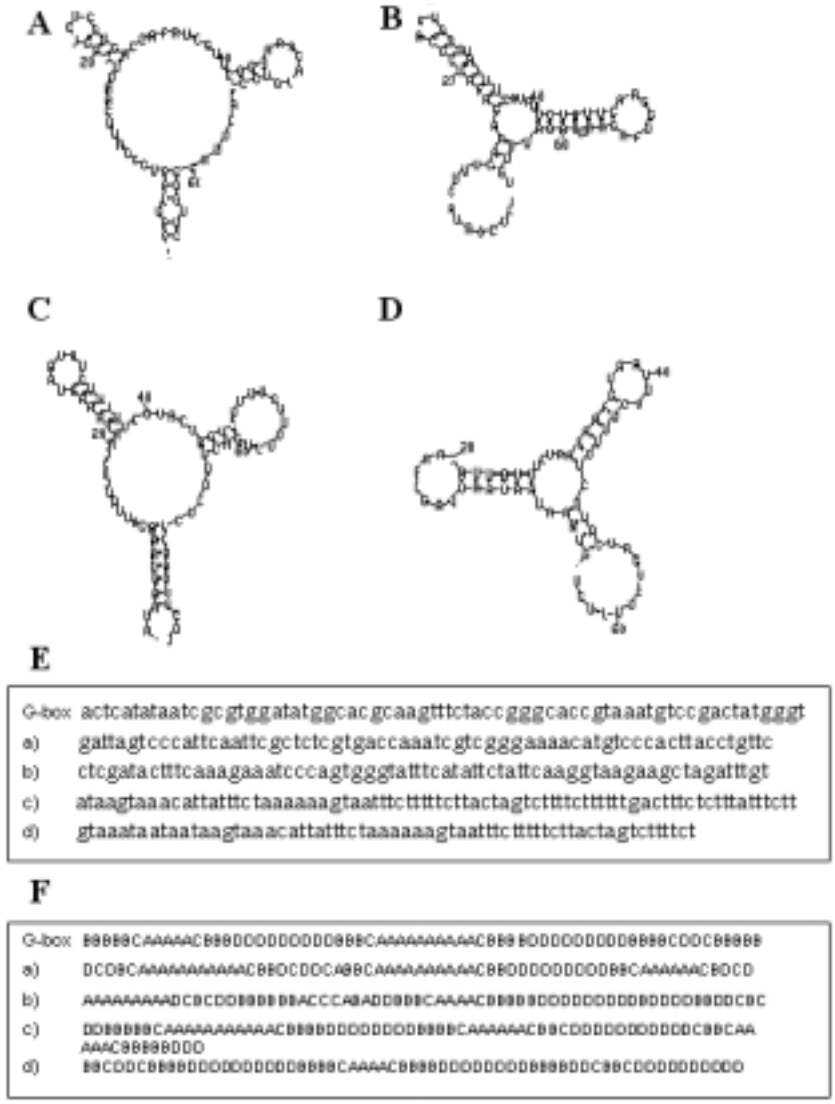

Fig. 4. Search Results. STR^2 results on the *Saccharomyces cerevisiae* chromosome XIV. (A) Predicted secondary structure from position 234048 to 234114. (B) Predicted secondary structure from position 235433 to 235501. (C) Predicted secondary structure from position 234219 to 234298. (D) Predicted secondary structure from position 234209 to 234274. (E) The sequence of G-box query extracted from *Bacillus subtilis-xpt* and the sequences of all matching positions found by STR^2. (F) The translated secondary structure of G-box query to a string, extracted from *Bacillus subtilis-xpt* and from all matching positions found by STR^2.

riboswitch with its G-box has not yet been successful to identify any potential riboswitch sequence candidate when using methods that mostly rely on sequence based searches (unpublished results), albeit the success of these searches to locate seven thiamine pyrophosphate riboswitch candidates in eukaryotes [Sudarsan *et al.*, 2003] which is considered a few. Thus, there is an obvious motivation to implement structure-based searches.

We note several limitations when applying the (STR^2) method for identifying new riboswitch candidates. First, because the string letter representation relies on a unique two-dimensional coordinate system

and various RNA secondary structures drawing schemes have been developed over the years [Shapiro *et al.*, 1982], all our secondary structure predictions should be performed with the same folding prediction graphical representation. Second, because of the limitations of ab-initio folding algorithms by energy minimization to predict the correct shape of several aptamers or box domains, the fully computational application of the (STR^2) method is currently limited to those box domains that are initially well predicted by available folding algorithms. Third, the STR^2 as an idea can be incorporated to other more sophisticated search methods that are used to detect structural motifs. But by itself, it should mostly be used to identify shape similarities with high resolution in selected or pre-scanned genes, since the signal to noise level will most likely make the STR2 useless when scanning entire genomes. The initial scan of an entire genome can be performed by packages such as RNAMotif [Macke *et al.*, 2001], RSEARCH [Klein and Eddy, 2003], or the recently introduced RNAProfile [Pavesi *et al.*, 2004] as well as sequence based methods such as BLAST [Altschul *et al.*, 1990], FASTA [Pearson and Lipman, 1988], or more specific ones [Bengert *et al.*, 2004]. For the post-processing, STR^2 can be used in the generation of candidates for screening before a laboratory experiment. The RNAMotif package is more sophisticated and was designed to search for motifs in large datasets using descriptors, whereas the STR^2 can simply be used to identify shapes in pre-selected genes and is conceptually tailored to concentrate on the transitions between the various motifs in the query and target, rather than the motif descriptors themselves. Finally, the Structure to String (STR^2) method is not limited to the purine riboswitch with its corresponding G-box, nor to domains of riboswitch structures in particular. The concept of representing the geometry of the secondary structure with a string of letters is broad in its scope. It can be used to search for shape similarities of various constructs, such as artificial aptamers, or other functional elements in the secondary structure of RNAs.

SUPPLEMENTARY

Mfold

For the folding predictions in this paper, we use Mfold version 3.1 web interface [Zuker *et al.*, 2003] found at: http://www.bioinfo.rpi.edu/applications/mfold/old/rna/. Suboptimal foldings [Zuker, 1989] are also taken into account. For reproducibility, we provide the list of default parameters that were used at the time of this research: RNA sequence- linear; folding temperature: 37°C; ionic conditions: 1M NaCl, no divalent ions; percent of suboptimality: 5%; maximum interior/bulge loop size: 30; maximum asymmetry of an interior/bulge loop: 30; distance between paired bases: no limit.

STR2 Parameter Sensitivity and Runtime

We present the validation results described previously using three sets of parameters that are substantially different from one another. We show that the selection of the parameter set does not alter the success of the method and we conclude that the STR^2 is not sensitive to its parameter set.

The two parameters for the STR^2 are the exact match minimal length and the inexact match shift. The exact match minimal length is the minimal number of exactly matching characters in a typical match. A value of '0' means eliminating the assumption that in two similar strings there is an exact match of any length. Therefore, in such a case the STR^2 will not perform a heuristic algorithm and its runtime will be similar to a badly implemented dynamic programming algorithm. Exact match minimal length with a value equal to the size of the target means that no inexact matches are allowed and the scoring

introduced previously are not needed. Thus, the offline runtime will be linear in the size of the target and the online (if the target is already encoded in a suffix tree) runtime is only linear in the size of the query. We use three different numbers: 5, 10 and 15 for the exact match minimal length parameter. The inexact match shift is used to calibrate between how long in size versus how similar the matches will occur. The lowest value of '1' means that a short but similar match will be preferred over a long but non-similar one, while a value of '10' is the opposite. Values above 10 would usually not make a difference since the matches are considered in an increasing order from the lowest score to the highest. We demonstrate three different selections for the inexact match shift: 1, 5 and 10. We repeat the search in bacteria as presented in the previous section with the following three sets of parameters:

(1) *Set A:* exact match minimal length 5 and inexact match shift 10, the structure in position 650324 is designated first as a result of the search (it has a similar fold to the one in 650328 that is designated third). In the second place comes the structure in position 648460. Thus, from a set of 7542 possibilities, we successfully found the matches without false positives.
(2) *Set B:* exact match minimal length 10 and inexact match shift 5, the structure in position 650328 is designated first (it has a similar fold to the one in position 650324 that is designated second). In the third place comes the structure in position 648460. Thus, with this set of parameters we again successfully found the matches without false positives.
(3) *Set C:* exact match minimal length 15 and inexact match shift 1, the structure in position 650328 is designated first. In the eighth place, the structure in position 648460 is designated. Thus, also taking into account the placement of the structure in position 650324, we experienced five false positive results.

The number of false positive results is mainly correlated to the shift parameter. In set C, a shift parameter value of '1' caused several short matches with a length of 45–55 nt to appear in prominent places. These matches are indeed very similar to the query but not long enough, hence they are false positives. The exact match parameter did not influence much the number of false positives but it did decrease the runtime from almost a minute in set A to a few seconds in set C. However, some care is needed when trying to speed up the search because in our case an exact match with values 17 and higher might miss good results, such as the one in position 648460 still appearing in set C. From the presented parameter sets above and additional ones, we conclude that the STR2 is robust with respect to parameter sensitivity.

REFERENCES

- Altschul, S. F., Warren, G., Webb, M., Eugene, W. M. and David, J. L. (1990). Basic local alignment search tool. J. Mol. Biol. **215**, 403-410.
- Bafna, V. and Zhang, S. (2004). FastR: Fast database search tool for non-coding RNA, Proceedings of the 3rd IEEE Computational Systems Bioinformatics Conference, 52-61, August 16-19, Stanford, CA, USA.
- Barrick, J. E., Corbino, K. A., Winkler, W. C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J. K. and Breaker R. R. (2004). New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc. Natl. Acad. Sci. USA **101**, 6421-6426.
- Bengert, P. and Dandekar, T. (2004). Riboswitch finder - a tool for identification of riboswitch RNAs. Nucleic Acids Res. **32**, W154-159.
- Bergig, O. and Kedem, K. (2004). Fast search for structural motifs on large databases using a translation of structure to string. Technical Report # 04-06, Department of computer science at Ben Gurion University, Israel.
- Grundy, F. J. and Henkin, T. M. (1998). The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. Mol. Microbiol. **30**, 737-749.
- Gusfield, D. (1997). Algorithms on Strings, Trees, and Sequences. Cambridge University Press, New York, NY, USA.

- Higgs, P. G. (1998). Compensatory neutral mutations and the evolution of RNA. Genetica **102/103**, 91-101.
- Hofacker, I. L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res. **31**, 3429-3431.
- Ji, Y., Xu, X. and Stormo, G. D. (2004). A graph theoretical approach to predict common RNA secondary structure motifs including pseudoknots in unaligned sequences. Bioinformatics **20**, 1591-1602.
- Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. Acta Cryst. **A34**, 827-828.
- Klein, R. J., and Eddy, S. R. (2003). RSEARCH: Finding homologs of single structured RNA sequences. BMC Bioinformatics **4**, 44.
- Kolodny, R., Koehl, P., Guibas, L. and Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. J. Mol. Biol. **323**, 297-307.
- Le, S.-Y., Maizel, J. V. jr. and Zhang, K. (2004). An algorithm for detecting homologues of known structured RNAs in genomes, Proceedings of the 3rd IEEE Computational Systems Bioinformatics Conference, 300-310, August 16-19, Stanford, CA, USA.
- Macke, T. J., Ecker, D. J., Gutell, R. R., Gautheret, D., Case, D. A. and Sampath, R. (2001). RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res. **29**, 4724-4735.
- Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. and Breaker, R. R. (2003). Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. Cell **113**, 577-586.
- Marks, V. D., van der Merwe, G. K. and van Vuuren, H. J. (2003). Transcriptional profiling of wine yeast in fermenting grape juice: regulatory effect of diammonium phosphate. FEMS Yeast Res. **3**, 269-287.
- Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. **288**, 911-940.
- Miranda-Rios, J., Navarro, M. and Soberon, M. (2001). From the Cover: A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. Proc. Natl. Acad. Sci. USA **98**, 9736-9741.
- Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A., Perumov, D. A. and Nudler, E. (2002). Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. Cell **111**, 747-756.
- Nou, X. and Kadner, R. J. (2000). Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA **97**, 7190-7195.
- Nudler, E. and Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. Trends Biochem. Sci. **29**, 11-17.
- Park, B. H. and Levitt, M. (1995). The complexity and accuracy of discrete state models of protein structure. J. Mol. Biol. **249**, 493-507.
- Pavesi, G., Mauri, G., Stefani, M. and Pesole, G. (2004). RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. Nucleic Acids Res. **32**, 3258-3269.
- Pearson, W. R. and Lipman, D. J. (1988). Improved Tools for Biological Sequence Analysis. Proc. Natl. Acad. Sci. USA **85**, 2444-2448.
- Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. and Gelfand, M. S. (2003). Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. J. Biol. Chem. **278**, 41148-41159.
- Shapiro, B. A., Lipkin, L. E. and Maizel, J. V. Jr. (1982). An interactive technique for the display of nucleic acid secondary structure. Nucleic Acids Res. **10**, 7041-7052.
- Sudarsan, N., Barrick, J. E. and Breaker, R. R. (2003). Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA **9**, 644-647.
- Vitreschak, A. G., Rodionov, D. A., Mironov, A. A. and Gelfand, M. S. (2004). Riboswitches: the oldest mechanism for the regulation of gene expression? Trends Genet. **20**, 44-50.
- Winkler, W., Nahvi, A. and Breaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. Nature **419**, 952-956.
- Winkler, W. C. and Breaker, R. R. (2003). Genetic control by metabolite-binding riboswitches. ChemBioChem **4**, 1024-1032.
- Zuker, M. (1989). On finding all suboptimal foldings of an RNA molecule. Science **244**, 48-52.
- Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31**, 3406-3415.