P-DCfold or how to predict all kinds of pseudoknots in RNA secondary structures

Fariza Tahi, Engelen Stefan, Mireille Régnier

P-DCFOLD OR HOW TO PREDICT ALL KINDS OF PSEUDOKNOTS IN RNA SECONDARY STRUCTURES

FARIZA TAHI
La.M.I.-UMR 8042, University of Evry
Evry, 91000, France
tahi@lami.univ-evry.fr

ENGELEN STEFAN
La.M.I.-UMR 8042, University of Evry
Evry, 91000, France
sengelen@lami.univ-evry.fr

MIREILLE REGNIER
INRIA Rocquencourt
Le Chesnay, 78153, France
mireille.regnier@inria.fr

Pseudoknots play important roles in many RNAs. But for computational reasons, pseudoknots are usually excluded from the definition of RNA secondary structures. Indeed, prediction of pseudoknots increase very highly the complexities in time of the algorithms, knowing that all existing algorithms for RNA secondary structure prediction have complexities at least of $O(n^3)$. Some algorithms have been developed for searching pseudoknots, but all of them have very high complexities, and consider generally particular kinds of pseudoknots.

We present an algorithm, called *P-DCFold* based on the comparative approach, for the prediction of RNA secondary structures including all kinds of pseudoknots. The helices are searched recursively using the "Divide and Conquer" approach, searching the helices from the "most significant" to the "less significant". A selected helix subdivide the sequence into two sub-sequences, the internal one and a concatenation of the two externals. This approach is used to search non-interleaved helices and allows to limit the space of searching. To search for pseudoknots, the processing is reiterated. Therefore, each helix of the pseudoknot is selected in a different step.

P-DCFold has been applied to several RNA sequences. In less than two seconds, their respective secondary structures, including their pseudoknots, have been recovered very efficiently.

Keywords: RNA secondary structure; pseudoknots; comparative approach; "Divide and conquer" approach.

1. Introduction

The concept of secondary structure was introduce by Doty and Fresco.[1] The secondary structure is composed of all the Watson-Crick pairings, AU, GC, and the

Wobble pairing GU. The consecutive pairing form helices and could be interleaved to make the pseudoknots. The knowledge of secondary structure is essential to understand the relations between structure and function of the RNA. The computational methods developed to predict the secondary structure of RNA have complexities in $O(n^3)$ or higher, where n is the length of the sequence under study. These high complexities allow to predict only small structures of RNA. These methods belong to two main approaches: the energy minimization methods[2–5] and the comparative sequence analysis methods.[6–9] The comparative methods rather have been generally more successful and robust than the energy methods on large RNA sequences.

Many important RNA molecules contain pseudoknots, which are usually excluded from the conventional definition of the secondary structure. Observations have suggested a role for pseudoknots as conformational switches or control elements in several biological functions.[10] In molecules that lack an overall three-dimensional fold, pseudoknots fold locally and their positions along the sequence reflect their function.[11] For example, pseudoknots that are folded at the 5'-end of mRNAs are frequently involved in translational control whereas those at the 3'-end maintain signals for replication. A database for RNA pseudoknots has been developed, called *Pseudobase*.[12]

Most algorithms developed for the RNA secondary structure prediction do not allow pseudoknots. The main reasons are computational. In Ref. 13, it has been proved that the general problem of predicting RNA secondary structures containing pseudoknots is NP-hard for a large class of reasonable models of pseudoknots. In Ref. 14, a modeling for pseudoknot searching is proposed (SCFG), using Stochastic Context Free Grammars. It is restricted to the search for pseudoknots with two helices and 3 simple loops (e.g. without other helices), on very small sequences. The pseudoknot grammar is decomposed into two SCFG's and allows $O(n^3)$ parsing algorithms instead of the $O(n^5)$ operations that would be required if the full pseudoknot grammar was used.

In Ref. 15, an algorithm based on the Maximum Weighted Matching (MWM) method[16] has been proposed. It remains an $O(n^2)$ memory and $O(n^3)$ time process when performing a 2-matching. In Ref. 17, a dynamic programming algorithm is presented, which includes some kinds of pseudoknots (simple pseudoknots). Its complexity is $O(n^6)$ in time and $O(n^4)$ in space. Because its complexity, this algorithm can be used only on very small sequences (less than 150 bases). In Ref. 13, an algorithm in time $O(n^5)$ and space $O(n^3)$ has been proposed, with a model that allows certain kinds of pseudoknots. In Ref. 18, a minimum free energy folding algorithm was implemented. It requires $O(mn^3)$ time and $O(mn^2)$ space, where m is a constant depending on the structural freedom approved to the pseudoknots. It searches only the most simple type of pseudoknots, called H-type pseudoknots.

In this paper, we propose an algorithm, called *P-DCFold*, for the prediction of RNA secondary structure including all kinds of pseudoknots. In Ref. 19, we proposed an algorithm, *DCFold*, for the RNA secondary structure prediction, based

on the comparative approach. The helices are searched recursively, from the more "likely" to the less "likely", using the "divide and conquer" approach. This approach, which allows to limit the amount of searching, was possible because pseudoknots were not searched. *P-DCFold* is an extension of $DCFold$ that allows to detect all the pseudoknots of the considered structure. It consists to search for pseudoknots in several steps. At each step, only "compatible helices", i.e. do not forming pseudoknots, are searched. Therefore, each helix of a pseudoknot is selected in a different step.

The algorithm has been tested on tmRNA, RNase P and SRP RNA which contain interesting pseudoknots. We also tested our algorithm on structures of RNA 5S and u1 RNA which do not have pseudoknots. The obtained results are very satisfactory, since all the pseudoknots and almost all of the helices are predicted. More important, our algorithm avoid the selection of false positive helices and pseudoknots. Finally, the complexity of *P-DCFold* is more interesting than other existing algorithms, since it is in the worse case of $O(log_4 n * n^2)$ in time and $O(n^2)$ in space.

2. RNA Structures and Pseudoknots — Definitions

Definition 2.1. A **RNA secondary structure** is composed of a set of **helices**, bulges and internal and external loops. A helix can be composed of smaller helices, separated by bulges or internal loops.

A helix is defined by a *palindrome* that represents a particular kind of repetition in the sequence.

Definition 2.2. A **palindrome** is a couple of words (p, p') such that:

$$(i)\ \ |p| = |p'| = m$$
$$(ii)\ \ p[k]R_c p'[m-k+1], \qquad \forall k, 1 \leq k \leq m,$$

where R_c is the relation between nucleotides: AR_cU, GR_cC and GR_cU.

Definition 2.3. A **structural palindrome** is a palindrome which defines a helix of the secondary structure.

Definition 2.4. Given a set of aligned sequences, a palindrome appearing in each sequence at the same aligned position is said **conserved**.

A palindrome (p, p') is not necessarily conserved with the same pairs of bases. Some mutations may occur. They are *compensated* when the pairing of the palindrome bases still remains possible.

Definition 2.5. Two palindromes (p, p') and (q, q') are **compatible** when they appear as follows:
— disjoint:

```
   p     p'    q     q'
...-->...<--...-->...<--...
```

- embedded:

```
   p     q    q'     p'
...-->...-->...<--...<--...
```

Definition 2.6. Two palindromes (p, p') and (q, q') form a **pseudoknot** when they appear interleaved:

```
      p     q     p'    q'
...-->...-->...<--...<--...
```

Definition 2.7. A **P-pseudoknot** is a pseudoknot composed of P interleaved palindromes.

When P is equal to 2, the pseudoknot is composed of two palindromes appearing as follows:

```
      p     q     p'    q'
...-->...-->...<--...<--...
```

When P is equal to 3, the pseudoknot is composed of three palindromes appearing as follows:

```
      p     q     r      p'    q'    r'
...-->...-->...-->... <--...<--...<--
```

Definition 2.8. A RNA secondary structure has a **complexity** of C, $C > 0$, if it contains at least one C-pseudoknot and no any $(C + k)$-pseudoknot, $k > 0$.

When C is equal to 1, the secondary structure does not contain any pseudoknot.

Almost all known secondary structures are of complexity 1, i.e. have no pseudoknot, or of complexity 2. But we know that structures with higher complexities exit. One is the *Escherichia coli α-operon mRNA*, which is of complexity 3.[20–22]

The complexity of a RNA secondary structure can be represented by a graph, called linked graph or linked diagram, where each pairing of two helices nucleotides is represented by an arc.[18,23] The complexity is related to the so-called *book-thickness p* of the linked graph in Ref. 18, or the *chromatic number* in Ref. 23.

3. *P-DCFOLD* Description

3.1. *Principle of DCFold*

P-DCFold is an extension of the algorithm *DCFold*.[19]

DCFold is an algorithm for RNA secondary structure prediction based on the comparative approach. Given a set of aligned sequences, the goal of the algorithm is to predict the secondary structure of one sequence, called the "*target sequence*", using informations from the other sequences, called the "*test sequences*". The search for helices is done in two steps: first, we search for helices in the target sequence,

then we check their conservation in the test sequences, in order to select the "most significant" ones to define common helices.

A conserved palindrome is not always structural. One of the central problems in our algorithm is to determine a heuristic criterion to select the ones that actually are structural. Two criteria are combined: the length of the palindromes and the number of compensated mutations. The palindromes are selected only if their length is greater than $log_4 n$, where n is the target sequence length, and when they present at least one compensated mutation per site.

We also attribute scores to palindromes according to thermodynamic parameters. The first parameter is on pairing stabilities. Indeed, GC pairings are more stable than AU pairings which are also more stable than GU ones. The second pairing is function of pairings close to helices. Indeed, we observe that GU pairings rarely close to helices.

The algorithm uses the "divide and conquer" approach. The principle is as follows:

- Search for palindromes that satisfy the length criterion and our thermodynamic parameters in the target sequence S.
- Select among them the conserved ones in the test sequences satisfying the mutation number criterion.
- Deduce a "valid set of anchoring points": palindromes selected above that are all mutually compatible.
- Iterate the process on sub-sequences of S (deduced from the subdivision of S by the anchoring points) that are long enough to contain palindromes.

As the length criterion is function of the sequences and sub-sequences length, the subdivision allows to search palindromes from the most significant to the less significant. The "divide and conquer" approach allows to reduce the space of research for the less significant palindromes, by forcing the structure with the most significant palindromes.

The obtained palindromes constitute a set of structural palindromes, i.e. defining the secondary structure helices. The subdivision of the sequence is possible because only compatible palindromes (so no pseudoknots) are searched in $DCFold$.

The complexity in time of $DCFold$ in the worse case is equal to $O(log_4 n * n^2)$, when n is the sequence length. Indeed,the search for palindromes in a sequence can be done without effort in $O(n^2)$ and the number of iterations is less than $log_4 n$. The complexity in space is in the worse case equal to $O(n^2)$.

3.2. *Principle of the pseudoknot searching*

The search for pseudoknots is done after finding the secondary structure without pseudoknots. Given a sequence S, $DCFold$ finds a list $L1$ of all compatible palindromes that satisfy our selection criteria. Re-launch $DCFold$ on S without

palindromes of L (S') allows to find another list $L2$ of all compatible palindromes which are not compatible with palindromes of $L1$. Therefore, a palindrome of $L2$ will form a 2-pseudoknot with a palindrome of $L1$. Now, relaunch again $DCFold$ on S' without palindromes of $L2$ allows to find a third list $L3$ of compatible palindromes which are not compatible with palindromes of the list $L1$ and with palindromes of $L2$ list. Therefore, a palindrome of $L3$ will form with a palindrome of $L1$ and a palindrome of $L2$ a 3-pseudoknot. And so on, until no palindromes are found. If $DCFold$ is launched C times, the secondary structure is found with a complexity equal to C.

Therefore, the principle of the algorithm is to search for pseudoknots in several steps, each helix of the pseudoknot being selected in a different step.

3.3. *The algorithm*

P-DCFold uses the procedure of *Structural_palindrome_search* defined in *DCFold*. This procedure is launched in a first time to find all compatible palindromes (helices with no pseudoknots). Then it is re-launched on the initial sequence without the sub-sequences corresponding to the selected palindromes.

The global procedure of structural palindrome searching including pseudoknots, called *All_Structural_Palindrome_Search*, is presented in Figure 1.

Procedure *All_Structural_Palindrome_Search*(S)
Begin
$L_{all} = \emptyset$ * L_{all}: global list of structural palindromes
$n = |S|$ * n: size of the target sequence S
$C = 0$ * C : complexity of the secondary structure $L_g = \emptyset$
Structural_Palindrome_Search(Ł$_g, S, n$)
$L_{all} \leftarrow L_{all} \cup L_g$
While ($L_g \neq \emptyset$)
  begin $C = C + 1$;
    Let S_g the global sequence S without the sub-sequences associated to palindromes of L_{all}
    $L_g = \emptyset$
    Structural_Palindrome_Search($L_g, S_g, |S_g|$)
    $L_{all} \leftarrow L_{all} \cup L_g$;
  end
Return(C,L_{all})
End

Fig. 1. Procedure of structural palindrome searching including pseudoknots

The procedure *All_Structural_palindrome_search* is based on the procedure *Structural_palindrome_search*, defined in Ref. 19, which searches for compatible structural palindromes (pseudoknots are not allowed).

This procedure is launched several times, until the list of the compatible palindromes obtained is empty. The number of calls is related to the complexity of the secondary structure. Indeed, when the procedure is launched only once, the predicted secondary structure is of complexity 1, i.e. without pseudoknots. If it is launched twice, then it is of complexity 2, i.e. it has 2-pseudoknots, etc. This is based on the principle that when the procedure *Structural_palindrome_search* is launched, all the structural compatible palindromes that satisfy our criteria are detected. Therefore, when the procedure is re-launched, a new list of compatible palindromes is detected, which is necessarily not compatible with the precedent lists. If the number of the precedent lists is equal to i, the palindromes of the new list will form i-pseudoknots with the ones of the precedent lists.

3.4. *Implementation*

The algorithm *P-DCFold* (and also $DCFold$) has been implemented in Java language. The input to *P-DCFold* is a set of aligned homologous RNA sequences, one of them is the target sequence, i.e. the sequence we want to predict the structure. The output is the positions of the target sequence palindromes, corresponding to the helices of the predicted secondary structure.

In order to align the input sequences, we have integrated to *P-DCFold* the software "ClustalW".[24] In order to visualize the results, we have integrated the software "RnaViz"[25] which, given a sequence and positions of helices, draws the corresponding secondary structure. We have therefore a complete RNA secondary structure prediction software.

4. Results

4.1. *Results on tmRNA structure*

The tmRNA (also known as 10Sa RNA or SsrA) plays an important role in translation. It combines both transfer and messenger RNA properties in order to solve problems arising from ribosomes stalled in translation.[26] Different informations about this RNA, such as its secondary structure, are provided in the tmRNA Website[27] and the tmRDB site.[28] These sites store more than two hundreds tmRNA sequences, and an alignment is provided for some of them. A survey on the tmRNA structure is given in Ref. 29.

The tmRNA secondary structure presents four pseudoknots, which make it interesting for our study. We have extracted from the tmRDB site five sequences, namely *Escherichia coli*, *Shewanella putrefaciens*, *Aquifex aeolicus*, *Thermotoga maritima* and *Enterococcus faecalis* initially aligned. The secondary structure was searched for the *Escherichia coli* sequence, which was the sequence where tmRNA was first identified.

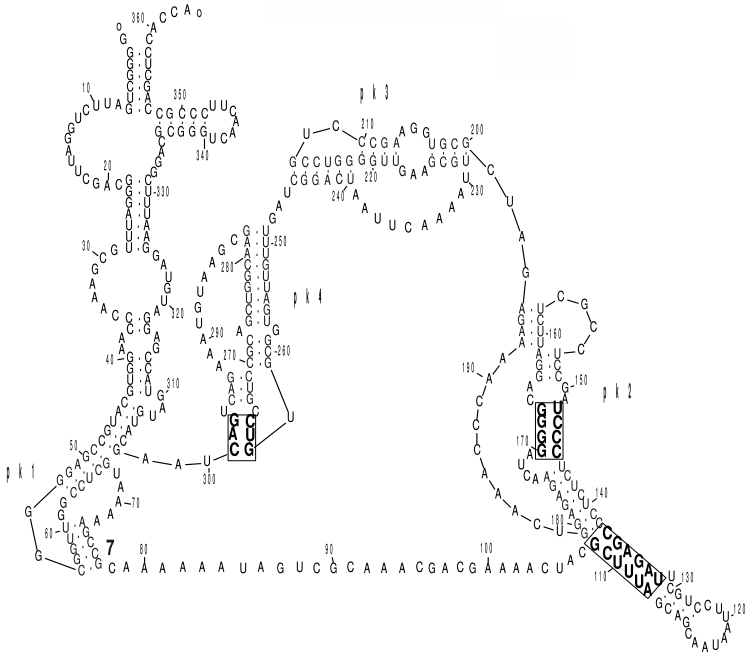

Fig. 2. The tmRNA secondary structure of Escherichia Coli predicted by *P-DCFold*. Bold pairings depicted in boxes correspond to false negative helices, i.e. helices that have not been detected by *P-DCFold*

We obtain very nice results, since the secondary structure predicted by *P-DCFold* corresponds to the known structure (see Figure 2). Indeed, *P-DCFold* predicts all the pseudoknots of the structure and almost all the helices. Only three helices have not been predicted. Also, *P-DCFold* does not predict any false positive helix or pseudoknot.

4.2. *Results on the Ribonuclease P structure*

Ribonuclease P is involved in processing all species of tRNA and is present in all cells and organelles that carry out tRNA synthesis.[30−32] A compilation of RNase P sequences, sequence alignments, secondary structures and three dimensional models are given in the RNase P Database.[33] The secondary structure of RNase P contains two pseudoknots. We applied our algorithm on the sequence of the RNase P of *Escherichia Coli*, compared to four test sequences: *Desulfovibrio desulfuricans*, *Rhodospirillum rubrum*, *Streptomyces bikiniensis* and *Deinococcus radiodurans*. On this example, we also obtained very nice results (see Figure 3).

Our algorithm detected almost all helices of RNase P of E. Coli, including the two pseudoknots. No false positive helices have been selected. There are five false negative helices, but among them, three are extensions of detected helices.

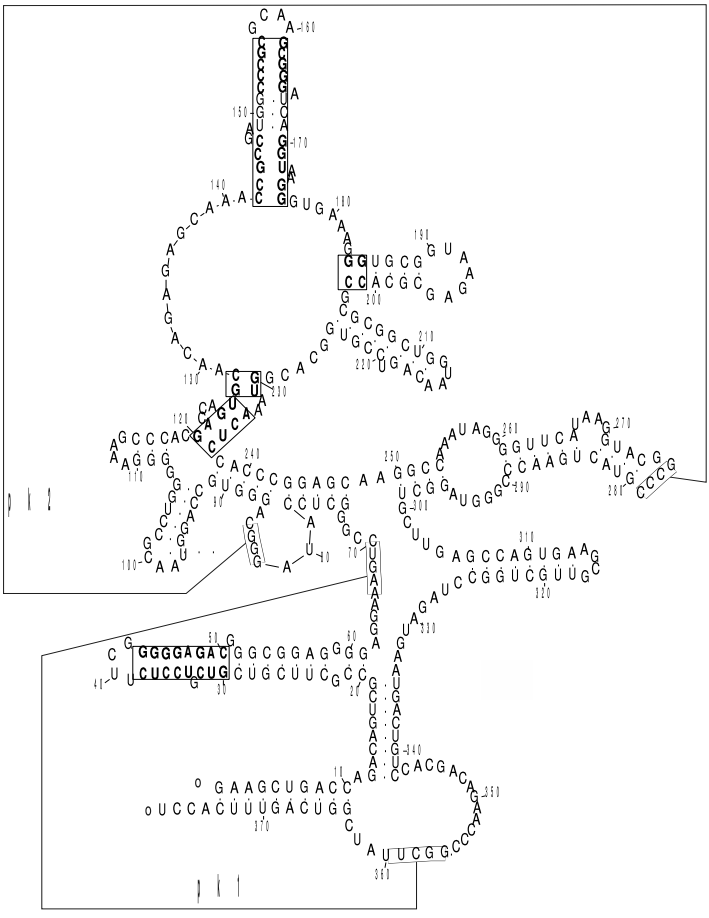

Fig. 3. The RNase P secondary structure of Escherichia Coli predicted by *P-DCFold*. Nucleotides in linked boxes represent the pseudoknots and bold pairings in boxes represent false negative helices (not predicted helices)

4.3. *Results on the SRPRNA structure*

The signal recognition particle (SRP) contains SRPRNA and associates with ribosomes that are in the process of translating the mRNA for a secretory protein. It allows to address secretory protein to Endoplasmic Reticulum membrane. We have used sequence alignment provided by the Signal Recognition Particle Database.[34] The SRPRNA has about 300 nucleotides. We applied our algorithm on SRPRNA of *Halobacterium halobium* using four tests sequences: *Haloferax volcanii, Methanococcus jannaschii, Methanothermus fervidus* and *Staphylococcus epidermidis*. The real structure of SRPRNA have twenty helices and our prediction gives only one false positive helix (pointed with an arrow) and one false negative helix (see Figure 4). *P-DCfold* success to predict the pseudoknot of the structure.

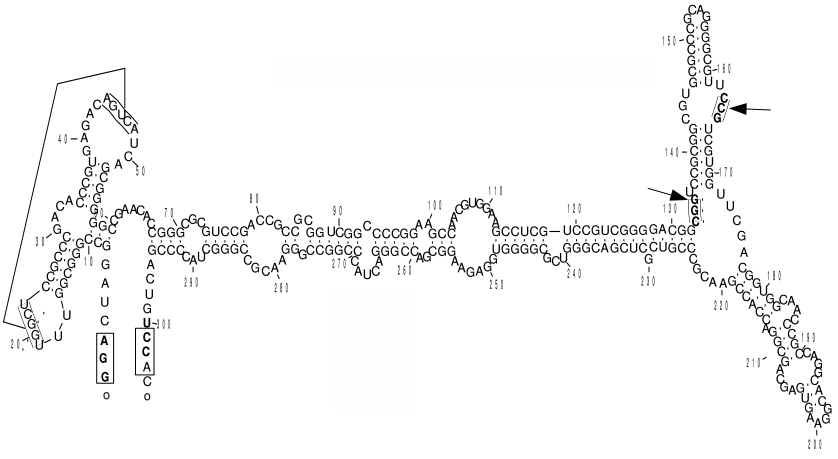

Fig. 4. The SRPRNA secondary structure of Halobacterium halobium predicted by *P-DCFold* contains one false positive helix (pointed by arrows) and one false negative (bold bases in boxes)

4.4. *Results on the u1RNA structure*

The u1RNA are component of the spliceosome which has a role in RNA splicing. This RNA has a small length and has not pseudoknots. We predicted the structure of the u1RNA of *Echinococcus multilocularis* (see Figure 5), using the following test sequences provided from the uRNA Database[35]: *Drosophila melanogaster, Caenorhabditis elegans, Physarum polycephalum* and *Tetrahymena thermophila*.

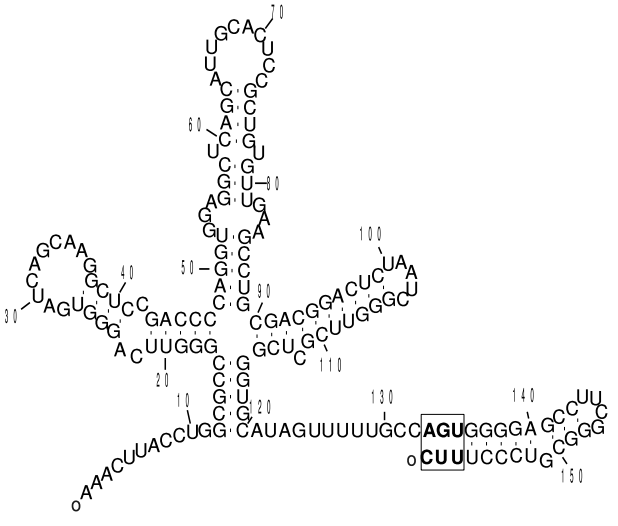

Fig. 5. The u1RNA secondary structure of Echinococcus multilocularis predicted by *P-DCFold*. In the box, the pairings not predicted by *P-DCFold*

P-DCFold prediction finds all the ten helices of the u1RNA structure. Only one helix has not been predicted in its whole.

4.5. *Results on the 5SRNA structure*

The 5SRNA is a component of the ribosome which allows the translation of the proteins. The length is of 120 nucleotides. The structure of this RNA is very simple and has only five long helices and no pseudoknots. The sequences used for the prediction are from the 5S ribosomal RNA database.[36] We predict the 5SRNA structure of Escherichia Coli with test sequences from *Helicobacter pylori, Clostridium carnis, Cytophaga aquatilis* and *Borrelia burgdorferi*. One helix of the structure is composed of non canonical pairings and couldn't be found by *P-DCfold*. All The other helices are well found by our algorithm (see Figure 6).

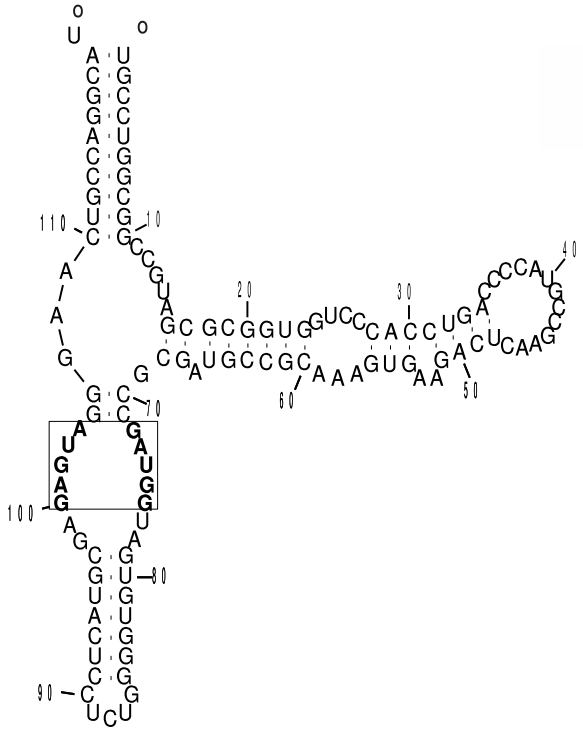

Fig. 6. The RNA5S secondary structure of Escherichia Coli predicted by *P-DCFold*

5. Conclusion

Our algorithm presents several strong points. The first one is the palindrome selection criteria used. Indeed, we have set length and mutation criteria that proved their highly efficiency. Only one false positive helix has been selected on the different RNA sequences used for our tests. Also, the algorithm detects almost all the

helices of the secondary structure. The very few exceptions concern regions with strong variability or regions highly conserved, i.e. without mutations. For example, in the tmRNA structure, among the three helices not detected, two correspond to a region with a strong variability (the helices do not appear in one of the considered test sequences) and one corresponds to a palindrome having a mismatch. In the RnaseP structure, among the four helices not detected, two correspond to a region with a strong variability and two to a region without mutations. Therefore, our algorithm insures a good prediction of the secondary structure including any kind of pseudoknots. In almost all cases, the complexity of the secondary structure has been well predicted. It was of 2 in our example (tmRNA, RNase P and SRP RNA).

Another strong point is the fact that very few sequences are necessary for our algorithm. Their number depends on the target sequence length l. It is set equal to $l/200$ when l is greater than 800 nucleotides, to 4 otherwise. Indeed, we consider that a palindrome must be conserved at least in four sequences in order to be valid, even when it concerns the treatment of a variable region.

Finally, the high performance of our algorithm is its ability to correctly predict a secondary structure in record time. For example, on the tmRNA structure, the running time is less than two seconds, the same on the RnaseP structure. Indeed, the search for pseudoknots do not increase the algorithm complexity, since it is just a re-launching of *DCFold*. The complexity in time in the worse case of *P-DCFold* is then equal to $O(log_4 n * n^2)$, when n is the sequence length.

Note that to compare our algorithms to others in results, this is difficult because no availability of RNA prediction algorithm including pseudoknots. Only the algorithm of Rivals and Eddy[17] is available but unfortunately, it can not be used on our sequences because its high complexity (it can be used only on very small sequences, less than 150 bases). Another algorithm (based on an other approach using the genetic algorithms) is available but is not free.[37] The most used algorithm for the RNA secondary structure prediction is *Mfold*, developed by Zucker.[4,38] This algorithm is based on the thermodynamic approach and the pseudoknots are not allowed. We have tested Mfold on our tmRNA and RNase P sequences and in all cases, we have obtained with our algorithm better results (Mfold finds less exact helices than *DCFold* and *P-DCFold* and a non unimportant number of false positive helices) and in less time.

A crucial and not yet solved problem in secondary structure prediction using the comparative approach is the selection of homologous sequences to use for the prediction. These sequences must be distant enough from the target sequence to have compensated mutations and close enough to have a minimum of differences in the secondary structure. Another reason to select the homologous sequences is the importance of the quality of the sequence alignment. Indeed, more the sequences are well aligned, more the prediction results are better. We are therefore developing an algorithm to select homologous sequences according to their variability and their alignment.

References

1. P. Doty, H. Boedtker, J.R. Fresco, R. Haselkorn, and M.Litt. Secondary structure in ribonucleic acids. *Proc. Natl. Acad. Sci.*, 1959.
2. M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9:133–148, 1981.
3. M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 4:591–621, 1984.
4. M. Zuker, D. H. Mathews, and D. H. Turner. Algorithms and thermodynamics for RNA secondary structure prediction. *A practical Guide*, 1999.
5. B. Cohen and S. Skiena. Designing RNA structures: Natural and artificial selection. In *Sixth Annual International Conference on Computational Biology*, pages 109–116. ACM Press, 2002.
6. R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acis. Res.*, 20:5785–5795, 1992.
7. J. Gorodkin, L. J. Heyer, and G. D. Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Research*, 25(18):3724–3732, 1997.
8. V. R. Akmaev, S. T. Kelley, and G. D. Stormo. A phylogenetic approach to RNA structure prediction. *Proc. Int. Conf. Intell. Syst. Mol. Bio.*, pages 10–7, 1999.
9. J. Gorodkin, S. L. Stricklin, and G. D. Stormo. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Research*, 29(10):2135–2144, 2001.
10. P. Schimmel. RNA pseudoknots that interact with components of the translation apparatus. *Cell*, 58:9–12, 1989.
11. R. Mans, C. Pleij, and L. Bosch. Transfer RNA-like structures: Structure, function and evolutionay significance. *Eur. J. Biochem.*, 201:303–324, 1991.
12. F. H. D. Van Batenburg, A. P. Gultyaev, and C. W. A. Pleij. Pseudobase: a database with RNA pseudoknots. *Nucl. Acids. Res*, 28:201–204, 2000.
13. R. B. Lyngso and C. N. S. Pedersen. Pseudoknots in RNA secondary structures. In *RECOMB*, pages 201–209, Tokyo, Japan, 2000. ACM 2000.
14. M. Brown and C. Wilson. RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. *Proceedings of the 1996 Pacific Symposium*, 1996.
15. J.E. Tabaska, R. B. Cary, H.N. Gabow, and G.D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.
16. R. B. Cary and G. D. Stormo. Graph-theoretic approach to RNA modeling using comparative data. In *Third International Conference on Intelligent Systems for Molecular Biology*, pages 75–80. AAAI Press, Menlo Park, CA, 1995.
17. E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.
18. C. Haslinger. *Prediction Algorithms for Restricted RNA Pseudoknots*. PhD thesis, Universitat Wien, March 2001.
19. F. Tahi, M. Gouy, and M. Regnier. Automatic RNA secondary structure prediction with a comparative approach. *Computers and Chemistry*, 26:521–530, 2002.
20. T. C. Gluick and D. E. Draper. Thermodynamics of folding a pseudoknotted mRNA fragment. *Journal of Molecular Biology*, 241:246–262, 1994.
21. C. K. Tand and D. E. draper. An unusal mRNA pseudoknot structure is recognized by a protein translation repressor. *Cell*, 57:531–536, 1989.
22. C. K. Tand and D. E. draper. Evidence for allosteric coupling between the ribo-

some and repressor binding sites for a translationally regulated mRNA. *Biochemistry*, 29:4434–4439, 1990.

23. C. Haslinger and P. F. Stadler. RNA structures with pseudo-knots: Graph-theoretical, combinatorial ans statistical properties. *Bul. Math. Biol.*, 61:437–467, 1999.
24. J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.
25. P. De Rijk and R. De Wachter. Rnaviz, a program for the visualisation of RNA secondary structure. *Nucleic Acids Res.*, 25(22):4679–4684, 1997.
26. B. Felden, H. Himeno, A. Muto, J. McCutcheon, J. Atkins, and R. Gesteland. Probing the structure of the Escherichia coli 10Sa RNA (tmRNA). *RNA*, 3:89–103, 1997.
27. K. P. Williams and D. P. Bartel. The tmRNA Website. *Nucleic Acids Res.*, 26:163–165, 1998.
28. B. Knudsen, J. Wower, C. Zwieb, and J. Gorodkin. tmRDB (tmRNA database). *Nucleic Acids Res.*, 29:171–172, 2001.
29. C. Zwieb, I. Wower, and J. Wower. Survey and summary. Comparative sequence analysis of tmRNA. *Nucleic acids Research*, 27(10):2063–2071, 1999.
30. S. Altman, L. Kirsebom, and S. Talbot. Recent studies of Ribonuclease P. *FASEB J.*, 7:7–14, 1993.
31. S. C. Darr, J. W. Brown, and N. R. Pace. The variations of Ribonuclease P. *Hermès, PariTrends Biochem. Sci.*, 17:178–182, 1992.
32. T. Pavlidis, R. Pace, and D. Smith. Ribonuclease P: function and variation. *J. Biol. Chem.*, 265:3587–3590, 1990.
33. J. W. Brown. The Ribonuclease P Database. *Nucleic Acids Research*, 27:314, 1989.
34. Alm Rosenblad M., Gorodkin J., Knudsen B., Zwieb C., and Samuelsson T. Srpdb (signal recognition particle database). *Nucl. Acids Res.*, 2003.
35. Zwieb C. The urna database. *Nucl. Acids Res.*, 2003.
36. Maciej Szymanski, Miroslawa Z. Barciszewska, Volker A. Erdmann, and Jan Barciszewski. 5s ribosomal rna database. *Nucl. Acids Res.*, 2002.
37. F. H. D. Van Batenburg, A. P. Gultyaev, and C. W. A. Pleu. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. theor. Biol.*, 174:269–280, 1995.
38. D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.