DP Algorithms for RNA Secondary Structure Prediction with Pseudoknots

Tatsuya Akutsu
takutsu@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

Abstract

This paper describes simple DP (dynamic programming) algorithms for RNA secondary structure prediction with pseudoknots, for which no explicit DP algorithm had been known. Results of preliminary computational experiments are described too.

1 Introduction

The problem of RNA secondary structure prediction is, given an RNA sequence, to compute its correct *secondary structure* (a tree-like structure) [7, 9]. Although it is still hard to compute (nearly) correct structures for all sequences, several methods have been developed and have been successfully applied to several RNA sequences. In most of such methods, RNA secondary structure prediction is defined as an energy minimization problem, in which an *optimal secondary structure* (i.e., a secondary structure with minimum pseudo energy) is to be computed.

Zuker and Stiegler developed a well-known DP algorithm [10], which can always find an optimal structure. However, their algorithm can not handle pseudoknots. For predicting RNA secondary structure with pseudoknots (see Fig. 1), several methods have been proposed. Abrahams et al. developed a local search method [1] and Akiyama et al. proposed a method using the Hopfield network [2]. However, these methods may miss optimal structures since they are not guaranteed to find optimal structures. Recently, Uemura et al. proposed a method using *tree adjunct grammar* [8]. Although their method can handle pseudoknots and can always find an optimal structure in polynomial time, it is complicated and hard to understand. Thus, we analyzed their method and we found that tree adjunct grammar was not crucial but the parsing procedure was crucial. Since the parsing procedure is intrinsically a DP procedure, we can re-formulate their method as a DP procedure without tree adjunct grammar. In this paper, we describe such a DP procedure. This DP procedure

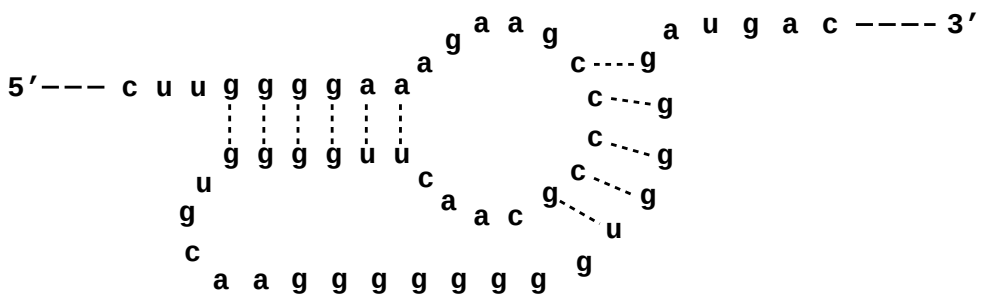

Figure 1: Example of a pseudoknot structure.

has some advantages over the original method: it is easier to understand, it is much simpler, it is easier to modify, and it is easier to cope with various score functions (i.e., pseudo energy functions). Some variants of the DP procedure, which cover most types of pseudoknots, are described as well as preliminary computational experiments.

Because of the high time complexity ($O(n^4)$, or $O(n^5)$) as in Ref. [8], the proposed algorithms are not yet practical. However, we believe that they will be made practical if some heuristics are combined with them (some ideas for such practical improvements are discussed in this paper). Moreover, we emphasize that the most important contribution of this paper is that it corrects the previous misunderstanding that pseudoknots can hardly be handled by a DP-based approach.

2 A DP Algorithm for Simple Pseudoknots

In order to explain the basic idea, we first describe a DP algorithm for a case that types of pseudoknots and limited to be simple ones.

Before considering pseudoknots, we briefly review a DP algorithm for RNA secondary prediction without pseudoknots [10]. In a simplest form, this problem is formalized as a problem of maximizing the number of base pairs (AU,GC,GU pairs), under the condition that secondary structure has a tree-like form. It is well known that the following simple DP procedure (recurrence) solves this problem in $O(n^3)$ time:

$$S(i,j) \;=\; \min\{\; S(i+1,j-1)+f(i,j),\; \min_{i\le k<j}\{S(i,k)+S(k+1,j)\}\;\},$$

where we omit the part of initialization, n denotes the length of an input sequence, and $f(i,j)=-1$ if i-th and j-th residues make a base pair, otherwise $f(i,j)=0$.

Next we consider pseudoknot substructures. Although no explicit definition of a pseudoknot is known, we consider such kinds of pseudoknots as (A) and (B) in Fig. 2 (i.e., recursive pseudoknot structure is not allowed). We call such a pseudoknot *a simple pseudoknot*.

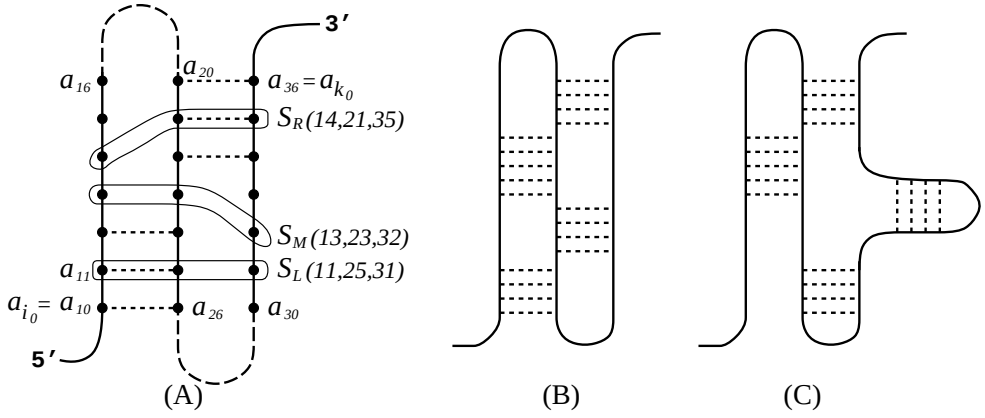

Figure 2: Three types of triplets used in a DP procedure are shown in (A), where turning points need not be fixed. The DP procedure can handle pseudoknots with forms (A) and (B), and can be modified for handling pseudoknots with form (C).

For finding a simple pseudoknot substructure whose endpoints are i_0-th and k_0-th residues, we consider triplet (i,j,k) $(i_0 \le i < j < k \le k_0)$ instead of (i,j). Moreover, we consider three types of triplets $S_L(i,j,k)$,$S_M(i,j,k)$, and $S_R(i,j,k)$. $S_L(i,j,k)$ (resp. $S_R(i,j,k)$) corresponds to a case that i-th and j-th (resp. j-th and k-th) residues make a base pair. Then, each triplet can be computed by the following recurrence:

$$
\begin{aligned}
S_L(i,j,k) &= \min\{\ S_L(i-1,j+1,k),\ S_M(i-1,j+1,k),\ S_R(i-1,j+1,k)\ \} + g(i,j),\\
S_R(i,j,k) &= \min\{\ S_L(i,j+1,k-1),\ S_M(i,j+1,k-1),\ S_R(i,j+1,k-1)\ \} + g(j,k),\\
S_M(i,j,k) &= \min\{\ S_M(i-1,j,k),\ S_M(i,j+1,k),\ S_M(i,j,k-1),\\
&\qquad S_L(i-1,j,k),\ S_L(i,j+1,k),\ S_R(i,j+1,k),\ S_R(i,j,k-1)\ \},
\end{aligned}
$$

where $g(i,j) = -1$ if i-th and j-th residues make a base pair, otherwise $g(i,j) = \infty$, and the initialization part is done by letting:

$$
\begin{aligned}
S_L(i_0,j,j+1) &= g(i_0,j) \text{ for all } j,\\
S_R(i_0,j,j+1) &= g(j,j+1) \text{ for all } j,\\
S_L(i_0,j,k) &= S_R(i_0,j,k) = S_M(i_0,j,k) = 0 \text{ for the other } j,k\\
&\qquad \text{satisfying } k=j \text{ or } k=j+1.
\end{aligned}
$$

For each (fixed) pair (i_0,k_0), we computes these triplets and we obtain a score for this pair by

$$
S_{pseudo}(i_0,k_0) = \min_{i_0 \le i<j<k \le k_0} \min\{S_L(i,j,k), S_R(i,j,k), S_M(i,j,k)\}.
$$

Finally, we obtain the minimum score by the following recurrence:

$$
S(i,j) = \min\{\ S_{pseudo}(i,j),\ S(i+1,j-1)+f(i,j),\ \min_{i\le k<j}\{S(i,k)+S(k+1,j)\}\ \}.
$$

It is almost obvious that an optimal structure can be computed using the above recurrences (see Fig. 3). Thus, we analyze the time complexity. For each pair (i_0,k_0), we must computes scores for $O(n^3)$ triplets. Therefore, scores for $O(n^5)$ triplets should be computed in total. However, scores computed for (i_0,k_0) can also be used for (i_0,k) such that $k \ge k_0$. Using this property, we can see that scores for $O(n^4)$ triplets are sufficient, where each score can be computed in constant time. Therefore, $O(n^4)$ time is sufficient in total, which matches with the time complexity in Ref. [8]. The space complexity is $O(n^3)$ because $O(n^3)$ space is required for computing $S_{pseudo}(i_0,k_0)$ (for each $S_{pseudo}(i_0,k_0)$, we can use the same memory space).

Theorem 1: An RNA secondary structure maximizing the number of base pairs can be computed in $O(n^4)$ time using $O(n^3)$ space, where a secondary structure may include simple pseudoknots.

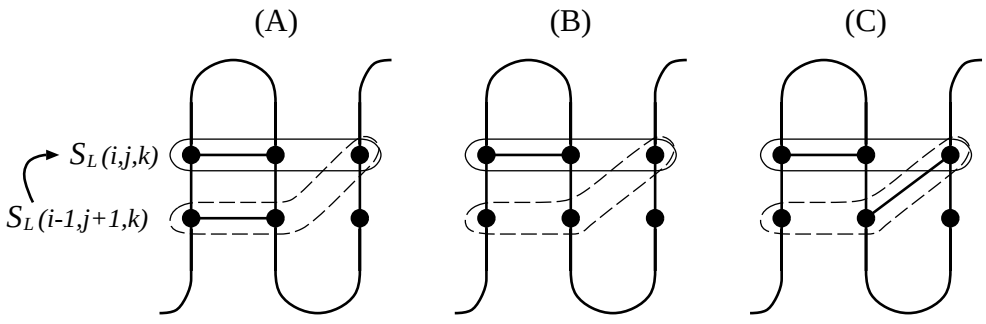

Figure 3: Illustration of the recurrence used in the DP procedure. Fig. (A) corresponds to a case that $S_L(i,j,k) = S_L(i-1,j+1,k)+g(i,j)$, Fig. (B) corresponds to $S_L(i,j,k) = S_M(i-1,j+1,k)+g(i,j)$, and Fig. (C) corresponds to $S_L(i,j,k) = S_R(i-1,j+1,k)+g(i,j)$.

3 Extensions and Limitations

Although the DP algorithm in Section 2 is simple, it is not practical and several modifications are required. Indeed, it can be modified in various ways. For example, it can be modified as follows:

(i) It can be modified for using (pseudo) energy defined by two base pairs [5, 8] without increasing the order of the time complexity.

(ii) It can be modified so that S_R type region occurs only once in a pseudoknot (i.e., such cases as Fig. 2(B) are inhibited) without increasing the order of the time complexity. This modification seems effective because S_R type region occurs only once in most pseudoknots in the literature [1, 8].

(iii) It can be modified so that recursive pseudoknot structures are allowed (i.e., some other (pseudoknot or usual) substructures can occur in a pseudoknot as in Fig. 2(C)) where the time complexity increases to $O(n^5)$ as in Ref. [8].

(iv) It can be modified so that energies for loop regions are taken into account, where the time complexity increases to $O(n^5)$ or more (depending on the forms of pseudo energy).

Since details of the above modifications are straight-forward but lengthy, we only give a short description about (iii) here (i.e., a case of recursive pseudoknots). For the simplicity, we only consider a case that recursive structures occur only in the right region as in Fig. 2(C) and we only show the recurrence for $S_R(i,j,k)$. However, the recurrence can be extended for covering the other cases in a straight-forward way without increasing the order of the time complexity. Of course, there is no limit on the depth of the recursion. Let

$$S'_R(i,j,k) = \min\{S_L(i,j,k), S_M(i,j,k), S_R(i,j,k)\}.$$

Then, $S_R(i,j,k)$ is computed by

$$\begin{aligned} S_R(i,j,k) \quad = \quad & \min\{ \quad S'_R(i,j+1,k-1) + g(j,k), \\ & \min_{k'<k-1} \{S'_R(i,j+1,k'-1) + S_{pseudo}(k',k-1) + g(j,k)\} \quad \}, \end{aligned}$$

where we can schedule DP procedure so that the required values of $S_{pseudo}(k',k-1)$ are already determined before determining $S_R(i,j,k)$.

By the above modifications, DP algorithms can cover almost all types of pseudoknots (we do not know what extent we should cover because no established definition of a pseudoknot is known). However, the forms of secondary structures can not be extended to the entire class of planar graphs. In such a case, we can prove an NP-hardness result (see Appendix).

4 Computational Experiments

We have made a computer program based on the proposed algorithm using C-language on SUN Ultra-1 workstation. Since no established method of computing energies of loop regions in pseudoknots is known, we only consider energies of stacked regions. We use the same energy parameters as in Refs. [5, 8].

In this experiment, we applied our computer program to HIV-2 gag-pol region and 16SrRNA. Note that we could not apply the program to the whole sequences because of the high ($O(n^4)$) time complexity. The results are shown in Fig. 4, where a solid line indicates a known stacked region, and a dashed line indicates a stacked region computed by our program. Note that our results are different from those in Ref. [8] because, in our program, GU pairs are taken into account, the minimum length

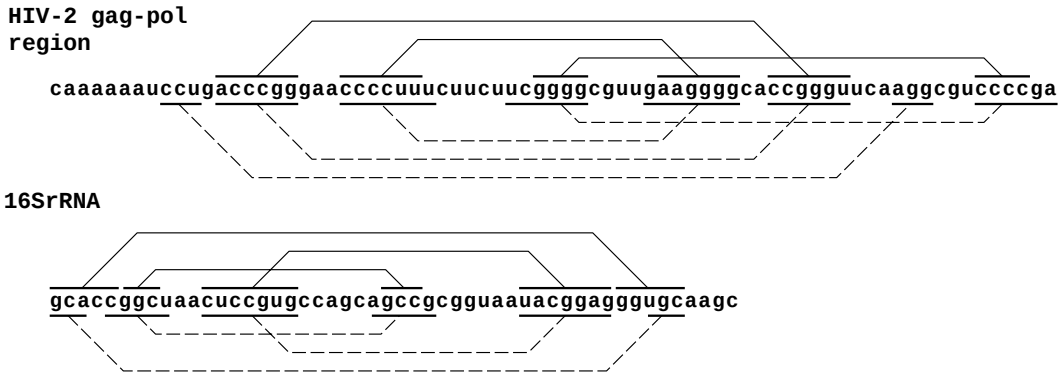

Figure 4: Experimental results on RNA secondary structure prediction, where solid lines indicate native structures and dashed lines indicate predicted structures.

of a stacked region is set to 3, and nested pseudoknots (such as Fig. 2(B)) are inhibited. However, owing to these modifications, we could obtain structures more similar to the native structures than in Ref. [8].

The CPU times for HIV-2 gag-pol region and 16SrRNA were 46.0 sec. and 10.0 sec., respectively. Since the time complexity of the proposed algorithm is $O(n^4)$, these CPU times are reasonable.

Although we can not yet succeed to improve the worst case time complexity, it is possible to reduce the average CPU time considerably. The following method seems the most feasible: first we enumerate candidates of stacked regions, and then we combine candidates so that combined regions do not violate the constraints. Although similar approaches have been already employed [1, 2], combining candidates can also be done using a DP-based procedure (even if pseudoknots are included) as in Section 2.

5 Concluding Remarks

In this paper, we have shown that DP is still useful for the RNA secondary structure prediction with pseudoknots. As mentioned in Introduction, the most important contribution of this paper is that it corrects the previous misunderstanding that pseudoknots can hardly be handled by a DP-based approach. As shown in Section 3, DP algorithms can cover most types of pseudoknots. However, the time complexity increases to $O(n^5)$ or more if complex pseudoknots must be handled. Therefore, improvements on the time complexities should be done. Since significant improvements have been done on Zuker and Stiegler's DP algorithm without pseudoknots [3], it seems possible to make significant improvements on the proposed algorithms.

Another important problem is that no establish definition of a pseudoknot is known. Although proposed DP algorithms can cover wide class of pseudoknots, we can not see from previously published references what extent we should cover. Thus, it would be helpful if a formal definition of a pseudoknot is discussed and given by biologists.

In computational experiments on RNA secondary structure prediction, we did not consider energies for loop regions because no established energy function was known for loop regions in pseudoknots. Thus, to develop and establish such energy functions is important for obtaining accurate results.

Acknowledgments

This work was supported in part by a Grant-in-Aid "Genome Science" for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] Abrahams, J. P., Berg, M., Batenburg, E. and Pleij, C., "Prediction of RNA secondary structure, including pseudoknotting by computer simulation," *Nucleic Acids Research*, 18:3035–3044, 1990.

[2] Akiyama, Y. and Kanehisa, M., "NeuroFold: an RNA secondary structure prediction system using a Hopfield neural network," *Proc. Genome Informatics Workshop III* (in Japanese), 199–202, 1992.

[3] Galil, Z. and Park, K., "Dynamic programming with convexity, concavity and sparsity," *Theoretical Computer Science*, 92:49–76, 1992.

[4] Garey, M. R. and Johnson, D. S., *Computers and Intractability: A Guide to the Theory of NP-completeness*, Freeman, NY, 1979.

[5] Gribskov, M. and Devereux, J. (eds), *Sequence Analysis Primer*, Sptockton Press, NY, 1991.

[6] Maier, R., The complexity of some problems on subsequences and supersequences, *J. ACM*, 25:322-336, 1978.

[7] Turner, D. H., Sugimoto N. and Freier, S. M., "RNA structure prediction," *Ann. Rev. Biophys, Biophys. Chem.*, 17:167–192, 1988.

[8] Uemura, Y., Hasegawa, A., Kobayashi, S. and Yokomori, T., "Grammatically modeling and predicting RNA secondary structures," *Proc. Genome Informatics Workshop VI*, 67–76, 1995.

[9] Zuker, M. and Sankoff, D., "RNA secondary structures and their prediction," *Bulletin of Mathematical Biology*, 46:591–621, 1984.

[10] Zuker, M. and Stiegler, P., "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic Acids Research*, 9:133–148, 1981.

Appendix

In the following, an RNA secondary structure with extended pseudoknots means a structure having any planar graph structure under the condition that each residue can not be connected with multiple residues.

Theorem: RNA secondary structure prediction with extended pseudoknots is NP-hard, where we assume that an arbitrary energy function determined by base-pairs and two base-pairs can be used.

(Proof Sketch) We use a reduction from LCS (Longest Common Subsequence), which is known to be NP-complete. For details about LCS, refer Refs. [4] and [6].

Let $L = \{s_1, s_2, \cdots, s_m\}$ be an instance of LCS over $\Sigma = \{\mathrm{A,U}\}$, where $|s_1| = |s_2| = \cdots = |s_m| = n$. It is easy to see that LCS remains NP-complete for such a case. Here, we consider a decision problem version of LCS that asks whether or not the length of LCS is greater than or equal to k.

For the simplicity, we explain the reduction using an example: $L = \{\mathrm{AAUA, AUAA, AUUA}\}$, $k = 3$ (i.e., LCS is AUA).

Let X^i denotes $\overbrace{XX\cdots X}^{i}$. Each shaded part in the figure is called a *bridge*, which consists of A^i and U^i, where different i's are used for different bridges, and i is sufficiently large ($i \gg n$). Then, we

construct an instance as in Fig. 5. Note that, in this reduction, substrings "CCC" and "GGG" are constructed in order to extract a subsequence with length $k = 3$, and two complementary substrings are constructed from each element in L. For example, substrings "cUcUcAcUc" and "gAgAgUgAg" correspond to "AAUA", where c and g denote C^{10} and G^{10} respectively.

Energy function is defined as follows: $f(\mathrm{A}, \mathrm{U}) = f(\mathrm{A}, \mathrm{C}) = f(\mathrm{U}, \mathrm{C}) = f(\mathrm{A}, \mathrm{G}) = f(\mathrm{U}, \mathrm{G}) = -1.0$, $f(\mathrm{AA}, \mathrm{UU}) = -100.0, f(\mathrm{CC}, \mathrm{GG}) = -100.0$, otherwise $f(X, Y) = 0.0$.

Then, bridges of the same length must make base-pairs in an optimal secondary structure. Moreover, C^i and G^i must make base-pairs too. From these, we can see that LCS has a solution if and only if each residue is connected with another residue in an optimal secondary structure as in Fig. 5.

Since the reduction can be done in polynomial time, the theorem holds. □

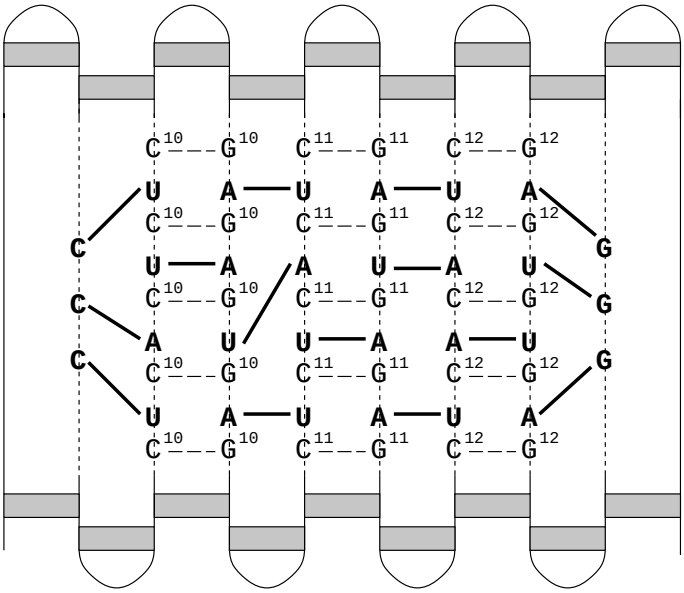

Figure 5: An optimal secondary structure with extended pseudoknots corresponding to LCS instance: {AAUA, AUAA, AUUA}, $k = 3$.