Article No. jmbi.1998.2436 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 285, 2053–2068

JMB

A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots

Elena Rivas and Sean R. Eddy*

Department of Genetics
Washington University
St. Louis, MO 63110, USA

We describe a dynamic programming algorithm for predicting optimal RNA secondary structure, including pseudoknots. The algorithm has a worst case complexity of $\mathcal{O}(N^6)$ in time and $\mathcal{O}(N^4)$ in storage. The description of the algorithm is complex, which led us to adopt a useful graphical representation (Feynman diagrams) borrowed from quantum field theory. We present an implementation of the algorithm that generates the optimal minimum energy structure for a single RNA sequence, using standard RNA folding thermodynamic parameters augmented by a few parameters describing the thermodynamic stability of pseudoknots. We demonstrate the properties of the algorithm by using it to predict structures for several small pseudoknotted and non-pseudoknotted RNAs. Although the time and memory demands of the algorithm are steep, we believe this is the first algorithm to be able to fold optimal (minimum energy) pseudoknotted RNAs with the accepted RNA thermodynamic model.

Keywords: RNA; secondary structure prediction; pseudoknots; dynamic programming; thermodynamic stability

*Corresponding author

Introduction

Many RNAs fold into structures that are important for regulatory, catalytic, or structural roles in the cell. An RNA's structure is dominated by base-pairing interactions, most of which are Watson-Crick pairs between complementary bases. The base-paired structure of an RNA is called its secondary structure. Because Watson-Crick pairs are such a stereotyped and relatively simple interaction, accurate RNA secondary structure prediction appears to be an achievable goal.

A rather reliable approach for RNA structure prediction is comparative sequence analysis, in which covarying residues (e.g. compensatory mutations) are identified in a multiple sequence alignment of RNAs with similar structures, but different sequences (Woese & Pace, 1993). Covarying residues, particularly pairs which covary to maintain Watson-Crick complementarity, are indicative of conserved base-pairing interactions. The accepted secondary structures of most structural and catalytic RNAs were generated by comparative sequence analysis.

If one has only a single RNA sequence (or a small family of RNAs with little sequence diversity), comparative sequence analysis cannot be applied. Here, the best current approaches are energy minimization algorithms (Schuster *et al.*, 1997). While not as accurate as comparative sequence analysis, these algorithms have still proven to be useful research tools. Thermodynamic parameters are available for predicting the ΔG of a given RNA structure (Freier *et al.*, 1986; Serra & Turner, 1995). The Zuker algorithm, implemented in the programs MFOLD (Zuker, 1989a) and ViennaRNA (Schuster *et al.*, 1994), is an efficient dynamic programming algorithm for identifying the globally minimal energy structure for a sequence, as defined by such a thermodynamic model (Zuker & Stiegler, 1981; Zuker & Sankoff, 1984; Sankoff, 1985). The Zuker algorithm requires $O(N^3)$ time and $O(N^2)$ space for a sequence of length N, and so is reasonably efficient and practical even for large RNA sequences. The Zuker dynamic programming algorithm was subsequently extended to allow experimental constraints, and to sample suboptimal folds (Zuker, 1989b). McCaskill's variant of the Zuker algorithm calculates probabilities (confidence estimates) for particular base-pairs (McCaskill, 1990).

Abbreviations used: MWM, maximum weighted matching; NP, non-deterministic polynomial; IS, irreducible surfaces.

E-mail address of the corresponding author: eddy@genetics.wustl.edu

0022-2836/99/052053–16 $30.00/0

One well-known limitation of the Zuker algorithm is that it is incapable of predicting so-called RNA pseudoknots. This is the problem that we address here.

The thermodynamic model for non-pseudoknotted RNA secondary structure includes some stereotypical interactions, such as stacked base-paired stems, hairpins, bulges, internal loops, and multiloops. Formally, non-pseudoknotted structures obey a ''nesting'' convention: that for any two base-pairs i, j and k, l (where $i < j$, $k < l$ and $i < k$), either $i < k < l < i$ or $i < j < k < l$. It is precisely this nesting convention that the Zuker dynamic programming algorithm relies upon to recursively calculate the minimal energy structure on progressively longer subsequences. An RNA pseudoknot is defined as a structure containing base-pairs which violate the nesting convention. An example of a simple pseudoknot is shown in Figure 1.

RNA pseudoknots are functionally important in several known RNAs (ten Dam *et al*., 1992). For example, by comparative analysis, RNA pseudoknots are conserved in ribosomal RNAs, the catalytic core of group I introns, and RNase P RNAs. Plausible pseudoknotted structures have been proposed (Pleij *et al*., 1985), and recently confirmed (Kolk *et al*., 1998) for the 3′ end of several plant viral RNAs, where pseudoknots are apparently used to mimic tRNA structure. *In vitro* RNA evolution (SELEX) experiments have yielded families of RNA structures which appear to share a common pseudoknotted structure, such as RNA ligands selected to bind HIV-1 reverse transcriptase (Tuerk *et al*., 1992).

Most methods for RNA folding which are capable of folding pseudoknots adopt heuristic search procedures and sacrifice optimality. Examples of these approaches include quasi-Monte Carlo searches (Abrahams *et al*., 1990) and genetic algorithms (Gultyaev *et al*., 1995; van Batenburg *et al*., 1995). These approaches are inherently unable to guarantee that they have found the ''best'' structure given the thermodynamic model, and consequently unable to say how far a given prediction is from optimality.

A different approach to pseudoknot prediction based on the maximum weighted matching (MWM) algorithm (Edmonds, 1965; Gabow, 1976) was introduced by Cary & Stormo (1995) and Tabaska *et al*. (1998). Using the MWM algorithm, an optimal structure is found, even in the presence of complicated knotted interactions, in $O(N^3)$ time and $O(N^2)$ space. However, MWM currently seems best suited to folding sequences for which a previous multiple alignment exists, so that scores may be assigned to possible base-pairs by comparative analysis. It is not clear to us that the MWM algorithm will be amenable to folding single sequences using the relatively complicated Turner thermodynamic model. However, we believe that this was the first work that indicated that optimal RNA pseudoknot predictions can be made with polynomial time algorithms. It had been widely believed, but never proven, that pseudoknot prediction would be an NP problem (NP, non-deterministic polynomial; e.g. only solvable by heuristic or brute force approaches).

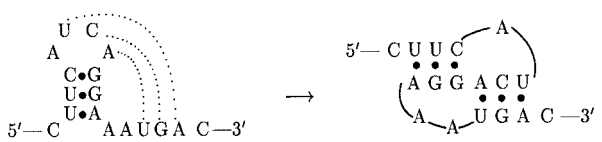

Figure 1. A simple pseudoknot. In a pseudoknot, nucleotides inside a hairpin loop pair with nucleotides outside the stem-loop.

Here, we describe a dynamic programming algorithm which finds optimal pseudoknotted RNA structures. We describe the algorithm using a diagrammatic representation borrowed from quantum field theory (Feynman diagrams). We implement a version of the algorithm that finds minimal energy RNA structures using the standard RNA secondary structure thermodynamic model (Freier *et al*., 1986, Serra & Turner, 1995), augmented by a few pseudoknot-specific parameters that are not yet available in the standard folding parameters, and by coaxial stacking energies (Walter *et al*., 1994) for both pseudoknotted and non-pseudoknotted structures. We demonstrate the properties of the algorithm by testing it on several small RNA structures, including both structures thought to contain pseudoknots and structures thought not to contain pseudoknots.

Algorithm

Here, we will introduce a diagrammatic way of representing RNA folding algorithms. We will start by describing the Nussinov algorithm (Nussinov *et al*., 1978), and the Zuker-Sankoff algorithm (Zuker & Sankoff, 1984; Sankoff, 1985) in the context of this representation. Later on we will extend the diagrammatic representation to include pseudoknots and coaxial stackings. The Nussinov and Zuker-Sankoff algorithms can be implemented without the diagrammatic representation, but this representation is essential to manage the complexity introduced by pseudoknots.

Preliminaries

From here on, unless otherwise stated, a flat continuous line will represent the backbone of an RNA sequence with its 5′-end placed in the left-hand side of the segment. N will represent the length (in number of nucleotides) of the RNA.

Secondary interactions will be represented by wavy lines connecting the two interacting positions in the backbone chain, while the backbone itself always remains flat. No more than two bases are allowed to interact at once. This representation does not provide insight about real (three-dimensional) spatial arrangements, but is very convenient for algorithmic purposes. When necessary

for clarification, single-stranded regions will be marked by dots, but when unambiguous, dots will be omitted for simplicity. Using this representation (Figure 2), we can describe hairpins, bulges, stems, internal loops and multiloops as simple nested structures; a pseudoknot, on the other hand, corresponds with a non-nested structure.

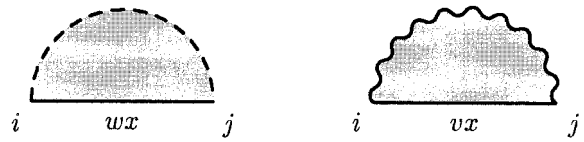

Figure 3. The wx and vx matrices.

Diagrammatic representation of nested algorithms

In order to describe a nested algorithm we need to introduce two triangular $N \times N$ matrices, to be called vx and wx. These matrices are defined in the following way: $vx(i, j)$ is the score of the best folding between positions i and j, provided that i and j are paired to each other; whereas $wx(i, j)$ is the score of the best folding between positions i and j regardless of whether i and j pair to each other or not. These matrices are graphically represented in the form indicated in Figure 3. The filled inner space indicates that we do not know how many interactions (if any) occur for the nucleotides inside, in contrast with a blank inner space which indicates that the fragment inside is known to be unpaired. The wavy line in vx indicates that i and j are definitely paired, and similarly the discontinuous line in wx indicates that the relation between i and j is unknown. Also part of our convention is that for a given fragment, nucleotide i is at the 5′-end, and nucleotide j is at the 3′-end, so that $i \leqslant j$.

The purpose of the nested dynamic programming algorithm is to fill the vx and wx matrices with appropriate numerical weights by means of some sort of recursive calculation.

The recursion for vx includes contributions due to: hairpins, bulges, internal loops, and multiloops. But what is special about hairpins, bulges, internal loops, and multiloops in this diagrammatic representation? To answer this question we have to introduce two more definitions: surfaces and irreducible surfaces (IS).

Roughly speaking a surface is any alternating sequence of continuous and wavy lines that closes on itself. An irreducible surface is a surface such that if one of the H-bonds (or secondary interactions) is broken, there is no other surface contained inside, that is, an IS cannot be ''reduced'' to any other surface. The order $\mathcal{O}$, of an IS is given by the number of wavy lines (secondary interactions), which is equal to the number of continuous-line intervals. It is easy to see that hairpin loops constitute the IS of $\mathcal{O}(1)$; stems, bulges and internal loops are all the IS of $\mathcal{O}(2)$, and what are referred to in the literature as ''multiloops'' are the IS of $\mathcal{O} > 2$.

For nested configurations, our ISs are equivalent to the ''k-loops'' defined by Sankoff (1985); however, the ISs are more general and also include non-nested structures. A technical report about irreducible surfaces is available from http://www.genetics.wustl.edu/eddy/publications/.

The actual recursion for vx is given in Figure 4, and can be expressed as:

$$vx(i,j) = \text{optimal} \begin{cases} EIS^1(i,\ j) \\ EIS^2(i,j:k,l) + vx(k,l) \\ EIS^3(i,j:k,l:m,n) + vx(k,l) + vx(m,n) \\ EIS^4(i,j:k,l:m,n:r,s) + vx(k,l) \\ \quad +vx(m,n) + vx(r,s) \\ \mathcal{O}(5) \end{cases} \tag{1}$$

$$[\forall k,l,m,n,r,s, \quad i \leqslant k \leqslant l \leqslant m \leqslant n \leqslant r \leqslant s \leqslant j]$$

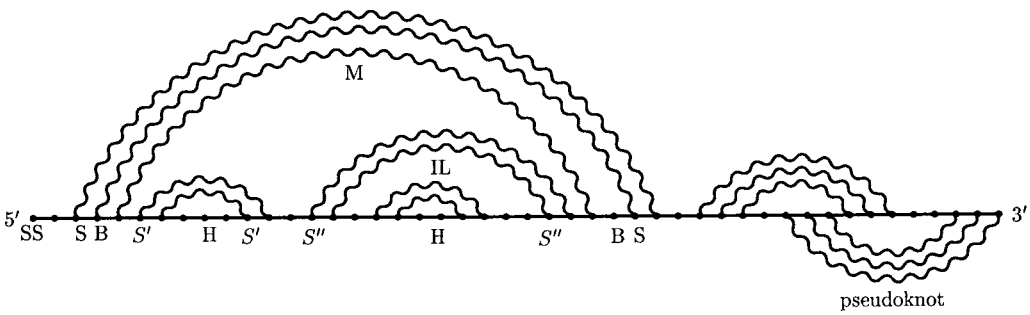

Figure 2. Diagrammatic representation of the most relevant RNA secondary structures, including a pseudoknot. The nucleotides of the sequence are represented by dots. Single-stranded regions (SS) are not involved in any secondary structure. A hairpin (H) is a sequence of unpaired bases bounded by one base-pair. Stems (S), bulges (B) and internal loops (IL) are all nested structures bounded by two base-pairs. In a stem, the two base-pairs are contiguous at both ends. In a bulge, the two base-pairs are contiguous only at one end. In an internal loop, the two base-pairs are not contiguous at all. Multiloops (M) refer to any structure bounded by three or more base-pairs. Any non-nested structure is referred to as a pseudoknot.

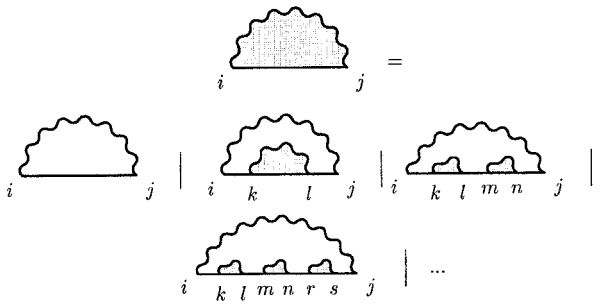

Figure 4. General recursion for vx in the nested algorithm.

Each line gives the formal score of one of the diagrams in Figure 4. The diagram on the left is calculated as the score of the best diagram on the right.

The initialization conditions are:

$$vx(i,i) = +\infty, \quad \forall i \quad 1 \leqslant i \leqslant N \tag{2}$$

The recursion (1) for vx is an expansion in ISs of successively higher order.

Here $EIS^n(i_1, j_1 : i_2, j_2 : \ldots : i_n, j_n)$. represents the scoring function for an IS of order n, in which i_k is paired to j_k. This general algorithm is quite impractical, because an IS^γ which has order γ, $\mathcal{O}(\gamma)$, adds a complexity of $\mathcal{O}(N^{2(\gamma-1)})$ to the calculation. (An IS^γ requires us to search through 2γ independent segments in the entire sequence of N nucleotides. To make it useful, we have to truncate the expansion in ISs at some order in the recursion for vx in Figure 4. The symbol $\mathcal{O}(\gamma)$ indicates the order of IS^γ at which we truncate the recursion.

These recursions are equivalent to those proposed by Sankoff (1985) in theorem 2. Notice also that in defining the recursive algorithm we have not yet had to specify anything about the particular manner in which the contribution from different ISs are calculated in order to obtain the most optimal folding.

The simplest truncation is to stop at order zero, $\mathcal{O}(0)$. In this approximation none of the ISs (hairpin, bulge, internal loop etc.) are given any specialized scores. We only have to provide a specific score for a base-pair, B. The recursion for vx then simplifies to Figure 5, and can be cast into the form:

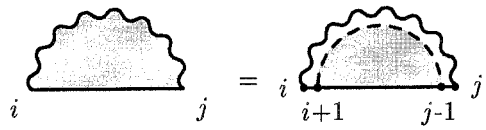

Figure 5. Recursion for vx truncated at $\mathcal{O}(0)$.

$$vx(i,j) = B + wx_I(i+1, j-1) \tag{3}$$

If we set $B = +1$, then we have the Nussinov algorithm (Nussinov *et al.*, 1978). The matrix wx_I is similar to wx defined before, with the specification of appearing inside a base-pair. This simple algorithm calculates the folding with the maximum number of base-pairs.

The next order of complexity we explore corresponds with a truncation at ISs of $\mathcal{O}(2)$. Hairpin loops, bulges, stems, and internal loops are treated with precision by the scoring functions EIS^1 and EIS^2. The rest of ISs, collected under the name of multiloops, which are much less frequent than the previous, are described in an approximate form. The diagrams of this approximation are given in Figure 6, and correspond with:

$$vx(i,j) = \text{optimal}\begin{cases} EIS^1(i,j) &] \quad IS^1 \\ EIS^2(i,j:k,l) + vx(k,l) &] \quad IS^2 \\ P_I + M + wx_I(i+1,k) + wx_I(k+1,j-1) &] \quad \text{multiloop} \end{cases} \tag{4}$$

$$[\forall k,l \quad i \leqslant k \leqslant l \leqslant j]$$

M stands for the score for generating a multiloop. The Turner thermodynamic rules also penalize an amount for each closing pair in a multiloop. By starting a multiloop we are specifying already one of its closing pairs; this closing-pair score is represented here by P_I.

The recursion relations used to fill the wx matrix include: single-stranded nucleotides, external pairs, and bifurcations. The actual recursion is easier to understand by looking at the diagrams involved (given in Figure 7) and the recursion can be expressed as:

$$wx(i,j) = \text{optimal}\begin{cases} P + vx(i,j) &] \quad \text{paired} \\ Q + wx(i+1,j) \\ Q + wx(i,j-1) &] \quad \text{single-stranded} \\ wx(i,k) + wx(k+1,j) \quad [\forall k, \quad i \leqslant k \leqslant j]. &] \quad \text{bifurcation} \end{cases} \tag{5}$$

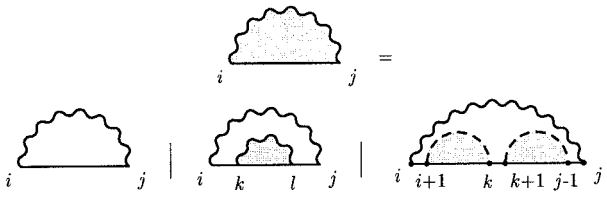

Figure 6. Recursion for vx truncated at $\mathcal{O}(2)$.

With the initialization condition:

$$wx(i,i) = 0, \quad \forall i \quad 1 \leqslant i \leqslant N \tag{6}$$

Note that we have two independent matrices, wx and wx_I, which have structurally identical recursions, but completely different interpretations. The matrix wx_I, used to truncate the recursion for vx in equation (4), is used exclusively for diagrams which will be incorporated into multiloops, whereas wx is only used when there are no external base-pairs. Therefore, the parameters controlling these two recursions will, in general, have very different values because they have very different meanings. Q_I is the penalty for an unpaired nucleotide in a multiloop, and P_I is the penalty for a closing base-pair (e.g. per stem) in a multiloop. On the other hand, Q represents the score for a single-stranded nucleotide, and P represents the score for an external base-pair. In Turner's thermodynamic rules both Q and P are approximated by zero.

Note also that the recursions for wx and wx_I always remain the same, independent of the order of irreducible surface to which the recursion for vx has been truncated.

This is the nested algorithm described by Sankoff (1985) in theorem 3, and is the approximation that MFOLD (Zuker & Stiegler, 1981) and ViennaRNA (Schuster *et al.*, 1994) implement. Higher orders of specificity of the general algorithm are possible, but are certainly more time consuming, and they have not been explored so far. One reason for this relative lack of development is that there is little information about the energetic properties of multiloops. The generalized nested algorithm provides a way to unify the currently available dynamic algorithms for RNA folding. At a given order, the error of the approximation is given by the difference between the assigned score to multiloops and the precise score that one of those higher-order ISs deserves.

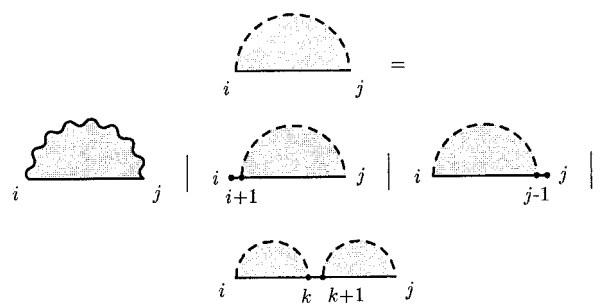

Figure 7. Recursion for wx in the nested algorithm.

Description of the pseudoknot algorithm

Pseudoknots are non-nested configurations and clearly cannot be described with just the wx and vx matrices we introduced in the previous section. The key point of the pseudoknot algorithm is the use of gap matrices in addition to the wx and vx matrices. Looking at the graphical representation of one of the simplest pseudoknots, Figure 8, we can see that we could describe such a configuration by putting together two gap matrices with complementary holes.

The pseudoknot dynamic programming algorithm uses one-hole or gap matrices (Figure 9) as a generalization of the wx and vx matrices (cf. Table 1). Let us define $whx(i, j:k, l)$ as the graph that describes the best folding that connects segments $[i, k]$ with $[l, j]$, $i \leqslant k \leqslant l \leqslant j$, such that the relation between i and j and k and l is undetermined. Similarly, we define $vhx(i, j:k, l)$ as the graph that describes the best folding that connects segments $[i, k]$ with $[l, j]$, $i \leqslant k \leqslant l \leqslant j$, such that i and j are base-paired and k and l are also base-paired. For completeness we have to introduce also

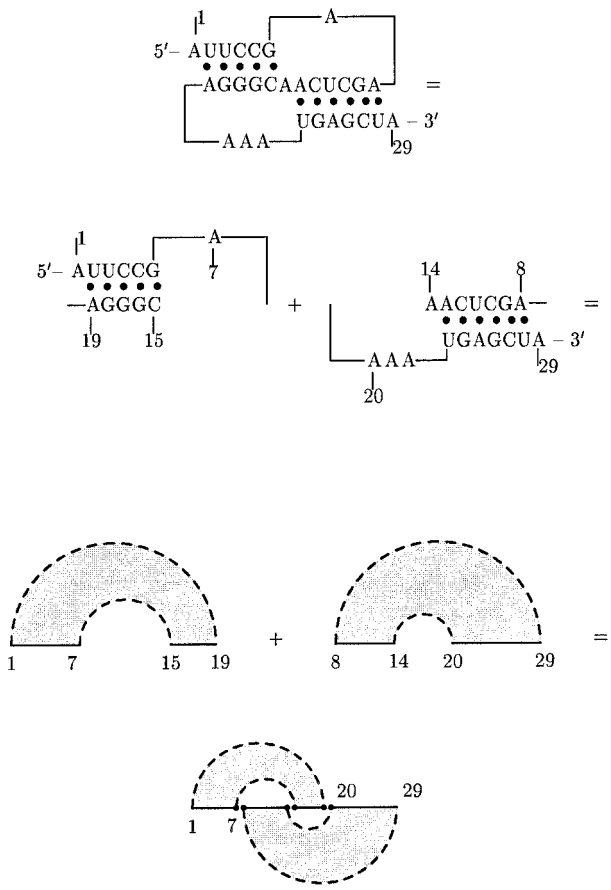

Figure 8. Construction of a simple pseudoknot using two gap matrices.

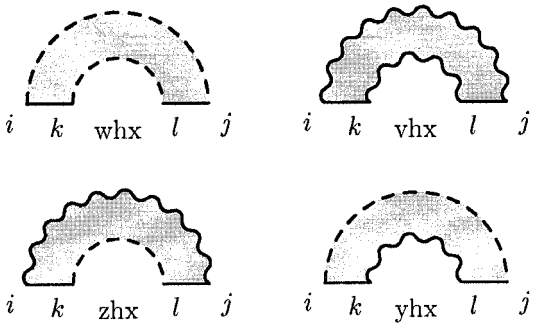

Figure 9. Representation of the gap matrices used in the algorithm for pseudoknots.

Table 1. Specifications of the matrices used in the pseudoknot algorithm

Matrix ($i \leqslant k \leqslant l \leqslant j$)	Relationship i, j	Relationship k, l
$vx(i, j)$	Paired	-
$wx(i, j)$	Undetermined	-
$vhx(i, j:k, l)$	Paired	Paired
$zhx(i, j:k, l)$	Paired	Undetermined
$yhx(i, j:k, l)$	Undetermined	Paired
$whx(i, j:k, l)$	Undetermined	Undetermined

matrix $yhx(i, j:k, l)$ in which k and l are paired, but the relation between i and j is undetermined, and its counterpart $zhx(i, j:k, l)$ in which i and j are paired, but the relation between k and l is undetermined.

The non-gap matrices wx, vx are contained as a particular case of the gap matrices. When there is no hole, $k = l - 1$, then by construction:

$$whx(i, j:k, k+1) = wx(i, j) \tag{7}$$

$$zhx(i, j:k, k+1) = vx(i, j) \quad \forall k, \quad i \leqslant k \leqslant j$$

We have introduced the gap matrices as the building blocks of the algorithm, but how do we establish a consistent and complete recursion relation? Here is where the analogy between the gap matrices and the Feynman diagrams of quantum field theory was of great help (Bjorken & Drell 1965).†

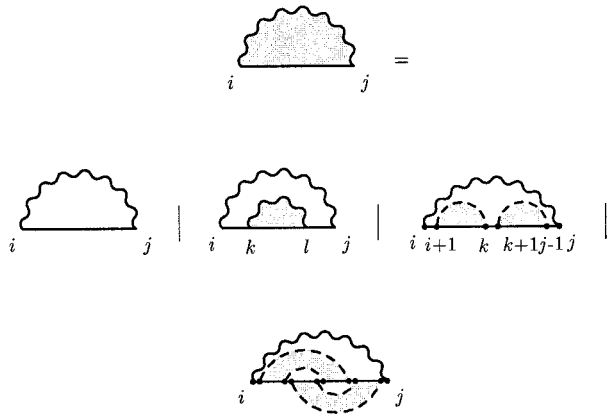

Figure 10. Recursion for vx in the pseudoknot algorithm truncated at $\mathcal{O}(whx + whx + whx)$. (Contiguous nucleotides are represented with explicit dots.)

Let us start with the generalization of the recursions for vx and wx in the presence of gap matrices. A non-gap matrix can be obtained by combining two gap matrices together, therefore the recursions for vx and wx add one more diagram with two gap matrices to recursions (4) and (5). Again the diagrammatic representation (Figures 10 and 11) is more helpful than words in explaining the recursions. (When possible, individual bases are labeled in the diagrams. Otherwise contiguous nucleotides are depicted with dots.) Note that the new term introduced in both recursions involves two gap matrices. In fact, the recursion is an expansion in the number of gap matrices.

The recursion for the non-gap matrix vx is given by (cf. Figure 10):

$$vx(i, j) = \text{optimal} \begin{cases} E\text{IS}^1(i, j) &] \quad \text{IS(1)} \\ E\text{IS}^2(i, j:k, l) + vx(k, l) &] \quad \text{IS(2)} \\ P_I + M + wx_I(i+1, k) + wx_I(k+1, j-1) &] \quad \text{nested multiloop} \\ \widetilde{P}_I + \tilde{M} + G_{wI} + whx(i+1, r:k, l) & \\ \quad + whx(K+1, j-1:l-1, r+1) &] \quad \text{non-nested multiloop} \end{cases} \tag{8}$$

$$[\forall i, k, l, r, j \quad i \leqslant k \leqslant l \leqslant r \leqslant j]$$

† More precisely, the analogy is more cleanly expressed in terms of Schwinger-Dyson diagrams which in QFT are used to represent full interacting vertices and propagators recursively in terms of elementary interactions.

The additional parameters for pseudoknots are: $\widetilde{P}_I$, the score for a pair in a non-nested multiloop; $\tilde{M}$, a generic score for generating a non-nested multiloop; and G_{wI} the score for generating an internal pseudoknot.

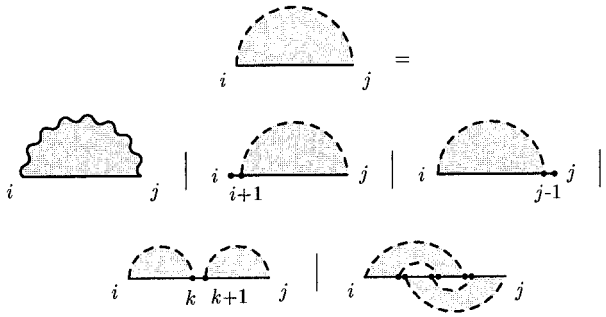

Figure 11. Recursion for wx in the pseudoknot algorithm truncated at $\mathcal{O}(whx + whx + whx)$. (Contiguous nucleotides are represented with explicit dots.)

a solvable configuration can be decomposed into a sum of gap matrices according to the rules provided by our recursions. A non-solvable configuration is one that requires diagrammatic topologies that involve three or more gap matrices. That is, a non-solvable configuration requires us to go to a higher orders in the expansion of the pseudoknot algorithm.

Our algorithm can solve ''overlapping pseudoknots'' (defined as those pseudoknots for which a planar representation does not require crossing lines) such as ABAB, ABACBC, ABACBDCD, etc. The algorithm can also find some ''non-planar pseudoknots'' (pseudoknots for which a planar representation requires crossing lines) such as

Similarly for wx (cf. Figure 11):

$$wx(i,j) = \text{optimal}\begin{cases} P + vx(i,j) & \left.\vphantom{\int}\right] \text{ paired} \\ Q + wx(i+1,j) & \left.\vphantom{\int}\right] \text{ single-stranded} \\ Q + wx(i,j-i) & \\ wx(i,k) + wx(k+1,j) & \left.\vphantom{\int}\right] \text{ nested bifurcation} \\ G_w + whx(i,r:k,l) & \left.\vphantom{\int}\right] \text{ non-nested bifurcation} \\ \quad + whx(k+1,j:l-1,r+1) & \end{cases} \tag{9}$$

Where G_w denotes the score for introducing a pseudoknot. We should also remember that the algorithm uses two different wx matrices depending on whether the subset $i \ldots j$ is free-standing (wx) or appears inside a multiloop (in which case we use wx_I). The two recursions are identical apart from having different parameter values as described in Table 2.

Practical considerations make us truncate the expansion at this stage; we will not include diagrams that require three or more gap matrices. This statement should not mislead one into thinking that we cannot deal with complicated pseudoknots. We define a solvable configuration as one that can be parsed by our algorithm. That is, ABCABC (the topology present in *Escherichia coli* α mRNA; Gluick *et al.*, 1994), and others. However, the algorithm is not able to solve all possible knotted configurations, as for instance a parallel β-sheet protein interaction ABCADBECD (see Figure 12 for some details.) For a given configuration we can decide unambiguously whether it is solvable or not by parsing it according to the model. However, we still lack a systematic *a priori* characterization of the class of configurations that this algorithm can solve.

Note that two approximations are involved in the algorithm. Apart from that just mentioned (truncating the infinite expansion in gap matrices to make the algorithm polynomial), we also use

Table 2. The parameters for which there is thermodynamic information provided by the Turner group

Symbol	Scoring parameter for	Value (kcal/mol)
EIS^1	Hairpin loops	Varies
EIS^2	Bulges, stems and internal loops	Varies
C	Coaxial stacking	Varies
P	External pair	0
Q	Single-stranded base	0
R, L	Base dangling off an external pair	Dangle + Q
P_I	Pair in a nested multiloop	0.1
Q_I	Non-paired base inside multiloop	0.4
R_I, L_I	Base dangling off a multiloop pair	Dangle + Q_I
M	Nested multiloop	4.6

These parameters are identical with those used in MFOLD (http://www.ibc.wustl.edu/~zuker/rna).

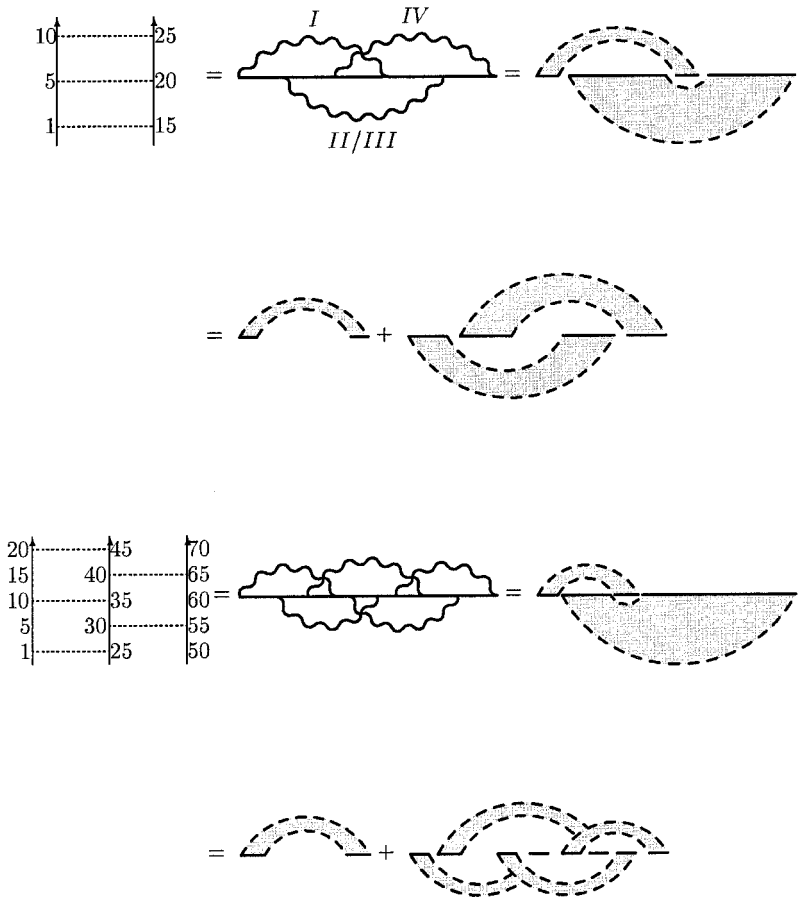

Figure 12. Top, the non-planar pseudoknot (ABCABC) presented in α mRNA and how to build it with gap matrices. The Roman numbers correspond with the numbering of stems introduced by Gluick *et al.* (1994). Bottom, an example of a pseudoknot that the algorithm cannot handle; interlaced interactions as seen in proteins in parallel β-sheet (ABCADBECDE). The assembly of this interaction using gap matrices would require us to use four gap matrices at once which is not allowed by the approximation at hand.

the approximation previously introduced for the nested algorithm (that ISs of $\mathcal{O}. > 2$ or multiloops are described in some approximated form). Despite these limitations, this truncated pseudoknot algorithm seems to be adequate for the currently known pseudoknots in RNA folding.

The algorithm is not complete until we provide the full recursive expressions to calculate the gap matrices. For a given gap matrix, we have to consider all the different ways that its diagram can be assembled using one or two matrices at a time. (Again, Feynman diagrams are of great use here.) The full description of those diagrams is quite involved and the many technical details will not add to the clarity of this exposition. In order to give the reader a feeling for the kind of topologies the pseudoknot algorithm allows, we provide in the Appendix a simplified version of the recursions for the gap matrices in which coaxial stacking or dangles are excluded (see below).

Coaxial stacking and dangles

It is quite frequent in RNA folding to create a more stable configuration when two independent configurations stack coaxially. This occurs, for instance, when two hairpin loops with their respective stems are contiguous. Then one of them can fall on top of the other, creating a more stable configuration than when the two hairpins just coexist without interaction of any kind.

The algorithm implements coaxial energies for both nested and non-nested structures. We adopt the coaxial energies provided by Walter *et al.* (1994) for coaxial stacking of nested structures. For coaxial stacking of non-nested structures we multiply these previous energies by an estimated (*ad hoc*) weighting parameter $g < 1$.

Using our diagrammatic representation it is possible to be systematic in describing the possible coaxial stacking that can occur. In the general recursion one has to look for contiguous nucleotides, and allow them to be explicitly paired (but not to each other). This is best understood with an example. Consider the recursion for wx in Figure 11, in particular the bifurcation diagram:

$$wx(i,j) \longrightarrow wx(i,k) + wx(k+1,j), \quad \forall k, i \leqslant k \leqslant j \tag{10}$$

In order to allow for the possibility of coaxial stacking, this bifurcation diagram has to be complemented with another one in which the nucleotides of the bifurcation are base-paired:

$$wx(i,j) \longrightarrow vx(i,k) + vx(k+1,j) + C(k,i:k+1,j), \quad \forall k, i \leqslant k \leqslant j \tag{11}$$

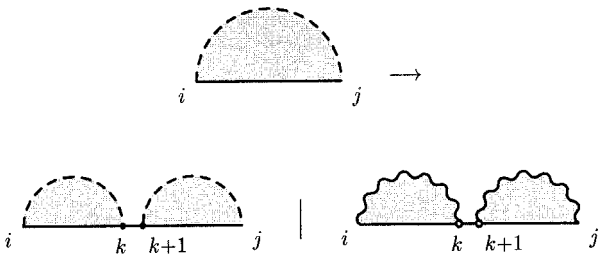

Figure 13. Coaxial stacking. Two base-pair interactions are energetically more favorable when they are contiguous with each other. Here, we indicate how to complement the regular bifurcation diagram in wx (left) with an additional diagram (right) to take into account such a coaxial stacking configuration. The coaxial scoring function depends on both base-pairs. (Coaxial diagrams can be recognized by the empty dots representing the contiguous coaxially stacking nucleotides.)

This new diagram (Figure 13) indicates that if nucleotides k and $k+1$ are paired to nucleotides i and j, respectively, that configuration is specially favored by an amount $C(k, i:k+1, j)$ (presumably negative in energy units) because both sub-structures, $vx(i, k)$ and $vx(k+1, j)$, will stack onto each other.

Similarly, unpaired nucleotides contiguous to a paired base seem to have a different thermodynamic contribution than other unpaired nucleotides. In order to take this fact into account, we have to systematically add dangle diagrams to the various recursions.

For instance, the dangle diagrams that we have to add for the recursion of the wx matrix are given in Figure 14, and correspond with the following terms in the recursion for wx:

$$wx(i,j) \longrightarrow \begin{cases} L^{i}_{i+1,\,j} + vx(i+1,j) \\ R^{j}_{i,\,j-1} + vx(i,j-1) \\ L^{i}_{i+1,\,j-1} + R^{j}_{i+1,\,j-1} + vx(i+1,j-1) \end{cases} \tag{12}$$

The dangle scoring functions, (R, L), depend both on the dangling bases and the contiguous base-pair. These dangle energies have been well characterized by the Turner group (Freier *et al*., 1986). Dangling bases can also appear inside multiloop diagrams. Notice also that the coaxial diagram in equation (11) really corresponds with four new diagrams because once we allow pairing, dangling bases also have to be considered, so the full nearest-neighbour interaction is taken into account.

† Since the implementation of the pseudoknot algorithm, the Turner group has produced a new complete and more accurate list of parameters (Mathews *et al*., 1998) which we have not yet implemented.

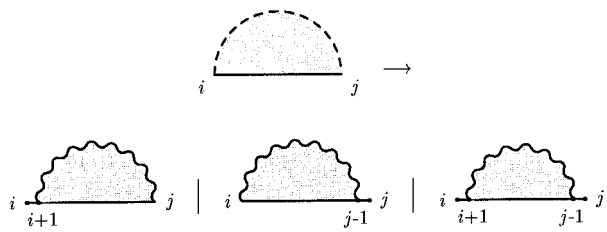

Figure 14. Dangles. The figures represent three types of dangling bases that can contribute to the ungapped matrix wx. The dangle score function associated with each of these diagrams depends both on the dangling bases and the base-pair adjacent to them.

Our pseudoknot algorithm implements both dangles and coaxial stackings. MFOLD currently implements only dangles, but will soon implement coaxials (Mathews *et al*., 1998). For purposes of clarity we will not explicitly show any of the additional diagrams to be included in the recursions to take care of coaxial stackings and dangles.

Minimum-energy implementation: thermodynamic parameters

We have implemented the pseudoknot algorithm using thermodynamic parameters in order to fill the scoring matrices, both gapped and ungapped. For the relevant nested structures, hairpin loops, bulges, stems, internal loops and multiloops, we have used the same set of energies as used in MFOLD.† Free energies for coaxial stacking, C, were those obtained by Walter *et al*. (1994). Table 2 provides a list of the parameters used for nested conformations.

For the non-nested configurations, there is not much thermodynamic information available (Wyatt *et al*., 1990; Gluick *et al*., 1994). This is not an untypical situation; there is very little thermodynamic information available for regular multiloops, let alone for pseudoknots. We had to tune by hand the parameters related to pseudoknots. For some non-nested structures we multiplied the nested parameters by an estimated weighting parameter $g < 1$. It would be very useful, in order to improve the accuracy of this thermodynamic implementation of the pseudoknot algorithm, to have more accurate, experimentally, based determinations of these parameters. Table 3 provides a list of the parameters we used for pseudoknot-related conformations.

Results

The main purpose of this work is to present an algorithm that solves optimal pseudoknotted RNA structures by dynamic programming. RNA structure prediction of single sequences with the nested algorithm already involves some approximation and inaccuracy (Zuker, 1995; Huynen *et al*., 1997).

Table 3. The new thermodynamic parameters specific for pseudoknot configurations which we had to estimate

Symbol	Scoring parameter for	Value (kcal/mol)
$\widetilde{EIS^2}$	IS^2 in a gap matrix	$EIS^2 \times g(0.83)$
$\tilde{C}$	Coaxial stacking in pseudoknots	$C \times g$
$\tilde{P}$	Pair in a pseudoknot	0.1
$\tilde{P}_I$	Pair in a non-nested multiloop	$\tilde{P} \times g$
$\tilde{Q}$	Non-paired base in pseudoknot	0.2
$\tilde{R}$, $\tilde{L}$	Base dangling off a pseudoknot pair	dangle $\times g + \tilde{Q}$
$\tilde{M}$	Non-nested multiloop	8.43
G_w	Generating a new pseudoknot	7.0
G_{wi}	Generating a pseudoknot in a multiloop	13.0
G_{wh}	Overlapping pseudoknots	6.0

We expect this inaccuracy to increase in our case, since the algorithm now allows a much larger configuration space. Therefore, our limited objective here is to show that on a few small RNAs that are thought to conserve pseudoknots, our program (a minimal-energy implementation of the pseudoknot algorithm using a thermodynamic model) will actually find the pseudoknots; and for a few small RNAs that do not conserve pseudoknots, our program finds results similar to MFOLD, and does not introduce spurious pseudoknots.

tRNAs

Almost all transfer RNAs share a common cloverleaf structure. We have tested the algorithm on a group of 25 tRNAs selected at random from the Sprinzl tRNA database (Steinberg *et al*., 1993). The program finds no spurious pseudoknot for any of the tested sequences. All but one (DT5090) of the tRNAs fold into a cloverleaf configuration. Of the 24 cloverleaf foldings, 15 are completely consistent with their proposed structures (that is, each helical region has at least three base-pairs in common with its proposed folding). The remaining nine cloverleaf foldings misplace one (six sequences) or two (three sequences) of the helical regions. On the other hand, MFOLD's lowest energy prediction for the same set of tRNA sequences includes only 19 cloverleaf foldings, of which 14 are completely consistent with their proposed structures. Performance for our program is, therefore, at least comparable with MFOLD; the inaccuracies found are the result of the approximations in the thermodynamic model, not a problem with the pseudoknot algorithm *per se*. The relevant result in relation to the pseudoknot algorithm is that its implementation predicts no spurious pseudoknots for tRNAs.

One should not think of this result as a trivial one, because when knots are allowed, the configuration space available becomes much larger than the observed class of conformations. This problem is particularly relevant for ''maximum-pairing-like'' algorithms, such as the MWM algorithm presented by Cary & Stormo (1995) or a Nussinov implementation of our pseudoknot algorithm (Figure 5). In both cases, the result is almost universal pairing because there is enough freedom to be able to coordinate any position with another one in the sequence.

Another important aspect of tRNA folding is coaxial energies. Most tRNAs gain stability by stacking coaxially two of the hairpin loops, and the third one with the acceptor stem. This aspect of tRNA folding is very important and in some cases crucial to determine the right structure. There are situations like tRNA DA0260 in which MFOLD does not assign the lowest energy to the correct structure (the MFOLD 3.0 prediction for DA0260 misses the acceptor stem, and has a free energy of −22.0 kcal/mol). Our algorithm, on the other hand, implements coaxial energies; as a result, the cloverleaf configuration becomes the most stable folding for tRNA DA0260 ($\Delta G = -24.3$ kcal/mol). The implementation of coaxial energies explains why we found more cloverleaf structures for tRNAs than MFOLD does.

HIV-1-RT-ligand RNA pseudoknots

High-affinity ligands of the reverse transcriptase of HIV-1 isolated by a SELEX procedure by Tuerk *et al*. (1992) seem to have a pseudoknot consensus secondary structure. These oligonucleotides have between 34 and 47 bases, and fold into a simple pseudoknot. Of a total of 63 SELEX-selected pseudoknotted sequences available from Tuerk *et al*. (1992), we found 54 foldings that agreed exactly with the structures derived by comparative analysis ($\Delta G = -9$ kcal/mol for sequence pattern I (3-2)). As expected, MFOLD predicts only one of the two stems ($\Delta G = -7.5$ kcal/mol for the same sequence).

Viral RNAs

Some virus RNA genomes (such as turnip yellow mosaic virus, TYMV; Guiley *et al*., 1979) present a tRNA-like structure at their 3′-end that includes a pseudoknot in the aminoacyl acceptor arm very close to the 3′-end (Kolk *et al*., 1998; Pleij

et al., 1985; Dumas *et al.*, 1987). Our program correctly predicts the TYMV tRNA-like structure with its pseudoknot for the last 86 bases at the 3′-end with $\Delta G = -30.4$ kcal/mol (the MFOLD 3.0 prediction for TYMV has a free energy of $\Delta G = -28.9$ kcal/mol). The tRNA-like 3′ terminal structure is conserved among tymoviruses, and also for the tobacco mosaic virus cowpea strain, another valine acceptor. Of the seven valine-acceptor tRNA-like structures proposed to date (Van Belkum *et al.*, 1987), we reproduce six of them, except for kennedya yellow mosaic virus.

Another interesting pseudoknot appears in the last 189 bases of the 3′ terminus of the tobacco mosaic virus (TMV; Van Belkum *et al.*, 1985). TMV also has a tRNA-like pseudoknot structure at the end, but it may have additional upstream pseudoknots, up to a total of five, forming a long quasi-continuous helix. We folded the upstream and downstream regions separately in a piece of 84 nucleotides (the folding requires 47 minutes and 9.8 Mb) and 105 nucleotides (the folding requires 235 minutes and 22.5 Mb), respectively. Our program predicts the 105 nucleotides downstream region exactly with $\Delta G = -32.5$ kcal/mol. For the 84 nucleotides upstream region we find four of the five helical regions with $\Delta G = -19.0$ kcal/mol.

Finally we have considered the recently crystallized ribozymes of the hepatitis delta virus (HDV; Ferré-D'Amaré *et al.*, 1998). Our program predicts correctly the structure of the 91 nt antigenomic HDV ribozyme ($\Delta G = -36.7$ kcal/mol). Our program also predicts the pseudoknot present in the 87 nt genomic ribozyme ($\Delta G = -43.9$ kcal/mol; in this case the prediction misses the short two-stem hairpin between positions 17-30).

Discussion

Here, we present an algorithm able to predict pseudoknots by dynamic programming. This algorithm demonstrates that using certain approximations consistent with the accepted Turner thermodynamic model, the prediction of pseudoknotted structures is a problem of polynomial complexity (although admittedly high). Having an optimal dynamic programming algorithm will enable extending other dynamic programming based methods that rigorously explore the conformational space for RNA folding (McCaskill, 1990; Bonhoeffer *et al.*, 1993) to pseudoknotted structures.

Apart from the usefulness of the algorithm in predicting pseudoknots, we also include coaxial energies (when two stems stack coaxially), a very common feature of RNA folding. We expect MFOLD will also include coaxial energies in the near future (Mathews *et al.*, 1998).

Our algorithm is presented in the context of a general framework in which a generic folding is expressed in terms of its elementary secondary interactions (which we have identified as the irreducible surfaces). This is a further generalization of the results reported by Sankoff (1985). The calculation of an optimal folding becomes an expansion in ISs of increasingly higher order. Our formalization incorporates all current dynamic programming RNA folding algorithms in addition to our pseudoknot algorithm. It also establishes the limitations of each approximation by determining at which order the expansion is truncated.

As for the thermodynamic implementation presented here, one of our major problems is the almost complete lack of thermodynamic information about pseudoknot configurations. The thermodynamic algorithm is also sensitive to the accuracy of the existing thermodynamic parameters. We expect to improve this aspect by implementing the more complete set of parameters provided by the Turner group (Mathews *et al.*, 1998).

The principal drawback is the time and memory constraints imposed by the computational complexity of the algorithm. At this early stage, we cannot analyze sequences much larger than 130-140 bases. For now, the program is adequate for folding small RNAs. A 100 nt RNA takes about four hours and 22.5 Mb to fold on an SGI R10K Origin200.

Due to practical limitations, at a given point in the recursion we only allow the incorporation of two gap matrices. However, since each of those gap matrices can in turn be assembled by other two of those matrices, it implies that the algorithm includes in its configuration space a large variety of knotted motifs. The limitations of this truncation appeared when we considered applying this approach to describe pairwise residue interactions in protein folding. A parallel β-sheet configuration in protein structure provides an example of a complicated knotted folding that cannot be handled by the pseudoknot algorithm presented here. However, all known RNA pseudoknots can be handled by the algorithm, which makes the approximation useful enough for RNA secondary structure.

Although we implemented the algorithm for energy minimization, extending MFOLD to pseudoknotted structures, the algorithm is not limited to energy minimization. Our algorithm can be converted into a probabilistic model for pseudoknot-containing RNA folding. Probabilistic models of RNA second structure based on ''stochastic context free grammar'' (SCFG) formalisms (Eddy *et al.*, 1994; Sakakibara *et al.*, 1994; Lefebvre, 1996) have been introduced both for RNA single-sequence folding and for RNA structural alignment and structural similarity searches. The Inside and CYK dynamic programming algorithms used for SCFG-based structural alignment are fundamentally similar to the Zuker algorithm (Durbin *et al.*, 1998), and have consequently also been unable to deal with pseudoknots. Heuristic approaches to applying SCFG-like structural alignment models to pseudoknots have been introduced (Brown, 1996;

Notredame *et al.*, 1997), and the maximum weighted matching algorithm has been applied to find optimal alignments (Tabaska & Stormo, 1997). An SCFG-like probabilistic version of our pseudoknot algorithm could be designed to obtain optimal structural alignment of pseudoknot-containing RNAs.

Methods

The algorithm was implemented in ANSI C on a Silicon Graphics Origin200. The algorithm has a theoretical worst-case complexity of $\mathcal{O}(N^6)$ in time and $\mathcal{O}(N^4)$ in storage. At its present stage, the program is empirically observed to run $\mathcal{O}(N^{6.8})$ in time and $\mathcal{O}(N^{3.8})$ in memory. For instance, a tRNA of 75 nt takes 20 minutes and uses 6.6 Mb of memory. The 3′-end of tobacco mosaic virus has 105 nucleotides and takes 235 minutes and uses 22.5 Mb. The program empirically scales above the theoretical complexity in time of the algorithm. This effect seems to have to do with the way the machine allocates memory for larger RNAs. The software and parameter sets are available by request from E. Rivas (elena@genetics.wustl.edu). A technical report giving the full algorithm is available from http://www.genetics.wustl.edu/eddy/publications/.

Acknowledgments

This work was supported by NIH grant HG01363 and by a gift from Eli Lilly. E.R. acknowledges the support of a fellowship by the Sloan Foundation. The idea for the algorithm came from a discussion with Gary Stormo at a meeting at the Aspen Center for Physics. Tim Hubbard suggested parallel β-strands in proteins as an example of a set of pairwise interactions that the algorithm cannot handle. We wish to thank the anonymous reviewers for very useful comments.

References

Abrahams, J. P., van der Berg, M., van Batenburg, E. & Pleij, C. W. A. (1990). Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.* **18**, 3035-3044.

Bjorken, J. D. & Drell, S. D. (1965). *Relativistic Quantum Fields*, McGraw-Hill, New York, NY.

Bonhoeffer, S., McCaskill, J. S., Stadler, P. F. & Schuster, P. (1993). Statistics of RNA secondary structure. *Eur. Biophys. J. (EHU),* **22**, 13-24.

Brown, M. (1996). RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. *Pacific Symposium on Biocomputing 1996*.

Cary, R. B. & Stormo, G. D. (1995). Graph-theoretic approach to RNA modeling using comparative data. In *ISMB-95* (Rawling, C., *et al.*, eds), pp. 75-80, AAAI Press.

Dumas, P., Moras, D., Florentz, C., Giegé, R., Verlaan, P., van Belkum, A. & Pleij, C. W. A. (1987). 3-D graphics modeling of the tRNA-like 3′ end of turnip yellow mosaic virus RNA: structural and functional implications. *J. Biomol. Struct. Dynam.* **4**, 707-728.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. J. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge UK.

Eddy, S. R. & Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucl. Acids Res.* **22**, 2079-2088.

Edmonds, J. (1965). Maximum matching and polyhedron with 0, 1-vertices. *J. Res. Nat. Bur. Stand.* **69B**, 125-130.

Ferré-D'Amaré, A. R., Zhou, K. & Doudna, J. A. (1998). Crystal structure of a hepatitis delta virus ribozyme. *Nature,* **395**, 567-574.

Freier, S., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T. & Turner, D. H. (1986). Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl Acad. Sci. USA,* **83**, 9373-9377.

Gabow, H. N. (1976). An efficient implementation of Edmonds' algorithm for maximum matching on graphs. *J. Asc. Com. Mach.* **23**, 221-234.

Gluick, T. C. & Draper, D. E. (1994). Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.* **241**, 246-262.

Guilley, H., Jonard, G., Kukla, B. & Richards, K. E. (1979). Sequence of 1000 nucleotides at the 3′ end of tobacco mosaic virus RNA. *Nucl. Acids Res.* **6**, 1287-1308.

Gultyaev, A. P., van Batenburg, F. H. & Pleij, C. W. A. (1995). The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.* **250**, 37-51.

Huynen, M., Gutell, R. & Konings, D. (1997). Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.* **267**, 1104-1112.

Kolk, M. H., van der Graff, M., Wijmenga, S. S., Pleij, C. W. A., Heus, H. A. & Hilbers, C. W. (1998). NMR structure of a classical pseudoknot: interplay of single- and double-stranded RNA. *Science,* **280**, 434-438.

Lefebvre, F. (1996). A grammar-based unification of several alignments and folding algorithms. *ISMB-96* (Rawlings, C., *et al.*, eds), pp. 143-154, AAAI Press.

Mathews, D. H., Andre, T. C., Kim, J., Turner, D. H. & Zuker, M. (1998). An updated recursive algorithm for RNA secondary structure prediction with improved free energy parameters. In *Molecular Modeling of Nucleic Acids* (Leontis, N. B. & SantaLucia, J., Jr, eds), American Chemical Society.

McCaskill, J. S. (1990). The equilibrium partition function and base pair bindings probabilities for RNA secondary structure. *Biopolymers,* **29**, 1105-1119.

Notredame, C., O'Brien, E. A. & Higgins, D. G. (1997). RAGA: RNA sequence alignment by genetic algorithm. *Nucl. Acids Res.* **25**, 4570-4580.

Nussinov, R., Pieczenik, G., Griggs, J. R. & Kleitman, D. J. (1978). Algorithms for loop matchings. *SIAM J. Appl. Math.* **35**, 68-82.

Pleij, C. W., Rietveld, K. & Bosch, L. (1985). A new principle of RNA folding based on pseudoknotting. *Nucl. Acids Res.* **13**, 1717-1731.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C. & Haussler, D.

(1994). Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.* **22**, 5112-5120.
Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* **45**, 810-825.
Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. (1994). From sequences to shapes and back: a case study in RNA secondary structure. *Proc. Roy. Soc. ser. B,* **255**, 279-284.
Schuster, P., Fontana, W., Stadler, P. F. & Renner, A. (1997). RNA structures and folding: from conventional to new issues in structure predictions. *Curr. Opin. Struct. Biol.* **7**, 229-235.
Serra, M. J. & Turner, D. H. (1995). Predicting the thermodynamic properties of RNA. *Methods Enzymol.* **259**, 242-261.
Steinberg, S., Misch, A. & Sprinzl, M. (1993). Compilation of RNA sequences and sequences of tRNA genes. *Nucl. Acids Res.* **21**, 3011-3015.
Tabaska, J. E. & Stormo, G. D. (1997). Automated alignment of RNA sequences to pseudoknotted structures. *ISMB-97,* **5**, 311-318.
Tabaska, J. E., Cary, R. B., Gabow, H. N. & Stormo, G. D. (1998). An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics,* **8**, 691-699.
ten Dam, E., Pleij, K. & Draper, D. (1992). Structural and functional aspects of RNA pseudoknots. *Biochemistry,* **31**, 11665-11676.
Tuerk, C., MacDougal, S. & Gold, L. (1992). RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc. Natl Acad. Sci. USA,* **89**, 6988-6992.
van Batenburg, F. H. D., Gultyaev, A. P. & Pleij, C. W. A. (1995). An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **174**, 269-280.
Van Belkum, A., Abrahams, J. P., Pleij, C. W. A. & Bosch, L. (1985). Five pseudoknots are present at the 204 nucleotides long 3′ non coding region of tobacco mosaic virus RNA. *Nucl. Acids Res.* **13**, 7673-7686.
Van Belkum, A., Bingkun, J., Pleij, C. W. A. & Bosch, L. (1987). Structural similarities among valine-accepting tRNA-like structures in tymoviral RNAs and elongator tRNAs. *Biochemistry,* **26**, 1144-1151.
Walter, A., Turner, D., Kim, J., Lyttle, M., Müller, P., Mathews, D. & Zuker, M. (1994). Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc. Natl Acad. Sci. USA,* **91**, 9218-9222.
Woese, C. R. & Pace, N. R. (1993). Probing RNA structure, function, and history by comparative analysis. *The RNA World* (Gesteland, R. F. & Atkins, J. F., eds), pp. 91-117, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Wyatt, J. R., Puglisi, J. D. & Tinoco, I., Jr (1990). RNA pseudoknots: stability and loop size requirements. *J. Mol Biol.* **214**, 455-470.
Zuker, M. (1989a). Computer prediction of RNA structure. *Methods Enzymol.* **180**, 262-288.
Zuker, M. (1989b). On finding all suboptimal foldings of an RNA molecule. *Science,* **244**, 48-52.
Zuker, M. (1995). ''Well-determined'' regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucl. Acids Res.* **23**, 2791-2798.
Zuker, M. & Sankoff, D. (1984). RNA secondary structure and their prediction. *Bull. Math. Biol.* **46**, 591-621.
Zuker, M. & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* **9**, 133-148.

Appendix: Recursions for the Gap Matrices in the Pseudoknot Algorithm

Here we provide simplified recursion relations for the gap matrices used in the pseudoknot algorithm, without including dangling and coaxial diagrams. (As before, contiguous nucleotides are given explicit dots in the diagrams.)

The recursion for the *vhx* matrix in the pseudoknot algorithm is given by (Figure A1):

$$vhx(i,j:k,l) = \text{optimal} \begin{cases} \widetilde{EIS}^2(i,j:k,l) \\ \widetilde{EIS}^2(i,j:r,s) + vhx(r,s:k,l) \\ \widetilde{EIS}^2(r,s:k,l) + vhx(i,j:r,s) \\ 2 * \tilde{P} + \tilde{M} + whx(i+1,j-1:k-1,l+1) \end{cases} \quad (1A)$$

$$[\forall i,r,k,l,s,j \quad i \leqslant r \leqslant k \leqslant l \leqslant s \leqslant j]$$

Here $\tilde{P}$ is the score for creating a pair in a pseudoknot, and $\tilde{M}$s; corresponds to the score given to a non-nested multiloop. $\tilde{P}$ and $\tilde{M}$ could be equal to P and M, the score for a pair in a nested structure and the score assigned to nested multiloops respectively, but it does not have to be. Similarly, the score for an irreducible surface of $\mathcal{O}(2)$, $\widetilde{EIS}^2$, could be the same as the score given for nested structures, EIS^2, but again, it does not have to be. We found the best fits by giving them values different from those used for nested foldings (cf. Tables 2 and 3).

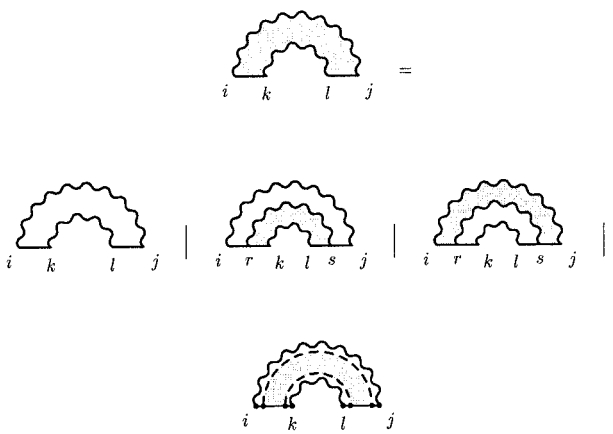

Figure A1. Recursion for the *vhx* matrix.

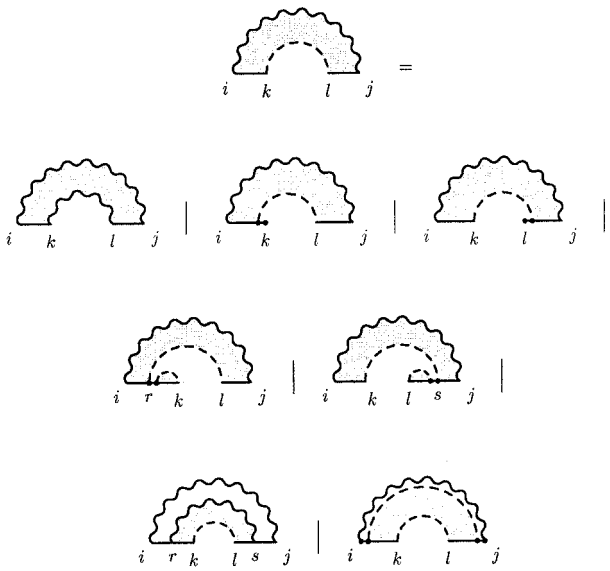

Figure A2. Recursion for the *zhx* matrix.

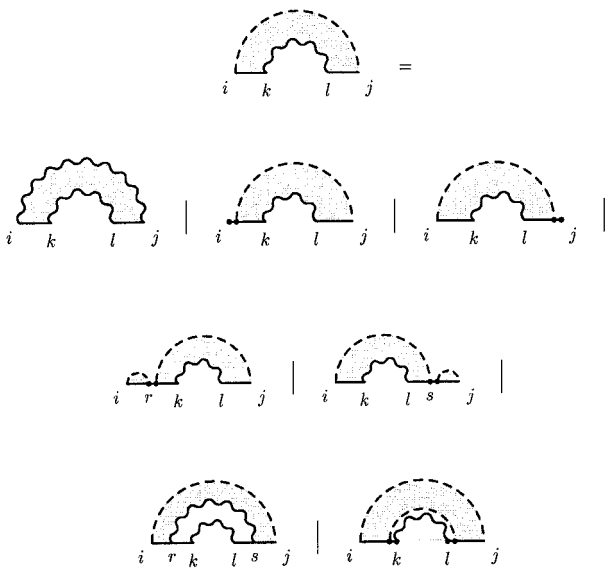

Figure A3. Recursion for the *yhx* matrix.

The recursions for the gap matrices *zhx* and *yhx* in the pseudoknot algorithm are complementary and given by (cf. Figures A2 and A3):

$$zhx(i,j:k,l) = \text{optimal}\begin{cases} \left.\begin{array}{l} \widetilde{P} + vhx(i,j:k,l) \end{array}\right] & \text{paired} \\ \left.\begin{array}{l} \widetilde{Q} + zhx(i,j:k-1,l) \\ \widetilde{Q} + zhx(i,j:k,l+1) \end{array}\right] & \text{single-stranded} \\ \left.\begin{array}{l} zhx(i,j:r,l) + wx_I(r+1,k) \\ zhx(i,j:k,s) + wx_I(l,s-1) \\ \widetilde{EIS}^{\,2}(i,j:r,s) + zhx(r,s:k,l) \\ \widetilde{P} + \widetilde{M} + whx(i+1,j-1:k,l) \end{array}\right] & \text{nested bifurcations} \end{cases} \tag{A2}$$

$$yhx(i,j:k,l) = \text{optimal}\begin{cases} \left.\begin{array}{l} \widetilde{P} + vhx(i,j:k,l) \end{array}\right] & \text{paired} \\ \left.\begin{array}{l} \widetilde{Q} + yhx(i+1,j:k,l) \\ \widetilde{Q} + yhx(i,j-1:k,l) \end{array}\right] & \text{single-stranded} \\ \left.\begin{array}{l} wx_I(i,r) + yhx(r+1,j:k,l) \\ yhx(i,s:k,l) + wx_I(s+1,j) \\ yhx(i,j:r,s) + \widetilde{EIS}^{\,2}(r,s:k,l) \\ \widetilde{P} + \widetilde{M} + whx(i,j:k-1,l+1) \end{array}\right] & \text{nested bifurcations} \end{cases} \tag{A3}$$

$$[\forall i,r,k,l,s,j \quad i \le r \le k \le l \le s \le j]$$

Finally, the recursion for the gap matrix *whx* appears in Figure A4, and is given by:

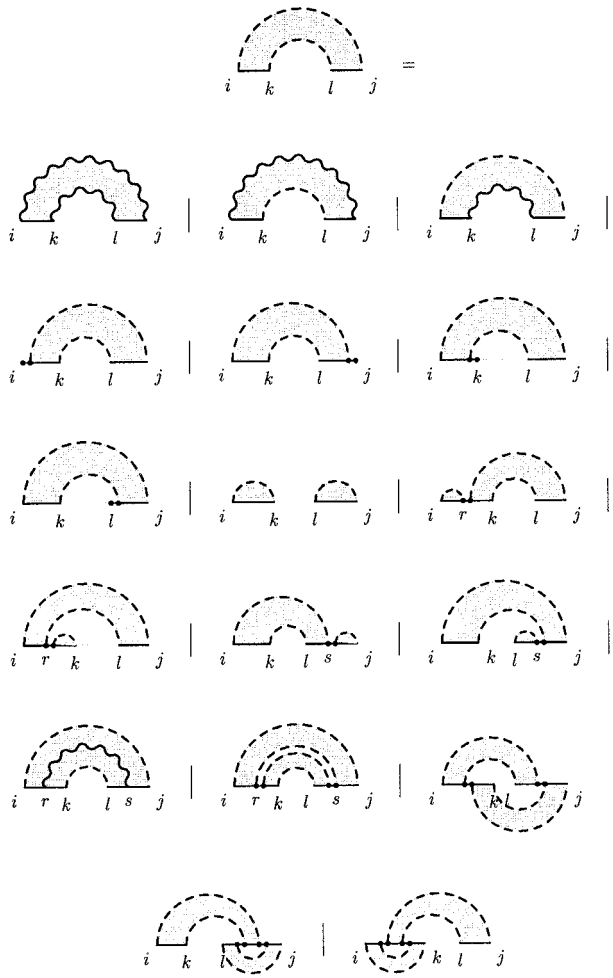

Figure A4. Recursion for the *whx* matrix.

$$
whx(i,j:k,l) = \text{optimal}\begin{cases}
\left.\begin{array}{l} 2*\widetilde{P} + vhx(i,j:k,l) \\ \widetilde{P} + zhx(i,j:k,l) \\ \widetilde{P} + yhx(i,j:k,l) \end{array}\right] & \text{paired} \\
\left.\begin{array}{l} \widetilde{Q} + whx(i+1,j:k,l) \\ \widetilde{Q} + whx(i,j-1:k,l) \\ \widetilde{Q} + whx(i,j:k-1,l) \\ \widetilde{Q} + whx(i,j:k,l+1) \end{array}\right] & \text{single-stranded} \\
\left.\begin{array}{l} wx_I(i,k) + wx_I(l,j) \\ wx_I(i,r) + whx(r+1,j:k,l) \\ whx(i,j:r,l) + wx_I(r+1,k) \\ whx(i,s:k,l) + wx_I(s+1,j) \\ whx(i,j:k,s+1) + wx_I(l,s) \\ yhx(i,j:r,s) + zhx(r,s:k,l) \\ \widetilde{M} + whx(i,j:r,s) + whx(r+1,s-1:k,l) \end{array}\right] & \text{nested bifurcations} \\
\left.\begin{array}{l} G_{wh} + whx(i,s:r,l) + whx(r+1,j:k,s+1) \\ G_{wh} + whx(i,s':k,s) + whx(l,j:s-1,s'+1) \\ G_{wh} + whx(r,j:r',l) + whx(i,k:r-1,r'+1) \end{array}\right] & \text{non-nested bifurcations}
\end{cases} \tag{A4}
$$

$$[\forall i,r,r',k,l,s,s',j \quad i \le r \le r' \le k \le l \le s \le s' \le j]$$

Here G_{wh} stands for the score given for finding overlapping pseudoknots, that is pseudoknots that appear within already existing pseudoknots.

The initialization conditions are:

$$\begin{aligned} whx(i,j:i,j) &= +\infty \\ vhx(i,j:k,k) &= +\infty \\ yhx(i,j:k,k) &= +\infty \\ whx(i,j:k,k) &= whx(i,j:k,k+1) = wx(i,j) \\ zhx(i,j:k,k) &= zhx(i,j:k,k+1) = vx(i,j) \end{aligned} \qquad \text{(A5)}$$

$$[\forall i,k,j \quad 1 \leqslant i \leqslant k \leqslant j \leqslant N]$$

Edited by I. Tinoco

(Received 27 July 1998; received in revised form 20 November 1998; accepted 22 November 1998)

http://www.academicpress.com/jmb

Supplementary material comprising 1 pdf file is available from JMB Online