BIOINFORMATICS

Vol. 14 no. 8 1998
Pages 691–699

An RNA folding method capable of identifying pseudoknots and base triples

Jack E. Tabaska[1,3], Robert B. Cary[1,4], Harold N. Gabow[2] and Gary D. Stormo[1]

[1]Department of Molecular, Cellular and Developmental Biology and [2]Department of Computer Science, University of Colorado, Boulder, CO 80309, USA, [3]Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA and [4]Present address: Los Alamos National Laboratory, Life Sciences Division, MS M888, Los Alamos, NM 87545, USA

Received on March 11, 1998; revised and accepted on June 18, 1998

Abstract

Motivation: *Recently, we described a Maximum Weighted Matching (MWM) method for RNA structure prediction. The MWM method is capable of detecting pseudoknots and other tertiary base-pairing interactions in a computationally efficient manner (Cary and Stormo, Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology, pp. 75–80, 1995). Here we report on the results of our efforts to improve the MWM method's predictive accuracy, and show how the method can be extended to detect base interactions formerly inaccessible to automated RNA modeling techniques.*

Results: *Improved performance in MWM structure prediction was achieved in two ways. First, new ways of calculating base pair likelihoods have been developed. These allow experimental data and combined statistical and thermodynamic information to be used by the program. Second, accuracy was improved by developing techniques for filtering out spurious base pairs predicted by the MWM program. We also demonstrate here a means by which the MWM folding method may be used to detect the presence of base triples in RNAs.*

Availability: *http://www.cshl.org/mzhanglab/tabaska/jax-page.html*

Contact: *tabaska@cshl.org*

Introduction

RNA molecules play important roles in cellular nucleic acid processing and gene expression, so gaining a deeper understanding of these processes can be aided by determining the tertiary structures of their RNA mediators. Unfortunately, the most popular computer methods for RNA structure prediction (Nussinov and Jacobson, 1980; Zuker, 1989) are limited to secondary structure prediction only. There have recently been described algorithms capable of predicting pseudoknots, but these often rely on heuristics that are not guaranteed to find the molecule's optimal structure (Gultyaev, 1991; Chen *et al*., 1992; van Batenburg *et al*., 1995), and may require the use of massively parallel computers that are not commonly available (Nakaya *et al*., 1995; Shapiro and Wu, 1997). In addition, while the list of RNAs of known crystal or NMR structure is growing (for a review, see Uhlenbeck *et al*., 1997), the acquisition of such structural data is still difficult, and cannot hope to keep pace with the rapid rate of new sequence discovery.

In a previous paper (Cary and Stormo, 1995), we described a new approach to RNA structure prediction, which is based on the Maximum Weighted Matching (MWM) algorithm of Gabow (1973). The results presented there demonstrated that given a set of base pairing likelihood scores, Gabow's algorithm can find an optimal set of base-pairing interactions, including pseudoknots and other tertiary pairs, in polynomial time and memory. However, structures predicted by MWM folding were generally only partially correct, and the procedure as a whole was very sensitive to noisy data.

This paper describes improvements that we have made to the MWM folding technique. New methods for scoring potential base pairs and the addition of noise filtration to the folding process allow the MWM algorithm to produce structures comparable with manually predicted structures. We have also implemented algorithmic enhancements that allow automatic detection of base triples. We believe that these changes will make MWM folding a valuable tool for RNA research.

Algorithms

Maximum Weighted Matching

RNA secondary structure prediction may be thought of as a matching problem: each base in a sequence is to be matched with the base it is paired with in the folded RNA, or left unmatched if the base is unpaired. To solve this problem, researchers collect various kinds of evidence—phylogenetic,

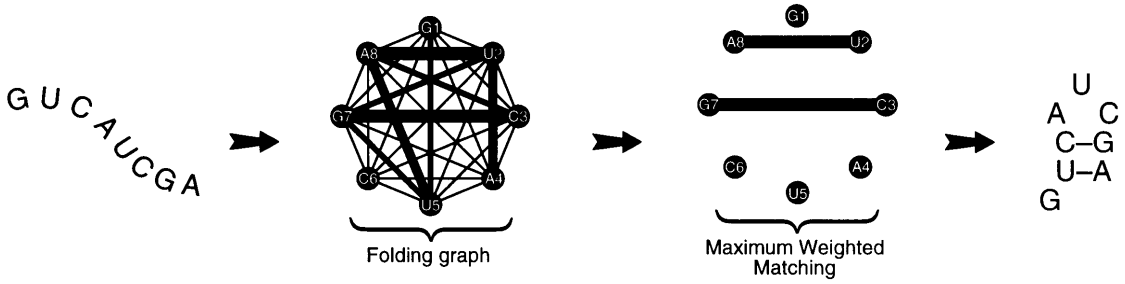

Fig. 1. The MWM RNA folding procedure. Starting with an RNA sequence, a folding graph is formed in which bases are represented by vertices and potential pairing interactions by edges. Edge weights, depicted here by line thickness, quantify the likelihood of a base pair's existence. From the folding graph, a maximum weighted matching is determined. This matching represents the best (i.e. most likely) structure for the RNA.

thermodynamic, NMR, chemical protection, etc.—which help them decide which base pairs are most likely to exist. The best structure is then the one consisting of base pairs for which the most corroborating evidence and least conflicting evidence exist.

Thus stated, RNA structure prediction may be cast directly into a well-known problem in the field of graph theory. We start by constructing a graph in which each vertex corresponds to a base in an RNA sequence, and potential base-pairing interactions are represented by edges that connect the vertices (bases) (Figure 1). We assign to each edge a weight which quantifies the strength of the evidence for that base pair's existence in the folded molecule. This graph is referred to as the folding graph. Every possible structure for the RNA is present in this graph in the form of a matching, i.e. a subgraph in which no vertex is connected to more than one other vertex. To find the best structure for the RNA in question, we then need to find the matching which has the highest total edge weight.

This is known as the MWM problem in graph theory, and it is provably solvable in $O(N^3)$ time and $O(N^2)$ memory (Gabow, 1973; also see below). This is perhaps somewhat surprising, in light of the fact that the foregoing statement of the RNA folding/MWM problem places no restrictions on the planarity (in the structural biological sense) of the base pairs comprising either the folding graph or the final RNA structure. In other words, a computationally efficient algorithm for solving the MWM problem can be used to predict an optimal RNA structure, including pseudoknots and other tertiary base-pairing interactions.

Such algorithms have existed for some time (e.g. Edmonds, 1965), and one of the authors (H.N.G.) has developed what is demonstrably the most efficient algorithm for solving the MWM problem on dense graphs (Gabow, 1973). While a detailed description of Gabow's algorithm is beyond the scope of this paper, a brief overview of the algorithm's operation will facilitate some of the discussion that follows. The MWM is constructed iteratively, by first finding the maximum weight matching consisting of exactly one edge, then finding the maximum weight matching of two edges, and so on, until the algorithm reaches a point beyond which further expansion of the matching cannot increase its total edge weight. Consequently, only those edges with weights greater than zero are included in the MWM (although it is possible to force the algorithm to include zero- or negatively weighted edges; the following section explains why one might want to do this).

Expansion of the intermediate matchings is accomplished through a process called *augmentation*. An augmentation replaces some number of edges, k, in the matching with $k + 1$ edges of higher total weight. If $k = 0$, the augmentation simply adds an edge to the previous matching. More generally, though, an augmentation makes two previously unmatched vertices into matched vertices, and changes the pairing partners of some other previously matched vertices; these latter vertices are said to have been *rematched*. Note that an augmentation never unmatches a previously matched vertex. Therefore, once a vertex becomes matched during the MWM construction, it stays in a matched state, although it may be rematched one or more times.

Prediction of RNA base triples. An RNA structure that contains base triples may be represented by a construct known as a 2-matching: a graph in which no vertex is connected to more than two other vertices (Figure 2a). This is simply a generalization of the above definition of a matching. Remarkably, Gabow's algorithm may be used to find maximum weighted 2-matchings on a graph without severe performance degradation, enabling us to predict structures that contain base triples.

Base triple prediction is accomplished by means of a special graph that is derived from the folding graph (Figure 2b and c). This graph is constructed as follows: First, each vertex, v, of the original folding graph is split into d_v 'internal' vertices, where d_v is the number of edges incident to v. The edges emanating from the original vertex are distributed

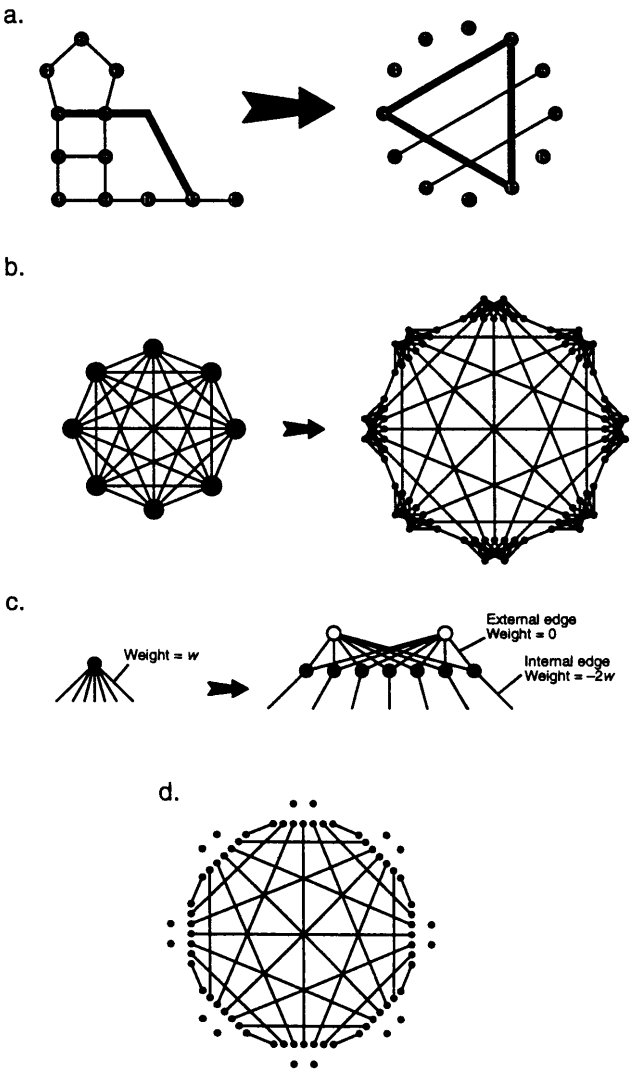

Fig. 2. (**a**) An RNA structure containing a base triple (thick line) and its representation as a 2-matching. Note that the base triple can be recognized in the 2-matching by the mutual pairing of vertices 3, 7 and 11 (cf. Gautheret *et al*., 1995). (**b**) A folding graph and the corresponding derived graph for maximum weighted 2-matching. (**c**) Detail of vertex splitting. The degree of the vertex shown is 7, so it is split into seven internal (filled) vertices and two external (open) vertices. Internal edges are distributed among the internal vertices, and external edges connect the external vertices to each internal vertex. (**d**) Starting point for 2-matching construction, consisting of all of the derived graph's internal edges.

among the internal vertices. The weights of these internal edges are multiplied by –2 (In principle, any negative number will work. We use –2 so that integrality of the given weights is preserved during the execution of the MWM algorithm). Then a pair of 'external' vertices is created for each set of internal vertices. A set of external edges connects each external vertex to its corresponding internal vertices. All of the external edges are assigned a weight of zero.

To find the desired 2-matching on the original folding graph, Gabow's algorithm is used to find the MWM of the derived graph, with the added constraint that each of the internal vertices must be paired with some other vertex in the final matching. This constraint is enforced by providing the algorithm with a starting point for matching construction (Figure 2d) which consists of all of the internal edges of the derived graph. Recalling that the MWM algorithm never unmatches a vertex, all of the internal vertices must be paired in the end, even if this means incorporating zero- or negatively weighted edges in the final matching. Use of this initial matching also speeds up the construction of the matching since it already contains most of the edges that will constitute the derived graph's MWM.

The rationale behind using this derived graph is as follows. Since the algorithm is constrained to keep each internal vertex matched, it can maximize the total edge weight of the final matching only by replacing the most negatively weighted set of internal edges with zero-weighted external edges. Since there are two external vertices adjacent to each set of internal vertices in the derived graph, two of the internal edges incident to each set of internal vertices may be exchanged for external edges. Now, recalling that the weight of each internal edge in the derived graph is –2 times the weight of an edge in the original graph, the best set of edges to exclude from the derived graph's MWM corresponds to the best set of edges to include in the maximum weighted 2-matching on the original graph. Therefore, by noting which internal edges are *not* present in the derived graph's MWM, one finds the edges that form the maximum weighted 2-matching on the original graph.

Performance of the MWM algorithm. The MWM algorithm's memory usage is dominated by arrays containing information on the edges of the input graph (Gabow, 1973). Since the number of edges in a dense graph (such as the folding graph) varies with the square of the number of vertices, this algorithm can fold an RNA containing N bases using $O(N^2)$ memory.

As stated above, the MWM algorithm builds the matching incrementally, up to a maximum of $N/2$ edges. In each iteration, the algorithm essentially makes a pass over the entire graph (Gabow, 1973). One such pass uses time $O(N^2)$. The MWM is therefore constructed in $O(N^3)$ time.

Note that during construction of the derived graph for 2-matching, each edge in the original graph gives rise to one internal edge and four external edges in the derived graph. The total number of edges under consideration is thus multiplied by a constant factor of five. The MWM algorithm therefore remains an $O(N^2)$ memory and $O(N^3)$ time process when performing a 2-matching.

Base pair scoring

The success of MWM RNA folding naturally depends on how potential base pairs are scored. Ideally, one would like to bring together a number of lines of evidence, both theoretical and experimental, and generate the best structure consistent with all of them. MWM folding is well suited to doing just that. Since edge weight calculation is separate from MWM construction, one can simply plug in sets of edge weights from different sources and obtain predicted structures for each set. Multiple sets of edge weights may even be combined into hybrid sets. We discuss here some of the methods we have tried for evaluating base pairs.

Phylogenetic and thermodynamic scores. Previously, we described the use of mutual information (MI) scores to score base pairs. In some ways, MI is ideal for use in RNA tertiary structure prediction: MI is sensitive to the presence of non-canonical base pairs and exotic structural elements such as base triples, parallel helices, and even base stacking (Gutell *et al*., 1992; Gautheret *et al*., 1995). In theory, then, MWM folding should be able to predict the entire structure of an RNA molecule, less backbone interactions, from a good set of MI scores. In practice, however, we have found that the use of MI scores alone tends to make the MWM algorithm miss many intuitively obvious base pairs, especially highly conserved pairs (for which MI≈0). Readers are referred to Cary and Stormo (1995) for a more detailed discussion of MI and its use in MWM folding.

We have also tested MWM folding using sets of thermodynamic scores similar to the energy dot plots described by Jacobson and Zuker (1993). The scores used were obtained by extracting from the MFOLD program (Zuker, 1989) the minimum energy matrix, which tabulates for every possible base pair the minimum energy of a structure containing that pair. Since pseudoknots are often represented in this matrix as alternative foldings of the RNA in question, the MWM algorithm can find non-planar interactions that are missed by MFOLD's dynamic programming algorithm. We have found, however, that these data can take prohibitively long to generate and tend to generate structures of lower quality than other thermodynamics-based scores such as helix plots (see below). MFOLD scores will therefore not be discussed further, except to say that they do perform well when used in combination with MI scores (Tabaska, 1996).

Helix plots. Helix plots are a means of scoring base pairs, which combines phylogenetic and thermodynamic information. Construction of a helix plot starts with an alignment of RNA sequences. For each sequence in the alignment, a square scoring matrix, not unlike a dot plot, is formed, but rather than a binary pair/no pair dot, each cell of the matrix receives a score based on whether the two bases corresponding to the cell can form a stable base pair. One could establish elaborate rules for assigning these scores, but we find a simple three-part scoring scheme works well: a small positive 'good pair' score for Watson–Crick and G-U pairs, a larger negative 'bad pair' score for every other type of base pair, and an even larger negative 'paired gap' score which penalizes potential single-stranded deletions within helices.

After this initial scoring matrix for a sequence has been established, it is scanned for potential helices. During scanning, scores of base pairs comprising helices shorter than some specified minimum length are changed to the bad pair score. Conversely, the scores of base pairs forming sufficiently long helices are increased by adding bonus scores that are proportional to the length of the helix they comprise. These bonus scores may be considered to be a sort of 'stacking energy' derived from the helices. After repeating this process for each RNA sequence in the alignment of interest, the individual scoring matrices are summed, yielding a set of scores for MWM folding.

There are several points worth noting about helix plot scores. First, base pairs that score well are those that are from long, highly conserved helices. Second, non-Watson–Crick/G-U pairs and single-stranded deletions are treated as evidence against the presence of a pairing interaction, as they receive negative scores. One can adjust how sensitive helix plot scores are to these types of negative evidence by adjusting the ratio of the good pair score to the bad pair and paired gap scores. For instance, if the good pair score is set to 1 and the bad pair score to –9, then a pair of alignment positions must contain canonical pairs over 90% of the time to be considered a true pair, paired gap penalties and helix bonuses notwithstanding. Note that since we use a paired gap penalty which is larger than the bad pair penalty, insertions or deletions in one strand of a potential helix are treated as the strongest evidence against the existence of that helix.

Finally, it is obvious that helix plot scores can only be used to predict structures composed of canonically paired, antiparallel double helices—in general, a molecule's secondary structure plus large pseudoknots. Helix plot scores are insensitive to other kinds of tertiary interactions.

Incorporating experimental data. Often, one has a limited amount of experimental structural data on an RNA—e.g. the results of a nuclease protection assay or a mutagenesis experiment—and wants to generate an optimal structure that is consistent with those data. We have developed a simple scheme for incorporating this information into folding graphs so that it will be reflected in MWM-predicted structures.

There are three types of structural information that are routinely obtained by experiments: a base or bases are known to be in a single-stranded region; a base or bases are shown to be paired with known partners; and a base or bases are known to be paired, but with unknown partners. The first of these, known unpaired bases, can be enforced in MWM folding by assigning a weight of zero to every edge incident to a vertex

that represents an unpaired base. Gabow's algorithm will then necessarily leave that vertex unpaired in the final matching.

In the second case, known base pairs, the folding graph is modified so that the edge representing a verified pair receives a positive score, and all other edges incident to the two vertices in question receive weights of zero. The final MWM must then include the experimentally determined pairs.

For the final case, paired bases with unknown partners, modifications to the folding graph are made as follows. If we define P as the set of all bases that are paired with unknown partners, N as the number of vertices in the folding graph, and let W be a number greater than the largest weight in the original folding graph, then the weight of each edge in the folding graph is increased by NpW, where p is 0, 1 or 2, depending on whether the edge connects zero, one or two vertices of P. Weighting the edges in this manner ensures that the total weight of any matching in which all of the vertices in P are matched will be greater than the maximum weight of any matching that leaves any vertex in P unmatched. This forces the MWM algorithm to produce a structure that conforms to the experimental data, while still allowing it to choose the best pairing partner for each base based on the original input weights.

Output filtration

One shortcoming of MWM folding has been that it usually generated structures containing many spurious base pairs, tending to predict pairs for bases that are actually single stranded. A little reflection reveals why this is so: the MWM algorithm seeks to maximize the total edge weight of its final matching, and even base pairs that are very poorly supported by the available data can add some small amount to this sum. To obtain useful structures from MWM folding, then, some form of output filtration must be applied.

In Cary and Stormo (1995), it was suggested that an offset score could be subtracted from the folding graph weights to eliminate low-scoring base pairs. While this can be effective, it is often difficult to select an offset score objectively. A second method is to simply discard all base pairs comprising helices shorter than some minimum length. This is a very effective and objective technique, especially when applied to structures predicted using helix plot scores, as it removes all structural elements which helix plot scores cannot be expected to reflect reliably. When working with MI scores, which are sensitive to unusual base interactions, though, it would be unadvisable simply to discard non-helical pairings as some of them may represent actual tertiary base pairs.

An elegant solution to this problem is the technique of i-matching. During testing of MWM folding, it was observed that once a true base pair is added to the emerging structure, it tends to remain constant through the rest of the folding process. Spurious pairs, however, tend to arise as the result of one or more rematching events. A detailed analysis of this effect has been made elsewhere (Tabaska, 1996), but conceptually it is similar to the P-num scores produced by the MFOLD programs (Zuker, 1989; Zuker and Jacobson, 1995): the more nearly equivalent interactions that a base may form, the less likely any one of them can be expected to exist. Therefore, an objective and very effective method for output filtering is to monitor the intermediate matchings (hence the term i-matching) produced as the MWM algorithm generates a structure, and disregard unstable pairings. Since i-matching makes no prior assumptions about relationships between base pairs, it may be used without fear of systematically discarding tertiary interactions.

Implementation

MWM RNA folding has been implemented in two programs: imatch and bmatch. imatch executes the standard (1-matching) form of Gabow's algorithm, and includes an i-matching monitor. bmatch finds maximum weighted 2-matchings (or higher-order matchings, if desired), and includes an i-matching monitor as well. Both imatch and bmatch are derived from the wmatch program written by Ed Rothburg.

We have written several base pair scoring modules for use in MWM folding:

hlxplot—generates helix plot scores for a set of aligned RNA sequences.

jmixy—calculates pairwise MI scores for an RNA sequence alignment.

exgraf—modifies folding graph files so as to force the MWM folding to conform to experimental data.

makegraf—combines two or more sets of folding graph edge weights into a single composite set, optionally applying scaling factors and an offset score.

All of the above programs are written in ANSI C.

A Perl script for translating the output from imatch into a set of XRNA input files is also available.

Discussion

Figure 3 illustrates the process of folding a tRNA by MWM. An alignment of 556 tRNA gene sequences (Steinberg *et al.*, 1993) was used to generate a combined MI/helix plot weight set. This set was constructed by generating separate MI and helix plot weights, scaling the weights so that they covered approximately the same range of values, and adding corresponding edge weights. The raw output from imatch is shown in Figure 3a. Application of i-matching to remove all of the rematched pairs produced the structure shown in Figure 3b. This structure consists of the secondary base pairs that form the cloverleaf structure of tRNA, plus two additional pairings. Both of these latter pairings represent tertiary

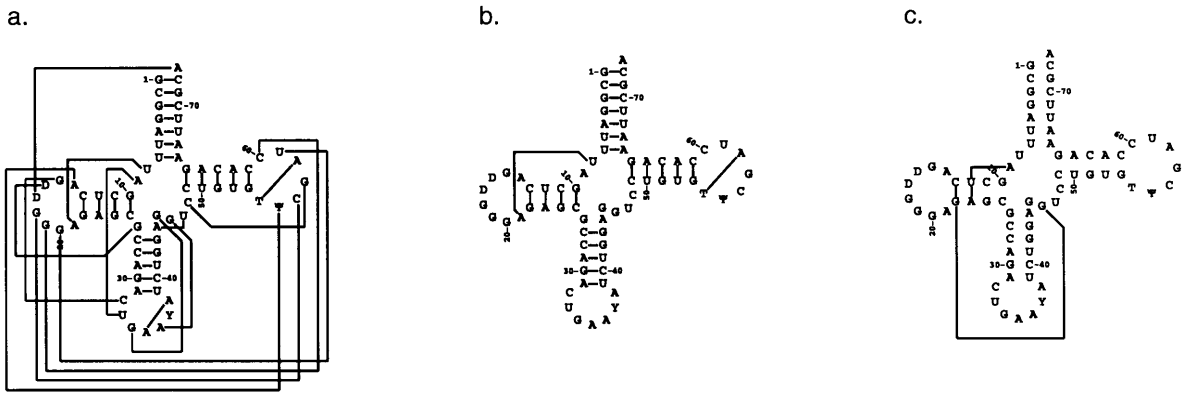

Fig. 3. tRNA folding. (**a**) Raw output from imatch, containing both true and spurious base pairs. (**b**) tRNA structure after application of i-matching output filtration. (**c**) Base triples found by bmatch.

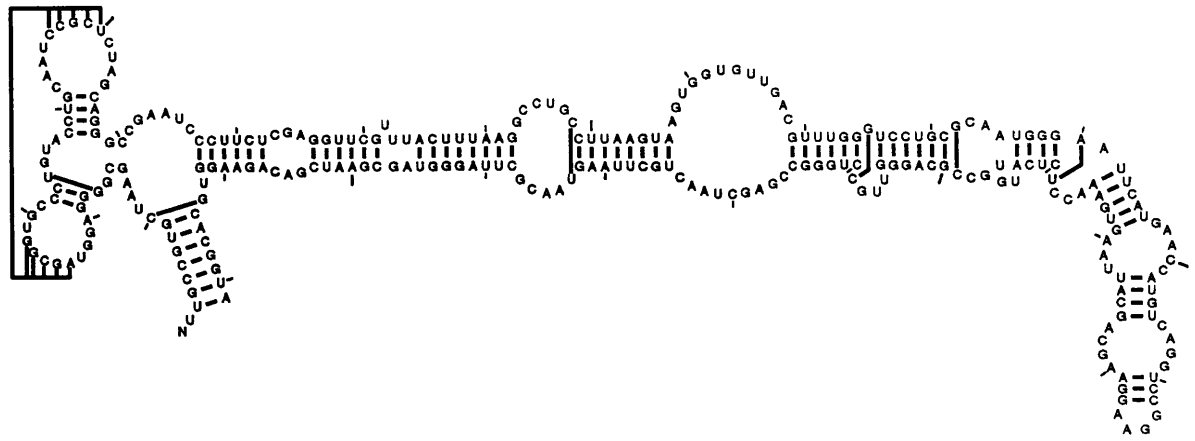

Fig. 4. Predicted structure of *B.subtilis* SRP RNA, compared with the accepted structure. The layout of bases reflects the molecule's structure as determined by Larsen and Zwieb (1990), and lines show base pairs predicted by imatch after removal of non-helical pairs.

base interactions that are generally observed in tRNAs, although the 8 – 21 interaction does not occur in some tRNAs (Westhof *et al*., 1988; Biou, 1994). i-matching did not cause any true base pairs to be discarded.

Prediction of base triples in tRNA was carried out by running bmatch on the MI data set alone. After i-matching, two base triples remained (Figure 3c): 9:12:23 and 13:22:46, which are observed in tRNA crystal structures (Westhof and Sundaralingam, 1986). A third base triple, 10:25:45, was not detected because base 45 is nearly invariant, resulting in low MI scores for pairings involving this base.

To demonstrate MWM folding on a larger molecule, we have generated a structure for *Bacillus subtilis* Signal Recognition Particle (SRP) RNA (Figure 4). This structure was constructed using a helix plot weight set based on an alignment of 33 eubacterial and archaebacterial SRP RNA sequences (Larsen *et al*., 1998). The structure was filtered by discarding helices shorter than hlxplot's default cut-off length of 3 bp. Comparison with the published structure of this molecule shows essentially complete agreement between the structures, including the large pseudoknot near the 5′ end. The most apparent difference between the two structures is the small helix (14 – 15:59 – 60) which was undetected because it was shorter than hlxplot's threshold. Three other base pairs were missed because they would represent bases paired with gaps in the alignment. Several additional base pairs were also found which extend helices in the accepted structure.

Aside from the similarity between the MWM generated structure and the manually built one, two other points about this demonstration bear mentioning. First, the input alignment contained three sequences which do not align well with the rest of the sequences, and may even differ structurally from other SRP RNAs (N.Larsen, on-line documentation for

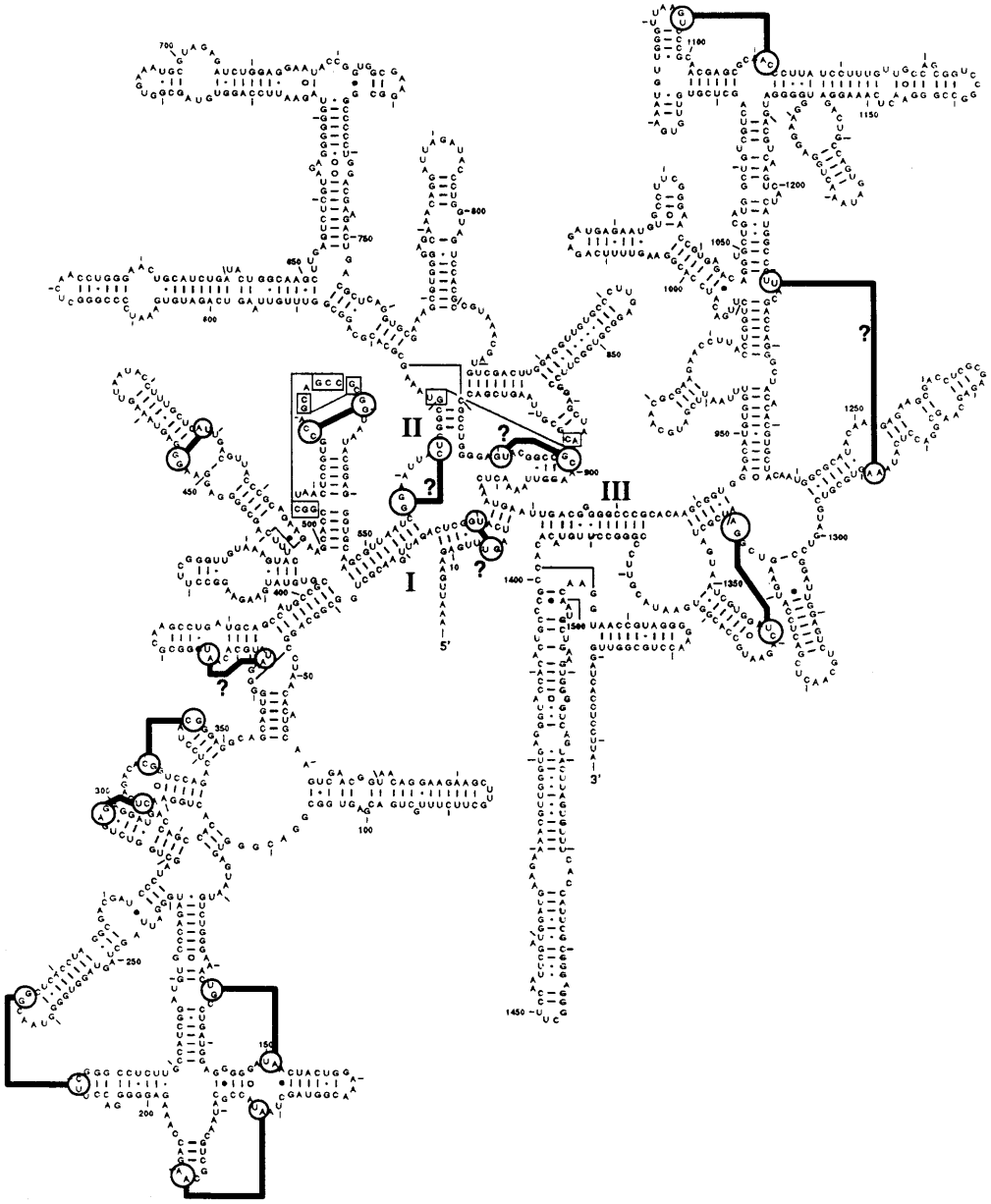

Fig. 5. Predicted helices in *E.coli* 16S rRNA. Thick lines represent predicted pairings. Question marks indicate pairings that conflict with suspected, but as yet unproven, base pairs.

SRP RNA database). Therefore, nearly 10% of the input data for this test were unreliably aligned and undoubtedly contributed some noise to the signal. Second, calculation of the helix plot weights and generation of the MWM required 1.33 s of CPU time on a Sun Ultra 1 computer, compared with ~2 min for MFOLD to fold the *B.subtilis* SRP RNA on the same machine. imatch's predicted structure is also more accurate than MFOLD's, particularly near the 5 end, where MFOLD fails to predict even a planar subset of the base pairs in the pseudoknot (data not shown).

Figure 5 illustrates how existing structural information may be incorporated into an MWM folding. Extensive sequence analysis has produced a nearly complete structure for the bacterial 16S rRNA. To see whether any additional structural elements could be detected, a combined MI/helix plot weight set was calculated from an alignment of 2849 sequences (Maidak *et al.*, 1994). The helix plot component of this weight set was used primarily to allow the detection of highly conserved base pairs, so the minimum helix length of hlxplot was set to 1 bp, and the helix bonus set to zero. exgraf

was used to modify this data set so as to force imatch to form all of the base pairs in the accepted structure of 16S rRNA (Gutell, 1994). Helices predicted by imatch are shown in Figure 5. In general, all of these pairings appear plausible, as they involve canonical base pairs, are fairly local interactions, and in two cases extend known helices. The proposed helices marked with question marks in Figure 5 conflict with suspected but unproven helices in the 16S structure (Gutell, 1996). These interactions may be considered dubious in that regard, although there is also a possibility that these represent base triples or alternate conformations. It is also notable that where pseudoknots are proposed by the MWM structure, they are short and have at least one strand immediately adjacent to another helix, as is typical of other rRNA pseudoknots (Gutell, 1996). Finally, a limited amount of experimental data exists supporting one of these proposed interactions: the bases 1339, 1340, 1358 and 1359, which form a short helix in Figure 5, have been localized to the P site of the bacterial ribosome (Mueller *et al*., 1997). These bases therefore share at least a functional, if not structural, relationship.

Conclusion

Automated RNA structure prediction has long been plagued by the pseudoknot problem: RNA folding programs had to sacrifice either optimality or computational efficiency to detect these structures. As illustrated in the examples above, though, the MWM method is now capable of predicting secondary and tertiary structural elements in RNAs with accuracies approaching those of manually predicted structures. Furthermore, as we have described elsewhere, MWM folding can be used in conjunction with our alignment program, Seq7, to generate structure-based RNA sequence alignments that include pseudoknots (Tabaska and Stormo, 1997). RNA pseudoknot prediction can now be done as routinely as secondary structure prediction.

Exciting possibilities remain for the further development of MWM folding. In theory, this algorithm should be able to detect any kind of base–base relationship in an RNA, as long as the edge weighting scheme used is sensitive to that type of interaction. For instance, in preparation of the demonstrations discussed above, many of the pairings that were discarded as ‘noise’ actually arose from weak covariances caused by functional relationships between bases or non-base-pairing structural features, such as base stacking interactions and bases that form one face of the folded molecule (Tabaska, 1996). Additional development of base pair scoring methods and output analysis should, therefore, not only allow researchers to obtain a detailed map of base interactions for an RNA, but also information on functional moieties, sites of intermolecular interaction, alternate conformations, and surface versus buried regions of the molecule.

Acknowledgement

This work was supported by DOE grant ER61606.

References

Biou,V., Yaremchuk,A., Tukalo,M. and Cusack,S. (1994) The 2.9 Å crystal structure of *T. thermophilus* seryl-tRNA synthetase complexed with $tRNA^{SER}$. *Science*, **263**, 1404–1410.

Cary,R.B. and Stormo,G.D. (1995) Graph-theoretic approach to RNA modeling using comparative data. In *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 75–80.

Chen,J.H., Le,S. and Maizel,J.V. (1992) A procedure for RNA pseudoknot prediction. *Comput. Applic. Biosci*., **8**, 243–248.

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

Gabow,H.N. (1973) Implementations of algorithms for maximum matching on nonbipartite graphs. PhD Dissertation, Stanford University Department of Computer Science.

Gautheret,D., Damberger,S.H. and Gutell,R.R. (1995) Identification of base-triples in RNA using comparative sequence analysis. *J. Mol. Biol*., **248**, 27–43.

Gultyaev,A.P. (1991) The computer simulation of RNA folding involving pseudoknot formation. *Nucleic Acids Res*., **19**, 2489–2494.

Gutell,R.R. (1994) Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994. *Nucleic Acids Res*., **22**, 3502–3507.

Gutell,R.R. (1996) Comparative sequence analysis and the structure of 16S and 23S rRNA. In Zimmerman,R.A. and Dahlberg,A.E. (eds), *Ribosomal RNA: Structure, Evolution, Processing and Function in Protein Biosynthesis*. CRC Press, Boca Raton, FL.

Gutell,R.R., Power,A., Hertz,G.Z., Putz,E.J. and Stormo,G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*., **20**, 5785–5795.

Jacobson,A.B. and Zuker,M. (1993) Structural analysis by energy dot plot of a large mRNA. *J. Mol. Biol*., **233**, 261–269.

Larsen,N. and Zwieb,C. (1990) SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Res*., **19**, 209–215.

Larsen,N., Samuelsson,T. and Zwieb,C. (1998) The Signal Recognition Particle Database (SRPDB). *Nucleic Acids Res*., **26**, 177–178.

Maidak,B.L., Larsen,N., McCaughey,M.J., Overbeek,R., Olsen,G.J., Fogel,K., Blandy,J. and Woese,C.R. (1994) The Ribosomal Database project. *Nucleic Acids Res*., **22**, 3485–3487.

Mueller,F. and Brimacombe,R. (1997) A new model for the three-dimensional folding of *Escherichia coli* 16 S ribosomal RNA III. The topography of the functional centre. *J. Mol. Biol*., **271**, 566–587.

Nakaya,A., Yamamoto,K. and Yonezawa,A. (1995) RNA secondary structure prediction using highly parallel computers. *Comput. Applic. Biosci*., **11**, 685–692.

Nussinov,R. and Jacobson,A.B. (1980) Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl Acad. Sci. USA*, **77**, 6309–6313.

Shapiro,B.A. and Wu,J.C. (1997) Predicting RNA H-type pseudoknots with the massively parallel genetic algorithm. *Comput. Applic. Biosci*., **13**, 459–471.

Steinberg,S., Misch,A. and Sprinzl,M. (1993) Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*., **21**, 3011–3015.

Tabaska,J.E. (1996) Improving automated RNA sequence analysis through applied graph theory. PhD Thesis, University of Colorado Department of Molecular, Cellular and Developmental Biology.

Tabaska,J.E. and Stormo,G.D. (1997) Automated alignment of RNA sequences to pseudoknotted structures. In *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 311–318.

Uhlenbeck,O.C., Pardi,A. and Feigon,J. (1997) RNA structure comes of age. *Cell*, **90**, 833–840.

van Batenburg,F.H., Gultyaev,A.P. and Pleij,C.W. (1995) An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol*., **174**, 269–280.

Westhof,E. and Kim,S. (1988) Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Crystallogr*., **A44**, 112–123.

Westhof,E. and Sundaralingam,M. (1986) Restrained refinement of the monoclinic form of yeast phenylalanine transfer RNA. Temperature factors and dynamics, co-ordinated waters, and base-pair propeller twist angles. *Biochemistry*, **25,** 4868–4878.

Zuker,M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, **244**, 48–52.

Zuker,M. and Jacobson,A.B. (1995) 'Well-determined' regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucleic Acids Res*., **23**, 2791–2798.