Prediction of common folding structures of homologous RNAs

Kyungsook Han and Hong-Jin Kim*
Departments of Computer Science and 1Microbiology and Molecular Genetics, Rutgers University, Piscataway, NJ 08855, USA

Received October 29, 1992; Revised and Accepted January 25, 1993

ABSTRACT
We have developed an algorithm and a computer program for simultaneously folding homologous RNA sequences. Given an alignment of M homologous sequences of length N, the program performs phylogenetic comparative analysis and predicts a common secondary structure conserved in the sequences. When the structure is not uniquely determined, it infers multiple structures which appear most plausible. This method is superior to energy minimization methods in the sense that it is not sensitive to point mutation of a sequence. It is also superior to usual phylogenetic comparative methods in that it does not require manual scrutiny for covariation or secondary structures. The most plausible 1–5 structures are produced in $O(M^2 N^2)$ time and $O(N^2)$ space, which are the same requirements as those of widely used dynamic programs based on energy minimization for folding a single sequence. This is the first algorithm provably practical both in terms of time and space for finding secondary structures of homologous RNA sequences. The algorithm has been implemented in C on a Sun SparcStation, and has been verified by testing on tRNAs, 5S rRNAs, 16S rRNAs, TAR RNAs of human Immunodeficiency virus type 1 (HIV-1), and RRE RNAs of HIV-1. We have also applied the program to cis-acting packaging sequences of HIV-1, for which no generally accepted structures yet exist, and propose potentially stable structures. Simulation of the program with random sequences with the same base composition and the same degree of similarity as the above sequences shows that structures common to homologous sequences are very unlikely to occur by chance in random sequences.

INTRODUCTION
After the first primary sequence of tRNA was published by Holley et al. (1) in 1963 with the cloverleaf form of secondary structure predicted from the sequence, several other secondary structures were proposed, and predicting structure by experimental or theoretical methods remains a challenging problem. The structure was confirmed by X-ray crystallography later in 1974 (2) and the over 300 known tRNA sequences have the same cloverleaf folding potential (3). Unfortunately, tRNA is the only type of RNA which has ever been crystallized and precise determination of tertiary structure is not possible without detailed three-dimensional structural data. However, the secondary structure of RNA can be predicted largely by its ribonucleotide sequence and biochemical or mutational support.

Existing theoretical approaches to RNA secondary structure prediction can be classified into two types: energy minimization and phylogenetic comparison. The energy minimization method uses thermodynamic estimates of structural stability to determine folding structures with minimum or near minimum free energy (4, 5, 6). Limitations of this method are partly due to the uncertainty of the underlying energy model. As we show below, the method is also overly sensitive both to point mutation and to range of a sequence.

The phylogenetic comparative method examines, usually manually, homologous sequences to identify potential helices, which maintain complementarity in the sequences. The current model of secondary structures of 16S and 23S of rRNA were obtained by Noller, Woese, and their collaborators with this method in conjunction with biochemical experiments (7, 8). Some of the Noller—Woese procedure lends itself to an explicit computer algorithm (9, 10), but the algorithm depends on storing all the patterns of interest and has limitations to be applied to sequences with potentially large number of patterns. The algorithm by Sankoff (11) for simultaneously aligning and folding M sequences of length N requires $O(N^2 M)$ time and $O(N^2 M)$ space, and is therefore not practical either in terms of time or space. Recently a Macintosh program for use in phylogenetic comparative analysis has been made available by Brown (12), which shows the result of the analysis in a covariation matrix. However, secondary structure cannot be determined directly from the covariation matrix, especially when many competing helices are present in the matrix. Manual scrutiny of the matrix for possible secondary structures is a laborious task and can overlook possibilities.

* Present address: Molecular Biology Section, Medical Research Division, American Cyanamid Company, Pearl River, NY 01695, USA
In this paper we describe an algorithm and a computer program for predicting common foldings conserved in all or most homologous RNAs. The program is not as sensitive to point mutation of a sequence as the energy minimization method and does not require manual examination for covariation or secondary structures as usual phylogenetic comparative analysis does.

SYSTEM AND METHODS

The operation of the program proceeds in two stages: (1) phylogenetic comparative analysis; (2) inference of structures from the analysis. Given an alignment of M homologous sequences of maximum length N, the program simultaneously analyzes the sequences to identify potential base pairings based on the following algorithm:

For each base i in the alignment,
for each base j(>i) in the alignment,
determine the relation BP(i,j).

The relation BP(i,j) represents the type of base pairing between base i and base j for all the M sequences in the alignment, not for a single sequence. Since different sequences can have different types of base pairing, the type of base pairing between two bases (such as A-U, G-C, or G-U) cannot be used for this relation. Instead, we define the relation as follows:

1. If base i pairs with base j in all the sequences and there is no variation in both i and j, is an exact-invariant pair (denoted by O).
2. If base i pairs with base j in all the sequences and there is compensating base changes in i and j, B(i,j) is an exact-variant pair (*).
3. If base i forms a G-U wobble pair with base j in most sequences, BP(i,j) is a wobble pair (w).
4. If base i pairs with base j in most sequences but not in all the sequences and the frequency of the mismatches do not exceed the specified number, BP(i,j) is an inexact pair (+).
5. If the frequency of mismatches between base i and base j exceeds the specified number, BP(i,j) is a mismatch (-).

The result of the comparative analysis is represented in an N x N covariation matrix, where entry (i,j) represents a relation BP(i,j). Since the relation BP(i,j) is symmetric, only upper triangular part of the matrix is used. In the covariation matrix potential helices of RNA are represented as diagonals of non-mismatch symbols (O, *, +, or +) from upper right to lower left.

In many cases inferring secondary structures from the covariation matrix is not straightforward due to conflicts among potential helices. Conflicts occur when: (1) more than one diagonal exist in the same rows or columns (e.g. b1−b2/b3−b4 and b1′−b2′/b3−b4 or b1−b2/b3−b4 and b1−b2′/b3′−b4′ where the notation b−b′ indicates a subsequence which overlaps a subsequence b−b′); (2) two base pairings exist in disjoint rows and columns but the row range of one overlaps the column range of the other, or vice versa (e.g. b1−b2/b3−b4 and b1′−b2′/b3′−b4′); (3) there are two base pairings b1−b2/b3−b4 and b1−b2/b5−b6 where b1−b2 precedes b3−b4 and b5−b6 precedes b3−b4 and so on; (4) A diagonal itself is self-conflicting because its row overlaps its column (e.g. b1−b2/b1′−b2′) or results in a hairpin loop with fewer than 3 bases. Hairpin loops with fewer than 3 unpaired bases make too sharp turns to be stable (13).

The system uses a heuristic algorithm with spatial constraint propagator incorporated in it. The algorithm first identifies all the potential helices which are not self-conflicting. For each potential helices identified, two types of scores are computed, which are weighted sums of five parameters: helix length, number of exact-variant pairs, number of wobble pairs, number of inexact pairs, and the length of a possible hairpin loop created by the helix. All the structures shown in this paper have been derived using default weight values: w11 = w21 = 1.0, w13 = w23 = −0.5, w14 = w24 = −2.0, w12 = 1.0, w15 = −0.05, w22 = 0.5, w25 = −0.01. The system also allows the user to use other weight values.

S1 = w11 length + w12 exact-variant-pairs + w13 wobble-pairs + w14 inexact-pairs + w15 length - hairpin-loop
S2 = w21 length + w22 exact-variant-pairs + w23 · w24 inexact-pairs + w25 length - hairpin-loop

S1 is computed for finding the most stable hairpin loop-stem structure of the overall structure, and S2 is for measuring the ‘goodness’ of a helix in general positions (e.g. helix next to a bulge loop, internal loop, multiple loop, or hairpin loop). A helix with the highest S1 value is the one which forms the most stable hairpin loop-stem structure and is called ‘best helix’ in the following algorithm. If there is any other helix which forms the same stem structure with the best helix (i.e. these helices are intervened only by a bulge or an internal loop), it becomes the supporting helix of the best helix. The best helix and its supporting helix (if any) constitute a skeleton of the overall structure. Once the skeleton is found, it is used to eliminate the helices which violate the spatial constraints with respect to the skeleton. Each of the remaining helices is examined with better one first, and is attached to currently maintained structures if it fits them. If it cannot fit, an alternative structure is created for it. The new structure contains the skeleton and the potential helix which had conflicts at the time it is first created. The heuristic algorithm is briefly stated below. Propagation of the folding constraints occur in step 4 and 6 of the algorithm.

1. bestscore = −∞
2. for each potential helix i identified
   compute S1 and S2
   if (S1 > bestscore)
   besthelix = helix i
   bestscore = S1 of helix i
3. skeleton = {besthelix} ∪ [any supporting helix]
4. Remove the helices which conflict with the skeleton
5. Sort the remaining helices with respect to S2
6. for each of the sorted helices
   if it fits currently maintained structures, add it to them
   otherwise, create an alternative structure for it
   
   The value of minimum length of a helix is controlled by the user with default value 3, and is one factor which affects the number of potential helices. The smaller the value, the more the potential helices. It also influences the prediction. If the number of minimum length of a helix is set to a larger value than 3, then a rough structure is obtained where helices shorter than the threshold value are missing but all the other helices stay. Therefore, increasing the minimum length of a helix does not affect much the prediction other than that shorter helices are ignored. Decreasing the number of minimum length of a helix from 3 to 2 has a slightly more complicated influence on the prediction; it can simply add new helices of length 2 to the previous structure while preserving all the existing helices in it;
or, helices of length 2 can sometimes defeat longer helices in the previous structure and replace them.

The total number of structures produced by the program is restricted to 5 or less. The value of 5 is a somewhat arbitrary choice. Generating more structures by a current version of the program is not difficult at all, but excessive number of possible structures seem to be as uninformative as none. Also, when a helix $h$ cannot fit any of the current structures (i.e. there is at least one helix in each existing structure which conflicts the helix $h$) and the program creates an alternative structure, generating more than one alternative structure due to the helix $h$ is prohibited. This is necessary to avoid generating all the 5 structures which contain identical helices except one helix.

RESULTS AND DISCUSSION

All the algorithms have been implemented in C on a Sun SparcStation running SunOS Release 4.1.1. The program is available from the first author upon request. Figure 1A shows the result of the comparative analysis on 10 tRNA sequences. The matrix contains only 7 potential helices which are at least 3 bp long. Four of them (1-7/83-89, 10-12/24-26, 28-32/40-44, and 66-70/78-82) correspond to the well-known 4 stems (acceptor, D, anti-codon, TyC) of tRNA (2, 3), and the program predicts only one structure, represented in a form of a sequence and matching parentheses (Figure 1B). This is a “quick and easy” representation requiring no special graphics device, and is directly readable by a drawing program such as LoopViewer or LoopDloop (14) for obtaining a better quality representation (Figure 1C). Test of the program on 34 prokaryotic 5S rRNA sequences shows a similar result. There are only 5 potential helices in the covariation matrix of the 5S rRNA sequences (data not shown) and 4 of them constitute the normally accepted stems of prokaryotic 5S rRNA structure (10, 15, 16, 17) (Figure 2A).

The program was also applied to parts of sequences in HIV-1, a causative agent of AIDS. cis -acting sequences required for the virus replication are present in the HIV-1 genome; TAR (trans-activation-responsive element) and RRE (rev-responsive element), two of the cis-acting sequences involved in gene regulation, lie downstream of the transcription start site in the viral long terminal repeat and in the env region, respectively (18, 19, 20); other cis-acting sequences involved in packaging are located between the primer binding site and gag initiation site (21, 22). These are highly conserved sequences and may not be good test cases of typical phylogenetic comparative analyses.

Figure 1. (A) Result of comparative analysis on 10 tRNA sequences (ANTTRLCAA, ANITRLCAG, BACTRLC, ECOVRLA1, ECOVRA1B, ECOVRL2, ECOVRL3, ECOVRS1, ECOVRS3). Minus signs (—) in the alignment indicate gaps. The last sequence in the alignment is a consensus sequence of the ten homologous sequences, derived by the program. Upper case letter in the consensus sequence represents a base conserved in all the sequences and lower case represents a base which is not conserved in all the sequences but most representative one in the position. A complementary base change is represented in an asterisk (*), an exact match in an open circle (O), a wobble pair in w, an inexact pair in a phi sign (+), a mismatch in dot (•). (B) tRNA structure inferred from the matrix of Figure 1A. It is represented in a form of a sequence and matching parentheses. The sequence and base numbering are based on the ANITRLCAA sequence (first sequence in the alignment). (C) tRNA structure drawn by LoopViewer given the input of Figure 1B.
There are 63, 145, and 561 potential helices in the covariation matrices of TAR sequences, cis-acting packaging sequences, and RRE sequences from HIV-1, respectively (data not shown). All the diagonals are potential helices, but they cannot coexist at the same time due to conflicts among them. The program predicts a stable stem-loop structure (Figure 2B) for the common structure of 10 HIV-1 TAR sequences, which corroborates the published data (18). Figure 2C shows the common structure of 10 HIV-1 RRE sequences (NL43, LAI, HXB2R, MN, JRCSF, SF2, NY5CG, CDC4, HAN, RF). The first base of the figure is base 7323 of the NL43 genomic sequence.

Zuker and Stiegler's folding program FOLD predicts the RRE structure correctly if Tinoco rules are used with proper range of the sequence (23).

Figure 3 shows two plausible structures for 10 HIV-1 cis-acting packaging sequences, which encompass a splice donor site. They are the most stable structures conserved in the 10 HIV-1 cis-acting packaging sequences. Unlike TAR and RRE, the structure of the cis-acting packaging sequences have not been stabilized yet. Due to little variation among sequences, compensating base changes are rare in the helices (Table 1). Figure 4A shows an optimal structure of the HIV$_{LAI}$ packaging sequence with minimum energy ($-24.5$ kcal/mol), predicted by FOLD (5) in version 7.1 of the GCG Sequence Analysis Software Package running on a VAX/VMS version V5.4. Figure 4B is an optimal structure of the HIV$_{NL43}$ packaging sequence with minimum energy ($-25.5$ kcal/mol), predicted by the same program. The cis-acting packaging sequences of HIV$_{LAI}$ and HIV$_{NL43}$ are the same except at two bases and are expected to have similar structures. Their structures predicted by FOLD, however, are radically different from each other. The two sequences are different at their first base (base 228 of their genomic sequences), but this difference did not affect the structural difference at all as shown in Figure 4. The actual and only reason for their structural difference is at the difference of base 281, which is A for HIV$_{LAI}$ and C for HIV$_{NL43}$. From the perspective of phylogenetic comparative analysis, the structure of Figure 4A is not as stable as that of Figure 4B. This result shows that the structure predicted by the energy minimization method is overly sensitive to point mutation of a sequence. Testing FOLD on HIV$_{LAI}$ or HIV$_{NL43}$ sequences with a slightly different range which encompasses the packaging sequences produces very different structures (data not shown), where not a single part of Figure 4A or Figure 4B is preserved.
This indicates that the structure predicted by the energy minimization is also sensitive to range which contains a sequence of interest. Recently Harrison and Lever have published a structure of the HTV-1 cis-packaging sequences (24), which is identical to the structure of Figure 4B, except that it has an additional G-C pair in stem III. In the structure of Figure 4B and the structure predicted by Harrison and Lever, there is uncertainty around the multiple loop and some of the helices (e.g. UCG/CGA) around the multiple loop may not exist in that form. Another structure of HTV packaging sequences proposed by Hayashi et al. (25), is not considered to be stable either in terms of phylogeny or free energy.

We have also tested the program with ten 16S rRNA sequences (E.coli, P.vulgars, M.capricol, Hf.volcard, Hc.morrhua, Mc.vanniel, Sul.solfat, Hb.cutirub, Hb.halobiu, Msp.hungat). Due to the limitation in the alignment file format of the program, the 16S rRNA sequences were split into 4 domains (5’ domain, central domain, 3’ major domain, and 3’ minor domain) based on (26). Each domain was tested separately with weight values, \( w_1 = 1.0, w_2 = 0.0, w_3 = 0.5, w_4 = -2.0, w_5 = 0.0 \) for \( S_j \) and the default weights for. All the helices with minimum 3 canonical or G-U pairs common to the individual 16S rRNA structures of (8, 27) are present in our predicted structures (data not shown).

Simulation of the program with random sequences has been performed to address whether the predicted common structures might occur by chance in random sequences with the same base composition and/or the same degree of similarity. We tested the program on two types of random sequences generated in a method described by Chan et al. (28) and reached a similar conclusion. Independent random sequences have much fewer potential helices than similar random sequences, which in turn have fewer potential helices than homologous RNA sequences. The number of potential helices with minimum 3 bp are summarized in Table

<table>
<thead>
<tr>
<th>organism</th>
<th>helix 12-15/59-61</th>
<th>helix 82-84/106-108</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIVNL43</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVLA1</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVHXB2R</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVMN</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVJRCSF</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVSF2</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVNY5CG</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVDC4</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVHAN</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>HIVRF</td>
<td>cgcg/cgc</td>
<td>uugu/ug</td>
</tr>
<tr>
<td>CONSENSUS</td>
<td>CC/GCG</td>
<td>UU/GGA</td>
</tr>
</tbody>
</table>

| Location of the helices are based on the base numbering in the alignment of the 10 sequences, which includes gaps. The last sequence in each helix is a consensus sequence of the 10 sequences.
Figure 4. (A) An optimal structure of HIV\_LAI packaging sequences (base 227—base 335 of HIV\_LAI genomic sequence) predicted by Zuker and Stiegler's folding program (5). Different bases from HIV\_M3 are boxed. SD, splice donor site. (B) An optimal structure of HIV\_W/43 packaging sequences (base 227—base 335 of HIV\_W/43 genomic sequence) predicted by Zuker and Stiegler's folding program. Different bases from HIV\_LAI are marked.

Table 2. Number of potential helices in the covariation analysis of random sequences.

<table>
<thead>
<tr>
<th></th>
<th>helix length</th>
<th>number of sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAR-I</td>
<td>≥ 3</td>
<td>4</td>
</tr>
<tr>
<td>TAR-S</td>
<td>≥ 3</td>
<td>33</td>
</tr>
<tr>
<td>PKG-I</td>
<td>≥ 3</td>
<td>10</td>
</tr>
<tr>
<td>PKG-S</td>
<td>≥ 3</td>
<td>97</td>
</tr>
<tr>
<td>RRE-I</td>
<td>≥ 3</td>
<td>40</td>
</tr>
<tr>
<td>RRE-S</td>
<td>≥ 3</td>
<td>388</td>
</tr>
</tbody>
</table>

Sequences with suffix I represent independent random sequences. Sequences with suffix S represent random sequences with 90 similarity. The results are the average of 10 computer runs.

Table 3. Lengths of potential helices in the covariation analysis of similar random sequences and homologous RNA sequences

<table>
<thead>
<tr>
<th>sequence</th>
<th>potential helix size</th>
<th>≥ 6</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAR-S</td>
<td>11</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>TAR</td>
<td>32</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>PKG-S</td>
<td>36</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>PKG</td>
<td>97</td>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>RRE-S</td>
<td>126</td>
<td>33</td>
<td>9</td>
</tr>
<tr>
<td>RRE</td>
<td>340</td>
<td>119</td>
<td>38</td>
</tr>
</tbody>
</table>

10 sequences were used for each analysis. The results shown for random sequences are the mean of 10 computer runs.

In this paper we have described an algorithm and a computer program for predicting common foldings conserved in homologous RNAs. The program finds 1—5 common foldings of M sequences of length N in \( O(MN^2 + N^3) \) time and \( O(N^2) \) space. In the first stage, the covariation matrix is generated in \( O(MN^2) \) time and \( O(N^2) \) space. The number of potential helices in the matrix is \( O(N^2) \) and sorting them takes \( O(N^2 \log N) \) time. Each helix is examined once and compared with a helix in structures \( k < 5 \). Since each structure contains nonconflicting helices, the maximum number of helices in each structure is \( O(N) \). Constructing common foldings takes \( O(N^2 \log N + N^2 \cdot k \cdot N) = O(N^3) \) time and \( O(N^2) \) space. Therefore, the total time and space for the algorithm are \( O(MN^2 + N^3) \) and \( O(N^2) \), respectively. The space requirement of the program is the same as that of the ordinary folding program for a single sequence. The time requirement is also the same as that of the ordinary program for a single sequence if \( M = O(N) \) or \( M = o(N) \), which is usually the case.

The program has been successfully tested on various types of RNAs. The execution time for predicting common foldings of 10 tRNAs of length 77, 34 5S rRNAs of length 143, 10 TAR RNAs of length 57, 10 RRE RNAs of length 196, and 10 HIV-1 cis-packaging sequences of length 113 are 0.1, 0.6, 0.1, 0.4, and 0.2 seconds, respectively. These lengths are the maximum lengths of the sequences in the alignments which include gaps. It is known that there is no reliable or automatic way of predicting a common folding even if the related sequences are already aligned and that one way to obtain a common folding is to fold each sequence separately using an ordinary folding program and to search for common structures (29). Our algorithm is the first one provably practical both in terms of time and space for finding folding structures of homologous RNA sequences. The algorithm is currently being extended for finding tertiary interactions and for modeling the process of the structure formation over time.

ACKNOWLEDGEMENTS

We are grateful to Mick Noordewier for comments on earlier version of this paper, and to Bryn Weiser for aligned 16S rRNA sequences.
REFERENCES