Inferring the conformation of RNA base pairs and triples from patterns of sequence variation

Daniel Gautheret* and Robin R. Gutell[1]

Structural and Genetic Information, CNRS EP91, 31 ch. Joseph Aiguier, 13402 Marseille Cedex 20, France and [1]Department of Chemistry and Biochemistry, Campus Box 215, University of Colorado, Boulder, CO 80309-0215, USA

Received December 31, 1996; Revised and Accepted February 26, 1997

ABSTRACT

The success of comparative analysis in resolving RNA secondary structure and numerous tertiary interactions relies on the presence of base covariations. Although the majority of base covariations in aligned sequences is associated to Watson–Crick base pairs, many involve non-canonical or restricted base pair exchanges (e.g. only G:C/A:U), reflecting more specific structural constraints. We have developed a computer program that determines potential base pairing conformations for a given set of paired nucleotides in a sequence alignment. This program (ISOPAIR) assumes that the base pair conformation is maintained through sequence variation without significantly affecting the path of the sugar–phosphate backbone. ISOPAIR identifies such 'isomorphic' structures for any set of input base pair or base triple sequences. The program was applied to base pairs and triples with known structures and sequence exchanges. In several instances, isomorphic structures were correctly identified with ISOPAIR. Thus, ISOPAIR is useful when assessing non-canonical base pair conformations in comparative analysis. ISOPAIR applications are limited to those cases where unusual base pair exchanges indeed reflect a non-canonical conformation.

INTRODUCTION

Comparative sequence analysis of RNA structure is based on the simple principle that homologous RNA molecules will adopt the same secondary and tertiary structures with different primary sequences. Practically, comparative studies identify secondary and tertiary structure base pairings by finding compensatory base changes (covariations) in alignments of homologous sequences. This approach has been successfully applied to several classes of RNA molecules, notably tRNA (1), 5S (2), 16S (3) and 23S (4) rRNA, group I (5,6) and II (7) introns and the RNA component of RNase P (8; see 9 for review). In tRNA for instance, every secondary base pair and several tertiary interactions were predicted before a crystal structure was available (1,10). There are now large numbers of sequences available for each of these RNA molecules and, with improved algorithms for the detection of covariations, secondary and tertiary structure models are being continuously refined. Comparatively derived models such as those of group I introns and ribosomal RNAs are supported by a considerable body of experimental data. In its search for a common structure, comparative analysis is now seeking to identify more detailed structural features.

Analyses of 16S and 23S rRNA sequence alignments (11) have revealed that many paired positions are restricted to certain types of pairing sequences, either subsets of the four Watson–Crick sequences or non-canonical sequences, such as A:C and G:U or A:G and G:A. These patterns of variation point out base pair conformations different from the canonical ones. Most of the restricted variations in rRNA base pair sequences are of the R:Y (purine:pyrimidine) type (11). For example, these sequences are either G:C or A:U, not U:A or C:G. Such events have been associated with base stacking constraints and specific deformations of A-helices involved in RNA recognition (12).

Although canonical Watson–Crick base pairs may occasionally be submitted to sequence constraints such as R:Y constraints, non-canonical base pairs should systematically result in manifest sequence biases. Sequences that are not compatible with the required base pairing conformation should be excluded, while other sequences could be freely explored during evolution. In this case, the base pairing structure could, in theory, be inferred by seeking structures that are common to the observed sequences. We present in this article a computer program, ISOPAIR, that automatically determines isomorphic structures for any observed pattern of sequence variation. We describe as 'isomorphic' a set of base pairs that can all be formed in a given structural environment. Practically, two base pairs that can form with a similar orientation of the sugar–phosphate backbone are considered isomorphic. For instance, all Watson–Crick pairs are isomorphic, but Watson–Crick and Hoogsteen pairs are not, as Hoogsteen pairs involve a different position of the RNA backbone. Isomorphism does not apply to every interaction in RNA structures, but can be a useful starting point in the study of non-canonical base pairs. We show that important tertiary interactions in tRNA display sequence variations consistent with this assumption.

*To whom correspondence should be addressed. Tel: +33 491 16 45 48; Fax: +33 491 16 45 49; Email: gauthere@igs.cnrs-mrs.fr

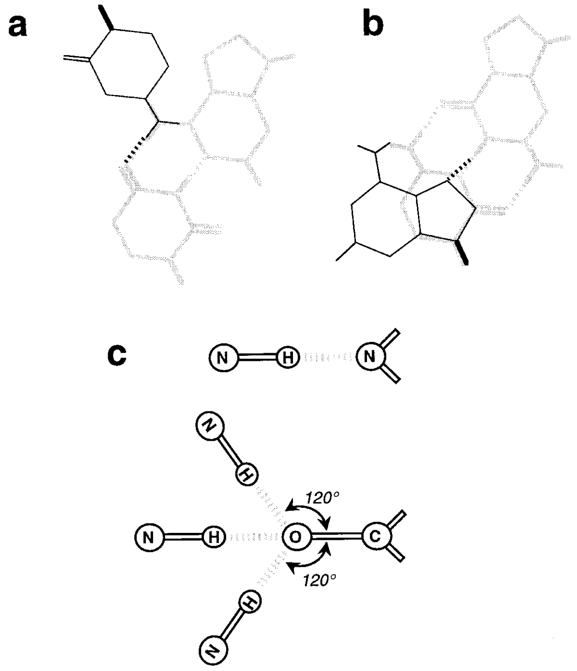


Figure 1. Construction of planar single H bond base pairs. (**a**) Stage 1. Superimposition of H bond donors and acceptors onto H bonds of initial base pairs. (**b**) Stage 2. Superimposition of glycosyl bonds onto glycosyl bonds of initial base pairs. (**c**) Stage 3. Systematic construction of N-H··N and N-H··O bonds, using the bond angles shown.

MATERIALS AND METHODS

The ISOPAIR program takes as input a set of base pair or triple sequences (e.g. {G:C, A:G, U:A} or {C:G:A, G:C:U}), typically obtained from covariation analysis. There is no limit to the number of input sequences. Isopair first generates internally a list of possible pairing conformations for each sequence and then seeks sets of isomorphic conformations that can be formed with every input sequence. Conformation sets are returned in the form of PostScript or Brookhaven 'pdb' files. The program is written in C. Unix executables are available through anonymous ftp at igs-server.cnrs-mrs.fr, in directory /pub/ISOPAIR, or through written request to gauthere@igs.cnrs-mrs.fr.

Initial generation of base pair conformations

An initial set of 28 double H-bond conformations available in the literature (13) was constructed using interactive molecular graphics. Single H-bond pairings are computer-generated. Their construction is limited to planar structures and proceeds as follows.

(i) A first collection of single H-bond pairs is obtained by superimposing each of the four bases onto the 28 double H-bond pairs built previously. ISOPAIR performs superimpositions in two ways. First, H-bond donors and acceptors are superimposed onto the H-bonds of the 28 initial pairs (Fig. 1a). New pairing structures that do not produce steric conflicts are stored. In the example shown, a new C:U pair is generated.

(ii) In a second stage, glycosyl bond of each four nucleotides are superimposed onto glycosyl bonds of the 28 initial base pairs (Fig. 1b). New structures that contain at least one H-bond and do not produce steric conflicts are stored. In the example shown, a new A:A pair is created. This stage ensures that no single H-bond conformation that is rigorously isomorphic to a known double H-bond conformation is omitted.

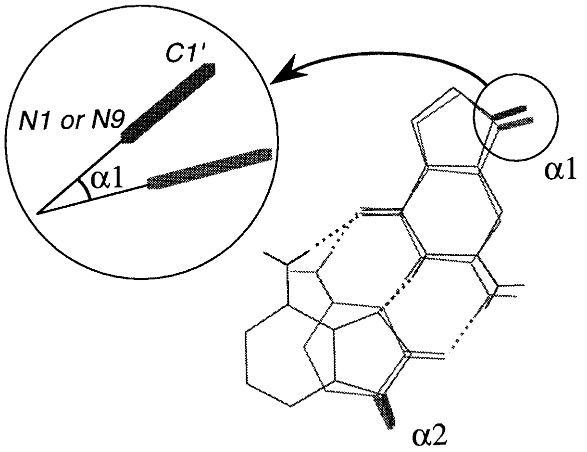


Figure 2. Base pair superimpositions used in the measure of isomorphism. Glycosyl bond atoms N1 (or N9) and C1′ from the two base pairs to be compared are superimposed and rms deviations are measured between these two sets of four atoms. Angles α1 and α2 are measured as well.

(iii) Finally, additional single H-bond base pairs are sought through a systematic connection of H-bond donors and acceptors in all four bases, using the H-bond angles shown in Figure 1c.

This procedure generates a total of 351 different pairing structures. Base triple structures are generated by combining Watson–Crick or wobble pairs with every non-canonical pair.

Defining the distance between two conformations

We define base pair isomorphism, *I*, as the ability to form while retaining similar sugar–phosphate backbone conformations. We must therefore quantify how well backbones from two different base pairs can be superimposed. As rotations about the glycosyl bond are possible, we can only compare the position of glycosyl bonds and the angle they form in the pairs. Glycosyl bond atoms (N1 or N9 and C1′) from the two pairs to be compared are thus superimposed and their root mean square (rms) deviation is measured, as well as the angles α1 and α2 represented in Figure 2. The hyperbolic functions *f*1(α) or *f*2(rms) represented in Figure 3 are then applied to convert rms and angle values into a distance value comprised between 0 and 1. These functions increase quasi-exponentially near zero, which quickly penalizes measures departing from ideal values but, unlike exponential functions, they are upper-bounded, which permits comparisons of independent measures. The plateau reached with high α or rms values is not a problem here since values in this range correspond to uninteresting non-isomorphic conformations. The final distance *D* is *f*1(α) × *f*2(rms). Base triple comparisons are performed similarly, using three glycosyl bonds instead of two.

Selection of isomorphic sets of conformations

Given the input base pair or triple sequences $\{s_1, s_2, ..., s_n\}$ where each sequence s_i has *m* possible conformations $C_i = \{c_{i,1}, c_{i,2}, ..., c_{i,m}\}$, we compute all the pairwise distances $D(c_i, c_j)$ where $c_i \in C_i$ and $c_j \in C_j$. Using a conventional branch and bound

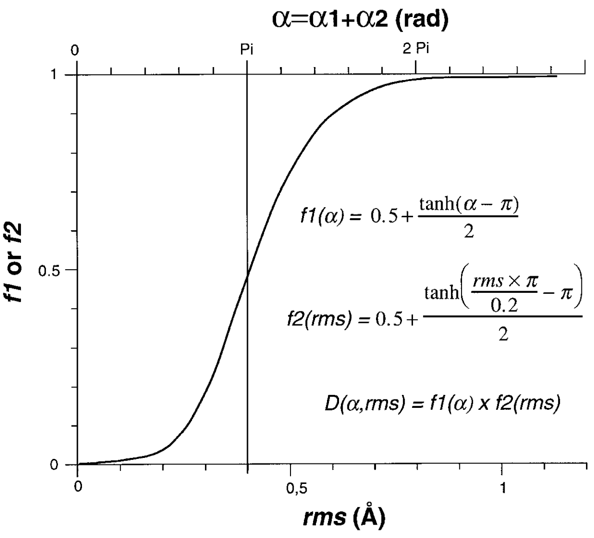


Figure 3. Functions used in the measure of isomorphism. α1, α2 and rms deviation are defined in here and text. The inflexion points of *f*1 and *f*2 for α = π and rms = 0.4 were chosen empirically. functions *f*1, *f*2 and *D* have the same shape.

algorithm, we then construct all the conformation sets of the form $\{c_1, c_2, ..., c_n\}$, $c_1 \in C_1$, $c_2 \in C_2$, ..., $c_n \in C_n$, satisfying:

$$I = \frac{\Sigma_{i=1..n, j=1..n} D(c_i, c_j)}{n^2} < t$$

This selects conformation sets for which the average pairwise distance *I* (or isomorphism) is lower than a fixed threshold *t*. Note that the lower *I*, the higher the isomorphism. After visual inspection of isomorphic structures, we empirically set the value of *t* at 3×10^{-3}. A set of conformations for which *I* is lower than this value is said 'isomorphic'.

Constraints

When the input set of base pair sequences yield several isomorphic solutions, further criteria are needed to distinguish the most interesting ones. ISOPAIR may optionally require that solutions contain at least one double H-bond pair. This is referred to as the 'double H-bond' constraint. Users can also prohibit conformations involving variable glycosyl bond orientations (*syn* and *anti*) in the same isomorphic set. This can be used for instance to avoid certain solutions containing *syn* pyrimidines.

ISOPAIR can also exclude pairings that can be formed by sequences that are not in the input set. For instance, if an isomorphic structure is found for the covariation {A:A, G:G} and this structure can also be formed with {C:G}, we may consider this structure as 'wrong', as it provides no strict rationale for the sequence observation. This constraint, which we term 'uniqueness', is useful when it can be reasonably argued that unobserved sequences are indeed counter selected. Uniqueness (*U*) of a isomorphic set is defined as the shortest average distance between the structures in that isomorphic set and any structure that can be formed with other base pair sequences. An isomorphic set with isomorphism *I* and uniqueness *U* is considered as unique if $U > 3 \times 10^{-3}$ or $U > 2 \times I$ (empirical thresholds). Due to the high number of base triple conformations, the current version of ISOPAIR cannot test uniqueness for base triples. Base triples, can optionally be constrained to occur in the major groove.

Base pair sequences

Base pair exchanges in tRNA were obtained from a sequence alignment adapted from that of Sprinzl (14), containing 895 type I nuclear tRNA and tRNA gene sequences. Type I and type II tRNAs differ in the size of their variable loop (4 or 5 nt in type I tRNAs, 10–24 nt in type II tRNAs) and in the position of their tertiary interactions.

RESULTS

A first simple question that can be addressed with ISOPAIR is the number of base pairs that may potentially adopt a common structure. There are 96 possible sets of two different base pairs ({A:A,A:C}, {A:A, A:G}, {A:A,A:U}, etc.) after removal of equivalent sets such as {A:A,A:U} and {A:A,U:A}. We ran ISOPAIR for each set and counted the number of solutions. For each set, ISOPAIR finds at least one isomorphic solution with an *I* value below 3×10^{-3}, that is with only mild differences in glycosyl bond angles and positions (data not shown). This result is not surprising if one considers the large variety of single H-bond conformations. Most of these common structures, however, are not unique to the input pairing set. For instance, the Watson–Crick conformation is a solution for the input set {U:A, A:U}, but this conformation may also be achieved with G:C or C:G. When studying sequence variations issued from comparative sequence analysis, this type of non-specific solution is questionable as it cannot explain why only certain base pair sequences are observed. This is why we introduced the 'uniqueness' constraint (see Materials and Methods), that discards solutions that can also be obtained with sequences not in the input set. Now of the same 96 sequence sets analyzed earlier, only 33 have a unique solution (Fig. 4). The uniqueness constraint thus considerably reduces the number of structures to be considered.

To test ISOPAIR's ability to reproduce known pairing geometries from typical covariations, the program was given canonical combinations of Watson–Crick and G:U sequences. As expected, the input set {A:U, U:A, G:C, C:G} produces the Watson–Crick conformation as the most isomorphic solution, with an *I* value of 5.3×10^{-6}. This result is obtained whether or not 'uniqueness' is imposed on solutions, confirming what we already knew about the conformation of this set of pairings. When G:U or U:G are added to the four Watson–Crick sequences, the most isomorphic solution has an *I* value of 7.3×10^{-5} and, surprisingly, does not contain Watson–Crick nor wobble conformations. This solution, shown in Figure 5a, only contains single H-bond base pairs. The expected Watson–Crick/wobble solution (Fig. 5b) ranks fourth but is the best solution involving double H-bond pairs. The number of H-bonds in base pairs could thus be important in ranking solutions. A correct prediction can be achieved here by choosing the solution that involves the highest number of H-bonds. The *I* value for this solution (8.2×10^{-5}) is an order of magnitude higher than for Watson–Crick sequences alone, due to the significant difference in glycosyl bond positions between Watson–Crick and wobble pairs.

Applying the uniqueness constraint to the {A:U, U:A, G:C, C:G, G:U} input set eliminates the Watson–Crick/wobble solution. This was expected, since a wobble structure can also form with an A:C pair. The input set {A:U, U:A, G:C, C:G, G:U,

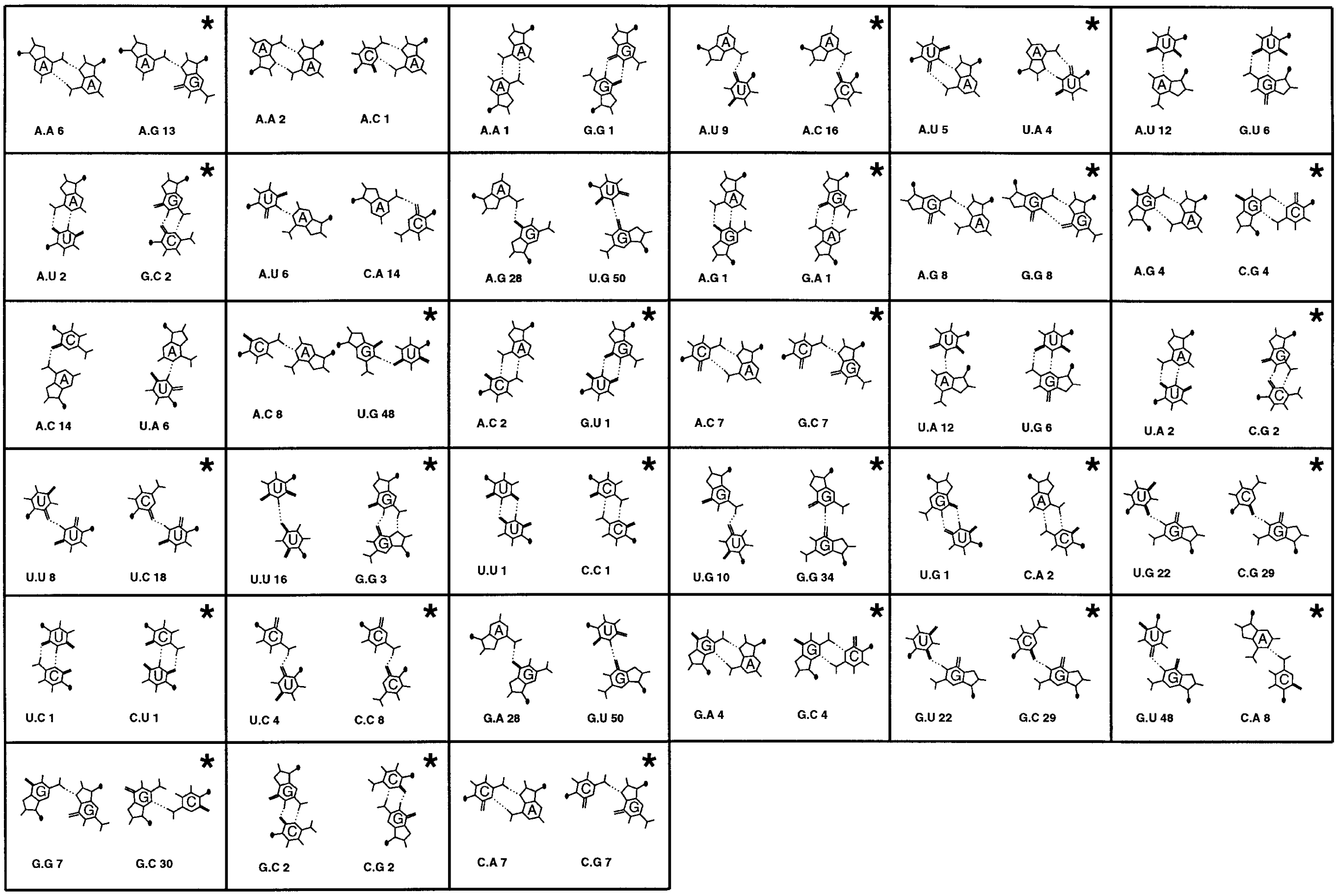


Figure 4. Unique isomorphic structures for combinations of two base pairing sequences. Asterisks indicate the presence of several isomorphic conformations for the input sequences. In this case, the solution shown was selected based on: (i) the presence of double H bond pairings; (ii) the lowest*I* value. The presence of both *syn* and *anti* conformations in the same solution was purposely not checked in these ISOPAIR runs. Numbers below each base pair refer to the internal numbering of base pair structures in ISOPAIR.

U:G} does not produce the Watson–Crick/wobble solution, whether or not the uniqueness constraint is used. This is consistent with the large deviation observed when superimposing wobble pairs G:U and U:G.

tRNA

Transfer RNA sequences provide a number of covariations that can be related to known pairing structures. An important unusual covariation is that found at position 15:48, known as the Levitt pair (10). The vast majority of tRNAs contain either G:C or A:U at this position, and Klug *et al.* (15) have suggested that this sequence constraint was consistent with the parallel reverse Watson–Crick pair present at this position, since bases other than G:C or A:U would induce a significant backbone displacement if paired similarly. ISOPAIR finds more than 10 different isomorphic structures for A:U and G:C, one of which is the reversed Watson–Crick pairing found in tRNA crystal structures. However, the only ‘unique’ solution containing a double H-bond pair is indeed the parallel reverse Watson–Crick pairing observed in tRNA crystal structures (Fig. 6a).

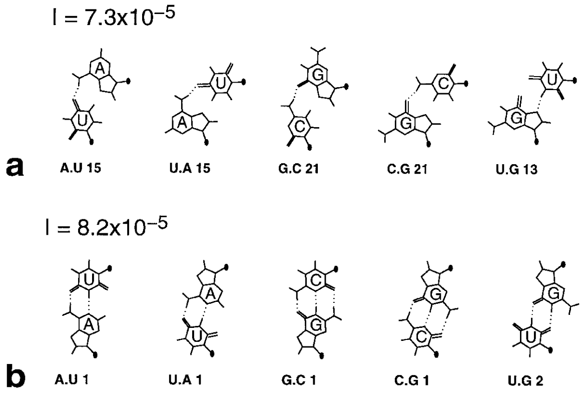


Figure 5. ISOPAIR results for input set {A:U, U:A, G:C, C:G, G:U}. (**a**) Highest ranking solution. (**b**) Fourth ranking solution. Numbers below each base pair refer to the internal numbering of base pair structures in ISOPAIR.

Certain cysteine tRNA do not have the usual G:C or A:U Levitt pair, but have instead a G:G pair (16) that cannot adopt a reverse-Hoogsteen conformation. This can be regarded as a threat to our initial assumption that sequences not compatible with a required base pairing would be excluded by selection. It has been shown, however, that this G:G Levitt pair is a determinant for the aminoacylation of tRNA^{Cys} (17). In this case, the absence of isomorphism is thus related to a variation in the structure and

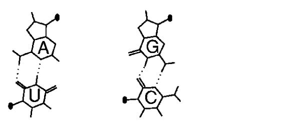
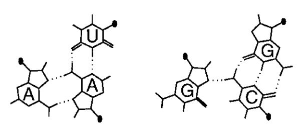
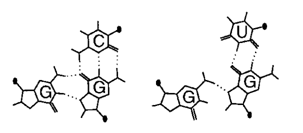
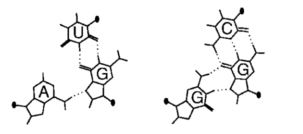
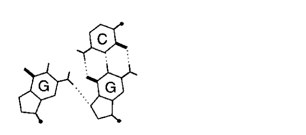
	Sequences (position, molecule)	Actual solution (as observed in crystal structure)	ISOPAIR rank
a	**A:U** **G:C** 15:48 All tRNAs		#1 with uniqueness and double H–bond constraints, rank > 10 otherwise
b	**U:A:A** **G:C:G** 12:23:9 tRNA^{Phe}		#1 with double H–bond constraint, #6 otherwise
c	**C:G:G** **U:G:G** 13:22:46 tRNA^{Phe}		#1 with double H–bond constraint, #4 otherwise
d	**U:G:A** **C:G:G** 13:22:46 tRNA^{Asp}		#4 with double H–bond constraint, #6 otherwise
e	**G:C:G** **G:U:G** 10:25:45 tRNA^{Phe}		Not found

Figure 6. Sequence variations in tRNA alignments and corresponding base pair or base triple structures. (**a**–**d**) The structures shown are isomorphic sets produced by ISOPAIR using the input sequence in the left column. The last column indicates the rank of this isomorphic solution in terms of *I* value, in the presence or absence of 'double H bond' constraint. (**e**) tRNA^{Phe} sequences for base triple 10:25:45 and the structure observed in yeast tRNA^{Phe}. ISOPAIR does not identify this structure.

function of the RNA molecule, consistent with our initial assumption.

Other unusual covariation patterns in tRNA are observed at base triple positions. Since the structure of these base triples varies considerably (18), they have been studied independently for each tRNA species (Asp, Phe, etc.). In yeast tRNA^{Phe}, base triples occur at positions 12:23:9, 13:22:46 and 10:25:45. At position 12:23:9, the two most predominant sequences are U:A:A and G:C:G in all 895 tRNA sequences in the database, as well as in each of the tRNA species for alanine, phenylalanine, asparagine and tryptophan. We sought isomorphic structures for these two triple sequences. When using the 'double H-bond' constraint (see Materials and Methods), the highest ranking solution (Fig. 6b) is that observed in the tRNA^{Phe} crystal structure (19). This solution ranks sixth without the 'double H-bond' constraint.

At positions 13:22:46, tRNA^{Phe} sequences are either C:G:G or U:G:G. The most isomorphic solution for these sequences is in agreement with the crystal structure, provided that the 'double H-bond' constraint is used (Fig. 6c). In tRNA^{Asp} species, the predominant sequences for this triple are U:G:A and C:G:G. The yeast tRNA^{Asp} crystal structure (20) is shown in Figure 6d (U:G:A sequence). ISOPAIR predicts this structure fourth in terms of *I* value for the sequence {U:G:A, C:G:G}, even when the 'double H-bond' constraint is used.

The predominant sequences at positions 10:25:45 in tRNA^{Phe} species are G:C:G and G:C:U. The yeast tRNA^{Phe} structure (Fig. 6e) cannot be identified by ISOPAIR using these sequences, whatever constraint is used. This result could be expected since (i) the peculiar single H-bond G:G interaction in this base triple does not follow ISOPAIR's rules for base pair construction, and (ii) the structure observed for G:C:G cannot form with sequence G:C:U, implying an absence of isomorphism at this position.

These test runs for tRNA triples (Fig. 6b–e) were all performed without using the 'major groove' constraint. This constraint imposes that solutions contain only major groove base triples (which is the case in all tRNA base triples). In the absence of the 'double H-bond' constraint, discarding minor groove triples slightly improves the ranking of the correct solutions in Figure 6b–d (data not shown).

DISCUSSION

We have presented a computed program (ISOPAIR) capable of seeking base pair conformations that are common to a given set of sequences. This program is intended for use in comparative sequence analysis when unusual base covariations are observed at specific RNA positions. Our underlying assumption was that any base covariation inferred from comparative analysis was amenable to a set of isomorphic base pair structures that could all form with a similar orientation of the sugar–phosphate backbone. This led to a definition of isomorphism based on comparisons of glycosyl bond positions and orientations.

Results obtained with tRNA base pairs and triples indicate that ISOPAIR may indeed be useful as an investigative tool for base pair conformations. Actual conformations often lie within the highest ranking solutions, although selecting the right solution cannot be guaranteed by the program. Most ISOPAIR runs generate multiple solutions (Fig. 4) and whether or not preference should be given to solutions with double H-bond pairs, pairs with *syn* or *anti* glycosyl bonds or pairs with a unique conformation remains an expert's task. Our tests with tRNA sequences suggest that correct solutions most often involve at least one double H-bond pair. However, parameters such as the number and variability of available sequences and prior knowledge of structural constraints in the vicinity of the base pair are essential in reaching a correct conclusion.

Another important factor to consider when interpreting sequence covariations is the nature of the constraints underlying an interaction. Our model explicitly seeks base pairs that can form with a similar orientation of the sugar–phosphate backbone. In many cases however, this constraint is not predominant or is combined with others. These involve purine:pyrimidine constraints, that account for covariations such as {A:U, G:C} or {A:U, G:C, G:U} observed in 16S and 23S rRNA (11), or the exposure of specific atoms to tertiary interactions or binding of external factors. Unusual covariations may also result from particular tertiary environments. For instance, certain pairing sequences could be excluded because they would result in unwanted tertiary interactions with surrounding residues. The ISOPAIR program can also be useful in identifying this variety of constraints, as it generates sets of similar structures that can be displayed or saved as three-dimensional coordinates for a detailed search of common structural properties.

ACKNOWLEDGEMENTS

R.R.G thanks the W. M. Keck Foundation for its support of RNA science on the Boulder campus. This work was supported in part by grants from the NIH (GM48207) and the Colorado RNA center to R.R.G.

REFERENCES

1 Holley,R., Agpar,J., Everett,G., Madison,J., Marquisee,M., Merrill,S., Penswick,J. and Zamir,A. (1965) *Science*, **147**, 1462–1465.
2 Fox,G. and Woese,C. (1975) *Nature*, **256**, 505–507.
3 Noller,H. and Woese,C. (1981) *Science*, **212**, 403–411.
4 Glotz,C., Zwieb,C. and Brimacombe,R. (1981) *Nucleic Acids Res.*, **9**, 3287–3306.
5 Michel,F., Jacquier,A. and Dujon,B. (1982) *Biochimie*, **64**, 867–881.
6 Davies,R., Waring,R., Ray,J., Brown,T. and Scazzocchio,C. (1982) *Nature*, **300**, 719–724.
7 Michel,F., Umesono,K. and Ozeki,H. (1989) *Gene*, **82**, 5–30.
8 James,B., Olsen,G., Liu,J. and Pace,N. (1988) *Cell*, **52**, 19–26.
9 Gutell,R.R. (1993) *Curr. Biol.*, **3**, 313–322.
10 Levitt,M. (1969) *Nature*, **224**, 759–763.
11 Gutell,R.R. (1996) Comparative sequence analysis and the structure of 16S and 23S RNA in Zimmerman,R.A. and Dahlberg,A.E. (eds) *Ribosomal RNA. Structure, Evolution, Processing and Function in Protein Biosynthesis.* CRC Press, pp. 111–128.
12 Bubienko,E., Cruz,P., Thomason,J.F. and Borer,P.N. (1983) *Prog. Nucleic Acids Res. Mol. Biol.*, **30**, 41–90.
13 Saenger,W. (1984) *Principles of Nucleic Acid Structure.* Springer Verlag, New York, p. 120.
14 Sprinzl,M., Dank,N., Nock,S. and Schon,A. (1991) *Nucleic Acids Res.*, **19**, 2127–2171.
15 Klug,A., Ladner,J. and Robertus,J. (1974) *J. Mol. Biol.*, **89**, 511–516.
16 Mazzara,G. and McClain,W. (1977) *J. Mol. Biol.*, **117**, 1061–1079.
17 Hou,Y.M., Westhof,E. and Giegé,R. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 6776–6780.
18 Gautheret,D., Damberger,S.H. and Gutell,R.R. (1995) *J. Mol. Biol.*, **248**, 27–43.
19 Quigley,G. and Rich,A. (1976) *Science*, **194**, 796–806.
20 Westhof,E., Dumas,P. and Moras,D. (1985) *J. Mol. Biol.*, **184**, 119–145.