Volume 16 Number 12 1988 Nucleic Acids Research

Fidelity of secondary and tertiary interactions in tRNA

Tim Haselman, Jack E.Chappelear and George E.Fox*

Department of Biochemical and Biophysical Sciences, University of Houston, Houston, TX 77004, USA

Received December 23, 1987; Revised and Accepted April 22, 1988

ABSTRACT

Contingency table analysis has previously been used to detect sequence correlations in RNAs caused by either secondary or tertiary interactions. An approach known as matrix reduction is developed here as an alternative to the usual Chi square test. This approach is especially sensitive to covariation between equivalent positions and is effective at detecting tertiary interactions that exhibit such covariation. Matrix reduction was also effective at detecting Watson-Crick base-pairs that exhibit a high degree of pairing fidelity. The method was applied to five closely related structural classes of tRNA and a number of tertiary interactions were detected in each class.

INTRODUCTION

The first successful application of comparative sequence analysis was the elucidation of tRNA secondary structure [1]. The ubiquitous cloverleaf secondary structure of these molecules provides a common sequence alignment for all areas of the molecule except for the somewhat variable D and V loops. In these more variable regions correct alignment relies on recognition of conserved bases as well [2]. As applied to regions of basepairing, the essence of the comparative approach is to locate pairs of positions at which the fidelity of base pairing is maintained throughout the aligned data set. This comparative approach has been extremely successful and beginning with 5S rRNA [3,4] has been used to elucidate secondary structure in all the rRNAs, as well as the RNA component of RNAse P [5]. The method and its limitations has been discussed in some detail on at least three occasions [6-8].

Less well known however is the fact that the base-base tertiary interactions of tRNA also exhibit covariations [9-11] similar to those seen in regions of secondary structure. In the case of the well known Levitt pair [12] the correlation is strong and follows the usual pairing rules, though the pair itself is of the reverse Watson-Crick type. Olsen [13] has pioneered an organized approach to recognizing such base-base interactions by analyzing four-by-

four contingency tables for each aligned pair of positions in a sequence set. A Chi square test was used to determine if the particular pair of positions shows significant deviation from random when base composition effects are accounted for. When applied to the tRNA sequences then available this method was useful in detecting both secondary and tertiary interactions [13]. The main limitation of this approach is that the number of correlations detected is too large. Effects such as historical sequence relationship (phylogeny) and amino acid specificity result in correlations that have no structural significance. When using this approach with a large molecule such as 16S rRNA it is necessary to closely examine a very large number of false positives to actually detect structurally interesting covariations. Alternatively a search for conformity with various rules at differing stringency levels has been used to attempt to identify base-base tertiary interactions in 16S rRNA [8]. In this paper we report a simple alternative to the Chi square test, matrix reduction, that decreases the noise encountered in the covarion analysis and use it to examine several structural classes of tRNA.

METHOD OF ANALYSIS

Design

The matrix reduction version of the covarion analysis conducted here begins with the usual four-by-four contigency table which summarizes the observed nucleotide occurrences at each pair of positions in the aligned sequence sets. The contigency table has four columns and four rows. The four possible nucleotides which can occur at position i in sequence n define the four columns. The rows are defined by the four possible nucleotides that can occur at a second column j. The nucleotide at position i determines which column is used to tally sequence n and the row is determined by the base at position j. Once the appropriate cell among the 16 possible is located the tally in that cell is incremented by one. The tallying process begins with the first sequence and is continued until all the sequences in the data set have been processed.

Once the tally is complete it is necessary to identify pairs of positions that behave in a correlated way. Intuitively the cell containing the largest tally is not interesting because it represents the baseline that all change is relative too. Tallys in those elements that are in the same row or column as the baseline element do not reflect simultaneous base change as one of the positions has not changed. The interesting elements from the covarion view then are those that are left after the baseline element and the first group containing nonsimultaneous change are eliminated. That is they are among the elements of a reduced three-by-three matrix. The most interesting of these is typically the largest element. Once

it is chosen several of the remaining elements again reflect nonsimultaneous change between the two columns. Further reduction to a two-by-two matrix eliminates these and again isolates the elements that are of most relevance from a covarion view. A final reduction reduces the matrix to one interesting element. This procedure is illustrated in Figure 1 for the base pair between positions 1 and 71 of the phenylalanine like tRNAs.

For each pair of positions, i and j, in the aligned sequences a covaration index , C_{ij}, is calculated according to the following formula:

$$C_{ij} = \frac{R_1 + R_2 + R_3}{N_t - B}$$

where R_1, R_2, and R_3 are the three reduction elements, B is the baseline element and N_t is the total number of comparisons for the pair of positions under consideration. At each stage of the reduction process the choice of the largest element may not ultimately be the one that leads to the largest value of the covaration index. Thus in the FORTRAN computer program used to calculate the covariation index this second order effect has

	BASE = 1				
		A	C	G	U	
	A	0	0	0	12	12
BASE = 71	C	0	0	43	0	43
	G	0	0	0	0	0
	U	15	0	4	0	19
		15	0	47	12	64

REDUCE MATRIX

	A	C	U	
A	0	0	12	12
G	0	0	0	0
U	15	0	0	19
	15	0	12	27

REDUCE MATRIX

	C	U	
A	0	12	12
G	0	0	0
	0	12	12

REDUCE MATRIX

	C	
G	0	0
	0	12

$$C(1,71) = \frac{15 + 12 + 0}{74 - 43} = 0.871$$

Figure 1. Example of matrix reduction routine

TABLE 1
VARIABLE LOOP AND STEM REGION OF tRNA

GROUP	D Loop (base)	D Stem (bpair)	V Loop (base)	Number of Sequences
1) D48V5	8	4	5	111
2) D48V4	8	4	4	60
3) D47V5	7	4	5	48 *
4) D47V4	7	4	4	59
5) D49V5	9	4	4	82 *

* excluding initiator tRNA

been allowed for by searching for the choices of B, R_1, R_2, and R_3 that maximize C_{ij}. The covariation number calculated in this way varies from zero (no simultaneous change) to one (perfect simultaneous change).

Sequences and alignment

tRNA sequences and tRNA gene sequences were obtained from the 1987 annual collections [2]. The sequences were subdivided into five major categories according to length variations in the D loop and variable region of the molecule. Additional sequences which typically contained a very large variable region or a D loop of unusual length, or both fell into a heterogeneous category that is not discussed further here because the data set for any structurally homogeneous subset is typically too small. Each category was screened for identical sequences and sequences exhibiting only one difference as these would have the effect of double weighting whatever sequence they displayed at each position. After removal of these duplicate sequences the 5 major structural categories, Table 1, contained a total of 360 sequences with the smallest class, D47V5, containing 48 sequences. In the case of Group 3 and 5, the initiator tRNAs were excluded. The assembly of the data base was facilitated by the availability of a specialized sequence editor (SEQEDT) that was developed and kindly provided to us by Dr. Gary Olsen of the Department of Microbiology at Indiana University. All programs were run on a Digital Equipment Corporation (Maynard, Mass) MicroVax II computer system with 9MB of main memory and 273 MB of storage under the VMS operating system.

Analyses Conducted

All five tRNA structural categories listed in Table 1 were separately analyzed by the matrix reduction routine. This distinction based on length variations was maintained in order to determine if such effects influenced the location of tertiary interactions. In addition, in order to obtain an indication of the number of sequences required to detect interest-

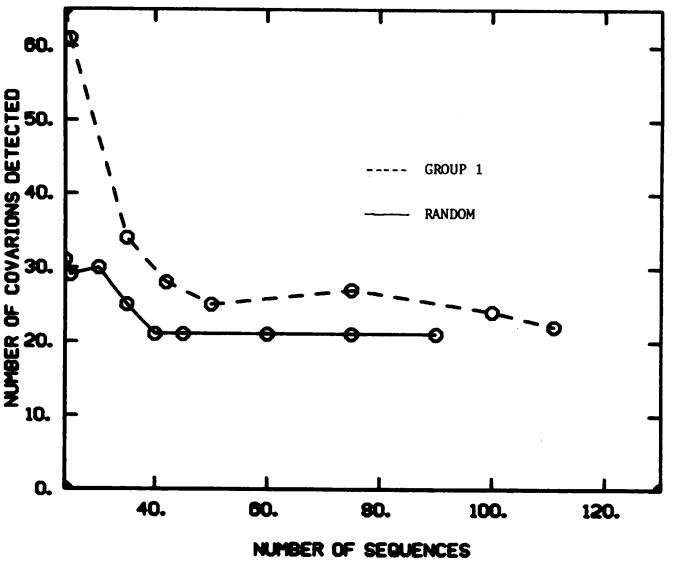

Figure 2. Effect of sequence number on covarion detection

ing correlations, a collection of randomly generated tRNA like sequences was examined. These sequences were all 72 bases long and all exhibited the usual cloverleaf secondary structure associated with the D48V5 tRNAs (phenylalanine like). This test sequence set was constructed by a random sequence generating program. Random numbers used in this program were obtained from the general random number generator in the Digital Equipment Corporation Fortran 4.0 Runtime library. This random number generator is of the multiplicative congruential type. The number of elapsed seconds since midnight was used to generate seed for the psuedo-number series.

RESULTS

MR Routine on Randomly Generated tRNA Sequences

Figure 2 graphically demonstrates the influence the quantity of tRNA sequences has on the number of covarions detected by the matrix reduction routine. As shown in this figure the number of correlations in which C_{ij} is greater than 0.5 quickly approaches 21, the number of base pairs in the test set, and reaches that level by the time 40 sequences are included in the data set. A similar comparison of the effect of data set size on the number of correlations found in D48V5 tRNAs is included in Figure 2. Covarions representing

TABLE 2

Distribution of Covariation Index Values for Various Groups of tRNA

C_{ij} Range	D48V5		D48V4		D47V5		D47V4		D49V5	
< 0.05	987		711		755		712		993	
.10	402		296		279		108		321	
.15	458		300		256		272		594	
.20	347		374		358		451		486	
.25	310		483		463		512		317	
.30	225		365		323		351		146	
.35	76		158		182		180		35	
.40	11		51		77		65		8	
.45	5	(5)	14		34		23		2	(2)
.50	2	(2)	1	(1)	15		4	* (3)	2	(1)
.55	5	(5)	0		11		3	* (1)	3	* (2)
.60	1	*	0		2	(2)	0		0	
.65	1		3	(2)	0		1		0	
.70	2		1		2		0		1	
.75	3	**	1		1		2		1	
.80	3		4	**	3		5		3	
.85	5		1		1		4	*	4	
.90	1		4		8	*	4		2	
.95	4		5		4		2		7	*
1.00	2	*	3		5	*	2		1	

(n) the number of false positives for categories near and below the breakpoint in the distribution

* a tertiary interaction is found in the associated category

no less then 95% of the crystallographically determined pairing fidelity (18 of 21 base pairs and 4 tertiary interactions) for Group 1 tRNAs are singled out after 50 or more tRNA sequences are examined. The somewhat slower decline in false positives encountered in real sequences presumeably reflects other biological effects. The obvious candidates being nearest neighbor effects, amino acid specificity and nonrandom distribution of the sequences (ie phylogenetic relationships).

The covariation index provided by the matrix reduction approach is relative. Higher values presumeably imply stronger correlation and hence greater probability of interaction than lower values. How does one tell what value of the index is interesting? The best starting place is with the distribution of index values. For a sequence of N bases the total number of binary comparisons is $\frac{N(N-1)}{2}$ which is 2850 for the D48V5 tRNAs. Table 2 shows the distribution of the correlation index values for all five tRNA classes. It is apparent in each case that the distribution is highly scewed with a breakpoint near 0.5. Beyond this there is an extended tail. Within this tail essentially every pair of correlated positions

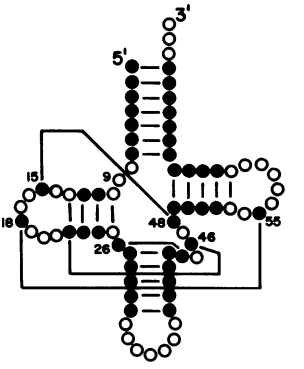

Figure 3. Group 1 sequences in secondary structure.
Open circle = no correlation, closed circle = correlated pair according to the matrix reduction covariation index.

is due to either a Watson-Crick base-pair or a tertiary interaction. Table 2 indicates false positives for all categories in the distribution tail up to the breakpoint. It is apparent from this that the false positives are small in number and localized near the breakpoint in the distribution. The tertiary interactions, as indicated in Table 2, are often among the strongest correlations. A cutoff value for the correlation index might be developed based on statistics of the distribution or a functional criterion such as the point at which 95% of the known secondary structure features are detected. In assessing actual results however it has been found prudent to evaluate the correlations with the actual distribution on hand.

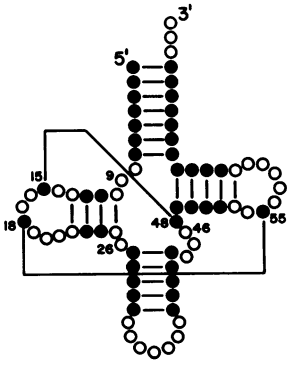

Figure 4. Group 2 sequences in secondary structure.
Open circle = no correlation, closed circle = correlated pair according to the matrix reduction covariation index.

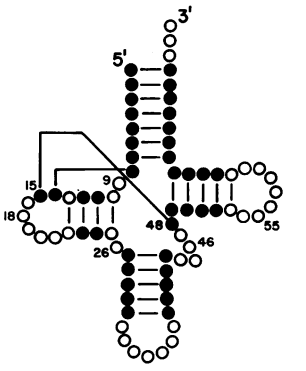

Figure 5. Group 3 sequences in secondary structure.
Open circle = no correlation, closed circle = correlated pair according to the matrix reduction covariation index.

MR routine on grouped tRNA sequences

The matrix reduction algorithm was applied to each of the five tRNA structure groups. Those pairs of positions which gave high covariation indexes (typically > 0.5) were examined in detail. As indicated on Figures 3-7 essentially all the standard base pairs and a number of tertiary interactions were detected. Typically the putative tertiary interac-

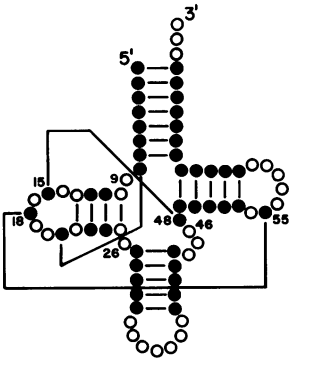

Figure 6. Group 4 sequences in secondary structure.
Open circle = no correlation, closed circle = correlated pair according to the matrix reduction covariation index.

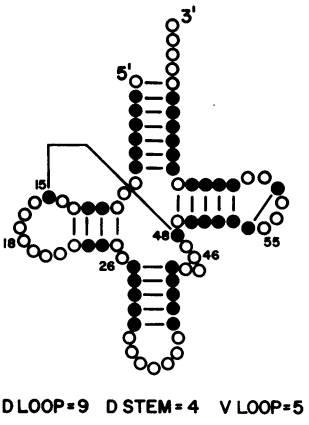

Figure 7. Group 5 sequences in secondary structure.
Open circle = no correlation, closed circle = correlated pair according to the matrix reduction covariation index.

tions detected in the various classes of tRNAs involved nucleotides that were equivalent to those involved in tertiary interactions in the known structures of class D48V4 and D48V5 tRNAs. A novel correlation that can be reasonably interpreted as a tertiary interaction was detected and is discussed in detail in the next section. Assuming that the tertiary interactions have been properly assigned it is clear that false positives are extremely rare in the tail of the distribution.

DISCUSSION

The matrix reduction approach is designed to identify the fidelity of covariation seen between each pair of positions in alignable RNA sequences. The results of the approach are discussed here with reference to the universal numbering system for tRNA [14]. Application to the phenylalanine like tRNAs (class D48V5) whose tertiary structure is well known was conducted first with the presumption that all the sequences in this group would exhibit the identical set of secondary and tertiary base-base interactions. The results, Figure 3, demonstrate that most base-pairs exhibit enough Watson-Crick pairing fidelity to be readily detected by the matrix reduction algorithm. Exceptions were of two types. The first was the conserved base-pair G53-C61 which could not be detected due to the complete absence of variability. The other are base pairs (10/25 and 13/22) that exhibit a high proportion of wobble pairing (G-U or U-G). In this latter situation only modest values of

the correlation index are obtained for all five groups because the replacement of a Watson-Crick pair with a wobble pair does not involve simultaneous change. In the case of tertiary interactions the approach proved very effective in identifying interactions whose essence is correlated change. Thus in the case of the D48V5 tRNAs four tertiary interactions were readily identifiable with essentially no background noise. The tertiary interactions that were not found included three (G19C56, U8A14, and T54A58) that exhibited no variability in the data set and two (G10G45 and A9A23) that followed either complex rules that encompass more than simultaneous variation or were not always present. One of these, (A9A23) was detected indirectly through a consequence of the pairing between U12 and A23 as a correlation between A9 and U12. This type of correlation has been referred to as transitive by Olsen [13] .

Comparison with the Chi square approach [13] emphasizes that different approaches will fare better in detecting different types of correlations. The Chi square analysis focuses on detecting deviations from random with no preference to the cause. It consequently fares better than the matrix reduction approach when applied to base pairs as it easily detects pairs which frequently involve wobble pairing. This success is at a price however. Correlations due to base pairing are invariably far stronger than any of the the correlations due to tertiary interaction. The tertiary interactions that are detected are at a substantially lower level where they comingle with a significant number of correlations that do not have an obvious structural orgin. When matrix reduction is used the tertiary interactions typically have correlation indexes comparable to the base-pairs, Table 2, and are not as often lost among false positives.

A different subset of the known tRNA tertiary interactions was detected in each of the five groups of tRNAs. The Levitt pair, G15C48, is the most easily detected and its analog was found in all five tRNA classes. The G18ψ55 pair between the D and T loops is quite conserved but what variability exists allows its detection in three groups, D48V5, D48V4 and D47V4. Likewise minor variability allowed the detection of U8A14 in the D47V5 tRNAs. The G26A44 and the triplet interaction involving G46 were both detected in the D48V5 tRNAs.

It is instructive to consider the correlation results in terms of known high resolution data. Crystallographic studies on aspartic acid tRNA [15] which belongs to the D48V4 group revealed that a set of tertiary interactions very similar to those seen in the phenylalanine tRNA were present. Two major exceptions were found. One of these involved the G18ψ55 interaction which appears to be missing in the aspartic acid tRNA. The authors attributed this to the existence of a codon-anticodon interaction between tRNAs in the

crystal structure. In subsequent solution studies [16] strong evidence for this interaction was obtained and this is supported by the correlation seen between these two positions in the D48V4 data set examined here. Of most relevance to the current work is the effect of the deletion of nucleotide 47 in the D48V4 tRNAs. The crystallographic data [15] indicates that the primary effect is localized to the neighboring bases as a rather substantial repositioning of nucleotides 46 and 48 occurs. The changes in these two residues induce different environments for several residues that adjoin these two in the three dimensional structure and thereby the effect of the deletion extends to residues that are distant in the primary sequence. Thus a new base-base interaction is found between A21 and U8 of the U8A14 interaction. In yeast phenylalanine tRNA these bases are close but the only interaction is between the sugar of U8 and the base of A14. This new tertiary interaction is not detectable among the D48V4 sequences by matrix reduction but is seen in the D47V4 sequences, Figure 6, that also have residue 47 deleted.

ACKNOWLEDGEMENTS

We are grateful to Dr. Gary Olsen for inspiring our initial interest in correlation analysis and for providing his programs for our use. Also, Dr. Dan Davison for the idea of randomly genereated tRNA like sequences. This work was supported by US ARMY grant DAAL03-86-G-0031, NIH grant GM37655 and NASA grant NSG7440 to G.E.F.

*To whom correspondence should be addressed

REFERENCES

1. Holley, R.W., Apgar, J., Everett, A., Madison, J.T., Marquisse, M., Merrill, S.H., Penwick, J.R., Zamir, A. (1965) Science, *147*, 1462.
2. Sprinzl, M., Hartmann, T., Meissner, F., Moll, J., Vorderwulbecke, T. (1987) Nucl. Acids Res. Supl. *15*, 53-175.
3. Fox, G.E., and Woese, C.R. (1975) Nature *256*, 505-507.
4. Nishikawa, K., and Takemura, S. (1974) FEBS Lett. *40*, 106-109.
5. James, B.D., Olsen, G.J., Liu, J., and Pace, N.R. (1988) Cell *52*, 19-26.
6. Noller, H.F., (1984) Ann. Rev. Biochem. *53*, 119-162.
7. Fox, G.E. (1985) The Bacteria (C.R. Woese & R.S. Wolfe eds.) *8*, 257-310.
8. Gutell, R., Weiser, B., Woese, C.R. and Noller, H.F. (1985) Prog. Nucl. Acids Res. & Mol. Biol. *32*, 155-216.
9. Rich, A. and RajBhandary, U.L. (1976) Annual Rev. Biochem., *45*. 805-860.
10. Clark, B.F.C. (1978) Transfer RNA, Mit Press, 14-47.
11. Levitt, M. (1969) Nature *224*, 759-763.
12. Brennan, T. and Sundaralingam, M. (1976) Nucl. Acids Res. *3*, 3235-3251.
13. Olsen, G.J. (1984) Comparative analysis of nucleotide sequence data, Ph.D. Dissertation, University Colorado Health Sciences Center, 88-137.

14. Schimmell, P. R., Söll, D., Abelson, J. N. (1979) Transfer RNA: Structure, Properties and Recognition, Cold Spring Harbor Laboratory, N.Y., 518-519.
15. Westof, E., Dumas, P., Moras, D. (1985) J. Mol. Biol. *184*, 119-145.
16. Romby, Pascale, Moras, D., Dumas, P., Ebel, J.B., Giege, R. (1987) J. Mol. Biol. *195* (1), 193-204.