Detection of Correlations in tRNA Sequences with Structural Implications

Tod M. Klingler and Douglas L. Brutlag

Department of Biochemistry and Section on Medical Informatics
Stanford University School of Medicine
Stanford, CA 94305-5307
klingler@cmgm.stanford.edu, brutlag@cmgm.stanford.edu

Abstract

Using an flexible representation of biological sequences, we have performed a comparative analysis of 1208 known tRNA sequences. We believe we our technique is a more sensitive method for detecting structural and functional relationships in sets of aligned sequences because we use a flexible representation (for sequences), as well as a general statistical method that can detect a wide range of relationships between positions in a sequence. Our method utilizes functional classifications of the sequence building-blocks (nucleotide bases and amino acids) based on physical or chemical properties. This flexibility in sequence representation improves the significance of finding sequence relationships mediated by the defining property. For example, using a purine/pyrimidine classification, we can detect base-stacking interactions in sets of nucleotide sequences that form base-paired helices. We use several statistical measures, including χ^2-tests, Monte Carlo simulations and an information measure to detect significant correlations in sequences. In this paper we illustrate our method by analyzing a set of tRNA sequences and showing that the correlations our program discovers, in each case, correspond to the known base-pairing and higher order interactions observed in tRNA crystal structures. Furthermore, we show that novel and interesting features of tRNAs are detected when sequence correlations with the charged amino acid (and anticodon) are evaluated. This technique is a powerful method for predicting the structure of RNAs and for analyzing specific functional characteristics.

Introduction

In the last few years, RNA molecules have reemerged as much more complex and dynamic molecules than previously thought. Functionally, the catalytic properties of self-splicing introns, RNase P, splicosomal RNAs and other ribozymes, as well as the structural properties of ribosomal RNAs, have captivated molecular biologists who had traditionally viewed RNAs simplistically. The discovery of pseudo-knots, tetra-loops and non-standard base-pairing have had a similar effect in challenging the view of RNA structure as being the simple composition of helical elements (Woese and Gutell 1989, Gutell, et al. 1992, Woese, et al. 1990). The result is a much more complicated view of RNA structures and their properties, and a realization that their importance for biological function has been underestimated.

As is the case with proteins, it has become imperative to understand the structure of RNAs in order to understand their function. The most successful technique for the prediction of RNA secondary structure is the analysis of an aligned set of sequences. This method of comparative analysis was first successfully demonstrated for the tRNAs by Holley *et al.* (Holley, et al. 1965). They observed regions of sequence that covaried according to a base-pairing scheme in just a few tRNA sequences, and from these were able to construct the clover-leaf model (Fig 1). The significance of their model (i.e. the probability of the sequence patterns arising randomly) was judged to be quite small, and was further validated by finding similar patterns in subsequent tRNA sequences. A few years later Levitt (Levitt 1969) used 14 tRNA sequences to generate a full tertiery model of the tRNA molecule. Although his overall model was not completely accurate, he correctly predicted a 3-way interaction and several isolated base-pairs between conserved positions in the structure. Full confirmation of Holley's clover-leaf model and Levitt's tertiery interactions came

with the crystal structures of the tRNA molecule (Sussman, et al. 1978).

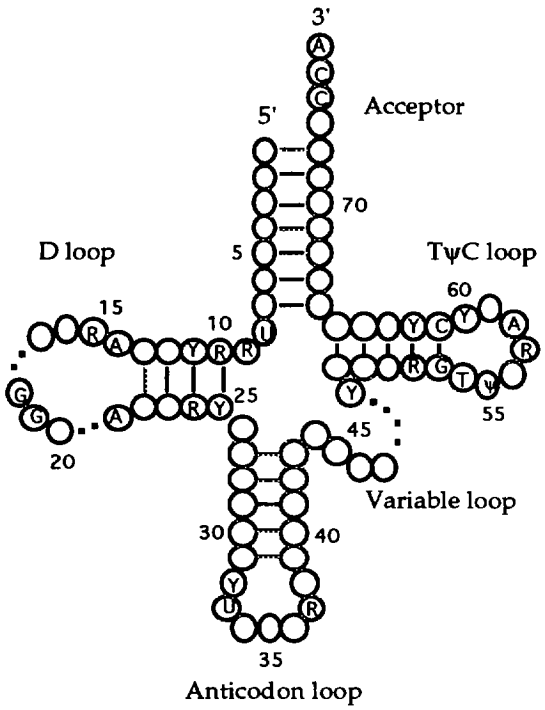

Figure 1: tRNA clover leaf diagram

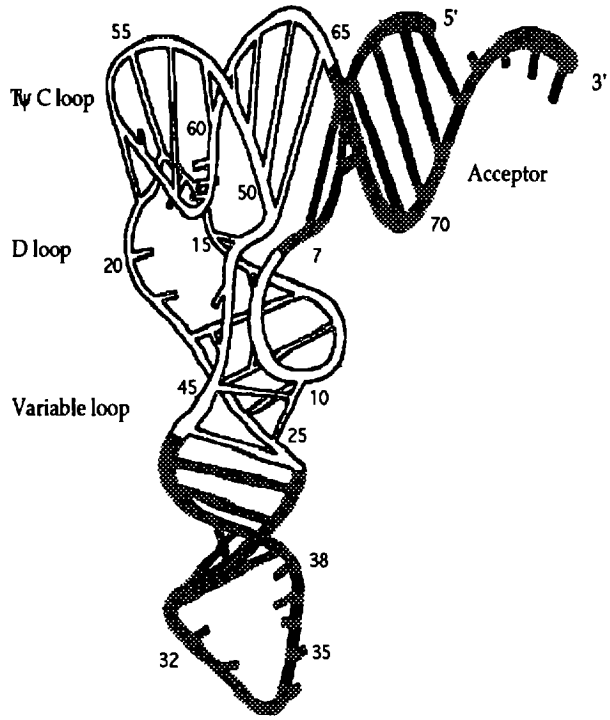

Figure 2: tRNA structure diagram

Although originally one was only able to generate predictions of secondary structure, i.e. the base-paired, helical regions represented in the clover-leaf diagram (Fig. 1), with the explosion of sequence information and the use of more sophisticated mathematical and statistical methods, elements of tertiary structure (Fig. 2) can now be automatically detected. Recently, the secondary structure of several ribosomal RNAs, both groups of self-splicing introns and other molecules have been predicted. Few predictive methods have offered reliable and quantitative information on tertiary interactions, however.

In a recent paper, Gutell, *et al.* (1992) are able to detect most of the tertiary interactions seen in the tRNA structures using a measure of mutual information (Gutell, et al. 1992). These included the two 3-way base interactions and several isolated base-pairs. In this paper, we confirm the results of their paper using an alternative algorithm. We also demonstrate the power of using multiple representations of sequences to detect finer structural details and functional properties.

Materials and Methods

Input for our correlation analysis program, MCSEQ, is a data set of aligned sequences. We used 1208 tRNA sequences compiled by Sprinzl *et al.* and retrieved electronically from the EMBL database (Sprinzl, et al. 1991). This set was originally compiled by aligning sequences based on maximum sequence similarity alone, i.e. positions were compared independently. Later sequences have been aligned using the knowledge of paired regions. In this sense, one could view parts of our analysis to be circular since we will report the detection of base-paired helical regions—knowledge of which was used to align some of the sequences in the first place. However, these regions are well understood and proven, both analytically and experimentally and are not the main focus of this work. Detecting the tertiary interactions, which were not used in any way in constructing the input, are the main result of this work.

MCSEQ performs an exhaustive search on all pairs of positions in the input, evaluating each for non-independence. We assume a null hypothesis that any two positions are uncorrelated, and reject this

hypothesis if a statistical test of independence exceeds a pre-chosen level of significance. Our primary test is the χ^2-test, but we also make use of Monte Carlo simulations and a measure of mutual information.

For each pair of positions we construct a 4x4 contingency table containing the numbers of each sequence pair seen in the two positions in the data set. For example, if two positions were perfectly (and uniformly) base-paired, the contingency table would look like:

	A	C	G	T	Rows
A	0	0	0	302	302
C	0	0	302	0	302
G	0	302	0	0	302
T	302	0	0	0	302
Cols	302	302	302	302	1208

A χ^2-statistic is calculated for each table to test for non-independence. For each data set mentioned below, we used a significance threshold of $1x10^{-50}$. This number was arbitrarily chosen to (1) overcome the loss of significance due to repeated application of statistical tests, and (2) to restrict the number of significant correlations to a manageable number. The tRNA sequences in our input are 96 bases long. 20 of these positions are inserts present in fewer than 10% of the data set. Thus, for the 76 core positions, we perfomed 76 x 75 / 2 = 2850 tests in looking for relationships between the positions of the tRNAs.

We also report several other measures of the strength of each significant correlation. The first two are transformations of the χ^2-statitistic: Cramer's V and the contingency coefficient C . We also report a symmetric uncertainty coefficient U_{xy} which is closely related to the parameter $R(x,y)$ of Gutell et al. (Gutell, et al. 1992). Each of these measures lie between zero and one, with values closer to one signifying strong correlations, and zero corresponding to the absence of a correlation. All of the statistical results are generated using algorithms based on those described in (Press, et al. 1988).

Unreported here is the Monte Carlo simulations we perform for each significant correlation in order to validate its significance. For the simulation we construct a large number (typically 1000) of simulated contingency tables from the base distributions of the two positions being analyzed, and using the independence assumption. Thus, we maintain base compositions at each position but randomize any covariation that might be present. Each simulated data set is tested as described above to determine the rate of detecting a correlation at the reported significance by random chance. For all correlations described in this paper, none of the 1000 simulated tables reached the significance of the actual sequence data, which confirms the significance at least the $p=10^{-3}$ level. Tables II through VI contain the results of running MCSEQ on the tRNA sequences looking for covariation in nucleotide sequences. Note that the analysis was run on the *gene* sequences for each tRNA, so the bases used consist of A, C, G and T (i.e. not U and the modified nucleotides).

MCSEQ only detects the presence of generic covariations in the data sets it is given. There is no inherent ability to detect specific types of interactions, such as base-pairing, from these correlations. However, one can infer the type of interaction by recognizing characteristic patterns in the contingency tables. MCSEQ displays the contingency tables for each significant correlation, with the log-likelihood in parentheses to give an indication of how far from independence each bin in the table is. In this paper we are evaluating the validity of our results in the context of the known tRNA structure. But we would like to suggest that the reverse is possible—inferring structure from the patterns of significant correlations.

Two variations of the basic analysis described above were also tested. In the first, a correlation search was performed on the same data set, but considered only purines and pyrimidines. An important feature of MCSEQ is the ability to define classifications of sequence characters. This has the effect of focusing the correlation search on specific types of sequence differences (strong/weak or amino/keto can also be used), often detecting otherwise unnoticed interactions. In a previous paper we describe and demonstrate the power of this idea for protein sequences (Klingler and Brutlag 1993). Table VII contains the results of this analysis.

Second, we performed two searches with MCSEQ looking for covariation between the tRNA sequence and the cognate amino acid. For each of these searches, 76 tests were performed—one for each position. The cognate amino acids were classified in two different ways based on the genetic code: (1) using second position groups, and (2) using first position groups. These two analyses also demonstrate the generality and power of MCSEQ in its ability to search for a wide range of correleations. The results of these two analyses from MCSEQ are reported in Tables VIII and IX, respectively.

Results

An example of the results from MCSEQ on base correlations in the tRNAs appear in Tables I, where 6 of the base-pairing correlations of the acceptor stem are listed. Similar tables were generated for the D stem, the anticodon stem, and the TψC stem. These correlations are the strongest observed for this data set, most having significances below the resolution of computer ($< 10^{-250}$).

Pos. 1 vs. 72	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	1(-2.9)	3(-3.6)	3(-1.5)	**182**(1.3)	189
Cytosine	2(-1.4)	2(-3.2)	**77**(2.6)	2(-2.3)	83
Guanine	2(-3.6)	**674**(0.4)	2(-3.3)	103(-0.6)	781
Thymine	**106**(2.3)	1(-4.2)	2(-1.4)	5(-1.7)	114
Col Sum	111	680	84	292	1167

p= 0.0000e+00, χ2= 2617.69, df= 9, V= 0.86, C= 0.83, Uxy= 0.69

Pos. 2 vs. 71	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	1(-3.1)	1(-4.1)	1(-4.1)	**176**(1.6)	179
Cytosine	5(-2.3)	2(-4.2)	**386**(1.1)	4(-3.0)	397
Guanine	2(-3.3)	**396**(1.0)	2(-4.3)	40(-0.8)	440
Thymine	**137**(2.0)	1(-4.0)	7(-2.0)	7(-1.4)	152
Col Sum	145	400	396	227	1168

p= 0.0000e+00, χ2= 2886.97, df= 9, V= 0.91, C= 0.84, Uxy= 0.81

Pos. 3 vs. 70	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	5(-2.0)	3(-3.1)	1(-4.0)	**201**(1.4)	210
Cytosine	4(-2.6)	2(-3.9)	**295**(1.3)	2(-3.6)	303
Guanine	1(-4.3)	**366**(1.0)	3(-3.7)	68(-0.4)	438
Thymine	**194**(1.6)	2(-3.5)	15(-1.4)	6(-2.1)	217
Col Sum	204	373	314	277	116

p= 0.0000e+00, χ2= 2712.72, df= 9, V= 0.88, C= 0.84, Uxy= 0.76

Pos. 4 vs. 69	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	6(-2.1)	10(-2.0)	1(-4.2)	**224**(1.3)	241
Cytosine	4(-2.6)	6(-2.5)	**246**(1.3)	3(-3.0)	259
Guanine	2(-3.6)	**327**(1.1)	2(-3.9)	52(-0.6)	383
Thymine	**220**(1.4)	2(-3.7)	59(-0.2)	4(-2.8)	285
Col Sum	232	345	308	283	1168

p= 0.0000e+00, χ2= 2456.44, df= 9, V= 0.84, C= 0.82, Uxy= 0.69

Pos. 5 vs. 68	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	5(-2.6)	7(-2.2)	1(-4.6)	**336**(1.0)	349
Cytosine	7(-2.1)	5(-2.4)	**269**(1.2)	3(-3.5)	284
Guanine	3(-2.8)	**204**(1.4)	4(-2.9)	55(-0.5)	266
Thymine	**209**(1.4)	2(-3.2)	49(-0.4)	8(-2.4)	268
Col Sum	224	218	323	402	1167

p= 0.0000e+00, χ2= 2448.44, df= 9, V= 0.84, C= 0.82, Uxy= 0.69

Pos. 6 vs. 67	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	2(-3.6)	1(-3.8)	4(-2.7)	**232**(1.3)	239
Cytosine	6(-2.4)	1(-3.8)	**212**(1.4)	4(-2.7)	223
Guanine	3(-3.3)	**222**(1.5)	1(-4.1)	34(-0.7)	260
Thymine	**346**(0.9)	3(-3.4)	66(-0.5)	30(-1.3)	445
Col Sum	357	227	283	300	1167

p= 0.0000e+00, χ2= 2460.99, df= 9, V= 0.84, C= 0.82, Uxy= 0.68

TABLE I: Acceptor stem correlations

Table II contains the correlations defining the two 3-way base interactions. These two 3-way interaction are detected as significant pairwise corelations in all three pairings of the three positions. Furthermore,

both of the triplet arrangements (T=A–A and C=G–G) are observed in both sets of contingency tables: the former is in italics and the latter is in bold.

A.

Pos. 13 vs. 22	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	100(0.4)	5(0.2)	15(-1.6)	43(0.9)	163
Cytosine	12(-2.6)	4(-0.9)	**372**(0.7)	4(-2.4)	392
Guanine	152(0.7)	10(0.8)	7(-2.5)	19(-0.1)	188
Thymine	*202*(0.2)	9(-0.1)	130(-0.3)	62(0.3)	403
Col Sum	466	28	524	128	1146

p= 9.3318e-138, χ2= 666.83, df= 9, V= 0.44, C= 0.61, Uxy= 0.29

Pos. 13 vs. 46	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	43(-0.2)	25(1.0)	24(-1.2)	73(1.0)	165
Cytosine	35(-1.2)	4(-1.7)	**345**(0.6)	8(-2.1)	392
Guanine	52(-0.1)	29(1.0)	44(-0.7)	63(0.7)	188
Thymine	*218*(0.6)	8(-1.1)	130(-0.4)	44(-0.4)	400
Col Sum	348	66	543	188	1145

p= 1.0616e-123, χ2= 601.37, df= 9, V= 0.42, C= 0.59, Uxy= 0.21

Pos. 22 vs. 46	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	*229*(0.5)	53(0.7)	48(-1.5)	135(0.6)	465
Cytosine	8(-0.1)	3(0.6)	13(0.0)	4(-0.1)	28
Guanine	66(-0.9)	6(-1.6)	**435**(0.6)	17(-1.6)	524
Thymine	45(0.2)	3(-0.9)	48(-0.2)	31(0.4)	127
Col Sum	348	65	544	187	1144

p= 5.6880e-111, χ2= 542.03, df= 9, V= 0.40, C= 0.57, Uxy= 0.24

B.

Pos. 9 vs. 12	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	96(0.0)	60(-0.9)	49(-0.7)	*525*(0.3)	730
Cytosine	5(0.0)	30(1.3)	4(-0.4)	2(-2.4)	41
Guanine	36(-0.2)	**119**(0.6)	105(0.8)	77(-0.9)	337
Thymine	7(0.4)	14(0.6)	4(-0.3)	13(-0.5)	38
Col Sum	144	223	162	617	1146

p= 4.2388e-77, χ2= 383.63, df= 9, V= 0.33, C= 0.50, Uxy= 0.16

Pos. 9 vs. 23	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	*522*(0.3)	45(-0.8)	63(-0.9)	101(0.2)	731
Cytosine	2(-2.4)	4(-0.3)	30(1.3)	4(-0.2)	40
Guanine	73(-0.9)	108(0.8)	**125**(0.6)	30(-0.3)	336
Thymine	15(-0.3)	2(-0.9)	19(0.9)	1(-1.5)	37
Col Sum	612	159	237	136	1144

p= 2.8760e-84, χ2= 417.23, df= 9, V= 0.35, C= 0.52, Uxy= 0.17

Pos. 12 vs. 23	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	8(-2.2)	6(-1.2)	3(-2.3)	125(2.0)	142
Cytosine	10(-2.5)	1(-3.4)	**214**(1.5)	1(-3.3)	226
Guanine	5(-2.9)	150(1.9)	3(-2.4)	4(-1.6)	162
Thymine	*590*(0.6)	2(-3.8)	20(-1.9)	6(-2.5)	618
Col Sum	613	159	240	136	1148

p= 0.0000e+00, χ2= 2816.67, df= 9, V= 0.90, C= 0.84, Uxy= 0.77

TABLE II: Tertiary pairing correlations for 13,22,46 (A) and 9,12,23 (B)

Table III lists the correlations detected that correspond to isolated base-pairs in the tRNA structures. Only one of the representative contingency tables is shown. The base-pairing pattern is not as easily recognizable for these correlations, although it is clearly a reflection of a

base-pairing interaction subject to the low sequence variability of the composite positions. The negative log-likelihoods both above and below, and to the left and right of the dominant bin indicate that indicate that the base-pair is conserved beyond the base composition.

Pos 8 vs. 14 **Pos 18 vs. 55** **Pos 54 vs. 58**
Pos 15 vs. 48 **Pos 19 vs. 56**

Pos. 15 vs. 48	Adenine	Cytosine	Guanine	Thymine	Row Sum
Adenine	27(0.2)	34(-1.9)	8(-0.1)	**313**(0.9)	382
Cytosine	1(-0.2)	6(-0.7)	5(2.3)	9(0.2)	21
Guanine	11(-1.2)	**583**(0.5)	8(-0.6)	16(-2.6)	618
Thymine	24(1.7)	12(-1.3)	4(0.9)	33(0.3)	73
Col Sum	63	635	25	371	1094

p= 1.6494e-193, χ2= 925.87, df= 9, V= 0.53, C= 0.68, Uxy= 0.48

TABLE III: Isolated base pair correlations

Table IV lists all of the neighboring correlations found when the purine/pyrimidine classification of bases was used in MCSEQ. Only one of the actual contingency tables is shown. Many more correlations were found to be significant in this representation. However, they are not reported here if they if they are redundant (i.e. appear as base correlations in any of the previously mentioned tables). For example, all base-pairing stem interactions were detected in this analysis as expected. Filtering out these correlations left a set consisting of mostly neighboring pairs, all but one with the same pattern—similar bases tend to stack next to each other. The one exception (positions 36 and 37) is not part of a helical stem where base-stacking has a greater affect in avoiding purine clashing.

Pos 2 vs. 3 **Pos 18 vs. 19** **Pos 68 vs. 69**
Pos 3 vs. 4 **Pos 36 vs. 37** **Pos 69 vs. 70**
Pos 4 vs. 5 **Pos 60 vs. 61** **Pos 70 vs. 71**

Pos 68 vs. 69	Purine	Pyrimidine	Row Sum
Purine	**370**(0.4)	184(-0.5)	554
Pyrimidine	165(-0.6)	**493**(0.3)	658
Col Sum	535	677	1212

p= 4.4911e-48, χ2= 212.23, df= 1, V= 0.42, C= 0.39, Uxy= 0.13

TABLE IV: Base-stacking correlations

Tables V contains correlations between positions in the tRNA sequences and the type of amino acid charged to those sequences. In the two analyses comprising this table we were looking for positions in the tRNA that vary with the amino acid it carries. We expected to see the anticodon positions show up as well as the discriminator base (Rould, et al. 1989). But we wanted to test other positions for functional significance with MCSEQ. For example, bases that confer some aspect of specficity in synthetase- or ribosome-binding would be detected with these analyses. In the table the amino acids are classified into four groups according to the prevalent amino acids encoded for by individual bases of the anticodon. Table V contains correlations with the amino acids grouped according to the base in the middle position of the anticodon (FLIMV, SPTA, YHQNKDE and CWRG), and correlations with the amino acids grouped according to the base in the first position of the anticodon (FSYCW, LPHQR, IMTNK and VADEG). As expected the strongest correlations found were the respective anticodon positions. The next strongest correlation involves position 73, which has been termed the discriminator base and is postulated to interact with the tRNA synthetase, affecting its specificity (Crothers, et al. 1972). The remaining significant correlations, listed in Table V, are non-overlapping and all occur in positions that could interact with the tRNA synthetase based on proximity in the co-crystal structure (Rould, et al. 1989).

Position 12	Adenine	Cytosine	Guanine	Thymine	Row Sum
VADEG	37(0.0)	79(0.4)	29(-0.3)	139(-0.1)	284
LPHQR	52(0.3)	82(0.3)	91(0.7)	93(-0.6)	318
FSYCW	10(-1.1)	18(-1.0)	8(-1.4)	207(0.5)	243
IMTNK	42(0.2)	43(-0.3)	30(-0.3)	175(0.1)	290
Col Sum	141	222	158	614	1135

p= 3.5936e-41, χ2= 214.13, df= 9, V= 0.25, C= 0.40, Uxy= 0.08

Position 13	Adenine	Cytosine	Guanine	Thymine	Row Sum
VADEG	43(0.1)	65(-0.4)	98(0.8)	77(-0.3)	283
LPHQR	69(0.4)	61(-0.6)	56(0.1)	132(0.2)	318
FSYCW	18(-0.7)	157(0.6)	17(-0.8)	51(-0.5)	243
IMTNK	32(-0.3)	105(0.0)	13(-1.3)	140(0.3)	290
Col Sum	162	388	184	400	1134

p= 1.7915e-49, χ2= 253.53, df= 9, V= 0.27, C= 0.43, Uxy= 0.08

Position 31	Adenine	Cytosine	Guanine	Thymine	Row Sum
VADEG	187(0.5)	27(-1.1)	27(-0.6)	59(0.0)	300
LPHQR	143(0.2)	61(-0.3)	61(0.1)	54(-0.1)	319
FSYCW	67(-0.3)	50(-0.2)	71(0.6)	55(0.2)	243
IMTNK	44(-0.9)	163(0.8)	32(-0.4)	51(-0.1)	290
Col Sum	441	301	191	219	1152

p= 1.1477e-55, χ2= 282.81, df= 9, V= 0.29, C= 0.44, Uxy= 0.09

Position 38	Adenine	Cytosine	Guanine	Thymine	Row Sum
VADEG	275(0.3)	18(-1.0)	2(-1.3)	5(-2.0)	300
LPHQR	175(-0.2)	49(-0.1)	9(0.1)	86(0.7)	319
FSYCW	210(0.2)	21(-0.7)	11(0.6)	1(-3.4)	243
IMTNK	120(-0.5)	109(0.8)	6(-0.2)	55(0.4)	290
Col Sum	780	197	28	147	1152

p= 2.6287e-60, χ2= 304.70, df= 9, V= 0.30, C= 0.46, Uxy= 0.12

Position 39	Adenine	Cytosine	Guanine	Thymine	Row Sum
VADEG	56(0.3)	36(-0.3)	30(-1.1)	178(0.4)	300
LPHQR	38(-0.1)	62(0.1)	65(-0.4)	154(0.2)	319
FSYCW	27(-0.2)	69(0.5)	71(0.0)	75(-0.2)	242
IMTNK	33(-0.2)	29(-0.5)	186(0.7)	42(-1.0)	290
Col Sum	154	196	352	449	1151

p= 8.9163e-57, χ2= 288.05, df= 9, V= 0.29, C= 0.45, Uxy= 0.09

Position 70	Adenine	Cytosine	Guanine	Thymine	Row Sum
CWRG	93(0.6)	77(-0.2)	85(0.1)	42(-0.5)	297
SPTA	20(-1.0)	70(-0.3)	69(-0.2)	141(0.7)	300
FLIMV	59(0.0)	119(0.1)	116(0.2)	52(-0.5)	346
YHQNKDE	28(-0.3)	103(0.4)	40(-0.3)	38(-0.3)	209
Col Sum	200	369	310	273	1152

p= 1.7948e-36, χ2= 191.72, df= 9, V= 0.24, C= 0.38, Uxy= 0.06

Position 72	Adenine	Cytosine	Guanine	Thymine	Row Sum
CWRG	6(-1.5)	179(0.0)	14(-0.4)	98(0.3)	297
SPTA	4(-1.9)	221(0.2)	37(0.6)	38(-0.7)	300
FLIMV	89(1.0)	170(-0.2)	27(0.1)	60(-0.4)	346
YHQNKDE	8(-0.9)	106(-0.1)	2(-2.0)	93(0.6)	209
Col Sum	107	676	80	289	1152

p= 7.7943e-50, χ2= 255.24, df= 9, V= 0.27, C= 0.43, Uxy= 0.09

TABLE V: Correlations with codon positions

Discussion

We have developed a method that has successfully detected most of the known structural interactions present in the tRNA structure. Using standard statistical tests we have detected significant correlations in an aligned set of 1208 tRNA sequence that correspond to base-pairing in helical regions, 3-way base interactions and isolated base-pairs. Additionally, we have detected base-stacking interactions using an alternate encoding of the tRNA sequences. Our method, which is primarily based on the χ^2-test with validation from information theory and Monte Carlo simulations, finds these structural correlations with high specificity.

The correlations detected for the base-paired helices in the tRNA molecule are unmistakable. There is a strong diagonal representing standard Watson-Crick base-pairing in all tables, and a minor contribution of acceptable G-U paring in some tables. These correlations patterns, the sequence in which they were detected and the base-stacking interactions are more than enough evidence to predict helical stems in sequences of unknown structure.

We also postulate that 3-way base interactions can be detected with this type of analysis. In our data, the two 3-way interactions were the only triples for which all component pairs were detected as significant correlations. Additionally, the patterns in the contingency tables suggest G=C–C and T=A–A interactions. To a lesser degree, we believe that isolated base-pairs can also be predicted. The data presented in Table III represent significant correlations detected beyond the strong sequence conservation. Although, the base-pairing pattern seen in the helical stems is not as obvious, one or two accepted base-pairings are preferred. As a control, we did not detect significant correlations between pairs of highly conserved positions that did not interact.

The ability to use arbitrary encodings of sequence information to detect sequence covariation is a powerful method for detecting otherwise latent structural or functional relationships. We have also developed a method for automatically translating the sequence variation and covariation information that is reported by MCSEQ into data structures that can be used for sequence classification and database search. Both of these properties are described elsewhere in the context of protein sequence analysis (Klingler and Brutlag 1993).

Finally, we have used MCSEQ to look at correlations between two different properties of tRNAs: their nucleotide sequence and their cognate amino acids. The flexibility in encoding categorical parameter of any type in our system allows one to investigate a wider range of questions. We have discovered a set of correlations in the tRNA molecule that have a striking spatial arrangement—all lie on the face of the molecule that interacts with the tRNA synthetase, suggesting a role in conferring specificity of interaction (Rould, et al. 1989). In fact, most of the positions we detect, including 34-37, 70-73 and positions in the D-stem play a role in recognition the tRNA synthetase. One
also expected interaction with the ribosome to produce correlations, but these cannot be confirmed at the present time.

We are currently analyzing other RNA sequence sets. Compilations of 5S rRNA sequences (Spect, et al. 1991), 23S rRNA sequences (Gutell, et al. 1992), small ribosomal subunit RNA sequences (De Rijk, et al. 1992), and other small RNA sequences (Shumyatsky and Reddy 1992) are easily obtained for analysis. Some of these sequence sets can be used as further validation of our method, while others have less well understood tertiary structures. We believe that comparative methods like ours can yield valuable information for the elucidation of RNA tertiary structures. However, a careful understanding of tertiary interactions in RNA and how those interactions will be reflected in contingency tables is essential. The contingency tables for base-paired helical stems have easily interpreted patterns which are reinforced by the base-stacking correlations. The coordinated patterns for T=A–A and C=G–G base triplets are less obvious, but can be recognized from strong pairwise correlations. The contingency tables for isolated base-pairs are the least interpretable because they can be highly dependent on the level of sequence variation in the contributing positions.

We are also exploring the possibility of translating definable sequence interactions into distance constraints that could be used in a distance geometry program to predict three-dimensional structures (Altman 1993).

Acknowledgements

This work is supported by the CAMIS grant from the National Library of Medicine LM05305 and in part by a seed grant from the Stanford Office of Technology Licensing.

References

R. B. Altman. 1993. A Three-dimensional tRNA structure from Sequence Correlation Data. *Proceedings of the Intelligent Systems for Molecualr Biology* (ISMB-93).

D. M. Crothers, T. Seno and D. G. Söll. 1972. Is There a Discriminator Site in Transfer RNA? *Proc. Natl. Acad. Sci. USA* 69: 3063-3067.

P. De Rijk, J. M. Neefs, Y. V. de Peer and R. De Wachter. 1992. Compilation of small ribosomal subunit RNA sequences. *Nucl. Acids. Res.* 20: 2075-2089.

R. R. Gutell, M. N. Schnare and M. W. Gray. 1992. A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures. *Nucl. Acids Res.* 20: 2095-2109.

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

R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick and A. Zamir. 1965. *Science* 147: 1462-1465.

T. M. Klingler, D. L. Brutlag. 1993. Bayesian Represenation of Protein Structure. Forthcoming.

M. Levitt. 1969. *Nature* 224: 759-763.

W. H. Press, B. P. Flannery, S. A. Teukolsky, W. T. Vetterling. 1988. Are Two Distributions Different? *Numerical Recipes in C.* Cambridge, UK: Cambridge University Press, Cambridge. 487-506.

M. A. Rould, J. J. Perona, D. Söll and T. A. Steitz. 1989. Structure of E. coli Glutaminyl-tRNA Synthetase Complexed with tRNA-Gln and ATP at 2.8 Å Resolution. *Science* 246: 1135-1142.

G. Shumyatsky and R. Reddy. 1992. Compilation of small RNA sequences. *Nucl. Acids Res.* 20: 2159-2165.

T. Spect, J. Wolters and V. A. Erdmann. 1991. Compilation of 5S rRNA and 5S rRNA gene sequences. *Nucl. Acids Res.* 19: 2189- 2191.

M. Sprinzl, N. Dank, S. Nock and A. Schön. 1991. Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.* 19: 2127-2171.

J. L. Sussman, S. R. Holbrook, R. W. Warrant, G. M. Church and S. H. Kim. 1978. Crystal Structure of Yeast Phenylalanine Transfer RNA. *J. Mol. Biol.* 123: 607-630.

C. R. Woese and R. R. Gutell. 1989. Evidence for several higher order structural elements in ribosomal RNA. *Proc. Natl. Acad. Sci. USA* 86: 3119-3122.

C. R. Woese, S. Winker, R. R. Gutell. 1990. Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops". *Proc. Natl. Acad. Sci. USA* 87: 8467-8471