BIOC218 Project
Paul Segars

Quantitative Analysis of tRNA Sequences Using Information Theory

Abstract

I performed a quantitative analysis on a large set of known transfer RNA (tRNA) sequences. Many tRNA have been found to conserve a secondary structure of base-pairing interactions more so than their primary sequence. This makes the sequence analysis of tRNA more complicated in that alignment algorithms focus on the primary similarity without considering the conserved relationships among the nucleotides in the sequence. The interactions between the nucleotides in a RNA sequence are an important factor in determining their structure and function. By failing to consider these relationships, alignment algorithms are unable to detect important structural information conserved in the correlations among the nucleotides. In this study, I examined the primary sequence similarity on an aligned set of tRNA sequences. In addition, I used measures derived from information theory to analyze the correlations or mutual information among distinct sites within the set of sequences. The correlations that I found were similiar to the ones found by previous studies, which correspond to the known secondary and tertiary structures of tRNA. From my analysis, I developed a predictive motif to describe the set of tRNAs. Different groups of tRNA sequences were then scored against this motif to see which were best represented by it. In addition, the sequences were checked to see how well the interrelationships of the nucleotides discovered in this analysis were conserved.

I. INTRODUCTION

Transfer RNA (tRNA), like proteins and other types of RNA, are transcribed from DNA. RNA molecules are composed of four different nucleotides: adenine (A), cytosine (C), guanine (G), and uracil (U). The function of tRNA is to transfer amino acids from the cytoplasm to a ribosome. The ribosome then adds each amino acid brought to it by the tRNA to the protein which it is synthesizing. The structure of tRNA is well suited for its function. A tRNA molecule consists of a single strand of about 80 nucleotides. Within the tRNA sequence, there is an amino acid attachment site and a region (anticodon) where the tRNA binds to a complementary codon on the messenger RNA (mRNA) coding for the protein. Due to base-pairing interactions between the nucleotides, the single strand folds back upon itself forming the secondary and tertiary structures. The Watson-Crick base pairs A-U and G-C form hydrogen bonded base pairs with the A-U pair bound with two hydrogen bonds and the G-C pair bound with three. In addition to the traditional Watson-Crick base pairs, other base pair interactions have been found. The most common of these is the G-U pair.

The base-pairing interactions are represented by the secondary structure of tRNA. The secondary structure of tRNA forms a cloverleaf as shown in Figure 1A. The cloverleaf structure of tRNA was formulated by Holley *et al*. (Holley *et al*. 1965) through a careful analysis of an aligned set of tRNA sequences. In their analysis, they noticed some base-pairing relationships that helped them to arrive at the cloverleaf model. The base-pairing interactions are illustrated in the figure as dotted lines. The amino acid binding site and the anticodon region are also shown. The tertiary structure of tRNA was formulated by Levitt (Levitt, 1969) using 14 tRNA sequences. The tertiary structure, shown in Figure 1B, is roughly L-shaped. Both structure models were verified through examination by crystallography (Sussman *et al*. 1978).

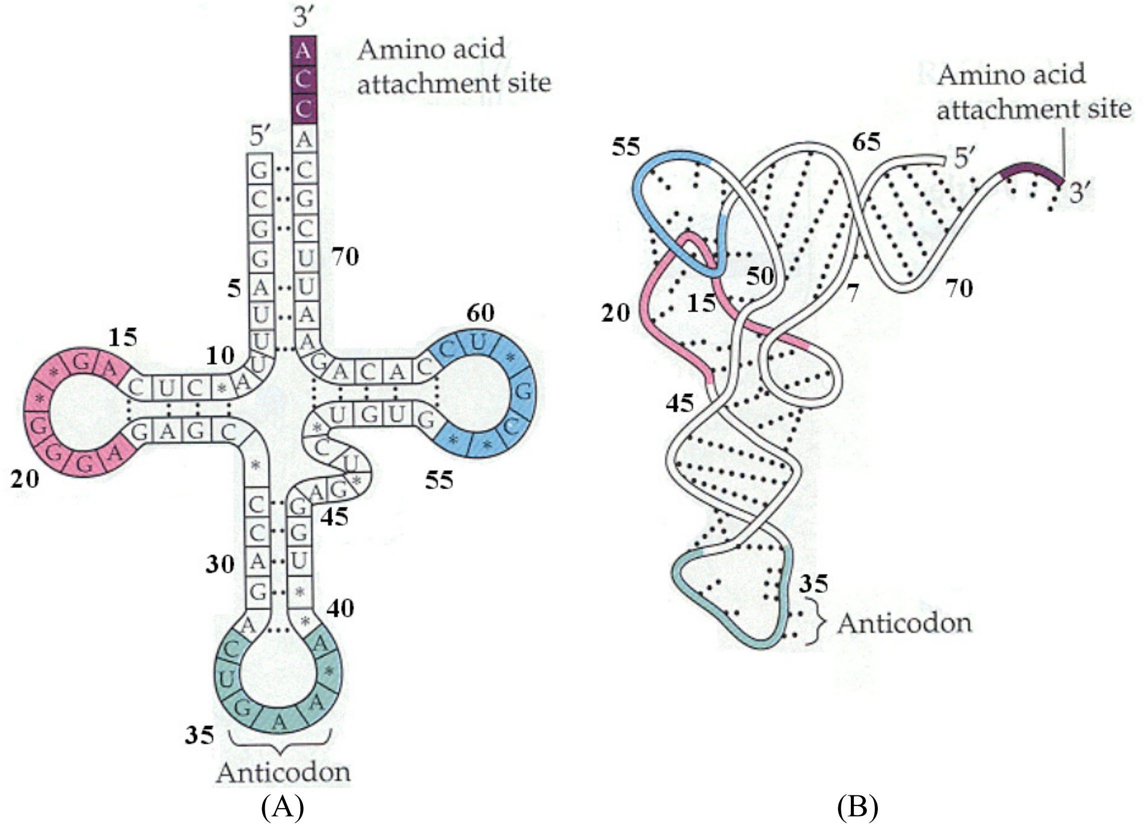


Figure 1: Two-dimensional secondary (A) and three-dimensional tertiary (B) structures of tRNA with the base-pairing interactions indicated by the dotted lines. (Campbell 1993)

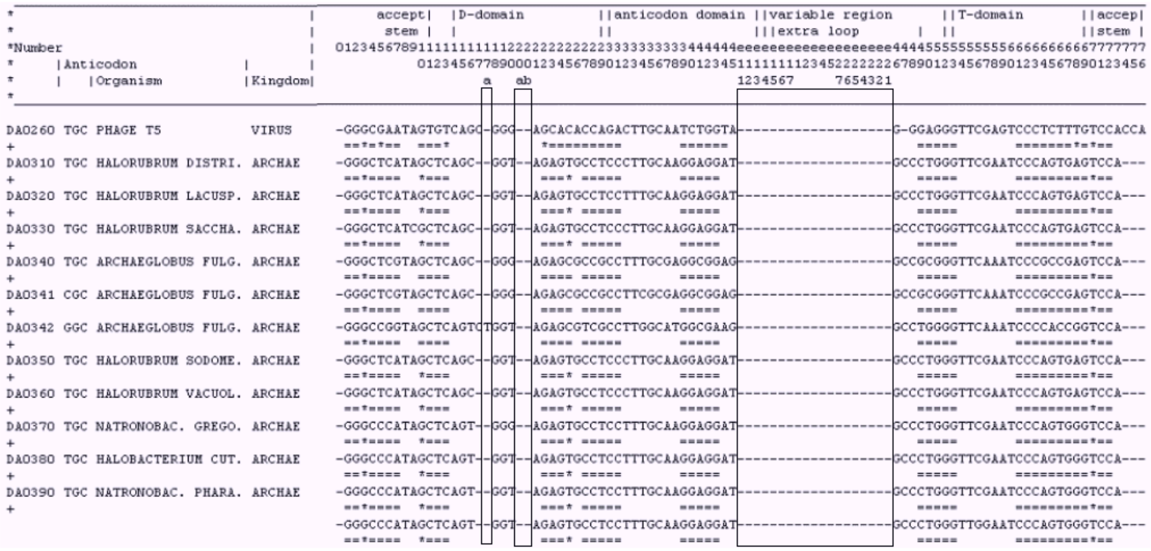
The first step in analyzing a set of sequences, whether they are composed of nucleotides or amino acids, is to perform a multiple sequence alignment. From the alignment, one can see important regions that are conserved between the various sequences. Since they are conserved, these indicate important structural and functional elements of the sequences. This is a much more difficult task in the case of tRNA. Alignment algorithms generally only consider each amino acid or nucleotide site in the sequences as independent of one another. They align the sequences based on primary sequence similarity without considering the relationships between sites within the sequences. Many homologous tRNA have been found to have a common structure without having much similarity in their primary sequences. The primary sequences may vary, but the important base-pairing interactions that determine the structure are conserved. Due to this, it is much more difficult to align a set of tRNA sequences. Different methods must be used that will check for conserved information within the sequences themselves.

Sprinzl *et al*. (Sprinzl *et al*. 1991) compiled a large set of aligned tRNA sequences taking these factors into consideration. The sequences were initially aligned based on their primary structure similarity. Secondary structure interactions were then identified and used to improve the alignment. This was done in an iterative fashion generating the final alignment available in the database.

Many studies have been done examining the mutual information content of tRNA molecules. Gutell *et al*. (Gutell *et al*. 1992) performed an analysis of 896 tRNA sequences from the Sprinzl database while Klingler and Brutlag (Klingler and Brutlag 1993) examined 1208 sequences. For this project, I follow the examples of these previous studies and perform a quantitative analysis on all 3600 of the tRNA sequences from the current Sprinzl database. I examine the sequences for primary structure similarity and for the mutual information content between the nucleotides. I then use this analysis to compose a predictive motif for the tRNA sequences and test the efficacy of this motif in describing the set of sequences.

II. METHODS

In my analysis, I used 3600 aligned tRNA sequences from the publicly available database compiled by Sprinzl *et al*. The sequences in the database are the genes (DNA) that code for the tRNA; therefore, the nucleotide base thymine (T) appears in the sequences instead of uracil (U). Figure 2 shows some representative sequences along with the base-pairing interactions that were used to align them. As can be seen, gaps have been inserted in the sequences in order to align them. These positions are indicated by the boxes in Fig. 2. These positions were not used in the analysis of the tRNA sequences.



```
*                                    |       accept|  |D-domain         ||anticodon domain ||variable region      ||T-domain       ||accep|
*                                    |        stem |  |                 ||                 |||extra loop       |  ||               ||stem |
*Number                              |  012345678911111111111222222222222333333333344444eeeeeeeeeeeeeeeeeee4444555555555566666666667777777
*     |Anticodon             |       |            01234567789000123456789012345678901234511111111234522222226789012345678901234567890123456
*     |   |Organism          |Kingdom|                    a   ab                         1234567     7654321
*

DA0260 TGC PHAGE T5           VIRUS     -GGGCGAATAGTGTCAGC-GGG--AGCACACCAGACTTGCAATCTGGTA-------------------G-GGAGGGTTCGAGTCCCTCTTTGTCCACCA
+                                        ==*=*==  ===*           *=========       ======                       =====       =======*=*==
DA0310 TGC HALORUBRUM DISTRI. ARCHAE    -GGGCTCATAGCTCAGC--GGT--AGAGTGCCTCCCTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGAGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
DA0320 TGC HALORUBRUM LACUSP. ARCHAE    -GGGCTCATAGCTCAGC--GGT--AGAGTGCCTCCTTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGAGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
DA0330 TGC HALORUBRUM SACCHA. ARCHAE    -GGGCTCATCGCTCAGC--GGT--AGAGTGCCTCCCTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGAGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
DA0340 TGC ARCHAEGLOBUS FULG. ARCHAE    -GGGCTCGTAGCTCAGC--GGG--AGAGCGCCGCCTTTGCGAGGCGGAG-------------------GCCGCGGGTTCAAATCCCGCCGAGTCCA---
+                                        ==*====  ====           ==== =====       =====                        =====       =========*==
DA0341 CGC ARCHAEGLOBUS FULG. ARCHAE    -GGGCTCGTAGCTCAGC--GGG--AGAGCGCCGCCTTCGCGAGGCGGAG-------------------GCCGCGGGTTCAAATCCCGCCGAGTCCA---
+                                        ==*====  ====           ==== =====       =====                        =====       =========*==
DA0342 GGC ARCHAEGLOBUS FULG. ARCHAE    -GGGCCGGTAGCTCAGTCTGGT--AGAGCGTCGCCTTGGCATGGCGAAG-------------------GCCTGGGGTTCAAATCCCCACCGGTCCA---
+                                        ==*====  ====           ==== =====       =====                        =====       =========*==
DA0350 TGC HALORUBRUM SODOME. ARCHAE    -GGGCTCATAGCTCAGC--GGT--AGAGTGCCTCCCTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGAGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
DA0360 TGC HALORUBRUM VACUOL. ARCHAE    -GGGCTCATAGCTCAGC--GGT--AGAGTGCCTCCCTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGAGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
DA0370 TGC NATRONOBAC. GREGO. ARCHAE    -GGGCCCATAGCTCAGT--GGG--AGAGTGCCTCCTTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGGGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
DA0380 TGC HALOBACTERIUM CUT. ARCHAE    -GGGCCCATAGCTCAGT--GGT--AGAGTGCCTCCTTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGGGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
DA0390 TGC NATRONOBAC. PHARA. ARCHAE    -GGGCCCATAGCTCAGT--GGT--AGAGTGCCTCCTTTGCAAGGAGGAT-------------------GCCCTGGGTTCGAATCCCAGTGGGTCCA---
+                                        ==*====  *===           ===* =====       =====                        =====       =========*==
                                        -GGGCCCATAGCTCAGT--GGT--AGAGTGCCTCCTTTGCAAGGAGGAT-------------------GCCCTGGGTTGGAATCCCAGTGGGTCCA---
                                         ==*====  *===           ===* =====       =====                        =====       =========*==
```

Figure 2: Aligned tRNA sequences obtained from the Sprinzl database. The different regions along the sequences are labeled at the top. Base-pairing relationships used to align the sequences are noted as follows. Nucleotides involved in Watson-Crick base-pairing interactions (A-T or G-C) are marked with '=', the G-U base pairing interactions are indicated with ''.*

Sequence Variability

The sequence variability at each position in the set of tRNA sequences was calculated in the following manner. For each position *v* defined on the set of aligned sequences, the probability for a nucleotide *A* to occupy the position was calculated by Equation 1

$$p(A/v) = \frac{Number \quad of \quad sequences \quad s \quad with \quad v(s) = A}{Number \quad of \quad sequences} \tag{1}$$

This calculation was performed for each nucleotide at each every position in the aligned sequences. The primary sequence variability for each position was then measured in terms of the Boltzmann entropy *E* as used by Shannon (Shannon and Weaver 1949). It measures the degree of variation among the categories of nucleotides at each position *j* in the domain and is defined by the following equation

$$E_j = -\sum_{N=(A,T,G,C)} p(N \mid j) \log_2 p(N \mid j) \tag{2}$$

where p(*N* | *j*) is the relative frequency of nucleotide *N* at position *j* as calculated by Equation 1. The entropy is computed as the sum of the probability calculations for each nucleotide. If the same nucleotide occupies the particular position for all sequences in the alignment, the entropy is zero. The entropy increases with an increase in the number of different nucleotides that occupy the position and their equal probabilities. For the case of DNA or RNA, it can reach a maximum of $\log_2 4 = 2$ where 4 (the number of nucleotides) is the maximum number of variations you could find at a position. The entropy calculation was performed for each position in the sequence alignment excluding gaps from the calculation.

Mutual Information

In addition to the primary sequence similarity, the mutual information content of the sequences was measured. Mutual information is a measure derived from information theory (Chiu and Kolodziejczak 1991; Gutell *et al*. 1992) and measures the correlation between nucleotide sites within the sequences. The mutual information $M_{v,w}$ between two positions *v* and *w* in an aligned set of sequences is defined as

$$M_{v,w} = \sum_{A,B} p(A,B \mid v,w) \times \log_2 \left(\frac{p(A,B \mid v,w)}{p(A \mid v) \times p(B \mid w)} \right) \tag{3}$$

where the probabilities *p*(*A*|*v*) x *p*(*B*|*w*) are calculated as in Equation 1. The probability *p*(*A*,*B*|*v*,*w*) is computed similarly. For any pair of positions *v* and *w*, the probability for the nucleotides *A* and *B* to occupy those respective positions is defined as *p*(*A,B*|*v,w*) which equals the number of sequences *s* where *v(s)* = *A* and *w(s)* = *B* divided by the total number of sequences. The mutual information goes to zero if the two positions are statistically independent of one another. In this case *p*(*A*|*v*) x *p*(*B*|*w*) equals *p*(*A*,*B*|*v*,*w*) for all *A*,*B*. The higher the mutual information calculation, the more likely it is that the two positions are correlated. The nucleotide at one position can be estimated with a high likelihood due to the presence of another nucleotide at a separate position.

The calculations for the primary sequence similarity and the mutual information content between all pairs of positions in the domain of the sequences were performed using a program written in C. To test the significance of the mutual information content found for the sequences, the nucleotides in each column in the alignment were randomized keeping the nucleotide distribution and, therefore, the entropy calculations the same. The maximum mutual information content was then calculated from the new randomized alignment. This was done several times each with a different set of randomized sequences. The maximum mutual information value in the case of the randomized sequences was found to be 0.0134. This was set as a significance threshold where values above this were determined to be statistically significant meaning they did not occur by chance.

Using the above analysis, a predictive motif was constructed to represent the tRNA sequences. Each tRNA sequence was tested against the motif, and the number of mismatches were computed and used as a measurement to score the ability of the motif to represent the sequences. The tRNA sequences from the different groups obtained in the Sprinzl database were then examined to see how well they were described by the motif and to see how well the secondary interactions were conserved.

III. RESULTS

The results from the primary sequence analysis are shown in Table 1. For each position in the aligned sequences, the frequency of each nucleotide is shown as well as the entropy calculation for the primary sequence variability. Each position was ranked according to its entropy calculation. Positions with a low degree of primary sequence variability were ranked lower while those with a high degree of variability were ranked higher. Positions that contained a large number of gaps were marked with an asterisk (*) in the first column to distinguish them.

As can be seen from Table 1, positions 8, 10, 14, 21, 33, 37, 53-55, and 58 are highly conserved ($E < 0.6$). Positions 2-6, 13, 27-29, 31, 35-36, 39, 41-43, 50-51, 59, 63-64, and 67-71 ($E = 1.27$ or higher) have a high degree of primary sequence variability. Over 50% of the positions in the sequence alignment were found to be quite variable ($E = 1.0$ or higher). This illustrates the ability of the tRNA sequences to maintain their higher-order structure and function despite a great deal of variability in their primary structure. The reason for this is due to the conserved relationships between the pairs of positions in the sequence alignment. This is discussed below in the analysis of the mutual information content of the tRNA sequences.

Table 1:Primary Sequence Analysis

Position	%A	%T	%G	%C	% GAPS	ENTROPY	RANK
*0	0.33	0.17	1.19	0.06	98.25	0.086722	0
1	20.25	9.97	60.36	8.42	1	1.066319	35
2	20.17	12.86	39.81	26.92	0.25	1.306607	59
3	18.56	19.25	38.89	23.19	0.11	1.335943	64
4	24.94	21.42	33.25	20.31	0.08	1.366235	71
5	30.86	23.22	24.17	21.69	0.06	1.376612	75
6	21.72	38.83	19.19	20.19	0.06	1.338858	65
7	34.36	20.89	41.89	2.81	0.06	1.158924	44
8	7.22	88.22	1.67	1.08	1.81	0.417633	6
9	65.42	3.56	25.22	4.64	1.17	0.886131	26
10	8.61	3.47	81.78	4.56	1.58	0.633068	14
11	3.31	32.83	8	54.19	1.67	1.012426	32
12	12.31	52.25	16.06	17.67	1.72	1.196909	48
13	13.86	38.64	14.58	31.17	1.75	1.285448	55

14	88.58	3.03	3.28	1.19	3.92	0.378215	5
15	40.42	7.31	46.56	2.83	2.89	1.014197	33
16	13.81	48.89	4.44	18.61	14.25	1.074528	37
*17	6.17	25.58	0.92	9.08	58.25	0.781459	18
18	7.97	6.47	69.19	3.94	12.42	0.761152	16
19	5.86	9.44	67.17	4.31	13.22	0.791878	19
20	8.56	52.08	3.56	8.72	27.08	0.881495	25
21	83.78	6.06	5.14	2.06	2.97	0.550502	9
22	41.44	10.83	43.06	2.97	1.69	1.073143	36
23	52.47	12.39	18.11	16.03	1	1.200009	49
24	32.03	3.47	55.75	7.92	0.83	1.007862	31
25	3.53	22.69	5.39	67.92	0.47	0.874722	24
26	48.17	7.94	37.33	5.5	1.05	1.080428	38
27	18.39	39.06	12.17	30.31	0.08	1.296683	58
28	18.64	38.31	10.14	32.83	0.08	1.27842	54
29	26.69	25.36	36.69	11.25	0	1.314168	60
30	13.69	3.94	63.97	18.33	0.06	0.996583	30
31	35.17	18.31	20.36	26.17	0	1.353205	66
32	2.11	40.22	1	56.58	0.08	0.816037	20
33	0.31	96.3	0.19	3.14	0.06	0.174759	1
34	3.72	46.03	34.47	15.78	0	1.138115	43
35	29.86	26.44	21.5	22.19	0	1.377227	76
36	22.39	32	21.58	24.03	0	1.373243	74
37	78.28	0.5	20.72	0.44	0.06	0.56844	10
38	67.22	12.69	3.94	15.89	0.25	0.948804	27
39	13.31	35.53	29.92	21.17	0.08	1.325717	63
40	2.61	19.5	19.17	58.72	0	1.043203	34
41	25.08	26.61	11.44	36.83	0.03	1.315139	61
42	34.44	20.53	36.14	8.58	0.31	1.270724	52
43	35.61	24.14	32.42	7.72	0.11	1.27373	53
44	48.81	25.31	12.81	12.39	0.69	1.219746	50
45	25.78	15.67	46.08	3.58	8.89	1.11616	39
46	33.83	13.58	44.64	3.58	4.36	1.117153	40
*47	4.25	35.86	2.69	4.11	53.08	0.730569	15
48	7.25	41.39	1.94	48.14	1.28	0.983914	29
49	28.31	7.28	48.2	14.89	1.33	1.1833	46
50	17.94	31.06	23.36	26.39	1.25	1.362679	69
51	29.39	16.31	39.44	13.56	1.31	1.293444	57
52	18.81	6.17	70.47	2.92	1.64	0.835768	22
53	7.67	3.42	81.39	2.61	4.92	0.575069	11
54	8.06	82.94	2.06	4.58	2.36	0.579166	12
55	5.94	84.08	5.25	3.94	0.78	0.59581	13
56	14.5	9.86	2.22	71.89	1.53	0.830305	21
57	44.72	4.83	43.94	3.36	3.14	0.98168	28
58	84.69	4.56	1.94	3.53	5.28	0.476024	7
59	39.72	21.53	19.86	10.42	8.47	1.254004	51

60	4.53	63.89	2.03	14	15.56	0.780675	17
61	2.89	7.44	1.64	83.08	4.95	0.517142	8
62	5.86	19.11	3.19	70.11	1.72	0.84151	23
63	13.36	33.67	15.03	36.83	1.11	1.288148	56
64	24.11	26.92	30.08	17.81	1.08	1.364863	70
65	6.25	38.5	15.44	38.44	1.36	1.196775	47
66	20	37.89	3.53	38.44	0.14	1.175102	45
67	30.36	26.11	25.86	17.53	0.14	1.367507	72
68	18.36	33.94	26.58	21.03	0.08	1.358049	67
69	18.14	28.28	24.25	29.25	0.08	1.369955	73
70	17.78	24.06	24.47	33.64	0.06	1.360774	68
71	12.5	23.75	27.03	36.61	0.11	1.322837	62
72	10.33	28.25	7.64	53.39	0.39	1.123163	41
73	54.67	16.33	20.83	6.08	2.08	1.123197	42
*74	0.19	0.19	0.06	13	86.56	0.293671	4
*75	0.11	0	0	12.53	87.36	0.267788	3
*76	12.39	0	0	0	87.61	0.258726	2

The correlations or mutual information between pairs of positions in the sequence alignment are illustrated by the image in Figure 3A. Bright intensities in the image indicate position pairs that have a high degree of correlation with one another. The image in Fig. 3A was regenerated using a threshold of 0.175, a value much higher than the maximum correlation (0.0134) found by randomizing the columns in the sequence alignment. This theshold was chosen since it revealed the relationships discussed by Klingler (Klingler and Brutlag 1993) and Gutell (Gutell *et al*. 1992). Figure 3B shows the thresholded image. The correlated pairs of positions and the mutual information calculation for them are listed in Table 2 along with the postulated relationships between the position pairs. The relative frequencies of the nucleotides for each correlated position (Table 1) were examined and used to determine the relationship for each pair. The most frequent nucleotides are listed for each position in Table 2.

Four distinct areas can be seen to have a high degree of mutual information. Colored arrows as well as numbers (1-4) distinguish these four areas in the image in Fig. 3B. The areas are also highlighted with the same colors in Tables 1 and 2. These regions of high mutual information appear to be the base-pairing interactions that form the secondary structure (cloverleaf) of the tRNA, Fig. 1A. In addition to these four areas, the two three-way interactions discovered previously (Klingler and Brutlag 1993 and Gutell *et al*. 1992) were found in this analysis. Position pairs (13, 22), (13, 46), and (22, 46) are involved in one three-way interaction while (9, 12), (9, 23), and (12, 23) are involved in the second. These position pairs are also distinguished in Fig. 3B by color-coded arrows and in Tables 1 and 2 with the same colors. For each three-way relationship, only two of the position pairs are shown with arrows. The third point, in each case, is involved in the base-pairing interactions of the cloverleaf.

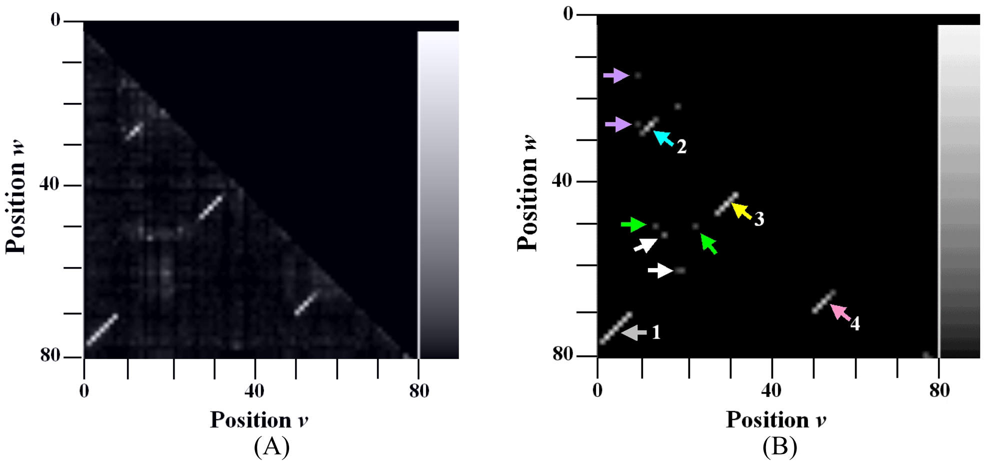


Figure 3: (A) Image of the mutual information calculation for each position pair in the tRNA alignment. Bright intensities indicate positions that are highly correlated. (B) Image of the mutual information calculations above a threshold of 0.175. Four areas of high correlation are numbered. These areas correspond to the four leafs in the cloverleaf model for the secondary structure of tRNA. The green and purple arrows point to pairs of positions that are involved in three-way interactions. The white arrows point to pairs of positions that may or may not be significant base-pairing interactions in the tertiary structure of the tRNA molecule.

Table 2: High Correlations Among Pairs of Positions in the tRNA Sequences

Position 1	Position 2	Type of Interaction	Mutual Information
1 (G)	72 (C)	Watson-Crick	0.708949
2 (ATGC)	71 (ATGC)	Any base pairing	1.091603
3 (ATGC)	70 (ATGC)	Any base pairing	1.026088
4 (ATGC)	69 (ATGC)	Any base pairing	0.967609
5 (ATGC)	68 (ATGC)	Any base pairing	0.933195
6 (ATGC)	67 (ATGC)	Any base pairing	0.931464
7 (AG)	66 (TC)	Any base pairing	0.906391
9 (A)	12 (T)	3-Way Interaction	0.175275
9 (A)	23 (A)	3-Way Interaction	0.186754
10 (G)	25 (C)	Watson-Crick	0.349585
11 (TC)	24 (AG)	Any base pairing	0.841822
12 (T)	23 (A)	3-Way Interaction	0.983671
13 (TC)	22 (AG)	3-Way Interaction	0.367795
13 (TC)	46 (AG)	3-Way Interaction	0.252385
15 (AG)	48 (TC)	Any base pairing	0.430615
18 (G)	19 (G)	None?	0.333390
18 (G)	56 (C)	Watson-Crick	0.190022
19 (G)	56 (C)	Watson-Crick	0.287858
22 (AG)	46 (AG)	3-Way Interaction	0.243109
27 (TC)	43 (AG)	Any base pairing	0.748799
28 (TC)	42 (AG)	Any base pairing	0.947045
29 (ATGC)	41 (ATGC)	Any base pairing	1.147633
30 (G)	40 (C)	Watson-Crick	0.770178
31 (ATGC)	39 (ATGC)	Any base pairing	1.000114
49 (AG)	65 (TC)	Watson-Crick	0.782718
50 (ATGC)	64 (ATGC)	Any base pairing	0.932983
51 (AG)	63 (TC)	Watson-Crick	0.956814
52 (G)	62 (C)	Watson-Crick	0.658236
53 (G)	61 (C)	Watson-Crick	0.335416

Other relationships that may or may not be significant were also discovered. The position pairs (15, 48), (18, 56), and (19, 56) appear to interact through Watson-Crick base-pairing as determined by the relative frequencies of the nucleotides at these positions. Judging by their relative positions in the tertiary structure of tRNA (Fig. 1), these interactions, as indicated by the white arrows in Fig. 3B, could be important determinants in the tertiary structure of tRNA. The position pair (18, 19) may not be an important interaction. Position 18 may be highly correlated with position 19 due to the fact that they both have a high frequency for the guanine (G) nucleotide. The mutual information calculation does not discriminate against correlations that are not base-pairing interactions. It finds any relationship that ties two positions together. These interactions may or may not be valid. Of course, the same could be said for the interaction between position pairs (22, 46) and (9, 23) which are both involved in a three-way relationship. They may be correlated only due to the fact that they generally contain the same nucleotides (Table 2). Position pairs (18, 19), (18, 56), and (19, 56) could be a three-way relationship. However, since positions 18 and 19 are next to each other in the sequence, this seems unlikely. Further study would need to be performed to check these positions out as well as position pairs (15,48), (18, 56), and (19, 56).

Some of the base-pairing interactions listed in Table 2 have a high degree of variability. One example of this is the base-pairing interactions that occur between positions 2-6 and positions 72-66 respectively. Any of the four nucleotides is as likely to be at any of these positions as can be seen in Table 1; therefore, there is a high degree of primary sequence variability. But, as can be seen in the mutual information calculations, these positions are highly correlated with one another through base-pairing interactions. If position 2 has an adenine (A), it is highly likely that position 72 will have a thymine (T) and so on. The tRNA sequences can vary greatly in their primary structure, but as long as the base-pairing interactions are maintained, their three-dimensional structure and function will be preserved.

Using the analysis presented above, a predictive motif was developed to best describe the set of tRNA sequences. The motif is listed position by position in Table 3 where R = [A,G], Y = [T,C], and K = [T,G]. All other symbols stand for the nucleotides themselves.

Table 3: Motif for the tRNA Sequences in the Study

Pos.	0	1	2-6	7	8	9	10	11	12	13	14
	-	G	-	R	T	A	G	Y	T	Y	A
Pos.	15	16	17	18	19	20	21	22	23	24	25
	R	Y	-	G	G	Y	A	R	A	R	C
Pos.	26	27-28	29	30	31	32	33	34	35-36	37-38	39
	R	Y	-	G	-	Y	T	K	-	A	-
Pos.	40	41	42-43	44	45-46	47	48	49	50	51	52-53
	C	-	R	-	R	-	Y	R	-	R	G
Pos.	54-55	56	57	58	59	60	61-62	63	64	65-66	67-71
	T	C	R	A	-	T	C	Y	-	Y	-
Pos.	72	73	74-76								
	C	R	-								

The motif was scored against the different groups in the Sprinzl database to see how well it described each group. The percentage of mismatches for each group was calculated by tabulating the number of mismatches and dividing it by the total number of sequences in the group times the number of nucleotides in each sequence (77). Table 4 shows the results obtained for each group in the Sprinzl database. The motif was the least accurate with the tRNA sequences from the animal mitochondria by only matching approximately 78% of each sequence. According to Gutell *et al*. (Gutell *et al*. 1992), the mitochondrial sequences exhibit a great amount of structural variation so this can be expected. However, 78% is not too bad. The motif was the most accurate in matching the tRNA from the Eubacteria. In this case, it matched approximately 90% of each sequence.

Table 4: Efficacy of the Motif in Representing the Different Groups of tRNA in the Sprinzl Database

Group	Number of sequences	% Mismatches
Virus or bacteriophage	53	13.08
Archaebacteria	160	10.55
Eubacteria	682	9.32
Cyanelle (Photosynthetic organella)	9	10.82
Chloroplast	383	10.9
Mitochondria of single cell or fungi	338	14.22
Mitochondria of plant	125	10.0
Mitochondria of animal	1440	22.1
Cytoplasm of single cell or fungi	174	12.57
Cytoplasm of plant	53	11.79
Cytoplasm of animal	186	11.79

The different groups were then analyzed to see if they abided by the structural relationships listed in Table 2. Each sequence in a group was checked for each relationship in the table. Each group is listed in Table 5 along with the different base-pairing and three-way interactions. For each group, the percentages of the sequences analyzed that had each relationship are shown. An average percentage is calculated for each relationship. Each base-pairing relationship was found to be in the majority of all the sequences (Average > 60%). The position pairs (1,72), (10, 25), (30,40), and (52,62) had the lowest likelihood. This was thought to be due to the fact that they were checked for specific base-pairing interactions (Table 2). For example, the position pair (1, 72) was only checked for a G-C relationship. These position pairs were checked to allow for any type of base-pairing relationship. The new average percentages for them are shown in parentheses. The average percentages for these positions increased to over 97% when all base-pairs were allowed.

The three-way relationships were found to be in only 40% of all sequences with the exception of the proposed relationship between positions 18, 19, and 56. It was found to be preserved in 90% of the sequences. It was not highly conserved in the animal mitochondria tRNA, but was extremely well conserved in the other groups. This relationship could be important, but it does not seem likely. The other positions were checked again allowing for any type of three-way interaction rather than the specific ones listed in Table 2. The new average percentages are shown in parentheses. They were found to only increase to about 55% which is not very significant. The three-way interactions do not seem to be highly conserved and may not be that significant in determining the structure of tRNA.

Table 5: Conservation of the Structural Interactions Among the Different Groups

Correlated Positions													
Group	(1,72)	(2,71)	(3,70)	(4,69)	(5,68)	(6,67)	(7,66)	(9,12,23)	(10,25)	(11,24)	(13,22,46)	(15,48)	(18,56)
Virus	64	100	100	98	98	98	100	51	75	98	47	81	96
Archae.	91	100	100	100	100	100	100	32	66	99	34	99	99
Eubact.	77	100	100	99	99	100	100	61	86	100	63	97	98
Cyanel.	89	89	100	100	89	100	100	56	100	100	56	100	100
Chroro.	69	99	100	98	97	96	99	51	78	99	54	97	100
Mit. Single cell	45	98	99	98	99	96	99	51	63	98	43	85	94
Mit. Plant	49	98	98	97	98	91	96	42	78	100	49	93	99
Mit. Animal	25	97	95	94	92	93	95	41	56	93	38	65	16
Cyto. Single cell	64	99	99	97	97	94	98	29	72	99	39	94	98
Cyto. Plant	91	100	96	98	96	92	98	25	47	96	34	92	96
Cyto. Animal	81	98	99	99	98	98	99	24	68	98	42	94	97
Average	68 (95)	98	99	98	97	96	99	42 (55)	72 (97)	98	45 (52)	91	90

Group	(18,56)	(19,56)	(18,19,56)	(27,43)	(28,42)	(29,41)	(30,40)	(31,39)	(49,65)	(50,64)	(51,63)	(52,62)	(53,61)
Virus	96	96	96	79	92	92	68	85	98	91	100	85	96
Archae.	99	99	99	100	100	100	79	100	99	99	100	93	100
Eubact.	98	97	97	97	99	99	74	100	100	95	98	87	99
Cyanel.	100	100	100	100	100	100	22	100	100	100	100	78	89
Chroro.	100	99	99	90	97	99	52	99	98	99	99	87	96
Mit. Single cell	94	96	93	91	97	99	48	94	96	93	97	52	94
Mit. Plant	99	99	99	94	97	96	58	98	99	100	97	90	100
Mit. Animal	16	18	13	91	94	96	47	93	91	90	90	52	55
Cyto. Single cell	98	98	97	94	99	96	68	97	98	97	98	72	98
Cyto. Plant	96	96	96	92	100	100	83	92	98	98	100	72	100
Cyto. Animal	97	96	96	98	99	100	82	99	99	99	99	84	97
Average	90	91	90	93	98	98	62 (98)	96	98	96	98	77 (98)	93

VI. CONCLUSIONS

I performed a quantitative analysis on a large set of aligned transfer RNA (tRNA) sequences from the Sprinzl database. The tRNA sequences were found to have a great deal of variability in their primary sequence structure, but were found to conserve base-pairing interactions among distinct sites within the alignment. These base-pairing relationships give rise to the secondary and tertiary structures of the tRNA molecules and thus their function. The base-pairing interactions found in this study agree with those found previously with the exception of the three-way interactions. These were found to only be conserved in about 50% of the tRNA sequences. This may mean that they are not an important factor in determining the structure of tRNA. In addition, some other relationships were found that may or may not be important determinants in the three-dimensional structure of tRNA. Further study must be done to examine these positions in more detail. From my analysis, I developed a predictive motif to describe the set of tRNAs. With a few minor changes to allow for more variety of base-paring interactions, the motif was found to reasonably represent the tRNA sequences from each group in the database. Future work would need to be done to test the motif on tRNA sequences that are not in the Sprinzl database. This would be a nice check against over-fitting. The motif may be a reasonable representation of the tRNA from which it was based, but it may not be adequate in describing other tRNA sequences.

VI. ACKNOWLEDGEMENTS

I would like to thank Dr. Brutlag for all of his help on this project and for guiding me through it over email. THANKS.

V. REFERENCES

T. M. Klingler and D. L. Brutlag. Detection of correlations in tRNA sequences with structural implications. *Intelligent Systems for Molecular Biology*. 1993. 1: p. 225 –333.

N. A. Campbell. Biology 3rd Edition. The Benjamin/Cummings Publishing Company Inc., New York, NY.

D. K. Y. Chui and T. Kolodziejczak. 1991. Inferring consensus structure from nucleic acid sequences. *Computer Applications in the Biosciences*. 7: 347-352.

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.

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

C. Shannon and W. Weaver. 1949. The mathematical theory of communication. University of Illinois Press, Urbana.

M. Sprinzl, N. Dank, S. Nock and A. Schon. 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. Kin. 1978. Crystal Structure of Yeast Phenylalanine Transfer RNA. *J. Mol. Biol.* 123: 607-630.