Submitted to Molecular Biology and Evolution on April 17, 1987, but not accepted for publication.

Transfer RNAs for chemically related amino acids show sequence similarities[1]

Hugh B. Nicholas, Jr. and William H. McClain
Department of Bacteriology,
University of Wisconsin, Madison

1.Key words: transfer RNA, tRNA, sequence comparison, multidimensional scaling, sequence similarity, two-way contingency tables.

Address for correspondence and reprints: William H. McClain, Department of Bacteriology, 1550 Linden Drive, University of Wisconsin, Madison, WI 53706.

Running head: Primitive tRNA Structure?

Abstract

Multidimensional scaling analysis of transfer RNA sequences from procaryotic and eucaryotic cell types reveals a clustering of transfer RNAs for chemically similar amino acids. Two-way contingency table analysis of these sequences indicates that the bases at nine positions are correlated with transfer RNA membership in one of six groups. The nine positions and bases associated with group membership may be remnants of primitive transfer RNA-like molecules.

Introduction

In an earlier paper we used multidimensional scaling to analyze 41 tRNA sequences representing 16 amino acid types from diverse organisms (Nicholas and Graves 1983). That analysis indicated that each amino acid acceptor type was correlated with its respective base sequence and that this correlation could be examined by a mathematical approach. The multidimensional scaling analysis reported here reveals an additional and unexpected feature in the data: tRNA sequences group according to the general chemical nature of the cognate amino acid side chains.

Methods

The set of 41 tRNA sequences representing 16 amino acid types from diverse cell types (bacteria, yeast, wheat, and rat) has been described (Nicholas and Graves 1983). For further analysis here, these 41 sequences were aligned so identical positions would be compared. For the multidimensional scaling analysis, we used the alignment and derived percent difference matrix given in Holmquist et al. (1973).

The set of 216 tRNA sequences examined here were taken from a compilation of tRNA and gene sequences (Sprinzl et al. 1985a,b). These sequences are encoded by nuclear genomes; no organelle sequences were used. Table 1 gives the number of tRNAs in each cell type and amino acid type category.

66 variable positions were analyzed; these included the 76 standard positions, minus 13 constant base positions, plus 3 positions to allow for tRNAs with longer D loops, as described below. Figure 1 shows the numbering system used (Schimmel et al. 1979) and the constant bases.

Alignment of sequences was achieved using stem and loop regions of the cloverleaf model, with further refinements in loop regions provided by the constant bases. Alignment in the D loop (positions 14 through 21) was made to increase agreement, first within isoacceptor sequences, then between acceptor groups. When present, added residues were at positions 17a, 20a, and 20b. Gaps were inserted in tRNA sequences not having the latter three residues; a gap was treated as a "fifth nucleotide" for comparative purposes. For tRNAs with more than five residues in the variable loop (positions 44 through 48), we used the three residues

adjacent to position 43 of the anticodon stem, and the two residues next to position 49 of the T stem; for tRNAs with four residues in the variable loop, alignment increased agreement as for the D loop.

Several unusual features of the tRNA sequences deserve further comments. The 5' ends of eight of the nine $tRNA^{His}$ (a Schizosaccharomyces pombe sequence was determined on DNA and the 5' base of the tRNA was not determined) sequences contain an additional base not present in other sequences; this additional base was not considered. Of the 216 sequences analyzed, that of phage T4 $tRNA^{Ile}$ is unusual in containing an added residue between positions 22 and 23. This residue (C) was deleted prior to the analysis, rather than creating position 22a and an alignment gap in the other 215 tRNA sequences. Some of the phage T5 tRNAs are unusual in showing variation at positions of the constant bases. This variation, involving a total of five bases at four positions, was not considered.

Two-way contingency tables of six rows and four columns (6 times 4 equals 24 cells) were formed for a position by labeling each row of the table with one of the six chemical groups and each column with one of the four nucleotides. Kullback's I statistic (Kullback

1959) was used to measures the discrepancy between the number of sequences expected in a cell of the table and the number actually observed, that is the degree of association between the bases of the sequences and the chemical groups. The six chemical groups of tRNAs are those with cognate amino acids having aliphatic, aromatic, carboxylic and amide, or amphoteric side chains in addition to the initiator and elongator methionine tRNAs.

Results

The percent difference between base sequences of each pair of tRNAs in the 41 tRNA data set (Holmquist et al. 1973) was arranged in a difference matrix and analyzed using nonmetric multidimensional scaling (Guttman 1968), as described in Nicholas and Graves (1983). A plot of the number of dimensions versus a residual goodness of fit measure (Kruskal 1964) showed an inflection point at five dimensions. Dimensions one through three were reported previously; figure 2 shows dimension four and five.

Figure 2 (left) labels the coordinate of each tRNA sequence with its cognate amino acid. An amino acid effect is apparent; for example, the four leucine tRNAs

(L) are grouped together, as are the respective isoacceptors for tRNAs for alanine (A), isoleucine (I), and valine (V).

An impression from fig. 2 is that amino acid types L, A, I, and V cluster in a region somewhat removed from other sequences ("A" in fig. 2, right). The side chains of these amino acid are aliphatic in nature. Analogous though less well defined clusters are seen for amino acid types with carboxylic and amide side chains, for amino acid types with aromatic side chains, and for amino acid types with amphoteric side chains (respectively "C", "O", and "H" in fig. 2, right). Methionine (M) and initiator methionine (iM) tRNA sequences segregrated from other sequences. No assignment could be made for the two glycine (G) tRNA sequences.

The 41 tRNA data set was then examined by two-way contingency table analysis in an attempt to identify the tRNA bases that correlate with the clusterings seen in fig. 2. However, the data set was too small for such an analysis to be meaningful. We therefore

examined a subset of the 216 tRNA sequence data set, which provided 134 new tRNA sequences corresponding to the same amino acid types as in fig. 2. (The 134 sequence subset = 216 sequences - (41 sequences in the original data + 15 unclassified glycine sequences + 26 asparagine, cysteine, proline and threonine sequences not yet examined)). The new sequences are encoded by procaryote and eucaryote nuclear genomes. The two-way contingency table results used Kullback's I statistic (Kullback 1959) to measure the extent of correlation.

A plot of the I statistics obtained from the two-way contingency table analyses against their expected chi-squared distribution indicated that bases in nine positions showed stronger than random association with the chemical groups. The nine positions are highlighted in fig. 1.

Two-way contingency tables for these nine positions, computed on the combined data set of 173 (134 + 41 - 2 Glycine tRNAs) tRNA sequences were used with Bayes theorem (Bayes 1958) to assign sequences accepting asparagine (6 tRNAs), proline (9 tRNAs), and

threonine (7 tRNAs) to one of the six chemical groups. In applying Bayes theorem it was assumed that the unassigned tRNAs had an initial probability of Ø.167 (one sixth) of belonging to any of the six chemical groups. The unassigned tRNAs were assigned to the chemical group with the highest probability of membership after the application of Bayes theorem, the highest posterior probability.

Asparagine tRNAs were assigned to the carboxylic and amide group with a posterior probability of Ø.272, and threonine tRNAs were assigned to the amphoteric group with a posterior probability of Ø.26Ø. These were, respectively, Ø.1Ø and Ø.Ø4 greater than the next largest posterior probability. Proline tRNAs were assigned to the aliphatic group with a posterior probability of Ø.218. Proline tRNA had a posterior probability of Ø.212 for assignment as aromatic; however this was caused in part by the coincidence of the same base, not present in other sequences, in the seven procaryotic proline tRNAs and the four eucaryotic tyrosine tRNAs. These eleven tRNAs are the only ones in this data set with G at position 72. After removing the four eucaryotic tyrosine tRNA sequences from the data

set, the posterior aromatic probability dropped to Ø.119, while the posterior aliphatic probability became Ø.219, clearly assigning the proline tRNAs to the aliphatic group.

Assignment was also attempted for the 17 glycine tRNA sequences and the 4 cysteine tRNA sequences, but the results were inconclusive. Glycine tRNAs had similar posterior probabilities for three groups: aliphatic (Ø.246), amphoteric (Ø.237), and carboxylic and amide (Ø.241). Cysteine tRNAs had similar posterior probabilities for amphoteric (Ø.291) and aromatic (Ø.3Ø7) groups.

The 195 tRNAs (216 - 17 glycine tRNAs - 4 cysteine tRNAs) for the 18 amino acid acceptors assigned to a chemical group were combined and two-way contingency tables computed on 66 variable positions. Figure 3 shows the plot of the observed I statistics against the expected chi-squared values (Kullback 1959). The coordinants should lie on a straight line for a random association of base positions and chemical groups. However, nine base positions deviate (fig. 3). These are the same nine positions identified in the earlier

analysis of 134 tRNAs. These positions are shaded in fig. 1.

Figure 4 summarizes our results, showing both the members of the chemical groups and the bases at the nine positions.

Cross-validation was performed on the 195 tRNA sequences by excluding one sequence from the data and recomputing two-way contingency tables on the remaining 194 sequences. This operation was performed 195 times. The two-way contingency tables were used with Bayes theorem (Bayes 1958) as described above to assign the excluded tRNA to one of the six chemical groups. 164 (84%) of the 195 tRNAs were assigned to the correct group by this cross-validation procedure.

Four instances of systematic incorrect assignment were observed in the cross-validation: all nine leucine tRNAs with YAA anticodons (Y = C or U) were assigned to the aromatic group; all three eucaryotic glutamine tRNAs were assigned to the amphoteric group; four of five eucaryotic histidine tRNAs were assigned to the carboxylic and amide group; and, four of seven

isoleucine tRNAs were assigned to the initiator methionine group. No pattern was discerned among the other eleven incorrectly assigned sequences.

Other cross-validations were performed. Using only the three anticodon bases (positions 34, 35, and 36), 131 (67%) of the 195 tRNA sequences were assigned correctly. Using only the six non-anticodon bases (positions 1, 31, 38, 39, 72, and 73), 135 (69%) of 195 tRNAs were assigned correctly. These results demonstrate that the six non-anticodon positions carry as much information about the six chemical groups as the anticodon positions. Also, neither subset of base positions is as effective as the complete set.

To provide a referent for the above cross-validation results, five sets of nine positions were randomly selected. Each random set contained four positions on two base pairs of stem regions and five positions on single-stranded or loop regions of the cloverleaf model; this provided the same structure as for the non-random set. The nine positions identified in the non-random set were excluded from the random sets. Each random set was then used for

cross-validation on the 195 tRNA sequences. We found correct assignments for 34%, 37%, 37%, 45% and 56% of the sequences, for an average of 41%. This compares with 84% correct assignment obtained for the non-random set, as reported above.

Another referent was provided by randomly assigning the amino acid acceptors to six artificial groups and analyzing these by two-way contingency table and cross-validation. In each random group, two amino acid acceptor types were excluded to mimic the unassigned glycine tRNA and cysteine tRNA acceptor groups in the non-random set. Also, the random groups were constrained to have the same number of acceptor classes as the non-random groups. In each of 500 random groups, the nine positions with the largest I statistics were identified and the sum of their I values were recorded. The average of this I value sum was 866. The five largest I value sums were 1390, 1198, 1190, 1182, and 1172. The sum for the non-random group was 1368. Thus, only one of 500 random groups had a higher I value sum than the chemical groups.

The five random groups with the largest summed I values were investigated further. For each group, the I statistics was plotted against the expected Chi-squared distribution to determine the number of positions which showed stronger than random association. The five groups had, respectively, 2, 2, 2, 2, and 3 positions with stronger than random association. These positions were then used for cross-validation of the sequences. These operations gave, respectively, 29%, 46%, 59%, 65%, and 73% correct reassignment of the sequences to their respective artificial group. These values compare with 84% correct reassignment using the nine positions of the natural groups shown in fig. 4, or 67% using only the three anticodon positions of these groups.

Discussion

Multidimensional scaling of tRNA sequences from numerous cell types has revealed an apparent clustering of these molecules according to the general chemical nature of the amino acid side chain (fig. 2). Two-way contingency table analysis revealed that bases at nine positions were sufficient to statistically identify a tRNA sequence as belonging to one of six chemical groups (fig. 4). These observations are novel, and their significance is unknown. No known biological process divides tRNAs or amino acids in this manner.

We suspect that the nine bases identified in fig. 4 were fixed in evolution prior to the divergence of the numerous cell types analyzed. These nine bases occupy two base pairs and single-stranded regions (fig. 1), and thus may have been part of a stem and loop secondary structure. Our analysis did not revealed other bases that may have been present in that structure.

The proteins produced by simple tRNA-like molecules as represented in figure 4 may have contained fewer than twenty amino acids or the proteins may have been heterogeneous in sequence, with only the chemical group of the amino acids specified. Clearly, this interpretation remains speculative without some type of laboratory experiment.

Acknowledgement

This work was supported by grant AI1Ø257 from the National Institutes of Health.

LITERATURE CITED

BAYES, T. 1958. Essay towards solving a problem in the doctrine of chance. Biometetrika, 45:293-315. Reprinted from Philosophical Transactions of the Royal Society. 53:370-418. (1763).

GUTTMAN, L. 1968. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika. 33:469-506.

HOLMQUIST, R., T.H. JUKES, and S. PANGBURN. 1973. Evolution of transfer RNA. J. Mol. Biol. 78:91:116.

KRUSKAL, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika. 29:1-27.

KULLBACK, S. 1959. Information Theory and Statistics. John Wiley & Sons. New York.

NICHOLAS, JR. H. B., and S. B. GRAVES. 1983. Clustering of transfer RNAs by cell type and amino acid specificity. J. Mol. Biol. 171:111-118.

SCHIMMEL, P. R., D SOLL, and J. N. ABELSON. eds. 1979. Transfer-RNA: Structure, properties and recognition. Cold Spring Harbor Laboratory. New York. pp. 518-519.

SPRINZL, M., J. MOLL, F. MEISSNER, and T. HARTMAN. 1985a. Compilation of tRNA sequences. Nucleic Acids Research. 13:r1-r49.

SPRINZL, M., T. VORDERWULBECKE, and T. HARTMAN. 1985b. Compilation of sequences of tRNA genes. Nucleic Acids Research. 13:r51-r1Ø4.

Table 1. Number of tRNAs from each amino acid and cell type category.

Chemical Group / Amino Acid	Procaryotes	Eucaryotes
Aliphatic	31	26
alanine	4	4
isoleucine	5	2
leucine	10	9
valine	5	9
proline	7	2
Amphoteric	17	24
arginine	5	8
lysine	2	7
serine	5	7
threonine	5	2
Aromatic	15	26
histidine	4	5
phenylalanine	6	13
tryptophan	1	4
tyrosine	4	4
Carboxylic and Amide	13	20
asparagine	3	7
aspartic acid	3	3
glutamine	4	3
glutamic acid	3	7
Initiator methionine	8	11
methionine	8	11
Elongator methionine	2	2
methionine	2	2
Unclassified	11	10
glycine	8	9
cysteine	3	1

Note: eucaryotic tRNAs are encoded by nuclear genomes.

Figure legends

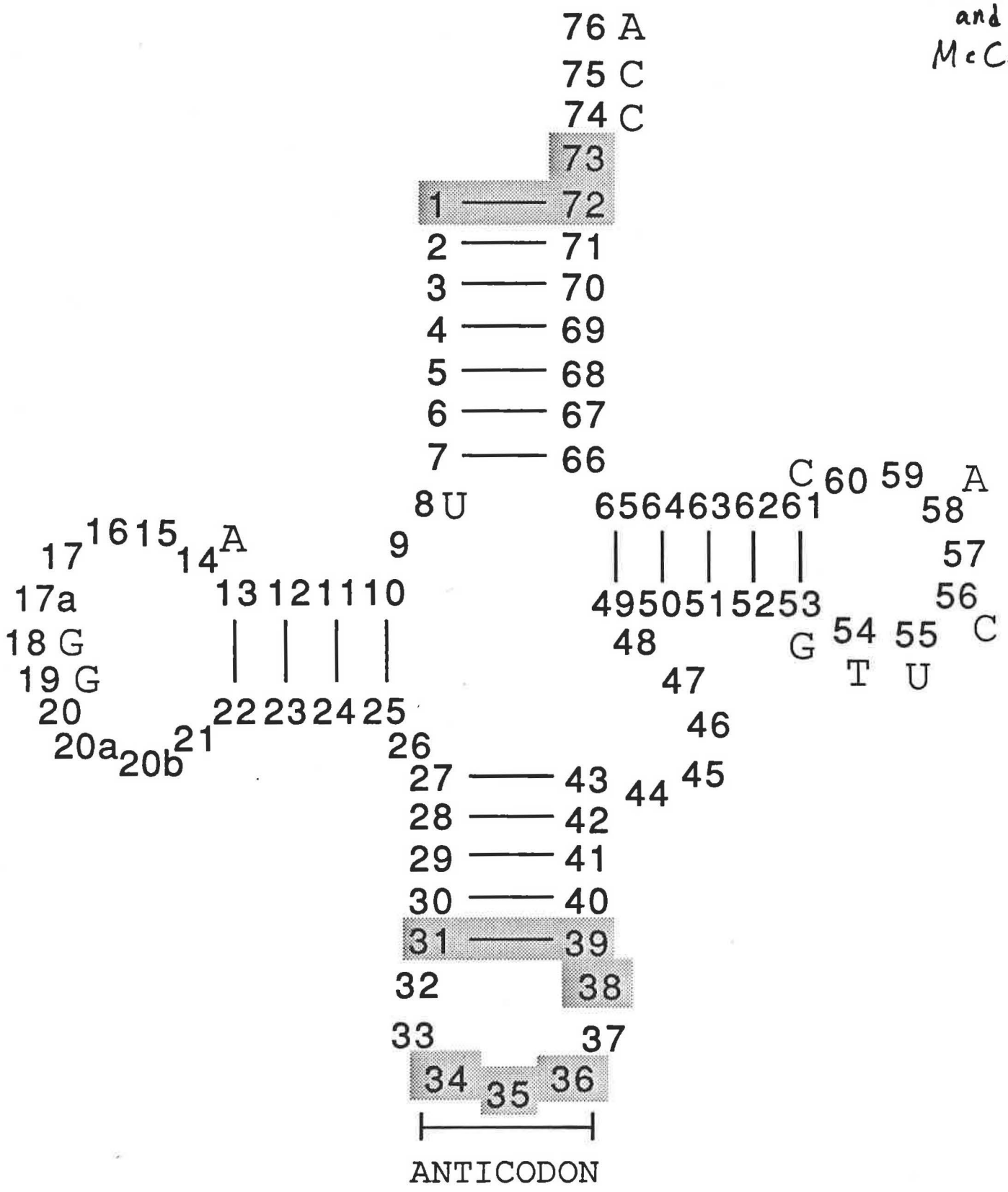
Fig. 1.--Numbering of base positions in tRNA. The 13 constant bases are noted. Shaded positions map tRNA sequences to the chemical groups shown in fig. 4.

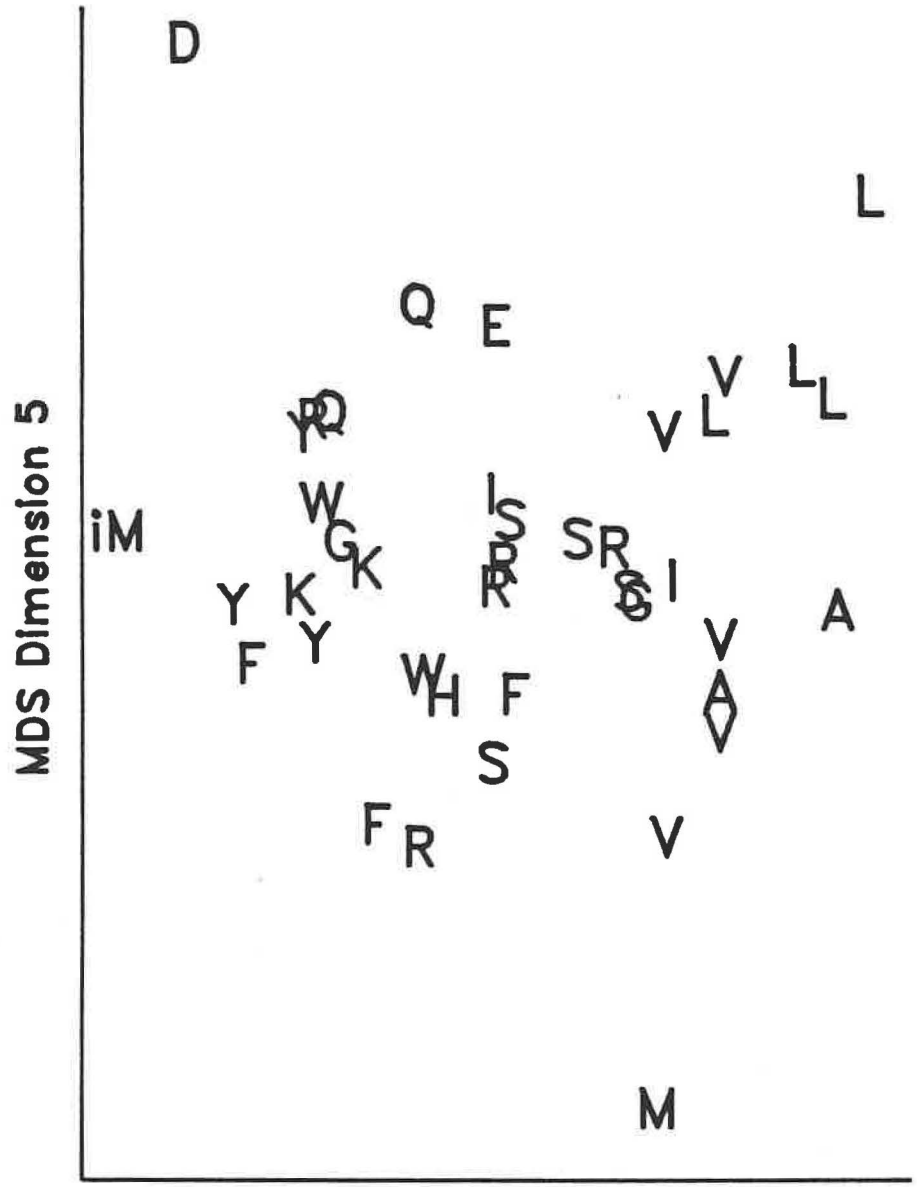
Fig. 2.--A plot of the MDS coordinates for dimensions four and five for the analysis of 48 positions in the 41 tRNA data set. Left: the location of each tRNA in the map is marked by the single-letter amino acid code; initiator methionine is marked "iM". The center of each letter marks the coordinate. Right: Same coordinates as Left, but replaces amino acid code with chemical group code. Abbreviations are: A, aliphatic; C, carboxylic; O, aromatic; and, H, amphoteric (G, M, and iM are not classified).

Fig. 3.--Plot of Kullback's I statistic against its expected Chi-square distribution for 66 positions in the 195 tRNA data set. The nine numbered positions show stronger than random association of base in the six chemical groups.

Fig. 4.--The panel for each chemical group shows the bases indicative of tRNAs in that group. The bases were obtained from a two-way contingency table analyses on the nine positions. A line marks positions where the bases do not provide information about that group of tRNAs. Codes for multiple bases are: R = A or G; S = C or G; W = A or U; M = A or C; and, Y = C or U. Glycine and cysteine tRNAs are not included in any group.

Fig. 1
Nicholas
and
McClain





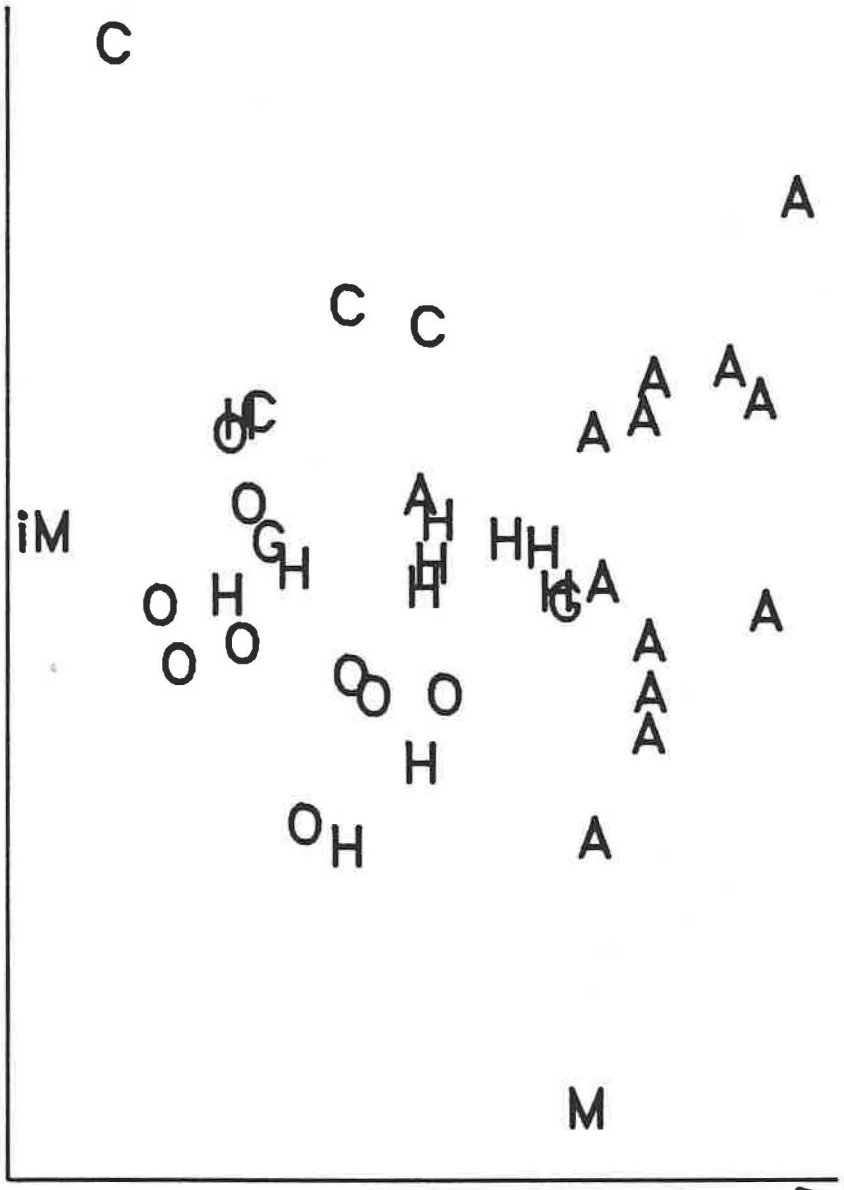


Fig. 2
Nicholas and McClain

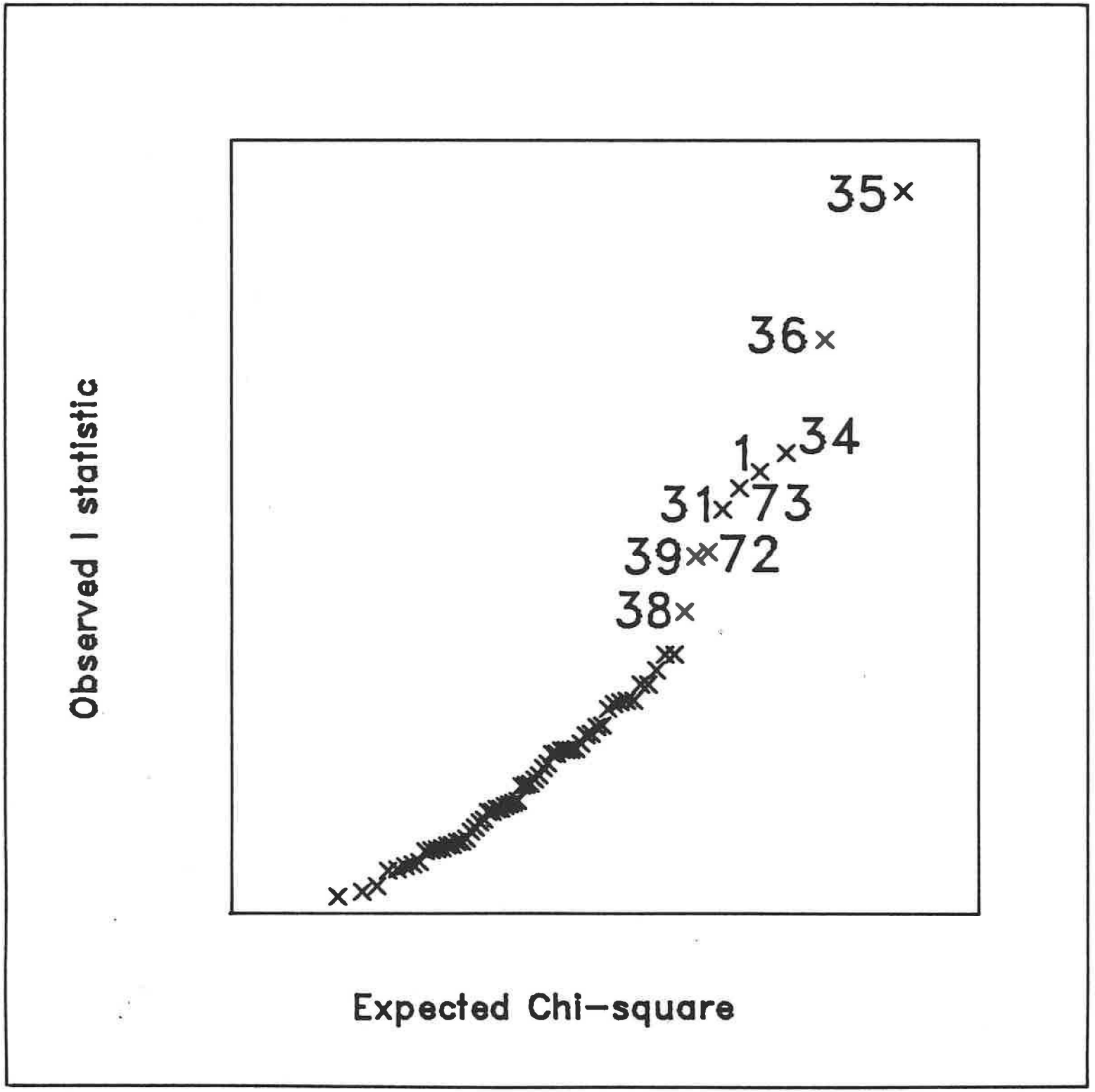
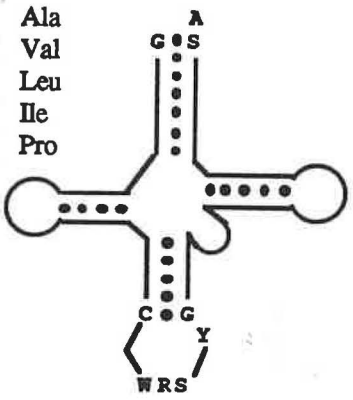
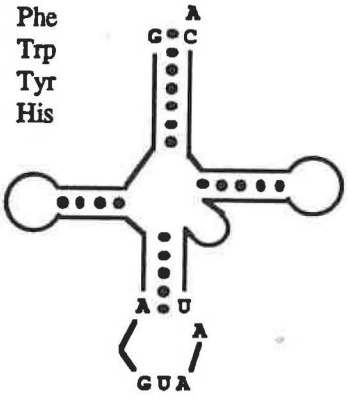


Fig. 3
Nicholas and McClain

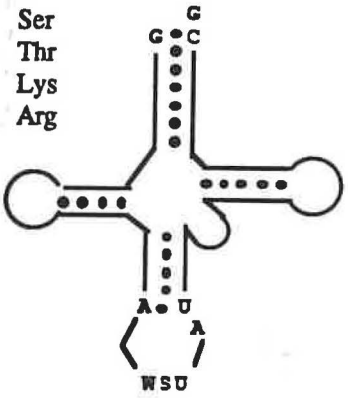
Fig. 4
Nicholas and McClain



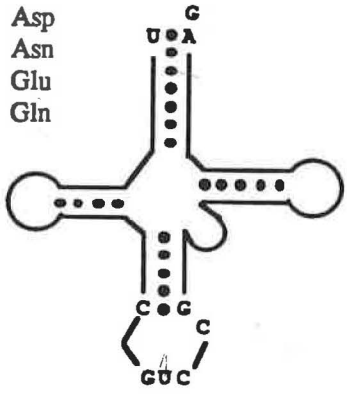
ALIPHATIC



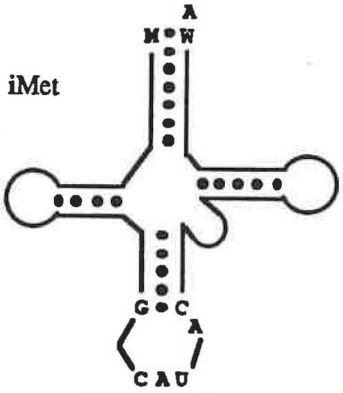
AROMATIC



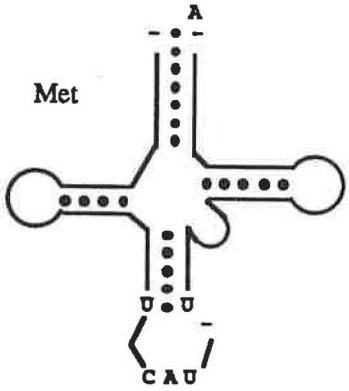
AMPHOTERIC



CARBOXYLIC



INITIATOR METHIONINE



ELONGATOR METHIONINE