A Method of Predicting tRNA Secondary Structure from Nucleotide Sequences

Li-Yeh Chuang, Yu-Da Lin, and Cheng-Hong Yang, *Member, IAENG*

Abstract—**The transfer RNA (tRNA) is an evolutionarily conserved important small molecule acting as central role to translation. All tRNAs have the characteristics that has structure resembles cloverleaves and having lengths mostly within 63-200 bases, moreover, the same anticodon of tRNAs from orthologous species usually fold into highly similar secondary structure. Hence, tRNAs have been extensively discussed in research of molecule evolution, besides that many research reports have directed the possibility that structure of tRNAs could lead to understanding some key points in the course of evolution. Although many systems provide functions such as prediction of secondary structure, it is still not satisfactory for biologists' needs on superior sensitivity. Hence, we introduce a new algorithm designed specifically for a better prediction of tRNA secondary structure.**

Index Terms—**tRNA, secondary structure, evolutionary.**

I. Introduction

The family of tRNAs, a kind of RNA molecules, has the special function to translate gene codons into amino acids, which will be concatenated simultaneously together by ribosome to form the proteins. Each tRNA molecule can recognize some specific combination of nucleic acid and carry the respective amino acid into the protein-building machinery. In order to successfully add amino acid to the end of a growing polypeptide, the tRNA must accurately read the coded segment from messenger RNA. Hence, predicting the anticodon of tRNA has become an important subject on researches. Furthermore, both the characteristics of central role played by tRNA to sustain every vital task in a cell and of its short sequence length have made it a popular tool in the field of evolution study. Recently, research reports have suggested that the conserved structure in tRNA is involved in some of the earliest and most profound evolutionary events.

A standard secondary structure of tRNA molecules takes the form of a cloverleaf comprising four stacked pairs (stem structure), four hairpin loops, one multi-loops and three spacer bases where the four stacked pairs contain acceptor stem (A-stem) from 7 bp long, dihydrouridine stem (D-stem) from 3 to 4 bp long, anticodon stem from 5 bp long, TΨC stem (T-stem) from 5 bp long and four hairpin loops contain TΨC loop (T-loop) from 5 to 7 bases long, variable loop (V-loop) from 3 to 24 bases long, anticodon loop from 7 bases long, and dihydrouridine loop (D-loop) from 4 to 11 bases long.

There are many tools that can provide the methods of prediction of tRNA secondary structure, e.g. tRNAscan-SE [1], ARAGORN [2] and tRNAfinder [3]. However, most of them cannot satisfy the users demand. In light of the above cases, we adjusted the tRNA secondary structure prediction method. This method provides an easy way to perform tRNA gene search and predict the secondary structure from found gene.

II. Method

A. Predicted tRNA gene and tRNA secondary structure method

The notions used in this paper are listed below:

AS: Acceptor Stem (1 to 7 and 66 to 72),
AD: The base between AS and DS (base 8 to 9),
DS: Dihydrouridine Stem (10 to 13 and 22 to 25),
DL: Dihydrouridine Loop (14 to 21),
DC: The base between DS and CS (27 to 31 and 39 to 43),
CS: Anticodon Stem (27 to 31 and 39 to 43),
CL: Anticodon Loop (32 to 38),
VL: Variable Loop (44 to 48),
TS: TΨC Stem (49 to 53 and 61 to 65),
TL: TΨC Loop (54 to 60),
mm: Any base pair different from Watson-Crick or GU base pairs.

Function of our proposed algorithm is dived into two parts, first part is directed at search of tRNA gene in the inputted complete genomic sequence and the other one is directed at prediction of the optimal tRNA secondary structure. The ability of global search for tRNA gene depends on the validity of tRNA secondary structure prediction, the consideration going in the predicting process will be detailed below. The goal of tRNA secondary structure prediction is to find an optimal structural configuration and the associated anticodon. There are multiple choices for the construction of tRNA secondary structure from one tRNA gene sequence and lack of precision of prediction results in failure in folding into tertiary structure and false anticodon detected, which implies loss of amino acid carried by that tRNA. Therefore,

L.Y. Chuang is with the Department of Chemical Engineering, I-Shou University , 84001 , Kaohsiung, Taiwan (E-mail: chuang@isu.edu.tw)
Y.D. Lin is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung, Taiwan (E-mail: e0955767257@yahoo.com.tw)
C.H. Yang is with the Department of Electronic Engineering, National Kaohsiung University of Applied Sciences, 80778, Kaohsiung Taiwan (phone: 886-7-3814526#5639; E-mail: chyang@cc.kuas.edu.tw)
C.H. Yang is also with the Network Systems Department, Toko University, 61363, Chiayi, Taiwan. (E-mail: chyang@cc.kuas.edu.tw).

we are taking into account of characteristics that seem to have impact on prediction, such as numbers of hydrogen bonds inside each portion of the structure and the total number of them, sequence length, loop length, intron length, GC% and significant pattern. Our parameters for constraints come from our experimental runs on tools, tRNAscan-SE, ARAGORN and tRNAfinder, besides that, observations on multiple databases, such as GtRNAdb [4] (http://rna.wustl.edu/GtRDB/), tRNADB-CE [5] (http://trna.nagahama-i-bio.ac.jp/cgi-bin/trnadb/index.cgi), tRNAdb [6] (http://trnadb.bioinf.uni-leipzig.de/), and literatures [7,8,9,10,11] both give us an insight into the approach. The selection of parameters and adjustment of their values have been optimized by reducing the incorrect predictions. The observations from characteristics of irregular tRNA structures are shown in Table I.

Table I. Length of each structure constraints.

Structure	Minimum length	Maximum length
AS	6	7
AD	2	2
DS	4	4
DL	4	11
DC	1	1
CS	5	5
CL	7	7
VL	4	21
TS	5	5
TL	4	7
Last base	1	1

Positions of DL at 14, 15, 18 and 19 play the determinative role in folding into tertiary structure because they support the critical L-shaped structure of tRNA molecule. Introns were found in the CL between sequence positions of 37 and 38, lengths of intron for various species are showed as follows:

$$\text{Intron Length} = \begin{cases} 6 \text{ to } 121, & \text{Archaea} \\ 0, & \text{Bacteria} \\ 1 \text{ to } 60, & \text{Eukarya} \end{cases} \tag{1}$$

In light of the above description, lengths of complete tRNA for various species are shown as follows:

$$\text{Length}(S) = \begin{cases} 63 \text{ to } 217, & \text{Archaea} \\ 63 \text{ to } 95, & \text{Bacteria} \\ 63 \text{ to } 155, & \text{Eukarya} \end{cases} \tag{2}$$

In predicting tRNA gene, our method will use GC% as a basis to filter out unfeasible sequence fragments, this preprocess will speed up the computation. The percentage of nucleotides G and C in sequence S is denoted as $GC_{ratio}(S)$, according to our experiments on tRNA sequence of *Saccharomyces cerevisiae* mitochondrion tRNA-Arg (accession number NC_001224, range of 69289 to 69362) from GtRNAdb database. The notions, $GC_{ratio}(S)$ and GC%(S), are defined as follows:

$$GC_{ratio}(S) = \frac{S_{total}(G) + S_{total}(C)}{|S|} \tag{3}$$

$$GC\%(S) = \begin{cases} \text{discard, if } GC_{ratio}(S) < 18\% \\ \text{retain, if } GC_{ratio}(S) \geq 18\% \end{cases} \tag{4}$$

From sequences that satisfy GC% requirement, an optimal structure will be constructed based on the characteristics of hydrogen bonds, i.e. one, two and three hydrogen bonds for AU, GC and GU pairs respectively. The following restrictions mainly dealing with throwing out unfeasible sequence were given as the direct discovery from literature:

$$AS_{GU}(S) = \begin{cases} \text{discard, if Numbers of GU pairings} > 3 \\ \text{retain, if Numbers of GU pairings} \leq 3 \end{cases} \tag{5}$$

$$AS_{mm}(S) = \begin{cases} \text{discard, if Numbers of mismatches} > 3 \\ \text{retain, if Numbers of mismatches} \leq 3 \end{cases} \tag{6}$$

$$DS_{GU}(S) = \begin{cases} \text{discard, if Numbers of GU pairings} > 2 \\ \text{retain, if Numbers of GU pairings} \leq 2 \end{cases} \tag{7}$$

$$DS_{mm}(S) = \begin{cases} \text{discard, if Numbers of mismatches} > 3 \\ \text{retain, if Numbers of mismatches} \leq 3 \end{cases} \tag{8}$$

$$CS_{GU}(S) = \begin{cases} \text{discard, if Numbers of GU pairings} > 2 \\ \text{retain, if Numbers of GU pairings} \leq 2 \end{cases} \tag{9}$$

$$CS_{mm}(S) = \begin{cases} \text{discard, if Numbers of mismatches} > 3 \\ \text{retain, if Numbers of mismatches} \leq 3 \end{cases} \tag{10}$$

$$TS_{GU}(S) = \begin{cases} \text{discard, if Numbers of GU pairings} > 3 \\ \text{retain, if Numbers of GU pairings} \leq 3 \end{cases} \tag{11}$$

$$TS_{mm}(S) = \begin{cases} \text{discard, if Numbers of mismatches} > 3 \\ \text{retain, if Numbers of mismatches} \leq 3 \end{cases} \tag{12}$$

$$Stem_{GU+mm}(S) = \begin{cases} \text{discard, if Numbers of GU pairings} \\ \qquad \text{and mismatches} > 7 \\ \text{retain, if Numbers of GU pairings} \\ \qquad \text{and mismatches} \leq 7 \end{cases} \tag{13}$$

$$\text{hydrogen bond score} = \begin{cases} 1 & \text{, if GU base pair} \\ 2 & \text{, if AU base pair} \\ 3 & \text{, if GC base pair} \end{cases} \tag{14}$$

Although the most stable structure for some sequence can be calculated, the problem that non-tRNA sequences are falsely predicted to fold into cloverleaf like structure still remains to be solved. To overcome this difficulty, we are giving score depends on graphical pattern arranged by Mark [10]. The tRNA gene candidates that are overlapped with each other will be searched according to their scores, then the one with highest score will be selected as the algorithm result.

$$\text{punished score} = \begin{cases} -4 & \text{,if base pair conform to fig.2} \\ 0 & \text{,if base pair not conform to fig.2} \end{cases} \qquad (15)$$

$$\text{Score} = \sum \text{hydrogen bonds score} + \sum \text{punished score} \qquad (16)$$

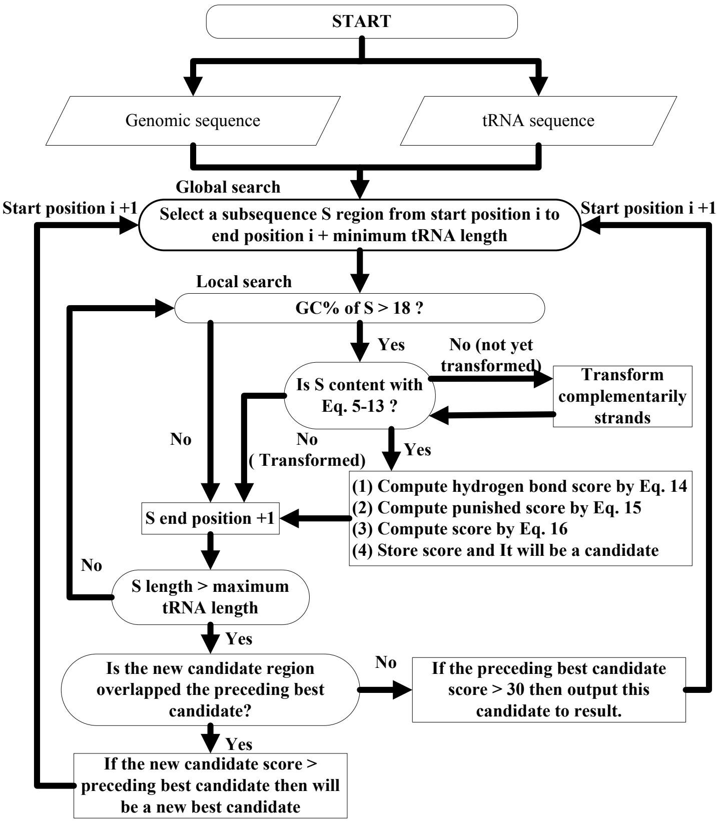
START
Genomic sequence
tRNA sequence
Global search
Start position i +1
Select a subsequence S region from start position i to end position i + minimum tRNA length
Start position i +1
Local search
GC% of S > 18 ?
Yes
No (not yet transformed)
Is S content with Eq. 5-13 ?
Transform complementarily strands
No
No (Transformed)
Yes
(1) Compute hydrogen bond score by Eq. 14
(2) Compute punished score by Eq. 15
(3) Compute score by Eq. 16
(4) Store score and It will be a candidate
S end position +1
No
S length > maximum tRNA length
Yes
Is the new candidate region overlapped the preceding best candidate?
No
If the preceding best candidate score > 30 then output this candidate to result.
Yes
If the new candidate score > preceding best candidate then will be a new best candidate

Fig.1 tRNA gene search and secondary structure prediction flowchart

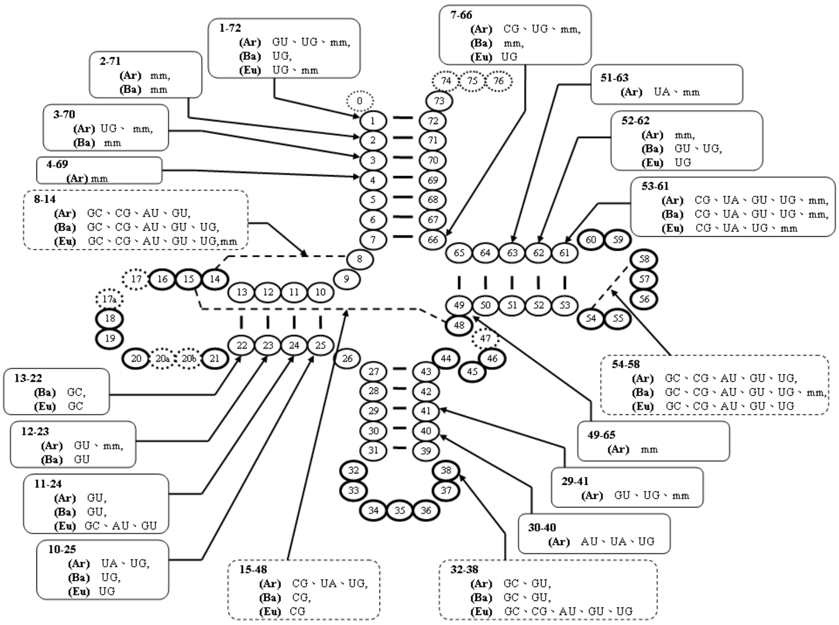
1-72 (Ar) GU、UG、mm, (Ba) UG, (Eu) UG、mm
2-71 (Ar) mm, (Ba) mm
3-70 (Ar) UG、 mm, (Ba) mm
4-69 (Ar) mm
8-14 (Ar) GC、CG、AU、GU, (Ba) GC、CG、AU、GU、UG, (Eu) GC、CG、AU、GU、UG,mm
7-66 (Ar) CG、UG、mm, (Ba) mm, (Eu) UG
51-63 (Ar) UA、mm
52-62 (Ar) mm, (Ba) GU、UG, (Eu) UG
53-61 (Ar) CG、UA、GU、UG、mm, (Ba) CG、UA、GU、UG、mm, (Eu) CG、UA、UG、mm
54-58 (Ar) GC、CG、AU、GU、UG, (Ba) GC、CG、AU、GU、UG、mm, (Eu) GC、CG、AU、GU、UG
49-65 (Ar) mm
29-41 (Ar) GU、UG、mm
30-40 (Ar) AU、UA、UG
32-38 (Ar) GC、GU, (Ba) GC、GU, (Eu) GC、CG、AU、GU、UG
15-48 (Ar) CG、UA、UG, (Ba) CG, (Eu) CG
10-25 (Ar) UA、UG, (Ba) UG, (Eu) UG
11-24 (Ar) GU, (Ba) GU, (Eu) GC、AU、GU
12-23 (Ar) GU、mm, (Ba) GU
13-22 (Ba) GC, (Eu) GC

Fig.2 tRNA patterns

A cloverleaf structure and all patterns which represent never found base pairs within over 4000 tRNA genes from 50 fully sequenced genomes [10]

III. Result and Discussion

The accuracy the test for our tRNA secondary structure prediction is based on tRNA gene sequences obtained from the database Sprinzl (http://www.old.uni-bayreuth.de/departments/biochemie/sprinzl/trna/), and the complete chromosome genome obtained from species (NC_*): NC_002695 [Escherichia coli O157:H7] was used to test the tRNA gene search method. The Sprinzl database provides a set of trusted true positives for testing the sensitivity. It contains the most comprehensive tRNA from wide variety of organisms, and divides into three different sets of tRNA gene, from Archaea (161 sequences), Bacteria (686 sequences) and Eukaryota (443 sequences).

According to Sprinzl database, our test result revealed the prediction sensitivity for species Archaea and Bacteria were both 100%, and that of species Eukarya was 99.3% (Table II). The incorrect prediction for Eukarya occurred at three: two sequences, Sprinzl ID DQ8510 and DA9360, were absent prediction, and one sequence was wrong anticodon predicted. For the prediction of tRNA gene from chromosome NC_002695, all genes were correctly predicted shown in table III. Thus, the results demonstrated our suggested method outperforms other methods in various areas.

In the development process, we observed that secondary structures predicted or obtained from many software or databases for an identical sequence can vary in different configuration even with the same anticodon. This unexpected finding sparked our interest that length of a secondary structure will not affect the stability of stem structure as long as the following rules were satisfied. Length of loop structures can be adjusted to fulfill construction of stems, even if there are non-pairing appeared at the stems. This situation occurs at only one stem most of the time, without violation of complementary at other stems. What of the above observation implied is that consideration of occurrences of non-pairing on stem structures should be made when predicting, thus flexibility was given to each stem for better prediction. In addition, some restriction were made to maintain integrity of overall structure, e.g. if DS appears to have only one base pair, then the other stems will not have multiple mismatches. Although structures from different compositions can have the same anticodon, the deciding factor is whether the prediction can be folded into a tertiary structure. Thus, we will limit the number of prediction being one by considering the restriction on tertiary structure. One of the common features in predicting tRNA secondary structure, nucleotides appearing at fixed positions analyzed by Mark with more than 4,000 sequences, has been much tone down its importance by our own experience from prediction results. The observation, no nucleotide is fixed at one position in secondary structure, is what separates our method from that of ARAGORN system, which appears to be inferior in predicting irregular tRNA gene because of its strict restriction of sequence motif “TRGYNAA” on T loop. The pattern used is effective to search tRNA genes, thus a modified system based on score was employed to provide a flexible structure prediction. The proposed score system allows predicted structure to have unusual portion stray away from canonical rules while penalties are given if this proportion exceeds some limit. After multiple tests conducted by ourselves, we decided that the threshold should be set as 30 for the optimal results. In order to process protein synthesis, tRNAs matured in the cytoplasm need to have 3’CCA terminus at the positions 74, 75 and 76. However, tRNAs in Eukarya lack this characteristic and many of the tRNAs from Bacteria and Archaea do not have 3’CCA neither, which leads us to abandon this feature as tRNA gene searching.

When we compared our method to the other popular tools today, tRNAscan-SE and ARAGORN, we concluded that there are no definite winners here. If computational resource is the most concerned criterion, then ARAGORN will lead others in this area, followed by our method and far ahead of tRNAscan-SE, the reason being ARAGORN mainly using sequence motif TRGYNAA as the search basis in the chromosome where we are using a combination of attributes, GC% and acceptor stem, to search the most fitting segment. In assessing the quality of prediction, our method leads in sensitivity because we are using extraordinary folding structure as the goal in secondary structure prediction, where as tRNAscan-SE wins in having better false positive rate. The most important contribution from our method is better tRNA secondary structure prediction can be guaranteed, though the unsatisfactory false positive rate has been planned to be our focus of improvement in the future.

In this paper, we have proposed a method capable of predicting accurate tRNA secondary structure, which was supported by the results obtained.

Table II. tRNA detection rates for tRNAscan-SE, ARAGORN and our method

Sequence soure	No. of	No. of tRNAs detected			Detection rate (%)		
	tRNAs	tRNAscan-SE [1]	ARAGORN [2]	Our	tRNAscan-SE [1]	ARAGORN [2]	Our
Archaea	161	160	**161**	**161**	99.38	**100.00**	**100.00**
Bacteria	686	682	684	**686**	99.42	99.71	**100.00**
Eukaryota	443	437	435	**440**	98.65	98.19	**99.32**
total	1290	1279	1280	**1287**	99.12	99.22	**99.77**

Table III. Information of count anticodons for searched NC_002695.

2nd base

1st base	A		G		C		U		3rd base
A	Phe	0	Ser	0	Cys	0	Tyr	0	A
G		2		2		1		3	
C	Leu	1	Ser	1	Trp	1	Stop	0	
U		1		1	Stop	0		0	
A	Leu	0	Pro	0	Arg	4	His	0	G
G		1		1		0		1	
C	Leu	3	Pro	1	Arg	1	Gln	2	
U		1		1		3		2	
A	Val	0	Ala	0	Gly	0	Asp	0	C
G		2		2		4		3	
C	Val	0	Ala	0	Gly	1	Glu	0	
U		5		3		1		4	
A	Ile	0	Thr	0	Ser	0	Asn	0	U
G		3		2		1		4	
C	Met	15	Thr	1	Arg	1	Lys	0	
U	Ile	0		1		0		5	

IV. Conclusion

Our method provides to predict the tRNA gene and the secondary structure. Users can use either complete chromosome or sequence fragments to predict the locations of tRNA gene and tRNA secondary structures.

We chose the chromosome genome from species *Escherichia coli* O157:H7 as the test set, to find the complete tRNA genes. We were adopting a method considering unique tRNA secondary structure to predict the tRNA location. We then constructed any possible structure and selected the most stable structure. Our test results, demonstrated this method not only can search the exact tRNA location but also predict the best structure, and the tRNA anticodon predicted conforms to the literature. Although the problem of high false positive rate still exists, we have collected reliable patterns which can help to improve the quality of our prediction. According to our analysis, this method provides an opportunity for biologists and researchers to locate tRNA gene with more efficiency.

References

[1] M.L. Todd and R.E. Sean, “tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence,” *Nucleic Acids Research*, Vol.25, 1997, pp.955-964.

[2] L. Dean and C. Bjorn, “ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequence,” *Nucleic Acids Research*, Vol. 32, 2004, pp.11-16.

[3] M. Kinouchi, K. Kurokawa, “tRNAfinder: A Software System To Find All tRNA Genes in the DNA Sequence Based on the Cloverleaf Secondary Structure,” *Journal of Computer Aided Chemistry*, Vol.7, 2006, pp.116-124

[4] P.P. Chan, T.M. Lowe, “GtRNAdb: a database of transfer RNA genes detected in genomic sequence,” *Nucleic Acids Research*, Vol.37, 2009, pp.93-97.

[5] T. Abe, T. lkemura, Y. Ohara, H. Uehara, M. Kinouchi, S. Kanaya, Y. Yamada, A. Muto and H. lnokuchi, “tRNADB-CE: tRNA gene database curated manually by experts,” *Nucleic Acids Research*, Vol.37, 2009, pp.163-168.

[6] J. Frank, M. Mario, K.H. Roland, S. Mathias, F.S. Peter and P. Joern, “tRNAdb 2009: compilation of tRNA sequences and tRNA genes,” *Nucleic Acids Research*, Vol. 37, 2009, pp. D159-D162.

[7] G. Dirheimer, G. Keith, P. Dumas and E. Westhof, “Primary secondary and tertiary structures of tRNAs,” *ln: Söll D, RajBhandary U, eds. tRNA: Structure, biosynthesis, and function. Washington DC*: ASM Press, 1995, pp.93-126.

[8] J.R. Macery, J.A. Schulte ll, A. Larson, B.S. Tuniyev, N. Orlov and T.J. Papenfuss, “Molecular Phylogenetics, tRNA Evolution, and Historical Biogeography in Anguid Lizards and Related Taxonomic Families,” *Molecular Phylogenetics and Evolution*, Vol.12, No.3, 1999, pp. 250-272.

[9] M. Kinouchi, S. Kanaya and Y. Kudo, “Detection of Transfer RNA Based on the Cloverleaf Secondary Structure,” *Journal of Computer Aided Chemistry*, Vol.1, 2000, pp.76-81.

[10] M. Christian and G. Henri, “tRNomics: Analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features,” *Bioinformatics*, Vol.8, 2002, pp.1189-1232.

[11] T. Vickie, M. Tom and A. David, “A novel method for finding tRNA genes,” *RNA*, Vol.9, 2003, pp.507-517.

[12] F.G. Johan, I.B. Clara and E.D. Edgar, “tRNA structure from a graph and quantum theoretical perspective,” *Journal of Theoretical Biology*, 2006, pp.574-582.