Molecular and Evolutionary Statistical Data Analysis Package: An Approach toward Investigating DNA Codon Frequency Parameters in Distributed Database

Miran H. Mohammed[1] and Shad A. Mohammed[2]

[1]Department of Basic Science, University of Sulaimani, Sulaimaniyah, Kurdistan Region – Iraq

[2]Department of Biology, University of Sulaimani, Sulaimaniyah, Kurdistan Region – Iraq

***Abstract*—*DNA* codon frequency is a parameter that can be used to compare the frequency of particular codons in the genome of living organisms. Certain organisms are biased toward using certain codons to code for certain amino acids. This bias is thought to be due to an evolutionary process that has shaped nearly every organism's genome. To get around this, one would need to fetch raw DNA data from already available databases such as nucleotide and GenBank. The obtained data will then would be automatically parsed into their corresponding codons and saved in distributed databases. Certain tools would be introduced to the database that allows instant statistical analysis as well as comparative genomic analysis. Thus, users would be able to retrieve codon frequency from the database of any gene of interest that can be found in GenBank and nucleotide databases, as well as analysis of codon frequency/amino acid frequency bias tables.**

***Index Terms*—Codon, Fasta, GenBank, Gene, Nucleotide.**

I. Introduction

Living organisms are rich of data, and the most important data can be found within their genetic makeup. Genomes are highly variable regarding the data they contain [1,2], that is, the nucleotides that span the whole genome of any given organism. One can write and hence use computer software to start investigating this massive data containing arena.

One of the known methods used to investigate genomes is to analyze certain parameters about DNA [3,4]. Such parameters include frequency parameters of DNA codons [5,6]. To investigate such parameters and the data they hold, one should postulate hypotheses and generate linear models such as substitution patterns [3].

Measuring frequency of codons in a DNA sequence will provide important insights about the quality and the history of the sequence [7]. One of the parameters that can be tested in this aspect is the measurement of codon usage bias of sequences [8]. Developing tools to analyze such parameters would provide critical measurements of mutation rates of the DNA sequence available in databases worldwide. Such measurements could be further analyzed using molecular statistics algorithms and summarize all the output in one figure.

II. Literature Review

In general, codons specify amino acids and any given protein is specified by its unique codon sequence. An amino acid is coded by more than one codon from the synonymous codon family. This is basically due to the degeneracy of the genetic code, especially at the third codon position [9]. For instance, glycine's synonymous codon family consists of four different codons, namely GGT, GGG, GGC, and GGA. These codons are not equivalent and depending on the usage of any given codon by an organism one codon is preferred over the other. GGC and GGT are highly preferred by *Escherichia coli* genome to code for glycine [10]. However, mammalian mitochondria tend to use GGA more often [11,12]. A common knowledge on the codon usage is that distantly related organisms have different patterns of synonymous codon usage, Table I demonstrates comparison of glycine codon usage of human genome, *E. coli* genome, and human mitochondrial genome. This is also applied in the different variants and genotypes of the same species. For instance, human metapneumovirus genotypes vary in their synonymous codon usage [13].

Simply, codon usage is characterized by the frequencies of the 64 codons. A comparison of codon usage among genes or genomes is simply a comparison of codon frequency tables. Such comparative studies are of prime importance to understand gene regulation. For instance, herpesviruses utilize unusual codon usage, significantly different from their hosts' codon usage bias, to make their glycoproteins [14]. This accounts for the reason of their persistence in staying infective in one's lifetime.

Furthermore, viruses that infect vertebrates would have a bias toward using certain codons similar to that of their hosts [15]. However, certain human viruses, particularly

Pure and Applied Science Conference | Koya University
Paper ID: ICPAS2018.SCT21, 5 pages
DOI: 10.14500/icpas2018.sct21
Received 15 February 2018; Accepted 27 March 2018
Conference paper: Published 01 August 2018
Conference track: Soft Computing Techniques (SCT)
Corresponding author's e-mail: miran.mohammed@univsul.edu.iq

human influenza-A virus, extensively use A-ending codons in their genomes [16]. This can be reasoned to the fact that ATP is highly abundant in the cellular environment, and hence, this increases transcription efficiency [17].

There is yet another reason for studying codon frequencies. It is a common knowledge that codon frequency parameters and rate ratio parameters are important to deduce substitution process. Thus, one would need to take the output generated by our online software and start analyzing substitution pattern of any given gene or genome on comparison.

III. Related Work

There are several similar projects available online; the majority of them require console installation. Furthermore, their fetching code gets interrupted once the software is installed on the platform. For instance, DAMBE is a wonderful software that can retrieve DNA sequences and users can start performing their meta-analysis offline. However, nearly every version of DAMBE fails to retrieve GenBank and FASTA files on request. Similarly, MEGA 7.0 is a powerful offline tool that can perform molecular data analysis, but it lacks data retrieval. On the other hand, our online tool, Molecular and Evolutionary Statistical Data Analysis (MESDA) Package, provides instant retrieval of FASTA files on entering accession codes. In addition, the software has been tested from different browsers and on different operating systems (that is, Windows 10 and Linux Ubuntu 16.04).

MESDA is built in such a way that can be installed as EXE file and it can be used online as well and it can retrieve DNA sequences from GenBank.

The software provides the following functionalities:

1. FASTA file retrieval
2. Generation of codon frequency table
3. Reverse complementation of a given DNA sequence
4. Starting from desired codon positions
5. Amino acid frequency table
6. Distributed database.

A. Objectives

The objectives of this study were as follows:

1. Generating a distributed database that can retrieve and share DNA sequence data.
2. Generate relevant *in silico* molecular toolkits for analyzing DNA sequences.
3. Write algorithms for software that performs statistical analysis of codon usage bias of the provided DNA data.

B. Hypothesis

A. Null hypotheses

- DNA sequences have equal proportions of nucleotides, and hence, the probability for any nucleotide to occur at any position is 0.25.
- There are equal proportions of codons to spread throughout any given DNA sequence.

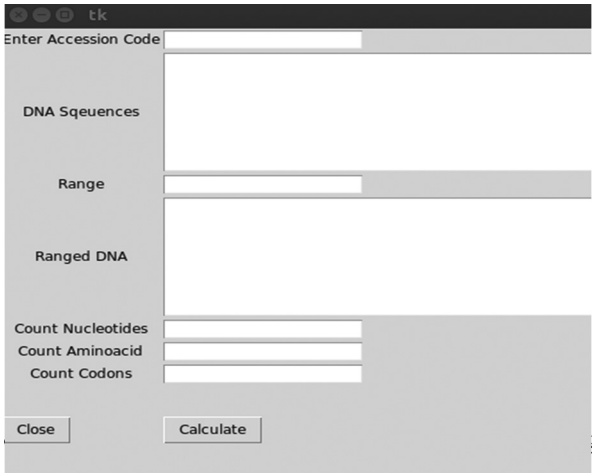

Fig. 1. General view of Molecular and Evolutionary Statistical Data Analysis package software framework

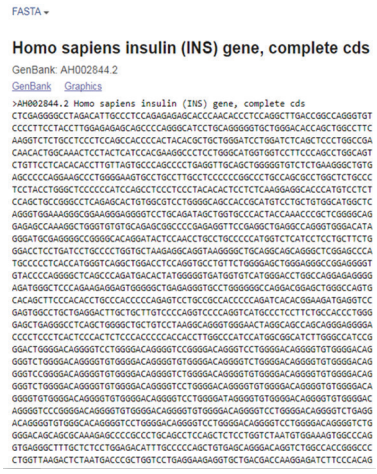

Fig. 2. FASTA File from GenBank to be retrieved.

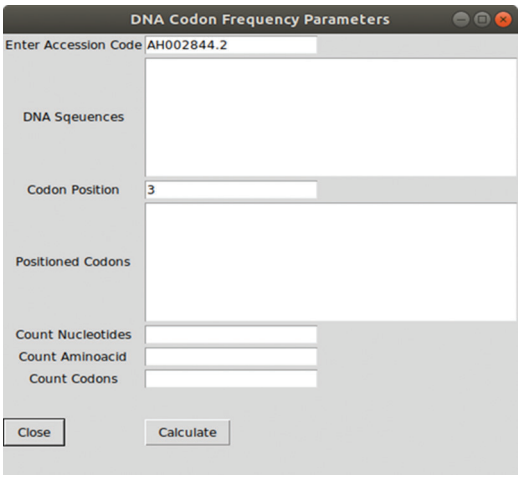

Fig. 3. Entering the accession code and ranging number.

B. Alternative hypotheses

- The probability of nucleotides' occurrence sums up to one.
- DNA codons are not equally spread throughout DNA, and it varies depending on the species.

DOI: http://dx.doi.org/10.14500/icpas2018

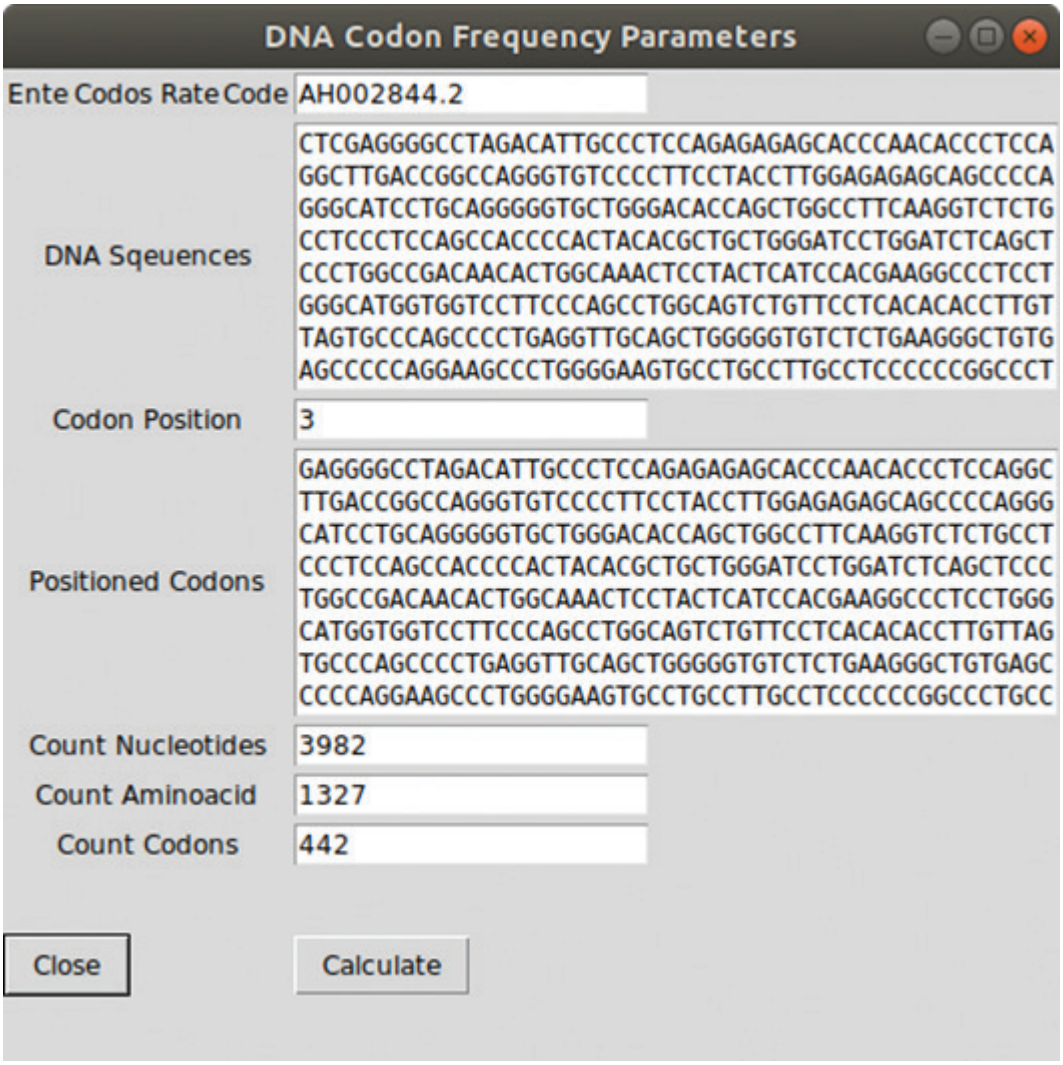

Fig. 4. First output of giving accession code (AH002844.2).

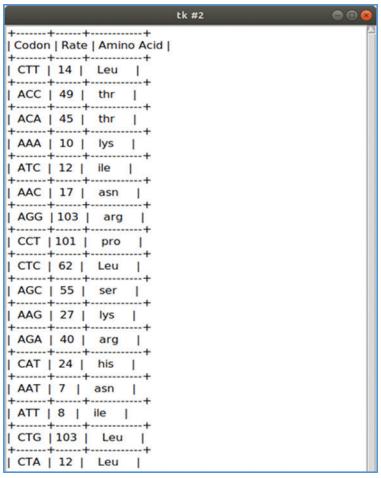

Fig. 5. Codon usage bias table generated by Molecular and Evolutionary Statistical Data Analysis.

IV. Materials and Methods

In this paper, two main programming languages have been used which are Python and Extensible Markup Languages (XML). Python is used as a control programming language that works on analyzing and parsing DNA into separate parts to calculate and rate each of the variables separately.

The XML part is used as a storage place that contains DNA information on separate files, and these data can be called from Python whenever they are required. Furthermore, XML is used to save the details of codons and comparing them with DNA nucleotide values, as well as their corresponding amino acids (for each codon).

Furthermore, the idea of the database retrieval (from National Center for Biotechnology Information [NCBI]/ GenBank) is implemented, and this is for storing nucleotide

TABLE I
Comparison of Glycine Synonymous Codon Usage Among Genomes of *Escherichia coli*, *Homo sapiens*, and *Human mitochondria*

Glycine usage		
Escherichia coli O157:H7	*Homo sapiens*	Mitochondrion of *Homo sapiens*
GGT 24.1 (39552)	GGT 10.8 (437126)	GGT 8.9 (80137)
GGC 27.9 (45695)	GGC 22.2 (903565)	GGC 20.8 (187077)
GGA 9.0 (14707)	GGA 16.5 (669873)	GGA 19.5 (175656)
GGG 11.9 (19534)	GGG 16.5 (669768)	GGG 9.6 (86524)

DOI: http://dx.doi.org/10.14500/icpas2018

TABLE II
STATISTICS ANALYSIS OF LEUCINE USAGE IN HUMAN INS GENE

	Observed_Leucine Counts						
	CTC	CTT	CTG	CTA	TTA	TTG	Total
	62	14	103	12	2	12	205
Percentage	0.30	0.07	0.50	0.06	0.01	0.06	1

Expected_Leucine Counts					
CTC	CTT	CTG	CTA	TTA	TTG
34.2	34.2	34.2	34.2	34.2	34.2

	Chi-test
P value	3.848E-48
Alpha	0.01

sequence details for each organism on requesting data through entering accession codes. Later, these data are called from Python for comparison and meta-analysis purposes, one of which is to find the rate of similarity and differences among such sequences. There are built-in functions in the databases we have created that calculate codon usage, codon frequency, nucleotide frequencies, so as to test the hypotheses postulated in this research.

V. Implementation and Design

In this research, their many steps have been taken to conduct the practical works and to get the accurate results from the different sources. The main practical has been created with using Python programming language, XML, and Excel sheet file. Moreover, each of those has their own usages for manipulating the software named MESDA Package.

The general idea of the software is to search for a particular DNA sequence based on the accession code that the user needs to enter in the field that is specified for accession code as a string, then hit the "Find Button" to start the process. In addition, the user must be aware that the software will not function if the accession code field is empty.

This software works with online distributed database - NCBI database, as it fetches the search result from there, gets the data as "FASTA" file, which is used to obtain nucleotide or peptide sequences. Although the internet connection is required for providing this connection, as the FASTA file has been downloaded to your drive it will not require to download it again as the software can function in offline mode as well.

After the software retrieves the FASTA file from the distributed server, other outputs will be shown and provided. For instance, user can bound the limitation of DNA sequences in two opposite bound sides, [1:] or [:−1]. In bioinformatics term, the former would mean the sense strand and the later would generate the antisense strand. Furthermore, it shows the frequency rate of nucleic acid and codons, in two different frames.

Furthermore, as the framework has been created using widgets from "Tkinter" library, it provided many facilities to create many widgets on the frames so that the software has the ability to show the exact number of calculated frequencies of nucleotides, amino acid, and codons on different fields.

VI. Results and Discussion

For test, the results from this research, some of accession code has been applied and to get different outputs in numbers of nucleotides, amino acid, and codons. Furthermore, the number of those finding number for each of them will be varied in the same species by giving different ranging numbers of the nucleotides.

Fig. 1 is the first windows of the software, and from this box the accession code "AH002844.2" is entered, which is an example that has been used to Homo sapiens insulin (INS) gene, and the result will return in FASTA type, and it will be download as txt file so that it can be used to parse and analyze the nucleotides.

The content of this file is that there are the basic information about the gene, accession code, and name of gene and lines of nucleotides. This will be the same for any testing accession codes with FASTA return type (Fig. 2).

Whenever, the accession code is entered into the specified field, with giving the ranging number, for example, "3" (Fig. 3). This would retrieve the DNA from the nucleotide position 3.

It will give an output like in the following Fig. 4, which returns count of nucleotides, amino acid, and codons.

The second output that the software generates includes codon usage bias table, as shown in following Fig. 5.

Given the output from the codon usage bias table, if one takes leucine amino acid usage into account, the comparison would be as shown in Table II.

As it is demonstrated in Table II, the human INS gene uses CTG to code for leucine more frequently than any other given codon in the leucine synonymous codon family. On the other hand, TTA is used the least. When the data are calculated using Chi-square test, the p value is much less than the alpha, which indicated a significant difference in the usage of different leucine coding codons.

VII. Conclusion

We generated important computer database and software that have the following functionalities:

1. Count codon frequencies of any given DNA sequence.
2. Output codon frequency tables and nucleotide frequency tables of a given DNA sequences.
3. Provide statistical analysis of DNA codon frequencies.

References

[1] P.D. Dixit, T.Y. Pang, F.W. Studier and S. Maslov. Recombinant transfer in the basic genome of *Escherichia coli*. *Proceedings of the National Academy of Sciences*, vol. 112, no. 29, pp. 9070-9075, 2015.

[2] J.L. Draper, L.M. Hansen, D.L. Bernick, S. Abedrabbo, J.G. Underwood, N. Kong, B.C. Huang, A.M. Weis, B.C. Weimer, A.H. Van Vliet and N. Pourmand. Fallacy of the unique genome: Sequence diversity within

DOI: http://dx.doi.org/10.14500/icpas2018

single *Helicobacter pylori* strains. *MBio*, vol. 8, no. 1, pp. e02321-e02316, 2017.

[3] S.K. Pond, W. Delport, S.V. Muse and K. Scheffler. Correcting the bias of empirical frequency parameter estimators in codon models. *PLoS One*, vol. 5, no. 7, p. e11230, 2010.

[4] X. Xia, T. Wei, Z. Xie and A. Danchin. Genomic changes in nucleotide and dinucleotide frequencies in *Pasteurella multocida* cultured under high temperature. *Genetics,* vol. 161, pp. 1385-1394, 2002.

[5] H.J. Thomas, C.R. Cantor and H.N. Munro. "Evolution of protein molecules". *Mammalian Protein Metabolism,* vol. 3, no. 21, pp. 132, 1969.

[6] X. Xia. How optimized is the translational machinery in *E. coli*, *S. typhimurium*, and *S. cerevisiae*? *Genetics*, vol. 149, pp. 37-44, 1998.

[7] S.K. Behura, B.K. Singh and D.W. Severson. Antagonistic relationships between intron content and codon usage bias of genes in three mosquito species: Functional and evolutionary implications. *Evolutionary Applications,* vol. 6, no. 7, pp. 1079-1089, 2013.

[8] R.P. Jeffrey and N.E. Moriyama. "Evolution of codon usage bias in Drosophila". *Proceedings of the National Academy of Sciences,* vol. 94, no. 15, 7784-7790, 1997.

[9] R. Oi and K. Ikehara. Direct evidence for GC-NSF (a) hypothesis on creation of entirely new gene/protein. *Current Proteomics*, vol. 15, no. 1, pp. 13-23, 2018.

[10] E.P. Rocha. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, vol. 14, no. 11, pp. 2279-2286, 2004.

[11] X.J. Min and D.A. Hickey. DNA asymmetric strand bias affects the amino acid composition of mitochondrial proteins. *DNA Research*, vol. 14, no. 5, pp. 201-206, 2007.

[12] X. Xia. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. *Molecular Biology and Evolution,* vol. 15, pp. 336-344, 1998.

[13] Q. Zhong, W. Xu, Y. Wu and H. Xu. Patterns of synonymous codon usage on human metapneumovirus and its influencing factors. *BioMed Research International*, vol. 2012, Article ID: 460837, 7 Pages, 2012.

[14] Y.C. Shin, G.F. Bischof, W.A. Lauer and R.C. Desrosiers. Importance of codon usage for the temporal regulation of viral gene expression. *Proceedings of the National Academy of Sciences*, vol. 112, no. 45, pp. 14030-14035, 2015.

[15] J.B. Miller, A.A. Hippen, S.M. Wright, C. Morris and P.G. Ridge. Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect. *Biomed Genet Genomics,* vol. 2, no. 2, pp. 1-5, 2017.

[16] E.H. Wong, D.K. Smith, R. Rabadan, M. Peiris and L.L. Poon. Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. *BMC Evolutionary Biology*, vol. 10, no. 1, p. 253, 2010.

[17] X. Xia. Maximizing transcription efficiency causes codon usage bias. *Genetics,* vol. 144, pp. 1309-1320, 1996.

DOI: http://dx.doi.org/10.14500/icpas2018