Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between human and 16 animals

Abstract
The mitochondrial DNA of vertebrates generally has same structure and functions, even share same numbers of genes, tRNAs, rRNAs and codon regions with different sequences in a narrow range of genome size. The resource of database was downloaded from the Genbank of National Center for Biotechnology Information (NCBI), and using a particular computational program to achieve the best results of alignment and statistical calculations to estimate the transition/transversion in the nucleotide and amino acid substitution, additionally, estimate the evolutionary distance rate between Human versus 16 animals. The results of maximum likelihood method show high rates of substitutions mainly from adenine and thymine to cytosine and guanine (A=>C; A=>G; T=>C; T=>G), respectively. Guanine (G) was the most conserved and stable nucleotide from changes in all over 17 organisms that may be related to the strength of the chemical bonds. The observation of evolutionary distance by the number of substitutions per site between sequences is shown for all three codon positions and non-codon regions. The scores put organisms in groups by comparing the numbers between pairs of sequence within difference and similarity, such as the human, chimpanzee, and gorilla had less distance among them, whilst, a remarkable likelihood between bison and the water buffalo were observed despite the historical and geographical distance between them. Additionally, studying the effect of substitution scores in nucleotide and amino acids on the synonymous/non-synonymous codon substitution in evolutionary distance bias was discussed.

Keywords: maximum likelihood, mitochondrial DNA, evolutionary distance, PERL, CPAN modules

Introduction
Mitochondrial DNA (mtDNA) is a masterpiece of polynucleotide intelligence provided as a double stranded circular DNA, to be the spirit and manager of molecular activities in eukaryotic cells. The nucleic DNA activity and regulation depend on the signals and the levels of the tRNA and rRNA which mtDNA produced in the cell. The foremost recognition-grabbing issue, mtDNA has the ability to adapt with each individual cell by modifying the sequence by slightly the initiation and termination points, likewise, the direction of transcription 5'→3' or 3'→5'.1 Mitochondria generate most of the cellular energy within the form of adenosine triphosphate ATP, regulate cellular oxidation-reduction state and integrate several of the signals for initiating necrosis. By means of retrograde signaling, mitochondrial communicate of these events to the nucleus and thus modulate nuclear organic phenomenon and cell cycle. In human, mitochondrial pathology leads to a massive array of pathologies, and many diseases result from various defects of mitochondrial biogenesis and maintenance, metabolism chain complexes or individual mitochondrial proteins.2

Perhaps, the estimation of the distance between two sequences is the simplest phylogenetic analysis, because the calculation of pairwise distances is the first step in distance-matrix methods of phylogeny reconstruction. Cluster algorithms used to convert a distance matrix into a phylogenetic tree. Markov-process models of nucleotide substitution used in distance estimation form the basis of likelihood and Bayesian analysis of multiple sequences on a phylogeny.3

To estimate the number of substitutions, it is needed a probabilistic model to describe changes between nucleotides this purpose. Continuous-time Markov chains are commonly used for the nucleotide sites in the sequence are normally measured to be evolving independently of each other. Substitutions at any particular site are described by a Markov chain, with the nucleotides to be the states of the chain. The main advantage of a Markov chain is that it has no memory given the present, likewise, the future does not depend on the past. In other words, the probability with which the chain jumps into different nucleotide states depends on the current state, but not on how the current state is reached. This is referred to as the Markovian property. Besides this basic assumption, it is often placed further constraints on substitution rates between nucleotides, leading to variable models of nucleotide substitution.4

The first application of a maximum likelihood method to tree construction was made by Cavalli-Sforza and Edwards (1967) for estimation gene frequency data. Later, Felsenstein (1973, 1981) developed maximum likelihood algorithms for amino acid and nucleotide sequence data. Because this approach involves fairly sophisticated statistical theory, that presented only some basic principles of the method without any mathematical details.5 A critical element is how the probabilities of the various changes are calculated. These probabilities depend on assumptions concerning the process of nucleotide substitution and the branch lengths, which in turn depend on the rate of substitution and the evolutionary time. These branch lengths are usually unknown and must be estimated as part of the process of computing the likelihood. The methods for

Volume 5 Issue 6 - 2017

Omar Esmail Hamad,1 Ismail Akyol,2 Mehmet Sait Ekinci,1 Emin Ozkose1
1Department of Animal Science, Kahramanmaras Sutcu Imam University, Turkey
2Biotechnology and Gene Engineering Laboratory, Turkey

Correspondence: Mehmet Sait Ekinci, Kahramanmaras Sutcu Imam University, Agriculture Faculty, Animal Science Department 46100/ Kahramanmaras, Turkey, Tel: +90 344 2802100, Fax: +90 344 2802109, Email: sekinci@ksu.edu.tr

Received: June 30, 2017 | Published: July 19, 2017
Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between human and 16 animals


discoversing the branch lengths that maximize the likelihood value usually involve an iterative approach also the likelihoods depend on the model of nucleotide substitution, a tree with the largest likelihood value under one substitution model. The maximum likelihood method is computationally extremely time-consuming, and so was not used often in the past. With the development of fast computers, the method is now used fairly often, although it is an exhaustive version it is still only applicable to a modest number of taxa.6,7

To outline some main points, as an observation from previous researchers have been documented about estimate the mitochondrial DNA evolutionary distance within maximum likelihood method among animals within various visions and scoring parameters. Started with the pronouncement by Irwin, Kocher, and Wilson (1991) studies on Evolution of the cytochrome-b gene of mammals that obtained 17 complete gene sequences representing three orders of hoofed mammals (ungulates) and dolphins (cetaceans). The fossil record of some unguulate lineages allowed estimation of the evolutionary rates for various components of the cytochrome DNA and amino acid sequences. The relative rates of substitution at first, second, and third positions within codons are in the ratio 10 to 1 to at least 33. For deep divergences (>5million years) it appears that both replacements and silent transversion in this mitochondrial gene can be used for phylogenetic inference. Phylogenetic genes include the association of (Drosophila 12 Genomes et al.) cetaceans, arctoidactyls, and perissodactyls to the exclusion of elephants and humans, (2) pronghorn and fallow deer to the exclusion of bovids like cow, sheep, and goat, (3) sheep and goat to the exclusion of other pecans such as cow, giraffe, deer, and pronghorn, and (4) advanced ruminants to the exclusion of the chevrotain and other articulated. Comparisons of these cytochrome sequences support current structure-function models for this membrane-spanning protein. Although there has been relatively a research results into mitochondrial DNA sequence divergence and diversity Chen & Li about genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees, The average sequence divergence was only 1.24% 5 0.07% for the human-chimpanzee pair, 1.62% 5 0.08% for the human-gorilla pair, and 1.63% 5 0.08% for the chimpanzee-gorilla pair relation. More importantly, the modern hypothesis of the evolutionary relationship between human and pig that based on the assumption of similarity in some organs tissues like kidneys and eyes.8,10

All in all, the huge similarity in structure and functions appears in the genomic mitochondrial DNA (mtDNA) of vertebrates. That was an encouraging point to consider about having a chance to make a comparative view among seventeen organisms including the human mtDNA within maximum likelihood method to estimate the evolutionary distance and the substitution effects of the nucleotides and amino acids on the codons frequencies. Moreover, may be generating hypothesis about the evolutionary of the organisms in interest within the convenient algorithmic method to compute the sequences of the complete genomic mitochondrial DNA in alignment method and applying particular mathematical methods. To accomplish the aims of discovery the evolutionary distances between Human versus other sixteen vertebrate organisms.

Materials and methods

The sources of database

For Maximum likelihood estimation of the evolutionary distance of the complete genomes of Mitochondrial DNA (mtDNA) between Human’s versus other 16 animals are investigated. The databases of all vertebrates for mitochondrial DNA (mtDNA) sequences were downloaded from the Genbank of National Center for Biotechnology Information NCBI11 (www.ncbi.nlm.nih.gov/GENOME).11,12 To find out the most trusted and proved sequences, by looking for same sequences could be found in International Nucleotide Sequence Database Collaboration (INSDC) (www.insdc.org). In this case, it is worth to mention the Human’s mtDNA is the Cambridge reference sequence (isogg.org/wiki/Cambridge_Reference_Sequence), is count as the central sequence which all researchers on mitochondrial DNA of human need to use it for comparison and studying the variation rate from this sequence.13

The reason behind choosing these organisms as it mentioned in Table 1, being in interest to get the genomic mtDNA and apply them in the comparative study, is the historical observation in the similarity of morphological and physiological characteristics which known as related to each other, like, Arabian camel with Bactrian camel, so between sheep and goat, Likewise, some of these similarities between organisms were causd the most controversial and debatable issues among the biologists, for the evolutionary relationship between human and chimpanze.14

<table>
<thead>
<tr>
<th>S. no</th>
<th>Taxa</th>
<th>Latin name</th>
<th>Accession numbers</th>
<th>INSDC Number</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human</td>
<td>Homo sapiens</td>
<td>NC_012920</td>
<td>J01415.2</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>Chimpanzee</td>
<td>Pan troglodytes</td>
<td>NC_001643</td>
<td>D38113.1</td>
<td>62</td>
</tr>
<tr>
<td>3</td>
<td>Gorilla</td>
<td>gorilla gorilla</td>
<td>NC_011120</td>
<td>X93347.1</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>Cattle</td>
<td>Bubalus bubalis</td>
<td>NC_006853</td>
<td>AYS26085.1</td>
<td>(Chung HY, Ha JM., 2005)*</td>
</tr>
<tr>
<td>5</td>
<td>Water buffalo</td>
<td>Bubalus bubalis</td>
<td>NC_006295</td>
<td>AY702618.1</td>
<td>56*</td>
</tr>
<tr>
<td>6</td>
<td>Bison</td>
<td>Bison bison</td>
<td>NC_012346</td>
<td>EU177871.1</td>
<td>64</td>
</tr>
<tr>
<td>7</td>
<td>Arabian camel</td>
<td>Camelus dromedarius</td>
<td>NC_009849</td>
<td>EU159113.1</td>
<td>(Huang X et al. 2007)*</td>
</tr>
<tr>
<td>8</td>
<td>Bactrian camel</td>
<td>Camelus bactrianus</td>
<td>NC_009628</td>
<td>EF212037.2</td>
<td>65</td>
</tr>
<tr>
<td>9</td>
<td>Horse</td>
<td>Equus caballus</td>
<td>NC_001640</td>
<td>X79547.1</td>
<td>66</td>
</tr>
<tr>
<td>10</td>
<td>Sheep</td>
<td>Ovis aries</td>
<td>NC_001941</td>
<td>AF010406.1</td>
<td>67</td>
</tr>
<tr>
<td>11</td>
<td>Goat</td>
<td>Capra hircus</td>
<td>NC_005044</td>
<td>GU295658.1</td>
<td>68</td>
</tr>
</tbody>
</table>

Table 1 List of organisms which involved in the evolutionary study, with the information of database of the complete genome, mtDNA
More importantly, including in the list some animals that considered as a high contrast with all, even out the cycle of mammals like chicken, then include the sequences in a parallel way with each other’s for comparison view between them evenly. Lastly, the combination of these organisms actually put this study in a unique position as far as it is concerned.15 As shown in Table 1, demonstrates the comparison of Human’s mitochondrial DNA versus other 16 vertebrate organisms with the accession number of NCBI, and the INSDC number. Furthermore, with publication in the Medline database of references and abstracts on life sciences and biomedical (PubMed), but three references of these sequences were unpublished and they have NCBI Project numbers only. Firstly, cattle’s project number is 13366 submitted in 22-February-2005 (www.ncbi.nlm.nih.gov/nuccore/NC_000845). Secondly, water buffalo with project number 13052 submitted in 02-August-2004 (www.ncbi.nlm.nih.gov/nuccore/NC_001323). Finally, Arabian camel with project number 20873 submitted in 17-September-2007 (www.ncbi.nlm.nih.gov/nuccore/NC_001700).

Computational approach

In the most trusted and depended websites which provide an open source bioinformatics tool services and databases resources. Practical extraction and report language known as Perl which is one of the major program applied in Bioinformatics for decades (https://www.perl.org/) supported by organization of Comprehensive Perl Archive Network (CPAN) that provides thousands of modules shared from scientists and computer programmers studying on bioinformatics.16–19 Nevertheless, needed to extract some mathematical functions from (megasoftware.net) which is academic open-public software for molecular evolutionary genetic analysis MEGA7-CC-Porto.20

Algorithm

A critical point is to decide choosing which algorithmic method would be used because it is related with the best way for interring data in a computer with choosing and designing the codes, then apply them to obtain the best results as it possible. The modules of Perl programming language which invented by the legendary computer programmer Larry Wall (en.wikipedia.org/wiki/LarryWall), were invented by the legendary computer programmer Larry Wall (en.wikipedia.org/wiki/LarryWall), were downloading from CPAN (www.cpantoolkit.org) that provide thousands of modules shared from scientists and computer programmers studying in bioinformatics.16–19 Nevertheless, needed to extract some mathematical functions from (megasoftware.net) which is academic open-public software for molecular evolutionary genetic analysis MEGA7-CC-Porto.20

The 17 sequences of mtDNA compiled and saved in a FASTA format (filename.fasta) then the codes were downloaded from the shell of CPAN by using the black window called command (CMD) in windows (http://www.bioperl.org/wiki/Installing_BioPerl_on_Windows) by using special codes for test and install in the computer, as an example (CPAN>test Bio::Tools::Run::Alignment::Muscle) and install the module if it works in this code (CPAN>install Bio::Tools::Run::Alignment::Muscle), next, open installed codes with a text editor like (ActiveState Komodo IDE8) and start to import the own data and mathematical problems by using some specific regular expressions to compile the all in one code like (Seq(x) ="<sequence (x)>") and (use <module>);. Then save the code in Perl format (filename.pl).21–24

Another essential point is the best modules were served the research poses. Firstly, an object for the calculation of an iterative multiple sequence alignment from a set of unaligned sequences or alignments using the MUSCLE program (Bio::Tools::Run::Alignment::Muscle) authored by Christopher Fields in 2011, (metacpan.org/pod/Bio::Tools::Run::Alignment::Muscle), the Representations for biological sequence alignment (Bio::Tools::Run::Alignment::Overview) announced by Felipe da Veiga Leprevost in 2014, (metacpan.org/pod/Bio::Tools::Run::Alignment::Overview). Thirdly, the interface for evolving sequences (Bio::Seq Evolution::EvolutionI) reported by Christopher Fields in 2014, (metacpan.org/pod/Bio::Seq Evolution::EvolutionI). Finally, the module of Maximum likelihood methods (Bio::Tools::Run::Phylo::ProtML) authored by Jason Stajich in 2011, (metal pan.org/pod/Bio::Tools::Run::Phylo::ProtML).

Alignment of 17 sequences

The 17 sequences of mtDNA were arranged in parallel depending on coding and noncoding regions of DNA even the proteins to distinguish regions of similarity and disparity. Consequently, get the distance and evolutionary relationships between the sequences. The dynamic programming algorithm of the multiple sequence alignment is by adding spaces (INDEL) or gaps in the sequences. Then calculate the highest scores of the alignment matrix were always being the diagonal arrows to yield an equal length sequences, in condition that obtain an optimum score value, then going to calculate the number of matches, mismatches and gaps, finally, apply the next model of maximum value.25–26

The computational multiple sequence alignment (MUSCLE) method used to provide high accuracy for creating different arrangements of high scale amino acids and nucleotide sequences.27–29 The velocity and precision of MUSCLE were contrasted with other three methods. Firstly, Tree-based Consistency Objective Function For alignment Evaluation (T-Coffee). Secondly, multiple sequence alignment program for amino acid or nucleotide sequences (MAFFT). Finally, with Clustal is a series of widely used computer programs for multiple sequence alignment (CLUSTALW). The achievement of

---

**Table Continued...**

<table>
<thead>
<tr>
<th>S. no</th>
<th>Taxa</th>
<th>Latin name</th>
<th>Accession numbers</th>
<th>INSDC Number</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Pig</td>
<td>Sus scrofa</td>
<td>NC_000845</td>
<td>AF034253.1</td>
<td>69</td>
</tr>
<tr>
<td>13</td>
<td>Chicken</td>
<td>Gallus gallus</td>
<td>NC_001323</td>
<td>X5292.1</td>
<td>70</td>
</tr>
<tr>
<td>14</td>
<td>Rabbit</td>
<td>Oryctolagus cuniculus</td>
<td>NC_001913</td>
<td>AJ001588.1</td>
<td>71</td>
</tr>
<tr>
<td>15</td>
<td>Dog</td>
<td>Canis lupus familiaris</td>
<td>NC_002008</td>
<td>U96639.2</td>
<td>72</td>
</tr>
<tr>
<td>16</td>
<td>Domestic cat</td>
<td>Felis catus</td>
<td>NC_001700</td>
<td>U20753.1</td>
<td>73</td>
</tr>
<tr>
<td>17</td>
<td>House mouse</td>
<td>Mus musculus</td>
<td>NC_005089</td>
<td>AY172335.1</td>
<td>74</td>
</tr>
</tbody>
</table>

---

**References:**

most elevated or joint highest rank in precision in all tests. At the point when utilized without refinement its precision is the same as T-Coffee or MAFFT and is the speediest at adjusting extensive sequences.\textsuperscript{30,31}

**Relative synonymous codon usage (RSCU)**

The numerous amino acids are coded by more than one codon, thus the several of multiple codons for a given amino acids are synonymous. Nevertheless, many genes display a nonrandom usage of synonymous codons for specific amino acids.\textsuperscript{12,32,33} In addition, the codes of the mathematical problem extracted from program MEGA7-cc-Porto (www.megasoftware.net) in a particular file format (filename.mao).\textsuperscript{34,35}

**Maximum likelihood**

The maximum likelihood method considered as the cornerstone of modern statistics depend on the parametric model of evolution appropriate for the characters and algorithm that will search through the trees. The model depends essentially on the nature of the characters under study, among the many possible models of character evolution.\textsuperscript{26}

The statement of the problem, suppose when to have a random sample $x_1, x_2, \ldots, x_n$ whose assumed probability distribution depends on some unknown parameter $\theta$. The primary goal here will be to find a point estimator $\hat{\theta}(x_1, x_2, \ldots, x_n)$ such that $\hat{\theta}(x_1, x_2, \ldots, x_n)$ is a “good” point estimate of $\theta$, where $x_1, x_2, \ldots, x_n$ are the observed values of the random sample. For example, if planned to take a random sample $x_1, x_2, \ldots, x_n$ for which the $x_i$ are assumed to be normally distributed with mean $\mu$ and variance $\sigma^2$, then our goal will be to find a good estimate of $\mu$, say, using the data $x_1, x_2, \ldots, x_n$ that we obtained from our specific random sample. The Basic Idea (onlinecourses.science.psu.edu; megasoftware.net).

**Estimating the evolutionary distances between genomic sequences**

The evolutionary distance between sequences usually is measured by the number of a polynucleotide or amino acid substitutions appear between them and the Alignment methods are used to compute evolutionary distances between DNA and protein sequences as a basis for phylogenetic reconstruction. It is calculated from the number of word matches between them, additionally; compute the substitutions of nucleotide, amino acids, and the synonymous-non-synonymous codes. Nucleotide sequences are compared nucleotide-by-nucleotide, these distances could be computed for protein coding and noncoding nucleotide sequences. Residue-by-residue for amino acid and Codon-by-codon for synonymous-non-synonymous codons with complete detection of gaps of missing data treatments and the substitution included the transition-transversion within maximum likelihood method.\textsuperscript{36-40}

**Results and discussion**

**Comparative view of the nucleotide and amino acids sequences in sizes**

The Figure 1 provides a vision about the difference of the genome sizes in mitochondrial DNA between human and the other vertebrates’ species, also the amino acid size numbers were around 5000 when the nucleotide sizes around 17000 bases, representing the total translation of the protein. However, the number of proteins is constant and same in all species that is 13, and even we have 22 tRNAs and 2 rRNAs, these numbers did not change between the 17 vertebrates. They have the same annotated structure with alternative lengths and sequences, to help to provide more functions for the same job, and this is what molecular evolution means.\textsuperscript{41-43}

![Figure 1](image)

**Figure 1** The length graph of nucleotide bases and amino acids number in Human’s mitochondrial DNA with other 16 vertebrates.

**Estimation of the codon usage bias**

The results of the codon bias, Table 2 show a prejudice in codon frequencies has been used for the conformity with previous results between nucleotide composition within amino acid composition, it is shown the top scores in count for Leucine, Isoleucine, Proline and Serine, and even with relative synonymous codon usage.\textsuperscript{44-48} The reason behind these results is due to tRNA corresponding to the codons CUA, UCA, AGC…..etc., are more abundant, because the translationary machinery tend to use abundant tRNA to produce proteins.\textsuperscript{49-50}

**Table 2** The frequency account of the codons and the relative synonymous codon usage (RSCU), in all over the 17 aligned mammalian mitochondrial sequences

<table>
<thead>
<tr>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU(F)</td>
<td>109</td>
<td>0.95</td>
<td>UCU(S)</td>
<td>100</td>
<td>1.19</td>
<td>UAU(Y)</td>
<td>130</td>
<td>1.05</td>
<td>UGU(C)</td>
<td>28.4</td>
<td>0.82</td>
</tr>
<tr>
<td>UUC(F)</td>
<td>120</td>
<td>1.05</td>
<td>UCC(S)</td>
<td>117</td>
<td>1.38</td>
<td>UAC(Y)</td>
<td>118</td>
<td>0.95</td>
<td>UGC(C)</td>
<td>40.8</td>
<td>1.18</td>
</tr>
<tr>
<td>UUA(L)</td>
<td>132</td>
<td>1.32</td>
<td>UCA(S)</td>
<td>126</td>
<td>1.49</td>
<td>UAA(*)</td>
<td>135</td>
<td>1.38</td>
<td>UGA(*)</td>
<td>74.6</td>
<td>0.76</td>
</tr>
<tr>
<td>UUG(L)</td>
<td>44.3</td>
<td>0.44</td>
<td>UCG(S)</td>
<td>33.5</td>
<td>0.4</td>
<td>UAG(*)</td>
<td>84.4</td>
<td>0.86</td>
<td>UGG(W)</td>
<td>28.6</td>
<td>1</td>
</tr>
<tr>
<td>CUU(L)</td>
<td>90.8</td>
<td>0.91</td>
<td>CUC(P)</td>
<td>138</td>
<td>1.27</td>
<td>CAU(H)</td>
<td>126</td>
<td>1.07</td>
<td>CGU(R)</td>
<td>27.1</td>
<td>0.78</td>
</tr>
<tr>
<td>CUC(L)</td>
<td>101</td>
<td>1.01</td>
<td>CCC(P)</td>
<td>138</td>
<td>1.27</td>
<td>CAC(H)</td>
<td>110</td>
<td>0.93</td>
<td>CGC(R)</td>
<td>29.9</td>
<td>0.86</td>
</tr>
<tr>
<td>CUA(L)</td>
<td>185</td>
<td>1.85</td>
<td>CCA(P)</td>
<td>120</td>
<td>1.11</td>
<td>CAA(Q)</td>
<td>145</td>
<td>1.39</td>
<td>CGA(R)</td>
<td>35.4</td>
<td>1.02</td>
</tr>
</tbody>
</table>

**Citation:** Hamadd OE, Akyol I, Ekinci MS, et al. Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between human and 16 animals. MOJ Proteomics Bioinform. 2017;5(6):186–193. DOI: 10.15406/mojpb.2017.05.00180
**Estimation of transition/transversion matrix by maximum composite likelihood (mcl)**

The results obtained by estimating the Maximum Likelihood substitution patterns called transition (inside the purine group or the pyrimidine group) and the transversion (between the purine and pyrimidine groups), by observing the changes in the nucleotide through the 17 mitochondrial genome sequences, are illustrated vertically in columns of Table 3. The Guanine (G) was the most conservative nucleotide in spite of showing substitution changes, and the most changeable nucleotide to others was the Adenine (A) in general, calculating the total of substitutions from adenine to the other nucleotides was the highest 33.7277 which came from (A=T 5.9789+A=G 11.5959+A=C 16.1529) and the lowest total score of substitutions were from guanine to other nucleotides 8.3552 which came from (G=A 7.0297+G=T 0.6117+G=G 0.7318), another essential point is the highest substitution shown the transition inside the pyrimidine group between two nucleotides T=C 19.9855 and C=T 20.3322.

**Table 3 Maximum likelihood estimation of transition/transversion bias**

<table>
<thead>
<tr>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
<th>Codon</th>
<th>Count</th>
<th>RSCU</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUG(L)</td>
<td>47.4</td>
<td>0.47</td>
<td>CCG(P)</td>
<td>37.6</td>
<td>0.35</td>
<td>CAG(Q)</td>
<td>63.2</td>
<td>0.61</td>
<td>CGG(R)</td>
<td>15.9</td>
<td>0.46</td>
</tr>
<tr>
<td>AUU(I)</td>
<td>140</td>
<td>0.97</td>
<td>ACU(T)</td>
<td>130</td>
<td>1.1</td>
<td>AAU(N)</td>
<td>138</td>
<td>0.96</td>
<td>AGU(S)</td>
<td>45.6</td>
<td>0.54</td>
</tr>
<tr>
<td>AUC(I)</td>
<td>134</td>
<td>0.93</td>
<td>ACC(T)</td>
<td>143</td>
<td>1.21</td>
<td>AAC(N)</td>
<td>149</td>
<td>1.04</td>
<td>AGC(S)</td>
<td>85.5</td>
<td>1.01</td>
</tr>
<tr>
<td>AUA(I)</td>
<td>158</td>
<td>1.1</td>
<td>ACA(T)</td>
<td>154</td>
<td>1.31</td>
<td>AAA(K)</td>
<td>181</td>
<td>1.45</td>
<td>AGA(R)</td>
<td>62.4</td>
<td>1.79</td>
</tr>
<tr>
<td>AUG(M)</td>
<td>66.5</td>
<td>1</td>
<td>ACG(T)</td>
<td>44.9</td>
<td>0.38</td>
<td>AGK(K)</td>
<td>69.3</td>
<td>0.55</td>
<td>AGG(R)</td>
<td>37.9</td>
<td>1.09</td>
</tr>
<tr>
<td>GUU(V)</td>
<td>36.2</td>
<td>0.85</td>
<td>GCU(A)</td>
<td>69.5</td>
<td>1</td>
<td>GAU(D)</td>
<td>56.1</td>
<td>1</td>
<td>GGU(G)</td>
<td>32.8</td>
<td>0.76</td>
</tr>
<tr>
<td>GUC(V)</td>
<td>39.6</td>
<td>0.93</td>
<td>GCC(A)</td>
<td>92.4</td>
<td>1.46</td>
<td>GAC(D)</td>
<td>55.9</td>
<td>1</td>
<td>GCC(G)</td>
<td>46</td>
<td>1.06</td>
</tr>
<tr>
<td>GUA(V)</td>
<td>68.5</td>
<td>1.6</td>
<td>GCA(A)</td>
<td>74.7</td>
<td>1.18</td>
<td>GAA(E)</td>
<td>71.7</td>
<td>1.16</td>
<td>GGA(G)</td>
<td>66.6</td>
<td>1.54</td>
</tr>
<tr>
<td>GUG(V)</td>
<td>26.7</td>
<td>0.62</td>
<td>GCG(A)</td>
<td>17.1</td>
<td>0.27</td>
<td>GAG(E)</td>
<td>51.5</td>
<td>0.84</td>
<td>GGG(G)</td>
<td>27.6</td>
<td>0.64</td>
</tr>
</tbody>
</table>

*Termination codes of transcription.*

Additionally, the results agreed with the next following research in nucleotide behaviors that may be related to strength of the chemical bonds in spite of different area investigations also some of them called guanine an ancestral nucleotide as the most conserved nucleotide.51–54

**Nucleotide pair frequencies from alignment of 17 sequences**

The calculation of the transition/transversion in a maximum probable number of 16 nucleotide pairs that could obtain from four different nucleotides, through alignment of 17 sequences in the positions 1st, 2nd and 3rd respectively. The R ratio used as a parameter score that equal 1, between transition and transversion that show harmony in levels of exchanges in all positions. In other words, the number of transitions is semi equal the transversion in all 16 pairs of nucleotides within the codon positions. More importantly, in the second part of Table 4, which illustrate the frequencies of the nucleotide pairs as a genome map of mtDNA estimating the probability of codon frequencies in the three positions respectively, also help to predict the sequences of proteins by this map. For instance, the highest levels of AA exemplifies, a high ratio of Asparagine N and Lysine K because they codon contain the AA.

Furthermore, as demonstrated the top number of observations in Table 4, it could be noticed that the harmony in the numbers of observation through the three positions. The line chart seems to be one line in spite of there are three lines in all over the 17 mammalian mitochondrial DNA sequences aligned against each other’s.55–56

**Nucleotide evolutionary distance**

The parameter of results depended on the numbers of base substitutions per site from between sequences as are shown in. The analyses were conducted using the Maximum composite likelihood model. Also, the rate variation among sites was modeled with a gamma distribution shape parameter score value is equaled 1. Codon positions included 1st+2nd+3rd+Noncoding. Additionally, all positions containing gaps and missing data were eliminated. There were a total of 14430 positions in the final dataset.

**Table 4 The transition/transversion calculated of 16 probable nucleotide pair frequencies by alignment of 17 sequences, in three codon positions**

<table>
<thead>
<tr>
<th>From\To</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>5.0463</td>
<td>9.9574</td>
<td>7.0297</td>
</tr>
<tr>
<td>T</td>
<td>5.9789</td>
<td>-</td>
<td>20.3332</td>
<td>0.6117</td>
</tr>
<tr>
<td>C</td>
<td>11.5959</td>
<td>19.9855</td>
<td>-</td>
<td>0.7138</td>
</tr>
<tr>
<td>G</td>
<td>16.1529</td>
<td>1.1863</td>
<td>1.4085</td>
<td>-</td>
</tr>
</tbody>
</table>


Deeply, in details, the results of evolutionary distance as shown in, separated the 17 organisms in several groups depending on the value of the minimum score. Firstly, the nearest animals to human are chimpanzee and gorilla (0.0913, 0.1157) respectively. Secondly, is the biggest group of animals led by water buffalo following by cattle 0.2681, also with bison and 0.1329 then Arabian camel 0.2689. The water buffalo could lead the major group of relationships by the highest scores to appear diagonally in the lower matrix between cattle, bison Arabian camel, Bactrian camel, horse, sheep, and goat. Moreover, results demonstrate the lowest divergence between an Arabian camel and a Bactrian camel, with cattle, and also shown among dog, rabbit, and sheep.

The most interesting results that relate with highest evolutionary distance in the pig with all 16 organisms in contrast whilst, chicken also had high divergence scores with all but significantly lower than the mtDNA of a pig. That is mean in spite of the highly morphological contrast between chicken and other organisms even it is not a mammal but shows a considerable similarity in mtDNA with all other animals including human.

The evident about molecular evolutionary by distance estimation, from, applying the Markov model of maximum likelihood method between pairs of sequence alignment results. It could be observed the evolutionary distance among all mammals’ organisms in spite of the variation in scores, the number of base substitutions per site from between sequences are shown for all three codon position and non-codon regions, also these scores put the organisms in groups by comparing the numbers between pairs of sequence, for instance, the human, chimpanzee and gorilla, likewise, the discovery of likelihood between bison and the water buffalo despite the historical and geographical distance between them. The highest distance ever was observed pig compared to the all other organisms.

**Amino acid substitution evolutionary distance**

The number of amino acid differences estimated with per sequence from other sequences is shown, in, which illustrates the results of involving 17 sequences of amino acid. The rate variation among sites was modeled with a gamma distribution score is equal to 1. Coding data translated assuming a vertebrate mitochondrial DNA genetic code table. All positions containing gaps and missing data were eliminated. Moreover, the coding data was translated assuming a Vertebrate Mitochondrial genetic code table. All positions containing gaps and missing data were eliminated. There was a total of 4091 positions in the final dataset. Similar to the previous results, the pig got the highest contrast distance against the others.

The results as demonstrated in, came in the same way with the substitution of nucleotide, evolutionary distance in the alignment of 17 sequences of mtDNA. Furthermore, the chicken as a bird is less divergence than pig, dog, cat, and mouse in compare with a human. Similar results are observed between chicken, rabbit, dog, cat, and mouse with a pig.

**Synonymous/non-synonymous codon substitution evolutionary distance**

The aim behind estimation of codon-based evolutionary divergence between sequences is to see the effect of substitutions in nucleotides on the codons of amino acids if that cause any changes in protein sequences that may cause a difference in annotation or function in the genome of mitochondrial DNA. The number of synonymous differences per sequence from shown, sequences involved in all positions containing gaps and missing data which eliminated. A total of 4091 positions in the final dataset.

The results in, demonstrate the effect of substitution levels on the frequencies of codon changes synonymously or synonymously. Firstly, the results between the human-chimpanzee pair were 445.33, human- gorilla pair 563.00 and chimpanzee-gorilla was 514.00. Secondly, the lowest score observed in Bactrian camel-horse pair 349.50. Finally, the huge change and divergence for a pig with all other organisms involved in this study Results in the respectively, explained and illustrated with results, the codons also changes show the harmony of the same rhythm with nucleotides and amino acids results, which provide same protein in another sequence. In fact, it was a shock if a study on mtDNA with 10 times more than nucleic DNA in substitution and could conserve itself through the time of evolution within animals. Since thousands of years mtDNA strict in the same function and annotation with keep changing its sequences. Actually, nowadays, this is a big foot step for human kind to explain or pretend understand the mechanism of mtDNA in evolution with all available sciences and Refutes all studies that talk about the evolutionary relationship between human and pig.

**Conclusion**

The current study opened a gate of huge question like the similarity between human with chimpanzee and gorilla, also how could be the divergence of the pig greater than chicken as a bird with other mammals even with a human. Moreover, this study generates a motivation to study the phylogenetic in deep and the de novo annotation looking for some confused answers that help to discover and understand more about mitochondrial DNA.

**Acknowledgements**

None.

**Conflict of interest**

The author declares no conflict of interest.

**References**


27. Xu CP, Lu YY, Yan JY, et al. [Molecular characteristics and its evolution of the complete genome of avian influenza H5N1 virus isolated in Zheji–


35. Xu CP, Lu YY, Yan JY, et al. [Molecular characteristics and its evolution of the complete genome of avian influenza H5N1 virus isolated in Zheji–


