



**T.R.**

**KAHRAMANMARAŞ SÜTÇÜ İMAM UNIVERSITY**  
**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE**

**COMPARATIVE BIOINFORMATICS ANALYSIS OF OMICS IN SOME ANIMALS**

**OMAR ESMAILL H. HAMAD**

**DOCTORATE THESIS**  
**DEPARTMENT OF ANIMAL SCIENCE**

**KAHRAMANMARAŞ 2016**

**T.R.**

**KAHRAMANMARAŞ SÜTÇÜ İMAM UNIVERSITY  
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE**

**COMPARATIVE BIOINFORMATICS ANALYSIS OF OMICS IN SOME ANIMALS**

**OMAR ESMAILL H. HAMAD**

**This thesis  
Prepared at the  
DEPARTMENT OF ANIMAL SCIENCE  
For the degree of  
*DOCTOR OF PHILOSOPHY***

**KAHRAMANMARAŞ 2016**

PhD thesis entitled “**COMPARATIVE BIOINFORMATICS ANALYSIS OF OMICS IN SOME ANIMALS**” and prepared by **OMAR ESMAILL H. HAMAD** who is student at Department of Animal Science, Graduate School of Natural and Applied Science Kahramanmaraş Sütçü İmam University, was certified by the majority jury members whose signatures are given below **17/10/2016**.

Prof. Dr. Emin OZKOSE (**Supervisor**) .....

Department of Animal Science

Kahramanmaraş Sütçü İmam University

Prof. Dr. M. Sait EKINCI (Member) .....

Department of Animal Science

Kahramanmaraş Sütçü İmam University

Assoc. Prof. Dr. Ismail AKYOL (Member) .....

Department of Agricultural Biotechnology

Kahramanmaraş Sütçü İmam University

Assoc. Prof. Dr. B. Devrim ÖZCAN (Member) .....

Department of Biology

Osmaniye Korkut Ata University

Asist. Prof. Dr. Bülent KAR (Member) .....

Organic Agricultural Program

Tunceli Vocational School

Munzur University

I confirm that the signatures above belong to mentioned academic members.

Assoc. Prof. Dr. Mustafa Şekkeli .....

Director of Graduate School of Natural and Applied Science

## THESIS NOTIFICATION

I declare and guarantee that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. Based on these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.



(Signature)

**(OMAR ESMAILL H. HAMAD)**

Note: Uses of the reports in this thesis, from original and other sources, tables, figures and photographs without citation, subject to the provisions of Law No. 5846 of Intellectual and Artistic Work

# BAZI HAYVANLARDA OMİKLERİN KARŞILAŞTIRMALI BİYOİNFORMATİK ANALİZİ

(DOKTORA TEZİ)

OMAR ESMAILL H. HAMAD

## ÖZET

Bu çalışma temel olarak insan mt-DNA gen dizi bilgileri ile 16 hayvana ait aynı bölgelerin biyoinformatik yaklaşımlarla nükleotid ve amino asit dizilerinin evolusyon süreci karşılaştırmaları ile sığırlarda *Brusella*'ya dayanıklılık olgusunun moleküler evolusyon ve immunomiks açılarından analiz edilmesini amaçlamıştır. Bu amaçla, bazı bilgisayar dilleri (PERL) ve yazılımlarla, matematiksel çözümler ve algoritmalar veri tabanlarında alınan gen dizi verileri üzerine uygulanmış ve sonuçlar dört ana başlık altında toplanmıştır. İlk olarak tamamlanmış insan mt-DNAsı ile diğer 16 hayvana ait aynı bölgelerin gen dizi verilerinin evolusyonel uzaklıkları maximum-likelihood estimation yöntemiyle analiz edilmiştir. Çalışma kapsamında kullanılan 17 organizmanın mt-DNA gendizi verilerinin yaklaşık 17000bp olduğu, 13 protein, 22 t-RNA ve 2 r-RNA sayılarının tüm organizmalar için sabit olduğu gözlemlenmiştir. İkinci bölümde insan ve 16 hayvanın tamamlanmış mt-DNA gen dizi verilerine göre maksimum olabilirlik metodu kullanılarak filogenetik soy ağaçlarının oluşturulması çalışılmıştır. Bu iki bölümde ayrıca insan, şempanze ve goril mt-DNA larına asit moleküler yapılarının benzerlikleri ile bu bölgelere ait gen dizi verilerine göre domuz ve tavukların diğerlerinden evolusyonel uzaklıkları belirlenmiştir. Sığırlarda patojen olan *brusella* nükleotid ve amino asit dizi analizleri tezin üçüncü bölümünde değerlendirilmiştir. Son bölümde farklı *Brucella* suşlarına ait her iki kromozom gen dizi verilerinin analizleri yapılmıştır. Sığırlarda *brusella* oluşumuyla ilgili olan makrofaj protein 1 (NRAMP) immunoinformatik kapsamında analizlere tabi tutulmuştur.

**Anahtar Kelimeler:** Maksimum olabilirlik, MT-DNA, Evolusyon, *Brusella*

Kahramanmaraş Sütçü İmam Üniversitesi  
Fen Bilimleri Enstitüsü  
Zotekni Anabilim Dalı, Ekim - 2016

Danışman: Prof. Dr. Emin OZKOSE  
Sayfa sayısı: 203

**COMPARATIVE BIOINFORMATICS ANALYSIS OF OMICS IN SOME ANIMALS  
(Ph.D. THESIS)**

**OMAR ESMAILL H. HAMAD  
ABSTRACT**

This study aimed fundamentally on bioinformatical approach to analysis the sequences of nucleotides and amino acids, through evolutionary comparison between human complete genomic of mt-DNA versus 16 animals, additionally, study the resistance of *Brucellosis* in cattle from the molecular evolution and immunomics points. For this aim, mathematical solutions and algorithms, within a particular computational language (PERL) and software, were applied on downloaded sequences from database resources and the results were categorized mainly in four chapters. Firstly, maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals were studied. The length of mt-DNA of 17 organisms were *ca* 17000 bp and 13 proteins, 22 t-RNAs and 2 r-RNAs were constant for all of them. Secondly, construction of phylogenetic tree of human's and 16 animals according to the complete sequence of mitochondrial DNAs using maximum likelihood method was studied. These two chapters concentrated also on the similarity between human with chimpanzee and gorilla, and the divergence of the pig and chicken from other mammals were also discussed. Nucleotide and amino acid sequences analysis in pathogen and host of brucellosis in cattle were evaluated in third chapter. Finally, immunoinformatics, antigenicity epitopes prediction in the solute carrier family 11 of the natural resistance associated macrophage protein 1 (NRAMP) related with *Brucellosis* in Cattle were studied using omics approaches.

**Key Words:** Maximum likelihood, Mt-DNA, Evolutionary Distance, *Brucellosis*,  
Antigenicity

Kahramanmaras Sütçü İmam University  
Graduate School of Natural and Applied Sciences  
Department of Animal Science, October - 2016

Supervisor: Prof. Dr. Emin OZKOSE

Page number: 203

## **ACKNOWLEDGEMENTS**

Firstly, I would prefer to categorical my sincere feeling to my advisor professor. Dr. Emin OZKOSE for the continual support of my Ph.D. study and related research, for his patience, motivation, and immense knowledge. His guidance helped me altogether the time of analysis and writing of this thesis. I couldn't have imaginary having a higher consultant and mentor for my Ph.D. study.

My sincere thanks also goes to Assoc. Prof. Dr. Ismail AKYOL and Prof. Dr. Mehmet Sait EKINCI, who provided me an opportunity to attend their lectures also to join their team as intern, and who gave access to the laboratory and research facilities. Without their precious support it would not be possible to conduct this research.

More importantly, I must express my very profound gratitude to my family: my parents and to my brothers and sisters for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

OMAR ESMAILL H. HAMAD

## TABLE OF CONTENTS

	<b>Page No.</b>
Pages of approval .....	ii
Student's page notification .....	iii
LIST OF ABBREVIATIONS.....	v
ABSTRACT .....	vii
ÖZET.....	viii
ACKNOWLEDGEMENTS.....	ix
Tables of Contents .....	x
<b>Chapter One</b> .....	1
1. Introduction.....	2
1.1. Bioinformatics in Molecular evolutionary and phylogenetic.....	3
1.1.1. Bacterial Genome.....	8
1.1.2. Mitochondrial DNA (mtDNA).....	9
1.1.3. Evolutionary Distance .....	11
1.2. Bioinformatics in Immunomics.....	13
1.2.1. Immunomics Database.....	17
1.2.2. Immunoinformatics Structure.....	18
1.2.3. Bovine's Immunogenomics.....	21
1.2.4. Bovine Genetics of Disease Resistance.....	28
<b>Chapter Two</b>	
2. Literature Review.....	32
2.1. Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals.....	33



2.2. Bovine's <i>Brucellosis</i> Antigenicity.....	36
2.3. The main aims s of the study.....	39
<b>Chapter Three</b>	
3. Material and Methods.....	41
3.1. The Sources of Database.....	41
3.2. Computational approach .....	45
3.3. Algorithm.....	51
3.4. Alignment of 17 sequences.....	52
3.5. Relative synonymous codon usage (RSCU).....	53
3.6. Estimating the evolutionary distances between genomic sequences.....	54
3.7. Markov models of nucleotide substitution and distance estimation .....	54
3.8. Estimate the probability of amino acid substitution.....	56
3.9. Disparity test of real substitution patterns Heterogeneity.....	56
3.10. Estimating the evolutionary distances between genomic sequences ...	56
3.11. Molecular Phylogenetic analysis by Maximum Likelihood method...	58
3.12. Predicting the antigenicity epitopes.....	58
<b>Chapter Four</b> .....	
4. The Results and Discussion.....	61
4.1. Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals.....	62
4.1.1. Computing the statistical quantities for sequence data .....	62
4.1.1.1. The nucleotide composition.....	62
4.1.1.2. The composition of amino acids.....	69
4.1.2. Estimation of the Codon Usage Bias.....	70

4.1.3. Probabilistic of nucleotide substitution with (ML).....	71
4.1.4. Probabilistic of the amino acid substitutions with (ML).....	72
4.1.5. Estimation of Transition/Transversion matrix by Maximum Composite Likelihood (MCL).....	78
4.1.6. Nucleotide Pair Frequencies from alignment of 17 sequences.....	79
4.1.7. Nucleotide evolutionary distance .....	80
4.1.8. Amino acid substitution evolutionary distance .....	83
4.1.9. Synonymous/non-synonymous codon substitution evolutionary distance.	85
4.2. Phylogenetic Tree Construction within Maximum Likelihood Method of complete genomes of mitochondrial DNA between Human's and 16 animals.....	88
4.2.1. Phylogenetic Tree Construction by Maximum likelihood method of 17 nucleotide sequences.....	89
4.2.2. Phylogenetic Tree Construction by Maximum likelihood method of 17 amino acid sequences.....	92
4.3. Nucleotide and amino acid sequences analysis in pathogen and host of <i>brucellosis</i> in cattle.....	95
4.3.1. Comparative view in DNA sequences between <i>Brucella abortus</i> and <i>Brucella melitensis</i> .....	96
4.3.2. Nucleotide Sequence analysis of SLC11A1 gene in cattle.....	99
4.3.3. Estimation of codon bias.....	102
4.3.4. The phylogenetic tree.....	104
4.4. Immunoinformatics, Antigenicity epitopes prediction in the solute carrier family 11 of the natural resistance associated macrophage protein 1 (NRAMP) related with <i>Brucellosis</i> in Cattle.....	105
4.4.1. The multiple sequence alignment.....	106
4.4.2. The Hydrophobic and Hydrophilic.....	107
4.4.3. The antigenic epitopes binding prediction.....	108

REFERENCES ..... 115

APPENDIX A..... 136

APPENDIX B ..... 166

APPENDIX C..... 177

SUMMARY..... 187

CURRICULUM VITAE..... 190



## THE LIST OF FIGURES

	<b>Page No.</b>
3.1. The flowchart explain the simple processing steps to apply by PERL.....	48
3.2. Shows the probability of substitution between nucleotides.....	55
4.1.1. The length graph of nucleotide bases and amino acids number in Human's mitochondrial DNA with other 16 vertebrates.....	63
4.1.2. The total nucleotide compositions of mitochondrial genome, occur among the Human's and other mammals.....	64
4.1.3. The Thymine (Uracil) frequencies within the protein coding regions of DNA at the 1 <sup>st</sup> , 2 <sup>nd</sup> and 3 <sup>rd</sup> position.....	65
4.1.4. The Cytosine frequencies within the protein coding regions of DNA at the 1 <sup>st</sup> , 2 <sup>nd</sup> and 3 <sup>rd</sup> position.....	66
4.1.5. The Adenine frequencies within the protein coding regions of DNA at the 1 <sup>st</sup> , 2 <sup>nd</sup> and 3 <sup>rd</sup> position.....	67
4.1.6. The Guanine frequencies within the protein coding regions of DNA at the 1 <sup>st</sup> , 2 <sup>nd</sup> and 3 <sup>rd</sup> position.....	68
4.1.7. The Frequencies (%) of amino acids occur in mitochondrial proteomes between Human and the average of all.....	69
4.2.1. The phylogenetic tree for the DNA by Maximum likelihood method with branch lengths.....	90
4.2.2. The phylogenetic tree for the DNA by Maximum likelihood method with ancestral states.....	91
4.2.3. Timetree phylogenetical analysis of DNA by Maximum Likelihood method.....	92
4.2.4. The phylogenetic tree for the amino acid sequences by Maximum likelihood method with branch lengths and ancestral states.....	93
4.2.5. Timetree phylogenetical analysis of amino acid sequences by Maximum Likelihood method.....	94

4.3.1. Dot plot matrix view of global alignment in chromosome I between <i>Brucella abortus</i> and <i>Brucella melitensis</i> .....	97
4.3.2. Dot plot matrix view of global alignment in chromosome II between <i>Brucella abortus</i> and <i>Brucella melitensis</i> .....	98
4.3.3. Estimation of codon bias for six sources of SLC11A1 gene in cattle.....	101
4.3.4. The four nucleotides frequencies codon in the first position .....	102
4.3.5. The four nucleotides frequencies codon in the second position .....	102
4.3.6. The four nucleotides frequencies codon in the third position.....	103
4.3.7. The phylogenetic tree for the nucleotide of SLC11A1 gene.....	104
4.4.1. The plot Presents high Similarity between the six proteins sequences.....	106
4.4.2. The amino acids frequencies distribution.....	108
4.4.3. The antigenicity plot for epitopes binding prediction.....	110
4.4.4. Predicting confidence score for the binding epitope fragments depending on protein's secondary structure.....	111

## THE LIST OF TABLES

	Page No.
3.1. List of organisms which involved in evolutionary study, with the information of database of complete genome, mtDNA.....	43
3.2. Database sources of SLC11A1 gene with the NRAMP proteins that produce.....	45
3.3. List of website names and the links were used for Genbank database and Bioinformatics tool services.....	46
3.4. List of the bioinformatics program names that used, and the original download links.....	48
4.1.1. The frequency account of the codons and the Relative Synonymous Codon Usage (RSCU), in all over the 17 aligned mammalian mitochondrial sequences.....	70
4.1.2. The lowest levels of probabilistic estimation in nucleotide substitution with (ML).....	71
4.1.3. The highest levels of probabilistic estimation in nucleotide substitution with (ML).....	71
4.1.4. Estimate the probabilistic of the amino acid substitutions with (ML), with the standard error.....	77
4.1.5. Maximum Likelihood Estimation of Transition/Transversion Bias.....	79
4.1.6. The transition/transversion calculated of 16 probable nucleotide pair frequencies by alignment of 17 sequences, in three codon positions.....	80
4.1.7. The substitution of nucleotide, evolutionary distance in alignment of 17 sequences of mtDNA. ....	82
4.1.8. The substitution of amino acids evolutionary distance in alignment of 17 sequences of mtDNA.....	84
4.1.9. Synonymous/non-synonymous codon substitution evolutionary distance in alignment of 17 sequences of mtDNA.....	86
4.3.1. Statistical nucleotide calculations of <i>Brucella abortus</i> and <i>Brucella melitensis</i> .	96

4.3.2. Statistical calculation comparative of the SLC11A1 gene that appear in six cattle stairs that resist to <i>brucellosis</i> .....	100
4.4.1. The antigenic epitopes binding prediction.....	109



## LIST OF ABBREVIATIONS

<b>BioPerl</b>	: The Bioperl Project is a world association of users & developers of open supply Perl tools for bioinformatics, genetics and bioscience
<b>CLASTAL W</b>	: Multiple Sequence Alignment tool
<b>CMD</b>	: Command prompt
<b>CPAN</b>	: The Comprehensive Perl Archive Network
<b>CRS</b>	: Cambridge Reference Sequence for human mitochondrial DNA
<b>EBI</b>	: European Institute for Bioinformatics
<b>EMBOSS</b>	: The Open Software Suite from European Molecular Biology
<b>EMMA</b>	: The European Mouse Mutant Archive
<b>GNU PSPP</b>	: GNU PSPP is a program for statistical analysis of sampled data. It is a Free replacement for the proprietary program SPSS
<b>INSDC</b>	: International Nucleotide Sequence Database Collaboration
<b>MAFT</b>	: MAFFT is a multiple sequence alignment program
<b>MHC</b>	: Major histocompatibility complex
<b>MtDNA</b>	: Mitochondrial DNA
<b>MUSCLE</b>	: Multiple Sequence Comparison by Log- Expectation
<b>NCBI</b>	: National Center for Biotechnology Information
<b>NRAMP</b>	: Solute carrier family 11 member 1
<b>PERL</b>	: Practical Extraction and Reporting Language
<b>PLOTCON</b>	: Plots the quality of conservation of a sequence alignment
<b>RSCU</b>	: Relative Synonymous Codon Usage
<b>SLC11A1</b>	: Solute carrier family 11 gene
<b>T-COFFEE</b>	: Multiple sequence alignment program
<b>XML</b>	: Extensible Markup Language





**CHAPTER ONE:**

**INTRODUCTION**

## 1. Introduction

Bioinformatics is known as “Is the area of studies which escapes easy definition inasmuch as of the fusion between science that attracts in the sciences of computational approach, and information technology to see and analyze genetic database and maths solutions” (Singh, 2014). The major merger of bioinformatics is between computational and biological sciences. This field has expanded to comprehend the data content and data flow in biological systems (SRINIVAS, 2005, Bingham et al., 2010, Ishida, 2004).

The earlier researches aimed to mapping individuals’ genomes and estimating variations to discover the population diversity. (Singh, 2014, Zvelebil and Baum, 2008). In fact, the genome of Human behind the importance of bioinformatics crucial by use applications of computer technology to the management of biological information thru gathering (Sharma, 2008), mining , examine and integrating the organic and genetically information with a purpose to be implemented to gene-based totally drug discovery and development (Assou et al., 2010).

Another essential point, the modality of bioinformatics created through fundamental motivated domains called the chain of reasoning, since the beginning emergence edge of the bioinformatics science. Firstly, the module of data observation questions (DOQ) which is the large scale molecular biology data accumulation considered as corpus of biological knowledge, also contain the biological simplification rules like the central dogma and the functional domain in genetics (Sadek, 2004). Secondly, the module of mathematical models that is known as a collection of mathematical and statistical methods for analyzing biological sequences which is majorly represented by the probability solutions in molecular evolutionary and phylogenetic by markovian models (Lorenz, 2010). Finally, the module of computational programing problems (Yang, 2010). , emphasis is placed on algorithms and their implementation in software (Isaev, 2006)

## **1.1. Bioinformatics in Molecular evolutionary and phylogenetic**

Evolution as it is known has few universal laws, but one of them is unassailable truth about each organism alive today had at least one parent, who in turn had either one or two folks depending on whether the lineage was vegetative or sexual, and so on extending back in time. Charles Darwin proposed a theory of evolution by means of natural selection in the 1858. Darwin's theory revolutionized not only biological thinking, but additionally politics, sociology, and moral philosophy. Besides the weakest part of Darwin's theory was its inability to account for the transfer of biological information from generation to generation (Rosenberg and Arp, 2009).

Evolutionary concepts underlie several of the ways used in bioinformatics, such as sequence alignments, identifying families of genes and proteins, and establishing homology between genes in completely different organisms. As well as evolutionary tree construction for example, the molecular phylogenetic. Itself a very massive field at intervals process biology. Since currently have several complete genomes, particularly in bacteria, will additionally begin to seem at biological process questions at the whole-genome level (Durbin et al., 1998).

The critical theme of DNA sequences of four sorts of nucleotide constructing called A, C, G, and T. it is the molecule that stores the genetic information of the cell. Which exists as a double helix composed of two precisely complementary strands. RNA is also composed of four nucleotide building blocks, but U is used instead of T (Sato et al., 2010, Richter et al., 2010).

Moreover, Evolution hypothesized that requires error in susceptible replication. consequently, the DNA replication is not perfect because errors are bound to occur, even if only rarely is called a mutation. Mutations may be single base substitutions, insertions, or deletions of one or more bases, or may involve large scale insertions, deletions, and rearrangements of the sequence. Mutations create new variant sequences in a population and

increase genetic diversity (Avice et al., 2010). The diversity could be quantified by measuring the fraction of polymorphic loci, or by measuring the average level of heterozygosity of loci.

Genetic diversity is reduced through natural selection. Which exclude lower fitness alleles from a inhabitants. Random driftage also reduces genetic diversity because some alleles can be lost by chance, even when there is no selection acting. The level of genetic diversity in a population is therefore determined by a balance between mutation (Ross et al., 2008). Gene sequences in a population are related to one another by descent from common ancestors. When the lines of descent of a gene from two individuals in a current population are traced back in time. Consequently, will have a common ancestor at some point in the past. The typical time back to the coalescence point will be of the order  $N$  generations, where  $N$  is the population size (Woodhams et al., 2015).

The probability of a neutral mutation is a beneficial mutation, with fitness, has an essentially larger probability of becoming fixed. Similarly, a deleterious mutation, with fitness, has a very small chance of fixation. However, both advantageous and deleterious mutations may be classed as nearly neutral. This means that random drift is more important than selection in determining their fate, and their probability of fixation is very close to that of a neutral mutation. Studies of human populations indicate the presence of many low-frequency mutant alleles. Information is available particularly for disease-linked genes. This suggests that most mutations are deleterious and will eventually be eliminated by selection (Bingham et al., 2010).

When compare the sequences between species, the differences appeared are the result of fixation of mutations in one lineage or the other. The most frequent types of change are conservative ones, for this reason synonymous substitutions occur more rapidly than non-synonymous ones, and amino acid changes occur more rapidly when the amino acids have similar properties (Moody, 2004). This suggests that the major mode of selection acting is

stabilizing, and that the changes that could notice those that were nearly neutral and were thus not selected against (Gibson and Baker, 2012).

Mutations create new variant sequences in a population and increase genetic diversity. This diversity can be quantified by measuring the fraction of polymorphic loci, or by measuring the average level of heterozygosity of loci. Genetic diversity is reduced by natural selection, which tends to eliminate lower fitness alleles from a population. Gene sequences in a population are related to one another by descent from common ancestors (Barbieri et al., 2014, Palanichamy et al., 2015). If the lines of descent of a gene from two individuals in a current population are traced back in time, they will coalesce, in this case will have a common ancestor at some point in the past. So the talking about typical time back to the coherence point will be of the order generations (Chauve et al., 2013).

Occurring a new mutation in a population, there will initially be only one copy. The number of copies of this mutation will increase and decrease due to the action of selection and drift. When new mutations will be eliminated from the population within a few generations for the reason of highly probability of chance when the copy number is small, even if they are selectively advantageous. Occasionally, a mutation will become fixed in the population, which it will spread through the population and reach high frequency (Hashizume et al., 2015, Gispert et al., 2015, Sevini et al., 2014).

Random aberration is more important than selection in locate their fate, and probability of solidification which is closed to neutral mutation. Researches in human populations mark the presence of many low frequency mutant alleles. The data is available especially for disease-linked genes (Ling et al., 2014). This suggests that most mutations are deleterious and will eventually be eliminated by selection. When compare sequences between species, the

differences has seen are the result of fixation of mutations in one lineage or the other (McMahon and LaFramboise, 2014).

Frequent types of change are conservative, with attention to synonymous substitutions occur more rapidly than nonsynonymous. Amino acid changes occur more rapidly when the amino acids have similar properties (Molnar et al., 2014, Hagen et al., 2013). This suggests that the major mode of selection acting is confirming (Bahitham et al., 2014).

The different ways in constructing phylogenetic trees, sometimes cause trees appearance to be different can have some of their groups interchanged, so that it becomes clear that they are actually the same (Röck et al., 2013). It is important to check even the branch lengths are drawn to scale and the tree is intended to be rooted or unrooted. A given unrooted tree can be rooted on any of its branches. So, trees provide different evolutionary interpretation (Sridhar et al., 2007).

Distance matrix approach in phylogenetic generated through accumulation of a matrix of pairwise distances between the sequences. Distances could be calculated using many diverse models for evolution of sequence. Values of the distance for pair of sequences depend on the model used (Fournier et al., 2012).

UPGMA, a hierarchical clustering method that assumes a constant rate of evolution in all the species, which is considered as a simplest method. The assumption is cause to unreliable results (Decottignies, 2005). The neighbor joining method use clustering method that works well if the input data are close to additive, as an illustration if the pairwise distances between the species can be expressed as the sum of the lengths of the branches on the tree that connect the pairs of species. NJ is useful for large data sets and for initial examination with new sequences (Bernt et al., 2013, Steele et al., 2012).

The method of maximum likelihood common for particular criterion, selection of the optimal tree. The likelihood also known observing a given set of sequences can be calculated on any proposed tree. This is a function of the tree topology, the branch lengths, and the values of the parameters that define the evolutionary model used (Ye et al., 2014). The principle is to choose the tree for which the likelihood is maximized. Both likelihood and parsimony methods require a tree-search program to produce candidate trees. Modern maximum likelihood used effectively with realistic data sets because allow fitting of model parameters to the sequence data being used (Xiong et al., 2014, Kumar et al., 2011). Also, allow tests to be made to identify between models or between alternative trees. Moreover, have low sensitive to problems of long branch attraction than some simpler methods. (Raharimalala et al., 2012).

A large number of at least a 100 randomized sequence data sets are generated, where every column of data is a copy of an indiscriminately selected column in the real sequence alignment. The repeated data set gives slightly different trees. The clade of the tree from the original data set provide the bootstrap percentage, when percentage of time the clade appears in the set of trees from the randomized data. High bootstrap percentages (>70%) indicate statistical support for the presence of the clade. Namely, bayesian phylogenetic are methods from the recent development of likelihood methods (Oliva et al., 1998).

Bayesian methods also estimate the likelihood of observing a given sequence set on a given tree, but instead of searching for a single tree that optimizes the likelihood, it takes an average over possible trees, weighting them according to their likelihood. In practice, a simulation technique known as Markov Chain Monte Carlo which used for generating sample of possible trees, the probability of proportional to its likelihood. The background of possibility of formation of given clades that calculated by averaging the properties of the trees (Greene and Hill, 2010).

### **1.1.1. Bacterial Genome**

Nowadays, there are abundant of available complete bacterial genomes, with comparisons across a large number of species. Bacterial genomes range in size from around half a million to over seven million bases (Zvelebil and Baum, 2008). The number of genes per genome varies almost in proportion to the length, with an average of just over 1000 bases between gene start points (Brown et al., 2016).

There is relatively little non-coding DNA between genes in bacteria (Moody, 2004). This suggests that selection for efficiency of genome replication is strong enough to prevent the widespread accumulation of repetitive sequences and transposable elements, in contrast to the situation in many eukaryotes (Bergeron, 2003, Nei and Kumar, 2000).

The genome size and content vary quite rapidly between related bacterial species. For instance, smallest genomes are in bacteria which considered as parasites or symbionts inside other cells (Yoshida et al., 2016). These species manage with a greatly reduced set of genes, perhaps the reason behind absorbed useful chemicals from host cell which synthesized from free living relatives (Brown et al., 2016).

There are several independent groups of parasitic bacteria where dramatic reduction of genome size has occurred in this way (Yang, 2010). This suggests that there is a tendency for genes that are no longer necessary for an organism to be deleted from genomes in a relatively short period. In cases where more than one strain of a bacterial species has been completely sequenced, there can be a surprising degree of variability in gene content between the genomes (Rosenberg and Arp, 2009).

### **1.1.2. Mitochondrial DNA (mtDNA)**

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 (Zhu et al., 2015b).

Mitochondria are semiautonomous organelles that contain their own DNA and are responsible for the bulk of ATP synthesis in the eukaryotic cell. Mitochondrial functions are linked to the aging process, apoptosis, sensitivities to anti-HIV drugs, and, possibly, some cancers. Mitochondria were first visualized as discrete organelles by light microscopy in 1840 (Achilli et al., 2008).

However, isolation of intact mitochondria had to wait until zonal centrifugation methods were developed in 1948. In the early 1960s it was determined that these cytoplasmic organelles contain their own DNA (Hassanin et al., 2010, Hiendleder et al., 1998). The DNA sequence of human mitochondrial DNA (mtDNA) was determined in 1981 and gene products were assigned by 1985, making it the first component of the human genome to be fully sequenced. Human mtDNA is a double-stranded 16,569 bp circular genome coding for 13 polypeptides required for oxidative phosphorylation and 22 tRNA and 2 ribosomal RNAs responsible for its synthesis. One noncoding segment, the displacement loop, contains several cis-acting elements required for initiation of transcription and replication (Ji et al., 2009, Achilli et al., 2008).

Mitochondrial DNA makes up only 1% of total cellular DNA, and the mtDNA polymerase activity accounts for less than 1% of the total DNA polymerase activity in the cell. Individual cells have up to 10,000 discrete mitochondrial genomes distributed within 10–1000 organelles. Heteroplasmy, mitochondrial genetic diversity within a single cell, can result from point mutations or deletions in mtDNA and usually increases exponentially with age (Bandelt et al., 2006).

Defects in mitochondrial function produce a wide range of human diseases and can be caused by mutations within the mtDNA. The first mutation discovered in mtDNA to be the cause of a mitochondrial disease, Leber's hereditary optic neuropathy, was first identified by Douglas Wallace and coworkers in 1988 (Durbin et al., 1998). Since that time several hundred point and deletion mutations in mtDNA have been described as the causes of mitochondrial disorders (St. John, 2013).

The Mitochondrial and Metabolic Disease Center reports that more than 1 in 4000 children born in the United States each year will develop a mitochondrial disease by age 10 with a mortality rate from 10 to 50%. Over 50 million people in the United States suffer from chronic degenerative disorders, and defects in mitochondrial function have been linked to several of the most common diseases of aging. Mutations within mtDNA and nuclear genes involved in the maintenance of mtDNA are the main cause of these mitochondrial diseases (Bolander, 2004, Singh, 1998).

The foremost attention-grabbing issue, mtDNA have the ability to adapt with each individual cell by modify the sequence by slightly the initiation and termination pints, likewise, the start direction of transcription  $5' \Rightarrow 3'$  or  $3' \Rightarrow 5'$  (Wilson and Hunt, 2002). 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 necrobiosis. By means of retrograde signaling, mitochondria 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 (Cízková et al., 2008).

### **1.1.3. Evolutionary Distance**

Perhaps, the estimation of the distance between two sequences is the simplest phylogenetic analysis, because calculation of pairwise distances as a premier step in distance matrix methods used for phylogeny reconstruction. Cluster algorithms used to convert a distance matrix into a phylogenetic tree (Lachowicz et al., 2009). The models of Markov process for estimating distance in nucleotide substitution. form the basis of likelihood and Bayesian analysis of multiple sequences on a phylogeny (Yang, 2014).

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 (Zhang et al., 2015). 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 (van Gisbergen et al., 2015, Szecsenyi-Nagy et al., 2015). Besides this basic assumption, it is often placed further constraints on substitution rates between nucleotides, leading to variable models of nucleotide substitution (Nielsen, 2005).

The first application of a maximum likelihood method to tree construction was made by Cavalli-Sforza and Edwards (1967) whom estimated the gene sequence frequency data. Following, 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 (Li and Graur, 1991). 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 (Kumar et al., 2011, Blanquart and Gascuel, 2011).

The methods for discovering 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 (Rosset et al., 2008, Marjoram et al., 2003).

To outline some main points, as a historical observation from previous researches have been documented about estimate the mitochondrial DNA evolutionary distance, within maximum likelihood method among animals within various visions and scoring parameters. First time started with the pronouncement by Irwin, Kocher, and Wilson (1991) when studied on evolution estimation cytochrome-b gene in mammals that acquired 17 complete gene sequences representing of mammals (ungulates) and dolphins (cetaceans) (Yang, 2006, SRINIVAS, 2005).

## **1.2. Bioinformatics in Immunomics**

Like many words, the term immunomics equates to different ideas contingent on context. For a brief span, immunomics meant the study of the Immunome, of which there were, in turn, several different definitions (Flower et al., 2010). Largely defunct meaning rendered the Immunome as the set of antigenic peptides or immunogenic proteins within a single

microorganism be that virus, bacteria, fungus, or parasite or microbial population, or antigenic or allergenic proteins and peptides derived from the environment as a whole, containing also proteins from eukaryotic sources (Lefranc, 2014a). However, times have changed and the meaning of immunomics has also changed. Other newer definitions of the Immunome have come to focus on the plethora of immunological receptors and accessory molecules that comprise the host immune arsenal (Garcia-Angulo et al., 2014).

Today, immunomics or immunogenomics is now most often used as a synonym for high-throughput genome-based immunology (Lefranc, 2014b). This is the study of aspects of the immune system using high-throughput techniques within a conceptual landscape borne of both clinical and biophysical thinking. Within an immunogenomic or immunomics framework,

How the phenotypic behavior of the immune system emerges from the interaction of its genome-encoded components should be of paramount interest to all involved in its investigation. Saying this is one thing; but actually achieving it is quite another (Flower et al., 2010).

Immunomics can stand as a synonym for system biology techniques applied to the study of Immunology. For many scientists, immunology is the pre-eminent example of systems behavior in biology (Flower, 2013).

Bayesian statistics can provide an insightful route to manifesting data which is both rigorous and of true utility. Clearly, the genome, the epitome, the proteome, the glycome, the metabolome, and all the rest of the omics that have come to dominate in current perceptions are of direct relevance to burgeoning understanding of immunology and immunological processes (Falus, 2009a). Genes, proteins, carbohydrates, lipids, glycoproteins and lipoproteins, together with the peptides and small molecules too, all take part in range of interactions that manifest themselves as an immune response to pathogen challenge. It is clear,

that a pivotal turning point has been reached; several key technologies have achieved long-awaited maturity, most notably predictive immunoinformatics methods and post-genomic strategies (Schönbach et al., 2008).

Of course, the whole of biology indeed the whole of the physical universe behaves as a system, and exhibits characteristic systems behavior. Since the immune system is innately hierarchical and exhibits confounding complexity at each tier of this cascading or branching hierarchy, as a system it can be said to exhibit emergent behavior at all levels (Viroj, 2008).

Since the discovery of antibodies and MHC restriction, humoral immunity and cellular immunologists have sought to understand the nature of these bio-macromolecular interactions, seeking to analyse them in the most fundamental way (Carbo et al., 2014).

Systems biology seeks to analyze higher levels of the immune system with the same degree of rigour, by both analyzing the system as it exhibits itself at these individual levels and by integrating detailed, low-level, small-scale molecular or mesoscopic information and more overtly macroscopic measurements with more intrinsically qualitative anatomical, functional, and phenotypic data. Thus, Systems Biology or, in this context, Systems Immunomics can be said to function at various length scales from the atomic to the macroscopic (Goodswen et al., 2013). Biological systems, of which immunological systems are an example, are seldom binary entities on the whole organism scale, any more than their cascading sub-systems be they organ, tissue, or cellular are binary entities at subsidiary levels. They operate stochastically, subject to random fluctuations and exhibit clear non-linear behavior (Tomar and De, 2010).

Immunology only truly manifests itself at the level of the whole organism, but at every intermediate level down to that of the molecule, significant and often unexpected emergent behavior within experimental systems is observed (Lane et al., 2010).

Many tools exist within systems biology and some tools are based on capitalizing on the latent power of simulation, be that simulations of abstract theoretical or mathematical models or molecular simulations of precise descriptions of molecular system. Other tools are analytical tools that can be used together to effect the synthesis of competing thesis and antithesis through the integration of measured data (Ehrenmann et al., 2010, Ansari et al., 2010).

The simplest types of systems model include network maps, which reticulate pathway components producing complex cellular representations akin to circuit diagrams, and so-called logical models, which describe immunological process in terms of sets of relatively simple rules. There are many other more complex and mathematically demanding models available; these include correlation models and kinetic modelling (Lefranc et al., 2009, Lefranc, 2009).

Multiple linear regression or Partial Least Squares or neural networks, or, any of a hundred other data mining techniques, can be used to identify commonalities of exchange or cooperation within or between the measured outputs of different signaling or regulatory pathways. Kinetic models, on the other hand, try to picture the spatiotemporal behavior of each and every individual component within the system (Viroj, 2008). They are the zenith and apotheosis of complexity with the currently available approaches within systems biology. It is also possible to combine these different kinds of model (Lefranc et al., 2008, Liu et al., 2006, Korber et al., 2006). This is particularly useful when one wishes to fuse data of different granularity. It is possible, for example, to build a detailed kinetic model for part of a pathway and then to fill in the lacuna within the available data by modelling the rest using much simpler Boolean models. The word bioinformatics, has formed part of the scientific lingua franca since the early 1990s; yet a simple and straightforward, and comprehensive and inclusive definition remains strangely elusive (Tung and Ho, 2007, Kurochkin et al., 2007). A particularly succinct

epitome of the discipline Bioinformatics is the application of informatics methods to biological macromolecules.

Bioinformatics has greatly expanded over the years, allowing for both new sub-disciplines to emerge within it and for bioinformatics to merge with other disciplines producing new and exciting hybrids. Sub-disciplines have tended to focus on areas of applications, such as neuroinformatics, transcriptomics, or proteomics, while hybrids have included text mining or statistical genetics (Moise et al., 2014). Immunoinformatics is another important sub-discipline. Which deals specifically with the unique problems of the immune system. Practically researchers look for key questions in the still highly experimental immunology (Jorgensen et al., 2014, Korber et al., 2006).

bioinformatics is constantly developing to include new frontier of application. However, it is concerns with medical, genomic, and biological information and supports both basic and clinical research (Giudicelli and Lefranc, 2012).

Bioinformatics is as much a fundamental technique as a branch of information. Operates at the level of protein and nucleic acid sequences, their structures and functions, through involving data from microarray experiments. Databases are main gate for research in bioinformatics and immunoinformatics (Sun et al., 2011).

### **1.2.1. Immunomics Database**

The algorithms tools for database mainly used to search, analyze, and interrogate biological information. Data handling in bioinformatics, mainly through the annotation of macromolecular sequence and structure databases (Tomar and De, 2010).

The application and development of databases within the immunoinformatics domain. Chapters IPD – The Immunopolymorphism Database and The IMGT/HLA Database, by Professor Marsh and co-workers, describe two world-leading resources: IPD and IMGT/HLA.



Ontology Development for the Immune Epitope Database by Bjorn Peters and colleagues neatly summarizes on-going development of the IEDB database (Singh, 2014, Flower et al., 2010, Schönbach et al., 2008).

Databases and Web Based Tools for inherent immunity extends and completes this strand by describing a variety of databases aimed at the archiving of data relating to the innate immune system (Falus, 2009b).

Attempting to address all of these possibilities in a systematic and effective manner using experiment only would be prohibitive to the point of intractability, in terms of time, resource, and that most precious quantity of all: human labor (Lane et al., 2010, Ehrenmann et al., 2010). The only practical and practicable solution is the deployment of bioinformatics. Which focuses on analyzing molecular sequence and structure data and molecular phylogenies. Also the analysis of post genomic data came from genomics, transcriptomic, and proteomics. seeking the solutions to two hypothesis challenges. First, the predicting the function from a sequence performed global homology searches, even more from the motif databases searches and the formation of multiple sequence alignments (Flower, 2002).

Discovery of Conserved Epitopes through Sequence Variability Analyses. The addresses the prediction of conserved epitopes within an immunomics and immunoinformatics context. Defining the Elusive Molecular Self picks up on this with its analysis of the molecular nature of the self-immune. Secondly, structure prediction from Sequence that attempted through applying the secondary structure prediction (Srivastava et al., 2014).

### **1.2.2. Immunoinformatics Structure**

Understanding MHC-Peptide-TR Binding provides a lucent and definitive description of the use of 3-dimensional structural data, as derived from experiment and computation, within the province of immunoinformatics investigation. As yet, the full power of 3-dimensional data

has not been realized (Falus, 2009a). Structure-based computation based on dynamic simulation and hypothesis-guided modelling has so much to reveal, but as yet the potential is not matched by available computing resources. The next few years should see this approach beginning to bear fruit as more and more studies are undertaken (Wise et al., 1998).

In reality, predictions of function depend on identifying similarity between sequences or between structures. High similarity give intrinsically reliable and useful inferences drawn. In contrast, similarity falls away in conclusions, become increasingly uncertain and potentially misleading (Singh, 2014). Thus, provenance is everything; and provenance and annotation. Bioinformatics still concerns handling and analyzing data. The classification into coherent groups on the strict annotation of macromolecular sequence and structure databases (Flower, 2002).

T-Cell Epitope Annotations addresses the integration of data sources for the rigorous and reliable annotation of T cell epitopes. Vaccines were for so long a moribund market, yet they have recently re-emerged as the most hopeful growth area for the Pharmaceutical Industry (Schönbach et al., 2008). Public health requirements safeguard vaccine supply of vaccines and in the absence of competition – Influenza apart, only two to three manufacturers target each vaccine-preventable disease this has led to a recent increase in unit price for specialty vaccines (Korber et al., 2006).

Pediatrics vaccines currently hold sway over the global market for vaccines, yet adult vaccines will help drive future growth. The cancer vaccine market, led by vaccines targeting cervical cancer, is the most lucrative area of vaccine development at 2012, cancer vaccines will account for around 30% of all vaccine revenues. As discussed in Computational Vaccinology, Immunomics and Systems Immunomics, at least in their informatics and computational guise,

have much to offer vaccine design and discovery and the still emergent science of Vaccinology (Moise and De Groot, 2006, Liu et al., 2006).

Returning to the first theme, the term vaccinology is said by many to have been coined by Jonas Salk to distinguish the systematic scientific study of vaccines – and thus how to develop and discover them from the practice of vaccination as a medical art (Flower, 2013). In recent times, another term, immune-vaccinology has been adopted by some to further differentiate the study of vaccine discovery and development based on a sound understanding of immunology, if such a thing exists, from what many might consider the highly empirical, microbiology-based science of vaccinology, as practiced in year gone by Davies and Flower give a concise examination of how immunoinformatics has and can impact upon the pursuance of a rational yet systematic approach to vaccine discovery (Kurochkin et al., 2007, Deluca and Blasczyk, 2007).

Despite the need for more accurate prediction algorithms, able to cover ever more MHC alleles in ever more species, the lack of persuasive evaluations of known methods continues to hamper and stymie uptake of this technology (Lefranc et al., 2008). In order that Immunoinformatics approaches might one day become universally used by experimental immunologists, methods should be tested over a wide range of alleles, species, and sequence-distinct peptides, with their accuracy reaching a high statistical significance (Flower et al., 2010). This will be greatly facilitated by adoption of a cyclically and progressive process of using and refining models and experiments (Feldhahn et al., 2009).

The effective implementation of immunoinformatics strategies within Immunomics and Systems Immunomics will deliver an unprecedented dividend of great if unquantifiable magnitude. Methods that accurately predict individual components of the immune response or allow us to model the behavior of the whole system or part thereof will be the most vital of

tools for tomorrow's immunologists (Falus, 2009b). Immunoinformatics prediction, within the broader system immunomics context, remains a scientific problem, being both challenging, and thus exciting, and of true practical value. Moreover, the proper realization of Systems Immunomics requires not only a deep appreciation of immunological mechanisms but also requires one to integrate many other disciplines, both experimental and theoretical (Lefranc et al., 2009, Lefranc, 2009).

The basic strategy of immunological investigation and it needs the confidence of experimentalists to commit laboratory work on this basis (Lefranc et al., 2015). The context of immunomics and systems immunomics, the synergy of experimental and informatics-based disciplines will enhance significantly the ability to understand and manipulate immunology process, leading to the augmented discovery of new laboratory reagents and diagnostics, in addition to new biomarkers and candidate vaccines (Ansari et al., 2010).

### **1.2.3. Bovine's Immunogenomics**

The immune system of jawed vertebrates evolved to provide innate and adaptive immunity against a diverse array of potentially harmful antigens. The adaptive immune effector cells are B and T lymphocytes also known as B cells and T cells, while innate system cells include those of the myeloid lineage (monocytes, macrophages, eosinophils, basophils, mast cells, neutrophils and dendritic cells) as well as primitive lymphoid cells known as natural killer cells (Seelye et al., 2016).

The cells of the innate immune system not only play their own direct role in immunity, for example, killing infectious microbes following phagocytosis, but in the case of macrophages and dendritic cells function as accessory cells for T cells by presenting antigenic peptides on major histocompatibility complex MHC, molecules and producing cytokines that

direct T cell functional responses (Ruan et al., 2016). The immune response has been historically broken into two aspects known as humoral and cell-mediated immunity. While B cells produce antibodies, which are mediators of humoral immunity, T cells can promote the B cell response through their production of specific soluble molecules known as cytokines, thereby facilitating humoral immunity (Obara et al., 2016).

Alternatively, T cells mediate cellular immunity by killing infected host cells and by their production of cytokines that activate macrophages to more effectively kill phagocytosed infectious organisms or inhibit viral replication. The host's immune system must differentiate between self and oneself antigens but still recognize a diverse array of potentially harmful antigens, estimated to be between  $10^8$  and  $10^{11}$  (Mishra et al., 2016, Schwartz and Hammond, 2015).

Significant advances have been made in describing the genetics of the bovine immune system receptors and MHC molecules that are involved in presenting peptides to T cells to engage their so-called T cell receptor (TCR) and will be reviewed here. It described in detail the genes that code for the T and B cell antigen-specific receptors (TCR and B cell receptor (BCR)) and the immunoglobulins (antibodies) that are secreted by B cells and which mirror the BCR of the secreting cell (Pandya et al., 2015).

These receptors and antibodies are formed by somatic gene rearrangements. In addition describing germline encoded multigene receptor families that are expressed by both innate and adaptive immune system cells and which interact with pathogen-associated molecular patterns (PAMP), host cell-derived damage associated molecule patterns (DAMP), as well as classical and non-classical MHC molecules (Konradsen et al., 2015, Thompson-Crispi et al., 2014).

Immunoglobulins are composed of two identical heavy H and two identical light L polypeptide chains in cattle. The heavy chains are known as  $\mu$ , d, g, e and a, while the light

chains are known as k or l, so-named for the genes that code for a portion of the chains referred to as the constant domains IGHC for the heavy chain; IGKC and IGLC for the k and l light chain, respectively. The standard IMGT nomenclature for immunoglobulin heavy and light chain genes has been used and explained and takes into consideration the historical gene designations widely cited in the literature (Kasahara and Yoshida, 2012).

Immunoglobulins are known as antibodies when secreted by B cells or as the BCR when bound to the membranes of B cells. Antibodies are the main effector molecules produced by B cells, while the BCR allows the cell to interact with antigens thereby becoming activated (Hammond et al., 2012). The immunoglobulin chains have terms for specific parts of the molecule: the part responsible for interacting with antigens is known as the 'variable domain' and occurs in both the heavy and light chains. The other parts of these chains are the constant domains and some of those in the heavy chains convey the functional differences among antibodies (Kataria et al., 2011).

The part of the antibody composed of heavy chain constant domains that convey function is known as the fragment-crystallizable or Fc piece. The variety of functions mediated by it include the ability of the antibody to interact with specific receptors on other cells known as Fc receptors or to activate an enzyme system in blood and interstitial fluids known as the complement system. Thus, immunoglobulins are divided into various classes previously termed isotypes according to their heavy constant regions as follows: IgM, IgD, IgG, IgA and IgE. For example, IgM means it is an immunoglobulin with a  $\mu$ -heavy chain encoded by the IGHM gene (Yassin et al., 2016, Tipu, 2016).

Immunoglobulins are coded for by a set of germline genes (previously referred to as 'gene segments' or exons because all parts are needed to create a functional transcript) that are 'rearranged' in a variety of possible combinations during lymphocyte development to give rise

to the two polypeptide chains which is known as heavy and light chains. Those genes are the so-called variable V or IGHV, diversity D or IGHD and joining J or IGHJ genes, with one to several hundred occurring in each set (Zheng et al., 2015).

The heavy chains are coded for by IGHV-IGHD-IGHJIGHC genes, where the variable domain, encoded by V-D-J gene recombination, is potentially more variable and complex (Lefranc et al., 2015). The k light chains are formed from rearrangement of IGKV-IGKJ and the constant C or IGKC gene, while l light chains are formed from rearrangement of IGLV-IGLJ and the constant C or IGLC gene. A gene is chosen from each group in a variety of combinations such that lymphocytes have the ability to recognize a nearly unlimited array of antigens especially when the additional mechanisms that contribute to diversity beyond the V-(D)-J-C recombination are considered (Lefranc et al., 2015). When the BCR engages the appropriate antigen, the B cell is activated and undergoes two genetic processes known as somatic hyper mutation, which affects the variable domains of both chains, and class switch recombination, which affects the constant domains of the heavy chain. These processes are mediated by activation-induced deaminase (Backert and Kohlbacher, 2015a).

Somatic hyper-mutations mean that additional random changes occur in the coding sequence for the variable domain, concentrated in regions known as complementarity determining regions. Some of these changes in coding sequence will make the interaction with the antigen stronger and, as a result, those B cells will be selected and stimulated to replicate and survive more efficiently (Aouinti et al., 2015).

This phenomenon is known as affinity maturation during the development of the antibody response. In contrast, class switching affects the constant domains of the heavy chain and means that the genes that code for those regions of the protein are changed or 'switched' leaving the variable region intact but making the class of antibody different. For example, IgM

is an immunoglobulin with a  $\mu$  heavy chain since the constant domains are coded for by the IGHM gene, but its variable region genes could become associated with genes that code for a different constant region, e.g. IGHA gene making it now an IgA class of antibody (Lohia and Baranwal, 2014).

The IgM-bearing B cells have been detected in the bovine fetus as early as 59 days into gestation. However, V-D-J and V-J recombination were observed in splenic B cells at 125 days of gestation and serum immunoglobulin was detectable in a 145-day-old fetus. At this developmental stage, some splenic B-cells may express V-D-J recombination alone, while others may secrete  $\lambda$  light chain only because of non-productive V-D-J recombination (Lefranc, 2014a).

In cattle, perinatal immunoglobulin diversification occurs in the ileal Peyer's patches, suggesting that the ileal Peyer's patches serve as the primary lymphoid organ in ruminants (Carvalho et al., 2011). The lymphoid follicles of ileal Peyer's patches consist mostly of IgM-bearing B cells that develop and expand oligo clonally, similar to bursal follicles in chicken. Nevertheless, *IGLVIGLJ* recombination-associated  $\lambda$  light chain diversification has been noted in bovine fetal spleen prior to the establishment of a diverse repertoire in the ileum (Flower, 2013).

B lymphopoiesis (as shown by the presence of so-called pre-B like cells that had intracellular  $\mu$  heavy chains) also has been observed in bovine fetal bone marrow and lymph node in parallel to ileal Peyer's patches. Thus, ileal Peyer's patches may not be the sole primary lymphoid organ in cattle (Herzig et al., 2006).

In general, variations with regard to B cell development across species seem to exemplify an outcome of divergent evolution. There are some known differences between immunoglobulin gene usages in fetal development versus the adult. Two *IGHV* genes



(*gI.110.20* and *BF2B5*) are preferentially used in the fetal *V-D-J* recombination (Schubert et al., 2016). In contrast to J-proximal conserved *IGHDQ52* gene, *IGHD7* and *IGHD5* genes are favorably expressed in both fetal and adult B cells. The bovine *IGHJ1* gene (*IGHJpB7S2*) expression is also predominant in both fetal and adult *V-D-J* recombination (Saini et al., 1997). Analysis of somatic hyper mutations in the CDRs revealed that transition nucleotide substitutions predominate over transversion. Further, somatic hyper mutations result in higher diversification in the third framework region of IgG as compared to IgM antibodies in cattle (Shi et al., 2015, Schubert et al., 2015).

The mechanisms of antibody diversification in species where immunoglobulins can be transferred across the placenta and into colostrum as well such as mice and humans' significant germline *IGHV*, *IGHD* and *IGHJ* gene divergence sequence and combinatorial diversity exists. In contrast, the primary antibody repertoire of cattle is composed of limited combinatorial diversity ( $1.5 \times 10^4$ ) because of restricted germline sequence divergence both at IGH and IGK or IGL loci. For example, while in mice and humans there are over 200 *IGHV* genes for the heavy chain, cattle have only 36 of which 10 are functional (Oany et al., 2015).

Thus, several other mechanisms compensate for this restricted combinatorial diversity in cattle including somatic hyper mutations, insertion of preserved short nucleotide sequences (CSNS) specifically at *V-D* junctions and extensive junctional flexibility in *V-D-J* recombination involving deletions and templated or untemplated nucleotide additions at the junctions. While evidence not exists for gene conversion in the heavy chain, it has been suggested to occur at the light chain variable region. Activation induced cytidine deaminase (AID), an enzyme crucial to somatic hyper-mutations, has been characterized in cattle. AID gene, located on chromosome 5, is expressed in neonatal and adult lymphoid tissue of cattle (Backert and Kohlbacher, 2015b, Aouinti et al., 2015).

The biased ‘hot spot’ triplets in the CDRs of bovine *V-D-J* recombination predispose them to somatic hyper-mutations similar to other species. Somatic hyper-mutations are also involved in diversifying the *V-J* recombination encoding *I*-light chains. Cattle have been shown to use somatic hyper-mutations without exposure to exogenous antigen to diversify the developing antibody repertoire during B cell ontogeny, with somatic hyper-mutations evident in the heavy chain CDR1 and CDR2 of 125-day-old fetus (Schönbach et al., 2008).

Finally, extensive size heterogeneity (3 to 66 codons) in the heavy chain CDR3 together with disulphide bridging between multiple even numbered cysteines leads to significant configurational diversity of this region, which constitutes the antigen-combining site however, cattle antibodies can express exceptionally long heavy chain CDR3s (>50 amino acids) with multiple even numbered cysteine residues, both in fetal and adult B cells (Saini *et al.*, 1999; (Suzuki et al., 2014, Jonsson et al., 2014, Seroussi et al., 2013).

The exceptionally long CDR3H occurs in 8–10% of circulating B cells and, while initially observed in IgM, it occurs in IgG, IgA and IgE classes of immunoglobulins. Recent crystallization of bovine antibodies with exceptionally long heavy chain CDR3 has revealed a unique ‘stalk and knob’ structure where configurational diversity is generated via creation of mini-domains through intra-CDR3H disulphide bridges between the cysteine amino acids (Yang et al., 2011, Osterhoff, 2010).

Such a structural diversity via mini-domains in the antigen-binding site is not yet known to exist in other species. Both fetal and adult antibodies with exceptionally long CDR3H originate from unique recombination of the germline *IGHV-gl.110.20*, longest *IGHD2* and *IGHJ1-pB7S2* genes. An insertion of 13–18 nucleotide long CSNS of unknown origin in adult *V-D-J* recombination, which has a disproportionate number of adenines, specifically at the *V-D* junction increases the CDR3 size to ~61 codons following encounter with antigen in the

periphery, providing a novel mechanism of antibody diversification (Weber et al., 2006, Norimine et al., 2006, Larson et al., 2006).

Such insertions at the *V-D* junction that contribute to the stalk structure of the antigen combining site are absent in *V-D-J* recombination in fetal B cells. Thus, the structure of the antigen-combining site of exceptionally long CDR3H encoded by fetal *V-D-J* recombination is likely to be different due to a relatively shorter or non-existent stalk. The B cells expressing immunoglobulin with exceptionally long heavy chain CDR3 undergo affinity maturation via somatic mutations upon antigen encounter and these heavy chains with unusually long CDR3s exclusively pair with light chains with *Ser90* conserved in the light chain CDR3, which provide minimal structural support without making contact with antigen. In conclusion, these exceptionally long heavy chain CDR3s found in all bovine antibody classes provide a distinct novel mechanism of antibody diversification (Herzig et al., 2015, Cui et al., 2015, Thompson-Crispi et al., 2014).

#### **1.2.4. Bovine Genetics of Disease Resistance**

Ongoing attempts to prevent, control and eradicate the most significant cattle diseases caused by different biological agents have been undertaken in many countries (Flower, 2013). A recent review of challenges and opportunities in USA shows the high complexity and variability of the situation. The bulk of these efforts involve management of cattle production systems and/or veterinary interventions (Shin et al., 2016).

The general question relevant to how knowledge of genetic resistance to certain diseases can be used in practice to complement the other approaches. The focus here is resistance to the deleterious consequences of the infected state, and more particularly, on the genetic basis of diversity in resistance within domestic cattle (Prabakaran et al., 2003). Genetics of Disease Resistance in Cattle Relevant Notes Studies in this field have been driven by two principal

objectives. The first of these, as with all scientific endeavor, is to increase knowledge and understanding. In this regard the genomic revolution dramatically widens opportunities for obtaining previously unavailable information on genes influencing resistance to different diseases (Robbertse et al., 2016, Langeveld et al., 2016). The second objective for research into disease resistance in cattle is the prospect of useful applications in agriculture to improve animal productivity, improve animal welfare or reduce risk of zoonoses. There is significant variation in cattle in terms of resistance to diseases, and this variation is of economic importance. Nevertheless, the application of selection for resistance in the field has been slow to develop (Behl et al., 2016).

Alternative options for disease control are common, such as efficient management practices including test and slaughter or isolation and quarantine, as are veterinary treatments, vaccination and control of infections. In contrast the genetic route to improving the disease resistance of entire breeds is slow and arduous due to long generation intervals, can be compromised by genetic change in the pathogen, and is difficult in many parts of the world that lack adequate animal breeding expertise and required infrastructure (Sundararaman et al., 2016). The extra effort is usually undertaken to minimize the disease exposure of elite animals and their heartmates, thereby reducing the opportunity for direct selection and limiting the information available for conventional prediction of breeding values (Kim et al., 2015a, Keele et al., 2015, Jonas et al., 2015).

For many diseases, there are not obvious phenotypic traits that are reliably correlated to the level of disease, such that the observed phenotypes are often categorical rather than continuous. Disease incidence can vary between herds and years, reducing the amount of information available for prediction when the incidence is low (MacIntyre, 2015).

Considering all pros and cons, one should not miss possible negative correlations between some productivity traits and disease resistance as well as cost of selection. The logistics of experimentation in disease resistance in cattle can pose a considerable challenge. However, the situation is changing and the stimulus to undertake research that will provide new options for disease control in cattle seems to be increasing (Kim et al., 2015b, Kim et al., 2015a, Keele et al., 2015).

The three major reasons. First, resistance among pathogens to chemotherapeutic and chemo-prophylactic drugs is apparently increasing. Compelling examples are resistance to anthelmintic and to trypanocidal compounds (Lupindu et al., 2015). In the case of trypanocidal resistance, it can be argued that development of the livestock sector in some of the poorest countries of the world is jeopardized. Second, safe, effective and inexpensive vaccines have not been developed yet for some economically important diseases. The comparative costs of non-genetic disease control options are also a consideration (Keele et al., 2015). Third, growing volumes of information relevant to genetic resistance to diverse diseases in cattle should provide a background for breeding and selection. Nevertheless, a realistic outlook is necessary. Obviously parasite and pathogen genomes will not remain unchanged while cattle genomes are modified by ongoing selection for resistance (Eidam et al., 2015, E et al., 2015). Still the fact that some livestock populations are relatively resistant to certain diseases, and have remained so for thousands of years in some cases, suggests that ‘agreements’ between pathogens and hosts can be brokered at various levels (Allan et al., 2015).

This view of the genetic option raises another important point. It implies that selection will usually be a means of disease control rather than a means of infection or parasite control per se. A subjective comparison of different disease control options, in terms of a variety of features. Some options like vaccination and movement control look particularly attractive and

are used in cattle populations very regularly (Rodriguez-Rivera et al., 2014, Mughini-Gras et al., 2014, Kizilkaya et al., 2014).

The option of selecting for disease resistance has one major problem, namely difficulty in creating such cattle. Except for a few rare examples, this option was not widely used in the past despite existence of genetic variability in different breeds to a variety of diseases. There are indications that the situation may change in the future due to new knowledge generated by genomics (Flower et al., 2010).

At this third level, potential pathogens may establish, but not cause a significant illness. A good example is provided by Trypanosome Congolese infection in resistant cattle types (Noyes et al., 2016, Lipkin and Strillacci, 2016).



**CHAPTER TWO:**

**LITERATURE REVIEW**

## **2. Literature Review**

### **2.1. Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals.**

For variation in mitochondrial genome and the origin of modern humans, Ingman et al. (2000) claimed about the analysis of mitochondrial DNA (mtDNA) has been a tool in understanding of human evolution. The studies of human evolution based on mtDNA sequencing have been confined to the control region, which constitutes less than 7% of the mitochondrial genome. Most comprehensive studies of the human mitochondrial molecule have been carried out through restriction-fragment length polymorphism analysis, providing data that are ill suited to estimations of mutation rate and therefore the timing of evolutionary events.

Also points out by Ingman and Gyllensten (2001) in studying the analysis of the complete human mtDNA genome through methodology and inferences for human evolution. The mitochondrial DNA hypervariable segment I (HVS-I) is widely used in studies of human evolutionary genetics, and therefore accurate estimates of mutation rates among nucleotide sites in this region are essential. Maximum likelihood methodology has been developed for estimating site-specific mutation rates from partial phylogenetic information, such as



haplogroup association. The resulting estimation problem is a generalized linear model, with a nonstandard link function. The development inference and bias correction tools for estimating and hypothesis-testing approach for site independence. Also demonstrated as methodology using 16,609 HVS-I samples from the Geno-graphic Project. The results suggest that mutation rates among nucleotide sites in HVS-I are highly variable. The 16,400–16,500 region exhibits significantly lower rates compared to other regions, suggesting potential functional constraints. Several loci identified in the literature as possible termination-associated sequences (TAS) do not yield statistically slower rates than the rest of HVS-I, casting doubt on their functional importance. To tests do not reject the null hypothesis of independent mutation rates among nucleotide sites, supporting the use of site-independence assumption for analyzing HVS-I. Potential extensions of methodology include its application to estimation of mutation rates in other genetic regions, like Y chromosome short tandem repeats.

In another hand, for genomes of cryptic chimpanzee plasmodium species reveal key evolutionary events leading to human malaria, Sundararaman et al. (2016) speculates through stating that African apes harbor at least six *Plasmodium* species of the subgenus *Laverania*, one of which gave rise to human *Plasmodium falciparum*. The selective amplification strategy to sequence the genome of chimpanzee parasites classified as *Plasmodium reichenowi* and *Plasmodium gaboni* based on the sub-genomic fragments. Genome-wide analyses show that these parasites indeed represent distinct species, with no evidence of cross-species mating. Both *P. reichenowi* and *P. gaboni* are 10-fold more diverse than *P. falciparum*, indicating a very recent origin of the human parasite. Also finding a remarkable *Laverania*-specific expansion of a multigene family involved in erythrocyte remodeling, and show that a short region on chromosome 4, which encodes two essential invasion genes, was horizontally transferred into a recent *P. falciparum* ancestor. Results validate the selective amplification

strategy for characterizing cryptic pathogen species, and reveal evolutionary events that likely predisposed the precursor of *P. falciparum* to colonize humans.

Moreover, the fossil record of some ungulate lineages allowed estimation of the evolutionary rates for various components of the 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 (>5 million years) it appears that both replacements and silent transversion in this mitochondrial gene can be used for phylogenetic inference. Phylogenetic findings include the association of (Drosophila 12 Genomes et al., 2007) cetaceans, artiodactyls, and perissodactyls to the exclusion of elephants and humans, pronghorn and fallow deer to the exclusion of bovid like cow, sheep, and goat, sheep and goat to the exclusion of other pecorans such as cow, giraffe, deer, and pronghorn, and advanced ruminants to the exclusion of the chevrotain and other artiodactyls (Scheffler, 2008).

Comparisons of these cytochrome sequences support current structure-function models for this membrane-spanning protein. Although there has been relatively results into mitochondrial DNA sequence divergence and diversity Chen and Li (2001) 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 (Balbinotti et al., 2012). More importantly, the modern hypothesis of the evolutionary relationship between human and pig that based on assumption of similarity in some organs tissues like kidneys and eyes (Brown, 2000, Rettenberger et al., 1995).

All in all, the huge similarity in structure and functions appears in the genomic mitochondrial DNA (mtDNA) of vertebrates. That was encouraging point to consider about

having a chance to make a comparative view among seventeen organism 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 (Sevini et al., 2014).

## **2.2. Bovine's *Brucellosis* Antigenicity**

*Brucellosis* is known as an infectious disease characterized by rising and lowering undulant fever, sweating, muscle and joint pains, and weakness. Also have other names like Bang's disease, Enzootic Abortion, Epizootic Abortion, Slinking of Calves, Ram Epididymitis and Contagious Abortion. The pathogen that caused brucellosis is the bacterium *Brucella*, which can be transmitted in unpasteurized milk from cattle, sheep, and goats; cheese made from this unpasteurized milk and contact with diseased animals (Nielsen and Duncan, 1990). Nowadays Antibiotics are available used to treat *Brucellosis*. Moreover, according to the American centers of disease controls and prevention (<http://www.cdc.gov/>) has declared *Brucella* as one of three major bioterrorist agents anthrax, tularemia and *Brucella* (Madkour, 2014, Corbel et al., 2006)

As well as, *Brucellosis* is the responsible for enormous economic losses as well as considerable human morbidity in endemic areas. The bacteria infects animals such as swine, cattle, goat, sheep, and dogs (Rossetti et al., 2013). Humans can become infected indirectly through contact with infected animals or by animal products consumption. *Brucellosis* occurs worldwide, but it is well controlled in most developed countries. The disease is rare in industrialized nations because of routine screening of domestic livestock and animal vaccination programmes (Wang et al., 2011, Wang et al., 2015b). Clinical disease is still common in the Middle East, Asia, Africa, South and Central America. This review article aims

to describe the prevalence of brucellosis in some countries these data are available around different regions of world, and risk factors associated infections according regression models (Ishida, 2004). The term brucellosis is applied to a group of closely related infectious diseases, which are caused by germs of the bacterial species *Brucella*. They occur all over the world. Man almost always receives the infection from infected animals, whereas transmission from man to man usually does not occur. Measures against this anthroozoonosis will therefore always have to aim at the control and eradication of the disease in the animal reservoir as well (Madkour, 2014).

The economy of abortions, infertility, loss of milk and meat in the case of domestic animals and to the public health through, chronic infections and absenteeism from work among the populace by this zoonosis amounts to millions of dollars in the countries affected by it. The bases for meaningful and economic measures in the national as well as in the international context are surveys and proven data on incidence (Nielsen and Duncan, 1990, Corbel et al., 2006). Geomedical maps provide a synopsis of the state of knowledge on the extent of the disease at the particular time, and at the same time maps of brucellosis distribution show the great changes that have taken place in the countries concerned since the last cartographic overview by W. Wundt in 1961 as a result of increased surveys and thus increased knowledge and intensified measures which were adopted in those states (Prabakaran et al., 2003).

Predicting the antigenic sites on proteins is of major importance for the production of synthetic an artificial peptide vaccines and peptide probes of antibody structure. Many predictive methods, based on various assumptions about the nature of the antigenic response have been proposed and tested. This review will discuss the principles underlying the different approaches to predicting antigenic sites and will attempt to answer the question of how well they work (Gwida et al., 2015).

As reviewed from Kolaskar & Tongaonkar method which coined in 1990. Analysis of data from experimentally determined antigenic sites on proteins has revealed that the hydrophobic residues Cys, Z\_XU and Val, if they occur on the surface of a protein, are more likely to be a part of antigenic sites. A semi empirical method which makes use of physiochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes was developed to predict antigenic determinants on proteins. Application of this method to a large number of proteins has shown the method can predict antigenic determinants with about 75% accuracy which is better than most of the known methods (Zinicola et al., 2015, Zhu et al., 2015a, Zhao et al., 2015).

The welling method for antigenicity prediction in 1985 came in contrast. Prediction of antigenic regions in a protein will be helpful for a rational approach to the synthesis of peptides which may elicit antibodies reactive with the intact protein. Earlier methods are based on the assumption that antigenic regions are primarily hydrophilic regions at the surface of the protein molecule (Thompson-Crispi et al., 2014, Jonsson et al., 2014, Hansen et al., 2014).

The method of antigenic prediction presented here is based on the amino acid composition of known antigenic regions in 20 proteins which is compared with that of 314 proteins Sequences and Structure. Antigenicity values were derived from the differences between the two data sets. The method was applied to bovine ribonuclease, the B-subunit of cholera toxin and herpes simplex virus type 1 glycoprotein D. There was a good correlation between the predicted regions and previously determined antigenic regions (Lipkin and Strillacci, 2016).

The most important point of this study is starting from scratch depending on a row data from like NCBI. All of sequences and programs that mentioned before, were examined by EMBOSS web servers then decided to choose the protein sequences that produced as antibodies

by the B-cell which associated with *Brucellosis* resistance in cattle (Usman et al., 2015, Spigelman et al., 2015, Singh et al., 2015).

### 2.3. The main aims s of the study

As a summarize of hole what mentioned in the chapters of Introduction and Literature review, from the historical background and the main categories of bioinformatics science and application in omics and proteomic sequences analysis in evolutionary distance and immunomics. Through four chapters of results and discussion in detail:

1. Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals.
2. Phylogenetic Tree Construction within Maximum Likelihood Method of complete. genomes of mitochondrial DNA between Human's and 16 animals.
3. Nucleotide and amino acid sequences analysis in pathogen and host of brucellosis in cattle.
4. Immunoinformatics, Antigenicity epitopes prediction in the solute carrier family 11 of the natural resistance associated macrophage protein 1 (NRAMP) related with *Brucellosis* in Cattle.



**CHAPTER THREE:**

**METHODOLOGY**

### **3. Materials and Methods**

#### **3.1. The Sources of Database**

For Maximum likelihood estimation of the evolutionary distance of the complete genomes of Mitochondrial DNA (mtDNA) between Human's versus 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 NCBI (Coordinators, 2015) ([www.ncbi.nlm.nih.gov/GENOME](http://www.ncbi.nlm.nih.gov/GENOME)); (Aali et al., 2014, Coordinators, 2015). 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](http://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](http://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 (Andrews et al., 1999).

The reason behind choosing these organisms as it mentioned in the Table 3.1, 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 caused the most controversial and debatable issues among the biologists, for the evolutionary relationship between human and chimpanzee (Jensen, 1993, 1895).

More importantly, including in the list some animals that considered as a highly 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 unique position as far as it is concerned(Beaz-Hidalgo et al., 2015). As below Table 3.1, shows the mitochondrial DNA of 17 vertebrate organisms with their accession number of NCBI, and the INSDC number used in this thesis. 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/nucore/60101824/](http://www.ncbi.nlm.nih.gov/nucore/60101824/)). Secondly, water buffalo with project number 13052 submitted in 02-Agust-2004 ([www.ncbi.nlm.nih.gov/nucore/NC\\_006295](http://www.ncbi.nlm.nih.gov/nucore/NC_006295) ). Finally, Arabian camel with project number 20873 submitted in 17-September-2007 ([www.ncbi.nlm.nih.gov/nucore/NC\\_009849](http://www.ncbi.nlm.nih.gov/nucore/NC_009849)).

Table 3.1. List of organisms which involved in evolutionary study, with the information of database of complete genome, mtDNA.

Taxa	<i>Latin name</i>	Accession numbers	INSDC number	References
1 Human	<i>Homo sapiens</i>	NC_012920	J01415.2	(Andrews et al., 1999)
2 Chimpanzee	<i>Pan troglodytes</i>	NC_001643	D38113.1	(Horai et al., 1995)
3 Gorilla	<i>gorilla gorilla</i>	NC_011120	X93347.1	(Xu and Arnason, 1996)
4 Cattle	<i>Bos taurus</i>	NC_006853	AY526085.1	(Chung HY, Ha JM.,2005) *
5 Water buffalo	<i>Bubalus bubalis</i>	NC_006295	AY702618.1	(Qian JX et all,2004) *
6 Bison	<i>Bison bison</i>	NC_012346	EU177871.1	(Achilli et al., 2008)
7 Arabian camel	<i>Camelus dromedarius</i>	NC_009849	EU159113.1	(Huang X et all, 2007) *
8 Bactrian camel	<i>Camelus bactrianus</i>	NC_009628	EF212037.2	(Ji et al., 2009)
9 Horse	<i>Equus caballus</i>	NC_001640	X79547.1	(Xu and Arnason, 1994)
10 Sheep	<i>Ovis aries</i>	NC_001941	AF010406.1	(Hiendleder et al., 1998)
11 Goat	<i>Capra hircus</i>	NC_005044	GU295658.1	(Hassanin et al., 2010)
12 Pig	<i>Sus scrofa</i>	NC_000845	AF034253.1	(Lin et al., 1999)
13 Chicken	<i>Gallus gallus</i>	NC_001323	X52392.1	(Valverde et al., 1994)
14 Rabbit	<i>Oryctolagus cuniculus</i>	NC_001913	AJ001588.1	(Gissi et al., 1998)
15 Dog	<i>Canis lupus familiaris</i>	NC_002008	U96639.2	(Kim et al., 1998)
16 Domestic cat	<i>Felis catus</i>	NC_001700	U20753.1	(Lopez et al., 1996)
17 House mouse	<i>Mus musculus</i>	NC_005089	AY172335.1	(Bayona-Bafaluy et al., 2003)

\*Refer to resources that unpublished yet as a research paper.

The database related to investigation in genomics and proteomics that associated with *Brucellosis* in cattle must be looking for both inside the pathogen which cause *brucellosis* disease and the animal that infected with.

There are two species of bacterial pathogens that recorded cause brucellosis disease in cattle. Firstly, the *Brucella abortus* with genome size 3,264,306 base pairs divided in two unequal size chromosomes (<https://www.ncbi.nlm.nih.gov/genome/?term=Brucella+%20abortus>). The whole genome that downloaded from genbank of National center of Biotechnology and Information (NCBI), within accession numbers NC\_007618.1 and NC\_007624.1 for chromosome I and II respectively (Chain et al., 2005). Secondly, the *Brucella melitensis* and its genome size 3,294,931 base pairs also divide in two unequal size chromosomes ([www.Ncbi.nlm.Nih.gov/genome/term/Brucella+melitensis](http://www.Ncbi.nlm.Nih.gov/genome/term/Brucella+melitensis)). The NCBI accession numbers of whole-genome chromosome I and II are NC\_003317.1 and NC\_003318.1 respectively (DelVecchio et al., 2002).

Also searching inside the database of the genes that related with producing the antibody to provide the diseases resistance against *brucellosis* in cattle as demonstrated in Table3.2. The most proved name and symbol is Solute Carrier Family 11 (SLC11A1) (<https://www.uniprot.org/uniprot/>; <https://www.Omim.Org/>), also known as Natural Resistance-Associated Macrophage Protein (NRAMP) (Coussens et al., 2004, Chen et al., 2007).

Table 3.2, Database sources of SLC11A1 gene with the NRAMP proteins that produce

Protein's Accession number	DNA's Accession number	Protein's References	DNA's References
1 NP_777077	AC_000159	(Hedges et al., 2013)	(Zimin et al., 2009)
2 ABF61463	DQ493965	*(Martinez et al., 2006)	*(Martinez et al., 2008)
3 ABM81484	DQ848779	*(Schutta et al., 2006)	*(Schutta et al., 2006)
4 ALC78257	KR002419	*(Zhang et al., 2015)	*(Zhang et al., 2015)
5 ALC78258	KR002420	*(Zhang et al., 2015)	*(Zhang et al., 2015)
6 ALC78259	KR002421	*(Shi et al., 2015)	*(Shi et al., 2015)

\*Refer to resources that unpublished yet as a research paper.

Equally important, get downloaded the natural resistance-associated macrophage protein 1, which translated from the SLC11A1 gene. As indicated in Table 3.2, the accession numbers resources of researches that worked on this protein and proved in the laboratory, also worth to mention that all these sources of protein share same sequences and size 548 amino acids (Zimin et al., 2009, Hedges et al., 2013).

### 3.1. 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) ([www.cpan.org](http://www.cpan.org)) that provide thousands of modules shared from scientists and computer programmers studying on bioinformatics (Bailey et al., 2015, Yu et al., 2011, Wu and Nacu, 2010, Vainshtein et al., 2010). Nevertheless, needed to extract some

mathematical functions from ([www.megasoftware.net](http://www.megasoftware.net) ) which is an academic open-public software for molecular evolutionary genetic analysis MEGA7-CC-Porto (Stecher et al., 2014).

As in Table 3.3, the study depended on the most trusted websites that provide bioinformatics tool services specially databases and the multiple sequence alignment, also , with mathematical tools of calculation models services that help to export the models to the software like, Practical Extraction and Report Language (PERL) (Jiang et al., 2015, Hokamp, 2015) and MEGA7 (Stecher et al., 2014).

Table 3.3.: List of website names and the links were used for Genbank database and bioinformatics tool services

	The website's name	URL
1	National Center for Biotechnology Information (NCBI)	<a href="http://www.ncbi.nlm.nih.gov/home/download.shtml">http://www.ncbi.nlm.nih.gov/home/download.shtml</a> <a href="http://www.ncbi.nlm.nih.gov/GENOME">http://www.ncbi.nlm.nih.gov/GENOME</a> <a href="http://blast.ncbi.nlm.nih.gov/Blast.cgi">http://blast.ncbi.nlm.nih.gov/Blast.cgi</a>
2	The European Bioinformatics Institute	<a href="http://www.ebi.ac.uk/services">http://www.ebi.ac.uk/services</a> <a href="http://www.ebi.ac.uk/services/dna-rna">http://www.ebi.ac.uk/services/dna-rna</a> <a href="http://www.ebi.ac.uk/services/proteins">http://www.ebi.ac.uk/services/proteins</a>
3	The Ensemble Project	<a href="http://www.ensembl.org/index.html">http://www.ensembl.org/index.html</a> <a href="http://www.ensembl.org/downloads.html">http://www.ensembl.org/downloads.html</a> <a href="http://www.ensembl.org/info/docs/tools/index.html">http://www.ensembl.org/info/docs/tools/index.html</a>
4	Cambridge Reference Sequence (CRS) for human mitochondrial DNA	<a href="http://isogg.org/wiki/Cambridge_Reference_Sequence">http://isogg.org/wiki/Cambridge_Reference_Sequence</a>
5	Bioinformatics resource portal	<a href="http://www.expasy.org/">http://www.expasy.org/</a> <a href="http://www.expasy.org/phylogeny_evolution">http://www.expasy.org/phylogeny_evolution</a>
6	Math works with PERL	<a href="http://www.mathworks.com/index.html?s_tid=gn_logo">http://www.mathworks.com/index.html?s_tid=gn_logo</a> <a href="http://csifdocs.cs.ucdavis.edu/">http://csifdocs.cs.ucdavis.edu/</a>

7. Perl programming language      <https://www.perl.org/>  
<http://bioperl.org/>
  8. CPAN      <http://www.cpan.org/>  
<http://www.code.org/>
- 

As a common knowledge, the approach of computing databases with programs, depend on three fundamental steps namely the input, run and the output. First, input step of the sequences and mathematical models into the programs, using sequences for DNA and protein through the GENBANK format (Annotated) for programs which provide a visual image and graphical figures, or, FASTA format (Not annotated) for gene discovery, as, sequence alignment (Zuo and Hao, 2015, Yonemoto et al., 2015, Xie et al., 2015, Pan et al., 2015) and apply the mathematical models, these two formats obtained from NCBI (Benson et al., 1998, Benson et al., 2000, Benson et al., 2013) by downloaded as a text in the Notepad, then saved with particular extensions serve the data depending on the program type as an input. To clarify, save the FASTA format as a [filename.pl] if wanted to apply the PERL program and [filename.Mas] for mega7 program. Second, running the programs, some programs provide tools and apply a particular mathematical models and the running process happen in side the software and these are recommended approach for small and medium fragment sizes between 300-1000 bases for each sequence.

moreover, a lot of limitations related with paid plugging or the specific function of the program which designed for that is mean you need to work with several programs to finish one complete research project, in the other hand, there are programs used for large scale of database and without any limitation just open the Command of Microsoft windows [CMD] and paste or drag selected your code script inside the program and apply the codes, depend on the speed and memory of the computer you use.

The major program which applied in Bioinformatics for decades is the Practical Extraction and Report Language [PERL] provide accumulate free thousands of codes in bioinformatics (Bailey et al., 2015, Yu et al., 2011, Wu and Nacu, 2010, Vainshtein et al., 2010). Third, the output, by exporting the results to the email or applied by associating with statistical or graphical software such as Bio. Excel or SPSS, as shown in figure 3.1. below.

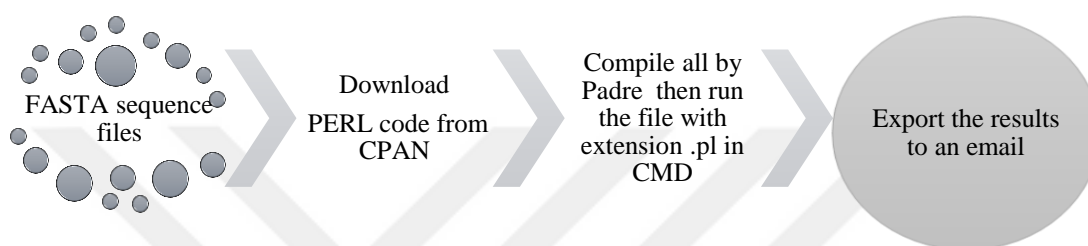


Figure 3.1. The flowchart explain the simple processing steps to apply by PERL

Table 3.4.:\_List of the bioinformatics programs names that used, and the original download links.

<b>SOFTWARE</b>	<b>URL</b>	<b>PURPOSE</b>
1 Snapgene	<a href="http://www.snapgene.com/">http://www.snapgene.com/</a>	Demonstration and graphics
2 MEGA-CC-PROTO	<a href="http://www.megasoftware.net/">http://www.megasoftware.net/</a>	Provide mathematical models to apply with command prompt CMD.
3 *Bio. Excel	<a href="https://bio.codeplex.com/">https://bio.codeplex.com/</a>	Mathematical genetics

The following table (3.4.), shows the list of the downloadable software that used in this research from the official responsible websites.

\*The Bio. Excel is a plugin extension adding bioinformatics tools to the Microsoft office Excel.

All packages of applications that used in this study are freely available for academia users and were compiled and/or under the Gnu/Linux operating system. The core applications used for deriving the antigenicity are in the European Molecular Biology Open Software Suite EMBOSS Stable released version 6.6.0, which is a free open source software analysis package specially developed for the needs of the molecular biology and Bioinformatics user community (<https://emboss.sourceforge.net/apps/>). Mainly, there are three programs that used under EMBOSS. Additionally, the codes of these applications are tested, converted and developed to serve this research specifically like adding the welling method 1985 that have not before with other functions by compiling applying all codes by using Perl programming language (version 5.22.1).

Firstly, the EMMA program, which designed for Multiple sequence alignment (ClustalW wrapper) by calculate the multiple alignment of nucleic acid or protein sequences according to the method of Thompson, J.D., Higgins, D.G. and Gibson, T.J (J. D. Thompson, Higgins, & Gibson, 1994) the usage of program through the command line of Unix terminal order (`% emma`) then paste the sequence in FASTA format this is an interface to the ClustalW distribution.

Secondly, the antigenic program for epitopes binding prediction and used by (`%antigenic`, and `% antigenic -rformat gff`) in the command line (Terminal) to find antigenic sites in proteins sequences. The algorithm of data analysis from experimentally determined antigenic sites on proteins has revealed that the hydrophobic residues Cys, Leu and Val, if they occur on the surface of a protein, are more likely to be a part of antigenic sites (<http://emboss.open-bio.org/rel/dev/apps/antigenic.html>). The method of Kolaskar and Tongaonkar also the welling method are applied to predict antigenic determinants in proteins is semi-empirical and makes use of physiochemical properties of amino acid residues and their



frequencies of occurrence in experimentally known segmental epitopes (Welling et al., 1985, Kolaskar and Tongaonkar, 1990)

Thirdly, the PLOTCON program for dot plot of Sequence conservation is calculated for windows of a specified length over the alignment. Within a window, the similarity of any one position is taken to be the average of all the possible pairwise substitution scores of the bases or residues at that position. The pairwise substitution scores are taken from the specified similarity matrix. The average of the position similarities within the window is plotted. Which depended on applying some results from ClustalW with a mathematical equation like (sequence weighting, matrix comparison table, number of sequences in the alignment and window size) within blocks substitution matrix. The application usage by mainly code of command line is (*%plotcon -sformat msf globins.msf -graph cps*) (Rice et al., 2011).

Furthermore, using the (GNU PSPP) and Gnumeric Spreadsheet 1.12.9 programs for statistical analysis of sampled data and display plots. Also they are a free replacement for the proprietary program SPSS, and appears very similar to it with a few exceptions. To display the plots of antigenicity chart also to apply the basic calculations of amino acids physiochemical properties for protein's Hydrophobic and Hydrophilic through the sequence (<https://www.gnu.org/software/pspp/>).

These sequences stored as XML files which contains the FASTA format to use as an input data. The terminal in (Bio-Linux 8.0/Ubuntu) distribution helps to apply the programs though typing the codes related with running the programs then paste the sequences to get the results saved specific folder within special format (Yang, 2010, de Souto and Kann, 2012).

### **3.2. 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 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/Larry\\_Wall](http://en.wikipedia.org/wiki/Larry_Wall)), were downloaded from CPAN ([www.cpan.org&metacpan.org](http://www.cpan.org&metacpan.org)) also from [www.github.com](http://www.github.com). It is worth mentioning, that programming languages should not be used directly after downloaded from the open source access websites because they are designed for general purposes and need manipulating with adding the private data and the mathematical problems serve the particular study (Wall et al., 2000).

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](http://www.bioperl.org/wiki/Installing_BioPerl_on_Windows)) by using especial 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) then 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) (Qing et al., 2014, Leimeister et al., 2014, Gao et al., 2014).

Another essential point, is the best modules were served the research proses. Firstly, 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](http://metacpan.org/pod/Bio::Tools::Run::Alignment::Muscle)). Secondly, the representation for biological sequence alignment (Bio::Tools::Alignment::Overview) announced by Felipe da Veiga Leprevost in 2014, ([metacpan.org/pod/Bio::Tools::Alignment::Overview](http://metacpan.org/pod/Bio::Tools::Alignment::Overview)). Thirdly, the interface for evolving sequences (Bio::SeqEvolution::EvolutionI) reported by Christopher Fields (2014), ([metacpan.org/pod/Bio::SeqEvolution::EvolutionI](http://metacpan.org/pod/Bio::SeqEvolution::EvolutionI)).

SeqEvolution::EvolutionI). Finally, the module of Maximum likelihood methods (Bio::Tools::Run::Phylo::Molphy::ProtML) authored by Jason Stajich in 2011, ([metacpan.org/pod/ Bio::Tools::Run::Phylo::Molphy::ProtML](http://metacpan.org/pod/Bio::Tools::Run::Phylo::Molphy::ProtML)).

### **3.3. Alignment of 17 mtDNA sequences**

The 17 sequences of mtDNA were arranged in parallel depending on coding and non-coding regions of DNA even the proteins to distinguish regions of similarity and disparity. Consequently, the distance and evolutionary relationships between the sequences were downloaded.

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 (Yang, 2006, Durbin et al., 1998).

The computational multiple sequence alignment (MUSCLE) method used to provide high accuracy for creating different arrangements of high scale amino acids and nucleotide sequences (Jia et al., 2012, Martinez-Perez et al., 2012, Hassanin et al., 2013). 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 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 (Bonhomme et al., 2011, Zhang et al., 2011).

### **3.4. 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 non-random usage of synonymous codons for specific amino acids (Wei et al., 2015, Kishino and Hasegawa, 1989). In addition, the codes of the mathematical problem extracted from program MEGA7-cc-Porto ([www.megasoftware.net](http://www.megasoftware.net)) in a particular file format (filename. Mao) (Xu et al., 2008, Wang et al., 2015a).

### **3.5. 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 (Yang, 2006). The statement of the problem, suppose when have a random sample  $x_1, x_2, \dots, x_n$  whose assumed probability distribution depends on some unknown parameter  $\theta$ . The primary goal here will be to find a point estimator  $u(x_1, x_2, \dots, x_n)$ , such that  $u(x_1, x_2, \dots, x_n)$  is a good point estimate of  $\theta$ , where  $x_1, x_2, \dots, x_n$  are the observed values of the random sample. For example, if planned to take a random sample  $x_1, x_2, \dots, x_n$  for which the  $x_i$  is assumed to be normally distributed with mean  $\mu$  and variance  $\sigma^2$ , then the goal will be to find a good estimate of  $\mu$ , say, using the data  $x_1, x_2, \dots, x_n$  that obtained from a specific random sample ([onlinecourses.science.psu.edu; megasoftware.net](http://onlinecourses.science.psu.edu/megasoftware.net)).

### **3.6. Estimating the evolutionary distances between genomic sequences**

The evolutionary distance between sequences usually is measured by the number of 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 of

phylogenetic reconstruction (Chauve et al., 2013). 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 non-coding 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 (Ross et al., 2008, Kari et al., 2015, Soares et al., 2012b, Soares et al., 2012a, Blair et al., 2013).

### **3.7. Markov models of nucleotide substitution and distance estimation with ML**

In the statistical genetic of bioinformatics, the probability is most used on to understand the changes that happened in DNA and Protein sequences by making the comparison between sequences within maximum likelihood method, additionally, also could predict or discover the mutations, functions and the evolutionary process in organisms. In fact there are many definitions to explain the probability (Wei et al., 2015, Tamura and Nei, 1993, Tamura, 1992, Felsenstein and Churchill, 1996). Simply, the traditional, established meaning of probability that is illustrate  $P(A)$ , an occasion A is resolved from the earlier without real experimentation. It is given by

Where N is the number of possible outcomes and NA is the number of outcomes that are favorable to the event A. Additionally, the Markov models in the DNA sequence work on the probability of changing between the four nucleotide letters Randomly, Conditionally or independently, related with the position and the strength of the bond as in the figure (2.3.) shown below(Kimura, 1980, Kishino and Hasegawa, 1989, Hasegawa et al., 1985, Felsenstein, 1981).

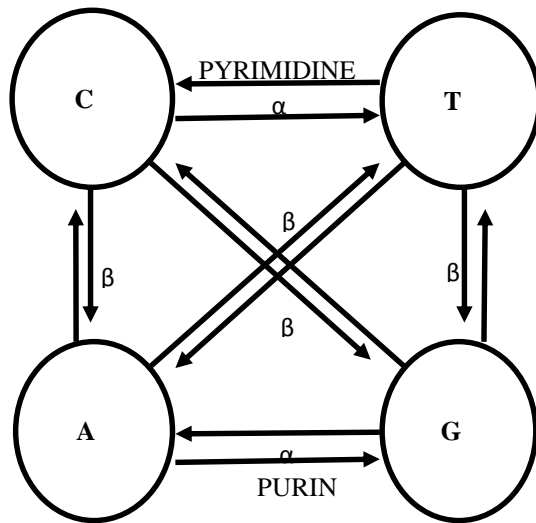


Figure 3.2. Shows the probability of substitution between nucleotides

In the figure (3.2.), relative substitution rates between nucleotides under 3 Markov-chain models of nucleotide substitution: namely (JC69) Jukes and Cantor in (1969), (K80) Kimura (1980), and (HKY85) Hasegawa et al. (1985). The thickness of the lines represents the substitution rates while the sizes of the circles represent the steady-state distribution (Wei et al., 2015, Felsenstein and Churchill, 1996). The Jukes-Cantor model assumes equal base frequencies and equal mutation rates; therefore, it does not have any free parameter.

The Kimura model assumes equal base frequencies and accounts for the difference between transitions and transversion with one parameter. But the HKY85 model does not assume equal base frequencies and accounts for the difference between transitions and transversion with one parameter. The Tamura-Nei model (1993) corrects for multiple hits, taking into account the differences in substitution rate between nucleotides and the inequality of nucleotide frequencies. It distinguishes between transitional substitutions rates between purines and transversion substitution rates between pyrimidine. It also assumes equality of substitution rates among sites (Tamura and Nei, 1993, Kimura, 1980, Felsenstein, 1981).

### 3.8. Estimation of the probability of amino acid substitution

This calculation depend on the results of estimation the probability substitution of nucleotide by markov model within maximum likelihood method, additionally, depending on genetic codon bias, by assuming the substitutions of amino acid codes after the maximum and minimum changes in aligned DNA sequences as well as, the process done by MEGA Software (Pan et al., 2015, Khan et al., 2013) .

### **3.9.Disparity test of real substitution patterns Heterogeneity**

The number of *Monte Carlo* replication and that was (500), as a simulation technique of what happened in the cell with errors that generate the disparity of heterogeneity in the molecular evolutionary between aligned sequences. The number of *Monte Carlo* replication is, doing econometric means estimating parameters, such as the mean of a Population, the coefficients in a linear regression or the auto correlation of time series given a sample of real world data. Besides the point estimate itself, which desired to know how close estimate is to the true Value (Soares et al., 2012b, Chen et al., 2012, Ross et al., 2008).

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The above elements in the bivariate ordinary least squares model, with  $u_i \sim N(0, \sigma^2)$ . The stochastic element in the model is  $u_i$ , the exogenous part is  $x_i$  is either fixed are also stochastic. Assuming values for the true parameters *Aali et al. (2014)* and  $\beta$  and drawing values for the stochastic element, that could simulate the endogenous variable (Adwan et al., 2013). The values of interest are then the least squares estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  in the simulated data set (Kim et al., 2012).

The parameter of estimating the nucleotide real substitution disparity test patterns in heterogeneity, is the number of *Monte Carlo* replication (Antunes and Ramos, 2005), complete detection of gaps or missing data treatment, finally, and selected the 1<sup>st</sup>, 2<sup>nd</sup> and the 3<sup>rd</sup> codon positions, plus, the non-coding sites, and also estimated by the evolutionary distance tree of the maximum likelihood. Nevertheless, which used the same parameters with estimate the disparity

of real substitution amino acid pattern heterogeneity, just with one extra parameter is the genetic codon table of the vertebrate mitochondrial genome (Lin et al., 2011, Pareek et al., 2011, Kabekkodu et al., 2014, Mehta et al., 2015).

### **3.10. Estimating the evolutionary distances between genomic sequences**

The evolutionary distance between a pair of sequences usually is measured by the number of polynucleotide or amino acid substitutions appear between them. Alignment methods are used to compute evolutionary distances between DNA and protein Sequences as a basis of phylogenetic reconstruction, evolutionary distance estimation in pairwise DNA sequences that is calculated from the number of word matches between them (Soares et al., 2012b, Soares et al., 2012a, Ross et al., 2008), additionally, computing the substitutions of nucleotide, amino acids and the synonymous-non-synonymous codes. Nucleotide Sequences are compared nucleotide-by-nucleotide, these distances can be computed for protein coding and non-coding nucleotide sequences, as well as, application of the 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 (Ross et al., 2008, Kari et al., 2015, Soares et al., 2012b, Soares et al., 2012a, Blair et al., 2013).

### **3.11. Molecular Phylogenetic analysis by Maximum Likelihood method**

The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Nei and Kumar, 2000, Yang, 2006). The initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and Bio-NJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site (next to the branches). The analysis involved 17 nucleotide sequences. Codon positions included were 1<sup>st</sup>,



2<sup>nd</sup>, 3<sup>rd</sup> and Non-coding. All positions containing gaps and missing data were eliminated (Corneli and Ward, 2000, Beerli and Felsenstein, 2001, Rosset et al., 2008).

### **3.12. Predicting the antigenicity epitops**

Predicting the antigenic sites on proteins is of major importance for the production of synthetic an artificial peptide vaccines and peptide probes of antibody structure. Many predictive methods, based on various assumptions about the nature of the antigenic response have been proposed and tested. This review will discuss the principles underlying the different approaches to predicting antigenic sites and will attempt to answer the question of how well they work (Stern, 1991).

Kolaskar & Tongaonkar method was coined in 1990. Analysis of data from experimentally determined antigenic sites on proteins has revealed that the hydrophobic residues Cys, Z\_XU and Val, if they occur on the surface of a protein, are more likely to be a part of antigenic sites. A semi empirical method which makes use of physiochemical properties of amino acid residues and their frequencies of occurrence in experimentally known segmental epitopes was developed to predict antigenic determinants on proteins. Application of this method to a large number of proteins has shown that method can predict antigenic determinants with about 75% accuracy which is better than most of the known methods (Amat-ur-Rasool, Saghir, & Idrees, 2015; Cai et al., 2015; Kolaskar & Tongaonkar, 1990) (Zygmunt et al., 2015)

In another hand, the welling method for antigenicity prediction in 1985 came in contrast. Prediction of antigenic regions in a protein will be helpful for a rational approach to the synthesis of peptides which may elicit antibodies reactive with the intact protein. Earlier methods are based on the assumption that antigenic regions are primarily hydrophilic regions at the surface of the protein molecule (Sun et al., 2002). The method based on the amino acid composition of known antigenic regions in 20 proteins which is compared with that of 314

proteins Sequences and Structure. Antigenicity values were derived from the differences between the two data sets. The method was applied to bovine ribonuclease, the B-subunit of cholera toxin and herpes simplex virus type 1 glycoprotein D. There was a good correlation between the predicted regions and previously determined antigenic regions (Welling et al., 1985, Rice et al., 2011).

The most important point of this study is starting from scratch depending on a row data from proved sources like NCBI. All of sequences and programs that mentioned before, were examined by EMBOSS web servers then decided to choose the protein sequences that produced as antibodies by the B-cell which associated with *Brucellosis* resistance in cattle.

## **CHAPTER FOUR:**

### **4. RESULTS AND DISCUSSIONS**



**4.1. MAXIMUM LIKELIHOOD ESTIMATION OF THE EVOLUTIONARY  
DISTANCE OF COMPLETE GENOMES OF MITOCHONDRIAL DNA  
BETWEEN HUMAN'S AND 16 ANIMALS**

## **4.1. Maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals**

After computing the sequences of mitochondrial complete genomes of human and other 16 mammals, the results came gradually, in order as will be demonstrated subsequently. It can see variable results numbers in spite of they have same function of mitochondrial DNA, and have same number of translated proteins, that explain how evolution could provide by having same protein with different lengths and sequences, also in this study try to understand the nucleotide behavior through the 17 species.

### **4.1.1. Computing the statistical quantities for sequence data**

#### **4.1.1.1. The nucleotide composition**

The Figure 4.1.1, 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 complete translated protein. But the number of proteins is constant and similar in all species

are 13 proteins. Even, have 22 tRNAs and 2 rRNAs, these numbers did not change between the 17 vertebrates, namely, that have the same function with alternative lengths and sequences, to help providing more functions to same job, and this is the molecular evolution besides can find a various mitochondrial DNA sizes even between breeds of each species.

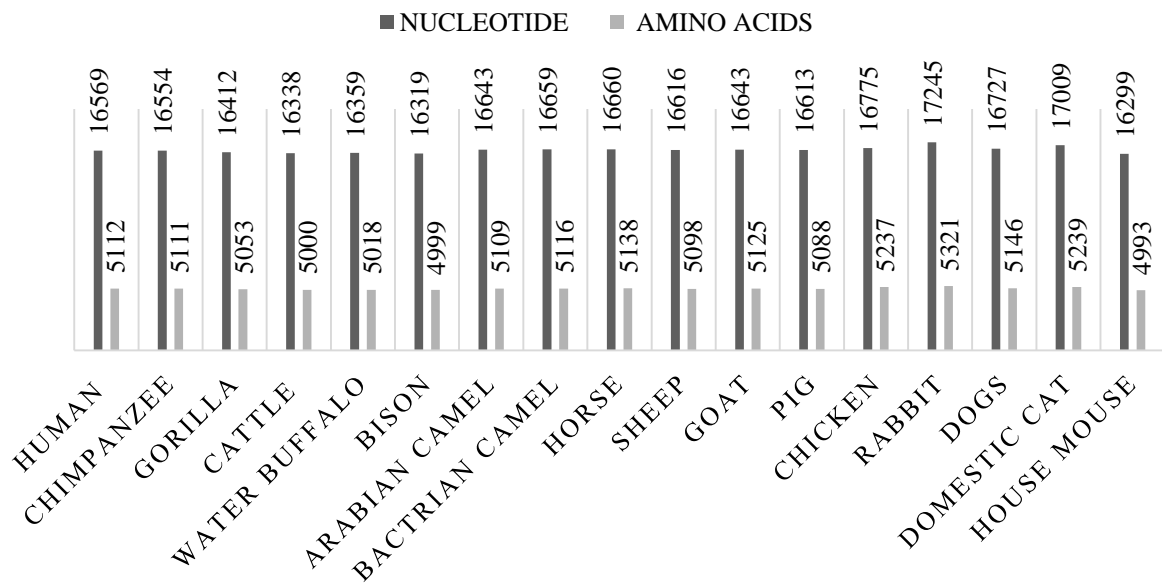


Figure 4.1.1.: The length graph of nucleotide bases and amino acids number in Human’s mitochondrial DNA with other 16 vertebrates

The nucleotide composition shows relative frequencies of the four nucleotides process for one particular sequence of each species as the Figure 4.1.2. which demonstrate the nucleotide frequencies in the mitochondrial DNA sequence of Human against 16 vertebrate species, as well as notice that are not whole of nucleotides as 100% are involved in the codon regions as the figures (4.1.3., 4.1.4., 4.1.5. and 4.1.6.). Moreover, the relationship between the nucleotide frequencies and the ratio of codes produced that helps to understand the nucleotide behavior in sequences.

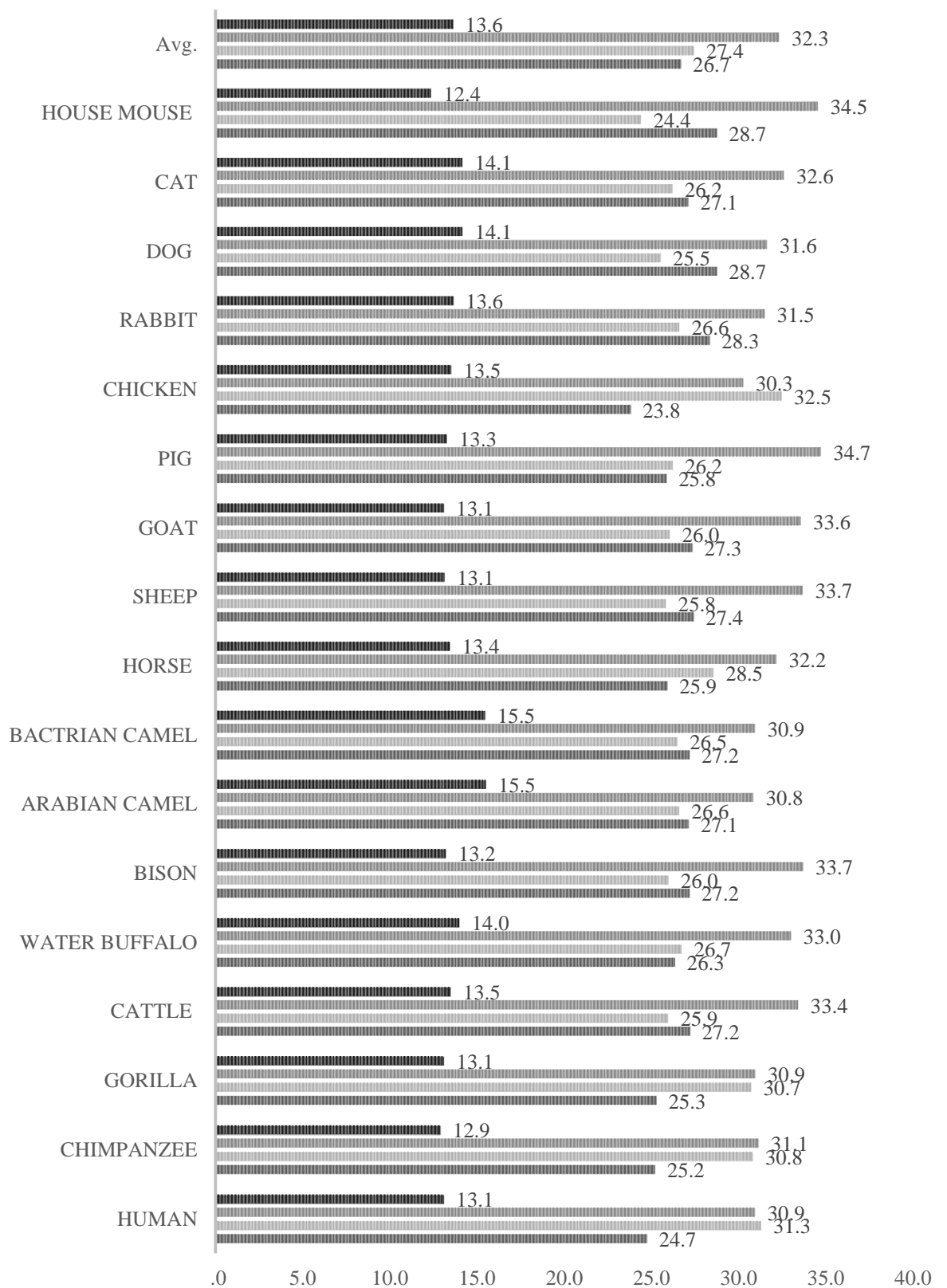


Figure 4.1.2.: The total nucleotide compositions of mitochondrial genome, occur among the Human's and other mammals

From figure (4.1.2.), the percentage levels occur for nucleotides. Adenine had the highest percentage in all sequences, about 32.3% as in average, and Guanine was the lowest, 13.6% as in average. The human's mitochondrial DNA was in the middle versus the 16 vertebrate mitochondrial DNA sequences. The highest level of nucleotide was 28.7% in [T (U)] for dogs, 32.5% in (C) for chicken, 34.7% in (A) for pig and 15.5% in (G) for Arabian camel. Additionally, the lowest frequencies of nucleotide percentages are T (U) 23.8% in chicken, C 24.4% in mouse, 30.3% in chicken and (G) 12.4% in mouse.

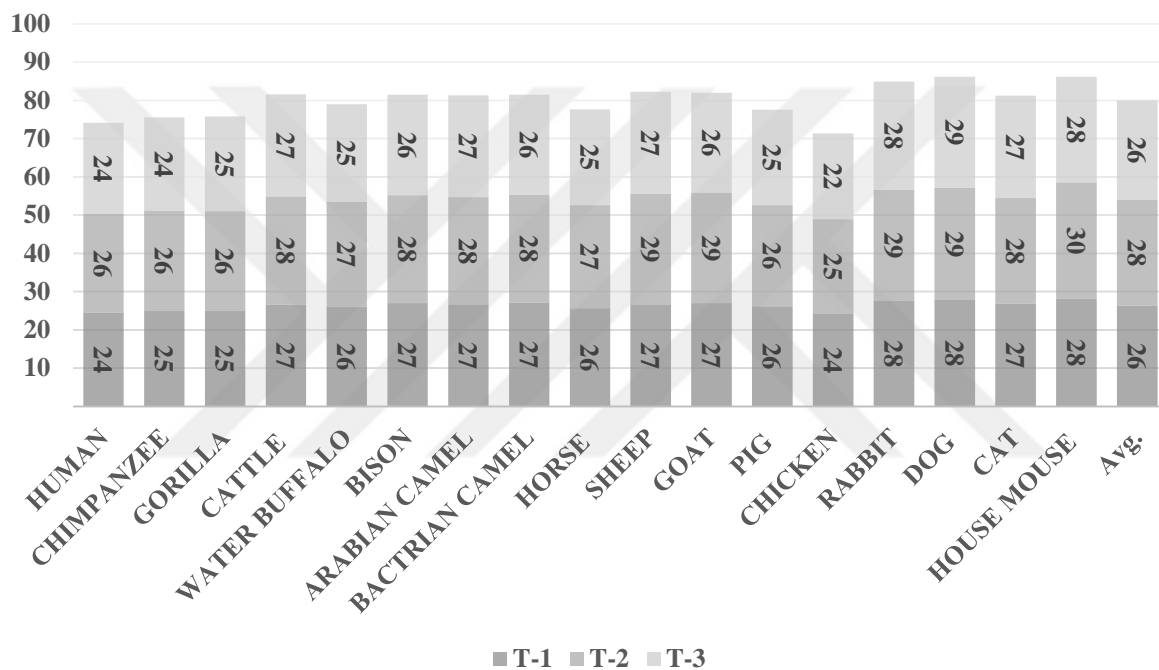


Figure 4.1.3.: The Thymine (Uracil) frequencies within the protein coding regions of DNA at the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> position.

As in figure (4.1.3.) shows, the distribution of the Thymine (Uracil) frequencies in percentage appeared through the three positions of codon regions in mitochondrial DNA. The chicken had the lowest ratios ever in the tree position one, two and three 22.4, 24.7 and 24.2% in order, then the Human become the second after chickens, but the highest percentages in the position one and two (T-2) found in house mouse 28.2, 30.4% in order, finally, the third position 29.0% in dog.

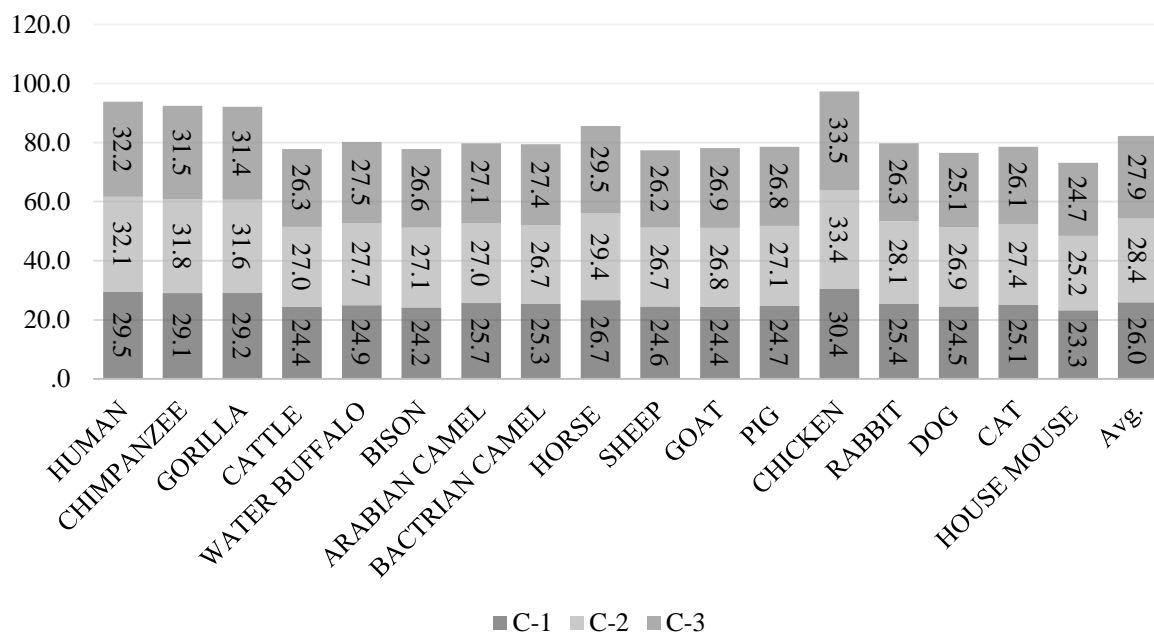


Figure 4.1.4. The Cytosine frequencies within the protein coding regions of DNA at the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> position.

The demonstration of the table (4.1.4.), about the distributions of the nucleotide Cytosine within the protein coding regions of mitochondrial DNA at three positions C-1, C-2 and C-3, so as the table, human's sequence was shown the frequencies of cytosine in the three positions as 29.5, 32.1, 32.2% in order. The most elevated levels of cytosine was occur in chicken at the all positions as 30.4, 33.4 and 33.5% in order, besides, the lowest levels of cytosine found in house mouse 23.3, 25.2 and 24.7% in order with the first, second and third position as reported before.



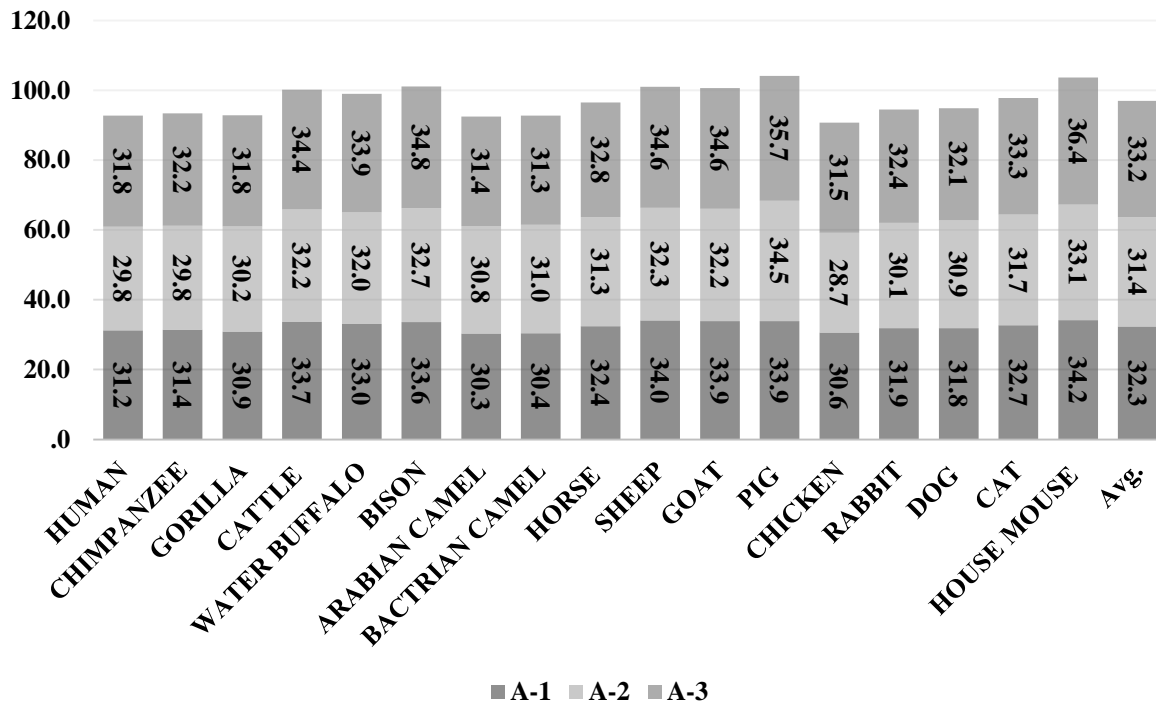


Figure 4.1.5.: The Adenine frequencies within the protein coding regions of DNA at the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> position.

At the figure (4.1.5.), the human's mitochondrial DNA sequence have the allocation of Adenine through the three positions A-1, A2 and A-3, are found as 31.2, 29.8 and 31.8 in order. The highest levels of adenine frequencies that occur in protein coding regions of DNA, at A-1 34.2% in mouse, at A-2 34.5% in pig, and at A-3 36.4% in mouse. Likewise, the lowest levels of adenine are observed, at A-1-30.3% in Arabian camel, at A-2-28.7% in chicken and finally, at A-3-31.3% in Bactrian camel.

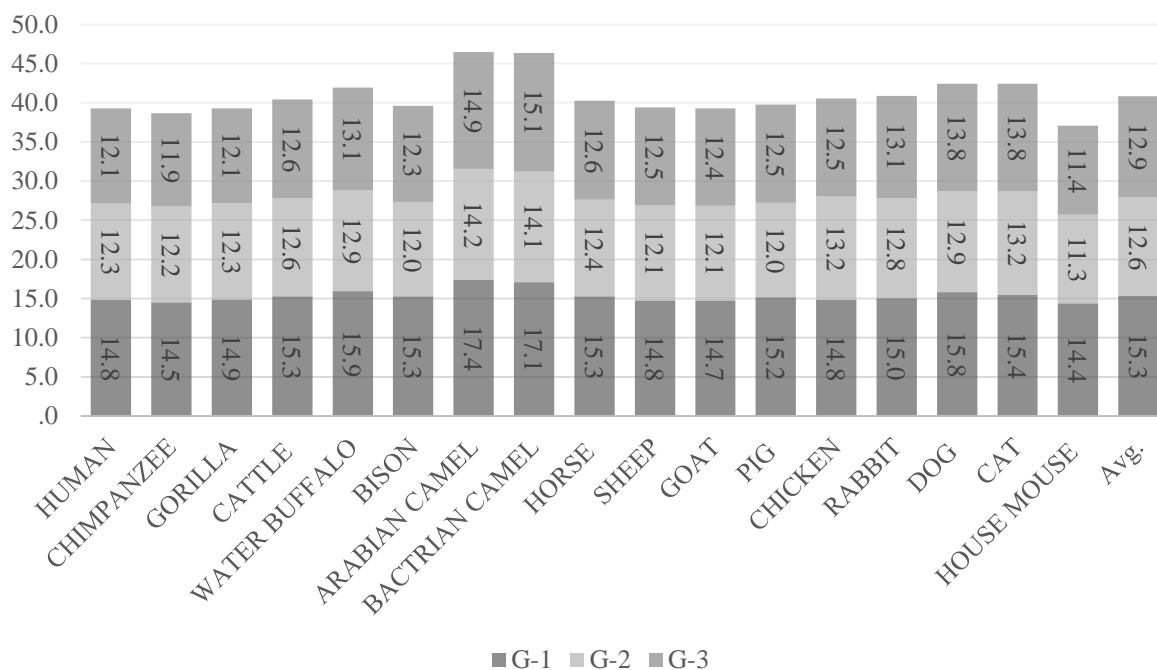


Figure 4.1.6.: The Guanine frequencies within the protein coding regions of DNA at the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> position.

The results in Table 4.1.6. as shown above, explain the frequencies if Guanine for human mtDNA in the positions G-1, G-2 and G-3 are 14.8, 12.3 and 12.1% in order, however, Arabian camel and Bactrian camel got the highest levels in guanine at the three positions of protein coding regions, in contrast to them, the lowest levels in all positions ever was in house mouse 14.4, 11.3 and 11.4% in order to the three positions.

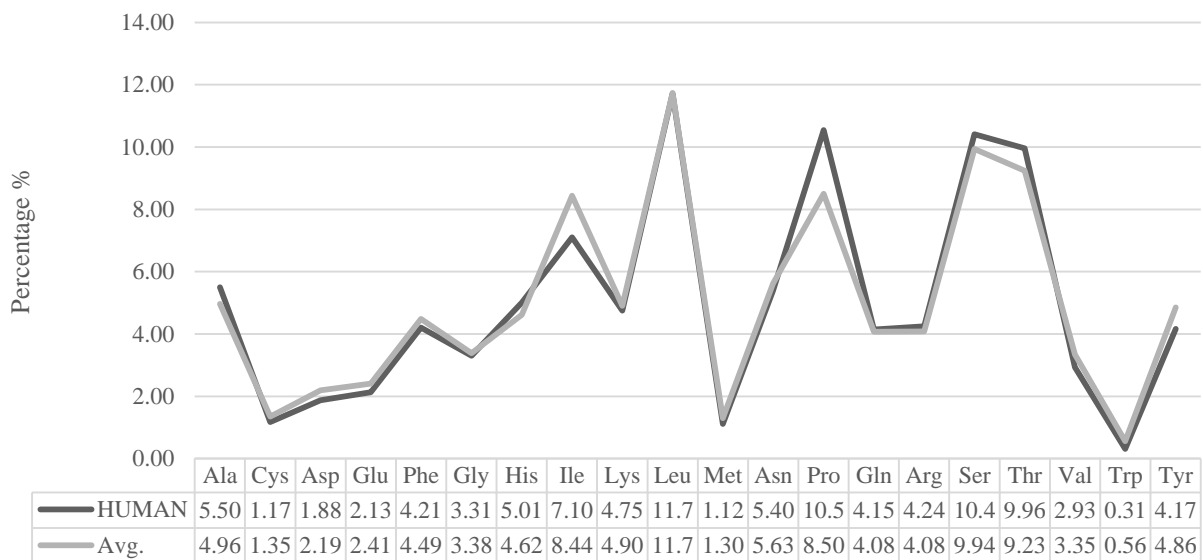


Figure 4.1.7.: The Frequencies (%) of amino acids occur in mitochondrial proteomes including human or average of rest of 16 vertebrates.

#### 4.1.1.2. The composition of amino acids

The amino acid composition of relative frequencies percentages results in the mitochondrial proteome, between human versus vertebrates, help to discover and understand the characteristics and mechanism of translation in mitochondrial DNA, and these results are come out after proceed the global multiple sequence alignment in proteome sequences of 17 taxa. The figure (4.1.7.) shows the comparison in general that realize high frequency levels, as in averages for Isoleucine (Ile) 8.44%, Leucine (Leu) 11.74%, Proline (Pro) 8.50%, Serine (Ser) 9.94% and Threonine (Thr) 9.23%, however, the lowest percentage levels are Cysteine (Cys) 1.35%, Aspartic acid (Asp) 2.19%, Glutamic acid (Glu) 2.41%, Glycine (Gly) 3.38% and Valine (Val) 3.35%, additionally, Leucine has the top level in all organisms. In addition, in Appendix B (tables B3 and B4) demonstrate the amino acid frequencies of each species shown the same harmony of levels as in human's mitochondrial proteome.

#### 4.1.2. Estimation of the Codon Usage Bias

The results of the codon bias, Table 4.1.1. 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. 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.

Table 4.1.1. The frequency account of the codons and the Relative Synonymous Codon

Usage (RSCU), in all over the 17 aligned mammalian mitochondrial sequences.

<b>Codon</b>	<b>Count</b>	<b>RSCU</b>	<b>Codon</b>	<b>Count</b>	<b>RSCU</b>	<b>Codon</b>	<b>Count</b>	<b>RSCU</b>	<b>Codon</b>	<b>Count</b>	<b>RSCU</b>
<b>UUU(F)</b>	109	0.95	<b>UCU(S)</b>	100	1.19	<b>UAU(Y)</b>	130	1.05	<b>UGU(C)</b>	28.4	0.82
<b>UUC(F)</b>	120	1.05	<b>UCC(S)</b>	117	1.38	<b>UAC(Y)</b>	118	0.95	<b>UGC(C)</b>	40.8	1.18
<b>UUA(L)</b>	132	1.32	<b>UCA(S)</b>	126	1.49	<b>UAA(*)</b>	135	1.38	<b>UGA(*)</b>	74.6	0.76
<b>UUG(L)</b>	44.3	0.44	<b>UCG(S)</b>	33.5	0.4	<b>UAG(*)</b>	84.4	0.86	<b>UGG(W)</b>	28.6	1
<b>CUU(L)</b>	90.8	0.91	<b>CCU(P)</b>	138	1.27	<b>CAU(H)</b>	126	1.07	<b>CGU(R)</b>	27.1	0.78
<b>CUC(L)</b>	101	1.01	<b>CCC(P)</b>	138	1.27	<b>CAC(H)</b>	110	0.93	<b>CGC(R)</b>	29.9	0.86
<b>CUA(L)</b>	185	1.85	<b>CCA(P)</b>	120	1.11	<b>CAA(Q)</b>	145	1.39	<b>CGA(R)</b>	35.4	1.02
<b>CUG(L)</b>	47.4	0.47	<b>CCG(P)</b>	37.6	0.35	<b>CAG(Q)</b>	63.2	0.61	<b>CGG(R)</b>	15.9	0.46
<b>AUU(I)</b>	140	0.97	<b>ACU(T)</b>	130	1.1	<b>AAU(N)</b>	138	0.96	<b>AGU(S)</b>	45.6	0.54
<b>AUC(I)</b>	134	0.93	<b>ACC(T)</b>	143	1.21	<b>AAC(N)</b>	149	1.04	<b>AGC(S)</b>	85.5	1.01
<b>AUA(I)</b>	158	1.1	<b>ACA(T)</b>	154	1.31	<b>AAA(K)</b>	181	1.45	<b>AGA(R)</b>	62.4	1.79
<b>AUG(M)</b>	66.5	1	<b>ACG(T)</b>	44.9	0.38	<b>AAG(K)</b>	69.3	0.55	<b>AGG(R)</b>	37.9	1.09
<b>GUU(V)</b>	36.2	0.85	<b>GCU(A)</b>	69.5	1.1	<b>GAU(D)</b>	56.1	1	<b>GGU(G)</b>	32.8	0.76
<b>GUC(V)</b>	39.6	0.93	<b>GCC(A)</b>	92.4	1.46	<b>GAC(D)</b>	55.9	1	<b>GGC(G)</b>	46	1.06
<b>GUA(V)</b>	68.5	1.6	<b>GCA(A)</b>	74.7	1.18	<b>GAA(E)</b>	71.7	1.16	<b>GGA(G)</b>	66.6	1.54
<b>GUG(V)</b>	26.7	0.62	<b>GCG(A)</b>	17.1	0.27	<b>GAG(E)</b>	51.5	0.84	<b>GGG(G)</b>	27.6	0.64

\*Termination codes of transcription.

#### 4.1.3. Probabilistic of nucleotide substitution with (ML)

By using markov models to estimate the probabilistic of nucleotide substitution, to find the best ratio in Nucleotide/Amino acid within maximum likelihood method (ML) and the tables (4.1.2.) Show the best and reliable results in probability of nucleotide substitution with maximum likelihood, in spite of was the lowest, the top was inside pyrimidine group for (C=>T) 0.35 and (T=>C) 0.35, but there is no substitution from (T=>G) and (C=>G), even so, Table 4.1.3. Shows the highest and only one probability score (0.08).

Table 4.1.2. The lowest levels of probabilistic estimation in nucleotide substitution with ML

From\To	A	T	C	G
A	-	0.02	0.03	0.05
T	0.02	-	0.35	0.00
C	0.03	0.35	-	0.00
G	0.13	0.01	0.01	-

Table 4.1.3.: The highest levels of probabilistic estimation in nucleotide substitution with ML

From\To	A	T	C	G
A	-	0.08	0.08	0.08
T	0.08	-	0.08	0.08
C	0.08	0.08	-	0.08
G	0.08	0.08	0.08	-

#### 4.1.4. Probabilistic of the amino acid substitutions with (ML)

The results of the Table 4.1.4., came after estimation the probability substitution of nucleotide by markov model within maximum likelihood method, additionally, depending on genetic codon bias, by assuming the substitutions of amino acid codes between the maximum and minimum changes in aligned DNA sequences. Using basic mathematical calculations after exporting data to Excel. Consequently, as noticed from the Table 4.1.4. the highest possible substitution could be found in the amino acids (I, L, V, A, T and S) with the standard error estimated at the upper group over the diagonals.



Table 4.1.4.: Estimate the probabilistic of the amino acid substitutions with ML, with the standard error

From\To	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	-	0.001	0.002	0.001	0.000	0.001	0.001	0.003	0.000	0.002	0.002	0.001	0.000	0.000	0.008	0.019	0.022	0.000	0.000	0.005
R	0.001	-	0.001	0.000	0.001	0.005	0.000	0.003	0.007	0.001	0.002	0.014	0.000	0.000	0.003	0.005	0.003	0.000	0.001	0.000
N	0.002	0.001	-	0.007	0.000	0.001	0.001	0.001	0.009	0.002	0.001	0.005	0.000	0.000	0.001	0.024	0.011	0.000	0.002	0.000
D	0.002	0.000	0.016	-	0.000	0.001	0.010	0.002	0.002	0.001	0.000	0.001	0.000	0.000	0.001	0.003	0.002	0.000	0.001	0.001
C	0.002	0.002	0.001	0.000	-	0.000	0.000	0.001	0.002	0.001	0.001	0.000	0.000	0.002	0.001	0.010	0.002	0.000	0.005	0.001
Q	0.002	0.006	0.002	0.001	0.000	-	0.005	0.000	0.013	0.000	0.004	0.006	0.000	0.000	0.007	0.003	0.002	0.000	0.001	0.000
E	0.003	0.001	0.002	0.009	0.000	0.005	-	0.002	0.001	0.001	0.001	0.004	0.000	0.000	0.001	0.002	0.002	0.000	0.000	0.001
G	0.005	0.003	0.002	0.002	0.000	0.000	0.001	-	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.009	0.002	0.000	0.000	0.001
H	0.001	0.006	0.012	0.001	0.000	0.009	0.000	0.000	-	0.001	0.003	0.001	0.000	0.001	0.005	0.004	0.002	0.000	0.013	0.000
I	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	-	0.014	0.000	0.003	0.002	0.000	0.002	0.012	0.000	0.001	0.017
L	0.001	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.011	-	0.000	0.003	0.006	0.004	0.003	0.001	0.000	0.001	0.003
K	0.001	0.012	0.007	0.000	0.000	0.005	0.002	0.001	0.001	0.001	0.001	-	0.000	0.000	0.001	0.002	0.005	0.000	0.000	0.000
M	0.001	0.001	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.022	0.023	0.001	-	0.001	0.001	0.001	0.010	0.000	0.000	0.005
F	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.015	0.000	0.000	-	0.001	0.005	0.001	0.000	0.012	0.001
P	0.005	0.001	0.000	0.000	0.000	0.002	0.000	0.000	0.003	0.000	0.006	0.000	0.000	0.000	-	0.013	0.006	0.000	0.000	0.000
S	0.011	0.002	0.015	0.001	0.001	0.001	0.000	0.004	0.002	0.002	0.003	0.001	0.000	0.002	0.012	-	0.023	0.000	0.001	0.001
T	0.013	0.001	0.007	0.001	0.000	0.001	0.000	0.001	0.001	0.012	0.002	0.002	0.001	0.000	0.005	0.023	-	0.000	0.000	0.002
W	0.000	0.002	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.001	0.003	0.000	0.000	0.001	0.000	0.002	0.000	-	0.002	0.000
Y	0.000	0.000	0.002	0.001	0.001	0.000	0.000	0.000	0.013	0.001	0.001	0.000	0.000	0.013	0.000	0.003	0.001	0.000	-	0.000
V	0.008	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.043	0.010	0.000	0.002	0.001	0.001	0.002	0.005	0.000	0.000	-

Caption: The number of base substitutions per site from between sequences are shown.

Standard error estimate(s) are shown above the diagonal.

#### 4.1.5. 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). The observation reported changes in the nucleotide through the 17 mitochondrial genome sequences, are illustrated vertically in columns of Table 4.1.5.

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=>C 11.5959 + A=>G 16.1529). 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=T 0.7138), another essential point is the highest substitution shown the transition inside the pyrimidine group between two nucleotides T=>C 19.9855 and C=>T 20.3332.

Additionally, the results were agreeing with the next following researches 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.

Table 4.1.5. Maximum Likelihood Estimation of Transition/Transversion Bias



From\To	A	T	C	G
A	-	5.0463	9.9574	7.0297
T	5.9789	-	20.3332	0.6117
C	11.5959	19.9855	-	0.7138
G	16.1529	1.1863	1.4085	-

#### 4.1.6. Nucleotide Pair Frequencies from alignment of 17 sequences.

The calculation of the transition/transversion in maximum probable number of 16 nucleotide pairs that could obtain from four different nucleotides, through alignment of 17 sequences in the positions 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> 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 the Table 4.1.6., 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 exemplify, a high ratio of Asparagine N and Lysine K because them codon contain the AA.

Furthermore, as demonstrated the top number of observations in the Table 4.1.6., 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.

Table 4.1.6. The transition/transversion calculated of 16 probable nucleotide pair frequencies by alignment of 17 sequences, in three codon positions.

	*ii	si	sv	R	TT	TC	TA	TG	CT	CC	CA	CG	AT	AC	AA	AG	GT	GC	GA	GG
Avg	12097	1975	1814	1	3188	730	322	65	619	3144	411	70	309	498	4047	316	59	80	310	1718
1 <sup>st</sup>	4184	585	517	1	1087	209	97	20	179	1027	113	22	89	139	1389	99	16	21	98	680
2 <sup>nd</sup>	4038	642	619	1	1124	239	106	22	206	1092	146	25	100	175	1297	100	19	28	97	524
3 <sup>rd</sup>	3876	749	678	1	977	283	119	23	234	1025	153	23	120	184	1360	117	23	31	115	513

\* ii: total of 16 nucleotide pairs identical pairs; si: total of 16 nucleotide pairs transition pairs;

sv: transversion pairs; R: the ratio of transition/ transversion ( $R=si/sv$ ) with total of 16 nucleotide pairs.

#### 4.1.7. Nucleotide evolutionary distance

The parameter of results depended on the numbers of base substitutions per site from between sequences as are shown in Table 4.1.7. 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 1<sup>st</sup>+2<sup>nd</sup>+3<sup>rd</sup>+Noncoding. Additionally, all positions containing gaps and missing data were eliminated. There were a total of 14430 positions in the final dataset.

Deeply, in details the results of evolutionary distance as shown in Table 4.1.7., separated the 17 organisms in several groups depending on the value of 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 appear diagonally in the lower matrix between cattle, bison Arabian camel, Bactrian camel, horse, sheep and goat. Moreover, results demonstrate lowest divergence between Arabian camel and Bactrian camel, with cattle, and also shown among dog, rabbit and sheep.

The most interesting results that relate with highest evolutionary distance in pig with all 16 organisms in contrast whilst, chicken also had high divergence scores with all but

significantly lower than the mtDNA of pig. That is mean in spite of the highly morphological contrast between chicken and other organisms even it is not 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 comparing to the all other organisms.

Table 4.1.7. The substitution of nucleotide, evolutionary distance in alignment of 17 sequences of mtDNA.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Human		0.0025	0.0035	0.0077	0.0096	0.0096	0.0084	0.0075	0.0074	0.0110	0.0103	0.0172	0.0091	0.0085	0.0091	0.0082	0.0105
2. Chimpanzee	0.0913		0.0035	0.0083	0.0103	0.0100	0.0094	0.0085	0.0082	0.0108	0.0101	0.0184	0.0086	0.0088	0.0082	0.0093	0.0103
3. Gorilla	0.1157	0.1108		0.0084	0.0104	0.0095	0.0090	0.0081	0.0076	0.0114	0.0109	0.0179	0.0101	0.0088	0.0092	0.0093	0.0098
4. Cattle	0.3819	0.3722	0.3788		0.0052	0.0051	0.0061	0.0063	0.0056	0.0059	0.0058	0.0158	0.0057	0.0055	0.0062	0.0051	0.0091
5. W. Buffalo	0.3965	0.3899	0.3991	0.2681		0.0032	0.0020	0.0059	0.0060	0.0038	0.0043	0.0141	0.0082	0.0059	0.0055	0.0059	0.0097
6. Bison	0.3920	0.3934	0.3986	0.2699	0.1329		0.0031	0.0067	0.0061	0.0047	0.0046	0.0147	0.0072	0.0058	0.0054	0.0056	0.0098
7. A. Camel	0.3988	0.3962	0.3943	0.2689	0.0597	0.1275		0.0051	0.0057	0.0034	0.0040	0.0134	0.0069	0.0057	0.0054	0.0051	0.0099
8. B. Camel	0.4310	0.4301	0.4324	0.3006	0.2947	0.2956	0.2935		0.0022	0.0064	0.0065	0.0151	0.0086	0.0059	0.0074	0.0060	0.0093
9. Horse	0.4309	0.4295	0.4335	0.3005	0.2970	0.2962	0.2926	0.0716		0.0062	0.0059	0.0141	0.0084	0.0056	0.0070	0.0062	0.0089
10. Sheep	0.3982	0.3932	0.3997	0.2669	0.1573	0.1577	0.1558	0.2960	0.2963		0.0029	0.0159	0.0073	0.0045	0.0062	0.0051	0.0078
11. Goat	0.3947	0.3938	0.3963	0.2646	0.1580	0.1584	0.1537	0.2974	0.2960	0.1049		0.0155	0.0073	0.0046	0.0060	0.0058	0.0078
12. Pig	0.6268	0.6271	0.6216	0.5978	0.5878	0.5938	0.5923	0.6428	0.6368	0.6013	0.5974		0.0143	0.0154	0.0133	0.0156	0.0152
13. Chicken	0.4070	0.4027	0.4053	0.3284	0.3419	0.3421	0.3435	0.3686	0.3634	0.3356	0.3315	0.6158		0.0079	0.0065	0.0087	0.0082
14. Rabbit	0.4004	0.3965	0.3979	0.2618	0.2618	0.2541	0.2560	0.2986	0.3007	0.2563	0.2550	0.5800	0.3474		0.0057	0.0064	0.0080
15. Dog	0.4257	0.4157	0.4277	0.2789	0.3035	0.3082	0.2964	0.3370	0.3382	0.2982	0.3039	0.6278	0.3538	0.2996		0.0055	0.0090
16. D. Cat	0.4118	0.4103	0.4157	0.2645	0.2916	0.2908	0.2905	0.3209	0.3187	0.2943	0.2929	0.6104	0.3410	0.2773	0.2576		0.0083
17. H. Mouse	0.4684	0.4644	0.4688	0.3963	0.3952	0.4034	0.3944	0.4344	0.4299	0.3992	0.3906	0.6500	0.4029	0.3935	0.4144	0.4068	

Demonstration of the nucleotide substitution evolutionary distance in the lower-left matrix, and the standard error which in the upper-right matrix.

In the middle diagonally shown empty boxes that related compare with the organism itself equal null.

#### **4.1.8. Amino acid substitution evolutionary distance**

The number of amino acid differences estimated with per sequence from other sequences are shown, in Table 4.18., 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 were 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 Table 4.1.8., came in the same way with the substitution of nucleotide, evolutionary distance in alignment of 17 sequences of mtDNA. Furthermore, the chicken as a bird is less divergence than pig, dog, cat and mouse in compare with human. Similar results are observed between chicken, rabbit, dog, cat and mouse with pig.

Table 4.1.8. The substitution of amino acids evolutionary distance in alignment of 17 sequences of mtDNA.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1.Human		22.6	22.4	29.4	29.0	30.0	28.4	30.7	31.5	25.5	28.4	32.8	32.7	29.4	28.0	28.3	27.4
2.Chimpanzee	523.0		23.3	31.2	29.9	28.2	28.8	28.0	29.5	26.6	25.8	28.2	30.5	28.7	25.7	27.0	28.8
3.Gorilla	620.0	617.0		33.3	30.9	28.2	30.7	27.1	28.2	26.3	26.7	30.3	31.4	30.7	27.6	29.3	29.3
4.Cattle	1484.0	1455.0	1466.0		28.1	28.2	25.2	26.4	26.4	26.0	27.4	31.2	27.9	30.2	25.7	25.7	27.6
5.W.Buffalo	1532.0	1502.0	1522.0	1157.0		23.9	20.0	29.9	28.8	25.1	25.9	31.1	27.7	28.8	25.0	28.5	26.2
6.Bison	1500.0	1500.0	1512.0	1167.0	679.0		24.0	31.1	30.1	28.5	25.2	33.0	28.3	31.1	27.7	28.8	29.1
7.A. Camel	1537.0	1523.0	1525.0	1190.0	356.0	674.0		28.7	28.9	24.3	24.4	32.5	26.8	26.7	24.9	29.2	26.6
8.B. Camel	1584.0	1549.0	1545.0	1246.0	1276.0	1241.0	1265.0		19.8	31.2	29.7	30.6	29.5	29.1	26.2	24.9	29.2
9.Horse	1572.0	1553.0	1565.0	1256.0	1290.0	1230.0	1261.0	450.0		30.1	29.2	30.0	27.1	26.9	25.5	26.3	29.0
10.Sheep	1517.0	1489.0	1524.0	1184.0	789.0	799.0	817.0	1256.0	1252.0		24.5	32.3	30.9	28.0	26.0	25.4	27.9
11.Goat	1502.0	1495.0	1516.0	1193.0	822.0	804.0	810.0	1273.0	1261.0	596.0		31.2	28.9	28.6	23.7	23.8	30.1
12.Pig	1908.0	1909.0	1903.0	1820.0	1857.0	1842.0	1872.0	1931.0	1920.0	1868.0	1885.0		29.7	33.1	29.0	28.8	31.2
13.Chicken	1540.0	1542.0	1544.0	1341.0	1386.0	1353.0	1398.0	1450.0	1437.0	1352.0	1349.0	1873.0		31.0	26.6	26.0	29.1
14.Rabbit	1507.0	1491.0	1496.0	1142.0	1164.0	1132.0	1135.0	1249.0	1245.0	1135.0	1135.0	1818.0	1355.0		27.7	28.4	27.7
15.Dog	1559.0	1517.0	1558.0	1181.0	1234.0	1248.0	1222.0	1344.0	1336.0	1226.0	1236.0	1872.0	1354.0	1234.0		28.7	26.7
16.D. Cat	1494.0	1524.0	1521.0	1108.0	1230.0	1235.0	1233.0	1314.0	1300.0	1254.0	1252.0	1862.0	1372.0	1180.0	1119.0		30.3
17.H. Mouse	1625.0	1625.0	1633.0	1483.0	1519.0	1509.0	1509.0	1570.0	1551.0	1487.0	1454.0	1920.0	1488.0	1444.0	1513.0	1518.0	

Demonstration of the amino acids substitution evolutionary distance in the lower-left matrix, and the standard error which in the upper-right matrix. In the middle diagonally shown empty boxes that related compare with the organism itself equal null.

#### **4.1.9. 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 of protein sequences that may be cause difference in annotation or function in the genome of mitochondrial DNA. The number of synonymous differences per sequence from shown, sequences involved by all positions containing gaps and missing data were eliminated. Total of 4091 positions in the final dataset.

The results in Table 4.1.9, demonstrate the effect of substitution levels on the frequencies of codon changes synonymously or nonsynonymously. Firstly, the results between human-chimpanzee pair was 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 pig with all other organisms involved in this study Results in the Tables 5, 6 and 7 respectively, explained and illustrated with results, the codons also change show the harmony of the same rhythm with nucleotides and amino acids results, which provide same protein in other sequence. In fact, it was a shock, if 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 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.

Table 4.1.9. Synonymous/non-synonymous codon substitution evolutionary distance in alignment of 17 sequences of mtDNA.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Human		18.84	22.79	26.91	27.26	28.88	26.10	28.65	27.13	32.47	28.05	23.65	29.20	28.04	28.21	26.54	28.04
2. Chimpanzee	445.33		21.75	28.59	27.16	27.81	26.57	29.38	28.79	30.82	28.71	26.28	29.51	28.50	27.63	29.45	29.25
3. Gorilla	563.00	514.00		28.30	30.17	29.51	29.41	29.50	28.77	31.18	27.30	24.66	28.64	27.14	28.45	29.90	27.10
4. Cattle	1086.58	1065.92	1108.92		23.36	26.56	23.23	26.78	27.93	23.92	23.06	27.82	27.27	26.86	25.42	25.35	24.42
5. W. Buffalo	1125.92	1113.83	1155.92	942.33		26.71	17.95	23.75	26.08	22.88	20.76	29.11	25.58	26.42	24.78	24.23	27.97
6. Bison	1109.42	1118.17	1142.25	952.33	650.67		23.89	26.31	27.93	22.09	23.57	29.55	28.61	25.82	26.59	25.00	31.06
7. A. Camel	1125.67	1119.42	1122.83	930.67	310.50	605.50		25.93	25.15	22.09	20.84	26.83	25.06	23.34	22.58	24.03	29.93
8. B. Camel	1186.33	1208.67	1211.58	1044.17	991.75	1013.50	994.50		16.89	26.76	27.06	28.56	29.08	29.27	26.15	26.97	31.65
9. Horse	1186.00	1201.67	1199.92	1037.00	998.42	1034.33	996.17	349.50		27.22	27.60	26.99	27.98	28.34	25.90	27.34	29.72
10. Sheep	1164.33	1151.00	1171.75	951.58	720.50	735.83	690.50	1003.67	1016.42		24.15	29.58	27.98	25.23	26.27	24.96	28.65
11. Goat	1147.58	1135.33	1138.17	919.00	698.67	718.17	689.17	1005.67	1007.92	525.83		29.16	27.49	24.71	25.70	21.21	28.60
12. Pig	1326.83	1339.08	1319.50	1309.67	1265.08	1299.75	1283.75	1336.67	1331.17	1322.92	1292.00		27.83	27.00	25.80	25.02	27.15
13. Chicken	1139.50	1113.25	1130.17	1047.33	1043.00	1077.33	1039.83	1075.00	1061.67	1053.83	1046.50	1325.75		27.30	25.79	24.35	25.35
14. Rabbit	1136.83	1128.25	1137.58	939.25	928.08	918.67	923.25	1036.08	1048.00	930.33	901.17	1268.17	1078.17		27.33	27.27	26.41
15. Dog	1227.42	1207.83	1243.25	990.08	1044.67	1084.00	1014.75	1103.50	1112.83	1054.42	1075.08	1386.33	1124.67	1037.00		25.97	26.63
16. D. Cat	1180.17	1153.00	1178.75	983.17	1016.92	1021.00	1015.92	1085.00	1086.83	1022.92	1014.33	1346.42	1055.50	1003.42	986.25		25.80
17. H. Mouse	1219.58	1206.17	1236.67	1102.33	1051.25	1093.08	1052.25	1145.58	1155.83	1106.33	1088.42	1319.50	1111.08	1120.08	1133.00	1123.83	

Demonstration of the Synonymous/non-synonymous codon substitution evolutionary distance in the lower-left matrix, and the standard error which

in the upper-right matrix. In the middle diagonally shown empty boxes that related compare with the organism itself equal null



regardless of the discussion was mentioned within the results, but it is worth to Be more highlighted. 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 human. Moreover , this study generate a motivation to study the phylogenetic in deep and the *de novo* annotation looking for some confised answers that help to discover and understand more about mitochondrial DNA.



**4.2. PHYLOGENETIC TREE CONSTRUCTION WITHIN MAXIMUM  
LIKELIHOOD METHOD OF COMPLETE GENOMES OF  
MITOCHONDRIAL DNA BETWEEN HUMAN'S AND 16 ANIMALS**

#### **4.2.1. Phylogenetic Tree Construction by Maximum likelihood method of 17 nucleotide sequences.**

The evolutionary history was inferred by using the Maximum Likelihood method based on the General Time Reversible model. The tree with the highest log likelihood (-129985.7031) is found by applying mathematical solutions to find the evolutionary factor. The percentage of trees in which the associated taxa clustered together is shown next to the branches the analysis involved 17 nucleotide sequences. Codon positions included were 1<sup>st</sup> +2<sup>nd</sup> +3<sup>rd</sup> +N noncoding. All positions containing gaps and missing data were eliminated. There were a total of 14430 positions in the final dataset.

Figure 4.2.1., demonstrated a rooted phylogenetic tree and demonstrate paraphyletic group of mitochondrial DNA sequences. human's sequence is the head base of the comparative. Results put the species in interest in two main monophyletic groups. The Glades by sharing the common ancestral point for each group except chicken is the out group of the tree. First group, shows the human and chimpanzee are descendants of gorilla and split in two different evaluated organisms. Second group, is the major and more complex, by observing the tree chronologically starting from the internal points to the terminal points can note cattle and bison share ancestor node and they are descendants of the water buffalo, also the dog and cat are descendants from horse.

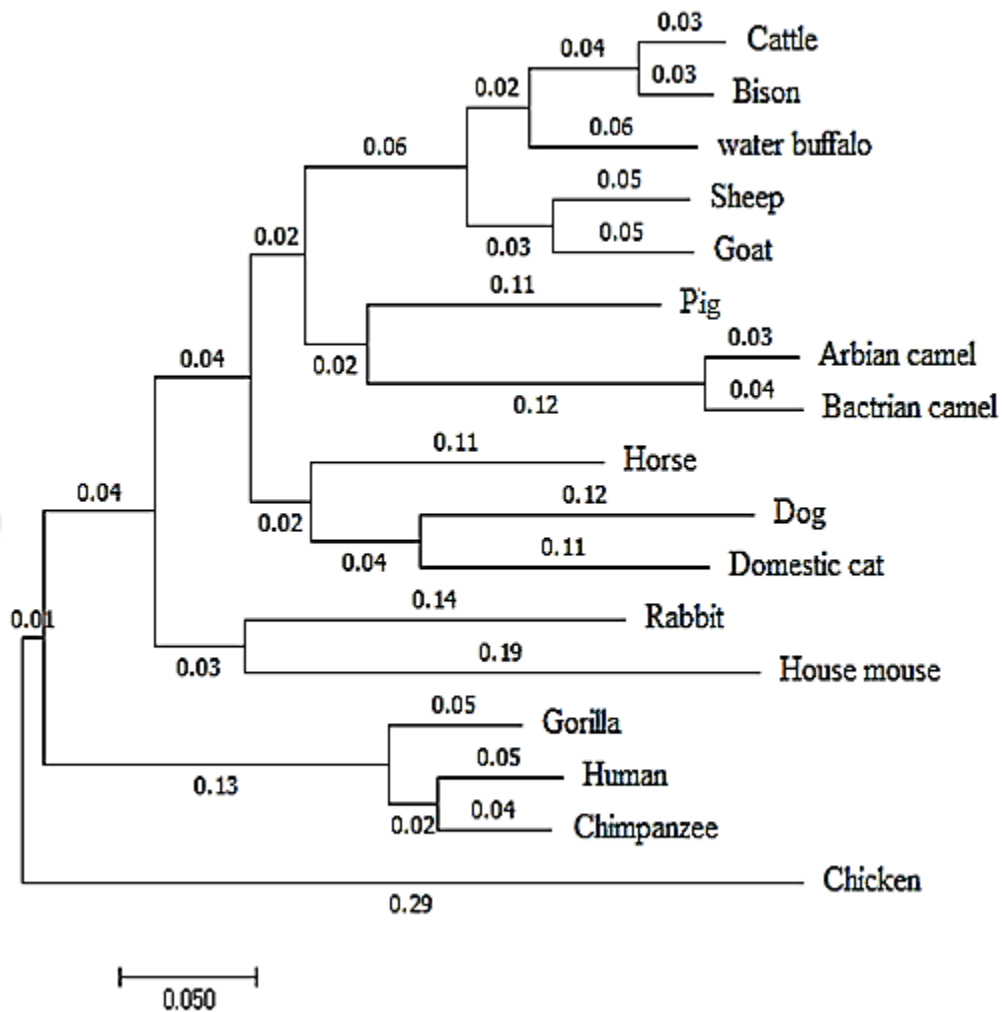


Figure 4.2.1. The phylogenetic tree for the DNA by Maximum likelihood method with branch lengths

Moreover, in Figure 4.2.1. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site next to the branches. Showing that human and chimpanzee are descended from gorilla, also dogs and cats are descendants of horse. In other words, the chicken shows as the ancestral species of all additionally, the human's group is considered as outgrouped.

As well as, figure 4.2.2. which observed the phylogenetic tree for the DNA by Maximum likelihood method with ancestral states. The major group of species that share the letter (A) starting from the chicken. Also the cattle, bison, goat and Arabian camel are the newest and developed in the evolutionary progress.

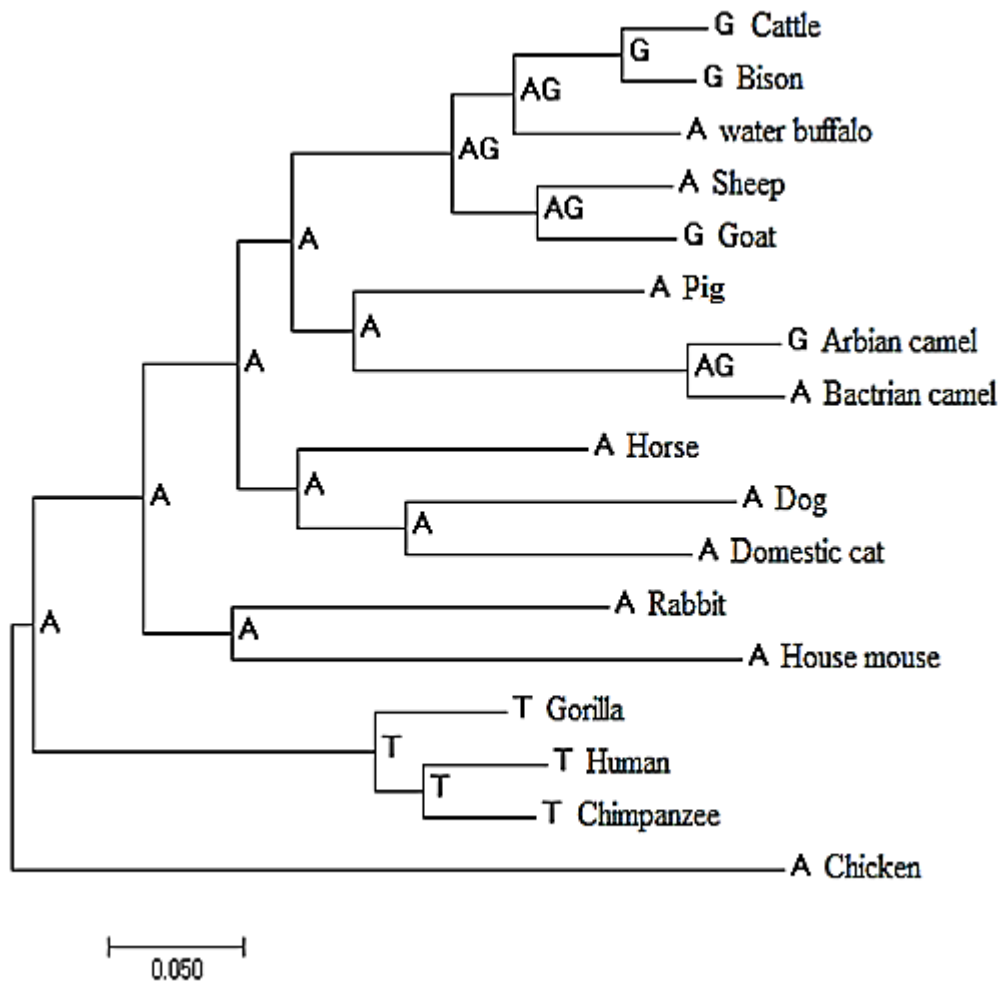


Figure 4.2.2. The phylogenetic tree for the DNA by Maximum likelihood method with ancestral states.

Most importantly, the results that came in the figure 4.2.3. The timetree shown was generated using the Real-time method. Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method based on the General Time Reversible model. The estimated log likelihood value of the topology shown is -131036.4810. The tree is drawn to scale, with branch lengths measured in the relative number of substitutions per site by next to the branches in timetree results of the 17 organism to explain and prove the previous results in figures 4.2.1. and 4.2.2. how the human and chimpanzee are outgrouped

became descendants of gorilla. Additionally, the chicken is the ancestor of the rest of species which included in this study.

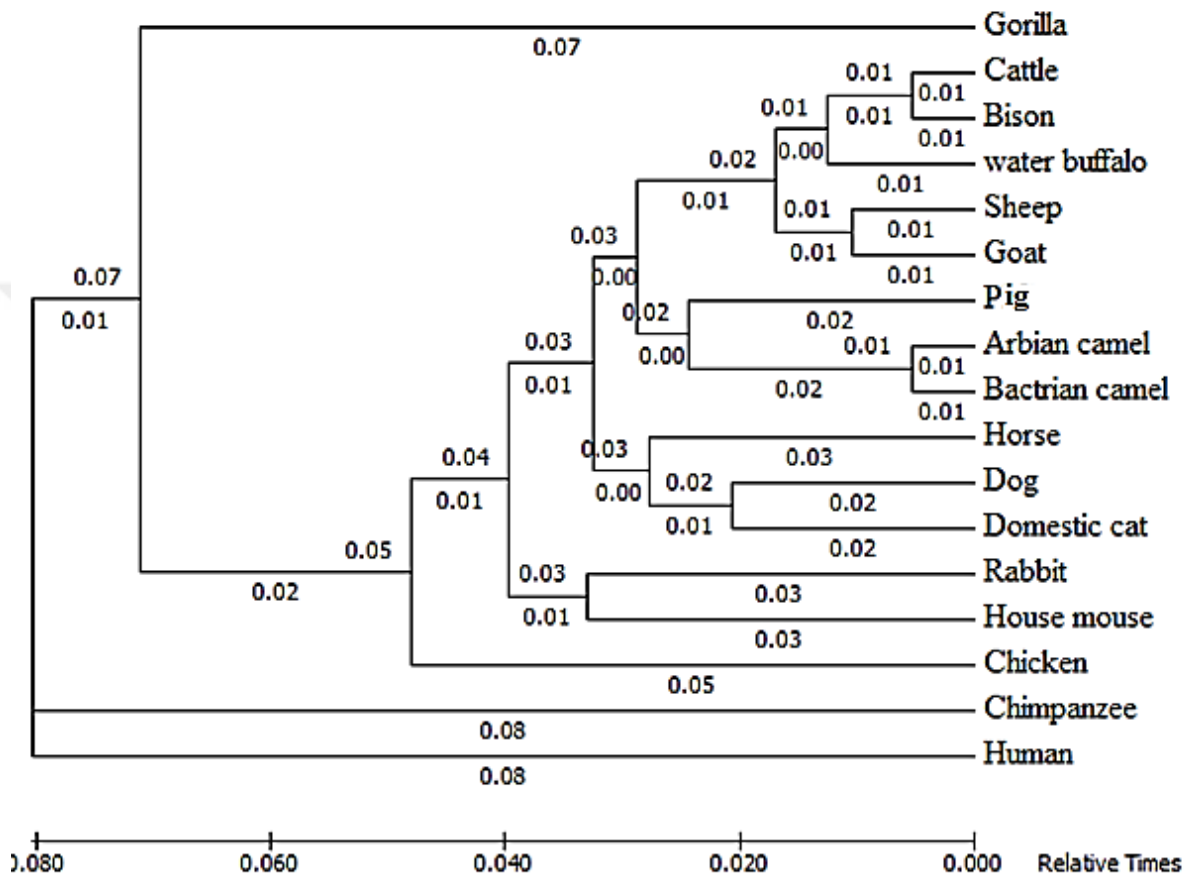


Figure 4.2.3. timetree phylogenetical analysis of DNA by Maximum Likelihood method.

#### **4.2.2. Phylogenetic Tree Construction by Maximum likelihood method of 17 amino acid sequences.**

About the rooted phylogenetic tree depended on the amino acids sequences through 17 species show slightly different results in evolutionary progress as shown in figure 4.2.4., the evolutionary history was inferred by using the Maximum Likelihood method based on the Equal Input model. The tree with the highest log likelihood (-75003.8228) is shown as time within evolution. Initial tree for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site next to the branches. The analysis involved 17 amino acid sequences. The coding data was translated assuming a Vertebrate Mitochondrial genetic code table. All positions containing gaps and missing data were eliminated. There were a total of 4005 positions in the final dataset. Actually demonstrate the pure evolution man between proteom sequences. Observation of the whole tree devided in two main common ancestral points *F* and *L* , human with chipanzee and gorila share the *F* node with same relation but unlike the previous figure chicken involved in the major monogroup by sharing the *L* node and shows relation with horse and pig in protien evolution.

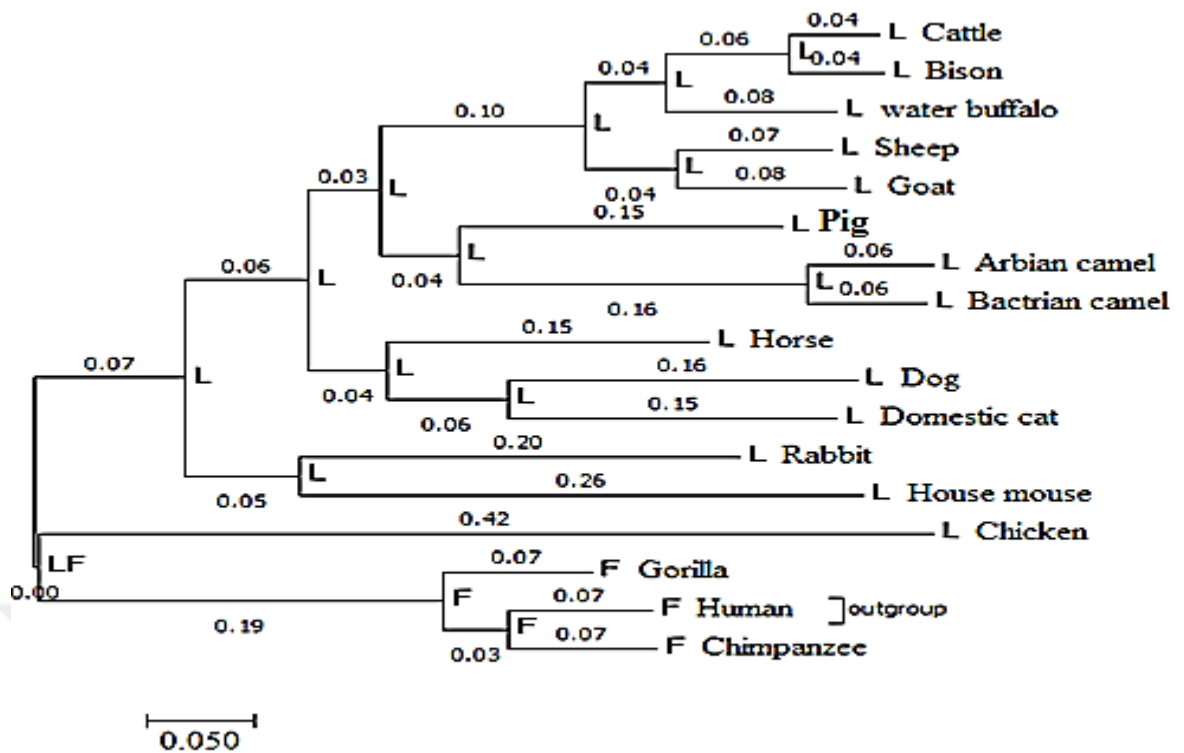


Figure 4.2.4. The phylogenetic tree for the amino acid sequences by Maximum likelihood method with branch lengths and ancestral states.

Moreover, as in Figure 4.2.5., the timetree shown was generated using the Real Time method. Divergence times for all branching points in the topology were calculated using the Maximum Likelihood method based on the Equal Input model. The estimated log likelihood value of the topology shown is -75003.8228. The tree is drawn to scale, with branch lengths measured in the relative number of substitutions per site of next to the branches. The analysis involved 17 amino acid sequences. The coding data was translated assuming a Vertebrate Mitochondrial genetic code table. Also all positions containing gaps and missing data were eliminated within total of 4005 positions in the final dataset.



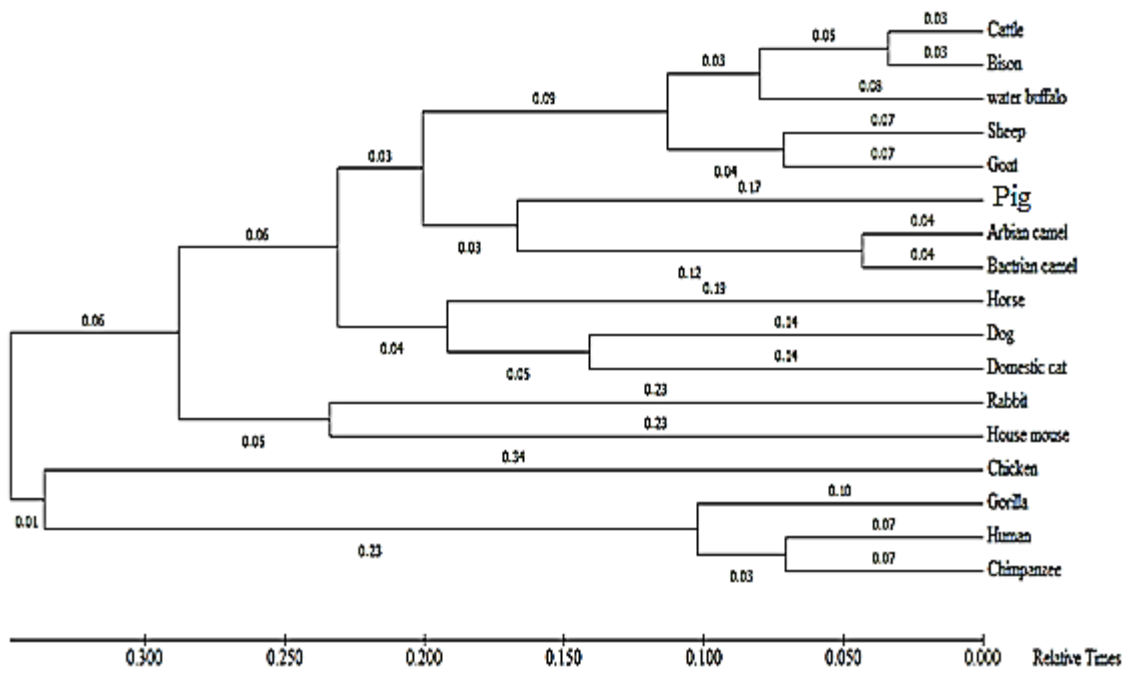


Figure 4.2.5. Timetree phylogenetical analysis of amino acid sequences by Maximum Likelihood method.

**4.3. NUCLEOTIDE AND AMINO ACID SEQUENCES ANALYSIS IN  
PATHOGEN AND HOST OF BRUCELLOSIS IN CATTLE**

#### **4.3.1. Comparative view in DNA sequences between *Brucella abortus* and *Brucella melitensis*.**

This study concentrated on the host of disease in the first place and pathogen came in second by using the database of nucleotide and amino acid sequences of *Brucella abortus* and *Brucella melitensis* for scanning the similarity and matches between sequences of amino acids which had produced from bovine's as some antibodies with the proteins of the pathogen species that observed the brucellosis in cattle.

The main Statistical nucleotide calculations of *Brucella abortus* and *Brucella melitensis* could give a general picture of the two species of bacteria are cause the bovine brucellosis. (Table 4.3.1.):

Table 4.3.1. Statistical nucleotide calculations of *Brucella abortus* and *Brucella melitensis*

\*MDa: mega Dalton, the unit of molecular mass weight MDa = (1,000,000 Da).

Information	<i>Brucella abortus</i>		<i>Brucella melitensis</i>	
	**Ch. I	***Ch. II	**Ch. I	***Ch. II
Genome size (bp)	1,156,948	2,107,358	2,117,144	1,177,787
*Weight (single-stranded) MDa	357.571	651.07	654.232	363.964
*Weight (double-stranded) MDa	714.955	1,302.273	1,308.32	727.833
Counts of nucleotides				
Adenine (A)	245,612	452,079	452,846	251,795
Cytosine (C)	329,101	603,950	603,116	336,601
Guanine (G)	334,302	600,369	607,001	338,717
Thymine (T)	247,933	450,960	454,173	250,672
Frequencies of nucleotides				
Adenine (A)	0.212	0.215	0.214	0.214
Cytosine (C)	0.284	0.287	0.285	0.286
Guanine (G)	0.289	0.285	0.287	0.288
Thymine (T)	0.214	0.214	0.215	0.213

\*\*Ch. I: Chromosome number one.

\*\*\*Ch. II: Chromosome number two.

So, the results show slightly different in genome size, molecular weight, counts of nucleotides and the frequencies of nucleotides. That means the highly similarity in DNA statistics is lead to the similar nucleotide behavior in central dogma process when they produce more than 3000 proteins for each in condition of similarity of sequences which going to be proved in figure 4.3.1. and 4.3.2.

As well as, the results of global alignment between *Brucella abortus* and *Brucella melitensis* in chromosome I and II. Figure 4.3.1. demonstrates the highest probable similarity because of the middle diagonal line shows the highest score could give within 99% identities

score presents. In DNA sequences by dot matrix view by showing regions of similarity based upon the BLAST results. The query sequence of the chromosome I of *Brucella abortus* is represented on the X-axis and the numbers represent the bases/residues of the query. Also, the chromosome I of *Brucella melitensis* represented on the Y-axis and again the numbers represent the bases/residues of the subject. Alignments are shown in the plot as lines. Moreover, strand and protein matches are slanted from the bottom left to the upper right corner, minus strand matches are slanted from the upper left to the lower right.

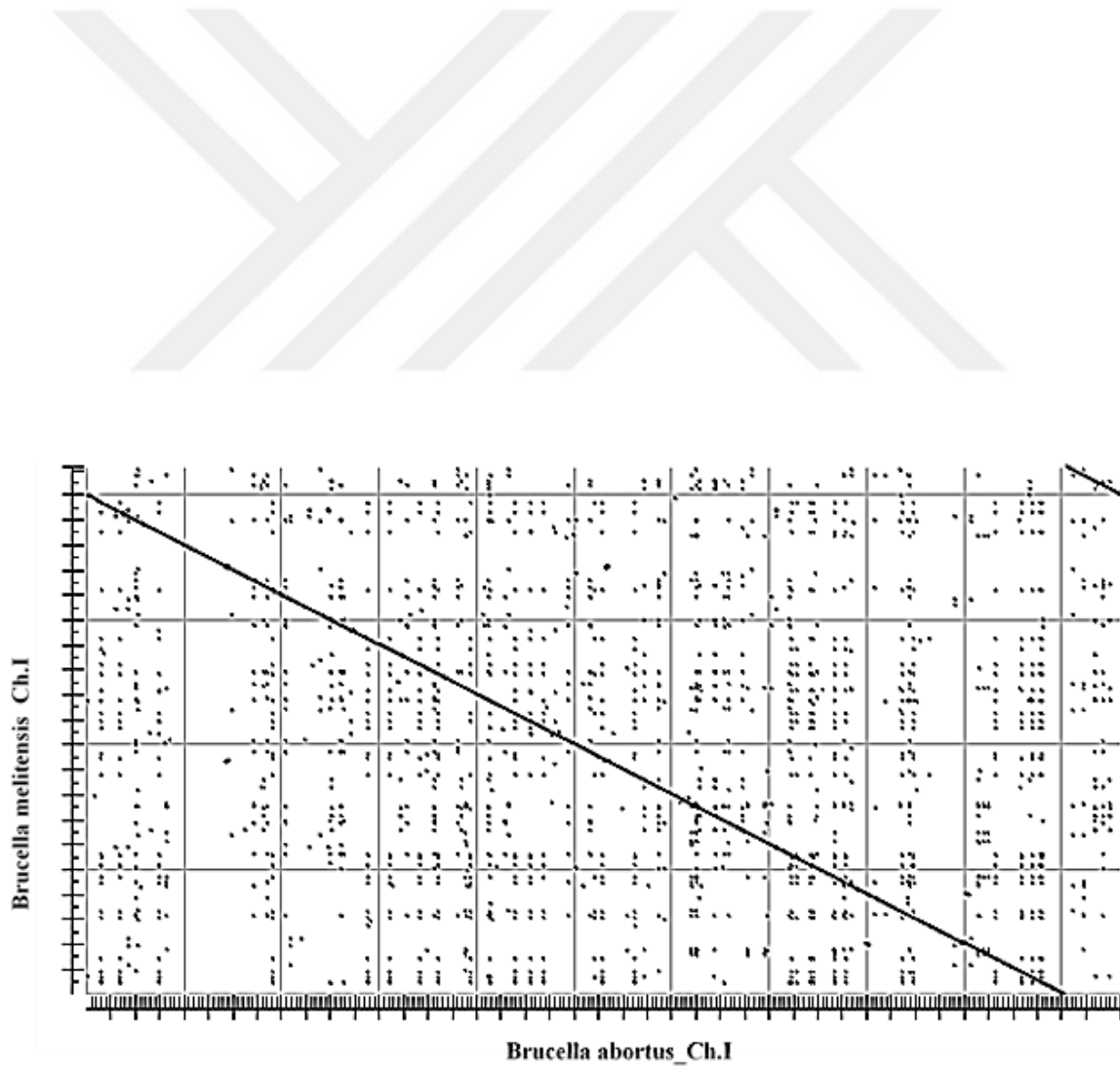


Figure 4.3.1. Dot plot matrix view of global alignment in chromosome I between *Brucella abortus* and *Brucella melitensis*. The number of lines shown in the plot is the same as the number of alignments found by BLAST.

Equally important, Figure 4.3.2., demonstrate the slightly lower similarity because of the middle diagonal line shows the highest score could give within identities score presents 95%. In DNA sequences by dot matrix view by showing regions of similarity based upon the BLAST results. The query sequence of the chromosome II of *Brucella abortus* is represented on the X-axis and the numbers represent the bases/residues of the query. Also, the chromosome II of *Brucella melitensis* represented on the Y-axis and again the numbers represent the bases/residues of the subject. Alignments are shown in the plot as lines. Moreover, strand and protein matches are slanted from the bottom left to the upper right corner, minus strand matches are slanted from the upper left to the lower right.

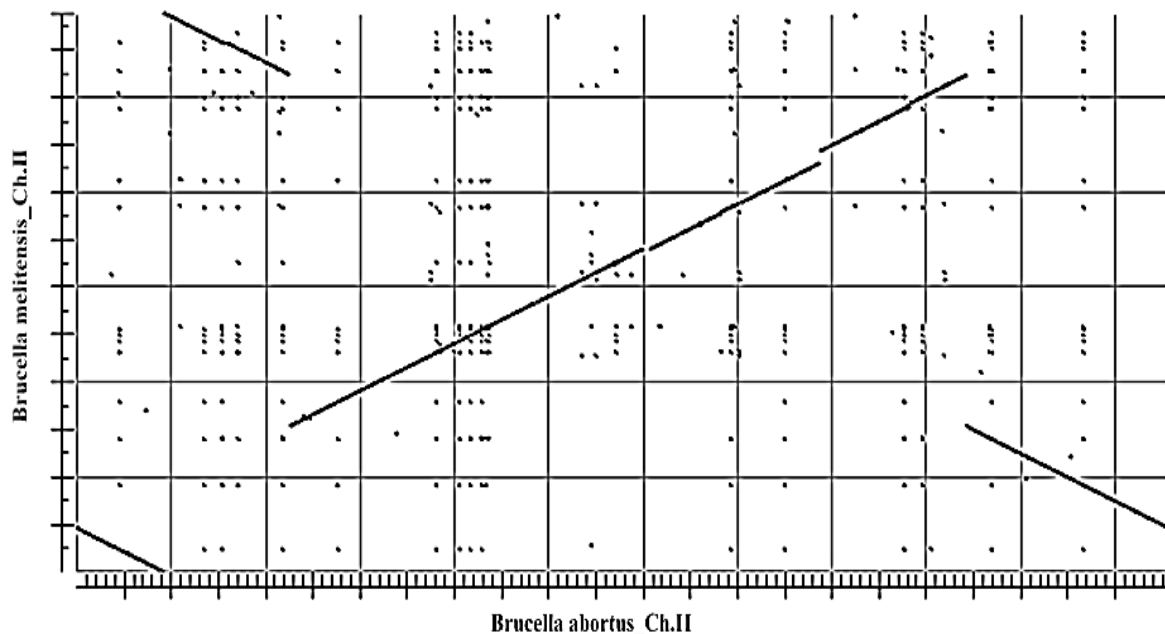


Figure 4.3.2. Dot plot matrix view of global alignment in chromosome II between *Brucella abortus* and *Brucella melitensis*. The number of lines shown in the plot is the same as the number of alignments found by BLAST.

### 4.3.2. Nucleotide Sequence analysis of SLC11A1 gene in cattle

The analysis of DNA sequences of SLC11A1 gene in cattle that showed resistance to *brucellosis* disease started by comparative the Statistical calculation of the SLC11A1 gene that appear in six cattle stairs.

Table 4.3.2. is demonstrated the major various in gene size from 10,665 to 13,543 nucleotide bases to produce the same Natural resistance-associated macrophage protein which is known as NRAMP protein of the antibody in b-cells to discover and destroy the antigens of *brucellosis* in the host body, what will be talking about it in depth at the seventh and also it is the last chapter of this thesis.

May be these results telling about the unnecessary DNA strings inserted through the evolution process by mistakes in replication process or any other recombinant DNA meanwhile the time, because the DNA meanwhile the time, because the source of SLC11A1 gene representing by the accession number Q493965 with the lowest gene size 10,665 bases and the least counts of the nucleotides adenine (A) 1,422, cytosine (C) 1,925, guanine (G) 1,848 and thymine (T) 1,402 that really involved in the transcriptional and translating process to produce NRAMP protein in spite of missing some fragments in the original sequence.

Table 4.3.2. Statistical calculation comparative of the SLC11A1 gene that appear in six cattle stairs that resist to *brucellosis*.

\*DQ493965: have missing fragments of un-known sequence in the database thus the counts

Information	AC_000159	KR002419	*DQ493965	KR002421	KR002420	DQ848779
-------------	-----------	----------	-----------	----------	----------	----------

and frequencies based on the available sequence count is 6597 bp instead of 10,665 bp.

Gene sizes (bp)	10,926	13,543	10,665	13,543	13,543	10,814
Weight (single-stranded) kDa	3,380.651	4,188.369	3,293.722	4,188.376	4,188.425	3,347.241
Weight (double-stranded) kDa	6,751.593	8,368.809	6,590.345	8,368.81	8,368.81	6,682.405
Counts of Nucleotides (bp)						
Adenine (A)	2,571	3,174	1,422	3,172	3,173	2,560
Cytosine (C)	2,873	3,632	1,925	3,632	3,631	2,807
Guanine (G)	3,078	3,781	1,848	3,782	3,783	3,068
Thymine (T)	2,404	2,956	1,402	2,957	2,956	2,379
Frequencies of nucleotides						
Adenine (A)	0.235	0.234	0.215	0.234	0.234	0.237
Cytosine (C)	0.263	0.268	0.291	0.268	0.268	0.260
Guanine (G)	0.282	0.279	0.280	0.279	0.279	0.284
Thymine (T)	0.220	0.218	0.212	0.218	0.218	0.220



### 4.3.3. Estimation of codon bias

The comparative in codon bias by estimating the code frequencies in each of the six sources of SLC11A1 gene. Figure 4.3.3. showing a highly similarity in frequencies of all codons usage except the GCA code which is represented the alanine A observed a significant difference in all over the six sources. Additionally, the codes CTA, GCA, GTA and ATA have the highest codon usage.

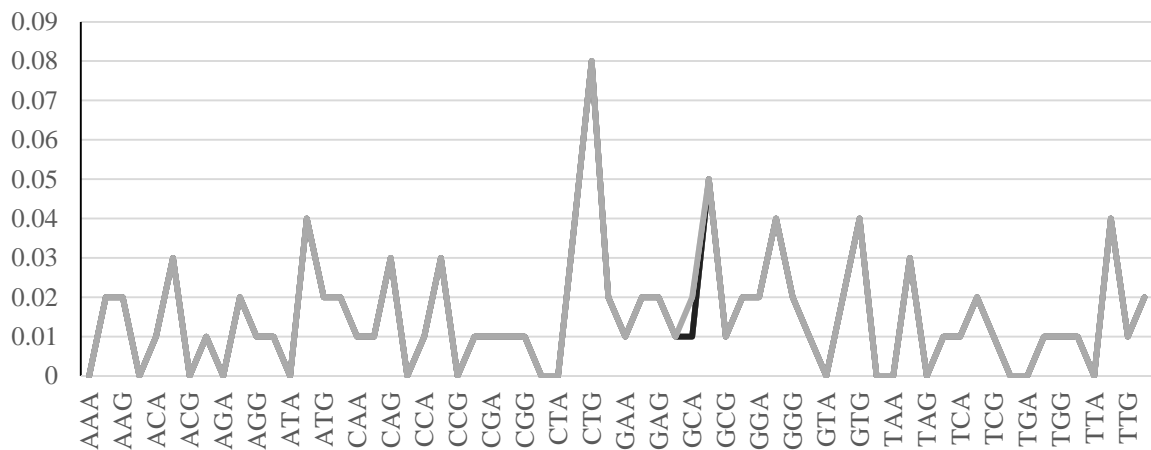


Figure 4.3.3. Estimation of codon bias for six sources of SLC11A1 gene in cattle.

Subsequently, to prove the results in figure 4.3.3. By going in deep of the code through estimate nucleotide frequencies in each position. Figure 4.3.4. the four nucleotides frequencies codon in the first position, having same frequencies for all nucleotides A-1 0.22, C-1 0.28, G-1 0.31 and T-1 0.18.

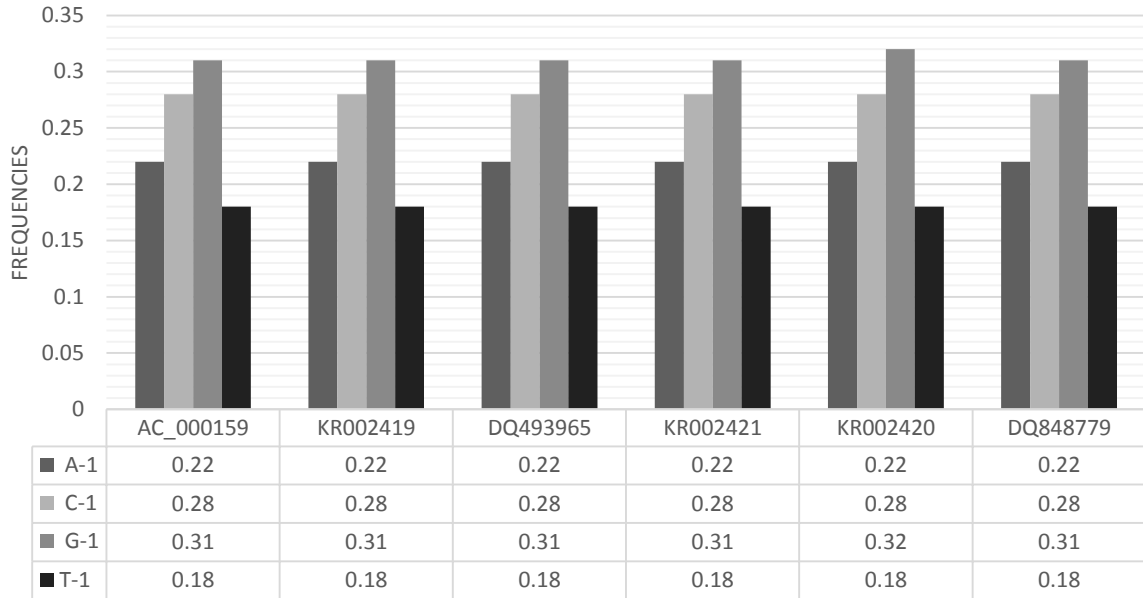


Figure 4.3.4. The four nucleotides frequencies codon in the first position

Moreover, Figure 4.3.4 in the four nucleotides frequencies codon in the second position, having same frequencies for all nucleotides A-2 0.2, C-2 0.25, G-2 0.19 and T-2 0.37.

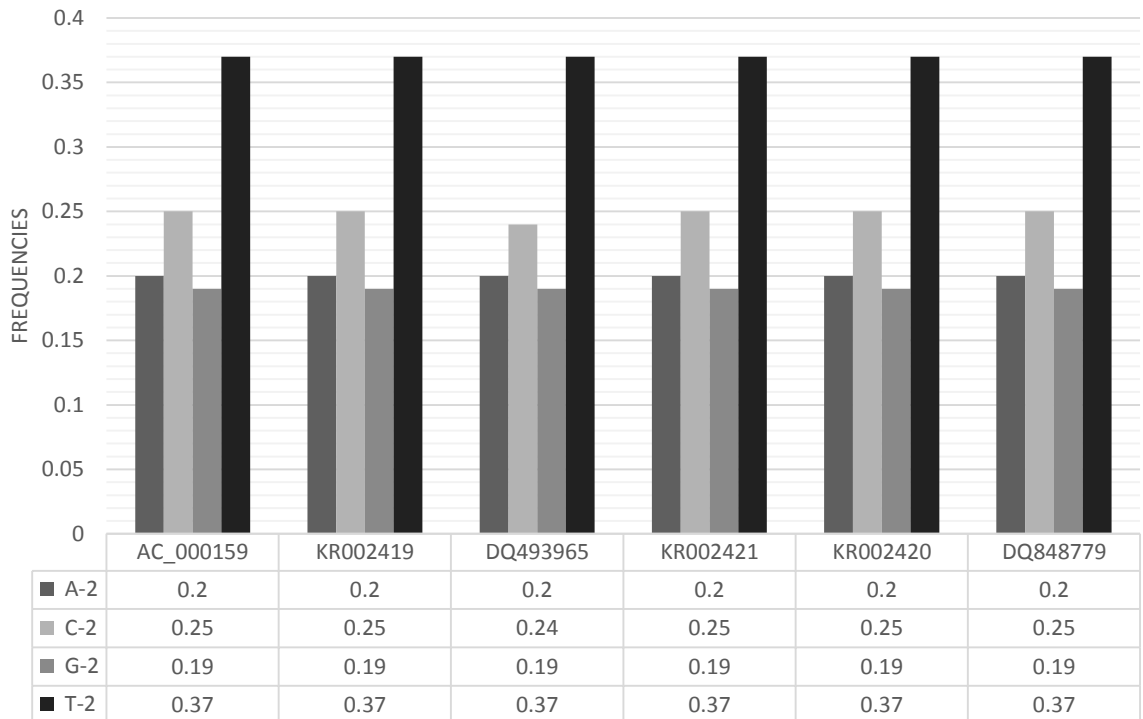


Figure 4.3.5. The four nucleotides frequencies codon in the second position

Finally, Figure 4.3.5. came in same rhythm of the four nucleotides frequencies codon in the third position, having same frequencies for all nucleotides A-3 0.2, C-3 0.25, G-3 0.19 and T-3 0.37.

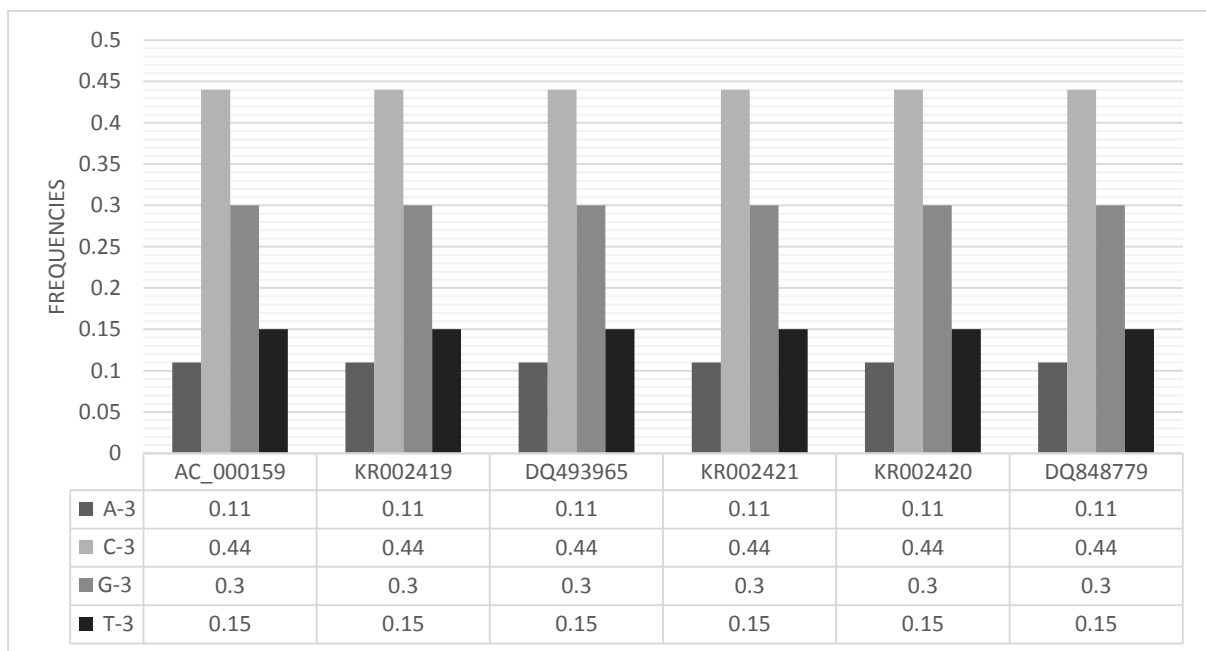


Figure 4.3.6. The four nucleotides frequencies codon in the third position

#### 4.3.4. The phylogenetic tree

The tree in Figure 4.3.7., is drawn to scale, with branch lengths measured in the number of substitutions per site next to the branches for the six sources of the SLC11A1 gene presented the superfamilies' that resisted the *brucellosis* disease by NCBI accession numbers which observed the phylogenetic tree for the DNA by minimum evolutionary method with ancestral states with estimation the mean of branch length equaled 0.001.

As well as, Figure 4.3.7. which show the phylogenetic tree for the DNA by minimum evolutionary method with ancestral states. The ancestral SLC11A1 gene is KR002421 and the other five sources became two main descendant's groups. The nearest nodes relation is KR002419 then KR002420

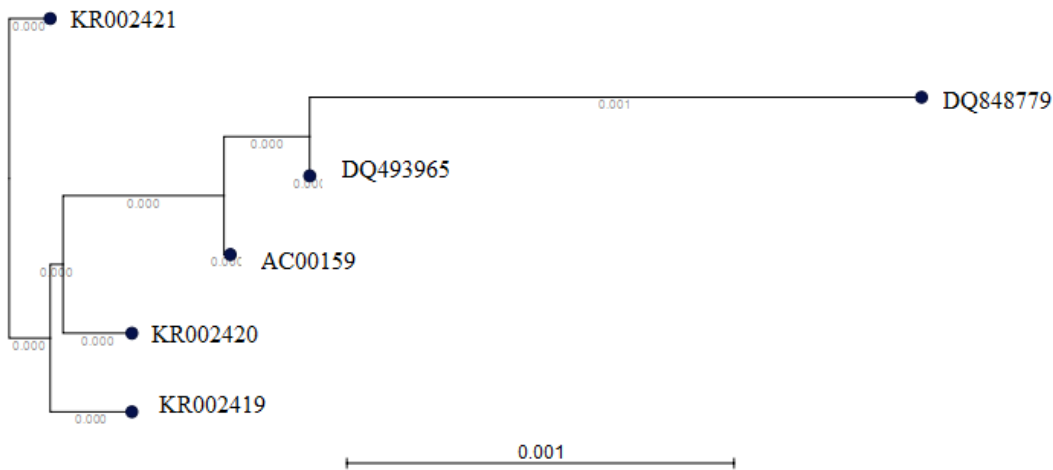
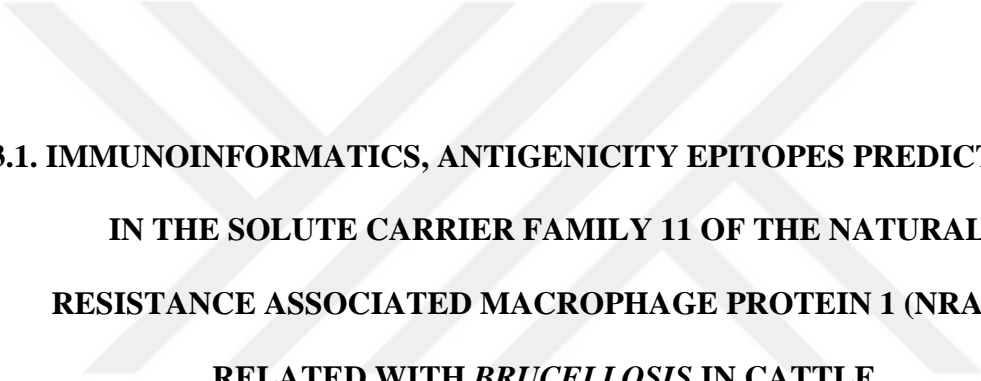


Figure 4.3.7. The phylogenetic tree for the nucleotide of SLC11A1 gene



**3.1. IMMUNOINFORMATICS, ANTIGENICITY EPITOPES PREDICTION  
IN THE SOLUTE CARRIER FAMILY 11 OF THE NATURAL  
RESISTANCE ASSOCIATED MACROPHAGE PROTEIN 1 (NRAMP)  
RELATED WITH *BRUCELLOSIS* IN CATTLE.**

#### 4.4.1. The multiple sequence alignment

The results of multiple sequence alignment by ClustalW wrapper between the six protein sequences which have same amino acid size number (548 a.a.) with substitution in one or two amino acids for each sequence. Provide identity score 542/548 (98.9%), similarity score 547/548 (99.8%) and gaps score 0/548 (0.0%). The reason behind these high scores are the similar number of amino acids that cause no gaps, because the gaps designed to produce same strings length after alignment, also the low value of substitutions to generate these values.

Furthermore, the sequence alignment leads to similarity plot as demonstrated in figure 4.4.1., the plot presents highest Similarity scores between the six proteins sequences for each residue of 10 -> 20 amino acids. Moreover, the highest scores recorded for 2.0 and more of the string positions (120 ->130; 170 -> 180; 220 -> 240; 410 -> 430 and 520 -> 550) also the highest results ever were in position 220 -> 260; in the other hand the lowest similarity that score recorded was 1.0 in the position 210 -> 218 but shown higher than the main of expect.

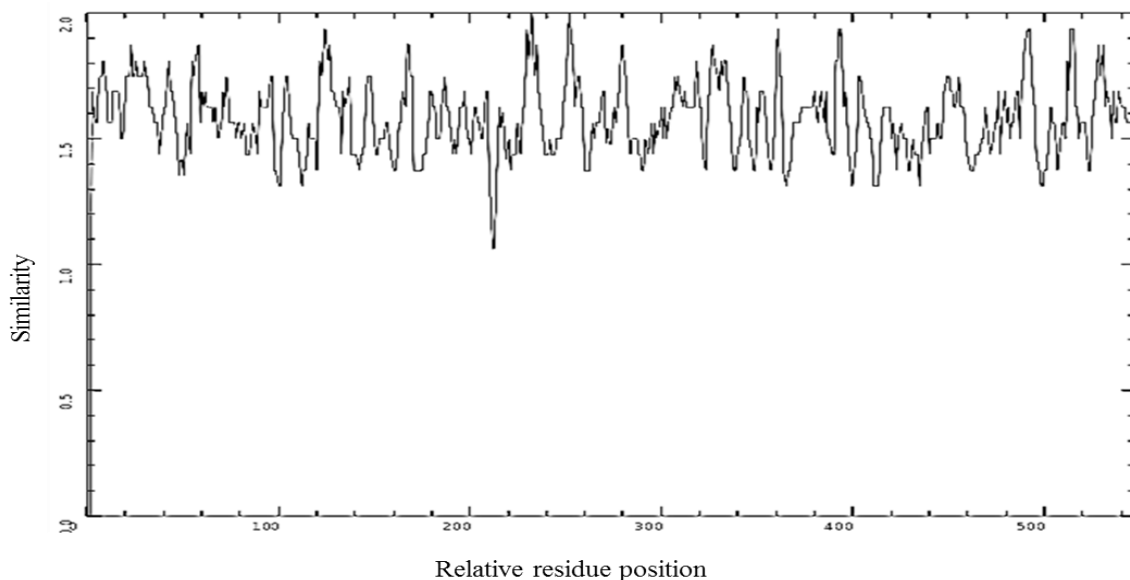


Figure 4.4.1. The plot presents high similarity between the six proteins sequences.

#### 4.4.2. The Hydrophobic and Hydrophilic

The importance of hydrophobic and hydrophilic estimation is related to the two different schools in Immunoinformatics. Firstly, the method of Kolaskar and Tongaonkar in 1990, declared the antigenic sites on proteins has revealed that the hydrophobic residues (Amat-ur-Rasool et al., 2015; Cai et al., 2015; Kolaskar & Tongaonkar, 1990; Sealey, Kirk, Walker, Rollinson, & Lawton, 2013). Secondly, welling method in 1985, assumed that antigenic regions are primarily hydrophilic regions at the surface of the protein molecule (Sun et al., 2002; Welling et al., 1985).

The results of hydrophobic and hydrophilic estimation depend on calculate the amino acids physiochemical properties. Figure 4.4.2., evince he amino acids frequencies distributions, shows the higher summation frequencies score for the amino acids which have Hydrophobic properties (A, F, G, I, L, M, P, V and W) that count 340/548 a.a. and the frequency scored 0.620 and the lower for the amino acids which have hydrophilic (C, N, Q, S, T and Y) count 133/548 a.a. with frequency score 0.243, with the remaining 75/548 a.a. and frequency 0.137 which have not hydrophobic or hydrophilic physiochemical properties.

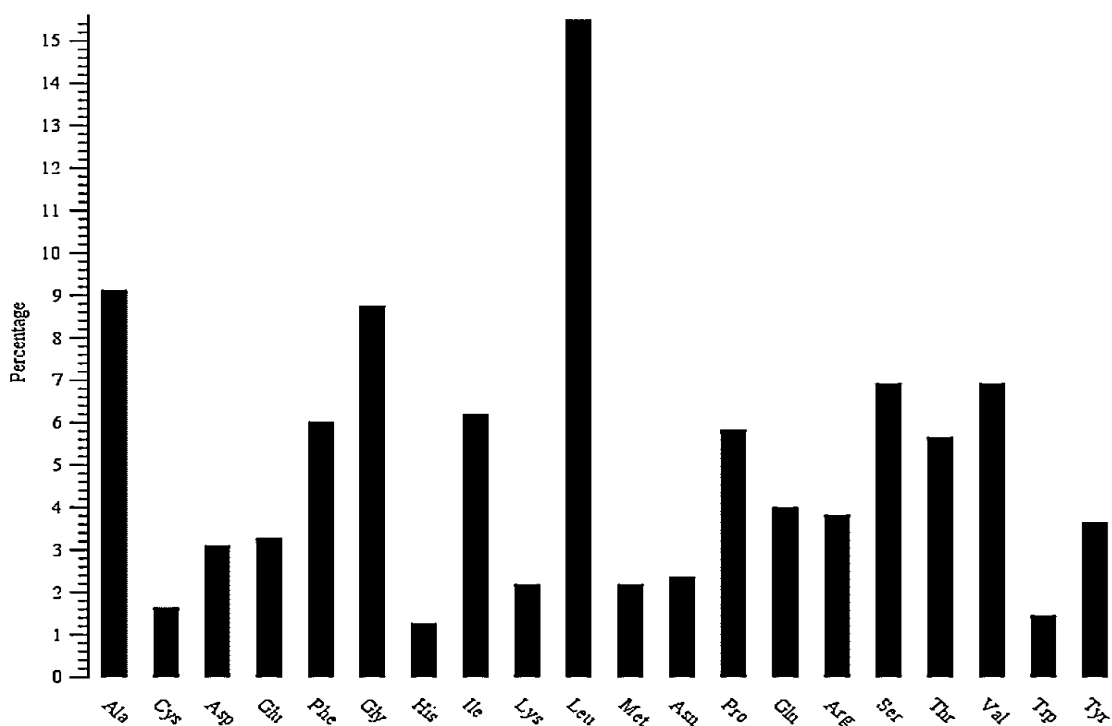


Figure 4.4.2. The amino acids frequencies distribution

#### 4.4.3. The antigenic epitopes binding prediction

Because of the highly similarity of the six protein sequences, the results of the antigenic epitopes binding prediction were coming in same for all over the six sequences. Thus, Table 4.4.1., illustrate the results with avoiding the repetition and indicated in sorted order from higher to lower score, also declare the position of amino acid within maximum score. The sequences of residues represented the peptides of epitopes that recognize and binding with antigen of bacterial pathogen which cause the *Brucellosis* in cattle.



Table 4.4.1. The antigenic epitopes binding prediction

	Score	Length	Maximum score position	Residue	Sequence
1	1.243	79	473	458->536	NGLVSKVITSSIMVLVCAVNLYFVISYLPSPHPAYF SLVALLAAAYLGLTTYLVWTCLITQGATLLAHSSH QRFLYGL
2	1.219	28	124	118->145	LGEVCHLYYPKVPRILLWLTIELAIVGS
3	1.205	37	101	80->116	QAGAVAGFKLLWVLLWATVLGLLCQRLAARLGV VTGK
4	1.202	35	350	335->369	NLTVAVDIYQGGVILGCLFGPPALYIWAVGLLAAG
5	1.201	24	433	424->447	LNDLLNVLQSLLLPFAVLPILTFT
6	1.191	20	173	168->187	WGGVLITVVDTFFFLFDNY
7	1.190	24	412	395->418	FARVLLTRSCAILPTVLLAVFRDL
8	1.168	76	258	189->264	LRKLEAFFGFLITIMALTFGYEYVVAQPAQGALLQ GLFLPSCPGCGPELLQAVGIIGAIIMPHNIYLHSSL VKSR
9	1.164	30	292	279->308	MYFLIEATIALSVSFLINLFVMAVFGQAFY
10	1.126	16	157	150->165	VIGTAIAFSLLSAGRI
11	1.110	16	321	316->331	FNICADSSLHDYAPIF
12	1.092	7	67	64->70	LMSIAFL
13	1.060	7	38	37->43	SEKIPIP
14	1.054	13	21	17->29	SISSPPSPEPQQA
15	1.052	8	385	384->391	MEGFLKLR
16	1.050	7	55	52->58	LRKLWAF

Figure 4.4.3. shows a linearly interpolate between multi-dimensional points display as a comparison between method of Kolaskar and Tongaonkar in 1990 versus welling method in 1985, The antigenicity plot starts with 0.0 score in the middle of vertical line and the score raise in two directions, up positively and down negatively; the positive results indicate the present to be a part of antigenic sites. By applying the converted amino acids sequence to a numerical scores represent the antigenic prediction for each single position within GNU PSPP program to show a significant supremacy and confidence with method of Kolaskar and Tongaonkar (1990) which display in black line which stay in the top score for all residues except one position 260 ->270.

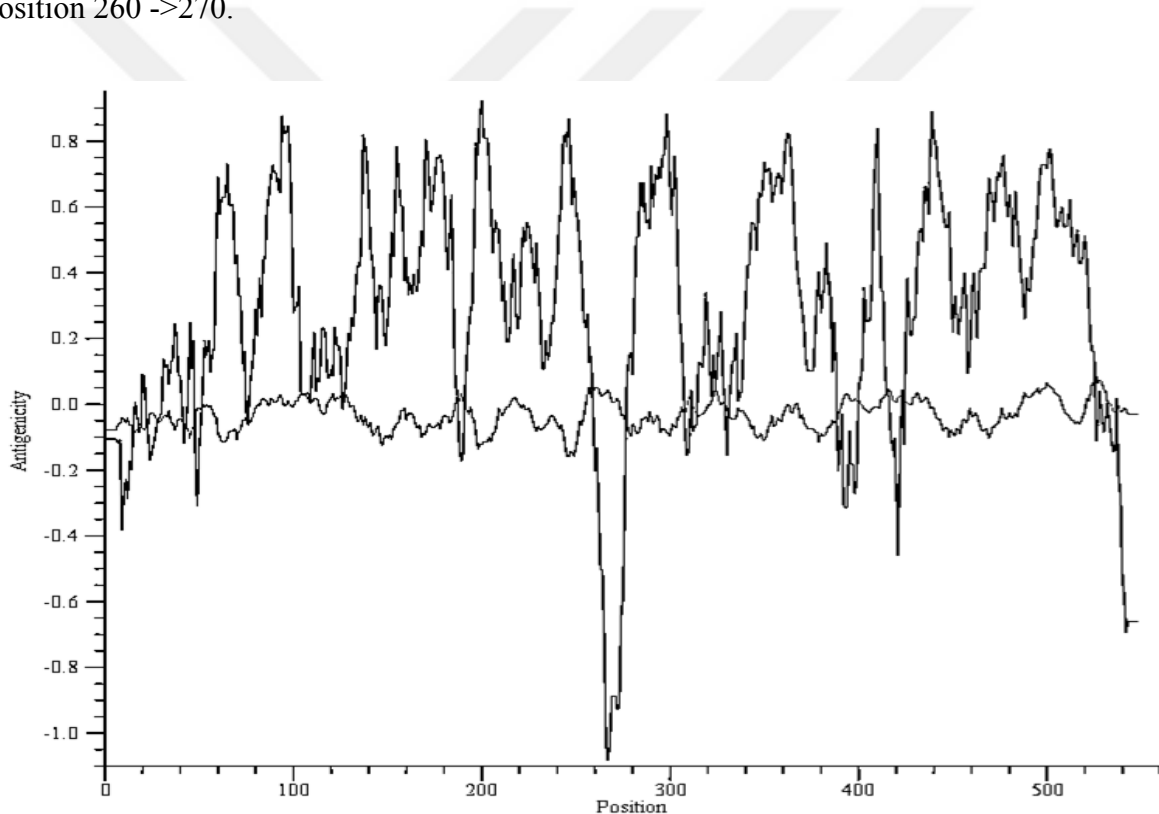


Figure 4.4.3. The antigenicity plot for epitopes binding prediction

\*The narrow curve linear in the middle near 0.0 represent the plot results with Welling-Wester method, (1985).

\*\*The major curve linear the plot results with Kolaskar and Tongaonkar method, (1990).

In the other hand, after searching at the BLAST and IMGT for the similar proteins also the database of immunomics observation through lab experiments, additionally studding the prediction protein secondary structure related with the disorder and confidence regions inside NRAMP protein as demonstrated in Figure 4.4.4., showing the high confidence scour significantly from the positions 1 to 46 and 535 to 548 of the peptide fragments that's represent the epitopes which binding with the antigen.

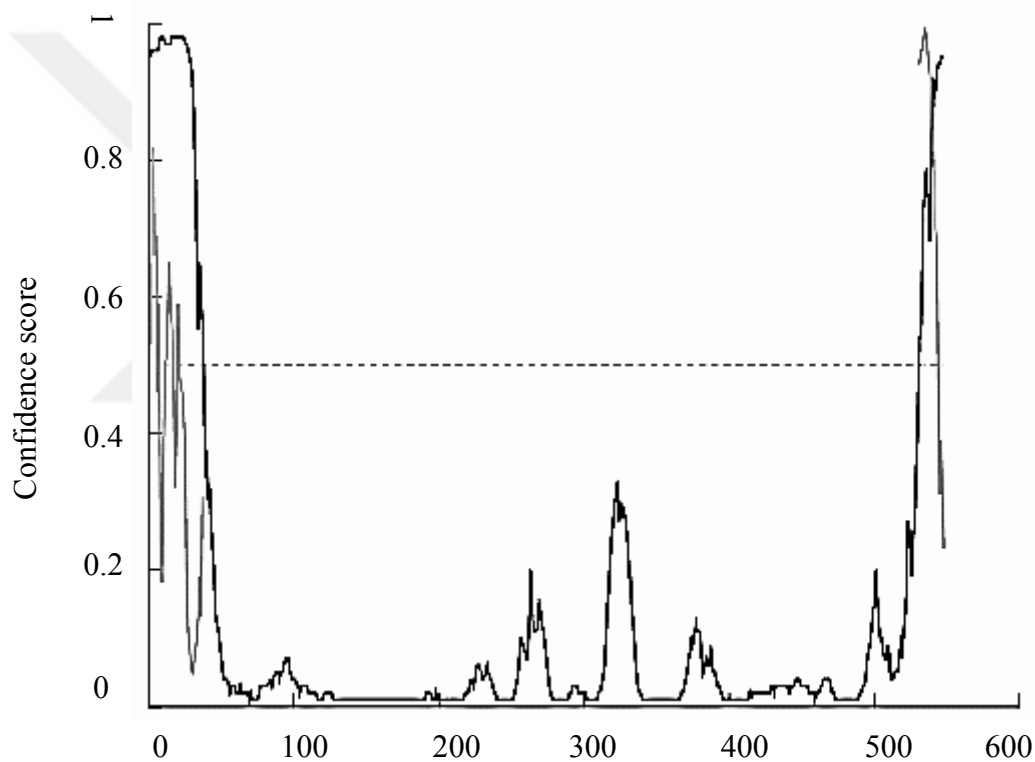


Figure 4.4.4. Predicting confidence score for the binding epitope fragments depending on protein's secondary structure.

The main aim behind prediction of antigenic epitopes is to find the maximum probability of potential peptides residues could recognize and bind with the antigen, which is become very handy to design drugs and looking for increasing the number of animals that have ability to

produce this protein of the natural resistance associated macrophage protein 1 (NRAMP) related with *Brucellosis* in Cattle.



## REFERENCES

1895. Thomas Henry Huxley. *Science*, 2, 85-7.
- AALI, M., MORADI-SHAHRBABA, M., MORADI-SHAHRBABA, H. & SADEGHI, M. 2014. Detecting novel SNPs and breed-specific haplotypes at calpastatin gene in Iranian fat- and thin-tailed sheep breeds and their effects on protein structure. *Gene*, 537, 132-9.
- ACHILLI, A., OLIVIERI, A., PELLECCIA, M., UBOLDI, C., COLLI, L., AL-ZAHERY, N., ACCETTURO, M., PALA, M., HOOSHIAR KASHANI, B., PEREGO, U. A., BATTAGLIA, V., FORNARINO, S., KALAMATI, J., HOUSHMAND, M., NEGRINI, R., SEMINO, O., RICHARDS, M., MACAULAY, V., FERRETTI, L., BANDELT, H. J., AJMONE-MARSAN, P. & TORRONI, A. 2008. Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr Biol*, 18, R157-8.
- ADWAN, G., ADWAN, K., BDIR, S. & ABUSEIR, S. 2013. Molecular characterization of *Echinococcus granulosus* isolated from sheep in Palestine. *Exp Parasitol*, 134, 195-9.
- ALLAN, A. J., SANDERSON, N. D., GUBBINS, S., ELLIS, S. A. & HAMMOND, J. A. 2015. Cattle NK Cell Heterogeneity and the Influence of MHC Class I. *J Immunol*, 195, 2199-206.
- ANDREWS, R. M., KUBACKA, I., CHINNERY, P. F., LIGHTOWLERS, R. N., TURNBULL, D. M. & HOWELL, N. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet*, 23, 147.
- ANSARI, H. R., FLOWER, D. R. & RAGHAVA, G. P. 2010. AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res*, 38, D847-53.
- ANTUNES, A. & RAMOS, M. J. 2005. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. *Genomics*, 86, 708-17.
- AOUINTI, S., MALOUCHE, D., GIUDICELLI, V., KOSSIDA, S. & LEFRANC, M. P. 2015. IMGT/HighV-QUEST Statistical Significance of IMGT Clonotype (AA) Diversity per Gene for Standardized Comparisons of Next Generation Sequencing Immunoprofiles of Immunoglobulins and T Cell Receptors. *PLoS One*, 10, e0142353.
- ASSOU, S., BOUMELA, I., HAOUZI, D., ANAHORY, T., DECHAUD, H., DE VOS, J. & HAMAMAH, S. 2010. Dynamic changes in gene expression during human early embryo development: from fundamental aspects to clinical applications. *Human Reproduction Update*, 17, 272-290.
- AVISE, J. C., AYALA, F. J. & NATIONAL ACADEMY OF, S. 2010. *In the light of evolution. Volume IV, Volume IV* [Online]. Washington, D.C.: National Academies Press. Available: <http://site.ebrary.com/id/10439401>.

- BACKERT, L. & KOHLBACHER, O. 2015a. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med*, 7, 119.
- BACKERT, L. & KOHLBACHER, O. 2015b. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med*, 7, 119.
- BAHITHAM, W., LIAO, X., PENG, F., BAMFORTH, F., CHAN, A., MASON, A., STONE, B., STOTHARD, P. & SERGI, C. 2014. Mitochondriome and cholangiocellular carcinoma. *PLoS One*, 9, e104694.
- BAILEY, S. D., VIRTANEN, C., HAIBE-KAINS, B. & LUPIEN, M. 2015. ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics*, 31, 3057-9.
- BALBINOTTI, H., SANTOS, G. B., BADARACO, J., AREND, A. C., GRAICHEN, D. Â. S., HAAG, K. L. & ZAHA, A. 2012. Echinococcus ortleppi (G5) and Echinococcus granulosus sensu stricto (G1) loads in cattle from Southern Brazil. *Vet Parasitol*, 188, 255-60.
- BANDELT, H.-J., MACAULAY, V. & RICHARDS, M. 2006. *Human mitochondrial DNA and the evolution of homo sapiens* [Online]. Berlin; New York: Springer. Available: <http://public.eblib.com/choice/publicfullrecord.aspx?p=304560>.
- BARBIERI, C., HEGGARTY, P., YANG YAO, D., FERRI, G., DE FANTI, S., SARNO, S., CIANI, G., BOATTINI, A., LUISELLI, D. & PETTENER, D. 2014. Between Andes and Amazon: the genetic profile of the Arawak-speaking Yaneshá. *Am J Phys Anthropol*, 155, 600-9.
- BAYONA-BAFALUY, M. P., ACIN-PEREZ, R., MULLIKIN, J. C., PARK, J. S., MORENO-LOSHUERTOS, R., HU, P., PEREZ-MARTOS, A., FERNANDEZ-SILVA, P., BAI, Y. & ENRIQUEZ, J. A. 2003. Revisiting the mouse mitochondrial DNA sequence. *Nucleic Acids Res*, 31, 5349-55.
- BEAZ-HIDALGO, R., HOSSAIN, M. J., LILES, M. R. & FIGUERAS, M. J. 2015. Strategies to avoid wrongly labelled genomes using as example the detected wrong taxonomic affiliation for aeromonas genomes in the GenBank database. *PLoS One*, 10, e0115813.
- BEERLI, P. & FELSENSTEIN, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A*, 98, 4563-8.
- BEHL, J. D., MISHRA, P., VERMA, N. K., NIRANJAN, S. K., DANGI, P. S., SHARMA, R. & BEHL, R. 2016. Nucleotide polymorphisms in the bovine lymphotoxin A gene and their distribution among Bos indicus zebu cattle breeds. *Gene*, 579, 82-94.
- BENSON, D. A., BOGUSKI, M. S., LIPMAN, D. J., OSTELL, J. & OUELLETTE, B. F. 1998. GenBank. *Nucleic Acids Res*, 26, 1-7.

- BENSON, D. A., CAVANAUGH, M., CLARK, K., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J. & SAYERS, E. W. 2013. GenBank. *Nucleic Acids Res*, 41, D36-42.
- BENSON, D. A., KARSCH-MIZRACHI, I., LIPMAN, D. J., OSTELL, J., RAPP, B. A. & WHEELER, D. L. 2000. GenBank. *Nucleic Acids Res*, 28, 15-8.
- BERGERON, B. P. 2003. *Bioinformatics Computing*, Prentice Hall.
- BERNT, M., BRABAND, A., MIDDENDORF, M., MISOF, B., ROTA-STABELLI, O. & STADLER, P. F. 2013. Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. *Mol Phylogenet Evol*, 69, 320-7.
- BINGHAM, N. H., GOLDIE, C. M. & KINGMAN, J. F. C. 2010. *Probability and mathematical genetics : [papers in honour of Sir John Kingman]*, Cambridge, UK; New York, Cambridge University Press.
- BLAIR, C., DAVY, C. M., NGO, A., ORLOV, N. L., SHI, H. T., LU, S. Q., GAO, L., RAO, D. Q. & MURPHY, R. W. 2013. Genealogy and Demographic History of a Widespread Amphibian throughout Indochina. *J Hered*, 104, 72-85.
- BLANQUART, S. & GASCUEL, O. 2011. Mitochondrial genes support a common origin of rodent malaria parasites and Plasmodium falciparum's relatives infecting great apes. *BMC Evol Biol*, 11, 70.
- BOLANDER, F. F. 2004. *Molecular endocrinology* [Online]. Amsterdam; Boston: Elsevier Academic Press. Available: [http://www.123library.org/book\\_details/?id=44304](http://www.123library.org/book_details/?id=44304).
- BONHOMME, F., ORTH, A., CUCCHI, T., RAJABI-MAHAM, H., CATALAN, J., BOURSOT, P., AUFRAY, J. C. & BRITTON-DAVIDIAN, J. 2011. Genetic differentiation of the house mouse around the Mediterranean basin: matrilineal footprints of early and late colonization. *Proc Biol Sci*, 278, 1034-43.
- BROWN, I. H. 2000. The epidemiology and evolution of influenza viruses in pigs. *Veterinary microbiology*, 74, 29-46.
- BROWN, T., DIDELOT, X., WILSON, D. J. & DE MAIO, N. 2016. SimBac: simulation of whole bacterial genomes with homologous recombination. *Microb Genom*, 2.
- CARBO, A., HONTECILLAS, R., ANDREW, T., EDEN, K., MEI, Y., HOOPS, S. & BASSAGANYA-RIERA, J. 2014. Computational modeling of heterogeneity and function of CD4+ T cells. *Front Cell Dev Biol*, 2, 31.
- CARVALHO, W. A., IANELLA, P., ARNOLDI, F. G., CAETANO, A. R., MARUYAMA, S. R., FERREIRA, B. R., CONTI, L. H., DA SILVA, M. R., PAULA, J. O., MAIA, A. A. & SANTOS, I. K. 2011. Haplotypes of the bovine IgG2 heavy gamma chain in tick-resistant and tick-susceptible breeds of cattle. *Immunogenetics*, 63, 319-24.
- CHAIN, P. S., COMERCI, D. J., TOLMASKY, M. E., LARIMER, F. W., MALFATTI, S. A., VERGEZ, L. M., AGUERO, F., LAND, M. L., UGALDE, R. A. & GARCIA,

- E. 2005. Whole-genome analyses of speciation events in pathogenic Brucellae. *Infect Immun*, 73, 8353-61.
- CHAUVE, C., EL-MABROUK, N. & TANNIER, E. 2013. *Models and algorithms for genome evolution* [Online]. Available: <http://public.eblib.com/choice/publicfullrecord.aspx?p=1466682>.
- CHEN, C., FRANKHOUSER, D. & BUNDSCHUH, R. 2012. Comparison of insertional RNA editing in Myxomycetes. *PLoS Comput Biol*, 8, e1002400.
- CHEN, F. C. & LI, W. H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet*, 68, 444-56.
- CHEN, S. L., ZHANG, Y. X., XU, J. Y., MENG, L., SHA, Z. X. & REN, G. C. 2007. Molecular cloning, characterization and expression analysis of natural resistance associated macrophage protein (Nramp) cDNA from turbot (*Scophthalmus maximus*). *Comp Biochem Physiol B Biochem Mol Biol*, 147, 29-37.
- CÍZKOVÁ, A., STRÁNECKÝ, V., IVÁNEK, R., HARTMANNOVÁ, H., NOSKOVÁ, L., PIHEROVÁ, L., TESAROVÁ, M., HANSÍKOVÁ, H., HONZÍK, T., ZEMAN, J., DIVINA, P., POTOCKÁ, A., PAUL, J., SPERL, W., MAYR, J. A., SENECA, S., HOUSTEK, J. & KMOCH, S. 2008. Development of a human mitochondrial oligonucleotide microarray (h-MitoArray) and gene expression analysis of fibroblast cell lines from 13 patients with isolated F1Fo ATP synthase deficiency. *BMC genomics*, 9.
- COORDINATORS, N. R. 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*.
- CORBEL, M. J., ORGANIZATION, W. H., FOOD, NATIONS, A. O. O. T. U. & EPIZOOTICS, I. O. O. 2006. *Brucellosis in Humans and Animals*, World Health Organization.
- CORNELI, P. S. & WARD, R. H. 2000. Mitochondrial genes and mammalian phylogenies: increasing the reliability of branch length estimation. *Mol Biol Evol*, 17, 224-34.
- COUSSENS, P. M., COUSSENS, M. J., TOOKER, B. C. & NOBIS, W. 2004. Structure of the bovine natural resistance associated macrophage protein (NRAMP 1) gene and identification of a novel polymorphism. *DNA Seq*, 15, 15-25.
- CUI, J., CHENG, Y. & BELOV, K. 2015. Diversity in the Toll-like receptor genes of the Tasmanian devil (*Sarcophilus harrisii*). *Immunogenetics*, 67, 195-201.
- DE SOUTO, M. C. P. & KANN, M. G. 2012. *Advances in Bioinformatics and Computational Biology: 7th Brazilian Symposium on Bioinformatics, BSB 2012, Campo Grande, Brazil, August 15-17, 2012, Proceedings*, Springer Berlin Heidelberg.
- DECOTTIGNIES, A. 2005. Capture of extranuclear DNA at fission yeast double-strand breaks. *Genetics*, 171, 1535-48.



- DELUCA, D. S. & BLASCZYK, R. 2007. The immunoinformatics of cancer immunotherapy. *Tissue Antigens*, 70, 265-71.
- DELVECCHIO, V. G., KAPATRAL, V., REDKAR, R. J., PATRA, G., MUJER, C., LOS, T., IVANOVA, N., ANDERSON, I., BHATTACHARYYA, A., LYKIDIS, A., REZNIK, G., JABLONSKI, L., LARSEN, N., D'SOUZA, M., BERNAL, A., MAZUR, M., GOLTSMAN, E., SELKOV, E., ELZER, P. H., HAGIUS, S., O'CALLAGHAN, D., LETESSON, J. J., HASELKORN, R., KYRPIDES, N. & OVERBEEK, R. 2002. The genome sequence of the facultative intracellular pathogen *Brucella melitensis*. *Proc Natl Acad Sci U S A*, 99, 443-8.
- DROSOPHILA 12 GENOMES, C., CLARK, A. G., EISEN, M. B., SMITH, D. R., BERGMAN, C. M., OLIVER, B., MARKOW, T. A., KAUFMAN, T. C., KELLIS, M., GELBART, W., IYER, V. N., POLLARD, D. A., SACKTON, T. B., LARRACUENTE, A. M., SINGH, N. D., ABAD, J. P., ABT, D. N., ADRYAN, B., AGUADE, M., AKASHI, H., ANDERSON, W. W., AQUADRO, C. F., ARDELL, D. H., ARGUELLO, R., ARTIERI, C. G., BARBASH, D. A., BARKER, D., BARSANTI, P., BATTERHAM, P., BATZOGLOU, S., BEGUN, D., BHUTKAR, A., BLANCO, E., BOSAK, S. A., BRADLEY, R. K., BRAND, A. D., BRENT, M. R., BROOKS, A. N., BROWN, R. H., BUTLIN, R. K., CAGGESE, C., CALVI, B. R., BERNARDO DE CARVALHO, A., CASPI, A., CASTREZANA, S., CELNIKER, S. E., CHANG, J. L., CHAPPLE, C., CHATTERJI, S., CHINWALLA, A., CIVETTA, A., CLIFTON, S. W., COMERON, J. M., COSTELLO, J. C., COYNE, J. A., DAUB, J., DAVID, R. G., DELCHER, A. L., DELEHAUNTY, K., DO, C. B., EBLING, H., EDWARDS, K., EICKBUSH, T., EVANS, J. D., FILIPSKI, A., FINDEISS, S., FREYHULT, E., FULTON, L., FULTON, R., GARCIA, A. C., GARDINER, A., GARFIELD, D. A., GARVIN, B. E., GIBSON, G., GILBERT, D., GNERRE, S., GODFREY, J., GOOD, R., GOTEA, V., GRAVELY, B., GREENBERG, A. J., GRIFFITHS-JONES, S., GROSS, S., GUIGO, R., GUSTAFSON, E. A., HAERTY, W., HAHN, M. W., HALLIGAN, D. L., HALPERN, A. L., HALTER, G. M., HAN, M. V., HEGER, A., HILLIER, L., HINRICHS, A. S., HOLMES, I., HOSKINS, R. A., HUBISZ, M. J., HULTMARK, D., HUNTLEY, M. A., JAFFE, D. B., et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450, 203-18.
- DURBIN, R., EDDY, S., KROGH, A. & MITCHISON, G. 1998. *Biological sequence analysis : probabilistic models of proteins and nucleic acids*, Cambridge, United Kingdom, Cambridge University Press.
- E, G. X., NA, R. S., ZHAO, Y. J., CHEN, L. P., QIU, X. Y. & HUANG, Y. F. 2015. Brief Note : Variability in the cathelicidin 6 (CATHL-6) gene in Tianzhu white yak from Tibetan area in China. *Genet Mol Res*, 14, 3129-32.
- EHRENMANN, F., KAAS, Q. & LEFRANC, M. P. 2010. IMGT/3Dstructure-DB and IMGT/DomainGapAlign: a database and a tool for immunoglobulins or antibodies, T cell receptors, MHC, IgSF and MhcSF. *Nucleic Acids Res*, 38, D301-7.
- EIDAM, C., POEHLEIN, A., LEIMBACH, A., MICHAEL, G. B., KADLEC, K., LIESEGANG, H., DANIEL, R., SWEENEY, M. T., MURRAY, R. W., WATTS,

- J. L. & SCHWARZ, S. 2015. Analysis and comparative genomics of ICEMh1, a novel integrative and conjugative element (ICE) of *Mannheimia haemolytica*. *J Antimicrob Chemother*, 70, 93-7.
- FALUS, A. 2009a. *Clinical applications of immunomics* [Online]. New York: Springer. Available: <http://public.eblib.com/choice/publicfullrecord.aspx?p=417148>.
- FALUS, A. 2009b. *Clinical Applications Of Immunomics. Immunomics Reviews* [Online]. Springer. Available: <http://www.myilibrary.com?id=195012>.
- FELDHahn, M., DONNES, P., THIEL, P. & KOHLBACHER, O. 2009. FRED--a framework for T-cell epitope detection. *Bioinformatics*, 25, 2758-9.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17, 368-76.
- FELSENSTEIN, J. & CHURCHILL, G. A. 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol*, 13, 93-104.
- FLOWER, D. R. 2002. *Drug Design: Cutting Edge Approaches*, Royal Society of Chemistry.
- FLOWER, D. R. 2013. *Immunomic discovery of adjuvants and candidate subunit vaccines* [Online]. New York, NY: Springer New York : Imprint: Springer.
- FLOWER, D. R., DAVIES, M. N. & RANGANATHAN, S. 2010. *Bioinformatics for immunomics* [Online]. New York: Springer. Available: <http://public.eblib.com/choice/publicfullrecord.aspx?p=510696>.
- FOURNIER, D., TINDO, M., KENNE, M., MBENOUN MASSE, P. S., VAN BOSSCHE, V., DE CONINCK, E. & ARON, S. 2012. Genetic structure, nestmate recognition and behaviour of two cryptic species of the invasive big-headed ant *Pheidole megacephala*. *PLoS One*, 7, e31480.
- GAO, J. F., ZHAO, Q., LIU, G. H., ZHANG, Y., ZHANG, Y., WANG, W. T., CHANG, Q. C., WANG, C. R. & ZHU, X. Q. 2014. Comparative analyses of the complete mitochondrial genomes of the two ruminant hookworms *Bunostomum trigonocephalum* and *Bunostomum phlebotomum*. *Gene*, 541, 92-100.
- GARCIA-ANGULO, V. A., KALITA, A., KALITA, M., LOZANO, L. & TORRES, A. G. 2014. Comparative genomics and immunoinformatics approach for the identification of vaccine candidates for enterohemorrhagic *Escherichia coli* O157:H7. *Infect Immun*, 82, 2016-26.
- GIBSON, R. & BAKER, A. 2012. Multiple gene sequences resolve phylogenetic relationships in the shorebird suborder Scolopaci (Aves: Charadriiformes). *Mol Phylogenet Evol*, 64, 66-72.
- GISPERT, S., BREHM, N., WEIL, J., SEIDEL, K., RUB, U., KERN, B., WALTER, M., ROEPER, J. & AUBURGER, G. 2015. Potentiation of neurotoxicity in double-mutant mice with Pink1 ablation and A53T-SNCA overexpression. *Hum Mol Genet*, 24, 1061-76.

- GISSI, C., GULLBERG, A. & ARNASON, U. 1998. The complete mitochondrial DNA sequence of the rabbit, *Oryctolagus cuniculus*. *Genomics*, 50, 161-9.
- GIUDICELLI, V. & LEFRANC, M. P. 2012. Imgt-Ontology 2012. *Front Genet*, 3, 79.
- GOODSWEN, S. J., KENNEDY, P. J. & ELLIS, J. T. 2013. A guide to in silico vaccine discovery for eukaryotic pathogens. *Brief Bioinform*, 14, 753-74.
- GREENE, W. & HILL, R. C. 2010. *Maximum simulated likelihood methods and applications* [Online]. Bingley, U.K.: Emerald. Available: <http://public.eblib.com/choice/publicfullrecord.aspx?p=647730>.
- GWIDA, M., EL-ASHKER, M., MELZER, F., EL-DIASTY, M., EL-BESKAWY, M. & NEUBAUER, H. 2015. Use of serology and real time PCR to control an outbreak of bovine brucellosis at a dairy cattle farm in the Nile Delta region, Egypt. *Ir Vet J*, 69, 3.
- HAGEN, C. M., AIDT, F. H., HAVNDRUP, O., HEDLEY, P. L., JESPERSGAARD, C., JENSEN, M., KANTERS, J. K., MOOLMAN-SMOOK, J. C., MOLLER, D. V., BUNDGAARD, H. & CHRISTIANSEN, M. 2013. MT-CYB mutations in hypertrophic cardiomyopathy. *Mol Genet Genomic Med*, 1, 54-65.
- HAMMOND, J. A., MARSH, S. G., ROBINSON, J., DAVIES, C. J., STEAR, M. J. & ELLIS, S. A. 2012. Cattle MHC nomenclature: is it possible to assign sequences to discrete class I genes? *Immunogenetics*, 64, 475-80.
- HANSEN, A. M., RASMUSSEN, M., SVITEK, N., HARND AHL, M., GOLDE, W. T., BARLOW, J., NENE, V., BUUS, S. & NIELSEN, M. 2014. Characterization of binding specificities of bovine leucocyte class I molecules: impacts for rational epitope discovery. *Immunogenetics*, 66, 705-18.
- HASEGAWA, M., KISHINO, H. & YANO, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22, 160-74.
- HASHIZUME, O., YAMANASHI, H., TAKETO, M. M., NAKADA, K. & HAYASHI, J. 2015. A specific nuclear DNA background is required for high frequency lymphoma development in transmitochondrial mice with G13997A mtDNA. *PLoS One*, 10, e0118561.
- HASSANIN, A., AN, J., ROPIQUET, A., NGUYEN, T. T. & COULOUX, A. 2013. Combining multiple autosomal introns for studying shallow phylogeny and taxonomy of Laurasiatherian mammals: Application to the tribe Bovini (Cetartiodactyla, Bovidae). *Mol Phylogenet Evol*, 66, 766-75.
- HASSANIN, A., BONILLO, C., NGUYEN, B. X. & CRUAUD, C. 2010. Comparisons between mitochondrial genomes of domestic goat (*Capra hircus*) reveal the presence of numts and multiple sequencing errors. *Mitochondrial DNA*, 21, 68-76.
- HEDGES, J. F., KIMMEL, E., SNYDER, D. T., JEROME, M. & JUTILA, M. A. 2013. Solute carrier 11A1 is expressed by innate lymphocytes and augments their activation. *J Immunol*, 190, 4263-73.

- HERZIG, C. T., BLUMERMAN, S. L. & BALDWIN, C. L. 2006. Identification of three new bovine T-cell receptor delta variable gene subgroups expressed by peripheral blood T cells. *Immunogenetics*, 58, 746-57.
- HERZIG, C. T., MAILLOUX, V. L. & BALDWIN, C. L. 2015. Spectratype analysis of the T cell receptor delta CDR3 region of bovine gammadelta T cells responding to leptospira. *Immunogenetics*, 67, 95-109.
- HIENDLEDER, S., LEWALSKI, H., WASSMUTH, R. & JANKE, A. 1998. The complete mitochondrial DNA sequence of the domestic sheep (*Ovis aries*) and comparison with the other major ovine haplotype. *J Mol Evol*, 47, 441-8.
- HOKAMP, K. 2015. Perl One-Liners: Bridging the Gap Between Large Data Sets and Analysis Tools. *Methods Mol Biol*, 1326, 177-91.
- HORAI, S., HAYASAKA, K., KONDO, R., TSUGANE, K. & TAKAHATA, N. 1995. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A*, 92, 532-6.
- INGMAN, M. & GYLLENSTEN, U. 2001. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. *J Hered*, 92, 454-61.
- INGMAN, M., KAESSMANN, H., PAABO, S. & GYLLENSTEN, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature*, 408, 708-713.
- ISAEV, A. 2006. *Introduction to Mathematical Methods in Bioinformatics*, Springer Berlin Heidelberg.
- ISHIDA, Y. 2004. *Immunity-based systems : a design perspective; and 14 tables*, Berlin; Heidelberg [u.a.], Springer.
- JENSEN, J. V. 1993. Thomas Henry Huxley's address at the opening of the Johns Hopkins University in September 1876. *Notes Rec R Soc Lond*, 47, 257-69.
- JI, R., CUI, P., DING, F., GENG, J., GAO, H., ZHANG, H., YU, J., HU, S. & MENG, H. 2009. Monophyletic origin of domestic bactrian camel (*Camelus bactrianus*) and its evolutionary relationship with the extant wild camel (*Camelus bactrianus ferus*). *Anim Genet*, 40, 377-82.
- JIA, W., YAN, H., LOU, Z., NI, X., DYACHENKO, V., LI, H. & LITTLEWOOD, D. T. 2012. Mitochondrial genes and genomes support a cryptic species of tapeworm within *Taenia taeniaeformis*. *Acta Trop*, 123, 154-63.
- JIANG, J., GU, J., ZHANG, L., ZHANG, C., DENG, X., DOU, T., ZHAO, G. & ZHOU, Y. 2015. Comparing Mycobacterium tuberculosis genomes using genome topology networks. *BMC Genomics*, 16, 85.
- JONAS, R., KITTL, S., OVERESCH, G. & KUHNERT, P. 2015. Genotypes and antibiotic resistance of bovine *Campylobacter* and their contribution to human campylobacteriosis. *Epidemiol Infect*, 143, 2373-80.

- JONSSON, N. N., PIPER, E. K. & CONSTANTINOIU, C. C. 2014. Host resistance in cattle to infestation with the cattle tick *Rhipicephalus microplus*. *Parasite Immunol*, 36, 553-9.
- JORGENSEN, K. W., RASMUSSEN, M., BUUS, S. & NIELSEN, M. 2014. NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology*, 141, 18-26.
- JUKES, T. H. 1995. A comparison of mitochondrial tRNAs in five vertebrates. *J Mol Evol*, 40, 537-40.
- KABEKKODU, S. P., BHAT, S., MASCARENHAS, R., MALLYA, S., BHAT, M., PANDEY, D., KUSHTAGI, P., THANGARAJ, K., GOPINATH, P. M. & SATYAMOORTHY, K. 2014. Mitochondrial DNA variation analysis in cervical cancer. *Mitochondrion*, 16, 73-82.
- KARI, L., HILL, K. A., SAYEM, A. S., KARAMICHALIS, R., BRYANS, N., DAVIS, K. & DATTANI, N. S. 2015. Mapping the space of genomic signatures. *PLoS One*, 10, e0119815.
- KASAHARA, M. & YOSHIDA, S. 2012. Immunogenetics of the NKG2D ligand gene family. *Immunogenetics*, 64, 855-67.
- KATARIA, R. S., TAIT, R. G., JR., KUMAR, D., ORTEGA, M. A., RODRIGUEZ, J. & REECY, J. M. 2011. Association of toll-like receptor four single nucleotide polymorphisms with incidence of infectious bovine keratoconjunctivitis (IBK) in cattle. *Immunogenetics*, 63, 115-9.
- KEELE, J. W., KUEHN, L. A., MCDANELD, T. G., TAIT, R. G., JONES, S. A., SMITH, T. P., SHACKELFORD, S. D., KING, D. A., WHEELER, T. L., LINDHOLM-PERRY, A. K. & MCNEEL, A. K. 2015. Genomewide association study of lung lesions in cattle using sample pooling. *J Anim Sci*, 93, 956-64.
- KHAN, A., ASIF, H., STUDHOLME, D. J., KHAN, I. A. & AZIM, M. K. 2013. Genome characterization of a novel *Burkholderia cepacia* complex genomovar isolated from dieback affected mango orchards. *World J Microbiol Biotechnol*, 29, 2033-44.
- KIM, E. S., SONSTEGARD, T. S., DA SILVA, M. V., GASBARRE, L. C. & VAN TASSELL, C. P. 2015a. Genome-wide scan of gastrointestinal nematode resistance in closed Angus population selected for minimized influence of MHC. *PLoS One*, 10, e0119380.
- KIM, K. S., LEE, S. E., JEONG, H. W. & HA, J. H. 1998. The complete nucleotide sequence of the domestic dog (*Canis familiaris*) mitochondrial genome. *Mol Phylogenet Evol*, 10, 210-20.
- KIM, R. N., KIM, D. S., CHOI, S. H., YOON, B. H., KANG, A., NAM, S. H., KIM, D. W., KIM, J. J., HA, J. H., TOYODA, A., FUJIYAMA, A., KIM, A., KIM, M. Y., PARK, K. H., LEE, K. S. & PARK, H. S. 2012. Genome analysis of the domestic dog (Korean Jindo) by massively parallel sequencing. *DNA Res*, 19, 275-87.

- KIM, Y. J., OH, D. H., SONG, B. R., HEO, E. J., LIM, J. S., MOON, J. S., PARK, H. J., WEE, S. H. & SUNG, K. 2015b. Molecular Characterization, Antibiotic Resistance, and Virulence Factors of Methicillin-Resistant *Staphylococcus aureus* Strains Isolated from Imported and Domestic Meat in Korea. *Foodborne Pathog Dis*, 12, 390-8.
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16, 111-20.
- KISHINO, H. & HASEGAWA, M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*, 29, 170-9.
- KIZILKAYA, K., FERNANDO, R. L. & GARRICK, D. J. 2014. Reduction in accuracy of genomic prediction for ordered categorical data compared to continuous observations. *Genet Sel Evol*, 46, 37.
- KOLASKAR, A. S. & TONGAONKAR, P. C. 1990. A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, 276, 172-4.
- KONRADSEN, J. R., FUJISAWA, T., VAN HAGE, M., HEDLIN, G., HILGER, C., KLEINE-TEBBE, J., MATSUI, E. C., ROBERTS, G., RONMARK, E. & PLATTS-MILLS, T. A. 2015. Allergy to furry animals: New insights, diagnostic approaches, and challenges. *J Allergy Clin Immunol*, 135, 616-25.
- KORBER, B., LABUTE, M. & YUSIM, K. 2006. Immunoinformatics comes of age. *PLoS Comput Biol*, 2, e71.
- KUMAR, S., BELLIS, C., ZLOJUTRO, M., MELTON, P. E., BLANGERO, J. & CURRAN, J. E. 2011. Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. *BMC Evol Biol*, 11, 293.
- KUROCHKIN, I. V., MIZUNO, Y., KONAGAYA, A., SAKAKI, Y., SCHONBACH, C. & OKAZAKI, Y. 2007. Novel peroxisomal protease Tysnd1 processes PTS1- and PTS2-containing enzymes involved in beta-oxidation of fatty acids. *EMBO J*, 26, 835-45.
- LACHOWICZ, M., MIEKISZ, J. & WORLD, S. 2009. *From genetics to mathematics* [Online]. Singapore; Hackensack, N.J.: World Scientific Pub. Co. Available: <http://site.ebrary.com/id/10361926>.
- LANE, J., DUROUX, P. & LEFRANC, M. P. 2010. From IMGT-ONTOLOGY to IMGT/LIGMotif: the IMGT standardized approach for immunoglobulin and T cell receptor gene identification and description in large genomic sequences. *BMC Bioinformatics*, 11, 223.
- LANGEVELD, J. P., JACOBS, J. G., HUNTER, N., VAN KEULEN, L. J., LANTIER, F., VAN ZIJDERVELD, F. G. & BOSSERS, A. 2016. Prion Type-Dependent Deposition of PRNP Allelic Products in Heterozygous Sheep. *J Virol*, 90, 805-12.

- LARSON, J. H., MARRON, B. M., BEEVER, J. E., ROE, B. A. & LEWIN, H. A. 2006. Genomic organization and evolution of the ULBP genes in cattle. *BMC Genomics*, 7, 227.
- LEFRANC, M. P. 2009. [Antibody databases: IMGT, a French platform of world-wide interest]. *Med Sci (Paris)*, 25, 1020-3.
- LEFRANC, M. P. 2014a. Immunoglobulin and T Cell Receptor Genes: IMGT((R)) and the Birth and Rise of Immunoinformatics. *Front Immunol*, 5, 22.
- LEFRANC, M. P. 2014b. Immunoglobulins: 25 years of immunoinformatics and IMGT-ONTOLOGY. *Biomolecules*, 4, 1102-39.
- LEFRANC, M. P., GIUDICELLI, V., DUROUX, P., JABADO-MICHALOUD, J., FOLCH, G., AOUINTI, S., CARILLON, E., DUVERGEY, H., HOULES, A., PAYSAN-LAFOSSÉ, T., HADI-SALJOQI, S., SASORITH, S., LEFRANC, G. & KOSSIDA, S. 2015. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res*, 43, D413-22.
- LEFRANC, M. P., GIUDICELLI, V., GINESTOUX, C., JABADO-MICHALOUD, J., FOLCH, G., BELLAHCENE, F., WU, Y., GEMROT, E., BROCHET, X., LANE, J., REGNIER, L., EHRENMANN, F., LEFRANC, G. & DUROUX, P. 2009. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res*, 37, D1006-12.
- LEFRANC, M. P., GIUDICELLI, V., REGNIER, L. & DUROUX, P. 2008. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *Brief Bioinform*, 9, 263-75.
- LEIMEISTER, C. A., BODEN, M., HORWEGE, S., LINDNER, S. & MORGENSTERN, B. 2014. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, 30, 1991-9.
- LI, W.-H. & GRAUR, D. 1991. *Fundamentals of molecular evolution*, Sunderland, Mass., Sinauer Associates.
- LIN, C. S., SUN, Y. L., LIU, C. Y., YANG, P. C., CHANG, L. C., CHENG, I. C., MAO, S. J. & HUANG, M. C. 1999. Complete nucleotide sequence of pig (*Sus scrofa*) mitochondrial genome and dating evolutionary divergence within Artiodactyla. *Gene*, 236, 107-14.
- LIN, R. Q., QIU, L. L., LIU, G. H., WU, X. Y., WENG, Y. B., XIE, W. Q., HOU, J., PAN, H., YUAN, Z. G., ZOU, F. C., HU, M. & ZHU, X. Q. 2011. Characterization of the complete mitochondrial genomes of five *Eimeria* species from domestic chickens. *Gene*, 480, 28-33.
- LING, J., DAOUD, R., LAJOIE, M. J., CHURCH, G. M., SOLL, D. & LANG, B. F. 2014. Natural reassignment of CUU and CUA sense codons to alanine in *Ashbya* mitochondria. *Nucleic Acids Res*, 42, 499-508.
- LIPKIN, E. & STRILLACCI, M. G. 2016. The Use of Kosher Phenotyping for Mapping QTL Affecting Susceptibility to Bovine Respiratory Disease. 11, e0153423.

- LIU, W., MENG, X., XU, Q., FLOWER, D. R. & LI, T. 2006. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics*, 7, 182.
- LOHIA, N. & BARANWAL, M. 2014. Conserved peptides containing overlapping CD4+ and CD8+ T-cell epitopes in the H1N1 influenza virus: an immunoinformatics approach. *Viral Immunol*, 27, 225-34.
- LOPEZ, J. V., CEVARIO, S. & O'BRIEN, S. J. 1996. Complete nucleotide sequences of the domestic cat (*Felis catus*) mitochondrial genome and a transposed mtDNA tandem repeat (Numt) in the nuclear genome. *Genomics*, 33, 229-46.
- LORENZ, T. 2010. *Mutational analysis a joint framework for cauchy problems in and beyond vector spaces* [Online]. Berlin: Springer. Available: <http://site.ebrary.com/id/10394603>.
- LUPINDU, A. M., DALSGAARD, A., MSOFFE, P. L., NGOWI, H. A., MTAMBO, M. M. & OLSEN, J. E. 2015. Transmission of antibiotic-resistant *Escherichia coli* between cattle, humans and the environment in peri-urban livestock keeping communities in Morogoro, Tanzania. *Prev Vet Med*, 118, 477-82.
- MACINTYRE, D. 2015. BSE resistance and maternal transmission. *Vet Rec*, 177, 80.
- MADKOUR, M. M. 2014. *Brucellosis*, Elsevier Science.
- MARJORAM, P., MOLITOR, J., PLAGNOL, V. & TAVARE, S. 2003. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A*, 100, 15324-8.
- MARTINEZ-PEREZ, J. M., ROBLES-PEREZ, D., ROJO-VAZQUEZ, F. A. & MARTINEZ-VALLADARES, M. 2012. Comparison of three different techniques to diagnose *Fasciola hepatica* infection in experimentally and naturally infected sheep. *Vet Parasitol*, 190, 80-6.
- MCCMAHON, S. & LAFRAMBOISE, T. 2014. Mutational patterns in the breast cancer mitochondrial genome, with clinical correlates. *Carcinogenesis*, 35, 1046-54.
- MEHTA, V., SEN, R., MOSHIRI, H. & SALAVATI, R. 2015. Mutational analysis of *Trypanosoma brucei* RNA editing ligase reveals regions critical for interaction with KREPA2. *PLoS One*, 10, e0120844.
- MISHRA, S. K., NIRANJAN, S. K., BANERJEE, B., DUBEY, P. K., GONGE, D. S., MISHRA, B. P. & KATARIA, R. S. 2016. High genetic diversity and distribution of Bubu-DQA alleles in swamp buffaloes (*Bubalus bubalis carabanesis*): identification of new Bubu-DQA loci and haplotypes. *Immunogenetics*.
- MOISE, L. & DE GROOT, A. S. 2006. Putting immunoinformatics to the test. *Nat Biotechnol*, 24, 791-2.
- MOISE, L., TERRY, F., GUTIERREZ, A. H., TASSONE, R., LOSIKOFF, P., GREGORY, S. H., BAILEY-KELLOGG, C., MARTIN, W. D. & DE GROOT, A. S. 2014. Smarter vaccine design will circumvent regulatory T cell-mediated evasion in chronic HIV and HCV infection. *Front Microbiol*, 5, 502.



- MOLNAR, J., NAGY, T., STEGER, V., TOTH, G., MARINCS, F. & BARTA, E. 2014. Genome sequencing and analysis of Mangalica, a fatty local pig of Hungary. *BMC Genomics*, 15, 761.
- MOODY, G. 2004. *Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business*, Wiley.
- MUGHINI-GRAS, L., SMID, J., ENSERINK, R., FRANZ, E., SCHOOLS, L., HECK, M. & VAN PELT, W. 2014. Tracing the sources of human salmonellosis: a multi-model comparison of phenotyping and genotyping methods. *Infect Genet Evol*, 28, 251-60.
- NEI, M. & KUMAR, S. 2000. *Molecular Evolution and Phylogenetics*, Oxford University Press.
- NIELSEN, K. & DUNCAN, J. R. 1990. *Animal Brucellosis*, Taylor & Francis.
- NIELSEN, R. 2005. *Statistical methods in molecular evolution*, New York, Springer.
- NORIMINE, J., HAN, S. & BROWN, W. C. 2006. Quantitation of Anaplasma marginale major surface protein (MSP)1a and MSP2 epitope-specific CD4+ T lymphocytes using bovine DRB3\*1101 and DRB3\*1201 tetramers. *Immunogenetics*, 58, 726-39.
- NOYES, N. R., YANG, X., LINKE, L. M., MAGNUSON, R. J., DETTENWANGER, A., COOK, S., GEORNARAS, I., WOERNER, D. E., GOW, S. P., MCALLISTER, T. A., YANG, H., RUIZ, J., JONES, K. L., BOUCHER, C. A., MORLEY, P. S. & BELK, K. E. 2016. Resistome diversity in cattle and the environment decreases during beef production. *Elife*, 5.
- OANY, A. R., AHMAD, S. A., HOSSAIN, M. U. & JYOTI, T. P. 2015. Identification of highly conserved regions in L-segment of Crimean-Congo hemorrhagic fever virus and immunoinformatic prediction about potential novel vaccine. *Adv Appl Bioinform Chem*, 8, 1-10.
- OBARA, I., NIELSEN, M., JESCHEK, M., NIJHOF, A., MAZZONI, C. J., SVITEK, N., STEINAA, L., AWINO, E., OLDS, C., JABBAR, A., CLAUSEN, P. H. & BISHOP, R. P. 2016. Sequence diversity between class I MHC loci of African native and introduced Bos taurus cattle in Theileria parva endemic regions: in silico peptide binding prediction identifies distinct functional clusters. *Immunogenetics*, 68, 339-52.
- OLIVA, M., MESSINA, A., RAGONE, G., CAGGESE, C. & DE PINTO, V. 1998. Sequence and expression pattern of the Drosophila melanogaster mitochondrial porin gene: evidence of a conserved protein domain between fly and mouse. *FEBS Lett*, 430, 327-32.
- OSTERHOFF, D. R. 2010. Research on animal blood groups and biochemical polymorphisms at Onderstepoort (1956-1990). *J S Afr Vet Assoc*, 81, 136-8.
- PALANICHAMY, M. G., MITRA, B., ZHANG, C. L., DEBNATH, M., LI, G. M., WANG, H. W., AGRAWAL, S., CHAUDHURI, T. K. & ZHANG, Y. P. 2015. West

- Eurasian mtDNA lineages in India: an insight into the spread of the Dravidian language and the origins of the caste system. *Hum Genet*, 134, 637-47.
- PAN, Y., SHAO, C., WANG, X., ZHU, Y., CHEN, J., CHEN, W., MA, S. & LIU, J. 2015. [Genomic characteristics of an echovirus 20 strain (KM/EV20/2010) isolated in Kunming, Yunnan, China]. *Zhonghua Liu Xing Bing Xue Za Zhi*, 36, 501-5.
- PANDYA, M., RASMUSSEN, M., HANSEN, A., NIELSEN, M., BUUS, S., GOLDE, W. & BARLOW, J. 2015. A modern approach for epitope prediction: identification of foot-and-mouth disease virus peptides binding bovine leukocyte antigen (BoLA) class I molecules. *Immunogenetics*, 67, 691-703.
- PAREEK, C. S., SMOCZYNSKI, R. & TRETYN, A. 2011. Sequencing technologies and genome sequencing. *J Appl Genet*, 52, 413-35.
- PRABAKARAN, R., FOOD & NATIONS, A. O. O. T. U. 2003. *Good Practices in Planning and Management of Integrated Commercial Poultry Production in South Asia*, Food and Agriculture Organization of the United Nations.
- QING, J., YAN, D., ZHOU, Y., LIU, Q., WU, W., XIAO, Z., LIU, Y., LIU, J., DU, L., XIE, D. & LIU, X. Z. 2014. Whole-exome sequencing to decipher the genetic heterogeneity of hearing loss in a Chinese family with deaf by deaf mating. *PLoS One*, 9, e109178.
- RAHARIMALALA, F. N., RAVAOMANARIVO, L. H., RAVELONANDRO, P., RAFARASOA, L. S., ZOUACHE, K., TRAN-VAN, V., MOUSSON, L., FAILLOUX, A.-B., HELLARD, E., MORO, C. V., RALISOA, B. O. & MAVINGUI, P. 2012. Biogeography of the two major arbovirus mosquito vectors, *Aedes aegypti* and *Aedes albopictus* (Diptera, Culicidae), in Madagascar. *Parasit Vectors*, 5, 56.
- RETTENBERGER, G., KLETT, C., ZECHNER, U., KUNZ, J., VOGEL, W. & HAMEISTER, H. 1995. Visualization of the conservation of synteny between humans and pigs by heterologous chromosomal painting. *Genomics*, 26, 372-378.
- RICE, P., BLEASBY, A. & ISON, J. 2011. *EMBOSS user's guide : practical bioinformatics*, Cambridge; New York, Cambridge University Press.
- RICHTER, R., PAJAK, A., DENNERLEIN, S., ROZANSKA, A., LIGHTOWLERS, R. N. & CHRZANOWSKA-LIGHTOWLERS, Z. M. 2010. Translation termination in human mitochondrial ribosomes. *Biochem Soc Trans*, 38, 1523-6.
- ROBBERTSE, L., BARON, S., VAN DER MERWE, N. A., MADDER, M., STOLTSZ, W. H. & MARITZ-OLIVIER, C. 2016. Genetic diversity, acaricide resistance status and evolutionary potential of a *Rhipicephalus microplus* population from a disease-controlled cattle farming area in South Africa. *Ticks Tick Borne Dis*.
- RÖCK, A. W., DÜR, A., VAN OVEN, M. & PARSON, W. 2013. Concept for estimating mitochondrial DNA haplogroups using a maximum likelihood approach (EMMA). *Forensic Sci Int Genet*, 7, 601-9.

- RODRIGUEZ-RIVERA, L. D., WRIGHT, E. M., SILER, J. D., ELTON, M., CUMMINGS, K. J., WARNICK, L. D. & WIEDMANN, M. 2014. Subtype analysis of Salmonella isolated from subclinically infected dairy cattle and dairy farm environments reveals the presence of both human- and bovine-associated subtypes. *Vet Microbiol*, 170, 307-16.
- ROSENBERG, A. & ARP, R. 2009. *Philosophy of Biology: An Anthology*, John Wiley & Sons.
- ROSS, H. A., MURUGAN, S. & LI, W. L. 2008. Testing the reliability of genetic methods of species identification via simulation. *Syst Biol*, 57, 216-30.
- ROSSET, S., WELLS, R. S., SORIA-HERNANZ, D. F., TYLER-SMITH, C., ROYYURU, A. K., BEHAR, D. M. & CONSORTIUM, G. 2008. Maximum-likelihood estimation of site-specific mutation rates in human mitochondrial DNA from partial phylogenetic classification. *Genetics*, 180, 1511-24.
- ROSSETTI, C. A., DRAKE, K. L., SIDDAVATAM, P., LAWHON, S. D., NUNES, J. E., GULL, T., KHARE, S., EVERTS, R. E., LEWIN, H. A. & ADAMS, L. G. 2013. Systems biology analysis of Brucella infected Peyer's patch reveals rapid invasion with modest transient perturbations of the host transcriptome. *PLoS One*, 8, e81719.
- RUAN, R., WAN, X. L., ZHENG, Y., ZHENG, J. S. & WANG, D. 2016. Assembly and characterization of the MHC class I region of the Yangtze finless porpoise (*Neophocaena asiaeorientalis asiaeorientalis*). *Immunogenetics*, 68, 77-82.
- SADEK, H. A. 2004. *Bioinformatics: Principles, Basic Internet Applications*, Trafford on Demand Pub.
- SATOH, T. P., SATO, Y., MASUYAMA, N., MIYA, M. & NISHIDA, M. 2010. Transfer RNA gene arrangement and codon usage in vertebrate mitochondrial genomes: a new insight into gene order conservation. *BMC Genomics*, 11, 479.
- SCHEFFLER, I. E. 2008. *Mitochondria*, Hoboken, N.J., Wiley-Liss.
- SCHÖNBACH, C., RANGANATHAN, S. & BRUSIC, V. 2008. *Immunoinformatics* [Online]. New York: Springer. Available: <http://public.eblib.com/choice/publicfullrecord.aspx?p=338029>.
- SCHUBERT, B., BRACHVOGEL, H. P., JURGES, C. & KOHLBACHER, O. 2015. EpiToolKit--a web-based workbench for vaccine design. *Bioinformatics*, 31, 2211-3.
- SCHUBERT, B., WALZER, M., BRACHVOGEL, H. P., SZOLEK, A., MOHR, C. & KOHLBACHER, O. 2016. FRED 2: an immunoinformatics framework for Python. *Bioinformatics*.
- SCHWARTZ, J. C. & HAMMOND, J. A. 2015. The assembly and characterisation of two structurally distinct cattle MHC class I haplotypes point to the mechanisms driving diversity. *Immunogenetics*, 67, 539-44.

- SEELYE, S. L., CHEN, P. L., DEISS, T. C. & CRISCITIELLO, M. F. 2016. Genomic organization of the zebrafish (*Danio rerio*) T cell receptor alpha/delta locus and analysis of expressed products. *Immunogenetics*, 68, 365-79.
- SEROUSSI, E., KLOMPUS, S., SILANIKOVE, M., KRIFUCKS, O., SHAPIRO, F., GERTLER, A. & LEITNER, G. 2013. Nonbactericidal secreted phospholipase A2s are potential anti-inflammatory factors in the mammary gland. *Immunogenetics*, 65, 861-71.
- SEVINI, F., GIULIANI, C., VIANELLO, D., GIAMPIERI, E., SANTORO, A., BIONDI, F., GARAGNANI, P., PASSARINO, G., LUISELLI, D., CAPRI, M., FRANCESCHI, C. & SALVIOLI, S. 2014. mtDNA mutations in human aging and longevity: controversies and new perspectives opened by high-throughput technologies. *Exp Gerontol*, 56, 234-44.
- SHARMA, V. 2008. *Bioinformatics*, Rastogi Publications.
- SHI, J., ZHANG, J., LI, S., SUN, J., TENG, Y., WU, M., LI, J., LI, Y., HU, N., WANG, H. & HU, Y. 2015. Epitope-Based Vaccine Target Screening against Highly Pathogenic MERS-CoV: An In Silico Approach Applied to Emerging Infectious Diseases. *PLoS One*, 10, e0144475.
- SHIN, Y., JUNG, H. J., JUNG, M., YOO, S. I., SUBRAMANIAM, S., MARKKANDAN, K., KANG, J. M., RAI, R., PARK, J. & KIM, J. J. 2016. Discovery of Gene Sources for Economic Traits in Hanwoo by Whole-Genome Resequencing. *Asian-Australas J Anim Sci*.
- SINGH, G. B. 2014. *Fundamentals of Bioinformatics and Computational Biology: Methods and Exercises in MATLAB*, Springer International Publishing.
- SINGH, J., MUKHOPADHYAY, C. S., ARORA, J. S. & KAUR, S. 2015. Biocomputational characterization and evolutionary analysis of bubaline dicer1 enzyme. *Asian-Australas J Anim Sci*, 28, 876-87.
- SINGH, K. K. 1998. *Mitochondrial DNA mutations in aging, disease, and cancer*, Berlin; New York; Georgetown, TX, Springer ; Landes Bioscience.
- SOARES, I., AMORIM, A. & GOIOS, A. 2012a. mtDNAoffice: a software to assign human mtDNA macro haplogroups through automated analysis of the protein coding region. *Mitochondrion*, 12, 666-8.
- SOARES, I., GOIOS, A. & AMORIM, A. 2012b. Sequence comparison alignment-free approach based on suffix tree and L-words frequency. *ScientificWorldJournal*, 2012, 450124.
- SPIGELMAN, M., DONOGHUE, H. D., ABDEEN, Z., EREQAT, S., SARIE, I., GREENBLATT, C. L., PAP, I., SZIKOSSY, I., HERSHKOVITZ, I., BAR-GAL, G. K. & MATHESON, C. 2015. Evolutionary changes in the genome of *Mycobacterium tuberculosis* and the human genome from 9000 years BP until modern times. *Tuberculosis (Edinb)*, 95 Suppl 1, S145-9.

- SRIDHAR, S., LAM, F., BLELLOCH, G. E., RAVI, R. & SCHWARTZ, R. 2007. Direct maximum parsimony phylogeny reconstruction from genotype data. *BMC Bioinformatics*, 8, 472.
- SRINIVAS, V. R. 2005. *BIOINFORMATICS: A MODERN APPROACH*, PHI Learning.
- SRIVASTAVA, A., SINGHAL, N., GOEL, M., VIRDI, J. S. & KUMAR, M. 2014. CBMAR: a comprehensive  $\beta$ -lactamase molecular annotation resource. *Database: The Journal of Biological Databases and Curation*, 2014, bau111.
- ST. JOHN, J. C. 2013. *Mitochondrial DNA, mitochondria, disease and stem cells* [Online]. New York: Humana Press. Available: <http://public.ebib.com/choice/publicfullrecord.aspx?p=994580>.
- STECHEER, G., LIU, L., SANDERFORD, M., PETERSON, D., TAMURA, K. & KUMAR, S. 2014. MEGA-MD: molecular evolutionary genetics analysis software with mutational diagnosis of amino acid variation. *Bioinformatics*, 30, 1305-7.
- STEELE, P. R., HERTWECK, K. L., MAYFIELD, D., MCKAIN, M. R., LEEBENS-MACK, J. & PIRES, J. C. 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. *Am J Bot*, 99, 330-48.
- SUN, P., CHEN, W., HUANG, Y., WANG, H., MA, Z. & LV, Y. 2011. Epitope prediction based on random peptide library screening: benchmark dataset and prediction tools evaluation. *Molecules*, 16, 4971-93.
- SUNDARARAMAN, S. A., PLENDERLEITH, L. J., LIU, W., LOY, D. E., LEARN, G. H., LI, Y., SHAW, K. S., AYOUBA, A., PEETERS, M., SPEEDE, S., SHAW, G. M., BUSHMAN, F. D., BRISSON, D., RAYNER, J. C., SHARP, P. M. & HAHN, B. H. 2016. Genomes of cryptic chimpanzee *Plasmodium* species reveal key evolutionary events leading to human malaria. *Nat Commun*, 7.
- SUZUKI, R., LEACH, S., LIU, W., RALSTON, E., SCHEFFEL, J., ZHANG, W., LOWELL, C. A. & RIVERA, J. 2014. Molecular editing of cellular responses by the high-affinity receptor for IgE. *Science*, 343, 1021-5.
- SZECSENYI-NAGY, A., BRANDT, G., HAAK, W., KEERL, V., JAKUCS, J., MOLLER-RIEKER, S., KOHLER, K., MENDE, B. G., OROSS, K., MARTON, T., OSZTAS, A., KISS, V., FECHER, M., PALFI, G., MOLNAR, E., SEBOK, K., CZENE, A., PALUCH, T., SLAUS, M., NOVAK, M., PECINA-SLAUS, N., OSZ, B., VOICSEK, V., SOMOGYI, K., TOTH, G., KROMER, B., BANFFY, E. & ALT, K. W. 2015. Tracing the genetic origin of Europe's first farmers reveals insights into their social organization. *Proc Biol Sci*, 282.
- TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*, 9, 678-87.
- TAMURA, K. & NEI, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10, 512-26.

- THOMPSON-CRISPI, K., ATALLA, H., MIGLIOR, F. & MALLARD, B. A. 2014. Bovine mastitis: frontiers in immunogenetics. *Front Immunol*, 5, 493.
- TIPU, H. N. 2016. Immunoinformatic Analysis of Crimean Congo Hemorrhagic Fever Virus Glycoproteins and Epitope Prediction for Synthetic Peptide Vaccine. *J Coll Physicians Surg Pak*, 26, 108-12.
- TOMAR, N. & DE, R. K. 2010. Immunoinformatics: an integrated scenario. *Immunology*, 131, 153-68.
- TUNG, C. W. & HO, S. Y. 2007. POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*, 23, 942-9.
- USMAN, T., WANG, Y., LIU, C., WANG, X., ZHANG, Y. & YU, Y. 2015. Association study of single nucleotide polymorphisms in JAK2 and STAT5B genes and their differential mRNA expression with mastitis susceptibility in Chinese Holstein cattle. *Anim Genet*, 46, 371-80.
- VAINSHTEIN, Y., SANCHEZ, M., BRAZMA, A., HENTZE, M. W., DANDEKAR, T. & MUCKENTHALER, M. U. 2010. The IronChip evaluation package: a package of perl modules for robust analysis of custom microarrays. *BMC Bioinformatics*, 11, 112.
- VALVERDE, J. R., MARCO, R. & GARESSE, R. 1994. A conserved heptamer motif for ribosomal RNA transcription termination in animal mitochondria. *Proc Natl Acad Sci U S A*, 91, 5368-71.
- VAN GISBERGEN, M. W., VOETS, A. M., STARMANS, M. H., DE COO, I. F., YADAK, R., HOFFMANN, R. F., BOUTROS, P. C., SMEETS, H. J., DUBOIS, L. & LAMBIN, P. 2015. How do changes in the mtDNA and mitochondrial dysfunction influence cancer and cancer therapy? Challenges, opportunities and models. *Mutat Res Rev Mutat Res*, 764, 16-30.
- VIROJ, W. 2008. *Medical biochemoinformatics*, New York, Nova Science Publishers.
- WALL, L., CHRISTIANSEN, T. & ORWANT, J. 2000. *Programming Perl*, Beijing; Cambridge, Mass., O'Reilly.
- WANG, F., HU, S., LIU, W., QIAO, Z., GAO, Y. & BU, Z. 2011. Deep-sequencing analysis of the mouse transcriptome response to infection with *Brucella melitensis* strains of differing virulence. *PLoS One*, 6, e28485.
- WANG, J., YU, X., HU, B., ZHENG, J., XIAO, W., HAO, Y., LIU, W. & WANG, D. 2015a. Physicochemical evolution and molecular adaptation of the cetacean osmoregulation-related gene UT-A2 and implications for functional studies. *Sci Rep*, 5, 8795.
- WANG, X., CHEN, X., SUN, Z. & XIA, J. 2015b. *Single cell sequencing and systems immunology* [Online]. Available: <http://public.eblib.com/choice/publicfullrecord.aspx?p=3109152>.

- WEBER, P. S., MADSEN-BOUSERSE, S. A., ROSA, G. J., SIPKOVSKY, S., REN, X., ALMEIDA, P. E., KRUSKA, R., HALGREN, R. G., BARRICK, J. L. & BURTON, J. L. 2006. Analysis of the bovine neutrophil transcriptome during glucocorticoid treatment. *Physiol Genomics*, 28, 97-112.
- WEI, W. Z., JONES, R. F., JUHASZ, C., GIBSON, H. & VEENSTRA, J. 2015. Evolution of animal models in cancer vaccine development. *Vaccine*, 33, 7401-7.
- WELLING, G. W., WEIJER, W. J., VAN DER ZEE, R. & WELLING-WESTER, S. 1985. Prediction of sequential antigenic regions in proteins. *FEBS Lett*, 188, 215-8.
- WILSON, J. H. & HUNT, T. 2002. *Molecular biology of the cell, 4th edition : a problems approach*, New York; London, Garland Science.
- WISE, D. L., WNEK, G. E., TRANTOLO, D. J., COOPER, T. M. & GRESSER, J. D. 1998. *Photonic Polymer Systems: Fundamentals: Methods, and Applications*, Taylor & Francis.
- WOODHAMS, M. D., FERNANDEZ-SANCHEZ, J. & SUMNER, J. G. 2015. A New Hierarchy of Phylogenetic Models Consistent with Heterogeneous Substitution Rates. *Syst Biol*, 64, 638-50.
- WU, T. D. & NACU, S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26, 873-81.
- XIE, X., GUAN, J. & ZHOU, S. 2015. Similarity evaluation of DNA sequences based on frequent patterns and entropy. *BMC Genomics*, 16 Suppl 3, S5.
- XIONG, H., CAMPELO, D., POLLACK, R. J., RAOULT, D., SHAO, R., ALEM, M., ALI, J., BILCHA, K. & BARKER, S. C. 2014. Second-generation sequencing of entire mitochondrial coding-regions (~15.4 kb) holds promise for study of the phylogeny and taxonomy of human body lice and head lice. *Med Vet Entomol*, 28 Suppl 1, 40-50.
- XU, C. P., LU, Y. Y., YAN, J. Y., FENG, Y., MAO, H. Y., LI, Z., CHEN, Y., WANG, S. K. & GAO, X. P. 2008. [Molecular characteristics and its evolution of the complete genome of avian influenza H5N1 virus isolated in Zhejiang province from 2002 to 2006]. *Zhonghua Liu Xing Bing Xue Za Zhi*, 29, 1114-8.
- XU, X. & ARNASON, U. 1994. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene*, 148, 357-62.
- XU, X. & ARNASON, U. 1996. A complete sequence of the mitochondrial genome of the western lowland gorilla. *Mol Biol Evol*, 13, 691-8.
- YANG, J., SANG, Y., MEADE, K. G. & ROSS, C. 2011. The role of oct-1 in the regulation of tracheal antimicrobial peptide (TAP) and lingual antimicrobial peptide (LAP) expression in bovine mammary epithelial cells. *Immunogenetics*, 63, 715-25.
- YANG, Z. 2006. *Computational molecular evolution*, Oxford, Oxford University Press.
- YANG, Z. 2014. *Molecular Evolution: A Statistical Approach*, OUP Oxford.

- YANG, Z. R. 2010. *Machine Learning Approaches to Bioinformatics*, World Scientific Publishing Company Pte Limited.
- YASSIN, G. M., AMIN, M. A. & ATTIA, A. S. 2016. Immunoinformatics Identifies a Lactoferrin Binding Protein A Peptide as a Promising Vaccine With a Global Protective Prospective Against *Moraxella catarrhalis*. *J Infect Dis*.
- YE, K., LU, J., MA, F., KEINAN, A. & GU, Z. 2014. Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc Natl Acad Sci U S A*, 111, 10654-9.
- YONEMOTO, N., TANAKA, S., FURUKAWA, T. A., KATO, T., MANTANI, A., OGAWA, Y., TAJIKA, A., TAKESHIMA, N., HAYASAKA, Y., SHINOHARA, K., MIKI, K., INAGAKI, M., SHIMODERA, S., AKECHI, T., YAMADA, M., WATANABE, N., GUYATT, G. H. & INVESTIGATORS, S. 2015. Strategic use of new generation antidepressants for depression: SUN(^\_^) D protocol update and statistical analysis plan. *Trials*, 16, 459.
- YOSHIDA, N., YANO, T., KEDO, K., FUJIYOSHI, T., NAGAI, R., IWANO, M., TAGUCHI, E., NISHIDA, T. & TAKAGI, H. 2016. A unique intracellular compartment formed during the oligotrophic growth of *Rhodococcus erythropolis* N9T-4. *Appl Microbiol Biotechnol*.
- YU, Q., RYAN, E. M., ALLEN, T. M., BIRREN, B. W., HENN, M. R. & LENNON, N. J. 2011. PriSM: a primer selection and matching tool for amplification and sequencing of viral genomes. *Bioinformatics*, 27, 266-7.
- ZHANG, J., ZHANG, Z. X., DU, P. C., ZHOU, W., WU, S. D., WANG, Q. L., CHEN, C., SHI, Q., CHEN, C., GAO, C., TIAN, C. & DONG, X. P. 2015. Analyses of the mitochondrial mutations in the Chinese patients with sporadic Creutzfeldt-Jakob disease. *Eur J Hum Genet*, 23, 86-91.
- ZHANG, W., YUE, B., WANG, X., ZHANG, X., XIE, Z., LIU, N., FU, W., YUAN, Y., CHEN, D., FU, D., ZHAO, B., YIN, Y., YAN, X., WANG, X., ZHANG, R., LIU, J., LI, M., TANG, Y., HOU, R. & ZHANG, Z. 2011. Analysis of variable sites between two complete South China tiger (*Panthera tigris amoyensis*) mitochondrial genomes. *Mol Biol Rep*, 38, 4257-64.
- ZHAO, L., OLIVER, E., MARATOU, K., ATANUR, S. S., DUBOIS, O. D., COTRONEO, E., CHEN, C. N., WANG, L., ARCE, C., CHABOSSEAU, P. L., PONSACOBAS, J., FRID, M. G., MOYON, B., WEBSTER, Z., ALDASHEV, A., FERRER, J., RUTTER, G. A., STENMARK, K. R., AITMAN, T. J. & WILKINS, M. R. 2015. The zinc transporter ZIP12 regulates the pulmonary vascular response to chronic hypoxia. *Nature*, 524, 356-60.
- ZHENG, W., RUAN, J., HU, G., WANG, K., HANLON, M. & GAO, J. 2015. Analysis of Conformational B-Cell Epitopes in the Antibody-Antigen Complex Using the Depth Function and the Convex Hull. *PLoS One*, 10, e0134835.
- ZHU, L., LEI, A. H., ZHENG, H. Y., LYU, L. B., ZHANG, Z. G. & ZHENG, Y. T. 2015a. Longitudinal analysis reveals characteristically high proportions of bacterial



- vaginosis-associated bacteria and temporal variability of vaginal microbiota in northern pig-tailed macaques (*Macaca leonina*). *Dongwuxue Yanjiu*, 36, 285-98.
- ZHU, T., DOS REIS, M. & YANG, Z. 2015b. Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. *Syst Biol*, 64, 267-80.
- ZIMIN, A. V., DELCHER, A. L., FLOREA, L., KELLEY, D. R., SCHATZ, M. C., PUIU, D., HANRAHAN, F., PERTEA, G., VAN TASSELL, C. P., SONSTEGARD, T. S., MARÇAIS, G., ROBERTS, M., SUBRAMANIAN, P., YORKE, J. A. & SALZBERG, S. L. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol*, 10, R42.
- ZINICOLA, M., HIGGINS, H., LIMA, S., MACHADO, V., GUARD, C. & BICALHO, R. 2015. Shotgun Metagenomic Sequencing Reveals Functional Genes and Microbiome Associated with Bovine Digital Dermatitis. *PLoS One*, 10, e0133674.
- ZUO, G. & HAO, B. 2015. CVTree3 Web Server for Whole-genome-based and Alignment-free Prokaryotic Phylogeny and Taxonomy. *Genomics Proteomics Bioinformatics*, 13, 321-31.
- ZVELEBIL, M. J. & BAUM, J. O. 2008. *Understanding Bioinformatics*, Garland Science.
- ZYGMUNT, M. S., BUNDLE, D. R., GANESH, N. V., GUIARD, J. & CLOECKAERT, A. 2015. Monoclonal Antibody-Defined Specific C Epitope of *Brucella* O-Polysaccharide Revisited. *Clin Vaccine Immunol*, 22, 979-82.

## APPENDIX A

### The database sources of DNA and protein sequences

#### 1. Homo sapiens mitochondrion, complete genome NCBI Reference Sequence:

##### NC\_012920.1

LOCUS NC\_012920 16569 bp DNA circular PRI 31-OCT-2014  
DEFINITION Homo sapiens mitochondrion, complete genome.  
ACCESSION NC\_012920 AC\_000021  
VERSION NC\_012920.1 GI: 251831106  
DBLINK BioProject: [PRJNA30353](#)  
KEYWORDS RefSeq.  
SOURCE mitochondrion Homo sapiens (human)  
ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
Catarrhini; Hominidae; Homo.  
REFERENCE 1 (bases 1 to 16569)  
AUTHORS Andrews,R.M., Kubacka,I., Chinnery,P.F., Lightowlers,R.N.,  
Turnbull,D.M. and Howell,N.  
TITLE Reanalysis and revision of the Cambridge reference sequence for  
human mitochondrial DNA  
JOURNAL Nat. Genet. 23 (2), 147 (1999)  
PUBMED [10508508](#)  
REFERENCE 2 (bases 324 to 743)  
AUTHORS Andrews,R.M., Kubacka,I., Chinnery,P.F., Lightowlers,R.N.,  
Turnbull,D.M. and Howell,N.  
TITLE Reanalysis and revision of the Cambridge reference sequence for  
human mitochondrial DNA  
JOURNAL Nat. Genet. 23 (2), 147 (1999)  
PUBMED [10508508](#)  
REFERENCE 3 (bases 1 to 16569)  
AUTHORS Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H.,  
Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A.,  
Sanger,F., Schreier,P.H., Smith,A.J., Staden,R. and Young,I.G.  
TITLE Sequence and organization of the human mitochondrial genome  
JOURNAL Nature 290 (5806), 457-465 (1981)  
PUBMED [7219534](#)  
REFERENCE 4 (bases 15888 to 15954)  
AUTHORS Anderson,S., Bankier,A.T., Barrell,B.G., de Bruijn,M.H.,  
Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A.,  
Sanger,F., Schreier,P.H., Smith,A.J., Staden,R. and Young,I.G.  
TITLE Sequence and organization of the human mitochondrial genome  
JOURNAL Nature 290 (5806), 457-465 (1981)  
PUBMED [7219534](#)

REFERENCE 5 (bases 1 to 16569)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (08-JUL-2009) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 6 (bases 1 to 16569)  
 AUTHORS Kogelnik,A.M. and Lott,M.T.  
 TITLE Direct Submission  
 JOURNAL Submitted (24-AUG-2006) Mitomap.org, Center for Molecular and Mitochondrial Medicine and Genetics (MAMMAG) University of California, University of California, Irvine, Irvine, CA 92697-3940, USA  
 REMARK Sequence update by submitter  
 REFERENCE 7 (bases 1 to 16569)  
 AUTHORS Kogelnik,A.M. and Lott,M.T.  
 TITLE Direct Submission  
 JOURNAL Submitted (18-APR-1997) Center for Molecular Medicine, Emory University School of Medicine, 1462 Clifton Road, Suite 420, Atlanta, GA 30322, USA

## 2. Pan troglodytes mitochondrion, complete genome NCBI Reference Sequence: NC\_001643.1

LOCUS NC\_001643 16554 bp DNA circular PRI 01-FEB-2010  
 DEFINITION Pan troglodytes mitochondrion, complete genome.  
 ACCESSION NC\_001643  
 VERSION NC\_001643.1 GI:5835121  
 DBLINK Project: [10627](#)  
 BioProject: [PRJNA10627](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion Pan troglodytes (chimpanzee)  
 ORGANISM [Pan troglodytes](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
 Catarrhini; Hominidae; Pan.  
 REFERENCE 1 (sites)  
 AUTHORS Horai,S., Hayasaka,K., Kondo,R., Tsugane,K. and Takahata,N.  
 TITLE Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs  
 JOURNAL Proc. Natl. Acad. Sci. U.S.A. 92 (2), 532-536 (1995)  
 PUBMED [7530363](#)  
 REFERENCE 2 (sites)  
 AUTHORS Horai,S., Satta,Y., Hayasaka,K., Kondo,R., Inoue,T., Ishida,T., Hayashi,S. and Takahata,N.  
 TITLE Man's place in Hominoidea revealed by mitochondrial DNA genealogy  
 JOURNAL J. Mol. Evol. 35 (1), 32-43 (1992)  
 PUBMED [1518083](#)  
 REFERENCE 3 (sites)  
 AUTHORS Foran,D.R., Hixson,J.E. and Brown,W.M.  
 TITLE Comparisons of ape and human sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis  
 JOURNAL Nucleic Acids Res. 16 (13), 5841-5861 (1988)  
 PUBMED [3399380](#)  
 REFERENCE 4 (sites)

AUTHORS Hixson, J.E. and Brown, W.M.  
 TITLE A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and humans: sequence, structure, evolution, and phylogenetic implications  
 JOURNAL Mol. Biol. Evol. 3 (1), 1-18 (1986)  
 PUBMED [3444394](#)  
 REFERENCE 5 (bases 1 to 16554)  
 CONSRM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (08-SEP-1999) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 6 (bases 1 to 16554)  
 AUTHORS Hayasaka, K.  
 TITLE Direct Submission  
 JOURNAL Submitted (02-SEP-1994) Human Genetics, National Institute of Genetics, 1,111 Yata, Mishima, Shizuoka 411, Japan  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The reference sequence was derived from [D38113](#).  
 COMPLETENESS: full length.

### 3. Gorilla gorilla mitochondrion, complete genome NCBI Reference Sequence: NC\_011120.1

LOCUS NC\_011120 16412 bp DNA linear PRI 14-FEB-2011  
 DEFINITION Gorilla gorilla gorilla mitochondrion, complete genome.  
 ACCESSION NC\_011120  
 VERSION NC\_011120.1 GI:195952353  
 DBLINK Project: [62967](#)  
 BioProject: [PRJNA62967](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion Gorilla gorilla gorilla (western lowland gorilla)  
 ORGANISM [Gorilla gorilla gorilla](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
 Catarrhini; Hominidae; Gorilla.  
 REFERENCE 1 (bases 1 to 16412)  
 AUTHORS Xu, X. and Arnason, U.  
 TITLE A complete sequence of the mitochondrial genome of the western lowland gorilla  
 JOURNAL Mol. Biol. Evol. 13 (5), 691-698 (1996)  
 PUBMED [8676744](#)  
 REFERENCE 2 (bases 1 to 16412)  
 CONSRM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (06-AUG-2008) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 16412)  
 AUTHORS Arnason, U.  
 TITLE Direct Submission  
 JOURNAL Submitted (16-NOV-1995) U. Arnason, Dept of Genetics, Division Evolutionary Molec. Systematics, University of Lund, Solvegatan 29,  
 S-223 62 LUND, SWEDEN  
 COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to final

NCBI review. The reference sequence was derived from [X93347](#).  
COMPLETENESS: full length.

#### 4. *Bos taurus* mitochondrion, complete genome NCBI Reference Sequence: NC\_006853.1

LOCUS NC\_006853 16338 bp DNA circular MAM 15-APR-2009  
DEFINITION *Bos taurus* mitochondrion, complete genome.  
ACCESSION NC\_006853  
VERSION NC\_006853.1 GI:60101824  
DBLINK Project: [13366](#)  
BioProject: [PRJNA13366](#)  
KEYWORDS RefSeq.  
SOURCE mitochondrion *Bos taurus* (cattle)  
ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.  
REFERENCE 1 (bases 1 to 16338)  
AUTHORS Chung,H.Y. and Ha,J.M.  
TITLE Haplotype analysis of mitochondrial DNA in Korean native cattle  
JOURNAL Unpublished  
REFERENCE 2 (bases 1 to 16338)  
CONSRTM NCBI Genome Project  
TITLE Direct Submission  
JOURNAL Submitted (22-FEB-2005) National Center for Biotechnology  
Information, NIH, Bethesda, MD 20894, USA  
REFERENCE 3 (bases 1 to 16338)  
AUTHORS Chung,H.Y. and Ha,J.M.  
TITLE Direct Submission  
JOURNAL Submitted (06-JAN-2004) Animal Genomics & Bioinformatics,  
National  
Livestock Research Institute, Omokchon don, Suwon, KY 441701,  
Korea  
COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
The  
reference sequence was derived from [AY526085](#).  
On Feb 27, 2006 this sequence version replaced gi:[60101823](#).  
COMPLETENESS: full length.

#### 5. *Bubalus bubalis* mitochondrion, complete genome NCBI Reference Sequence:

NC\_006295.1

LOCUS NC\_006295 16359 bp DNA circular MAM 01-FEB-2010  
DEFINITION *Bubalus bubalis* mitochondrion, complete genome.  
ACCESSION NC\_006295

VERSION NC\_006295.1 GI:52220982  
 DBLINK Project: [13052](#)  
 BioProject: [PRJNA13052](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion Bubalus bubalis (Swamp buffalo)  
 ORGANISM [Bubalus bubalis](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
 Ruminantia;  
 Pecora; Bovidae; Bovinae; Bubalus.  
 REFERENCE 1 (bases 1 to 16359)  
 AUTHORS Qian,J.X., Dong,K.J., Huang,Y.J., Yang,B.Z., He,M., Liu,Z.J.  
 and  
 Li,J.  
 TITLE Complete sequence of Bubalus bubalis mitochondrial DNA  
 JOURNAL Unpublished  
 REFERENCE 2 (bases 1 to 16359)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (17-SEP-2004) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 16359)  
 AUTHORS Qian,J.X., Dong,K.J., Huang,Y.J., Yang,B.Z., He,M., Liu,Z.J.  
 and  
 Li,J.  
 TITLE Direct Submission  
 JOURNAL Submitted (02-AUG-2004) Transgenic Laboratory, Hainan Medical  
 College, Chengxi Road, Haikou, Hainan 571101, China  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The  
 reference sequence was derived from [AY702618](#).  
 COMPLETENESS: full length.

## 6. Bison mitochondrion, complete genome NCBI Reference Sequence: NC\_012346.1

LOCUS NC\_012346 16319 bp DNA circular MAM 13-APR-  
 2009  
 DEFINITION Bison bison mitochondrion, complete genome.  
 ACCESSION NC\_012346  
 VERSION NC\_012346.1 GI:225622211  
 DBLINK Project: [36339](#)  
 BioProject: [PRJNA36339](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion Bison bison (American bison)  
 ORGANISM [Bison bison](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
 Ruminantia;  
 Pecora; Bovidae; Bovinae; Bison.  
 REFERENCE 1 (bases 1 to 16319)  
 AUTHORS Achilli,A., Olivieri,A., Pellecchia,M., Uboldi,C., Colli,L.,  
 Al-Zahery,N., Accetturo,M., Pala,M., Kashani,B.H., Perego,U.A.,  
 Battaglia,V., Fornarino,S., Kalamati,J., Houshmand,M.,  
 Negrini,R.,  
 Semino,O., Richards,M., Macaulay,V., Ferretti,L., Bandelt,H.J.,

Ajmone-Marsan,P. and Torroni,A.  
 TITLE Mitochondrial genomes of extinct aurochs survive in domestic  
 cattle  
 JOURNAL Curr. Biol. 18 (4), R157-R158 (2008)  
 PUBMED [18302915](#)  
 REFERENCE 2 (bases 1 to 16319)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (09-APR-2009) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 16319)  
 AUTHORS Achilli,A., Olivieri,A., Pellecchia,M., Uboldi,C., Colli,L.,  
 Al-Zahery,N., Accetturo,M., Pala,M., Hooshiar  
 Kashani,B.H.B.H.B.,  
 Perego,U.A., Battaglia,V., Fornarino,S., Houshmand,M.,  
 Negrini,R.,  
 Semino,O., Richards,M., Macaulay,V., Ferretti,L., Bandelt,H.-J.  
 Jr., Ajmone-Marsan,P. and Torroni,A.  
 TITLE Direct Submission  
 JOURNAL Submitted (26-SEP-2007) Dipartimento di Genetica e  
 Microbiologia,  
 University of Pavia, Via Ferrata 1, Pavia 27100, Italy  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The  
 reference sequence was derived from [EU177871](#).  
 COMPLETENESS: full length.

## 7. Camelus dromedarius mitochondrion, complete genome NCBI Reference Sequence:

### NC\_009849.1

LOCUS NC\_009849 16643 bp DNA circular MAM 14-APR-  
 2009  
 DEFINITION Camelus dromedarius mitochondrion, complete genome.  
 ACCESSION NC\_009849  
 VERSION NC\_009849.1 GI:157690784  
 DBLINK Project: [20873](#)  
 BioProject: [PRJNA20873](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion Camelus dromedarius (Arabian camel)  
 ORGANISM [Camelus dromedarius](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Tylopoda;  
 Camelidae; Camelus.  
 REFERENCE 1 (bases 1 to 16643)  
 AUTHORS Huang,X., Shah,R.S. and Khazanehdari,K.A.  
 TITLE Complete nucleotide sequence of mitochondrial genome of the  
 dromedary camel, Camelus dromedarius: Structure and the control  
 region  
 JOURNAL Unpublished  
 REFERENCE 2 (bases 1 to 16643)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (27-SEP-2007) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA

REFERENCE 3 (bases 1 to 16643)  
AUTHORS Huang,X.  
TITLE Direct Submission  
JOURNAL Submitted (17-SEP-2007) Molecular Biology & Genetics, Central Veterinary Research Laboratory, Dubai P.O.Box 597, United Arab Emirates

COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
The reference sequence was derived from [EU159113](#).  
COMPLETENESS: full length.

## 8. Camelus bactrianus mitochondrion, complete genome NCBI Reference Sequence: NC\_009628.2

LOCUS NC\_009628 16659 bp DNA circular MAM 29-JUL-2009

DEFINITION Camelus bactrianus mitochondrion, complete genome.  
ACCESSION NC\_009628  
VERSION NC\_009628.2 GI:157011955  
DBLINK Project: [19999](#)  
BioProject: [PRJNA19999](#)

KEYWORDS RefSeq.  
SOURCE mitochondrion Camelus bactrianus (Bactrian camel)  
ORGANISM [Camelus bactrianus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Tylopoda;  
Camelidae; Camelus.

REFERENCE 1 (bases 1 to 16659)  
AUTHORS Ji,R., Cui,P., Ding,F., Geng,J., Gao,H., Zhang,H., Yu,J., Hu,S. and Meng,H.  
TITLE Monophyletic origin of domestic bactrian camel (Camelus bactrianus) and its evolutionary relationship with the extant wild camel (Camelus bactrianus ferus)  
JOURNAL Anim. Genet. 40 (4), 377-382 (2009)  
PUBMED [19292708](#)

REFERENCE 2 (bases 1 to 16659)  
AUTHORS Cui,P., Ji,R., Ding,F., Qi,D., Gao,H., Meng,H., Yu,J., Hu,S. and Zhang,H.  
TITLE A complete mitochondrial genome sequence of the wild two-humped camel (Camelus bactrianus ferus): an evolutionary history of camelidae  
JOURNAL BMC Genomics 8, 241 (2007)  
PUBMED [17640355](#)  
REMARK Publication Status: Online-Only

REFERENCE 3 (bases 1 to 16659)  
CONSRM NCBI Genome Project  
TITLE Direct Submission  
JOURNAL Submitted (03-JUL-2007) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA

REFERENCE 4 (bases 1 to 16659)  
AUTHORS Ji,R., Cui,P., Gao,H., Meng,H., Hu,S. and Zhang,H.  
TITLE Direct Submission  
JOURNAL Submitted (09-JAN-2007) College of Food Science and Engineering, Laboratory of Dairy Biotechnology and Engineering Ministry of



Education, Inner Mongolia Agricultural University, Zhaowuda  
 Road 306, Huhhot, Inner Mongolia 010018, China  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The reference sequence was derived from [EF212037](#).  
 On Sep 6, 2007 this sequence version replaced gi:[150375649](#).  
 COMPLETENESS: full length.

**9. Equus caballus mitochondrion, complete genome NCBI Reference Sequence:  
 NC\_001640.1**

LOCUS NC\_001640 16660 bp DNA circular MAM 01-FEB-2010  
 DEFINITION Equus caballus mitochondrion, complete genome.  
 ACCESSION NC\_001640  
 VERSION NC\_001640.1 GI:5835107  
 DBLINK Project: [19129](#)  
 BioProject: [PRJNA19129](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion Equus caballus (horse)  
 ORGANISM [Equus caballus](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Perissodactyla; Equidae;  
 Equus.  
 REFERENCE 1 (bases 1 to 16660)  
 AUTHORS Xu,X. and Arnason,U.  
 TITLE The complete mitochondrial DNA sequence of the horse, Equus caballus: extensive heteroplasmy of the control region  
 JOURNAL Gene 148 (2), 357-362 (1994)  
 PUBMED [7958969](#)  
 REFERENCE 2 (bases 1 to 16660)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (28-OCT-1999) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 16660)  
 AUTHORS Arnason,U.  
 TITLE Direct Submission  
 JOURNAL Submitted (06-JUN-1994) University of Lund, Dept. of Genetics, Division Evolutionary Molec. Systematics, Solvegatan 29, 223 62 Lund, Sweden  
 COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to final  
 well  
 NCBI review. The reference sequence was derived from [X79547](#). Users are requested to refer to the citation of this entry as  
 as the accession number in their publications.  
 COMPLETENESS: full length.

**10. Ovis aries mitochondrion, complete genome NCBI Reference Sequence: NC\_001941.1**

LOCUS NC\_001941 16616 bp DNA circular MAM 01-FEB-2010  
 DEFINITION Ovis aries mitochondrion, complete genome.

ACCESSION NC\_001941  
 VERSION NC\_001941.1 GI:5835554  
 DBLINK Project: [10764](#)  
 BioProject: [PRJNA10764](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion *Ovis aries* (sheep)  
 ORGANISM [Ovis aries](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
 Ruminantia;  
 Pecora; Bovidae; Caprinae; *Ovis*.  
 REFERENCE 1 (bases 1 to 16616)  
 AUTHORS Hiendleder,S., Lewalski,H., Wassmuth,R. and Janke,A.  
 TITLE The complete mitochondrial DNA sequence of the domestic sheep  
 (*Ovis aries*) and comparison with the other major ovine haplotype  
 JOURNAL J. Mol. Evol. 47 (4), 441-448 (1998)  
 PUBMED [9767689](#)  
 REFERENCE 2 (bases 1 to 16616)  
 AUTHORS Hiendleder,S.  
 TITLE A low rate of replacement substitutions in two major *Ovis aries*  
 mitochondrial genomes  
 JOURNAL Anim. Genet. 29 (2), 116-122 (1998)  
 PUBMED [9699271](#)  
 REFERENCE 3 (bases 1 to 16616)  
 AUTHORS Hiendleder,S., Mainz,K., Plante,Y. and Lewalski,H.  
 TITLE Analysis of mitochondrial DNA indicates that domestic sheep are  
 derived from two different ancestral maternal sources: no  
 evidence  
 for contributions from urial and argali sheep  
 JOURNAL J. Hered. 89 (2), 113-120 (1998)  
 PUBMED [9542158](#)  
 REFERENCE 4 (bases 1 to 16616)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (12-JUL-2004) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 5 (bases 1 to 16616)  
 AUTHORS Hiendleder,S., Wassmuth,R. and Lewalski,H.  
 TITLE Direct Submission  
 JOURNAL Submitted (19-AUG-1998) Animal Breeding and Genetics,  
 Justus-Liebig-University, Ludwigstr. 21B, Giessen 35390,  
 Germany  
 REMARK Sequence update by submitter  
 REFERENCE 6 (bases 1 to 16616)  
 AUTHORS Hiendleder,S., Wassmuth,R. and Lewalski,H.  
 TITLE Direct Submission  
 JOURNAL Submitted (26-JUN-1997) Animal Breeding and Genetics,  
 Justus-Liebig-University, Ludwigstr. 21B, Giessen 35390,  
 Germany  
 COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to  
 final  
 NCBI review. The reference sequence was derived from [AF010406](#).  
 COMPLETENESS: full length.

## 11. *Capra hircus* mitochondrion, complete genome NCBI Reference Sequence:

NC\_005044.2

LOCUS NC\_005044 16643 bp DNA circular MAM 05-JAN-2011  
DEFINITION *Capra hircus* mitochondrion, complete genome.  
ACCESSION NC\_005044  
VERSION NC\_005044.2 GI:316926505  
DBLINK Project: [12170](#)  
BioProject: [PRJNA12170](#)  
KEYWORDS RefSeq.  
SOURCE mitochondrion *Capra hircus* (goat)  
ORGANISM [Capra hircus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Caprinae; Capra.  
REFERENCE 1 (bases 1 to 16643)  
AUTHORS Hassanin,A., Bonillo,C., Nguyen,B.X. and Cruaud,C.  
TITLE Comparisons between mitochondrial genomes of domestic goat  
(*Capra hircus*) reveal the presence of numts and multiple sequencing  
errors  
JOURNAL Mitochondrial DNA 21 (3-4), 68-76 (2010)  
PUBMED [20540682](#)  
REFERENCE 2 (bases 1 to 16643)  
CONSRM NCBI Genome Project  
TITLE Direct Submission  
JOURNAL Submitted (04-JAN-2011) National Center for Biotechnology  
Information, NIH, Bethesda, MD 20894, USA  
REFERENCE 3 (bases 1 to 16643)  
AUTHORS Hassanin,A. and Cruaud,C.  
TITLE Direct Submission  
JOURNAL Submitted (09-DEC-2009) Systematique & Evolution, MNHN, 55, rue  
Buffon, Paris 75005, France  
COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to  
final  
NCBI review. The reference sequence is identical to [GU295658](#).  
On Jan 4, 2011 this sequence version replaced gi:[33285125](#).  
COMPLETENESS: full length.

## 12. *Sus scrofa* mitochondrion, complete genome NCBI Reference Sequence: NC\_000845.1

LOCUS NC\_000845 16613 bp DNA circular MAM 24-OCT-2013  
DEFINITION *Sus scrofa* mitochondrion, complete genome.  
ACCESSION NC\_000845  
VERSION NC\_000845.1 GI:5835862  
DBLINK BioProject: [PRJNA28993](#)  
KEYWORDS RefSeq.  
SOURCE mitochondrion *Sus scrofa* (pig)  
ORGANISM [Sus scrofa](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;

Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Suina;  
 Suidae;  
 Sus.  
 REFERENCE 1 (bases 1 to 16613)  
 AUTHORS Lin,C.S., Sun,Y.L., Liu,C.Y., Yang,P.C., Chang,L.C.,  
 Cheng,I.C.,  
 Mao,S.J. and Huang,M.C.  
 TITLE Complete nucleotide sequence of pig (*Sus scrofa*) mitochondrial  
 genome and dating evolutionary divergence within Artiodactyla  
 JOURNAL Gene 236 (1), 107-114 (1999)  
 PUBMED [10433971](#)  
 REFERENCE 2 (bases 1 to 16613)  
 AUTHORS Lin,C.S., Liu,C.Y., Wu,H.T., Sun,Y.L., Chang,L.C., Yen,N.T.,  
 Yang,P.C., Huang,M.C. and Mao,S.J.T.  
 TITLE SSCP analysis in the D-loop region of porcine mitochondrial DNA  
 as  
 confirmed by sequence diversity  
 JOURNAL J. Anim. Breed. Genet. 115, 73-78 (1998)  
 REFERENCE 3 (bases 1 to 16613)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (08-SEP-1999) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 4 (bases 1 to 16613)  
 AUTHORS Lin,C.S.  
 TITLE Direct Submission  
 JOURNAL Submitted (12-NOV-1997) Comparative Medicine, Pig Research  
 Institute Taiwan, P.O. Box 23, Chunan, Miaoli 350, Taiwan, ROC  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The  
 reference sequence was derived from [AF034253](#).

### 13. *Gallus gallus* mitochondrion, complete genome NCBI Reference Sequence:

#### NC\_001323.1

LOCUS NC\_001323 16775 bp DNA circular VRT 14-APR-  
 2009  
 DEFINITION *Gallus gallus* mitochondrion, complete genome.  
 ACCESSION NC\_001323  
 VERSION NC\_001323.1 GI:5834843  
 DBLINK Project: [10808](#)  
 BioProject: [PRJNA10808](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion *Gallus gallus* (chicken)  
 ORGANISM [Gallus gallus](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;  
 Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes;  
 Phasianidae; Phasianinae; Gallus.  
 REFERENCE 1 (bases 1 to 16775)  
 AUTHORS Valverde,J.R., Marco,R. and Garesse,R.  
 TITLE A conserved heptamer motif for ribosomal RNA transcription  
 termination in animal mitochondria  
 JOURNAL Proc. Natl. Acad. Sci. U.S.A. 91 (12), 5368-5371 (1994)  
 PUBMED [7515499](#)  
 REFERENCE 2 (bases 1 to 16775)  
 AUTHORS Desjardins,P. and Morais,R.

TITLE Sequence and gene organization of the chicken mitochondrial genome.  
 A novel gene order in higher vertebrates  
 JOURNAL J. Mol. Biol. 212 (4), 599-634 (1990)  
 PUBMED [2329578](#)  
 REFERENCE 3 (bases 1 to 16775)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (13-SEP-2005) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 4 (bases 1 to 16775)  
 AUTHORS Morais,R.  
 TITLE Direct Submission  
 JOURNAL Submitted (03-APR-1990) Morais R., Departement de Biochemie, Universite de Montreal, C.P. 6128, Succ. A, Montreal (Quebec), H3C #j7, Canada  
 COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to final NCBI review. The reference sequence was derived from [X52392](#).  
 COMPLETENESS: full length.

#### 14. *Oryctolagus cuniculus* mitochondrion, complete genome NCBI Reference Sequence:

NC\_001913.1

LOCUS NC\_001913 17245 bp DNA circular MAM 01-FEB-2010  
 DEFINITION *Oryctolagus cuniculus* mitochondrion, complete genome.  
 ACCESSION NC\_001913  
 VERSION NC\_001913.1 GI:5835526  
 DBLINK Project: [11845](#)  
 BioProject: [PRJNA11845](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion *Oryctolagus cuniculus* (rabbit)  
 ORGANISM [Oryctolagus cuniculus](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Glires; Lagomorpha;  
 Leporidae; *Oryctolagus*.  
 REFERENCE 1 (bases 1 to 17245)  
 AUTHORS Gissi,C., Gullberg,A. and Arnason,U.  
 TITLE The complete mitochondrial DNA sequence of the rabbit, *Oryctolagus cuniculus*  
 JOURNAL Genomics 50 (2), 161-169 (1998)  
 PUBMED [9653643](#)  
 REFERENCE 2 (bases 1 to 17245)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (08-SEP-1999) National Center for Biotechnology Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 17245)  
 AUTHORS Gissi,C.

TITLE Direct Submission  
 JOURNAL Submitted (17-SEP-1997) Gissi C., Dept. of Genetics, Division  
 of Evolutionary Systematics, University of Lund, Solvegatan 29,  
 Lund,  
 223 62, SWEDEN  
 COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to  
 final NCBI review. The reference sequence was derived from [AJ001588](#).  
 COMPLETENESS: full length.

### 15. *Canis lupus familiaris* mitochondrion, complete genome NCBI Reference Sequence:

#### NC\_002008.4

LOCUS NC\_002008 16727 bp DNA circular MAM 14-APR-  
 2009  
 DEFINITION *Canis lupus familiaris* mitochondrion, complete genome.  
 ACCESSION NC\_002008  
 VERSION NC\_002008.4 GI:17737322  
 DBLINK Project: [12384](#)  
 BioProject: [PRJNA12384](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion *Canis lupus familiaris* (dog)  
 ORGANISM [Canis lupus familiaris](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Carnivora; Caniformia;  
 Canidae;  
 Canis.  
 REFERENCE 1 (bases 1 to 16727)  
 AUTHORS Kim,K.S., Lee,S.E., Jeong,H.W. and Ha,J.H.  
 TITLE The complete nucleotide sequence of the domestic dog (*Canis*  
*familiaris*) mitochondrial genome  
 JOURNAL Mol. Phylogenet. Evol. 10 (2), 210-220 (1998)  
 PUBMED [9878232](#)  
 REFERENCE 2 (bases 1 to 16727)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (22-JUN-2007) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 16727)  
 AUTHORS Kim,K.S., Lee,S.E., Jeong,H.W., Jeong,S.Y., Sohn,H.S. and  
 Ha,J.H.  
 TITLE Direct Submission  
 JOURNAL Submitted (07-APR-1997) Genetic Engineering, Animal Genetics,  
 1370 Sankyuk-dong, Pukgu, Taegu 702-701, Korea  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The reference sequence was derived from [U96639](#).  
 On Dec 14, 2001 this sequence version replaced gi:[15805032](#).  
 COMPLETENESS: full length.

### 16. *Felis catus* mitochondrion, complete genome NCBI Reference Sequence: NC\_001700.1

LOCUS NC\_001700 17009 bp DNA circular MAM 21-APR-2009  
 DEFINITION *Felis catus* mitochondrion, complete genome.  
 ACCESSION NC\_001700  
 VERSION NC\_001700.1 GI:5835205  
 DBLINK Project: [10762](#)  
 BioProject: [PRJNA10762](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion *Felis catus* (domestic cat)  
 ORGANISM [Felis catus](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Carnivora; Feliformia;  
 Felidae;  
 Felinae; Felis.  
 REFERENCE 1 (bases 1 to 17009)  
 AUTHORS Lopez, J.V., Cevario, S. and O'Brien, S.J.  
 TITLE Complete nucleotide sequences of the domestic cat (*Felis catus*)  
 mitochondrial genome and a transposed mtDNA tandem repeat  
 (Numt) in  
 the nuclear genome  
 JOURNAL Genomics 33 (2), 229-246 (1996)  
 PUBMED [8660972](#)  
 REFERENCE 2 (bases 1 to 17009)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (08-SEP-1999) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 17009)  
 AUTHORS Lopez, J.V.  
 TITLE Direct Submission  
 JOURNAL Submitted (07-FEB-1995) Jose V. Lopez, Laboratory of Viral  
 Carcinogenesis, PRI/DynCorp, Biological Carcinogenesis and  
 Development Prog, Bldg 560, Room 11-21, NCI-Frederick Cancer  
 Research and Development Center, Frederick, MD 21702-1201, USA  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The  
 reference sequence was derived from [U20753](#).  
 COMPLETENESS: full length.

## 17. *Mus musculus* mitochondrion, complete genome NCBI Reference Sequence:

### NC\_005089.1

LOCUS NC\_005089 16299 bp DNA circular ROD 31-OCT-2014  
 DEFINITION *Mus musculus* mitochondrion, complete genome.  
 ACCESSION NC\_005089  
 VERSION NC\_005089.1 GI:34538597  
 DBLINK BioProject: [PRJNA169](#)  
 KEYWORDS RefSeq.  
 SOURCE mitochondrion *Mus musculus* (house mouse)  
 ORGANISM [Mus musculus](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;

Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.  
 REFERENCE 1 (bases 1 to 16299)  
 AUTHORS Bayona-Bafaluy, M.P., Acin-Perez, R., Mullikin, J.C., Park, J.S.,  
 Moreno-Loshuertos, R., Hu, P., Perez-Martos, A., Fernandez-  
 Silva, P.,  
 Bai, Y. and Enriquez, J.A.  
 TITLE Revisiting the mouse mitochondrial DNA sequence  
 JOURNAL Nucleic Acids Res. 31 (18), 5349-5355 (2003)  
 PUBMED [12954771](#)  
 REFERENCE 2 (bases 1 to 16299)  
 CONSRTM NCBI Genome Project  
 TITLE Direct Submission  
 JOURNAL Submitted (09-SEP-2003) National Center for Biotechnology  
 Information, NIH, Bethesda, MD 20894, USA  
 REFERENCE 3 (bases 1 to 16299)  
 AUTHORS Mullikin, J.C. and Enriquez, J.A.  
 TITLE Direct Submission  
 JOURNAL Submitted (04-NOV-2002) Bioquimica y Biologia Molecular y  
 Celular,  
 Universidad de Zaragoza, Miguel Servet, 177, Zaragoza 50013,  
 Espana  
 COMMENT REVIEWED [REFSEQ](#): This record has been curated by NCBI staff.  
 The  
 reference sequence was derived from [AY172335](#).  
 On Sep 11, 2003 this sequence version replaced gi:[5834953](#).  
 COMPLETENESS: full length.

## 18. *Brucella melitensis* biovar Abortus 2308 chromosome I, complete sequence, strain 2308

### NCBI Reference Sequence: NC\_007618.1

LOCUS NC\_007618 2121359 bp DNA circular CON 18-AUG-  
 2015  
 DEFINITION *Brucella melitensis* biovar Abortus 2308 chromosome I, complete  
 sequence, strain 2308.  
 ACCESSION NC\_007618  
 VERSION NC\_007618.1 GI:82698932  
 DBLINK BioProject: [PRJNA224116](#)  
 BioSample: [SAMEA3138256](#)  
 Assembly: [GCF\\_000054005.1](#)  
 KEYWORDS RefSeq; complete genome.  
 SOURCE *Brucella abortus* 2308  
 ORGANISM [Brucella abortus 2308](#)  
 Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;  
 Brucellaceae; *Brucella*.  
 REFERENCE 1 (bases 1 to 2121359)  
 AUTHORS Chain, P.S., Comerci, D.J., Tolmasky, M.E., Larimer, F.W.,  
 Malfatti, S.A., Vergez, L.M., Aguero, F., Land, M.L., Ugalde, R.A.  
 and  
 Garcia, E.  
 CONSRTM Microbial Genomics Group, Lawrence Livermore National  
 Laboratory,  
 and the Genome Analysis Group, Oak Ridge National Laboratory  
 TITLE Whole-genome analyses of speciation events in pathogenic  
 Brucellae  
 JOURNAL Infect. Immun. 73 (12), 8353-8361 (2005)  
 PUBMED [16299333](#)  
 REFERENCE 2 (bases 1 to 2121359)



AUTHORS Larimer, F.  
 CONSRTM Microbial Genomics Group, Lawrence Livermore National  
 Laboratory,  
 and the Genome Analysis Group, Oak Ridge National Laboratory  
 TITLE Direct Submission  
 JOURNAL Submitted (21-JUN-2006) Larimer F., Oak Ridge National  
 Laboratory,  
 1 Bethel Valley Road, Bldg 5700 A201 Oak Ridge, TN 37831, USA  
 COMMENT [REFSEQ INFORMATION](#): The reference sequence was derived from  
[AM040264](#).  
 Submitted on behalf of the Microbial Genomics Group, Lawrence  
 Livermore National Laboratory, and the Genome Analysis Group,  
 Oak  
 Ridge National Laboratory;  
 chain2@llnl.gov, larimerfw@ornl.gov.  
 Annotation was added by the NCBI Prokaryotic Genome Annotation  
 Pipeline (released 2013). Information about the Pipeline can be  
 found here: [http://www.ncbi.nlm.nih.gov/genome/annotation\\_prok/](http://www.ncbi.nlm.nih.gov/genome/annotation_prok/)

```

##Genome-Annotation-Data-START##
Annotation Provider      :: NCBI
Annotation Date         :: 08/18/2015 04:28:24
Annotation Pipeline     :: NCBI Prokaryotic

Genome
Annotation Method      :: Best-placed reference
                        protein set;
GeneMarkS+
Annotation Software revision :: 3.0
Features Annotated     :: Gene; CDS; rRNA;
                        tRNA;
                        ncRNA; repeat_region
Genes                  :: 3,185
CDS                    :: 3,084
Pseudo Genes          :: 33
rRNAs                  :: 3, 3, 3 (5S, 16S,
23S)
complete rRNAs        :: 3, 3, 3 (5S, 16S,
23S)
tRNAs                  :: 55
ncRNA                  :: 4
Frameshifted Genes     :: 20
Frameshifted Genes On Monomer Runs :: 3
Frameshifted Genes Not On Monomer Runs :: 4
##Genome-Annotation-Data-END##
COMPLETENESS: full length.

```

## 19. *Brucella melitensis* biovar Abortus 2308 chromosome II, complete sequence, strain 2308

### NCBI Reference Sequence: NC\_007624.1

LOCUS NC\_007624 1156948 bp DNA circular CON 18-AUG-2015  
 DEFINITION *Brucella melitensis* biovar Abortus 2308 chromosome II, complete  
 sequence, strain 2308.  
 ACCESSION NC\_007624  
 VERSION NC\_007624.1 GI:83268957  
 DBLINK BioProject: [PRJNA224116](#)  
 BioSample: [SAMEA3138256](#)

Assembly: [GCF\\_000054005.1](#)

KEYWORDS RefSeq; complete genome.

SOURCE *Brucella* abortus 2308

ORGANISM [Brucella abortus 2308](#)  
 Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales;  
 Brucellaceae; *Brucella*.

REFERENCE 1 (bases 1 to 1156948)

AUTHORS Chain,P.S., Comerici,D.J., Tolmasky,M.E., Larimer,F.W.,  
 Malfatti,S.A., Vergez,L.M., Agüero,F., Land,M.L., Ugalde,R.A.  
 and  
 Garcia,E.

CONSRTM Microbial Genomics Group, Lawrence Livermore National  
 Laboratory,  
 and the Genome Analysis Group, Oak Ridge National Laboratory

TITLE Whole-genome analyses of speciation events in pathogenic  
 Brucellae

JOURNAL Infect. Immun. 73 (12), 8353-8361 (2005)

PUBMED [16299333](#)

REFERENCE 2 (bases 1 to 1156948)

AUTHORS Larimer,F.

CONSRTM Microbial Genomics Group, Lawrence Livermore National  
 Laboratory,  
 and the Genome Analysis Group, Oak Ridge National Laboratory

TITLE Direct Submission

JOURNAL Submitted (21-JUN-2006) Larimer F., Oak Ridge National  
 Laboratory,  
 1 Bethel Valley Road, Bldg 5700 A201 Oak Ridge, TN 37831, USA

COMMENT [REFSEQ INFORMATION](#): The reference sequence was derived from  
[AM040265](#).  
 Submitted on behalf of the Microbial Genomics Group, Lawrence  
 Livermore National Laboratory, and the Genome Analysis Group,  
 Oak  
 Ridge National Laboratory;  
 chain2@llnl.gov, larimerfw@ornl.gov.  
 Annotation was added by the NCBI Prokaryotic Genome Annotation  
 Pipeline (released 2013). Information about the Pipeline can be  
 found here: [http://www.ncbi.nlm.nih.gov/genome/annotation\\_prok/](http://www.ncbi.nlm.nih.gov/genome/annotation_prok/)

##Genome-Annotation-Data-START##

Annotation Provider :: NCBI

Annotation Date :: 08/18/2015 04:28:24

Annotation Pipeline :: NCBI Prokaryotic

Genome

Annotation Method :: Annotation Pipeline  
 :: Best-placed reference  
 protein set;

GeneMarkS+

Annotation Software revision :: 3.0

Features Annotated :: Gene; CDS; rRNA;

tRNA;

ncRNA; repeat\_region

Genes :: 3,185

CDS :: 3,084

Pseudo Genes :: 33

rRNAs :: 3, 3, 3 (5S, 16S,

23S)

complete rRNAs :: 3, 3, 3 (5S, 16S,

23S)

tRNAs :: 55

ncRNA :: 4

Frameshifted Genes :: 20  
 Frameshifted Genes On Monomer Runs :: 3  
 Frameshifted Genes Not On Monomer Runs :: 4  
 ##Genome-Annotation-Data-END##  
 COMPLETENESS: full length.

20. *Brucella melitensis* bv. 1 str. 16M chromosome I, complete sequence NCBI Reference

Sequence: NC\_003317.1

LOCUS NC\_003317 2117144 bp DNA circular CON 30-JUL-2015  
 DEFINITION *Brucella melitensis* bv. 1 str. 16M chromosome I, complete sequence.  
 ACCESSION NC\_003317 NZ\_AE009444-NZ\_AE009638  
 VERSION NC\_003317.1 GI:17986284  
 DBLINK BioProject: [PRJNA224116](#)  
 BioSample: [SAMN02603416](#)  
 Assembly: [GCF\\_000007125.1](#)  
 KEYWORDS RefSeq.  
 SOURCE *Brucella melitensis* bv. 1 str. 16M  
 ORGANISM [Brucella melitensis](#) bv. 1 str. 16M  
 Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Brucellaceae; *Brucella*.  
 REFERENCE 1 (bases 1 to 2117144)  
 AUTHORS DelVecchio,V.G., Kapatral,V., Redkar,R.J., Patra,G., Mujer,C., Los,T., Ivanova,N., Anderson,I., Bhattacharyya,A., Lykidis,A., Reznik,G., Jablonski,L., Larsen,N., D'Souza,M., Bernal,A., Mazur,M., Goltsman,E., Selkov,E., Elzer,P.H., Hagius,S., O'Callaghan,D., Letesson,J.J., Haselkorn,R., Kyrpides,N. and Overbeek,R.  
 TITLE The genome sequence of the facultative intracellular pathogen *Brucella melitensis*  
 JOURNAL Proc. Natl. Acad. Sci. U.S.A. 99 (1), 443-448 (2002)  
 PUBMED [11756688](#)  
 REFERENCE 2 (bases 1 to 2117144)  
 AUTHORS DelVecchio,V.G., Redkar,R.J., Patra,G. and Mujer,C.  
 TITLE Direct Submission  
 JOURNAL Submitted (13-NOV-2001) Institute of Molecular Biology and Medicine, University of Scranton, Scranton, PA 18510, USA  
 REFERENCE 3 (bases 1 to 2117144)  
 AUTHORS Elzer,P.H. and Hagius,S.  
 TITLE Direct Submission  
 JOURNAL Submitted (13-NOV-2001) Department of Veterinary Science, LSU Ag Center, 111 Dalrymple Building, Baton Rouge, LA 70803, USA  
 REFERENCE 4 (bases 1 to 2117144)  
 AUTHORS Kapatral,V., Los,T., Ivanova,N., Anderson,I., Bhattacharyya,A., Lykidis,A., Reznik,G., Jablonski,L., Larsen,N., D'Souza,M., Bernal,A., Mazur,M., Goltsman,E., Selkov,E., Haselkorn,R., Kyrpides,N. and Overbeek,R.  
 TITLE Direct Submission  
 JOURNAL Submitted (13-NOV-2001) Integrated Genomics, Inc., 2201 W. Campbell Park Drive, IL 60612, USA  
 REFERENCE 5 (bases 1 to 2117144)  
 AUTHORS Letesson,J.-J.  
 TITLE Direct Submission

JOURNAL Submitted (13-NOV-2001) Unite de Recherche en Biologie  
Moleculaire,  
Laboratoire d'Immunologie et de Microbiologie, Universite of  
Namur,  
61 rue de Bruxelles, Namur 5000, Belgium

REFERENCE 6 (bases 1 to 2117144)

AUTHORS O'Callaghan,D.

TITLE Direct Submission

JOURNAL Submitted (13-NOV-2001) Faculte de Medecine, INSERM U431,  
Avenue  
Kennedy, Nimes 30900, France

COMMENT [REFSEQ INFORMATION](#): The reference sequence was derived from  
[AE008917](#).  
Annotation was added by the NCBI Prokaryotic Genome Annotation  
Pipeline (released 2013). Information about the Pipeline can be  
found here: [http://www.ncbi.nlm.nih.gov/genome/annotation\\_prok/](http://www.ncbi.nlm.nih.gov/genome/annotation_prok/)

##Genome-Annotation-Data-START##

Genome	Annotation Provider	:: NCBI
	Annotation Date	:: 07/30/2015 14:12:55
	Annotation Pipeline	:: NCBI Prokaryotic
	Annotation Method	:: Best-placed reference protein set;
GeneMarkS+	Annotation Software revision	:: 3.0
	Features Annotated	:: Gene; CDS; rRNA;
tRNA;		ncRNA; repeat_region
	Genes	:: 3,149
	CDS	:: 2,972
	Pseudo Genes	:: 113
	rRNAs	:: 3, 3, 3 (5S, 16S,
23S)	complete rRNAs	:: 3, 3, 3 (5S, 16S,
23S)	tRNAs	:: 54
	ncRNA	:: 1
	Frameshifted Genes	:: 87
	Frameshifted Genes On Monomer Runs	:: 15
	Frameshifted Genes Not On Monomer Runs	:: 18
	##Genome-Annotation-Data-END##	
	COMPLETENESS: full length.	

## 21. *Brucella melitensis* 16M chromosome II, complete sequence NCBI Reference Sequence:

### NC\_003318.1

LOCUS NC\_003318 1177787 bp DNA circular CON 30-JUL-2015

DEFINITION *Brucella melitensis* 16M chromosome II, complete sequence.

ACCESSION NC\_003318 NZ\_AE009639-NZ\_AE009745

VERSION NC\_003318.1 GI:17988344

DBLINK BioProject: [PRJNA224116](#)  
BioSample: [SAMN02603416](#)  
Assembly: [GCF\\_000007125.1](#)

KEYWORDS RefSeq.

SOURCE *Brucella melitensis* bv. 1 str. 16M

ORGANISM [Brucella melitensis](#) bv. 1 str. 16M  
 Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; Brucellaceae; *Brucella*.

REFERENCE 1 (bases 1 to 1177787)

AUTHORS DelVecchio,V.G., Kapatral,V., Redkar,R.J., Patra,G., Mujer,C., Los,T., Ivanova,N., Anderson,I., Bhattacharyya,A., Lykidis,A., Reznik,G., Jablonski,L., Larsen,N., D'Souza,M., Bernal,A., Mazur,M., Goltsman,E., Selkov,E., Elzer,P.H., Hagijs,S., O'Callaghan,D., Letesson,J.J., Haselkorn,R., Kyrpides,N. and Overbeek,R.

TITLE The genome sequence of the facultative intracellular pathogen *Brucella melitensis*

JOURNAL Proc. Natl. Acad. Sci. U.S.A. 99 (1), 443-448 (2002)

PUBMED [11756688](#)

REFERENCE 2 (bases 1 to 1177787)

AUTHORS DelVecchio,V.G., Redkar,R.J., Patra,G. and Mujer,C.

TITLE Direct Submission

JOURNAL Submitted (13-NOV-2001) Institute of Molecular Biology and Medicine, University of Scranton, Scranton, PA 18510, USA

REFERENCE 3 (bases 1 to 1177787)

AUTHORS Elzer,P.H. and Hagijs,S.

TITLE Direct Submission

JOURNAL Submitted (13-NOV-2001) Department of Veterinary Science, LSU Ag Center, 111 Dalrymple Building, Baton Rouge, LA 70803, USA

REFERENCE 4 (bases 1 to 1177787)

AUTHORS Kapatral,V., Los,T., Ivanova,N., Anderson,I., Bhattacharyya,A., Lykidis,A., Reznik,G., Jablonski,L., Larsen,N., D'Souza,M., Bernal,A., Mazur,M., Goltsman,E., Selkov,E., Haselkorn,R., Kyrpides,N. and Overbeek,R.

TITLE Direct Submission

JOURNAL Submitted (13-NOV-2001) Integrated Genomics, Inc., 2201 W. Campbell Park Drive, IL 60612, USA

REFERENCE 5 (bases 1 to 1177787)

AUTHORS Letesson,J.-J.

TITLE Direct Submission

JOURNAL Submitted (13-NOV-2001) Unite de Recherche en Biologie Moleculaire, Laboratoire d'Immunologie et de Microbiologie, Universite of Namur, 61 rue de Bruxelles, Namur 5000, Belgium

REFERENCE 6 (bases 1 to 1177787)

AUTHORS O'Callaghan,D.

TITLE Direct Submission

JOURNAL Submitted (13-NOV-2001) Faculte de Medecine, INSERM U431, Avenue Kennedy, Nimes 30900, France

COMMENT [REFSEQ INFORMATION](#): The reference sequence was derived from [AE008918](#).  
 Annotation was added by the NCBI Prokaryotic Genome Annotation Pipeline (released 2013). Information about the Pipeline can be found here: [http://www.ncbi.nlm.nih.gov/genome/annotation\\_prok/](http://www.ncbi.nlm.nih.gov/genome/annotation_prok/)

##Genome-Annotation-Data-START##  
 Annotation Provider :: NCBI  
 Annotation Date :: 07/30/2015 14:12:55  
 Annotation Pipeline :: NCBI Prokaryotic

Genome Annotation Pipeline

Annotation Method :: Best-placed reference protein set;  
 GeneMarkS+ Annotation Software revision :: 3.0  
 Features Annotated :: Gene; CDS; rRNA;  
 tRNA; ncrRNA; repeat\_region  
 Genes :: 3,149  
 CDS :: 2,972  
 Pseudo Genes :: 113  
 rRNAs :: 3, 3, 3 (5S, 16S,  
 23S)  
 complete rRNAs :: 3, 3, 3 (5S, 16S,  
 23S)  
 tRNAs :: 54  
 ncrRNA :: 1  
 Frameshifted Genes :: 87  
 Frameshifted Genes On Monomer Runs :: 15  
 Frameshifted Genes Not On Monomer Runs :: 18  
 ##Genome-Annotation-Data-END##  
 COMPLETENESS: full length.

## 22. Bos taurus breed Hereford chromosome 2, Bos\_taurus\_UMD\_3.1.1, whole genome

### shotgun sequence NCBI Reference Sequence: AC\_000159.1

LOCUS AC\_000159 137060424 bp DNA linear CON 26-JAN-2016  
 DEFINITION Bos taurus breed Hereford chromosome 2, Bos\_taurus\_UMD\_3.1.1, whole genome shotgun sequence.  
 ACCESSION AC\_000159 GPC\_000000171  
 VERSION AC\_000159.1 GI:258513365  
 DBLINK BioProject: [PRJNA33843](#)  
 BioSample: [SAMN02898106](#)  
 Assembly: [GCF\\_000003055.6](#)  
 KEYWORDS WGS; RefSeq.  
 SOURCE Bos taurus (cattle)  
 ORGANISM [Bos taurus](#)  
 Euteleostomi; Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
 Ruminantia; Pecora; Bovidae; Bovinae; Bos.  
 REFERENCE 1 (bases 1 to 137060424)  
 AUTHORS Zimin,A.V., Delcher,A.L., Florea,L., Kelley,D.R., Schatz,M.C., Puiu,D., Hanrahan,F., Pertea,G., Van Tassell,C.P., Sonstegard,T.S., Marcais,G., Roberts,M., Subramanian,P., Yorke,J.A. and Salzberg,S.L.  
 TITLE A whole-genome assembly of the domestic cow, Bos taurus  
 JOURNAL Genome Biol. 10 (4), R42 (2009)  
 PUBMED [19393038](#)  
 COMMENT [REFSEQ INFORMATION](#): The reference sequence is identical to [GK000002.2](#).  
 Assembly Name: Bos\_taurus\_UMD\_3.1.1  
 The genomic sequence for this RefSeq record is from the whole genome reassembly released by the Center for Bioinformatics and

whole  
and  
using  
project  
and  
assemble  
The

Computational Biology, University of Maryland. The original genome shotgun project has the project accession DAAA00000000.2 was submitted in December 2009. The assembly was generated using genomic traces submitted by the bovine genome sequencing project (Project ID 12555). In this assembly synteny between the cow and human genomes and independent mapping data were used to assemble roughly 99% of the genome onto the 30 *Bos taurus* chromosomes. The *Bos\_taurus\_UMD\_3.1.1* version of the assembly was created by excluding 173 contaminant contigs from *Bos\_taurus\_UMD\_3.1*.

```
##Genome-Assembly-Data-START##
Assembly Provider      :: Center for Bioinformatics and
                        Computational Biology, University of
                        Maryland
Assembly Method       :: UMD Overlapper v. 2009; additional
                        processing
Assembly Name         :: Bos_taurus_UMD_3.1.1
Genome Coverage       :: 9x
Sequencing Technology :: Sanger
##Genome-Assembly-Data-END##

##Genome-Annotation-Data-START##
Annotation Provider   :: NCBI
Annotation Status     :: Full annotation
Annotation Version    :: Bos taurus Annotation Release
Annotation Pipeline   :: NCBI eukaryotic genome
                        pipeline
Annotation Software Version :: 6.5
Annotation Method     :: Best-placed RefSeq; Gnomon
Features Annotated    :: Gene; mRNA; CDS; ncRNA
##Genome-Annotation-Data-END##
```

[105](#)

annotation

### 23. *Bos taurus* solute carrier family 11 member 1 (SLC11A1) gene, complete cds GenBank:

DQ493965.1

```
LOCUS       DQ493965                10665 bp    DNA     linear   MAM 30-APR-
2007
DEFINITION  Bos taurus solute carrier family 11 member 1 (SLC11A1) gene,
complete cds.
ACCESSION  DQ493965
VERSION    DQ493965.1  GI:99030402
KEYWORDS   .
SOURCE     Bos taurus (cattle)
  ORGANISM Bos taurus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi;
            Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;
Ruminantia;
            Pecora; Bovidae; Bovinae; Bos.
```

REFERENCE 1 (bases 1 to 10665)  
 AUTHORS Martinez,R., Barrera,G., Dunner,S. and Canon,J.  
 TITLE Novel polymorphisms in the SLC11A1 gene detected by SSCP in Zebu and Colombian Creole cattle  
 JOURNAL Unpublished

REFERENCE 2 (bases 1 to 10665)  
 AUTHORS Martinez,R., Barrera,G., Dunner,S. and Canon,J.  
 TITLE Direct Submission  
 JOURNAL Submitted (17-APR-2006) Animal Genetic Resource Program, Colombian Corporation of Livestock Research CORPOICA, Km 14 Via Mosquera, Bogota, Cundinamarca 57, Colombia

#### 24. *Bos taurus* solute carrier 11A1 (SLC11A1) gene, complete cds GenBank: DQ848779.1

LOCUS DQ848779 10814 bp DNA linear MAM 14-JUL-2007  
 DEFINITION *Bos taurus* solute carrier 11A1 (SLC11A1) gene, complete cds.  
 ACCESSION DQ848779  
 VERSION DQ848779.1 GI:123979203  
 KEYWORDS .  
 SOURCE *Bos taurus* (cattle)  
 ORGANISM [Bos taurus](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia;  
 Pecora; Bovidae; Bovinae; Bos.

REFERENCE 1 (bases 1 to 10814)  
 AUTHORS Schutta,C.J., Feng,J., Niu,S., Crider,B.P., Adams,L.G. and Templeton,J.W.  
 TITLE Complete Genomic Sequence of Bovine SLC11A1 Isolated from a Bovine Genomic BAC Library  
 JOURNAL Unpublished

REFERENCE 2 (bases 1 to 10814)  
 AUTHORS Schutta,C.J., Feng,J., Niu,S., Crider,B.P., Adams,L.G. and Templeton,J.W.  
 TITLE Direct Submission  
 JOURNAL Submitted (13-JUL-2006) Veterinary Pathobiology, Texas A&M University, College Station, TX 77843-4467, USA

#### 25. *Bos taurus* natural resistance-associated macrophage protein 1 (SLC11A1) gene, complete cds GenBank: KR002419.1

LOCUS KR002419 13543 bp DNA linear MAM 08-SEP-2015  
 DEFINITION *Bos taurus* natural resistance-associated macrophage protein 1 (SLC11A1) gene, complete cds.  
 ACCESSION KR002419  
 VERSION KR002419.1 GI:924919500  
 KEYWORDS .  
 SOURCE *Bos taurus* (cattle)  
 ORGANISM [Bos taurus](#)



Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.

REFERENCE 1 (bases 1 to 13543)  
AUTHORS Zhang,Y., Zhang,B., Qin,B., Wang,Y. and Liu,K.  
TITLE Study of the association between Nramp1 Polymorphisms and susceptibility to tuberculosis in dairy cattle  
JOURNAL Unpublished

REFERENCE 2 (bases 1 to 13543)  
AUTHORS Zhang,Y., Zhang,B., Qin,B., Wang,Y. and Liu,K.  
TITLE Direct Submission  
JOURNAL Submitted (15-MAR-2015) Faculty Of Annimal Science and Technology,  
Yunnan Agricultural University, FengYuan Road, Beishi District, Kunming, Yunnan 650201, China

COMMENT ##Assembly-Data-START##  
Sequencing Technology :: Sanger dideoxy sequencing  
##Assembly-Data-END##

## 26. Bos taurus natural resistance-associated macrophage protein 1 (Nramp1) gene, complete

cds GenBank: KR002420.1

LOCUS KR002420 13543 bp DNA linear MAM 08-SEP-2015

DEFINITION Bos taurus natural resistance-associated macrophage protein 1 (Nramp1) gene, complete cds.

ACCESSION KR002420

VERSION KR002420.1 GI:924919502

KEYWORDS .

SOURCE Bos taurus (cattle)  
ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.

REFERENCE 1 (bases 1 to 13543)  
AUTHORS Zhang,B., Zhang,Y., Qin,B., Wang,Y. and Liu,K.  
TITLE Study of the association between Nramp1 Polymorphisms and susceptibility to tuberculosis in dairy cattle  
JOURNAL Unpublished

REFERENCE 2 (bases 1 to 13543)  
AUTHORS Zhang,B., Zhang,Y., Qin,B., Wang,Y. and Liu,K.  
TITLE Direct Submission  
JOURNAL Submitted (16-MAR-2015) Faculty Of Annimal Science and Technology,  
Yunnan Agricultural University, FengYuan Road, Beishi District, Kunming, Yunnan 650201, China

COMMENT ##Assembly-Data-START##  
Sequencing Technology :: Sanger dideoxy sequencing  
##Assembly-Data-END##

## 27. *Bos taurus* natural resistance-associated macrophage protein 1 (NRAMP1) gene,

complete cds GenBank: KR002421.1

LOCUS KR002421 13543 bp DNA linear MAM 08-SEP-2015  
DEFINITION *Bos taurus* natural resistance-associated macrophage protein 1 (NRAMP1) gene, complete cds.  
ACCESSION KR002421  
VERSION KR002421.1 GI:924919504  
KEYWORDS .  
SOURCE *Bos taurus* (cattle)  
ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.  
REFERENCE 1 (bases 1 to 13543)  
AUTHORS Shi,X., Zhang,Y., Qin,B., Wang,Y., Liu,K. and Zhang,B.  
TITLE Study of the association between Nramp1 Polymorphisms and susceptibility to tuberculosis in dairy cattle  
JOURNAL Unpublished  
REFERENCE 2 (bases 1 to 13543)  
AUTHORS Shi,X., Zhang,Y., Qin,B., Wang,Y., Liu,K. and Zhang,B.  
TITLE Direct Submission  
JOURNAL Submitted (16-MAR-2015) Faculty Of Animal Science and Technology,  
Yunnan Agricultural University, FengYuan Road, Beishi District, Kunming, Yunnan 650201, China  
COMMENT ##Assembly-Data-START##  
Sequencing Technology :: Sanger dideoxy sequencing  
##Assembly-Data-END##

## 28. natural resistance-associated macrophage protein 1 [*Bos taurus*] NCBI Reference

Sequence: NP\_777077.1

LOCUS NP\_777077 548 aa linear MAM 24-APR-2016  
DEFINITION natural resistance-associated macrophage protein 1 [*Bos taurus*].  
ACCESSION NP\_777077  
VERSION NP\_777077.1 GI:27807177  
DBSOURCE REFSEQ: accession [NM\\_174652.2](#)  
KEYWORDS RefSeq.  
SOURCE *Bos taurus* (cattle)  
ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.  
REFERENCE 1 (residues 1 to 548)  
AUTHORS Hedges JF, Kimmel E, Snyder DT, Jerome M and Jutila MA.  
TITLE Solute carrier 11A1 is expressed by innate lymphocytes and augments their activation

JOURNAL J. Immunol. 190 (8), 4263-4273 (2013)  
PUBMED [23509347](#)  
REMARK GeneRIF: Preferential expression of SLC11A1 transcripts in gammadelta T cells is detected in bovine, human, and mouse gammadelta T cells.

REFERENCE 2 (residues 1 to 548)  
AUTHORS Hasenauer FC, Caffaro ME, Czibener C, Commerci D, Poli MA and Rossetti CA.  
TITLE Genetic analysis of the 3' untranslated region of the bovine SLC11A1 gene reveals novel polymorphisms  
JOURNAL Mol. Biol. Rep. 40 (1), 545-552 (2013)  
PUBMED [23065223](#)  
REMARK GeneRIF: Analysis of allelic variants in the first and second microsatellite at the 3;UTR region of the SLC11A1 gene in cattle breeds present in Argentina.

REFERENCE 3 (residues 1 to 548)  
AUTHORS Cheng X and Wang H.  
TITLE Multiple targeting motifs direct NRAMP1 into lysosomes  
JOURNAL Biochem. Biophys. Res. Commun. 419 (3), 578-583 (2012)  
PUBMED [22382021](#)  
REMARK GeneRIF: NRAMP1 consists of multiple targeting motifs for trafficking into lysosomes.

REFERENCE 4 (residues 1 to 548)  
AUTHORS Pinedo PJ, Buergelt CD, Donovan GA, Melendez P, Morel L, Wu R, Langae TY and Rae DO.  
TITLE Candidate gene polymorphisms (BoIFNG, TLR4, SLC11A1) as risk factors for paratuberculosis infection in cattle  
JOURNAL Prev. Vet. Med. 91 (2-4), 189-196 (2009)  
PUBMED [19525022](#)  
REMARK GeneRIF: A tendency toward statistical significance for the effect of polymorphisms in the odds of infection in cattle was only found for alleles SLC11A1.

REFERENCE 5 (residues 1 to 548)  
AUTHORS Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marcais G, Roberts M, Subramanian P, Yorke JA and Salzberg SL.  
TITLE A whole-genome assembly of the domestic cow, Bos taurus  
JOURNAL Genome Biol. 10 (4), R42 (2009)  
PUBMED [19393038](#)

REFERENCE 6 (residues 1 to 548)  
AUTHORS Coussens PM, Coussens MJ, Tooker BC and Nobis W.  
TITLE Structure of the bovine natural resistance associated macrophage protein (NRAMP 1) gene and identification of a novel polymorphism  
JOURNAL DNA Seq. 15 (1), 15-25 (2004)  
PUBMED [15354350](#)  
REMARK GeneRIF: Identification of a novel polymorphism within the bovine NRAMP 1 gene intron X.

REFERENCE 7 (residues 1 to 548)  
AUTHORS Ables GP, Nishibori M, Kanemaki M and Watanabe T.  
TITLE Sequence analysis of the NRAMP1 genes from different bovine and buffalo breeds  
JOURNAL J. Vet. Med. Sci. 64 (11), 1081-1083 (2002)  
PUBMED [12499702](#)  
REMARK GeneRIF: DNA sequence analysis of NRAMP1 gene polymorphisms from 5

different cattle breeds  
 REFERENCE 8 (residues 1 to 548)  
 AUTHORS Smith TP, Grosse WM, Freking BA, Roberts AJ, Stone RT, Casas E, Wray JE, White J, Cho J, Fahrenkrug SC, Bennett GL, Heaton MP, Laegreid WW, Rohrer GA, Chitko-McKown CG, Pertea G, Holt I, Karamycheva S, Liang F, Quackenbush J and Keele JW.  
 TITLE Sequence evaluation of four pooled-tissue normalized bovine  
 cDNA libraries and construction of a gene index for cattle  
 JOURNAL Genome Res. 11 (4), 626-630 (2001)  
 PUBMED [11282978](#)  
 REFERENCE 9 (residues 1 to 548)  
 AUTHORS Horin P, Rychlik I, Templeton JW and Adams LG.  
 TITLE A complex pattern of microsatellite polymorphism within the  
 bovine NRAMP1 gene  
 JOURNAL Eur. J. Immunogenet. 26 (4), 311-313 (1999)  
 PUBMED [10457896](#)  
 REFERENCE 10 (residues 1 to 548)  
 AUTHORS Feng J, Li Y, Hashad M, Schurr E, Gros P, Adams LG and  
 Templeton JW.  
 TITLE Bovine natural resistance associated macrophage protein 1  
 (Nramp1) gene  
 JOURNAL Genome Res. 6 (10), 956-964 (1996)  
 PUBMED [8908514](#)  
 COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to  
 final NCBI review. The reference sequence was derived from [U12862.1](#).  
 Publication Note: This RefSeq record includes a subset of the  
 publications that are available for this gene. Please see the  
 Gene record to access additional publications.  
 ##Evidence-Data-START##  
 Transcript exon combination :: U12862.1 [ECO:0000332]  
 introns RNAseq introns :: single sample supports all  
 SAMN02822088, SAMN02822091  
 [ECO:0000348]  
 ##Evidence-Data-END##

## 29. solute carrier family 11 member 1 [Bos taurus] GenBank: ABF61463.1

LOCUS ABF61463 548 aa linear MAM 30-APR-  
 2007  
 DEFINITION solute carrier family 11 member 1 [Bos taurus].  
 ACCESSION ABF61463  
 VERSION ABF61463.1 GI:99030403  
 DBSOURCE accession [DQ493965.1](#)  
 KEYWORDS .  
 SOURCE Bos taurus (cattle)  
 ORGANISM [Bos taurus](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
 Euteleostomi;  
 Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
 Ruminantia;

Pecora; Bovidae; Bovinae; Bos.

REFERENCE 1 (residues 1 to 548)

AUTHORS Martinez,R., Barrera,G., Dunner,S. and Canon,J.

TITLE Novel polymorphisms in the SLC11A1 gene detected by SSCP in Zebu and Colombian Creole cattle

JOURNAL Unpublished

REFERENCE 2 (residues 1 to 548)

AUTHORS Martinez,R., Barrera,G., Dunner,S. and Canon,J.

TITLE Direct Submission

JOURNAL Submitted (17-APR-2006) Animal Genetic Resource Program, Colombian Corporation of Livestock Research CORPOICA, Km 14 Via Mosquera, Bogota, Cundinamarca 57, Colombia

COMMENT Method: conceptual translation supplied by author.

### 30. solute carrier 11A1 [Bos taurus] GenBank: ABM81484.1

LOCUS ABM81484 548 aa linear MAM 14-JUL-2007

DEFINITION solute carrier 11A1 [Bos taurus].

ACCESSION ABM81484

VERSION ABM81484.1 GI:123979204

DBSOURCE accession [DQ848779.1](#)

KEYWORDS .

SOURCE Bos taurus (cattle)

ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla; Ruminantia; Pecora; Bovidae; Bovinae; Bos.

REFERENCE 1 (residues 1 to 548)

AUTHORS Schutta,C.J., Feng,J., Niu,S., Crider,B.P., Adams,L.G. and Templeton,J.W.

TITLE Complete Genomic Sequence of Bovine SLC11A1 Isolated from a Bovine Genomic BAC Library

JOURNAL Unpublished

REFERENCE 2 (residues 1 to 548)

AUTHORS Schutta,C.J., Feng,J., Niu,S., Crider,B.P., Adams,L.G. and Templeton,J.W.

TITLE Direct Submission

JOURNAL Submitted (13-JUL-2006) Veterinary Pathobiology, Texas A&M University, College Station, TX 77843-4467, USA

### 31. natural resistance-associated macrophage protein 1 [Bos taurus] GenBank: ALC78257.1

LOCUS ALC78257 548 aa linear MAM 08-SEP-2015

DEFINITION natural resistance-associated macrophage protein 1 [Bos taurus].

ACCESSION ALC78257

VERSION ALC78257.1 GI:924919501

DBSOURCE accession [KR002419.1](#)

KEYWORDS .

SOURCE Bos taurus (cattle)

ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.

REFERENCE 1 (residues 1 to 548)  
AUTHORS Zhang,Y., Zhang,B., Qin,B., Wang,Y. and Liu,K.  
TITLE Study of the association between Nrampl Polymorphisms and susceptibility to tuberculosis in dairy cattle  
JOURNAL Unpublished

REFERENCE 2 (residues 1 to 548)  
AUTHORS Zhang,Y., Zhang,B., Qin,B., Wang,Y. and Liu,K.  
TITLE Direct Submission  
JOURNAL Submitted (15-MAR-2015) Faculty Of Annimal Science and Technology,  
Yunnan Agricultural University, FengYuan Road, Beishi District, Kunming, Yunnan 650201, China

COMMENT Method: conceptual translation supplied by author.

### 32. natural resistance-associated macrophage protein 1 [Bos taurus] GenBank: ALC78258.1

LOCUS ALC78258 548 aa linear MAM 08-SEP-2015

DEFINITION natural resistance-associated macrophage protein 1 [Bos taurus].

ACCESSION ALC78258

VERSION ALC78258.1 GI:924919503

DBSOURCE accession [KR002420.1](#)

KEYWORDS .

SOURCE Bos taurus (cattle)  
ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.

REFERENCE 1 (residues 1 to 548)  
AUTHORS Zhang,B., Zhang,Y., Qin,B., Wang,Y. and Liu,K.  
TITLE Study of the association between Nrampl Polymorphisms and susceptibility to tuberculosis in dairy cattle  
JOURNAL Unpublished

REFERENCE 2 (residues 1 to 548)  
AUTHORS Zhang,B., Zhang,Y., Qin,B., Wang,Y. and Liu,K.  
TITLE Direct Submission  
JOURNAL Submitted (16-MAR-2015) Faculty Of Annimal Science and Technology,  
Yunnan Agricultural University, FengYuan Road, Beishi District, Kunming, Yunnan 650201, China

COMMENT Method: conceptual translation supplied by author.

### 33. natural resistance-associated macrophage protein 1 [Bos taurus] GenBank: ALC78259.1

LOCUS ALC78259 548 aa linear MAM 08-SEP-2015

DEFINITION natural resistance-associated macrophage protein 1 [Bos taurus].

ACCESSION ALC78259  
VERSION ALC78259.1 GI:924919505  
DBSOURCE accession [KR002421.1](#)  
KEYWORDS .  
SOURCE Bos taurus (cattle)  
ORGANISM [Bos taurus](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Euteleostomi;  
Mammalia; Eutheria; Laurasiatheria; Cetartiodactyla;  
Ruminantia;  
Pecora; Bovidae; Bovinae; Bos.  
REFERENCE 1 (residues 1 to 548)  
AUTHORS Shi,X., Zhang,Y., Qin,B., Wang,Y., Liu,K. and Zhang,B.  
TITLE Study of the association between Nramp1 Polymorphisms and  
susceptibility to tuberculosis in dairy cattle  
JOURNAL Unpublished  
REFERENCE 2 (residues 1 to 548)  
AUTHORS Shi,X., Zhang,Y., Qin,B., Wang,Y., Liu,K. and Zhang,B.  
TITLE Direct Submission  
JOURNAL Submitted (16-MAR-2015) Faculty Of Annimal Science and  
Technology,  
Yunnan Agricultural University, FengYuan Road, Beishi District,  
Kunming, Yunnan 650201, China  
COMMENT Method: conceptual translation supplied by author.

## APPENDIX B

This appendix related with the results of chapter three about the comparative between the sequences of 17 vertebrate organisms including human to discover the evolutionary distance by maximum likelihood.

Table B.1.: Vertebrate mitochondrial genetic code

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu	TCA S Ser	TAA * Ter	TGA W Trp
TTG L Leu	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile i	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile i	ACC T Thr	AAC N Asn	AGC S Ser
ATA M Met i	ACA T Thr	AAA K Lys	AGA * Ter
ATG M Met i	ACG T Thr	AAG K Lys	AGG * Ter
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val i	GCG A Ala	GAG E Glu	GGG G Gly

(i)Means the alternative initiation codes in vertebrates mitochondria, and the \* show the termination codons.

[Source (<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG2> )]

Table B.2.: The genetic code that differ in vertebrate mitochondrial DNA from slandered codes



	UGA	AUA	AGR
STANDARD	Ter	Ile	Arg
VERTEBRATE MT-DNA	Trp	Met	Ter

Show differ amino acids between standard and vertebrate MT-DNA by with same code.

[Source (<http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG2> )]

Table B.3.: The first ten of amino acid composition, frequencies shown in percentage (%)

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu
HUMAN	5.50	1.17	1.88	2.13	4.21	3.31	5.01	7.10	4.75	11.72
CHIMPANZEE	5.34	1.43	2.05	2.13	4.34	3.01	4.68	7.40	4.75	11.82
GORILLA	5.60	1.27	1.94	2.24	4.55	3.23	4.99	7.30	4.77	11.58
CATTLE	4.84	1.32	2.20	2.58	4.36	3.30	4.44	9.06	5.24	12.00
WATER BUFFALO	4.82	1.47	2.23	2.55	4.28	3.65	4.74	8.39	4.88	11.46
BISON	4.86	1.30	2.12	2.68	4.52	3.30	4.44	9.02	5.40	11.88
ARABIAN CAMEL	5.44	1.61	2.47	2.39	4.60	3.91	4.82	7.99	4.56	11.43
BACTRIAN CAMEL	5.24	1.84	2.48	2.50	4.48	3.81	4.40	8.21	4.57	11.53
HORSE	4.83	1.46	2.34	2.49	4.07	3.27	4.85	8.70	4.53	11.39
SHEEP	4.59	1.18	2.26	2.39	5.00	3.24	4.37	9.36	5.26	11.69
GOAT	4.35	1.05	2.24	2.36	4.82	3.24	4.43	9.35	5.29	11.67
PIG	4.76	1.26	2.18	2.73	4.26	3.22	5.13	8.77	5.33	10.59
CHICKEN	5.33	1.49	2.00	2.25	3.65	3.63	4.68	7.29	4.45	11.57
RABBIT	4.92	1.13	2.10	2.35	5.41	3.27	4.06	8.34	5.02	12.40
DOG	4.96	1.38	2.33	2.35	4.51	3.42	4.26	8.98	4.43	12.44
CAT	4.83	1.64	2.21	2.50	4.08	3.47	4.92	8.40	4.73	11.74
HOUSE MOUSE	4.13	0.98	2.22	2.36	5.13	3.24	4.31	9.87	5.37	12.64
Avg.	4.96	1.35	2.19	2.41	4.49	3.38	4.62	8.44	4.90	11.74

Table B.4.: The second ten of amino acid composition, frequencies shown in percentage (%)

	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
HUMAN	1.12	5.40	10.54	4.15	4.24	10.41	9.96	2.93	0.31	4.17
CHIMPANZEE	1.12	5.32	10.49	4.15	4.27	10.00	10.00	2.84	0.53	4.34
GORILLA	1.09	5.42	10.47	4.08	4.12	9.95	9.60	2.85	0.49	4.47
CATTLE	1.14	6.08	7.66	3.94	3.76	10.40	9.02	3.38	0.56	4.72
WATER BUFFALO	1.38	6.04	8.03	4.07	3.67	10.02	9.37	3.61	0.58	4.76
BISON	1.14	6.22	7.60	4.00	3.50	10.14	9.10	3.32	0.58	4.86
ARABIAN CAMEL	1.45	4.89	8.24	3.93	4.54	9.08	8.49	4.23	0.92	5.01
BACTRIAN CAMEL	1.49	5.08	8.19	3.91	4.79	8.95	8.35	4.18	0.74	5.26
HORSE	1.23	5.47	8.60	4.34	3.97	10.55	9.60	3.25	0.56	4.52
SHEEP	1.41	6.18	7.45	4.08	3.98	9.85	9.14	3.20	0.35	5.02
GOAT	1.39	5.81	7.36	4.25	3.96	10.36	9.07	3.38	0.47	5.15
PIG	1.24	6.27	7.37	4.40	4.09	9.34	9.77	3.38	0.41	5.50
CHICKEN	1.16	4.72	11.50	4.37	4.30	10.94	9.72	2.44	0.61	3.90
RABBIT	1.37	5.26	8.38	4.00	4.19	10.19	8.93	3.23	0.68	4.75
DOG	1.46	5.48	7.50	3.65	4.29	9.70	8.78	3.81	0.60	5.67
CAT	1.58	5.36	8.05	4.08	4.01	9.43	9.56	3.42	0.65	5.31
HOUSE MOUSE	1.36	6.73	6.87	3.89	3.65	9.69	8.51	3.44	0.44	5.17
Avg.	1.30	5.63	8.50	4.08	4.08	9.94	9.23	3.35	0.56	4.86

Table B.5.: Nucleotide probability

Model	#Param	BIC	AICc	lnL	Invariant	Gamma	R
GTR+G+I	41	268608.7	268181.9	-134050	0.194054	0.240001	6.759319
GTR+G	40	268854.5	268438.1	-134179	n/a	0.174724	6.896178
TN93+G+I	38	270047.3	269651.8	-134788	0.237408	0.334542	2.883097
HKY+G+I	37	270399.5	270014.3	-134970	0.234479	0.329194	2.797287

TN93+G	37	271052.8	270667.7	-135297	n/a	0.225806	2.359947
HKY+G	36	271242.4	270867.6	-135398	n/a	0.221002	2.410799
T92+G+I	35	278350.7	277986.4	-138958	0.225705	0.331256	2.856028
T92+G	34	278899.9	278545.9	-139239	n/a	0.227476	2.474889
K2+G+I	34	280466.8	280112.9	-140022	0.216252	0.309883	3.198215
GTR+I	40	280477.6	280061.2	-139991	0.442152	n/a	1.816152
K2+G	33	280980.3	280636.8	-140285	n/a	0.228237	2.487192
TN93+I	37	282855.4	282470.2	-141198	0.441942	n/a	1.885039
HKY+I	36	282997.8	282623	-141276	0.442577	n/a	1.859754
T92+I	34	289200.2	288846.2	-144389	0.442263	n/a	1.898807
JC+G+I	33	290830.6	290487.1	-145211	0.340281	0.776371	0.5
JC+G	32	291203.2	290870.1	-145403	n/a	0.298753	0.5
K2+I	33	291310.4	290966.8	-145450	0.442762	n/a	1.959017
JC+I	32	297818.8	297485.7	-148711	0.442178	n/a	0.5
GTR	39	309490.5	309084.6	-154503	n/a	n/a	1.227333
TN93	36	313954.6	313579.8	-156754	n/a	n/a	1.739579
HKY	35	314517.2	314152.9	-157041	n/a	n/a	1.722791
T92	33	318009.8	317666.3	-158800	n/a	n/a	1.758968
K2	32	319938.3	319605.1	-159771	n/a	n/a	1.814744
JC	31	325484.5	325161.8	-162550	n/a	n/a	0.5
Average	35.33333	289458	289090.2	-144510	0.341838	0.308104	2.283547

Table B.6.: the probability of nucleotide substitution

Model	A=>T	A=>C	A=>G	T=>A	T=>C	T=>G	C=>A	C=>T	C=>G	G=>A	G=>T	G=>C
GTR+G+I	0.02	0.03	0.05	0.02	0.35	0	0.03	0.35	0	0.13	0.01	0.01
GTR+G	0.02	0.03	0.06	0.02	0.35	0	0.03	0.35	0	0.13	0.01	0.01
TN93+G+I	0.03	0.03	0.08	0.04	0.24	0.02	0.04	0.24	0.02	0.19	0.03	0.03

HKY+G+I	0.03	0.03	0.1	0.04	0.2	0.02	0.04	0.2	0.02	0.24	0.03	0.03
TN93+G	0.04	0.04	0.07	0.05	0.24	0.02	0.05	0.23	0.02	0.17	0.04	0.04
HKY+G	0.04	0.04	0.1	0.05	0.2	0.02	0.05	0.19	0.02	0.23	0.04	0.04
T92+G+I	0.04	0.03	0.15	0.04	0.15	0.03	0.04	0.22	0.03	0.22	0.04	0.03
T92+G	0.04	0.03	0.15	0.04	0.15	0.03	0.04	0.21	0.03	0.21	0.04	0.03
K2+G+I	0.03	0.03	0.19	0.03	0.19	0.03	0.03	0.19	0.03	0.19	0.03	0.03
GTR+I	0.05	0.08	0.08	0.05	0.21	0.01	0.1	0.21	0.01	0.18	0.02	0.01
K2+G	0.04	0.04	0.18	0.04	0.18	0.04	0.04	0.18	0.04	0.18	0.04	0.04
TN93+I	0.05	0.05	0.07	0.05	0.21	0.02	0.05	0.21	0.02	0.17	0.05	0.05
HKY+I	0.05	0.05	0.09	0.05	0.18	0.02	0.05	0.18	0.02	0.21	0.05	0.05
T92+I	0.05	0.03	0.14	0.05	0.14	0.03	0.05	0.19	0.03	0.19	0.05	0.03
JC+G+I	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
JC+G	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
K2+I	0.04	0.04	0.17	0.04	0.17	0.04	0.04	0.17	0.04	0.17	0.04	0.04
JC+I	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
GTR	0.05	0.1	0.07	0.06	0.18	0.01	0.12	0.18	0.02	0.15	0.01	0.04
TN93	0.05	0.05	0.07	0.06	0.2	0.02	0.06	0.2	0.02	0.17	0.05	0.05
HKY	0.05	0.05	0.09	0.06	0.18	0.02	0.06	0.17	0.02	0.21	0.05	0.05
T92	0.05	0.04	0.13	0.05	0.13	0.04	0.05	0.19	0.04	0.19	0.05	0.04
K2	0.04	0.04	0.16	0.04	0.16	0.04	0.04	0.16	0.04	0.16	0.04	0.04
JC	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.08
Average	0.05	0.05	0.11	0.05	0.18	0.03	0.055	0.19	0.03	0.167	0.04	0.042

Table B.7.: the probability of amino acid substitution 1

Model	#Param	BIC	AICc	lnL	Invariant	Gamma
JTT+G+I+F	52	139902.9	139427.2	-69661.5	0.21077	1.619979
JTT+G+F	51	140057.3	139590.7	-69744.3	n/a	0.725311
JTT+I+F	51	141422.6	140956.1	-70427	0.2816	n/a

mtREV24+G+I+F	52	143160.9	142685.2	-71290.5	0.161662	1.083008
mtREV24+G+F	51	143222.7	142756.1	-71327	n/a	0.640252
JTT+G+I	33	143257.2	142955.3	-71444.6	0.210148	1.674195
JTT+G	32	143410.9	143118.1	-71527	n/a	0.749437
cpREV+G+F	51	143411.7	142945.1	-71421.5	n/a	0.629424
cpREV+G+I+F	52	143513.1	143037.4	-71466.6	0.166538	1.085458
WAG+G+I+F	52	144160.2	143684.4	-71790.2	0.204546	1.477074
WAG+G+F	51	144291.6	143825	-71861.5	n/a	0.70107
JTT+I	32	144694.5	144401.7	-72168.8	0.281037	n/a
mtREV24+G+I	33	144752.1	144450.2	-72192.1	0.140656	0.94263
mtREV24+G	32	144788.3	144495.6	-72215.8	n/a	0.610448
JTT+F	50	145517.8	145060.4	-72480.2	n/a	n/a
WAG+I+F	51	145796.2	145329.7	-72613.8	0.282182	n/a
Dayhoff+G+I+F	52	145834.5	145358.7	-72627.3	0.191001	1.232074
Dayhoff+G+F	51	145940.2	145473.6	-72685.8	n/a	0.645748
cpREV+I+F	51	146227.7	145761.1	-72829.5	0.225593	n/a
mtREV24+I+F	51	146375.4	145908.8	-72903.4	0.227302	n/a
LG+G+I+F	52	146692.3	146216.6	-73056.3	0.189454	1.156458
cpREV+G	32	146786.8	146494	-73215	n/a	0.636809
LG+G+F	51	146789.2	146322.7	-73110.3	n/a	0.616778
cpREV+G+I	33	146929.1	146627.2	-73280.6	0.157649	1.061002
mtREV24+I	32	147690.6	147397.8	-73666.9	0.26154	n/a
Dayhoff+I+F	51	147946.2	147479.6	-73688.8	0.280944	n/a
WAG+G+I	33	148574.3	148272.4	-74103.2	0.202727	1.537331
JTT	31	148637.4	148353.7	-74145.9	n/a	n/a
WAG+G	32	148706.6	148413.8	-74174.9	n/a	0.729716
rtREV+G+I+F	52	148715.8	148240.1	-74068	0.189612	1.146467
rtREV+G+F	51	148813.9	148347.3	-74122.6	n/a	0.61125
LG+I+F	51	148870.2	148403.7	-74150.8	0.281977	n/a

mtREV24+F	50	149566.8	149109.3	-74504.6	n/a	n/a
cpREV+I	32	149736.7	149443.9	-74690	0.218974	n/a
WAG+F	50	149926.3	149468.9	-74684.4	n/a	n/a
WAG+I	32	150150.2	149857.4	-74896.7	0.280476	n/a
cpREV+F	50	150223.2	149765.8	-74832.8	n/a	n/a
Dayhoff+G+I	33	150499.8	150197.9	-75065.9	0.185896	1.293156
Dayhoff+G	32	150609.3	150316.5	-75126.3	n/a	0.682968
rtREV+I+F	51	150979.9	150513.4	-75205.7	0.281386	n/a
LG+G+I	33	151075.8	150773.9	-75353.9	0.192095	1.25906
LG+G	32	151179.7	150886.9	-75411.5	n/a	0.6548
mtREV24	31	151715.6	151432	-75685	n/a	n/a
Dayhoff+F	50	152263.1	151805.7	-75852.8	n/a	n/a
Dayhoff+I	32	152527.2	152234.5	-76085.2	0.273888	n/a
rtREV+G+I	33	152716.8	152414.8	-76174.4	0.18969	1.254856
rtREV+G	32	152817.5	152524.8	-76230.4	n/a	0.656178
LG+I	32	153048.7	152755.9	-76345.9	0.281465	n/a
LG+F	50	153331.9	152874.5	-76387.2	n/a	n/a
cpREV	31	153488.7	153205.1	-76571.5	n/a	n/a
WAG	31	154104.7	153821.1	-76879.5	n/a	n/a
rtREV+I	32	154763.2	154470.4	-77203.2	0.279077	n/a
rtREV+F	50	155557.6	155100.2	-77500.1	n/a	n/a
Dayhoff	31	156531	156247.4	-78092.7	n/a	n/a
LG	31	157292.6	157008.9	-78473.5	n/a	n/a
rtREV	31	159035.6	158752	-79345	n/a	n/a

Table B.8.: the probability of amino acid substitution 2

Model	Freq L	Freq K	Freq M	Freq F	Freq P	Freq S	Freq T	Freq W	Freq Y	Freq V
JTT+G+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
JTT+G+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04

JTT+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
mtREV24+G+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
mtREV24+G+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
JTT+G+I	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.03	0.07
JTT+G	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.03	0.07
cpREV+G+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
cpREV+G+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
WAG+G+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
WAG+G+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
JTT+I	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.03	0.07
mtREV24+G+I	0.17	0.02	0.05	0.06	0.06	0.07	0.09	0.03	0.03	0.04
mtREV24+G	0.17	0.02	0.05	0.06	0.06	0.07	0.09	0.03	0.03	0.04
JTT+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
WAG+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
Dayhoff+G+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
Dayhoff+G+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
cpREV+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
mtREV24+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
LG+G+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
cpREV+G	0.10	0.05	0.02	0.05	0.04	0.06	0.05	0.02	0.03	0.07
LG+G+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
cpREV+G+I	0.10	0.05	0.02	0.05	0.04	0.06	0.05	0.02	0.03	0.07
mtREV24+I	0.17	0.02	0.05	0.06	0.06	0.07	0.09	0.03	0.03	0.04
Dayhoff+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
WAG+G+I	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.04	0.07
JTT	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.03	0.07
WAG+G	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.04	0.07
rtREV+G+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
rtREV+G+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04

LG+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
mtREV24+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
cpREV+I	0.10	0.05	0.02	0.05	0.04	0.06	0.05	0.02	0.03	0.07
WAG+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
WAG+I	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.04	0.07
cpREV+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
Dayhoff+G+I	0.09	0.08	0.01	0.04	0.05	0.07	0.06	0.01	0.03	0.06
Dayhoff+G	0.09	0.08	0.01	0.04	0.05	0.07	0.06	0.01	0.03	0.06
rtREV+I+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
LG+G+I	0.10	0.06	0.02	0.04	0.04	0.06	0.05	0.01	0.03	0.07
LG+G	0.10	0.06	0.02	0.04	0.04	0.06	0.05	0.01	0.03	0.07
mtREV24	0.17	0.02	0.05	0.06	0.06	0.07	0.09	0.03	0.03	0.04
Dayhoff+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
Dayhoff+I	0.09	0.08	0.01	0.04	0.05	0.07	0.06	0.01	0.03	0.06
rtREV+G+I	0.10	0.08	0.02	0.03	0.07	0.05	0.06	0.03	0.03	0.06
rtREV+G	0.10	0.08	0.02	0.03	0.07	0.05	0.06	0.03	0.03	0.06
LG+I	0.10	0.06	0.02	0.04	0.04	0.06	0.05	0.01	0.03	0.07
LG+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
cpREV	0.10	0.05	0.02	0.05	0.04	0.06	0.05	0.02	0.03	0.07
WAG	0.09	0.06	0.02	0.04	0.05	0.07	0.06	0.01	0.04	0.07
rtREV+I	0.10	0.08	0.02	0.03	0.07	0.05	0.06	0.03	0.03	0.06
rtREV+F	0.12	0.04	0.01	0.05	0.08	0.10	0.10	0.00	0.05	0.04
Dayhoff	0.09	0.08	0.01	0.04	0.05	0.07	0.06	0.01	0.03	0.06
LG	0.10	0.06	0.02	0.04	0.04	0.06	0.05	0.01	0.03	0.07
rtREV	0.10	0.08	0.02	0.03	0.07	0.05	0.06	0.03	0.03	0.06

Table B.9.: Fisher temporary nucleotide



---

HUMAN																		
CHIMPANZEE	1.0																	
GORILLA	1.0	1.0																
CATTLE	1.0	1.0	1.0															
WATER BUFFALO	1.0	1.0	1.0	1.0														
BISON	1.0	1.0	1.0	1.0	1.0													
ARABIAN CAMEL	1.0	1.0	1.0	1.0	1.0	1.0												
BACTRIAN CAMEL	1.0	1.0	1.0	1.0	1.0	1.0	1.0											
HORSE	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0										
SHEEP	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0									
GOAT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0								
PIG	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0							
CHICKEN	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0						
RABBIT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0					
DOG	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0				
CAT	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0			
HOUSE MOUSE	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

---

Table B.10.: Maximum Likelihood Estimate of Gamma Parameter for Site Rates nucleotide

From\To	A	T	C	G
A	-	2.9435	4.0045	6.7647
T	3.4875	-	30.2602	0.5095
C	4.6635	29.7427	-	0.3673
G	15.5440	0.9880	0.7247	-



## APPENDIX C

This appendix related with chapter six and seven in results.

### 1. Counts of annotation for nucleotides sequences

Feature type	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
CDS	1	1	1	1	1	1
Exon	0	0	15	0	0	0
Gap	0	0	4	0	0	0
Gene	1	1	1	1	1	1
Misc. feature	0	0	1	0	0	0
Source	1	1	1	1	1	1
Variation	0	0	4	0	0	0
mRNA	1	1	1	1	1	1

### 2. Counts of atoms

2.1. As single-stranded, Ambiguous residues are omitted in atom counts.

Atoms	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
hydrogen (H)	133,516	165,474	80,568	165,475	165,474	132,149
carbon (C)	106,387	131,798	64,045	131,798	131,799	105,333
nitrogen (N)	41,672	51,583	24,929	51,580	51,585	41,319
oxygen (O)	65,389	81,041	39,563	81,044	81,042	64,704
phosphorus (P)	10,926	13,543	6,597	13,543	13,543	10,814

2.2. As double-stranded Ambiguous residues are omitted in atom counts.

Atoms	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
hydrogen (H)	267,199	331,166	161,156	331,165	331,165	264,479
carbon (C)	212,569	263,447	128,167	263,446	263,446	210,405
nitrogen (N)	82,433	102,214	49,952	102,215	102,215	81,573
oxygen (O)	131,112	162,518	79,166	162,518	162,518	129,770
phosphorus (P)	21,852	27,086	13,194	27,086	27,086	21,628

### 3. Frequencies of atoms

3.1.As single-stranded Ambiguous residues are omitted in atom counts.

Atoms	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
hydrogen (H)	0.373	0.373	0.374	0.373	0.373	0.373
carbon (C)	0.297	0.297	0.297	0.297	0.297	0.297
nitrogen (N)	0.116	0.116	0.116	0.116	0.116	0.117
oxygen (O)	0.183	0.183	0.183	0.183	0.183	0.183
phosphorus (P)	0.031	0.031	0.031	0.031	0.031	0.031

3.2.As double-stranded Ambiguous residues are omitted in atom counts.

Atoms	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
hydrogen (H)	0.374	0.374	0.373	0.374	0.374	0.374
carbon (C)	0.297	0.297	0.297	0.297	0.297	0.297
nitrogen (N)	0.115	0.115	0.116	0.115	0.115	0.115
oxygen (O)	0.183	0.183	0.183	0.183	0.183	0.183
phosphorus (P)	0.031	0.031	0.031	0.031	0.031	0.031

### 4. Counts of nucleotides

Nucleotide	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
Adenine (A)	2,571	3,174	1,422	3,172	3,173	2,560
Cytosine (C)	2,873	3,632	1,925	3,632	3,631	2,807
Guanine (G)	3,078	3,781	1,848	3,782	3,783	3,068
Thymine (T)	2,404	2,956	1,402	2,957	2,956	2,379
Purine (R)	0	0	0	0	0	0
Pyrimidine (Y)	0	0	0	0	0	0
Adenine or cytosine (M)	0	0	0	0	0	0
Guanine or thymine (K)	0	0	0	0	0	0
Cytosine or guanine (S)	0	0	0	0	0	0

Adenine or thymine (W)	0	0	0	0	0	0
Not adenine (B)	0	0	0	0	0	0
Not cytosine (D)	0	0	0	0	0	0
Not guanine (H)	0	0	0	0	0	0
Not thymine (V)	0	0	0	0	0	0
Any nucleotide (N)	0	0	4,068	0	0	0
C + G	5,951	7,413	3,773	7,414	7,414	5,875
A + T	4,975	6,130	2,824	6,129	6,129	4,939

## 5. Frequencies of nucleotides

Nucleotide	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
Adenine (A)	0.235	0.234	0.133	0.234	0.234	0.237
Cytosine (C)	0.263	0.268	0.180	0.268	0.268	0.260
Guanine (G)	0.282	0.279	0.173	0.279	0.279	0.284
Thymine (T)	0.220	0.218	0.131	0.218	0.218	0.220
Purine (R)	0.000	0.000	0.000	0.000	0.000	0.000
Pyrimidine (Y)	0.000	0.000	0.000	0.000	0.000	0.000
Adenine or cytosine (M)	0.000	0.000	0.000	0.000	0.000	0.000
Guanine or thymine (K)	0.000	0.000	0.000	0.000	0.000	0.000
Cytosine or guanine (S)	0.000	0.000	0.000	0.000	0.000	0.000
Adenine or thymine (W)	0.000	0.000	0.000	0.000	0.000	0.000
Not adenine (B)	0.000	0.000	0.000	0.000	0.000	0.000

Not cytosine (D)	0.000	0.000	0.000	0.000	0.000	0.000
Not guanine (H)	0.000	0.000	0.000	0.000	0.000	0.000
Not thymine (V)	0.000	0.000	0.000	0.000	0.000	0.000
Any nucleotide (N)	0.000	0.000	0.381	0.000	0.000	0.000
C + G	0.545	0.547	0.354	0.547	0.547	0.543
A + T	0.455	0.453	0.265	0.453	0.453	0.457

## 6. Codon statistics from coding regions

Codon	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
AAA	2	2	2	2	2	2
AAC	11	11	11	11	11	12
AAG	10	10	10	10	10	10
AAT	2	2	2	2	2	2
ACA	4	4	4	4	4	4
ACC	18	18	18	18	18	18
ACG	2	2	2	2	2	2
ACT	6	7	6	7	7	6
AGA	1	1	1	1	1	1
AGC	12	12	12	12	12	12
AGG	3	3	3	3	3	3
AGT	4	4	4	4	4	4
ATA	1	1	1	1	1	1
ATC	24	24	24	24	24	24
ATG	12	12	12	12	12	12
ATT	9	9	9	9	9	9
CAA	6	6	6	6	6	6
CAC	5	5	5	5	5	5
CAG	16	16	16	16	16	16
CAT	2	2	2	2	2	2
CCA	8	8	8	8	7	7
CCC	15	15	15	15	15	15

CCG	2	2	2	2	2	2
CCT	7	7	7	7	7	7
CGA	5	5	5	5	5	5
CGC	4	4	4	4	4	4
CGG	8	8	8	8	8	8
CGT	0	0	0	0	0	0
CTA	2	2	2	2	2	2
CTC	22	22	22	22	22	22
CTG	44	44	44	44	44	44
CTT	9	9	9	9	9	9
GAA	5	5	5	5	5	5
GAC	13	13	13	13	13	12
GAG	13	13	13	13	13	13
GAT	4	4	4	4	4	4
GCA	8	8	8	8	9	9
GCC	26	26	26	26	26	26
GCG	4	4	3	4	4	4
GCT	13	12	13	13	13	13
GGA	12	12	12	12	12	12
GGC	21	21	21	21	21	21
GGG	9	9	9	9	9	9
GGT	6	6	6	6	6	6
GTA	2	2	2	2	2	2
GTC	13	13	13	13	13	13
GTG	22	22	23	22	22	22
GTT	1	2	1	1	1	1
TAA	0	0	0	0	0	0
TAC	17	16	17	16	16	17
TAG	0	0	0	0	0	0
TAT	3	3	3	3	3	3
TCA	4	4	4	4	4	4
TCC	12	12	12	12	12	12
TCG	4	4	4	4	4	4
TCT	2	2	2	2	2	2
TGA	0	0	0	0	0	0
TGC	6	6	6	6	6	6
TGG	8	8	8	8	8	8
TGT	3	3	3	3	3	3

TTA	0	0	0	0	0	0
TTC	24	24	24	24	24	24
TTG	8	8	8	8	8	8
TTT	9	9	9	9	9	9

---

## 7. Frequency of codons

Codon	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
AAA	0.00	0.00	0.00	0.00	0.00	0.00
AAC	0.02	0.02	0.02	0.02	0.02	0.02
AAG	0.02	0.02	0.02	0.02	0.02	0.02
AAT	0.00	0.00	0.00	0.00	0.00	0.00
ACA	0.01	0.01	0.01	0.01	0.01	0.01
ACC	0.03	0.03	0.03	0.03	0.03	0.03
ACG	0.00	0.00	0.00	0.00	0.00	0.00
ACT	0.01	0.01	0.01	0.01	0.01	0.01
AGA	0.00	0.00	0.00	0.00	0.00	0.00
AGC	0.02	0.02	0.02	0.02	0.02	0.02
AGG	0.01	0.01	0.01	0.01	0.01	0.01
AGT	0.01	0.01	0.01	0.01	0.01	0.01
ATA	0.00	0.00	0.00	0.00	0.00	0.00
ATC	0.04	0.04	0.04	0.04	0.04	0.04
ATG	0.02	0.02	0.02	0.02	0.02	0.02
ATT	0.02	0.02	0.02	0.02	0.02	0.02
CAA	0.01	0.01	0.01	0.01	0.01	0.01
CAC	0.01	0.01	0.01	0.01	0.01	0.01
CAG	0.03	0.03	0.03	0.03	0.03	0.03
CAT	0.00	0.00	0.00	0.00	0.00	0.00
CCA	0.01	0.01	0.01	0.01	0.01	0.01
CCC	0.03	0.03	0.03	0.03	0.03	0.03
CCG	0.00	0.00	0.00	0.00	0.00	0.00
CCT	0.01	0.01	0.01	0.01	0.01	0.01
CGA	0.01	0.01	0.01	0.01	0.01	0.01
CGC	0.01	0.01	0.01	0.01	0.01	0.01
CGG	0.01	0.01	0.01	0.01	0.01	0.01



CGT	0.00	0.00	0.00	0.00	0.00	0.00
CTA	0.00	0.00	0.00	0.00	0.00	0.00
CTC	0.04	0.04	0.04	0.04	0.04	0.04
CTG	0.08	0.08	0.08	0.08	0.08	0.08
CTT	0.02	0.02	0.02	0.02	0.02	0.02
GAA	0.01	0.01	0.01	0.01	0.01	0.01
GAC	0.02	0.02	0.02	0.02	0.02	0.02
GAG	0.02	0.02	0.02	0.02	0.02	0.02
GAT	0.01	0.01	0.01	0.01	0.01	0.01
GCA	0.01	0.01	0.01	0.01	0.02	0.02
GCC	0.05	0.05	0.05	0.05	0.05	0.05
GCG	0.01	0.01	0.01	0.01	0.01	0.01
GCT	0.02	0.02	0.02	0.02	0.02	0.02
GGA	0.02	0.02	0.02	0.02	0.02	0.02
GGC	0.04	0.04	0.04	0.04	0.04	0.04
GGG	0.02	0.02	0.02	0.02	0.02	0.02
GGT	0.01	0.01	0.01	0.01	0.01	0.01
GTA	0.00	0.00	0.00	0.00	0.00	0.00
GTC	0.02	0.02	0.02	0.02	0.02	0.02
GTG	0.04	0.04	0.04	0.04	0.04	0.04
GTT	0.00	0.00	0.00	0.00	0.00	0.00
TAA	0.00	0.00	0.00	0.00	0.00	0.00
TAC	0.03	0.03	0.03	0.03	0.03	0.03
TAG	0.00	0.00	0.00	0.00	0.00	0.00
TAT	0.01	0.01	0.01	0.01	0.01	0.01
TCA	0.01	0.01	0.01	0.01	0.01	0.01
TCC	0.02	0.02	0.02	0.02	0.02	0.02
TCG	0.01	0.01	0.01	0.01	0.01	0.01
TCT	0.00	0.00	0.00	0.00	0.00	0.00
TGA	0.00	0.00	0.00	0.00	0.00	0.00
TGC	0.01	0.01	0.01	0.01	0.01	0.01
TGG	0.01	0.01	0.01	0.01	0.01	0.01
TGT	0.01	0.01	0.01	0.01	0.01	0.01
TTA	0.00	0.00	0.00	0.00	0.00	0.00
TTC	0.04	0.04	0.04	0.04	0.04	0.04
TTG	0.01	0.01	0.01	0.01	0.01	0.01
TTT	0.02	0.02	0.02	0.02	0.02	0.02

---

## 8. Nucleotide count in codon positions

Nucleotide	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
per position						
1. pos. A	121	122	121	122	122	122
1.pos. C	155	155	155	155	154	154
1.pos. G	172	172	172	172	173	172
1.pos. T	100	99	100	99	99	100
2. pos. A	109	108	109	108	108	109
2.pos. C	135	135	134	136	136	135
2.pos. G	102	102	102	102	102	102
2.pos. T	202	203	203	202	202	202
3.pos. A	60	60	60	60	60	60
3.pos. C	243	242	243	242	242	243
3.pos. G	165	165	165	165	165	165
3.pos. T	80	81	80	81	81	80

## 9. Nucleotide frequency in codon positions

Nucleotide	AC_000159	KR002419	DQ493965	KR002421	KR002420	DQ848779
per position						
1.pos. A	0.22	0.22	0.22	0.22	0.22	0.22
1.pos. C	0.28	0.28	0.28	0.28	0.28	0.28
1.pos. G	0.31	0.31	0.31	0.31	0.32	0.31
1.pos. T	0.18	0.18	0.18	0.18	0.18	0.18
2.pos. A	0.20	0.20	0.20	0.20	0.20	0.20
2.pos. C	0.25	0.25	0.24	0.25	0.25	0.25
2.pos. G	0.19	0.19	0.19	0.19	0.19	0.19
2.pos. T	0.37	0.37	0.37	0.37	0.37	0.37
3.pos. A	0.11	0.11	0.11	0.11	0.11	0.11
3.pos. C	0.44	0.44	0.44	0.44	0.44	0.44
3.pos. G	0.30	0.30	0.30	0.30	0.30	0.30
3.pos. T	0.15	0.15	0.15	0.15	0.15	0.15

## 10. Protein Secondary structure

Type	Region
------	--------

---

Beta strand	35..36
Alpha helix	52..57
Beta strand	63..69
Alpha helix	79..81
Beta strand	87..97
Alpha helix	98..109
Beta strand	112..113
Beta strand	120..125
Beta strand	131..143
Beta strand	149..151
Beta strand	153..159
Beta strand	171..184
Alpha helix	189..195
Beta strand	196
Beta strand	198..207
Beta strand	209..214
Beta strand	221
Beta strand	225..226
Alpha helix	237..241
Beta strand	242..249
Beta strand	254..256
Beta strand	261
Alpha helix	271..277
Beta strand	280..303
Beta strand	306..308
Alpha helix	313..315
Beta strand	316
Beta strand	318..319
Beta strand	336..343
Beta strand	347..352
Beta strand	358..366
Beta strand	382..384
Beta strand	387
Alpha helix	388..390
Alpha helix	392..396
Beta strand	397..401
Alpha helix	409..435
Beta strand	439..441

Beta strand	443..446
Alpha helix	450..458
Beta strand	461
Beta strand	463..484
Alpha helix	492..504
Beta strand	508..518
Beta strand	522..524
Alpha helix	530..532
Beta strand	533

---



## **COMPARATIVE BIOINFORMATICS ANALYSIS OF OMICS IN SOME ANIMALS**

### **SUMMARY**

Bioinformatics is known as “The area of studies which escapes easy definition in as much as of the fusion between science that attracts in the sciences of computational approach, and information technology to see and analyze genetic database and maths solutions” The major merger of bioinformatics is between computational and biological sciences. This field has expanded to comprehend the data content and data flow in biological systems.

As a summarize of that mentioned in the chapters of Introduction and Literature review, from the historical background and the main categories of bioinformatics science and

application in omics and proteomic sequences analysis in evolutionary distance and immunomics. In the first section, the maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals were analyzed. Secondly, construct the Phylogenetic Tree using Maximum Likelihood Method of complete. genomes of mitochondrial DNA between Human's and 16 animals were conducted. Thirdly, Nucleotide and amino acid sequences analysis in pathogen and host of *brucellosis* in cattle. Whilst investigating the immunoinformatics Antigenicity epitopes prediction in the solute carrier family 11 of the natural resistance associated macrophage protein 1 (NRAMP) related with *Brucellosis* in Cattle. Was carried out in final result section of the thesis.

The methodology process chronologically passed through protocols of the fundamental three steps. Firstly, with mining database in the official resources that related with evolutionary study incomplete genome of mitochondrial DNA between human versus 16 animals. Additionally, the database related to investigation in genomics and proteomics that associated with *Brucellosis* in cattle must be looking for both inside the pathogen which cause brucellosis disease and the animal that infected with. Secondly, 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) [www.cpan.org/](http://www.cpan.org/) that provide thousands of modules shared from scientists and computer programmers studying on bioinformatics Nevertheless, needed to extract some mathematical functions from [www.megasoftware.net](http://www.megasoftware.net) which is an academic open-public software for molecular evolutionary genetic analysis MEGA7-CC-Porto. Finally, the 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 computer with choosing and designing the codes, then apply them to obtain the best

results as it possible. It is worth mentioning, that programming languages should not be used directly after downloaded from the open source access websites because they are designed for general purposes and need manipulating with adding the private data and the mathematical problems serve the particular study.

The main results of this study for maximum likelihood estimation of the evolutionary distance of complete genomes of mitochondrial DNA between Human's and 16 animals and phylogenetic tree 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 comparing to the all other organisms. Moreover, demonstrated a rooted phylogenetic tree and demonstrate paraphyletic group of mitochondrial DNA sequences. human's sequence is the head base of the comparative. Results put the species in interest in two main monophyletic groups. The Glades by sharing the common ancestral point for each group except chicken is the out group of the tree. First group, shows the human and chimpanzee are descendants of gorilla and split in two different evaluated organisms. Second group, is the major and more complex, by observing the tree chronologically starting from the internal points to the terminal points can note cattle and bison share ancestor node and they are descendants of the water buffalo, also the dog and cat are descendants from horse.

Equally important The main aim behind prediction of antigenic epitopes is to find the maximum probability of potential peptides residues could recognize and bind with the antigen,

which is become very handy to design drugs and looking for increasing the number of animals that have ability to produce this protein of the natural resistance associated macrophage protein 1 (NRAMP) related with *Brucellosis* in Cattle.



## CURRICULUM VITAE Personal Information

**Name, Surname** : OMAR ESMAILL H. HAMAD  
Nationality : IRAQI - KURDISH  
Birth date and place : 12.07.1977 NINEVEH  
Marriage : Single  
Telephone : 0536 964 77 62  
Fax : -  
e-mail : [omarehamad7@gmail.com](mailto:omarehamad7@gmail.com)

---

### **Education**

<b><u>Degree</u></b>	<b>Education Division</b>	<b>Graduate date</b>
Master	Mosul University/ Department of Animal resources	2001
Undergraduate	Mosul University/ Department of Animal resources	1999
High School	Al-Sharqiya High School	1995

### **Work Experience**

<b><u>Year</u></b>	<b>Place</b>	<b>Title</b>
2002- 2007	University of Salahaddin - Erbil, KRG.	University lecturer
2007- Till now	University of Soran, KRG.	University lecturer

### **Foreign Language**

English  
ARABIC  
TURKISH  
RUSSIAN  
PERSIAN

### **Hobbies**

Reading, Yoga, Computer programming and playing guitar



