

# **A Bioinformatics Analysis of Theory of Devolution**

*A Major Project dissertation submitted in partial fulfillment of  
the requirement for the degree of*

**Master of Technology**

**In**

**Bioinformatics**

*Submitted by*

**Prateek Sukumar**

**2K12/BIO/020**

**Delhi Technological University, Delhi, India**

*Under the supervision of*

**Dr. Yasha Hasija**



Department Of Biotechnology  
Delhi Technological University  
(Formerly Delhi College of Engineering)  
Shahbad Daultapur, Main Bawana Road,  
Delhi -110042, India



## **CERTIFICATE**

This is to certify that the dissertation entitled “**A Bioinformatics Analysis of Theory of Devolution**” in the partial fulfillment of the requirements for the reward of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the **Prateek Sukumar (2K12/BIO/020)** own work and is carried out by him under my guidance.

The information and data enclosed in this thesis is original and has not been submitted elsewhere for honoring of any other degree.

**Date:**

**Dr. Yasha Hasija**  
**Assistant Professor & Associate Head**  
Department of Bio-Technology  
Delhi Technological University

# DECLARATION

I, **Prateek Sukumar**, hereby declare that the work entitled “**A Bioinformatics Analysis of Theory of Devolution.**” has been carried out by me under the guidance of Dr. Yasha Hasija, in Delhi Technological University, Delhi.

This dissertation is part of partial fulfillment of requirement for the degree of M.Tech. in Bioinformatics. This is the original work and has not been submitted for any other degree in any other university.

Prateek Sukumar

Roll no. : 2K12/BIO/020

## **ACKNOWLEDGEMENT**

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely fortunate to have got all this along the completion of my project work. Whatever I have done is only due to such guidance and assistance and I would not forget to thank them.

First, of all, I am extremely thankful to Prof P.B. Sharma, Vice –Chancellor, Delhi Technological University, for providing me an opportunity to study in one of the premier institutes of the country. My warm thanks to Dr. B.D. Malhotra, Head of Department of Biotechnology, DTU for his never ending support and constant encouragement.

I owe great thanks to my guide Dr.Yasha Hasija (Asst. Prof., Dept. Of Biotechnology, DTU) for her exemplary guidance, monitoring and constant encouragement throughout the course of the thesis.

I would also like to thank Mr. Prashant Vaishla and my group mates for giving help and advice whenever needed.

Last but not the least, I am very grateful towards my friends and family whose constant support and love gave me strength to complete my work.

**Prateek Sukumar**  
**2K12/BIO/020**

# LIST OF FIGURES

Figure 1: Showing either of the two ways of survival. Either increase the brain complexity or decrease the body complexity or do both in order to survive.

Figure 2: View of MEGA6 – GUI home page.

Figure 3: File HK\_genes.txt containing the Gene name and Ref\_seq id in tab separated format.

Figure 4: File HK\_genes\_refseq\_ids.txt formed after parsing HK\_genes.txt

Figure 5: Each single file contains a pair of homologous sequence

Figure 6: View of file after step no. 11

Figure 7: Making the analysis settings .mao file for the alignment.

Figure 8: Executing MEGA-CC for alignment of input file h1.fasta

Figure 9: The standard settings selected for Z test of positive selection for pair of 2 sequences

# LIST OF TABLES

Table 1: Showing result of 10 homologous sequence pairs between Human and Chimpanzee.

# CONTENTS

<b>S. No</b>	<b>Title</b>	<b>Page</b>
1	ABSTRACT	1
2	INTRODUCTION	2
3	REVIEW OF LITERATURE	4
	3.1 Less is more hypothesis	4
	3.1.1 Selection for Year-Round Fecundity in Mice	4
	3.1.2 Gene losses during Human Origin	5
	3.2 Implausible nature of Darwinian Evolution in practice	6
	3.2.1 Irreducible complexity	6
	3.2.2 Evolution of TTSS system	7
	3.3 Various other examples supporting theory of devolution	8
	3.4 Neural Theory of Evolution	10
	3.5 Story of evolution of Humans from Chimpanzee	12
	3.6 MEGA	12
	3.7 Background of the experiment	14
4	MATERIALS AND METHOD	15
	4.1 Using MEGA-CC for alignment of the homologous sequence pair	20
	4.2 Using MEGA-CC for positive selection test of the homologous sequence pair	22
5	RESULTS AND DISCUSSION	24
6	CONCLUSION	29
7	FUTURE PERSPECTIVE	30
8	REFERENCES	31
9	APPENDIX	35

# A Bioinformatics analysis of Theory of Devolution

Prateek Sukumar

Delhi Technological University, Delhi, India

E-mail ID: [prateek.dtu@hotmail.com](mailto:prateek.dtu@hotmail.com)

## 1. ABSTRACT

The standard theory of evolution by Darwin attributes the development of all complexities manifested by different life forms to forces of natural selection and survival of the fittest. It also states that the need for adaptation to survive and produce off springs has caused the life to develop from low complexity single cell organism to multi cellular highly complex human being.

But in this thesis we argue that if in fact natural selection shapes different forms of life then life must have devolved from higher complex form to the lower complex form. The major premise being losing complexity improves your chances at survival. We try to prove it through documenting various kind of evidences through examples. Moreover an experiment has been carried out in which we determine the positive selection between *Pan troglodytes* and *Homo sapiens* house keeping genes.

The house keeping genes were tested using Codon based z -test of positive selection by forming a pipeline through MEGA software. A test hypothesis was alternative hypothesis :  $dN$ (rate of non synonymous substitutions per non synonymous site)  $>$   $dS$ (rate of synonymous substitution per synonymous site) was tested against the alternative hypothesis.  $dN > dS$  is condition for positive selection. For all the genes the hypothesis was not rejected i.e. alternative hypothesis of positive selection was not selected. This suggests that our genes are not under positive selection as suggested by Darwin theory of evolution.

## 2.INTRODUCTION

*“The origin of life on the earth has been one of the most outstanding problems in science. Although it is still a formidable problem, with the development of molecular genetics there is a much better hope that substantial progress will be made in the near future. I imagine that in the coming century, studies on the origin of life will become more popular and will be regarded as a more respectable field than it is now. “ – Motoo Kimura (The neutral theory of molecular evolution and the world view of the neutralist).*

With this view in mind we suppose that we should not adhere so rigidly to Darwin theory of evolution as far as origin of life and tree of life is considered. Hence inquiry into evolution and origin of life is not obsolete. Indeed it is contrary, as it is the beginning. Such studies shall shed light on big questions like who we are or from where we have come from, the process through which we are the way we are and finally what shall be our future or where are we heading to. The problem of origin of life thus should be taken as a complete field in itself and inquiries should be made in it, not to find an immediate solution but as a small step towards development of this field.

Many a times, a look forward to future can provide us clue to the journey of past to the present. Just imagine some thousands years down the lane - whom do one reason to inhabit the earth? Would the only survivors will be these simple single cell organism like bacteria and virus or the complex organism like Elephants and Tigers?. If we have a good chance of bacteria being the survivors then perhaps the life on earth is becoming only less complex.

This theory of devolution proposes that life came into being with highly complex body and most lowly developed brain. Now in order to survive this life form shall had taken one step towards either increasing its brain complexity or losing its body complexity or both. And in this process should have formed various creatures and species with intermittent brain and body complexity as survival and chance would have allowed them. The *Homo sapiens* are towards one end of the chain where they maintain high body complexity through high complexity brain. While bacteria or virus on the other end, where they have lost the whole body complexity to become most simple organism in order to survive. It is like body and brain maintain a balance together.



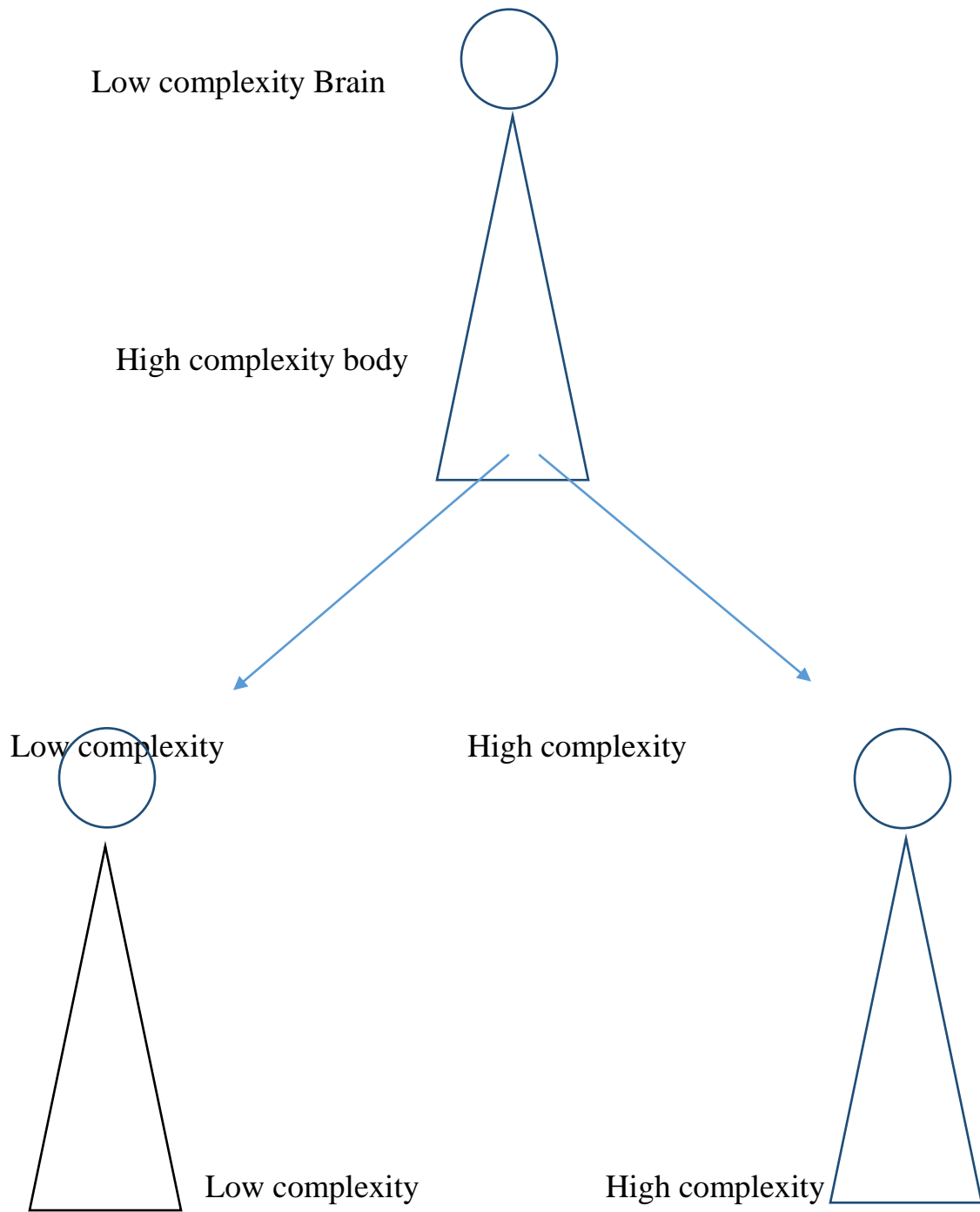


Fig 1: Showing either of the two ways of survival. Either increase the brain complexity or decrease the body complexity or do both in order to survive.

## 3. REVIEW OF LITERATURE

### 3.1 Less is more hypothesis

This hypothesis is coined by Maynard V. Olson. Through this hypothesis he proposes a testable view that gene loss is a major motif of molecular evolution (Olson, M.V., 1999). This is carried out by taking several examples from mice, yeast and human genomes that support adaptation and survival through gene loss.

#### 3.1.1 Selection for Year-Round Fecundity in Mice

This hypothesis is made concrete by observing the difference between the reproductive behavior of wild and laboratory mice. Wild strains of *Mus musculus*, the species from which the laboratory mice were derived, show a seasonal manner or pattern of reproduction. These wild strains show the similar diurnal cycles of melatonin synthesis that occur in the pineal gland as shown nearly by all mammals that have central role in the regulation of seasonal reproduction.<sup>1</sup>(Tamarkin et al 1985). This mechanism is evolutionary conserved, this mechanism monitors the changes in the length of daylight and adjust the reproductive behavior accordingly in that response. As this is evolutionary conserved mechanism it must be the ancestral state. But when we study the laboratory mice whose reproduction is uncoupled from seasonal change it now show know pineal-melatonin synthesis. This feature is because of the occurrence of recessive mutations in two genes, which code for the two enzyme required for the conversion of serotonin to melatonin. (Ebihara et al. 1986; Goto et al .1994). A plausible hypothesis is that these mutations due to selection for trait of unregulated breeding, a highly desirable characteristic of domesticated mice.

In Humans, the instances of adaptive gene loss included Duffy-negative blood group and its relationship with resistance towards *Plasmodium vivax*(Tournamille et al., 1995). This involves loss of chemokine receptor that are essential for entry of the pathogens into target cells. In this example, there is the occurrence of recessive mutation in a promoter element required for expressing the receptor DARC in erythroid lineages. In some regions of western Africa, there is 100% allele frequency of Duffy-negative mutation.

The strength of this hypothesis is that it can be readily tested apart from the fact its genetic plausibility. The testing of this loss-is-more hypothesis rests on the ease with which gene loss mutations could be easily recognized through sequence analysis. Many human genetic diseases, as phenylketonuria, cystic fibrosis, some types of breast cancer 1 that involve early onset of BRAC1 and BRAC2 genes, seem simply to require loss of the relevant gene function (Olson. M. V., et al., 2003). Existing data that proves less is more hypothesis are indeed good but are limited. The best described differences between *Homo sapiens* and *Pan troglodytes* adhere to this hypothesis. The example of it is in one major biochemical difference that humans cannot

synthesize a form of the cell-surface SALIC ACID called N-glycolyl-neuraminic acid.(Muchmore, E.A., 1998).

### 3.1.2 Gene losses during Human Origin

A study has been conducted that checks whether there is process of adaptive pseudogenization involved with human origin (Wang, X., et al., 2006).The adaptive pseudogenization would actually mean selection by loss of genes. A comparative genomic analysis was carried out to identify 80 non processed pseudogenes that got inactivated in *Homo sapiens* after its separation from chimpanzee lineage. The functions involving chemoreception and immune response were found over representing. However to study adaptive pseudogenization the focus was on CASPASE12 gene, a cysteinyl aspartate proteinase that participates in inflammatory and innate immune response to endotoxin. Through population genetic evidence it has been found that there is nearly complete fixation of a null allele at CASPASE12. And this process is being driven by positive selection as null allele would provide protection against sepsis. It was also estimated that pseudogenization of CASPASE12 started shortly after out-of-Africa migration of Humans. Interestingly, two more genes that were also associated with sepsis were also pseudogenized in humans. Thus the identification and analysis of human –specific pseudogenes open the door for understanding the roles of gene losses in human origins, and the finding that gene loss is in itself an adaption that supports “less-is-more” hypothesis.

The gene loss hypothesis is particularly more intriguing in human evolution, many gene losses have been proposed that provide adaptations and are responsible for specific human phenotypes. As, the pseudogenization of MYH16 in human is responsible for reduction in the size of hominin masticatory muscles that provided space for brain size expansion. Hence it is fascinating to identify and analyse all of the human-specific gene losses i.e. the gene losses that occurred after the human-chimpanzee divergence event. In this study this would the gene loss may have occurred independently in other species also except chimpanzee. First step is the identification of human – specific gene losses by the comparison of human nonprocessed pseudogenes with the chimpanzee genome sequence. Such human specific pseudogenes were formed after the separation event of human – chimp in the last 6-7 million years(Brunet, M., et al., 2002).

The genome of human has abundance of pseudogenes.(Zhang, Z., et al., 2003)but most of them are found to be processed. Processed pseudogenes are DNA sequences that are reversed transcribed and then randomly inserted in the genome.( Torrents, D., et al., 2003) Such genes never had any function and hence eliminated from the study. In contrast, nonprocessed pseudogenes are those sequence that once had a function but now have their coding sequence interrupted. But many of these non processed pseudogenes are formed after event of gene duplication to avoid genetic redundancy. (Zhang, J., 2003) .

Thus such nonprocessed genes also won't have any functional evolutionary consequences. Hence the study takes only those human specific nonprocessed pseudogenes that do not result from human specific gene duplicates.

The study identification of pseudogenes that are human specific give information about evolution of human specific features. There is overrepresentation of genes involved in the chemoreception and immunity that had gone pseudogenized in humans. This indicates substantive changes in these two areas of physiology in course of evolution of *Homo sapiens*. The study should be better completed by also analysing non processed chimpanzee specific pseudogenes. However due to lack of accuracy in sequencing of chimpanzee genome it was not possible.

### **3.2 Implausible nature of Darwinian Evolution in practice**

Darwin theory of evolution seem quite plausible at morphological level. The gaining of new morphology or phenotype by addition of slight variations in phenotype of the species looks reasonable. But a further inquiry at genetic or protein level opens the huge complexity involved that would be required to gain a new characteristic. It has been shown that the vast majority of possible protein sequences formed for a given size shall be unstable and cannot be maintained inside the body of living organism. Let apart being functionally beneficial.(Durstun, K.K. et al. , 2007), (Bowie, J.U. et al., 1989), (Bowie, J.U. et al. , 1990), (Reidhaar-Olson, J.F. et al. , 1990), (R.T. Sauer et al. , 1989). Take a comparatively small protein sequence of 150 aa in length. The total number of sequences that can be formed by random permutations and combinations shall be 20<sup>150</sup>. Out of these total possible structures how many protein sequence can possibly produce a stable protein, let apart a beneficial function. Experimental studies have shown that only 1 in 10<sup>74</sup> sequences of 150aa in six are capable for the formation of a stable protein and this ratio decreases exponentially with increase in size of the protein. (Thirumalai,D., et al., 1999), (Meyer, S. C. , 2014). To produce a functional protein the ration becomes 1 in 10<sup>77</sup> protein sequences.

When one starts to consider the protein that require above 1000 aa with an average degree of specificity, then the least distance between this protein and the next functional protein evolved from this protein would be 150 aa non selectable modifications. (Sean Pitman, 2010). It means that it would take around 150 modifications on average in a protein of size 1000 aa to produce a new protein carrying a new function. So how much time it would take to bring 150 non selectable mutations within a large population. Now total number of bacteria that are on earth is 1e<sup>30</sup> and the total number of living organism that ever lived on earth are less than 1e<sup>70</sup>. (Meyer, S. C. , 2014). To carry 150 mutations, the search space size is 20<sup>150</sup> = 1e<sup>195</sup>. , that would bring a new function to a 1000 aa long sequence. Now if we assume that a random mutation occur every 30 minutes in every one of our organism , then it would take 1e<sup>140</sup> years to achieve success to bring new function in one protein – a huge amount of time.

#### **3.2.1 Irreducible complexity**

Darwin himself acknowledged that - "If it could be demonstrated that any complexorgan existed which could not possibly have been formed by numerous, successive, slight modifications, my

theory would absolutely break down." In this light comes the concept of irreducible complexity (Behe, M. J. , 2009). By irreducible complexity it is meant that there is a single system that is formed or composed of several interacting parts that together contribute to a single basic function and the removal of any one of the part shall cause the cease of function of the system.

There are as such many examples of biological complex system but most famous is the bacteria flagellar motility system. It consists of 50 genes that include genes for the sensory apparatus that turns flagellum clockwise or anti clockwise according to the environment and 40 other structural genes that builds the whole flagellum. The DNA required for building of flagellum is over 10000 codons. (Macnab, R. M., et al., 1999). And this requirement is irreducible in order to attain the function of flagellar motility. The argument is that how can nature bring all the parts together gradually when the system does not work until all the parts are in their unique place at the same time.

### **3.2.2 Evolution of TTSS system**

Dr. Kenneth Miller proposed that the Type III secretory system (TTSS : is a toxin injector) is the argument that goes against the concept of irreducible complexity as forwarded by Michael Behe. (Pallen, M. J., 2009). It is argued that the TTSS system demonstrates that how the complex system like that of flagellar motility system comes to existence from the smaller simple systems like the TTSS. TTSS 's 10 protein parts are found to be contained in the 50 or so protein parts of the flagellar motility system. But it turned out to be the other way round – that is a form of devolution and not evolution. The question that arises is which system has evolved earlier ? The simpler TTSS system or the complex system of flagellar motility. It can be reasonable predicted that actually the simpler TTSS system evolved or rather devolved from the more complex flagellar motility function.

1. Bacterial flagellum is found in many different kinds like mesophilic, thermophilic, spirochete, gram-negative and gram-positive bacteria but TTSS are only found in certain gram negative bacteria only. TTSS system is restricted to only pathogenic gram negative bacteria that attacks animals and plants and obviously animals and plants are believed to be evolved after million of years of evolution of flagellar motility system of bacteria. (He, S. Y., 1998.)
2. The GC content of TTSS genes is typically low then the GC content of surrounding chromosome. Moreover TTSS genes are found commonly on large virulence plasmids suggesting their spreading by horizontal gene transfer. (Kim, J. F., 2001)
3. Moreover TTSS system show very little homology with any other bacterial transport system. The observations points that the TTSS or anything homologous did not exist in pre flagellar time . Therefore TTSS should have arisen from the complex flagellar motility system by the removal of other parts.( Sukhan, A., 2001), (Plano, G. V., et al. , 2001), (Nguyen, L., et al. , 2000.)

### **3.3 Various other examples supporting theory of devolution**

### **Example of Cancer**

Activation of EGFR related pathways induces genes implicated in tumor progression. While removal or blocking of these pathways would mean better survival as this would prevent cancer to occur. Infact we already have in place the EGFR targeted therapies ( eg. Cetuximab for colon cancer) block the activation of this signaling pathway. (Lievre, A., et al., 2006).

### **Example of AIDS**

Clinicians have noticed that a small fraction of people engaged in high risk behavior did not develop AIDS. The reason deduced is the functionality of co receptor of CD4 being lost. As HIV binds to CD4 and its co- receptor, the loss of co-receptor would mean that HIV won't be able to bind and hence enter inside the cell to cause infection. (Libert ,F.,et al. 2002 ), ( Sephens , J.C. et al .1998)

### **Example of Malaria**

Haemoglobin mutation at position 6 from glutamine to valine, helps to prevent malaria.(Fairhurst, R. M et al. , 2005), (Luzzatto, L. , 2012).

The above example, one may argue that does not support the corollary that addition of proteins or pathways would cause disease. True enough. But the point is we don't have single example of addition or formation of new proteins to prevent diseases. And similarly we see that removal of proteins or pathways is preventing diseases, although not always, but it does. The losing of protein functionality cause disease but that has no relevance to evolution. The only deduction is thus, that reduction in total protein complexes of the organism has prevented it from diseases and given it a chance to survive.

With no proper evidence of past, the only phenomena that pro Darwin theory people cite is the bacterial strains getting drug resistance. The argument is that variant bacteria that are resistance to drug are being selected and this is evolution. Quite true that they are selected, but have they evolved or devolved. See the example with genetic and molecular mechanism of the drug resistance by M Tuberculosis.

**Drug Resistance of M. Tuberculosis** – The insertion sequence IS6110 has been associated with new resistance emerging through the inactivation of critical genes. Transposition of the insertion sequence IS6110, was identified in the pncA gene from 19 pyrazinamide-resistant Mycobacterium tuberculosis strains. Alignment of the PncA protein from homologous proteins from different bacteria species revealed three highly conserved regions in PncA which play an important role in the processing of pyrazinamide. (Kim, H. J., et al. , 2012).

### **Low occurence of Cancer in Chimpanzee**

Cancer is one of the major disease in modern societies. But, the occurence of this disease in non-human primates is very low. The availability of the chimpanzee genome sequence gives an good opportunity to find the genetic basis for the biological differences between *Homo sapiens* and our most closely related organism .(Mikkelsen TS, et al. , 2005) An striking finding in this regard

is the observation that non-human primates show a lower incidence of cancer than humans.( Varki, A. ,2000). ( Beniashvili DS, 1989).

A study was carried out to find whether the genetic differences between human and chimpanzee genome is the reason causing low susceptibility of cancer in chimps. Examination of a set of 333 human cancer genes with its orthologous genes in chimpanzee was carried out.

If we count the number of nucleotide differences between human and chimp at BRCA1 , it is very few. The number of nucleotide differences between human and chimpanzee sequences at the BRCA1 is low and difference percentage is 1.15%. However, humans and chimpanzee sequences differ by an 8 Kb insertion/deletion within the partially duplicated region. NBR2 gene is prematurely truncated due to this 8 KB deletion. The sequencing of the BRCA1 gene has revealed an 8 Kb deletion in the chimpanzee sequence that would prematurely truncate the co-regulated NBR2 gene (Puente, X. S., et al. , 2006).

### **Crippling of APOC3 cut heart disease risk by 40%**

By studying and mining the DNAs of thousands of patients, researchers at the Broad Institute, Massachusetts General Hospital found the gene mutations significantly reduce a person's risk of coronary heart disease by lowering the levels of triglycerides, a type of fat in the blood.

The four rare mutations all cripple the same gene, called APOC3. The APOC3 protein is mainly made in the liver and pours out into the blood stream.

There, it is thought to prevent the removal of triglyceride-rich lipoproteins from the blood in a few distinct ways, particularly by delaying their clearance following a meal.

This is the most recent study published in June 2014 in the New England Journal of Medicine. The scientists analysis more than 110.000 patient samples and genotyped their necessary parts of APOC3 gene. A comparison was made between the heart attack rates in those patients that carried the mutations and those without the mutations.

It study found that there is 40% lower risk of coronary heart disease in the people who carried the mutations that crippled the activity of APOC3 gene. Hence, suggesting inhibition of APOC3 as a new potential strategy for therapeutic development (Geach, T. , 2014), (Jørgensen, A. B., 2014).

### 3.4 Neural Theory of Evolution

As against the Darwinian theory of evolution using natural selection, the neutral theory of evolution gives emphasis on the value of random genetic drift and the mutation pressure as the main thing causing molecular evolution.

When the neutral theory was proposed (Kimura, M., 1968), the data that was available consisted of sequences of amino acids of only few protein in related organism, as haemoglobin sequences. The examination of neutral theory was thus restricted due to limited availability of DNA data.

This situation would change with the emergence in huge amount of DNA sequence data to perform experiments to validate the neutral theory. This DNA data came through development of new sequencing techniques.

According to the neutral theory, most of the mutant substitution causing evolution at the molecular level were done by random fixation, through sample drift of neutral mutants (selectively neutral to be precise) being under continued mutation pressure. (Kimura, M., 1983)

It is quite in contrast with the neo-Darwin or synthetic theory of evolution, which says that positive natural selection causes the spread of mutants within the species.

The neutral theory also propounds that intra specific variations produced at the molecular level is selectively neutral. This variability is maintained within the species by creation of a balance between mutation input and the random extinction of that mutation.

In other words, the neutral theory regards protein and DNA polymorphisms as a transient phase of molecular evolution (Kimura, M. et al., 1971). The neutral theory rejects the idea of majority of polymorphisms being adaptive and that they are maintained actively in the species through some act of balancing selection.

The neutral theory is apart from other traditional theories of evolution in the fact of it being quantitative- there are formulae for such quantities - as the rate of evolution and the amount of intraspecific variability. Moreover the validity of the formulae by comparing theoretical predictions with actual data can be checked.

First, let us consider a process which is cumulative, in which neutral mutants are getting substituted sequentially at a site by the continued input of new mutations through random genetic drift. Then we have this formula, for the rate of evolution per generation -

$$K_g = V_o, [1]$$

where  $K_g$  represents the long-term average per generation of the number of mutants that spread through the population and  $V_o$  is the rate of production of neutral mutants per locus (or site) per generation.

A property of neutral mutations that long term mutation substitution rate is equal to mutation rate is the basis of the above equation.



If we denote by  $V_T$  the total mutation rate, and if  $f_o$  is the fraction of neutral mutations at the time of occurrence, so that

$f_o = V_O/V_T$ , then Eq. 1 may be rewritten as

$$k_g = V_T f_o. [2].$$

Now, beneficial mutations can occur, but the neutral theory assumes that they are so rare and therefore they can be neglected in our quantitative consideration. Thus,  $(1 - f_o)$  is representing that fraction of definitely deleterious mutants which are removed or eliminated from the population and had not contributed to either evolution or polymorphism, even when the selective disadvantages involved are very small. The above formulation is so remarkably simple in the fact that the evolutionary rate (on the long-term basis) is independent of population size and environmental conditions of each organism.

In studies of molecular evolutionary, it is routine to measure the evolutionary rate in terms of years (i.e., the unit length of time is year) and not in terms of the generations.

Therefore, it is relevant to modify Eq. 2 so that it provide the evolutionary rate per year.

$$k_1 = (V_T/g)f_o. [3].$$

In the above equation  $g$  stands for the generation span (in years), and  $V_T/g$  is the total mutation rate per year.

### Evolution of DNA Sequences

It is now quite established as a fact that synonymous base substitutions that do not cause change in amino acid are occurring at much higher rate in evolution than the non synonymous substitution that are amino acid altering. Moreover it is now also established that substitutions occurring in the region of intronic region occur at rate equal to higher than the synonymous substitution rate.

When we consider the fact that natural selection acts upon the phenotype of the organism, these observations that are inclined towards the more occurrence of synonymous and other silent substitutions suggests that the molecular changes that are less likely to become the subject of natural selection occur more rapidly in evolution. Now, such molecular changes have the higher chance of being selectively neutral and hence can be better explained by neutral theory.

Neutral Theory predicts that the maximum evolutionary rate is set by the mutation rate ( $k_g \leq V_T$ ) and maximum rate will be attained when all the occurring mutation would be selectively neutral. ( $f_o = 1$ ). This point is validated through the discovery of high evolutionary rates in the pseudo genes. (Miyata T. et al., 1980) has showed this by doing a careful study of the evolutionary rate of pseudo alpha – globin gene in the mouse. Another study including a statistical analysis of the evolutionary rates of pseudo genes is carried out by (Li et al., 1981). An interesting fact is revealed by these studies. It is that the rate of substitution are equally high in all three codon positions. The estimated rate of substitution in globin pseudo genes is  $\sim k = 5 * 10^{-9}$

substitutions per nucleotide site per year in mammals. Now this is almost twice higher than the rate of substitution at 3<sup>rd</sup> codon position in normal globin genes (most of which are synonymous). This implies that the rate of synonymous mutation is subject to weak negative selection. It is therefore evident that pseudo genes are liberated from the constraint of negative selection and can thus accumulate deletions and additions at much higher rates.

### **3.5 Story of evolution of Humans from Chimpanzee**

About ten million years ago, in Africa there used to be quite a wetter climate than what we have presently. (Brunet, M., et al., 2002). From Atlantic to Indian Ocean there was an unbroken piece of low land covered wholly with Tropical rain forests. Starting from eight million years ago, the Africa was split into two due to development of tectonic forces. An east African rift valley was formed. An uplift was formed due to the tectonic forces that hindered the easterly flow of the rain clouds i.e. a rain shadow area was formed over east Africa. East Africa started to become dry.

These geological occurrences of events have caused the split of the common ancestors of modern chimpanzee and Humans into two geographically separate populations. One part of the population remained in wet, west Africa while the other in dry east Africa rift valley. This other part of population of dry East Africa, started to adapt to increasingly open and dry habitats of east and north central regions of Africa. This group of east Africa eventually evolved into modern Humans of today. (Alls, D. L., et al., 2003).

### **3.6 MEGA**

Availability : <http://www.megasoftware.net/>

The molecular evolution genetics analysis (MEGA) tool is an integrated software or suite that performs statistical based test on comparative analysis of molecular sequence data and is based on the evolutionary concepts. (Kumar et al.2008; Tamura et al.2011). This software is widely used by the biologists for reconstruction of evolutionary history of the species and to find the extent and nature of selection forces that have shaped the evolution of species and genes. MEGA is widely used as a tool in evolutionary bioinformatics.

MEGA key features include its Graphical User Interface (GUI), which provides visualization in detail and good exploration of sequence data in interactive way, phylogenetic trees and other analysis results. With increasing popularity and usefulness of MEGA the use and need of users requires multi-gene and genome-scale data analysis. This requires high throughput analysis. For this purpose, MEGA has re-engineered its source code i.e. computational core – which implements the algorithms for all analysis in MEGA, and now provides it as a stand alone program. And can be executed through command line by use of scripting language as perl.

MEGA-CC comes with MEGA-Proto. MEGA – Proto allows the generation of a configuration file. It ask the user to provide parameters for the analysis and it is the mirror of the GUI based MEGA application. As soon as the user is done with the selection of parameters and configuration, this selection is saved as the file. Using the MEGA-Proto application, the user first generates an analysis options file that specifies the chosen settings for the substitution model, genetic code table, gaps/missing data treatment, distribution of rates, topology search approach etc.

The analysis options file made by the MEGA-Proto has to be saved on the disk. Once this is done we can now execute MEGA-CC to perform our required task. There are many options for processing of the alignment files. MEGA-CC provides a simple method of ‘File Iterator’ system. File Iterator will search in the user specified directory for the input files and choose all the compatible files and process them iteratively.

Alternatively we can provide a text file for ‘File Iterator’ system which should contain the list of names of all the input files. But in order to use this feature, the user will have to launch MEGA-CC executable from a command prompt along with the appropriate parameter flag, and pass it a directory name or the name of a text file which lists all target input files. Alternatively, the user can generate their own script that iterates over the names of the alignment files and launches the MEGA-CC executable for each file. Supported input data files for MEGA-CC are FastA and MEGA files for sequence data, Newick files for phylogenies and MEGA files for distance matrices. Choose the desired settings in this dialogbox and saves them into the analysis options file in an appropriate directory for use with MEGA-CC.

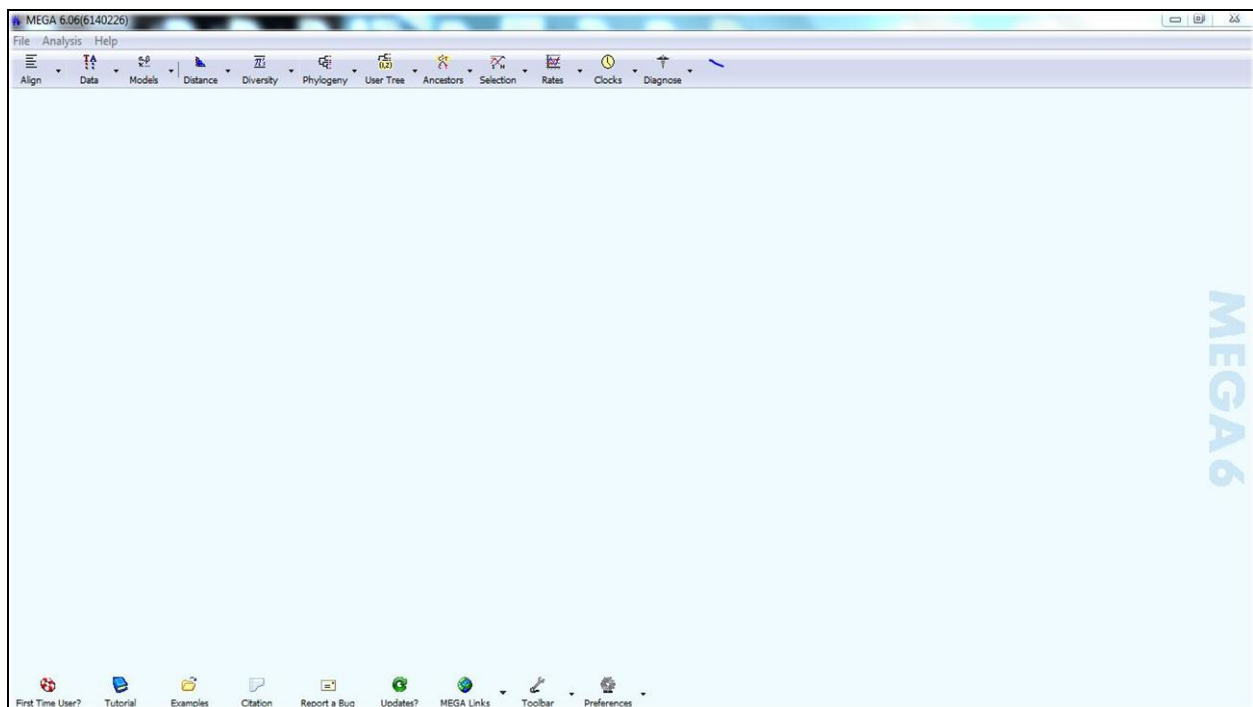


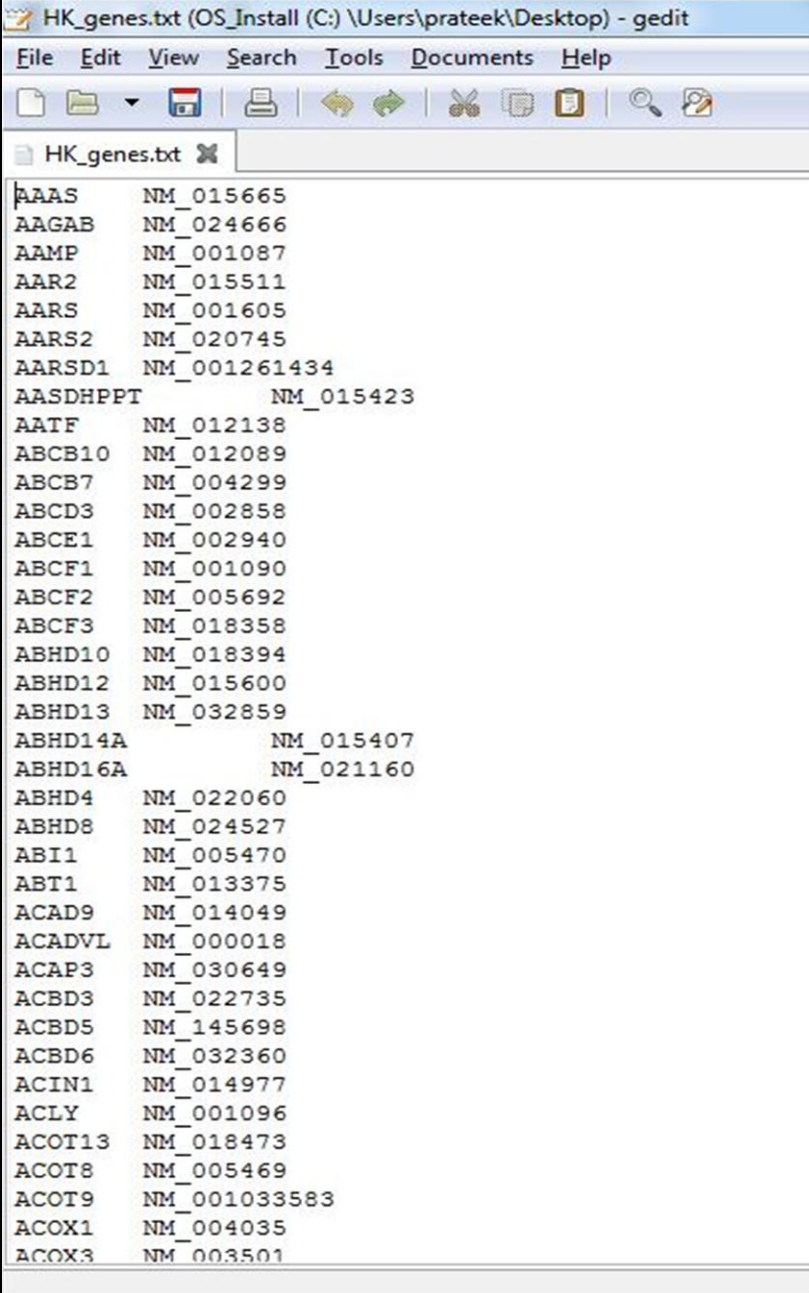
Fig 2: View of MEGA6 – GUI home page.

### **3.7 Background of the experiment**

Non synonymous substitutions are those nucleotide substitution that produce change in the codon in a manner that it now codes for a different amino acid and are hence subjected for evolution selection. While synonymous substitutions do not produce amino acid change. Synonymous substitutions are hence neutral as they do not bring any change in functional unit i.e. protein. Therefore synonymous substitution always occur as neutral rate because they are not subject to selection forces of nature. Hence if  $dN > dS$  , then  $dN/dS > 1$  , it means that non synonymous mutations are being fixed at a rate greater than that of synonymous mutations, hence there is occurring a positive selection. Similarly if  $dN < dS$ , then  $dN/dS < 1$ , it means that rate of non synonymous mutation is less than that of synonymous mutation which means that non synonymous mutations are deleterious and hence being eliminated by nature. This is called purifying selection. And similarly if  $dN/dS = 1$ , then no selection is under way and hence this is in accordance with kimura's neutral theory of evolution. (Yang, Z., 1997 ; Zhang, L., et al., 2005)

## 4.MATERIAL AND METHODS

- 1- Went to [www.tau.ac.il/~elieis/HKG](http://www.tau.ac.il/~elieis/HKG) website to collect all housekeeping genes of *Homo sapiens*. (Eisenberg, E. et al, 2013)
- 2- The file contains the gene symbol and RefSeq id in tab separated format.



```
HK_genes.txt (OS_Install (C:) \Users\prateek\Desktop) - gedit
File Edit View Search Tools Documents Help
HK_genes.txt
AAAS NM_015665
AAGAB NM_024666
AAMP NM_001087
AAR2 NM_015511
AARS NM_001605
AARS2 NM_020745
AARSD1 NM_001261434
AASDHPPT NM_015423
AATF NM_012138
ABCB10 NM_012089
ABCB7 NM_004299
ABCD3 NM_002858
ABCE1 NM_002940
ABCF1 NM_001090
ABCF2 NM_005692
ABCF3 NM_018358
ABHD10 NM_018394
ABHD12 NM_015600
ABHD13 NM_032859
ABHD14A NM_015407
ABHD16A NM_021160
ABHD4 NM_022060
ABHD8 NM_024527
ABI1 NM_005470
ABT1 NM_013375
ACAD9 NM_014049
ACADVL NM_000018
ACAP3 NM_030649
ACBD3 NM_022735
ACBD5 NM_145698
ACBD6 NM_032360
ACIN1 NM_014977
ACLY NM_001096
ACOT13 NM_018473
ACOT8 NM_005469
ACOT9 NM_001033583
ACOX1 NM_004035
ACOX3 NM_003501
```

Fig 3 : File HK\_genes.txt containing the Gene name and Ref\_seq id in tab separated format.

- 3- Wrote perl script 1 (see appendix) to take only the second column having just the RefSeq ids of the human house keeping genes.

```
HK_genes_refseq_ids (Data (D:) \project) - gedit
File Edit View Search Tools Documents Help
HK_genes_refseq_ids
NM_015665
NM_024666
NM_001087
NM_015511
NM_001605
NM_020745
NM_001261434
NM_015423
NM_012138
NM_012089
NM_004299
NM_002858
NM_002940
NM_001090
NM_005692
NM_018358
NM_018394
NM_015600
NM_032859
NM_015407
NM_021160
NM_022060
NM_024527
NM_005470
NM_013375
NM_014049
NM_000018
NM_030649
NM_022735
NM_145698
NM_032360
NM_014977
NM_001096
NM_018473
NM_005469
NM_001033583
NM_004035
NM_003501
```

Fig 4:- File HK\_genes\_refseq\_ids.txt formed after parsing HK\_genes.txt

- 4- There are total of 3805 human house keeping genes in the list. Now converted these total 3805 genes in text file each containing 500 ids because Genbank can accept 500 ids given together. This was done using perl script 2 (see appendix) .
- 5- Pasted the Refseq ids in NCBI nucleotide database search bar, 500 ids together at a time to get nucleotide sequences for the 3805 human house keeping genes. Downloaded all the sequences by going to download tab.
- 6- Installed BioPerl module of the perl. Typed CPAN Bio::Perl to install BioPerl on Command Prompt with access to internet.
- 7- Installed module to access BLAST remotely from the local system. Typed Bio::Tools::Run::RemoteBlast on cmd to install this module on the system with access to internet.
- 8- To check whether the module is installed correctly on the system , typed the following command on the cmd window –  
  
C:\> perl -MBio::Tools::Run::RemoteBlast -e 1
- 9- Used to standard module Bio::Tools::Run::Remote blast for getting the homologous sequences of Human house keeping genes in *Pan troglodytes*.
- 10- Now had files for all the Homologous sequence pairs between *Homo sapiens* and *Pan troglodytes* that are for human house keeping genes. For some genes a relevant homologous sequence was not found in *Pan troglodytes*.
- 11- As MEGA do not accept very long identification line , Wrote a perl script -3 (see appendix) to cut short the first identification line upto RefSeq id.
- 12- Converted the text files into fasta format as MEGA accepts only fasta format for the alignment . Wrote perl script 3 for it (see appendix).



```

*seqdump (2).txt X
>gi|331999976|ref|NM_013375.3| Homo sapiens activator of basal transcription 1 (ABT1), mRNA
AAGGCACCTTTACGGCCGTCGTGCCGCTCGTGTCAAGTGGAGGCAGAGGAATCGGAGAAGGCCGCA
ACGGAGCAAGAGCCGCTGGAAGGGACAGAACAGACACTAGATGCCGAGGAGGAGCAGGAGGAATCCGAAG
AAGCGGCTGTGGCAGCAAGAAACGGGTAGTCCAGGTATTGTGTACCTGGGCCATATCCCGCCGCGCTT
CCGGCCCTGCACGTCGCCAACCTTCTCAGCGCCTATGGCGAGGTCGGACGCGTCTTCTTTTCAGGCTGAG
GACCGGTTTCGTGAGACGCAAGAAGAAGGCAGCAGCAGCTGCCGGAGGAAAAAGCGGTCTACACCAAGG
ACTACACCGAGGGATGGGTGGAGTTCGGTGACAAGCGCATAGCCAAGCGCGTGGCGGCCAGTCTACACAA
CACGCCATGGGTGCCCGCAGGGCAGCCCTTCCGTTATGATCTTTGGAACCTCAAGTACTGCACCGT
GGCCGATGGGGACCTGCTCGCCAGATGGCTCCTGGACATTTGCCAGCGTCTACTGAGCAGGAACTG
AGGGCCCGTAAAGCAGCAGGGCCAGGGGACGTGAACGGGCTCGCCTGGCAACTGCCAGGACAAGGCC
GCTCCAACAAGGGCTCCTGGCCAGGATCTTTGGAGCCCGCCACCCTCAGAGAGCATGGAGGGACCTTC
CCTGTGACGGGACTCCTGAGGGCTGGGTGGCCCTTCCATTTCTGGCCCTGCTCTGCTTCTGTCTAC
CTCATACTAGAATGATCGTGACTACCCGGGCAGACATTTACTGTGTTTCTCAGACCAAGTGTCTACTGA
TGGC.....
.....TAATTC
TGTTGCTTGTGCTGTTTGTGTTTTCATCTGTCAATGTGATGATCTGTGTTTATAGGGTAGAGT
GGATTTGCTACTTTGGCTGTAAAATACCCTAATCACATTATGATCTTGACAGGTGCACCTTACTGGGGA
GAATAAAAAGGACCATACGGTAAA
>gi|410040395|ref|XM_001173335.3| PREDICTED: Pan troglodytes activator of basal transcription 1 (ABT1), mRNA
CTTTACGGCCGTCGTGCCGCTCGTGTCAAGTGGAGGCAGAGGAATCGGAGAAGGCCCAACGGAGCAAGAGCCGC
TGGAAGGGACAGAACAGACACTAGATGCCGAGGAGGAGCAGGAGGAATCCGAAGAAGCGGCCTGTGGCAGCAAGAAGCGG
GTAGTGGCAGGTATTGTGTACCTGGGCCATATCCCGCCGCGCTTCCGGCCCTGCACGTCGCGCAACCTTCTCAGCGCCTA
TGGCGAGGTGGACGCGTCTTCTTTTCAGGCTGAGGACCGGTTTCGTGAGACGCAAGAAGAAGGCAGCAGCAGCTGCCGGAG
GGAAAAAGCGGTCTACACCAAGGACTACACCGAGGGATGGGTGGAGTTCGGTGACAAGCGCATAGCCAAGCGCGTGGCG
GCCAGTCTACACAACACGCCTATGGGTGCCCGCAGGGCAGCCCTTCCGTTATGATCTTTGGAACCTCAAGTACTTGCA
CCGTTTACCTGGTCCACCTCAGCGAGCACCTCGCCTTTGAGCGCCAGGTGCGCAGGACGCGCTTGAGAGCGGAGGTTG
CTCAAGCCAAGCGTGAGACCGACTTCTATCTTCAAAGTGTGGAACGGGGACAACGCTTCTTTCGGCCGATGGGGACCCCT
CCCGCCACCCTCAGAGAGCATGGAGGGACCTTCCCTTGTGACGGGACTCCTGAGGGCCTGGGTGGCCCTTCCATTTCT
GGCCCTGCTCTGCTTCTGTCTACCTCATACTAGAATGATCGTGACTACCCGGGCAGACATTTACTGTGTTTCTCAGAC
ATTTTTATTGGTGGGGAGGTGTTGGACAAGTTCCAACCTTTCATCTTGTGTTCCCTTCCACTTCAATATCTGATCTTAGA
GTATTAAGAGTTTTACAGGCCACTAGTATAAGATAGGCATCGTGGTAGATGCGTATAAAGGGTGGGAATGGGAGCCATGGC
AGGTCA.....
.....TTAAG
CCCATTGAGTGTAGAAAATAATTTTTACATCCACAATTTTGTCTATTTAAACAGCACTTGTTTTTTTATATGAAGT
AACTCATTTAATCCACAACATATAATATTGTTATCCCTCCATTAATAACTGAGCTCAGAGAGATGACTCAA

```

Fig 5:- Each single file contains a pair of homologous sequence



```

seqdump (2).fasta X
>gi|331999976|ref|NM_013375.3|
AAGGCACCTTTACGGCCGTCGTGCCGCTCGTGTCAACATGGAGGCAGAGGAATCGGAGAAGGCCGCA
ACGGAGCAAGAGCCGCTGGAAGGGACAGAACAGACACTAGATGCGGAGGAGGAGCAGGAGGAATCCGAAG
AAGCGCCTGTGGCAGCAAGAAACGGGTAGTGCCAGGTATTGTGTACCTGGGCCATATCCCGCCGCGCTT
CCGGCCCCTGCACGTCCGCAACCTTCTCAGCGCTATGGCGAGGTCGGACGCGTCTTCTTTCAGGCTGAG
GACCGGTTTCGTGAGACGCAAGAAGAAGGCAGCAGCAGCTGCCGGAGGAAAAAGCGGTCTACACCAAGG
ACTACACCGAGGGATGGGTGGAGTTCGGTGACAAGCGCATAGCCAAGCGGTGGCGGCCAGTCTACACAA
CACGCCATATGGGTGCCCGCAGGCGCAGCCCCTTCCGTTATGATCTTTGGAACCTCAAGTACTTGACCCTG
GGCCGATGGGGACCCTGCTCGCCAGATGGCTCCTGGACATTTGCCAGCGTCTACTGAGCAGGAAGTGG
AGGGCCCGTAAAGCAGCAGGCCAGGGGACGTGAACGGGCTCGCCTGGCAACTGCCAGGACAAGGCC
GCTCCAACAAAGGGCTCCTGGCCAGGATCTTTGGAGCCCCGCCACCCTCAGAGAGCATGGAGGGACCTT
CCTTGTGAGGGACTCCTGAGGGCCTGGGTGGCCCCTTCCATTTCTGGCCCTGCTCTGCTTCTGTCTAC
CTCATACTAGAATGATCGTGACTACCGGGCAGACATTTTACTGTGTTTCTCAGACCAAGTGTCTACTGA
TGGC.....
.....TAATTTT
TGTTTGCTTGTGCTGTTGTTTTGTTTTTTCATCTGTCAATGTGATGATCTGTGTTTTATAGGGTAGAGT
GGATTTGTCTACTTTGGCTGTAAAATACCCTAATCACATTATGATCTTGACAGGTGCACTTTACTGGGGA
GAATAAAAAGGACCATACGGTAAA
>gi|410040395|ref|XM_001173335.3|
CTTTACGGCCGTCGTGCCGCTCGTGTCAACATGGAGGCAGAGGAATCGGAGAAGGCCGCAACGGAGCAAGAGCCGC
TGGAAGGGACAGAACAGACACTAGATGCGGAGGAGGAGCAGGAGGAATCCGAAGAAGCGCCTGTGGCAGCAAGAAGCGG
GTAGTGCCAGGTATTGTGTACCTGGGCCATATCCCGCCGCGCTCCGGCCCCTGCACGTCCGCAACCTTCTCAGCGCCTA
TGGCGAGGTCGGACGCGTCTTCTTTCAGGCTGAGGACCGGTTTCGTGAGACGCAAGAAGAAGGCAGCAGCAGCTGCCGGAG
GGAAAAAGCGGTCTACACCAAGGACTACCCGAGGGATGGGTGGAGTTCGGTGACAAGCGCATAGCCAAGCGGTGGCG
GCCAGTCTACACAACACGCTATGGGTGCCCGCAGGCGCAGCCCCTTCCGTTATGATCTTTGGAACCTCAAGTACTTGCA
CCGTTTCACTGGTCCCACCTCAGCGAGCACCTCGCCTTTGAGCGCCAGGTGCGCAGGCAGCGCTTGAGAGCGGAGGTTG
CTCAAGCCAAGCGTGAGACCGACTTCTATCTTCAAAGTGTGGAACGGGGACAACGCTTCTTTCGGGCCGATGGGGACCCT
CCCCGCCACCCTCAGAGAGCATGGAGGGACCTTCCCTTGTGAGGGACTCCTGAGGGCCTGGGTGGCCCCTTCCATTTCT
GGCCCTGCTCTGCTTCCGTCTACCTCATACTAGAATGATCGTGACTACCGGGCAGACATTTTACTGTGTTTCTCAGAC
ATTTTTATTGGTGGGGAGGTGTTGGACAAGTTCCAACCTTTCATCTGTGTTCCCTTACCTTCATATCCTGATCTTAGA
GCCCCCTCCCCCTGCCACCCACCTTACTGTTAACTGGATTTTTTTTTTCTATTTAATTTTTGTCTAATATCTTAGCCC
AGTTTATCAATCAGTTATCTTAAGTCAGCATTCTTAAGCCATTGTTTGGAGAAACAGTGACAATAGGTAACACATCTTA
GTATTAAGAGTTTTACAGGCCACTAGTATAAGATAGGCATCGTGGTAGATGCGTATAAAGGGTGAATGGGAGCCATGGC
AGGICA.....
.....TTAAG
CCCATTGAGTGTAGGAAAAATAATTTTTTACATCCACAATTTTGTCTTATTTAAACAGCACTTGTTTTTTTATATGAAGT
AACTCATTAAATCCACAACCTATATAATATTGTTATCCCTCCATTAATAACTGAGCTCAGAGAGATGACTCAA

```

Fig 6: View of file after step no. 11

- 13- Now our files are ready to be fed into MEGA software.
- 14- Downloaded and Installed MEGA-CC software. MEGA-CC is computational core package to automate the process through command line or scripting language. MEGA GUI has to be used manually analyzing one sequence at a time.

15- MEGA-CC contains two application M6CC and M6Proto.

#### 4.1 Using MEGA-CC for alignment of the homologous sequence pair

16- Opened M6Proto. This is same as MEGA-GUI except it cannot run algorithms. It is used to save parameter settings with which you want to run the algorithm.

17- First the pair of sequences needed to be aligned. So selected the parameters for alignment of two nucleotide sequences as shown in figure below.

18- Now saved these settings with the name –align.mao. A file is formed with the name ‘align’ and having extension .mao

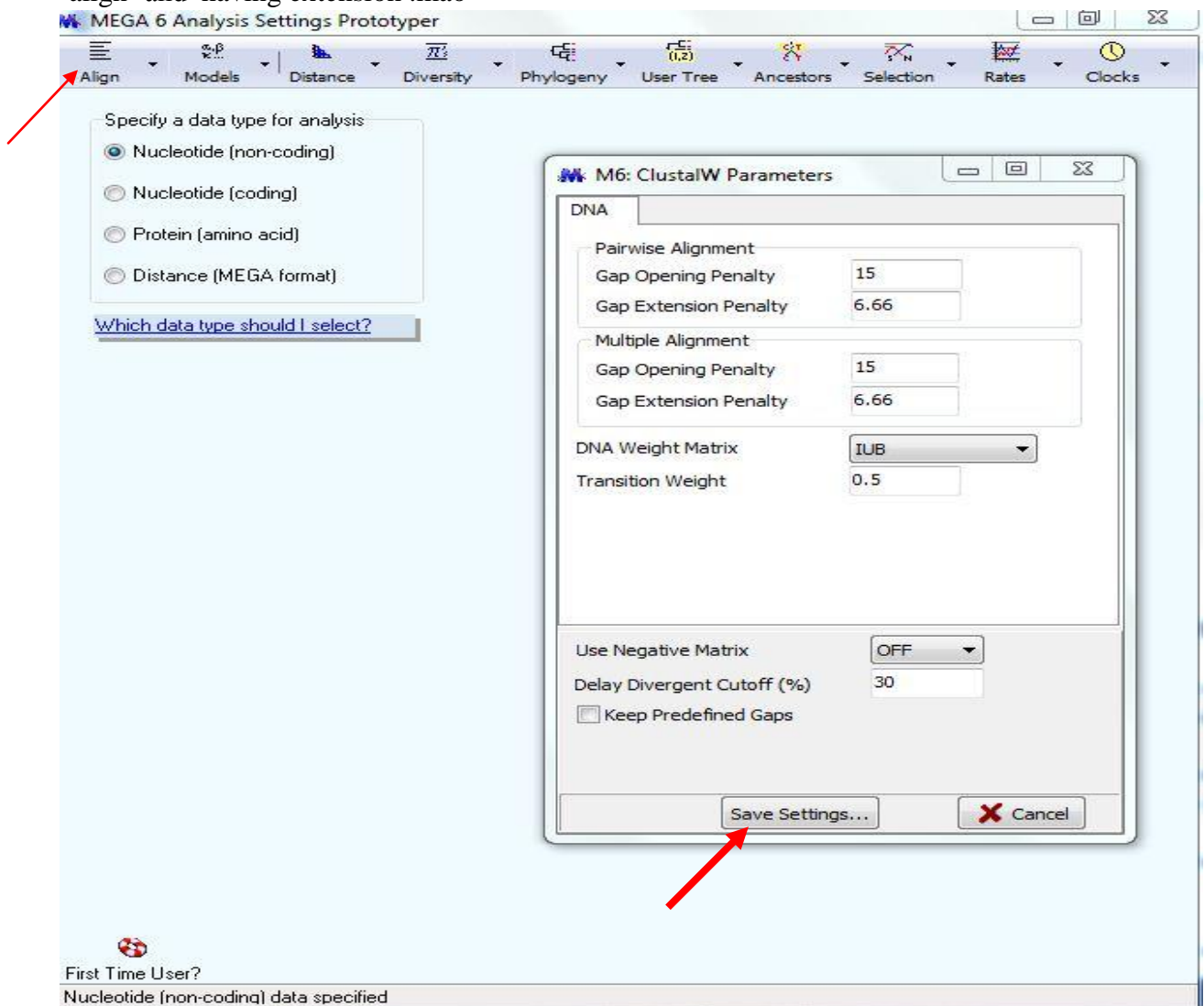
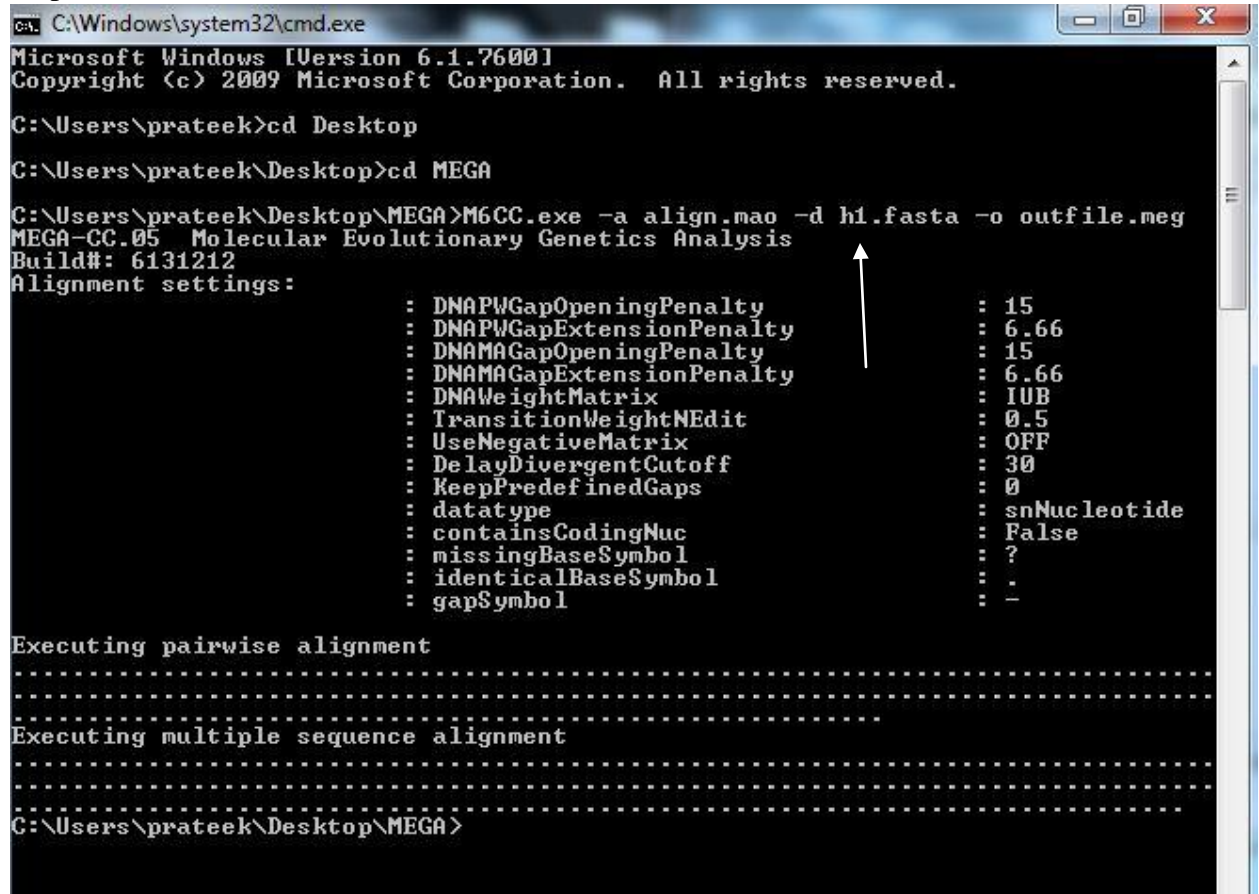


Fig 7:- Making the analysis settings .mao file for the alignment.

19- Have to execute M6CC for the alignment using align.mao as parameter settings. The command is –

```
C:/> M6CC.exe -a align.mao -d datafile.fasta -o outFile.meg
```

Where -a is the flag indicating analysis settings file, -d indicates your file that contains data to be analysed and -o for output file. As we want output to be in .meg extension (mega format) as input for selection test we use .meg extension in name of output file.



```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7600]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\prateek>cd Desktop
C:\Users\prateek\Desktop>cd MEGA
C:\Users\prateek\Desktop\MEGA>M6CC.exe -a align.mao -d h1.fasta -o outfile.meg
MEGA-CC.05 Molecular Evolutionary Genetics Analysis
Build#: 6131212
Alignment settings:
      : DNAPWGapOpeningPenalty           : 15
      : DNAPWGapExtensionPenalty        : 6.66
      : DNAMAGapOpeningPenalty          : 15
      : DNAMAGapExtensionPenalty        : 6.66
      : DNAWeightMatrix                 : IUB
      : TransitionWeightNEdit           : 0.5
      : UseNegativeMatrix                : OFF
      : DelayDivergentCutoff             : 30
      : KeepPredefinedGaps              : 0
      : datatype                        : snNucleotide
      : containsCodingNuc                : False
      : missingBaseSymbol                : ?
      : identicalBaseSymbol              : .
      : gapSymbol                        : -

Executing pairwise alignment
.....
Executing multiple sequence alignment
.....
C:\Users\prateek\Desktop\MEGA>
```

Fig 8:- Executing MEGA-CC for alignment of input file h1.fasta

20- Now wrote a perl script 4 (see appendix) to automate the procedure of alignment for many files.

21- Using script of step no. 20 also formed a new folder that contained all the alignment output files with extension '.meg'. See figure below to see example of output file generated after alignment.



## 4.2 Using MEGA-CC for positive selection test of the homologous sequence pair

- 22- Opened M6CC-Proto and selected the option of 'coding nucleotide'.
- 23- Went to "Selection" tab in the menu bar and clicked on Codon-based Z test of selection.
- 24- Chose parameters as shown in the figure below. Saved these settings for selection test analysis with the file name analysis\_settings.mao.

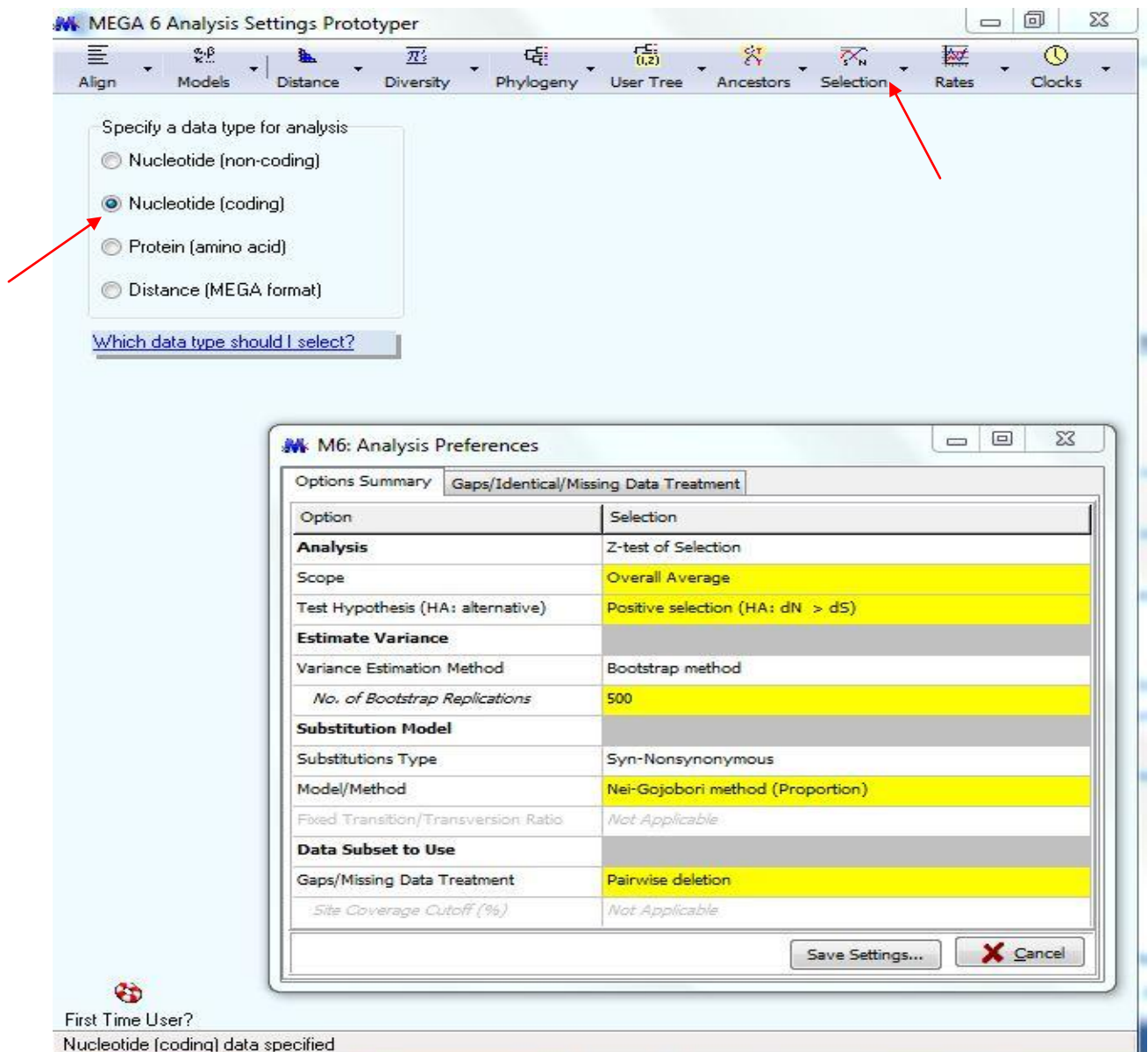


Fig 9 : The standard settings selected for Z test of positive selection for pair of 2 sequences

25- Wrote perl script 5 (see appendix) to automate the execution of M6CC using analysis\_settings.mao as analysis settings file and alignment output file as data files.

26- Output files created are stored in a new folder that is formed using the script of step no. 2

## 5.RESULT AND DISCUSSION

- After the execution of MEGA- CC for Z test of positive selection, 2 result files were generated.
- 1st file is the text summary file and 2<sup>nd</sup> file is mega file that contains the value of probability and statistic ( $d_N - d_S$ ), that is used to calculate the value of probability.
- The value of probability is  $> 0.05$  for all the homologous pairs and for most of the pairs it is equal to 1.00.
- The probability of rejecting the null hypothesis of strict-neutrality ( $d_N = d_S$ ) in favor of the alternative hypothesis ( $d_N > d_S$ ) (in the Probability column) is shown. Values of  $P$  less than 0.05 are considered significant at the 5% level and are highlighted.
- The test statistic ( $d_N - d_S$ ) is shown in the Statistic column.  $d_S$  and  $d_N$  are the numbers of synonymous and nonsynonymous substitutions per site, respectively.
- The table below reports the result of 10 such homologous pairs.
- The result is in corroboration with a similar study conducted using 161 human house keeping genes. There was no positive selection detected. (Zhang, L., et al., 2007)

Human Gene RefSeq Id	Chimp gene RefSeq Id	Probability	Statistic
NM_013375.3	XM_001173335.3	1.00	-0.797
NM_013333.3	XM_001137261.3	1.00	-0.62907953
NM_004446.2	XM_001172425.2	1.00	-0.25211827
NM_005702.2	XM_511365.3	1.00	-1.27289933
NM_016442.3	XM_003310766.2	1.00	-0.71010391
NM_001983.3	AC193935.3	1.00	-0.64578082
NM_000400.3	NM_001246590.1	1.00	-0.17127437
XM_005252652.1	XM_507693.2	0.31511748	0.48263392
NM_000122.1	XM_003828264.1	1.00	-2.15705690
NM_000123.3	XM_003314205.1	1.00	-2.18771552

Table 1: Showing result of 10 homologous sequence pairs between Human and Chimpanzee

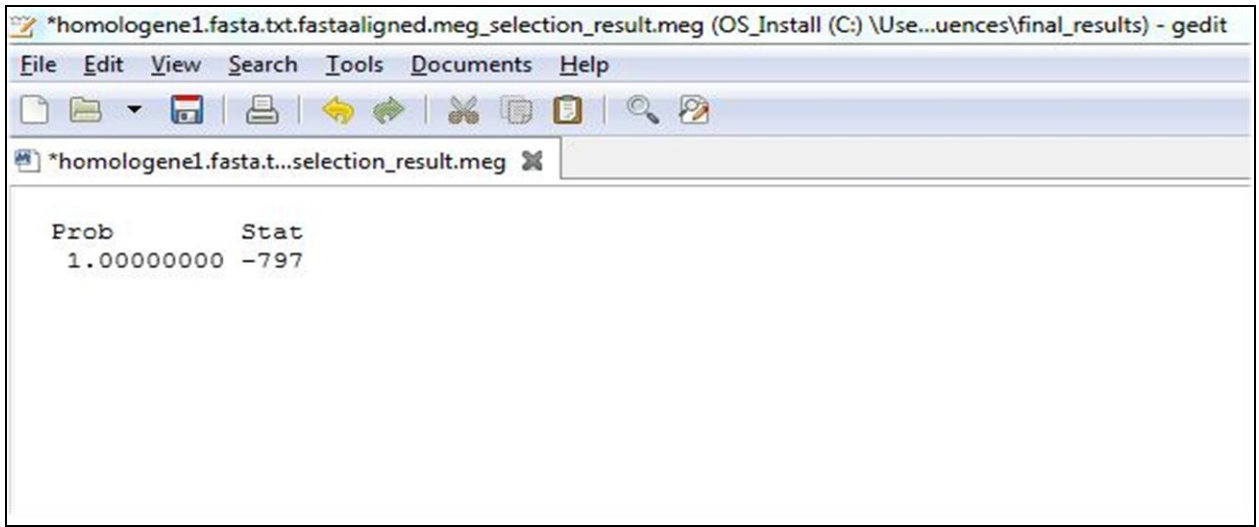


Fig 9: Output file for Sequence pair NM\_013375.3 and XM\_001173335.

**Text of Output summary file**

; Suggested Citation for MEGA-CC:

; MEGA-CC: Computing Core of Molecular Evolutionary Genetics

; Analysis Program for Automated and Iterative Data Analysis.

; Bioinformatics (2012) doi:10.1093/bioinformatics/bts507

;

\*\*\*\*\*  
 \*\*\*\*

; Please see the important message(s) at the bottom of this file

;

\*\*\*\*\*  
 \*\*\*\*

[General Info]

Data Type = nucleotide (non-coding)

No. of Taxa = 2

No. of Sites = 2038

Data File =  
'C:/Users/prateek/Desktop/MEGA/homologene\_fasta/aligned\_sequences/homologene1.fasta.txt.fastaaligned.meg'

Settings File = 'C:/Users/prateek/Desktop/MEGA/analysis\_settings.mao'

Command Line = C:/Users/prateek/Desktop/MEGA/M6CC.exe -a  
C:/Users/prateek/Desktop/MEGA/analysis\_settings.mao -d  
C:/Users/prateek/Desktop/MEGA/homologene\_fasta/aligned\_sequences/homologene1.fasta.txt.fastaaligned.meg -o  
C:/Users/prateek/Desktop/MEGA/homologene\_fasta/aligned\_sequences/final\_results/homologene1.fasta.txt.fastaaligned.meg\_selection\_result.meg

[Analysis Settings]

Analysis = Z-test of Selection

Scope = Overall Average

Test Hypothesis (HA: alternative) = Positive selection (HA:  $dN > dS$ )

Variance Estimation Method = Bootstrap method

No. of Bootstrap Replications = 500

Substitutions Type = Syn-Nonsynonymous

Model/Method = Nei-Gojobori method (Proportion)



Fixed Transition/Transversion Ratio = Not Applicable  
Gaps/Missing Data Treatment = Pairwise deletion  
Site Coverage Cutoff (%) = Not Applicable  
datatype = snNucleotide  
containsCodingNuc = True  
missingBaseSymbol = ?  
identicalBaseSymbol = .  
gapSymbol = -  
Select Codon Positions = Select Codon Positions=1st, 2nd, 3rd, Non-Coding

[Analysis Statistics]

Execution Time = 0.671 (seconds)  
Peak Memory Used(Working Set) = 12.984 (MB)

;  
\*\*\*\*\*  
\*\*\*\*

; Please note the following important messages:

;  
\*\*\*\*\*

\*\*\*\*

; Warning: Stop codon(s) were found in your alignment. They were removed and the analysis continued.

## 6. CONCLUSION

As all results did not reject the hypothesis to favor the alternative hypothesis – HA:  $dN > dS$ , it can be safely said that human housekeeping genes are not under positive selection. By this it means that non synonymous mutations that bring change in gene function in order to adapt are not under any selection force of nature as proposed by Darwin. It essentially means that the formation of new amino acid is either disfavored or neutrally favored. This is not in line with the Darwin theory of evolution according to which both of the species must be adapting according to their environment and in order to accomplish this their genes must be under force of gradual natural selection to positively select for variations that favor in adaptation.

Here we cannot possibly give the specific nature or mechanism of path and manner in which all species were formed from complex structure organism to simple structure organism. But all the evidences of present day indicate that the path of devolution is quite possible. From drug resistance in bacteria , adaption in laboratory mice for much desired characteristic of unregulated reproduction, several disease resistance in humans, development of drug that would cripple a particular gene, to phenomenon of adaptive pseudogenization, all show the motif of loss of genes. On the other hand, there is huge improbability calculated in formation of new stable and functional protein as well as there is not a single gene or protein that is newly formed as a part of adaption by an organism and can be given as a supportive example for Darwin theory of evolution.

## 7.FUTURE PERSPECTIVE

Darwin easily explained the path of evolution by citing morphological evidence. But with knowledge of sequencing at genome and proteome level, several scientists have started to doubt the formation of different life forms in a manner according to theory of natural selection. Hence inquiries in origin of life and tree of life hold lots of promise.

Role of Bioinformatics can be crucial in the development of the subject of evolution. There is a scope in simulation of process of randomization in mutations that happen in genome. The phenomenon of chance associated with evolution can be programmed and tested over the virtual biological data available. And with the completion of GWAS (Genome Wide Association Studies) project, the SNP variation association data with disease is available. To associate SNP variation with evolution is a potential possibility. Two efforts are already made in this direction. Cheng, F., et al.,2009 ; Voight, B. F., et al., 2006)

Overall computational statistics and programming can play a crucial information in mining patterns and knowledge from large amounts of sequence data available.

Apart from the significance of larger questions like “from where we have come?” or “who are we?” and “where we are heading to?”, the studies on evolution could be hugely beneficial in the field of medicine. There can be a study to find a pattern or parameters involved in loss of genes and complexity that have formed different species in order to survive. Only which kind of genes can be lost or if two particular genes have to be lost together or if there is a sequence to loss of gene?. Or if we can exploit the intra human species variation data and then point out the adaption with geography. Answer to all these questions can be useful for the field of medicine.

## 8. REFERENCES

- Alles, D. L., & Stevenson, J. C. (2003). Teaching Human Evolution. *The American Biology Teacher*, Vol. 65, No. 5(May), 333-339.
- Behe, M. J. (2009). Irreducible complexity: Obstacle to Darwinian evolution. *Philosophy of Biology: An Anthology*, 427.
- Blanpain, C., Libert, F., Vassart, G., & Parmentier, M. (2002). CCR5 and HIV infection. *Receptors and Channels*, 8(1), 19-31.
- Bowie, J. U., & Sauer, R. T. (1989). Identifying determinants of folding and activity for a protein of unknown structure. *Proceedings of the National Academy of Sciences*, 86(7), 2152-2156.
- Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A., & Sauer, R. T. (1990) Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitution, *Science* 247, 1306-1310.
- Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., ... & Zollikofer, C. (2002). A new hominid from the Upper Miocene of Chad, Central Africa. *Nature*, 418(6894), 145-151.
- Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., ... & Zollikofer, C. (2002). A new hominid from the Upper Miocene of Chad, Central Africa. *Nature*, 418(6894), 145-151.
- Cheng, F., Chen, W., Richards, E., Deng, L., & Zeng, C. (2009). SNP@ Evolution: a hierarchical database of positive selection on the human genome. *BMC evolutionary biology*, 9(1), 221.
- Cohorts, D. Loss-of-Function Mutations in APOC3, Triglycerides, and Coronary Disease.
- Durston, K. K., Chiu, D. K., Abel, D. L., & Trevors, J. T. (2007). Theoretical Biology and Medical Modelling. *Theoretical Biology and Medical Modelling*, 4, 47.
- Eisenberg, E. , Levanon, E.Y., (2013) *Trends in Genetics*, 29 .
- Fairhurst, R. M., Baruch, D. I., Brittain, N. J., Ostera, G. R., Wallach, J. S., Hoang, H. L., ... & Wellems, T. E. (2005). Abnormal display of PfEMP-1 on erythrocytes carrying haemoglobin C may protect against malaria. *Nature*, 435(7045), 1117-1121.
- Fink, A. L., Calciano, L. J., Goto, Y., Kurotsu, T., & Palleros, D. R. (1994). Classification of acid denaturation of proteins: intermediates and unfolded states. *Biochemistry*, 33(41), 12504-12511.
- Geach, T. (2014). Genetics: APOC3 mutations lower CVD risk. *Nature Reviews Cardiology*.

- He, S. Y. (1998). Type III protein secretion systems in plant and animal pathogenic bacteria. *Annual review of phytopathology*, 36(1), 363-392.
- JÃrgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G., & TybjÃrg-Hansen, A. (2014). Loss-of-Function Mutations in APOC3 and Risk of Ischemic Vascular Disease. *New England Journal of Medicine*.
- Kim, H. J., Kwak, H. K., Lee, J., Yun, Y. J., Lee, J. S., Lee, M. S., ... & Lee, K. H. (2012). Patterns of pncA mutations in drug-resistant Mycobacterium tuberculosis isolated from patients in South Korea. *The International Journal of Tuberculosis and Lung Disease*, 16(1), 98-103.
- Kim, J. F. (2001). Revisiting the chlamydial type III protein secretion system: clues to the origin of type III protein secretion. *Trends in Genetics*, 17(2), 65-69.
- Kimura, M. & Ohta, T. (1971) *Nature (London)* 229, 467-469.
- Kimura, M. (1968) *Nature (London)* 217, 624-626.
- Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, U.K.)
- Kleinman, H. K., Ebihara, I., Killen, P. D., Sasaki, M., Cannon, F. B., Yamada, Y., & Martin, G. R. (1987). Genes for basement membrane proteins are coordinately expressed in differentiating F9 cells but not in normal adult murine tissues. *Developmental biology*, 122(2), 373-378.
- Kumar, S., Stecher, G., Peterson, D., & Tamura, K. (2012). MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics*, 28(20), 2685-2686.
- Lievre, A., Bachet, J. B., Le Corre, D., Boige, V., Landi, B., Emile, J. F., ... & Laurent-Puig, P. (2006). KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer research*, 66(8), 3992-3995.
- Luzzatto, L. (2012). Sick cell anaemia and malaria. *Mediterranean journal of hematology and infectious diseases*, 4(1).
- Macnab, R. M. (1999). The bacterial flagellum: reversible rotary propeller and type III export apparatus. *Journal of bacteriology*, 181(23), 7149-7153.
- Meyer, S. C. (2014). *Darwin's doubt: The explosive origin of animal life and the case for intelligent design*. HarperOne.
- Miyata, T., & Yasunaga, T. (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *Journal of Molecular Evolution*, 16(1), 23-36.

Muchmore, E. A., Diaz, S., & Varki, A. (1998). A structural difference between the cell surfaces of humans and the great apes. *American journal of physical anthropology*, 107(2), 187-198.

Nei, M., & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, 3(5), 418-426.

Nguyen, L., Paulsen, I. T., Tchieu, J., Hueck, C. J., & Saier Jr, M. H. (2000). Phylogenetic analyses of the constituents of type III protein secretion systems. *J. Mol. Microbiol. Biotechnol*, 2(2), 125-144.

Olson, M. V. (1999). When less is more: gene loss as an engine of evolutionary change. *The American Journal of Human Genetics*, 64(1), 18-23.

Olson, M. V., & Varki, A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. *Nature Reviews Genetics*, 4(1), 20-28.

Pallen, M. J., & Matzke, N. J. (2006). From *The Origin of Species* to the origin of bacterial flagella. *Nature Reviews Microbiology*, 4(1), 784-790.

Puente, X. S., Velasco, G., Guti rrez-Fern ndez, A., Bertranpetit, J., King, M. C., & L pez-Ot n, C. (2006). Comparative analysis of cancer genes in the human and chimpanzee genomes. *BMC genomics*, 7(1), 15.

R.T. Sauer, James U Bowie, John F.R. Olson, and Wendall A. Lim, (1989) *Proceedings of the National Academy of Science* USA 86, 2152-2156.

Reidhaar-Olson, J. F., & Sauer, R. T. (1990). Functionally Acceptable Substitutions in Two - Helical Regions of Repressor, Proteins: Structure, Function, and Genetics 7, 306-316.

Sean Pitman, *The Evolution of the Flagellum* (2010)

Sequencing, T. C., & Analysis Consortium. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055), 69-87.

Stephens, J. C., Reich, D. E., Goldstein, D. B., Shin, H. D., Smith, M. W., Carrington, M., ... & Dean, M. (1998). Dating the Origin of the *CCR5* AIDS-Resistance Allele by the Coalescence of Haplotypes. *The American Journal of Human Genetics*, 62(6), 1507-1515.

Sukhan, A., Kubori, T., Wilson, J., & Gal n, J. E. (2001). Genetic Analysis of Assembly of the *Salmonella enterica* Serovar Typhimurium Type III Secretion-Associated Needle Complex. *Journal of bacteriology*, 183(4), 1159-1167.

Tamarkin, L., Baird, C. J., & Almeida, O. F. (1985). Melatonin: a coordinating signal for mammalian reproduction?. *Science*, 227(4688), 714-720.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10), 2731-2739.

Tamura, K., Stecher, G., Peterson, D., Filipowski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* 30: 2725-2729

Thirumalai, D., & Klimov, D. K. (1999). Emergence of stable and fast folding protein structures. *arXiv preprint cond-mat/9910248*.

Torrents, D., Suyama, M., Zdobnov, E., & Bork, P. (2003). A genome-wide survey of human pseudogenes. *Genome research*, 13(12), 2559-2567.

Tournamille, C., Colin, Y., Cartron, J. P., & Le Van Kim, C. (1995). Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy negative individuals. *Nature genetics*, 10(2), 224-228.

Varki, A. (2000). A chimpanzee genome project is a biomedical imperative. *Genome research*, 10(8), 1065-1070.

Voight, B. F., Kudravalli, S., Wen, X., & Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS biology*, 4(3), e72.

Wang, X., Grus, W. E., & Zhang, J. (2006). Gene losses during human origins. *PLoS biology*, 4(3), e52.

Wang, X., Grus, W. E., & Zhang, J. (2006). Gene losses during human origins. *PLoS biology*, 4(3), e52.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13:555-556.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in ecology & evolution*, 18(6), 292-298.

Zhang, L., & Li, W. H. (2005). Human SNPs reveal no evidence of frequent positive selection. *Molecular biology and evolution*, 22(12), 2504-2507.

Zhang, Z., Harrison, P. M., Liu, Y., & Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome research*, 13(12), 2541-2558.



## 9. APPENDIX

### Perl script 1

```
#!/usr/bin/perl -w
$filename = 'gene_result.txt' ;
open(fh1, $filename) ;
@a = <fh1> ;
close fh1 ;
$string = join ("", @a) ;
#print $string ;
while ($string =~ /(NC_.*?)\t/g)
{
$gene_id = $& ;
$gene_id =~ s/\s//g ;
print $gene_id, "\n" ;
}
```

### Perl script 2

```
#!/usr/bin/perl/ -w
$filename = "HK_genes.txt" ;
open(fh1, $filename);
@array = <fh1> ;
close fh1 ;
$prateek = "p" ;
$prateek1 = "p1" ;
$prateek2 = "p2" ;
$prateek3 = "p3" ;
$prateek4 = "p4" ;
$prateek5 = "p5" ;
$prateek6 = "p6" ;
$prateek7 = "p7" ;
$q = 1;
foreach $line (@array)
{
@vals = split(' ', $line);
print $vals[1], "\n" ;

if ($q <= 500)
{
```

```
open (OUTPUTFD, ">$prateek");  
print OUTPUTFD $vals[1], "\n" ;  
}
```

```
elsif ($q <= 1000)  
{  
open (OUTPUTFD1, ">$prateek1");  
print OUTPUTFD1 $vals[1], "\n" ;  
}
```

```
elsif ($q <= 1500)  
{  
open (OUTPUTFD2, ">$prateek2");  
print OUTPUTFD2 $vals[1], "\n" ;  
  
}
```

```
elsif ($q <= 2000)  
{  
open (OUTPUTFD3, ">$prateek3");  
print OUTPUTFD3 $vals[1], "\n" ;  
}
```

```
elsif ($q <= 2500)  
{  
open (OUTPUTFD4, ">$prateek4");  
print OUTPUTFD4 $vals[1], "\n" ;  
}
```

```
elsif ($q <= 3000)  
{  
open (OUTPUTFD5, ">$prateek5");  
print OUTPUTFD5 $vals[1], "\n" ;  
}
```

```
elsif ($q <= 3500)  
{
```

```
open (OUTPUTFD6, ">$prateek6");
print OUTPUTFD6 $vals[1], "\n" ;
}
```

```
elsif ($q <= 4000)
{
open (OUTPUTFD7, ">$prateek7");
print OUTPUTFD7 $vals[1], "\n" ;
}
++$q ;
}
```

### Perl script 3

```
#!/usr/bin/perl -w
@files = ();
$folder = 'homologenes' ;
opendir(FOLDER, $folder);
@files = readdir(FOLDER) ;
closedir(FOLDER) ;
mkdir "homologene_fasta" ;
foreach $file (@files)
{
open(fh, "$folder/$file") ;
@array = <fh>;
close fh ;
$string = join(" ", @array) ;
$string =~ s/(|\s.*?)\n\n/g ;
open (OUTPUTFD, ">homologene_fasta/$file.fasta");
print OUTPUTFD $string ;
}
```

### Perl script 4

```
#!/usr/bin/perl -w
$folder = "homologene_fasta" ;
opendir(FOLDER, $folder) ;
@files = () ;
@files = readdir(FOLDER) ;
closedir(FOLDER);
```

```

shift @files ;
shift @files ;
mkdir "$folder/aligned_sequences" ;
foreach $file (@files)
{
$output_name = $file."aligned" ;
system("C:/Users/prateek/Desktop/MEGA/M6CC.exe -a
C:/Users/prateek/Desktop/MEGA/align.mao -d C:/Users/prateek/Desktop/MEGA/$folder/$file -
o C:/Users/prateek/Desktop/MEGA/$folder/aligned_sequences/$output_name.meg");
}
exit ;

```

### Perl script 5

```

#!/usr/bin/perl -w
$folder = "homologene_fasta/aligned_sequences" ;
opendir(FOLDER, $folder) ;
@files = ( ) ;
@files = readdir(FOLDER) ;
closedir(FOLDER);
shift @files ;
shift @files ;
mkdir "$folder/final_results" ;
foreach $file (@files)
{
$output_name = $file."_selection_result" ;
system("C:/Users/prateek/Desktop/MEGA/M6CC.exe -a
C:/Users/prateek/Desktop/MEGA/analysis_settings.mao -d
C:/Users/prateek/Desktop/MEGA/$folder/$file -o
C:/Users/prateek/Desktop/MEGA/$folder/final_results/$output_name.meg");
}
exit ;

```

### Perl Script 6 #if we chose option sequence pairs instead of overall average

```

#!/usr/bin/perl -w
@files = ( );
$folder = 'final_results' ;
opendir(FOLDER, $folder);
@files = readdir(FOLDER) ;

```

```

closedir(FOLDER) ;
shift @files ;
shift @files ;
$count = 0 ; # number of prob val <0.05
$count1 = 0 ; #total count of prob val

for($i = 0 ; $i < scalar @files ; $i= $i+3)
{
open (fh, "$folder/$files[$i]");
@text = <fh> ;
close fh ;
$text_as_string = join(" ", @text) ;
while($text_as_string =~ /\[d\]\s*(\d.*?)\n/g)
{
$string = $1 ;
$string =~ s/[.*/]//g ;
@array = split (' ', $string) ;
foreach $arr (@array)
{
if ($arr =~ /0\04|0\03|0\02|0\01|0\00/g)
{
++$count;
}
++$count1 ;
}
print $string ;
$array = ();
}
}
print "\n", "number of prob val <0.05 = ", $count, "\n" ;
print "total count of prob val = ", $count1 ;
$per = $count/$count1 *100 ;
print "\n", "percentage of positive selection of genes= ", $per ;
exit ;

```

---