International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 5 Issue4, April-2016

A Natural Selection Analysis of Amino Acid Residues

Sangharsh Saini

Department of Computer Science & Applications, S.D. College (Lahore) Ambala Cantt., Haryana, India

ABSTRACT

Natural selection is the basic mechanism underlying all adaptive change in any species on planet earth. The efficient and effective functioning of proteins, genes, and their interactors is largely depends on the action of natural selection. Thus, we can gain great deal of insight not only into how proteins and genes evolve during any time by the study of natural selection in species, but can also identify the genetic changes in species responsible for specific adaptations and mutations by identifying the patterns left by natural selection on any genome of a species . The creation of phylogenetic trees and alignment of homologous bacterial genes, allow us to make inferences about how these genes may have evolved and how they are related to each other. In this paper, we will attempt to determine what kind of selection is acting, and how the sequences changed because of the impact of natural selection on these genes. We care about the action of natural selection because it is important to understand if you are strictly interested in studying gene function. We analyze various sequences of the Pseudomonas syringae HrpZ gene. The P. syringae is a plant pathogenic bacterium that causes disease in many important crops important for food production. For our purposes, we are going to focus on the most common test, dN/dS Ratio Test, to identify positive and negative selection. We perform FEL and SLAC Analysis in this paper.

Keywords: Natural selection, amino acid, codons, proteins, selection analysis, phylogenetics.

1. INTRODUCTION

Heritable genetic variants change in frequency because of their impact on the fitness of the species carrying them. This process is very simply termed as Natural selection. Natural selection can perform its action only when there is heritable genetic variation to act upon [2]. So, variation is the key for natural selection. Some of this variation may be preferable, and therefore may increase in frequency, while other variation may not be preferable, and may therefore decrease in frequency. Preferable should not be over-interpreted. As an example, it may simply mean that a specific variant of a protein performs functions more effectively in a particular environment [3]. In a different environment, that same variant may in fact function less effectively. So, this means, natural selection is not only dependent upon the presence of heritable genetic variation in a species, but also upon the particular environment, where that variation is found. Natural selection should also not be thought of as some kind of acting entity [4]. It is, very simply, the differential survival and transmission of genetic variants to following generations.

2. THE MAJOR TYPES OF NATURAL SELECTION

Natural selection can act in many different manners. In this paper we consider the simplest and most relevant ones that are relevant to our analysis. These are listed below:

A. Positive selection

Positive selection occurs when a favorable mutation occurs in a population of a species and that mutation increases in frequency. Of course, other variation will be concurrently decreasing in their frequency, if this mutation is increasing in frequency [5]. A classic example of a positive selection is the occurrence of an antibiotic resistance allele in a bacterial population which is being exposed to that antibiotic.

B. Negative (purifying) selection

The other type of selection is Negative (purifying) selection. It is the opposite of positive selection. A negative selection occurs when a detrimental or unfavorable mutation is selected out of a population. Most mutations in genome that cause

a change in a protein coding sequence are believed to be detrimental since most proteins have undergone millions or billions of years of evolution. These harmful mutations are to be removed from the population by negative selection [5]. The fraction of mutations for any protein that is detrimental is directly related to the evolutionary conservation of that protein in a species.

C. Balancing or diversifying selection

Balancing or diversifying selection is the selection which favors the maintenance of genetic variation at a locus in a species. While both positive and negative selection purge variation either selecting for or against variants, balancing or diversifying selection actually maintains variation in a species by selecting for multiple genetic variants [5]. The simplest way to imagine this happening is to consider the case of different environments that select for multiple allelic forms of a protein. For example, a receptor protein which is strongly beneficial when a pathogen is present, but does nothing but put a load on the system when there is no pathogen present in a population.

To detect and measure natural selection we use characteristic marks or footprints on the genome. Fortunately, all evolutionary processes including mutation, natural selection, recombination, gene flow, and genetic drift leave their footprints or characteristic marks on the genome. Some of these marks and footprints are obvious and long lasting, while others are transient and / or very obscure. If we know what to look for, and where to look then you we very often reconstruct the evolutionary history of a genetic region in a species. There are many approaches for characterizing and identifying the footprints and marks left behind by natural selection [6]. For our purposes, we are going to focus on the most common test to identify positive and negative selection. As already mentioned, positive selection is selection for a genetic change that increases fitness and benefits the organism, while negative selection is typically associated with evolutionary conservation to preserve an essential function by a protein.

3. DN/DS RATIO TEST

In this analysis we use dN/dS Ratio Test. It is perhaps the most widely used method for detecting the pattern of natural selection in a genome from nucleotide sequence data. This method is particularly useful because it is able to infer natural selection acting on a genome all the way down to the level of the nucleotide or codon.

Using dN and dS to Infer Selection

The dN/dS test, also known at the Ka/Ks or ω test, calculates the ratio of the rate of no synonymous substitutions (dN stands for the number of non-synonymous substitutions per no synonymous site) to the rate of synonymous substitutions (dS stands for the number of synonymous substitutions per synonymous site) [7]. Non-synonymous substitutions are those mutations that result in a change in the protein sequence, while synonymous substitutions are those that change the DNA sequence, but not the protein sequence due to the degeneracy of the genetic code. Note that we are interested in the rate of these substitutions, not their absolute number.

Synonymous substitutions are generally not exposed to strong selective pressures since they don't result in a change to the protein sequence; therefore, they tend to accumulate at roughly a constant rate. We can assume this rate as the baseline by which we will compare the rate of substitutions that change the protein sequence i.e. non-synonymous substitutions. In the case of a completely neutral sequence, one that is free to change with no constraints, we would expect dN to be the same as dS, or the ratio dN/dS = 1.

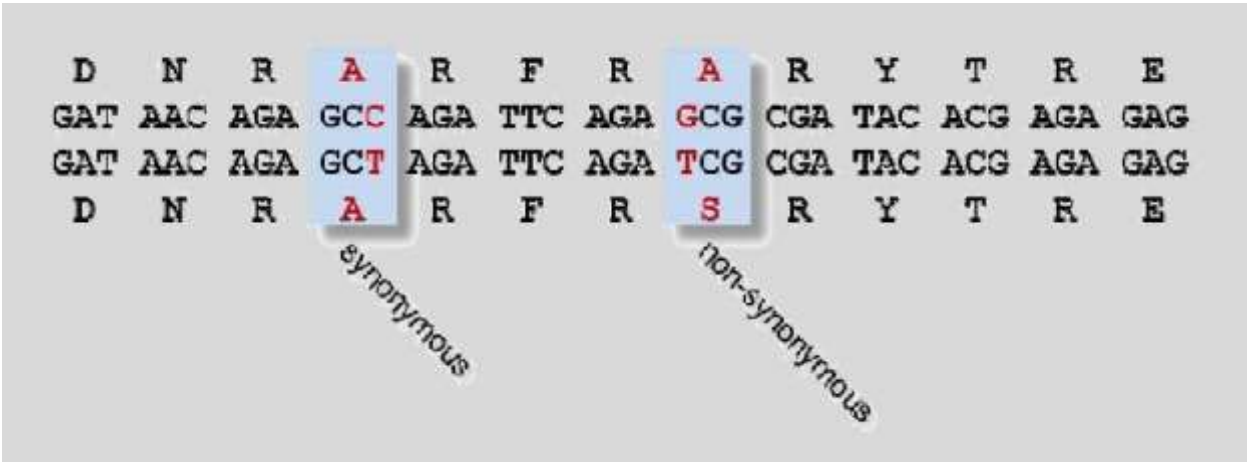


Fig 1 Synonymous and non-synonymous substitutions

When there are selective constraints on a sequence, in case of negative selection, we would expect fewer substitutions that change the protein sequence, or a lower dN; therefore, dN/dS < 1 [8].

In the case of positive selection, we would expect to see a higher proportion of amino acid substitutions in our population because they are being increased by positive selection, so a higher dN; therefore, dN/dS > 1.

We can determine if a gene is under positive or negative selection by measuring the dN/dS ratio. dN/dS > 1 is a strong indicator of positive selection. dN/dS < 1 is a strong indicator of negative selection. A neutral sequence should, in theory, have a dN/dS = 1. The easiest way to think about this is that the rate of synonymous or silent substitutions should remain constant since it is not exposed to selection, but the rate of non-synonymous substations can go up or down depending on whether there is selection to change the amino acid sequence or keep it the same, respectively.

4. METHODOLOGY

We use the online tool Data monkey at www.datamonkey.org to look for selection in our sequences [1]. Data monkey is a powerful and very straightforward tool that provides access to complex and sophisticated evolutionary analyses of sequences. These analyses are strictly carried out on a remote server so analysts don't need access to a high-end workstation. For analysis, Data monkey uses a HyPhy package, which is a very powerful multiplatform package to carry out likelihood analyses of patterns and rates of sequence evolution.

We analyzed 51 sequences of the Pseudomonas syringae HrpZ gene in a data file: Lab5_Psy_hrpZ.fas. The HrpZ gene encodes part of the type III secretion system. This secretion system is used by the bacterium to inject harmful proteins into the plant host cell in order to suppress the plant immune response. To do an initial analysis we uploaded the file to Data monkey analysis site www.datamonkey.org/dataupload.php. and examined our data. Data monkey reports the number of sequences, columns (codon sites), partitions and an amino-acid translation of the alignment in multiple formats.

SLAC Analysis Setup

Data monkey assigns a Job ID to each successfully uploaded file. These IDs can be used to track all the analyses performed on the alignment. Data monkey provides a range of different analyses. In this analysis we focused on the two most straightforward analysis, SLAC and FEL. We selected SLAC (single-likelihood ancestor counting) Method for identifying selection acting on each codon position in a protein sequence [9].

We ran the model selection procedure on the data file and verified that the TrN model is the best fitting model for our data. After the model selection analysis is finished, we returned to the analysis setup page and selected TrN93 for the nucleotide substitution model. The rest of the parameters were left to the default values. After we submit the job Data monkey would display a page with more detailed results. With our results page we analyze the average dN/dS ratio for this sequence and based on this, we describe how selection is acting on this sequence. We performed selection analysis to determine whether or not a set of sequences were under positive selection, balancing selection, or purifying selection.

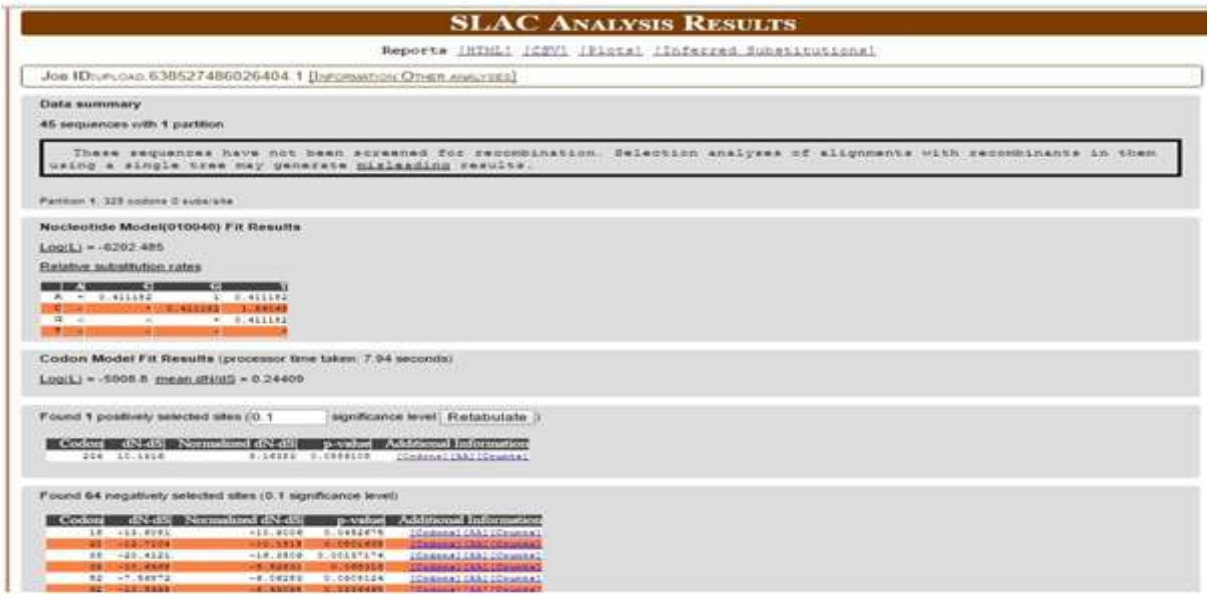


Figure 2 SLAC analysis results page

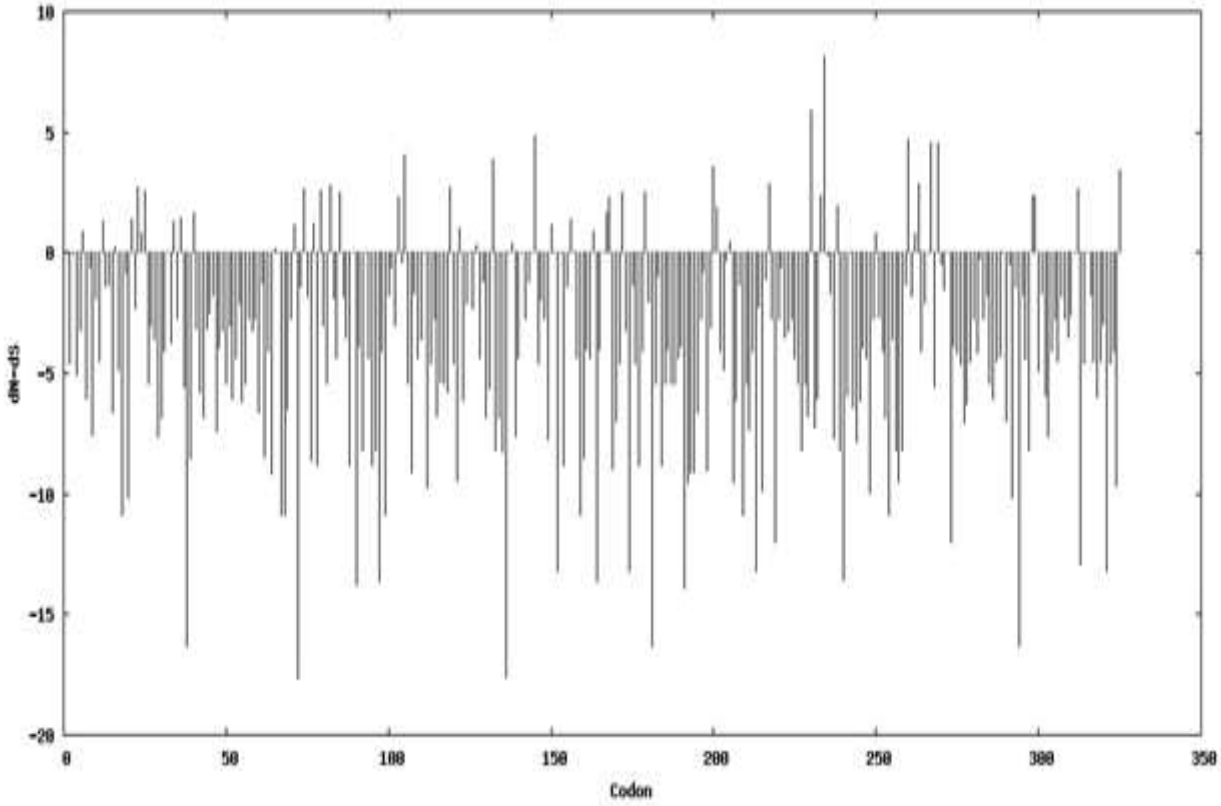


Figure 3 dN-dS plot

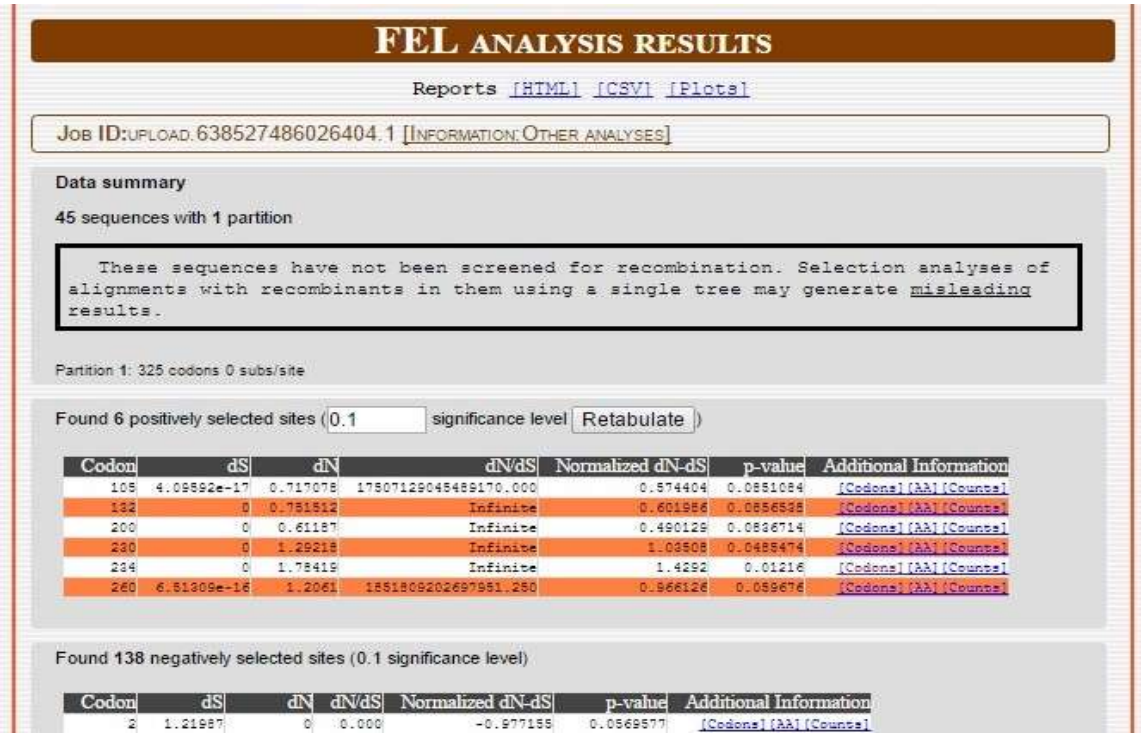
The average dN/dS ratio for these sequences is .24. We would say that the selection acting on the sequence in general is purifying. There's just one amino acid 234, which is under positive selection, and there are 64 amino acids under negative selection with a 0.1 significance level. In case, we changed the significance level to 0.5, we got more amino acids being under positive selection and more amino acids being under negative selection. In the case of positively selected sites, we get 16 at p=.5 and 216 negatively selected sites at p=.5. The amino acid under the strongest positive selection is amino acid 234 with normalized dN-dS score of 8.16, and the amino acid, which is at position 72 is under the strongest negative selection, with a normalized dN-dS score of minus 17.69.

In our analysis, Data monkey presents all positively (dN/dS>1) and negatively (dN/dS <1) selected codons. We have to recall that dN is the rate of no synonymous (amino acid changing) substitutions, while dS is the rate of synonymous substitutions; therefore, a significantly positive dN-dS value indicates positive selection, while a significantly negative value indicates negative selection. We generate a plot from the Reports section and start with the default plot with the parameters shown in Fig. 3. The resulting plot presents the difference between dN and dS (along y-axis) for each codon along the sequence (along x-axis).

In general you would think that a dN-dS score of greater than zero means that residue is under positive selection. A dN-dS score of less than 0 is under purifying, or negative selection. Whereby, a dN-dS score equal to 0 would be a neutrally evolving site. We call something as significant depends on the P value for the particular amino acid residue. Basically at P=.1, only the one site is above the significance threshold and that's the 234 where as at the lower significant threshold to P=.5, we get more sites being above the cut off. Similarly, more sites being below the negative cut off for negative selection.

FEL Analysis

FEL (fixed effects likelihood) is a bit more powerful than SLAC, but is slower since it is an order of magnitude more computationally expensive [10]. We run the FEL analysis with default options using the same nucleotide substitution bias model, TrN93, as for the SLAC analysis.



FEL ANALYSIS RESULTS

Reports [HTML] [CSV] [Plots]

JOB ID:UPLOAD.638527486026404.1 [INFORMATION: OTHER ANALYSES]

Data summary

45 sequences with 1 partition

These sequences have not been screened for recombination. Selection analyses of alignments with recombinants in them using a single tree may generate misleading results.

Partition 1: 325 codons 0 subs/site

Found 6 positively selected sites (0.1 significance level Retabulate)

Codon	dS	dN	dN/dS	Normalized dN-dS	p-value	Additional Information
105	4.09592e-17	0.717078	17507129045489170.000	0.574404	0.0851084	[Codons] [AA] [Counts]
132	0	0.751512	Infinite	0.601986	0.0856535	[Codons] [AA] [Counts]
200	0	0.61187	Infinite	0.490129	0.0836714	[Codons] [AA] [Counts]
230	0	1.29218	Infinite	1.03508	0.0489474	[Codons] [AA] [Counts]
234	0	1.78419	Infinite	1.4292	0.01216	[Codons] [AA] [Counts]
260	6.51309e-16	1.2061	1851809202697951.250	0.966126	0.059676	[Codons] [AA] [Counts]

Found 138 negatively selected sites (0.1 significance level)

Codon	dS	dN	dN/dS	Normalized dN-dS	p-value	Additional Information
2	1.21987	0	0.000	-0.977155	0.0569577	[Codons] [AA] [Counts]

Figure 1 FEL Analysis Results

We generated a dN/dS Plot by exporting the Data monkey data and performing some simple analysis in MS Excel. The file has the following columns: Codon, dS, dN, dN/dS, Normalized dN-dS, dS (when dN=dS), Log(L), LRT, p-value. We sorted the data by dS by selecting all of the columns.

We observed about 50 rows with extremely small dS values. These extremely small values would result in nonsensical dN/dS values. To avoid this we simply deleted the dN/dS values for any row with these extremely low dS values and resort the data by codon position. We selected the dN/dS column and ploted it by inserting a column chart. Then we marked the two sets of values (position and dN/dS) by highlighting the first set and the second set. The Figure 6 shows the resulted plot.

The positively selected residues have dN/dS > 1, while negatively selected residues have dN/dS<1.

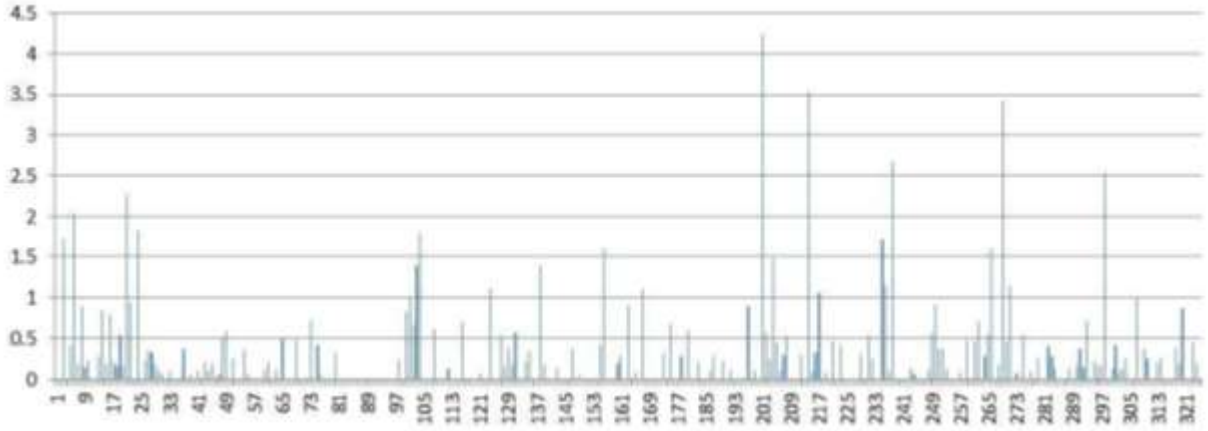


Figure 5 dN/dS Plot

CONCLUSION

By the end of this analysis we can conclude that there are more sites under positive selection in, with the default level of significance and we note that there are more sites under negative selection with the default level selected at significance of 0.1. This is because the FEL analysis is computationally more intensive and more sensitive than the SLAC analysis. Any major differences between SLAC and EFL plots are because we've actually filtered out the cases where our dN/dS score would be infinite and those ones actually happen to be it's significant residues so in this case

We don't see the amino acid at 234 showing up. The value for that peak would be infinite, it can't be plotted because the dS value would be 0. We do see a couple of other residues here above a ratio of one. However, when you look at the P values for them, these aren't actually significant. So, there are differences but certainly in general the trend is that there are far fewer apparently positively selected sites. And there are those under negative selection with a value of close to zero. Knowing which sites are under positive selection may be useful in designing vaccines or drugs that can act at that site in any case when a pathogen tries to evade the host's defenses.

5. ACKNOWLEDGMENT

The author would like to thank Mrs. Monika Saini Department of Computer Science, Guru Jambheshwar University, Hisar for her valuable help in the experiment. The author is also thankful to the staff members of Computer Science department at S.D. College (Lahore) Ambala Cantt. for their valuable guidance.

REFERENCES

[1]. http://www.datamonkey.org/
[2]. "The Influence of Selective Pressure on the Observed Frequency of Synonymous and Nonsynonymous Mutations" in Chapter 7 "Recovering Evolutionary History" in Understanding Bioinformatics by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 240-241.
[3]. Misawa K, Tajima F (1997) Estimation of the Amount of DNA Polymorphism When the Neutral Mutation Rate Varies Among Sites. Genetics 1997 147: 1959-1964..
[4]. Pond SL, Frost SD, and Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676-679.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
[5]. "Recovering Evolutionary History" in Understanding Bioinformatics by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 223-264.
[6]. "Building Phylogenetic Trees" in Understanding Bioinformatics by Marketa Zvelebil and Jeremy Baum, Garland Science, 2008. pp 267-311.
[7]. K Tamura, J Dudley, M Nei, S Kumar (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24(8):1596-9.
[8]. WF Doolittle (1999) Phylogenetic classification and the universal tree. Science 284: 2124-2128. RDM Page and MA Charleston (1997) From gene to organismal phylogeny: reconcidle trees and the gene / species tree problem. Mol. Phylogenet. Evol. 7:231-240.
[9]. N Saitou and M Nei (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.
[10]. S Guidon and O Gascuel (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696-704.