

Analysis of Evolutionary Pressure and Pathogenicity of Missense Variations

Master's thesis

Imrul Faisal

Institute of Biomedical Technology (IBT)

**Institute of Biosciences and Medical Technology
(BioMediTech)**

University of Tampere, Finland

September 2012

DEDICATION

This work is dedicated to my father **Late Mozammel Haque.**

ACKNOWLEDGEMENT

I am grateful to almighty Allah for his mercies and grace. Thank you so much indeed for giving me sound health, and mental strength throughout the entire period of this research.

This thesis work has been carried out at the Bioinformatics research group of Institute of Biomedical Technology (IBT), University of Tampere lead by Professor Mauno Vihinen. I would like to thank Professor Vihinen for giving me the opportunity to work in his group. Your moral support, inspiration and guidelines have helped me a lot to carry out this research.

My special thanks go to Acting Professor Csaba Peter Ortutay for overall support, guidance and inspiration throughout the entire thesis project. Without your strong and efficient supervision, this work might have been very difficult. I am also grateful to Jouni Väliaho for assisting me in software issues. Your guidance and assistance have helped me a lot to overcome each obstacle in programming and database issues.

I am really thankful and grateful to Martti Tolvanen for guiding and inspiring me throughout the entire period of this Master of Science studies. Your profound suggestions and recommendations have helped me to manage everything nicely during this degree study. I can't but say that it might have not been possible to pursue this degree without your support as I could have not been able to join this degree according to the schedule time. My thanks also go to Study Secretary Mira Pihlström for entire support in academic matters. I am grateful to Abhishek Niroula for his excellent support from the beginning of this Master of Science study as my tutor. He continued motivating and inspiring me anytime whenever needed. I am grateful to all the teaching staff of University of Tampere, University of Turku and Tampere University of Technology.

My special thanks go to my wife Sohana Parvin for her endless support, excellent inspiration and limitless cooperation in conducting the entire thesis. Thank you so much indeed. I am happy and really lucky to have you beside me. Finally, I would like to thank my mother Shahnaj Begum and my brother Kamrul Faisal for their overall support.

Imrul Faisal.

MASTER'S THESIS

Place	UNIVERSITY OF TAMPERE Institute of Biomedical Technology (IBT)
Author	IMRUL FAISAL
Title	Analysis of evolutionary pressure and pathogenicity of missense variations
Pages	53
Supervisors	Acting Professor Csaba Ortutay; Professor Mauno Vihinen
Reviewers	Acting Professor Csaba Ortutay; Professor Mauno Vihinen
Time	September 2012

Abstract

Background and aims: Gene and protein sequences are subject to variation upon reproduction where favorable alleles are expressed in further generations by positive selection and bad alleles are being eliminated by means of purifying selection. Thus, selective pressures acting on sequences can result positive or negative selection upon evolution. K_a/K_s ratio is the ratio of non-synonymous and synonymous substitution rate. Calculation of K_a/K_s ratio returns selective evolutionary pressure acting on a particular gene. However, codon wise site specific K_a/K_s ratios calculated for each amino acid differ from the overall selective pressure active on the respective gene. This study was aimed to calculate site specific evolutionary pressure of specific human missense variations of both neutral and pathogenic type and to retrieve hidden facts underlying in it. In addition, analysis of pathogenicity of variants was also another major aspect in this research.

Methods: Variation data were obtained from *VariBench* database. Gene and respective protein sequences reported in the dataset were downloaded along with their orthologs to calculate site specific evolutionary pressures. After that, K_a/K_s values were calculated by *Selecton* software for each amino acid. Finally, statistical analyses were performed with *R* to examine selective pressure distribution separately for pathogenic and neutral variations. In addition, pathogenicity was also calculated for individual amino acids and for different amino acid groups.

Results: Overall distribution of K_a/K_s values was found to be lower for pathogenic variations than neutral ones. Means, medians, and quartiles were also lower for pathogenic variations for each amino acid and for different amino acid groups. Cysteine, glycine, and tryptophan are among the amino acids which were most likely substituted into pathogenic variants whereas threonine, asparagine and isoleucine were more frequent to result neutral variants.

Conclusion: Overall K_a/K_s distribution was lower for pathogenic variants. Neutral variants were less conserved in comparison with the pathogenic type. Moreover, aromatic amino acids were most likely to be converted into pathogenic variations and hence can be marked as having highest pathogenicity. In contrast, negatively charged acidic amino acids showed least pathogenicity.

CONTENTS

1. Introduction	1
2. Review of literature	3
2.1 Amino acids	3
2.1.1 Classification of amino acids	3
2.2 Natural selection	4
2.2.1 Fitness	6
2.2.2 Positive selection	6
2.2.3 Negative selection	7
2.3 K_a/K_s ratio	8
2.4 Methods used to calculate K_a/K_s ratio	9
2.4.1 Method classes	9
2.4.1.1 Approximate methods	9
2.4.1.2 Maximum-likelihood methods	10
2.4.2 Substitution models	10
2.4.3 Selected Method	12
2.5 Statistical aspects	12
2.5.1 Non-parametric statistics	12
2.5.2 Goodness of fit	13
2.5.3 Kolmogorov-Smirnov test	13
2.6 Pathogenicity prediction	14
2.6.1 Pathogenicity prediction methods	15
2.6.2 Ongoing project	15
2.6.3 PON-P	16
3. Objectives	18
4 Materials and methods	19
4.1 Materials	19
4.1.1 Variation dataset	19
4.1.2 Sequences	20
4.1.3 Tool for calculating K_a/K_s ratio	21
4.1.4 Software for statistical analysis	21
4.2 Methods	21
4.2.1 Preparation of datasets	21

4.2.2 Building pipeline	22
4.2.3 Calculation of site-specific K_a/K_s values	22
4.2.4 Statistical analysis	24
5. Results	25
5.1 Ka/Ks value	25
5.1.1 Overall comparison of K_a/K_s values	25
5.1.1.1 Two-sample Kolmogorov-Smirnov test	25
5.1.2 K_a/K_s value comparisons for individual amino acids	26
5.1.2.1 Two-sample Kolmogorov-Smirnov test	26
5.1.3 K_a/K_s value comparisons according to amino acid groups	28
5.1.3.1 Two-sample Kolmogorov-Smirnov test	29
5.2 Pathogenicity comparisons	31
5.2.1 Comparison of individual amino acid variability	31
5.2.2 Analysis of pathogenicity for individual amino acids	33
5.2.3 Group wise comparison of amino acid variability	33
5.2.4 Analysis of pathogenicity for different amino acid groups	35
6. Discussion	37
6.1 K_a/K_s value aspects	37
6.1.1 Pathogenic and neutral variation data sets	37
6.1.2 K_a/K_s results	37
6.2 Amino acid variability aspects	40
6.2.1 Variability of individual amino acids	40
6.2.2 Variability of different amino acid groups	42
6.3 Future perspectives	42
6.3.1 PON-P	43
7. Conclusion	44
8. References	45
9. Appendix	53

ABBREVIATIONS

AA	Amino acid
BLAST	Basic Local Alignment Search Tool
BTK	Bruton tyrosine kinase
CAGI	Critical Assessment of Genome Interpretation
CDF	Cumulative distribution function
CNIO	Centro Nacional de Investigaciones Oncológicas (Spanish National Cancer Research Centre)
cDNA	Complementary deoxyribonucleic acid
DNA	Deoxyribonucleic acid
IBT	Institute of Biomedical Technology
KS	Kolmogorov-Smirnov
ML	Maximum-Likelihood
MSA	Multiple Sequence Alignment
NGS	Next-generation sequencing
PON-P	Pathogenic-or-Not-Pipeline
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
nsSNP	Nonsynonymous SNP

1. INTRODUCTION

In genomic sequence of any species, mutation or variation is the change of nucleotide (A, T, C, or G) in DNA sequence. Variations are also occurring in RNA sequence of some lower organisms, for example viruses, which do not contain DNA in their genomic sequence. At the DNA replication stage of meiosis cell division, these variations appear for different factors and replication errors. Viruses are considered as another cause of variations. Sometimes these variations do not have any bad effect; however, very often these can be very harmful. Gene products might alter, or become inactive due to variations in protein coding region of the gene. When variations are harmful, they can be pathogenic. On the other hand, neutral variations do not show any harmful effect (Bertram, 2000; Burrus and Waldor, 2004; Aminetzach, Macpherson, and Petrov, 2005; Sawyer *et al.*, 2007).

When the variation occurs in a single nucleotide position of DNA, it is generally termed as single nucleotide polymorphism (SNP). As genes are having coding and non-coding sequences, SNPs can be found in either of the regions. In addition, inter-regions of genes might also contain SNPs. These single variations are found more frequent in non-coding regions of genes. Some SNPs change the respective proteins; however, due to degeneracy of the genetic code, some SNPs do not change the protein (Barreiro *et al.*, 2008; Stenson *et al.*, 2009; Varela and Amos, 2010). Synonymous variations do not change the amino acid. However, if the variation changes the respective amino acid, it is termed as non-synonymous variation. A non-synonymous variation can affect protein function in numerous ways. For example, it could occur at a critical site in a protein such as at a catalytic site or in a ligand interaction surface, or it may affect protein structural properties, leading to improper folding, structural instability, or protein aggregation (Thusberg and Vihinen, 2009; Olatubosun *et al.*, 2012).

Calculating non-synonymous (K_a , or d_N) and synonymous (K_s , or d_S) substitution rates is of great significance in reconstructing phylogeny and understanding evolutionary dynamics of protein-coding sequences across closely related and yet diverged species (Zhang *et al.*, 2006). Evolutionary pressures on genes or proteins are often quantified by the ratio of substitution rates at non-synonymous and synonymous sites (Kryazhimskiy and Plotkin, 2008).

By the process of natural selection, beneficial alleles are reproduced in subsequent generation whereas bad or harmful alleles are being eliminated naturally. Thus, natural selection of genes result positive or negative selection, or remain neutral. Gene variants having higher fitness reproduce nicely in further generation. So, evolutionary pressures are active on individual genes and they are subject to positive, negative or neutral evolution (Bell, 1997; Hughes, 1999; Ridley, 2004; Forsdyke 2006).

K_a/K_s ratio (alternatively known as ω , or d_N/d_S) estimates the evolutionary pressure acting on a gene where K_a is the rate of non-synonymous substitutions. K_s represents another rate in the same way for synonymous changes in synonymous sites (Miyata and Yasunaga, 1980). In general, K_a/K_s ratio is a measure of selective pressure on a protein and it differentiates codon-based analyses from the more general tests of neutrality proposed in population genetics (Kreitman and Akashi 1995; Wayne and Simonsen 1998). Analyzing evolutionary pressure of sequences at the codon level, as opposed to the amino acid level, enables detecting positive, neutral or purified selection sites for each codon. Thus, by contrasting silent (synonymous) substitutions against amino acid altering (non-synonymous) substitutions, it is possible to detect the different selection forces operating on each amino acid site (Doron-Faigenboim *et al.*, 2005; Glaser *et al.*, 2003; Gu and Vander Velden, 2002). Therefore, site specific K_a/K_s ratio for each amino acid position differs from the overall K_a/K_s value of the entire gene or protein.

The aim of this thesis work was to estimate codonwise K_a/K_s ratio for each amino acid of large set of human proteins. This calculation retrieves selection pressure for thousands of variations of both neutral and pathogenic types which were collected from VariBench database containing hundreds of genes and proteins (Nair and Vihinen, 2012). Statistical analysis of the selection pressure is important to understand key facts of site variation. For instance, it is significant to analyze whether disease causing variations are more frequent in positive selection pressure sites (K_a/K_s value is higher) than in conserved sites (K_a/K_s value is lower). The same query applies to variations of neutral types. Therefore, it was an objective in this study to examine whether pathogenicity of novel variants is predictable or not based on evolutionary pressure acting on it. Additionally, this thesis work has real interest in analyzing pathogenicity of different amino acids.

2 REVIEW OF LITERATURE

2.1. Amino acids

Amino acids are building blocks of proteins. Each amino acid is composed of an amine (-NH_2), a carboxylic acid group (-COOH) and a side chain which is unique for each amino acid. C (carbon), H (hydrogen), O (oxygen), and N (nitrogen) are vital elements of an amino acid. Although amino acids are fundamental macromolecules of protein formation, they also play vital role in many physiological activities. For instance, glutamic acid is considered as an excitatory neurotransmitter which is facilitating conduction of nerve impulse. Some amino acids are synthesized by human cells and are referred as non-essential amino acids. However, other amino acids are not synthesized by cells and hence must be supplied by means of nutrition etc. (Kyte and Doolittle, 1982; Pamela *et al.*, 2004; Nelson and Cox, 2005; Ambrogelly *et al.*, 2007).

2.1.1 Classification of amino acids

Functional and structural properties of amino acids vary a lot depending on the side chains. Some amino acids show polarity whereas others are non-polar. Some of them contain aliphatic chains and some have aromatic ring. Moreover, some amino acids show acidic properties, some are basic and the rests show neutral property. Short overview of amino acid classification depending on their properties is given below.

- A, C, G, I, L, M, F, P, W, and V are belonging to non-polar amino acids.
- R, N, D, E, Q, H, K, S, T, and Y amino acids show polar properties.
- R, H and K are positively charged (basic) amino acids.
- D and E are acidic amino acids containing negative charge.
- A, N, C, Q, G, I, L, M, F, P, S, T, W, Y, and V are neutral (uncharged) amino acids.
- F, Y, and W are having aromatic ring in their structure

(Kyte and Doolittle, 1982; Pamela *et al.*, 2004; Nelson and Cox, 2005; Ambrogelly *et al.*, 2007; Meierhenrich, 2008). Some of them show mixed properties and hence might be considered member of more than one group. Amino acids were divided in five groups in this thesis work in order to compare pathogenicity of different groups. These five groups were acidic (polar negatively charged), basic (polar positively charged), neutral (no charge), aliphatic (non-polar) and amino acids containing aromatic ring. Phenylalanine (F), tyrosine (Y) and tryptophan (W) are the amino acids which contain aromatic ring in their structure. Only two amino acids are acidic. They are aspartic acid (D) and glutamic acid (E). In addition, a few of them contain positive charge and marked as basic. These are lysine (L), arginine (R) and histidine (H). S, T, C, Q, and N do not contain any charge (uncharged) in their structure.

2.2 Natural selection

Selection is nonrandom differential survival or reproduction of classes of phenotypically different entities. At the molecular level, selection occurs when a particular DNA variant becomes more common because of its effect on the organisms that carry it. Individuals of a population vary to each other and their offsprings get variation in genome. Thus, variations are occurring by nature. By the process of this variation, certain biological traits are found more or less frequent in further generations (Darwin, 1872; Hartl, 1981; Maynard-Smith, 1989; Ridley, 2004; Schaffner and Sabeti, 2008).

The association of a phenotype with change in frequency, separated from other forces that change phenotype, is one abstract way to describe natural selection (Frank, 2012). Natural selection is one of the evolutionary mechanisms, in which relative frequencies of genotypes change according to their relative fitness in the population (Figure 2.1).

There are several outside environmental forces which are acting on genes. These forces affect phenotypic characteristics of individuals. Darwin (1872) had described these environmental forces as selection pressure.

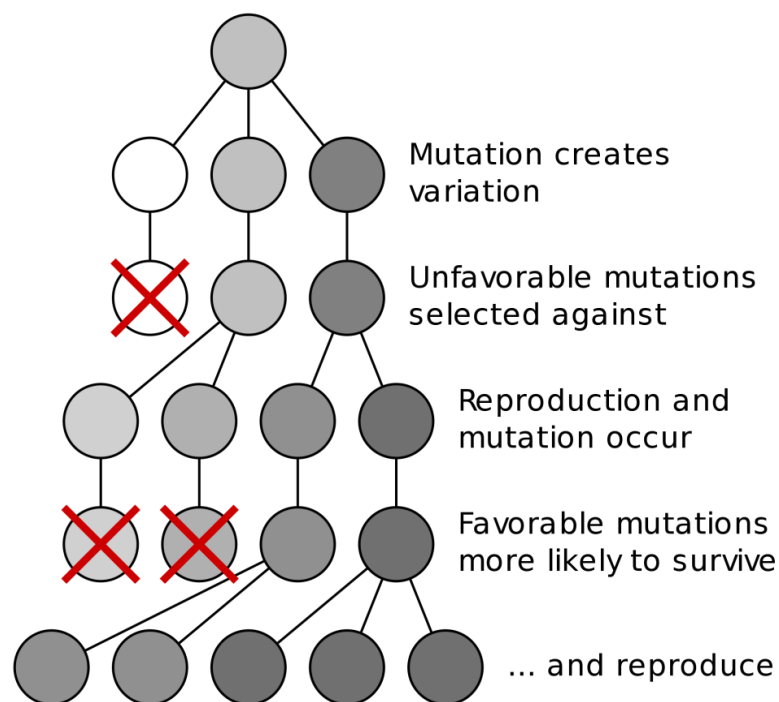


Figure 2.1: Process of natural selection by which favorable variations survive in next generations. Thus, alleles having lower fitness are eliminated gradually by means of purified selection (image downloaded from <http://freethinkerperspective.blogspot.fi/2012/07/how-natural-selection-selects.html>).

“Darwin’s process of natural selection has four components (Evolution notes, 2010).

1. *Variation:* Organisms (within populations) exhibit individual variation in appearance and behavior. These variations may involve body size, hair color, facial markings, voice properties, or number of offspring. On the other hand, some traits show little to no variation among individuals, for example, number of eyes in vertebrates.
2. *Inheritance:* Some traits are consistently passed on from parent to offspring. Such traits are heritable, whereas other traits are strongly influenced by environmental conditions and show weak heritability.
3. *High rate of population growth:* Most populations have more offspring each year than local resources can support leading to a struggle for resources. Each generation experiences substantial mortality.
4. *Differential survival and reproduction:* Individuals possessing traits well suited for the struggle for local resources will contribute more offspring to the next generation”.

2.2.1 Fitness

Fitness is the core idea of natural selection in any population. In a given environment, fitness is described as both genotypic and phenotypic characteristics of individuals of certain population. Upon reproduction, it is the average contribution of an allele or genotype to the next generation or to succeeding generations. Therefore, fitness can be described as the success of an entity in reproducing; hence, the average contribution of an allele or genotype to the next generation or to succeeding generations (Darwin, 1872; Hartl, 1981; Maynard-Smith, 1989; Charlesworth *et al.*, 1995; Burch and Chao, 1999; Loewe, 2008).

2.2.2 Positive selection

The natural selection can be divided into positive (or Darwinian) and negative (or purifying) selections (Suzuki and Gojobori, 1999). Shortly, positive selection is the selection for an allele that increases fitness. Positive selection refers to the type of natural selection that promotes the spread of beneficial alleles (Page and Holmes, 1998; Zhang, 2008; Loewe, 2008).

Individuals of same species or different species of a certain environment compete to others for resources. This competition might be described as an interaction between individuals of the same species or different species whereby resources used by one are made unavailable to others. Genotype frequency of the winner of this competition increases and it might be indicated as positively selected (Charlesworth *et al.*, 1995; Burch and Chao, 1999; Loewe, 2008). As advantageous alleles that are under positive selection increase in prevalence, these alleles leave distinctive signatures, or patterns of genetic variation, in the DNA sequence. For a trait to undergo positive selection, it must have two characteristics. (1) The trait must be beneficial; in other words, it must increase the organism's probability of surviving and reproducing. (2) The trait must be heritable so that it can be passed to an organism's offspring (Schaffner and Sabeti, 2008; Byrk *et al.*, 2008; Bersaglieri *et al.*, 2004).

Certain alleles are having higher fitness value than others. Thus, variations on specific alleles may lead positive or negative selection into next generations. Positive selection promotes the spread of beneficial alleles. For instance, if functionality of certain enzyme

increases upon variation, it will increase fitness of the respective gene. It will be subject to positive selection. This variation is considered as beneficial for life. The outside forces of the environment responsible for the positive selection are termed as positive selection pressure (Bell, 1997; Hughes, 1999; Ridley, 2004; Forsdyke 2006).

Positive selection promotes the emergence of new phenotypes. Of the many phenotypic traits that define our species are the enormous brain, advanced cognitive abilities, complex vocal organs, bipedalism and opposable thumbs. Most of them are likely the product of strong positive selection. Positive selection can leave a set of telltale signatures in the genes under its influence. For instance, the rapid divergence of functional sites between species and the depression of polymorphism within species might be given as example (Eric and Bruce, 2004; Bamshad and Wooling, 2003; Kreitman, 2000; Yang and Belawski, 2000). This study was aimed to study evolutionary pressures (positive and negative) at molecular level only.

2.2.3 Negative selection

Some DNA variations are beneficial and some variations are harmful for individuals. Deleterious variations are being eliminated by means of negative selection as they are considered harmful. Genotype proportions of those individuals having lower fitness are decreased in their offsprings in further generations. As this selection is very important for higher stability of biological structures, it is often referred as purifying selection. Thus, negative selection or purifying selection refers to the type of natural selection that prohibits the spread of deleterious alleles (Page and Holmes, 1998; Zhang, 2008; Loewe, 2008).

Purifying selection prevents deleterious variation so that it cannot take over a population. Any improved structure which is once fixed in a population is maintained as long as it is needed. Frequently, ecological circumstances also play a role in determining mutational effects. For instance, if the niche of a species stays the same, some mutations that would be beneficial in other niches will be under negative selection (Loewe, 2008).

Less adaptive variants are subject to extinction by the role of negative selection. Moreover, if some variants are considered best-adapted and do not change to maintain a stable local optima, the role of negative selection would be to eliminate all new variants

for that optimum trait gradually. The outside forces of the environment involved in the negative selection process are collectively known as negative selection pressure (Charlesworth *et al.*, 1995; Bell, 1997; Burch and Chao, 1999; Hughes, 1999; Ridley, 2004; Forsdyke 2006; Loewe, 2008).

2.3 K_a/K_s ratio

K_a/K_s ratio is the ratio of non-synonymous and synonymous substitution rate. Certain amino acids are subject to change upon variation. Therefore, synonymous (S) variations result same amino acid (unchanged) at protein production level whereas non-synonymous (N) variation will change the amino acid. Selective evolutionary pressure acting on a particular gene depends on the rate at which the sequences are changed. As variations are unequal, sequences of certain genes are usually under pressure to change, drift randomly or to remain almost neutral. Selective forces operating at the amino acid sequence level have been detected mainly by comparing the number of non-synonymous substitutions per site with that of synonymous substitutions per site (Hughes and Nei 1988; Endo *et al.*, 1996; Tsunoyama and Gojobori 1998). Usually, K_a (or, d_N) value is calculated from the rate of non-synonymous changes from the number of potential non-synonymous sites. Calculation in the same way results K_s (or, d_S) value for synonymous sites (Miyata and Yasunaga, 1980; Ina, 1994; Comeron, 1995; Yang and Nielsen, 2000).

Calculation of K_a/K_s ratio (or ω , d_N/d_S) is an optimum way to estimate evolutionary pressure acting on a gene. If the K_a/K_s ratio is unity ($K_a/K_s = 1$), the amino acid change is considered as neutral. Deviation of this value from one clearly indicates the selective pressure acting on the protein of the corresponding gene. The protein sequence is conserved if the ratio is smaller ($K_a/K_s < 1$). However, a higher ratio ($K_a/K_s > 1$) tells us a positive selective pressure. Statistics of these two variables and their ratio in genes or proteins from different evolutionary lineages provides a powerful tool for quantifying molecular evolution (Zhang *et al.*, 2006). K_a/K_s ratio significantly higher than one is convincing evidence for diversifying selection (Yang and Bielawski, 2000).

K_a/K_s value can be estimated either for an entire gene or for each codon of its protein coding region. Estimation of K_a/K_s value for entire gene requires only a single reference gene. Two classes of methods are available to estimate K_a/K_s ratio of coding region

between two protein coding gene sequences (details are discussed in section 2.4). Between the two sequences, one is our target gene for which we are interested to calculate K_a/K_s ratio. The other gene sequence is used as a reference.

As an example, K_a/K_s ratio for a human gene can be calculated in respect to same gene in chimpanzee. Average K_a/K_s value for all human-chimpanzee ortholog pairs is 0.23 (De Magalhães and Church, 2007). On the other hand, site-specific K_a/K_s value can be calculated for each codon of the entire protein coding region of certain gene. Such calculation of K_a/K_s value requires an alignment of many ortholog coding gene sequences. Fitch *et al.* (1997) used an alignment of multiple protein-coding sequences to reconstruct a phylogenetic tree. Then, for each codon site, they compared the total number of nonsynonymous changes throughout the phylogenetic tree with that of synonymous changes to detect positively selected amino acid sites (Suzuki and Gojobori, 1999). In this thesis work, K_a/K_s value was calculated codon wise for a large set of human genes to analyze evolutionary pressure for both pathogenic and neutral human variations. Thus, sets of ortholog sequences were needed to estimate site-specific K_a/K_s values.

2.4 Methods used to calculate K_a/K_s ratio

2.4.1 Method classes

There are many methods available to calculate K_a/K_s ratio. These methods can be divided in two major classes. These are called approximate methods and maximum likelihood methods (Suzuki and Gojobori, 1999; Yang and Bielawski, 2000; Sergei *et al.*, 2005; Zhang *et al.*, 2006). The major two classes of methods are outlined here.

2.4.1.1 Approximate methods

The approximate methods involve three basic steps: (1) counting the numbers of synonymous and non-synonymous sites, (2) calculating the numbers of synonymous and non-synonymous substitutions, and (3) correcting for multiple substitutions (Zhang *et al.*, 2006). This type of methods have been developed from parsimony (Suzuki and Gojobori 1999) or likelihood-based methods (Nielsen 2002; Nielsen and Huelsenbeck 2002; Suzuki 2004). These methods are very suitable for large data sets because they are fast to

compute. A major drawback of these methods is that these are not suitable for small data sets containing few sequences or containing lower divergence (Sergei and Simon, 2005).

2.4.1.2 Maximum-likelihood methods

This class of methods was nicely illustrated by Nielsen and Yang (1998). The maximum likelihood method integrates evolutionary features (reflected in nucleotide models) into codon-based models and uses the probability theory to finish all the three steps in one go (Yang and Bielawski, 2000; Zhang *et al.*, 2006). This class of methods involves fitting a distribution of substitution rates across sites and then inferring the rate at which individual sites evolve. When this site-by-site inference is based on the maximum likelihood estimates of the rate parameters, this inference is known as empirical Bayes (Nielsen and Yang 1998; Yang *et al.* 2000; Sergei and Simon, 2005).

Maximum likelihood methods are based on explicit models of codon substitution (Yang and Bielawski, 2000). Parameters in the model (i.e. sequence divergence, transition/transversion rate ratio and the K_a/K_s ratio) are estimated from the data by ML, and are used to calculate K_a and K_s according to their definitions (Goldman and Yang, 1994; Muse, 1996; Yang, 2000; Yang and Nielsen, 2000; Yang and Bielawski, 2000).

2.4.2 Substitution models

Substitution models play a significant role in phylogenetics and evolutionary analyses of protein coding sequences by integrating diverse processes of sequence evolution through various assumptions and providing approximations to datasets (Zhang *et al.* 2006). Some models are built based on equal base frequencies with uniform substitution rates. Jukes-Cantor (JC) model (Jukes and Cantor, 1969) might be considered as an example of this type. On the other hand, some models are considered as more advanced, using different substitution rates with unequal nucleotide frequencies. These features might be observed in general time-reversible (GTR) model (Tavare, 1986). Some features of different substitution models (Jukes and Cantor, 1969; Kimura, 1980; Felsenstein, 1981; Kimura 1981; Hasegawa *et al.*, 1985; Tavare, 1986; Tamura and Nei, 1993; Zharkikh, 1994) are presented in table 2.1.

So, individual models have been made upon modifications. It is possible to implement codon-based models in a maximum-likelihood (ML) framework upon combining maximum-likelihood scores obtained from specific candidate or substitution models (Goldman and Yang, 1994; Muse and Gaut, 1994; Zhang *et al.*, 2006).

A general formula of a substitution rate for any sense codon i to j ($i \neq j$) can be expressed as:

$$q_{ij} = 0; \text{ if } i \text{ and } j \text{ differ by more than one difference}$$

$$q_{ij} = k_{xy}\pi_j; \text{ if } i \text{ and } j \text{ differ by a synonymous substitution of } x \text{ for } y$$

$$q_{ij} = \omega k_{xy}\pi_j; \text{ if } i \text{ and } j \text{ differ by a nonsynonymous substitution of } x \text{ for } y$$

where, π_j is the frequency of codon j , ω is the K_a/K_s ratio, k_{xy} is the ratio of r_{xy} to r_{CA} , and $x, y \in \{A, C, G, T\}$ (Goldman and Yang, 1994; in Zhang *et al.*, 2006)

Table 2.1: Substitution models used in phylogenetic and evolutionary analysis

Model	Description	Nucleotide frequency	Substitution rate*
JC	Jukes-Cantor model	Equal	$r_{TC} = r_{AG} = r_{TA} = r_{CG} = r_{TG} = r_{CA}$
F81	Felsenstein's model	Unequal	
K2P	Kimura's two-parameter model	Equal	$r_{TC} = r_{AG} \neq r_{TA} = r_{CG} = r_{TG} = r_{CA}$
HKY	Hasegawa-Kishino-Yano model	Unequal	
TNEF	TN model (equal nucleotide frequencies)	Equal	$r_{TC} \neq r_{AG} \neq r_{TA} = r_{CG} = r_{TG} = r_{CA}$
TN	Tamura-Nei model	Unequal	
K3P	Kimura's three-parameter model	Equal	$r_{TC} = r_{AG} \neq r_{TA} = r_{CG} \neq r_{TG} = r_{CA}$
K3PUF	K3P model (unequal nucleotide frequencies)	Unequal	
TIMEF	Transition model (equal frequencies)	Equal	$r_{TC} \neq r_{AG} \neq r_{TA} = r_{CG} \neq r_{TG} = r_{CA}$
TIM	Transition model Unequal	Unequal	
TVMEF	Transversion model (equal) frequencies	Equal	$r_{TC} = r_{AG} \neq r_{TA} \neq r_{CG} \neq r_{TG} \neq r_{CA}$
TVM	Transversion model	Unequal	
SYM	Symmetrical model	Equal	$r_{TC} \neq r_{AG} \neq r_{TA} \neq r_{CG} \neq r_{TG} \neq r_{CA}$
GTR	General time-reversible model	Unequal	

* r_{ij} indicates the rate of substitution of i for j , where $i, j \in \{A, C, G, T\}$ (Zhang *et al.* 2006).

2.4.3 Selected Method

In this thesis work, Selecton software (Doron-Faigenboim *et al.*, 2005; Stern *et al.*, 2007) was used to calculate K_a/K_s ratio for selected set of human genes. Selecton uses Maximum-likelihood (ML) approach to calculate K_a/K_s ratio. Estimation of parameters such as codon equilibrium frequencies, the transition transversion ratio and the phylogenetic tree branch lengths are incorporated. Codon equilibrium frequencies are calculated from the observed nucleotide frequencies of dataset following the previously reported methods (Yang, 1997; Yang *et al.*, 2000). M7, M8, M8a, M5, and MEC models are introduced in later version of Selecton (Stern *et al.*, 2007). Each of these models assumes different biological phenomena. One of the main advantages of these models is that they enable contrasting different hypotheses, by testing which model better fits the data at hand. A more detailed description of these methods is provided at <http://selecton.bioinfo.tau.ac.il/overview.html>

Branch length optimization was performed using expectation maximization technique (Dempster *et al.*, 1977). As ML approach resulted significant false positive rates, Selecton had replaced the ML method to an empirical Bayesian method (Mayrose *et al.*, 2004) in order to calculate K_a/K_s ratio more accurately. In order to test whether positive selection is operating on a protein, it is custom to perform two steps: (1) Perform a likelihood ratio test between a null model (which doesn't account for sites under positive selection), and an alternative model that does; (2) Predict whether a site is undergoing positive selection using a Bayesian approach (Doron-Faigenboim *et al.*, 2005; Stern *et al.*, 2007).

2.5 Statistical aspects

2.5.1 Non-parametric statistics

Non-parametric statistics covers at least two important aspects. The first fact is that these techniques are not associated with any specific distribution. Alternatively, data belong to this group do not follow any pre-defined probability distributions. The second aspect here is that these models do not have any fixed structure. The properties of these models are changeable depending on complexity of datasets (Hettmansperger and McKean, 1998;

Gibbons and Chakraborti, 2003; Wasserman, 2007; Corder and Foreman, 2009; Bagdonavicius *et al.*, 2011).

2.5.2 Goodness of fit

In a statistical model, goodness of fit illustrates compatibility of a set of observations with the model. It generalizes the association between expected and observed trends under the model of selection. Goodness of fit can be tested in many ways. For instance, normality test of a sample distribution, comparison of distribution between two samples, comparison of distribution of a sample with the reference (specified) distribution etc. all can be given as example of this goodness to fit test (John R.T., 1997; Corder and Foreman, 2009).

Goodness of fit can be expressed by the following formula (Charlie and Tonya, n.d.):

$$X^2 = \sum \frac{(O - E)^2}{\sigma^2}$$

Where, σ^2 is the variance of observation, O is the observed data, and E is the theoretical data.

2.5.3 Kolmogorov-Smirnov test

There are some nonparametric tests available in Statistics. Kolmogorov-Smirnov test is one of them which can be used both for one sample and two samples to compare equality of data distributions. This test is also known as goodness-of-fit test (Boes *et al.*, 1974; DeGroot, 1991; Corder and Foreman, 2009).

For one sample Kolmogorov-Smirnov test, the probability distribution of the sample is compared with the reference. In contrast, probability distribution of two samples is compared to each other in two sample Kolmogorov-Smirnov test. This test measures the difference (D) between empirical distribution function and cumulative distribution function (CDF) between the sample and reference distribution (one sample test) or within the two samples (two sample test). The null distribution assumed for one sample Kolmogorov-Smirnov test is that the sample is taken from the reference distribution.

Similarly, null distribution in case of two sample Kolmogorov-Smirnov test is assumed that both samples are having identical distribution. Depending on the calculated probability value, it is concluded whether we accept or reject the null hypothesis. Similar distribution will result greater p-value than the significance level (95 percent). On the other hand, null distribution can be easily rejected if the calculated p-value is small enough (Eadie *et al.*, 1971; Stephens, 1979; Stuart *et al.*, 1999; Corder and Foreman, 2009).

This thesis work has dealt with two different sets of variations and their respective K_a/K_s values. Two sample Kolmogorov-Smirnov test was performed to check whether there are similarities in distributions of K_a/K_s values between those two samples. In addition, differences of K_a/K_s distribution between pathogenic and neutral variations were tested using the KS test for both individual amino acids and amino acid groups.

2.6 Pathogenicity prediction

Many SNPs are not considered as harmful. However, many of them are involved in phenotypic differences of individuals. Non-synonymous SNPs (nsSNPs) are more interesting from medical viewpoint as they are found in protein coding region of genes and expose phenotypic differences.

Prediction of the possible disease-association of missense variants is a difficult problem because an amino acid substitution can affect the biological function of a gene product in a number of ways (Thusberg and Vihinen, 2009; Thusberg *et al.*, 2011). An amino acid substitution may disrupt sites that are critical in protein function, such as catalytic residues or ligand-binding pockets. A missense variation may as well lead to alterations in the structure, folding, or stability of the protein product, thereby altering or preventing the function of the protein. On the other hand, amino acid substitutions do not necessarily affect protein function. Effects of missense variations are often the most difficult to predict while the consequences of most deletions, insertions, and nonsense mutations are rather self-evident (Thusberg *et al.*, 2011).

2.6.1 Pathogenicity prediction methods

Total quantity of human variants is increasing dramatically by the application of high-throughput sequencing techniques. Gathering experimental knowledge about the possible disease association of variants is laborious, costly and time-consuming. Several computational methods have been developed for the classification of SNPs according to their predicted pathogenicity. Thusberg *et al.* (2011) used and evaluated nine widely used pathogenicity prediction methods with a set of over 40,000 pathogenic and neutral variants. Those methods were MutPred (Li *et al.*, 2009), nsSNPAnalyzer (Bao *et al.*, 2005), Panther (Thomas *et al.*, 2003), PhD-SNP (Capriotti *et al.*, 2006), PolyPhen (Ramensky *et al.*, 2002), PolyPhen2 (Adzhubei *et al.*, 2010), SIFT (Ng and Henikoff, 2001), SNAP (Bromberg and Rost, 2007), and SNPs&GO (Calabrese *et al.*, 2009). They have found SNPs&GO and MutPred as better performing methods in their study.

2.6.2 Ongoing project

Powerful next-generation sequencing (NGS) approaches produce variation information at an ever-increasing rate. Given the size and complexity of the variation data, and the rate of data generation, experimentally characterizing the disease association of each of these variations, or their effect on protein function would be expensive, difficult, time consuming, and in practice impossible. This reflects the need for computational approaches in interpreting the data. The output of computational models can be highly useful for preprocessing and prioritization of variants, and to further guide laboratory and clinical experiments (Thusberg and Vihinen, 2009; Khan and Vihinen, 2010; Olatubosun *et al.*, 2012).

The ability to discriminate between pathogenic and benign variants computationally could significantly aid targeting disease-causing variations by helping in the selection and prioritization of likely candidates from a pool of data (Thusberg *et al.*, 2011). Thus, bioinformatic approaches are needed to identify and predict types of variations. Different aspects of variations are being tested in our research lab to observe which of them are good in distinguishing pathogenic and non-pathogenic (neutral) variants. This thesis study is a part of a longer ongoing project where the objective is to develop tool that will be efficient in analyzing and predicting novel variants.

2.6.3 PON-P

Pathogenicity prediction methods described in the previous section differ in training datasets, training features, and method of prediction. Recent evolutions indicate that these predictions are still suboptimal (Khan and Vihinen, 2010; Thusberg *et al.*, 2011; Olatubosun *et al.*, 2012). All indications point to the need for improvement in prediction performance and better information integration and utilization (Olatubosun *et al.*, 2012). Olatubosun *et al.* (2012) have developed a novel tool called Pathogenic-or-Not-Pipeline (PON-P) to improve the performance of pathogenicity prediction methods.

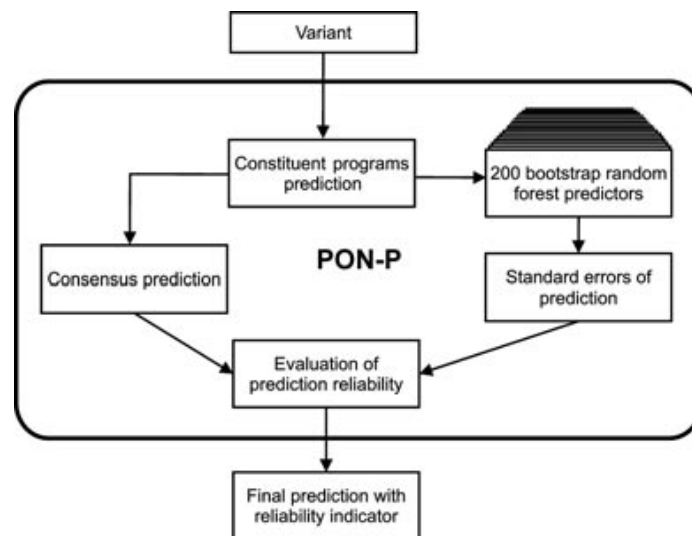


Figure 2.2: Conceptual framework for the determination of prediction reliability. PON-P derives a consensus prediction from five predictors utilizing a trained random forest. It additionally uses 200 random forests trained on bootstrapped dataset to derive the standard error of prediction. These are subsequently combined to evaluate the prediction reliability and derive the final prediction (Olatubosun *et al.*, 2012).

PON-P integrates 5 predictors to predict the probability that non-synonymous variations affect protein function and may consequently be disease related (Figure 2.2). The PON-P is based on a random forest (machine learning method composed of classification and regression trees), composed of 800 trees, using the Random Forest package in R (Liaw and Wiener, 2002). Random forest methodology-based PON-P shows consistently improved performance in cross-validation tests and on independent test sets, providing

ternary classification and statistical reliability estimate of results. PON-P provides high quality prediction of the effect of missense variations, by: (1) aggregating predictions of the five constituent predictors and deducing a consensus prediction, and (2) determining the reliability of the consensus prediction and based on that, classifying the cases as neutral, pathogenic, or unclassified variant (Olatubosun *et al.*, 2012). PON-P has a great interest in this thesis study.

3. OBJECTIVES

The main objective of this thesis work is to evaluate if K_a/K_s value is efficient for distinguishing pathogenic and neutral variants, therefore if it can be used for predicting the pathogenicity of novel variants. Objectives of this research include:

- Calculating K_a/K_s values of each amino acid position for a large set of human protein sequences
- Retrieving selective pressures of all selected neutral and pathogenic variations
- Statistical analysis to understand differences (if any) of selective pressure distributions between pathogenic and neutral variants

Additionally, discovering abundances and pathogenicity of individual amino acids and amino acid groups are objectives of this research.

4 MATERIALS AND METHODS

4.1 Materials

4.1.1 Variation dataset

All human variations downloaded and collected from VariBench (Nair and Vihinen, 2012), a benchmark database for human protein variations created and maintained by Institute of Biomedical Technology (IBT) of University of Tampere (currently maintained by Department of Experimental Medical Science, Lund University, Lund, Sweden). It contains datasets of experimentally verified high-quality variation data carefully chosen from literature and relevant databases. For instance, VariBench datasets can be used to test performance of prediction tools as well as to train novel machine learning-based tools. Five different categories of reference variation datasets are included in the current version of VariBench database. Tolerance-related datasets (Thusberg *et al.*, 2011; Olatubosun *et al.*, 2012) contain information on whether missense variants are tolerated (i.e., benign) or not (functionally impaired) in proteins (Nair and Vihinen, 2012).

Table 4.1: Summary of selected dataset from VariBench dataset:

Dataset	Pathogenic variations	Total number of genes (pathogenic)	Neutral Variations	Total number of genes (neutral)
VariBench dataset	19335	1190	21170	9011
Selected dataset	5958	439	1123	439

Two different datasets were used from the database. The first one reports large set of pathogenic variants whereas the second one contains variants of neutral types. All variants which are part of common proteins were selected from the two datasets. Alternatively, all selected proteins contained variations of both neutral and pathogenic types. An overview of the dataset is given in table 4.1. Among all genes which contained pathogenic and neutral variations, 439 common genes have been found. These 439 genes were selected in this thesis work for analysis. All selected genes contained both type of

variants. In the original dataset, total numbers of variations are higher for neutral type variations. Total number of genes contained neutral variants is also very large in comparison with the pathogenic type. So, on average each gene in the pathogenic type variants contained bigger amount of variations compared to a gene of neutral type. In our selected dataset, all selected 439 genes contained only 1123 neutral variations whereas the amount of pathogenic variants (5958) has been found at least 5 times more into the same genes.

4.1.2 Sequences

All selected proteins sequences and their corresponding gene sequences were downloaded from Ensembl (Flicek *et al.*, 2010) database in FASTA format. This was done in multiple steps. At first, Entrez gene IDs were converted in corresponding Ensembl IDs by “Clone/Gene ID Converter”, software used to convert gene and clone IDs maintained by Bioinformatics unit of CNIO (Spanish National Cancer Research Centre). After that, a set of ortholog sequences were collected and downloaded from Ensembl database using those Ensembl IDs. This downloading of ortholog sequences was performed by Perl script. Number of ortholog sequences varied for different genes. A distribution plot of the number of ortholog sequences are shown in Figure 9.1 (appendix). Individual data files were created in order to download and store individual gene sequences and their corresponding ortholog sequences in individual files. Similarly, human protein sequences and their respective orthologs were downloaded and collected in different individual sequence files.

From the individual sequence files containing proteins and their orthologs, multiple sequence alignments (MSA) were made for each sequence IDs by ClustalW (Larkin *et al.*, 2007). These sequence alignments were required to serve as input in Pal2nal (Mikita S. *et al.*, 2006) software. This software requires respective DNA sequences in FASTA format as reference. Finally, this software returns a codon alignment (alignment of DNA sequences based on respective protein alignment) as output. This whole procedure is repeated for each human gene together with its respective protein. These are supplied by means of single files containing human genes and their orthologs in DNA FASTA files and human proteins and their orthologs in separate protein FASTA files.

4.1.3 Tool for calculating K_a/K_s ratio

Site specific K_a/K_s value calculation was the main aspect of this thesis work. Although there are multiple tools or software available for calculating K_a/K_s values, Selecton (Doron-Faigenboim *et al.*, 2005; Stern *et al.*, 2007) was selected in our analysis as an ideal one. Selecton takes codon-aligned cDNA sequences. It produces output files containing amino acids with their calculated K_a/K_s values for each amino acid of the sequence specified with the input parameters. Here, the human sequence among all homolog sequences was our main interest. Thus the human sequence was specified from all ortholog sequences for each gene. This specification of human sequence was indicated using input parameter of this software. Thus, site specific K_a/K_s values were calculated by Selecton software from the codon alignments. A wrapper script or pipeline was developed in Perl programming language to automate the calculation of codon wise K_a/K_s values for each selected human genes given the codon aligned ortholog sequences (section 4.2.2).

4.1.4 Software for statistical analysis

Good statistical software are available for computational analysis of statistical aspects in scientific research. Among the efficient statistical software, *R* was chosen here mainly for data management, statistical analysis, and data visualization.

4.2 Methods

4.2.1 Preparation of datasets

From the variation dataset VariBench (Nair and Vihinen, 2012) those protein sets were selected which contained variations for both neutral and pathogenic types. Alternatively, selected protein sequences in this thesis work contained both neutral and pathogenic types of variations. However, quantity of pathogenic variations was higher than that of neutral type. 5,958 pathogenic and 1,123 neutral variations were selected for further analysis from 439 common protein or gene sequences (details are given in section 4.1.1).

All protein sequences together with their orthologs were stored in separate files. It should be noted here that all cDNA and protein sequences were collected and stored in FASTA format. Codon alignment was an input requirement for the K_a/K_s analysis software

Selecton (Doron-Faigenboim *et al.*, 2005; Stern *et al.*, 2007). So, all the cDNA sequences were aligned into codon alignment by Pal2nal (Mikita S. *et al.*, 2006). Thus, input requirement of site specific calculation of K_a/K_s value was ensured.

4.2.2 Building pipeline

The main aim of building this pipeline was to automate the actual analysis stepwise and collecting the output results in suitable formats. This programming algorithm or so called pipeline estimated K_a/K_s values for all selected proteins one by one automatically. The input parameters were also specified with the input files containing codon-aligned DNA sequences. Reference sequence name among all homolog sequences (human sequence) was specified. Upon calculating the K_a/K_s values for all amino acids of a particular gene sequence, the pipeline continued receiving the outputs and stored each output file according to the respective sequence ID in specified locations.

The workflow of this pipeline can be outlined by the following algorithm:

1. Provide 2 files as input; one containing nucleotide sequences and the other having respective protein sequences (both of these files containing human sequences and its orthologs were downloaded from Ensembl database by Perl script)
2. Make MSA of proteins
3. Supply the MSA of protein sequences and the respective nucleotide sequences into Pal2nal for codon alignment
4. Provide the codon aligned nucleotide sequences created in step 3 into Selecton; fix the input parameters
5. Collect the output files containing K_a/K_s ratio together with other aspects created by Selecton which are created in the specified location
6. Repeat step 1 to 5 for all selected human gene sequence files

4.2.3 Calculation of site-specific K_a/K_s values

Selecton was selected for site-specific calculation of K_a/K_s value for individual amino acids of each gene sequence. Selecton read all the codon aligned sequence files of human cDNA sequences together with its orthologs and returned respective selective pressure or K_a/K_s values. It also returned a numeric grading scale between 1 and 7 specifying the conservation rank of each amino acid (shown in Figure 4.1 and Figure 4.2).

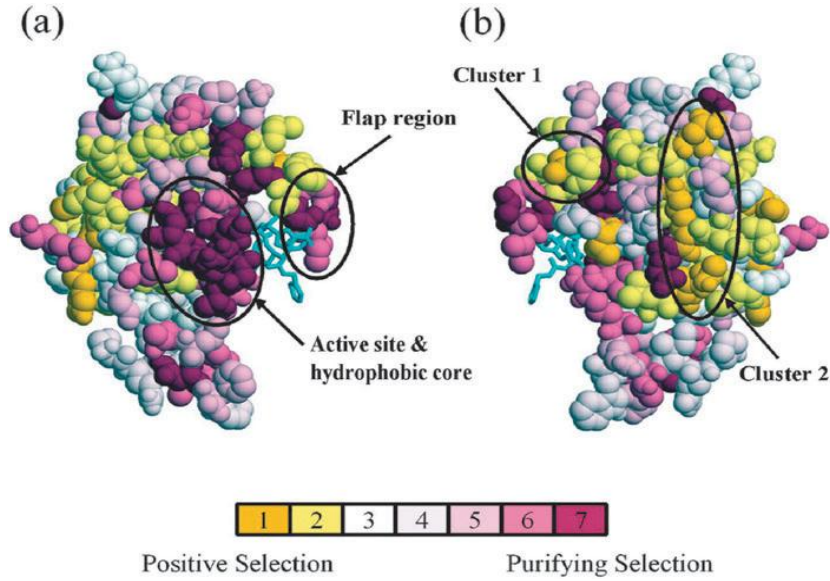


Figure 4.1: Selecton results for HIV-1 protease chain A complexed with the inhibitor. K_a/K_s scores are color-coded onto its Van-der-Vaals surface. The inhibitor (Ritonavir) is shown in light blue as a backbone model. Significant purifying and positive selected sites (p -value < 0.05) are colored in bordeaux (color number 7) and dark yellow (color number 1) respectively. (a) View of the active site (residues 22–33), flap region (residues 47–52) and hydrophobic core (residues 74–87); (b) View of two clusters of positively selected sites. Cluster number 1 contains residues Met46, Phe53, Ile54 and Pro79. Cluster number 2 contains residues Leu10, Val11, Thr12, Leu19, Lys20, Met36 and Asn37 (Figure and caption taken from Doron-Faigenboim *et al.*, 2005).



Figure 4.2: Selecton result for “DNA mismatch repair protein MLH1” upon color shaded according to evolutionary pressure estimated for each amino acid.

The example of this grading scale is shown for the human immunodeficiency virus type 1 (HIV-1) which is as essential enzyme for viral replication and thus is the target for design of drug inhibitors (Peng et al., 1989; Flexner, 1998; Doron-Faigenboim *et al.*, 2005). This numerical rank returned by Selecton software nicely shows the conservation state of the specific amino acid which indicates whether the amino acid will conserve, remain neutral or change. Selecton also returns a phylogenetic tree of homolog sequences for each gene sequence.

4.2.4 Statistical analysis

Upon calculating K_a/K_s values for site specific amino acids of selected genes, distribution of K_a/K_s values were analyzed statistically. For instance, distributions of these values were plotted in graphs to visualize and compare. Several boxplots were drawn using the obtained values upon previous calculations. These Figures were made not only to compare overall distribution of neutral and pathogenic variations but also to observe the differences between each amino acid of both types.

In addition, statistical test for distribution comparison of both variations were performed in order to examine distribution pattern between them. Smirnov-Kolmogorov test was chosen in this thesis work. These comparisons with this particular test were also done in both individual amino acid levels, and for different amino acid groups.

5. RESULTS

5.1 K_a/K_s values

5.1.1 Overall comparison of K_a/K_s values

Calculated K_a/K_s values were collected for the selected variation dataset for both pathogenic and neutral types. The distribution of those K_a/K_s values were analyzed statistically and plotted.

The distribution of K_a/K_s values for all selected variations (amino acids) of both pathogenic and neutral type are shown in Figure 5.1. Average K_a/K_s score was lower for pathogenic variations than that of neutral type. Inter-quartile distances are also higher for neutral variations. First quartile, median and third quartile K_a/K_s values for pathogenic variations are 0.032, 0.11, and 0.25; and for neutral variations are 0.14, 0.3, and 0.56 respectively. 259 outliers were found in pathogenic dataset, however, no outlier was found in neutral dataset. Average K_a/K_s ratio in hominids has been found 0.20 (De Magalhães and Church, 2007). Thus, a horizontal line was drawn at 0.2 level of y-axis in Figure 5.1 and it is clearly visible from the Figure that the median K_a/K_s value is lower for pathogenic variations.

5.1.1.1 Two-sample Kolmogorov-Smirnov test

Two-sample Kolmogorov-Smirnov test was performed to examine the overall distribution of pathogenic and neutral types of variations according to their K_a/K_s value. This test was performed to check and distinguish distribution of K_a/K_s values between those two types of variations. The null hypothesis was considered as identical distributions of K_a/K_s values between two groups. On the other hand the alternative distribution was considered as different distribution of K_a/K_s values both for neutral and pathogenic type variations. This two sided test has proven the fact that there are differences in distributions between two types of variations (p -value $< 2.2\text{e-}16$, $D = 0.3487$) and the null hypothesis was rejected. Thus, significant differences in distribution of K_a/K_s values were found. The boxplot (Figure 5.1) clearly indicates lower distribution of K_a/K_s values for pathogenic type variations from the neutral type.

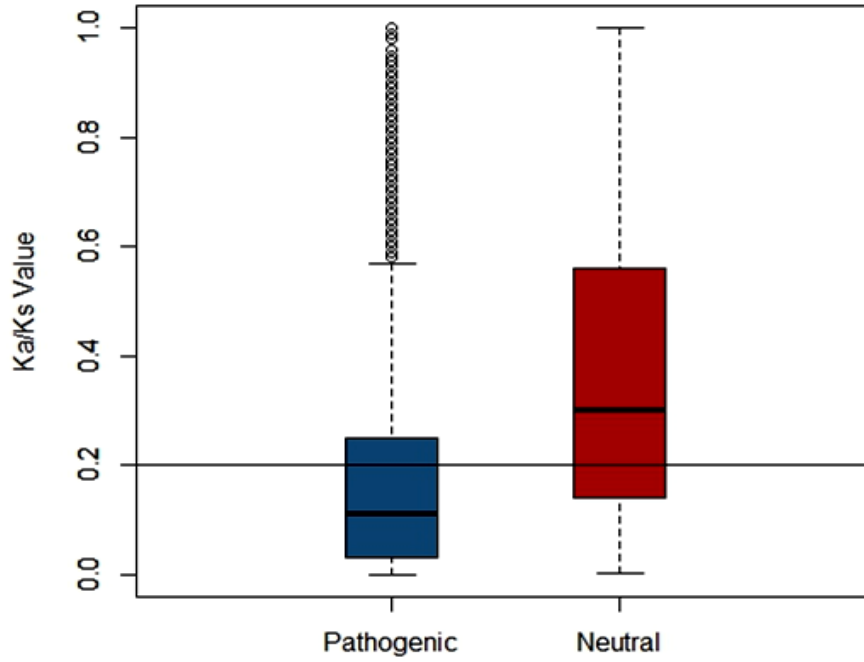


Figure 5.1: Distribution of K_a/K_s values of amino acids for all selected neutral and pathogenic types of variations. Here the boxes represent inter-quartile distances and the thick horizontal marks inside the boxes indicate the median of the respective box. The dots at the lower and upper portion of boxes represent distance from minimum value to the first quartile and from third quartile to maximum value respectively. The circles which are drawn outside the maximum value of inter-quartile range represent outliers.

5.1.2 K_a/K_s value comparisons for individual amino acids

As the overall distribution of amino acid variations were different in pathogenic and neutral type variants, distribution of individual amino acids were plotted to observe an insight view of their distribution property. Figure 5.2 demonstrates distribution of K_a/K_s values of individual amino acids for both neutral and pathogenic type variations.

5.1.2.1 Two-sample Kolmogorov-Smirnov test:

As this thesis work had two sample datasets, neutral and pathogenic variations, two sample KS test was performed to analyze and compare distribution of K_a/K_s values between those two samples. The results of these tests between those two samples for each amino acids are shown in Table 5.1. Null hypothesis was assumed as significant

differences between distributions of pathogenic and neutral variations. All p -values lower than 0.05 reject the null hypothesis and proof that there are significant differences between those two groups of variations. Such amino acids are shaded with yellow color. Important phenomenon has been observed here that all the amino acids for which null hypothesis are accepted (higher p -value) and differences are insignificant, these amino acids had less than 20 observations for neutral variants. Thus, more observations may change the distribution pattern of these exceptional amino acids which gave identical K_a/K_s distributions for pathogenic and neutral variants.

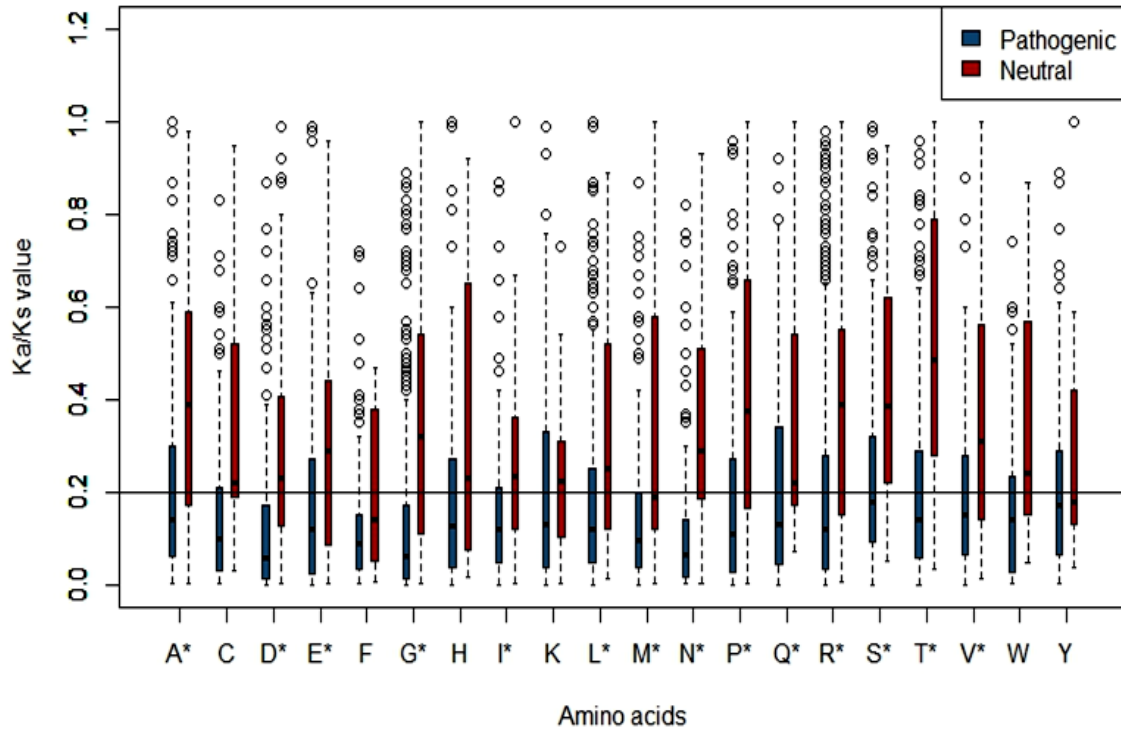


Figure 5.2: Distribution of K_a/K_s values of individual amino acids separately for both neutral and pathogenic types of variations. Here the boxes represent inter-quartile distances and the thick horizontal marks inside the boxes indicate the median of the respective box. The dots at the lower and upper portion of boxes represent distance from minimum value to the first quartile and from third quartile to maximum value respectively. The circles which are drawn outside the maximum value of inter-quartile range represent outliers. Asterisk (*) indicates amino acids having significant K_a/K_s value distribution differences between pathogenic and neutral variants.

Table 5.1: Two sample (pathogenic and neutral variations) KS test results for individual amino acids. Amino acid rows shaded by yellow color indicate significant differences (p -value < 0.05) of K_a/K_s value distributions between pathogenic and neutral type variants.

Amino acid	D value	p -value	Pathogenic variations	Neutral variations
A	0.3558	5.572e-06	295	61
C	0.5081	0.1606	209	5
D	0.397	0.0001479	219	35
E	0.342	0.000203	241	47
F	0.3776	0.06865	132	13
G	0.4831	2.719e-08	724	41
H	0.2805	0.1847	140	17
I	0.2917	0.01026	169	38
K	0.1919	0.8209	109	12
L	0.3394	0.001796	395	33
M	0.3864	0.02295	125	17
N	0.5804	2.16e-09	143	39
P	0.3797	4.519e-05	238	44
Q	0.4713	0.001482	111	19
R	0.3641	2.967e-08	730	75
S	0.4126	7.475e-05	254	34
T	0.4751	4.355e-08	181	50
V	0.3364	0.0003358	233	46
W	0.3659	0.5508	88	5
Y	0.2567	0.4091	152	13

5.1.3 K_a/K_s value comparisons according to amino acid groups

Distributions of the K_a/K_s values have been found to have same trend in both individual (Figure 5.2) and overall (Figure 5.1) comparison. Distributions of K_a/K_s values according to amino acid groups were plotted (Figure 5.3). These groups were selected according to their structural property (more details in review of literature).

Figure 5.3 reveals comparison of K_a/K_s values according to five subgroups of amino acids for both pathogenic and neutral types. In this group-wise comparison, all amino acid groups have scored lower K_a/K_s value distributions for pathogenic variations. Hardly any outlier was found in neutral dataset whereas pathogenic set contained outliers for all amino acid groups. As average K_a/K_s value is 0.2 for hominids, a horizontal line at this level (0.2 in y-axis) is also drawn here. All amino acid groups have scored lower median value than this horizontal line for pathogenic variants whereas the average median value was higher than this standard level for neutral variants. However, only aromatic amino acid group has shown exception in the sense that median K_a/K_s value of neutral variants of this group has scored lower than the horizontal line. Inter-quartile differences are wider in other groups in comparison with the aromatic group. For instance, polar (uncharged) and non-polar (aliphatic) groups have higher quartile distances.

5.1.3.1 Two-sample Kolmogorov-Smirnov test

Like the previous studies, two sample KS tests were performed for different amino acid groups between pathogenic and neutral variations (Table 5.2). Here null hypothesis is identical K_a/K_s value distributions between pathogenic and neutral type of variants. The alternative hypothesis is considered that the K_a/K_s value distributions between these two groups are different or there are significant differences between their distribution properties.

Here, all p -value lower than 0.05 will reject the null hypothesis. Table 5.2 clearly demonstrates that null hypothesis for all amino acid groups has been rejected and hence it has been proved that all amino acid groups have different K_a/K_s value distributions between pathogenic and neutral types of variants. The difference has been found highest (lowest p -value) for non-polar (aliphatic) and polar (uncharged) groups whereas aromatic group has scored least difference (highest p -value). Another interesting phenomenon was seen that the groups which have scored highest difference contained larger amount of observations (variations) and the lowest K_a/K_s value difference group had lowest number of observations. Thus, more observations might change the p -value of the aromatic group.

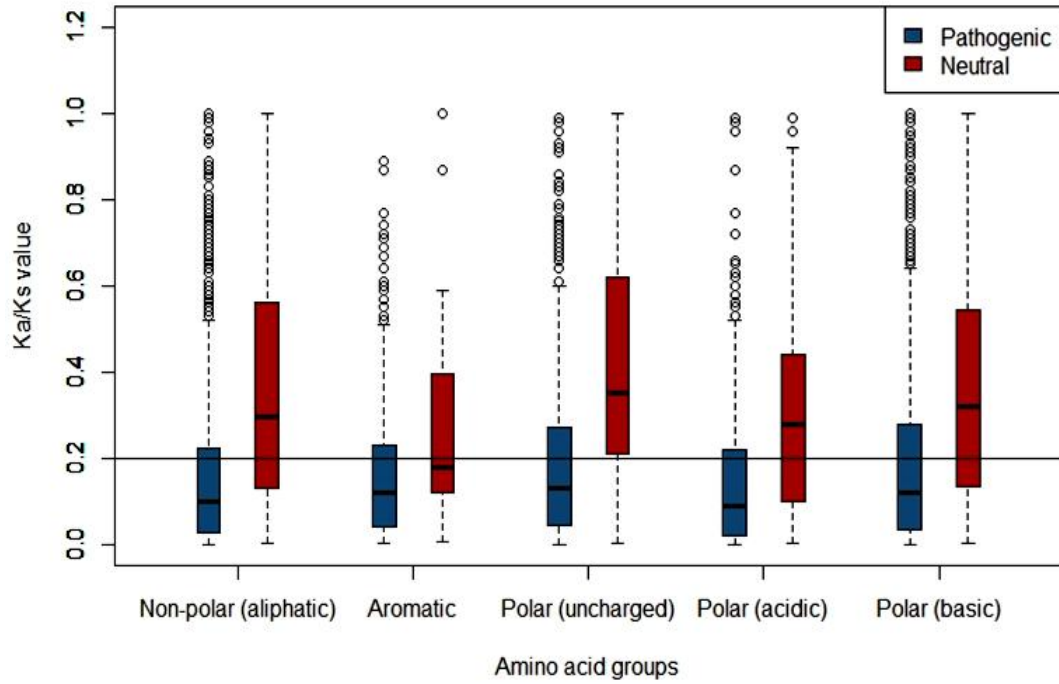


Figure 5.3: Distribution of K_a/K_s values of different amino acid groups for both neutral and pathogenic types of variations. Here the boxes represent inter-quartile distances and the thick horizontal marks inside the boxes indicate the median of the respective box. The dots at the lower and upper portion of boxes represent distance from minimum value to the first quartile and from third quartile to maximum value respectively. The circles which are drawn outside the maximum value of inter-quartile range represent outliers.

Table 5.2: Two sample (pathogenic and neutral variations) KS test results for different amino acid groups:

AA group	D value	p -value	Pathogenic variations	Neutral variations
Non-polar (aliphatic)	0.3677	$< 2.2e-16$	2179	280
Aromatic	0.2661	0.03483	372	31
Polar (uncharged)	0.4203	$< 2.2e-16$	898	147
Polar (acidic)	0.33	5.234e-07	460	81
Polar (basic)	0.3008	8.184e-08	979	104

5.2 Pathogenicity comparisons

5.2.1 Comparison of individual amino acid variability

Analysis of amino acids was also a major aspect in this thesis work. Original amino acids and their changes upon variations were collected and analyzed. Quantity of variations on particular amino acids varied a lot and more specifically, certain amino changes have been found more frequent than others. However, if few amino acids changes are more frequent than others, it is possible that their proportions in sequences are also higher than others. So, comparisons of amino acid changes also require their respective proportions in sequences.

Amino acid changes upon variation and their respective proportions in original sequences are shown in Figure 5.4 both for pathogenic and neutral variations. Here, changes are calculated according to their respective percentages. Exact proportions of amino acid variation rates were calculated upon normalization according to their contributions in original protein sequences. As amino acid variations have been compared with their sequence proportions, percentages of variations were divided by their relative sequence proportions. In this way, normalization was performed for each group of individual amino acids. So, pathogenicity of amino acid changes have been found in comparison with their proportion in sequences where the original sequence entity of each amino acid is considered unity (or 1).

These proportions of variations upon normalization are visualized in Figure 5.5. Both types of variations have been found more frequent for some amino acids, for instance, R, M, H etc. in comparison with their relative sequence proportions. In contrast, K, L, Q, S etc. amino acids have shown inverse behavior in this point of view. They have shown less variations compared to their sequence proportions. Moreover, some amino acids show higher pathogenic variants than neutral and vice versa.

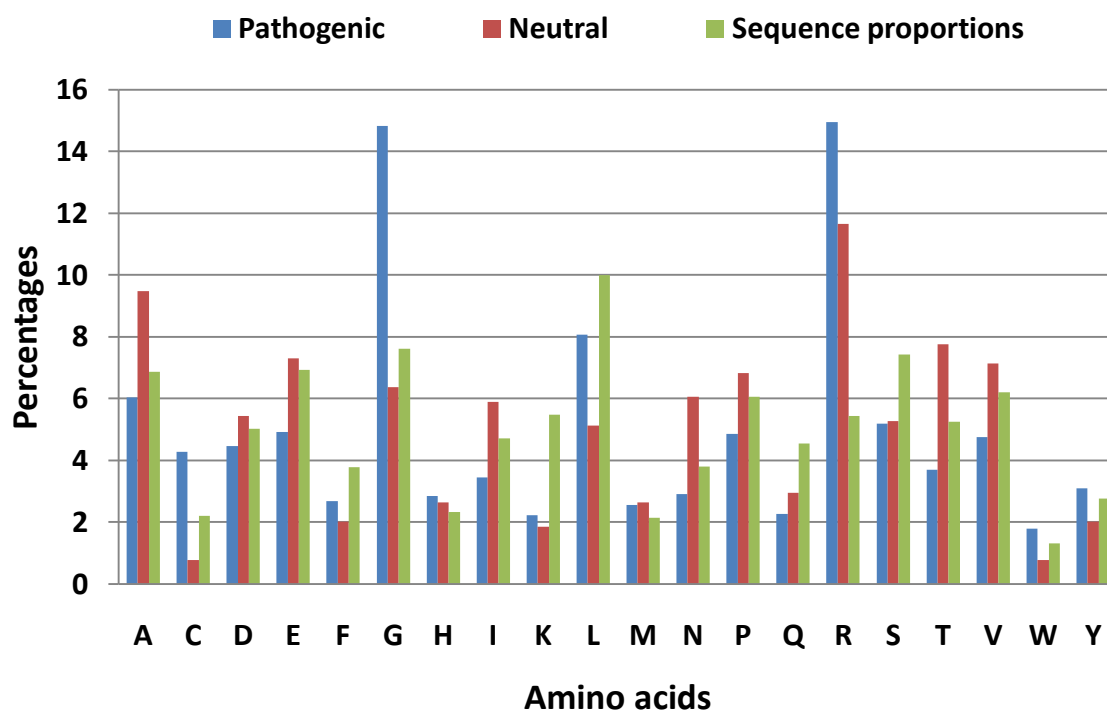


Figure 5.4: Proportion of amino acid changes for both pathogenic and neutral types together with their corresponding quantities in original protein sequences. Here, sequence proportions are total quantity of respective amino acids in all the protein sequences of selected dataset.

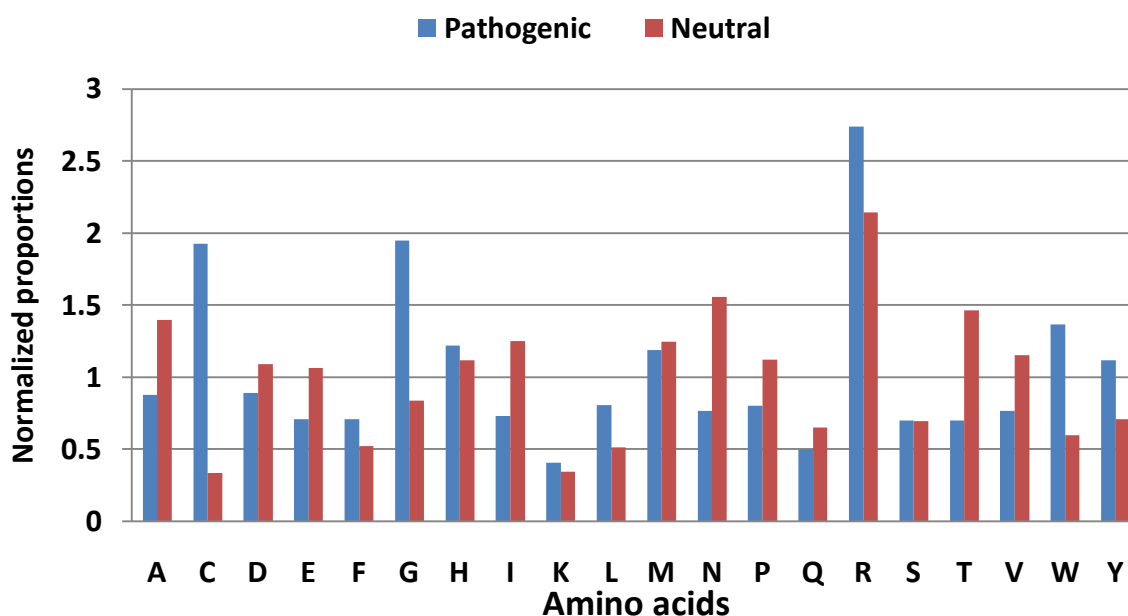


Figure 5.5: Proportion of amino acid changes after normalization for both pathogenic and neutral types

5.2.2 Analysis of pathogenicity for individual amino acids

For some amino acids, relative proportion of one type of variant is found higher (>1) in Figure 5.5 than their sequence proportions whereas the proportion of other variant was lower (<1). Thus, it is obvious to calculate pathogenicity. Pathogenicity of individual amino acids was calculated in order to investigate their pathogenic property. Some amino acids have revealed higher trend to be converted into pathogenic type upon variation whereas others showed opposite characters to be changed into neutral type.

Pathogenicity of amino acids is calculated by dividing the proportion of pathogenic variants by the respective neutral one. Thus, the property of being converted into pathogenic type variations are shown in Figure 5.6 in descending order where amino acids associated with least pathogenic property are shown at the last. C, G, W etc. have scored highest pathogenic property whereas I, N, T etc. have shown lowest pathogenic property. H, S, M etc. can be considered as almost medium in this contest.

5.2.3 Group wise comparison of amino acid variability

Upon studying individual amino acid variations, another aspect in this work was to observe group wise comparison of amino acid variability. The main aim of this comparison was to see variability characteristic for different amino acid groups.

Group wise comparison of amino acid variations is shown in Figure 5.7 with their corresponding proportion in protein sequences. As variability is unequal for different groups, it was also required to see their relative sequence proportions. Some groups have shown higher variation than their sequence proportions whereas some have shown equal or lower variation. For instance, polar (basic) and polar (uncharged) groups have shown opposite characteristics for variation in comparison with their sequence proportions. To measure exact variation proportions in comparison with the sequence quantity, normalization was performed.

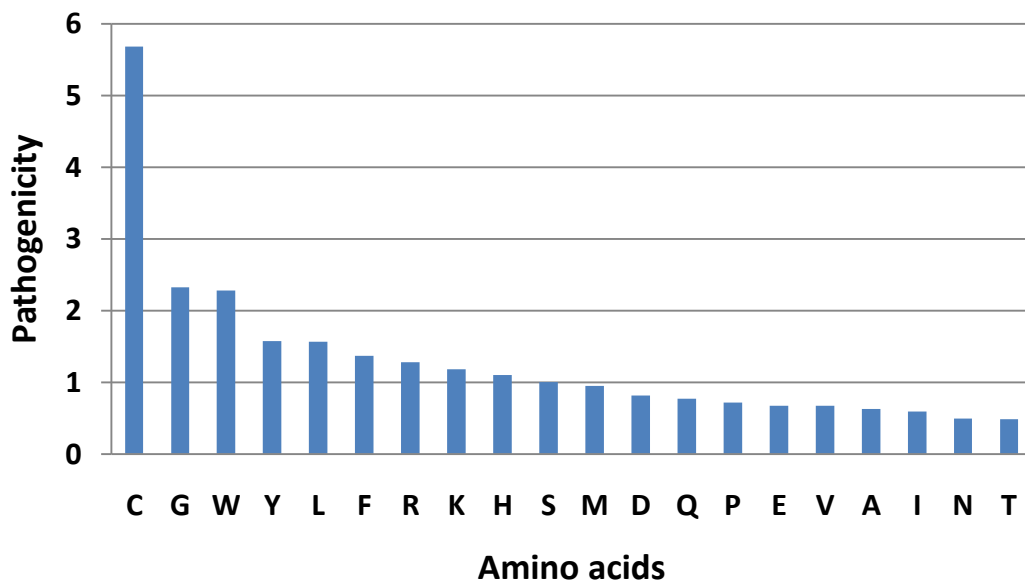


Figure 5.6: Pathogenicity of amino acids (in descending order)

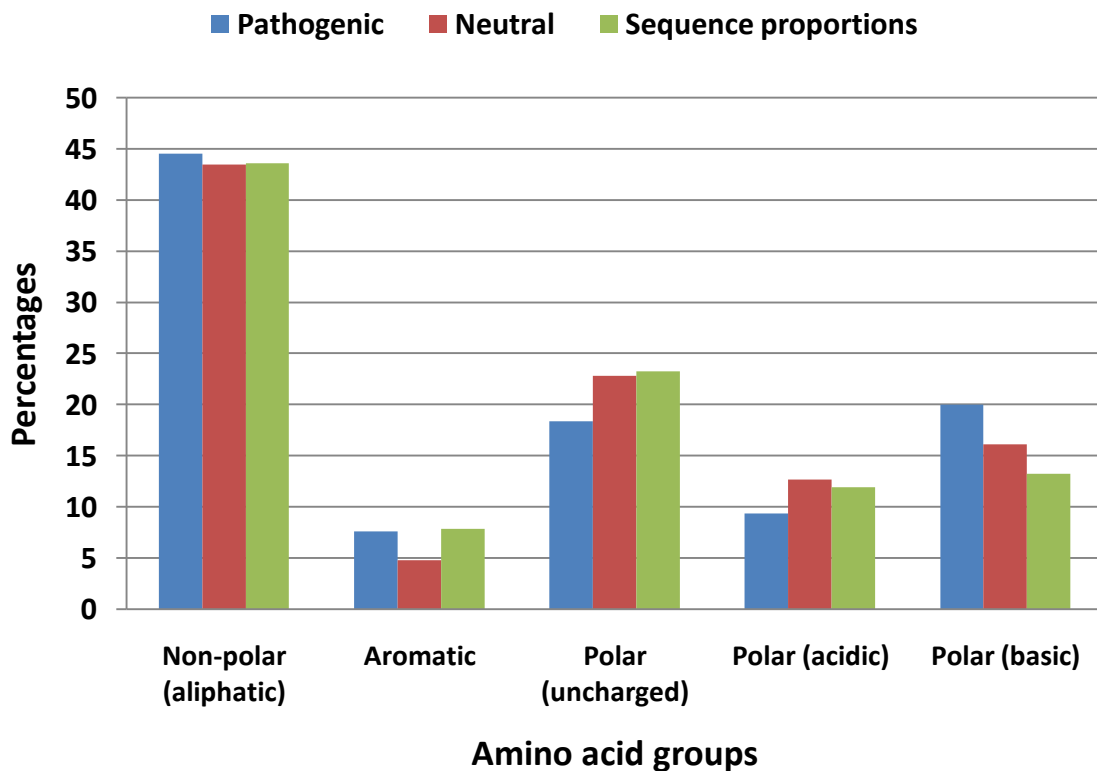


Figure 5.7: Proportion of amino acid changes into both pathogenic and neutral types. Here, sequence proportions are cumulative quantity of amino acids (in groups) in all the protein sequences of selected dataset.

The relative proportions of amino acid groups in comparison with their respective quantities into protein sequences are shown in Figure 5.8. Proportions of both pathogenic and neutral variations were divided by their respective sequences proportions individually. Thus, their relative proportions of variation were calculated in respect to their respective sequence proportions.

In Figure 5.8, it is clearly visible that polar (basic) amino acids have showed highest variation both for neutral and pathogenic types in compared to others. Aromatic group have scored lowest for neutral variation but not for pathogenic. In contrast, polar (uncharged) and polar (acidic) amino acids have shown less pathogenic variations, however, their neutral variant proportion is not equal. Non-polar aliphatic groups have equal variation proportions for both variations and also their relative proportions of variation are almost equal to sequence proportions (1 in this case).

5.2.4 Analysis of pathogenicity for different amino acid groups

As some amino acid groups showed higher proportion for pathogenic variations and others scored higher for neutral types, it was required to calculate their pathogenicity of pathogenicity to have individual exact proportions in comparison with others. Thus, pathogenicity for different amino acid groups was calculated. These values are represented in Figure 5.9 in descending order of pathogenicity.

Aromatic amino acid group has shown highest pathogenicity. Amino acids belong to this group have been converted more frequently into pathogenic variations. In contrast, polar (acidic) group has shown less pathogenic property. Non-polar aliphatic group has behaved almost neutrally in this point of view. They have almost equal probability to be changed into either pathogenic or neutral type variants. Moreover, their variation proportions have also been found almost equal to their sequence proportion (1 in this case). Although they have almost equal possibility to become any of the two types of variants upon variation, they have scored slightly higher for pathogenic variation.

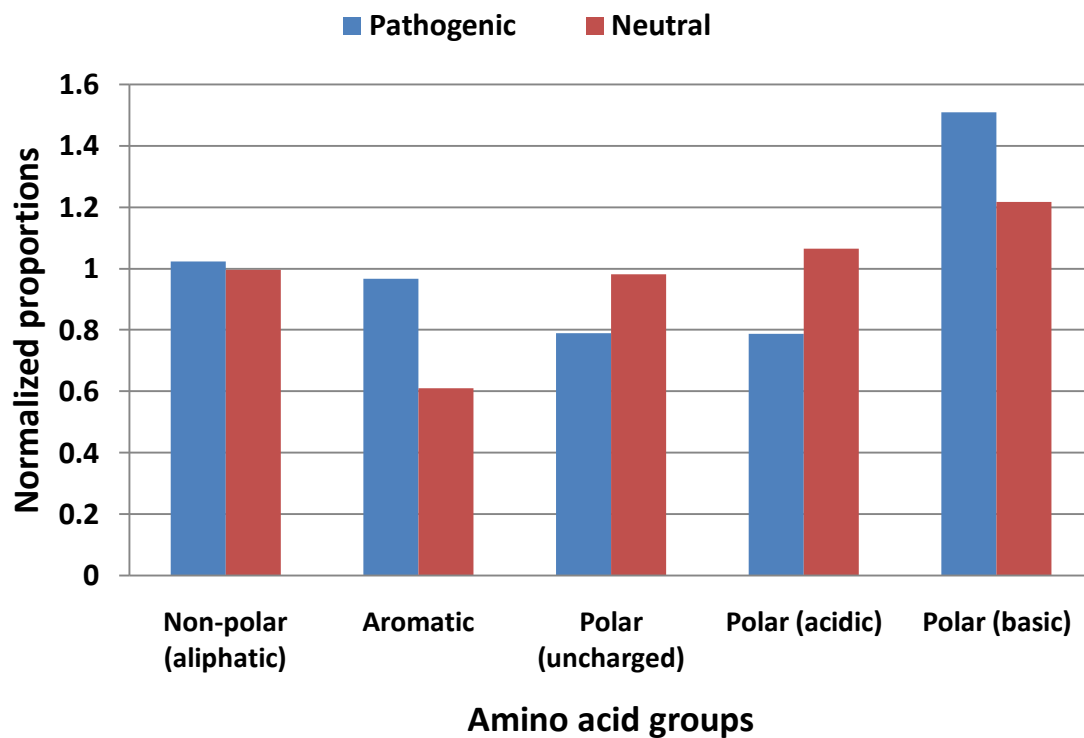


Figure 5.8: Relative proportion of amino acid changes upon normalization in different groups for both pathogenic and neutral types

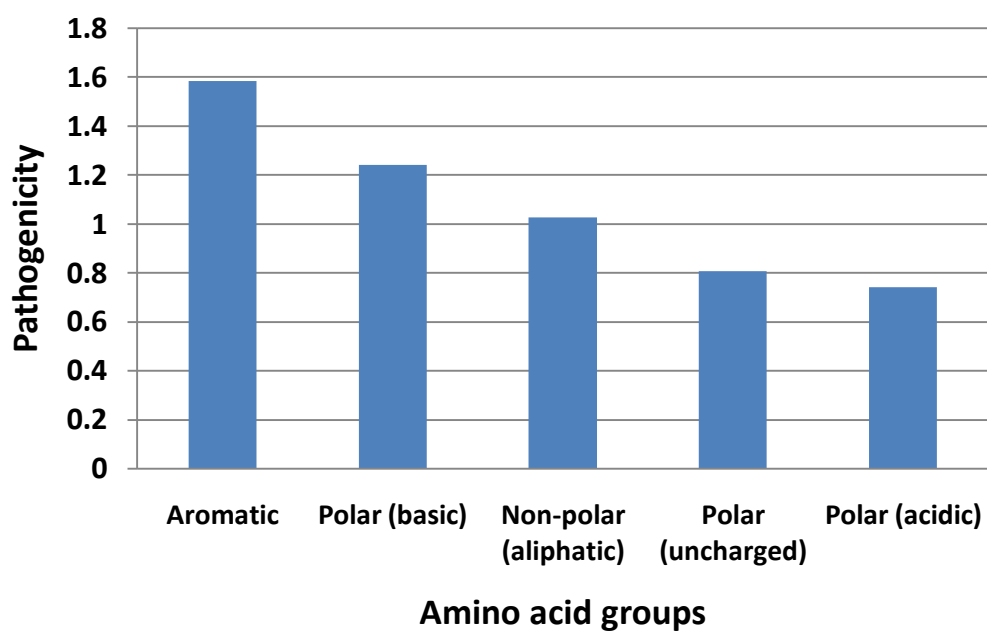


Figure 5.9: Pathogenicity of amino acid groups (in descending order)

6. DISCUSSION

Evolutionary pressure estimation and further analysis were main aspect of this research. Natural selection can result positive and negative selection upon evolution. K_a/K_s ratio returns the evolutionary pressure acting on a gene. As we are studying here evolutionary pressure for human missense variations, site-specific codon wise calculation of K_a/K_s ratio was needed. Upon calculation, evolutionary pressure of each amino acid of the selected proteins was available. Thus, K_a/K_s ratio for each selected missense variants were retrieved from the protein sequences for both pathogenic and neutral variants.

Human missense variations of both pathogenic and neutral type were selected from VariBench database. All human gene and respective protein sequences containing all the selected missense variants were downloaded from Ensembl. Nucleotide sequences were aligned according to respective protein alignments (codon-alignment) and then K_a/K_s ratio for each codon was calculated from the codon-alignment of all orthologs. This calculation resulted K_a/K_s values for all amino acids of each selected protein. Thus, evolutionary pressure of all selected variants was identified and collected from the results.

6.1 K_a/K_s value aspects

6.1.1 Pathogenic and neutral variation data sets

Among all the missense variations reported in VariBench database 5,958 pathogenic and 1,123 neutral variations were selected in this research. Selected pathogenic variation set was larger than the neutral set. Alternatively, total quantity of selected pathogenic variations is approximately five times larger than the selected neutral variations. This quantity difference between two sets of variants might slightly affect the K_a/K_s value distribution results shown in Figure 5.1.

6.1.2 K_a/K_s results

De Magalhães and Church (2007) have found average K_a/K_s ratio 0.2 for human-rhesus gene pairs (9,857 pairs), 0.13 for human-mouse gene pairs (12,063), and 0.13 for human-rat gene pairs (11,594). In addition, the average K_a/K_s ratio in hominids has been found 0.20 (De Magalhães and Church, 2007). Therefore, a horizontal line was drawn at 0.2 in

y-axis of K_a/K_s ratio to compare site specific K_a/K_s value distributions for both variations in this study.

From Figure 5.1, distribution of K_a/K_s values has been found lower in pathogenic variations than those neutral variations. Not only mean and median values are higher but also distributions of inter-quartile ranges are larger in neutral variations in comparison with pathogenic variations. From the theoretical background, it is the fact that amino acids (or sites) having higher selective pressure are more likely to be changed. Alternatively, sites of genes or proteins having lower selective pressure will be conserved, i.e. not likely to be changed easily. As pathogenic variations have scored lower K_a/K_s value distributions, this phenomenon clearly indicates that pathogenic variations are more conserved than the neutral type. Alternatively, neutral variations are more likely to be changed in next generations than that of pathogenic type. The horizontal line at 0.2 level of y-axis clearly differentiates inter-quartile ranges of evolutionary pressure acting on pathogenic and neutral variations. The median K_a/K_s value for pathogenic variations has been found much lower than the horizontal line. In contrast, median of pathogenic variations is much higher than this average line.

Two-sample Kolmogorov-Smirnov test was performed to check the K_a/K_s distribution property between these two types of variation sets. Null hypothesis (H_0) was assumed that the both sample distributions are identical. On the other hand, alternative hypothesis (H_a) was considered that their evolutionary pressure distribution pattern is different. This two sample KS test has also revealed clear difference of K_a/K_s distribution values between pathogenic and neutral types (p -value $< 2.2\text{e-}16$, $D = 0.3487$). As p -value was much lower than 0.05, the null hypothesis (H_0) was rejected. Thus, it is clearly marked here that conservation property of pathogenic variation is not similar to the pathogenic type.

This fact has also been observed in case of distribution K_a/K_s values for individual amino acids separately for both types of variations (Figure 5.2). As median values of K_a/K_s distribution for all amino acids have been found lower for pathogenic type of variants, all amino acids have been found more conserved for pathogenic type than that of neutral type. Alternatively, all amino acids scored higher inter-quartile K_a/K_s value distribution for neutral type variations. Some amino acids have really significant differences in K_a/K_s value distribution between the two type variations whereas others have shown small

differences. For instance, K_a/K_s value distribution of threonine is very different in two groups. In fact, K_a/K_s value quartiles vary a lot for threonine and median values have highest difference in threonine among others. In addition, amino acids like proline, and glycine have shown same type of characteristics to have different K_a/K_s value distributions. They had wider quartile differences in K_a/K_s value distribution among both types of variations. In contrast, few amino acids have shown least difference in mean and median values for K_a/K_s distribution between two variation groups. Tyrosine, phenylalanine, and tryptophan might be considered as this type of amino acids where K_a/K_s value distributions has not shown significant differences. Interesting phenomenon found here is that F, Y, and W belong to the second type and they all contain aromatic ring in their structure. Figure 5.3 also clearly indicates least difference in the distribution of K_a/K_s value between pathogenic and neutral variation class for aromatic amino acid group.

In Figure 5.2, the horizontal line drawn at 0.2 level of K_a/K_s value clearly differentiates inter-quartile boxes of almost all boxplots. Major portion of the inter-quartile boxes have fallen below the horizontal line for pathogenic variations. In contrast, this feature has been found opposite for the neutral variations. Most part of the inter-quartile distances are positioning above the horizontal line. Median values for most of the amino acids are also found below the 0.2 line for pathogenic variations and above for the neutral variations. Only F, M, and Y showed exception where median of neutral variations have been found lower than this line. However, median of all amino acids are lower than the horizontal line.

Two sample KS test was performed for each amino acid to examine distribution property between both types of variants. All the parameters, for instance, significance level, null hypothesis, and alternative hypothesis were same like the overall distribution analysis. However, few amino acids have shown deviation. Null hypothesis for C, F, H, K, W, and Y was accepted as they scored higher p -value (>0.05). Thus, distributions of selective pressures have been found quite similar to each other between pathogenic and neutral type. More data samples might change the trend found in these amino acids. Nevertheless, all other amino acids have followed the main trend found in overall distribution and their null hypothesis was rejected.

Based on the information found for individual amino acids, it was obvious to check differences in K_a/K_s value distribution in different amino acid groups according to their properties. In Figure 5.3, inter-quartile differences have been found significantly higher in amino acids which are polar but having no charge in their structure. Same Figure clearly and nicely demonstrates least difference in K_a/K_s distribution for aromatic amino acids (having aromatic ring in their structure). However, all groups have showed identical results to have lower K_a/K_s value distributions among pathogenic type variations and hence pathogenic variations have been found more conserved than the neutral type. In addition, all amino acid groups have showed same behavior to this horizontal line at 0.2 level of K_a/K_s value. Median and inter-quartile areas for almost all groups were lower than this horizontal level for pathogenic variations whereas it was seen opposite for neutral variations. Aromatic group has shown deviation here. Median value for neutral variations of this group was lower than the horizontal line.

Like all the previous analysis, two sample KS test was performed for different amino acid groups with same parameters. Although their p -values varied greatly, null hypothesis for all groups was rejected and hence all groups showed different distribution property between pathogenic and neutral type variants. The distribution difference of K_a/K_s value was less for aromatic group; however, their observation points were also least among all groups. More observation points or variations might increase the difference the distribution pattern between pathogenic and neutral variants of this group. In contrast, non-polar (aliphatic) and neutral polar groups have shown highest difference of evolutionary pressure distribution pattern between both types of variants.

6.2 Amino acid variability aspects

6.2.1 Variability of individual amino acids

This study for individual amino acids shows that some amino acids (R, H, and M) are very likely to be mutated into either pathogenic or neutral variants (Figure 5.4). These amino acids have been found to have the higher tendency to be converted into any type of variations although their proportions in protein sequences are relatively less. Arginine and glycine are examples of this kind of amino acids where variations have been found more frequent than others. However, their degree of conversion into pathogenic and neutral

type variation has been found different. Some are more likely to be changed into pathogenic type and rests are more likely to become neutral variant. On the other hand, some amino acids have been found less frequent to become variants although their proportions are higher in original protein sequences. Lysine, glutamine and leucine can be taken as examples of this category.

Amino acid variability comparison was performed upon normalization in respect to their proportion in protein sequences. Figure 5.5 clearly indicates that amino acid variability is not equal for individual amino acids. In fact, it varies a lot for both types of variants. Both variations have been found more frequent in case of arginine although proportion of pathogenic type has been found higher than the neutral type. Glycine has shown same phenomenon like arginine and it is also more likely to be converted into pathogenic type rather than neutral. In contrast, certain amino acids have lower trend to be changed into pathogenic type, rather, they are more likely to become neutral type and hence they show lower pathogenicity. Alanine, isoleucine, asparagine, and threonine might be given as example to this second type which is found less pathogenic in comparison with the first type. However, some amino acids like lysine and glutamine are unlikely to be changed. These amino acids show lower variability irrespective of pathogenic or neutral type.

Amino acids which show higher variability differ in degree of pathogenicity. Some of them show lower degree of variability for pathogenic type, whereas others show higher pathogenicity. So, their pathogenicity was calculated to observe pathogenicity which can be observed in Figure 5.6. Among all amino acids, cysteine, glycine, and tryptophan show higher pathogenicity. They are very likely to be converted into pathogenic variations. On the other hand, certain amino acids like threonine, asparagine, and isoleucine show very low pathogenicity. They are more likely to be converted into neutral type variants. Thus, degree of pathogenicity of certain amino acids has been found very different. From the Figure 5.6, it is clearly visible that cysteine (pathogenicity 5.68) has shown more than ten times higher pathogenicity than threonine (pathogenicity 0.48). Thus, cysteine has shown higher probability of changing into a pathogenic variant rather than neutral. Some amino acids show almost equal pathogenicity. These are changing into either pathogenic or neutral types upon variation. Amino acids which fall into this category are histidine, serine, and methionine.

6.2.2 Variability of different amino acid groups

Figure 5.7 describes overview of amino acid variability in different groups. Some group has showed higher variation rate, but some are less abundant to be mutated compared to their proportions in protein sequences. For example, positively charged basic amino acids have expressed higher variability than their sequence proportions, but uncharged polar amino acid group has shown opposite characteristics in this point of view. In contrast, non-polar aliphatic amino acids have remained neutral in this occasion.

Variability of amino acid groups have been more clearly observed upon normalization with their respective proportions in protein sequences. Figure 5.8 nicely shows more variability of polar basic amino acids than that of aromatic group. Uncharged (neutral) and polar acidic amino acid groups have revealed almost equal variability, except the later one has scored slightly higher for neutral variation. These two groups have shown less pathogenic property. Others, for instance, polar basic and aromatic groups are opposite in this point of view. They show higher pathogenicity. However, the non-polar aliphatic group has acted neutrally. This group has been found equal probability to be changed into any of the two types of variants.

However, pathogenicity property upon variation can be compared nicely among the groups from Figure 5.9. Aromatic group was found most pathogenic. Alternatively, amino acids belong to this group are changing more frequently into pathogenic type than neutral. In this view point, polar acidic and polar uncharged amino acid group was found less pathogenic than others. Pathogenicity of aromatic group was found almost 2 times higher than polar acidic group. Pathogenicity of non-polar aliphatic group was close to 1. Thus, it has tendency of changing into any of the two types of variants.

6.3 Future perspectives

This study was initiated to evaluate if K_a/K_s value is efficient for distinguishing pathogenic and neutral variants; and it was also a great interest here if it can be used for predicting pathogenicity of novel variants. As this study has given excellent results, it will help in further bioinformatics research in the group where this research was performed. It will definitely inspire calculation or estimation of K_a/K_s values on a larger scale so that

site-specific (codon wise) evolutionary pressure would be known for each protein coding site of human genome.

6.3.1 PON-P

In the next version of PON-P, K_a/K_s feature will be incorporated for pathogenicity prediction. This evolutionary pressure evaluating feature is not used at the current version of PON-P. The results and methodology used for K_a/K_s value in this study will inspire and assist to implement the pathogenicity prediction feature for novel variation(s) using K_a/K_s aspects to next PON-P version.

7. CONCLUSION

The overall objective of this study was to analyze selective evolutionary pressure for human missense variations of both pathogenic and neutral types. This task was performed by observing overall distribution of K_a/K_s values and related aspects in individual amino acid level as well as for different amino acid groups. Importance was also given to find out amino acid characteristics and pathogenicity aspects.

Pathogenic variations have been found to have lower K_a/K_s values in average whereas the neutral type variations have scored higher K_a/K_s values. This different K_a/K_s value distribution difference has revealed the fact that pathogenic variations are more conserved than that of neutral variations. This conservation property of pathogenic variations has demonstrated same feature in case of separate studies both for individual amino acids and for different amino acid group level.

In addition, amino acid variability studies showed some interesting aspects for specific amino acids and different groups of amino acids. Arginine (R) has been found most abundant to be mutated irrespective of type, whereas lysine (K) was opposite in this point of view. In addition, cysteine (C) has been found most frequent to be converted into a pathogenic type variation followed by glycine (G) and tryptophan (W). However, isoleucine (I), asparagine (N), and threonine (T) were least in this consideration. Hence, threonine (T) has been found less frequent to become a pathogenic variation.

Polar basic amino acids have been found to be changed into both pathogenic and neutral type variants. In contrast, aromatic amino acids have shown two times more pathogenicity than polar acidic amino acids.

8. REFERENCES

- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS and Sunyaev SR (2010). A method and server for predicting damaging missense mutations. *Nat Methods* **7**: 248–449
- Ambrogelly A, Palioura S and Söll D (2007). Natural expansion of the genetic code. *Nat Chem Biol* **3**: 29–35
- Aminetzach YT, Macpherson JM and Petrov DA (2005). Pesticide resistance via transposition-mediated adaptive gene truncation in *Drosophila*. *Science* **309**: 764–767
- Bagdonavicius V, Kruopis J and Nikulin MS (2011). Non-parametric tests for complete data", *Iste & Wiley: London & Hoboken*
- Bamshad M and Wooding SP (2003). Signatures of natural selection in the human genome. *Nat. Rev. Genet* **4**: 99–111
- Bao L, Zhou M and Cui Y (2005). nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res* **33**: 480–482
- Barreiro LB, Laval G, Quach H, Patin E and Quintana-Murci L (2008). Natural selection has driven population differentiation in modern humans. *Nature Genet* **40**: 340–345
- Bell G (1997). *Selection: The Mechanism of Evolution*
- Bersaglieri T, *et al.* (2004). Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**: 1111–1120
- Bertram J (2000). The molecular biology of cancer. *Mol. Aspects Med.* **21**: 167–223
- Boes DC, Graybill FA and Mood AM (1974). *Introduction to the Theory of Statistics, 3rd ed.* New York: McGraw-Hill
- Bromberg Y and Rost B (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **35**: 3823–3835
- Burch CL and Chao L (1999). Evolution by small steps and rugged landscapes in the RNA virus phi6. *Genetics* **151**: 921–927
- Burrus V and Waldor M (2004). Shaping bacterial genomes with integrative and conjugative elements. *Res. Microbiol.* **155**: 376–86
- Bryk J, *et al.* (2008). Positive selection in East Asians for an *EDAR* allele that enhances NF-kappaB activation. *PLoS ONE* **3**: e2209

- Calabrese R, Capriotti E, Fariselli P, Martelli PL and Casadio R (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* **30**:1237–1244
- Capriotti E, Calabrese R and Casadio R (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**: 2729–2734
- Charlesworth, D *et al.* (1995). The pattern of neutral molecular variation under the background selection model. *Genetics* **141**: 1619–1632
- Charlie L and Tonya LK. *Chi-Squared Data Fitting*. University California, Davis (http://neutrons.ornl.gov/workshops/sns_hfir_users/posters/Laub_Chi-Square_Data_Fitting.pdf)
- Comeron JM (1995). A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* **41**: 1152–1159
- Corder GW and Foreman, DI (2009). *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Wiley
- Darwin C, (1872). *The Origin of Species*
- DeGroot MH (1991). *Probability and Statistics, 3rd ed.* Reading, MA: Addison-Wesley
- Dempster AP, Laird NM and Rubin DB (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion) . *J. Roy. Stat. Soc. Ser. B.*, **39**: 1–38
- De Magalhães JP and Church GM (2007). Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. *Mech Ageing Dev*, **128**: 355-364
- Doron-Faigenboim A, Stern A, Mayrose I, Bacharach E and Pupko T (2005). Selecton: a server for detecting evolutionary forces at a single amino-acid site . *Bioinformatics*, **21**: 2101-2103
- Eadie WT, Drijard J, Roos and Sadoulet (1971). *Statistical Methods in Experimental Physics*. Amsterdam: North-Holland. 269–271
- Endo T, Ikeo and Gojobori (1996). Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13**: 685–690
- Eric JV and Bruce TL (2004). Positive selection on the human genome. *Hum. Mol. Genet.* **13**: 245-254

Evolution and Natural Selection notes from University of Michigan (2010); downloaded from

<http://www.globalchange.umich.edu/globalchange1/current/lectures/selection/selection.html>.

Felsenstein J (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**: 368-376

Fitch WM, Bush B and Cox (1997). Long term trends in the evolution of H(3) HA1 human influenza type A . *Proc. Natl. Acad. Sci.* **94**: 7712–7718

Flexner C (1998). HIV-protease inhibitors . *N. Engl. J. Med.* **338**: 1281–1292

Flicek P, Aken BL, Ballester B, Beal , Bragin E, Brent , Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Gräf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, Megy K, Overduin B, Pritchard B, Rios D, Ruffier M, Schuster M, Slater G, Smedley D, Spudich G, Tang YA, Trevanion S, Vilella A, Vogel J, White S, Wilder SP, Zadissa A, Birney E, Cunningham F, Dunham I, Durbin R, Fernández-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Smith J and Searle SM (2010). Ensembl's 10th year. *Nucleic Acids Res.* **38**: D557-D562

Forsdyke DR (2006). *Evolutionary Bioinformatics*. Springer, New York

Frank SA (2012). Natural selection. III. Selection versus transmission and the levels of selection. *J. Evol. Biol.* **25**: 227-243

Gibbons DJ, Chakraborti S (2003). *Nonparametric Statistical Inference*, 4th Ed. CRC

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E and Ben-Tal N (2003). ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**: 163–164

Goldman N and Yang Z (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**: 725–736

Gu X and Vander Velden K (2002). DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. *Bioinformatics*, **18**: 500–501

Hartl DL (1981). A Primer of Population Genetics. *AJMG* **17**: 869

Hasegawa, M, et al. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA . *J. Mol. Evol.* **22**: 160-174

- Hettmansperger TP and McKean JW (1998). *Robust nonparametric statistical methods. Kendall's Library of Statistics. 5* (First ed.). London: Edward Arnold
- Hughes AL and Nei (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**: 167–170
- Hughes AL (1999). *Adaptive Evolution of Genes and Genomes*. Oxford University Press, New York
- Ina Y (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* **40**: 190–226
- John RT (1997). *An introduction to error analysis*. University Science Books
- Jukes TH and Cantor CR (1969). *Evolution of protein molecules. In Mammalian Protein Metabolism* (ed. Munro, H.N.), pp. 21-123. Academic Press, New York, USA
- Khan S and Vihinen M (2010). Performance of protein stability predictors. *Hum Mutat* **31**: 675–684
- Kimura M (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**: 111-120
- Kreitman M and Akashi H (1995). Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* **26**: 403–422
- Kreitman M (2000). Methods to detect selection in populations with applications to the human. *Annu. Rev. Genomics Hum. Genet.* **1**: 539–559
- Kryazhimskiy S and Plotkin JB (2008). The Population Genetics of dN/dS. *PLoS Genet* **4**: e1000304
- Kyte J and Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**: 105–132
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**: 2947-2948
- Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD and Radivojac P (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**: 2744–2750
- Liaw A and Wiener M (2002). Classification and regression by randomForest. *R News* **2**: 18–22

- Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipurksy SL and Darnell J (2004). *Mol Cell Bio* (5th ed.). New York, New York: WH Freeman and Company
- Loewe L (2008). Negative selection. *Nature Education* **1**(1);
<http://www.nature.com/scitable/topicpage/Negative-Selection-1136>
- Maynard-Smith J (1989). *Evolutionary Genetics*. Oxford University Press, Oxford
- Mayrose I, Graur D, Ben-Tal N and Pupko T (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.*, **21**: 1781–1791
- Meierhenrich and Uwe J (2008). *Amino acids and the asymmetry of life* (1st ed.). Springer
- Mikita S, David T and Peer B (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**: 609-612
- Miyata T and 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 applications. *J. Mol. Evol.* **16**: 23–36
- Mount DW (2004). *Bioinformatics: Sequence and Genome Analysis* (2nd ed.)
- Muse SV and Gaut BS (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**: 715-724
- Muse SV (1996). Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* **13**: 105–114
- Nair PS and Vihinen M (2012). VariBench: A benchmark database for variations. *Hum Mutat.* 2012
- Nei M and Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Mol. Biol. Evol.* **3**: 418–426
- Nei M and Kumar S (2000). *Molecular Evolution and Phylogenetics*, Oxford University Press
- Nelson DL and Cox MM (2005). *Lehninger's Principles of Biochemistry* (4th ed.). New York, New York: W. H. Freeman and Company
- Ng PC and Henikoff S (2001). Predicting deleterious amino acid substitutions. *Genome Res* **11**: 863–874
- Nielsen R and Yang (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936

- Nielsen R (2002). Mapping variations on phylogenies. *Syst. Biol.* **51**: 729–739
- Nielsen R and Huelsenbeck (2002). Detecting positively selected amino acid sites using posterior predictive p-values. *Pac. Symp. Biocomput.* **7**: 576–588
- Olatubosun A, Väliäho J, Härkönen J, Thusberg J, and Vihinen M (2012). PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat.* **33**:1166-1174
- Page R and Holmes E (1998). *Molecular Evolution: A Phylogenetic Approach*. Oxford: Blackwell Science
- Pamela CC, Richard AH and Denise RF (2004). *Lippincott's Illustrated Reviews: Biochemistry*. Lippincott Williams & Wilkins
- Peng C, Ho BK, Chang TW and Chang NT (1989). Role of human immunodeficiency virus type 1-specific protease in core protein maturation and viral infectivity. *J. Virol.* **63**: 2550–2556
- Ramensky V, Bork P and Sunyaev S (2002). Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**: 3894–3900
- Ridley M (2004). *Evolution*. 2nd edition, Oxford University Press
- R Development Core Team (2011). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL <http://www.R-project.org/>
- Sawyer SA, Parsch J, Zhang Z and Hartl DL (2007). Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.* **104**: 6504–6510
- Schaffner S and Sabeti P (2008). Evolutionary adaptation in the human lineage. *Nature Education* **1**(1)
- Sergei LKP and Simon DWF (2005). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics* **21**: 2531-2533
- Stephens MA (1979). Test of fit for the logistic distribution based on the empirical distribution function, *Biometrika*, **66**, 591-5
- Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS and Cooper DN (2009). The Human Gene Variation Database: 2008 update. *Genome medicine* **1**: 13
- Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E and Pupko T (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Research*. **35**: 506-511

- Stuart A, Ord K and Arnold SF (1999). *Classical Inference and the Linear Model*. Kendall's Advanced Theory of Statistics. **2A** (Sixth ed.). London: Arnold. pp. 25.37–43
- Suzuki Y and Gojobori T (1999). A method for detecting positive selection at single amino acid sites. *Mol Biol Evol.* **16**: 1315-28
- Suzuki Y (2004). New methods for detecting positive selection at single amino acid sites. *J. Mol. Evol.* **59**: 11–19
- Tamura K and 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**: 512-526
- Tavare S (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**: 57-86
- Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A and Narechania A (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**:2129–2141
- Thusberg J and Vihinen M (2009). Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *HumMutat* **30**: 703–714
- Thusberg J, Olatubosun A and Vihinen M (2011). Performance of variation pathogenicity prediction methods on missense variants. *HumMutat.* **32**: 358-368
- Tsunoyama K and Gojobori (1998). Evolution of nicotinic acetylcholine receptor subunits. *Mol. Biol. Evol.* **15**: 518–527
- Varela MA and Amos W (2010). Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence. *Genomics* **95**: 151–159
- Wasserman L (2007). *All of nonparametric statistics*, Springer
- Wayne ML and Simonsen KL (1998). Statistical tests of neutrality in the age of weak selection. *Trends Ecol. Evol.* **13**: 236–240
- Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**: 555–556
- Yang Z (2000). Adaptive molecular evolution. In *Handbook of Statistical Genetics* (Balding, D. et al., eds), Ch. 12, Wiley
- Yang ZH and Bielawski JP (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**: 496–503

- Yang Z and Nielsen R (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**: 32–43
- Yang Z, Nielsen R, Goldman N and Pedersen AM (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, **155**: 431–449
- Yang Z (2002). Inference of selection from multiple species alignments. *Curr. Opin. Genet. Develop.*, **12**: 688–694
- Zhang Z, Li J, Zhao XQ, Wang J, Wong GK and Yu J (2006). K_a K_s Calculator: Calculating K_a and K_s through model selection and model averaging. *Genomics Proteomics Bioinformatics* **4**: 259-263
- Zhang J (2008). Positive selection, not negative selection, in the pseudogenization of *rcaA* in *Yersinia pestis*. *Proc Natl Acad Sci.* 105: E69
- Zharkikh A (1994). Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39**: 315-329

9. APPENDIX

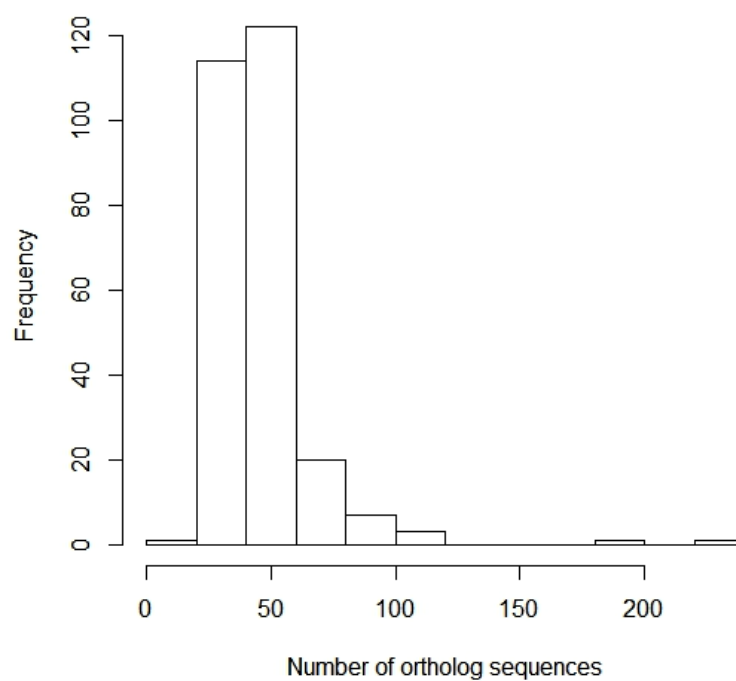


Figure 9.1: Histogram of number of ortholog sequences in individual gene and protein sequence files. The ortholog sequence numbers are total number of orthologs downloaded from Ensembl for each gene.