12-2016

TESTING THE INDEPENDENCE HYPOTHESIS OF ACCEPTED MUTATIONS FOR PAIRS OF ADJACENT AMINO ACIDS IN PROTEIN SEQUENCES

Jyotsna Ramanan
University of Nebraska-Lincoln, jor.compscie@gmail.com

Ramanan, Jyotsna, "TESTING THE INDEPENDENCE HYPOTHESIS OF ACCEPTED MUTATIONS FOR PAIRS OF ADJACENT AMINO ACIDS IN PROTEIN SEQUENCES" (2016). *Computer Science and Engineering: Theses, Dissertations, and Student Research*. 118.
http://digitalcommons.unl.edu/computerscidiss/118

TESTING THE INDEPENDENCE HYPOTHESIS OF ACCEPTED MUTATIONS FOR PAIRS OF ADJACENT AMINO ACIDS IN PROTEIN SEQUENCES

by

Jyotsna Ramanan

A THESIS

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfilment of Requirements
For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Peter Z. Revesz

Lincoln, Nebraska

December, 2016

TESTING THE INDEPENDENCE HYPOTHESIS OF ACCEPTED MUTATIONS FOR PAIRS OF ADJACENT AMINO ACIDS IN PROTEIN SEQUENCES

Jyotsna Ramanan, MS
University of Nebraska, 2016

Adviser: Peter Z. Revesz

Evolutionary studies usually assume that the genetic mutations are independent of each other. However, that does not imply that the observed mutations are independent of each other because it is possible that when a nucleotide is mutated, then it may be biologically beneficial if an adjacent nucleotide mutates too.

With a number of decoded genes currently available in various genome libraries and online databases, it is now possible to have a large-scale computer-based study to test whether the independence assumption holds for pairs of adjacent amino acids. Hence the independence question also arises for pairs of adjacent amino acids within proteins. The independence question can be tested by considering the evolution of proteins within a closely related sets of proteins, which are called protein families.

In this thesis, we test the independence hypothesis for three protein families from the PFAM library, which is a publicly available online database that records a growing number of protein families. For each protein family, we construct a hypothetical common ancestor, or consensus sequence. We compare the hypothetical common ancestor of a protein family with each of the descendant protein sequences in the family to test where the mutations occurred during evolution. The comparison yields actual probabilities for each pair of amino acids changing into another pair of amino acids. By comparing the actual probabilities with the theoretical probabilities under the independence assumption, we identify anomalies that indicate that the independence

assumption does not hold for many pairs of amino acids.

DEDICATION

In loving memory of my father.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartfelt gratitude and appreciation to my adviser Dr. Peter Z. Revesz for all the guidance and immense motivation that he endowed upon me during several phases of my research. I am obliged to thank him for his humility and patience during my highs and lows all along this research.

I am also grateful to Dr. Juan Cui and Dr. Stephen Scott for investing their time to be a part of my examination comittee and for sharing constructive opinion about my work which can be thought-through as I plan to improvize my work in the future. I am thankful to Dr. Stephen D. Kachman and his students from the Department of Statistics at University of Nebraska- Lincoln for their pertinent help in this research.

I would like to use this opportunity to thank my loving mother, for her affection, encouragement and for taking a stand all these years for me to attend this stature. I am gratified to my grandparents for their wishes and prayers. Lastly, I wish to show appreciation to my friends for being there for me and for cordially lending their helping hands when I needed the most.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Biological evolution depends on random mutations accompanied by natural selection for the more fit genes. That simple statement does not imply that the observed mutations are independent from each other. It is possible that if a nucleotide changes, then it is biologically beneficial to have some of the adjacent or nearby nucleotides change as well. For example, if in some protein-coding region within some triplet that encodes a hydrophilic amino acid, a nucleotide changes such that the triplet would encode a hydrophobic amino acid, then a mutation of another nucleotide in the same triplet may be advantageous if with that mutation the triplet would again encode a hydrophilic amino acid (or preserve another key property of amino acids). In other words, some mutations within a triplet slightly increase the probability that some accompanying mutation with a readjusting effect would survive in the offspring.

1.2 Problem Statement

With the greatly increasing number of decoded genes currently available in a number of genome libraries and online databases, it is now possible to have a large-scale computer-based study to test whether the independence assumption holds. One difficulty, however, is to find the coding regions and coding triplets. Hence it seems more convenient to investigate proteins derived from the coding regions. The mutations in the coding regions of the DNA are usually reflected in the mutations of amino acids. Therefore, instead of the evolution of genes, one may talk about the evolution of proteins within a closely related set of proteins, which is called a protein family.

1.3 Objective

The PFAM library [2] records a growing number of protein families. Each protein in a protein family can be assumed to be genetically related to the other proteins in that family and to have evolved from a single ancestor protein. For any set of DNA strings and any set of proteins, there are several algorithms that can be used to find a hypothetical evolutionary tree [3] and [17]. Revesz [16] has proposed recently a new phylogenetic tree-building algorithm called the Common Mutation Similarity Matrixes (CMSM) algorithm. The first step of the CMSM algorithm is to find a hypothetical common ancestor, which is denoted by μ. In this research, we will use the idea of a hypothetical common ancestor. We can compare the hypothetical common ancestor of a family of proteins with each of the proteins in the family to test where the mutations occur. We also can test for each adjacent pair of amino acids how many times that pair changed into another pair of amino acids. The resulting experimental statistics can be compared with the theoretical probability under the independence assumption. If the deviation from the theoretical probability is significant, then the independence assumption fails to provide a satisfying explanation for the experimental results.

1.4 Contribution

As a part of the research, we have developed an efficient technique that could be used to test the independence hypothesis for pairwise mutations in a set of protein sequences that belong to a family. For each Protein family that we have considered for the experiments for this thesis, we have devised the following:

- Hypothetical Common Ancestor for the protein families. Constructing the hypothetical common ancestor for protein families are explained in detail in Chapter 2. The hypothetical common ancestor is also called the consensus sequence which is mostly the first sequence of the protein family in thesis. Also note that the terms 'hypothetical common ancestor' and 'consensus sequence' are used interchangeable throughout the thesis.

- The Mutation Probability Matrices for individual protein families showing the actual mutations for every single amino acid in each of the protein families were calculated. This matrix is of size 20 x 20 showing all the actual probabilities of one amino acid in the consensus sequence mutating into another amino acid in its descendent sequences. This mutation probability matrix could also be considered similar to the PAM 250 scoring matrix, which is explained in Chapter 2 in detail.

- Based on the mutation probability matrices that stores the mutations of a single amino acid in an individual protein sequence mutating into another amino acid in its descendant sequence, we calculated the theoretical probabilities that shows all possible pairwise mutations of amino acids in the protein sequences. The size of the matrix that shows the theoretical mutation probabilities is 400 x 400 since the pairs are the possible combinations of all the 20 amino acids that exists in nature. The total number of elements in this matrix is about 160,000. The detailed explanation on calculating theoretical probabilities are described in Chapter 4.

- For every set of sequences of the protein family, we calculated the actual probability of mutations of every adjacent amino acid pairs in the consensus sequence mutating into another pair in the following descendant sequences. The frequencies and the indices of the occurrence of all the adjacent pairs in the consensus sequences are found, and then we check those pairs in the consensus sequence, we check for the mutations in the descendant sequences in the corresponding window of the column. The mechanism of calculating the actual pairwise mutation probabilities for adjacent amino acids of the consensus sequences are explained in detail in Chapter 4.

- The percentage probability differences between the theoretical pairwise probabilities and actual pairwise probabilities for the corresponding top 30 pairs in each of the individual protein families are considered for analysis and test the independence hypothesis. Used these results to analyze and infer the independence hypothesis that is currently the subject of this thesis.

- A part of this research of testing the independence hypothesis for pairwise adjacent amino acids of a protein sequence has been presented in INASE Conference, during the academic year October ’15 and successively published in the proceedings [14].

1.5 Outline of Thesis

The thesis is outlined in the following manner:

Chapter 1 briefly introduces the idea of the thesis such as the problem statement, objectives, the strategies that will be used in the future chapters and contributions of this research. Chapter 2 reveals the related work and some popular background concepts that this research topic was developed on. Chapter 3 explains in detail about the large datasets which in this case are three protein families that were downloaded from the PFAM Library. The sections introduce the aligned sequences of the protein family and a brief summary of description of the protein families. Chapter 4 demonstrates the independence testing method, which is the prime intent of this research. In Chapter 5 presents the experimental results that were attained as the outcome of our methodology in the previous chapter. Some of the inferences are showcased based on the final results with bar charts for improving readability. Chapter 6 analyzes the inference and summarizes the conclusion and possible future enhancements.

Chapter 2

Background Concepts and Related Work

2.1 Fundamentals of Biology

In biology, amino acids are organic compounds composed of the functional groups amine and carboxylic acid, with a specific side chain. The key elements of an amino acid are carbon, hydrogen and nitrogen. So far, about five hundred amino acids has been identified. These amino acids are classified according to the structural functions and properties like – polar, charged, aliphatic, aromatic, hydrophilic and hydrophobic. The amino acids are classified based on its properties. Basically, there are twenty basic essential amino acids into existence. Table 2.1 shows the twenty different amino acids under respective classification.

Deoxyribonucleic acid or the DNA is considered the blueprint of all living organisms [15]. The DNA encodes the genetic material composed of the four main nucleotides that are:

Adenine (A)
Thymine (T)
Cytosine (C)
Guanine (G)

These nucleotides form long strands using peptide bonds. The structure of a DNA is double stranded and helical where the chain of nucleotides run through these strands [11].

Table 2.1: Classification of Twenty Amino Acids

Charged	Polar	Hydrophobic
Arginine (R)	Glutamine (Q)	Alanine (A)
Lysine (K)	Asparagine (N)	Isoleucine (I)
Aspartic Acid (D)	Histidine (H)	Leucine (L)
Glutamic Acid (E)	Serine (S)	Phenylalanine (F)
	Threonine (T)	Valine (V)
	Tyrosine (Y)	Proline (P)
	Cysteine (C)	Glycine (G)
	Tryptophan (W)	

The DNA contains coding regions that stores information about the proteins. Proteins are composed of a sequence of amino acids (Revesz, Introduction to Databases: From Biological to Spacio-Temporal, 2010). The sequences of nucleotides are translated into a sequence of amino acids using a genetic code. The translation of nucleotides into amino acids are carried out using triplets of nucleotides called codons. These sequences are then aligned using some tools online so that the protein sequences could be used for various testing. In the protein sequences, mutations occur during the process of DNA replication when errors occur in the polymerization of the DNA strand. These errors could possible affect the phenotype of the organism, if they occur within the protein code sequence of a gene. It is implied that mutations are rare events as error rates are usually very low.

2.2 Phylogenic Trees

Phylogenic trees or evolutionary trees are used to show the relationship among the genes and organisms [17]. There are several types of diagrams that are into existence to depict these kinds of relationships. Phylogenic trees could be of two types – rooted or unrooted. Since these resemble the structure of a tree, the terms referring to various parts of these diagrams are also similar to that of a tree. Biologists are often interested in the time of common origin of a group or a taxon [12]. Some of the phylogenetic tree analyses lets us to calculate the most recent common ancestor for all the genes.

Phylogenic trees can also be called as gene trees since the show the evolutionary history of a gene or a set of DNA sequence. The relationships between ancestor and descendants could be represented using phylogram, where the branch length represents the evolutionary distances between a group of genes [22].

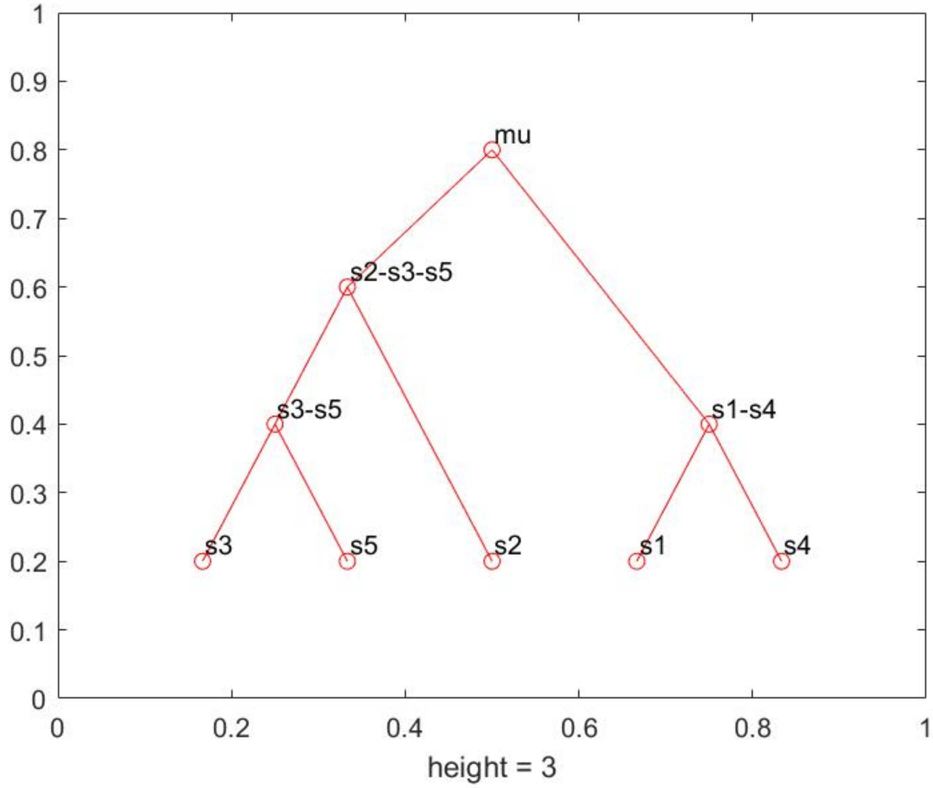


Figure 2.1: A Phylogenic Tree

2.3 Constructing the Hypothetical Common Ancestor

As can be seen from the sections above, which explains about the phylogenic trees, it is understood that a phylogenic tree has a common ancestor. There are several ways to calculate this common ancestor. The reconstruction of the original sequence in a protein family is made harder by the fact that different branches of the evolutionary tree evolve by different rates of mutations. Shortridge et al.[18] study the different rates of mutations in various bacterial phyla. For this thesis, we use the idea of hypothetical common ancestor (µ) which is mentioned by Revesz [16] in a paper that talks about constructing an evolutionary tree based on the number of common mutations happening in a set of sequences (CMSM).

Suppose there are seven DNA sequences that are related, we can find the hypothetical common ancestor (µ) as the mode of each column. If there is no most frequent nucleotide in a column, then we arbitrarily choose one of the most frequent nucleotides in the sequence. We can think that in each sequence Si, the nucleotides that do not match the corresponding nucleotide in µ indicates to have undergone mutation at some point during evolution. The more common mutations two sequences share, the closer they are like to appear in the evolutionary tree. The hypothetical common ancestor µ is also referred as the consensus sequence at some places in this thesis. Further demonstration of calculating the common ancestor µ are shown in Chapter 4 when we talk about the independence testing method.

2.4 Sequence Similarity Matrix

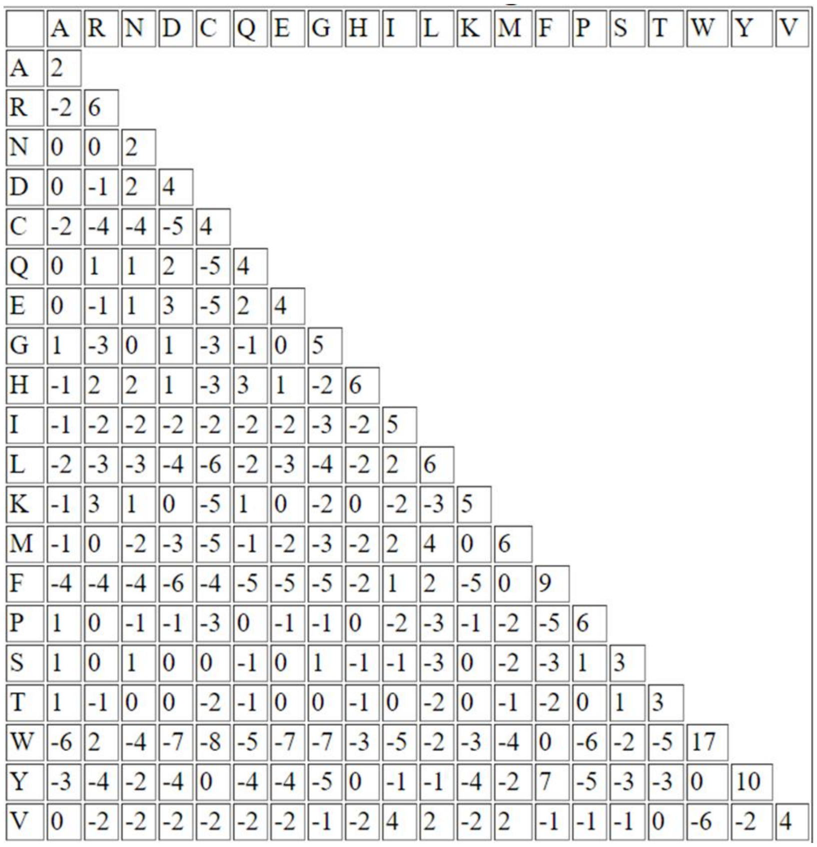
Sequences are aligned using one of the techniques like BLAST [8] or FASTA [13, 6] before they could be used for any experiment. The sequences are assigned with similarity scores after alignment. The score of an alignment is the sum of the scores for each position in the alignment [19]. This is an example of dynamic programming paradigm, as we need to find the highest scoring alignment.

2.4.1 PAM 250 Matrix

The most commonly used scoring matrix is the PAM matrix which records the scores for the mutations that occur in a sequence.

PAM – Point Accepted Mutation.

The term “accepted” denotes that a particular sequence has accepted that mutation has been embraced by one of the amino acids. PAM 250 means that about 250 mutations has occurred per 100 amino acids [23]. PAM matrices comprise of both positive and negative values. If the alignment score is greater than zero, then the sequences are considered to be related. If the scores are negative, then it means that the sequences are not related. Hence these scores represent the relationship between the sequences of a protein family. The PAM 250 scoring matrix obtained from the website mentioned earlier, is shown in Figure 2.4.1.1 below.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

Figure 2.2: PAM 250 Scoring Matrix

Chapter 3

Data Source

3.1 Protein Families

For methods for testing the independence hypothesis which we will see in the future chapters, were also conducted on real world datasets that contains about more than a hundred of sequences for each family. The sequences for each protein family were obtained from the PFAM library [2]. The sequences were aligned using FASTA sequencing algorithm. Note that the independence hypothesis of pairwise mutations were tested on seed sequences rather than full sequences as the number of proteins in the seed sequences remain the same at all times wherease the number of full sequences tend to vary as there could be additions of protein sequences according to the mutations that may take place with time. The list of the three protein families used in this research for testing the independence hypothesis are the following:

- DAGK_cat (PF00781)
- IL17 (PF06083)
- KA1 (PF02149)

The experimental results showing the theoretical probabilities and actual probabilities are mentioned in later chapters under Experimental Results and Discussion.

3.2 Description of Protein Families

3.2.1 DAGK_cat (PF00781)

The protein family used here to test the method on large data set is the Diacylglycerol kinase catalytic domain (DAGK_cat) whose sequences can be referred from the PFAM Library. This domain consists of 31217 sequences, out of which 110 seed sequences were used for the experiment in this paper. The common mutation ancestor µ was calculated to be:

KALVIVNPKSGTARGGKGKKLLERKVRPLLEEAGVSDDELDLRLTENPGPGDVLRRGYGNLEKLKSNAL

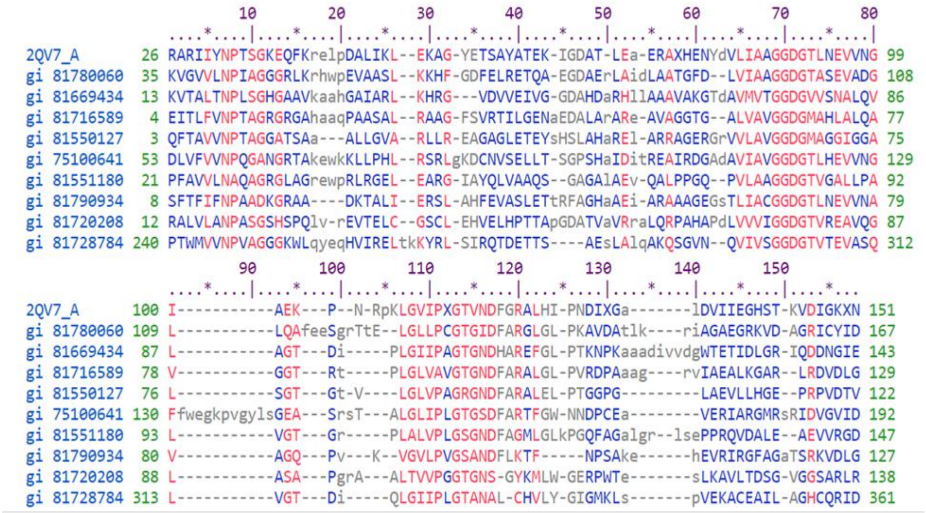
ELLAGAAREAAEANEQSDGDTLLPWSENLAYGYCPDLIVAAGGDGTVNEVLNGLAGNARRDDLELATRN

HPRAVLVPSSPPLGIIPLGRTGNDFARALNAHGGFEEGIPLGYDPEEAARAALELIKKIKGQTRPVDVGKV

In chemistry, Diacylglycerol kinase (DGK or DAGK) is a family of enzymes that catalyzes the conversion of diacylglycerol (DAG) to phosphatidic acid (PA) utilizing ATP as a source of the phosphate [10].

Protein Sequences

As can be seen in Figure 3, some parts of the sequences of the protein family DAGK_cat (PF 00781) are shown in intervals of 10 sequences per row with types of nucleotides those are diverse among the members themselves. These sequences are generated in Hypertext format using the tool provided by the NCBI and it is accessible publicly online at the official NCBI website [5].



```
                      10        20        30        40        50        60        70        80
              ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
2QV7_A       26 RARIIYNPTSGKEQFKrelpDALIKL--EKAG-YETSAYATEK-IGDAT-LEa-ERAXHENYdVLIAAGGDGTLNEVVNG 99
gi 81780060  35 KVGVVLNPIAGGGRLKrhwpEVAASL--KKHF-GDFELRETQA-EGDAErLAidLAATGFD--LVIAAGGDGTASEVADG 108
gi 81669434  13 KVTALTNPLSGHGAAVkaahGAIARL--KHRG---VDVVEIVG-GDAHDaRHllAAAVAKGTdAVMVTGGDGVVSNALQV 86
gi 81716589   4 EITLFVNPTAGRGRGAhaaqPAASAL--RAAG-FSVRTILGENaEDALArARe-AVAGGTG--ALVAVGGDGMAHLALQA 77
gi 81550127   3 QFTAVVNPTAGGATSAa---ALLGVA--RLLR-EAGAGLETEYsHSLAHaREl-ARRAGERGrVVLAVGGDGMAGGIGGA 75
gi 75100641  53 DLVFVVNPQGANGRTAkewkKLLPHL--RSRLgKDCNVSELLT-SGPSHaIDitREAIRDGAdAVIAVGGDGTLHEVVNG 129
gi 81551180  21 PFAVVLNAQAGRGLAGrewpRLRGEL--EARG-IAYQLVAAQS--GAGAlAEv-QALPPGQ--PVLAAGGDGTVGALLPA 92
gi 81790934   8 SFTFIFNPAADKGRAA----DKTALI--ERSL-AHFEVASLETtRFAGHaAEi-ARAAAGEGsTLIACGGDGTLNEVVNA 79
gi 81720208  12 RALVLANPASGSHSPQlv-rEVTELC--GSCL-EHVELHPTTApGDATVaVRraLQRPAHAPdLVVVIGGDGTVREAVQG 87
gi 81728784 240 PTWMVVNPVAGGGKWLqyeqHVIRELtkKYRL-SIRQTDETTS----AEsLAlqAKQSGVN--QVIVSGGDGTVTEVASQ 312
                      90       100       110       120       130       140       150
              ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*...
2QV7_A      100 I-----------AEK---P--N-RpKLGVIPXGTVNDFGRALHI-PNDIXGa-------lDVIIEGHST-KVDIGKXN 151
gi 81780060 109 L-----------LQAfeeSgrTtE--LGLLPCGTGIDFARGLGL-PKAVDAtlk----riAGAEGRKVD-AGRICYID 167
gi 81669434  87 L-----------AGT---Di-----PLGIIPAGTGNDHAREFGL-PTKNPKaaadivvdgWTETIDLGR-IQDDNGIE 143
gi 81716589  78 V-----------GGT---Rt-----PLGLVAVGTGNDFARALGL-PVRDPAaag----rvIAEALKGAR--LRDVDLG 129
gi 81550127  76 L-----------SGT---Gt-V----LGLVPAGRGNDFARALEL-PTGGPG---------LAEVLLHGE--PRPVDTV 122
gi 75100641 130 FfwegkpvgylsGEA---SrsT---ALGLIPLGTGSDFARTFGW-NNDPCEa--------VERIARGMRsRIDVGVID 192
gi 81551180  93 L-----------VGT---Gr-----PLALVPLGSGNDFAGMLGLkPGQFAGalgr--lsePPRQVDALE--AEVVRGD 147
gi 81790934  80 V-----------AGQ---Pv---K--VGVLPVGSANDFLKTF-----NPSAke------hEVRIRGFAGaTSRKVDLG 127
gi 81720208  88 L-----------ASA---PgrA---ALTVVPGGTGNS-GYKMLW-GERPWTe-------sLKAVLTDSG-VGGSARLR 138
gi 81728784 313 L-----------VGT---Di-----QLGIIPLGTANAL-CHVLY-GIGMKLs-------pVEKACEAIL-AGHCQRID 361
```

Figure 3.1: Highlighting a part of the aligned sequences of the protein family DAGK_cat

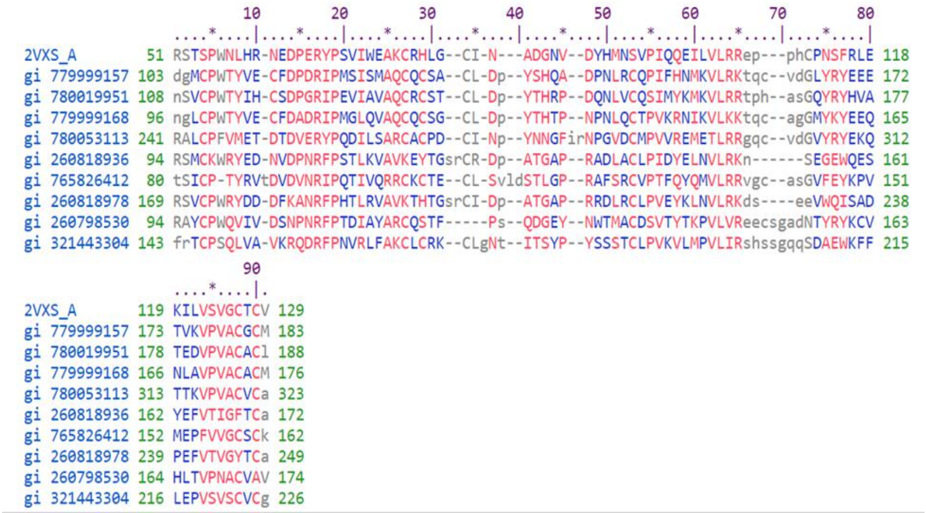
3.2.2 IL 17 (PF06083)

The second protein family used here to test the method on large data set is Interleukin (IL 17) whose sequences can be referred from the PFAM Library. This family consists of 531 sequences in total, where around 102 sequences were used for the experiment discussed in this paper. The common mutation ancestor µ was calculated to be:

RSLSPWDYREIDPHDPNRYPRVIAEARCLLCSGGSRCIGDLNPATGQGEDDIAELQGLRRSLNSVPIYQE ILVAFLDGGGKLRRLCDKPCSRPKTHEPCAGCRYSYRLEPVKETVTVGCTV

Protein Sequences

As can be seen in Figure 4, some parts of the sequences of the protein family Interleukin 17 (PF 06083) are shown in intervals of 10 sequences per row with types of nucleotides those are diverse among the members themselves. These sequences are generated in Hypertext format using the tool provided by the NCBI and it is accessible publicly online at the official NCBI website [10].



```
                        10        20        30        40        50        60        70        80
                 ....*....|....*....|....*....|....*....|....*....|....*....|....*....|....*....|
2VXS_A        51 RSTSPWNLHR-NEDPERYPSVIWEAKCRHLG--CI-N---ADGNV--DYHMNSVPIQQEILVLRRep---phCPNSFRLE 118
gi 779999157 103 dgMCPWTYVE-CFDPDRIPMSISMAQCQCSA--CL-Dp--YSHQA--DPNLRCQPIFHNMKVLRKtqc--vdGLYRYEEE 172
gi 780019951 108 nSVCPWTYIH-CSDPGRIPEVIAVAQCRCST--CL-Dp--YTHRP--DQNLVCQSIMYKMKVLRRtph--asGQYRYHVA 177
gi 779999168  96 ngLCPWTYVE-CFDADRIPMGLQVAQCQCSG--CL-Dp--YTHTP--NPNLQCTPVKRNIKVLKKtqc--agGMYKYEEQ 165
gi 780053113 241 RALCPFVMET-DTDVERYPQDILSARCACPD--CI-Np--YNNGFirNPGVDCMPVVREMETLRRgqc--vdGVYRYEKQ 312
gi 260818936  94 RSMCKWRYED-NVDPNRFPSTLKVAVKEYTGsrCR-Dp--ATGAP--RADLACLPIDYELNVLRKn------SEGEWQES 161
gi 765826412  80 tSICP-TYRVtDVDVNRIPQTIVQRRCKCTE--CL-SvldSTLGP--RAFSRCVPTFQYQMVLRRvgc--asGVFEYKPV 151
gi 260818978 169 RSVCPWRYDD-DFKANRFPHTLRVAVKTHTGsrCI-Dp--ATGAP--RRDLRCLPVEYKLNVLRKds----eeVWQISAD 238
gi 260798530  94 RAYCPWQVIV-DSNPNRFPTDIAYARCQSTF-----Ps--QDGEY--NWTMACDSVTYTKPVLVReecsgadNTYRYKCV 163
gi 321443304 143 frTCPSQLVA-VKRQDRFPNVRLFAKCLCRK--CLgNt--ITSYP--YSSSTCLPVKVLMPVLIRshssgqqSDAEWKFF 215
                        90
                 ....*....|.
2VXS_A       119 KILVSVGCTCV 129
gi 779999157 173 TVKVPVACGCM 183
gi 780019951 178 TEDVPVACACl 188
gi 779999168 166 NLAVPVACACM 176
gi 780053113 313 TTKVPVACVCa 323
gi 260818936 162 YEFVTIGFTCa 172
gi 765826412 152 MEPFVVGCSCk 162
gi 260818978 239 PEFVTVGYTCa 249
gi 260798530 164 HLTVPNACVAV 174
gi 321443304 216 LEPVSVSCVCg 226
```

Figure 3.2: Highlighting a part of the aligned sequences of the protein family IL 17

3.2.3 KA 1 (PF02149)

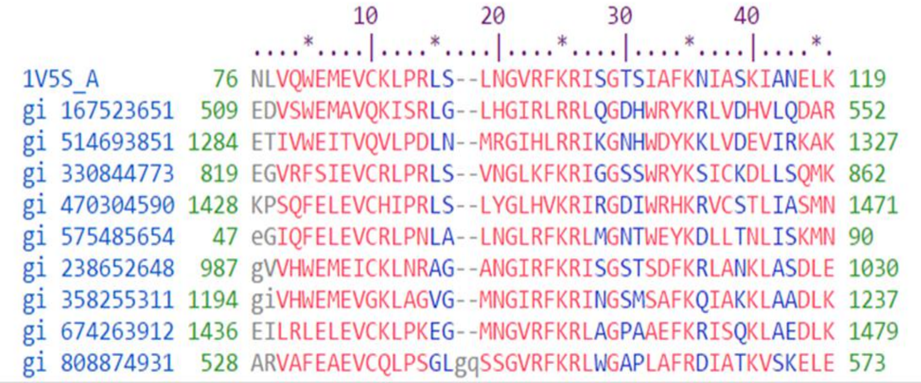
The third protein family used here to test the method on large data set is the Kinase Domain (KA 1) whose sequences can be referred from the PFAM Library. This family consists of 1349 sequences in total, where around 105 sequences were used for the experiment discussed in this paper. The common mutation ancestor µ was calculated to be:

LVVKFEIEVCKVPLLSGNSNSQEHLYGVQFKRINSGDTWQYKNLASKILSELKL

In molecular biology, the functions of the KA1 domain is not yet known clearly, but there are classes of mammalian proteins that contain the domain KA1. Members if the Kinase family are present in various biological processes that involve cells and their control, ans also in protein stability [21].

Protein Sequences

As can be seen in Figure 3.3, some parts of the sequences of the protein family KA 1 (PF 02149) are shown in intervals of 10 sequences per row with types of nucleotides those are diverse among the members themselves. These sequences are generated in Hypertext format using the tool provided by the NCBI and it is accessible publicly online at the official NCBI website [10].



```
                                10        20        30        40
                          ....*....|....*....|....*....|....*....|....*.
1V5S_A             76 NLVQWEMEVCKLPRLS--LNGVRFKRISGTSIAFKNIASKIANELK 119
gi 167523651      509 EDVSWEMAVQKISRLG--LHGIRLRRLQGDHWRYKRLVDHVLQDAR 552
gi 514693851     1284 ETIVWEITVQVLPDLN--MRGIHLRRIKGNHWDYKKLVDEVIRKAK 1327
gi 330844773      819 EGVRFSIEVCRLPRLS--VNGLKFKRIGGSSWRYKSICKDLLSQMK 862
gi 470304590     1428 KPSQFELEVCHIPRLS--LYGLHVKRIRGDIWRHKRVCSTLIASMN 1471
gi 575485654       47 eGIQFELEVCRLPNLA--LNGLRFKRLMGNTWEYKDLLTNLISKMN 90
gi 238652648      987 gVVHWEMEICKLNRAG--ANGIRFKRISGSTSDFKRLANKLASDLE 1030
gi 358255311     1194 giVHWEMEVGKLAGVG--MNGIRFKRINGSMSAFKQIAKKLAADLK 1237
gi 674263912     1436 EILRLELEVCKLPKEG--MNGVRFKRLAGPAAEFKRISQKLAEDLK 1479
gi 808874931      528 ARVAFEAEVCQLPSGLgqSSGVRFKRLWGAPLAFRDIATKVSKELE 573
```

Figure 3.3: Highlighting a part of the aligned sequences of the protein family KA 1

Chapter 4

The Independence Testing Method

4.1 An Example Artificial Dataset

In this section, we describe the step-by-step procedure that we used to test whether among the surviving descendants of the hypothetical common ancestor µ the adjacent pairs of amino acids are mutated independently of each other.

As an artificial and simplified example, suppose that there exists an ancestor protein µ that is made up of only the amino acids A, D, N and R as shown in Table 2. Further assume during evolution each of these four amino acids either remains unchanged or is mutated into only one of the other three amino acids within this group of four amino acids. Suppose that the seven descendants are S1… S7 as shown also in Table 2.

Table 4.1: A set of seven artificial sequences for sample

S_1	`RNARDANDRADNRDANRARA`
S_2	`NRARDANRADADNANARNAD`
S_3	`RADNRANDANDRANDRDRAN`
S_4	`DNARDNARDRNARDANRANR`
S_5	`RNDRANRDRDANDNANDRAN`
S_6	`RNARDANDRADNRDANRARA`
S_7	`RNARDADDRADNRDANDADA`

4.2 Algorithm for Testing the Independence Hypothesis

Our testing method consists of the following five steps.

Step 1:

Construct the hypothetical common ancestor for the proteins in the given set of protein family using the method that is also used by the Common Mutation Similarity Matrix. In the case of amino acid sequences, the hypothetical common ancestor, μ, is constructed by taking an alignment of the amino acid sequences, and in each column of the alignment finding the amino acid (out of the twenty possible amino acids that are used in almost every protein in all organisms) that is overall closest to the all the amino acids in that column. The overall closest amino acid is by definition the amino acid that occurs most number of times. That is, we take the mode of the amino acids with the highest mode. If there are two or more values that are minimal, then we make a random selection. For the example in Table 4.1, consisting of seven artificial sequences from S1, S2, ... S7, each with a length of twenty nucleotides, the consensus sequence is:

Table 4.2: The consensus sequence for the artificial protein family in Figure 4.1

µ	`RNARDANDRADNRDANRNAA`

Step 2:

Next, we calculate a mutation probability matrix. The mutation probability matrix contains the probabilities of any amino acid changing into another amino acid. For the running example with the data shown in Table 4.1, the mutation probability matrix is shown in Table 4.3.

Table 4.3: The mutation probability matrix for the data in Figure 4.1

	A	**R**	**N**	**D**	**Total**
A	24	4	8	6	42
R	3	23	3	6	35
N	6	6	21	2	35
D	4	3	3	18	28

The mutation Probability Matrix in Table 4.2.1 shows the frequencies of the each of the four amino acid changes into one of the other three amino acids or remains the same. The column 'Total' shows the total number of the possibility of one amino acid can mutate into another amino acid, or remain the same throughout the entire sequence (S1 to S7).

Step 3:

Based on the mutation probability matrix values, we estimate the probability of the changes of any adjacent pair of amino acids into another pair of amino acids assuming that the mutations are independent of each other. For example, the probability of AN changing into DR can be computed as follows:

$$Prob(AN, DR) = Prob(A, D) * Prob(N, R) = \frac{6}{42} * \frac{6}{35} = \frac{6}{245} \approx 0.0245$$

Hence the theoretical probability corresponding to the amino acid pair AN changing to DR is approximately 0.0245. The theoretical probabilities for all possible combinations of amino acid pairs of the artificial sequence in Table 4.1 mutating into another possible pair of the same set are shown in Table 4.3. Note that the table values are in decimal format for the purpose of calculation.

Step 4:

Now, we calculate the actual probabilities of changes for each pair of amino acids in the consensus sequence. Starting from the first pair to the end of the consensus string, we first calculate the number of times and the index, each pair in the consensus string occurs. We then calculate the frequencies of that specific pair in the consensus string mutating into another pair among the rest of the descendent sequences in that column. If the current adjacent amino acid pair of the consensus string happens to appear in another index of the same consensus string, then we repeat the step to check for frequencies of that pair mutating into other possible pairs in that column, for the rest of the descendant sequences. We then slide the window of the current pair in the consensus string to the adjacent consecutive pair of the same consensus string, to calculate their respective frequencies of mutations among the descendent pairs of that column. The steps mentioned in the above paragraph are repeated until we encounter the last possible pair of the consensus sequence. The results for the example in Table 4.1 of the seven artificial sequences, are shown in Table 4.6. Note that in Table 4.6, the column ‘Total’ refers to the total number of ways in which a pair of the consensus sequence can mutate into another possible pair in its descendant sequence, whose value is the product of the number of times a single pair appears in the consensus string and the total number of sequences in the protein family. For example, in consensus string µ for the artificial sequence in Table 4.1, NR appears in two indices as highlighted in the Figure 4.1 below. In this case, the total number of possibilities of NR changing into another pair is $2 * 7 = 14$, where 7 denotes the total number of sequences of the protein family.

µ	`RNARDANDRADNRDANRNAA`

Figure 4.1: Recurring amino acid pairs of the consensus string are highlighted

The algorithm devised for calculating the actual probabilities for adjacent amino acid pairs are mentioned in the following paragraphs, in which we pass the protein sequences as a parameter to the algorithm.

Algorithm 1 ACTUAL-PROBABILITY-PAIRWISE(*sequence*)

```
INPUT: Read the sequences of a protein family that is in FASTA format and aligned appropriately. The sequences are numbered as $S_1, S_2, \ldots, S_n$ where $n$ denotes the total number of sequences.
//TOT gives the overall total number of possible ways a particular pair can mutate to another pair
protein := Consensus_Sequence //read the consensus sequence
m := Consensus_Sequence.size
n := sequence.length
for $i \rightarrow 1$ to $m$-1 do
    Calculate the count and index of all the adjacent pairs in the consensus sequence
    TOT := count * n
end for
for $i \rightarrow 1$ to $m$-1 do
    for $j \rightarrow 2$ to $n$ do
        calculate the occurrences of possible pairs in the descendent sequences corresponding to the column sequence[i][i+1] which is the consensus sequence
    end for
end for
```

Theorem. *The running time of the algorithm is* $O(n^2m)$ *where* $m \leq n$*, and* m *is the size of the consensus sequence and* n *is the length of the sequences of the protein family.*

Proof. The algorithm ACTUAL-PROBABILITY-PAIRWISE mentioned above, falls under the paradigm of dynamic programming in computer algorithm. We iterate through the consensus sequence m number of times for each adjacent pair in the consensus sequence and for each of those iterations we count the frequencies of the pair in that window which may or may not mutate into another pair in their descendant sequences of the corresponding window, which takes about n number of comparisons.

This operation can be seen under the nested loops of line 8 and line 9 in the algorithm above. Line 10 calculates the occurrences of the pair in the consensus mutating into one of the possible 400 pairs in the descendent sequences. This takes about n times of comparisons depending on the number of sequences that the protein is made up of. □

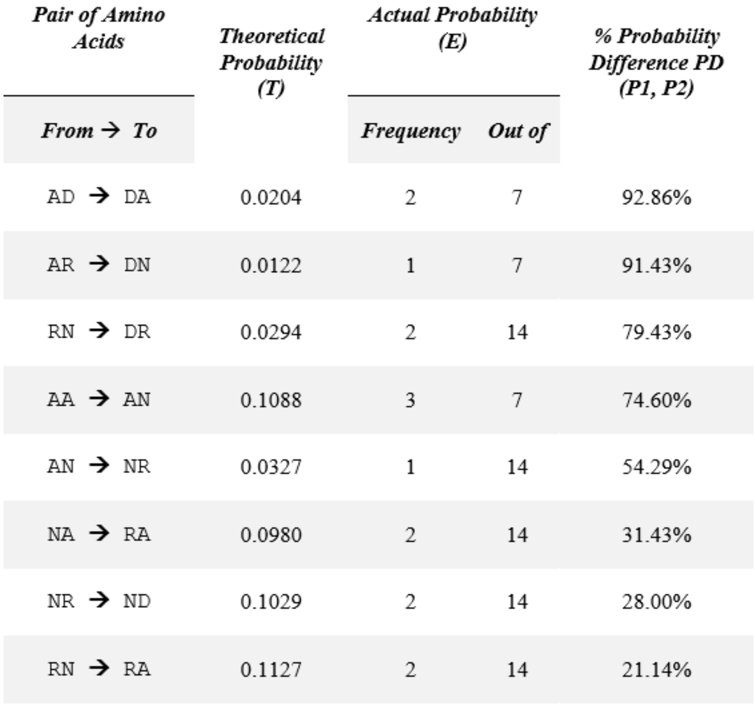
Step 5:

We compare the theoretical and the actual probabilities and note the most important discrepancies. The *percentage probability difference* in the theoretical and actual probabilities of the mutations of amino acid pairs is the absolute value of the difference between the two types of probabilities divided by the maximum of the two probabilities. Let $T(p1, p2)$ and$E(p1, p2)$ be the theoretical and the experimental probabilities, respectively, that the amino acid pair $p1$ changes into the amino acid pair$p2$. Let also $PD(p1, p2)$ be the percent probability difference defined as follows:

$$PD(p_1, p_2) = \frac{|T(p_1,p_2)-E(p_1,p_2)|}{Max(T(p_1,p_2),E(p_1,p_2))}$$

The percentage Probability Difference (PD) or the anomalous probabilities for the top eight pairs of the consensus sequence mutating into other pairs in the descendant sequences of the artificial protein family is shown in Table 4.4 below.

Table 4.4: Probability Differences for the artificial protein sequence in Figure 4.1

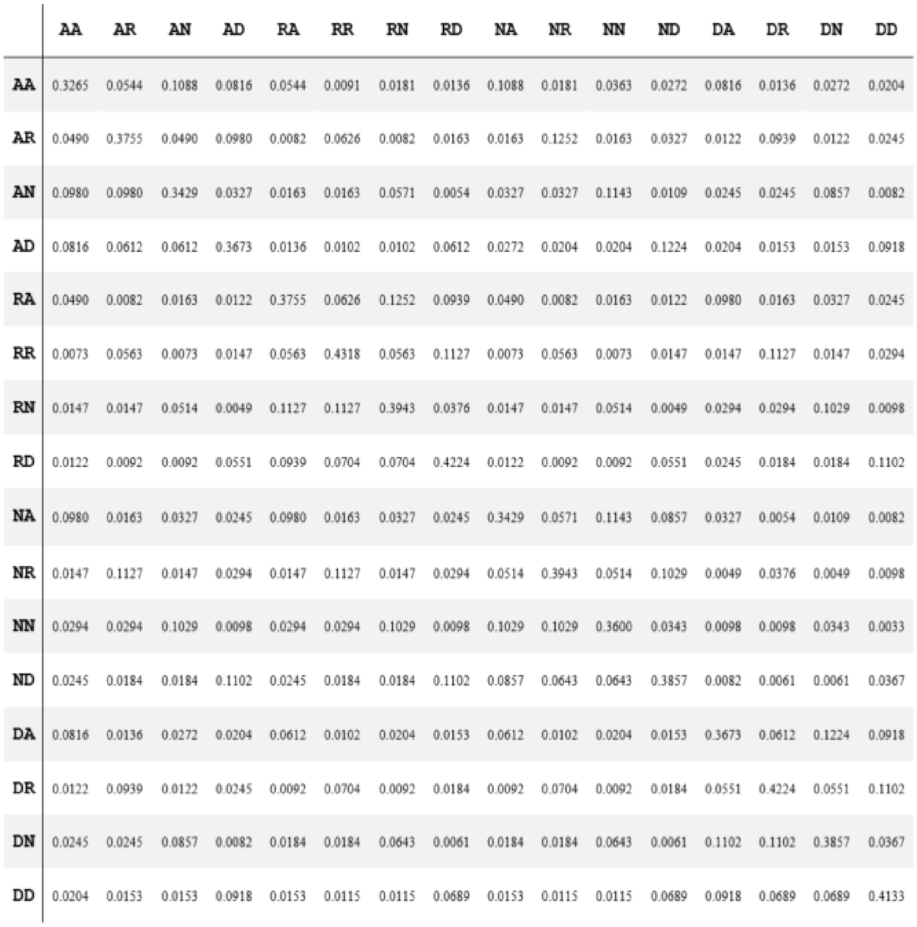


Pair of Amino Acids	***Theoretical Probability (T)***	***Actual Probability (E)***		***% Probability Difference PD (P1, P2)***
From → To		***Frequency***	***Out of***	
AD → DA	0.0204	2	7	92.86%
AR → DN	0.0122	1	7	91.43%
RN → DR	0.0294	2	14	79.43%
AA → AN	0.1088	3	7	74.60%
AN → NR	0.0327	1	14	54.29%
NA → RA	0.0980	2	14	31.43%
NR → ND	0.1029	2	14	28.00%
RN → RA	0.1127	2	14	21.14%

4.3 Applying the Algorithm to the Artificial Dataset

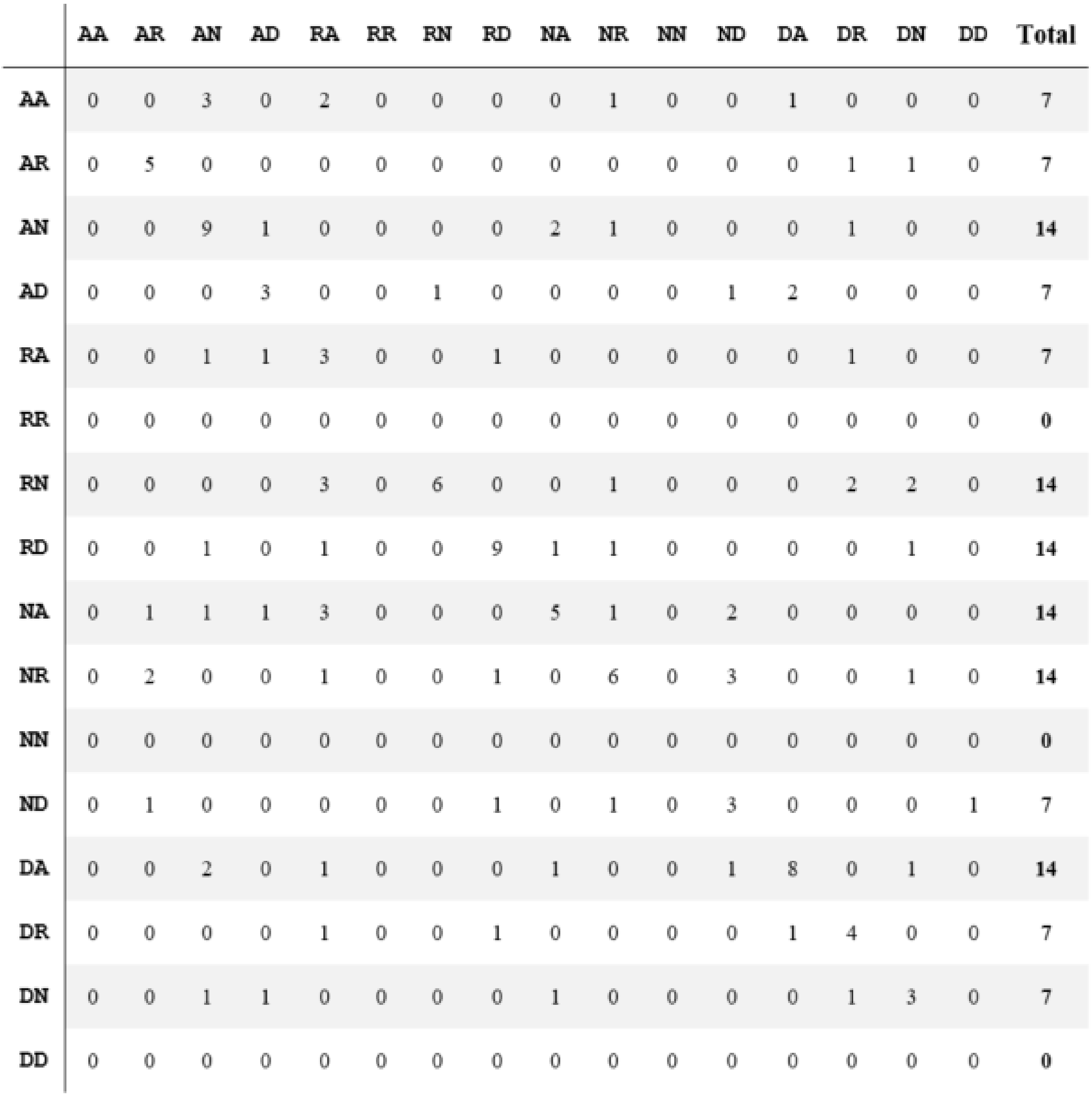
The following tables show the experimental results that were obtained as a result of running the proposed independence testing method on a set of artificial set of sequences that we had showcased in the previous sections.

Table 4.5: The theoretical probabilities of changes for each pairof amino acids for the artificial sample protein family



	AA	AR	AN	AD	RA	RR	RN	RD	NA	NR	NN	ND	DA	DR	DN	DD
AA	0.3265	0.0544	0.1088	0.0816	0.0544	0.0091	0.0181	0.0136	0.1088	0.0181	0.0363	0.0272	0.0816	0.0136	0.0272	0.0204
AR	0.0490	0.3755	0.0490	0.0980	0.0082	0.0626	0.0082	0.0163	0.0163	0.1252	0.0163	0.0327	0.0122	0.0939	0.0122	0.0245
AN	0.0980	0.0980	0.3429	0.0327	0.0163	0.0163	0.0571	0.0054	0.0327	0.0327	0.1143	0.0109	0.0245	0.0245	0.0857	0.0082
AD	0.0816	0.0612	0.0612	0.3673	0.0136	0.0102	0.0102	0.0612	0.0272	0.0204	0.0204	0.1224	0.0204	0.0153	0.0153	0.0918
RA	0.0490	0.0082	0.0163	0.0122	0.3755	0.0626	0.1252	0.0939	0.0490	0.0082	0.0163	0.0122	0.0980	0.0163	0.0327	0.0245
RR	0.0073	0.0563	0.0073	0.0147	0.0563	0.4318	0.0563	0.1127	0.0073	0.0563	0.0073	0.0147	0.0147	0.1127	0.0147	0.0294
RN	0.0147	0.0147	0.0514	0.0049	0.1127	0.1127	0.3943	0.0376	0.0147	0.0147	0.0514	0.0049	0.0294	0.0294	0.1029	0.0098
RD	0.0122	0.0092	0.0092	0.0551	0.0939	0.0704	0.0704	0.4224	0.0122	0.0092	0.0092	0.0551	0.0245	0.0184	0.0184	0.1102
NA	0.0980	0.0163	0.0327	0.0245	0.0980	0.0163	0.0327	0.0245	0.3429	0.0571	0.1143	0.0857	0.0327	0.0054	0.0109	0.0082
NR	0.0147	0.1127	0.0147	0.0294	0.0147	0.1127	0.0147	0.0294	0.0514	0.3943	0.0514	0.1029	0.0049	0.0376	0.0049	0.0098
NN	0.0294	0.0294	0.1029	0.0098	0.0294	0.0294	0.1029	0.0098	0.1029	0.1029	0.3600	0.0343	0.0098	0.0098	0.0343	0.0033
ND	0.0245	0.0184	0.0184	0.1102	0.0245	0.0184	0.0184	0.1102	0.0857	0.0643	0.0643	0.3857	0.0082	0.0061	0.0061	0.0367
DA	0.0816	0.0136	0.0272	0.0204	0.0612	0.0102	0.0204	0.0153	0.0612	0.0102	0.0204	0.0153	0.3673	0.0612	0.1224	0.0918
DR	0.0122	0.0939	0.0122	0.0245	0.0092	0.0704	0.0092	0.0184	0.0092	0.0704	0.0092	0.0184	0.0551	0.4224	0.0551	0.1102
DN	0.0245	0.0245	0.0857	0.0082	0.0184	0.0184	0.0643	0.0061	0.0184	0.0184	0.0643	0.0061	0.1102	0.1102	0.3857	0.0367
DD	0.0204	0.0153	0.0153	0.0918	0.0153	0.0115	0.0115	0.0689	0.0153	0.0115	0.0115	0.0689	0.0918	0.0689	0.0689	0.4133

Table 4.6: The actual probabilities of changes for each pair of amino acids for the artificial sample protein family



	AA	**AR**	**AN**	**AD**	**RA**	**RR**	**RN**	**RD**	**NA**	**NR**	**NN**	**ND**	**DA**	**DR**	**DN**	**DD**	**Total**
AA	0	0	3	0	2	0	0	0	0	1	0	0	1	0	0	0	**7**
AR	0	5	0	0	0	0	0	0	0	0	0	0	0	1	1	0	**7**
AN	0	0	9	1	0	0	0	0	2	1	0	0	0	1	0	0	**14**
AD	0	0	0	3	0	0	1	0	0	0	0	1	2	0	0	0	**7**
RA	0	0	1	1	3	0	0	1	0	0	0	0	0	1	0	0	**7**
RR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
RN	0	0	0	0	3	0	6	0	0	1	0	0	0	2	2	0	**14**
RD	0	0	1	0	1	0	0	9	1	1	0	0	0	0	1	0	**14**
NA	0	1	1	1	3	0	0	0	5	1	0	2	0	0	0	0	**14**
NR	0	2	0	0	1	0	0	1	0	6	0	3	0	0	1	0	**14**
NN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
ND	0	1	0	0	0	0	0	1	0	1	0	3	0	0	0	1	**7**
DA	0	0	2	0	1	0	0	0	1	0	0	1	8	0	1	0	**14**
DR	0	0	0	0	1	0	0	1	0	0	0	0	1	4	0	0	**7**
DN	0	0	1	1	0	0	0	0	1	0	0	0	0	1	3	0	**7**
DD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Chapter 5

Experimental Results and Discussions

5.1 Definition

This chapter initially focuses on defining the terms that are an integral part of the algorithm in the previous chapter. For better understanding, we first highlight the key points about each eminent term that we may come across later in this chapter.

5.1.1 Mutation Probabilty Matrix

The following tables in this section show the Mutation Probability Matrices that were generated for every single amino acid for each of the protein families. According to the methodology that was elucidated in Chapter 4, the mutation probability matrices for every single amino acid or nucleotides in each of the protein families separately, that are shown in the tables (Table 5.1 – Table 5.3) are used in the further steps where we generate the theoretical mutation probability matrix for every possible pair of amino acids. The resulting theoretical probability matrix in this case is a matrix of size 400 x 400 as there are 20 possible amino acid and hence not presented as tables here due to space constraints.

We then calculate the actual mutation probability for every pair of amino acids for each of the three families separately, which is also a huge set of results that contain all the possible probabilities of one pair in the consensus sequence of the protein family mutating into another pair. The number of resulting probabilities might be any number up to 400 x 400 as there are twenty amino acids in existence and there might be any pair of nucleotide mutating into another pair in their descendent sequences.

5.1.2 Mutations with Anomalous Probabilty

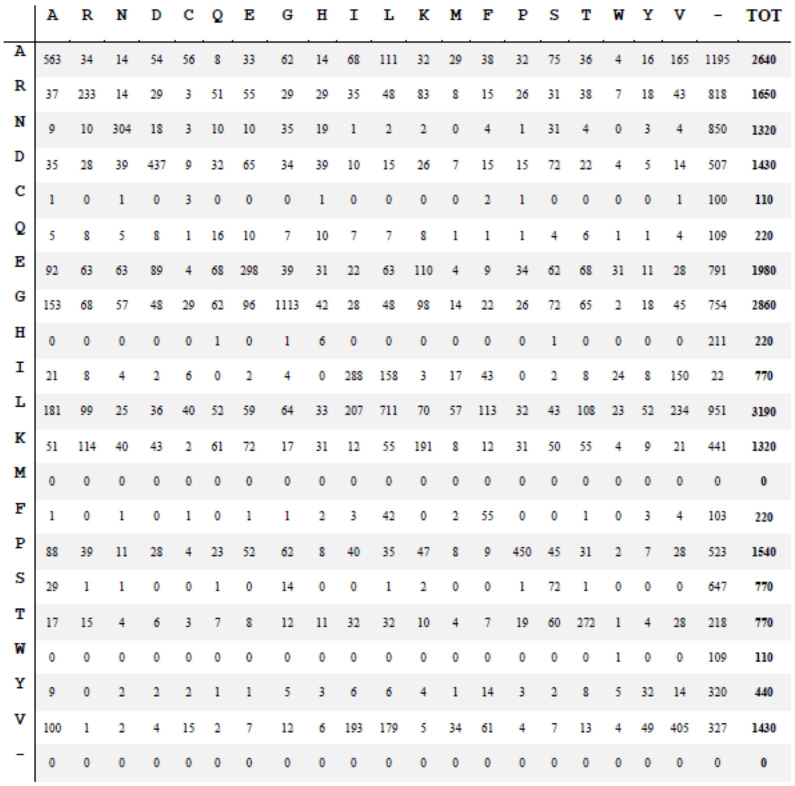
After the generation the mutation probability matrix corresponding to the theoretical and actual probabilities, we can check for pairwise mutations in the protein family that tends to have anomalous probability. Note that pairs that do not undergo mutations are also considered to be analyzed for anomalous probability. For all the pairwise mutations, we check the deviations of the actual probability of pairwise mutations with that of the theoretical probability. If the difference between them are significantly small, then it means that the independence hypothesis fails. In this thesis we consider the amino acid pairs that goes as low as 10%.

5.2 Results

This section lists the outcome of running the independence testing algorithm on the large data sets of protein sequences that was mentioned in Chapter 3. The Mutation Probability Matrix for single amino acid in a protein sequences are shown in subsection 5.2.1. The Theoretical Probability calculated using the mutation probability matrix are shown in the subsection 5.2.2 where we show the first fifteen pairs in rows and columns only, as the size of the original matrix is about the size of 400 x 400 in dimension.

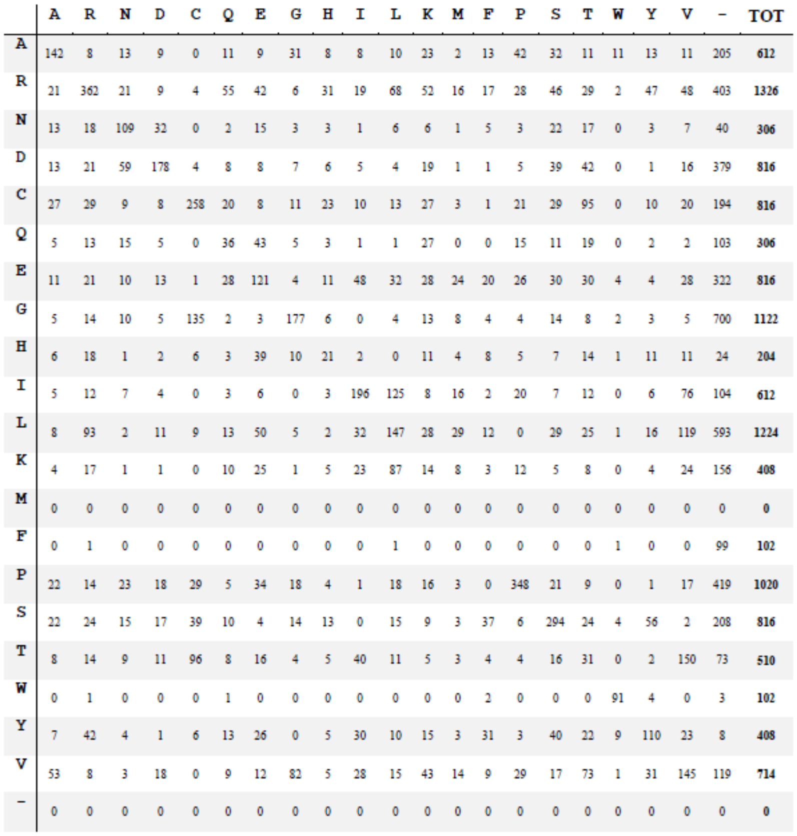
5.2.1 Mutation Probabilty Matrix for Single Amino Acids

Table 5.1: The actual probabilities of changes for each amino acid for the protein family DAGK_cat



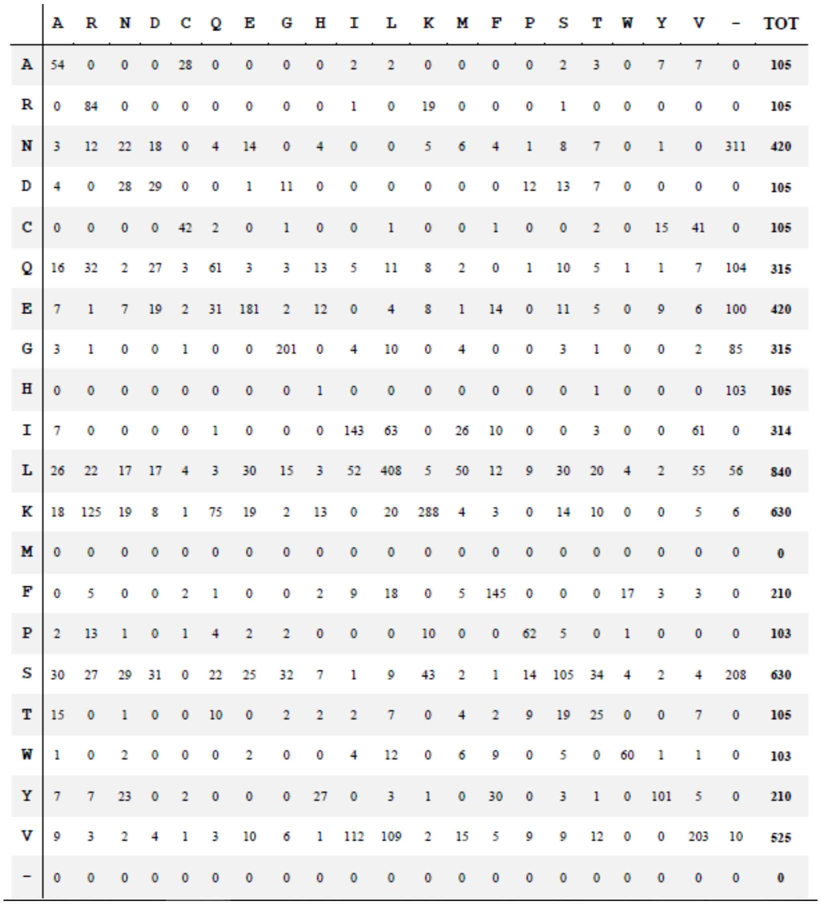
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	-	TOT
A	563	34	14	54	56	8	33	62	14	68	111	32	29	38	32	75	36	4	16	165	1195	**2640**
R	37	233	14	29	3	51	55	29	29	35	48	83	8	15	26	31	38	7	18	43	818	**1650**
N	9	10	304	18	3	10	10	35	19	1	2	2	0	4	1	31	4	0	3	4	850	**1320**
D	35	28	39	437	9	32	65	34	39	10	15	26	7	15	15	72	22	4	5	14	507	**1430**
C	1	0	1	0	3	0	0	0	1	0	0	0	0	2	1	0	0	0	0	1	100	**110**
Q	5	8	5	8	1	16	10	7	10	7	7	8	1	1	1	4	6	1	1	4	109	**220**
E	92	63	63	89	4	68	298	39	31	22	63	110	4	9	34	62	68	31	11	28	791	**1980**
G	153	68	57	48	29	62	96	1113	42	28	48	98	14	22	26	72	65	2	18	45	754	**2860**
H	0	0	0	0	0	1	0	1	6	0	0	0	0	0	0	1	0	0	0	0	211	**220**
I	21	8	4	2	6	0	2	4	0	288	158	3	17	43	0	2	8	24	8	150	22	**770**
L	181	99	25	36	40	52	59	64	33	207	711	70	57	113	32	43	108	23	52	234	951	**3190**
K	51	114	40	43	2	61	72	17	31	12	55	191	8	12	31	50	55	4	9	21	441	**1320**
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
F	1	0	1	0	1	0	1	1	2	3	42	0	2	55	0	0	1	0	3	4	103	**220**
P	88	39	11	28	4	23	52	62	8	40	35	47	8	9	450	45	31	2	7	28	523	**1540**
S	29	1	1	0	0	1	0	14	0	0	1	2	0	0	1	72	1	0	0	0	647	**770**
T	17	15	4	6	3	7	8	12	11	32	32	10	4	7	19	60	272	1	4	28	218	**770**
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	109	**110**
Y	9	0	2	2	2	1	1	5	3	6	6	4	1	14	3	2	8	5	32	14	320	**440**
V	100	1	2	4	15	2	7	12	6	193	179	5	34	61	4	7	13	4	49	405	327	**1430**
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Table 5.2: The actual probabilities of changes for each amino acid for the protein family IL17



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	-	TOT
A	142	8	13	9	0	11	9	31	8	8	10	23	2	13	42	32	11	11	13	11	205	**612**
R	21	362	21	9	4	55	42	6	31	19	68	52	16	17	28	46	29	2	47	48	403	**1326**
N	13	18	109	32	0	2	15	3	3	1	6	6	1	5	3	22	17	0	3	7	40	**306**
D	13	21	59	178	4	8	8	7	6	5	4	19	1	1	5	39	42	0	1	16	379	**816**
C	27	29	9	8	258	20	8	11	23	10	13	27	3	1	21	29	95	0	10	20	194	**816**
Q	5	13	15	5	0	36	43	5	3	1	1	27	0	0	15	11	19	0	2	2	103	**306**
E	11	21	10	13	1	28	121	4	11	48	32	28	24	20	26	30	30	4	4	28	322	**816**
G	5	14	10	5	135	2	3	177	6	0	4	13	8	4	4	14	8	2	3	5	700	**1122**
H	6	18	1	2	6	3	39	10	21	2	0	11	4	8	5	7	14	1	11	11	24	**204**
I	5	12	7	4	0	3	6	0	3	196	125	8	16	2	20	7	12	0	6	76	104	**612**
L	8	93	2	11	9	13	50	5	2	32	147	28	29	12	0	29	25	1	16	119	593	**1224**
K	4	17	1	1	0	10	25	1	5	23	87	14	8	3	12	5	8	0	4	24	156	**408**
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
F	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	99	**102**
P	22	14	23	18	29	5	34	18	4	1	18	16	3	0	348	21	9	0	1	17	419	**1020**
S	22	24	15	17	39	10	4	14	13	0	15	9	3	37	6	294	24	4	56	2	208	**816**
T	8	14	9	11	96	8	16	4	5	40	11	5	3	4	4	16	31	0	2	150	73	**510**
W	0	1	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	91	4	0	3	**102**
Y	7	42	4	1	6	13	26	0	5	30	10	15	3	31	3	40	22	9	110	23	8	**408**
V	53	8	3	18	0	9	12	82	5	28	15	43	14	9	29	17	73	1	31	145	119	**714**
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

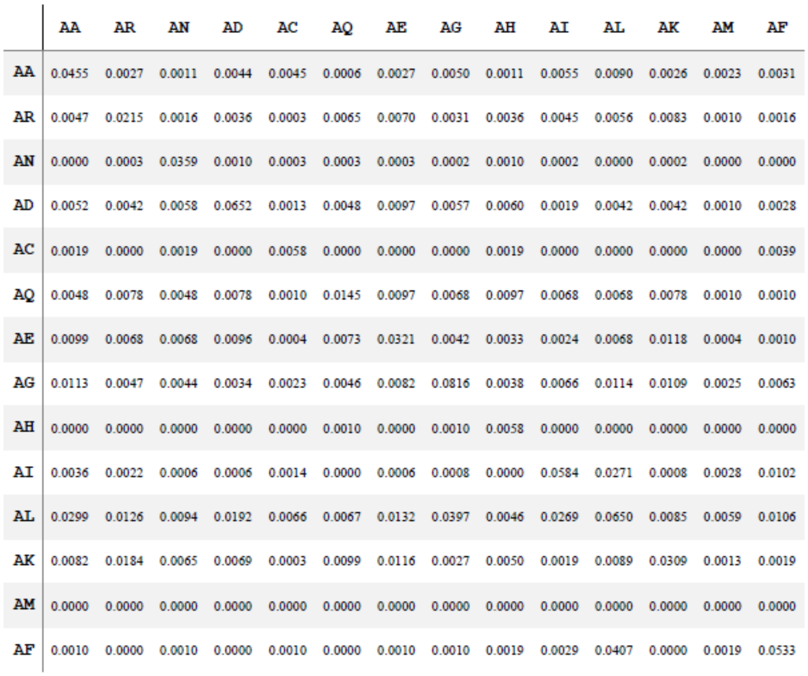
Table 5.3: The actual probabilities of changes for amino acids for the protein family KA1



	A	**R**	**N**	**D**	**C**	**Q**	**E**	**G**	**H**	**I**	**L**	**K**	**M**	**F**	**P**	**S**	**T**	**W**	**Y**	**V**	**-**	**TOT**
A	54	0	0	0	28	0	0	0	0	2	2	0	0	0	0	2	3	0	7	7	0	**105**
R	0	84	0	0	0	0	0	0	0	1	0	19	0	0	0	1	0	0	0	0	0	**105**
N	3	12	22	18	0	4	14	0	4	0	0	5	6	4	1	8	7	0	1	0	311	**420**
D	4	0	28	29	0	0	1	11	0	0	0	0	0	0	12	13	7	0	0	0	0	**105**
C	0	0	0	0	42	2	0	1	0	0	1	0	0	1	0	0	2	0	15	41	0	**105**
Q	16	32	2	27	3	61	3	3	13	5	11	8	2	0	1	10	5	1	1	7	104	**315**
E	7	1	7	19	2	31	181	2	12	0	4	8	1	14	0	11	5	0	9	6	100	**420**
G	3	1	0	0	1	0	0	201	0	4	10	0	4	0	0	3	1	0	0	2	85	**315**
H	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	103	**105**
I	7	0	0	0	0	1	0	0	0	143	63	0	26	10	0	0	3	0	0	61	0	**314**
L	26	22	17	17	4	3	30	15	3	52	408	5	50	12	9	30	20	4	2	55	56	**840**
K	18	125	19	8	1	75	19	2	13	0	20	288	4	3	0	14	10	0	0	5	6	**630**
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**
F	0	5	0	0	2	1	0	0	2	9	18	0	5	145	0	0	0	17	3	3	0	**210**
P	2	13	1	0	1	4	2	2	0	0	0	10	0	0	62	5	0	1	0	0	0	**103**
S	30	27	29	31	0	22	25	32	7	1	9	43	2	1	14	105	34	4	2	4	208	**630**
T	15	0	1	0	0	10	0	2	2	2	7	0	4	2	9	19	25	0	0	7	0	**105**
W	1	0	2	0	0	0	2	0	0	4	12	0	6	9	0	5	0	60	1	1	0	**103**
Y	7	7	23	0	2	0	0	0	27	0	3	1	0	30	0	3	1	0	101	5	0	**210**
V	9	3	2	4	1	3	10	6	1	112	109	2	15	5	9	9	12	0	0	203	10	**525**
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

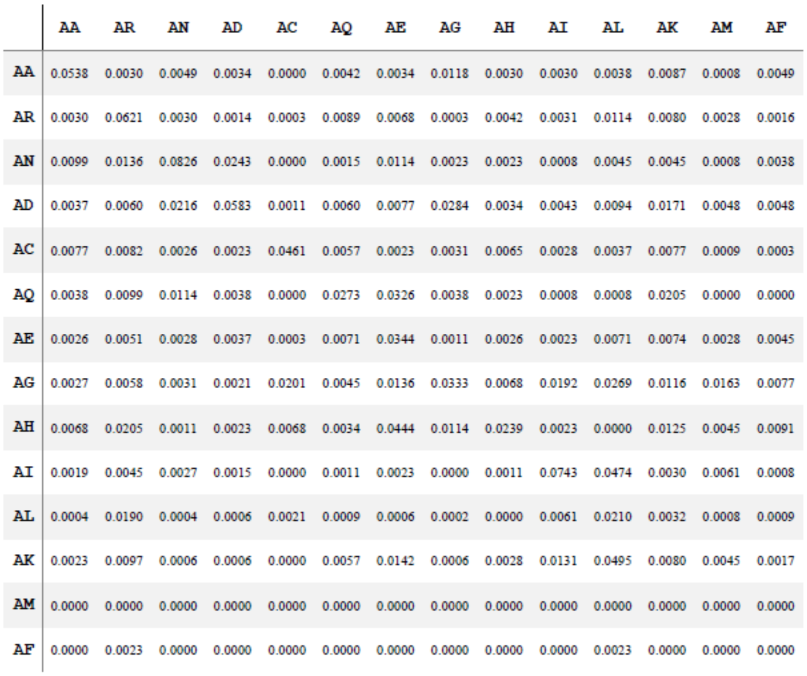
5.2.2 Mutation Probabilty Matrix for Amino Acid Pairs

Table 5.4: Theoretical Probabilities for amino acid pairs for the protein family DAGK_cat



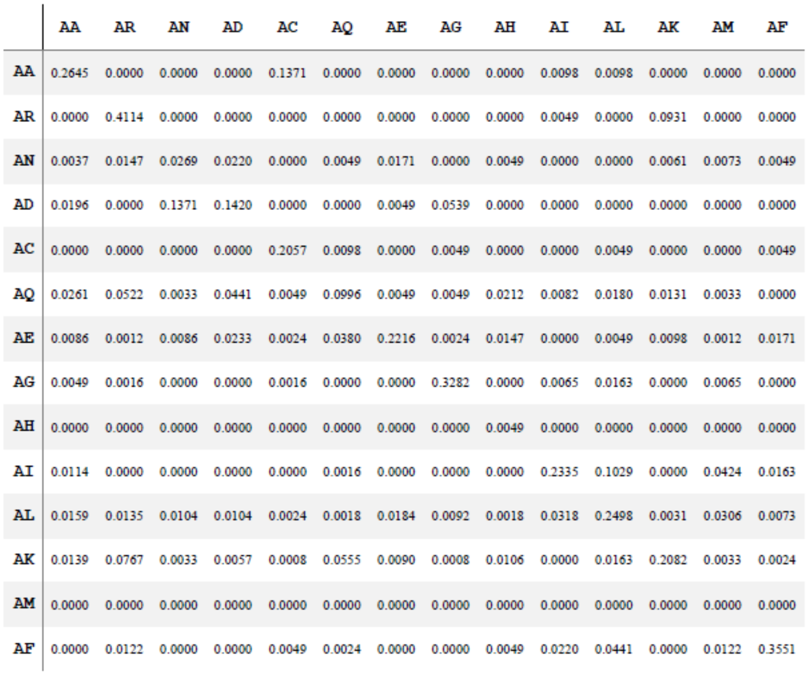
	AA	AR	AN	AD	AC	AQ	AE	AG	AH	AI	AL	AK	AM	AF
AA	0.0455	0.0027	0.0011	0.0044	0.0045	0.0006	0.0027	0.0050	0.0011	0.0055	0.0090	0.0026	0.0023	0.0031
AR	0.0047	0.0215	0.0016	0.0036	0.0003	0.0065	0.0070	0.0031	0.0036	0.0045	0.0056	0.0083	0.0010	0.0016
AN	0.0000	0.0003	0.0359	0.0010	0.0003	0.0003	0.0003	0.0002	0.0010	0.0002	0.0000	0.0002	0.0000	0.0000
AD	0.0052	0.0042	0.0058	0.0652	0.0013	0.0048	0.0097	0.0057	0.0060	0.0019	0.0042	0.0042	0.0010	0.0028
AC	0.0019	0.0000	0.0019	0.0000	0.0058	0.0000	0.0000	0.0000	0.0019	0.0000	0.0000	0.0000	0.0000	0.0039
AQ	0.0048	0.0078	0.0048	0.0078	0.0010	0.0145	0.0097	0.0068	0.0097	0.0068	0.0068	0.0078	0.0010	0.0010
AE	0.0099	0.0068	0.0068	0.0096	0.0004	0.0073	0.0321	0.0042	0.0033	0.0024	0.0068	0.0118	0.0004	0.0010
AG	0.0113	0.0047	0.0044	0.0034	0.0023	0.0046	0.0082	0.0816	0.0038	0.0066	0.0114	0.0109	0.0025	0.0063
AH	0.0000	0.0000	0.0000	0.0000	0.0000	0.0010	0.0000	0.0010	0.0058	0.0000	0.0000	0.0000	0.0000	0.0000
AI	0.0036	0.0022	0.0006	0.0006	0.0014	0.0000	0.0006	0.0008	0.0000	0.0584	0.0271	0.0008	0.0028	0.0102
AL	0.0299	0.0126	0.0094	0.0192	0.0066	0.0067	0.0132	0.0397	0.0046	0.0269	0.0650	0.0085	0.0059	0.0106
AK	0.0082	0.0184	0.0065	0.0069	0.0003	0.0099	0.0116	0.0027	0.0050	0.0019	0.0089	0.0309	0.0013	0.0019
AM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AF	0.0010	0.0000	0.0010	0.0000	0.0010	0.0000	0.0010	0.0010	0.0019	0.0029	0.0407	0.0000	0.0019	0.0533

Table 5.5: Theoretical Probabilities for amino acid pairs for the protein family IL17



	AA	AR	AN	AD	AC	AQ	AE	AG	AH	AI	AL	AK	AM	AF
AA	0.0538	0.0030	0.0049	0.0034	0.0000	0.0042	0.0034	0.0118	0.0030	0.0030	0.0038	0.0087	0.0008	0.0049
AR	0.0030	0.0621	0.0030	0.0014	0.0003	0.0089	0.0068	0.0003	0.0042	0.0031	0.0114	0.0080	0.0028	0.0016
AN	0.0099	0.0136	0.0826	0.0243	0.0000	0.0015	0.0114	0.0023	0.0023	0.0008	0.0045	0.0045	0.0008	0.0038
AD	0.0037	0.0060	0.0216	0.0583	0.0011	0.0060	0.0077	0.0284	0.0034	0.0043	0.0094	0.0171	0.0048	0.0048
AC	0.0077	0.0082	0.0026	0.0023	0.0461	0.0057	0.0023	0.0031	0.0065	0.0028	0.0037	0.0077	0.0009	0.0003
AQ	0.0038	0.0099	0.0114	0.0038	0.0000	0.0273	0.0326	0.0038	0.0023	0.0008	0.0008	0.0205	0.0000	0.0000
AE	0.0026	0.0051	0.0028	0.0037	0.0003	0.0071	0.0344	0.0011	0.0026	0.0023	0.0071	0.0074	0.0028	0.0045
AG	0.0027	0.0058	0.0031	0.0021	0.0201	0.0045	0.0136	0.0333	0.0068	0.0192	0.0269	0.0116	0.0163	0.0077
AH	0.0068	0.0205	0.0011	0.0023	0.0068	0.0034	0.0444	0.0114	0.0239	0.0023	0.0000	0.0125	0.0045	0.0091
AI	0.0019	0.0045	0.0027	0.0015	0.0000	0.0011	0.0023	0.0000	0.0011	0.0743	0.0474	0.0030	0.0061	0.0008
AL	0.0004	0.0190	0.0004	0.0006	0.0021	0.0009	0.0006	0.0002	0.0000	0.0061	0.0210	0.0032	0.0008	0.0009
AK	0.0023	0.0097	0.0006	0.0006	0.0000	0.0057	0.0142	0.0006	0.0028	0.0131	0.0495	0.0080	0.0045	0.0017
AM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AF	0.0000	0.0023	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0023	0.0000	0.0000	0.0000

Table 5.6: Theoretical Probabilities for amino acid pairs for the protein family KA1



	AA	AR	AN	AD	AC	AQ	AE	AG	AH	AI	AL	AK	AM	AF
AA	0.2645	0.0000	0.0000	0.0000	0.1371	0.0000	0.0000	0.0000	0.0000	0.0098	0.0098	0.0000	0.0000	0.0000
AR	0.0000	0.4114	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0049	0.0000	0.0931	0.0000	0.0000
AN	0.0037	0.0147	0.0269	0.0220	0.0000	0.0049	0.0171	0.0000	0.0049	0.0000	0.0000	0.0061	0.0073	0.0049
AD	0.0196	0.0000	0.1371	0.1420	0.0000	0.0000	0.0049	0.0539	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AC	0.0000	0.0000	0.0000	0.0000	0.2057	0.0098	0.0000	0.0049	0.0000	0.0000	0.0049	0.0000	0.0000	0.0049
AQ	0.0261	0.0522	0.0033	0.0441	0.0049	0.0996	0.0049	0.0049	0.0212	0.0082	0.0180	0.0131	0.0033	0.0000
AE	0.0086	0.0012	0.0086	0.0233	0.0024	0.0380	0.2216	0.0024	0.0147	0.0000	0.0049	0.0098	0.0012	0.0171
AG	0.0049	0.0016	0.0000	0.0000	0.0016	0.0000	0.0000	0.3282	0.0000	0.0065	0.0163	0.0000	0.0065	0.0000
AH	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0049	0.0000	0.0000	0.0000	0.0000	0.0000
AI	0.0114	0.0000	0.0000	0.0000	0.0000	0.0016	0.0000	0.0000	0.0000	0.2335	0.1029	0.0000	0.0424	0.0163
AL	0.0159	0.0135	0.0104	0.0104	0.0024	0.0018	0.0184	0.0092	0.0018	0.0318	0.2498	0.0031	0.0306	0.0073
AK	0.0139	0.0767	0.0033	0.0057	0.0008	0.0555	0.0090	0.0008	0.0106	0.0000	0.0163	0.2082	0.0033	0.0024
AM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AF	0.0000	0.0122	0.0000	0.0000	0.0049	0.0024	0.0000	0.0000	0.0049	0.0220	0.0441	0.0000	0.0122	0.3551

5.3 Discussions

In this section, we discuss the findings that were generated as a result of the independence testing algorithm proposed in the previous chapter. Some of the key areas that we are interested to talk about, are about the anomalous probabilities of pairwise mutations and also about the chances of finding a single common pairwise mutations among all the three protein families.

5.3.1 Probability of Finding a Single Common Pairwise Mutation

The common or similar pairwise mutations can be deduced from the percentage probabilities that are shown in Table 5.8 to Table 5.10. The following Table 5.7 shows five pairwise mutations that are common in at least two of the three protein families that we studied. The first three mutations occur exactly the same in the corresponding protein families. In the fourth and the fifth mutations, the pairs are interchanged. For example, when we take the IP→VP mutation, which occurs in the DAGK_cat protein, and interchange the pairs on both the left and the right hand sides, then we get the symmetric mutation PI→PV, which occurs in the IL17 protein. These two mutations are very similar to each other because proteins are amino acid chains, and the two mutations simple "read" these amino acid chains from different directions.

There are a total of 400 x 400 = 160,000 possible pairwise mutations. The probability of finding a common pairwise mutation out of the top 31 of IL17 mutations and the top 18 KA1 mutations, can be calculated as:

Prob (out of the 18 new pairs picked from 160,000 at least one will match with one of the 31 pairs picked before)

Prob (out of the 18 new pairs picked from 160,000 at least one will match with one of the 31 pairs picked before)
= 1 – Prob (none of 18 new picked matches 31 picked before)

Considering this probabiltiy in terms of permutations, this problem could be solved as follows:

$$1 - \frac{{}_{n}P_{r}}{{}_{m}P_{r}} = 1 - \frac{\frac{n!}{(n-r)!}}{\frac{m!}{(m-r)!}} \quad \text{where, } m = 160000, n = 160000 - 31, r = 18$$

On substitution respectively, we get,

$$1 - \frac{{}_{(160000-31)}P_{18}}{{}_{160000}P_{18}} \approx 0.0035$$

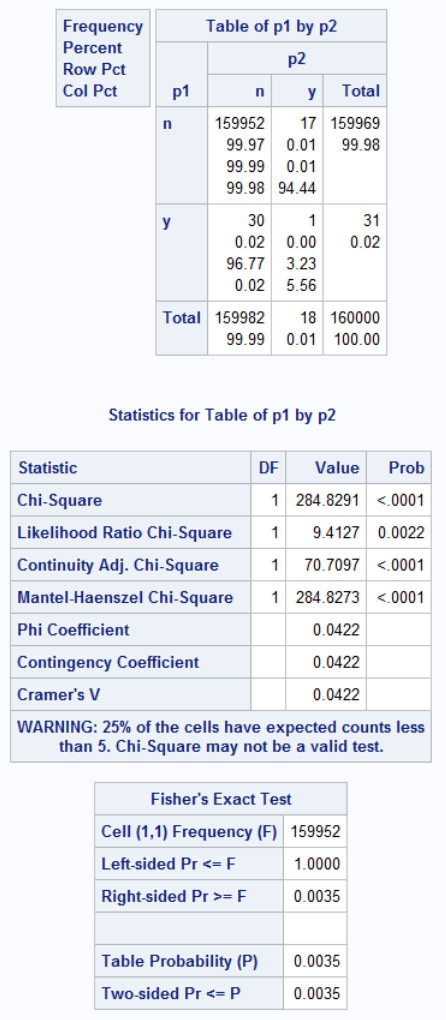
Let us set this to be our P-value.

The common or similar mutations for the three protein families are shown under Table 5.7.

Table 5.7: Common or similar mutations in the three protein families

Mutation	**DAGK_cat**	**IL17**	**KA1**
1	`EV`→`EV`		`EV`→`EV`
2		`LS`→`LS`	`LS`→`LS`
3		`VP`→`LP`	`VP`→`LP`
4	`IP`→`VP`	`PI`→`PV`	
5	`VL`→`VV`	`LV`→`VV`	

As can be seen, in this case there are three pairs that are common in at least two protein families, and there are two pairs that are complement of each other, which could be treated to be similar. Statistically, the probability of finding five common mutations in at least two of the protein families was calculated to be about ≤ 0.0001 which is significantly lesser than the P-value. The following figures show the statistical results generated using SAS for our example.



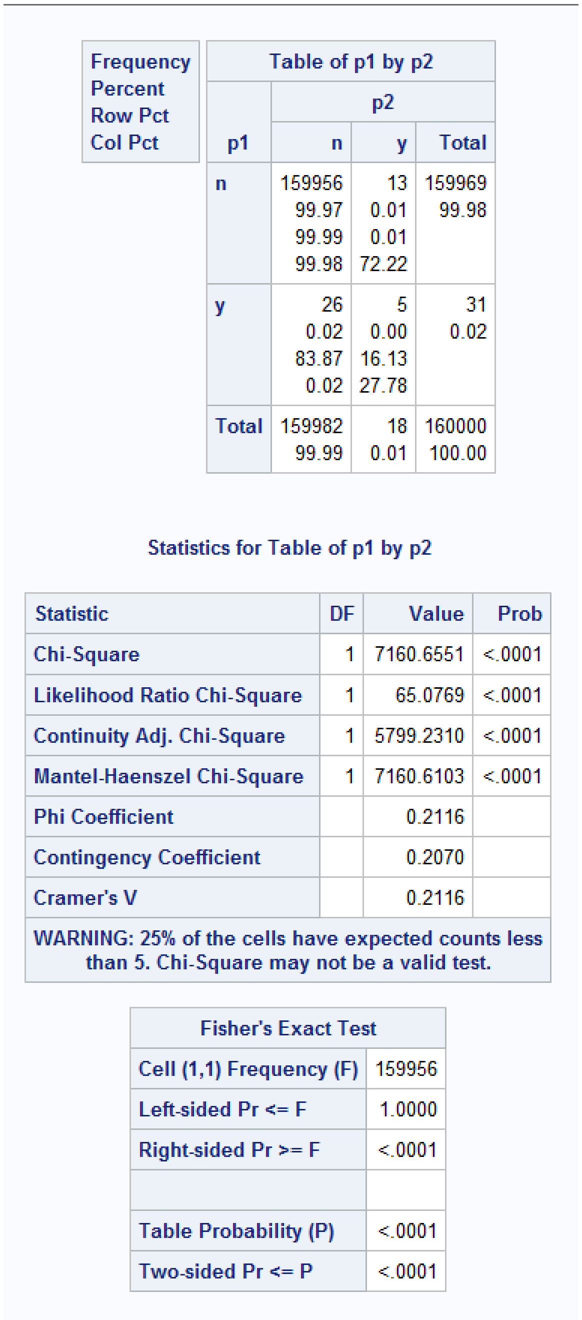
Frequency Percent Row Pct Col Pct	Table of p1 by p2		
	p2		
p1	n	y	Total
n	159952 99.97 99.99 99.98	17 0.01 0.01 94.44	159969 99.98
y	30 0.02 96.77 0.02	1 0.00 3.23 5.56	31 0.02
Total	159982 99.99	18 0.01	160000 100.00

Statistics for Table of p1 by p2

Statistic	DF	Value	Prob
Chi-Square	1	284.8291	<.0001
Likelihood Ratio Chi-Square	1	9.4127	0.0022
Continuity Adj. Chi-Square	1	70.7097	<.0001
Mantel-Haenszel Chi-Square	1	284.8273	<.0001
Phi Coefficient		0.0422	
Contingency Coefficient		0.0422	
Cramer's V		0.0422	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Fisher's Exact Test	
Cell (1,1) Frequency (F)	159952
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	0.0035
Table Probability (P)	0.0035
Two-sided Pr <= P	0.0035

Figure 5.1: SAS results showing the probability of finding at least one common pairwise muation out of the top 31 of IL17 mutations and the top 18 KA1 mutations



Frequency Percent Row Pct Col Pct

Table of p1 by p2			
	p2		
p1	n	y	Total
n	159956 99.97 99.99 99.98	13 0.01 0.01 72.22	159969 99.98
y	26 0.02 83.87 0.02	5 0.00 16.13 27.78	31 0.02
Total	159982 99.99	18 0.01	160000 100.00

Statistics for Table of p1 by p2

Statistic	DF	Value	Prob
Chi-Square	1	7160.6551	<.0001
Likelihood Ratio Chi-Square	1	65.0769	<.0001
Continuity Adj. Chi-Square	1	5799.2310	<.0001
Mantel-Haenszel Chi-Square	1	7160.6103	<.0001
Phi Coefficient		0.2116	
Contingency Coefficient		0.2070	
Cramer's V		0.2116	
WARNING: 25% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

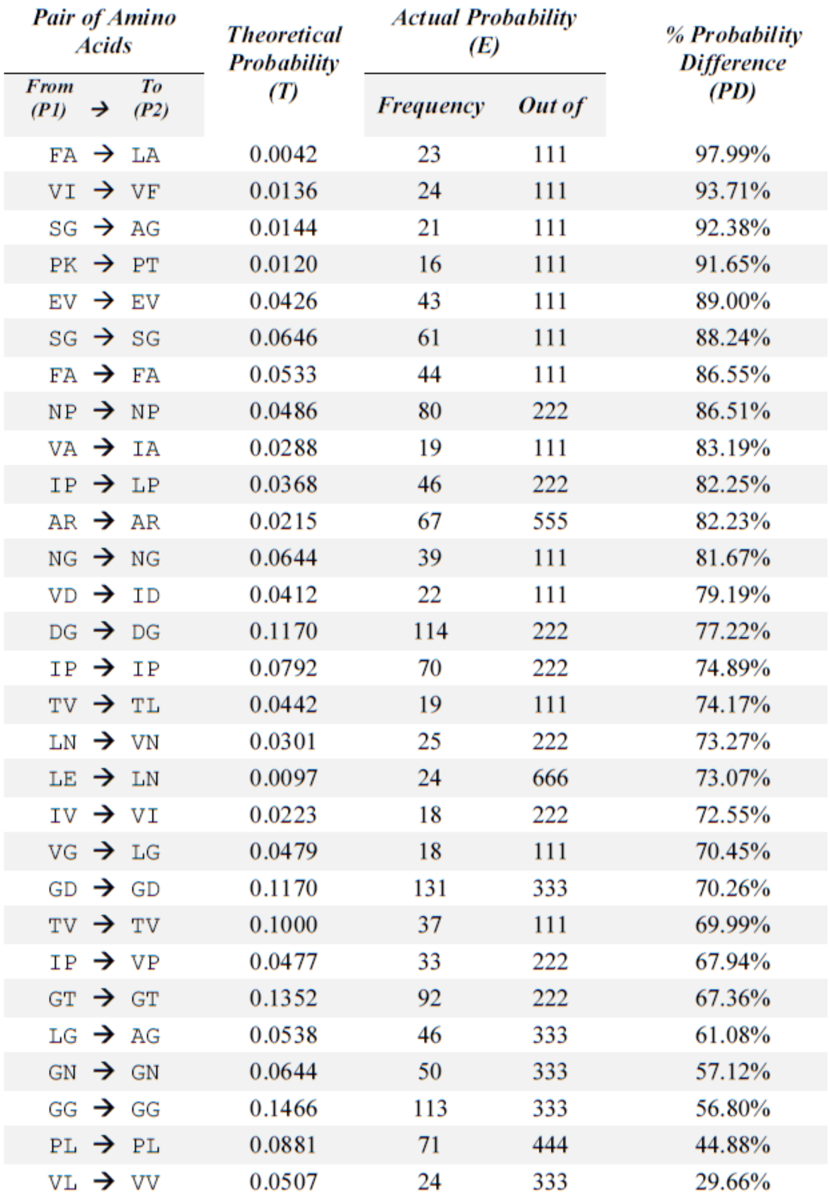
Fisher's Exact Test	
Cell (1,1) Frequency (F)	159956
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Figure 5.2: Finding 5 common pairwise mutations out of the top 31 IL17 mutations, the top 18 KA1 mutations and the top 31 DAGK_cat mutations

5.3.2 Anomalously Frequent Mutations

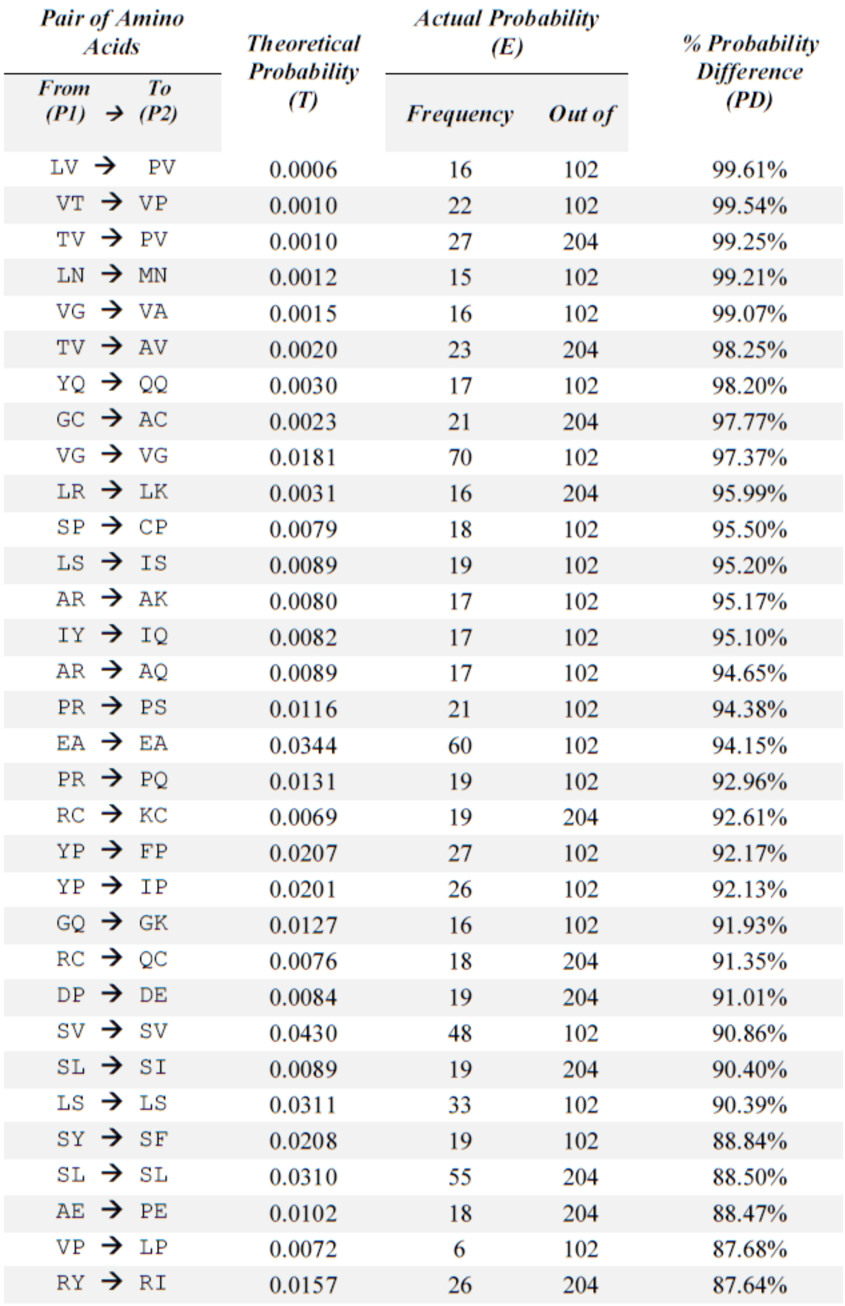
The following tables show the probability differences in percentage (%) or the anomalous probabilities for one pair mutating into another pairs. The Anomalous probability is calculated based on the theoretical probability and actual probability of the top fifteen amino acid pairs and it can be deduced that the higher percentage probabilities mean that the actual probabilities are less deviated from the theoretical probabilities and hence imply that the mutations of those pairs satisfy the independence hypothesis. In this section we represent the mutation pairs with anomalous probabilities in Table 5.8 through Table 5.10.

Table 5.8: Experimental results using the amino acid sequences in the DAGK_cat protein family



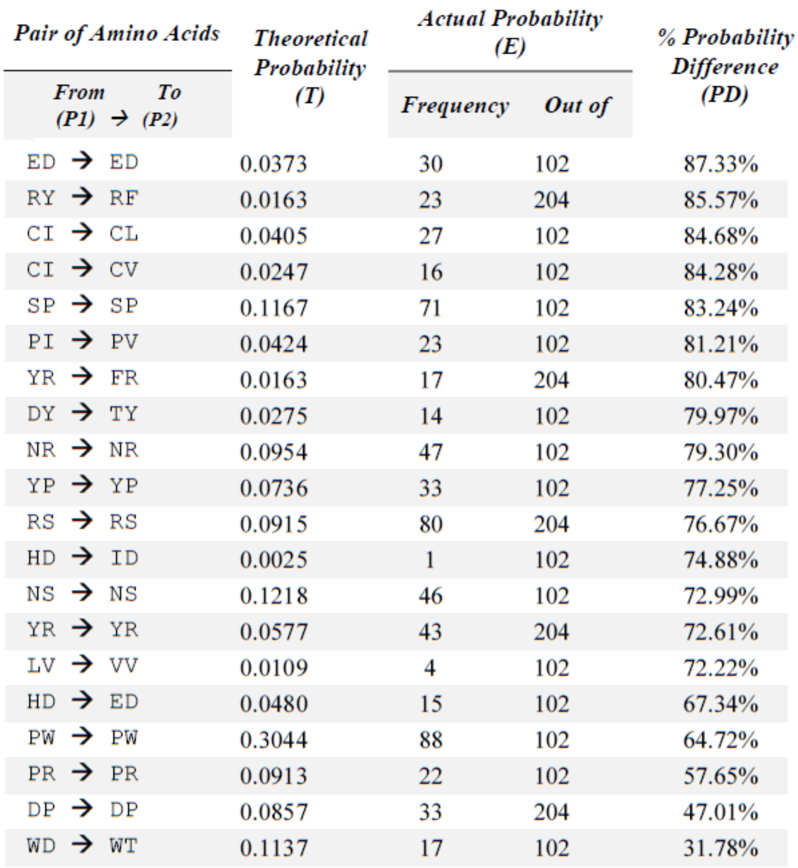
Pair of Amino Acids			Theoretical Probability (T)	Actual Probability (E)		% Probability Difference (PD)
From (P1)	→	To (P2)		Frequency	Out of	
FA	→	LA	0.0042	23	111	97.99%
VI	→	VF	0.0136	24	111	93.71%
SG	→	AG	0.0144	21	111	92.38%
PK	→	PT	0.0120	16	111	91.65%
EV	→	EV	0.0426	43	111	89.00%
SG	→	SG	0.0646	61	111	88.24%
FA	→	FA	0.0533	44	111	86.55%
NP	→	NP	0.0486	80	222	86.51%
VA	→	IA	0.0288	19	111	83.19%
IP	→	LP	0.0368	46	222	82.25%
AR	→	AR	0.0215	67	555	82.23%
NG	→	NG	0.0644	39	111	81.67%
VD	→	ID	0.0412	22	111	79.19%
DG	→	DG	0.1170	114	222	77.22%
IP	→	IP	0.0792	70	222	74.89%
TV	→	TL	0.0442	19	111	74.17%
LN	→	VN	0.0301	25	222	73.27%
LE	→	LN	0.0097	24	666	73.07%
IV	→	VI	0.0223	18	222	72.55%
VG	→	LG	0.0479	18	111	70.45%
GD	→	GD	0.1170	131	333	70.26%
TV	→	TV	0.1000	37	111	69.99%
IP	→	VP	0.0477	33	222	67.94%
GT	→	GT	0.1352	92	222	67.36%
LG	→	AG	0.0538	46	333	61.08%
GN	→	GN	0.0644	50	333	57.12%
GG	→	GG	0.1466	113	333	56.80%
PL	→	PL	0.0881	71	444	44.88%
VL	→	VV	0.0507	24	333	29.66%

Table 5.9: Experimental results using the amino acid sequences in the IL17 protein family



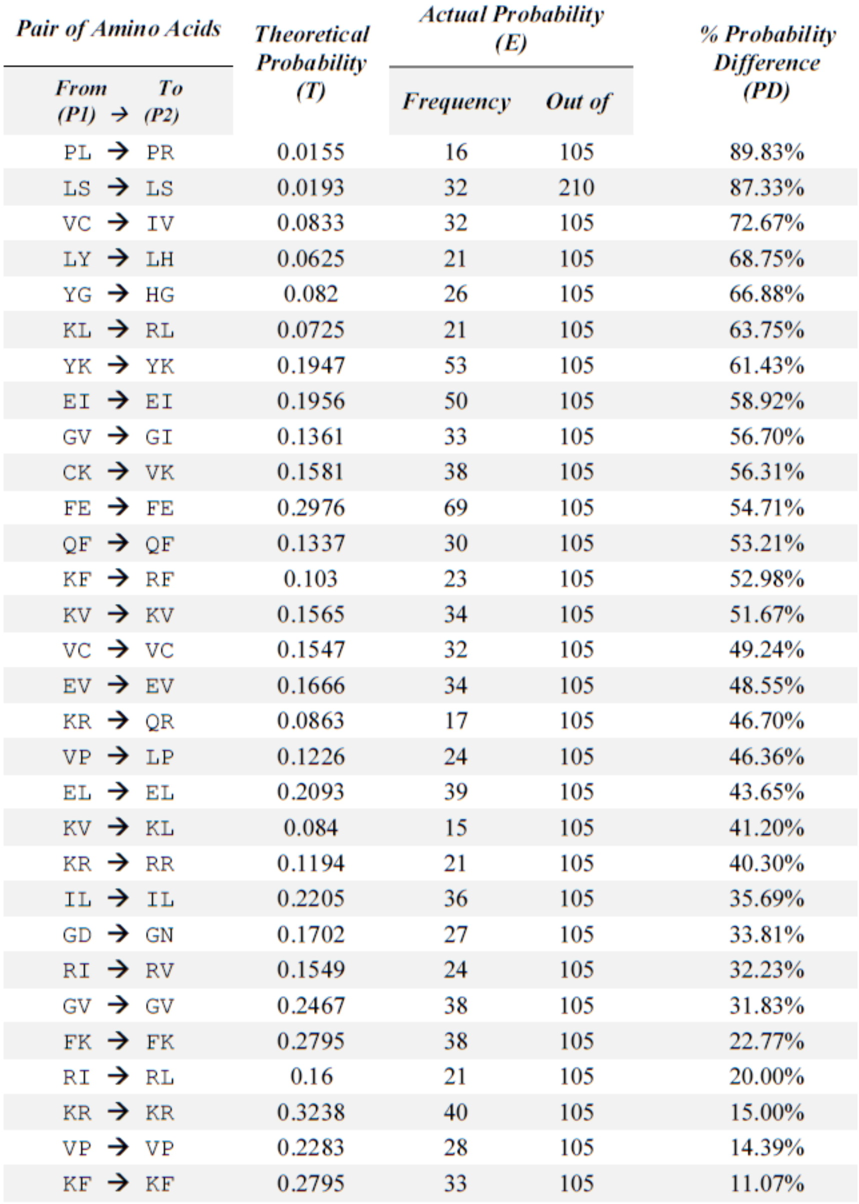
Pair of Amino Acids			***Theoretical Probability (T)***	***Actual Probability (E)***		***% Probability Difference (PD)***
From (P1)	→	***To (P2)***		***Frequency***	***Out of***	
LV	→	PV	0.0006	16	102	99.61%
VT	→	VP	0.0010	22	102	99.54%
TV	→	PV	0.0010	27	204	99.25%
LN	→	MN	0.0012	15	102	99.21%
VG	→	VA	0.0015	16	102	99.07%
TV	→	AV	0.0020	23	204	98.25%
YQ	→	QQ	0.0030	17	102	98.20%
GC	→	AC	0.0023	21	204	97.77%
VG	→	VG	0.0181	70	102	97.37%
LR	→	LK	0.0031	16	204	95.99%
SP	→	CP	0.0079	18	102	95.50%
LS	→	IS	0.0089	19	102	95.20%
AR	→	AK	0.0080	17	102	95.17%
IY	→	IQ	0.0082	17	102	95.10%
AR	→	AQ	0.0089	17	102	94.65%
PR	→	PS	0.0116	21	102	94.38%
EA	→	EA	0.0344	60	102	94.15%
PR	→	PQ	0.0131	19	102	92.96%
RC	→	KC	0.0069	19	204	92.61%
YP	→	FP	0.0207	27	102	92.17%
YP	→	IP	0.0201	26	102	92.13%
GQ	→	GK	0.0127	16	102	91.93%
RC	→	QC	0.0076	18	204	91.35%
DP	→	DE	0.0084	19	204	91.01%
SV	→	SV	0.0430	48	102	90.86%
SL	→	SI	0.0089	19	204	90.40%
LS	→	LS	0.0311	33	102	90.39%
SY	→	SF	0.0208	19	102	88.84%
SL	→	SL	0.0310	55	204	88.50%
AE	→	PE	0.0102	18	204	88.47%
VP	→	LP	0.0072	6	102	87.68%
RY	→	RI	0.0157	26	204	87.64%

Table 5.9 (Continued..)



Pair of Amino Acids			***Theoretical Probability (T)***	***Actual Probability (E)***		***% Probability Difference (PD)***
From (P1)	→	***To (P2)***		***Frequency***	***Out of***	
ED	→	ED	0.0373	30	102	87.33%
RY	→	RF	0.0163	23	204	85.57%
CI	→	CL	0.0405	27	102	84.68%
CI	→	CV	0.0247	16	102	84.28%
SP	→	SP	0.1167	71	102	83.24%
PI	→	PV	0.0424	23	102	81.21%
YR	→	FR	0.0163	17	204	80.47%
DY	→	TY	0.0275	14	102	79.97%
NR	→	NR	0.0954	47	102	79.30%
YP	→	YP	0.0736	33	102	77.25%
RS	→	RS	0.0915	80	204	76.67%
HD	→	ID	0.0025	1	102	74.88%
NS	→	NS	0.1218	46	102	72.99%
YR	→	YR	0.0577	43	204	72.61%
LV	→	VV	0.0109	4	102	72.22%
HD	→	ED	0.0480	15	102	67.34%
PW	→	PW	0.3044	88	102	64.72%
PR	→	PR	0.0913	22	102	57.65%
DP	→	DP	0.0857	33	204	47.01%
WD	→	WT	0.1137	17	102	31.78%

Table 5.10: Experimental results using the amino acid sequences in the KA1 protein family



Pair of Amino Acids	*Theoretical Probability (T)*	*Actual Probability (E)*		*% Probability Difference (PD)*
From (P1) → To (P2)		*Frequency*	*Out of*	
PL → PR	0.0155	16	105	89.83%
LS → LS	0.0193	32	210	87.33%
VC → IV	0.0833	32	105	72.67%
LY → LH	0.0625	21	105	68.75%
YG → HG	0.082	26	105	66.88%
KL → RL	0.0725	21	105	63.75%
YK → YK	0.1947	53	105	61.43%
EI → EI	0.1956	50	105	58.92%
GV → GI	0.1361	33	105	56.70%
CK → VK	0.1581	38	105	56.31%
FE → FE	0.2976	69	105	54.71%
QF → QF	0.1337	30	105	53.21%
KF → RF	0.103	23	105	52.98%
KV → KV	0.1565	34	105	51.67%
VC → VC	0.1547	32	105	49.24%
EV → EV	0.1666	34	105	48.55%
KR → QR	0.0863	17	105	46.70%
VP → LP	0.1226	24	105	46.36%
EL → EL	0.2093	39	105	43.65%
KV → KL	0.084	15	105	41.20%
KR → RR	0.1194	21	105	40.30%
IL → IL	0.2205	36	105	35.69%
GD → GN	0.1702	27	105	33.81%
RI → RV	0.1549	24	105	32.23%
GV → GV	0.2467	38	105	31.83%
FK → FK	0.2795	38	105	22.77%
RI → RL	0.16	21	105	20.00%
KR → KR	0.3238	40	105	15.00%
VP → VP	0.2283	28	105	14.39%
KF → KF	0.2795	33	105	11.07%

5.4 A Partial Explanation of Anomalies in Pairwise Mutations

In order to better understand why the pairwise mutations that we found are anomalously more frequent than expected, we investigated the frequency distribution of the various amino acids in the proteins. The following figures (Figure 5.1 to Figure 5.3) are probability bar charts showing the total number of possible outcomes of each amino acid in the sample protein family sequences. The amino acids are along the x-axis and the total possible outcomes (in numbers) are along the y-axis.

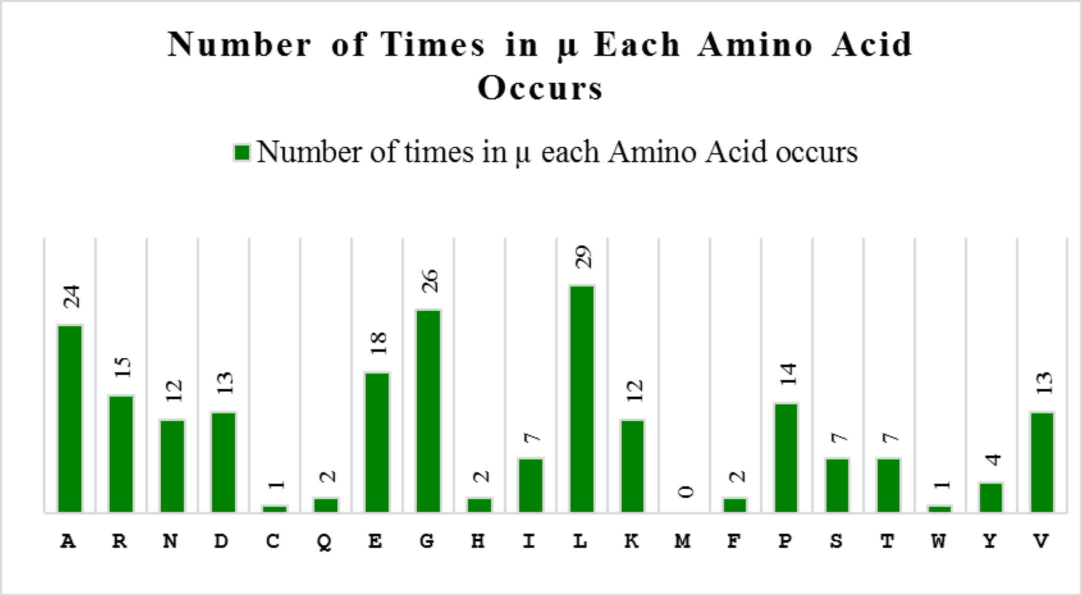


Figure 5.3: A Bar chart showing the number of times in µ each amino acid appears for the protein family DAGK_cat

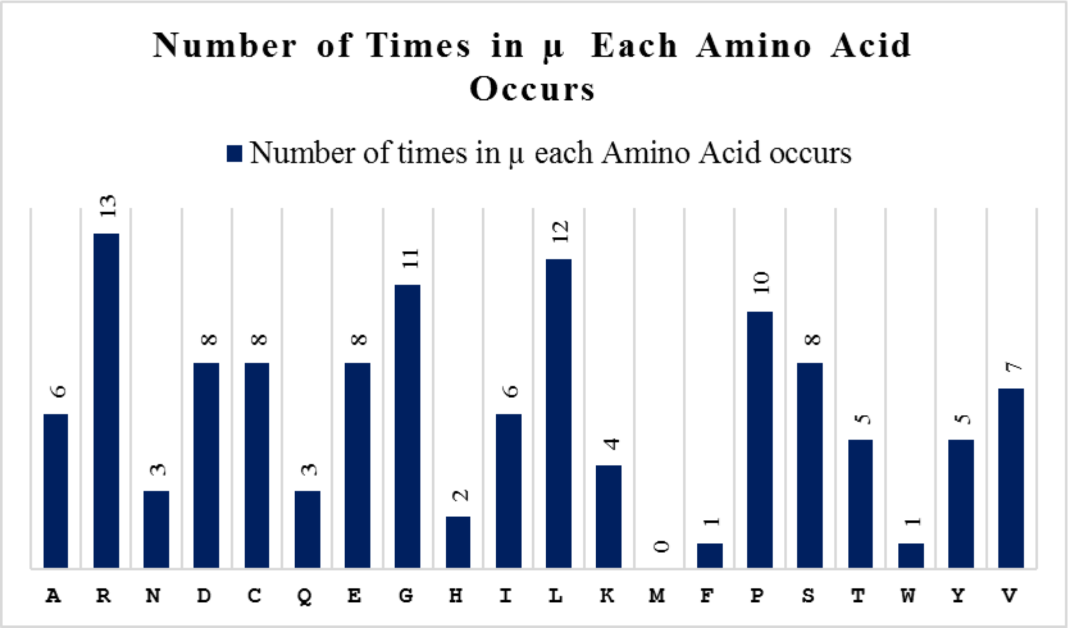


Figure 5.4: A Bar chart showing the number of times in µ each amino acid appears for the protein family IL17

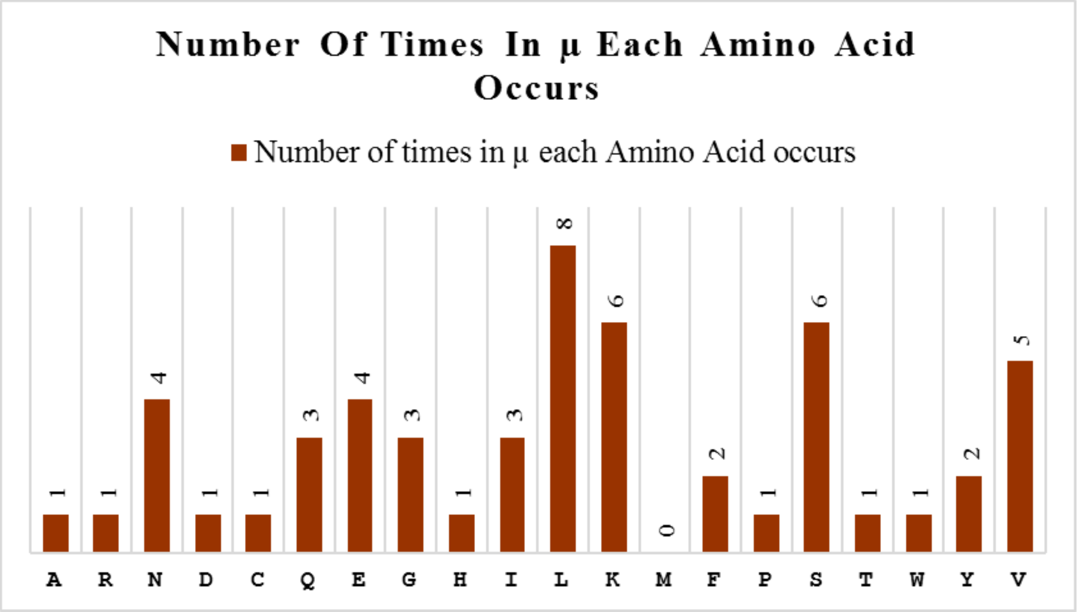


Figure 5.5: A Bar chart showing the number of times in µ each amino acid appears for the protein family KA1

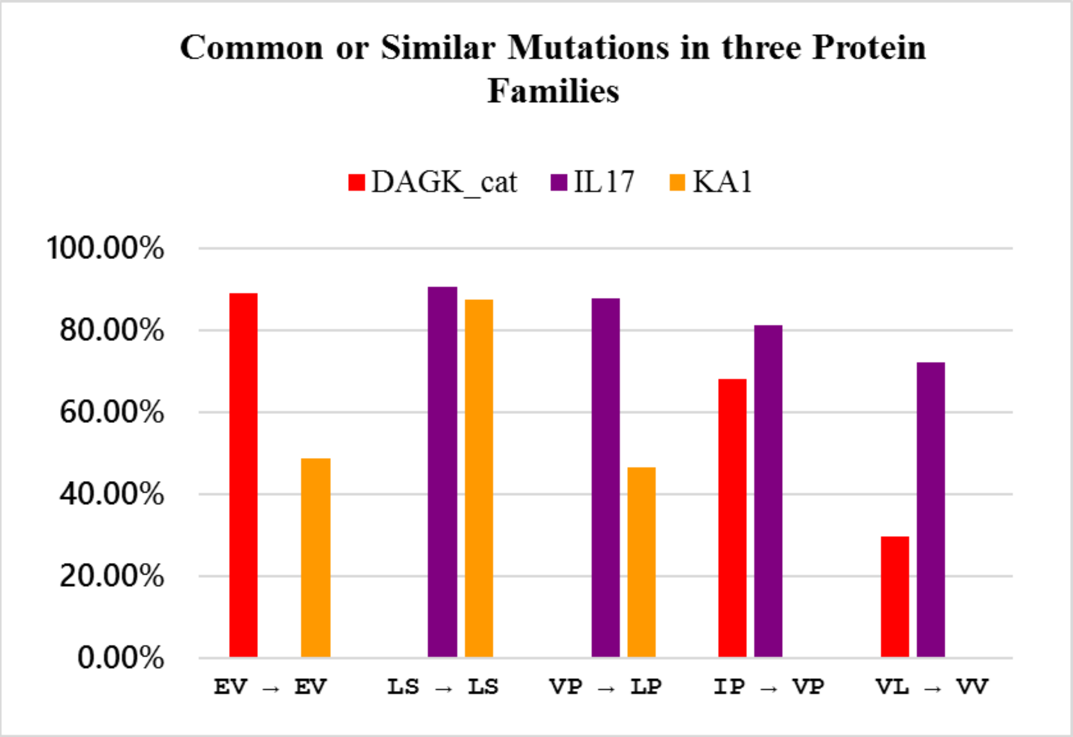


Figure 5.6: A Bar chart showing the Common or Similar Mutations in Three Protein Families

The figure above is a pictorial representation of the findings shown in Table 5.7. This table shows all the pairwise mutations that had seemed to be preserved in at least one of the other protein family in our data source, with range of anomalous probability in each of the protein families, shown with different color components. An interesting question is to know why these pairs occur in two protein families which probably might be due to the chemical properties of the nucleotides or the evolutionary distances among them.

Chapter 6

Conclusions and Future Work

The experimental results in Chapter 5 suggest that adjacent pairs of amino acids in the surviving descendants are sometimes mutated in a dependent way instead of an independent way. Since the probability of overlap mentioned under Section 5.2.3 seems to be small about ≤ 0.0001 and evidently lesser than out P-value which about ≤ 0.0035 implies that we have a concrete proof that our findings cannot be explained as a random event. This shows that the anomalies we found are not accidental but are some consequence of the chemical nature of these particular amino acid pairs and evolutionary forces acting on those pairs. Moreover, the above low probability is just for finding at least one common pairwise mutation whereas we have found three of them plus two other pairs that are complements of each other. From the overall set of experiments, we can infer that the pairwise mutations of a protein sequence in a protein family does not have to be independent all the time. However, the experimental data is based only on three protein families.

In the future we plan to use our independence testing method on other protein families that has more than a thousand see sequences. We plan to experiment with the sequences aligned with formats other than FASTA and also considering other evolutionary distances among the sequences apart from PAM 250. We also plan to look at longer sequences, that is, consider adjacent N-mers of amino acids for N > 2. The results can be analyzed in depth by considering the biological factors of the amino acids such as its properties - hydrophilic/hydrophobic, aliphatic/aromatic and see how such properties impact the independence assumption that is the key idea in this research.

Bibliography

[1] “The pam 250 matrix,” http://www.bioinformatics.org/wiki/Scoring_matrix.

[2] “The pfam protein library,” http://pfam.xfam.org/.

[3] D. A. Baum and S. D. Smith, *Tree thinking: an introduction to phylogenetic biology.* Roberts, 2013.

[4] C. Brocker, D. Thompson, A. Matsumoto, D. W. Nebert, and V. Vasiliou, “Evolutionary divergence and functions of the human interleukin (il) gene family,” *Human genomics*, vol. 5, no. 1, p. 1, 2010.

[5] M. K. Derbyshire, N. R. Gonzales, S. Lu, J. He, G. H. Marchler, Z. Wang, and A. Marchler-Bauer, “Improving the consistency of domain annotation within the conserved domain database,” *Database*, vol. 2015, p. bav012, 2015.

[6] E. S. Donkor, N. T. Dayie, and T. K. Adiku, “Bioinformatics with basic local alignment search tool (blast) and fast alignment (fasta),” *Journal of Bioinformatics and Sequence Analysis*, vol. 6, no. 1, pp. 1–6, 2014.

[7] B. G. Hall, *Phylogenetic trees made easy: A how to manual.* Sinauer,, 2011, no. 576.88 H174p.

[8] M. Krauthammer, A. Rzhetsky, P. Morozov, and C. Friedman, “Using blast for identifying gene and protein names in journal articles,” *Gene*, vol. 259, no. 1, pp. 245–252, 2000.

[9] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz *et al.*, “Cdd: Ncbi’s conserved domain database,” *Nucleic acids research*, p. gku1221, 2014.

[10] I. Mérida, A. Ávila-Flores, and E. Merino, “Diacylglycerol kinases: at the hub of cell signalling,” *Biochemical Journal*, vol. 409, no. 1, pp. 1–18, 2008.

[11] K. Nishikawa, Y. Kubota, and O. Tatsuo, “Classification of proteins into groups based on amino acid composition and other characters. ii. grouping into four types,” *Journal of biochemistry*, vol. 94, no. 3, pp. 997–1007, 1983.

[12] L. R. Novick, K. M. Catley, and E. G. Schreiber, “Understanding evolutionary history: An introduction to tree thinking,” 2012.

[13] W. R. Pearson, “Finding protein and nucleotide similarities with fasta,” *Current protocols in bioinformatics*, pp. 3–9, 2004.

[14] J. Ramanan and P. Z. Revesz, “Mutations of adjacent amino acid pairs are not always independent,” in *Mathematical Models and Computational Methods.* INASE, Oct. 2015.

[15] P. Revesz, *Introduction to database: From Biological to Spatio-Temporal.* Springer, 2010.

[16] P. Z. Revesz, “An algorithm for constructing hypothetical evolutionary trees using common mutation similarity matrices,” in *Proc. 4th ACM International Con-*

ference on Bioinformatics and Computational Biology (ACM BCB). Bethesda, MD, USA, September 2013: ACM Press, 2013, pp. 731–734.

[17] M. Salemi and A.-M. Vandamme, *The phylogenetic handbook: a practical approach to DNA and protein phylogeny.* Cambridge University Press, 2003.

[18] M. D. Shortridge, T. Triplet, P. Revesz, M. A. Griep, and R. Powers, "Bacterial protein structures reveal phylum dependent divergence," *Computational biology and chemistry*, vol. 35, no. 1, pp. 24–33, 2011.

[19] V. K. Sohpal, A. Singh, and A. Dey, "Optimization of substitution matrix for sequence alignment of major capsid proteins of human herpes simplex virus," *International Journal of BioAutomation*, vol. 15, pp. 277–284, 2012.

[20] G. D. Stormo, "An introduction to sequence similarity (âĂIJhomologyâĂİ) searching," *Current Protocols in Bioinformatics*, pp. 3–1, 2009.

[21] J.-P. Tassan and X. Goff, "An overview of the kin1/par-1/mark kinase family," *Biology of the Cell*, vol. 96, no. 3, pp. 193–199, 2004.

[22] C.-H. Tsai and R. E. Fordyce, "Ancestor–descendant relationships in evolution: origin of the extant pygmy right whale, caperea marginata," *Biology letters*, vol. 11, no. 1, p. 20140875, 2015.

[23] W. J. Wilbur, "On the pam matrix model of protein evolution." *Molecular biology and evolution*, vol. 2, no. 5, pp. 434–447, 1985.