Summer 7-15-2017

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

Peter Revesz
University of Nebraska-Lincoln, prevesz1@unl.edu

Testing the Independence Hypothesis of Accepted Mutations for Pairs of Adjacent Amino Acids in Protein Sequences

Jyotsna Ramanan and Peter Z. Revesz

***Abstract*—**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.

***Keywords*—**amino acid, independent probabilities, nucleotide, genetic mutation, protein.

I. Introduction

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.

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*.

The PFAM library [4] 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 (see the textbooks by Baum and Smith [1], Hall [2], and Lerney et al. [3] for an overview of these algorithms.) Revesz [5] has recently proposed a new phylogenetic tree-building algorithm called the *Common Mutation Similarity Matrixes* (CMSM) algorithm. This algorithm finds a hypothetical evolutionary tree. The first step of the CMSM algorithm is to find a hypothetical common ancestor, which is denoted by μ.

In this paper, 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 can also 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

The work of the second author was supported in part by a Fulbright Scholarship sponsored by the U.S. Department of State while he was on leave from the University of Nebraska-Lincoln.

Jytosna Ramanan was with the Department of Computer Science and Engineering of the University of Nebraska-Lincoln, while she was an M.S. student. Now she is working as a software engineer in Chicago, IL, USA (email: jor.compscie@gmail.com).

Peter Z. Revesz is a professor in the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588 USA (phone: 402-472-3488; fax: 402-472-7767; email: revesz@cse.unl.edu). For part of this research he was on a Fulbright Scholarship leave at the Aquincum Institute of Technology in Budapest, Hungary.

deviation from the theoretical probability is significant, then the independence assumption fails to provide a satisfying explanation for the experimental results.

Evolutionary studies usually assume that the genetic mutations are independent of each other. This paper tests the independence hypothesis for genetic mutations with regard to protein coding regions. As discussed in Section III, according to our experimental results the independence assumption generally holds, but there seem to be certain exceptions. We give examples in Section III of some particular adjacent amino acid pairs that seem to change in ways that deviate significantly from the expected theoretical probability under the independence assumption.

This paper is organized as follows. Section II describes our method with an extended example. Section III describes the protein families that were used in the experiments. Section IV presents our experimental results. Finally, Section V gives some conclusions and directions for further research.

II. The Independence Testing Method

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 Fig. 1. 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 S_1… S_7 are as shown in Fig. 1.

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

Fig. 1 A set of seven artificial sequences

Our testing method consists of five steps that are explained in Sections (A-E).

A. Find Consensus Sequence

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 one for which the sum of the PAM250 matrix distance values between it and the amino acids in the column considered is minimal. If there are two or more values that are minimal, then we make a random selection.

For the example in Fig. 1, consisting of seven artificial sequences from S_1, S_2, ... S_7, each with a length of twenty nucleotides, the consensus sequence is as shown in Fig. 2.

μ	RNARDANDRADNRDANRNAA

Fig. 2 The consensus sequence for the artificial protein family in Fig. 1

B. Calculate Mutation Probability Matrix

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 Fig. 1, the mutation probability matrix is shown in Table 1.

The mutation Probability Matrix in Table 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 (S_1 to S_7).

Table 1 The mutation probability matrix for the data in Fig.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

C. Find Theoretical Probabilities

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 Fig. 1 mutating into another possible pair of the same set are shown in Table 2. Note that the table values are in decimal format for the purpose of calculation.

D. Find Actual Probabilities

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 Fig. 1 with the seven artificial sequences are shown in Table 3.

Note that in Table 3, 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 Fig. 1, NR appears in two indices as underlined in Fig. 3 below.

μ	RNARDANDRADNRDANRNAA

Fig. 2 The consensus sequence for the artificial protein family in Fig. 1

In this case, the total number of possibilities of NR changing into another pair is 2 * 7 = 14, where seven is the total number of sequences of the protein family.

Algorithm ACTUAL-PROBABILITY (S, n, m)

INPUT: The set S of aligned sequences of a protein family represented as a matrix. The sequences are S[1][1…m], S[2][1…m], …, S[n][1…m] where *n* denotes the total number of sequences and *m* denotes their lengths.

//TOT is the overall total number of possible ways a //particular pair can mutate to another pair.

//The auxiliary function *Find_Consensus_Sequence(S)* finds //the consensus sequence of S.

```
1     C := Find_Consensus_Sequence(S)

2     for i → 1 to m-1 do
3         Calculate the count and index of all the
          adjacent pairs in the consensus sequence
4         TOT := count * n
5     end for

6     for i → 1 to m-1 do
7       for j → 2 to n do
8         Calculate the occurrences of possible pairs
          in the descendent sequences corresponding
          to the column S[i][i+1] which is the
          consensus sequence.
14      end for
15    end for
```

The algorithm ACTUAL-PROBABILITY is a dynamic programming computer algorithm. We iterate through the consensus sequence *m* number of times for each adjacent pair in the consensus sequence. During 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 that start on lines 6 and 7 in the algorithm. Line 8 calculates the occurrences of the pair in the consensus mutating into one of the possible 400 pairs in the descendent sequences. That takes about *n* number of comparisons in the worst case. Hence we can prove the following theorem.

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

Table 2 The theoretical probabilities of changes for each pair of amino acids for the artificial example protein family in Fig. 3.

	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 3 The actual probabilities of changes for each pair of amino acids for the artificial example protein family in Fig. 3.

	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**

E. Find Discrepancies between the Actual and Theoretical Probabilities

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 PD (p1, p2) be the percent probability difference defined as follows:

$$PD(p1,p2) = \frac{|T(p1,p2) - E(p1,p2)|}{Max(T(p1,p2), E(p1,p2))}$$

The percentage Probability Difference (PD) 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 below.

Table 4 Probability Differences for the protein sequences in Fig. 1

Pair of Amino Acids	*Theoretical Probability (T)*	*Actual Probability (A)*		*Probability Difference PD (T, A)*
From → To		*Freq*	*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%

III. The Protein Families Used in the Experiments

Our testing methods were also conducted on three other protein families for obtaining robust outcomes. We present the experimental values of the following protein families below. The PF series number indicates the serial number of the protein in the PFAM library (pfam.xfam.org) [4]. The PFAM library can be also browsed using the PROFESS protein database system [12] and references in the InterPro protein classification system [9].

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

Note that the theoretical probabilities and actual probabilities are presented in Tables 4- 6. In that, the theoretical probabilities of certain amino acids that had a negligible probability value (<0.05) were rounded to zeros as they seemed to be meager for large sets of protein sequences. The actual mutation probability matrices for the respective protein families are presented in Tables 7 – 9 at the end of the article. are using *Word,* use either the Microsoft Equation Editor or the *MathType* add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation *or* MathType Equation). "Float over text" should *not* be selected.

In the next three sections we describe the protein families that were used in the experiments.

A. The DAGK_cat Protein Family (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:

KALVIVNPKSGTARGGKGKKLLERKVRPLLEEAGVS
DDELDLRLTENPGPGDVLRRGYGNLEKLKSNALELL
AGAAREAAEANEQSDGDTLLPWSENLAYGYCPDLIV
AAGGDGTVNEVLNGLAGNARRDDLELATRNHPRAV
LVPSSPPLGIIPLGRTGNDFARALNAHGGFEEGIPLGY
DPEEAARAALELIKKIKGQTRPVDVGKV

B. The KA Protein Family (PF02149)

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:

LVVKFEIEVCKVPLLSGNSNSQEHLYGVQFKRINSGD
TWQYKNLASKILSELKL

C. The IL17 Protein Family (PF06083)

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:

RSLSPWDYREIDPHDPNRYPRVIAEARCLLCSGGSRCI
GDLNPATGQGEDDIAELQGLRRSLNSVPIYQEILVAF
LDGGGKLRRLCDKPCSRPKTHEPCAGCRYSYRLEPV
KETVTVGCTV

IV. Experimental Results

Table 5 shows the experimental results for the DAGK_cat protein family (PF00781).

Table 5 Anomalous probability differences for the DAGK_cat protein family for pairs of amino acids with or without mutations

Pair of Amino Acids	*Theoretical Probability (T)*	*Actual Probability (A)*		*Probability Difference (PD)*
From → To		*Freq*	*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%

Next, Table 6 shows the experimental results for the KA protein family (PF02149).

Table 6 Anomalous probability differences for the KA protein family for pairs of amino acids with or without mutations

Pair of Amino Acids	*Theoretical Probability (T)*	*Actual Probability (A)*		*Probability Difference (PD)*
From → To		*Freq*	*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%

Next, Table 7 shows the experimental results for the IL17 protein family (PF06083).

Table 7 Anomalous probability differences for the IL17 protein family for pairs of amino acids with or without mutations

Pair of Amino Acids	*Theoretical Probability (T)*	*Actual Probability (A)*		*Probability Difference (PD)*
From → To		*Freq*	*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%
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 8 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 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.

Table 8 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	

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, is:

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

In terms of number of permutations, this problem could be solved as:

$$1 - \frac{{}_nP_r}{{}_m P_r} = 1 - \frac{\frac{n!}{(n-r)!}}{\frac{m!}{(m-r)!}}$$

where m = 160000, n = 160000-31 and r =18. After the substitution of the respective values, we get:

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

Let us set the above-calculated value 0.0035 to be our P-value. As can be seen from Table 8, 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. Figs. 4 and 5 show the calculations and the statistical results generated using SAS for our proof.

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

Fig. 4 SAS results showing the probability of finding at least one common pairwise mutation 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

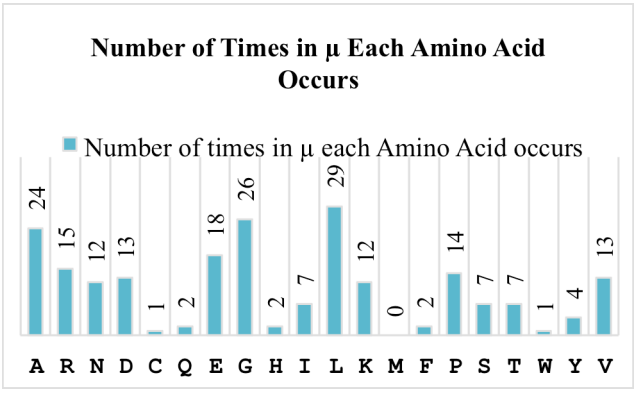
Statistics for Table of p1 by p2

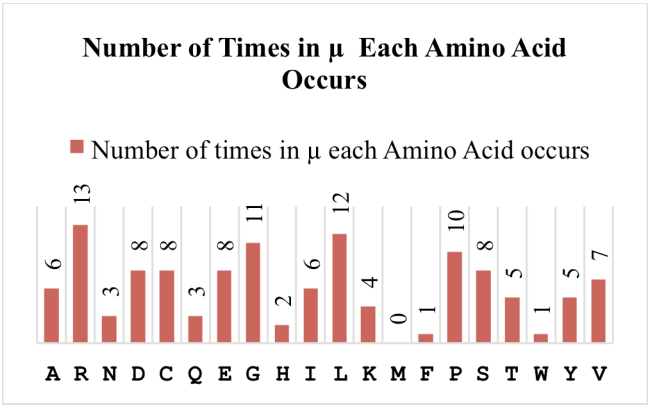
Statistic	DF	Value	Prob
Chi-Square	1	7160.6551	<.0001
Likelihood Ratio Chi-Square	1	65.0769	<.0001
Continuity Adj. Chi-Square	1	5799.2310	<.0001
Mantel-Haenszel Chi-Square	1	7160.6103	<.0001
Phi Coefficient		0.2116	
Contingency Coefficient		0.2070	
Cramer's V		0.2116	
WARNING: 25% of the cells have expected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

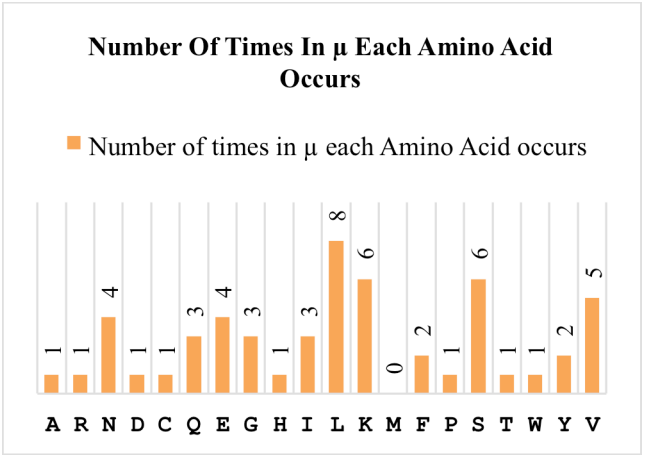
Fisher's Exact Test	
Cell (1,1) Frequency (F)	159956
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Fig. 5 Finding 5 common pairwise mutations out of the top 31 IL17 mutations, the top 18 KA1 mutations and the top 31 DAGK_cat mutation

Since the above probability of overlap is so small, our finding 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.







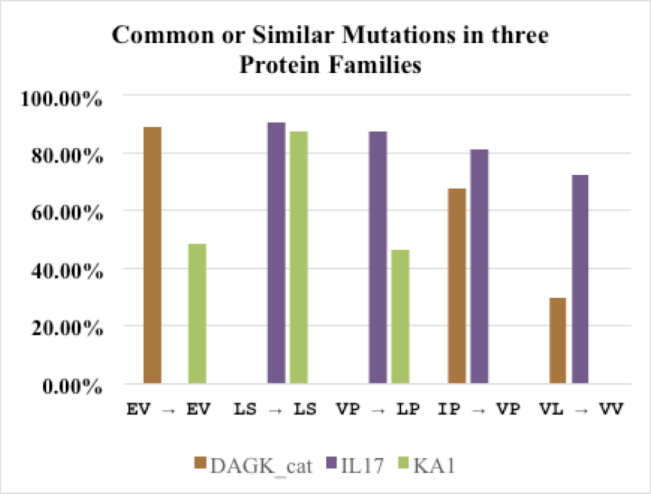


Fig. 6 Histograms showing the number of times each amino acid appears in μ for the protein family DAGK_cat (blue), IL17 (red) and KA (orange), and the histogram showing the percent probability differences (PDs) for the mutations that are commonly anomalous in at least two of the three protein families.

A. Charts

Fig. 6 shows the histograms of the probability of each amino acid in the sample protein families. The amino acids are along the x-axis and the total possible outcomes (in numbers) are along the y-axis.

V. Conclusion and Future Work

Large The experimental results suggest that adjacent pairs of amino acids in the surviving descendants are sometimes mutated in a dependent instead of an independent way. Since the probability of overlap 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 for many other protein families. We also plan to experiment with using other amino acid substitution matrixes beside the PAM250 matrix [6]. We also plan to look at longer sequences, that is, consider adjacent N-mers of amino acids for $N > 2$.

Finally, it is an intriguing question what the changed view of the amino acid similarity table implies about evolution [17].

Acknowledgment

We thank Dr. Stephen D. Kachman of the Department of Statistics at University of Nebraska- Lincoln for his pertinent

help in this research. One of us (J.R) would like to thank Dr. Juan Cui and Dr. Stephen Scott for investing their time to serve in the M.S. thesis committee.

REFERENCES

[1] D. Baum and S. Smith, *Tree Thinking: An Introduction to Phylogenetic Biology*, Roberts and Company Publishers. 2012.
[2] B. G. Hall, *Phylogenetic Trees Made Easy: A How to Manual*, 4th edition, Sinauer Associates, 2011.
[3] P. Lerney, M. Salemi, and A.-M Vandamme, editors. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, 2nd edition, Cambridge University Press, 2009.
[4] The PFAM Protein Library. Available: http://pfam.xfam.org/family/PF16506
[5] P. Z. Revesz, "An algorithm for constructing hypothetical evolutionary trees using common mutations similarity matrices," *Proc. 4th ACM International Conference on Bioinformatics and Computational Biology*, ACM Press, Bethesda, MD, USA, September 2013, pp. 731-734.
[6] P. Z. Revesz, *Introduction to Databases: From Biological to Spacio-Temporal*, Springer, 2010.
[7] H. M. Bakali, M. D. Herman, K. A. Johnson, A. A. Kelly, A. Wieslander, B. M. Hallberg, and P. Nordlund; "Crystal structure of YegS, a homologue to the mammalian diacylglycerol kinases, reveals a novel regulatory metal binding site," *J. Biol Chem* 282:19644-19652, . 2007.
[8] I. Merida, A. Avila-Flores, and E. Merino, "Diacylglycerol kinases: at the hub of cell signaling," *Biochem. J.* 409 (1): 1–18. doi:10.1042/BJ20071040, 2008.
[9] InterPro Protein sequence analysis & classification. Available: http://www.ebi.ac.uk/interpro/references.html
[10] C. Brocker, D. Thompson, A. Matsumoto, D. W. Nebert, V. Vasiliou "Evolutionary divergence and functions of the human interleukin (IL) gene family." *Human Genomics* 5 (1), 2010.
[11] S. Aggarwal and A. L. Gurney "IL-17: prototype member of an emerging cytokine family," *J. Leukoc. Biol.* 71(1), 2002.
[12] T. Triplet, M. D. Shortridge, M. A. Griep, R. Powers and P. Revesz "PROFESS: A PROtein Function, Evolution, Structure and Sequence database." *Database*, 2010:baq011. *PMC20624718, 2010.*
[13] J. P. Tassan and X. Le Goff "An overview of the KIN1/PAR-1/MARK kinase family," *Biol. Cell* 96 (3): 193–9, 2004.
[14] J. Biernat, Y. Z. Wu, T. Timm, Q. Zheng-Fischhofer, E. Mandelkow, L. Meijer, E. M. Mandelkor (November 2002). "Protein kinase MARK/PAR-1 is required for neurite outgrowth and establishment of neuronal polarity". Mol. Biol. Cell 13 (11): 4013–28.
[15] S. Guo and K. J. Kemphues (May 1995). "par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed". Cell 81 (4): 611–20.
[16] M. Elbert, G. Rossi and P. Brennwald "The yeast par-1 homologs kin1 and kin2 show genetic and physical interactions with components of the exocytic machinery," *Mol. Biol. Cell* 16 (2): 532–49, 2005.
[17] M. D. Shortridge, T. Triplet, P. Revesz, M. Griep and R. Powers "Bacterial Protein Structures Reveal Phylum Dependent Divergence," *Comput. Biol. Chem.*, 35:24-33. PMC3049983, 2011.

Jyotsna Ramanan (MS'16) was a graduate student in the Department of Computer Science and Engineering at the University of Nebraska-Lincoln and studied previously at Pondicherry University, Pondicherry, India.

Her research interests are in bioinformatics, big data and database system concepts. She is currently working as a software engineer in Chicago.

Peter Z. Revesz (Ph.D'91) holds a Ph.D. degree in Computer Science from Brown University and was a postdoctoral fellow at the University of Toronto.

He is an expert in databases, data mining, big data analytics and bioinformatics. He is the author of *Introduction to Databases: From Biological to Spatio-Temporal* (Springer, 2010) and *Introduction to Constraint Databases* (Springer, 2002). He is currently a professor in the Department of Computer Science and Engineering at the University of Nebraska-Lincoln, Lincoln, NE 6815, USA.

Dr. Revesz also held visiting appointments at the Aquincum Institute of Technology, the IBM T. J. Watson Research Center, INRIA, the Max Planck Institute for Computer Science, the University of Athens, the University of Hasselt, the U.S. Air Force Office of Scientific Research and the U.S. Department of State. He is a recipient of an AAAS Science & Technology Policy Fellowship, a J. William Fulbright Scholarship, an Alexander von Humboldt Research Fellowship, a Jefferson Science Fellowship, a National Science Foundation CAREER award, and a "Faculty International Scholar of the Year" award by *Phi Beta Delta*, the Honor Society for International Scholars.