10-1-2015

Mutations of Adjacent Amino Acid Pairs are not Always Independent

Jyotsna Ramanan
University of Nebraska-Lincoln, jramanan@cse.unl.edu

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

Mutations of Adjacent Amino Acid Pairs are not Always Independent

Jyotsna Ramanan and Peter Z. Revesz

***Abstract*—**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. According to the new experimental results the independence assumption generally holds, but there are certain exceptions. In particular, the coding regions that represent two adjacent amino acids seem to change in ways that sometimes deviate significantly from the expected theoretical probability under the independence assumption.

***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 near by 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 order 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.

Jyotsna Ramanan (jramanan@cse.unl.edu) and Peter Z. Revesz (revesz@cse.unl.edu). are with the Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588, USA

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 at al. [3] for an overview of these algorithms.) Revesz [5] has proposed recently 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 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.

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 IV, according to our experimental results the independence assumption generally holds, but there seem to be certain exceptions. We give examples in Section IV 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 some background concepts about hypothetical common ancestors. Section III describes our method with an extended example. Section IV presents our experimental results. Finally, Section V gives some conclusions and directions for further research.

ISBN: 978-1-61804-350-4

II. Background Concepts Related to Hypothetical Common Ancestor

Consider the seven amino acid sequences, S_1...S_7 shown in Figure 1 below. Theses seven amino acid sequences belong to the protein family DiSB-ORF2_chro (PFAM library identification number PF16506) [1]. The sequences as shown in Figure 1 are already aligned with each other.

S_1	SPYMFDRSCLNVYRTNDYLFGECTLPPNCSEPSVVKLDKTFYGQETVVCHS
S_2	TPYVFDRECLSVYRTNDWFFSQCSLPPNCTNPSVVKLERTFFGQETVVCHS
S_3	SPFEFDPECIEVHRTHSWFFQGCTLPPSCGDVHTKILDSSF-GFKELMCYS
S_4	SPYMFDRSCLNVYRTNDYLFGECTLPPNCSEPSVIKLDKTFYGQETVVCHS
S_5	SPYHTDPTCVSVYRTNDWFFAGCELPPHCLGKVVSIIEKKWYGQETVFCHS
S_6	SPFEFDPECIEVHRTHSWFFQGCTLPPSCGDVHTKILDSSF-GFKELMCYS
S_7	SPYVFDRSCLNVYRTNDYLFGECTLPSNCSEP----------

Fig.1 Seven example proteins from the protein family DiSB-ORF2_chro

For the above set of amino acid sequences, the CMSM algorithm [5] generates the hypothetical evolutionary tree shown in Figure 2.

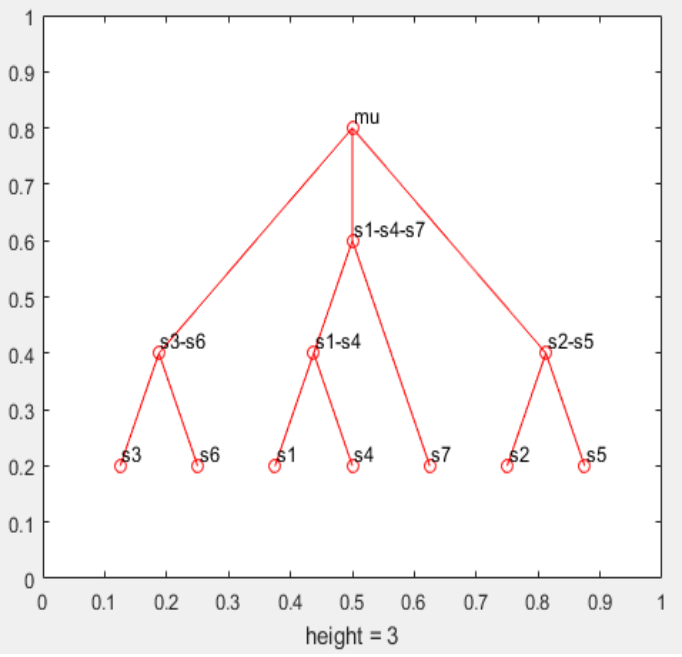

Fig. 2 The hypothetical evolutionary tree for the DiSB-ORF2_chro generated using the CMM algorithm

In Figure 2, the variables S_1, S_2 ..S_7 correspond to the seven amino acid sequences that are listed in Figure 1. The variable mu (μ) is the hypothetical common ancestor that was generated using the CMSM algorithm. The value of μ was found to be the following string of amino acids.

μ= SPYEFDRECLNVYRTNDWFFGECTLPPNCSEPSVK ILDKTFYGQETVVCHS

III. 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 Figure 3. 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 S_1, …, S_7 as shown also in Figure 3.

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

Fig. 3 A set of seven artificial sequences and their hypothetical common ancestor

Our testing method consists of the following five steps.

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 algorithm in the case of amino acid sequences. 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 organizes) 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.

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 Figure 3, the mutation probability matrix is shown in Table 1.

Table 1 The mutation probability matrix for the data in Figure 3.

	A	D	N	R
A	18/35	6/35	7/35	4/35
D	7/35	11/35	4/35	15/35
N	12/35	3/35	18/35	5/35
R	4/35	11/35	3/35	11/35

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 an NN can be computed as follows:

$$Prob(AD, NR) = Prob(A, N) * Prob(N, N) = \frac{3}{35}$$

4. The actual probabilities of changes are calculated for each pair of amino acids. The results for our example are shown in Table 2.

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:

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

IV. Experimental Results

The experimental results shown in Table 3 are based on the protein family DiSB-ORF chro, which has a PFAM identification number PF16506. Table 3 displays only the top ten highest percentage probability differences that we found.

Table 3 Experimental results using the amino acid sequences in the DiSB-ORF chro protein family

Pair of amino acids		Theoretical probability under independence assumption T(p1, p2)	Actual probability E(p1, p2)	Percent probability difference PD(p1, p2)
From p1	To p2			
SP	SP	289/19600	6/7	1.0528
SP	PP	289/19600	6/7	1.0528
SP	TP	34/19600	1/7	0.329
SP	PS	119/19600	1/7	0.975
ER	ER	17/19600	2/7	0.997
ER	DV	17/19600	3/7	0.998
ER	GK	12/19600	1/7	0.995
DR	DP	260/19600	3/7	0.969
DR	DR	281/19600	4/7	0.975

V. Conclusion and Future Work

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. However, the experimental data is based only on one protein family. 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.

ISBN: 978-1-61804-350-4

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

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

ISBN: 978-1-61804-350-4

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.

Jyotsna Ramana is currently a graduate student in the Department of Computer Science and Engineering at the University of Nebraska-Lincoln. Her research interests are in bioinformatics, big data and database systems.

Peter Z. Revesz holds a Ph.D. degree in Computer Science from Brown University. He was a postdoctoral fellow at the University of Toronto before joining the University of Nebraska-Lincoln, where he is a professor in the Department of Computer Science and Engineering. Dr. Revesz 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). Dr. Revesz held visiting appointments at 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.