Prospect

ibc Interdisciplinary Bio Central | Open Access, Open Review Journal

Controversies on governing the rates of protein evolution

Sun Shim Choi[1,2,*]

[1]Department of Molecular and Medical Biotechnology, Kangwon National University, Chunchon 200-701, Korea
[2]Institute of Bioscience & Biotechnology, Kangwon National University, Chunchon 200-701, Korea

Subject areas: Bioinformatics, evolutionary genomics

Author contribution: Choi SS designed and wrote this manuscript.

***Correspondence** and requests for material should be addressed to S.S.C. (schoi@kangwon.ac.kr).

Reviewer: Keun Woo Lee, Gyeongsang National University, Republic of Korea. Hyun Joo, Inje Univ. College of Medicine, Republic of Korea.

Editor: Dongsup Kim, KAIST, Republic of Korea

Received July 23, 2009;
Accepted July 30, 2009;
Published August 05, 2009

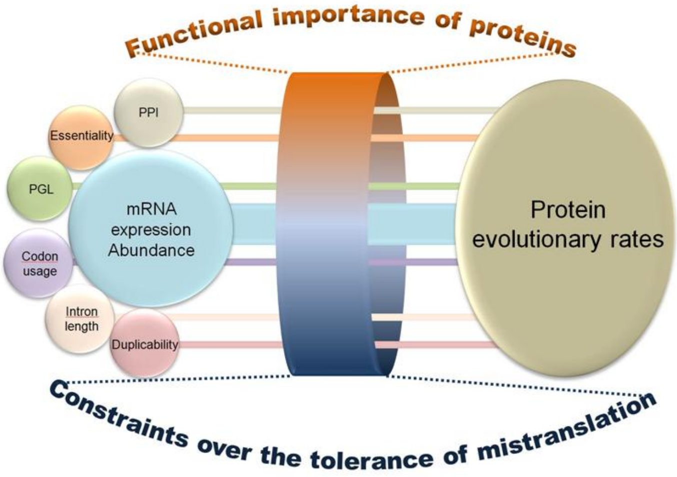
Citation: Choi, S. Comparison of the two different hypotheses on protein evolutionary rates. IBC 2009, 1(3):11, 1-5. doi:10.4051/ibc.2009.3.0011

Funding: This work was supported by the Korea Research Foundation Grant funded by the Korea Government (MEST, Basic Research Promotion Fund).

Competing interests: All authors declare no financial or personal conflict that could inappropriately bias their experiments or writing.

SYNOPSIS

One of the main issues of molecular evolution is to reveal the principles dictating protein evolutionary rates. A traditional hypothesis posits that protein evolutionary rates are mostly determined by the average functional importance of amino acids in a given protein. Thus the correlations of evolutionary rates with different variables such as PPI, gene essentiality and expression abundance have been studied to test the traditional hypothesis. Recently, mRNA expression abundance among the variables has drawn much attention, not only because it shows relatively strong correlation with protein evolutionary rates, but also because of the controversies surrounding an alternative hypothesis against the traditional one. Here, I will give an overview over the traditional hypothesis, and summarize the different variables that have been found to correlate with protein evolutionary rates. Then I will introduce pros and cons on the two different hypotheses.


Functional importance of proteins
PPI
Essentiality
PGL
mRNA expression Abundance
Codon usage
Intron length
Duplicability
Protein evolutionary rates
Constraints over the tolerance of mistranslation

Keywords: evolutionary rates, mRNA expression, functional importance

Introduction

Classically, it has long been hypothesized that the functional importance of amino acids and their density in a protein has a pivotal role in regulating the protein evolutionary rates. To address this, several groups have tried to show that the genome variables related to the functional importance of proteins such as gene essentiality, protein-protein interaction (PPI), and expression level correlate with the protein evolutionary rates. In fact, several variables show negative or positive correlations with the protein evolutionary rates even if the signals are not strong (Huang, Winter et al. 2004; Kumar 2005; Medina 2005). Among the variables, expression abundance has recently drawn much attention for two reasons (Pal, Papp et al. 2001; Drummond, Bloom et al. 2005; Wall, Hirsh et al. 2005; Koonin and Wolf 2006; Pal, Papp et al. 2006). First, the correlation of expression abundance with the evolutionary rates is quite strong as compared to other variables (Pal, Papp et al. 2001; Drummond, Bloom et al. 2005; Wall, Hirsh et al. 2005). Several studies have consistently shown that abundantly expressed genes evolve slowly (Pal, Papp et al. 2001; Krylov, Wolf et al. 2003; Subramanian and Kumar 2004; Wright, Yau et al. 2004). Second, an interesting novel hypothesis inferred from the prominent correlation for explaining the mechanisms underlying the correlations raised controversies (Akashi 2001; Rocha and Danchin 2004; Koonin and Wolf 2006; Rocha 2006). In this perspective, I will try to address two different views based on the two different hypotheses on explaining why the abundantly expressed proteins evolve slowly. The traditional hypothesis claims that the highly expressed proteins evolve slowly because they have greater fitness effect, while the novel hypothesis reject the idea that the rates are mostly controlled by the functional importance of proteins. Then I will present pros and cons about the two hypotheses and try to suggest what evidence should be presented to prove the hypotheses in the future studies.

Birth of the traditional hypothesis (on the protein evolutionary rates)

Evolution happens when a new mutation arises in a population and finally is fixed. In other words, it is called evolution when a new mutation completely replaces original allele in a population. Therefore, evolutionary rate is defined as the number of substitutions per site per year (Graur and Li). Since Darwin, it had long been believed that molecular evolution occurred in DNA level is mostly determined by the fitness effect of the new allele, i.e., if a new mutation is beneficial to population it will be fixed, otherwise it will disappear (Bamshad and Wooding 2003; Lynch 2007).

However, comparing the genomes of existing species, Kimura proposed that the vast majority of molecular differences are selectively "neutral", and the molecular changes represented by these differences therefore do not influence the fitness of the individual organism (Kimura 1968). The neutral theory originally suggested by Kimura argued that evolutionary rates are mostly determined by genetic drift not by selection. Given the random nature of genetic drift, the neutral theory predicts that protein sequence divergence correlates evolutionary distance between two species. Along the line, Zuckerkandl and Pauling (1965) demonstrated that the number of amino acid differences in globin protein between lineages scales roughly with divergence time, as is estimated from fossil evidence. They generalized this observation as 'molecular clock' hypothesis to assert that there is a uniform rate of molecular evolution over time and over different lineages. The existence of uniform molecular clock would strongly support the idea of neutral evolution (Kumar 2005).

Although evolutionary rates of a protein are relatively uniform among different lineages, evolutionary rates of different proteins vary in more than 3 orders of magnitude (Koonin and Wolf 2006; Rocha 2006; Plotkin and Fraser 2007). To explain this big difference of evolutionary rate, the neutral theory was revised into 'nearly neutral theory' allowing the effect of natural selection including purifying selection and positive selection in the original model (Ohta 1973; Kimura 1983). I would not cover positive selection issue in this perspective, even though it also contributes at least 20-45% of all amino acid substitution rates (Fay, Wyckoff et al. 2002; Bierne and Eyre-Walker 2004; Pal, Papp et al. 2006). According to the nearly neutral theory, proteins evolve with different rates because each protein has different proportion of amino acid residues of different functional importance (Zuckerkandl 1976; Kimura 1983; Wolf 2006). Purifying selection, a selection removing deleterious mutation in population, more strongly acts on more important amino acid residues (Zuckerkandl 1976; Koonin and Wolf 2006; Rocha 2006). The idea that functional importance dictate protein evolutionary rates has been a paradigm for over 30 years in molecular evolution ever since (Graur, Hide et al. 1991; McInerney 2006). That amino acid resides of high importance in protein function evolves slowly, so that proteins containing high proportion of important amino acid residues (i.e. more important proteins) evolve more slowly.

Variables associated with functional importance of proteins

The importance of proteins has been estimated by genetically engineered loss-of function mutant experiments using model organisms such as yeast, C. elegans, drosophila or mouse. The genes considered to be essential if the effect of their deletions causes the death of the mutants (i.e. lethality) (Hirsh and Fraser 2001; Jordan, Rogozin et al. 2002; Liang and Li 2007). More broadly, it is also considered to be important genes if the deletion of a gene affects on model animal's fitness (Hirsh and Fraser 2001). Based on phenotypic consequences of mutations on genes in model organisms, many studies have shown that non-essential genes evolve more rapidly than essential genes just as what the traditional hypothesis expects (Hirsh and Fraser 2001; Jordan, Rogozin et al. 2002; Yang, Gu et al. 2003; Wall, Hirsh et al. 2005; Park, Park et al. 2008). The correlations of evolutionary rates with other variables are also known to be fit to the traditional idea. For example, proteins interacting with more partners, proteins expressing in more diverse tissues, and proteins exerting more multiple functions evolve slowly. Different variables controlled by functional importance are known to be positively or negatively correlated with each other (Jeong, Mason et al. 2001; Wagner 2001; Fraser, Wall et al. 2003; Krylov, Wolf et al. 2003; Liang and Li 2007) (Figure1).

However, the traditional hypothesis fails to explain some correlations. For example, some variables which are barely related to protein functions such as codon usage bias or intron size also show a correlation with the protein evolutionary rates: proteins with a high level of codon usage bias, (Sharp and Li 1986; Sharp and Li 1987) or proteins with a smaller size of intron are known to evolve slowly (Marais, Nouvellet et al. 2005) (Figure 1). Since there is no experimental evidence that the proteins with a higher codon usage bias, or a smaller size of intron are more important, it is not easy to understand the correlation between those variables and the protein evolutionary rates. One clue can be drawn from the observation that highly expressed proteins are more likely to use biased codons during translation, and to have a smaller intron (Marais and Duret 2001) (Figure 1).

Expression abundance is the most important parameter determining evolutionary rates.

There are some controversies on whether several variables have independent effects, or only one key variable carrying subsidiary other variables have major effects on the evolutionary rates. Wall et al. (2005) have argued that dispensability and expression have independent, significant effects on protein evolutionary rates: the

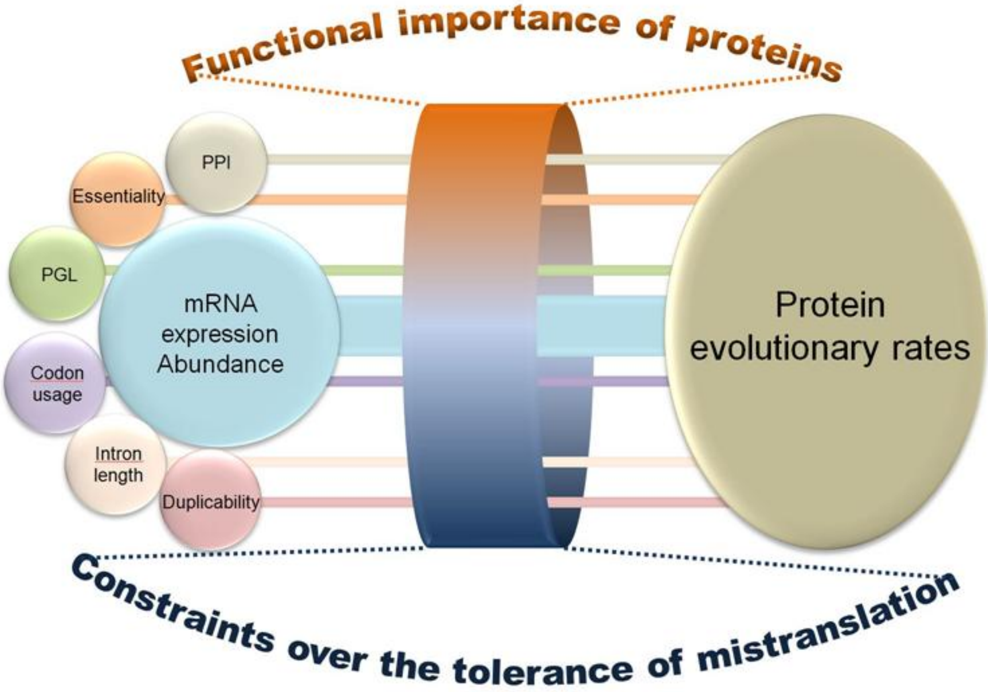

Functional importance of proteins
PPI
Essentiality
PGL
mRNA expression Abundance
Codon usage
Intron length
Duplicability
Protein evolutionary rates
Constraints over the tolerance of mistranslation

Figure 1. Schematic representations of two different views on the protein evolutionary rates.

variables (e.g., dispensability, number of protein-protein interactions, codon preference, expression breadth, dispensability) may exert independently small but cumulatively severe constraining effects on protein evolution. On the contrary, Drummond et al. (2005) have argued that mRNA expression abundance is the most prominent causing variable in determining evolutionary rates among all the variables (Figure 1). They argued that a single principal component, mRNA expression level, account for 43% of the variance in amino acid evolutionary rates. Against this Drummond group's argument, Plotkin and Fraser (2007) have claimed that the PCR analysis Drummond group performed was confounded by noise in biological data. One consistent finding in these controversies, however, is that abundantly expressed genes strongly negatively correlate with the evolutionary rates (Pal, Papp et al. 2001; Krylov, Wolf et al. 2003; Subramanian and Kumar 2004; Wright, Yau et al. 2004; Drummond, Bloom et al. 2005).

Two different views for explaining the correlation between expression abundance and evolutionary rates

a. The view of the traditional hypothesis: the more highly a protein is expressed, the more important its function is

I already mentioned above a traditional explanation, which is that highly expressed proteins evolve slowly because they have a higher importance in organism's fitness and survival. Rocha and Danchin (2004) posits that each protein molecule contributes a small amount to organism fitness by performing its function, so deleterious effects of mutations could be bigger in more abundant, causing the more abundant protein to evolve slower. However, Drummond et al. (2005) rejected this hypothesis by the observation that genes of highly abundant proteins from a low copy number of mRNA transcripts do not evolve more slowly than genes of lowly expressed proteins from a high copy number of mRNA transcripts.

b. The view of new hypothesis: highly expressed proteins evolve slowly not because they are more functionally important, but because they are tolerant to mistranslation

It is not as easy to understand as it first looks for why highly expressed genes evolve slowly Drummond et al. (2005;2006) suggested a very interesting translation-driven hypothesis, called 'translational robustness' hypothesis. The main idea of the hypothesis is that the natural selection pressure for tolerance to translational errors affects more strongly on highly expressed proteins than on lowly expressed proteins. The hypothesis is based on the observation that missense translation error rate is relatively high, up to 20%, and inactivated or misfolded proteins resulting from mistranslation are eventually toxic to the cell (Bucciantini, Giannoni et al. 2002). It is expected that the toxic effect can be small for proteins with low abundance, but might result in a serious problem for the highly expressed proteins. Therefore, this hypothesis predicts that the natural selection might favor rare amino acid sequences that reduce the risk of incorrect folding caused by error-prone translation process (Drummond, Bloom et al. 2005; Drummond, Raval et al. 2006; Wilke and Drummond 2006). The amino acids that contribute to tolerance would become highly conservative resulting in lower nonsynonymous evolutionary rates (dN or Ka).

There is another translation-driven hypothesis claiming that the abundant proteins evolve slowly by different selection constraints from the functional importance of proteins, which is based on the observation that synonymous substitution rates usually designated by dS or Ks of highly expressed genes become lower than those of lowly expressed genes (Akashi 2001; Akashi 2003). This hypothesis reasons that increased expression level leads to selection for optimal synonymous and nonsynonymous codons (i.e. codon usage bias), and may slow down the rates of proteins (so-called 'translational accuracy' hypothesis) (Akashi 2001; Akashi 2003). Certainly, these two translation-driven hypotheses challenge the traditional hypothesis, and may argue that highly expressed proteins evolve slowly not due to their functional importance.

Controversies surrounding the new hypothesis

A great deal of studies have shown that some specific amino acids in a given protein have more important function than the other amino acids, so that the genetic deletion or replacement experiments generate different consequences to cells or organisms depending on the importance of the amino acid residues in the protein. It has been observed that the amino acid residues located at the disease causing mutations are relatively conservative, and the genes that are related to human diseases are known to evolve slowly (Lopez-Bigas and Ouzounis 2004). The traditional hypothesis acknowledging amino acid functional importance rather than the new hypothesis is consistent with these experimental observations.

Furthermore, the translation-driven hypotheses were mostly derived from single cell organismal data such as S. cerevisiae (Drummond, Bloom et al. 2005). It is sure that multicellular organisms hold more complicated system not only in mRNA abundance regulation but also in temporal or spatial gene expression regulation and protein-protein interaction.

Another problem in the translation-driven hypotheses is that it is hard to know whether the mistranslated or misfolded proteins cause toxicity to cell, or whether misfolded proteins acquire new functions due to mutations and become deleterious to cell. If the resistance on the toxicity of misfolded proteins is the main subject for natural selection, then expression abundance can be only prominent determinant of evolutionary rates. Otherwise, other variables related to functional importance such as dispensability, pleiotropy, or PPI also provide independent contribution to the evolutionary rates of proteins.

Further, the negative correlation between nonsynonymous evolutionary rates and mRNA expression level still sustain even after all the preferred codons are removed in the calculation of evolutionary rates (Drummond, Bloom et al. 2005).

Conclusions

Determining the rate of protein evolution is one of the most powerful ways to understand how present genome make-ups in billions of organisms have been shaped. It also provides an opportunity to study the identification of functionally important sites, peptides that are involved in human genetic diseases, drug targets or protein interaction partners (Pal, Papp et al. 2006). Searching the variables for determining evolutionary rates is just one aspect of interesting encounters between evolution and a large-scale data. Among the variables, expression level has shown to be the most consistent and strong correlation with the rates of evolution. There are two different views on explaining the correlation. In this perspective, two different hypotheses on explaining the correlation between the expression abundance and the protein evolutionary rates were discussed. It would be essential to compare stability and aggregation risk between highly and lowly expressed genes through mutagenesis experiments (Pal, Papp et al. 2006) to reveal purifying selection pressure against protein misfolding.

Acknowledgements

I thank Donghyun Park and Seung Gu park for critically reading the manuscript and helpful comments. This work was supported by the program of Nuclear R&D program (2008-00903) of the Korea Science and Engineering Foundation (KOSEF).

References

Akashi, H. (2001). Gene expression and molecular evolution. Curr. Opin. Genet. Dev. 11(6), 660-6.

Akashi, H. (2003). Translational selection and yeast proteome evolution. Genetics 164(4), 1291-303.

Bamshad, M. and S. P. Wooding (2003). Signatures of natural selection in the human genome. Nat. Rev. Genet. 4(2), 99-111.

Bierne, N. and A. Eyre-Walker (2004). The genomic rate of adaptive amino acid substitution in Drosophila. Mol. Biol. Evol. 21(7), 1350-60.

Bucciantini, M., E. Giannoni, et al. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature 416(6880), 507-11.

Drummond, D. A., J. D. Bloom, et al. (2005). Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. USA 102(40), 14338-43.

Drummond, D. A., A. Raval, et al. (2006). A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 23(2), 327-37.

Fay, J. C., G. J. Wyckoff, et al. (2002). Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature 415(6875), 1024-6.

Fraser, H. B., D. P. Wall, et al. (2003). A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol. Biol. 3, 11.

Graur, D., W. A. Hide, et al. (1991). Is the guinea-pig a rodent? Nature 351(6328), 649-52.

Graur, D. and W. H. Li Fundamentals of molecular evolution, Sinauer Associates.

Hirsh, A. E. and H. B. Fraser (2001). Protein dispensability and rate of evolution. Nature 411(6841), 1046-9.

Huang, H., E. E. Winter, et al. (2004). Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol. 5(7), R47.

Jeong, H., S. P. Mason, et al. (2001). Lethality and centrality in protein networks. Nature 411(6833), 41-2.

Jordan, I. K., I. B. Rogozin, et al. (2002). Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12(6), 962-8.

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet Res. 11(3), 247-69.

Kimura, M. (1983). Diffusion model of intergroup selection, with special reference to evolution of an altruistic character. Proc. Natl. Acad. Sci. U S A 80(20), 6317-6321.

Koonin, E. V. and Y. I. Wolf (2006). Evolutionary systems biology: links between gene evolution and function. Curr. Opin. Biotechnol. 17(5), 481-7.

Krylov, D. M., Y. I. Wolf, et al. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13(10), 2229-35.

Kumar, S. (2005). Molecular clocks: four decades of evolution. Nat. Rev. Genet. 6(8), 654-62.

Liang, H. and W. H. Li (2007). Gene essentiality, gene duplicability and protein connectivity in human and mouse. Trends Genet. 23(8), 375-8.

Lopez-Bigas, N. and C. A. Ouzounis (2004). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. 32(10), 3108-14.

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. Nat. Rev. Genet. 8(10), 803-13.

Marais, G. and L. Duret (2001). Synonymous codon usage, accuracy of translation, and gene length in Caenorhabditis elegans. J. Mol. Evol. 52(3), 275-80.

Marais, G., P. Nouvellet, et al. (2005). Intron size and exon evolution in Drosophila. Genetics 170(1), 481-5.

McInerney, J. O. (2006). The causes of protein evolutionary rate variation. Trends Ecol. Evol. 21(5), 230-2.

Medina, M. (2005). Genomes, phylogeny, and evolutionary systems biology. Proc. Natl. Acad. Sci. U S A 102 Suppl. 1, 6630-5.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. Nature 246(5428), 96-8.

Pal, C., B. Papp, et al. (2001). Highly expressed genes in yeast evolve slowly. Genetics 158(2), 927-31.

Pal, C., B. Papp, et al. (2006). An integrated view of protein evolution. Nat. Rev. Genet. 7(5), 337-48.

Park, D., J. Park, et al. (2008). Analysis of human disease genes in the context of gene essentiality. Genomics.

Plotkin, J. B. and H. B. Fraser (2007). Assessing the determinants of evolutionary rates in the presence of noise. Mol. Biol. Evol.

24(5), 1113-21.

Rocha, E. P. (2006). The quest for the universals of protein evolution. Trends Genet. 22(8), 412-6.

Rocha, E. P. and A. Danchin (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Biol. Evol. 21(1), 108-16.

Sharp, P. M. and W. H. Li (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24(1-2), 28-38.

Sharp, P. M. and W. H. Li (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15(3), 1281-95.

Subramanian, S. and S. Kumar (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168(1), 373-81.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol. Biol. Evol. 18(7), 1283-92.

Wall, D. P., A. E. Hirsh, et al. (2005). Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. USA 102(15), 5483-8.

Wilke, C. O. and D. A. Drummond (2006). Population genetics of translational robustness. Genetics 173(1), 473-81.

Wolf, Y. I. (2006). Coping with the quantitative genomics 'elephant': the correlation between the gene dispensability and evolution rate. Trends Genet. 22(7), 354-7.

Wright, S. I., C. B. Yau, et al. (2004). Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol. Biol. Evol. 21(9), 1719-26.

Yang, J., Z. Gu, et al. (2003). Rate of protein evolution versus fitness effect of gene deletion. Mol. Biol. Evol. 20(5), 772-4.

Zuckerkandl, E. (1976). Evolutionary processes and evolutionary noise at the molecular level. II. A selectionist model for random fixations in proteins. J. Mol. Evol. 7(4), 269-311.