Conservation of Physicochemical Properties during Protein Evolution

Masashi Fujita `fujita@kuicr.kyoto-u.ac.jp`

Masumi Itoh `itoh@kuicr.kyoto-u.ac.jp`

Minoru Kanehisa `kanehisa@kuicr.kyoto-u.ac.jp`

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Keywords: amino acid composition, physicochemical property, active compensation

1 Introduction

Amino acid composition is one of the most basic characteristics of the proteome, and will provide various insights into protein chemistry and evolution. Bacterial organisms are known to have strikingly diversified amino acid compositions depending on the GC-content, habitat environment and energetic costs for biosynthesis. Therefore it is a fundamental question whether any kind of physicochemical properties are kept constant in spite of such diverse amino acid contents.

Here, we analyzed the physicochemical properties of amino acid compositions for 61 complete bacterial and archaeal genomes. Comparison of these properties with random mixtures of the compositions revealed that the average hydrophobicity and charge are well conserved during evolution, while average residue volumes are quite different among organisms. Further analysis on closely related species proved that such conservation could not be explained by a simple accumulation of independent mutation events. This implies that each protein actively compensates for any deviation in hydrophobicity and total charge in the course of evolution in order to stabilize itself or maintain total environment in the cell.

2 Materials and Methods

2.1 Composition Analysis

We selected 13 archaeal and 48 eubacterial organisms from NCBI Clusters of Orthologous Groups (COG) [4] and computed the amino acid composition of these organisms. Amino acid indices were downloaded from the KEGG AAindex database [3] and the average scores were calculated for each of these indices and organisms.

To determine whether an amino acid property p is conserved during evolution, we computed the deviation of the average p among organisms (σ_p). If σ_p is 'relatively small,' one can argue that the property p is conserved. The relative magnitude of σ_p was determined by comparison with 100 random models. A random model is defined as a deviation of the property p among 61 artificial amino acid compositions. Based on these 100 models, σ_p was transformed into a Z-score value and their significance was estimated. Although the method to generate such artificial amino acid compositions may be controversial, we employed a simple mixing procedure. The frequency of alanine in the imaginary model was set equal to the frequency of alanine in a randomly chosen organism. The same procedure was iterated 19 times for the other amino acids, and finally normalized to 1.

2.2 Mutation Pattern Analysis

Escherichia coli K12 and *Salmonella typhimurium* LT2 were analyzed because of their close evolutionary relationship. 1101 orthologous protein pairs were retrieved from COG and aligned by MAFFT [2]. The differences in physicochemical properties between orthologs were calculated. To compare the difference with the random model, artificial homologs of the *E. coli* proteins were generated. The generation of artificial sequences was based on the amino acid evolution model of Adachi and Hasegawa [1]. The distance d

between two orthologous proteins was measured in the PAM (percent accepted mutation) unit, and the PAM-*d* transition probability matrix was computed and applied to the mutated positions of the *E.coli* protein. To determine whether the difference between true orthologs is smaller than the difference between natural protein and artificial sequence, one-sided Wilcoxon signed rank test was performed.

3 Results and Discussions

We analyzed the amino acid compositions of 61 bacterial organisms in the context of net charge, hydrophobicity (transfer free energy from cyclohexane to water) and volume. Charge and hydrophobicity showed significant conservation. Their Z-scores were -6.24 and -8.19, respectively. On the contrary, residue volume was quite variable among species. (Z-score is 6.52) This suggests that using residue volume as a similarity measure could be potentially problematic when two proteins are distantly related.

Next, we investigated the underlying evolutionary mechanism for the conservation of charge and hydrophobicity. Could it be explained by independent mutation events that follow the matrix-based amino acid substitution model, or do proteins actively compensate the perturbing substitutions to minimize the evolutionary fluctuation in their physicochemical properties? We compared 1101 ortholog pairs of two closely related species, *Escherichia coli* K12 and *Salmonella typhimurium* LT2, and calculated the differences in the properties. The absolute values of the observed differences were significantly smaller than that of the independent mutation model. (Table 1.) This implies that each protein keeps their total charge and hydrophobicity by the active compensation mechanism. Although residue volume was also conserved among *E. coli* and *S. typhimurium*, the statistical significant was rather low.

Various constraints are imposed onto proteins in order for them to maintain biological function during evolution. These constraints work on not only the functional activity and the stability of an isolated molecule. Unnecessary increment in the hydrophobic content at the surface will cause an aggregation of molecules in the cell. Changes in charge distribution may disturb well-designed kinetics between enzymes and substrates. Thus, we propose a hypothesis that the active compensation is a universal trend in the course of evolution.

Table 1. Differences in the physicochemical properties

	Average (natural)	Average (artificial)	P-value
Charge	1.68	2.41	$<2.2\times10^{-16}$
Hydrophobicity	21.2	34.3	$<2.2\times10^{-16}$
Volume	158.8	181.1	7.2×10^{-5}

Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, the Japan Society for the Promotion of Science and the Japan Science and Technology Agency. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Adachi J, Hasegawa M., Model of amino acid substitution in proteins encoded by mitochondrial DNA, J Mol Evol. 1996 Apr;42(4):459-68.

[2] Katoh K, Misawa K, Kuma K, Miyata T., MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 2002 Jul 15;30(14):3059-66.

[3] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV *et al.*, The COG database: an updated version includes eukaryotes, BMC Bioinformatics. 2003 Sep 11;4(1):41.

[4] Tomii K, Kanehisa M., Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins, Protein Eng. 1996 Jan;9(1):27-36.