The Distributions of Amino Acid Compositions of Proteins in an Organism's Proteome Uniformly Approximate Binomial Distributions

Genshiro Esumi
Department of Pediatric Surgery, Hospital of the University of Occupational and Environmental Health, Kitakyushu, Japan

Abstract

Advances in the life science technologies have made it possible to access the genomic information of organisms, providing a bird's eye view of the "proteome," the entire set of proteins encoded by each genome. There have been many reports and discoveries in this research field. However, the shape of the distribution of the amino acid composition of the proteins in each proteome has not been reported.

In this study, I used NCBI proteome data. I calculated the amino acid compositions from the downloaded protein amino acid sequences. Then, I examined their distributions within each species and found that they all have bell-shaped distributions without exception.

Assuming that binomial distributions could explain these distribution shapes, I compared these proteome distributions with their adjusted binomial distributions, adjusted for lengths and means. These pairs of distributions were in fairly good agreement.

From these results, I speculated that the amino acid compositions of the proteins in each organism's proteome are in a state of mutual convergence with the amino acid compositions of the organism's cell bodies, which are composed of the proteome proteins, and this is the reason why their proteome distributions approximate binomial distributions.

Keywords: proteome, amino acid composition, distribution, binomial distribution, resource

Email: esumi@clnc.uoeh-u.ac.jp

1. Background

Advances in life science technology have made it possible to access the genomic information of organisms, providing a bird's eye view of the entire set of proteins encoded by each genome. These advances have already revealed, for example, that membrane proteins make up approximately 23% of prokaryotic proteomes and that this proportion is conserved across prokaryotes [1]. However, the shapes of the distributions of the amino acid compositions of the proteins in each proteome have not yet been reported, except by me [2].

In this report, I will show some unknown features of the distributions of the amino acid compositions of the proteins in an organism's proteome, and then give a putative explanation of their background.

2. The distributions of amino acid compositions of proteome proteins

2.1. Materials and Methods

In this study, I used proteome data from the NCBI website. For this work, I selected three species from each domain of organisms. I chose Homo sapiens, the human, as a representative of eukaryotes, Escherichia coli as a representative of bacteria, and Methanocaldococcus jannaschii, a thermophilic methanogenic archaean, as a representative of archaea [3-5].

In the NCBI genome data set, a file named "protein.faa" contains all protein names and amino acid sequence data encoded in each organism's genome.

From this file, I extracted the names of the proteins and their amino acid sequences, and then counted each amino acid residue in each protein. The amino acid compositions were then calculated by dividing the residue numbers of each 20 amino acid by the residue sums of all 20 amino acids.

As a result, each amino acid composition took a value from 0 to 1, and the total sum of the 20 amino acid compositions was 1.

The amino acid composition distribution was analyzed with a step resolution of 0.005. This resolution was considered sufficient to obtain a rough overview of the amino acid composition distributions.

In this and the following study, I used Microsoft® Excel for Mac v16.73 (Microsoft Corporation, Redmond, WA, USA) to calculate compositions and distributions. I also used JMP® 17.1.0 (SAS Institute Inc., Chicago, IL, USA) to generate graphs and figures.

2.2. Results

The distributions of the amino acid compositions of the proteins within each species are shown in Figure 1.

Figure 1A shows those of a human, a eukaryote. Figure 1B shows those of Escherichia coli, a bacterium. Figure 1C shows those of Methanocaldococcus jannaschii, an archaeum.

Without exception, each distribution shows a similar bell-shaped distribution.

2.3. Discussion of these distribution shapes

The distributions shown in Figures 1A, 1B, and 1C all look like Gaussian distributions, and they also look like binomial distributions. The binomial distribution is an estimated frequency distribution that describes the number of successes in a given number of independent trials, and they are known to approximate Gaussian distributions.

For the present analysis, if we define "success" as the incorporation of a given amino acid residue into the protein sequence, and assume that the probability of incorporation of the amino acid is fixed in each proteome, then the amino acid composition of a given protein in the proteome could follow a binomial distribution of the length of the protein, its amino acid residue numbers.

Based on the above, the next section examines whether each distribution follows binomial distributions.

3. Comparison with the adjusted binomial distributions

3.1. Materials and Methods

Assuming that proteome distributions follow binomial distributions with certain probabilities of event occurrence, these probabilities are considered equal to the mean of all values in the distribution due to the nature of the binomial distribution itself. Therefore, I calculated the means of each distribution of amino acid compositions within each proteome and used them as the probabilities of event occurrence for each binomial distribution.

On the other hand, since the proteome is a collection of proteins with different numbers of amino acid sequence lengths, the comparison target distribution must also be based on the same length distribution of its target. To generate an appropriate comparison distribution, a binomial distribution corresponding to each number of amino acid residue lengths was pre-calculated, and then their products with their sequence length compositions were summed. I refer to these as "adjusted binomial distributions" in this paper. To compare these distributions with the proteome distributions, I put them on the same sheets.

3.2. Results of the comparisons

Comparisons of the distributions of the amino acid compositions of the proteins within each organism's proteome with their comparative adjusted binomial distributions are shown in Figure 2. The proteome distributions (the same data as in Figure 1) are shown as gray areas, and the adjusted binomial distributions are shown as black outlines.

Figure 2A compares human distributions with their adjusted binomial distributions. Figure 2B compares Escherichia coli distributions with their adjusted binomial distributions. Figure 2C compares Methanocaldococcus jannaschii distributions with their adjusted binomial distributions.

Each pair of distributions was in fairly good agreement.

3.3. Discussion of the result of the comparison

I have shown that the two distributions compared above are in good agreement. One is the distribution of amino acid composition in the proteome, and the other is the adjusted binomial distribution based on the above assumptions. Since the comparison of all these distributions is unlikely to coincide by chance at this level, I concluded that the first assumption, that the distributions of amino acid compositions in the proteome follow binomial distributions, and the second assumption, that the probability of amino acid incorporation into the protein amino acid sequence is fixed in each proteome, may both be correct. I will discuss the implications of these two assumptions in the following section.

4. Discussion

In this paper, I showed that the distributions of amino acid composition in the proteomes of different organisms uniformly follow bell-shaped distributions. I then assumed that these distributions approximate binomial distributions, examined the validity of this assumption, and confirmed that the two distributions are pretty close. Thus, the assumption that the distribution of amino acid composition of proteome proteins follows binomial distributions may be correct.

The binomial distribution is an estimated frequency distribution that describes the number of successes in a given number of independent trials with a given probability of success. One might expect that it would be unreasonable to apply such distributions to the composition of amino acids in the amino acid sequence of a protein. Then why does this analysis show some similarity between the two distributions? The presumed reason is discussed below.

In living cells, even when there is no externally observable biological activity, it is generally accepted that protein degradation and synthesis are constantly taking place. In this process, the resources for their amino acid synthesis are expected to be the degradation products of their own intracellular proteins [6]. Therefore, I assumed that the probability fixations of the binomial distributions considered here originate from the amino acid composition of the degradation products of their own intracellular proteins.

To further discuss why each amino acid composition approximates a binomial distribution, the following four conditions must be considered. First, all protein sequences are encoded in the proteome of the organism's genome. Second, the proteins of the organism are synthesized primarily from the degradation products of its own intracellular proteins. Third, the efficiency of amino acid utilization within cells is expected to be extremely well optimized, and fourth, natural selection during evolution must have selected only highly adaptive mutations among the randomly occurring variations, in other words, protein genes with amino acid compositions that are difficult to synthesize from their resource composition are unlikely to be selected. As a consequence of these four conditions and under incomplete induction, I concluded that there must be mutual constraints between the amino acid composition of proteomes and the composition of whole intracellular proteins. I then estimated that these compositions are in a state of convergence.

In this convergence, the distribution of the amino acid composition of the proteome does not necessarily follow a binomial distribution. However, in the point of maximizing the efficiency of amino acid resource utilization, "protein amino acid compositions that are likely to be synthesized by chance from a given amino acid composition resource" can be paraphrased as "protein amino acid compositions that are easy to synthesize from a given amino acid composition resource". As a result, it would be reasonable to assume that the amino acid composition distribution of a proteome constrained by the amino acid resource composition would approximate the amino acid composition distribution likely to be synthesized from the same amino acid resource.

On the other hand, in Figure 2, the two distributions did not match perfectly and had some residual differences that varied by amino acid and species. All observed residual differences were in the direction of broadening the composition distributions. If the above assumption of a binomial distribution is correct, these compositions found in the residual differences should be more difficult to synthesize than compositions that follow binomial distributions. Why would they bother to synthesize proteins with such outlier compositions? I hypothesized two reasons. One is that some proteins in the proteome may achieve their particular protein functions by assembling specific amino acids in their sequences. The other is that the amino acid composition of their cell bodies may vary to some extent, especially in multicellular organisms whose cells differentiate into different cell types.

5. Conclusion

In this report, I described my finding that the distributions of the amino acid compositions of the proteins in an organism's proteome approximate binomial distributions. And I also described that these distributions can be explained by mutual constraint between the amino acid compositions of the proteome protein genes and the actual intracellular protein amino acid compositions. These findings are simple and seem easy to mention, but they have never been mentioned before. I believe that the reports I have made here describe some fundamental properties of biological science, and I also believe that they will explain some other properties that have never been explained before. I hope they will help us understand how organisms live in nature.

6. References

1. Sawada, R., Ke, R., Tsuji, T., Sonoyama, M., & Mitaku, S. (2007). Ratio of membrane proteins in total proteomes of prokaryota. *BIOPHYSICS*, *3*, 37–45. https://doi.org/10.2142/biophysics.3.37
2. Esumi, G. (2022). Proteome and cellular amino acid compositions may be mutually constrained and in a state of narrow convergence. *Jxiv*. https://doi.org/10.51094/jxiv.95
3. Genome dataset, “Homo sapiens (human) / Genome assembly T2T-CHM13v2.0", NCBI website, https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/
4. Genome dataset, "Escherichia coli str. K-12 substr. MG1655 / Genome assembly ASM584v2", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000005845.2/
5. Genome dataset, "Methanocaldococcus jannaschii DSM 2661 / Genome assembly ASM9166v1", NCBI website, https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000091665.1/
6. Esumi, G. (2020). Autophagy: Possible origin of essential amino acids. *Cambridge Open Engage*. https://doi.org/10.33774/coe-2020-1ll03

Figure 1

Figure 1A

The distributions of amino acid compositions of the proteins within Homo sapiens proteome

Homo sapiens (human) / Genome assembly T2T-CHM13v2.0 n = 64942

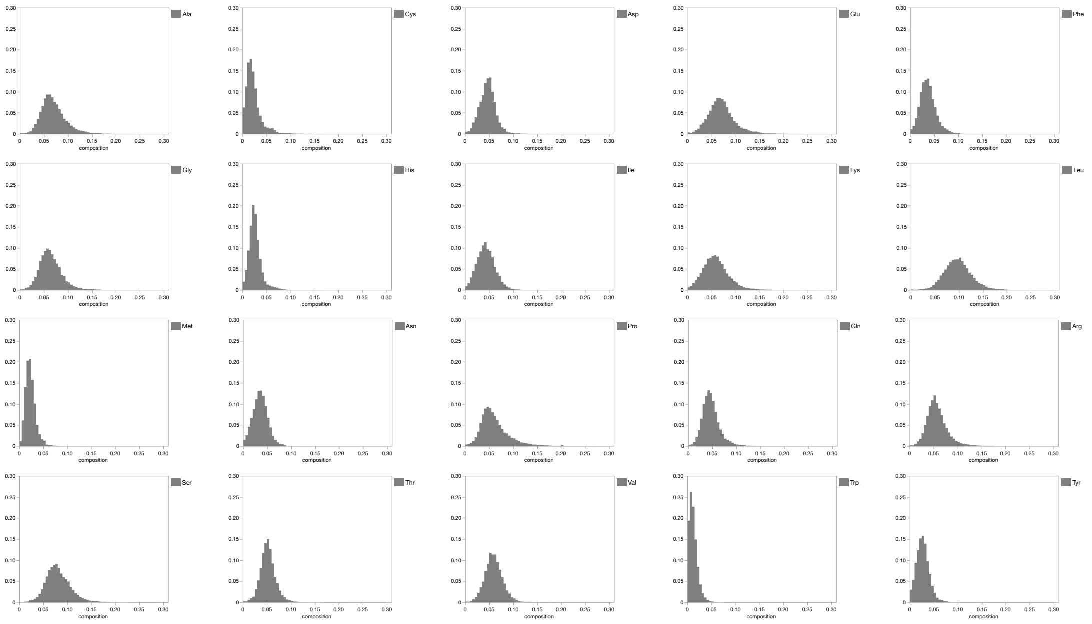

Figure 1B

The distributions of amino acid compositions of the proteins within Escherichia coli proteome

Escherichia coli str. K-12 substr. MG1655 / Genome assembly ASM584v2 n = 4448

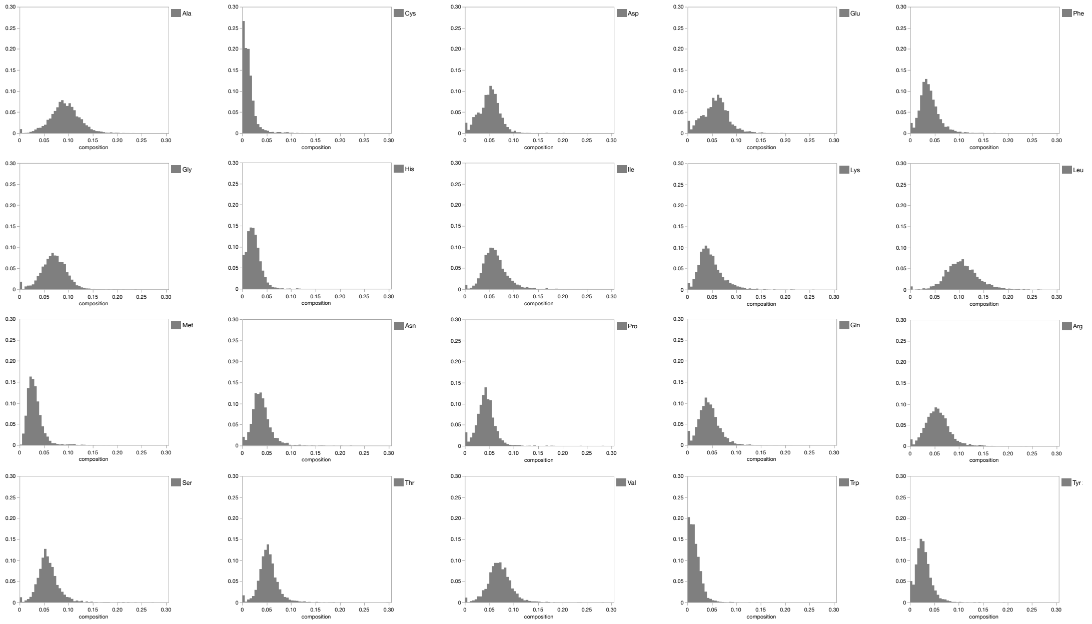

Figure 1C

The distributions of amino acid compositions of the proteins within Methanocaldococcus jannaschii proteome

Methanocaldococcus jannaschii DSM 2661 / Genome assembly ASM9166v1 n = 1817

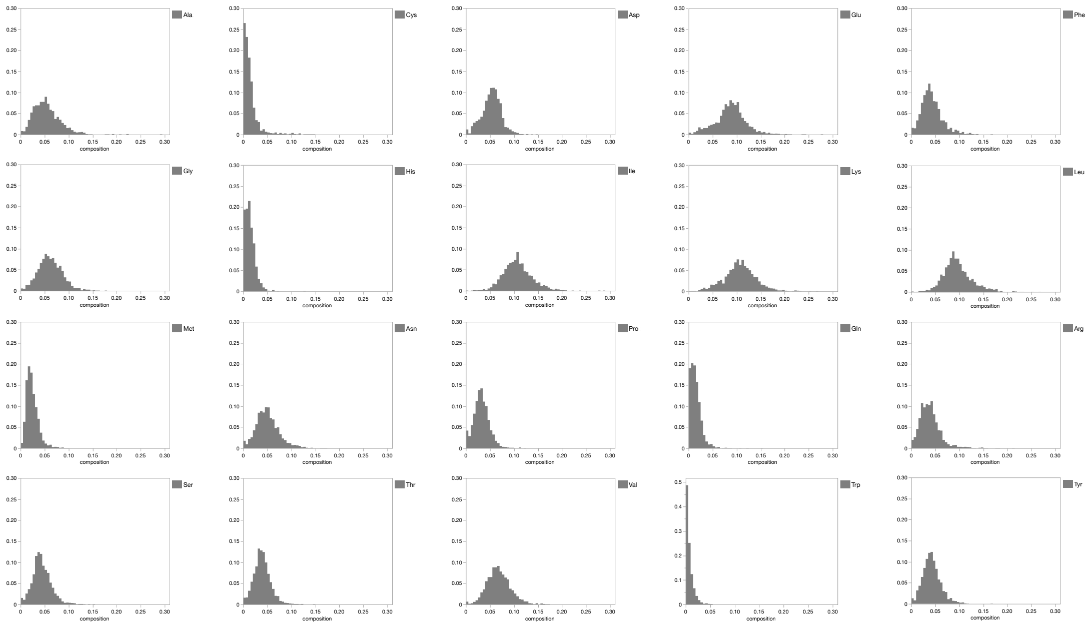

Figure 1 legend

The distributions of the amino acid compositions of the proteins within each species are shown. These distributions are ordered by the one-letter alphabetic code corresponding to each amino acid in these figures.

Figure 1A shows those of humans, a eukaryote.

Figure 1B shows those of Escherichia coli, a bacterium.

Figure 1C shows those of Methanocaldococcus jannaschii, an archaeum.

Without exception, each distribution shows a similar bell-shaped distribution.

Figure 2

Figure 2A

Comparisons of the distributions of the amino acid compositions of the proteins within Escherichia coli proteome with their adjusted binomial distributions

Homo sapiens (human) / Genome assembly T2T-CHM13v2.0 n = 64942

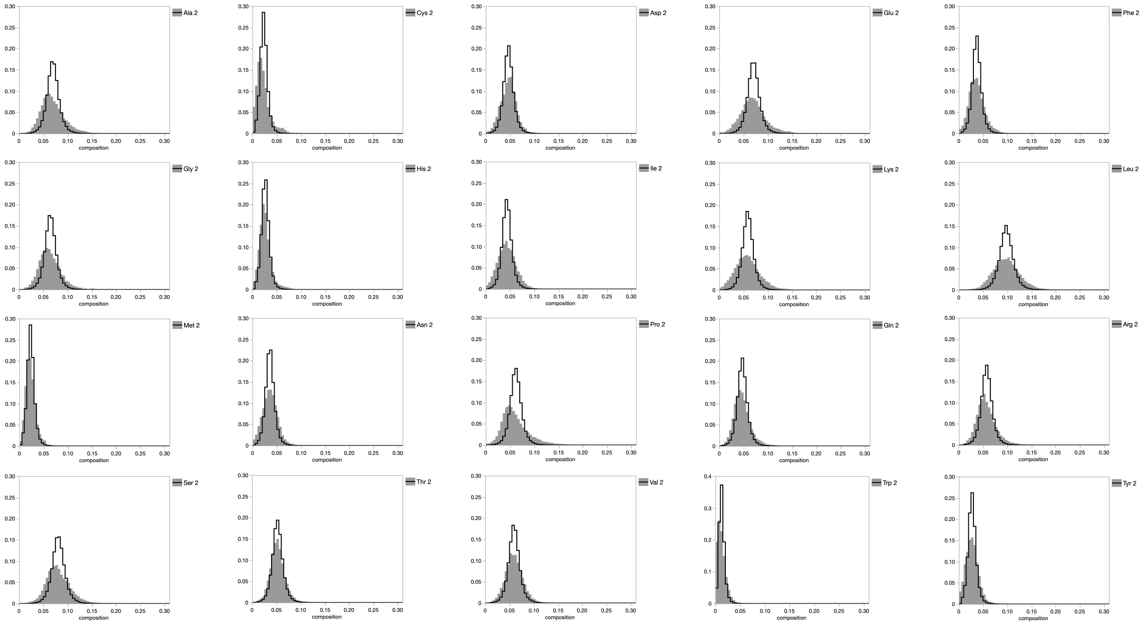

Figure 2B

Comparisons of the distributions of the amino acid compositions of the proteins within Homo sapiens proteome with their adjusted binomial distributions

Escherichia coli str. K-12 substr. MG1655 / Genome assembly ASM584v2 n = 4448

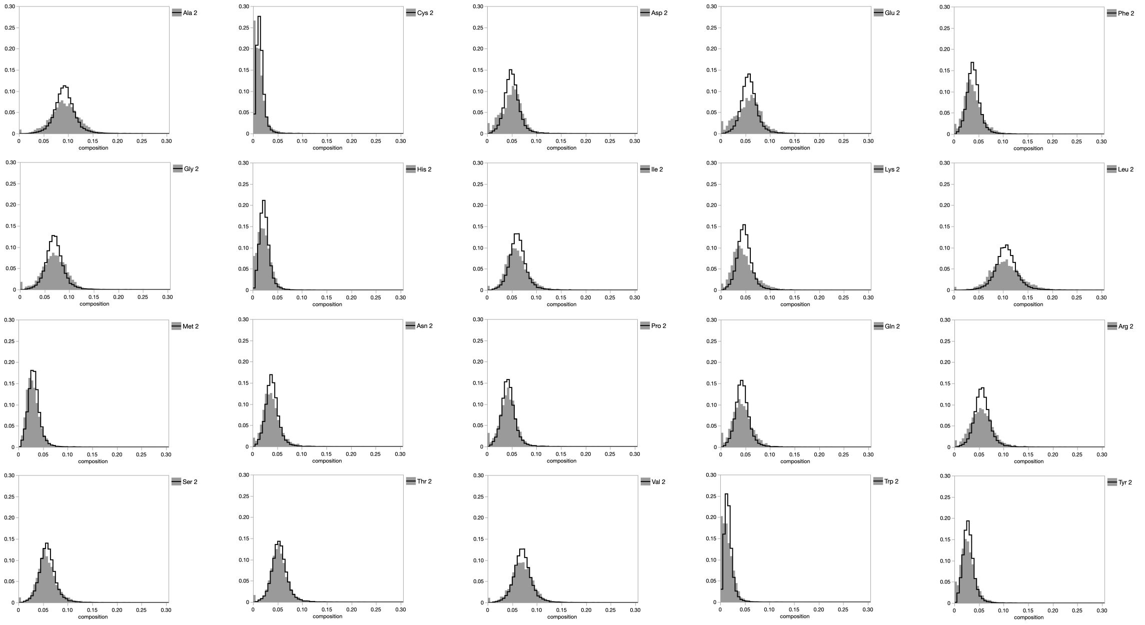

Figure 2C

Comparisons of the distributions of the amino acid compositions of the proteins within Methanocaldococcus jannaschii proteome with their adjusted binomial distributions

Methanocaldococcus jannaschii DSM 2661 / Genome assembly ASM9166v1 n = 1817

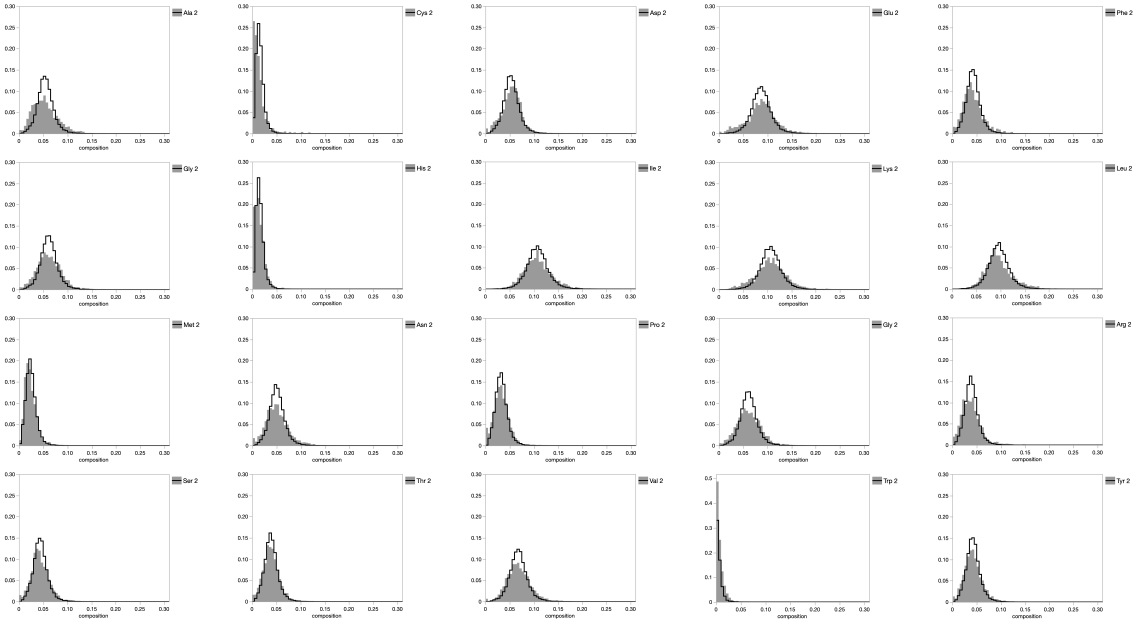

Figure 2 legend

Comparisons of the distributions of the amino acid compositions of the proteins within each species proteome with their comparative adjusted binomial distributions are shown. The proteome distributions (the same data as in Figure 1) are shown as gray areas, and the adjusted binomial distributions are shown as solid outlines.

Figure 2A compares human distributions with their adjusted binomial distributions.

Figure 2B compares Escherichia coli distributions with their adjusted binomial distributions.

Figure 2C compares Methanocaldococcus jannaschii distributions with their adjusted binomial distributions.

Each pair of distributions was in fairly good agreement.