In-depth analysis of amino acid and nucleotide sequences of Hsp60: how conserved is this protein?

Tatyana Tikhomirova[1], Maxim Matyunin[2], Mikhail Lobanov[3], and Oxana Galzitskaya[3]

[1]1Institute for Biological Instrumentation, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences"
[2]3 Institute of Protein Research, Russian Academy of Sciences
[3]Institute of Protein Research

September 25, 2021

Abstract

Chaperonin Hsp60, as a protein found in all organisms, is of great interest in medicine, since it is present in many tissues and can be used both as a drug and as an object of targeted therapy. Hence, Hsp60 deserves a fundamental comparative analysis to assess its evolutionary characteristics. It was found that the percent identity of Hsp60 amino acid sequences both within and between phyla was not high enough to identify Hsp60s as highly conserved proteins. In turn, their amino acid composition remained relatively constant. At the same time, the analysis of the nucleotide sequences showed that GC content in the Hsp60 genes was comparable to or greater than the genomic values, which may indicate a high resistance to mutations due to tight control of the nucleotide composition by DNA repair systems. Natural selection plays a dominant role in the evolution of Hsp60 genes. The degree of mutational pressure affecting the Hsp60 genes is quite low, and its direction does not depend on taxonomy. Interestingly, for the Hsp60 genes from Chordata, Arthropoda, and Proteobacteria the exact direction of mutational pressure could not be determined. However, upon further division into classes, it was found that the direction of the mutational pressure for Hsp60 genes from Fish differs from that for other chordates. The direction of the mutational pressure affects the synonymous codon usage bias. The number of high and low represented codons increases with increasing GC content, which can improve codon usage.

In-depth analysis of amino acid and nucleotide sequences of Hsp60: how conserved is this protein?

Tatyana S. Tikhomirova[1¶], Maxim A. Matyunin[2¶], Michail Yu. Lobanov[2] and Oxana V. Galzitskaya[2,3*¶]

[1]Institute for Biological Instrumentation, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", Pushchino, Moscow Region, Russia

[2]Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, Russia.

[3]Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Moscow Region, Russia

*Corresponding author

E-mail: ogalzit@vega.protres.ru (O.V.G.)

Abstract

Chaperonin Hsp60, as a protein found in all organisms, is of great interest in medicine, since it is present in many tissues and can be used both as a drug and as an object of targeted therapy. Hence, Hsp60 deserves

a fundamental comparative analysis to assess its evolutionary characteristics. It was found that the percent identity of Hsp60 amino acid sequences both within and between phyla was not high enough to identify Hsp60s as highly conserved proteins. In turn, their amino acid composition remained relatively constant. At the same time, the analysis of the nucleotide sequences showed that GC content in the Hsp60 genes was comparable to or greater than the genomic values, which may indicate a high resistance to mutations due to tight control of the nucleotide composition by DNA repair systems. Natural selection plays a dominant role in the evolution of Hsp60 genes. The degree of mutational pressure affecting the Hsp60 genes is quite low, and its direction does not depend on taxonomy. Interestingly, for the Hsp60 genes from Chordata, Arthropoda, and Proteobacteria the exact direction of mutational pressure could not be determined. However, upon further division into classes, it was found that the direction of the mutational pressure for Hsp60 genes from Fish differs from that for other chordates. The direction of the mutational pressure affects the synonymous codon usage bias. The number of high and low represented codons increases with increasing GC content, which can improve codon usage.

Keywords: chaperonin, bioinformatics, codon usage bias, amino acid compositions, database construction

Abbreviations: IL-8 (6), interleukin-8 (6); IgG (A), immunoglobulin G (A); Hsp60, heat shock protein (60 kDa); PID, percent identity; UPGMA, unweighted pair group method with arithmetic mean; ENC, effective number of codons; RSCU, relative synonymous codon usage.

Introduction

Hsp60 (˜60 kDa) is a member of the heat shock protein family. The main function of Hsp60 is to capture newly synthesized or denatured protein and promote its folding. It should be noted that Hsp60 exhibits the properties of a molecular chaperonin not only at high temperatures, as its name suggests, but also under conditions of moderately acidic[1] or high salinity[2]. Because the cells of organisms are under incessant stress, Hsp60 is a highly expressed protein. Thus, its ubiquitous distribution makes Hsp60 a promising object for research and use.

Today Hsp60 plays an important role in medicine. In particular, Hsp60, which is involved in the pro-inflammatory response, up-regulates the production of IL-8[3] and IL-6[4] in human bronchial epithelial cells and microglia, respectively. In addition, Hsp60 as an adipokine can be released from adipose tissue[5]. The level of chaperonins in the blood of obese patients positively correlates with markers of inflammation, which may indicate the development of cardiovascular disease. Hsp60 is also involved in the autoimmune response as an antigen[6]. Since Hsp60 is overexpressed in tumor cells, it can be used both to detect the early stage of cancer[7] and to develop immunogenicity against it[8]. On the other hand, chaperonin can be used as a target not only for immunotherapy, but also for antibiotic therapy of bacterial diseases such as African sleeping sickness[9,10] and *Mycobacterium tuberculosis*infections[11]. In such cases, Hsp60 of pathogenic microorganisms is inhibited by various inhibitors[9–11].

At present, in-depth analysis of Hsp60 sequences *in silico* is quite rare, although it is a very interesting object for research. Since Hsp60 is ubiquitous, it can be used in evolutionary analysis of organisms belonging to different taxonomic groups[12–15]. Using bioinformatics, epitopes in Hsp60 can be predicted[16,17]. In particular, an epitope-based vaccine containing Hsp60 epitopes from *Helicobacter pylori* was tested in a model of a Mongolian gerbil infected with *H. pylori* [16]. Oral immunization with this vaccine reduced*H. pylori* colonization due to the T helper, IgG, and IgA responses and antibodies against various *H. pylori* antigens. In another study, B-cell epitopes were identified *in silico* in Hsp60 overexpressed by tumor cells[17]. In addition, it was proposed to use the sequence of Hsp60 gene for the species-specific identification of organisms belonging to the genus *Acetobacter sp* .[18], *Helicobacter sp* .[19],*Staphylococcus sp* .[20], *Bifidobacterium sp* .[21,22], and *Bacteroidetes sp.* [23] The interaction patterns of Hsp60 and the Aβ(1-42) peptide were predicted by molecular dynamics simulation and protein-peptide docking[24]. These results are important because Hsp60 affects oligomers of the Aβ peptide reducing their cytotoxicity in patients with Alzheimer's disease[25].

As you can see, Hsp60 has a fairly wide range of applications and requires in-depth research. Therefore, the purpose of this study is a comprehensive bioinformatic analysis of the amino acid and nucleotide sequences

of Hsp60 and the compilation of convenient databases on its basis.

Materials and methods

Databases preparation

Database 1, containing 29360 amino acid sequences, was built with NCBI BLAST at the following request:

- Database: non-redundant protein sequences (nr);
- Algorithm: blastp (protein-protein BLAST);
- Expected threshold: 10^{-20}.

Database 2, containing 2416 nucleotide sequences, was created using the NCBI Gene Database (*https://www.ncbi.nlm.nih.gov/gene/*).

The original databases contained a lot of junk items such as duplicates of certain sequences and truncated sequences.

Truncated sequences were excluded according to the following criteria: 450 aa residues [?] number of amino acid residues in Hsp60 (amino acid sequences) [?] 650 aa residues, 1350 bp [?] number of nucleotides in Hsp60 gene (nucleotide sequences) [?] 1950 bp. The criterion 99% [?] PID [?] 100% was used to remove duplicate amino acid sequences, where the PID is the percent identity of two compared sequences.

Definition of taxonomy

The search for taxonomic rank was carried out automatically using the Selenium WebDriver library (Python 3.6;*https://github.com/SeleniumHQ/selenium*). The databases contain Hsp60 sequences of 19 phyla that were sorted by taxonomic rank using the NCBI Taxonomy Database (https://www.ncbi.nlm.nih.gov/taxonomy) (Table 1). Some of the sequences belonging to the super kingdom Viruses have not been identified by phylum.

Multiple sequence alignment

Multiple alignment of Hsp60 amino acid sequences was performed using the ClustalOmega software[26] and the BioPython 1.68 library[27]. Multiple alignment of Hsp60 nucleotide sequences was performed using the MUSCLE program[28] and the MEGA-CC software[29]. The nucleotide sequences were aligned using the MUSCLE Codons option. Here, all codons in the sequences have been converted to amino acids using the standard genetic code. The obtained amino acid sequences were aligned with subsequent reverse translation of amino acids into codons. Note that this option is valid for encoding nucleotide sequences in multiples of three and without reading frame displacement.

Percent identity of sequences

The percent of sequence identity (PID) (Supplementary, PID and SD) was calculated according to equation (1):

$$\mathrm{PID} = \frac{N_{\mathrm{eq}} \bullet 100}{N_1 + N_2 - N_{\mathrm{eq}}}, \ \% \ (1)$$

where N_{eq} is the number of identical aligned non-gapped characters; N_1 and N_2 are the length of aligned sequences[30]. Note that equation 1 is applied to full-sized sequences. In this study, the number of amino acid residues (hereinafter referred to as the length of the sequences) ranged from 450 to 650 aa residues.

Hierarchical clustering of sequences

Clustering of values from a symmetrical matrix was performed using the UPGMA algorithm and the SciPy library[31,32]. The UPGMA algorithm uses the Euclidean distance matrix as input. At each stage of the process, the minimum value corresponding to the nearest clusters (A and B) must be found in this matrix. Clusters A and B should be merged into a new cluster C. The columns and rows corresponding to clusters A and B should be replaced with a new column and row from cluster C. This procedure should be repeated until all clusters are merged. During clustering, the UPGMA algorithm builds a rooted-tree (dendrogram). The

length of the branches of this dendrogram corresponds to the Euclidean distance between the two nearest clusters.

To separate clusters, you need to set the cophenetic distance. The cophenetic distance is the height of the dendrogram where two branches, including two objects, merge into one branch[33]. In this work, a cophenetic distance of 70% of the final confluence height on the dendrogram was used.

Min-max normalization

The min-max normalization strategy was used to compare two or more samples by bringing the data in them to the same scale using the following equation (2):

$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$ (2),

where x_{norm} is the normalized value in the sample in the range from 0 to 1; x is the real value in the sample;x_{min} and x_{max} are the minimum and maximum values in the sample, respectively.

Statistical analysis

Statistical analysis was performed using the SciPy library (Python 3.6; www.scipy.org[31]). Data was presented as mean and standard deviation and compared with Student's t-test. According to the tests of statistical hypothesis used, the average values of the compared samples differ non-significant (the difference between their mean values is random) if the p-value of the statistical test is higher than the significance level (alpha-error). An alpha error of 5% was accepted for all cases.

Comparison of the amino acid composition of Hsp60 with the proteomic composition

Amino acid composition (X) as a percentage of 20 amino acid residues in the sequence was determined for each of the Hsp60 sequences in database 1. Then the average amino acid composition (X) of Hsp60s was determined for each of the 19 phyla.

To estimate the tendency to decrease or increase in the content of certain amino acids in the Hsp60 sequences, it is necessary to calculate the amino acid profile (X_n). For this, the average amino acid composition of Hsp60 (X) for each of the 19 phyla was compared with the average amino acid composition of the corresponding proteomes (X_p) (Supplementary, AA composition of proteomes) using the following conditions:

- If $X_i > X_{pi}$and p-value < alpha error, average percentage (X_i) of i -amino acid for Hsp60s is considered greater than proteomic value (X_{pi}),$X_{ni} = 1$;
- If $X_i < X_{pi}$and p-value < alpha error, average percentage content (X_i) of i -amino acid for Hsp60s is considered lower than proteomic value (X_{pi}),X_{ni} = -1;
- If the p-value of the statistical test is higher than the alpha error, the average percentage (X_i) of i -amino acid for Hsp60s is considered comparable to the proteomic value (X_{pi}), $X_{ni} = 0$,

where X_{ni} is the normalized value of the average percentage of i -amino acid for Hsp60s (X_i) relative to the average proteomic value (X_{pi}) for the corresponding phylum.

The amino acid profile of phylum (X_n) includesX_{ni} values for all amino acid residues. Average proteomic amino acid compositions (X_p) were calculated using 8747 reference proteomes obtained from the UniProt Proteomes database (*https://www.uniprot.org/proteomes/*).

Using the normalized values X_{ni} of 19 phyla, the average normalized value $<X_{ni}>$ can be calculated. The summary amino acid profile includes $<X_{ni}>$values for all phyla. In should be noted that such an amino acid profile is only meaningful if the groups included have similar amino acid profiles.

In this work, the amino acid profiles of the Hsp60 sequences for 19 phyla and the total amino acid profile were calculated.

GC content

The GC content reflects the percentage of guanine and cytosine bases in the nucleotide sequence. The average GC content in Hsp60 sequences and genomes for 17 phyla was calculated using database 2 and 20808 reference genomes obtained from the NCBI Genome database (*https://www.ncbi.nlm.nih.gov/genome/*), respectively (Supplementary, GC-content and ENC).

The GC content of the first, second and third codon positions (GC_1, GC_2, and GC_3, respectively) is required to evaluate their contribution to the GC content and is calculated by equation (3):

$GC_n = \frac{G_n + C_n}{N}$ (3),

where n is the base position in the codon (n = 1, 2, 3);G_n and C_n are the number of guanine and cytosine bases in the n-position of the codons; $N$is the number of codons in the sequence.

Neutrality analysis

Neutrality plot analysis was used to assess codon usage bias caused by mutational pressure and natural selection. Codon neutrality is expressed through the relationship between $GC_{1,2}$ (average GC content between GC_1 and GC_2) and GC_3, represented by a regression line (neutrality plot)[34,35]. The slope ϵ of this regression line, calculated using least squares regression analysis, is a neutrality value that ranges from 0 when there is no effect of mutation pressure (full selective constraints) to 1 when mutation bias is the main force affecting the codon usage (complete neutrality of codon usage). The equilibrium point E_p was also defined. This point is at the intersection of the neutrality plot (regression line) and the line of complete equilibrium between the average GC content and the directional mutation pressure with neutrality equal to one ($\epsilon = 1$). The E_p value was computed as$\frac{A_0}{(1-\varepsilon)}$, where A_0 is the value calculated from the regression of $GC_{1,2}$ and GC_3 at GC_3 = 0[34]. The direction of the mutational pressure in the gene was defined according to the position of this gene on the neutrality plot relative to the equilibrium point.

Relative synonymous codon usage (RSCU) analysis

Relative synonymous codon usage (RSCU) is a measure for assessing the uneven use of synonymous codons in nucleotide sequences or in the genome and is calculated using equation (4):

$RSCU = \frac{X_i}{\left(\frac{1}{n}\right)\sum_{i=1}^{n} X_i}$ (4),

where n is the number of synonymous codons for the X -amino acid (1 [?] n [?] 6); X_i is the number of occurrences of the i -codon in the nucleotide sequence. Codons with RSCU $>$ 1 are considered as frequent, and a higher RSCU corresponds to a larger bias in codon usage. RSCU $<$ 1 indicates the opposite situation. If the RSCU is one, no codon usage bias is observed. The latter corresponds to the Met and Trp residues, since these amino acids have no synonymous codons in accordance with the standard genetic code (Supplementary, Genetic Code).

In this work, RSCU values were determined for each nucleotide sequences from database 2. Average RSCU values were determined for 17 phyla, excluding Viruses and Platyhelminthes (Supplementary, RSCU).

Effective number of codons

Effective number of codons (ENC) is a measure for assessing codon usage bias. The ENC value ranges from 20 (maximum bias, i.e. only one codon is used to encode an amino acid) to 61 (minimum bias, i.e. codons are used randomly)[36,37]. In contrast to the RSCU value, the ENC value characterizes the overall codon usage bias for a gene/genome.

The calculation of codon homozygosity (F_k) for all amino acids having synonymous codons, which is required to calculate the ENC value, was carried out according to the following equations (5):

$S = \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2$; $\quad F_k = \frac{nS-1}{n-1}$ (5),

where n_i is the occurrence of the k -codon for the i -amino acid; n is the total number of codons for the i -amino acid; k is the number of codons for the$i$ -amino acid (k = 2, 3, 4, 6). In this work,n_i and n were

represented by the average values for Hsp60 sequences of 17 phyla excluding Viruses and Platyhelminthes (Supplementary, GC-content and ENC).

The ENC values were calculated for 17 phyla according to equations (6):

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}; \qquad \mathrm{ENC_{exp}} = 2.5 + s + \frac{29.5}{s^2+(1-s)^2} \quad (6)$$

where *2* is a sum of F for Met and Trp, since these amino acid residues have no synonymous codons; $F_{2/3/4/6}$is an average F_k for those amino acid residues that have k synonymous codons. The ENC_{exp}value is the expected ENC value for various GC contents at the third synonymous codon position (GC_3 content) in the Hsp60 gene for 17 phyla. Here, s is the given value of GC_3 content (Supplementary, GC-content and ENC). The data obtained were used to construct the Nc-plot, where the ENC values are presented as a scatter plot, and the ENC_{exp}values are presented as a solid curve. Gene codons are unbiased if the corresponding point is on the expected curve[38].

Results and discussion

Databases

According to the main criteria (see Databases preparation) databases were compiled containing amino acid (database 1) and nucleotide (database 2) sequences of Hsp60. The size of database 1 was 19220 amino acid sequences and the size of database 2 was 1925 nucleotide sequences. The composition of these databases is described in details in Table 1.

Table 1. General composition of databases

Run	Kingdom	Phylum	Number of AA sequences in the group (database 1)	Average PID±SD, %	Number of subclusters[a]	Number of nucleotide sequences in the group (database 2)
1	Viruses	Viruses	79	28±4	4	-
2	Archaea	Euryarchaeota	151	42±15	3	63
3	Bacteria	Chlamydiae	122	31±10	2	12
4	Bacteria	Bacteroidetes	1358	58±4	2	22
5	Bacteria	Proteobacteria	7973	51±9	2	336
6	Bacteria	Cyanobacteria	682	53±6	3	8
7	Bacteria	Actinobacteria	4309	52±7	3	73
8	Bacteria	Firmicutes	3132	56±8	2	132
9	Protozoa[b]	Apicomplexa	47	42±14	3	16
10	Protozoa	Euglenozoa	25	50±14	2	5
11	Plantae	Streptophyta	717	44±17	4	431
12	Plantae	Chlorophyta	26	41±8	3	8
13	Fungi	Ascomycota	228	64±7	2	136
14	Fungi	Basidiomycota	79	63±3	4	31
15	Metazoa	Nematoda	50	72±19	3	4
16	Metazoa	Arthropoda	162	50±10	3	255
17	Metazoa	Platyhelminthes	9	66±5	5	-
18	Metazoa	Chordata	61	84±6	3	384
19	Metazoa	Mollusca	10	67±4	7	9
Total	Total	Total	19220		60	1925

Run	Kingdom	Phylum	Number of AA sequences in the group (database 1)	Average PID±SD, %	Number of subclusters[a]	Number of nucleotide sequences in the group (database 2)
[a] Clustering was carried out in each of 19 phyla using the UPGMA algorithm. [b] The designation "Protozoa" was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy.	[a] Clustering was carried out in each of 19 phyla using the UPGMA algorithm. [b] The designation "Protozoa" was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy.	[a] Clustering was carried out in each of 19 phyla using the UPGMA algorithm. [b] The designation "Protozoa" was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy.	[a] Clustering was carried out in each of 19 phyla using the UPGMA algorithm. [b] The designation "Protozoa" was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy.	[a] Clustering was carried out in each of 19 phyla using the UPGMA algorithm. [b] The designation "Protozoa" was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy.	[a] Clustering was carried out in each of 19 phyla using the UPGMA algorithm. [b] The designation "Protozoa" was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy.	[a] Clustering was carried out in each of 19 phyla using the UPGMA algorithm. [b] The designation "Protozoa" was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy.

Comparative analyses of amino acid sequences of Hsp60 from 19 phyla

PID estimation

Average PID values were calculated for 19220 Hsp60 sequences in 19 phyla (Table 1), where they ranged from 28±4% (Viruses) to 84±6% (Chordata). The maximum average PID value was determined for the Hsp60 sequences belonging to Chordata, which was expected in accordance with our previous work[14]. The minimum average PID value corresponds to Viruses, which can be explained by their high mutation rate. For groups with a higher taxonomic rank, the average PID values varied within a narrow range from 51±9% to 58±4% for Bacteria (Proteobacteria, Cyanobacteria, Actinobacteria, Firmicutes, and Bacteroidetes, with the exception of Chlamydiae), from 63±3% to 64±7% for Fungi (Basidiomycota and Ascomycota), and from 41±18% to 44±17% for Plantae (Chlorophyta and Streptophyta). Apicomplexa and Euglenozoa can be combined into Protozoa with the average PID values ranging from 42±14% to 50±14%. For Hsp60 sequences belonging to Metazoa (Arthropoda, Mollusca, Platyhelminthes, Nematoda, and Chordata), the range of average PID values was wider: from 50±10% for Arthropoda to 84±6% for Chordata.

Clustering was performed on 19 phyla, and the average PID value was calculated for each of the sub-clusters (Supplementary, PID, and SD). The number of sub-clusters ranged from two to seven in 19 phyla (Table 1). The calculation of the average PID values and their clustering were carried out between 60 sub-clusters (Figure 1).

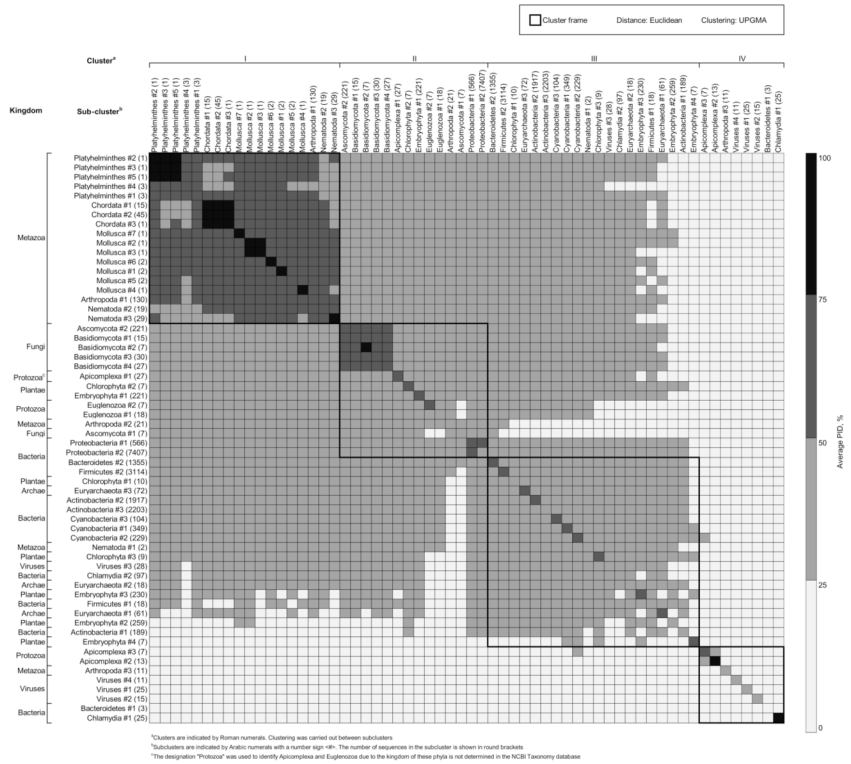

Figure 1. Symmetric matrix of the average PID values. The matrix contains 60 sub-clusters of Hsp60 sequences from 19 phyla. The X- and Y-axis items "Sub-cluster" are represented in the following format "Phylum #sub-cluster (number of sequences in a sub-cluster)". Sub-clusters of Viruses have no phylum labels. The Y-axis "Kingdom" represents sub-clusters united by a higher taxonomic rank (Kingdom). The black frames and the X-axis "Cluster" show four clusters and their numbers (Roman numerals), which were obtained by clustering 60 sub-clusters. Clustering was performed using the UPGMA algorithm.

Apparently, the symmetric matrix contains four clusters. Cluster I contains 258 Hsp60 sequences from Metazoa (multicellular animals). It should be noted that, as a rule, in this cluster the average PID values are more than 60% and 50% within and between sub-clusters, respectively (Supplementary, PID, SD). The number of identical amino acid residues in the sequences reflects the degree of conservatism. Thus, the Hsp60 sequences belonging to cluster I can be considered as intermediate and highly conserved, as noted in our previous work[14]. Cluster II includes 608 Hsp60 sequences of Fungi, Plantae, Protozoa, and Metazoa. It should be noted that fungal Hsp60s were clustered into a small group with average PID values greater than 60% and 50% within and between sub-clusters, respectively, as observed for cluster I. Thus, the Hsp60 amino acid sequences from Fungi can also be classified as intermediate and highly conserved. Others sub-clusters in cluster II demonstrate intermediate and low sequence conservatism, with average PID values ranging from 40±30% to 62±13% within sub-clusters and from 24±4% to 47±19% between them.

The largest cluster III contains 18244 sequences (22 sub-clusters), mainly including Hsp60 of Bacteria, Plantae, and Archaea. It should be noted that within the 11 sub-clusters in cluster III, the average PID values vary from 53±10% (Firmicutes #2) to 73±8% (Cyanobacteria #1), indicating intermediate and highly conserved Hsp60 sequences. However, between the sub-clusters of cluster III, these values are less than 50%. Thus, it can be assumed that, in the whole, the level of conservatism in cluster III is low.

Finally, there is the smallest cluster IV containing 110 Hsp60 sequences from Viruses, Bacteria, Protozoa, and

Metazoa. In this cluster, sub-clusters of Viruses #1, Apicomplexa #3, Chlamydia #1, and Apicomplexa #2 show average PID values of more than 50%. On the other hand, the PID values between the sub-clusters of cluster IV are quite low and range from 10±1% (Chlamydiae #1/Apicomplexa #2) to 33±2% (Apicomplexa #2/Apicomplexa #3). Moreover, the average PID values between cluster IV and other clusters are also low (Figure 1). This extremely low level of conservatism of Hsp60 sequences in cluster IV may indicate how these Hsp60s have evolved to distance themselves from others Hsp60.

In some studies[2,4,17,25] Hsp60 is called a highly conserved protein. But, according to the obtained data, the percent of identical amino acid residues varies widely. Thus, Hsp60 is not a highly conserved protein.

It should be noted that metazoan Hsp60 sequences belonging to sub-clusters Arthropoda #2 and #3, and Nematoda #1 were not included in cluster I. To explain this phenomenon using a symmetric PID matrix (Supplementary, PID), the average PID values between Hsp60 sequences of these sub-clusters and others Hsp60 sequences belonging to each of 19 phyla were calculated (Figure 2; Supplementary, Metazoan artifacts).

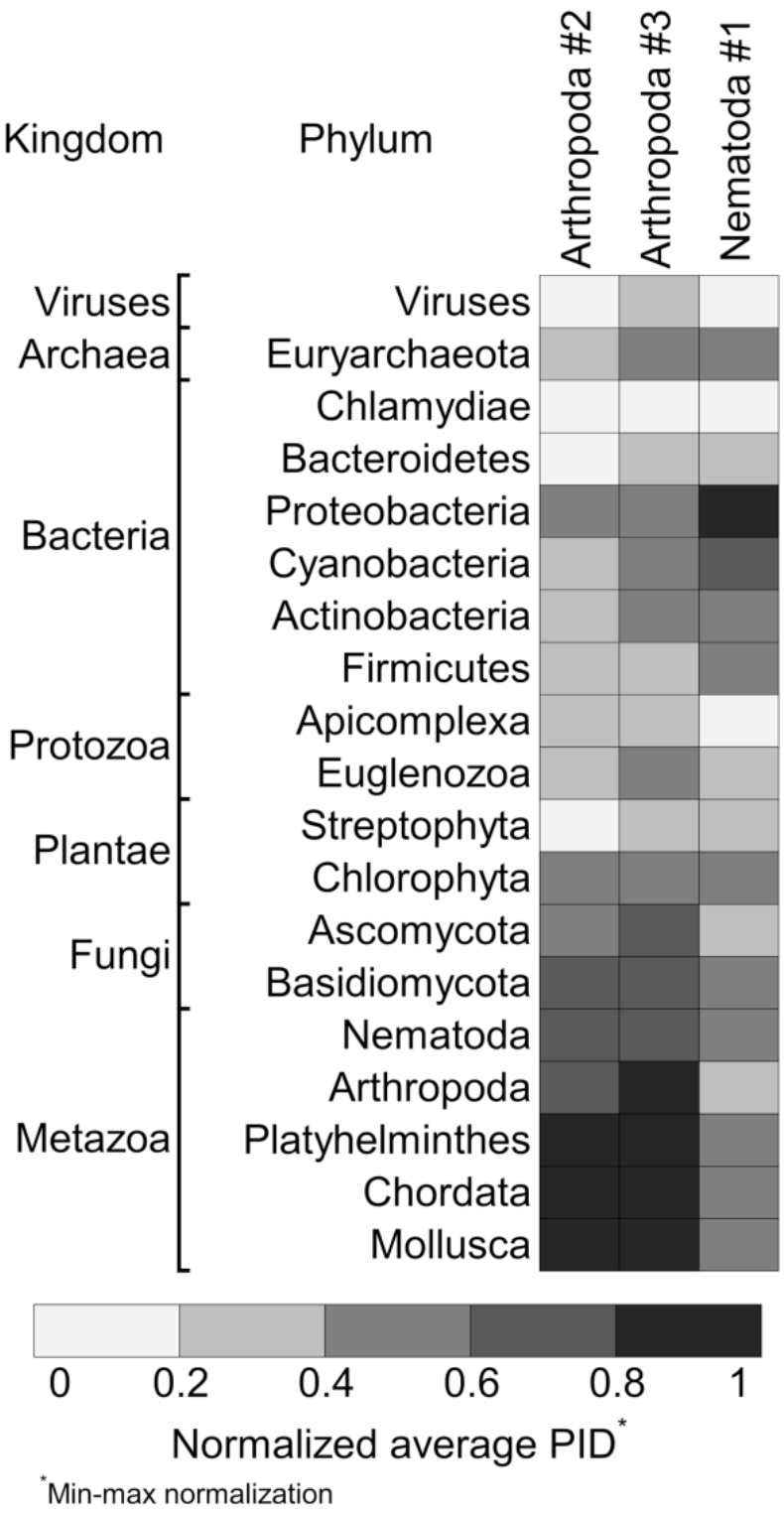

Figure 2. Heat map showing normalized average PID values between Hsp60 sequences belonging to sub-clusters Arthropoda #2, Arthropoda #3, and Nematoda #1 and Hsp60 sequences belonging to each of 19 phyla. The average PID values were normalized using the min-max normalization method. The 19 phyla were sorted by NCBI Taxonomy.

According to Figure 2, the highest average PIDs were observed between Arthropoda sub-clusters #2 and #3 and phyla Mollusca and Arthropoda, respectively. Since these phyla belong to Metazoa, the presence of these sub-clusters in clusters II and IV can be explained as a feature of dendrogram construction and the selected percent of cophenetic distance (see Hierarchical clustering of sequences).

On the other hand, the average PID value between the Nematoda #1 sub-cluster and Proteobacteria phylum was 37.8±0.3% (Supplementary, Metazoan artifacts), which was higher than for all metazoan phyla including Nematoda (28.7±0.9%). This feature may be the result of a database error.

In turn, this may be explained by horizontal gene transfer[39]. Accordingly, possible cases of horizontal Hsp60 gene transfer were identified using the obtained data (Supplementary, Horizontal Hsp60 gene transfer). It was found that horizontal Hsp60 gene transfer is most common in Bacteria, which is obvious. In turn, the Hsp60 sequences from the Nematoda #1 sub-cluster have the maximum PID in the following pairs:

Onchocerca volvulus (CAA70570.1, Nematoda)/*Wolbachia sp.*(WP_014869024.1, Proteobacteria), PID 78.3%;

Trichuris trichiura (CDW56974.1, Nematoda)/*Enterobacter sp.* (WP_015572446.1, Proteobacteria), PID 88.6%.

Hence, the transmission paths can be tracked. Endosymbiotic bacteria *Wolbachia sp.* are present in filarial nematodes[40], to which *Onchocerca volvulus* belongs. At the same time, *Trichuris trichiura* roundworms parasitize in human intestine, where *Enterobacter sp.* bacteria are also present. These data are not conclusive evidence of horizontal Hsp60 gene transfer between these organisms and require in-depth review.

Amino acid composition of Hsp60

The amino acid composition of 19220 Hsp60 sequences in database 1 and the average amino acid compositions of Hsp60s for 19 phyla were calculated (Supplementary, AA composition of Hsp60). Despite the low average PID values between sub-clusters (Figure 1), the amino acid composition was quite similar (Figure 3a).

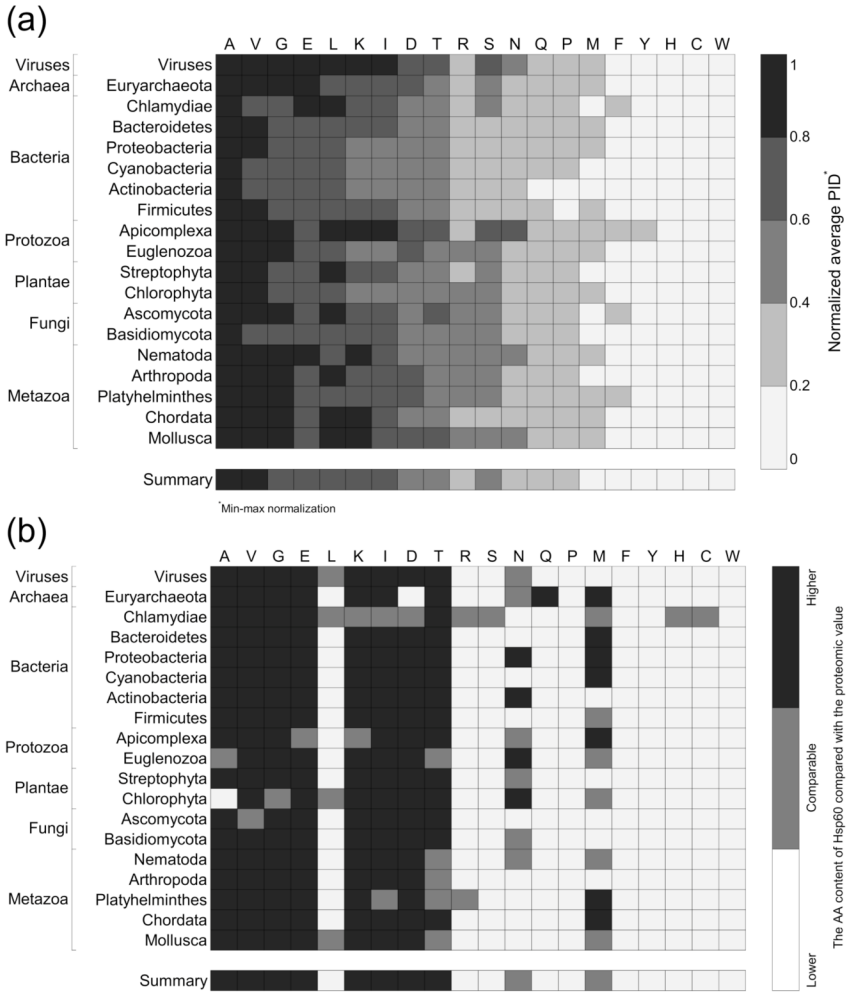

Figure 3. The average amino acid composition of the Hsp60 sequences from 19 phyla: a - Heat map displaying normalized average PID values between 19 phyla of Hsp60 sequences; b - The average amino acid composition of the Hsp60 sequences for each of the 19 phyla compared to the corresponding proteomic values. In Figure 3a the average values were normalized using the min-max normalization method. The line "Summary" presents the average normalized amino acid composition of Hsp60 for 19 phyla. In Figure 3b the amino acid profiles were represented as the average amino acid composition of Hsp60 for each of 19 phyla compared to the average amino acid composition of the respective proteomes. The structure of color scale is as follows: Higher/Lower - the amino acid content in Hsp60 is higher/lower than in proteomes, respectively; Comparable - the amino acid content in Hsp60 is comparable to the average proteomic value. The "Summary" line shows the average amino acid profile of Hsp60 for 19 phyla. The groups were sorted using the NCBI Taxonomy. Amino acids were sorted using an average amino acid composition of 19220 Hsp60 sequences.

As can be seen, aliphatic (Ala, Val, Gly, Leu, and Ile), charged (Glu, Lys, Asp), and polar (Thr) amino acid residues are overrepresented in the Hsp60 sequences of all 19 phyla (Figure 3a). The average content of these amino acid residues for 19220 Hsp60 sequences ranges from 6.2±1.1% for Thr to 12.7±1.9 for Ala (Supplementary, AA composition of Hsp60). In turn, aromatic amino acid residues (Phe, Tyr, His, and Trp) and Cys are underrepresented and their content ranges from 0.2±0.3% (Cys) to 1.6±0.4% (Phe) (Figure 3a; Supplementary, AA composition of Hsp60).

To assess the amino acid composition of Hsp60s relative to proteomic values (Supplementary, AA Composition of proteomes), amino acid profiles of Hsp60 sequences were determined for each of 19 phyla (Figure 3b).

The average content of aliphatic (Ala, Val, Gly, and Ile), charged (Glu, Lys, and Asp), and polar (Thr) amino acid residues in Hsp60 sequences is not only high, but also higher than the average proteomic values for almost all phyla (Figure 3b). The high content of aliphatic amino acid residues may indicate the structural stability of these proteins[15]. The content of Pro, aromatic (Phe, Tyr, His, and Trp) and polar (Ser and Cys) amino acid residues was low and lower compared to the average proteomic values. In general, according to the summary amino acid profile, no matter how evolutionarily distant the 19 phyla are from each other, the amino acid composition of their Hsp60 sequences remained the same, as its correlation with the amino acid composition of the corresponding proteome.

Despite the low PID values both within and between phyla, functionality and domain structure[41–45] of Hsp60 persisted over time. It can be assumed that these features largely depended on the amino acid composition, i.e. on the percentage content of amino acids in Hsp60.

Comparative analyses of Hsp60 genes from 17 phyla

GC content and mutation pressure for codon usage

In the present study, the average GC content and $GC_{1/2/3}$for the Hsp60 sequences and the average genomic GC content were calculated for 17 phyla (Figure 4a; Supplementary, GC-content and ENC).

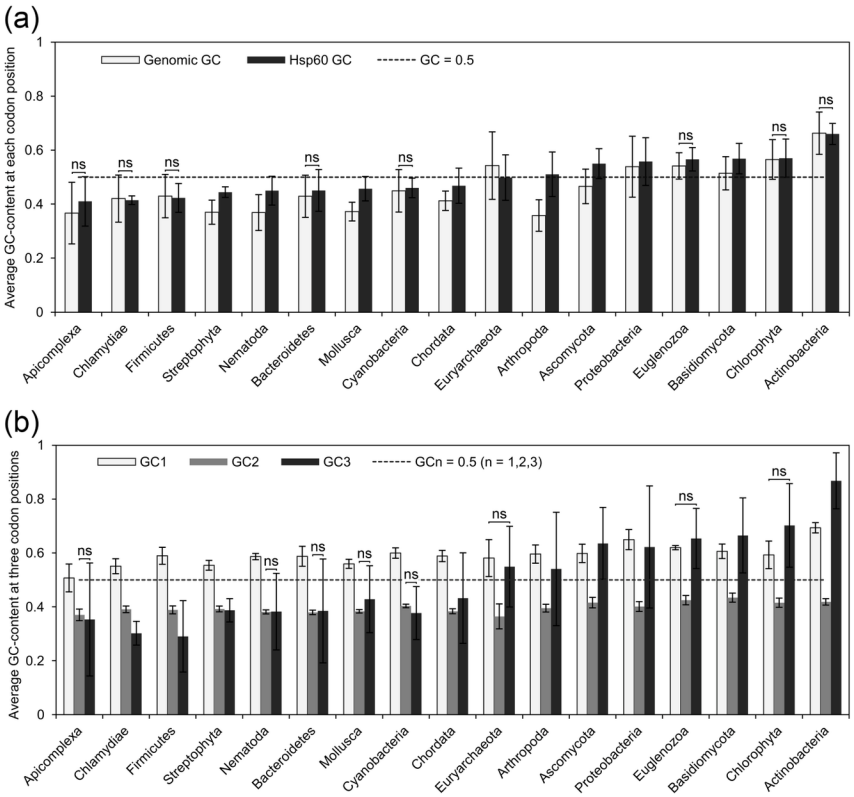

Figure 4. The average nucleotide composition of the Hsp60 genes from 17 phyla: a – The average total GC contents at each positions of codon of Hsp60 sequences and corresponding genomes; b - The average content of GC_1, GC_2, and GC_3 in Hsp60 genes. Phyla were sorted by average total GC content of Hsp60 sequences.

Student's t-test was used to compare the average GC content of the Hsp60 sequences and the average GC content of the corresponding genomes. The difference between two independent samples of GC values is considered statistically significant if the p-value is less than 0.05. Statistically indistinguishable average GC values are marked with "ns" (non-significant).

The average GC content in the Hsp60 genes ranges from 0.41±0.13 (Apicomplexa) to 0.67±0.04 (Actinobacteria) (Figure 4a). As can be seen, the average GC content of almost all Hsp60 genes is comparable to or exceeds the average genomic background. In turn, the opposite is observed for Euryarchaeota. The upward trend in the GC content in the Hsp60 genes may be associated with recombination (GC-biased gene conversion)[46,47], repair[48], and the environmental changes[49,50], in which there is an increase in the frequency of AT-GC substitutions. Thus, it can be assumed that the Hsp60 gene is tightly controlled by DNA repair systems that protect the genetic material from mutations.

To determine the contribution of each of three codon positions to the total GC content of Hsp60 genes from 17 phyla, the GC_1, GC_2, and GC_3 contents were calculated. The average GC_1 values vary from 0.49+-0.06 (Apicomplexa) to 0.69+-0.02 (Actinobacteria), and their contribution to the total GC content of Hsp60 genes is moderate (Figure 4b). At the same time, GC_2 in the range from 0.36+-0.03 (Apicomplexa) to 0.44+-0.02 (Basidiomycota) were the least variable and practically did not affect the GC composition of the Hsp60 genes. These results are obvious since the second codon position is the most conserved. GC_3 values vary from 0.25+-0.15 (Firmicutes) to 0.9+-0.11 (Actinobacteria), which indicates a mutation bias[51]. It should be noted that starting from Euryarchaeota, the average GC_3 values of the Hsp60 genes increase sharply (Figure 4b), and the average GC content becomes more than 0.5 (Figure 4a).

The substitution of nucleotides at the third position of the codon, caused by point mutations or repair processes, does not change the amino acid, but only indicates the mutational pressure for codon usage. According to the theory[35], mutational pressure tends to push the GC content in a gene/genome towards equilibrium (neutrality of codon usage), reducing the heterogeneity caused by natural selection[34]. Equilibrium of the nucleotide composition of the gene/genome, in which selective constraints (factors that reduced the evolutionary divergence of the functional sequence) do not affect the GC content, is achieved when the frequencies of the AT-GC and GC-AT mutations are equal[34]. These mutations can be fixed or removed from the population by natural selection and random genetic drift[35]. The frequencies of the AT-GC and GC-AT mutations at the third position of the codons also reflect the direction of the mutational pressure. In general, the GC_3 value is less than 0.5 when the gene is under the influence of AT pressure, and *vice versa* [52]. Thus, we can initially identify two groups of Hsp60 genes that differ in the direction of mutational pressure (Figure 4b). The AT-group includes Apicomplexa, Chlamydiae, Firmicutes, Streptophyta, Nematoda, Bacteroidetes, Mollusca, Cyanobacteria, and Chordata, which have an average total GC content of less than 0.5 in the Hsp60 genes. In turn, the phyla Euryarchaeota, Arthropoda, Ascomycota, Proteobacteria, Euglenozoa, Basidiomycota, Chlorophyta, and Actinobacteria form a GC-group with an average total GC content of more than 0.5. However, the threshold of 0.5 is nominal due to the imbalance between the rates of mutation and repair processes[53]. Therefore, a neutrality analysis was carried out to clarify the direction of the mutational pressure and to reveal the degree of its influence on the codon usage with the determination of the equilibrium point (Figure 5).

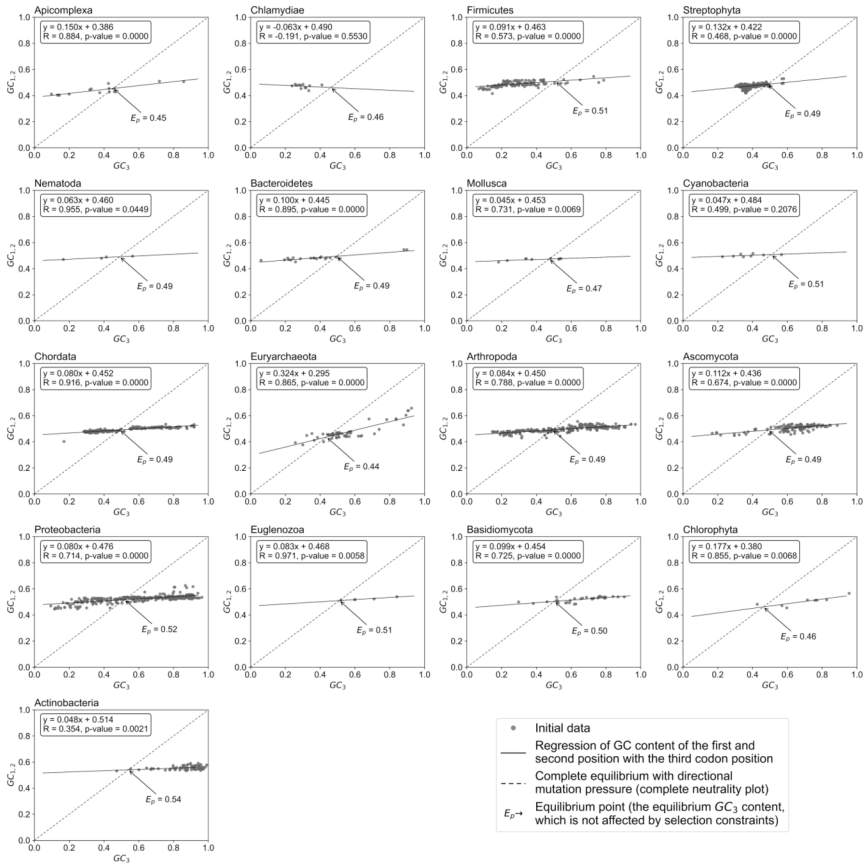

Figure 5. Neutrality plots ($GC_{1,2}$ vs. GC_3) for Hsp60 genes from 17 phyla. The $GC_{1,2}$ values represent the average GC content at the first and second positions of codon (GC_1 and GC_2), while GC_3 values represent the GC content at the third synonymous codon position. The solid line represents the linear regression of $GC_{1,2}$ versus GC_3, the correlation of which is described by the regression coefficient R and its p-value. The correlation coefficient R reflects the strength of the impact of GC_3 on $GC_{1,2}$. The p-value characterizes the significance of R. Changes in GC_3 values actually affect the $GC_{1,2}$ values when the p-value of R is less than 0.05. In turn, changes in GC_3 are considered random, and the R coefficient is not irrelevant when the p-value is greater than 0.05, i.e. GC_3 and $GC_{1,2}$ values are not correlated. The slope ϵ of the regression line indicates the neutrality of the codon usage. Neutrality values were determined by equation [$\epsilon \times 100$, %]. Slope values ranging from 0 to 1 were calculated using the least-squares regression analysis. The dashed line is a complete neutrality plot, which reflects the complete equilibrium of the nucleotide composition of the gene/genome with directional mutation pressure. The equilibrium point E_p was defined as the intersection point of the neutrality plot (regression line) and the complete neutrality plot. The $E_p$value reflects the GC_3 content of the gene/genome when the mutation frequencies (AT-GC and GC-AT) are equal. The direction of the mutational pressure, indicating an imbalance in the frequencies of the AT-GC and GC-AT mutations, was determined in accordance with the following conditions: the average GC content value less than theE_p value reflects the AT mutational pressure; the average GC content value greater than the E_p value reflects the GC mutational pressure. Phyla were sorted by average total GC content of Hsp60 genes.

The results of the neutrality analysis reflect statistically significant correlations between the $GC_{1,2}$ and GC_3values of the Hsp60 genes from 15 phyla (Figure 5). Taking into account the results of neutrality analysis and difference between the average GC_3 values and the corresponding values at equilibrium points, 17 phyla were divided into three groups according to the direction of mutational pressure (Table 2).

Table 2. The results of neutrality plot analysis

Run	Kingdom	Phylum[b]	Average GC_3 content	Neutrality, $\epsilon \times 100\%$	Equilibrium point E_p^c	Direction of mutational pressure
1	Protozoa[a]	Apicomplexa	0.35±0.21	15	0.45	AT
2	Bacteria	Chlamydiae	0.3±0.04	6.3*	0.46	AT
3	Bacteria	Firmicutes	0.29±0.13	9.1	0.51	AT
4	Plantae	Streptophyta	0.39±0.04	13.2	0.49	AT
5	Metazoa	Nematoda	0.38±0.14	6.3	0.49	AT
6	Bacteria	Bacteroidetes	0.38±0.19	10	0.49	AT
7	Metazoa	Mollusca	0.4±0.13	4.5	0.47	AT
8	Bacteria	Cyanobacteria	0.38±0.1	4.7*	0.51	AT
9	Metazoa	Chordata	0.45±0.17	8	0.49	AT/GC
10	Metazoa	Arthropoda	0.55±0.15	8.4	0.49	AT/GC
11	Bacteria	Proteobacteria	0.54±0.21	8	0.52	AT/GC
12	Archaea	Euryarchaeota	0.64±0.13	32.4	0.44	GC
13	Fungi	Ascomycota	0.62±0.23	11.2	0.49	GC
14	Protozoa	Euglenozoa	0.65±0.11	8.3	0.51	GC
15	Fungi	Basidiomycota	0.67±0.14	9.9	0.50	GC
16	Plantae	Chlorophyta	0.7±0.16	17.7	0.46	GC
17	Bacteria	Actinobacteria	0.87±0.1	4.8	0.54	GC

Run	Kingdom	Phylum[b]	Average GC_3 content	Neutrality, $\epsilon \times 100\%$	Equilibrium point E_p^c	Direction of mutational pressure
Note: The neutrality values with statistically non-significant correlation coefficient R (p-value > 0.05) are marked with symbol (*). [a] The designation “Protozoa” was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy. [b] Phyla were divided by the direction of the mutational pressure and sorted by average GC content. [c] The value of equilibrium point was calculated as $\frac{A_0}{(1-\varepsilon)}$, where A_0 is a value computed from regression of $GC_{1,2}$ and GC_3 at $GC_3 = 0$	Note: The neutrality values with statistically non-significant correlation coefficient R (p-value > 0.05) are marked with symbol (*). [a] The designation “Protozoa” was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy. [b] Phyla were divided by the direction of the mutational pressure and sorted by average GC content. [c] The value of equilibrium point was calculated as $\frac{A_0}{(1-\varepsilon)}$, where A_0 is a value computed from regression of $GC_{1,2}$ and GC_3 at $GC_3 = 0$	Note: The neutrality values with statistically non-significant correlation coefficient R (p-value > 0.05) are marked with symbol (*). [a] The designation “Protozoa” was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy. [b] Phyla were divided by the direction of the mutational pressure and sorted by average GC content. [c] The value of equilibrium point was calculated as $\frac{A_0}{(1-\varepsilon)}$, where A_0 is a value computed from regression of $GC_{1,2}$ and GC_3 at $GC_3 = 0$	Note: The neutrality values with statistically non-significant correlation coefficient R (p-value > 0.05) are marked with symbol (*). [a] The designation “Protozoa” was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy. [b] Phyla were divided by the direction of the mutational pressure and sorted by average GC content. [c] The value of equilibrium point was calculated as $\frac{A_0}{(1-\varepsilon)}$, where A_0 is a value computed from regression of $GC_{1,2}$ and GC_3 at $GC_3 = 0$	Note: The neutrality values with statistically non-significant correlation coefficient R (p-value > 0.05) are marked with symbol (*). [a] The designation “Protozoa” was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy. [b] Phyla were divided by the direction of the mutational pressure and sorted by average GC content. [c] The value of equilibrium point was calculated as $\frac{A_0}{(1-\varepsilon)}$, where A_0 is a value computed from regression of $GC_{1,2}$ and GC_3 at $GC_3 = 0$	Note: The neutrality values with statistically non-significant correlation coefficient R (p-value > 0.05) are marked with symbol (*). [a] The designation “Protozoa” was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy. [b] Phyla were divided by the direction of the mutational pressure and sorted by average GC content. [c] The value of equilibrium point was calculated as $\frac{A_0}{(1-\varepsilon)}$, where A_0 is a value computed from regression of $GC_{1,2}$ and GC_3 at $GC_3 = 0$	Note: The neutrality values with statistically non-significant correlation coefficient R (p-value > 0.05) are marked with symbol (*). [a] The designation “Protozoa” was used to identify Apicomplexa and Euglenozoa, since kingdom of these phyla is not determined in the NCBI Taxonomy. [b] Phyla were divided by the direction of the mutational pressure and sorted by average GC content. [c] The value of equilibrium point was calculated as $\frac{A_0}{(1-\varepsilon)}$, where A_0 is a value computed from regression of $GC_{1,2}$ and GC_3 at $GC_3 = 0$

Run	Kingdom	Phylum[b]	Average GC_3 content	Neutrality, $\epsilon \times 100\%$	Equilibrium point E_p^c	Direction of mutational pressure

According to the data obtained, the main direction of mutational pressure was determined for Hsp60 genes of 14 phyla, with the exception of Chordata, Arthropoda, and Proteobacteria (Table 2). However, the peculiarities of the neutrality plot of Chordata should be noted (Figure 6).

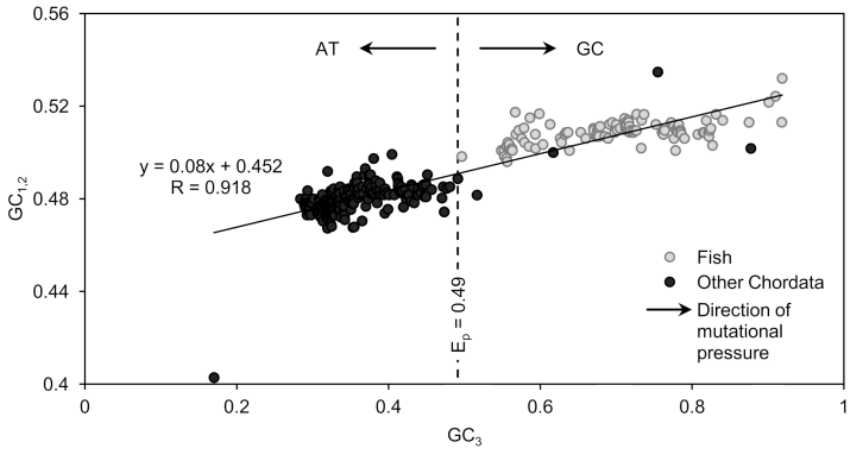

Figure 6. Neutrality plot for Hsp60 genes of Chordata

As can be seen, the Hsp60 genes can be divided into two groups. The AT-biased group includes the Hsp60 genes from Mammalia, Aves, Reptilia, and Amphibia. In turn, the second group with GC-bias includes all Hsp60 genes from Fish and five Hsp60 genes belonging to Hyperoartia (*Petromyzon marinus* , GC_3 0.92), Mammalia (*Ornithorhynchus anatinus* , GC_3 0.88;*Lipotes vexillifer* , GC_3 0.76), and Leptocardii (*Branchiostoma floridae* , GC_3 0.62; *Gekko japonicus* , GC_3 0.52). Interestingly, four of these organisms, with the exception of *Gekko japonicus* , are aquatic. On the other hand, there is the Hsp60 gene from *Boleophthalmus pectinirostris* , belonging to Fish, with GC_3 of 0.496, which is closest to the equilibrium GC_3 of Chordata. This mudskipper is an amphibious fish capable of feeding on land. It has been suggested that codon usage bias due to substitutions of synonymous codons may be associated with the lifestyle of organisms[54–56]. Based on this, we can assume that the habitat or specific diet can influence the synonymous codon substitution in the Hsp60 genes of the mentioned organisms.

Neutrality of codon usage for all phyla, with the exception of Euryarchaeota, was less than 20%, indicating that natural selection dominates mutational pressure at the third position of codons in these Hsp60 genes (Table 2). At the same time, the neutrality of 32.4% for the Hsp60 genes from Euryarchaeota is the highest among all 17 phyla and indicates a rather high degree of mutational pressure. The high correlation coefficient R of 0.865 reflects the important role of mutational pressure as a factor affecting codon usage[57].

Thus, the high resistance to spontaneous mutations provided by DNA repair systems, low mutational pressure, and constant amino acid composition (see Amino acid composition of Hsp60) suggest that the Hsp60 gene is a trait inherent in all organisms[58].

Codon usage biases

The degree of codon usage bias, or, in other words, the size of the “codon dictionary” needs to be assessed. The effective number of codons (ENC) is one of the most widely used metrics for this purpose. In this study, ENC values were calculated using the GC_3content of the Hsp60 genes for each of the 17 phyla (Figure 7).

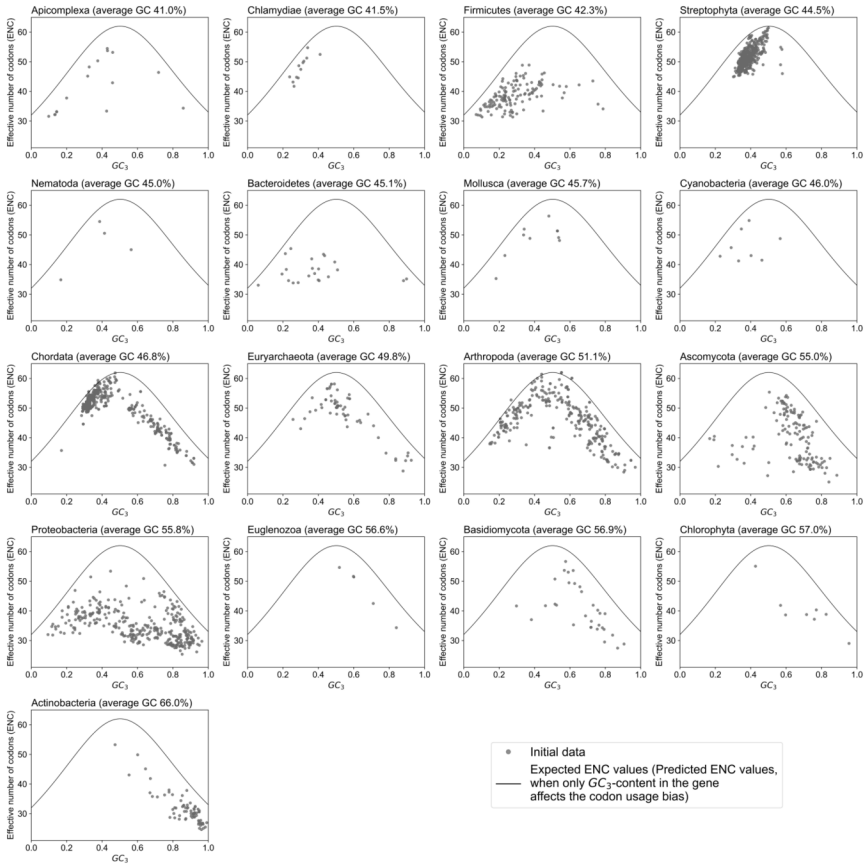

Figure 7. Nc-plots of codon usage bias in Hsp60 genes from the 17 phyla. Gray scatter plots represent ENC values versus GC_3 content for Hsp60 genes from 17 phyla. The black bell-shaped curves represent the expected effective number of codons (ENC_{exp}), i.e. predicted ENC values if codon usage bias is influenced by GC_3 content (GC content at the third synonymous position of codons) in the Hsp60 gene only. Phyla were sorted by the average total GC content of Hsp60 genes.

Almost all Hsp60 genes have lower ENC values than expected, suggesting that the pressure of mutations affects codon usage[59]. As with the neutrality plot analysis (see above), attention should be paid to the Nc-plot for Chordata, where the ENC values are divided. Accordingly, the ENC values of the Hsp60 genes from Chordata were grouped using taxonomy (class) (Figure 8).

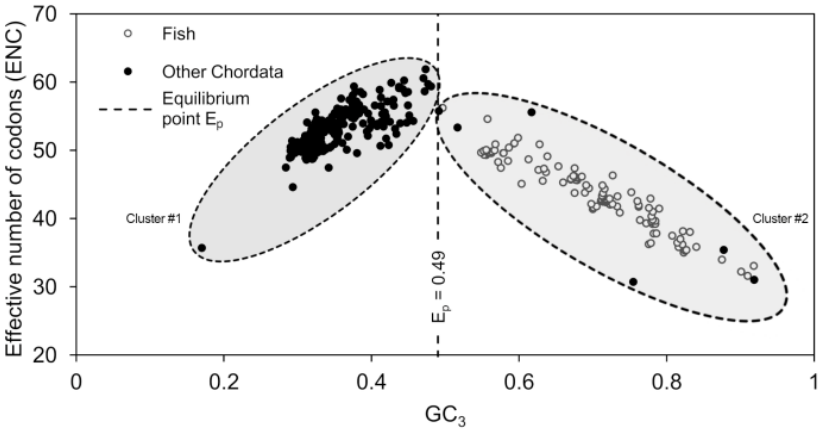

Figure 8. Clustering of ENC values for Hsp60 genes from Chordata. Cluster #1 includes the ENC values of Hsp60 genes of Mammalia, Aves, Reptilia, and Amphibia. Cluster #2 includes ENC values of Hsp60 genes of Fish. Clustering was performed using the value of the GC_3 content corresponding to the equilibrium pointE_p , which was determined earlier (see GC-content and mutation pressure for codon usage).

As can be seen, the ENC values of the Hsp60 genes from Chordata form two groups (Figure 8). Hsp60 genes from cluster #1 (GC_3 is less than E_p) belong to Mammalia, Aves, Reptilia, and Amphibia. In turn, cluster #2 (GC_3 is larger thanE_p) consists only of Hsp60 genes from Actinopterygii (ray-finned fishes). The ENC values and the GC_3 content of the Hsp60 genes in cluster #2 differ more than in cluster #1. This feature can be explained by the fact that Hsp60 from Fish is evolutionarily far from Hsp60 from other Chordata classes[14].

It should be noted that the ENC value of the gene can correlate with the level of its expression in the cell[59–62]. Based on current research, the following ENC thresholds have been established for determining the level of expression of the Hsp60 gene: ENC $<$ 40 for highly expressed genes[59,63,64]; 40 $<$ ENC [?] 55 for moderately expressed genes[59,63,65]; ENC $>$ 55 for lowly expressed genes[60,63]. These conditions can be used to assess the main trends in the average codon usage bias for Hsp60 genes (Figure 9).

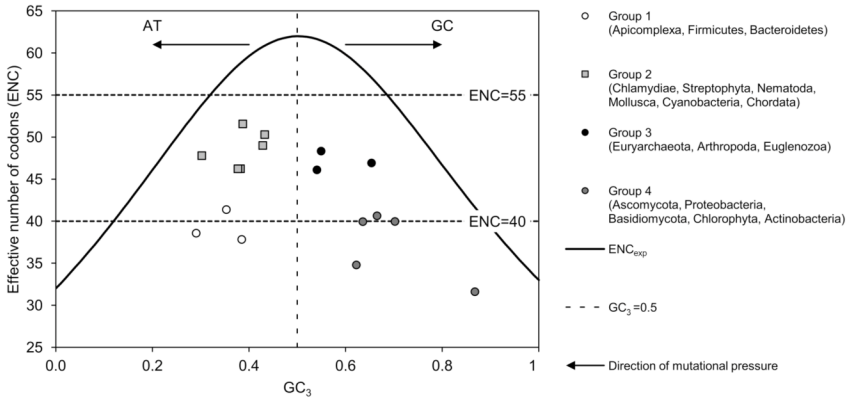

Figure 9. Nc-plot of the average codon usage bias in the Hsp60 genes of 17 phyla. The plot space was divided into six quadrants using the ENC and GC_3 thresholds. The ENC thresholds, reflecting the level of the Hsp60 gene expression, were as follows: ENC $<$ 40 for genes with high expression; 40 $<$ ENC [?] 55 for moderately

expressed genes; ENC > 55 for low expressed genes. The GC_3 thresholds reflecting the direction of the mutational pressure were as follows: $GC_3 < 0.5$ represents the AT-mutation pressure; $GC_3 > 0.5$ represents the GC-mutation pressure. The average ENC values were grouped according to the obtained quadrants: Group 1 (Apicomplexa, Firmicutes, and Bacteroidetes); Group 2 (Chlamydiae, Streptophyta, Nematoda, Mollusca, Cyanobacteria, and Chordata); Group 3 (Euryarchaeota, Arthropoda, and Euglenozoa); Group 4 (Ascomycota, Proteobacteria, Basidiomycota, Chlorophyta, and Actinobacteria). The black bell-shaped curve represents the expected effective number of codons (ENC_{exp}), i.e. predicted ENC values if the codon bias is influenced only by the GC content at the third synonymous position of codons (GC_3) in the Hsp60 gene. The horizontal dashed lines (ENC=40 and ENC=55) indicate ENC thresholds for determining the codon usage bias and gene expression level. The vertical dashed line indicates the GC_3 content of 0.5. The position of the ENC value regarding this line indicates the direction of the mutation pressure affecting the Hsp60 genes (depicted by arrows).

The maximum and minimum average ENC values were observed for Streptophyta (ENC 51.6±3.2) and Actinobacteria (ENC 31.6±5.3), respectively. Using the convention of 40 < ENC [?] 55 to determine the degree of codon usage bias[60,66,67], it can be assumed that Hsp60 genes of 17 phyla are highly and moderately biased. The Hsp60 genes from Group 1 (Apicomplexa, Firmicutes, and Bacteroidetes) with AT-mutation pressure and Group 4 (Ascomycota, Proteobacteria, Basidiomycota, Chlorophyta, and Actinobacteria) with GC-mutation pressure have a high level of expression. In turn, a moderate level of Hsp60 gene expression is characteristic of Group 2 (Chlamydiae, Streptophyta, Nematoda, Mollusca, Cyanobacteria, Chordata) with AT-mutation pressure and Group 3 (Euryarchaeota, Arthropoda, and Euglenozoa) with GC-mutation pressure.

However, it should be noted that the individual ENC values of the Hsp60 genes from Streptophyta, Mollusca, Chordata, Euryarchaeota, Arthropoda, Basidiomycota exceed 55 (Figure 7). In addition, there was a high statistical variance in the average ENC values ranging from 3.2 (Streptophyta) to 8.8 (Apicomplexa) (Supplementary, ENC) for Hsp60 genes of 17 phyla, which may indicate that overall ENC values are not conserved for these genes within the phylum. Therefore, the ENC values were compared using Student's t-test (Figure 10; Supplementary, ENC).

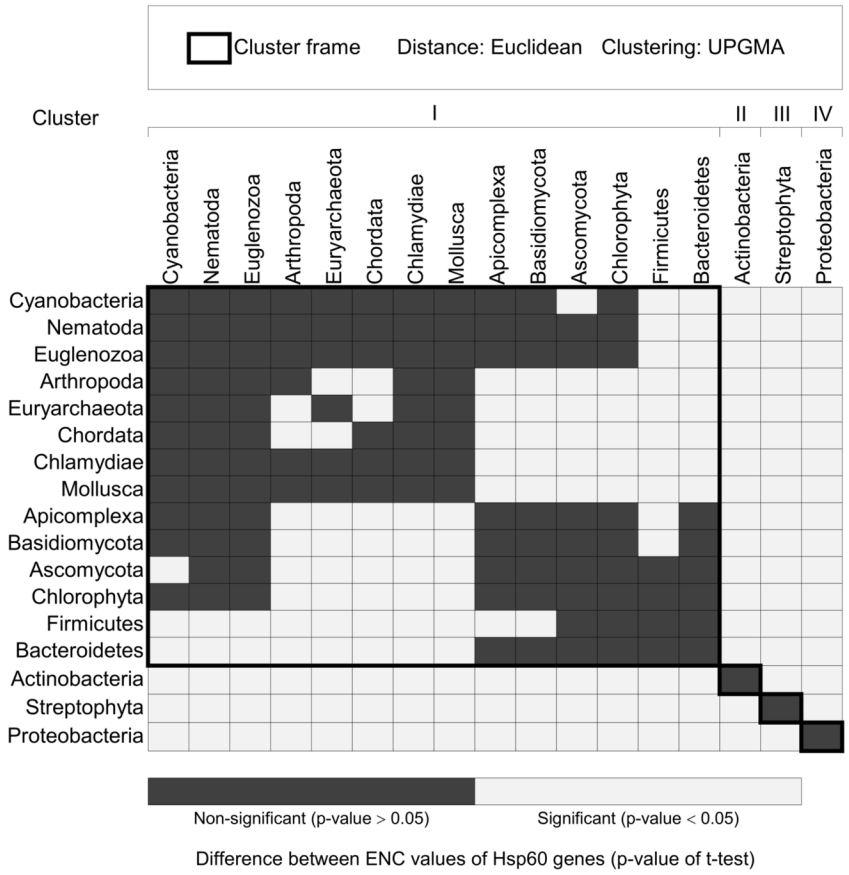

Figure 10. Symmetric matrix of p-values of the t-test between the ENC values of the Hsp60 genes from 17 phyla. The statistically indistinguishable ENC values of Hsp60 genes of the two phyla, having a t-test with a p-value greater than 0.05, are marked in black. Statistically different ENC values of the Hsp60 genes of the two phyla, having a t-test with a p-value less than 0.05, are marked in white. The black frames and the "Cluster" X-axis represent the four clusters and their numbers (Roman numerals). Clustering was carried out using the UPGMA algorithm.

According to the results of the t-test (Supplementary, ENC (t-test results)), the ENC values of Hsp60 genes of 14 phyla (cluster I) are statistically indistinguishable (Figure 10). In turn, the ENC values for Actinobacteria (cluster II), Streptophyta (cluster III), and Proteobacteria (cluster IV) are statistically different from all other phyla. It can be assumed that the expression level of Hsp60 genes from Proteobacteria and Actinobacteria is high (34.8 and 31.6, respectively), while the expression level of the Hsp60 genes from Streptophyta is close to low (51.6). It should be noted that more accurate conclusions about the level of Hsp60 gene expression should be made based on the results obtained by conventional laboratory methods.

Patterns of synonymous codon usage for Hsp60 genes

To understand the peculiarities of codon usage, it is necessary to assess not only the degree of codon usage bias, but also the representation of synonymous codons in a gene, or, in other words, the patterns of synonymous codon usage. The approach widely used for this purpose is the calculation of synonymous codon usage bias, reflecting the peculiarities of amino acid coding, when some codons in a gene are used more often than other synonymous codons[68]. Under the influence of the mutational pressure and natural selection[69], the representation of various codons in a gene plays an important role in the processes of its transcription[70]

and translation[71]. Determination of patterns of synonymous codon usage is necessary for understanding the features of gene/genome evolution. Thus, the relative synonymous codon usage (RSCU) of 64 codons in the Hsp60 genes was calculated (Figure 11; Supplementary, RSCU).

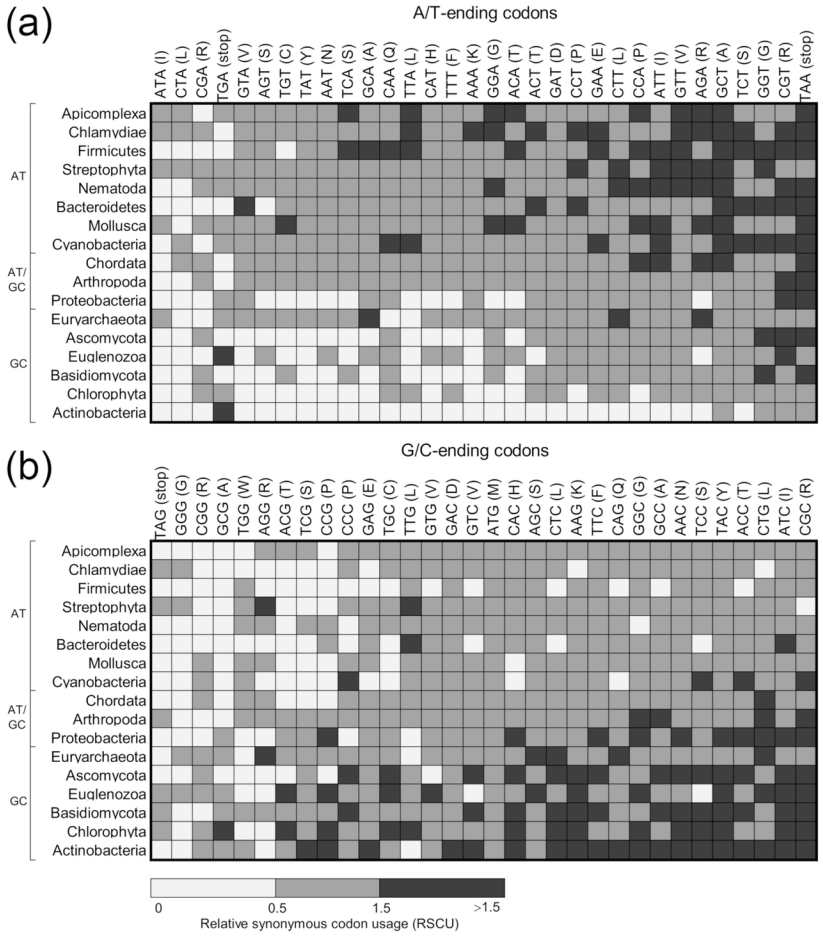

Figure 11. Average values of relative synonymous codon usage (RSCU) for Hsp60 genes from 17 phyla. The RSCU values for each of 17 phyla were divided into two main groups according to the type of base at the third synonymous position of codon: a - A/T-ending codons; b - G/C-ending codons. Phyla were divided into three groups by the direction of the mutational pressure (see Table 2) and sorted by the average total GC content of Hsp60 genes. The codons were sorted by the average RSCU value between 17 phyla.

The RSCU values of A/T-ending codons were higher for the Hsp60 genes under the influence of AT mutation pressure, and *vice versa*(Figure 11), which was to be expected[72]. The RSCU values greater than 1.5 and less than 0.5 correspond to high and low represented codons, respectively. Among the stop codons, TAA is the most widely represented stop codon with an average RSCU value of 1.8±0.7. In turn, TGA codon is widely used in the Hsp60 genes of Euglenozoa and Actinobacteria with the average RSCU values of 1.8±1.6 and 1.8±1.5, respectively. This may be due to the direction of the mutational pressure affecting the Hsp60 genes of these phyla, since with an increase in the GC content, the frequency of TGA codon usage increases[73]. This may also explain the decrease in the average RSCU values for TAA for GC-biased phyla (Figure 11).

To assess the main trends in the synonymous codon usage in Hsp60 genes from phyla with different directional

mutation pressure, summary patterns of RSCU values were compiled (Figure 12).

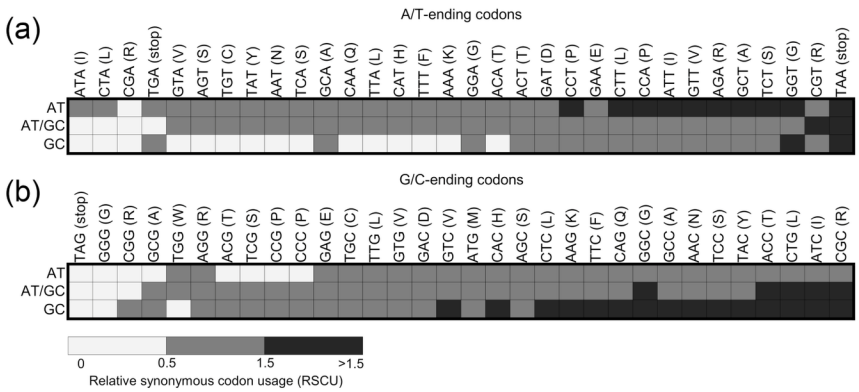

Figure 12. Summary patterns of relative synonymous codon usage for Hsp60 genes being under the different mutational pressure. The average RSCU values of Hsp60 genes from phyla with the AT, AT/GC, and GC mutational pressure were divided into two main groups according to the type of base at the third synonymous position of codon: a - A/T-ending codons; b - G/C-ending codons. The codons were sorted by the average RSCU value between all 17 phyla.

As can be seen, the number of high and low represented codons that correspond to phyla under the GC mutational pressure is greater than for AT and AT/GC-biased phyla. Such results for AT/GC-biased phyla can be explained by their GC content at the border between the AT and GC mutational pressure groups (Table 2). Codons encoding Ala, Arg, Gly, Ile, Leu, Ser, Thr, and Val were widely represented for almost all Hsp60 genes. For each of 17 phyla, the set of these codons is consistent with the direction of the mutational pressure (predominance of A/T- or G/C-ending codons) affecting the Hsp60 genes. CGA (R), GGG (G), and TAG (stop) codons have the lowest average RSCU values for all 17 phyla (less than 0.5) regardless of the direction of the mutational pressure.

Taking these results into account, it can be assumed that with an increase in the GC content in the Hsp60 gene due to mutations at the third synonymous codon position, the selectivity of codon usage increases.

Conclusion

In the present study, a comprehensive analysis of the amino acid and nucleotide Hsp60 sequences from 19 phyla was carried out. For this, databases 1 and 2 were built, consisting of 19220 amino acid and 1925 nucleotide Hsp60 sequences, respectively.

Multiple alignment, calculation of the percent identity (PID) of amino acid sequences from database 1, and subsequent clustering of the obtained data were performed to establish the level of conservatism of Hsp60. It turned out that Hsp60 cannot be considered a highly conserved protein, since the average PID values vary widely from 10.1±0.5% (Chlamydiae #1 / Apicomplexa #2) to 97.9±0.0% (Mollusca #2 / Mollusca #3). However, the result of component analysis indicates a relatively constant amino acid composition of Hsp60, showing a high content of aliphatic (Ala, Val, Gly, Leu, and Ile), charged (Glu, Lys, and Asp), and polar (Thr) amino acid residues. Thus, it can be assumed that the functional features of Hsp60 are determined not only by its sequence, but also by its amino acid composition.

The nucleotide sequences from database 2 were analyzed to determine the genetic and evolutionary characteristics of Hsp60 genes from 17 phyla using conventional metrics. The GC content of the analyzed Hsp60 genes was comparable to or higher than the corresponding genomic values. It can be assumed that the Hsp60 genes are tightly controlled by DNA repair systems, providing high resistance to spontaneous mutations in the third position of codons and thereby increasing their GC content. It was further found that natural se-

lection plays a dominant role in the evolution of Hsp60 genes. According to the results of the neutrality plot analysis, the percent of impact of mutational pressure on the codon usage of Hsp60 genes in 16 phyla does not exceed 20%, with the exception of Euryarchaeota, for which Hsp60 genes are characterized by high mutational pressure with neutrality values of 32.4%. In addition, the direction of mutational pressure affecting the third position of codons was determined. Accordingly, the Hsp60 genes from Apicomplexa, Chlamydiae, Firmicutes, Streptophyta, Nematoda, Bacteroidetes, Mollusca, and Cyanobacteria are under AT mutational pressure. In turn, GC mutational pressurized Hsp60 genes belong to Euryarchaeota, Ascomycota, Euglenozoa, Basidiomycota, Chlorophyta, and Actinobacteria phyla. It should be noted that the Hsp60 genes from Chordata, Arthropoda, and Proteobacteria phyla cannot be assigned to any of these groups. However, further division by class showed an interesting result for Chordata. The Hsp60 genes from Fish were found to be under GC mutational pressure, while Hsp60 genes from other classes were AT-biased. Also noteworthy are four representatives of Hyperoartia, Mammalia, and Leptocardii classes, whose Hsp60 genes belonging to the Fish's subgroup. This feature may be due to the fact that they are all aquatic animals, which makes them related to Fish.

The values of effective number of codons (ENC) and relative synonymous codon usage (RSCU) were used to assess codon usage and level of codon bias. According to the ENC values, a moderate or high level of Hsp60 gene expression was observed, which is evident since Hsp60 is a ubiquitous protein. At the same time, the direction of mutational pressure in the Hsp60 genes did not affect the size of a "codon dictionary" that is used to encode genes. However, the results of the synonymous codons bias analysis showed that the average RSCU values for A/T-ending codons were higher for Hsp60 genes under AT mutation pressure and *vice versa* . TAA codon was the most preferred stop codon for the Hsp60 genes. Using division by the direction of mutational pressure and the average RSCU values for Hsp60 genes from these groups, it was found that the number of high (RSCU>1.5) and low (RSCU<0.5) represented codons for the Hsp60 genes under GC mutation pressure is greater, than for AT-biased Hsp60 genes. It can be assumed that an increase in the GC content of Hsp60 genes ensures the optimization of codon usage.

Thus, the present study demonstrates that Hsp60 is a protein inherent in all living organisms, characterized by a relatively constant amino acid composition and low sequence conservatism, the feature of which is the dominance of natural selection forces in evolution of the gene and its high resistance to spontaneous mutations in the third synonymous position of codons.

References

1. Mendoza JA, Weinberger KK, Swan MJ. The Hsp60 protein of helicobacter pylori displays chaperone activity under acidic conditions.*Biochem Biophys Rep* . 2016;9:95–99.

2. Yer EN, Baloglu MC, Ayan S. Identification and expression profiling of all Hsp family member genes under salinity stress in different poplar clones. *Gene* . 2018;678:324–336.

3. Sangiorgi C, Vallese D, Gnemmi I, Bucchieri F, et al. Hsp60 activity on human bronchial epithelial cells. *Int J Immunopathol Pharmacol* . 2017;30(4):333–340.

4. Swaroop S, Sengupta N, Suryawanshi AR, Adlakha YK, et al. Hsp60 plays a regulatory role in IL-1β-induced microglial inflammation via TLR4-p38 MAPK axis. *J Neuroinflammation* . 2016;13.

5. Sell H, Poitou C, Habich C, Bouillot J-L, et al. Heat shock protein 60 in obesity: effect of bariatric surgery and its relation to inflammation and cardiovascular risk. *Obesity (Silver Spring)* . 2017;25(12):2108–2114.

6. Wick C. Tolerization against atherosclerosis using heat shock protein 60. *Cell Stress Chaperones* . 2016;21(2):201–211.

7. Hong Y, Long J, Li H, Chen S, et al. An analysis of immunoreactive signatures in early stage hepatocellular carcinoma. *EBioMedicine* . 2015;2(5):438–446.

8. Cappello F, Angileri F, de Macario EC, Macario AJL. Chaperonopathies and chaperonotherapy. Hsp60 as therapeutic target in cancer: potential benefits and risks. *Curr. Pharm. Des.* 2013;19(3):452–457.

9. Abdeen S, Salim N, Mammadova N, Summers CM, et al. Targeting the Hsp60/10 chaperonin systems of *Trypanosoma brucei* as a strategy for treating African sleeping sickness. *Bioorg. Med. Chem. Lett.*2016;26(21):5247–5253.

10. Stevens M, Abdeen S, Salim N, Ray A-M, et al. Hsp60/10 chaperonin systems are inhibited by a variety of approved drugs, natural products, and known bioactive molecules. *Bioorg. Med. Chem. Lett.*2019;29(9):1106–1112.

11. Washburn A, Abdeen S, Ovechkina Y, Ray A-M, et al. Dual-targeting GroEL/ES chaperonin and protein tyrosine phosphatase B (PtpB) inhibitors: A polypharmacology strategy for treating *Mycobacterium tuberculosis* infections. *Bioorg. Med. Chem. Lett.*2019;29(13):1665–1672.

12. Brocchieri L, Karlin S. Conservation among Hsp60 sequences in relation to structure, function, and evolution. *Protein Science* . 2000;9(3):476–486.

13. Karlin S, Brocchieri L. Heat shock protein 60 sequence comparisons: Duplications, lateral transfer, and mitochondrial evolution. *Proc Natl Acad Sci U S A* . 2000;97(21):11348–11353.

14. Tikhomirova TS, Galzitskaya OV. Functionally significant amino acid motifs of heat shock proteins: structural and bioinformatics analyses of Hsp60/Hsp10 in five classes of Chordata. *Mol Biol* . 2018;52(5):761–778.

15. Seddigh S. Proteomics analysis of two heat shock proteins in insects. *J. Biomol. Struct. Dyn.* 2019;37(10):2652–2668.

16. Guo L, Yang H, Tang F, Yin R, et al. Oral immunization with a multivalent epitope-based vaccine, based on NAP, urease, Hsp60, and HpaA, provides therapeutic effect on *H. pylori* infection in Mongolian gerbils. *Front Cell Infect Microbiol* . 2017;7.

17. Marchan J. *In silico* identification of epitopes present in human heat shock proteins (HSPs) overexpressed by tumour cells. *J. Immunol. Methods* . 2019.

18. Huang C-H, Chang M-T, Huang L, Chua W-S. Molecular discrimination and identification of *Acetobacter* genus based on the partial heat shock protein 60 gene (Hsp60) sequences. *J. Sci. Food Agric.*2014;94(2):213–218.

19. Puri A, Rai A, Dhanaraj PS, Lal R, et al. An in silico approach for identification of the pathogenic species, *Helicobacter pylori* and its relatives. *Indian J. Microbiol.* 2016;56(3):277–286.

20. Kwok AYC, Su S-C, Reynolds RP, Bay SJ, et al. Species identification and phylogenetic relationships based on partial Hsp60 gene sequences within the genus Staphylococcus. *Int J Syst Evol Microbiol* . 1999;49(3):1181–1192.

21. Stenico V, Michelini S, Modesto M, Baffoni L, et al. Identification of *Bifidobacterium spp* . using Hsp60 PCR-RFLP analysis: an update.*Anaerobe* . 2014;26:36–40.

22. Zhu L, Li W, Dong X. Species identification of genus Bifidobacterium based on partial Hsp60 gene sequences and proposal of*Bifidobacterium thermacidophilum* subsp. porcinum subsp. nov.*Int J Syst Evol Microbiol* . 2003;53(5):1619–1623.

23. Sakamoto M, Suzuki N, Benno Y. Hsp60 and 16S rRNA gene sequence relationships among species of the genus *Bacteroides* with the finding that *Bacteroides suis* and *Bacteroides tectus* are heterotypic synonyms of *Bacteroides pyogenes* . *Int J Syst Evol Microbiol* . 2010;60(12):2984–2990.

24. Padmadas N, Panda PK, Durairaj S. Binding patterns associated Aß-Hsp60 P458 conjugate to HLA-DR-DRB allele of human in Alzheimer's disease: an *in silico* approach. *Interdiscip Sci Comput Life Sci* . 2016:1–12.

25. Marino C, Krishnan B, Cappello F, Taglialatela G. Hsp60 protects against amyloid β oligomer synaptic toxicity via modification of toxic oligomer conformation. *ACS Chem Neurosci* . 2019.

26. Sievers F, Higgins DG. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.*2014;1079:105–116.

27. Cock PJA, Antao T, Chang JT, Chapman BA, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* . 2009;25(11):1422–1423.

28. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* . 2004;32(5):1792–1797.

29. Kumar S, Stecher G, Peterson D, Tamura K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* . 2012;28(20):2685–2686.

30. Lobanov MYu, Galzitskaya OV. Disordered patterns in clustered protein data bank and in eukaryotic and bacterial proteomes. *PLoS One* . 2011;6(11).

31. Virtanen P, Gommers R, Oliphant TE, Haberland M, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* . 2020;17(3):261–272.

32. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for Python. 2001.

33. Sokal RR, Rohlf FJ. The comparison of dendrograms by objective methods. *Taxon* . 1962;11(2):33–40.

34. Sueoka N. Directional mutation pressure, selective constraints, and genetic equilibria. *J. Mol. Evol.* 1992;34(2):95–114.

35. Sueoka N. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.* 1988;85(8):2653–2657.

36. Fuglsang A. The "effective number of codons" revisited.*Biochem. Biophys. Res. Commun.* 2004;317(3):957–964.

37. Liu X. A more accurate relationship between 'effective number of codons' and GC3s under assumptions of no selection. *Computational Biology and Chemistry* . 2013;42:35–39.

38. Li X, Song H, Kuang Y, Chen S, et al. Genome-wide analysis of codon usage bias in *Epichloë festucae* . *Int J Mol Sci* . 2016;17(7).

39. Boto L. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proc Biol Sci* . 2014;281(1777).

40. Bazzocchi C, Jamnongluk W, O'Neill SL, Anderson TJ, et al. Wsp gene sequences from the *Wolbachia* of filarial nematodes. *Curr. Microbiol.* 2000;41(2):96–100.

41. Bartolucci C, Lamba D, Grazulis S, Manakova E, et al. Crystal structure of wild-type chaperonin GroEL. *J. Mol. Biol.*2005;354(4):940–951.

42. Douglas NR, Reissmann S, Zhang J, Chen B, et al. Dual action of ATP hydrolysis couples lid closure to substrate release into the group II chaperonin chamber. *Cell* . 2011;144(2):240–252.

43. Clare DK, Vasishtan D, Stagg S, Quispe J, et al. ATP-triggered conformational changes delineate substrate-binding and -folding mechanics of the GroEL chaperonin. *Cell* . 2012;149(1):113–123.

44. Nisemblat S, Yaniv O, Parnas A, Frolow F, et al. Crystal structure of the human mitochondrial chaperonin symmetrical football complex.*PNAS* . 2015;112(19):6044–6049.

45. Shimamura T, Koike-Takeshita A, Yokoyama K, Masui R, et al. Crystal structure of the native chaperonin complex from *Thermus thermophilus* revealed unexpected asymmetry at the cis-cavity.*Structure* . 2004;12(8):1471–1480.

46. Lassalle F, Périan S, Bataillon T, Nesme X, et al. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLOS Genetics* . 2015;11(2):e1004941.

47. Niu Z, Xue Q, Wang H, Xie X, et al. Mutational biases and GC-biased gene conversion affect GC content in the plastomes of *Dendrobium*genus. *Int J Mol Sci* . 2017;18(11).

48. Weissman JL, Fagan WF, Johnson PLF. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLoS Genet* . 2019;15(11).

49. Villada JC, Duran MF, Lee PKH. Genomic evidence for simultaneous optimization of transcription and translation through codon variants in the pmoCAB operon of type Ia methanotrophs. *mSystems* . 2019;4(4).

50. Seward EA, Kelly S. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol* . 2016;17(1):226.

51. Tatarinova T, Kerton O. GC3 biology in Eukaryotes and Prokaryotes. In: *DNA Methylation - From Genomics to Technology* . London: IntechOpen; 2012:55–68.

52. Khrustalev VV, Barkovsky EV. Study of completed archaeal genomes and proteomes: hypothesis of strong mutational at pressure existed in their common predecessor. *Genomics, Proteomics & Bioinformatics* . 2010;8(1):22–32.

53. Brown TA. Mutation, Repair and Recombination. In: *Genomes* . 2nd ed. Oxford: Wiley-Liss; 2002:1–35.

54. Botzman M, Margalit H. Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome Biol* . 2011;12(10):R109.

55. Arella D, Dilucca M, Giansanti A. Codon usage bias and environmental adaptation in microbial organisms. *Mol Genet Genomics* . 2021;296(3):751–762.

56. Carbone A, Képès F, Zinovyev A. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Mol Biol Evol* . 2005;22(3):547–561.

57. Khandia R, Singhal S, Kumar U, Ansari A, et al. Analysis of Nipah virus codon usage and adaptation to hosts. *Front. Microbiol.*2019;10.

58. Martiny AC, Treseder K, Pusch G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J* . 2013;7(4):830–838.

59. Wright F. The "effective number of codons" used in a gene.*Gene* . 1990;87(1):23–29.

60. Butt AM, Nasrullah I, Tong Y. Genome-wide analysis of codon usage and influencing factors in Chikungunya viruses. *PLoS ONE* . 2014;9(3):e90905.

61. Guan D-L, Ma L-B, Khan MS, Zhang X-X, et al. Analysis of codon usage patterns in *Hirudinaria manillensis* reveals a preference for GC-ending codons caused by dominant selection constraints. *BMC Genomics* . 2018;19.

62. Yi S, Li Y, Wang W. Selection shapes the patterns of codon usage in three closely related species of genus *Misgurnus* .*Genomics* . 2018;110(2):134–142.

63. Chamani Mohasses F, Solouki M, Ghareyazie B, Fahmideh L, et al. Correlation between gene expression levels under drought stress and synonymous codon usage in rice plant by *in-silico* study.*PLoS One* . 2020;15(8).

64. Xu Q, Chen H, Sun W, Zhu D, et al. Genome-wide analysis of the synonymous codon usage pattern of *Streptococcus suis* .*Microbial Pathogenesis* . 2021;150:104732.

65. Hussain S, Rasool ST, Asif AH. A detailed analysis of synonymous codon usage in human bocavirus. *Arch Virol* . 2019;164(2):335–347.

66. Cho M, Kim H, Son HS. Codon usage patterns of LT-Ag genes in polyomaviruses from different host species. *Virol J* . 2019;16.

67. Tyagi A, Kumar BTN, Singh NK. Genome dynamics and evolution of codon usage patterns in shrimp viruses. *Arch Virol* . 2017;162(10):3137–3142.

68. Behura SK, Severson DW. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. *Biol Rev Camb Philos Soc* . 2013;88(1):49–61.

69. Chen Y. A comparison of synonymous codon usage bias patterns in DNA and RNA virus genomes: quantifying the relative importance of mutational pressure and natural selection. *Biomed Res Int* . 2013;2013:406342.

70. Zhou Z, Dang Y, Zhou M, Li L, et al. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. *PNAS* . 2016;113(41):E6117–E6125.

71. Frumkin I, Lajoie MJ, Gregg CJ, Hornung G, et al. Codon usage of highly expressed genes affects proteome-wide translation efficiency.*PNAS* . 2018;115(21):E4940–E4949.

72. Ermolaeva MD. Synonymous codon usage in bacteria. *Curr Issues Mol Biol* . 2001;3(4):91–97.

73. Korkmaz G, Holm M, Wiens T, Sanyal S. Comprehensive analysis of stop codon usage in bacteria and its correlation with release factor abundance. *J Biol Chem* . 2014;289(44):30334–30342.

Figure captions

Figure 1. Symmetric matrix of the average PID values. The matrix contains 60 sub-clusters of Hsp60 sequences from 19 phyla. The X- and Y-axis items "Sub-cluster" are represented in the following format "Phylum #sub-cluster (number of sequences in a sub-cluster)". Sub-clusters of Viruses have no phylum labels. The Y-axis "Kingdom" represents sub-clusters united by a higher taxonomic rank (Kingdom). The black frames and the X-axis "Cluster" show four clusters and their numbers (Roman numerals), which were obtained by clustering 60 sub-clusters. Clustering was performed using the UPGMA algorithm.

Figure 2. Heat map showing normalized average PID values between Hsp60 sequences belonging to sub-clusters Arthropoda #2, Arthropoda #3, and Nematoda #1 and Hsp60 sequences belonging to each of 19 phyla. The average PID values were normalized using the min-max normalization method. The 19 phyla were sorted by NCBI Taxonomy.

Figure 3. The average amino acid composition of the Hsp60 sequences from 19 phyla: a - Heat map displaying normalized average PID values between 19 phyla of Hsp60 sequences; b - The average amino acid composition of the Hsp60 sequences for each of the 19 phyla compared to the corresponding proteomic values. In Figure 3a the average values were normalized using the min-max normalization method. The line "Summary" presents the average normalized amino acid composition of Hsp60 for 19 phyla. In Figure 3b the amino acid profiles were represented as the average amino acid composition of Hsp60 for each of 19 phyla compared to the average amino acid composition of the respective proteomes. The structure of color scale is as follows: Higher/Lower - the amino acid content in Hsp60 is higher/lower than in proteomes, respectively; Comparable - the amino acid content in Hsp60 is comparable to the average proteomic value. The "Summary" line shows the average amino acid profile of Hsp60 for 19 phyla. The groups were sorted using the NCBI Taxonomy. Amino acids were sorted using an average amino acid composition of 19220 Hsp60 sequences.

Figure 4. The average nucleotide composition of the Hsp60 genes from 17 phyla: a – The average total GC contents at each positions of codon of Hsp60 sequences and corresponding genomes; b - The average content of GC_1, GC_2, and GC_3 in Hsp60 genes. Phyla were sorted by average total GC content of Hsp60 sequences. Student's t-test was used to compare the average GC content of the Hsp60 sequences and the average GC

content of the corresponding genomes. The difference between two independent samples of GC values is considered statistically significant if the p-value is less than 0.05. Statistically indistinguishable average GC values are marked with "ns" (non-significant).

Figure 5. Neutrality plots ($GC_{1,2}$ vs. GC_3) for Hsp60 genes from 17 phyla. The $GC_{1,2}$ values represent the average GC content at the first and second positions of codon (GC_1 and GC_2), while GC_3 values represent the GC content at the third synonymous codon position. The solid line represents the linear regression of $GC_{1,2}$ versus GC_3, the correlation of which is described by the regression coefficient R and its p-value. The correlation coefficient R reflects the strength of the impact of GC_3 on $GC_{1,2}$. The p-value characterizes the significance of R. Changes in GC_3 values actually affect the $GC_{1,2}$ values when the p-value of R is less than 0.05. In turn, changes in GC_3 are considered random, and the R coefficient is not irrelevant when the p-value is greater than 0.05, i.e. GC_3 and $GC_{1,2}$ values are not correlated. The slope ϵ of the regression line indicates the neutrality of the codon usage. Neutrality values were determined by equation [$\epsilon \times 100$, %]. Slope values ranging from 0 to 1 were calculated using the least-squares regression analysis. The dashed line is a complete neutrality plot, which reflects the complete equilibrium of the nucleotide composition of the gene/genome with directional mutation pressure. The equilibrium point E_p was defined as the intersection point of the neutrality plot (regression line) and the complete neutrality plot. The $E_p$value reflects the GC_3 content of the gene/genome when the mutation frequencies (AT-GC and GC-AT) are equal. The direction of the mutational pressure, indicating an imbalance in the frequencies of the AT-GC and GC-AT mutations, was determined in accordance with the following conditions: the average GC content value less than theE_p value reflects the AT mutational pressure; the average GC content value greater than the E_p value reflects the GC mutational pressure. Phyla were sorted by average total GC content of Hsp60 genes.

Figure 6. Neutrality plot for Hsp60 genes of Chordata

Figure 7. Nc-plots of codon usage bias in Hsp60 genes from the 17 phyla. Gray scatter plots represent ENC values versus GC_3 content for Hsp60 genes from 17 phyla. The black bell-shaped curves represent the expected effective number of codons (ENC_{exp}), i.e. predicted ENC values if codon usage bias is influenced by GC_3 content (GC content at the third synonymous position of codons) in the Hsp60 gene only. Phyla were sorted by the average total GC content of Hsp60 genes.

Figure 8. Clustering of ENC values for Hsp60 genes from Chordata. Cluster #1 includes the ENC values of Hsp60 genes of Mammalia, Aves, Reptilia, and Amphibia. Cluster #2 includes ENC values of Hsp60 genes of Fish. Clustering was performed using the value of the GC_3 content corresponding to the equilibrium pointE_p , which was determined earlier (see GC-content and mutation pressure for codon usage).

Figure 9. Nc-plot of the average codon usage bias in the Hsp60 genes of 17 phyla. The plot space was divided into six quadrants using the ENC and GC_3 thresholds. The ENC thresholds, reflecting the level of the Hsp60 gene expression, were as follows: ENC < 40 for genes with high expression; $40 <$ ENC [?] 55 for moderately expressed genes; ENC > 55 for low expressed genes. The GC_3 thresholds reflecting the direction of the mutational pressure were as follows: $GC_3 < 0.5$ represents the AT-mutation pressure; $GC_3 > 0.5$ represents the GC-mutation pressure. The average ENC values were grouped according to the obtained quadrants: Group 1 (Apicomplexa, Firmicutes, and Bacteroidetes); Group 2 (Chlamydiae, Streptophyta, Nematoda, Mollusca, Cyanobacteria, and Chordata); Group 3 (Euryarchaeota, Arthropoda, and Euglenozoa); Group 4 (Ascomycota, Proteobacteria, Basidiomycota, Chlorophyta, and Actinobacteria). The black bell-shaped curve represents the expected effective number of codons (ENC_{exp}), i.e. predicted ENC values if the codon bias is influenced only by the GC content at the third synonymous position of codons (GC_3) in the Hsp60 gene. The horizontal dashed lines (ENC=40 and ENC=55) indicate ENC thresholds for determining the codon usage bias and gene expression level. The vertical dashed line indicates the GC_3 content of 0.5. The position of the ENC value regarding this line indicates the direction of the mutation pressure affecting the Hsp60 genes (depicted by arrows).

Figure 10. Symmetric matrix of p-values of the t-test between the ENC values of the Hsp60 genes from 17 phyla. The statistically indistinguishable ENC values of Hsp60 genes of the two phyla, having a t-test with

a p-value greater than 0.05, are marked in black. Statistically different ENC values of the Hsp60 genes of the two phyla, having a t-test with a p-value less than 0.05, are marked in white. The black frames and the "Cluster" X-axis represent the four clusters and their numbers (Roman numerals). Clustering was carried out using the UPGMA algorithm.

Figure 11. Average values of relative synonymous codon usage (RSCU) for Hsp60 genes from 17 phyla. The RSCU values for each of 17 phyla were divided into two main groups according to the type of base at the third synonymous position of codon: a - A/T-ending codons; b - G/C-ending codons. Phyla were divided into three groups by the direction of the mutational pressure (see Table 2) and sorted by the average total GC content of Hsp60 genes. The codons were sorted by the average RSCU value between 17 phyla.

Figure 12. Summary patterns of relative synonymous codon usage for Hsp60 genes being under the different mutational pressure. The average RSCU values of Hsp60 genes from phyla with the AT, AT/GC, and GC mutational pressure were divided into two main groups according to the type of base at the third synonymous position of codon: a - A/T-ending codons; b - G/C-ending codons. The codons were sorted by the average RSCU value between all 17 phyla.