International Journal of
Molecular Sciences

MDPI

Article

Multidimensional Gene Space as an Approach for Analyzing the Organization of Genomes

Konstantin Zaytsev, Natalya Bogatyreva and Alexey Fedorov *

Bach Institute of Biochemistry, Federal Research Center of Biotechnology of the Russian Academy of Sciences, 119071 Moscow, Russia
* Correspondence: a.fedorov@fbras.ru

Abstract

Genomic organization and its comparative analysis throughout all major kingdoms of life are extensively studied across multiple scales, ranging from individual gene-level analyses to system-wide investigations. This work introduces a novel framework for characterizing genetic architecture through a new integral genomic parameter. We propose the concept of a multidimensional Gene Space to enable holistic quantification of genome organization principles. Gene Space—a multidimensional space based on the frequencies of nucleotide tokens, such as individual nucleotides, codons, or codon pairs. We demonstrate that in this space, genes from each of the studied microorganism species occupy a limited region, and individual genes from different species can be effectively separated with more than 95% accuracy. Consequently, a specific Genome Subspace can be defined for each species, which constrains the organism's evolutionary pathways, thereby determining the constraints on gene optimization for these species. Further in-depth analysis is required to test if it is true for other organisms as well. The Gene Space framework offers a novel and powerful approach for genome analysis at the most basic levels, with promising applications in comparative genomics, evolutionary biology, and gene optimization.

Keywords: gene space; genome subspace; codon bias; gene optimization

Academic Editor:
Candice Brinkmeyer-Langford

Received: 28 October 2025
Revised: 4 December 2025
Accepted: 5 December 2025
Published: 10 December 2025

Citation: Zaytsev, K.; Bogatyreva, N.; Fedorov, A. Multidimensional Gene Space as an Approach for Analyzing the Organization of Genomes. *Int. J. Mol. Sci.* **2025**, *26*, 11926. https://doi.org/10.3390/ijms262411926

1. Introduction

The genetic code is a universal apparatus essential to all living organisms. Each codon determines which amino acid will be incorporated into the synthesized protein. With 61 codons encoding 20 distinct amino acids [1], the genetic code is redundant: the same amino acid can be encoded by one to six synonymous codons [2]. Consequently, numerous nucleotide sequences can encode identical amino acid sequences.

Despite this diversity, organisms have preferences for specific nucleotide combinations. For example, genes from different organisms encoding homologous proteins have significant differences in codon usage bias: in every organism, certain codons are used more frequently than others, forming species-specific codon usage profiles [3,4]. A key driver for these differences is the GC content [5,6], which describes the difference in usage of A + T and G + C nucleotides [7].

K-mer based genome analysis methods count occurrences of short DNA substrings (k-mers) to enable alignment-free comparisons, taxonomy, phylogenomics, and trait association without the reference genomes [8,9]. This approach excels at handling large-scale genomic data by revealing diversity, evolutionary patterns, and functional elements through k-mer abundance profiles [10].

Physical parameters of the DNA double strands, such as melting enthalpy, directly reflect functional information encoded in the primary nucleotide sequence [11]. The signal theory approach detects novel DNA similarities by comparing profiles of these physical properties [12,13].

During protein synthesis, each codon is recognized by a specific tRNA. Thus, codon frequencies in the genome correlate with the abundance of corresponding tRNAs [14,15]. Codon bias directly impacts translation efficiency: overuse of a codon can deplete its cognate tRNA, reducing overall protein synthesis efficiency and affecting organismal viability [16].

In recombinant gene expression, the primary goal is maximizing target protein yield. However, disrupting cellular process balance may induce cell stress [17]. Therefore, a common gene optimization strategy involves adjusting codon frequencies to match host organism preferences. Crucially, not only individual codon frequencies but also their combinations (correlations between adjacent codons) influence protein synthesis efficiency [18].

Optimization typically is based on genome-wide averaged codon bias. Yet codon bias varies significantly not only among genes but also within different regions of the same gene [4,19]. Notably, genes with varying expression levels exhibit distinct codon bias differences [4,20]. Our previous research further demonstrates that codon frequency optimization characterizes all organismal genes—not just highly expressed ones [21]. Thus, even low-expression genes possess unique codon usage profiles differing from highly expressed genes.

In this study, we propose a novel approach to investigating codon bias ranges through the concept of Genome Subspaces, which defines the range of sequence variations across the certain organism. Examining this approach with four distinct unicellular species have brought captivating observations of separate gene clusters occurring for each species. In addition, we propose that our organism-specific Genome Subspace analysis, owing to its simplicity, will provide new insights into the origin and organization of primitive living organisms and life itself.

2. Results

2.1. Gene Space Organization

Any nucleotide sequence can be split into a series of short subsequences—such as individual nucleotides, codons, codon pairs, etc., also called k-mers in the literature [22–24]; however, for the sake of simplicity and clarity for the common audience we propose the term token. Consequently, any gene can be characterized by the frequencies of its tokens.

In this study, we introduce the concept of a multidimensional Gene Space capable of describing any possible set of genes. Within this space, a gene is represented as an n-dimensional vector, where n denotes the total number of possible tokens (e.g., 4 for nucleotides, 61 for codons, 3721 for codon pairs). For this work, we exclusively use amino acid-encoding codons; thus, any combination containing a stop codon is excluded. Gene Space is based on n orthogonal axes, each corresponding to the frequency of a specific token, with values ranging from 0 to 1, representing the frequency of that token in a given nucleotide sequence normalized to the number of tokens in the sequence. A visual example for the nucleotide-based Gene Space is presented in Figure 1.

Since token frequencies are normalized (summing up to 1), the Gene Space constitutes an (n–1)-dimensional simplex containing n vertices. Each vertex lies at a point where one token's frequency equals 1 and all others are 0. Thus, Gene Space is a bounded region within the n-dimensional space, constrained by the interval from 0 to 1 along each axis.

Every nucleotide sequence maps to a specific point in Gene Space, yet multiple sequences can occupy identical positions if they share token frequencies. For example:

sequences ATGGCA and ATGATGGCAGCA coincide in nucleotide-based and codon-based Gene Spaces (identical codon frequencies) but diverge in codon-pair-based space.

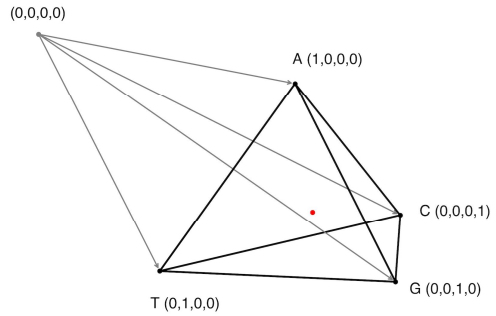

Figure 1. Grey point with coordinates (0, 0, 0, 0) shows the origin and grey arrows represent four orthogonal axes for the four nucleotides (A, T, G, C). Black dots show the points equal to 1 on each of the axes. The wireframe pyramid shows the Gene Space for individual nucleotides. The red dot with coordinates (0.25, 0.25, 0.25, 0.25) shows the geometric center of the pyramid, representing a sequence with equiprobable nucleotides.

Conversely, ATGGCAATG and ATGGCAATGGCAATG coincide in codon-pair-based space but differ in nucleotide/codon spaces.

To illustrate Gene Space organization, we use four arbitrarily selected unicellular species belonging to evolutionary distant groups (see Table 2) with comparable genome sizes. Among three bacteria, one, *E. coli*, belongs to the kingdom Psuedomonadati, while the two others, *R. castenholzii* and *S. tropica*, shares the same Bacillati kingdom, but belong to different phyla, Chloroflexota and Actinomycetota, respectively. *S. cerevisiae* is eukaryotic fungi. Full phylogeny for the selected organisms is presented in the Supplementary File S1.

Let us consider the simplest nucleotide Gene Space, which can be visualized as a 3-dimensional simplex (tetrahedron) (Figure 2). Subsequent analyses use higher-dimensional codon and codon-pair Gene Spaces for increased precision, though nucleotide Gene Space remains optimal for intuitive visualization.

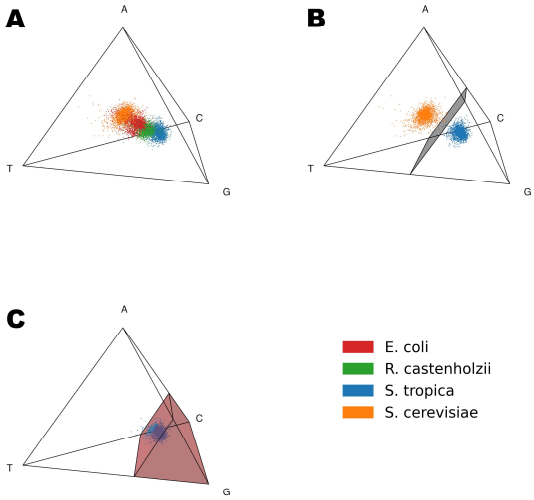

Figure 2. (**A**) A visualization of the 3-dimensional space based on the frequencies of nucleotides A, T, G, and C (k = 1). Each point corresponds to a single gene from one of the organisms: orange points represent *Saccharomyces cerevisiae*, red points for *Escherichia coli*, green points for *Roseiflexus castenholzii*, and blue points for *Salinispora tropica*. (**B**) Hyperplane (grey polygon) separation of genes from different organisms in the 3-dimensional Gene Space based on nucleotide frequencies for *S. cerevisiae* (orange points) and *S. tropica* (blue points) organisms. (**C**) *S. tropica* Genome Subspace (red area) and isolation of its genes by the subspace boundary inside the nucleotide—based Gene Space.

Figure 2A demonstrates the 3D Gene Space based on the nucleotide frequencies (A, T, G, C). Each point represents a single gene from one of the four studied organisms: *Saccharomyces cerevisiae* (orange), *Escherichia coli* (red), *Roseiflexus castenholzii* (green), *Salinispora tropica* (blue). Even inside the simplest nucleotide-based Gene Space, genes form distinct organism-specific clusters.

2.2. Separation of Genes from Different Organisms

From Figure 2A, it is evident that genes from different organisms form distinct clusters and can be separated even at the nucleotide level. In order to do so we used the Support Vector Machine (SVM) method to construct a decision boundary separating a pair of organisms. Figure 2B shows the decision plane built using the SVM to separate *S. cerevisiae* and *S. tropica* genes inside the nucleotide Gene Space. While there is a clear dependency between positions of the genes inside the nucleotide Gene Space and GC content of the organism, there is a lot of variation between the positions of the individual genes, meaning that it is not the only factor influencing the positioning of genes inside the Gene Space. There are also lots of codon combinations that could achieve a similar GC content, which can only be identified by models with longer tokens.

We applied the Support Vector Machine (SVM) method to separate each pair of the test organisms in the nucleotide, codon and codon pair Gene Spaces (Table 1). As shown in Table 1, the increase in Gene Space dimensionality improves the accuracy of gene classification for each pair of organisms. Specifically, in the codon-pair space, the probability of the correct gene to organism matching exceeds 97% for all the organism pairs. Thus, longer tokens that form higher-dimensional Gene Spaces enable more precise compartmentalization of individual organisms. Based on these observations, we hypothesize that each organism occupies a limited region within the Gene Space.

Table 1. The probability of correct identification of genes of an organism from a test set of genes when separating two organisms using SVM at three different tokenization levels.

Organism	Token	*E. coli*	*R. castenholzii*	*S. tropica*	*S. cerevisiae*
E. coli	nucleotides	-	0.903	0.989	0.893
	codons	-	0.968	0.990	0.974
	codon pairs	-	0.982	0.996	0.978
R. castenholzii	nucleotides	0.903	-	0.910	0.979
	codons	0.968	-	0.981	0.985
	codon pairs	0.982	-	0.989	0.991
S. tropica	nucleotides	0.989	0.910	-	0.998
	codons	0.990	0.981	-	0.998
	codon pairs	0.996	0.989	-	0.998
S. cerevisiae	nucleotides	0.893	0.979	0.998	-
	codons	0.974	0.985	0.998	-
	codon pairs	0.978	0.991	0.998	-

Thus, we have demonstrated that the proposed Gene Space allows for effective and accurate separation of genomes.

2.3. Genome Subspace Identification

Based on this, we hypothesize that the genes of each organism occupy a specific bounded region in the Gene Space, which we will refer to as the Genome (organism's) Subspace. When developing a method for identification of such subspaces within the Gene

Space, there are two main constraints: the majority of the organism's genes should fall within its corresponding subspace and the size of the allocated Genome Subspace should be minimized around the actual genes.

Unlike the binary classification problem of separating genes from two organisms, Genome Subspace identification lacks a negative dataset, thus requiring a different approach. For this task, we employed a one-class Support Vector Machine (one-class SVM), trained to identify a certain percentage (we used 90%) of genes from the reference set. As a result, the one-class SVM produced a boundary of the subspace represented by a hyperplane dividing the Gene Space into two parts, one of which constitutes the organism's subspace.

Figure 2C illustrates the Genome Subspace boundary for *S. tropica* in the nucleotide Gene Space: red region represents the subspace, dark blue points correspond to genes inside the subspace, and light blue points indicate the 10% of *S. tropica* genes excluded from the subspace.

Thus, we can delineate the boundary and identify the Genome Subspace of a specific organism within the Gene Space. It is important to note that when applying machine learning methods, especially for high-dimensional data, the issue of overfitting inevitably arises. To assess overfitting, we calculated overfitting ratios, defined as the ratio between the model's loss on disjoint training and validation datasets. An overfitting ratio equal to one indicates that the validation loss matches the training loss, while higher values signify more pronounced overfitting.

The results of this test are presented in Table 2. As observed, for both nucleotide- and codon-based spaces, the overfitting ratios are close to 1, indicating virtually no overfitting. For codon-pair spaces, the values approach 1.3, meaning that compared to the 10% of genes excluded from the Genome Subspaces in the training set, approximately 13% of genes fall outside of the Genome Subspaces in the validation set. However, this level of overfitting remains manageable and should not pose significant issues.

Table 2. Overfitting ratios for the Genome Subspace models trained on four different organisms. 90% of genes from each organism are inside the subspace.

Trained On	Family	Number of Genes	Nucleotide	Codon	Codon Pair
E. coli	Prokaryote; Bacteria; Pseudomonadati	4852	1.002	1.037	1.335
R. castenholzii	Prokaryote; Bacteria; Bacillati	4855	1.023	1.014	1.285
S. tropica	Prokaryote; Bacteria; Bacillati	4666	1.006	1.024	1.141
S. cerevisiae	Eukaryota	6600	1.009	1.017	1.370

It is also worth noting that the datasets were split in half for this test, implying that the number of genes in the training sets was smaller than the number of tokens in the codon-pair model. Therefore, training on the full gene set could reduce the model's level of overfitting, supporting our recommendation to use both codon-based Gene Space and codon-pair-based Gene Space.

2.4. Characterization of the Genome Subspaces

The Genome Subspace for each organism represents a specific and unique mode of genome optimization. Identification and comparison of the Genome Subspaces for different organisms can be highly informative. The most straightforward metric for a subspace

would be its volume; however, calculating volume in high-dimensional spaces such as those based on codons and codon pairs is computationally challenging due to extreme complexity. Therefore, we propose an estimate for the volume of an organism's Genome Subspace from the number of the Gene Space vertices n_a contained inside of it. These estimates would depend on the total Gene Space size N, which equals the number of tokens used to construct it and the variability between genes of a specific organism. The resulting volume estimates for the Genome Subspaces are presented in Table 3.

Table 3. Estimated volumes of the Genome Subspaces of the four studied organisms in the Gene Space.

	The Total Size of the Gene Space	***E. coli***	***R. castenholzii***	***S. tropica***	***S. cerevisiae***
Nucleotides	4	1	2	2	2
Codons	61	20	18	15	20
Codon Pairs	3721	778	758	375	992

Despite the fact that Table 3 shows that each organism's Genome Subspace occupies a substantial portion of the entire Gene Space, genes distribution within these subspaces is highly uneven. Since we conceptualize the Gene Space as a simplex—where each vertex corresponds to the sequences entirely composed of a single token (whether it be a nucleotide, codon, or codon pair)—the geometric center of the simplex represents a sequence with all tokens present in equal proportions. Therefore, the position of each gene can be interpreted as the distance from this center point, with the maximum possible distance being the distance to any vertex, which is approximately equal to 1. As the vertices represent sequences composed entirely of a single token, this distance serves as a measure of inequality in token usage within the gene. Figure 3 illustrates the distributions of distances from each gene of the studied organisms to the center of Gene Space for both codon-based and codon-pair-based Gene Spaces.

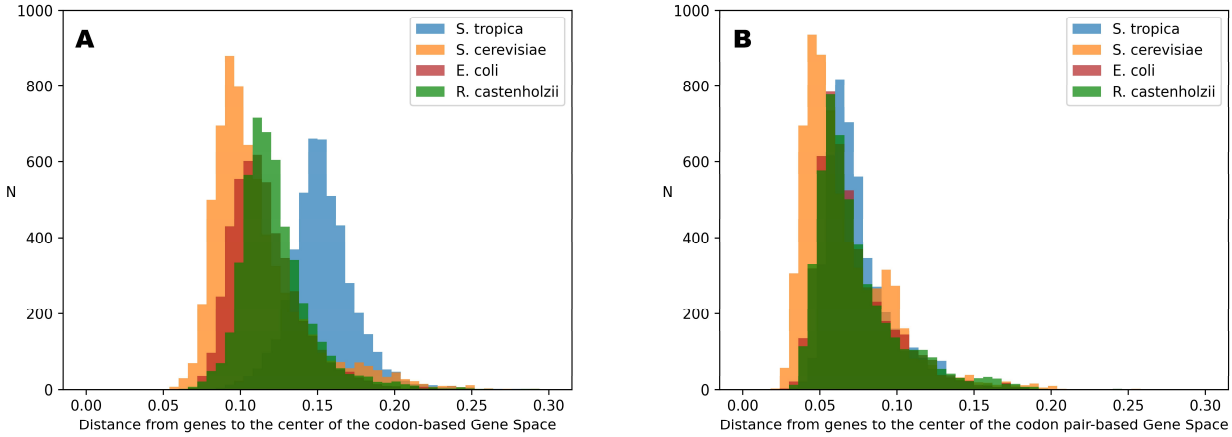

Figure 3. Distributions of distances from genes to the centers of the codon (**A**) and codon pair (**B**) Gene Spaces for the four studied organisms.

Figure 3 shows that, despite the minor differences, all distributions occupy the same range of distances in the codon-pair space. Moreover, all genes are clustered closer to the center point (0 on the distance scale) than to the vertices, with the most distant gene (i.e., the gene with the most non-random codon-pair composition) located at approximately 20% of the maximum possible distance (equals 1 for the Gene Space vertices). For codon-based spaces, there are greater differences between the distributions for different organisms; however, they still occupy about 20% of the possible distance range (from 5% to 25% from the center point).

This means that although Gene Space is quite sparse and much of it is empty, the gene-occupied region extends at most 25% away from the center point. Nevertheless, a relatively simple machine learning approach can successfully identify token variants specific to particular organisms.

While volume estimates for the Genome Subspace work well for comparisons between different organisms, they should not be interpreted as the proportions of the entire Gene Space.

The differences observed in the distributions shown in Figure 3 can be explained by variations in genome optimization, differences in amino acid composition, and gene length. These factors influence the minimal possible distance from a gene point to the center of the Gene Space.

2.5. Commonalities Between the Subspaces for the Four Studied Organisms

Let us examine the commonalities and differences among all the organisms studied. Figure 4 shows the intersections of the simplex vertices belonging to the Genome Subspace of each of the four studied organisms. It can be seen that 39 vertices in the codon-pair Gene Space are common to all four organisms, while only one vertex in the codon space is shared by all of them (corresponding to glutamic acid). Among the bacteria, seven codons are shared, corresponding to such amino acids as alanine, glycine, isoleucine, leucine, valine, and glutamine.

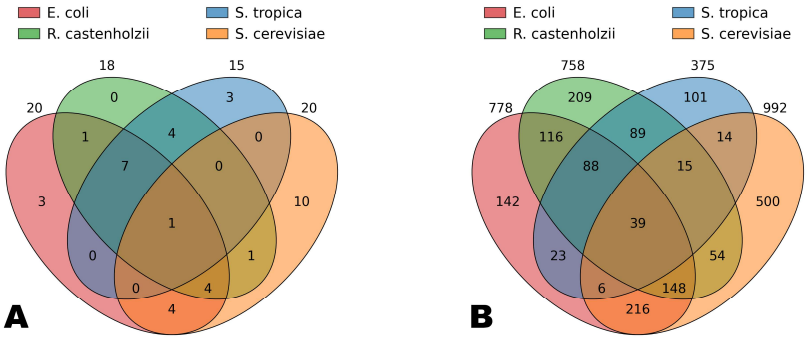

Figure 4. Venn diagrams for the volumes of the Genome Subspace intersections for the four studied organisms in the codon (**A**) and codon pair (**B**) Gene Spaces.

It is important to note that the simplex vertices included in the Genome Subspaces are arranged such that the diversity of amino acids within a Genome Subspace is consistently proportionally higher than the diversity of codons. For example, in the codon Gene Space, each organism's subspace includes vertices corresponding to more than 10 amino acids, accounting for over half of the total amino acid diversity. In contrast, these codons represent less than one-third of the total codon diversity. This suggests that Genome Subspaces tend to encompass the maximal or necessary number of amino acids while preserving the unique codon composition of each organism.

Notably, cysteine and tryptophan vertices are not included in any of the studied subspaces, and histidine, serine, and tyrosine are absent from the bacterial Genome Subspaces. Meanwhile, amino acids such as alanine, aspartic acid, glutamic acid, isoleucine, and leucine are present in the Genome Subspaces of all four organisms, often each encoded by their organism-specific codons.

Furthermore, the Venn diagrams presented in Figure 4 allow conclusions to be drawn about the evolutionary relationships among the studied organisms.

The distribution of distances from each gene to the center of Gene Space (see Figure 3) reveals non-uniform density and suggests genomic organization into primary and secondary genes within this spatial context. To quantify the subspace filling uniformity, we

artificially constrained Genome Subspace volume so that it would fit only a fraction of the organism's genes, and estimated the subspace volumes. Results (Figure 5) demonstrate a nonlinear dependence, confirming the unevenness of the gene distributions. Vertex analysis across ν values showed an invariant composition in >95% of cases, validating Genome Subspaces as a robust, organism-specific genomic characteristic.

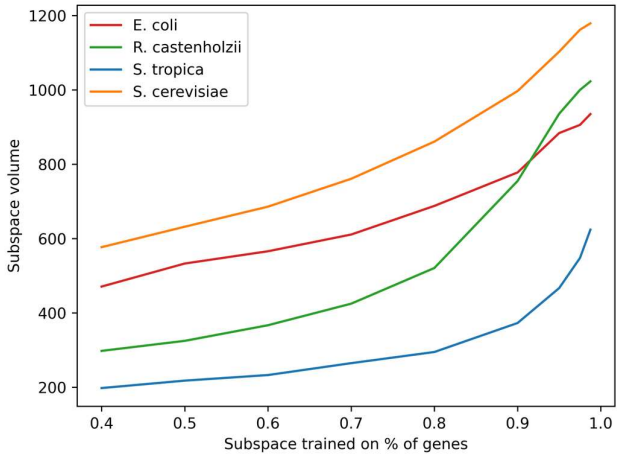

Figure 5. Dependence of the volume of Genome Subspaces of organisms on the percentage of genes of this organism falling into the distribution core.

2.6. Python Module and Source Code

We created a Python module for calculation of the Genome Subspaces from a set of nucleotide sequences and analysis of such subspaces. Package is available for download from PyPI as "genespace". Source code for the module is available on Github at https://github.com/conzaytsev/GeneSpace (accessed on 27 October 2025).

3. Discussion

The fundamental law of life dictates that all genes of an organism are encoded using the same universal principle: each amino acid can be specified by 1 to 6 synonymous codons. As a result, each amino acid sequence can be represented by an astronomically large number of synonymous nucleotide sequences. But despite all the diversity, organisms select certain configurations that are unique for each species due to the different optimization paths [25,26]. This phenomenon is well known and researched and the most popular index that describes it is the codon bias [3,27], which shows codon preferences of the organism.

In this study, we introduced the concept of a multidimensional Gene Space which is based on the normalized frequencies of the nucleotide tokens—ranging from single nucleotides to codons and codon pairs—and demonstrated its utility for characterizing and distinguishing genes across different organisms. Our results show that genes from each species occupy a distinct, bounded region within the Gene Space, which we defined as an organism-specific confined Genome Subspace. This finding supports the hypothesis that each organism's genome is optimized as a whole in a unique manner, being reflected in the specific distribution of token frequencies that shape its subspace. Codon-based Genome Subspaces practically reflect the codon bias range, showing not only the codon bias preferred by the organism, but also the possible variations in codon usage across the genes.

The application of machine learning methods, particularly the One-class Support Vector Machine (OcSVM) [28], allowed us to effectively delineate these subspaces especially in the high-dimensional codon and codon-pair Gene Spaces. The classification accuracy improved with the increase in dimensionality, but the method becomes extremely com-

putationally complex for tokens longer than six nucleotides due to the computational complexity. Importantly, overfitting tests revealed that our models generalize well, with overfitting ratios close to one for nucleotide and codon Gene Spaces, and still manageable levels for the codon-pair Gene Spaces. This robustness justifies the use of both codon and codon-pair Gene Spaces for genome analysis.

While longer tokens allow for more effective genome separation, there is a practical limit to the length. With longer tokens there are more possible token variations and so, the Gene Space contains even more dimensions. For example, for 9 nucleotide long tokens the Gene Space would be 226,980-dimensional, significantly exceeding the number of genes in any living organism. And if a Genome Subspace model was to be trained in such Gene Space, it would suffer from extreme overfitting. Thus, we limited ourselves to codon and codon pair-based models.

Volume determination for these subspaces, while challenging due to the computational complexity of high-dimensional spaces, was approached as an estimate from the number of simplex vertices included in each organism's Genome Subspace. This proxy measure revealed that subspaces cover substantial portions of the overall Gene Space but are unevenly filled by genes. Analysis of distances from the center of the Gene Space—a point representing equal token usage—indicated that all genes cluster closer to the center than to vertices, suggesting that genomes maintain a balance between variation and specificity in token usage. The maximum observed distance was about 20–25% of the total possible distance, reflecting constraints imposed by the amino acid composition and gene length. The furthest points from the Gene Space center are the vertices that represent sequences consisting of a single token. Therefore, the distance from a gene to the Gene Space center shows the degree of heterogeneity in the token distribution, linked to the optimization of the organism's genes for its life conditions [29,30].

Comparative analysis of the Genome Subspace intersections highlighted both shared and unique features among the organisms. For example, while 39 vertices were common to all four studied species in the codon-pair Gene Space, only one vertex was shared by all four studied organisms in the codon Gene Space. Bacterial subspaces shared codons corresponding to a core set of amino acids, indicating evolutionary conservation, whereas other amino acids were organism-specific. This pattern suggests that subspaces strive to encompass a broad and necessary set of amino acids while preserving distinct codon usage profiles, reflecting both functional requirements and evolutionary history.

Finally, investigation of gene distribution density within the Genome Subspaces revealed a nonlinear relationship between the fraction of genes included and the occupied volume, confirming that gene distribution is heterogeneous rather than uniform. This heterogeneity likely reflects some biological factors such as gene expression levels, functional clustering, and evolutionary pressures.

Overall, the Gene Space framework and associated machine learning methods provide a novel and powerful approach for understanding genome organization, optimization, and evolution. By capturing the complex interplay of token usage patterns, gene length, and amino acid composition, this approach offers new opportunities for comparative genomics and synthetic biology, including improved gene optimization strategies tailored to the specific organisms like genetically modified crops, allowing the adaptation of genes from different sources to the species of interest [31,32].

4. Materials and Methods

4.1. Dataset

In this study we analyzed coding sequence regions from 4 different species: *Escherichia coli* k12, *Roseiflexus castenholzii* ASM1780v1, *Salinispora tropica* ASM1642v1 and

Saccharomyces cerevisiae R64-1-1, which were obtained from Ensembl database [https://bacteria.ensembl.org/Escherichia_coli_k_12_gca_004802935/Info/Index (accessed on 14 November 2024), https://bacteria.ensembl.org/Roseiflexus_castenholzii_dsm_13941_gca_000017805/Info/Index (accessed on 17 March 2025), https://bacteria.ensembl.org/Salinispora_tropica_cnb_440_gca_000016425/Info/Index (accessed on 16 January 2025), https://fungi.ensembl.org/Saccharomyces_cerevisiae/Info/Index (accessed on 12 December 2024)].

4.2. Gene Space

We analyze gene sequences from the point of token frequencies. Here tokens are short subsequences of a certain fixed length, for example, individual nucleotides, codons or codon pairs. To construct the Gene Space, we created n orthogonal axes, each corresponding to the specific token, where n is the total number of possible tokens of such length. As we only analyze amino acid-coding codons, n = 4 for nucleotides; n = 61 for codons; n = 3721 for codon pairs. Each axis spans from 0 (in the origin) to 1. For each gene we calculated the number of token occurrences and divided them by the total number of tokens in the gene, thus receiving the token frequencies of the gene. Each gene can be represented as a point, with n coordinates (token frequencies) describing its location on each of the axes. Due to the normalization of the codon frequencies, the gene position range represents an n-1 dimensional simplex with n vertices located at the points where a single axis value equals 1. This range was denoted as the Gene Space. Any nucleotide sequence can be represented as a point in the Gene Space with the coordinates described by its token frequencies.

4.3. Separation of Genes from Two Species

In order to separate clusters of genes from two different species we used the Support Vector Machine (SVM) [33] method. It creates an n-2 dimensional hyperplane that splits the Gene Space into two parts. When training the SVM, we utilized a linear kernel so that the boundary would become a hyperplane.

In order to examine the efficiency of species separation, we split each organism's gene dataset into two disjoint equal size sets. One of the sets from both organisms was used for model training and the other one was used to calculate the probability of the correct identification of the genes. Then we swapped the sets and repeated the process (trained the model on the second sets and validated on the first ones). The resulting probability value is a mean of the two results.

4.4. Genome Subspace

Genome Subspace is an area of the Gene Space that contains the majority of genes from that organism. In order to define the subspace, we utilized the One Class SVM [28] method as it does not require a negative dataset, which is absent in the unary classification task. The model was trained using a linear kernel and the upper bound for the number of training errors was set to 0.1, so the subspace would contain at least 90% of the genes from the training set. The result is a hyperplane that splits the Gene Space into two parts. One of them contains the majority of genes and is the organism's Genome Subspace.

4.5. Overfitting Ratio

In order to control the level of overfitting for the used models, we randomly split the dataset into two separate sets with the same number of genes in each one. The first set was used for the Genome Subspace model training, and the second was used for testing the prediction accuracy. When training the One Class SVM model, some genes are left outside

the subspace, we refer to them as the Training loss. Genes from the testing set that are left outside the subspace are referred to as the Validation loss.

To eliminate any discrepancies due to the specific split of the dataset, the model then was trained on the second set, and tested on the first. The resulting overfitting ratio is calculated as an average of the two measurements.

4.6. Distance Between Points in the Gene Space

If we know coordinates of the two points in the Gene Space, then the distance between them can be calculated as the Euclidean distance.

The largest possible distance inside the Gene Space is between any two vertices and is equal to $\sqrt{2} \approx 1.41$. The center point of the Gene Space is located in the geometric center of the simplex and corresponds to the equal distribution of tokens in the nucleotide sequence. The furthest points of the Gene Space from the center point are simplex vertices, and so the maximum distance is $\sqrt{\frac{n}{n+1}}$, where n is the number of vertices in the Gene Space. Therefore, for codons (n = 61) and codon pairs (n = 3721) this distance can be approximated as 1.

4.7. Subspace Volume

Genome Subspace boundary is a hyperplane that divides the Gene Space into two parts. It is located in the relative neighborhood of the center of the Gene Space. The Gene Space is highly dimensional (60 dimensions for codons and 3720 for codon pairs), therefore calculating the volume of the cut off part of the space is highly problematic due to the computational complexity of the process. The most optimal way is to estimate the volume of the cut off part from the number of vertices that fit inside the subspace compared to the total number of simplex vertices. There are more accurate estimates, but due to the unevenness of the gene distribution inside the Gene Space, they do not make much sense for comparisons between different organisms.

5. Conclusions

We developed a new framework for the analysis of nucleotide sequences through a geometric representation in a multidimensional Gene Space. This concept allows to quantify differences between nucleotide sequences, which enables accurate classification of genomes. Each genome occupies only a minor part of the Gene Space, and genes from different organisms can be effectively separated with more than 95% accuracy. Such areas, which we called Genome Subspaces, represent the patterns occurring in genes of the specific organisms. When used with the codon-based Gene Space, these subspaces represent the available codon bias range of an organism, showing the limits to the codon frequency variations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms262411926/s1.

Author Contributions: Conceptualization, K.Z. and A.F.; methodology, K.Z. and N.B.; software, K.Z.; validation, K.Z., N.B. and A.F.; formal analysis, K.Z. and N.B.; investigation, K.Z.; resources, A.F.; data curation, K.Z. and N.B.; writing—original draft preparation, K.Z. and N.B.; writing—review and editing, K.Z., N.B. and A.F.; visualization, K.Z. and N.B.; supervision, A.F.; project administration, A.F.; funding acquisition, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Ministry of Science and Higher Education of the Russian Federation (Federal Scientific and Technical Program for the Development of Genetic Technologies for 2019–2030, agreement №075-15-2025-470 dated 29 May 2025).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CDS sequences for the studied organisms were derived from Ensembl database: [https://bacteria.ensembl.org/Escherichia_coli_k_12_gca_004802935/Info/Index (accessed on 14 November 2024), https://bacteria.ensembl.org/Roseiflexus_castenholzii_dsm_13941_gca_000017805/Info/Index (accessed on 17 March 2025), https://bacteria.ensembl.org/Salinispora_tropica_cnb_440_gca_000016425/Info/Index (accessed on 16 January 2025), https://fungi.ensembl.org/Saccharomyces_cerevisiae/Info/Index (accessed on 12 December 2024)]. Python module for the Gene Space analysis is available in PyPI at https://pypi.org/project/genespace/. Source code for the module is available on Github at https://github.com/conzaytsev/GeneSpace.

Acknowledgments: We are grateful to Maria Yurkova and Dmitrii Kostenko for their seminar discussion and support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Crick, F.H.; Barnett, L.; Brenner, S.; Watts-Tobin, R.J. General Nature of the Genetic Code for Proteins. *Nature* **1961**, *192*, 1227–1232. [CrossRef] [PubMed]
2. Söll, D.; Ohtsuka, E.; Jones, D.S.; Lohrmann, R.; Hayatsu, H.; Nishimura, S.; Khorana, H.G. Studies on Polynucleotides, XLIX. Stimulation of the Binding of Aminoacyl-sRNA's to Ribosomes by Ribotrinucleotides and a Survey of Codon Assignments for 20 Amino Acids. *Proc. Natl. Acad. Sci. USA* **1965**, *54*, 1378–1385. [CrossRef] [PubMed]
3. Grantham, R.; Gautier, C.; Gouy, M.; Mercier, R.; Pavé, A. Codon Catalog Usage and the Genome Hypothesis. *Nucleic Acids Res.* **1980**, *8*, 197. [CrossRef] [PubMed]
4. Plotkin, J.B.; Kudla, G. Synonymous but Not the Same: The Causes and Consequences of Codon Bias. *Nat. Rev. Genet.* **2011**, *12*, 32–42. [CrossRef]
5. Lynn, D.J. Synonymous Codon Usage Is Subject to Selection in Thermophilic Bacteria. *Nucleic Acids Res.* **2002**, *30*, 4272–4277. [CrossRef]
6. Chen, S.L.; Lee, W.; Hottes, A.K.; Shapiro, L.; McAdams, H.H. Codon Usage between Genomes Is Constrained by Genome-Wide Mutational Processes. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 3480–3485. [CrossRef]
7. Muto, A.; Osawa, S. The Guanine and Cytosine Content of Genomic DNA and Bacterial Evolution. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 166–169. [CrossRef]
8. Sievers, A.; Bosiek, K.; Bisch, M.; Dreessen, C.; Riedel, J.; Froß, P.; Hausmann, M.; Hildenbrand, G. K-Mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes* **2017**, *8*, 122. [CrossRef]
9. Bussi, Y.; Kapon, R.; Reich, Z. Large-Scale k-Mer-Based Analysis of the Informational Properties of Genomes, Comparative Genomics and Taxonomy. *PLoS ONE* **2021**, *16*, e0258693. [CrossRef]
10. Pornputtapong, N.; Acheampong, D.A.; Patumcharoenpol, P.; Jenjaroenpun, P.; Wongsurawat, T.; Jun, S.-R.; Yongkiettrakul, S.; Chokesajjawatee, N.; Nookaew, I. KITSUNE: A Tool for Identifying Empirically Optimal K-Mer Length for Alignment-Free Phylogenomic Analysis. *Front. Bioeng. Biotechnol.* **2020**, *8*, 556413. [CrossRef]
11. Deyneko, I.V.; Kalybaeva, Y.M.; Kel, A.E.; Blöcker, H. Human–Chimpanzee Promoter Comparisons: Property-Conserved Evolution? *Genomics* **2010**, *96*, 129–133. [CrossRef] [PubMed]
12. Deyneko, I.V.; Kel, A.E.; Bloecker, H.; Kauer, G. Signal-Theoretical DNA Similarity Measure Revealing Unexpected Similarities of *E. Coli* Promoters. *Silico Biol.* **2005**, *5*, 547–555. [CrossRef]
13. Deyneko, I.V.; Bredohl, B.; Wesely, D.; Kalybaeva, Y.M.; Kel, A.E.; Blocker, H.; Kauer, G. FeatureScan: Revealing Property-Dependent Similarity of Nucleotide Sequences. *Nucleic Acids Res.* **2006**, *34*, W591–W595. [CrossRef] [PubMed]
14. Ikemura, T. Correlation between the Abundance of *Escherichia coli* Transfer RNAs and the Occurrence of the Respective Codons in Its Protein Genes. *J. Mol. Biol.* **1981**, *146*, 1–21. [CrossRef]
15. Ikemura, T. Correlation between the Abundance of *Escherichia coli* Transfer RNAs and the Occurrence of the Respective Codons in Its Protein Genes: A Proposal for a Synonymous Codon Choice That Is Optimal for the E. Coli Translational System. *J. Mol. Biol.* **1981**, *151*, 389–409. [CrossRef]
16. Frumkin, I.; Lajoie, M.J.; Gregg, C.J.; Hornung, G.; Church, G.M.; Pilpel, Y. Codon Usage of Highly Expressed Genes Affects Proteome-Wide Translation Efficiency. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E4940–E4949. [CrossRef]
17. Hoffmann, F.; Rinas, U. Stress Induced by Recombinant Protein Production in *Escherichia coli*. In *Physiological Stress Responses in Bioprocesses*; Advances in Biochemical Engineering/Biotechnology; Springer: Berlin/Heidelberg, Germany, 2004; Volume 89, pp. 73–92. ISBN 978-3-540-20311-7.

18. Huang, Y.; Lin, T.; Lu, L.; Cai, F.; Lin, J.; Jiang, Y.; Lin, Y. Codon Pair Optimization (CPO): A Software Tool for Synthetic Gene Design Based on Codon Pair Bias to Improve the Expression of Recombinant Proteins in Pichia Pastoris. *Microb. Cell Fact.* **2021**, *20*, 209. [CrossRef]
19. Powell, J.R.; Moriyama, E.N. Evolution of Codon Usage Bias in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 7784–7790. [CrossRef]
20. Liu, Y.; Yang, Q.; Zhao, F. Synonymous but Not Silent: The Codon Usage Code for Gene Expression and Protein Folding. *Annu. Rev. Biochem.* **2021**, *90*, 375–401. [CrossRef]
21. Zaytsev, K.; Bogatyreva, N.; Fedorov, A. Link Between Individual Codon Frequencies and Protein Expression: Going Beyond Codon Adaptation Index. *Int. J. Mol. Sci.* **2024**, *25*, 11622. [CrossRef]
22. Moeckel, C.; Mareboina, M.; Konnaris, M.A.; Chan, C.S.Y.; Mouratidis, I.; Montgomery, A.; Chantzi, N.; Pavlopoulos, G.A.; Georgakopoulos-Soares, I. A Survey of K-Mer Methods and Applications in Bioinformatics. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2289–2303. [CrossRef]
23. Smith, R.P.; Riesenfeld, S.J.; Holloway, A.K.; Li, Q.; Murphy, K.K.; Feliciano, N.M.; Orecchia, L.; Oksenberg, N.; Pollard, K.S.; Ahituv, N. A Compact, in Vivo Screen of All 6-Mers Reveals Drivers of Tissue-Specific Expression and Guides Synthetic Regulatory Element Design. *Genome Biol.* **2013**, *14*, R72. [CrossRef]
24. Ghandi, M.; Lee, D.; Mohammad-Noori, M.; Beer, M.A. Enhanced Regulatory Sequence Prediction Using Gapped K-Mer Features. *PLoS Comput. Biol.* **2014**, *10*, e1003711. [CrossRef] [PubMed]. Erratum in *PLoS Comput. Biol.* **2014**, *10*, e1004035. [CrossRef]
25. Tuller, T.; Waldman, Y.Y.; Kupiec, M.; Ruppin, E. Translation Efficiency Is Determined by Both Codon Bias and Folding Energy. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 3645–3650. [CrossRef] [PubMed]
26. Quax, T.E.F.; Claassens, N.J.; Söll, D.; van der Oost, J. Codon Bias as a Means to Fine-Tune Gene Expression. *Mol. Cell* **2015**, *59*, 149–161. [CrossRef] [PubMed]
27. Parvathy, S.T.; Udayasuriyan, V.; Bhadana, V. Codon Usage Bias. *Mol. Biol. Rep.* **2022**, *49*, 539–565. [CrossRef]
28. Schölkopf, B.; Williamson, R.C.; Smola, A.; Shawe-Taylor, J.; Platt, J. Support Vector Method for Novelty Detection. In *Proceedings of the Advances in Neural Information Processing Systems*, Denver, CO, USA, 29 November–4 December 1999; Solla, S., Leen, T., Müller, K., Eds.; MIT Press: Cambridge, MA, USA, 1999; Volume 12.
29. Korotkov, E.; Zaytsev, K.; Fedorov, A. Use of 6 Nucleotide Length Words to Study the Complexity of Gene Sequences from Different Organisms. *Entropy* **2022**, *24*, 632. [CrossRef]
30. Botzman, M.; Margalit, H. Variation in Global Codon Usage Bias among Prokaryotic Organisms Is Associated with Their Lifestyles. *Genome Biol.* **2011**, *12*, R109. [CrossRef]
31. Jones, J. Why Crop Plants Need Recombinant DNA for Increased Disease Resistance. *Nat. Biotechnol.* **1999**, *17*, 19. [CrossRef]
32. Khan, S.; Ullah, M.W.; Siddique, R.; Nabi, G.; Manan, S.; Yousaf, M.; Hou, H. Role of Recombinant DNA Technology to Improve Life. *Int. J. Genom.* **2016**, *2016*, 2405954. [CrossRef]
33. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A Training Algorithm for Optimal Margin Classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; ACM: New York, NY, USA, 1992; pp. 144–152.