ARTICLE

OPEN ACCESS

The grand biological universe: A comprehensive geometric construction of genome space

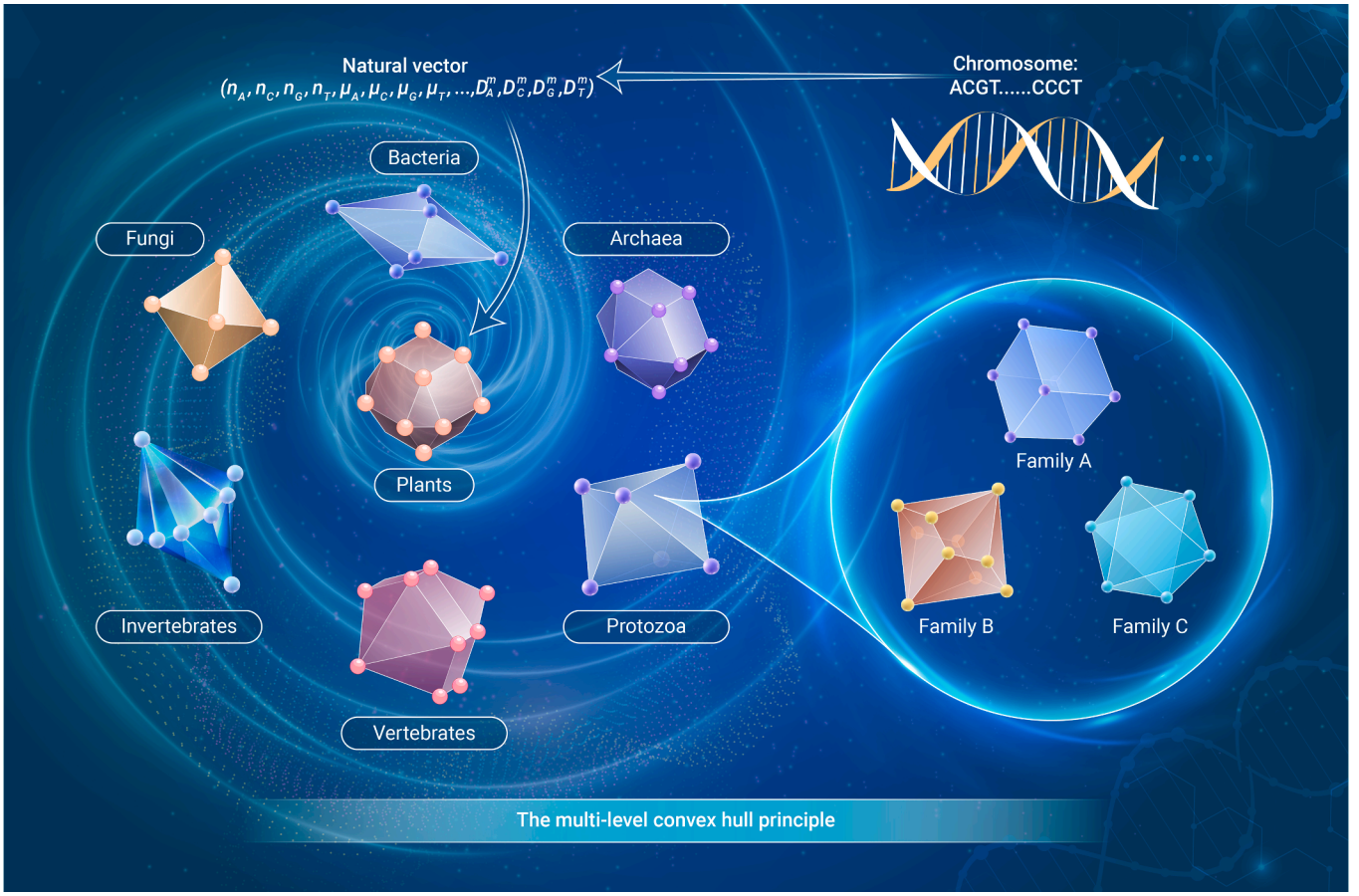
Hongyu Yu,[1,5] Nan Sun,[2,5] Ruohan Ren,[3,4,5] Tao Zhou,[1,5] Mengcen Guan,[1,5] Leqi Zhao,[1,5] and Stephen S.-T. Yau[1,2,*]

*Correspondence: yau@uic.edu

Received: October 22, 2024; Accepted: April 27, 2025; Published Online: April 30, 2025; https://doi.org/10.1016/j.xinn.2025.100937

www.the-innovation.org

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- The grand biological universe integrates natural vectors derived from all reliable reference sequences.
- The multi-level convex hull principle is validated within the grand biological universe.
- Convex hull separation originates from biological traits rather than high-dimensional space characteristics.
- The optimal metrics for classification are constructed within the grand biological universe.

OPEN ACCESS

ARTICLE

The Innovation

The grand biological universe: A comprehensive geometric construction of genome space

Hongyu Yu,[1,5] Nan Sun,[2,5] Ruohan Ren,[3,4,5] Tao Zhou,[1,5] Mengcen Guan,[1,5] Leqi Zhao,[1,5] and Stephen S.-T. Yau[1,2,*]

[1]Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

[2]Beijing Institute of Mathematical Sciences and Applications (BIMSA), Beijing 101408, China

[3]Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

[4]Tri-Institutional Training Program in Computational Biology and Medicine, Weill Cornell Medicine, New York, NY 10065, USA

[5]These authors contributed equally

*Correspondence: yau@uic.edu

Received: October 22, 2024; Accepted: April 27, 2025; Published Online: April 30, 2025; https://doi.org/10.1016/j.xinn.2025.100937

Citation: Yu H., Sun N., Ren R., et al., (2025). The grand biological universe: A comprehensive geometric construction of genome space. The Innovation **6(8)**, 100937.

Analyzing the geometric relationships among genomic sequences from a mathematical perspective and revealing the laws hidden within these relationships is a crucial challenge in bioinformatics. The natural vector method constructs a genome space by extracting statistical moments of *k*-mers to illuminate the relationships among genomes. This approach highlights a fundamental law in biology known as the convex hull principle, which states that natural vectors corresponding to different types of biological sequences form distinct, non-overlapping convex hulls. Previous studies have validated this important principle across various datasets. However, they often focused on specific kingdoms and did not thoroughly analyze the significance of the dimensions required for the convex hull separation. In this study, we integrate all reliable sequences from different kingdoms to construct the grand biological universe, within which we comprehensively validate the multi-level convex hull principle. We demonstrate that the separation of convex hulls arises from biological properties rather than mathematical characteristics of high-dimensional spaces. Furthermore, we develop suitable metrics within the grand biological universe to facilitate efficient sequence classification. This research advances the convex hull principle through both theoretical development and experimental validation, making significant contributions to the understanding of the geometric structure of genome space.

INTRODUCTION

Imitating Hilbert, who proposed 23 problems in mathematics in 1900, the Defense Advanced Research Projects Agency (DARPA) introduced 23 mathematical challenges in 2008. These challenges have proven to be highly influential for the development of mathematics in the 21st century. Many of these challenges significantly intersect with the field of biology. From a mathematical perspective, biological problems can be understood more deeply and quickly than relying solely on expensive experiments. Notably, two intimately connected challenges—"Geometry of Genome Space" and "What are the Fundamental Laws of Biology?"—have garnered widespread research interest.

To explore the geometric structure of genomes from a mathematical perspective, genomes must first be mapped into what is referred to as genome space, a specific type of metric space. This mapping concept forms the foundation for many alignment-free methods. Traditionally, biological sequence comparison relies on alignment techniques that assess sequence similarity using dynamic programming algorithms.[1–4] These methods align sequences position by position, making the process computationally demanding. In response to the growing number of discovered sequences, search-based alignment methods have emerged.[5–7] While the concept of searching has enhanced the computational efficiency of alignment methods, these methods heavily depend on large databases and often analyze sequences in fragments rather than understanding their overall structure. Consequently, alignment-free methods have been developed to overcome these constraints in recent years.[8,9] The predominant idea in alignment-free methods involves transforming genomes of diverse sizes into fixed-dimension vectors, allowing their distances to be calculated using the L^p norm. The key advantage of this strategy is its computational speed and independence from databases, enhancing efficiency when analyzing extensive sequence datasets.

Various alignment-free methods, based on different theories such as Fourier transforms, Markov chains, and deep learning, have been successfully implemented in fields like sequence similarity search, clustering, classification, and phylogenetic analysis.[10–14] These methods introduce novel concepts and tools for bioinformatics research. However, not all alignment-free methods can accurately depict the geometry of genome space, as they must fulfill specific mathematical properties. Most fundamentally, the genetic relatedness between sequences should be well represented by the distance between vectors, and different types of sequences should exhibit good separation properties in the vector space. Furthermore, we seek a one-to-one transformation from sequences to vectors, which is a necessary condition in genome space. Using deep learning-based alignment-free methods as an instance, due to their black-box nature, it is typically difficult to mathematically prove the one-to-one correspondence between sequences and points, making it challenging for rigorous mathematical analysis. Additionally, the embedding space of deep learning models can change significantly depending on the training data, making it unsuitable for analyzing stable genome space properties. Among the various alignment-free methods, the natural vector method, which embeds sequences based on statistical moments of *k*-mers, stands out for fulfilling these criteria.[15–20]

Building on the characterization of the geometric structure of genome space, the natural vector method further reveals a fundamental law of biology: the convex hull principle. This principle asserts that the natural vectors corresponding to different types of biological sequences form distinct, non-overlapping convex hulls in high-dimensional space. This property has been extensively validated across various types of biological data, suggesting that it represents a widely applicable theorem in the biological domain.[17–20] The convex hull principle has broad applications: not only can we classify unknown sequences by determining to which convex hull they belong, but we can also reverse the process to identify potential natural vectors within the convex hull that correspond to sequences yet to be discovered experimentally.[21,22] This principle, derived from a mathematical perspective, encapsulates one of the fundamental laws of biology.

However, the responses to DARPA's two questions remain incomplete. Specifically, there are two main issues. First, previous alignment-free methods, including the natural vector approach, have often been limited to classifying sequences within specific datasets, such as those confined to a single kingdom, without integrating all types of biological data. This has resulted in an incomplete genome space, meaning that the inferred fundamental laws have only been tested within localized genome spaces rather than within a comprehensive genome space that encompasses all biological sequences.[17–20] Second, spaces of different dimensions possess distinct geometric properties, and challenges such as the curse of dimensionality become particularly significant in high dimensions.[23] Previous interpretations of the convex hull principle have not quantitatively analyzed the dimensions required for the non-overlapping property of the convex hulls, nor have they distinguished between the biological and mathematical mechanisms underlying this non-overlapping nature. Therefore, there is still room for further development of the convex hull principle.

In this paper, we will utilize the natural vector method to construct a complete genome space using all reliable sequences from seven datasets (bacteria, archaea, fungi, plants, protozoa, invertebrates, and vertebrates) available in NCBI, which we will refer to as the grand biological universe. This grand biological universe will be employed to validate the convex hull principle, offering a

comprehensive verification of this important law. Furthermore, we will demonstrate that the dimensionality required for the non-overlapping property of the convex hulls is sufficiently small and is based on biological mechanisms rather than solely the mathematical mechanisms of high-dimensional spaces. Additionally, we develop metrics that integrate various sequence features within the grand biological universe to measure similarities between sequences, facilitating efficient sequence classification. In conclusion, we have constructed a marvelous geometric picture: within the grand biological universe, there exist seven mutually disjoint biological galaxies. Within each galaxy, there are also numerous mutually disjoint star clusters, and the distances between stars can be measured by metrics distorted by the gravitational influence of the stars. The work presented in this paper provides a geometric characterization of the complete genome space and makes significant contributions to the theoretical development and experimental validation of the convex hull principle.

MATERIALS AND METHODS

Materials

Seven types of genomes were downloaded from NCBI to construct the grand biological universe: bacteria, archaea, fungi, plants, protozoa, invertebrates, and vertebrates. The dataset was sourced from the RefSeq database (https://ftp.ncbi.nlm.nih.gov/refseq/release/), the Assembly GCF database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/), and the Assembly GCA database (https://ftp.ncbi.nlm.nih.gov/genomes/genbank/) in March 2022.

This study applies the following filtering criteria: (1) removal of duplicate sequences, (2) exclusion of unassembled sequences or organelle sequences, (3) removal of sequences containing degenerate bases, and (4) omission of sequences lacking classification labels. After the filtering process, 30,121 reliable sequences remain.

In the phylogenetic analysis of fungi, plants, protozoa, vertebrates, and invertebrates, we incorporated organelle data as supplemental information separately. These data originate from the same sources as the previously mentioned datasets, but instead of being constrained to chromosomal sequences, they are based on complete organelle sequences. After the filtering process, 16,577 reliable sequences remain.

To verify the stability of the grand biological universe, we additionally introduced the latest plant reference sequences as a supplement. This dataset was sourced from the RefSeq database (https://ftp.ncbi.nlm.nih.gov/refseq/release/) in March 2025, and the same data-cleaning methods as before were applied. A total of 170 new chromosomes were included.

The accession numbers and names of all reliable sequences can be found in the GitHub repository (https://github.com/BobYHY/GrandUniverse).

The natural vector method

The natural vector method is an alignment-free approach that transforms DNA sequences into vectors of moments.[15] Let $s_i, \alpha \in \{A, C, G, T\}$ and consider the sequence $S = s_1 s_2 \ldots s_n$, we can define

$$w_\alpha(s_i) = \begin{cases} 1, s_i = \alpha \\ 0, \text{otherwise.} \end{cases}$$

Then the j-th ordered element D_α^j of the natural vector can be defined as

$$\begin{cases} D_\alpha^0 = n_\alpha = \sum_{i=1}^{n} w_\alpha(s_i) \\ D_\alpha^1 = \mu_\alpha = \sum_{i=1}^{n} \frac{i}{n_\alpha} w_\alpha(s_i) \\ D_\alpha^j = \sum_{i=1}^{n} \frac{(i-\mu_\alpha)^j}{n_\alpha^{j-1} n^{j-1}} w_\alpha(s_i), (j = 2, 3, 4, \ldots) \end{cases}$$

where $n = n_A + n_T + n_C + n_G$. D_α^0 and D_α^1 are known as the frequency and the mean position of nucleotide α. We define the natural vector of order m as

$$\left(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, \ldots, D_A^m, D_C^m, D_G^m, D_T^m\right).$$

It is proven that one-to-one correspondence between sequences and natural vectors exists rigorously when the order is sufficiently high.

The k-mer natural vector method is an extension of the traditional natural vector method.[16] A k-mer is a string composed of k nucleotides, resulting in 4^k possible k-mers, denoted as $l_1, \ldots, l_{4^k}$. For a sequence $S = s_1 s_2 \ldots s_n$, we can view it as a sequence consisting of $n - k + 1$ k-mers; i.e., $(s_1 \ldots s_k), \ldots, (s_{n-k+1} \ldots s_n)$. We can define the j-th ordered moments of k-mers similarly to those of nucleotides and thereby define the k-mer natural vector as follows:

$$\left(n_{l_1}, \ldots, n_{l_{4^k}}, \mu_{l_1}, \ldots \mu_{l_{4^k}}, \ldots, D_{l_1}^m, \ldots, D_{l_{4^k}}^m\right).$$

Since the k-mer natural vector method has already extended the information content of the vectors from the perspective of k-mer length, there is no need to further expand it in terms of moments. Therefore, in the k-mer natural vector, the order m is typically set to 2. Additionally, if $D_{l_i}^0 = 0$, we set $D_{l_i}^1 = D_{l_i}^2 = \ldots = 0$.

The convex hull principle

The convex hull is one of the most basic concepts in computational geometry. Mathematically, the convex hull is the minimal convex set that contains the finite point set $C = \{x_1, x_2, \ldots, x_k\}, x_i \in \mathbb{R}^K$:

$$\mathrm{Conv}(C) = \left\{ \sum_{i=1}^{k} \theta_i x_i \middle| \theta_i \geq 0, \sum_{i=1}^{k} \theta_i = 1 \right\}.$$

The determination of convex hull disjointness can be solved using optimization methods. Suppose that $A = \mathrm{Conv}(\{a_1, \ldots, a_m\})$ and $B = \mathrm{Conv}(\{b_1, \ldots, b_n\})$ are convex hulls in $\mathbb{R}^K$. Any point in A and B can be represented by a linear combination of their point sets. If A and B intersect, then the convex combination of these points satisfies the equation

$$\sum_{i=1}^{m} \lambda_i a_i = \sum_{j=1}^{n} \theta_j b_j,$$

where non-negative numbers λ_i, θ_j satisfy $\sum_{i=1}^{m} \lambda_i = 1$ and $\sum_{j=1}^{n} \theta_j = 1$.

Therefore, the problem of determining the disjointness of convex hulls can be transformed into an optimization problem. The objective is to find coefficients $\{\lambda_1, \ldots, \lambda_m, \theta_1, \ldots, \theta_n\}$ so that the optimal solution of the following optimization problem is zero:

$$\min_{\lambda_i, \theta_j} \left\| \sum_{i=1}^{m} \lambda_i a_i - \sum_{j=1}^{n} \theta_j b_j \right\|$$

subject to:

$$\begin{cases} \sum_{i=1}^{m} \lambda_i = 1, \lambda_i \geq 0 \quad \text{for all } i \\ \sum_{j=1}^{n} \theta_j = 1, \theta_j \geq 0 \quad \text{for all } j \end{cases}$$

If the optimal value is greater than zero, then the convex hulls A and B are disjoint. It is worth noting that, when certain dimensional values are relatively small, numerical errors in optimization can arise. To mitigate these numerical errors caused by differing magnitudes across dimensions, we employed maximum normalization.

In this study, the convex hull principle is applied to sets of points x_i, where each point represents the 1-mer natural vector corresponding to a nucleic acid sequence. These sequences are grouped by shared biological classifications, such as belonging to the same family or originating from the same biological dataset, referred to as the biological galaxy. The convex hull principle posits that the natural vectors derived from sequences of different classifications form distinct, disjoint convex hulls in the vector space. Furthermore, the low dimensionality of this space implies that the separation of these convex hulls is highly dependent on the correctness of the biological classification labels. This indicates that the mechanism underlying the convex hull principle is driven by biological factors rather than purely mathematical properties, underscoring the importance of accurate classification in determining disjointness. This is the rationale for using lower-dimensional 1-mer natural vectors in this analysis.

After establishing the method for determining convex hull intersections, we introduce the concept of the significance ratio to further prove that the dimensionality used is biologically meaningful. The significance ratio is computed by randomly shuffling the classification labels of the sequences and reapplying the convex hull principle. This process is repeated N times. Let n represent the number of instances where the convex hull principle still holds after the randomization. The significance ratio is then defined as $\frac{N-n}{N}$. A higher significance ratio indicates that the convex hull disjointness relies more on biologically meaningful labels rather than mathematical properties alone. In this study, $N = 100$.

The natural metrics

In the k-mer biological universe, we define a series of selectable distance metrics based on the k-mer natural vectors. The general structure of these metrics is represented as

$$\sum_{k=1}^{K} a_k d_k^p,$$

where d_k^p denotes the distance metric derived from the k-mer natural vector in the L^p space. In this study, p is set to either 1 or 2, and K ranges from 1 to 9. The coefficients a_k represent different weighting strategies, which are developed from previous work,[20] and consist of the following seven choices:

$$\begin{cases} A1: a_k = 1_{\{k=K\}} \\ A2: a_k = \dfrac{1}{1.5^k} \\ A3: a_k = \dfrac{1}{2^k} \\ A4: a_k = \dfrac{1}{3^k} \\ A5: a_k = \dfrac{1}{k^{1.5}} \\ A6: a_k = \dfrac{1}{k^2} \\ A7: a_k = \dfrac{1}{k^3} \end{cases}$$

The natural metrics are selected based on which of the 126 possible combinations achieves the highest 1-nearest neighbor classification accuracy under the leave-one-out strategy.

t-distributed stochastic neighbor embedding

t-distributed stochastic neighbor embedding (t-SNE) is a non-linear dimensionality reduction technique specifically designed to map high-dimensional data to 2D or 3D space for visualization. It emphasizes preserving local structures and partially addresses the crowding problem inherent in high-dimensional data visualization.

The principle of t-SNE can be summarized in the following steps. First, it models the relationships in high-dimensional data using Gaussian distributions. Then, it reconstructs similar probability relationships in the low-dimensional space using the t distribution. By minimizing the difference between the high-dimensional and low-dimensional distributions, it achieves a low-dimensional visualization that closely approximates the high-dimensional structure.

t-SNE has been widely adopted in various fields, such as bioinformatics and natural language processing, due to its ability to reveal patterns and structures in complex datasets. It is important to know that, while t-SNE effectively captures local relationships, it fails to sufficiently maintain global relationships. Moreover, the distances in the t-SNE plot may not accurately reflect the true pairwise distances in the original high-dimensional space. Therefore, we do not expect that the convex hull principle holds after dimension reduction. Instead, we conduct t-SNE to indirectly observe separation properties of natural vectors.

RESULTS

Overview of the grand biological universe

The goal of the grand biological universe is to embed all known reliable biological genomes into a vector space and analyze them from a unified geometric perspective. This allows us to prove the important convex hull principle and perform effective classification.

As shown in Figure 1, the basic framework of modern classification, based on the Linnaean system, consists of domain, kingdom, phylum, class, order, family, genus, and species.[24,25] In this study, we divide the grand biological universe into seven parts based on the classification, analogous to galaxies in the cosmos. These biological galaxies correspond to bacteria, archaea, fungi, plants, protozoa, invertebrates, and vertebrates. Our analysis covers both the relationships between these seven biological galaxies and the internal properties and structures within each biological galaxy.

The method employed for embedding genomes into vector space is based on the natural vector, demonstrating a one-to-one correspondence in high-dimensional space. Following previous research, we adopt two different strategies to extract sequence features. The first considers only 1-mers and increases the order of the moments. This approach results in a lower-dimensional space, within which we will demonstrate profound geometric properties, such as the convex hull principle. In addition, this space allows for rapid classification. We refer to it as the 1-mer biological universe. The second method considers k-mers, embedding sequences into a higher-dimensional space that enables more refined classification. We refer to it as the k-mer biological universe. Both approaches are finite-dimensional truncations of the same feature extraction method, collectively forming the grand biological universe.

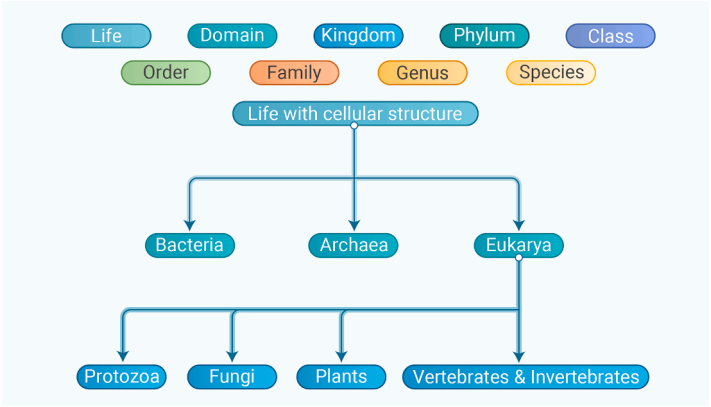


Figure 1. The basic scheme of modern classification for lives

The convex hull principle

The concept of the convex hull illustrated in Figure 2 is fundamental in mathematics. The convex hull of a point set characterizes the region corresponding to the set. The convex hull principle is an important empirical law in the field of biology, stating that the natural vectors corresponding to biological sequences of different classifications form distinct, non-overlapping convex hulls in space. In Figure 3, we present a toy example that demonstrates the process of transforming sequences into the vector form where the convex hulls can be tested for non-overlapping properties. Previous validations of this principle were typically based on relatively homogeneous datasets. Here, we provide a complete validation using the grand biological universe. We aim to achieve this non-overlapping property in a lower-dimensional space, which is why we consider the 1-mer biological universe.

The results show that the convex hulls formed by the seven biological galaxies (or datasets) are non-overlapping in the order 16 1-mer natural vector space, which is a 68-dimensional vector space. Furthermore, within this 68-dimensional space, the convex hulls formed by the families within each of the seven biological galaxies are also non-overlapping. (Due to the limited number of reliable sequences in the archaea, invertebrate, and vertebrate datasets, where the number of sequences from individual families is small, we instead verified the convex hull principle using classifications at the phylum, order, and class levels for these three datasets, respectively.) This result demonstrates the multi-level convex hull principle. As illustrated in Figure 4, the larger circles represent biological galaxies, while the smaller circles correspond to families within each galaxy. Not only are the biological galaxies mutually non-overlapping, but the families within each galaxy also do not intersect. In fact, within each biological galaxy, the number of dimensions required for the convex

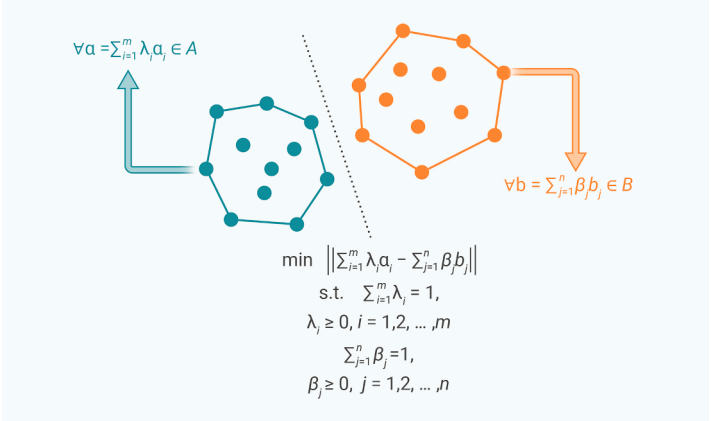


Figure 2. The convex hulls and the determination of their intersection The intersection of convex hulls is equivalent to the optimization problem having a minimum value of zero.

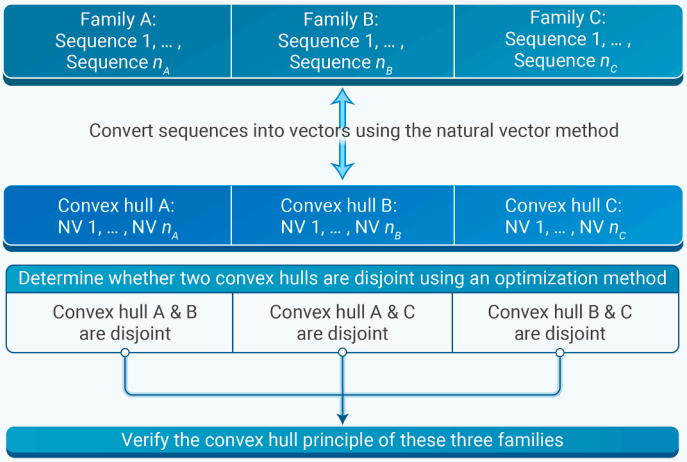


Figure 3. Flowchart of the convex hull analysis for a toy example

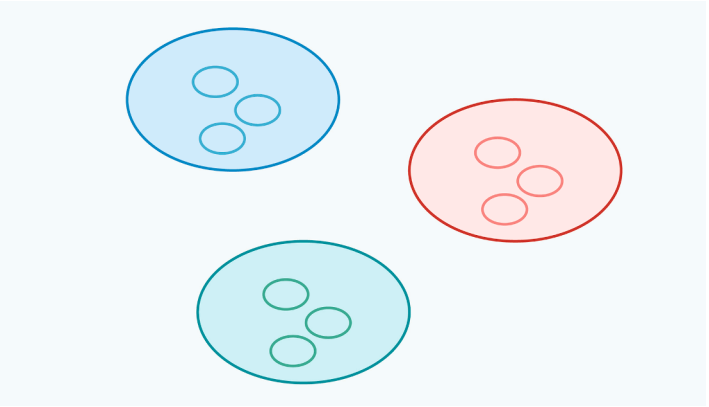


Figure 4. Demonstration of the multi-level convex hull principle

hulls to be disjoint is lower than the 68 dimensions required for the full space. For example, only 20 dimensions are required for the convex hull principle within the galaxy of plants. These details are provided in Table 1. This comprehensive validation, using all reliable data, confirms the convex hull principle across multiple taxonomic levels.

Since dimensionality reduction leads to the loss of high-dimensional information, the convex hulls are not necessarily non-overlapping in two-dimensional space, making direct planar visualization of mutually exclusive groups difficult. Therefore, we use an example to indirectly illustrate the spatial differences between various families. In Figure 5, we select the three largest families of bacteria and use the t-SNE method to visualize their natural vectors.[26] Due to the information loss from the dimensionality reduction, we do not expect that the convex hull principle holds in this 2-dimensional space. This is merely used as an example to demonstrate the clustering tendency and the separation properties of natural vectors. The effectiveness of our method is strictly validated by the convex hull principle instead of the visualization.

We further analyze the value of the dimensions required for the convex hull principle. In extreme cases, if we need thousands of dimensions to ensure that the convex hull does not intersect, such a high dimensionality clearly lacks value. We aim for the separation of the convex hull to be based on biological mechanisms rather than mathematical properties of high-dimensional spaces. We apply the concept of the significance ratio introduced before to assess the magnitude of dimensions. The rationale behind the significance ratio is that if the convex hull principle strictly relies on true labels rather than merely on high-dimensional properties, then we consider it to have biological significance. The results indicate that the significance ratios for the seven galaxies are all close to 100%, further supporting the biological relevance of the convex hull principle.

We can describe the convex hull principle in the grand biological universe using cosmological concepts: within the grand biological universe, there exist seven biological galaxies that are mutually disjoint, and within each galaxy, there are numerous star clusters that are also mutually disjoint.

The natural metrics

In addition to illustrating the geometric structure of the grand biological universe, we aim to define metrics on this space to measure distances between biological sequences. Once pairwise distances between sequences are established, we can apply straightforward classification methods, such as k-nearest neighbors, to classify new sequences based on these distances.[27] Using the Euclidean metric within the 1-mer biological universe, we can rapidly determine the biological galaxy to which a sequence belongs, achieving an accuracy of 94.1% under the 1-nearest neighbor method with leave-one-out strategy. For more granular classification within a galaxy (typically down to the family level, though for datasets with fewer sequences in the same family, such as archaea, invertebrates, and vertebrates, classification is performed at the phylum, order, and class levels, respectively), the reduced number of sequences enables us to utilize more refined metrics offered by the *k*-mer biological universe.

The *k*-mer biological universe integrates statistical information from *k*-mers across varying values of *k*. Unlike the 1-mer biological universe, the *k*-mer biological universe offers a richer set of metrics. In addition to selecting different upper bounds for *k*, we can apply various weighting schemes during the integration of *k*-mer statistical information. Furthermore, both L_1 and L_2 metrics are available for measuring distances. By enumerating combinations of these three adjustable parameters, we determined the optimal metric for each biological galaxy, referred to as the natural metric. The natural metrics for the seven biological galaxies, along with their corresponding classification accuracies, are presented in Table 2. In the table, d_k^p denotes the distance metric based on the *k*-mer natural vector in the L^p space. Furthermore, to better demonstrate the advantages of the natural metric, we also included the accuracy obtained using a simple natural vector approach (i.e., the Euclidean metric of the 12-dimensional natural vector). It is evident that, by integrating information from various *k*-mers, the accuracy is significantly improved. More detailed results are provided in the GitHub repository.

We observed that, while all possible metrics result in relatively high classification accuracy, the optimal natural metric varies across galaxies. This variation can be attributed to inherent differences in the genome characteristics of different biological galaxies. However, an alternative hypothesis can be drawn from a cosmological analogy. If we were to incorporate all real-world sequences into the grand biological universe, perfect classification (100% accuracy) might be attainable. However, because not all real-world sequences have been sequenced, and those that have been sequenced may contain errors not filtered out, the data we are currently working with represent only a small subset of the full set of sequences. The unobserved sequences, akin to dark matter, are not part of the observable dataset but still influence the natural metrics, much like dark matter in cosmology distorts space-time metrics.

Furthermore, the method for constructing natural metrics can also determine the optimal distance calculation method in phylogenetic analysis. In phylogenetic analysis, if a species' genome is distributed across multiple chromosomes, with each chromosome sequence containing only partial genetic information, the excessive sequence divergence can lead to poor phylogenetic analysis results. Therefore, for species with multiple nuclear sequences (fungi, plants, protozoa, invertebrates, and vertebrates), organellar sequences are used as the analysis target. Although the grand biological universe focuses on sequences of nuclear or nucleoid DNA, we can still apply the same method to determine natural metrics within the organellar space. We calculate the distance between two families in phylogenetic analysis by using the natural metric between the average natural vector of all sequences within each family. Subsequently, we employ the BIONJ algorithm implemented in FastME for the phylogenetic analysis.[28,29] Some biologically meaningful conclusions can be drawn from the results. For instance, we discovered a close relationship between Skeletonemataceae and Bacillariaceae. Despite this finding, they are classified under different classes in the NCBI database. The classification of Skeletonemataceae has been a matter of debate. In the NCBI database, it is classified under the class Coscinodiscophyceae. However, other sources, such as the

Table 1. Dimension required for convex hull disjointness within each biological galaxy and the corresponding significance ratio

	Dimension for disjointness	Significance ratio
Bacteria	48	100%
Archaea	68	97%
Fungi	40	100%
Plants	20	100%
Protozoa	36	100%
Invertebrates	40	100%
Vertebrates	28	100%

WoRMS (World Register of Marine Species; https://www.marinespecies.org/aphia.php?p=taxdetails&id=622500) and GBIF (Global Biodiversity Information Facility; https://www.gbif.org/search?q=skeletonemataceae), classify it under the class Bacillariophyceae. Furthermore, in some databases, like AlgaeBase,[30] it is grouped into the class Mediophyceae. This inconsistency seems to stem from different classification decisions. Based on our comprehensive analysis of mitochondrial sequences based on natural metrics, we argue that Skeletonemataceae and Bacillariaceae should belong to the same class, as the results of WoRMS and GBIF. We hope this provides additional information for taxonomists from the perspective of mitochondrial sequence characteristics. This analysis is not a result of the grand biological universe itself but rather an application of the same method in the organellar space. Therefore, we will not elaborate further; for more details, such as the structure of the organellar space as well as the phylogenetic trees, please refer to the GitHub repository.

In the process of constructing the phylogenetic tree, it is worth noting the method for measuring the distance between sets of vectors. In previous studies, the Hausdorff distance was commonly used to compare the distances between vector sets. In contrast, we directly use the distance between the average natural vectors in this study for two main reasons. First, our study clearly demonstrates the geometric relationships among different biological universes and among families within the same biological universe. As shown in Figure 4, natural vector sets are not irregularly shaped; rather, they occupy distinct positions with non-overlapping convex hulls. In such cases, the distance between average vectors sufficiently captures the inter-set differences. Second, when dealing with large datasets, calculating the Hausdorff distance becomes very time consuming, making it unsuitable for rapid phylogenetic analysis. The computation of the average natural vector, however, is much faster and better meets the demands of the big data era. Therefore, we chose to use the distance between the average natural vectors to evaluate our phylogenetic analysis results.

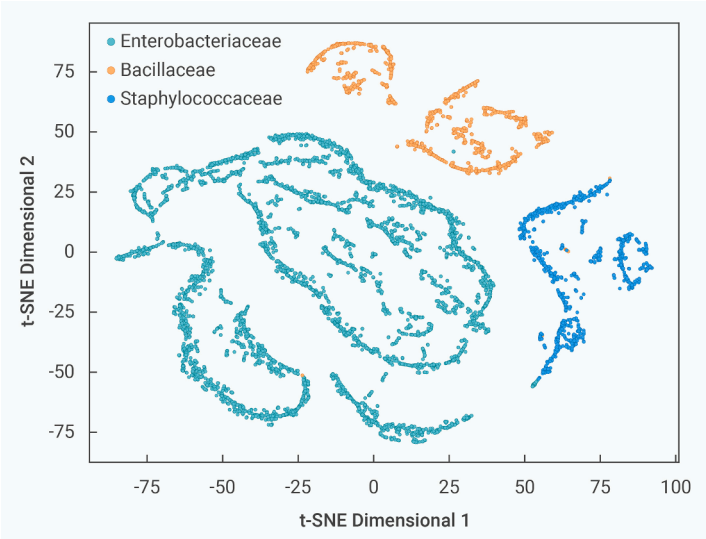


Figure 5. The t-SNE visualization of 48-dimensional natural vectors for the three largest bacterial families

Table 2. The natural metric performance and simple natural vector performance within each biological galaxy

	Natural metric Acc	Simple NV Acc
Bacteria	d_9^1: 90.3%	77.7%
Archaea	d_9^1: 90.6%	72.2%
Fungi	d_9^1: 90.6%	86.5%
Plants	d_2^1: 82.2%	79.7%
Protozoa	$\sum_{k=1}^{9} \frac{1}{k^{1.5}} d_k^1$: 92.7%	88.7%
Invertebrates	$\sum_{k=1}^{9} \frac{1}{k^{2}} d_k^1$: 87.8%	82.6%
Vertebrates	$\sum_{k=1}^{2} \frac{1}{2^{k}} d_k^1$: 85.9%	82.6%

NV, natural vector; Acc, accuracy.

Stability of the grand biological universe

We further evaluated the stability of the grand biological universe by incorporating an additional dataset comprising the latest plant reference chromosome sequences. After integrating these new chromosome data, the multi-level convex hull principle continued to hold at the originally determined dimensions. Specifically, the seven biological galaxies remained non-overlapping in the 68-dimensional space, and the convex hulls in the plant galaxy persisted as non-overlapping in the 20-dimensional space. The unchanged dimensionality at which these convex hulls remain disjoint underscores a strong correlation between the dimensionality and the biological universes. Moreover, on the expanded plant dataset, the accuracy of the natural metric reached 82.4%, which is in close agreement with the previous result of 82.2%. These findings confirm the robustness and stability of the grand biological universe framework. To facilitate future validations with new datasets, we integrated the convex hull analysis code into our GitHub repository, enabling researchers to quickly assess the applicability of the convex hull principle on their own data.

DISCUSSION

The natural vector method allows genomes to be mapped one-to-one into a vector space using statistical techniques, enabling the study of the geometry of genome space. The convex hull principle is a key geometric characterization in this context. It states that the natural vectors corresponding to sequences from different classifications form disjoint convex hulls in the vector space. This principle is hierarchical, applying across multiple levels of classification. Previous studies have partially validated this conclusion, but they were limited to smaller datasets and did not further investigate the significance of the dimensionality of the space.[17,18,20]

In this paper, we extend the convex hull principle both in practice and theory. For the first time, we validated this principle at multiple taxonomic levels using reliable sequences from all species. Additionally, we introduced a novel algorithm to demonstrate that the separability observed in the vector space is driven by biological characteristics rather than purely mathematical properties. This approach adds a new layer of understanding to the convex hull principle, emphasizing its biological foundation.

The grand biological universe we constructed not only validates important biological principles from a theoretical standpoint but also offers promising applications. For instance, when new, unknown sequences are discovered in the future, we can apply the optimal metric to determine the category to which these sequences should belong. Moreover, past research has demonstrated that it is possible to reverse-engineer sequences by identifying points within the convex hull corresponding to natural vectors, leading to the discovery of previously unknown sequences.[21,22] With the convex hull principle now validated in the grand biological universe, this methodology for detecting new sequences can be further refined and developed, opening up new avenues for biological discovery.

In conclusion, this paper introduces the concept of the grand biological universe, modeled after cosmological principles. Within this universe, we identified seven disjoint biological galaxies, each representing a major taxonomic group. Within each galaxy, there exist multiple star clusters, each forming a disjoint set

in the universe. These clusters are composed of points that represent individual sequences, and the relationships between these points are defined by natural metrics. Interestingly, the natural metric varies slightly across different biological galaxies, reflecting how the metric is influenced by the unique characteristics of each galaxy. This nuanced variation demonstrates how the geometry of the space is subtly shaped by the biological properties of the sequences, offering profound insight into the structure and diversity of life.

RESOURCE AVAILABILITY

Materials availability

This study did not generate new unique materials/reagents.

Data and code availability

All data and code are available in the GitHub repository (https://github.com/BobYHY/GrandUniverse).

FUNDING AND ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (12171275) and the Tsinghua University Education Foundation fund. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. S.S.-T.Y. is grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was done.

AUTHOR CONTRIBUTIONS

S.S.-T.Y. designed the experiment and supervised the project. H.Y., N.S., R.R., T.Z., M.G., and L.Z. performed research, analyzed data, and wrote the paper. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used ChatGPT in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Needleman, S.B. and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**:443–453. DOI:https://doi.org/10.1016/0022-2836(70)90057-4.
2. Smith, T.F. and Waterman, M.S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**:195–197. DOI:https://doi.org/10.1016/0022-2836(81)90087-5.
3. Higgins, D.G. and Sharp, P.M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**:237–244. DOI:https://doi.org/10.1016/0378-1119(88)90330-7.
4. Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**:1792–1797. DOI:https://doi.org/10.1093/nar/gkh340.
5. Altschul, S.F., Gish, W., Miller, W. et al. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**:403–410. DOI:https://doi.org/10.1016/S0022-2836(05)80360-2.
6. Remmert, M., Biegert, A., Hauser, A. et al. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**:173–175. DOI:https://doi.org/10.1038/nmeth.1818.
7. Steinegger, M., Mirdita, M. and Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods* **16**:603–606. DOI:https://doi.org/10.1038/s41592-019-0437-4.
8. Vinga, S. and Almeida, J. (2003). Alignment-free sequence comparison—a review. *Bioinformatics* **19**:513–523. DOI:https://doi.org/10.1093/bioinformatics/btg005.
9. Bonham-Carter, O., Steele, J. and Bastola, D. (2014). Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Brief. Bioinform.* **15**:890–905. DOI:https://doi.org/10.1093/bib/bbt052.
10. Yin, C. and Yau, S.S.-T. (2015). An improved model for whole genome phylogenetic analysis by Fourier transform. *J. Theor. Biol.* **382**:99–110. DOI:https://doi.org/10.1016/j.jtbi.2015.06.033.
11. Qi, J., Wang, B. and Hao, B.I. (2004). Whole Proteome Prokaryote Phylogeny Without Sequence Alignment: A K-String Composition Approach. *J. Mol. Evol.* **58**:1–11. DOI:https://doi.org/10.1007/s00239-003-2493-7.
12. Jun, S.-R., Sims, G.E., Wu, G.A. et al. (2010). Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. *Proc. Natl. Acad. Sci. USA* **107**:133–138. DOI:https://doi.org/10.1073/pnas.0913033107.
13. Zheng, W., Yang, L., Genco, R.J. et al. (2019). SENSE: Siamese neural network for sequence embedding and alignment-free comparison. *Bioinformatics* **35**:1820–1828. DOI:https://doi.org/10.1093/bioinformatics/bty887.
14. Ren, R., Yin, C. and S-T Yau, S. (2022). kmer2vec: A Novel Method for Comparing DNA Sequences by word2vec Embedding. *J. Comput. Biol.* **29**:1001–1021. DOI:https://doi.org/10.1089/cmb.2021.0536.
15. Deng, M., Yu, C., Liang, Q. et al. (2011). A Novel Method of Characterizing Genetic Sequences: Genome Space with Biological Distance and Applications. *PLoS One* **6**:e17293. DOI:https://doi.org/10.1371/journal.pone.0017293.
16. Wen, J., Chan, R.H.-F., Yau, S.-C. et al. (2014). K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* **546**:25–34. DOI:https://doi.org/10.1016/j.gene.2014.05.043.
17. Zhao, X., Tian, K., He, R.L. et al. (2019). Convex hull principle for classification and phylogeny of eukaryotic proteins. *Genomics* **111**:1777–1784. DOI:https://doi.org/10.1016/j.ygeno.2018.11.033.
18. Tian, K., Zhao, X. and Yau, S.S.-T. (2018). Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. *J. Theor. Biol.* **456**:34–40. DOI:https://doi.org/10.1016/j.jtbi.2018.07.035.
19. Wang, Y., Tian, K. and Yau, S.S.-T. (2019). Protein Sequence Classification Using Natural Vector and Convex Hull Method. *J. Comput. Biol.* **26**:315–321. DOI:https://doi.org/10.1089/cmb.2018.0216.
20. Sun, N., Pei, S., He, L. et al. (2021). Geometric construction of viral genome space and its applications. *Comput. Struct. Biotechnol. J.* **19**:4226–4234. DOI:https://doi.org/10.1016/j.csbj.2021.07.028.
21. Jiao, X., Pei, S., Sun, Z. et al. (2021). Determination of the nucleotide or amino acid composition of genome or protein sequences by using natural vector method and convex hull principle. *Fundam. Res.* **1**:559–564. DOI:https://doi.org/10.1016/j.fmre.2021.08.010.
22. Zhao, R., Pei, S. and Yau, S.S.-T. (2022). New Genome Sequence Detection via Natural Vector Convex Hull Method. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**:1782–1793. DOI:https://doi.org/10.1109/TCBB.2020.3040706.
23. Verleysen, M. (2003). Learning high-dimensional data. *NATO Sci. Ser. III Comput. Syst. Sci.* **186**:141–162.
24. Woese, C.R. and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74**:5088–5090. DOI:https://doi.org/10.1073/pnas.74.11.5088.
25. Woese, C.R., Kandler, O. and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**:4576–4579. DOI:https://doi.org/10.1073/pnas.87.12.4576.
26. van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**:2579–2605.
27. Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13**:21–27. DOI:https://doi.org/10.1109/TIT.1967.1053964.
28. Gascuel, O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**:685–695. DOI:https://doi.org/10.1093/oxfordjournals.molbev.a025808.
29. Lefort, V., Desper, R. and Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol. Biol. Evol.* **32**:2798–2800. DOI:https://doi.org/10.1093/molbev/msv150.
30. Liu, S., Xu, Q., Liu, K. et al. (2021). Chloroplast Genomes for Five Skeletonema Species: Comparative and Phylogenetic Analysis. *Front. Plant Sci.* **12**:774617. DOI:https://doi.org/10.3389/fpls.2021.774617.