



# The geometry of genome space and its applications

Stephen S.-T. Yau

Department of Mathematical Sciences, Tsinghua University  
Yanqi Lake Beijing Institute of Mathematical Sciences and Applications

( Oct. 2022 )

# CONTENTS

一、 Research Background

二、 Research Objective

三、 Research Methods

四、 Results

五、 Application



# ABSTRACT

Imitating Hilbert who proposed twenty-three problems in mathematics in 1900, DAPRA proposed twenty-three problems in pure and applied mathematics in 2008. These problems will be proven to be very influential for the development of mathematics in 21st-century. In the number 15 of DAPRA problems, we are asked to understand "The Geometry of Genome Space". A genome space consists of all known genomes of living beings and provides insights into their relationships, reflecting the important nature of the genomic universe. Mathematically, the genome space can be considered as the moduli space in mathematics. In this talk, we shall show that genome sequences can be canonically embedded in a high-dimensional Euclidean space by means of their natural vectors which describe the nucleotides distribution information within the genome sequence. In this way, we construct genome space as a subspace in a high-dimensional Euclidean space. In this space, a genome sequence is uniquely represented as a point, and how sequences are distributed in the genome space is determined. The similarity of sequences can be measured by the natural metric which is different from the induced metric from the ambient Euclidean space. Like our physical world, the dark matter / dark energy plays a crucial role in the construction of the correct natural metric in genome space. Here, we report the construction of genome spaces of virus, bacteria, and plants with natural metrics. These metrics are quite different in each genome space because different dark matter / dark energy may bend the space-time as predicted by Einstein theory.

DAPRA problem # 23 asks: What are the Fundamental Laws of Biology? Our convex hull principle for molecular biology states that the convex hull formed from natural vectors of one biological group does not intersect with the convex hull formed from any other biological group. This can be viewed as one of the Fundamental Laws of Biology for which DAPRA has been looking for since 2008. As applications, we provide the first mathematical method to find undiscovered genome sequence. Our theory allows us to explore where SARS-CoV-2 originated from. It provides a novel geometric perspective to study molecular biology. It also gives accurate way for large-scale sequences comparison in real-time manner.



## PART 01

# Research Background



## What is a Genome ?

- Genome is the complete set of genetic materials in an organism.
- In mathematical terminology, genome is a complete set of invariants of organism.

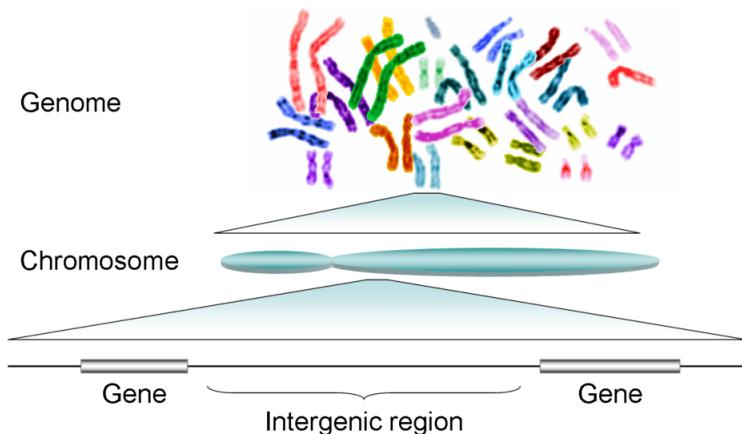


Figure : The genome is the sum of all genetic material in the body. Genes are carried on chromosomes.

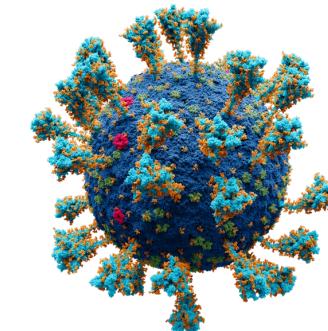
<https://www.biologyonline.com/dictionary/genome>





## Genome Example: SARS-CoV-2 (Virus)

- SARS-CoV-2 is a positive-sense, single-strand RNA virus that belongs to the beta coronavirus genus. It has a genome size of 29903 bp, which encodes multiple non-structural and structural proteins.



SARS-CoV-2 reference genome (NC\_045512.2)



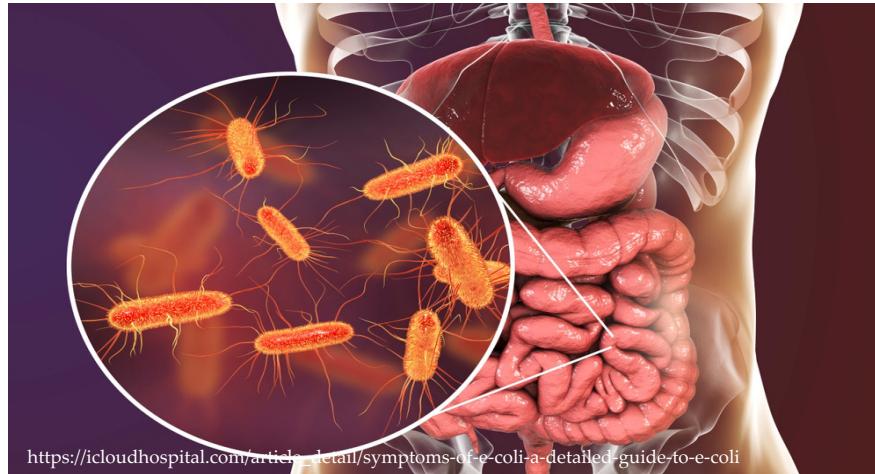
```
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAGTTTACCTCCCAAGTAACAAACCAACCAACTTCGATCTTGTAGATCTGTTCTAA
CGAACTTAAATCTGTGGCTGCACTCGGCTCATGCTTAGTGACTCACGAGTATAATTAAAC
TAATTAAGTCGTTGACAGGACACGAGTAACCTGCTATCTTCGAGGCTGCTTACGGTTCTGCGT
TTGCAGCCGATCATCAGCACACTAGGTTCTGTCGGGTGTGACCAGAAAGGTAAGATGGAGAGCCTGTC
CCTGGTTCAACGAGAAAACACAGCTCAACTCAGTTGCTTACAGGTCGGACGTGCTGTC
GTGGCTTGGAGACTCCGTGGAGGAGGCTTATCAGAGGACGCTAACATCTAAAGATGGCACTGTGG
CTTAGTAGAAGTTAAAGGCGTTTGCCTCAACTGAACAGCCCTATGTGTTATCAAACGTTGGAT
GCTCGAACTGACCTCATGGTGTAGTGGAGCTGAGCAGAAGGCAATTCTAGTACGGTC
GTAGTGGTAGAGACACTTGTGCTCTCATGGGCAATACCGTGGCTTACCGCAAGGTTCT
TCTTCGTAAGAACGTTAAAGGAGCTGGTGGCCATAGTTACGGGCCGATCTAAAGTCATTTACTTA
GGGACGAGCTTGGACTGATCCTTAAGAGTTCAAGAAAACCTGGAACACTAAACATAGCAGTGGT
TTACCCGTGAACTCATGGTGAACGTTAACGGAGGGCATACACTCGCTATGTCGATAACAACTTGTGG
CCCTGATGGTACCCCTTGGACTGCAATTAAAGACCTCTAGCACGTGCTGTTAAAGCTTATGCACTTG
TCCGAACAACTGGACTTATTGACACTAAAGAGGGGTATACTGCTGCCGTGAAACATGAGCATGAAATTG
CTTGGTACCGGAACGTTGAAAAGAGCTATGAATTGAGACACCTTTGAAATTAAATGGCAAGAA
ATTGACACCTTCAATGGGAATGTCAAATTGATTTCCATTAAATTCCATAATCAAGACTATTCAA
CCAAGGGTGAAGAAAAGCTTGTGATGGTGAAGTGTGATCATTGTTGAAACTTGTGCGTAC
CAAATGAATGCAACAAATGTGCCCTCAACTCTCATGAAGTGTGATCATTGTTGAAACTTGTGCA
GACGGGGCATTTGTTAAAGCCATTGCAATTGTCGACTGAGAATTGACTAAAGAAGGTGCCACT
ACTTGTGGTACTTACCCAAATGCTGTTAAAATTATTGTCCAGCATGTACAATTGAGTAG
```



NCBI: <https://www.ncbi.nlm.nih.gov/data-hub/taxonomy/2697049/>



## Genome Example: Escherichia coli (Bacteria)



- *E. coli* is commonly found in the lower intestine of warm-blooded organisms. In most cases, this bacteria helps digest the food you eat. However, certain strains of *E. coli* can cause symptoms including diarrhea, stomach pain. Some *E. coli* infections can be dangerous.
- The first complete DNA sequence of an *E. coli* genome (strain K-12 substr. MG1655) was published in 1997. It is a circular DNA molecule 4.6 million base pairs in length.
- The length of another *E. coli* reference genome recorded in NCBI is 5.6 million base pairs.

### Reference genome ASM584v2

Univ. Wisconsin (2013). Strain: K-12 substr. MG1655.  
RefSeq GCF\_000005845.2

[Download](#)

Genome size	4.6 Mb
Contig N50	4.6 Mb
Genes	4,639

### Reference genome ASM886v2

GIRC (2018). Strain: Sakai substr. RIMD 0509952.  
RefSeq GCF\_000008865.2

[Download](#)

Genome size	5.6 Mb
Contig N50	5.5 Mb
Genes	5,417



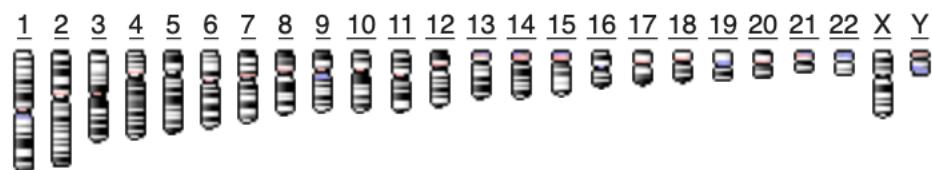
## Genome Example: *Homo sapiens* (Vertebrate)

- The Human Genome Project reported the sequencing of the entire genome for *Homo sapiens* in April 2003 [1], although only 92% of the DNA was actually decoded. With advancements in technology, scientists reported the first end-to-end human genome sequence in March, 2022.[2]
- In humans, the nuclear genome comprises approximately 3.2 billion nucleotides of DNA, divided into 24 linear molecules, the shortest 50 000 000 nucleotides in length and the longest 260 000 000 nucleotides, each contained in a different chromosome.[3]

### **Homo sapiens reference genome GRCh38.p14**

Submitted by Genome Reference Consortium (February 2022) [4]

RefSeq: GCF\_000001405.40



[1] Roth, Stephanie Clare (2019). What is genomic medicine?. Journal of the Medical Library Association. University Library System, University of Pittsburgh. 107 (3): 442–448.

[2] Hartley, Gabrielle. The Human Genome Project pieced together only 92% of the DNA – now scientists have finally filled in the remaining 8%. TheConversation.org. The Conversation US, Inc. Retrieved 4 April 2022.

[3] Human genome. <https://whatisdna.net/wiki/genetic-genealogy-understanding-ancestry-dna/>

[4] <https://www.ncbi.nlm.nih.gov/genome/?term=human%20genome>



## An important problem is how to compare these sequences ?

- In bioinformatics, there are alignment-based and alignment-free sequence analysis approaches to molecular sequence and structure data [1].
  - The pioneering approaches for sequence analysis were based on sequence alignment either global or local, pairwise or multiple sequence alignment [2, 3]. Alignment-based approaches generally give excellent results when the sequences under study are closely related and can be reliably aligned, but when sequences are divergent, a reliable alignment cannot be obtained. Another limitation of alignment-based approaches is their computational complexity and time-consuming. Given set A of sequences, if you add a new sequence {s} in to the set A, you need to re-align the set  $A \cup \{s\}$ , i.e. the previous result in A is not helpful. The similarity measure given by alignment does not satisfy triangular inequality.
  - Alignment-free methods can broadly be classified into five categories: a) methods based on k-mer/word frequency, b) methods based on the length of common substrings, c) methods based on the number of (spaced) word matches, d) methods based on micro-alignments, e) methods based on information theory and f) methods based on graphical representation.
  - Alignment-free approaches have been used in sequence similarity searches,[5] clustering and classification of sequences,[6] and more recently in phylogenetics [7, 8].
- None of the previous alignment-free methods are useful since a lot of sequence information are loss. In 2011, our team proposed natural vector to compare genomic sequences [9], which has been successfully applied to many studies.

[1] Vinga S, Almeida J (2003). Alignment-free sequence comparison-a review. Bioinformatics. 19 (4): 513–23.

[2] Batzoglou S (2005). The many faces of sequence alignment. Briefings in Bioinformatics. 6 (1): 6–22.

[3] Mullan L (2006). Pairwise sequence alignment—it's all about us!. Briefings in Bioinformatics. 7 (1): 113–5.

[4] Kemeny C, Notredame C (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics. 25 (19): 2455–65.

[5] Hide W, Burke J, Davison DB (1994). Biological evaluation of d2, an algorithm for high-performance sequence comparison. Journal of Computational Biology. 1 (3): 199–215.

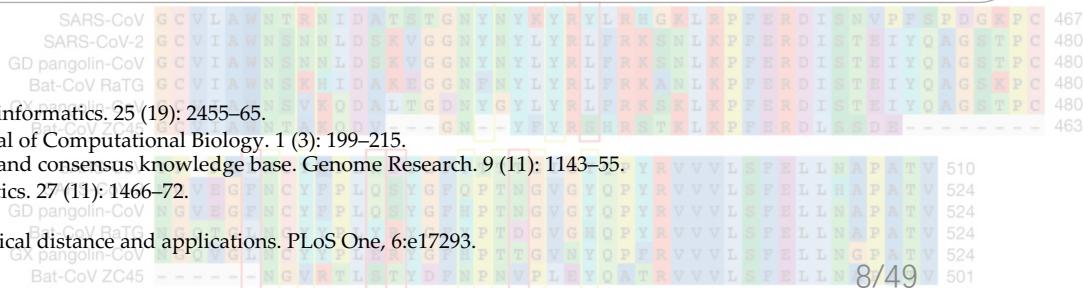
[6] Miller RT, et al. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. Genome Research. 9 (11): 1143–55.

[7] Domazet-Lošo M, Haubold B (2011). Alignment-free detection of local similarity among viral and bacterial genomes. Bioinformatics. 27 (11): 1466–72.

[8] Chan CX, Ragan MA (2013). Next-generation phylogenomics. Biology Direct. 8: 3.

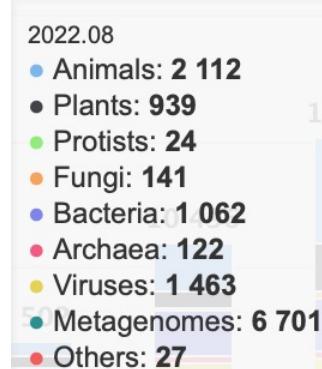
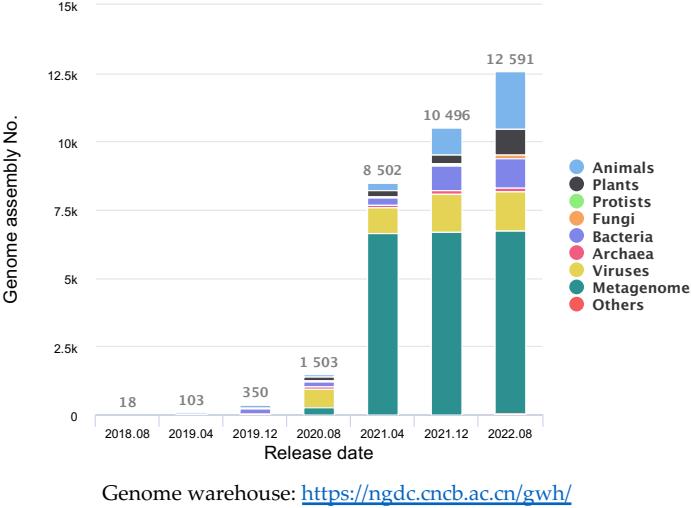
[9] Deng M, Yu CL, Liang Q, He RL, Yau SST (2011). A novel method of characterizing genetic sequences: genome space with biological distance and applications. PLoS One, 6:e17293.

[10] Wikipedia: [https://en.wikipedia.org/wiki/Alignment-free\\_sequence\\_analysis#cite\\_note-2](https://en.wikipedia.org/wiki/Alignment-free_sequence_analysis#cite_note-2)

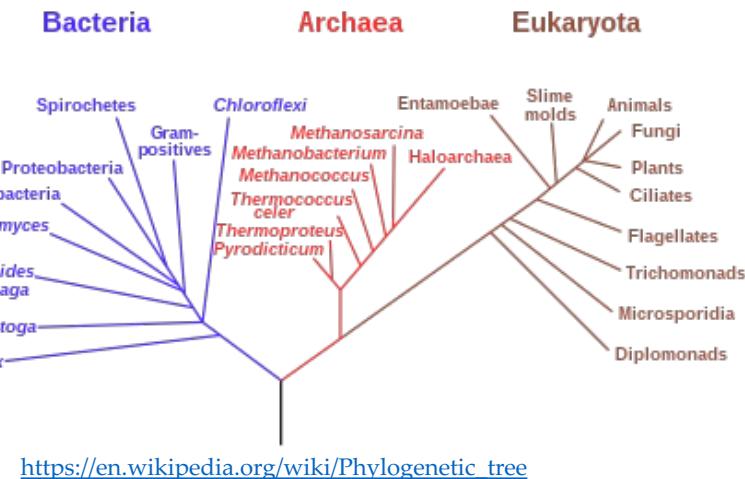




## Grand Biological Universe

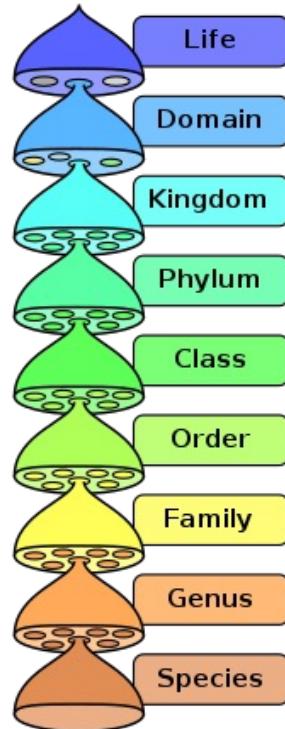


- With the progress of sequencing technology, more and more genomic sequences have been sequenced. According to National Genomics Data Center (NGDC, <https://ngdc.cncb.ac.cn/gwh/>), the number of genomes increases exponentially.
- What can we study as a bioinformatics researcher with a mathematical background?
- In 2011, after I introduced the natural vector method, I conceived the concept of **Grand Biological Universe** of life including virus and 7 kingdoms with cellular structures.
- Let us think about our own universe. You see the galaxies in the sky. Each galaxy has its position, so we want to make a clear positioning (e.g. longitude and latitude) for every organism in a proper mathematical space.
- The classical phylogenetic tree representation of organisms has no clear position of these organisms, which is unsatisfactory. The Grand Biological Universe will give position of every organism.
- We shall use our natural vector method and convex hull principle to construct this Grand Biological Universe.

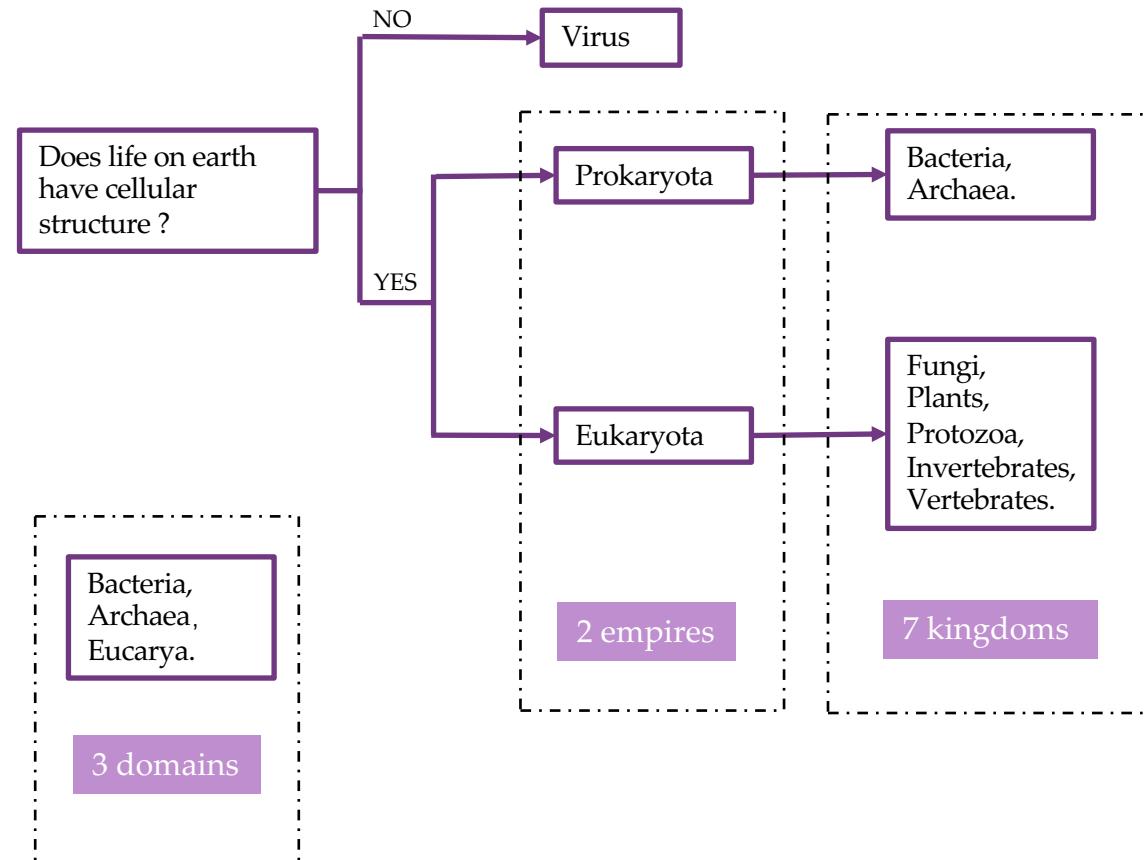




## Taxonomy (biology)



Furthermore



# Research Background: 23 problems proposed by DAPRA



1. The Mathematics of the Brain
2. The Dynamics of Networks
3. Capture and Harness Stochasticity in Nature
4. 21st Century Fluids
5. Biological Quantum Field Theory
6. Computational Duality
7. Occam's Razor in Many Dimensions
8. Beyond Convex Optimization
9. What are the Physical Consequences of Perelman's Proof of Thurston's Geometrization Theorem?
10. Algorithmic Origami and Biology
11. Optimal Nanostructures
12. The Mathematics of Quantum Computing, Algorithms, and Entanglement
13. Creating a Game Theory that Scales
14. An Information Theory for Virus Evolution
15. **The Geometry of Genome Space**
16. What are the Symmetries and Action Principles for Biology?
17. Geometric Langlands and Quantum Physics
18. Arithmetic Langlands, Topology, and Geometry
19. Settle the Riemann Hypothesis
20. Computation at Scale
21. Settle the Hodge Conjecture
22. Settle the Smooth Poincare Conjecture in Dimension 4
23. **What are the Fundamental Laws of Biology?**

- In 2008, the Defense Advanced Research Projects Agency (DAPRA) proposed 23 most important mathematical problems to be solved in the 21st century.
- Among them, the two major problems related to biomathematics are questions: “15. The Geometry of Genome Space” and “23. What are the Fundamental Laws of Biology”.



PART 02  
**Research Objective**



Construction of genome space.



Fundamental Principle of Molecular Biology: The Convex Hull Principle in Genome Space.

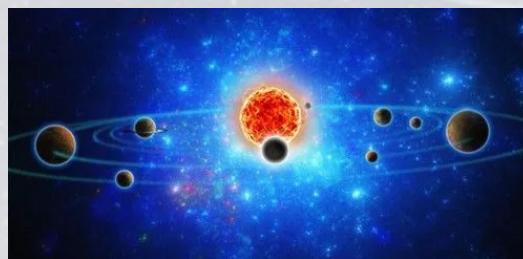


Natural Metric describing the geometry of genome space.



## Genome space

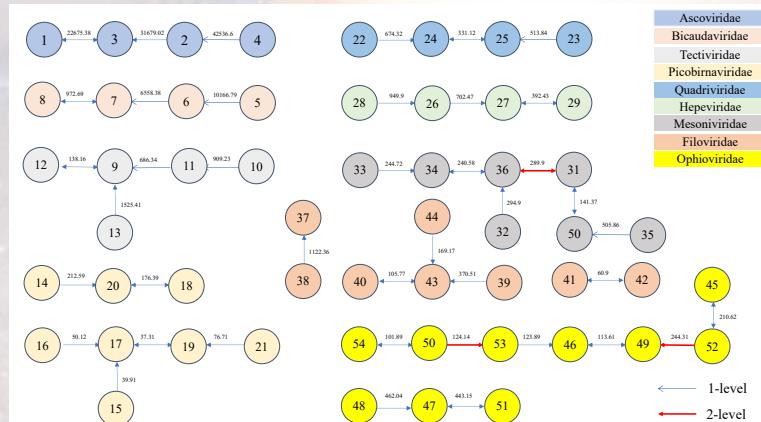
- A genome space consists of all known genomes and provides insights into their relationships, reflecting the important nature of the genomic universe.
  - Mathematically, the genome space can be considered to be the moduli space and constructed as a subspace in a high-dimensional Euclidean space. In this space, a genome sequence is uniquely represented as a point, and how sequences are arranged in the genome space is determined.
- Moduli space: Mathematically, every point in moduli space represents a equivalent class of this kind of algebraic objects.





## Natural metric

- Think about our own universe. We can see the galaxies in the sky. But there are things that we cannot see , for example the dark material or dark energy which contribute to gravitation force. According to Einstein theory, it will bend the space-time. So Euclidean distance may not correct.
- In order to find a proper natural metric to describe the geometric structure of the genome space, we need to understand what is the dark matter or dark energy in genome space. We propose that the k-mer ( $k>1$ ) and its distribution corresponds to dark matter and dark energy.



Natural graph based on the natural metric.

# Research Objective



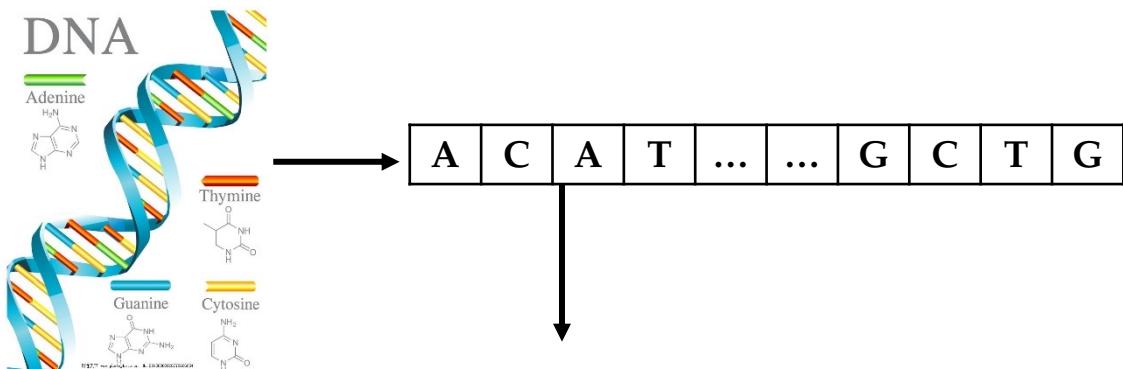
The genome space with a proper metric is a powerful means of determining the phylogenetics and classification of genomes.



PART 03  
**Methods & Materials**



## Natural Vector



- The counts of nucleotide A, C, G, T in S:  $(n_A, n_C, n_G, n_T)$
- The average location of letter A, C, G, T:  $(\mu_A, \mu_C, \mu_G, \mu_T)$
- the j-th central moment of position of letter A, C, G, T ( $j \geq 2$ ):  $(D_j^A, D_j^C, D_j^G, D_j^T)$

$(4 + 4 \cdot j)$ -dimensional Natural Vector:

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T, \dots, D_j^A, D_j^C, D_j^G, D_j^T)$$



## Natural Vector

Let  $S = s_1 s_2 s_3 \dots s_n$  be a genomic sequence of length  $n$ , and  $L = \{A, C, G, T/U\}$ . For  $k \in L$ , we define the indicator functions:  
 $w_k(\cdot): L \rightarrow \{0, 1\}$ , i.e.:

$$w_k(s_i) = \begin{cases} 1, & \text{if } s_i = k, \\ 0, & \text{otherwise.} \end{cases}$$

Where  $s_i \in L, i = 1, 2, 3, \dots, n$ .

- Let  $n_k = \sum_{i=1}^n w_k(s_i)$  denote the counts of nucleotide  $k$  in  $S$ .
- Let  $\mu_k = \sum_{i=1}^n i \frac{w_k(s_i)}{n_k}$  specify the average location of letter  $k$ .
- Let  $D_j^k = \sum_{i=1}^n \frac{(i - \mu_k)^j w_k(s_i)}{n_k^{j-1} n_{k-1}}$  be the  $j$ -th central moment of position of letter  $k$ .

Then we can get  $(4 + 4 \cdot j)$ -dimensional Natural Vector:

$$(n_A, n_C, n_G, n_T, \mu_A, \mu_C, \mu_G, \mu_T, D_2^A, D_2^C, D_2^G, D_2^T, \dots, D_j^A, D_j^C, D_j^G, D_j^T)$$



## Natural Vector

Here we give an example. If the genomic sequence is ACGGTAGTCC, the indicator functions are shown as follows:

Sequence	A	C	G	G	T	A	G	T	C	C
Position	1	2	3	4	5	6	7	8	9	10
$w_A(i)$	1	0	0	0	0	1	0	0	0	0
$w_C(i)$	0	1	0	0	0	0	0	0	1	1
$w_G(i)$	0	0	1	1	0	0	1	0	0	0
$w_T(i)$	0	0	0	0	1	0	0	1	0	0

The corresponding components of distribution vector are calculated as follows:

- $n_A = 2, n_C = 3, n_G = 3, n_T = 2.$
- $\mu_A = 1 \cdot \frac{1}{2} + 6 \cdot \frac{1}{2} = 3.5; \mu_C = 2 \cdot \frac{1}{3} + 9 \cdot \frac{1}{3} + 10 \cdot \frac{1}{3} = 7;$
- $\mu_G = 3 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 7 \cdot \frac{1}{3} = 4.67; \mu_T = 5 \cdot \frac{1}{2} + 8 \cdot \frac{1}{2} = 6.5.$
- $D_2^A = \frac{\left(1 - \frac{7}{2}\right)^2}{2 \cdot 10} + \frac{\left(6 - \frac{7}{2}\right)^2}{2 \cdot 10} = 0.63;$
- $D_2^C = \frac{(2-7)^2}{3 \cdot 10} + \frac{(9-7)^2}{3 \cdot 10} + \frac{(10-7)^2}{3 \cdot 10} = 1.27;$
- $D_2^G = \frac{\left(3 - \frac{14}{3}\right)^2}{3 \cdot 10} + \frac{\left(4 - \frac{14}{3}\right)^2}{3 \cdot 10} + \frac{\left(7 - \frac{14}{3}\right)^2}{3 \cdot 10} = 0.29;$
- $D_2^T = \frac{\left(5 - \frac{13}{2}\right)^2}{2 \cdot 10} + \frac{\left(8 - \frac{13}{2}\right)^2}{2 \cdot 10} = 0.23;$

Then the 12-dimensional Natural Vector is: (2,3,3,2,3.5,7,4.67,6.5,0.63,1.27,0.29,0.23).



## Natural Vector

Input Data:

	Length
HIV	About 8000bp
Escherichia coli	About 5Mbp
Bird	About 900Mbp
Homo sapiens	About 3000Mbp

Output Data:

	Output dimension
Alignment methods	The length of sequences
Natural vector	12-dimension

For example:

48 modern birds	Time
Alignment method	About 1-2 years by 9 supercomputing centers
Natural vector	About 4 days by a small server with 384GB

The calculation speed is improved, and the output data is normalized to the same dimension.

- Genome can be abstractly viewed as a sequence of nucleotides A, C, G and T. The distribution of these nucleotides uniquely determines the genome. We use the information of nucleotides distributions within the virus genome (which we call it as natural vector) to canonically embed this virus genome as a point in Euclidean space. This can be viewed as the discrete analog of Kodaira embedding.
- Any sequence can be represented as a point in this space, then different genomes can be compared. Genomes with similar nucleotides distributions lie closely.



## k-mer

- If a genome sequence is  $S = s_1s_2s_3 \dots s_N$ ,  $s_i \in \{A, C, G, T/U\}$ .
- k-mer  $l_i$  ( $i = 1, 2, \dots, 4^k$ ) is a segment of length k.
  - For example:
    - 1-mers indicate A, C, G, T;
    - 2-mers include AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT;
    - 3-mers include AAA, AAC, AAG, AAT, ACA, ACC, ACG, ACT, AGA, AGC, AGG, AGT, ATA, ATC, ATG, ATT, CAA, CAC, CAG, CAT, CCA, CCC, CCG, CCT, CGA, CGC, CGG, CGT, CTA, CTC, CTG, CTT, GAA, GAC, GAG, GAT, GCA, GCC, GCG, GCT, GGA, GGC, GGG, GGT, GTA, GTC, GTG, GTT, TAA, TAC, TAG, TAT, TCA, TCC, TCG, TCT, TGA, TGC, TGG, TGT, TTA, TTC, TTG, TTT
  - For each given k, the number of k-mer is fixed. According to combinatorial mathematics, k-mers consist of  $4^k$  sequence segments.



## k-mer Natural Vector

Suppose that  $l_i[j]$  is the location of the  $j$ -th occurrence of a k-mer  $l_i$  in  $S$  ( $i = 1, 2, \dots, 4^k$ ), the distributions of a k-mer  $l_i$  can be described by three components:

- $n_{l_i}$  denotes the counts of k-mer  $l_i$  in  $S$ ;
- $\mu_{l_i} = \sum_{j=1}^{n_{l_i}} \frac{l_i[j]}{n_{l_i}}$  specifies the average location of k-mer  $l_i$ ;
- $D_{l_i}^m = \sum_{m=1}^{n_{l_i}} \frac{(l_i[j] - \mu_{l_i})^m}{n_{l_i}^{m-1} (n-k+1)^{m-1}}$  ( $m = 2, \dots, n, \dots$ ) is the m-order central moment of emergence position of letter k-mer  $l_i$ .

Then the k-mer Natural Vector with high order central moment (k-mer NVH) for sequence  $S$  is defined by:

$$(n_{l_1}, \dots, n_{l_{4^k}}, \mu_{l_1}, \dots, \mu_{l_{4^k}}, D_{l_1}^2, \dots, D_{l_{4^k}}^2, \dots, D_{l_1}^n, \dots, D_{l_{4^k}}^n).$$

The dimension of central moment is  $4^k \cdot (n - 1)$ , and the dimension of counts and average location of k-mers are both  $4^k$ , so the complete k-mer dimensional natural vector is  $4^k \cdot (n + 1)$ -dimensional.

k-mer Natural Vector with second central moment (k-mer NVS) has been verified to be enough to represent the sequence and satisfies one-to-one mapping and it is  $4^k \cdot 3$ -dimensional:

$$(n_{l_1}, \dots, n_{l_{4^k}}, \mu_{l_1}, \dots, \mu_{l_{4^k}}, D_{l_1}^2, \dots, D_{l_{4^k}}^2).$$



## k-mer Natural Vector

Here we give an example. If the genomic sequence is ACATACTG, the 2-mer sequences and their positions are as follows:

2-mer	AC	CA	AT	TA	AC	CT	TG
position	1	2	3	4	5	6	7

The corresponding components of distribution vector are calculated as follows:

- $n_{AC} = 2, n_{CA} = n_{AT} = n_{TA} = n_{CT} = n_{TG} = 1$
- $\mu_{AC} = \frac{1+5}{2} = 3, \mu_{CA} = 2, \mu_{AT} = 3, \mu_{TA} = 4, \mu_{CT} = 6, \mu_{TG} = 7$
- $D_2^{AC} = \frac{(1-3)^2 + (5-3)^2}{7 \times 2} = \frac{8}{14}, D_2^{CA} = D_2^{AT} = D_2^{TA} = D_2^{CT} = D_2^{TG} = 0$

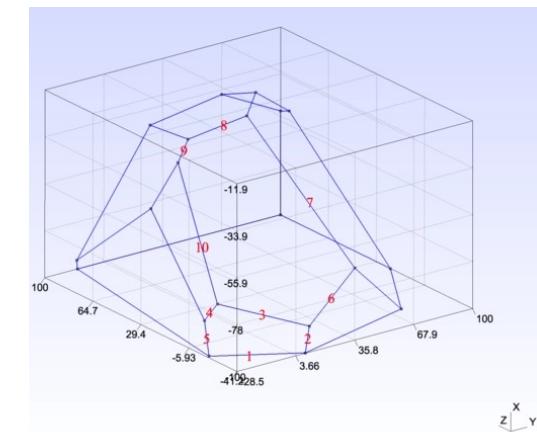
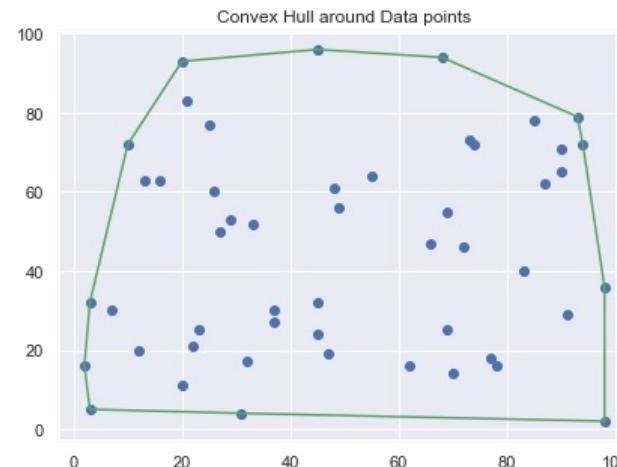
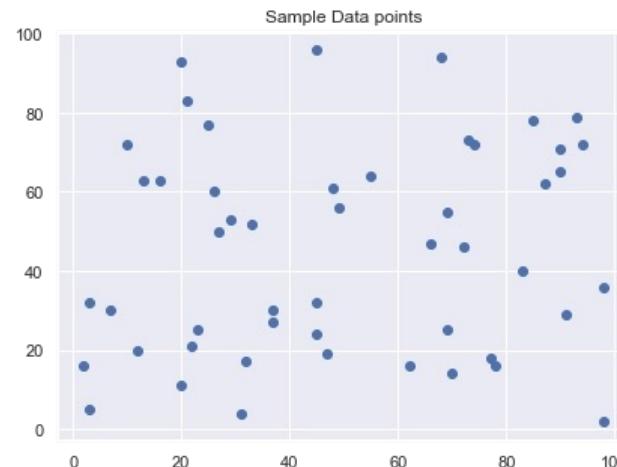
Then the 2-mer Natural Vector is  $4^2 \cdot 3$ -dimensional vector.



## Convex hull

- Convex hull is one of the most fundamental concepts in computational geometry.
- Mathematically, the convex hull of a point set  $C = \{x_1, x_2, \dots, x_k\}, x_i \in R^n$  is the minimal convex set that contains these points.
- By the concept of convex combinations, the convex hull of a finite point set C is equivalently defined as the set of all convex combinations of points in C:

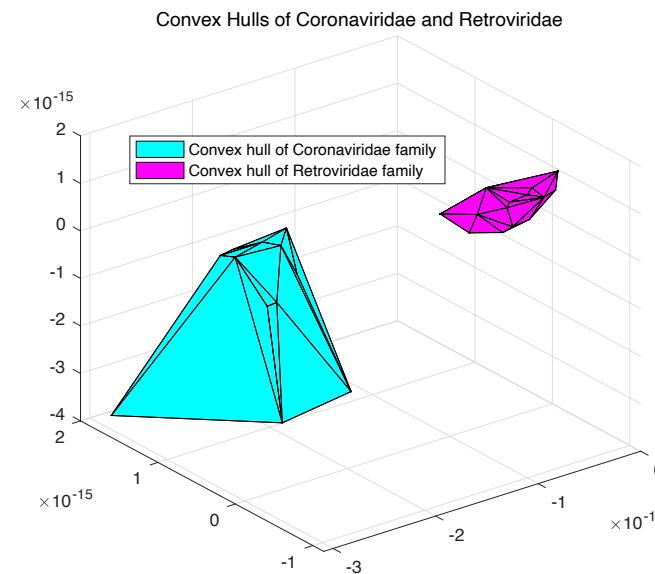
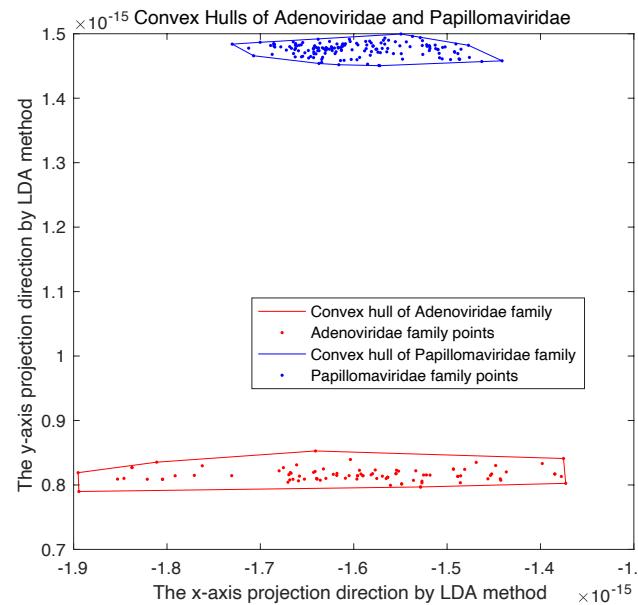
$$convC = \{\theta_1x_1 + \theta_2x_2 + \dots + \theta_kx_k, x_i \in C, \theta_1 + \theta_2 + \dots + \theta_k = 1, \theta_i \geq 0, i = 1, 2, \dots, k\}.$$





## Convex hull principle of molecular biology

- In this study,  $x_i$  is the natural vector and we propose a convex hull principle of molecular biology, pointing out that convex hulls corresponding to different biological group (family or genera) do not overlap with each other.
- In this way each biological group corresponds to a point cloud, which reflects the genetic variety of this biological group .

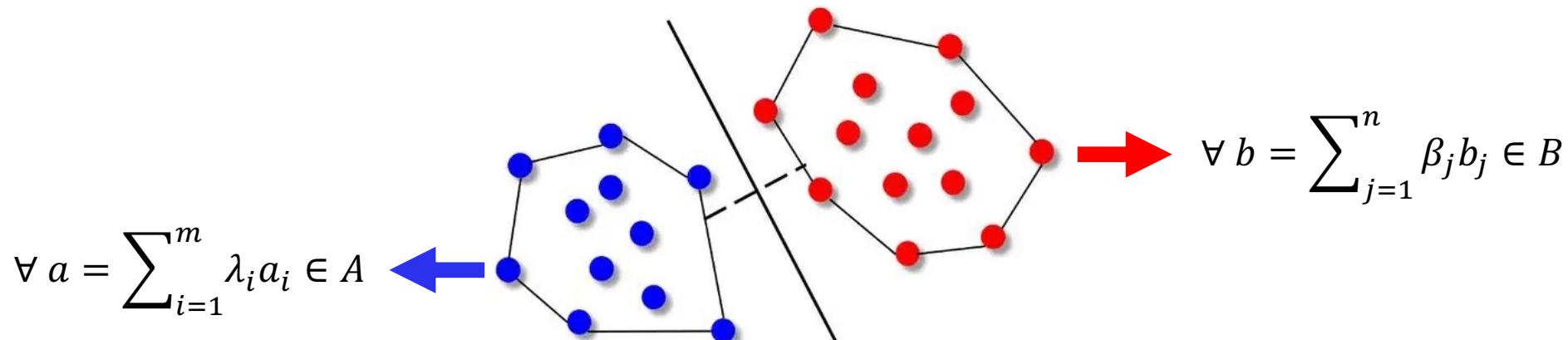




## Optimization method

□ We use optimization method to determine the separateness of two convex convex hull:

- If A is the convex hull of point set  $\{a_1, a_2, \dots, a_m\}$ , and B is the convex hull of point set  $\{b_1, b_2, \dots, b_n\}$ . The mathematical principle is that if A and B intersect, then  $\sum_{i=1}^m \lambda_i a_i = \sum_{j=1}^n \beta_j b_j$ , where  $\sum_{i=1}^m \lambda_i = 1, \lambda_i \geq 0, i = 1, 2, \dots, m, \sum_{j=1}^n \beta_j = 1, \beta_j \geq 0, j = 1, 2, \dots, n, a_i, b_j \in R^k$ .



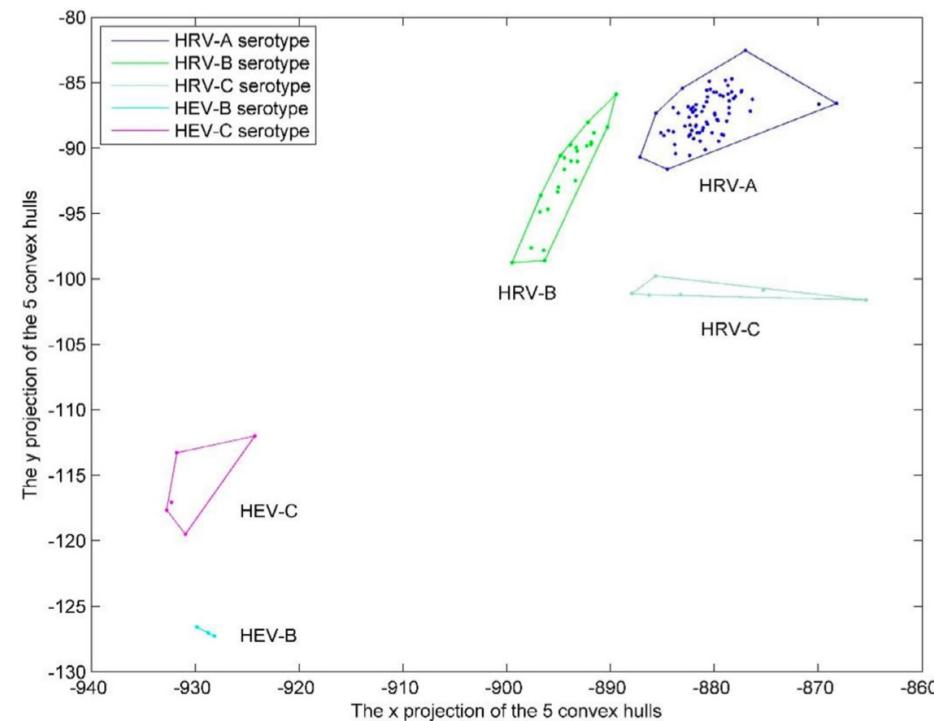
- It can be transformed a optimization problem: if there exists non-zero coefficients  $\{\lambda_1, \lambda_2, \dots, \lambda_m, \beta_1, \beta_2, \dots, \beta_n\}$  in feasible domain such that the minimum value of the following optimization problem is 0, then A and B intersect:

$$\begin{aligned}
 & \min \left| \left| \sum_{i=1}^m \lambda_i a_i - \sum_{j=1}^n \beta_j b_j \right| \right| \\
 & s.t. \quad \sum_{i=1}^m \lambda_i = 1, \\
 & \quad \lambda_i \geq 0, i = 1, 2, \dots, m \\
 & \quad \sum_{j=1}^n \beta_j = 1, \\
 & \quad \beta_j \geq 0, j = 1, 2, \dots, n
 \end{aligned}$$



## Convex hull analysis

For a certain category of viruses, the sequences can be converted into a series of points in a high-dimensional Euclidean space by natural vector method. The convex hull of the points can be constructed, and it can be projected into two dimensions, as shown in the right figure.



The two-dimensional projection of the convex hulls composed of 3 HRV and 2 HEV serotypes. The convex hulls of the five families are mutually disjoint.



## Dataset: Virus & Bacteria & Plant

	Database access link and description	Data filtering	Sequence Information	Date
Virus	RefSeq: <a href="http://ftp.ncbi.nlm.nih.gov/genomes/Viruses">http://ftp.ncbi.nlm.nih.gov/genomes/Viruses</a>	We removed three types of sequences: <ul style="list-style-type: none"> <li>• Viruses without Baltimore class label;</li> <li>• Viruses without family label;</li> <li>• Families including one or two sequences.</li> </ul>	<input type="checkbox"/> 7382 sequences <ul style="list-style-type: none"> <li>• 83 families</li> <li>• 304 genera</li> <li>• 7 Baltimore classes</li> </ul>	March 2020
Bacteria	RefSeq: <a href="https://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/">https://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/</a>	We removed two types of sequences: <ul style="list-style-type: none"> <li>• Sequences with degenerate bases;</li> <li>• Families <u>including one sequence</u>.</li> </ul>	<input type="checkbox"/> 23409 complete genomes <ul style="list-style-type: none"> <li>• 313 families</li> </ul>	March 2022
Plant	RefSeq: <a href="https://ftp.ncbi.nlm.nih.gov/refseq/release/">https://ftp.ncbi.nlm.nih.gov/refseq/release/</a>	We removed two types of sequences: <ul style="list-style-type: none"> <li>• Sequences with degenerate bases;</li> <li>• Families including one sequence.</li> </ul>	<input type="checkbox"/> 5934 complete genomes: <ul style="list-style-type: none"> <li>• 216 families;</li> </ul>	March 2022

Degenerate base symbol:

Symbol	R	Y	M	K	S	W	H	B	V	D	N
Bases represented	A/G	C/T	A/C	G/T	G/C	A/T	A/T/C	G/T/C	G/A/C	G/A/T	A/T/C/G



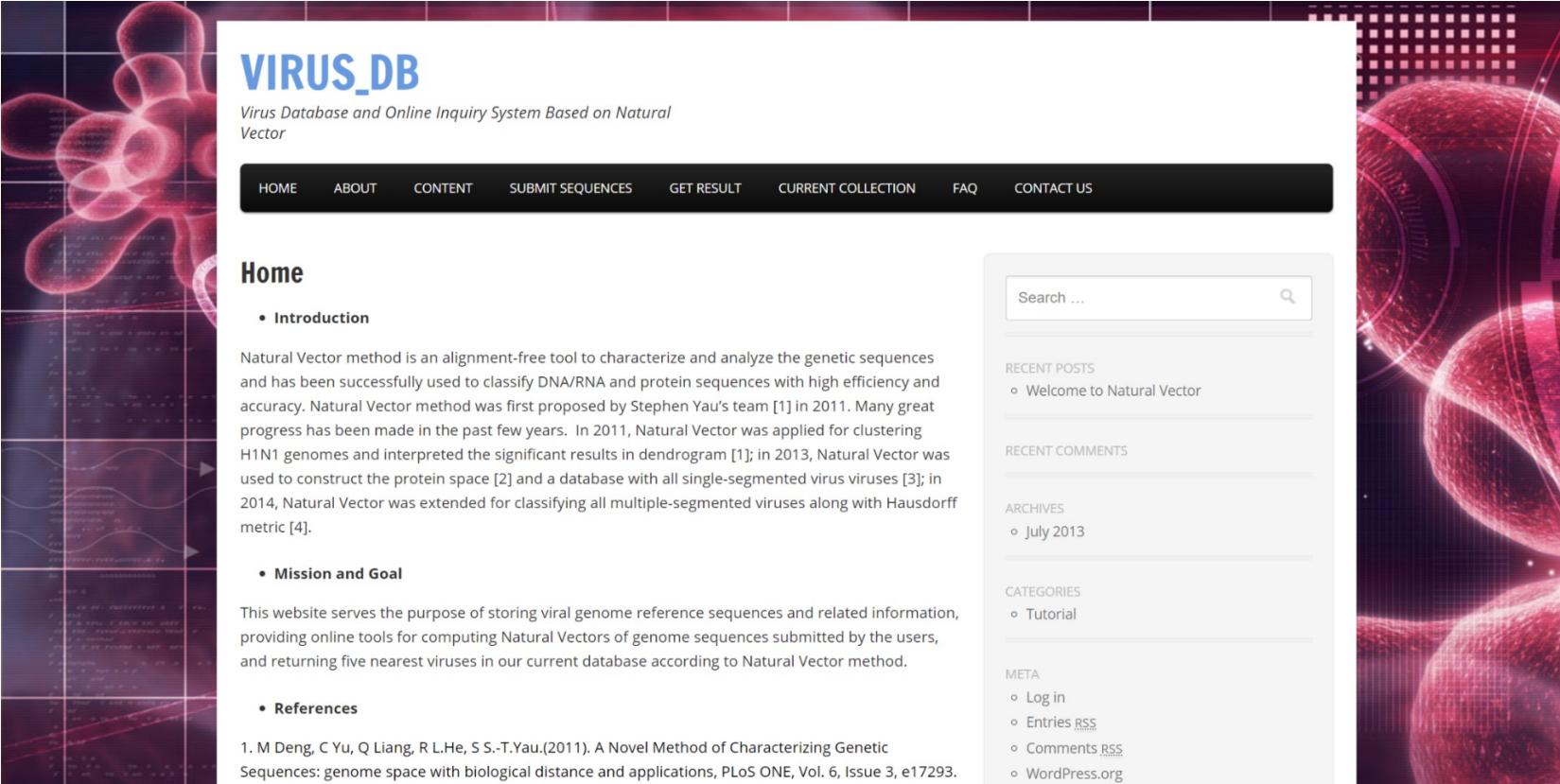
PART 04

## Results

### Virus Database



## Update the virus database based on natural vector method: VirusDB



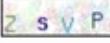
The screenshot shows the homepage of the VirusDB website. The background features a dark, abstract design with red glowing spheres and lines. The header contains the text "VIRUS\_DB" and "Virus Database and Online Inquiry System Based on Natural Vector". Below the header is a navigation bar with links: HOME, ABOUT, CONTENT, SUBMIT SEQUENCES, GET RESULT, CURRENT COLLECTION, FAQ, and CONTACT US. The main content area is titled "Home" and includes sections for "Introduction", "Mission and Goal", and "References". The "Introduction" section describes the Natural Vector method and its applications. The "Mission and Goal" section explains the purpose of the website. The "References" section lists a single publication: "M Deng, C Yu, Q Liang, R LHe, S S-T.Yau.(2011). A Novel Method of Characterizing Genetic Sequences: genome space with biological distance and applications, PLoS ONE, Vol. 6, Issue 3, e17293." To the right of the main content is a sidebar with a search bar, recent posts (including "Welcome to Natural Vector"), recent comments, archives (July 2013), categories (Tutorial), and meta-links (Log in, Entries RSS, Comments RSS, WordPress.org).

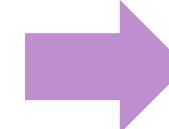
<http://yaulab.math.tsinghua.edu.cn/VirusDB/>

# Virus Database

## Submission

Reply

	<input checked="" type="radio"/> Single Segment <input type="radio"/> Multiple Segments
Email	serenadong1993@outlook.com
Database Version	15.1: 2017-04-14
Baltimore	-Choose one from the list-
Family	-Choose one from the list-
Sequence	>NC_012636.1 Aedes aegypti densovirus 2 strain 0814616, complete genome TATAAGTCATATTCCATATAAGAAATTATTCGTGATACGGATACTGTAAGATACAGTTCTATTAG AAACGATGTATTACATCTGTATCTTACAGTATCCGTACAGAAATAATTTTTATGGATTATGGA CTTATATCAAATTCTATATGGATCACTGGAGGTGGAAAATAAGGGAAAACATAAGGCCGAAATTAACT TATTCTCACACACAAATAAACCTAATTCCACTACCACATGGTCCACCCCTATAAAGGAGTACAA AGGAGAGGCCAATCGAGTAATGAATTCTAGTCTGTTTGAAACATTGCCGTGTGAACACGGAAACCTATAT TGTGAGTGCATATATTGTTGGAGCATGACAGGCCAGTGCAGGGGAAAAAACTGGATTGGGAGAATCAA CTGGAATCGAAGGAAGACTGGCAACGATAACCAACCAACCCAGGGCTCTCAGATTTATTCGACCGAGAC AATACATTTGCAACTACAGTACCGGAAAGAAGTCATCAATCGAGAAAGATTACGTCAAGGATTTCGCT GGTCAAACCGTTGGTGACCTCTACCCACAATTACAAGGCAGCACCGGAGCCTCTGAACCAATTGATTG ... ...
Fasta File	选择文件 未选择任何文件
Verification code	ZSVP 
	*Click the picture again if you can't see it clearly
	<input type="button" value="Submit"/> <input type="button" value="Reset"/>
	*There is no distinction between the lower case letters and capital letters



<http://yaulab.math.tsinghua.edu.cn/VirusDB/>

Dear serenadong1993@outlook.com,

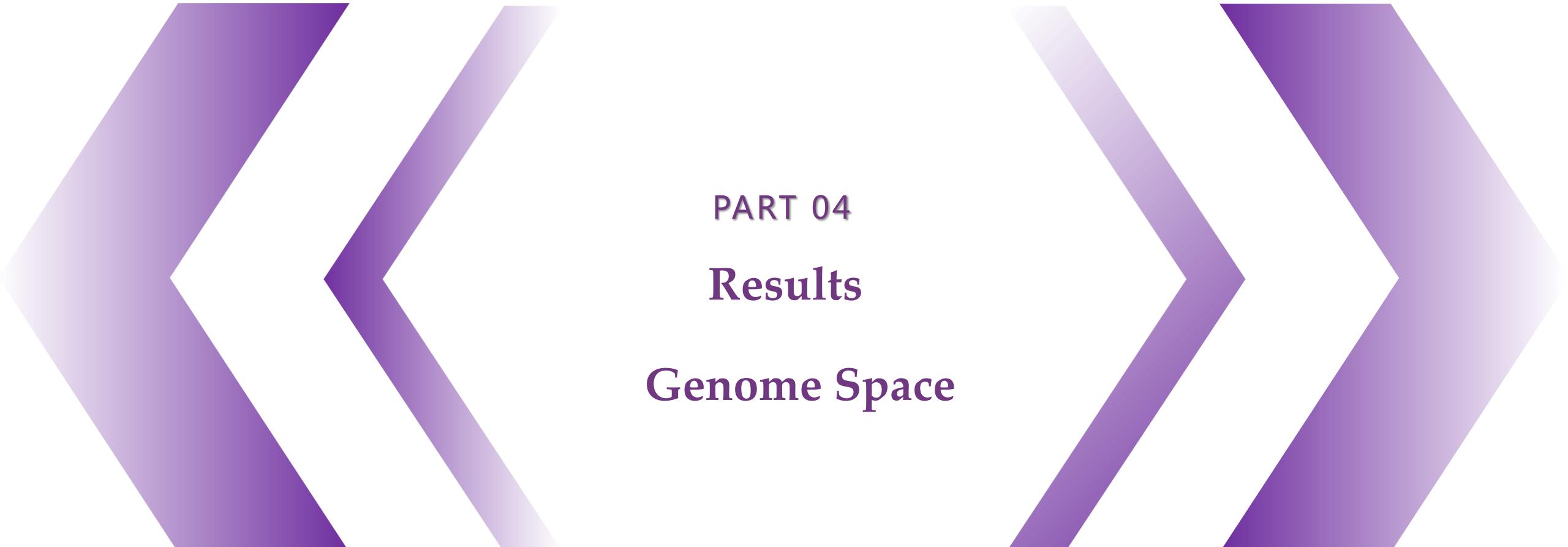
With refer to your calculation request (Reference Number : 3EIVQK) submitted on 2017-05-31 22:13:50, we would like to reply you the calculated result.

Based on current version of Database, we predict the sequence that you have submitted belongs to Baltimore = 'II' and Family = 'Parvoviridae'. The following are the five closest viruses in our data set.

Neighbor Index : 1  
Virus Name : Aedes\_aegypti\_densovirus\_uid37821  
Virus Order :  
Baltimore : II  
Family : Parvoviridae  
SubFamily : Densovirinae  
Genus : Brevidensovirus  
Accession : NC\_012636  
GI : 229342011  
GenPar1 : ssDNA viruses  
GenPar2 :  
Shape : linear  
Neighbor Distance : 2.00096432327775  
---

Neighbor Index : 2  
Virus Name : Mosquito\_densovirus\_BR\_07\_uid62639  
Virus Order :  
Baltimore : II  
Family : Parvoviridae  
SubFamily : Densovirinae  
Genus : unclassified Brevidensovirus  
Accession : NC\_015115  
GI : 322688186  
GenPar1 : ssDNA viruses  
GenPar2 :  
Shape : linear  
Neighbor Distance : 150.206534660969  
---

Neighbor Index : 3  
Virus Name :  
Infectious\_hypodermal\_and\_hematopoietic\_necrosis\_virus\_uid14436  
Virus Order :  
Baltimore : II  
Family : Parvoviridae  
SubFamily : Densovirinae  
Genus : Penstyldensovirus  
Accession : NC\_002190  
GI : 294441960  
GenPar1 : ssDNA viruses  
GenPar2 :  
Shape : linear  
Neighbor Distance : 354.699429615121  
(Excerpt)

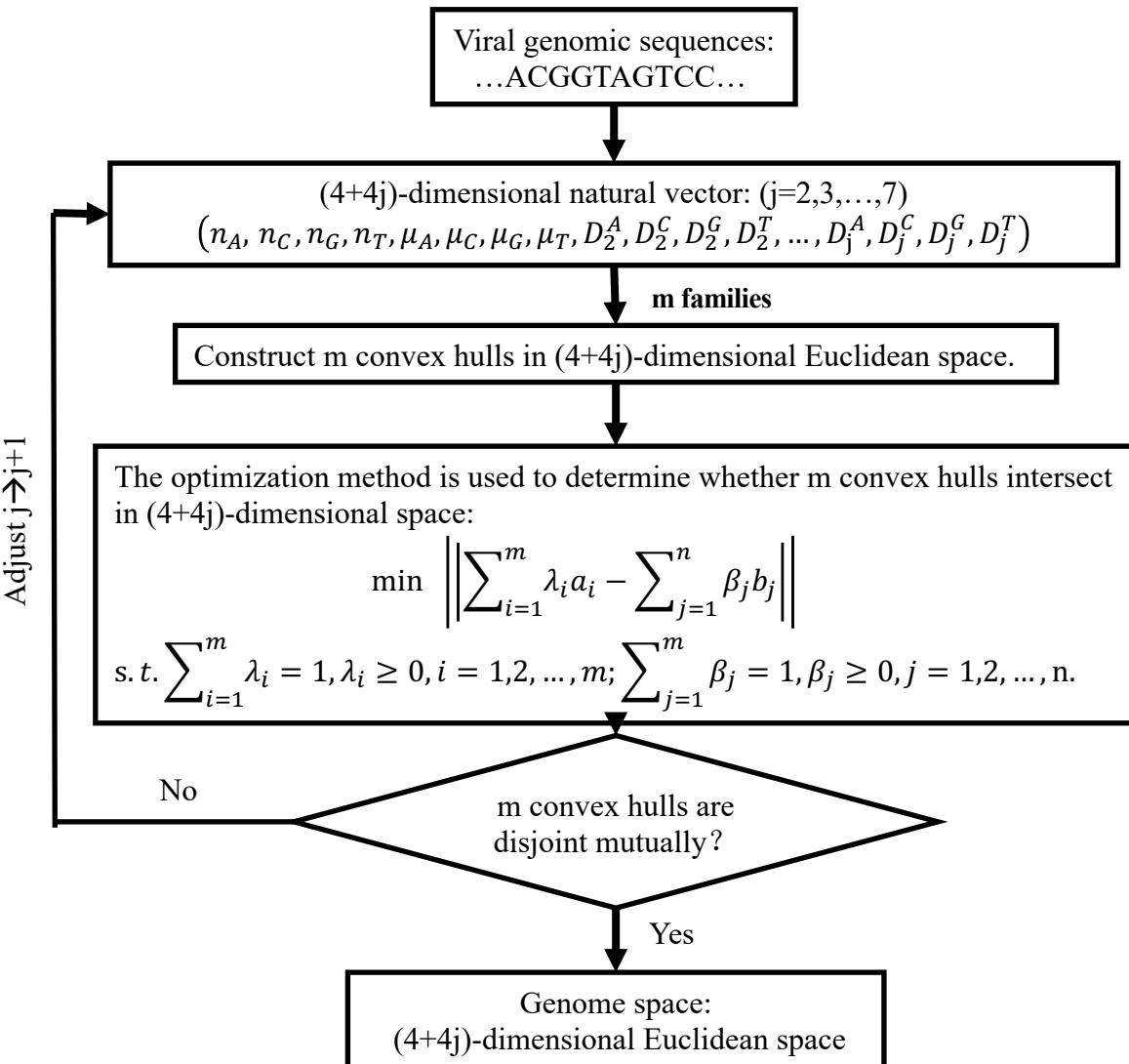


PART 04

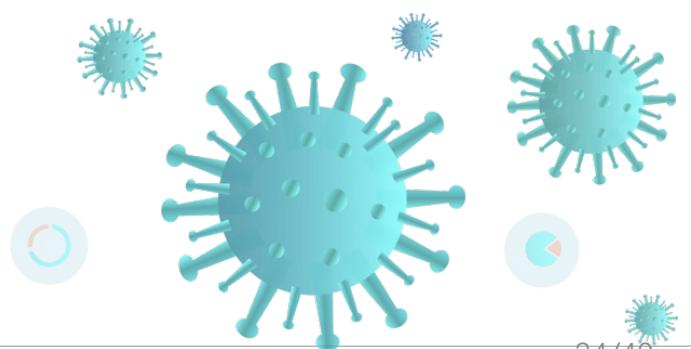
## Results

# Genome Space

# The genome space of virus



According to the downloaded viral genome sequence data, the  $(4+4j)$ -dimensional natural vector of each sequence is calculated (starting from  $j=2$ ), and the convex hulls of different families are established. We can determine whether all convex hulls in the  $(4+4j)$ -dimensional space are disjoint through optimization method. We repeat the judgment step until all the convex hulls are disjoint. Then the viral genome space is sitting in the  $(4+4j)$ -dimensional Euclidean space, and the convex hull principle of viral genomes holds in  $R^{4+4j}$ .



# The convex hull principle of viral genomes holds in $R^{32}$



**Table:** The number of disjoint convex hull pairs changes with the increase in the dimension of the Euclidean space. Total convex hull pairs of 83 families are 3404. When the dimension of the natural vector is more than 32 ( $j \geq 7$ ), there are no intersecting convex hull pairs. According to the definition of embedding dimension of the moduli space, we chose the space with the lowest dimension, which indicates that the viral genome space is sitting in a 32-dimensional Euclidean space.

Euclidean space	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9
	$R^{12}$	$R^{16}$	$R^{20}$	$R^{24}$	$R^{28}$	$R^{32}$	$R^{36}$	$R^{40}$
No. of disjoint convex hull pairs	3221	3291	3338	3354	3395	3403	3403	3403
No. of intersectant convex hull pairs	182	112	65	49	8	0	0	0

# The convex hull principle of bacterial genomes holds in $R^{48}$



**Table:** The number of disjoint convex hull pairs changes with the increase in the dimension of the Euclidean space. Total convex hull pairs of 313 families are 48828. When the dimension of the natural vector is no less than 48 ( $j \geq 11$ ), there are no intersecting convex hull pairs. We choose the space with the lowest dimension, which indicates that the bacterial genome space is sitting in a 48-dimensional Euclidean space.

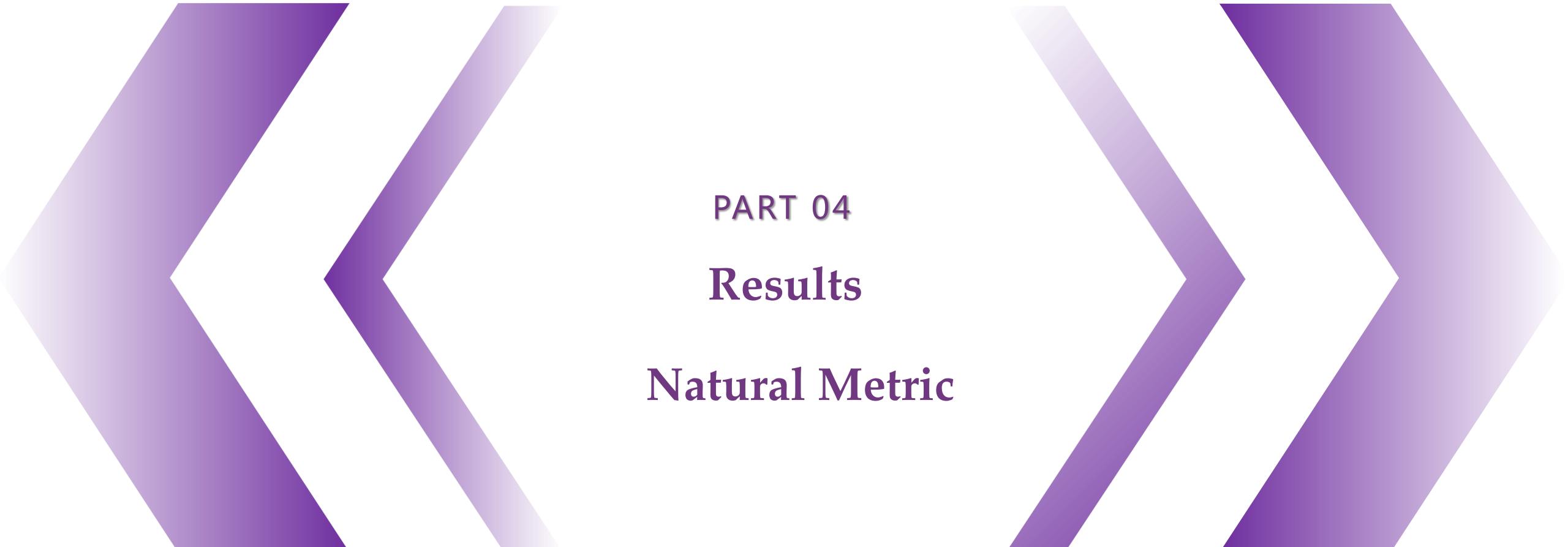
Euclidean Space	$R^{12}$ $j=2$	$R^{16}$ $j=3$	$R^{20}$ $j=4$	$R^{24}$ $j=5$	$R^{28}$ $j=6$	$R^{32}$ $j=7$	$R^{36}$ $j=8$	$R^{40}$ $j=9$	$R^{44}$ $j=10$	$R^{48}$ $j=11$	$R^{52}$ $j=12$	$R^{56}$ $j=13$
No. of disjoint convex hull pairs	47032	47660	48213	48374	48483	48551	48785	48825	48826	48828	48828	48828
No. of intersectant convex hull pairs	1796	1168	615	454	345	277	43	3	2	0	0	0

# The convex hull principle of plant genomes holds in $R^{24}$



**Table:** The number of disjoint convex hull pairs changes with the increase in the dimension of the Euclidean space. Total convex hull pairs of 439 families are 96141. When the dimension of the natural vector is no less than 24 ( $j \geq 5$ ), there are no intersecting convex hull pairs. We choose the space with the lowest dimension, which indicates that the plant genome space is sitting in a 24-dimensional Euclidean space.

Euclidean Space	$R^{12}$ $j=2$	$R^{16}$ $j=3$	$R^{20}$ $j=4$	$R^{24}$ $j=5$	$R^{28}$ $j=6$	$R^{32}$ $j=7$	$R^{36}$ $j=8$	$R^{40}$ $j=9$	$R^{44}$ $j=10$
No. of disjoint convex hull pairs	95883	96059	96092	96141	96141	96141	96141	96141	96141
No. of intersectant convex hull pairs	258	82	49	0	0	0	0	0	0



PART 04

## Results

### Natural Metric

# Natural metric in viral genome space



- L1-norm
- L2-norm
- $d = d_1 + \frac{1}{2^2}d_2 + \frac{1}{3^2}d_3 + \dots + \frac{1}{n^2}d_n$
- $d = d_1 + \frac{1}{2}d_2 + \frac{1}{2^2}d_3 + \dots + \frac{1}{2^{n-1}}d_n$
- $d = a_1d_1 + a_2d_2 + a_3d_3 + \dots + a_nd_n$ , which has the best coefficients  $a_i$  for the natural metric, maybe the natural metric in the genome space

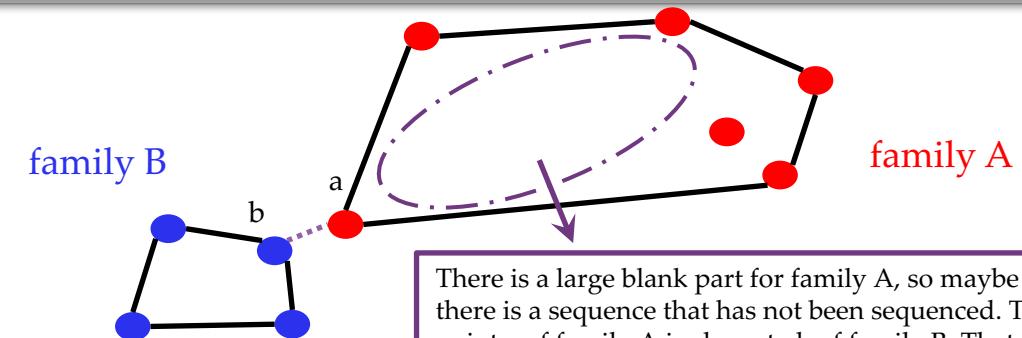


Table: The nearest neighborhood classification accuracies of virus family based on the new natural metric for different n. For weight  $\frac{1}{2^k}$  ( $d = \sum_{k=1}^n \frac{1}{2^{k-1}}d_k$ ), the classification is more accurate with the increase in n. For weight  $\frac{1}{k^2}$  ( $d = \sum_{k=1}^n \frac{1}{k^2}d_k$ ) the accuracy decreases when n=9, indicating that this definition is unstable. The natural metric is defined as  $d = d_1 + \frac{1}{2}d_2 + \frac{1}{2^2}d_3 + \dots + \frac{1}{2^{n-1}}d_n$ .

Weight	n	1	2	3	4	5	6	7	8	9
$\frac{1}{2^k}$	Accuracy	79.9%	82.8%	83.3%	83.3%	84.1%	85.8%	86.9%	87.4%	88.3%
$\frac{1}{k^2}$	Accuracy	79.9%	82.8%	83.3%	83.3%	84.4%	86.3%	87.7%	88.0%	85.6%

L2-norm

# Natural metric in bacterial genome space



Table. The nearest neighborhood classification accuracies of **bacterial** family based on the k-mer natural vector for different k. And **the accuracy is the greatest (87.28%) when k = 3 and L2-distance (that is  $D_3 = \frac{1}{1.5} d_1 + \frac{1}{1.5^2} d_2 + \frac{1}{1.5^3} d_3$ ).**

		$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$
		78.33%	85.00%	82.04%	81.44%	82.39%	81.61%	80.11%	79.21%	82.08%
L2-norm	Weight	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	$D_6$	$D_7$	$D_8$	$D_9$
	$\frac{1}{2^k}$	85.00%	85.37%	87.07%	86.92%	86.87%	86.86%	86.82%	86.80%	86.80%
	$\frac{1}{k^2}$	85.00%	85.38%	86.66%	86.54%	86.52%	86.48%	86.51%	86.53%	86.51%
	$\frac{1}{1.5^k}$	85.00%	85.29%	87.28%	86.93%	86.82%	86.68%	86.62%	86.51%	86.58%

- $d_k$  is the L2-distance of k-mer natural vector with second central moment.
- $D_n = d_1 + \frac{1}{2}d_2 + \frac{1}{2^2}d_3 + \dots + \frac{1}{2^{n-1}}d_n$  or  $D_n = d_1 + \frac{1}{2^2}d_2 + \frac{1}{3^2}d_3 + \dots + \frac{1}{n^2}d_n$  or  $D_n = \frac{1}{1.5}d_1 + \frac{1}{1.5^2}d_2 + \frac{1}{1.5^3}d_3 + \dots + \frac{1}{1.5^n}d_n$ .

# Natural metric in plant genome space

Table. The nearest neighborhood classification accuracies of **plant** family based on different metrics. For direct L1-distance or L2-distance of k-mer NVs, the accuracy is the greatest when  $k = 8$  (0.9564 for L1-distance, 0.9543 for L2-distance). For both weight  $\frac{1}{2^k} \left( D_n = d_1 + \frac{1}{2} d_2 + \frac{1}{2^2} d_3 + \dots + \frac{1}{2^{n-1}} d_n \right)$  and weight  $\frac{1}{k^2} \left( D_n = d_1 + \frac{1}{2^2} d_2 + \frac{1}{3^2} d_3 + \dots + \frac{1}{n^2} d_n \right)$ , the accuracy increases with the increase of n (Both L1-distance and L2-distance), and **the accuracy is the greatest (0.9617) when  $k = 9$  and L1-distance (that is  $D_9 = d_1 + \frac{1}{2^2} d_2 + \frac{1}{3^2} d_3 + \dots + \frac{1}{9^2} d_9$ )**.

		L1-24d NVH	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$
L1-norm	Weight	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	$D_6$	$D_7$	$D_8$	$D_9$	
		0.7986	0.7986	0.8739	0.8953	0.9154	0.9309	0.9422	0.9498	0.9564	0.9542
		$\frac{1}{2^k}$	0.7986	0.8652	0.8879	0.9085	0.9230	0.9368	0.9476	0.9569	0.9617
		$\frac{1}{k^2}$	0.7986	0.8561	0.8832	0.9068	0.9238	0.9383	0.9488	0.9575	0.9604
		L2-24d NVH	$d_1$	$d_2$	$d_3$	$d_4$	$d_5$	$d_6$	$d_7$	$d_8$	$d_9$
L2-norm	Weight	$D_1$	$D_2$	$D_3$	$D_4$	$D_5$	$D_6$	$D_7$	$D_8$	$D_9$	
		0.7983	0.7983	0.8682	0.8913	0.912	0.9311	0.9387	0.9464	0.9543	0.9538
		$\frac{1}{2^k}$	0.7983	0.8552	0.8783	0.8953	0.9063	0.9221	0.9355	0.943	0.9481
		$\frac{1}{k^2}$	0.7983	0.8443	0.8694	0.8773	0.8787	0.9267	0.9398	0.9491	0.9572

- $d_k$  is the L1-distance or L2-distance of k-mer natural vector with second central moment.
- $D_n = d_1 + \frac{1}{2} d_2 + \frac{1}{2^2} d_3 + \dots + \frac{1}{2^{n-1}} d_n$  or  $D_n = d_1 + \frac{1}{2^2} d_2 + \frac{1}{3^2} d_3 + \dots + \frac{1}{n^2} d_n$ .
- “L1-24d NVH” means the L1-distance of 24-dimensional natural vector with five-order central moment;
- “L2-24d NVH” means the L2-distance of 24-dimensional natural vector with five-order central moment.

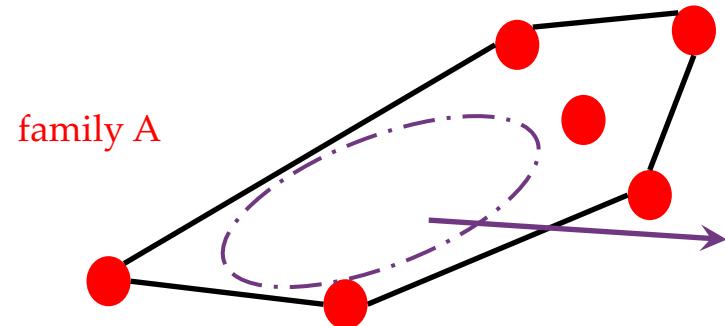


## PART 05

# Application

- A. Find points with biological significance in convex hull
- B. The early transmission of SARS-CoV-2

# A. Find points with biological significance in convex hull



There is a large blank part for family A, so maybe there is a sequence that has not been sequenced. An important problem is to find points with biological significance in convex hull.

- It is a challenging problem how to find existing but undiscovered genome sequence or predict potential genome sequence mutations based on real sequence data.
  - A genome sequence can be represented as a point in a finite dimensional space, and each family corresponds to a convex hull.
- Motivated by this, we developed approaches to detect new, undiscovered genome sequences in convex hulls by exploring points with biological significance in convex hull. For example, our heuristic algorithms, Random-permutation Algorithm with Penalty (RAP) and Random-permutation Algorithm with Penalty and COs-trained Search (RAPCOS), et al [1, 2].

[1] New Genome Sequence Detection via Natural Vector Convex Hull Method (with Ruzhang Zhao and Shaojun Pei), IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.19 (2022), 1782-1793.

[2] Determination of the nucleotide or amino acid composition of genome or protein sequences by using natural vector method and convex hull principle (with Xiaopei Jiao\*, Shaojun Pei\*, Zeju Sun and Jiayi Kang), Fundamental Research, Vol. 1 (2021), 559-564.

# B. The early transmission of SARS-CoV-2

## Background

- Currently, SARS-CoV-2 was reported in Wuhan first. However, many studies have detected SARS-CoV-2 in earlier preserved biological or environmental samples in other many countries. The first cases of COVID-19 could be earlier than that officially reported in Wuhan.
- Bats are considered the most likely natural hosts for SARS-CoV-2. A few bat-derived coronaviruses with genomes similar to SARS-CoV-2, such as RaTG13 and RmYN02, have been reported.

### Sequence similarity of bat coronaviruses compared to SARS-CoV-2:

	Complete Genome	ORF1ab	S	ORF3a	E	M	ORF6	ORF7a	ORF7b	ORF8	N	ORF10
RmYN02	93.3%	97.2%	71.9%	96.4%	98.7%	94.8%	96.8%	96.2%	92.4%	45.8%	97.3%	99.1%
RaTG13	96.1%	96.5%	92.9%	96.3%	99.6%	95.4%	98.4%	95.6%	99.2%	97.0%	96.9%	99.1%

- The most closely related virus to SARS-CoV-2 is RaTG13, with a genome-wide nucleotide identity of 96.1 % and a 92.9 % identity in the S gene [1] **in 2003**, which may form a distinct lineage from SARS-CoV-2.
- Another bat-derived coronavirus, RmYN02, identified in 227 bats from the Yunnan Province in China **between May and October 2019** shares whole genome nucleotide identity of 93.3 % with SARS-CoV-2 and 97.2 % identity for the 1ab gene [2].
- Therefore, we selected two bat coronaviruses, RaTG13 and RmYN02, as the references to identify the SARS-CoV-2 genome sequence that exhibits the highest similarity with bat coronaviruses.

[1] Zhou P, Yang X L, Wang X G, Hu B, Zhang L, Zhang W, Si H, Zhu Y, Li B, Huang C, et al. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 20579: 270–273.

[2] Zhou H, Chen X, Hu T, Li J, Song H, Liu Y, Wang P, Liu D, Yang J, Holmes E C, et al. 2020. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr. Biol.* 30: 2196–2203.

# B. The early transmission of SARS-CoV-2

## Dataset:

- All complete genome sequences of SARS-CoV-2 were downloaded from GISAID before December 31, 2020. In order to ensure the accuracy of analysis, low-quality sequences containing letters other than A, C, G, and T were eliminated from the dataset. Finally, there were 144,566 sequences in the analysed dataset.
- All reference sequences of ss-RNA viruses were downloaded from NCBI before March 23, 2020. In this study, the following three types of sequences were excluded: (1) virus sequences without a family label; (2) families including one or two sequences; and (3) viruses including letters other than A, C, G, and T. In total, 2,051 sequences belonging to 40 families were retained.

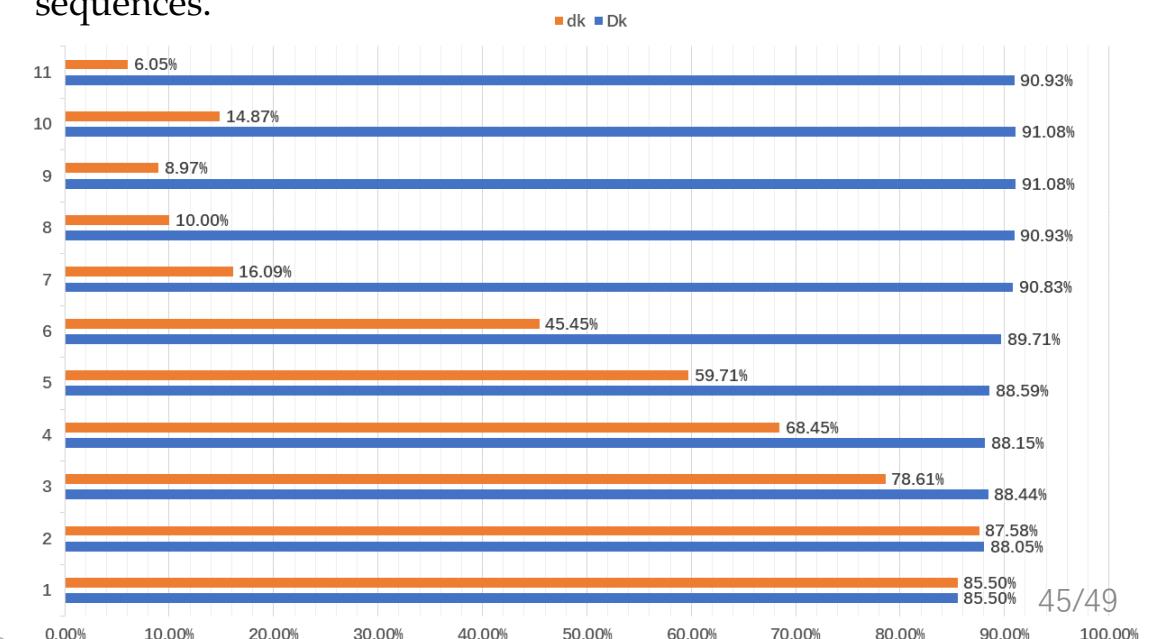
## Natural Metric:

- Let  $d_k(s_1, s_2)$  be the Euclidean distance between two k-mer natural vectors of two genome sequences  $s_1, s_2$  for  $\forall k \geq 1$ , then the new natural metric of two genome sequences  $s_1, s_2$  is defined as  $D_k(s_1, s_2) = d_1(s_1, s_2) + \frac{1}{2^2}d_2(s_1, s_2) + \frac{1}{2^3}d_3(s_1, s_2) + \dots + \frac{1}{2^k}d_k(s_1, s_2)$ .

Shaojun Pei, Stephen S.-T. Yau. The genomic distance between bat coronaviruses and SARS-CoV-2 analysed using a new metric reveals multiple origins of COVID-19. (in submission)

## The choice of the most accurate natural metric based on the nearest neighbour classification of ss-RNA virus

- For  $k$ , from 1 to 11, the new metric  $D_k$  was used as the distance for the nearest neighbour classification of virus families. The results are illustrated in the following Figure, represented by blue-colored bars. **The highest classification accuracy was found to be 91.08 % for  $k=9$ .**
- For comparison, the nearest neighbour classification accuracies we also calculated using  $d_1$  to  $d_{11}$  individually, which are indicated by orange-colored bars in Figure.
- This new natural metric was evidently more accurate; therefore,  $D_9$  was selected for the subsequent analysis of SARS-CoV-2 genome sequences.



# B. The early transmission of SARS-CoV-2

- The distances  $D_9$  between the genome sequence of RaTG13/RmYN02 and all the genome sequences of SARS-CoV-2 in our dataset are ranked.
- The first five SARS-CoV-2 genome sequences with the shortest distance are shown in the right Table.
- In **Table A**: The distance between SARS-CoV-2 collected in Wuhan and bat coronavirus RaTG13 is 25429.13 (ranking 346);
- In **Table B**: The distance between SARS-CoV-2 collected in Wuhan and bat coronavirus RmYN02 is 28405.08 (ranking 7964);
- In **Table C**: The value of the metric between the reference sequence and the SARS-CoV-2 sequence collected from Wuhan was found to be 14.719, which ranked 892. (we also calculated the natural metrics for all the coding regions. to balance the magnitude of the natural metric corresponding to different coding sequences, the natural metric of each coding sequence was divided based on the length of the coding sequence and then added together. )
- In summary, these results indicated that **the distances between the first SARS-CoV-2 sequence collected in Wuhan and the sequences of bat coronaviruses are distant**. Some collected SARS-CoV-2 sequences obtained from France, India, Netherlands, Singapore, and the United States were found to be more similar to bat coronaviruses. Therefore, it is highly unlikely that **China was the first country where the first human-to-human transmission of SARS-CoV-2 occurred**.

## The top five SARS-CoV-2 sequences with a shorter metric.

### A. RaTG13 at the complete genome scale

Rank	Information	Value
1	hCoV-19/ <a href="#">France</a> /B5434/2020 EPI_ISL_443279 2020-04-01	25311.95
2	hCoV-19/ <a href="#">Netherlands</a> /Utrecht_10024/2020 EPI_ISL_454773 2020-03-26	25322.00
3	hCoV-19/ <a href="#">USA</a> /VA-DCLS-0392/2020 EPI_ISL_467788 2020-04-19	25324.18
4	hCoV-19/ <a href="#">USA</a> /CruiseA-1/2020 EPI_ISL_413606 2020-02-17	25327.38
5	hCoV-19/ <a href="#">Sichuan</a> /SC-WCH4-288/2020 EPI_ISL_451390 2020-01-23	25343.60
346	hCoV-19/ <a href="#">Wuhan</a> /WH01/2019 EPI_ISL_406798 2019-12-26	25429.13

### B. RmYN02 at the complete genome scale

1	hCoV-19/ <a href="#">Singapore</a> /14/2020 EPI_ISL_414380 2020-02-13	28049.28
2	hCoV-19/ <a href="#">Singapore</a> /23/2020 EPI_ISL_420100 2020-03-02	28058.59
3	hCoV-19/ <a href="#">Singapore</a> /13/2020 EPI_ISL_414379 2020-02-18	28085.93
4	hCoV-19/ <a href="#">Singapore</a> /30/2020 EPI_ISL_420107 2020-03-09	28093.80
5	hCoV-19/ <a href="#">USA</a> /VA_NIDDL_3216/2020 EPI_ISL_491943 2020-04-16	28094.80
7964	hCoV-19/ <a href="#">Wuhan</a> /WH01/2019 EPI_ISL_406798 2019-12-26	28405.08

### C. totaling all the coding regions

1	hCoV-19/ <a href="#">Japan</a> /PG-0241 EPI_ISL_479932 2020-03-13	14.511
2	hCoV-19/ <a href="#">USA</a> /WA-S187 EPI_ISL_430225 2020-03-20	14.526
3	hCoV-19/ <a href="#">USA</a> /NM-DOH-2020017512 EPI_ISL_831912 2020-04-07	14.548
4	hCoV-19/ <a href="#">Switzerland</a> /BL-UHB-42192943 EPI_ISL_528137 2020-03-16	14.55
5	hCoV-19/ <a href="#">Hungary</a> /1029-28 EPI_ISL_1095616 2020-04-30	14.557
892	hCoV-19/ <a href="#">Wuhan</a> /WH01/2019 EPI_ISL_406798 2019-12-26	14.719



- In this study, we use pure mathematics concept to study important problems in biology. Imitating Hilbert who proposed twenty-three problems in mathematics in 1900, DAPRA also proposed twenty-three problems in pure and applied mathematics. They will be proved to be very influential for 21th-century mathematics.
  - In the number 15 of DAPRA problems, we are asked to understand "**The Geometry of Genome Space**".
  - David Mumford used his geometric invariant theory to construct moduli space of curves. Many people followed his idea to construct moduli space of high dimensional varieties. What we established is the discrete analog of Mumford theory.
  - The natural metric can accurately reflect the distribution of oligonucleotides of different lengths.
  - The genome space with a proper metric is a powerful means of determining the phylogenetics and classification of genomes.
- Sequences from the same biological family have similar nucleotide distribution. Our **convex hull principle** for genome states that the convex hull formed from natural vectors from the same biological groups does not intersect with the convex hulls formed from natural vectors from other biological groups. This can be viewed as **one of the Fundamental Laws of Biology** for which DAPRA is looking for since 2008.
- As **applications**, we explore: (1) the existing but undiscovered genome sequence; (2) the early transmission of SARS-CoV-2 in this moduli space. Furthermore, we apply the concepts in protein, i.e. constructing the protein space.

## Main references:

- Deng M, Yu CL, Liang Q, He RL, Yau SST. A novel method of characterizing genetic sequences: genome space with biological distance and applications. *PLoS One* 2011;6:e17293.
- Yu CL, Hernandez T, Zheng H, Yau SC, Huang HH, He RL, et al. Real time classification of viruses in 12 dimensions. *PLoS One*. 2013;8:E64328.
- Wen J, Chan RHF, Yau SC, He RL, Yau SST. K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene* 2014;546:25–34.
- Sun N, Dong R, Pei S, Yin C, Yau SST. A new method based on coding sequence density to cluster bacteria. *J Comput Biol* 2020;27:1688–98.
- Yu CL, Deng M, Cheng SY, Yau SC, He RL, Yau SST. Protein space: a natural method for realizing the nature of protein universe. *J Theor Biol* 2013;318:197–204.
- Zhao X, Tian K, He RL, Yau SST. Convex hull principle for classification and phylogeny of eukaryotic proteins. *Genomics* 2019;111:1777–84.
- Dong R, Zheng H, Tian K, Yau SC, Mao WG, Yu WP, et al. Virus database and online inquiry system based on natural vectors. *Evolutionary Bioinformatics*. 2017;13. 1176934317746667.
- Boyd S, Lieven V. Convex optimization. Cambridge 2004.
- Defense Advanced Research Projects Agency (DARPA) 2008 proposal of the 23 mathematical challenges.  
<http://www.darpa.mil/dso/personnel/mann.htm>.
- Zhao R, Pei S, Yau SST. New genome sequence detection via natural vector convex hull method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi: 10.1109/TCBB.2020.3040706.

## Acknowledgment

- This work is joint with my students Rui Dong, Shaojun Pei, Nan Sun, Tao Zhou et al. at Tsinghua University.
- This work is supported by National Natural Science Foundation of China (NSFC) grant (91746119), Tsinghua University Spring Breeze Fund (2020Z99CFY044), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (042202008).





清华大学  
Tsinghua University

Thanks