Grand Biological Universe: The geometric construction 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

2023

Abstract

- **Two DARPA problems**

  Imitating Hilbert who proposed 23 problems in mathematics in 1900, Defense Advanced Research Projects Agency (DARPA) proposed 23 problems in pure and applied mathematics in 2008. These problems will prove to be very influential for the development of mathematics in the 21st century. In the DARPA problems, we are asked to understand "The Geometry of Genome Space" (the number 15) and "What are the Fundamental Laws of Biology" (the number 23).

- **Convex hull principle for molecular 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 DARPA has been looking for since 2008.

- **Genome space**

  On the basis of the convex hull principle, we can construct the geometry of the genome space. A genome space consists of all known genomes of living beings and provides insights into their relationships. The genome space can be considered as the moduli space in mathematics, and genome sequences can be canonically embedded in a high-dimensional Euclidean space by means of Natural Vectors. In this space, a sequence is uniquely represented as a point by the nucleotide distribution information of the sequence. Similar sequences lie closely, and convex hulls of different groups are disjoint according to the convex hull principle.

- **Natural Metric**

  The geometry of space is reflected in the similarity of sequences. 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, dark matter and dark energy play a crucial role in the construction of the correct Natural Metric in genome space. Our goal is to construct the genome spaces of seven kingdoms with Natural Metrics. These metrics are quite different in each genome space because different dark matter and dark energy may bend space-time as predicted by Einstein's theory.

- **Applications**

  As applications, we provide the first mathematical method to find undiscovered genome sequence. Our theory allows us to explore the phylogenetic relationships of biological sequences and where SARS-CoV-2 originated from. It provides a novel geometric perspective to study molecular biology. It also gives an accurate way for large-scale sequences comparison in real-time manner.

CONTENTS

PART 01

Research Background

The Biotechnology Revolution in the 21st Century Brings Massive Biological Observation Data

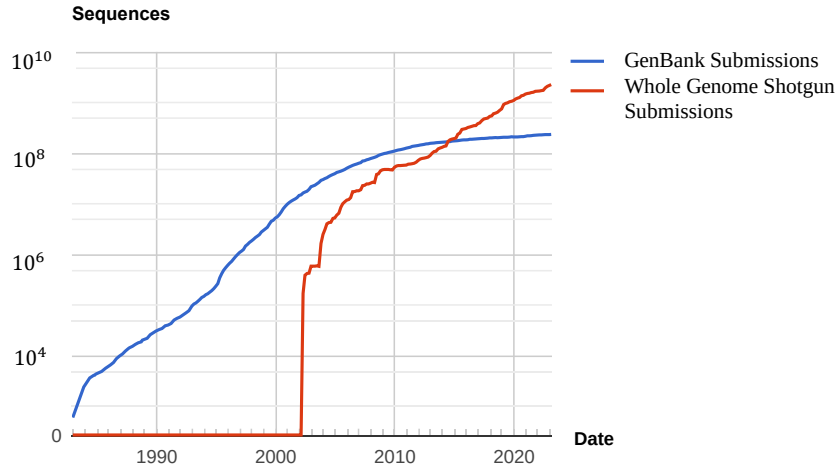
The first stage: Sanger sequencing technology.

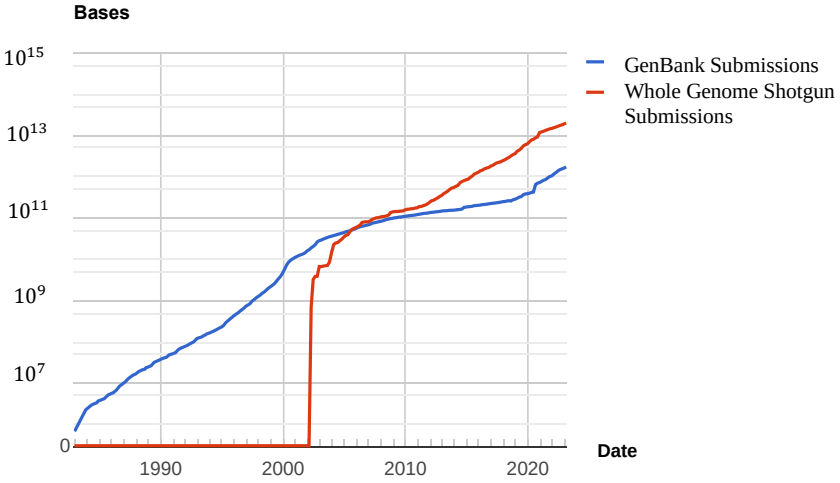
The second stage: Next generation sequencing technology (NGS), including Illumina sequencing etc.

The third stage: The third-generation sequencing technology, including Pacific Biosciences' single-molecule real-time sequencing (SMRT) and Oxford Nanopore Technologies' nanopore sequencing.

1970s

Advances in high-throughput sequencing technology drive sequence development





https://www.ncbi.nlm.nih.gov/genbank/statistics/

Genomic Data is Diverse

The genome size and the chromosome have enormous differences:

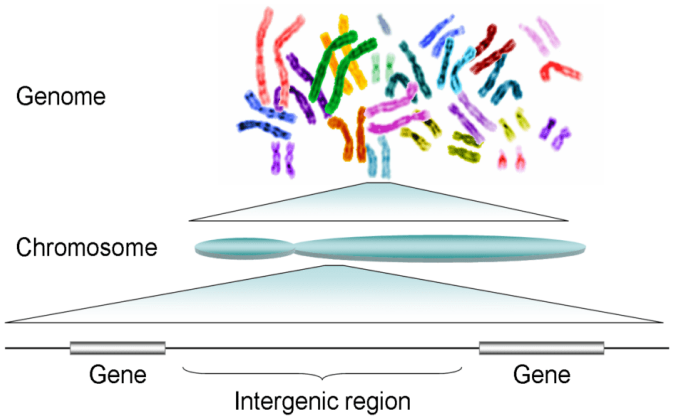
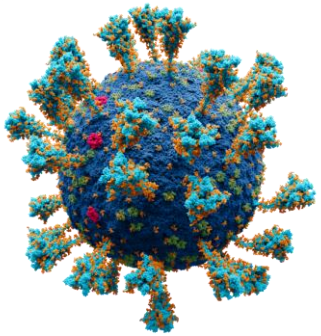


Figure: The genome is the sum of all genetic material in the body. Genes are carried on chromosomes. https://www.biologyonline.com/dictionary/genome



Virus Example:

SARS-CoV-2 is a (+)ssRNA virus. It has a genome size of 29903 bp. (NC_045512)

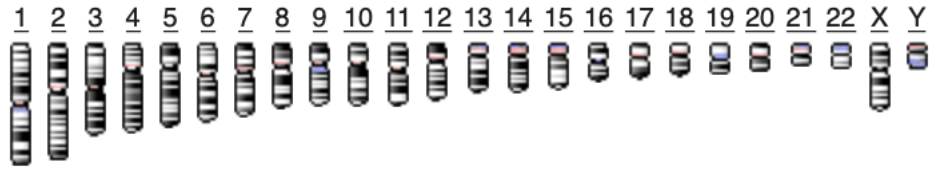
Vertebrate Example:

Diploid human has 46 chromosomes, of which 24 linear molecules comprise 3.1 billion nucleotides (GCF_000001405.40 GRCh 38.p14).

Homo sapiens reference genome GRCh38.p14

Submitted by Genome Reference Consortium (February 2022)

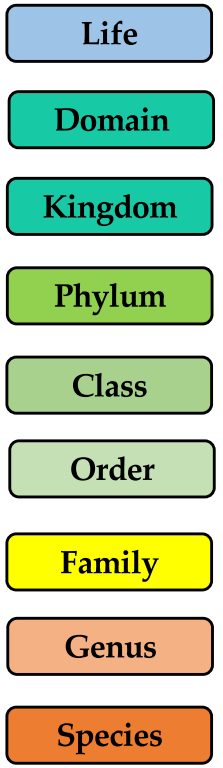
RefSeq: GCF_000001405.40



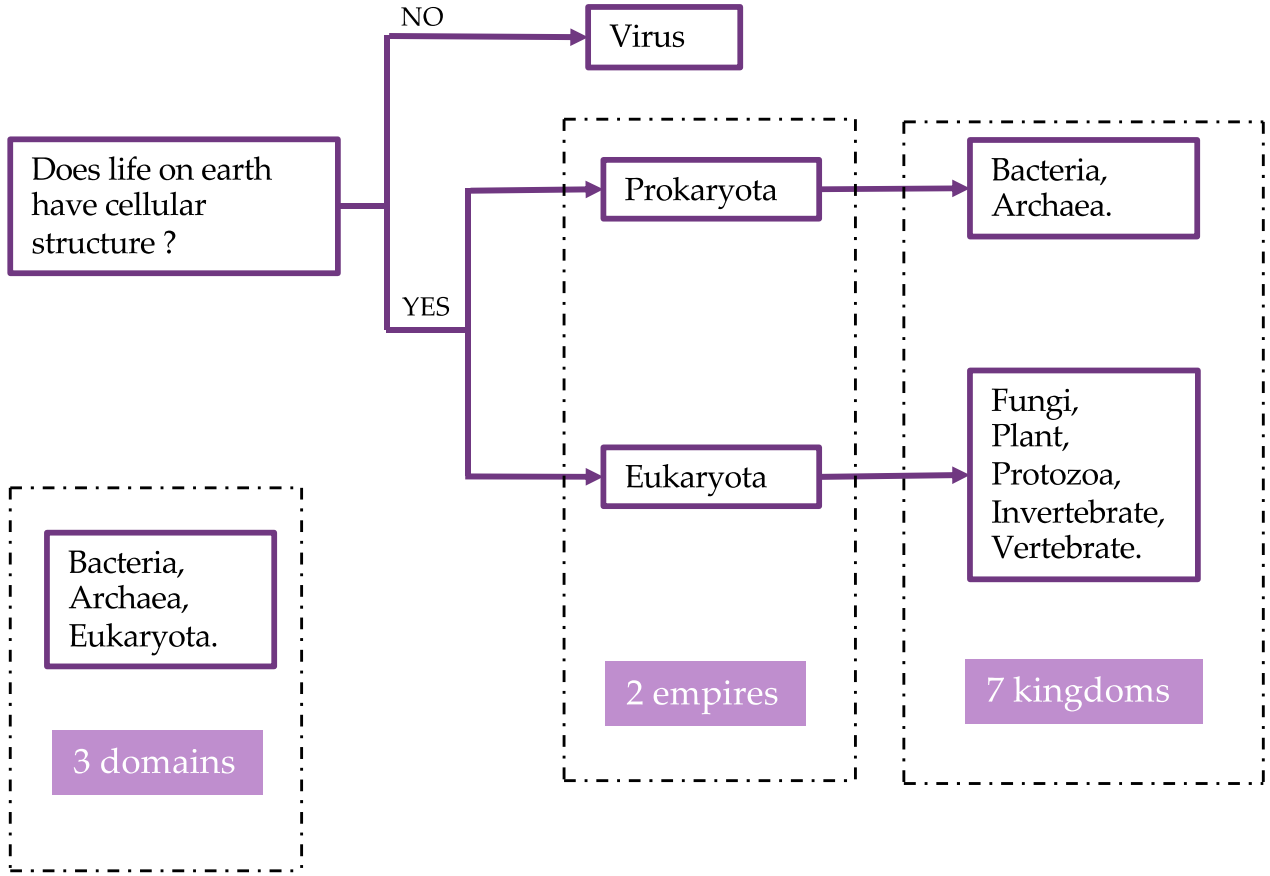
Bacteria Example:

E.coli has 1 chromosomes, and its genome size is 5.6 Mbp. (GCF_000008865.2)

Background: Biological Taxonomy



Furthermore

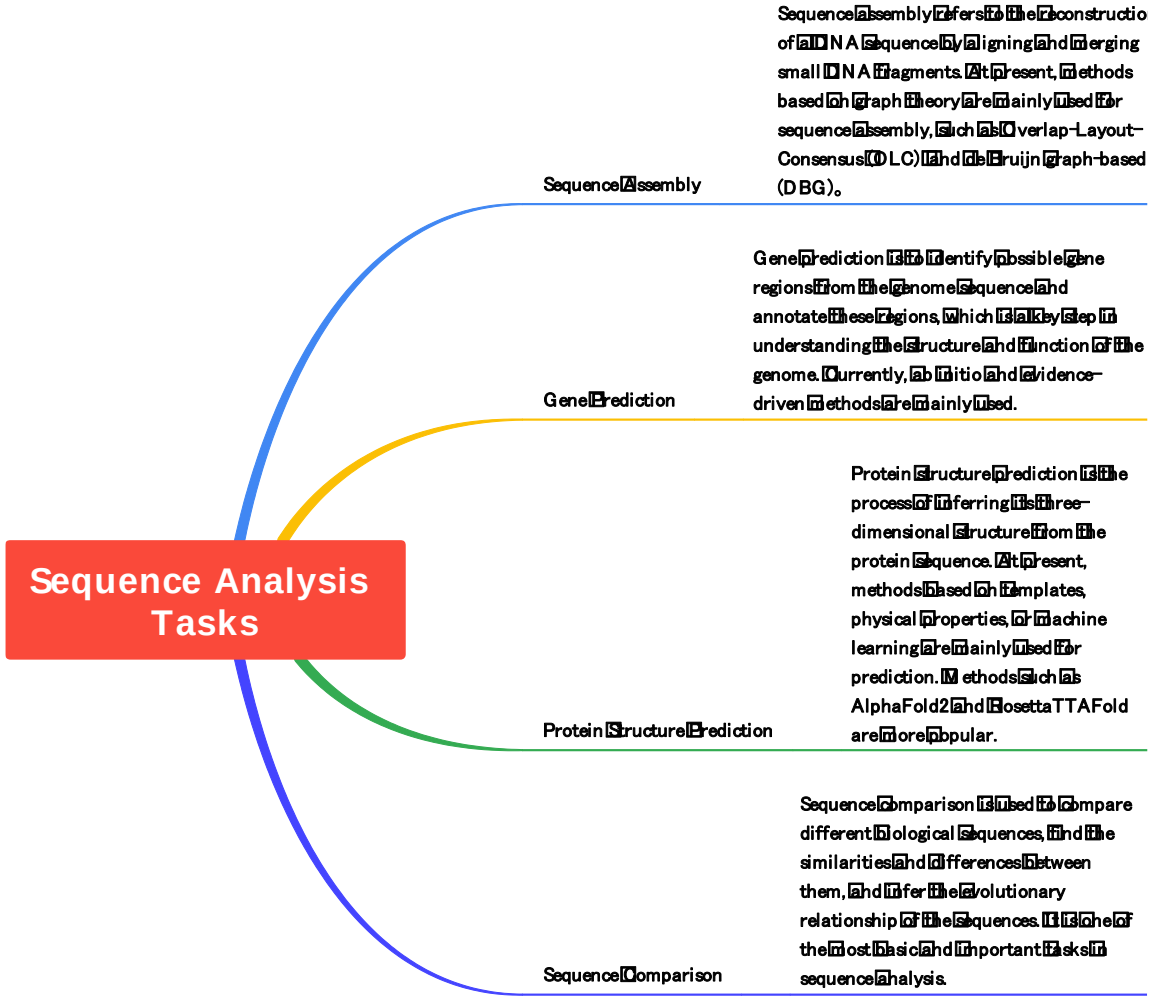


The basic scheme of modern classification.

The principal ranks of modern classification and the classification label according to whether life on earth has cellular structure.

Sequence Analysis Methods

Sequence analysis tasks include sequence assembly, gene prediction, protein structure prediction, and sequence comparison.



Sequence comparison methods include alignment-based algorithms and alignment-free algorithms：

- Alignment-based methods use known sequences as a reference to predict the location and characteristics of new sequences.
- Alignment-free methods do not rely on reference sequences, but directly compare the similarity between two or more sequences.

Sequence Analysis Methods

- Alignment-based methods are limited by the sequence data size, and reliable alignments cannot be obtained when the sequences are much diverse. So alignment-free methods have been widely used in recent years.
- Alignment-free methods include methods based on k-mer frequency, length of common substring, number of word matches, micro-alignments, information theory, and graphical representation.
- These methods have been widely used in the fields of sequence similarity search, sequence clustering and classification, and phylogenetics of molecular biology.

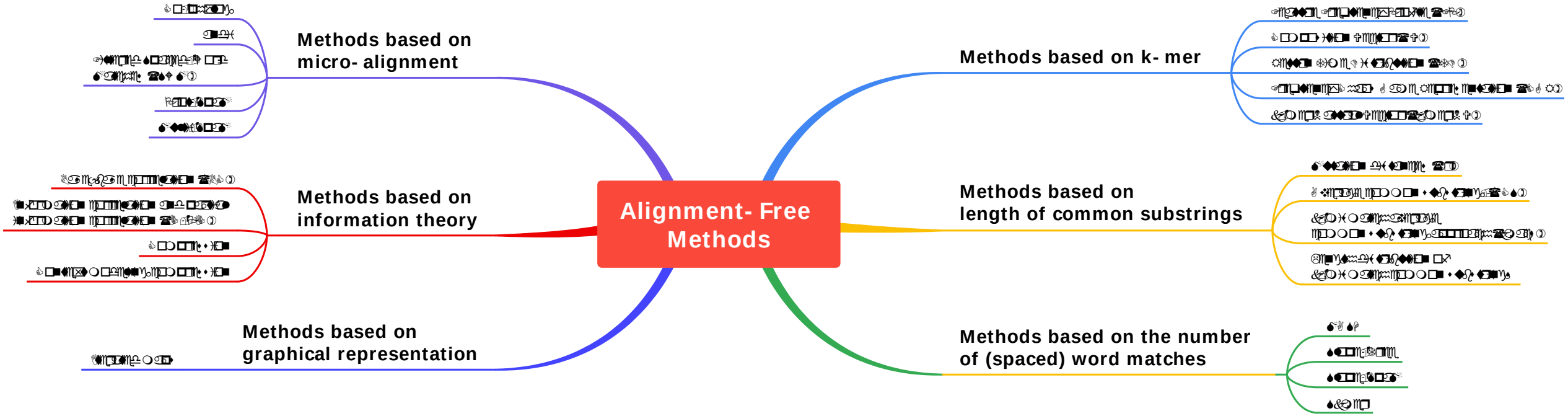


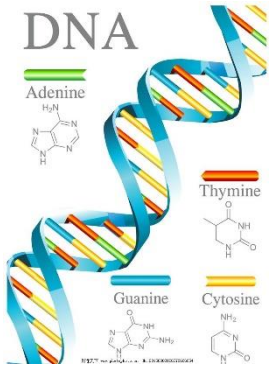
Figure: Alignment-Free Methods overview. Summarized from https://en.wikipedia.org/wiki/Alignment-free_sequence_analysis。

Sequence Analysis Methods (Our Team Works)

Our team has made many cutting-edge achievements in alignment-free methods.

- Developed a series of algorithms to efficiently and accurately compare large biological datasets
  - Graphical representation of DNA sequence
  - Numerical Represetation Based on k-mer
    - Traditional natural vector method describing four base position distributions
    - Improved methods of adding base position covariance information
    - k-mer natural vector method describing the k-mer position distribution
    - Vector representation using the global distribution of bases
  - Sequence Signal Coding Method Based on Fourier Transform
    - Method using Fourier transform power spectrum
    - Method based on cumulative Fourier power spectrum
    - Cumulative Fourier Power-Phase Spectrum Method Introducing Phase Spectrum
  - Sequence Representation Method Based on Machine Learning
    - word2vec
    - Neural Network based method
- Most of these methods have been successfully applied to important practical problems such as
  - Human origin
  - Traceability of SARS-CoV-2
  - Global reflection of HIV genome sequence mutation and local detection of highly mutated gene regions
  - Protein detection
- It also performed well on many theoretical problems, such as
  - A theoretical definition of genome space that reveals the geometric structure of biological sequences
  - A theoretical definition of protein space that reveals the geometric structure of biological sequences
  - A theoretical definition of Natural Distance for Measuring Sequence Similarity

Natural Vector Encoding (Our Team Key Method)



ACAGCTCT……GCTCACATG

Natural Vector Definition:

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

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

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

- 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 $D_j^k = \sum_{i=1}^{n} \frac{(i-\mu_k)^j w_k(s_i)}{n_k^{j-1} n^{j-1}}$ be the j-th central moment of position of letter k.

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

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

Natural Vector Example:

If the genomic sequence is **ACGGTAGTCC**, the indicator functions are shown as follows:

Seq.	A	C	G	G	T	A	G	T	C	C
Pos.	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

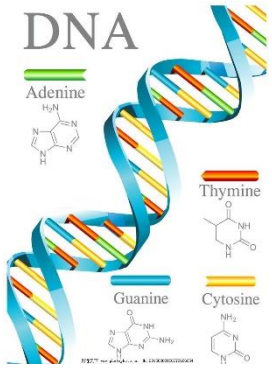
The corresponding components of distribution vector are calculated as follows:

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

Then the 12-dimensional Natural Vector is:

$(2, 3, 3, 2, 3.5, 7, 4.67, 6.5, 0.63, 1.27, 0.29, 0.23)$.

k-mer Natural Vector Encoding (Our Team Key Method)



ACAGCTCT……GCTCACATG

k-mer l_i 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.

K-mer Natural Vector Definition:

Let $S = s_1 s_2 s_3 \ldots s_n$ be a genomic sequence of length n. Suppose that $l_i[j]$ is the indicator function of the j-th occurrence of a k-mer l_i in S $(i = 1, 2, \ldots, 4^k)$: $l_i[j] = \begin{cases} 1, & if\ \mathrm{L_j} = l_i, \\ 0, & otherwise. \end{cases}$ the distributions of a k-mer l_i can be described by three components:

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

Then the k-mer Natural Vector with high order central moment (k-mer NVHO) for sequence S is defined by a $4^k \cdot (\mathrm{m}+1)$-dimensional vector ($n = 2$ is enough) :

$$(\mathrm{n}_{l_1}, \ldots, \mathrm{n}_{l_{4^k}}, \mu_{l_1}, \ldots, \mu_{l_{4^k}}, \mathrm{D}_{l_1}^2, \ldots, \mathrm{D}_{l_{4^k}}^2, \ldots, \mathrm{D}_{l_1}^n, \ldots, \mathrm{D}_{l_{4^k}}^n)$$

K-mer Natural Vector 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
Pos.	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:

$(\mathrm{n_{AA}}, \mathrm{n_{AC}}, n_{AG}, n_{AT}, n_{CA}, n_{CC}, n_{CG}, n_{CT}, n_{\mathrm{GA}}, \mathrm{n_{GC}}, \mathrm{n_{GG}}, \mathrm{n_{GT}}, \mathrm{n_{TA}}, \mathrm{n_{TC}}, \mathrm{n_{TG}}, \mathrm{n_{TT}},$
$\mu_{\mathrm{AA}}, \mu_{\mathrm{AC}}, \mu_{AG}, \mu_{AT}, \mu_{CA}, \mu_{CC}, \mu_{CG}, \mu_{CT}, \mu_{\mathrm{GA}}, \mu_{\mathrm{GC}}, \mu_{\mathrm{GG}}, \mu_{\mathrm{GT}}, \mu_{\mathrm{TA}}, \mu_{\mathrm{TC}}, \mu_{\mathrm{TG}}, \mu_{\mathrm{TT}},$
$\mathrm{D_2^{AA}}, \mathrm{D_2^{AC}}, \mathrm{D_2^{AG}}, \mathrm{D_2^{AT}}, \mathrm{D_2^{CA}}, \mathrm{D_2^{CC}}, \mathrm{D_2^{CG}}, \mathrm{D_2^{CT}}, \mathrm{D_2^{GA}}, \mathrm{D_2^{GC}}, \mathrm{D_2^{GG}}, \mathrm{D_2^{GT}}, \mathrm{D_2^{TA}}, \mathrm{D_2^{TC}}, \mathrm{D_2^{TG}}, \mathrm{D_2^{TT}}$
)

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.

Natural Vector Advantage

清華大學 Tsinghua University

Advantage 1: The fast calculation speed

Input Data

- HIV: About 8000bp
- E. coli: About 5Mbp
- Bird: About 900Mbp
- Homo sapiens: About 3000Mbp

Output Data

- Alignment methods:
  Output dimension is the length of sequences
- Natural Vector:
  Output dimension is 12- dimension

For example, there are 48 sequences belonging to modern birds.

- Alignment method:
  About 1- 2 years by 9 supercomputing centers
- Natural Vector:
  About 4 days by a small server with 384GB

Advantage 2: Interpretability

Reason

- The distribution of ATCG uniquely determines the genome.
- The information of nucleotides distributions within the virus genome (Natural Vector) can be used to canonically embed this virus genome as a point in Euclidean space. This can be viewed as the discrete analog of Kodaira embedding.

Jarvis, E. D., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. Science.

Motivation: 23 Problems Proposed by DARPA

Tsinghua University

Problem	Brief explanation
1	The continuum hypothesis (that is, there is no set whose cardinality is strictly between that of the integers and that of the real numbers)
2	Prove that the axioms of arithmetic are consistent.
3	Given any two polyhedra of equal volume, is it always possible to cut the first into finitely many polyhedral pieces that can be reassembled to yield the second?
4	Construct all metrics where lines are geodesics.
5	Are continuous groups automatically differential groups?
6	Mathematical treatment of the axioms of physics: (a) axiomatic treatment of probability with limit theorems for foundation of statistical physics (b) the rigorous theory of limiting processes "which lead from the atomistic view to the laws of motion of continua"
7	Is a^b transcendental, for algebraic a ≠ 0,1 and irrational algebraic b ?
8	The Riemann hypothesis ("the real part of a non-trivial zero of the Riemann zeta function is 1/2") and other prime-number problems, among them Goldbach's conjecture and the twin prime conjecture
9	Find the most general law of the reciprocity theorem in any algebraic number field.
10	Find an algorithm to determine whether a given polynomial Diophantine equation with integer coefficients has an integer solution.
11	Solving quadratic forms with algebraic numerical coefficients.
12	Extend the Kronecker–Weber theorem on Abelian extensions of the rational numbers to any base number field.
13	Solve 7th-degree equation using algebraic (variant: continuous) functions of two parameters.
14	Is the ring of invariants of an algebraic group acting on a polynomial ring always finitely generated?
15	Rigorous foundation of Schubert's enumerative calculus.
16	Describe relative positions of ovals originating from a real algebraic curve and as limit cycles of a polynomial vector field on the plane.
17	Express a nonnegative rational function as quotient of sums of squares.
18	(a) Are there only finitely many essentially different space groups in n-dimensional Euclidean space? (b) Is there a polyhedron that admits only an anisohedral tiling in three dimensions? (c) What is the densest sphere packing?
19	Are the solutions of regular problems in the calculus of variations always necessarily analytic?
20	Do all variational problems with certain boundary conditions have solutions?
21	Proof of the existence of linear differential equations having a prescribed monodromic group
22	Uniformization of analytic relations by means of automorphic functions
23	Further development of the calculus of variations

Hilbert's 23 problems

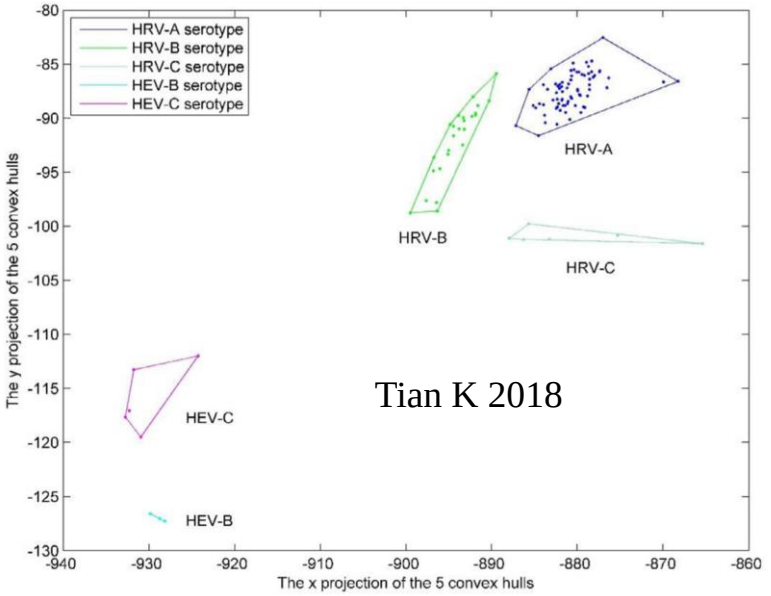
DARPA's 23 problems

Problem	Brief explanation
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 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 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?

Imitating Hilbert who proposed 23 problems in mathematics in 1900, Defense Advanced Research Projects Agency (DARPA) proposed 23 problems in pure and applied mathematics in 2008. These problems will be proven to be very influential for the development of mathematics in the 21st-century.

Convex Hull Principle for Molecular Biology & Genome Space & Natural metric

- 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.



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

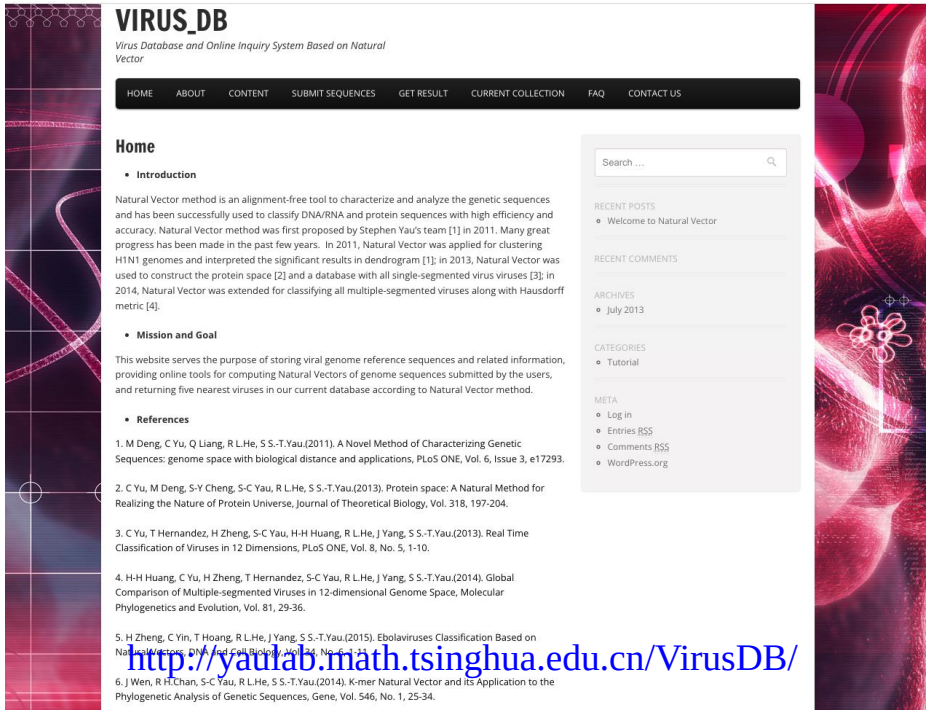
- On the basis of the convex hull principle, we can construct the geometry of the genome space.
- A **genome space** consists of all known genomes of living beings. It can be considered as the moduli space in mathematics, and genome sequences can be canonically embedded in a high-dimensional Euclidean space by means of Natural Vectors.

- The geometry of genomic space is reflected in sequence similarity, and proper **Natural Metric** is important for inferring genome phylogeny and taxonomy.

Biological Sequence Datasets Analyzed

To construct the genome space, we downloaded all reliable genomes of virus and 7 kingdoms.

Dataset	Data Information
Virus	7382 sequence: 83 families
Bacteria	24719 nucleoid sequences: 425 families
Archaea	440 complete genomes: 7 phyla
Fungi	2628 chromosome sequences: 23 families.
Plants	399 chromosome sequences: 22 families.
Protozoa	1200 chromosome sequences: 20 families.
Vertebrates	390 chromosome sequences: 4 class
Invertebrates	345 chromosome sequences: 3 orders



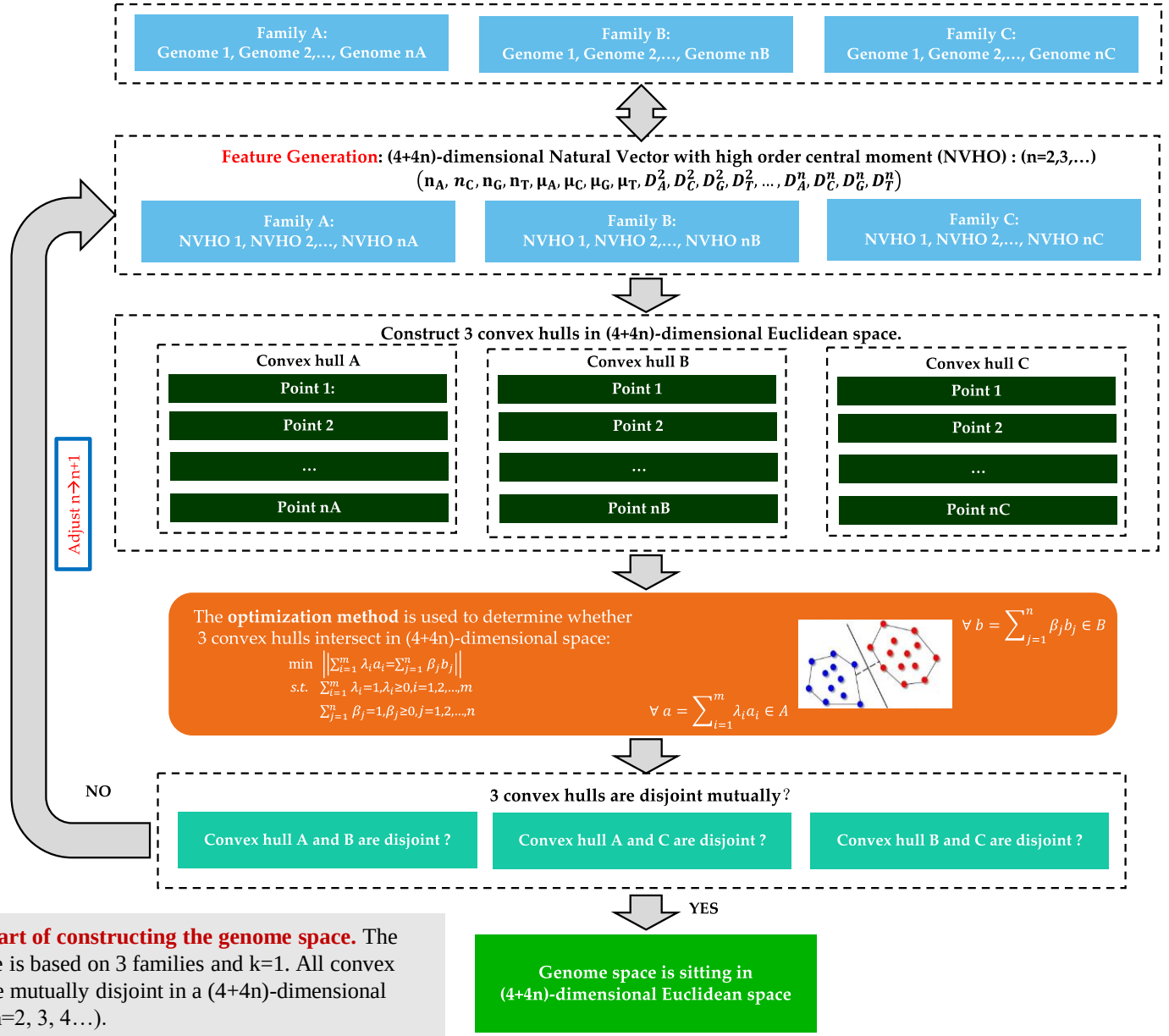
http://yaulab.math.tsinghua.edu.cn/VirusDB/

Update the Virus Database Based on Natural Vector Method: VirusDB

PART 02

Methods & Materials

Research Target 1: The Genome Space Construction



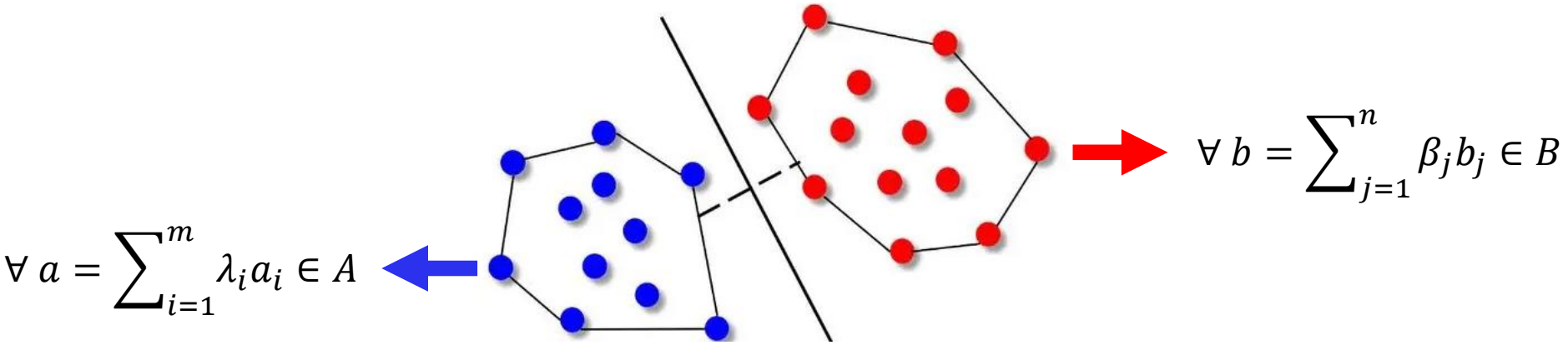
Flowchart of constructing the genome space. The example is based on 3 families and k=1. All convex hulls are mutually disjoint in a (4+4n)-dimensional space (n=2, 3, 4…).

Review genome space Definition:

- The genome space contains all known genomes and reflects the important nature of the genome universe.
- Mathematically, the genome space can be regarded as a moduli space and constructed as a subspace in a high-dimensional Euclidean space.
- If the convex hull principle holds in R^K, the genome space exists and the genome space is located in a K-dimensional Euclidean space. Here K is the minimum dimension of the Euclidean space where the convex hull principle holds.

Optimization Method is Used to Check Whether Two Convex Hulls Intersect.

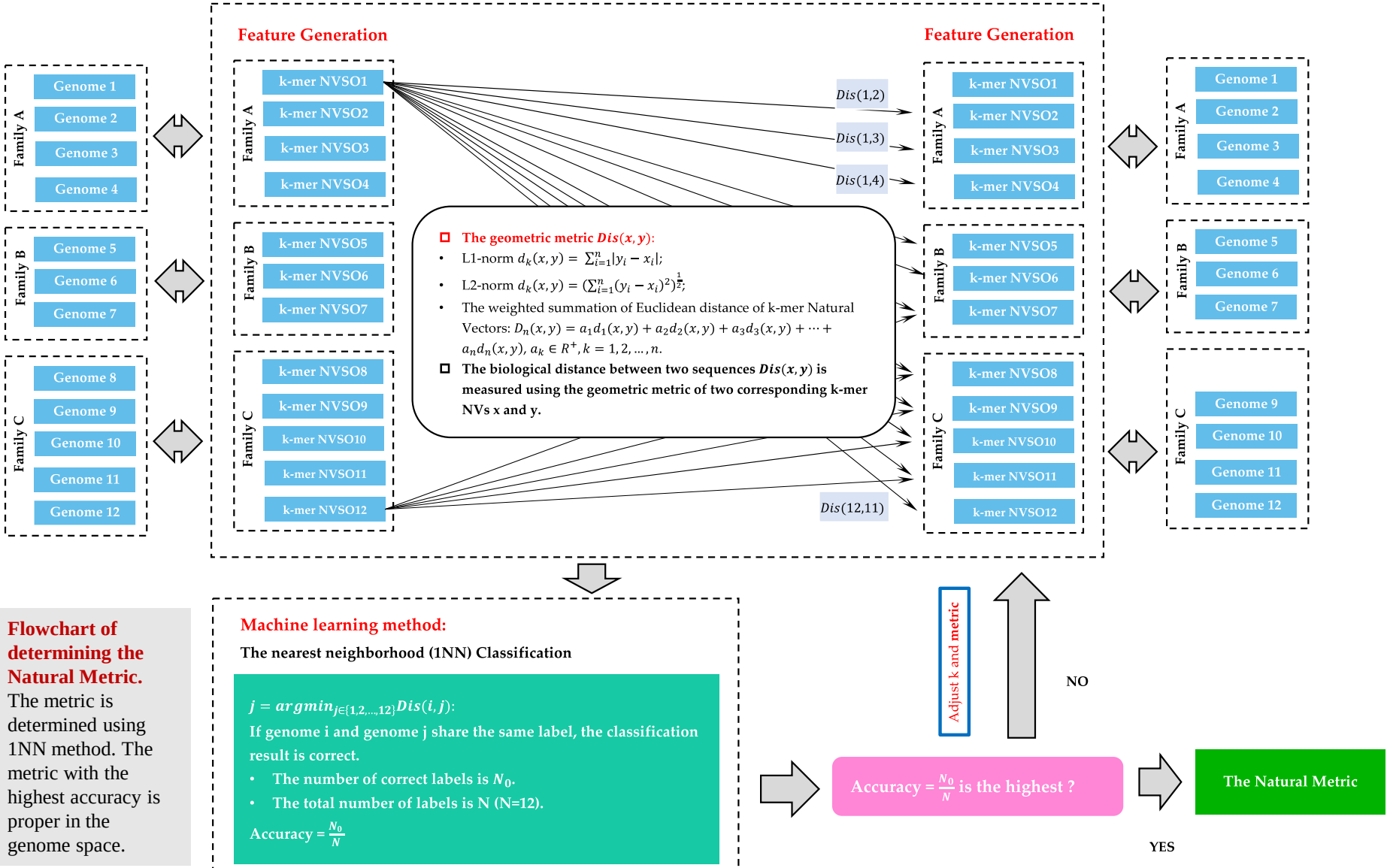
- If A is the convex hull of point set $\{a_1, a_2, \dots, a_m\}$, and B is the convex hull of point set $\{b_1, b_2, \dots, b_n\}$. The mathematical principle is that if A and B intersect, then $\sum_{i=1}^{m} \lambda_i a_i = \sum_{j=1}^{n} \beta_j b_j$, where $\sum_{i=1}^{m} \lambda_i = 1, \lambda_i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



- 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:

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

Research Target 2: Determination of Natural Metrics in Genome Space



Flowchart of determining the Natural Metric. The metric is determined using 1NN method. The metric with the highest accuracy is proper in the genome space.

Review Natural Metric Definition:

- The geometric metric with the highest 1NN accuracy is the Natural Metric in the genome space. The Natural Metric can measure the similarity of the sequences, and determine the phylogenetics and classification of genomes.
- The uncertainty of k gives the space to adjust the weights of D_n and improve the classification accuracy using the metric definition.
- The weight a_k of the weighted distance D_n reflects the contribution of the corresponding k-mer NV to the description of sequence similarity.

The Geometric Metrics

- L1-distance of k-mer Natural Vectors: $d_k(x,y) = \sum_{i=1}^{n}|y_i - x_i|$;
- L2-distance of k-mer Natural Vectors: $d_k(x,y) = (\sum_{i=1}^{n}(y_i - x_i)^2)^{\frac{1}{2}}$;
- The weighted summation of Euclidean distance of k-mer Natural Vectors: $D_n(x,y) = a_1 d_1(x,y) + a_2 d_2(x,y) + a_3 d_3(x,y) + \cdots + a_n d_n(x,y)$, $a_k \in R^+, k = 1,2,\ldots,n$. The weight a_k reflects the contribution of the corresponding k-mer NV to the description of sequence similarity. The uncertainty of k and weight a_k provides space for improving classification accuracy. Normally, we can test the following a_k:

  ① If $a_k = \frac{1}{1.5^{k-1}}$, $D1_n = d_1 + \frac{1}{1.5}d_2 + \frac{1}{1.5^2}d_3 + \cdots + \frac{1}{1.5^{n-1}}d_n$;

  ② If $a_k = \frac{1}{2^{k-1}}$, $D2_n = d_1 + \frac{1}{2}d_2 + \frac{1}{2^2}d_3 + \cdots + \frac{1}{2^{n-1}}d_n$;

  ③ If $a_k = \frac{1}{3^{k-1}}$, $D3_n = d_1 + \frac{1}{3}d_2 + \frac{1}{3^2}d_3 + \cdots + \frac{1}{3^{n-1}}d_n$;

  ④ If $a_k = \frac{1}{k^{1.5}}$, $D4_n = d_1 + \frac{1}{2^{1.5}}d_2 + \frac{1}{3^{1.5}}d_3 + \cdots + \frac{1}{n^{1.5}}d_n$;

  ⑤ If $a_k = \frac{1}{k^2}$, $D5_n = d_1 + \frac{1}{2^2}d_2 + \frac{1}{3^2}d_3 + \cdots + \frac{1}{n^2}d_n$;

  ⑥ If $a_k = \frac{1}{k^3}$, $D6_n = d_1 + \frac{1}{2^3}d_2 + \frac{1}{3^3}d_3 + \cdots + \frac{1}{n^3}d_n$;

PART 03

Results

Result 1: Geometric Construction of Genome Space

Domain	Kingdom	Column A	Column B		Column C	
		Data Information	The embedded dimension of genome space	The corresponding vector	The Natural Metric	The corresponding maximum accuracy
Virus		7382 sequence: 83 families	32	NV with 7-order moment	$D2_9 = d_1 + \frac{1}{2}d_2 + \frac{1}{2^2}d_3 + \cdots + \frac{1}{2^8}d_9$; L2-distance	0.883
Prokaryota	Bacteria	24719 nucleoid sequences: 425 families	48	NV with 11-order moment	d_9; L1-distance	0.9030
	Archaea	440 complete genomes: 7 phyla	48	2-mer NV with 2-order moment	$\mathrm{D6}_{10} = d_1 + \frac{1}{2^3}d_2 + \frac{1}{3^3}d_3 + \cdots + \frac{1}{10^3}d_{10}$; L1-distance	0.9384
Eukaryota	Fungi	2628 chromosome sequences: 23 families.	40	NV with 9-order moment	d_9; L1-distance	0.9056
	Plants	399 chromosome sequences: 22 families.	24	NV with 5-order moment	d_2; L1-distance	0.8216
	Protozoa	1200 chromosome sequences: 20 families.	36	NV with 8-order moment	$\mathrm{D4}_9 = d_1 + \frac{1}{2^{1.5}}d_2 + \cdots + \frac{1}{9^{1.5}}d_9$; L1-distance	0.9274
	Vertebrates	390 chromosome sequences: 4 class	28	NV with 6-order moment	$\mathrm{D2}_2 = d_1 + \frac{1}{2}d_2$; L1-distance	0.8590
	Invertebrates	345 chromosome sequences: 3 orders	48	2-mer NV with 2-order moment	$\mathrm{D5}_9 = d_1 + \frac{1}{2^2}d_2 + \frac{1}{3^2}d_3 + \cdots + \frac{1}{9^2}d_9$; L1-distance	0.8783

Calculation Example of Genome Space (Virus)

Table: **The convex hull principle of viral genomes holds in R^{32}.**

The results are based on 7382 sequences of 83 families. The number of disjoint convex hull pairs with the increase of the order of NVHO. There are $C_{83}^{2} = 3404$ convex hull pairs totally, and all pairs are disjoint when the order is more than 7. According to the definition of embedding dimension of the moduli space, we choose the space with the lowest dimension, which indicates that the viral genome space is sitting in a 32-dimensional Euclidean space.

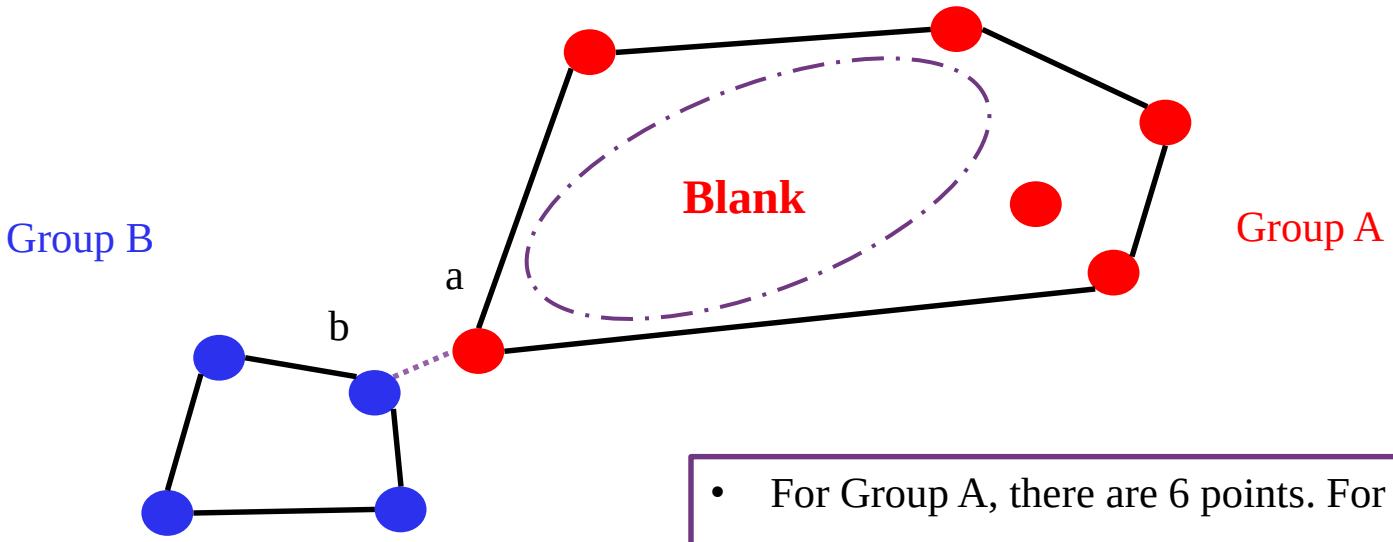
Natural Vector with n-order central moment	n=2	n=3	n=4	n=5	n=6	n=7	n=8	n=9	n=10	n=11
Euclidean space	R^{12}	R^{16}	R^{20}	R^{24}	R^{28}	R^{32}	R^{36}	R^{40}	R^{44}	R^{48}
No. of disjoint convex hull pairs	3221	3291	3338	3354	3395	3403	3403	3403	3403	3403
No. of intersecting convex hull pairs	182	112	65	49	8	0	0	0	0	0

Calculation Example of Natural Metric (Virus)

Table: Natural metric in viral genome space. The nearest neighborhood classification accuracies of virus family based on the new Natural Metric for different n. For weight $\frac{1}{2^k}$ $\left(d = \sum_{k=1}^{n} \frac{1}{2^{k-1}} d_k\right)$, 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 $d = d_1 + \frac{1}{2} d_2 + \frac{1}{2^2} d_3 + \cdots + \frac{1}{2^{n-1}} d_n$.

Weight a_k	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
$\frac{1}{2^k}$	79.9%	82.8%	83.3%	83.3%	84.1%	85.8%	86.9%	87.4%	88.3%
$\frac{1}{k^2}$	79.9%	82.8%	83.3%	83.3%	84.4%	86.3%	87.7%	88.0%	85.6%

Analysis With 1NN Accuracy Less Than 100%: Incomplete genome space



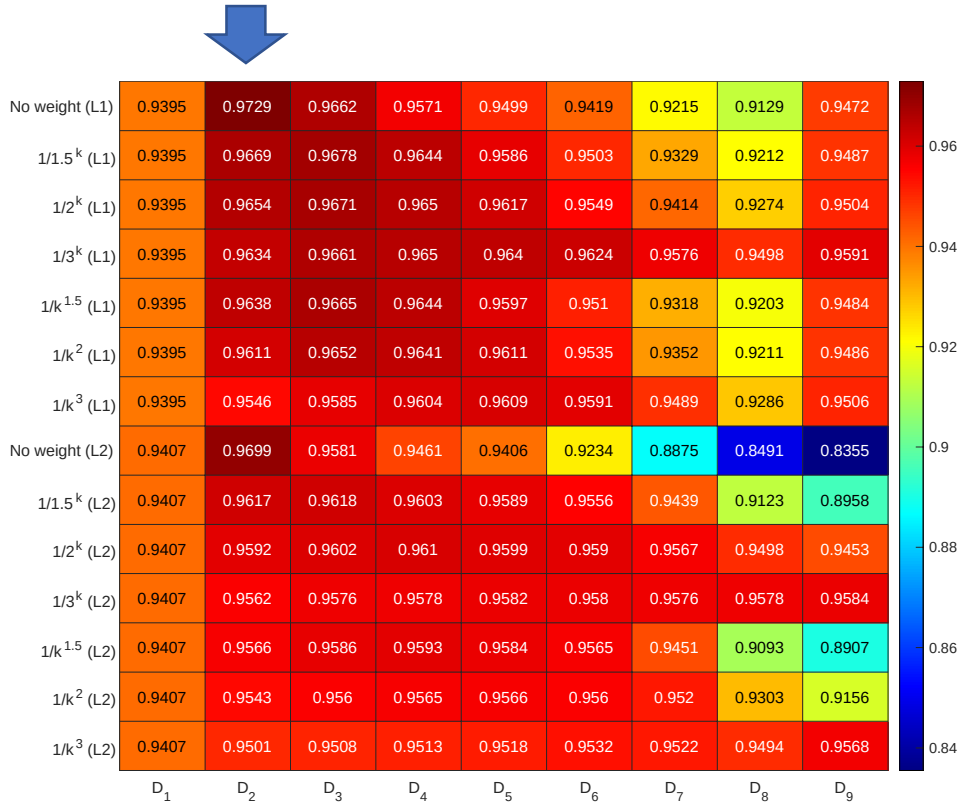
- For Group A, there are 6 points. For Group B, there are 4 points.
- There is a large blank part for Group A, so maybe there is a sequence that has not been sequenced. Then point **a** of Group A is closer to point **b** of Group B. That is to say, the 1NN accuracy does not have to be 100%.

Result 2: The Grand Biological Universe

- **Grand Biological Universe Definition:** Grand Biological Universe should contain all reliable biological sequences. In this Grand Universe, the universes of seven kingdoms are mutually disjoint, and the convex hull principle of genomes for each kingdom holds. The metric of the Grand Biological Universe can reflect the distance between two universes.

The geometric structure of the genome space corresponding to each kingdom may be different. In order to accommodate all seven kingdoms, we intuitively consider the maximum value of each genome space dimension to determine the arrangement of each sequence:

	The embedded dimension of genome space	The corresponding vector
Bacteria	48	NV with 11-order moment
Archaea	48	2-mer NV with 2-order moment
Fungi	40	NV with 9-order moment
Plants	24	NV with 5-order moment
Protozoa	36	NV with 8-order moment
Vertebrates	28	NV with 6-order moment
Invertebrates	48	2-mer NV with 2-order moment

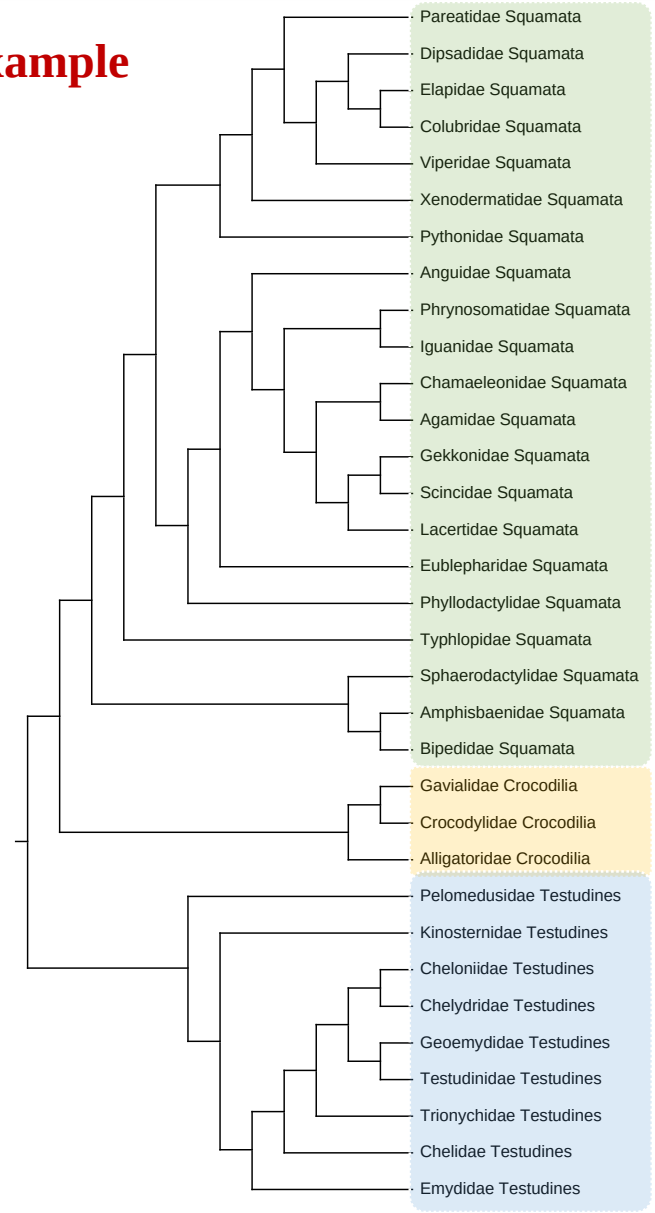


PART 04

Applications

Application 1 (k-mer Natural Vector): Phylogenetic Tree Construction

Example

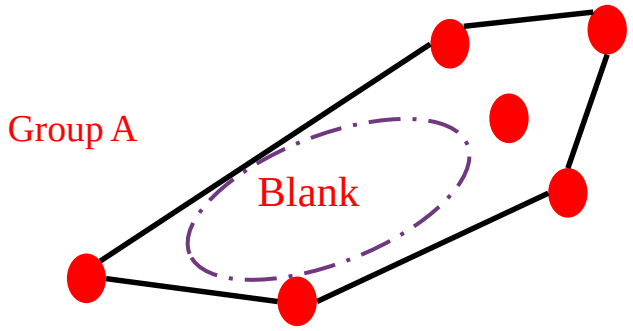


	Data Information	The embedded dimension of genome space	The corresponding vector	The Natural Metric	The corresponding maximum accuracy
Vertebrates	390 chromosome sequences: 4 class	28	NV with 6-order moment	$D_7 = d_1 + \frac{1}{2}d_2$; L1-distance	0.8590

- We use FastME software (http://www.atgc-montpellier.fr/fastme/) to draw a tree and analyze the phylogenetic relationship.

- Here we give an example of invertebrate mitochondrial sequences. The tree is based on sequences belonging to 54 families and 4 orders of reptile. Here we only present a part of the tree. The name type is “family_order”.
- **Result Analysis:** Sequences that are from the same order are clustered together. The result suggests that the k-mer NV has the potential to do phylogenetics analysis.

Application 2 (Convex Hull Principle): Genome Sequence Detection



- An established convex hull may have a large blank part, so there may be sequences with biological significance. These sequences may be mutated sequences or already exist but have not yet been discovered.

- By listing the equations that satisfy the convex hull conditions and solving them, a series of heuristic methods are developed to find the biologically meaningful points in the convex hull of the virus, and then deduce the biologically meaningful sequences.

Algorithm 1. Random-Permutation Algorithm With Penalty (RAP)

M_{Iter}: maximal iteration number; ϵ: preset limit.
Initialization Step : Randomly generate a sequence $\mathbf{S}^{\text{seq}}$ with $n_k, k \in \mathbb{K}$ requirement. And let Iter = 0.
Iteration Step

```
Get the positions of [A, C, G, T] based on S^seq.
while loss(S^seq) > ε and iter < M_Iter do
    Iter = Iter + 1.
    Get [P_A, P_C, P_G, P_T](S^seq, S^tg) based on Equation (7).
    Let S^new = S^seq.
    while loss(S^new) ≥ loss(S^seq) do
        Randomly Select two nucleotides, k, q ∈ K, k ≠ q based on Probabilities [P_A, P_C, P_G, P_T].
        Randomly Select a position from the positions of k and q, separately: pos_k and pos_q.
        Get S^new from S^seq (Do a permutation between pos_k and pos_q).
    end while
    Remove pos_k from positions of k and remove pos_q from positions of q.
    Add pos_q to positions of k and add pos_k to positions of q.
    Let S^seq = S^new.
end while
The S^seq is the sequence which minimizes our loss function with the preset limit.
```

Algorithm 2. Random-permutation Algorithm With Penalty and Constrained Search (RAPCOS)

M_{Iter}: maximal iteration number; ϵ: preset limit.
Initialization Step : Randomly generate a sequence $\mathbf{S}^{\text{seq}}$ with $n_k, k \in \mathbb{K}$ requirement. And let Iter = 0.
Iteration Step

```
while loss(S^seq) > ε and iter < M_Iter do
    Iter = Iter + 1.
    Get the positions of [A, C, G, T] based on S^seq.
    Get [P_A, P_C, P_G, P_T](S^seq, S^tg) based on Equation (7).
    Let S^new = S^seq.
    while loss(S^new) ≥ loss(S^seq) do
        Randomly select a nucleotide k ∈ K based on probability [P_A, P_C, P_G, P_T].
        Randomly select pos_k from the positions of k.
        Denote K_c = K\{k}.
        for k_1 ∈ K_c do:
            if (μ_k − μ_k^tg)(μ_k1 − μ_k1^tg) > 0 then:
                K_c = K_c\{k_1}.
            end if
        end for
        while K ≠ ∅ do
            Randomly select a nucleotide q ∈ K_c.
            Find the relation of D_2^k and D_2^q according to Theorem 3 and constrain the search region.
            if The requirement for constrained search region is satisfied then
                Randomly select pos_q from the constrained search region. Break.
            end if
            K_c = K_c\{q}.
        end while
    end while
    Remove pos_k from positions of k and remove pos_q from positions of q.
    Add pos_q to positions of k and add pos_k to positions of q.
    Get S^new from S^seq (Permute pos_k and pos_q).
    S^seq = S^new.
end while
The S^seq is the sequence which minimizes our loss function with preset limit.
```

[1] Zhao R, Pei S, and Yau SST, New genome sequence detection via natural vector convex hull method. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020.
[2] Jiao X*, Pei S*, Sun Z, Kang J, and Yau SST. Determination of the nucleotide or amino acid composition of genome or protein sequences by using natural vector method and convex hull principle, Fundamental Research, Vol. 1 (2021), 559-564.

Application 3 (Natural Metric): The Early Transmission of SARS-CoV-2

- The Natural Metric in the SARS-CoV-2 subspace is $D_9(s_1, s_2) = d_1(s_1, s_2) + \frac{1}{2^2} d_2(s_1, s_2) + \frac{1}{2^3} d_3(s_1, s_2) + \cdots + \frac{1}{2^9} d_9(s_1, s_2)$.
- Then we downloaded all reliable genomes from GISAID (144,566 sequences), and compared with RaTG13 (Sequence similarity 96.1%) and RmYN02 (Sequence similarity 93.3%) [1, 2].

- The 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, 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.

Table: The top five SARS-CoV-2 sequences with the shortest metric.

Rank	Information	Metric Value
A. RaTG13 at the complete genome scale		
1	hCoV-19/France/B5434/2020|EPI_ISL_443279|2020-04-01	25311.95
2	hCoV-19/Netherlands/Utrecht_10024/2020|EPI_ISL_454773|2020-03-26	25322.00
3	hCoV-19/USA/VA-DCLS-0392/2020|EPI_ISL_467788|2020-04-19	25324.18
4	hCoV-19/USA/CruiseA-1/2020|EPI_ISL_413606|2020-02-17	25327.38
5	hCoV-19/Sichuan/SC-WCH4-288/2020|EPI_ISL_451390|2020-01-23	25343.60
346	hCoV-19/Wuhan/WH01/2019|EPI_ISL_406798|2019-12-26	25429.13
B. RmYN02 at the complete genome scale		
1	hCoV-19/Singapore/14/2020|EPI_ISL_414380|2020-02-13	28049.28
2	hCoV-19/Singapore/23/2020|EPI_ISL_420100|2020-03-02	28058.59
3	hCoV-19/Singapore/13/2020|EPI_ISL_414379|2020-02-18	28085.93
4	hCoV-19/Singapore/30/2020|EPI_ISL_420107|2020-03-09	28093.80
5	hCoV-19/USA/VA_NIDDL_3216/2020|EPI_ISL_491943|2020-04-16	28094.80
7964	hCoV-19/Wuhan/WH01/2019|EPI_ISL_406798|2019-12-26	28405.08

[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.
[3] 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)

Summary

- In this study, we use pure mathematics concept to study important problems in biology. 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 genome space with a proper metric is a powerful means of determining the phylogenetics and classification of genomes.

- 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) Phylogenetics tree construction; (2) Genome sequence detection; (2) the early transmission of SARS-CoV-2.

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 Nan Sun, Tao Zhou, Ruohan Ren, Hongyu Yu, Leqi Zhao, Mengcen Guan 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).


清華大學
Tsinghua University
Thanks