Fundamental Research 1 (2021) 559–564

Contents lists available at ScienceDirect

Fundamental Research

journal homepage: http://www.keaipublishing.com/en/journals/fundamental-research/

Article

Determination of the nucleotide or amino acid composition of genome or protein sequences by using natural vector method and convex hull principle

Xiaopei Jiao[a,1], Shaojun Pei[a,1], Zeju Sun[a], Jiayi Kang[a], Stephen S.-T. Yau[a,b,*]

[a] Department of Mathematical Sciences, Tsinghua University, Beijing, 100084, China
[b] Yanqi Lake Beijing Institute of Mathematical Sciences and Applications, Beijing, 101408, China

ARTICLE INFO

Article history:
Received 25 April 2021
Received in revised form 10 August 2021
Accepted 12 August 2021
Available online 3 September 2021

Keywords:
Natural vector
Convex hull
New sequences
Optimization

ABSTRACT

Although with the continuous development of sequencing technology, the number of genome and protein sequences has grown rapidly, these sequences are only a small part of nature. Biologically, it is still a challenging and important problem to detect and predict some new genome or protein sequences based on real sequence data, which motivates us to solve the problem mathematically. The first step to predict the new sequences is determining the nucleotide or amino acid composition of them. In this paper, we apply natural vector method and convex hull principle to determine the nucleotide or amino acid composition of new genome or protein sequences. Our algorithm is based on optimization strategy. The SARS-CoV-2 genome and protein datasets are used to verify the feasibility of our algorithm. Numerical experiments show that our algorithm can detect and predict possible number of each nucleotide or amino acid of genome and protein sequence with respect to the second order natural vectors.

1. Introduction

With the continuous development of sequencing technology, the genome and protein sequences of many species have been well sequenced. The genome and protein sequences play a key role in biological organisms because they can determine some internal factors, such as adaption, evolution and phenotypes of life. Thus, these sequences can serve as identification for different species. To distinguish genome and protein sequences of different species, many sequence coding methods are proposed to analyze the similarity of the sequences in the past decades [1–4]. We proposed a fast and accurate alignment-free method called natural vector method [5]. It associates a high dimensional vector in Euclidean space to a genome or protein sequence, which performs well in different biological species, especially in viruses.

The severe acute respiratory syndrome COVID-19 was discovered on December 31, 2019 in China and is caused by a new type of coronavirus called SARS-CoV-2. Subsequently, many COVID-19 cases were reported in the almost all over the world. A lot of SARS-CoV-2 genome and protein data have been measured and many different subtypes of the virus have been found [6]. However, the sequences of SARS-CoV-2 mutate consistently [7]. Due to this fact, detecting the variants of SARS-CoV-2 is very meaningful to study the properties of viruses and control the pandemic of COVID-19. The first step to predict the variants is determining their nucleotide or amino acid composition. Therefore, we want to propose a data-driving algorithm to determine the nucleotide or amino acid composition mathematically. The nucleotide or amino acid composition detected by our mathematical method will be beneficial and give a guide to biological experimental research.

Based on natural vector method, the sequences are transformed as the points in the Euclidean space. And the first component of a natural vector is the number of each nucleotide or amino acid. Then the convex hulls of natural vectors from different species are constructed to study their relationships [8]. Mathematically, convex hull is the smallest convex polygon formed by a set of points, where the convex polygon encloses all of the points in the set. The convex hull principle shows that the convex hulls of the natural vectors of the sequences from different species are disjoint [9]. The genome or protein sequences with natural vectors in the same convex hull share similar properties and are highly likely from the same species. If new genome or protein sequences whose natural vectors fall in the convex hull formed by known ones, those sequences are likely from the same species as known ones. Based on this principle, we propose an algorithm to predict the number of each nucleotide or amino acid of new sequences. The genome and protein sequences of SARS-CoV-2 are used to test the algorithm and we get the possible nucleotide or amino acid composition of genome or protein of SARS-CoV-2.

* Corresponding author.
E-mail address: yau@uic.edu (S.S.-T. Yau).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.fmre.2021.08.010

2. Problem formulation

2.1. Natural vector method

$$\vec{v} = (n_K, \mu_K, D_K^2, \cdots, D_K^m), \tag{1}$$

where the first component $\{n_K, K \in \mathcal{K}\}$ is number of K in the sequence and:

$$\begin{aligned} \mu_K &= \frac{1}{n_K}\sum_{i=1}^{n_K} S[K][i], \\ D_K^j &= \sum_{i=1}^{n_K} \frac{\left(S[K][i]-\mu_K\right)^j}{n_K^{j-1} n^{j-1}}, \quad 2 \le j \le m, m \le \min\left(n_K\right) \end{aligned} \tag{2}$$

where $S[K][i]$ represents the i-th position of genome composition K in the sequence. For example, for DNA sequence, the second order natural vector is:

$$\vec{v} = (n_A, n_C, n_T, n_G, \mu_A, \mu_C, \mu_T, \mu_G, D_A^2, D_C^2, D_T^2, D_G^2) \tag{3}$$

Definition 1. (Integer point) Integer point of a sequence is a vector $(n_K, K \in \mathcal{K})$ which consists of the numbers of nucleotides for a DNA sequence or amino acids for a protein sequence.

For example, for a DNA sequence, corresponding integer point is $(n_A, n_C, n_T, n_G) \in \mathbb{R}^4$ which consists of the numbers of nucleotides. For a protein sequence, corresponding integer point is $(n_A, n_R, n_N, \cdots, n_V) \in \mathbb{R}^{20}$ which consists of the numbers of amino acids. Therefore, the integer point of a DNA or protein sequence is essentially first 4 or 20 dimensional components of its corresponding natural vector.

Next we introduce some elementary properties satisfied by natural vector. Some basic computations show that all moments of natural vector should satisfy the following conditions:

$$\begin{aligned} \sum_{K\in\mathcal{K}} n_K \mu_K &= \sum_{j=1}^{n} j, \\ \sum_{K\in\mathcal{K}}\sum_{i=1}^{n_K} S[K][i]^l &= \sum_{K\in\mathcal{K}}\sum_{p=1}^{l-1} D^{l-p+1} K n^{l-p} n_K^{l-p} \mu_K^{p-1} + \sum_{K\in\mathcal{K}} n_K \mu_K^l \\ &= \sum_{j=1}^{n} j^l, \ 2 \le l \le m. \end{aligned} \tag{4}$$

For convenience, we denote second equation in (4) as $F(\vec{\alpha}, l) = \sum_{j=1}^{n} j^l$.

Next we explain the relationship between a sequence and its corresponding natural vector. If we select a biological sequence, we can calculate the corresponding m-order natural vector by the formula (1) and (2). Conversely, by the knowledge of statistics, if all the moments up to order $m \le \min(n_K)$ are obtained, its corresponding discrete sequence is fully determined. Hence we show that a biological sequence (such as DNA or protein) is corresponding to an order-m natural vector uniquely.

Definition 2. ([10]) Natural vector space up to order m is defined as:

$$\begin{aligned} \mathcal{NV} = \{\overline{nv} = (n_K, \mu_K, D_K^2, \cdots, D_K^m) | K \in \mathcal{K}, n_K \in \mathbb{Z}, \sum_{K\in\mathcal{K}} n_K \mu_K = \sum_{j=1}^{n} j, \\ \sum_{K\in\mathcal{K}}\sum_{p=1}^{l-1} \binom{m}{l-1} D^{l-p+1} K n^{l-p} n_K^{l-p} \mu_K^{p-1} + \sum_{K\in\mathcal{K}} n_K \mu_K^l \\ = \sum_{j=1}^{n} j^l, 2 \le l \le m.\} \end{aligned} \tag{5}$$

which is a set of all natural vectors satisfying up to m-th order moment condition.

2.2. Convex hull principle

Based on the natural vector method, each sequence is transformed as a vector in the Euclidean space. Then convex hulls of natural vectors from different biological groups can be constructed. Tian et al. [8] and Zhao et al. [9] show the convex hulls of natural vectors from different biological groups are disjoint, called convex hull principle. So based on the principle, the problem of detecting some new genome or protein sequence from a biology group is turned into searching some new sequences whose natural vectors are in the convex hull of it.

We consider one biology group consisting of N sequences (genome or protein). For every sequence, the corresponding natural vector is denoted by $\vec{v}_i = (n_{K,i}, \mu_{K,i}, D_{K,i}^2, \cdots, D_{K,i}^m)$, $1 \le i \le N$, where $K \in \mathcal{K}$. We define the convex hull generated by $\{\vec{v}_1, \cdots, \vec{v}_N\}$ as below:

$$\mathrm{Conv}(\vec{v}_1, \cdots, \vec{v}_N) := \left\{ \sum_{i=1}^{N} \alpha_i \vec{v}_i \middle| \alpha_i \ge 0, \sum_{i=1}^{N} \alpha_i = 1 \right\} \tag{6}$$

where $\vec{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$ denotes the coefficient of convex linear combination.

We want to find a natural vector $\vec{v}_\alpha = (n_{K,\alpha}, \mu_{K,\alpha}, D_{K,\alpha}^2, \cdots, D_{K,\alpha}^m) \in \mathcal{NV}$ contained in the $\mathrm{Conv}(\vec{v}_1, \cdots, \vec{v}_N)$. We denote $n_\alpha = \sum_{K\in\mathcal{K}} n_{K,\alpha}$.

We denote following vectors:

$$\begin{aligned} \vec{n}_K &= (n_{K,1}, n_{K,2}, \cdots, n_{K,N})^T, \\ \vec{\mu}_K &= (\mu_{K,1}, \mu_{K,2}, \cdots, \mu_{K,N})^T, \\ \vec{D}_K^l &= (D_{K,1}^l, D_{K,2}^l, \cdots, D_{K,N}^l)^T, 2 \le l \le m. \end{aligned} \tag{7}$$

Then $\vec{v}_\alpha \in \mathrm{Conv}(\vec{v}_1, \cdots, \vec{v}_N)$ is equivalent to existence of $\vec{\alpha}$ such that:

$$\begin{aligned} n_{K,\alpha} &= \vec{\alpha} \cdot \vec{n}_K, \\ \mu_{K,\alpha} &= \vec{\alpha} \cdot \vec{\mu}_K, \\ D_{K,\alpha}^l &= \vec{\alpha} \cdot \vec{D}_K^l, 2 \le l \le m. \end{aligned} \tag{8}$$

Considering $\vec{v}_\alpha \in \mathcal{NV}$ and Definition 2, $\vec{v}_\alpha$ must satisfy following equations:

$$\begin{aligned} n_{K,\alpha} &= \vec{\alpha} \cdot \vec{n}_K \\ \sum_{K\in\mathcal{K}} n_{K,\alpha}(\vec{\alpha} \cdot \vec{\mu}_K) &= \sum_{j=1}^{n_\alpha} j \\ F(\vec{\alpha}, l) &= \sum_{j=1}^{n_\alpha} j^l, 2 \le l \le m \end{aligned} \tag{9}$$

Definition 3. (Admissible integer point) If natural vector $\vec{v}_\alpha$ is contained in $\mathrm{Conv}(\vec{v}_1, \cdots, \vec{v}_N)$, (i.e., Eq. 9 hold), then for DNA sequence, the admissible point satisfying q-th moment condition is formed by the first 4-dimensional components of $\vec{v}_\alpha$ (i.e., $\{n_{K,\alpha}, K \in \mathcal{K}\}$). For protein sequence, the admissible point is formed by the first 20 dimensional components of $\vec{v}_\alpha$.

In this paper, we will focus on the second order natural vector space, i.e., $m = 2$ in Eq. 9. For second order natural vector space, the third and fourth equations in Eq. 9 become:

$$\begin{aligned} \sum_{K\in\mathcal{K}} n_K \left(\sum_{i=1}^{N} \alpha_i \mu_{K,i}\right) &= \frac{1}{2} n_\alpha (n_\alpha + 1) \\ \sum_{K\in\mathcal{K}} n n_K \sum_{i=1}^{N} \alpha_i D_2^{K,i} + n_K \left(\sum_{i=1}^{N} \alpha_i \mu_{K,i}\right)^2 &= \frac{1}{6} n_\alpha (n_\alpha + 1)(2n_\alpha + 1) \end{aligned} \tag{10}$$

In order to detect possible new natural vectors in convex hull, the first step is to find potential admissible integer points. Eq. 9 are hard to solve directly because this is a large scale underdetermined equations of $\vec{\alpha}$. Therefore, for a particular integer point $\{n_{K,\alpha}, K \in \mathcal{K}\}$, we consider to transform the existence of solution of Eq. 9 to following optimization

problem:

$$\begin{cases} \min_{\vec{\alpha}}(\text{or}\max_{\vec{\alpha}}) & \sum_{K\in\mathcal{K}} n_\alpha n_{K,\alpha} \sum_{i=1}^N \alpha_i D_2^{K,i} + n_{K,\alpha}\left(\sum_{i=1}^N \alpha_i \mu_{K,i}\right)^2 \\ s.t. & \sum_{i=1}^N \alpha_i = 1 \\ & \sum_{i=1}^N \alpha_i n_{K,i} = n_{K,\alpha} \quad \forall K \in \mathcal{K} \\ & \sum_{K\in\mathcal{K}} n_{K,\alpha}\left(\sum_{i=1}^N \alpha_i \mu_{K,i}\right) = \frac{n_\alpha(n_\alpha+1)}{2} \\ & 0 \le \alpha_i \le 1, \quad i = 1, 2, \cdots, N \end{cases} \tag{11}$$

If target value $\frac{1}{6}n_\alpha(n_\alpha+1)(2n_\alpha+1)$ lies between minimum and maximum, then by Definition 3, $\{n_{K,\alpha}, K \in \mathcal{K}\}$ is an admissible integer point satisfying second moment condition.

In the following, we will take two main steps to solve the integer point detection problem. Step 1 is to find integer points in the convex hull of integer points of known sequences (dataset). Step 2 is to check integer points found in step 1 and verify admissible integer points by solving optimization problem Eq. 11. We will discuss the details of these two steps in the next section.

3. Methods

3.1. Find integer points in the convex hull of integer points of known sequences

In order to find admissible integer points, first we need to find integer points contained in the convex hull generated by known integer $\{n_K, K \in \mathcal{K}\}$ in dataset. So we present following two methods.

Geometric Method [12]: Consider the convex hull generated by a set of points $A = \{a_i\}_{i=1}^N \subset \mathbb{R}^n$. Let $\Gamma \subset \mathbb{R}^n$ be an arbitrary face of the convex hull of point set A, then Γ is a hyper plane, which can be denoted by $\Gamma = \{x \in \mathbb{R}^n : a^\top x + b = 0, a \in \mathbb{R}^n, b \in \mathbb{R}\}$.

Let $y_0 = \frac{1}{N}\sum_{i=1}^N a_i$ be the center of the point set A, which is in the convex hull of A. Therefore, for any point $x_0 \in \mathbb{R}^n$, if x_0 is inside the convex hull of A, then x_0 and y_0 must be on the same side of hyper plane Γ, namely:

$$(a^\top x_0 + b) \cdot (a^\top y_0 + b) \ge 0 \tag{12}$$

Checking Eq. 12 for each face Γ, we can conclude whether the point x_0 is inside the convex hull of A.

Linear Programming Method [8]: Consider the same case with geometric method. In order to test whether x_0 is in convex hull, we consider the following linear programming:

$$\begin{cases} \min_\lambda(\text{or}\max_\lambda) & \sum_{i=1}^N \lambda_i a_i, \\ s.t. & \sum_{i=1}^N \lambda_i = 1, \lambda_i \ge 0 \end{cases} \tag{13}$$

If target value (component of x_0) lies between minimum and maximum of Eq. 13, x_0 is in convex hull.

For genome sequences, the integer point components are four dimensional. The relationship between an integer point and the convex hull can be easily detected with the geometric method.

For protein sequences, the integer point components are 20 dimensional, and it is hard to apply the geometric method directly. In order to reduce computational cost furthermore, we first separate the 20 integer components in the natural vector into 3 parts, $(\mathcal{K}_1, \mathcal{K}_2, \mathcal{K}_3)$, with each part 7, 7 and 6 amino acids, respectively. For each part, the points we need to check are all points in the rectangle $\prod_{K\in\mathcal{K}_i}[m_K, M_K]$, $i = 1, 2, 3$, where m_K, M_K are the minimal and maximal number of amino acid K in one of the known proteins. For each part, we use geometric method to find all the integer points in the convex hull. Next, the Cartesian product of these parts can be used. Finally we use linear programming to check these points. This decomposition method will have higher efficiency than using linear programming directly.

3.2. Verify admissible integer point based on optimization theory

In this section, we will focus on Eq. 11 and provide corresponding optimization theory and method. By some basic computation, Eq. 11 can be written as standard quadratic programming form:

$$\text{QP}: \quad \begin{cases} \min_x(\text{or}\max_x) & \frac{1}{2}x^T Ax + x^T c \\ s.t. & Bx - b = 0 \\ & Mx \le f \end{cases} \tag{14}$$

where $x \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}, c \in \mathbb{R}^n, B \in \mathbb{R}^{m\times n}, M \in \mathbb{R}^{p\times n}, x \in \mathbb{R}^m$. Feasible set is defined by $\Omega = \{x \in \mathbb{R}^n | Bx - b = 0, Mx \le f\}$. Correspondingly, we have:

$$\begin{cases} A = 2\sum_{K\in\mathcal{K}} n_{K,\alpha}\vec{\mu}_K\vec{\mu}_K^T, c = n_\alpha \sum_{K\in\mathcal{K}} n_{K,\alpha}\vec{D}_K, \\ B = (\vec{n}_K, \vec{1}, \sum_{K\in\mathcal{K}} n_{K,\alpha}\vec{\mu}_K)^T, b = (n_{K,\alpha}, 1, \frac{1}{2}n_\alpha(n_\alpha+1))^T, \\ M = (I_n, -I_n)^T, f = (1, \cdots, 1, 0, \cdots, 0)^T, \end{cases} \tag{15}$$

where I_n is an identity matrix. Notice that A is a positive semidefinite matrix, i.e., $A \ge 0$. Next we will analyze the property of optimal solution of (14). Lagrangian function can be calculated as below:

$$L(x, \lambda, \mu) = \frac{1}{2}x^T Ax + x^T c + (Bx - b)^T\lambda + (Mx - f)^T\mu, \ \lambda \in \mathbb{R}^m, \ \mu \in \mathbb{R}_+^p \tag{16}$$

Dual function is defined by $g(\lambda, \mu) = \min_{x\in\mathbb{R}^n} L(x, \lambda, \mu)$ and dual optimization problem is as below:

$$\text{Dual problem}: \quad \begin{cases} \max_{\lambda,\mu} & g(\lambda, \mu) \\ s.t. & \mu \in \mathbb{R}_+^p, \lambda \in \mathbb{R}^m \end{cases} \tag{17}$$

Based on dual theory, dual problem always provides a lower bound for primal problem.

We focus on the case when optimal solution of dual problem is equal to one of primal problem (strong duality). In this case, well known Karush-Kuhn-Tucker (KKT) condition [13,14] holds.

Theorem 3.1. *(Karush-Kuhn-Tucker)[13, 14] For an optimization problem:*

$$\begin{cases} \min_x & f(x) \\ s.t. & g^{(i)}(x) = 0, i = 1, \cdots, n \\ & h^{(j)}(x) \le 0, j = 1, \cdots, m \end{cases} \tag{18}$$

with strong duality, its local optimal solution satisfies:

$$\begin{cases} \frac{\partial f}{\partial x_k} + \sum_{i=1}^n \lambda_i \frac{\partial g^{(i)}(x)}{\partial x_k} + \sum_{j=1}^m \mu_j \frac{\partial h^{(j)}(x)}{\partial x_k} = 0, k = 1, \cdots, l \\ g^{(i)}(x) = 0, i = 1, \cdots, n \\ \mu_j h^{(j)}(x) = 0, j = 1, \cdots, m \\ \mu_j \ge 0, j = 1, \cdots, m \end{cases} \tag{19}$$

where condition corresponding to inequality constraints $\mu_j h^{(j)}(x) = 0, j = 1, \cdots, m$ is called complementary slackness.

Next we prove optimal solution of Eq. 14 satisfies KKT condition.

Theorem 3.2. *Any local minimizer $x^* \in \Omega$ of the problem Eq. 14 satisfies the KKT conditions.*

Proof. Since constraints of problem Eq. 14 consist of affine functions, based on Theorem 3.3 in the result of Eustaquio et al. [11], then x^* satisfies the KKT conditions.

Theorem 3.2 shows the optimization problem corresponding to integer point detection has good mathematical property.

Based on strong duality property in Theorem 3.2, we choose different optimization algorithms to solve the problem Eq. 14. For minimum, since A is a positive semidefinite matrix, this is a convex optimization.

Table 1
An example to check first order condition of integer points. In column 'Result', '1' represents the integer point satisfies first moment condition, otherwise '0'.

(n_A, n_C, n_T, n_G)	Min Value	Target Value	Max Value	Result
(8800,5379,9420,5731)	432933349	432930025	432936851.8	0
(8800,5382,9423,5731)	433778037	433783785	433785876	1

Table 2
Proportion of interger points satisfying first moment condition.

Trail	1	2	3	4	5
Proportion	65.8%	68.5%	67.9%	66.5%	68.3%

Active set method can be applied. For maximum, this is a non-convex optimization problem, sequential quadratic programming can be applied.

4. Results

4.1. Integer point detection in the convex hull of genome sequences

4.1.1. Dataset pretreatment and numerical algorithm

We download all the complete genome sequences of SARS-CoV-2 from GISAID until July 19, 2020 (https://www.gisaid.org/). To ensure the accuracy of analysis, the low-quality sequences which contain letters other than A, C, G and T are eliminated from the dataset. All the sequences are in 'vertex.fasta' which contains 8522 genome sequences. To reduce the calculation load, we only use a subset of the whole dataset to construct the convex hull. We design following pretreatment algorithm to make the convex hull large enough so that new admissible integer points can be found more easily.

By Algorithm 1, we select 1030 known SARS-CoV-2 natural vectors from whole dataset, i.e., a subset of dataset.

In the following, we find integer points in the convex hull generated by known integer $\{n_K, K \in \mathcal{K}\}$ by method in Section 3.1. In our numerical experiments, we will verify integer points in which component 'A' is 8800.

Next we introduce numerical algorithm. In order to calculate the optimization problem Eq. 11, we first check whether the feasible set consisting of constraints is empty (first moment condition). Therefore we propose following Algorithm 2.

After checking feasible set, we will solve the problem Eq. 11. By Section 3.2, we can use active set method for minimum problem of Eq. 14 and sequential quadratic programming for maximum. In summary, we obtain following Algorithm 3.

4.1.2. Admissible integer points in convex hull of genome sequences

In this subsection, we will show some numerical results of integer point detection. Through Algorithm 2, we can screen integer points satisfying first moment condition effectively. In Table 1, we show an example.

In order to test robustness of the Algorithm 2, we run this algorithm for 5 times. In every trial, number of random integer points is set 1000. The proportion of integer points satisfying first moment condition is listed in Table 2.

The results are stable and it shows that the integer points satisfying the first moment condition is many enough. We can estimate the total number of integer points satisfying first moment condition is $\approx \frac{(65.8+68.5+67.9+66.5+68.3)}{5\times100} \times 304760 = 205408$ and this is still a large number.

In the following, we make numerical experiments by Algorithm 3. Let Toler $= 10^{-10}$ and MaxIterNum $= 10^4$. Result shows that there exist 11 integer points satisfying second order moment condition. In the following Table 3, we list them.

Table 3
New detected integer points satisfying second order moment condition and components of reference sequence.

Integer point	n_A	n_C	n_T	n_G
1	8800	5406	9429	5757
2	8800	5404	9427	5759
3	8800	5422	9454	5790
4	8800	5403	9434	5757
5	8800	5405	9424	5761
6	8800	5406	9425	5757
7	8800	5406	9423	5755
8	8800	5406	9426	5755
9	8800	5407	9419	5763
10	8800	5404	9424	5761
11	8800	5404	9427	5758
Reference	8954	5492	9594	5863

Table 4
Nine integer points in the 20 dimensional convex hull.

Integer point	W	M	C	R	H	Q	V	F	I	T
New detected one	12	14	40	42	17	62	97	77	76	97
Reference	12	14	40	42	17	62	97	77	76	97
Integer point	G	P	K	L	A	S	E	D	Y	N
New detected one	**82**	58	61	108	79	99	48	61	54	88
Reference	**83**	58	61	108	79	99	48	61	54	88

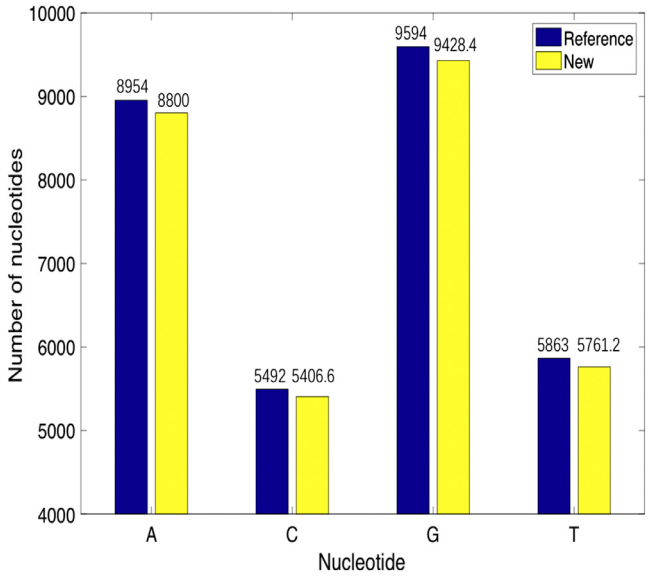

Fig. 1. Integer point comparison. Blue bar denotes nucleotide number of reference sequence in the dataset. Yellow bar denotes average integer point of new detected 11 admissible points. The 'y-axis' represents the corresponding nucleotide number. Result shows that there are high similarities between reference sequence and new detected integer points. Biologically, new detected integers may have similar properties with known sequences in dataset.

Total time consumed is about 13.8 hours. The proportion of admissible integers satisfying second moment condition is approximately 0.11%. Upper bound of admissible integer points can be roughly estimated (≈ 335). It can be seen that with considering higher order moment condition, number of admissible integer points will rapidly decrease. In order to show difference between a reference sequence and detected integer points, we give an illustration in Fig. 1.

Furthermore, in order to verify the feasibility of our detection method, we make following numerical experiments. By detection algorithm proposed in this section, we can verify integer point (8792,5389,9466,5761) is an admissible integer point falling in the convex hull.

In the database of SARS-CoV-2 DNA sequences, we can find there exists a real sequence whose number of nucleotides A, C, T, G is indeed this integer point. The sequence information is '$hCoV-19/Australia/NSW342/2020|\ EPI\_ISL\_451604|2020-03-29$'. Complete genome sequence is listed in Appendix 2. This shows the feasibility and robustness of our integer detection algorithm and our method can detect real sequence and potential new sequence.

4.2. Integer point detection in the convex hull of protein sequences

4.2.1. Dataset pretreatment and numerical algorithm

The spike protein, which is critical for SARS-CoV-2 infection and differs CoV types, is responsible for ACE2 receptor binding and membrane fusion. So the amino acid sequences of spike protein of SARS-CoV-2 are downloaded for analysis. All the sequences are in 'protein_S_1.fasta' which contains 15586 amino acid sequences of spike proteins from the SARS-CoV-2. However, most of the amino acid sequences are identical and there are only 841 different kinds of amino acid sequences in the dataset. Different kinds of proteins are still similar to each other. In fact, 10046 of 15586 proteins has the same amino acid sequence, which can be regarded as a reference sequence. Other 840 types of amino acid sequences can be regarded as a mutation from the reference sequence at several amino acid sites. For example, the only difference between the second most abundant amino acid sequence, 'D614G', which exists 3124 times in the dataset, and the reference sequence is that the 614th amino acid changes from 'D' to 'G'. Therefore, the convex hull generated by sequence in dataset is very small.

Our main goal is to find admissible integer points in the convex hull of protein sequences. The detection algorithm is similar to Section 4.1 and is summarized as below.

4.2.2. Admissible integer points in convex hull of protein sequences

For protein sequences dataset, the range of the number of each amino acid is shown in Appendix 1. By the method mentioned in Section 3.1, we separate the 20 amino acids into three parts and detect each part with geometric method. Result shows that only 9 integer points in the convex hull generated by known integer points. All 9 integer points are shown in Appendix 1.

For these 9 integer points, we calculate the optimization problem Eq. 11 by Algorithm 2 and 3. Result shows that there exists unique admissible integer point satisfying second order moment condition. In Fig. 2, we show new detected admissible integer point and corresponding range of each component of known sequences. The new detected admissible integer point is highly similar to integer point of reference sequence except mild difference in amino acid 'G'.

Algorithm 1 Data pretreatment

Step 1. Calculate minimum and maximum in first 4 dimensional components of known natural vectors.
A: 8718~9077,C: 5338~5690,T: 9346~9775,G: 5702~6101.
Step 2. Determine the scope of first 4 dimensional integer, which contains minimum and maximum in first 4 dimensional integers of known natural vectors.
A: 8718~8795, 8955~9077
C: 5338~5393, 5491~5690
T: 9346~9467, 9601~9775
G: 5702~5762, 5864~6101
Step 3. Select known natural vectors satisfying Step 2 condition.

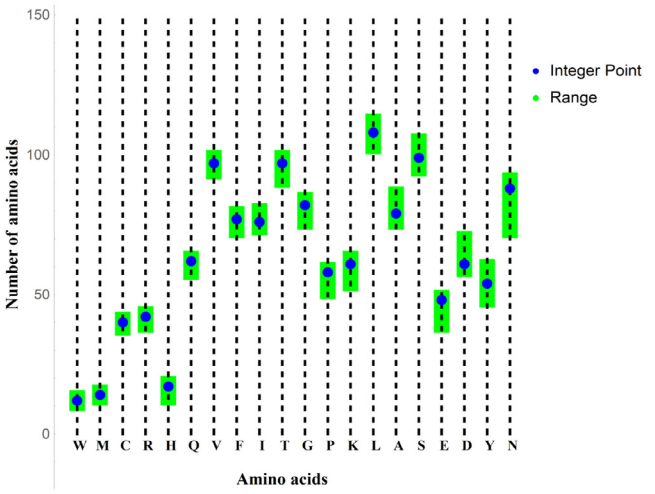

Fig. 2. New detected admissible integer point in the convex hull of protein sequences. 'x'-axis represents different amino acid. 'y'-axis represents the corresponding amino acid number. For illustration, the range of 'y'-axis is rescaled. Blue points are admissible integer point. Green bar denotes the scope of number of each amino acid in the dataset.

Algorithm 2 Verification of first order moment condition

Step 1. Randomly select 1000 integer points in dataset.
Step 2. Calculate following optimization problem and obtain its minimum and maximum.

$$\begin{cases} \min_\alpha(\text{or}\max_\alpha) & \sum_{K\in\mathcal{K}} n_{K,\alpha}\mu_K \\ s.t. & \alpha\cdot n_K = n_{K,\alpha} \\ & 0\le\alpha_i\le 1, \quad \forall 1\le i\le N \\ & \sum_{i=1}^N \alpha_i = 1 \end{cases}$$

Step 3. Check whether following condition holds,

$$\sum_{j=1}^{n_\alpha} j \in [\min_\alpha \sum_{K\in\mathcal{K}} n_{K,\alpha}\mu_K, \max_\alpha \sum_{K\in\mathcal{K}} n_{K,\alpha}\mu_K]$$

If above condition holds, corresponding particular integer point allows nonempty feasible set and can be verified second order moment condition further. Otherwise, this integer point can be excluded.

Algorithm 3 Verification of Second order moment condition

Step 1. Import 1030 genome sequences information and calculate corresponding natural vectors. Next, we choose integers in which $n_{A,\alpha} = 8800$. Number of integer points is 304760. Set tolerance 'Toler' and maximum iteration number 'MaxIterNum'.
Step 2. If IterNum>MaxIterNum, exit the program. Otherwise, randomly select an integer point $(n_{A,\alpha}, n_{C,\alpha}, n_{T,\alpha}, n_{G,\alpha})$ among 304760 items.
Step 3. Calculate optimization problem (20). IterNum = IterNum + 1. If target value $\frac{n_\alpha(n_\alpha+1)}{2}$ lies between minimum and maximum of (20), it turns to next step. Otherwise, return to **Step 2.**
Step 4. Calculate (14) and obtain the minimum and maximum of $F(x)$. If the following condition holds, output corresponding integer point $(n_{A,\alpha}, n_{C,\alpha}, n_{T,\alpha}, n_{G,\alpha})$.

$$\sum_{j=1}^{n} j^2 \in [\min F(x), \max F(x)]$$

$$\frac{\|B\cdot x_1 - b\|_2}{\|b\|_2} + \frac{\|B\cdot x_2 - b\|_2}{\|b\|_2} < \text{Toler}$$

where x_1, x_2 represent minimum point and maximum point calculated by active set method and sequential quadratic programming.

Algorithm 4

Step 1. Find all integer points in the convex hull generated by integer points of known sequences.
Step 2. Check the optimization problem (11) using the same method in Section 4.1 and verify admissible integer points of protein sequences.

5. Conclusion

With the development of biological technologies, more and more genome sequences are measured. However, for many species, only a small part of genome sequences are known and studied. For example, there appear some new mutations in SARS-CoV-2 consistently. So if some new, undiscovered genome sequences can be detected in advance, scientists can better and more comprehensively conduct drug research on the virus. In order to find potential new genome sequences based on known sequences dataset, it is the first and important step to find the possible number of compositions of a potential sequence.

The natural vector method can describe the DNA or protein sequence mathematically. Convex hull principle can find quantitative relationship between different sequences. In our paper, in order to solve detection problem in the convex hull, we introduce optimization method which largely reduces the difficulty of solving the equations directly. Optimization has theoretical support and is a quick, efficient and robust algorithm to verify the potential number of genome compositions in DNA or protein. Our detection algorithm can be used in different sequence datasets theoretically. In this paper, we use DNA and protein sequence datasets of SARS-CoV-2 to verify robustness of algorithm for different datasets numerically.

However, our work still has some limitations and challenges remain open to solve in the future. First, our algorithms only determine 4-dimensional integer points by the mean positions and the second order of central moments of natural vector. Higher order central moments can be used and the mean positions can be determined in the future work. As well, this work provides a new algorithm to determine the number of each nucleotide or amino acid of sequences, but the distributions of nucleotides or amino acids in the sequences cannot be determined based on current proposal, which need further to study.

Declaration of Competing Interest

The authors declare that they have no conflict of interest in this work.

Acknowledgements

This work is supported by National Natural Science Foundation of China (Grant No. 11961141005), Tsinghua University Spring Breeze Fund (Grant No. 2020 Z99CFY044), Tsinghua University start-up fund, and Tsinghua University Education Foundation fund (Grant No. 042202008). Professor Stephen Shing-Toung Yau is grateful to the National Center for Theoretical Sciences for providing an excellent research environment while part of this research was done.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.fmre.2021.08.010.

References

[1] G.E. Sims, S.R. Jun, G.A. Wu, et al., Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions, Proceedings of the National Academy of Sciences 106 (2009) 2677–2682.
[2] G.E. Sims, S.R. Jun, G.A. Wu, et al., Whole-genome phylogeny of mammals: evolutionary information in genic and non-genic regions, Proceedings of the National Academy of Sciences 106 (2009) 17077–17082.
[3] M.R. Kantorovitz, G.E. Robinson, S. Sinha, A statistical method for alignment-free comparison of regulatory sequences, Bioinformatics 23 (2007) 249–255.
[4] S. Vinga, J. Almeida, Alignment-free sequence comparison: a review, Bioinformatics 19 (2003) 513–523.
[5] M. Deng, C. Yu, Q. Liang, et al., A novel method of characterizing genetic sequences: Genome space with biological distance and applications, PLoS ONE 6 (3) (2011) e17293.
[6] X. Tang, C. Wu, X. Li, et al., On the origin and continuing evolution of SARS-cov-2, National Science Review 7 (2020) 1012–1023.
[7] P. Wang, M.S. Nair, L. Liu, et al., Antibody resistance of SARS-cov-2 variants b.1.351 and b.1.1.7, Nature, 2021. DOI:10.1038/s41586-021-03398-2
[8] K. Tian, X. Zhao, S.S.T. Yau, Convex hull analysis of evolutionary and phylogenetic relationships between biological groups, Journal of theoretical biology 456 (2018) 34–40.
[9] X. Zhao, K. Tian, R.L. He, S.S.T. Yau, Convex hull principle for classification and phylogeny of eukaryotic proteins, Genomics 111 (2019) 1777–1784.
[10] R.Z. Zhao, S.J. Pei, S.S.T. Yau, New genome sequence detection via natural vector convex hull method, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020). Doi: 10.1109/TCBB.2020.3040706
[11] R.G. Eustaquio, E.W. Karas, A.A. Ribeiro, Constraint qualifications for nonlinear programming, Federal University of Parana, 2008.
[12] F.P. Preparata, M. Shamos, Computational Geometry, 1st ed, Springer-Verlag, New York, 1985.
[13] H. Kuhn, A. Tucker, in: Nonlinear programming In Proceedings of 2nd Berkeley symposium, (pp. 481-492). Berkeley: University of California Press, 1951.
[14] W. Karush, Minima of functions of several variables with inequalities as side constraints, M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, 1939.

Xiaopei Jiao received the B.S. degree from the Zhiyuan college, Shanghai Jiao Tong University, Shanghai, China, in 2017. He is currently pursuing the Ph. D. degree in Applied Mathematics from the Department of Mathematical Sciences, Tsinghua University, Beijing, China. His research interests include nonlinear filtering, estimation algebras, bioinformatics.

Shaojun Pei received the B.S. degree in the School of Mathematical Sciences, Beijing Normal University, Beijing, China, in 2017. She is currently pursuing the Ph.D. degree in Applied Mathematics from the Department of Mathematical Sciences, Tsinghua University, Beijing, China.

Stephen S. T. Yau received the Ph.D. degree in mathematics from the State University of New York at Stony Brook, NY, USA in 1976. In 2012, he joined Tsinghua University, Beijing, China, where he is a full-time professor in the Department of Mathematical Sciences. His research interests include bioinformatics, computational biology, nonlinear filtering, complex algebraic geometry, CR geometry and sigularities theory. He was awarded the Sloan Fellowship in 1980, the Guggenheim Fellowship in 2000, and the AMS Fellow Award in 2013.