genes

MDPI

Article

Classification of Protein Sequences by a Novel Alignment-Free Method on Bacterial and Virus Families

Mengcen Guan [1], Leqi Zhao [1] and Stephen S.-T. Yau [1,2,*]

[1] Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China
[2] Yanqi Lake Beijing Institute of Mathematical Sciences and Applications (BIMSA), Huairou District, Beijing 101400, China
* Correspondence: yau@uic.edu

Abstract: The classification of protein sequences provides valuable insights into bioinformatics. Most existing methods are based on sequence alignment algorithms, which become time-consuming as the size of the database increases. Therefore, there is a need to develop an improved method for effectively classifying protein sequences. In this paper, we propose a novel accumulated natural vector method to cluster protein sequences at a lower time cost without reducing accuracy. Our method projects each protein sequence as a point in a 250-dimensional space according to its amino acid distribution. Thus, the biological distance between any two proteins can be easily measured by the Euclidean distance between the corresponding points in the 250-dimensional space. The convex hull analysis and classification perform robustly on virus and bacteria datasets, effectively verifying our method.

Keywords: accumulated natural vector; convex hull method; proteins; alignment-free; classification

Citation: Guan, M.; Zhao, L.; Yau, S.S.-T. Classification of Protein Sequences by a Novel Alignment-Free Method on Bacterial and Virus Families. *Genes* **2022**, *13*, 1744. https://doi.org/10.3390/genes13101744

Academic Editor: Quan Zou

Received: 28 August 2022
Accepted: 23 September 2022
Published: 27 September 2022

1. Introduction

Proteins, the highly complex substances that are the basic organic matter of life, have been a hot topic for bioinformatics [1]. Proteins play different roles in the processes of life [2], such as enzymes [3]. Proteins carry out the duties specified by gene information. Thus, the research on proteins and protein sequences can reveal the evolutionary relationships of different species. From the molecular biology perspective, a protein sequence is a series of amino acids bonded via peptide bonds, and protein structures vary. There are 20 different types of amino acids that can be combined to make a protein [4]; the sequence of the amino acids determines each protein's unique three-dimensional structure and specific function. These sequences contain the distribution information of 20 types of amino acids, and a single nucleotide polymorphism (SNP) may result in a change in a protein's function [5]. There are many reasons for protein diversity [6]. One reason is the diversity of the sequences; proteins are diverse both within and between families, which makes classifying different virus or bacteria families based on protein sequences reliable.

Studying the relationship among different protein sequences is now a matter of great concern in related research. The methods for sequence similarity analysis commonly depend on a multiple sequence alignment, which usually requires a long computation time to obtain results. Therefore, alignment-free methods are proposed to overcome this ineffectiveness. Currently published alignment-free methods include graphical representation [7,8], probabilistic measure [9,10], k-mer [11,12], etc. Furthermore, sequence vector representation methods without alignment are also popular, such as the moment vector [13] and natural vector [14,15].

In this paper, we propose a novel alignment-free method for protein sequences. The accumulated natural vector for genome sequences is a previously published method that performs well on many datasets [16]. However, it ignores the vast field of protein sequences. Similar to the 18-dimensional vector of nucleotide sequences [16], the accumulated natural

vector of protein sequences we designed also covers the number, average position, variance, and covariance information of amino acids. Detailed information is introduced in the next chapter. Our method not only considers the basic properties of each amino acid but also the covariance between them. Each protein sequence is in one-to-one correspondence with a point in a 250-dimensional space. Subsequently, the biological distance between two protein sequences is measured by the Euclidean distance between the corresponding points. Therefore, this approach can classify sequences into the correct cluster at a lower time cost without reducing accuracy. We also performed a convex hull analysis, which states that the convex hull formed from the same family's protein corresponding points do not intersect with other families' convex hulls. The classification results with high and robust accuracy further validate our methods effectively.

2. Materials and Methods

2.1. Accumulated Natural Vector for Protein Sequences

Since the distribution of amino acids determines a protein sequence, the use of appropriate models to describe the distribution of amino acids is an important issue. Former discrete models, such as the "pseudo amino acid composition" (PseAAC) model [17], have been applied to the prediction of various protein attributes. The information that can be extracted from the distribution of amino acids is very diverse. Our accumulated natural vector method is a natural description of amino acid distribution information. Assume $S = (s_1, s_2, s_3, \cdots, s_N)$ is a protein sequence of length N, i.e., $s_i \in \{A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$, $i = 1, 2, \cdots, N$. Each letter represents a type of amino acid, and there are 20 letters representing 20 types of amino acids. For convenience, the set of these 20 letters is denoted as $\mathcal{A}$. This subsection defines the accumulated natural vector of the protein sequence S.

2.1.1. Related Definitions

We first define the indicator function of these 20 amino acids, respectively:

$$I_\alpha(i) = \begin{cases} 1 & \text{if } s_i = \alpha \\ 0 & \text{if } s_i \neq \alpha \end{cases} \tag{1}$$

where $\alpha \in \mathcal{A}, i = 1, 2, \cdots, N$

Next, we define the accumulated indicator function of each amino acid:

$$\tilde{I}_\alpha(k) = \sum_{i=1}^{k} I_\alpha(i) \qquad \alpha \in \mathcal{A} \tag{2}$$

We have defined the indicator function and accumulated indicator function and noticed that there is an obvious property: n_α, the total amount of the amino acid α in the sequence S, is the last column.

$$n_\alpha = \sum_{i=1}^{N} I_\alpha(i) = \tilde{I}_\alpha(N) \qquad \alpha \in \mathcal{A} \tag{3}$$

Now we define the average position of the amino acid α in the sequence S:

$$\zeta_\alpha = \frac{\sum_{i=1}^{N} \tilde{I}_\alpha(i)}{n_\alpha} \qquad \alpha \in \mathcal{A} \tag{4}$$

For the two different amino acids α and β, define their covariance in the protein sequence S as:

$$cov(\alpha, \beta) = \sum_{i=1}^{N} \frac{(\tilde{I}_\alpha(i) - \theta_\alpha) \times (\tilde{I}_\beta(i) - \theta_\beta)}{n_\alpha \times n_\beta} \tag{5}$$

where $\alpha, \beta \in \mathcal{A}, \theta_\alpha = \sum_{i=1}^{N} \tilde{I}_\alpha(i)/N, \theta_\beta = \sum_{i=1}^{N} \tilde{I}_\beta(i)/N$. Note that the definition of θ_α here is different from the average position above.

Then, the variance of the amino acid α is the special case of $\alpha = \beta$ in (5):

$$D_\alpha = cov(\alpha, \alpha) = \sum_{i=1}^{N} \frac{(\tilde{I}_\alpha(i) - \theta_\alpha)^2}{n_\alpha^2} \qquad \alpha \in \mathcal{A} \tag{6}$$

Thus, we have defined every concept needed in the accumulated natural vector.

2.1.2. Accumulated Natural Vector

Now we can build up the accumulated natural vector of the protein sequence S. The first 20 dimensions describe the amount of 20 amino acids, the second 20 dimensions describe the average positions of the 20 amino acids, and the third 20 dimensions describe the variances of the 20 amino acids. The final $\binom{20}{2} = 190$ dimensions describe the covariances between each two amino acids. The total number of dimensions is 250.

$$\begin{aligned}(&n_A, n_R, \cdots, n_V, \zeta_A, \zeta_R, \cdots, \zeta_V, D_A, D_R, \cdots, D_V,\\ &cov(A,R), cov(A,N), \cdots, cov(Y,V))\end{aligned} \tag{7}$$

For example, take the protein sequence $S = (ARRNADCDCC)$ of length 10. The indicator functions and the accumulated indicator functions are shown in Table 1 and Table 2, respectively. Here, except for A, R, N, D, and C, the functions of 15 amino acids are all 0.

Table 1. The indicator functions of S.

Sequence S	**A**	**R**	**R**	**N**	**A**	**D**	**C**	**D**	**C**	**C**
Position(i)	1	2	3	4	5	6	7	8	9	10
$I_A(i)$	1	0	0	0	1	0	0	0	0	0
$I_R(i)$	0	1	1	0	0	0	0	0	0	0
$I_N(i)$	0	0	0	1	0	0	0	0	0	0
$I_D(i)$	0	0	0	0	0	1	0	1	0	0
$I_C(i)$	0	0	0	0	0	0	1	0	1	1

Table 2. The accumulated indicator functions of S.

Sequence S	**A**	**R**	**R**	**N**	**A**	**D**	**C**	**D**	**C**	**C**
Position(i)	1	2	3	4	5	6	7	8	9	10
$\tilde{I}_A(i)$	1	1	1	1	2	2	2	2	2	2
$\tilde{I}_R(i)$	0	1	2	2	2	2	2	2	2	2
$\tilde{I}_N(i)$	0	0	0	1	1	1	1	1	1	1
$\tilde{I}_D(i)$	0	0	0	0	0	1	1	2	2	2
$\tilde{I}_C(i)$	0	0	0	0	0	0	1	1	2	3

According to the tables, we calculate:

$$n_A = \tilde{I}_A(10) = 2 \tag{8}$$

$$\begin{aligned}\zeta_A &= \frac{\sum_{i=1}^{10} \tilde{I}_A(i)}{n_A}\\ &= \frac{1+1+1+1+2+2+2+2+2+2}{2}\\ &= 8\end{aligned} \tag{9}$$

Similarly, we can obtain $n_R = 2, n_N = 1, n_D = 2, n_C = 3, \zeta_R = 8.5, \zeta_N = 7, \zeta_D = 4$, and $\zeta_C = 2.333$.

Next, we calculate the variance and covariance.

$$\begin{aligned} \theta_A &= \frac{\sum_{i=1}^{10} \tilde{I}_A(i)}{10} \\ &= \frac{1+1+1+1+2+2+2+2+2+2}{10} \\ &= 1.6 \end{aligned} \tag{10}$$

$$\begin{aligned} D_A &= \sum_{i=1}^{10} \frac{(\tilde{I}_A(i) - \theta_A)^2}{n_A^2} \\ &= \frac{4 \times (1-1.6)^2 + 6 \times (2-1.6)^2}{2^2} \\ &= 0.6 \end{aligned} \tag{11}$$

Similarly, we can obtain $D_R = 1.025, D_N = 2.1, D_D = 1.9$, and $D_C = 1.122$

$$\begin{aligned} cov(A,R) &= \sum_{i=1}^{10} \frac{(\tilde{I}_A(i) - \theta_A) \times (\tilde{I}_R(i) - \theta_R)}{n_A \times n_R} \\ &= \frac{1}{2 \times 2}((1-1.6) \times (0-1.7) + \\ &\quad (1-1.6) \times (1-1.7) + (1-1.6) \times (2-1.7) \\ &\quad \times 2 + (2-1.6) \times (2-1.7) \times 6) \\ &= 0.45 \end{aligned} \tag{12}$$

As well as $cov(A,N) = 0.9, cov(A,D) = 0.8, cov(A,C) = 0.467, cov(R,N) = 1.05$, $cov(R,D) = 0.6, cov(R,C) = 0.35, cov(N,D) = 1.2, cov(N,C) = 0.7$, and $cov(D,C) = 1.233$.

Finally, the accumulated natural vector of S is $(2, 2, 1, 2, 3, 0, \cdots, 0, 8, 8.5, 7, 4, 2.333, 0, \cdots, 0, 0.6, 1.025, 2.1, 1.9, 1.122, 0, \cdots, 0, 0.45, 0.9, 0.8, 0.467, 0, \cdots, 0, 1.05, 0.6, 0.35, 0, \cdots, 0, 1.2, 0.7, 0, \cdots, 0, 1.233, 0, \cdots, 0)$.

2.2. Convex Hull Method

In the previous section, we introduce how a protein sequence is represented by a 250-dimensional vector. Therefore, the distance between two vectors represents the biological distance of the corresponding two protein sequences. The convex hull principle for protein states that convex hulls corresponding to different families are disjoint with each other [15,18]. The relationship among point sets can be better observed by constructing their convex hulls.

Given a finite point set $A = \{a_1, a_2, \cdots, a_n\}$ in R^k space, we define the convex hull of A as:

$$C(A) = \{p | p = \sum_{i=1}^{n} \lambda_i a_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0, 1 \leq i \leq n\} \tag{13}$$

Generally speaking, the convex hull of a given point set is the smallest convex set that contains this point set.

There are many ways to judge whether two convex hulls intersect. Considering that the construction of a convex hull in a high-dimensional space is computationally intensive and time-consuming, we do not directly consider two convex hulls here. Instead, we consider two point sets to determine whether the convex hull formed by them intersects.

Given two finite point sets in the R^k spaces $A = \{a_1, a_2, \cdots, a_n\}$ and $B = \{b_1, b_2, \cdots, b_m\}$, we want to determine whether the convex hulls of A and B intersect. From the definition given in (13), if some sets of coefficients satisfy:

$$\begin{aligned} \sum_{i=1}^{n} \lambda_i a_i &= \sum_{j=1}^{m} \mu_j b_j, \\ \sum_{i=1}^{n} \lambda_i &= 1, \\ \sum_{j=1}^{m} \mu_j &= 1, \\ 0 \le \lambda_i, \mu_j &\le 1, \\ 1 \le i \le n, 1 \le j &\le m, \end{aligned} \tag{14}$$

Then the two convex hulls of A and B intersect; otherwise, the two convex hulls are disjoint [18].

2.3. Convex Hull Distance

The distance between the two sequences we have defined is the Euclidean distance between their corresponding natural vectors. Then, the distance between two convex hulls corresponding to the two families is the Euclidean distance between the centers of the two convex hulls constructed by these natural vectors.

$$distance = \|meanvector(A) - meanvector(B)\| \tag{15}$$

A, B are the two sets of 250-dimensional accumulated natural vectors of the two families. Here, *meanvector*(A) represents the center of the convex hull composed of set A. The expression of *meanvector*(A) is:

$$meanvector(A) = \frac{1}{n} \sum_{i=1}^{n} A_i \tag{16}$$

where n is the number of vectors in set A, A_i is a vector in set A. With this convex hull distance definition, it is easy to calculate the distance of two convex hulls if we know their point sets.

2.4. Linear Discriminant Analysis

With the help of the accumulated natural vector and the convex hull method, we can construct convex hulls and study their intersection in a 250-dimensional space. When we want to visualize convex hulls and research their properties in a lower-dimensional space, especially in a 2-dimensional space, the dimensional reduction method in cluster analysis is indispensable. First, we introduce linear discriminant analysis, which is the widely known method.

We define $X_j (j = 0, 1)$ as the set of samples in Class j. In our experiment, the two sets of samples are two point sets of the convex hulls whose intersection is our study subject.

Our goal is to find the projection direction ω that maximizes the sample distance between the two projected categories (0,1).

We define $N_j (j = 0, 1)$ as the number of samples in Class j, $\mu_j (j = 0, 1)$ as the mean value of the samples in Class j, and $\sum_j$ as the covariance matrix of the samples in Class j. $\mu_j (j = 0, 1)$ can be calculated by:

$$\mu_j = \frac{1}{N_j} \sum_{x \in X_j} x (j = 0, 1) \tag{17}$$

$\sum_j$ can be calculated by:

$$\sum_j = \sum_{x \in X_j} (x - \mu_j)(x - \mu_j)^T (j = 0, 1) \tag{18}$$

As previously mentioned, the LDA (short for linear discriminant analysis) algorithm should make two projected classes of datasets (X_1, X_2) in the projection direction ω as far as possible, so we need to maximize $\|\omega^T\mu_0 - \omega^T\mu_1\|_2^2$ (using the Euclidean distance here) and minimize the within-group covariance in the meantime.

The within-class scatter matrix S_ω is defined as:

$$S_\omega = \sum_0 + \sum_1 = \sum_{x \in X_0} (x - \mu_0)(x - \mu_0)^T = \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^T \tag{19}$$

The between-class scatter matrix S_b is defined as:

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^T \tag{20}$$

Then, the two-class LDA can be formulated as an optimization problem to find a set of linear combinations with the coefficient ω.

$$argmax J(\omega) = \frac{\omega^T S_b \omega}{\omega^T S_\omega \omega} \tag{21}$$

The expression is the Rayleigh Quotient R(A,x). The solution ω to the above optimization function $J(\omega)$ is the eigenvector corresponding to the maximum eigenvalue of the matrix $S_\omega^{-1} S_b$. In the case of projecting to two dimensions, we need to calculate the two eigenvectors that correspond to the two largest eigenvalues of the matrix $S_\omega^{-1} S_b$. This step can be accomplished by MATLAB programming or another program.

2.5. Maximum Margin Criterion

When we use LDA to complete the dimensionality reduction, we may meet the small sample size problem [19], which is caused by the singular S_ω. The maximum margin criterion (MMC) is proposed to avoid the small sample size problem and calculate the most discriminant vectors. The MMC method is simple, efficient, and stable compared to the PCA+LDA method [20].

Consider a linear mapping $W \in R^{D \times d}$, where D and d are the dimensionalities of the data before and after the projection. The MMC introduces a new objective function:

$$J(\omega) = tr(\omega^T (S_b - S_\omega)\omega) \tag{22}$$

We can suppose that $\omega^T \omega = 1$ because ω can be multiplied by any constant to make it be the unit vector. Then, we just need to solve the following constrained optimization:

$$\sum_{k=1}^{d} \omega^T (S_b - S_\omega)\omega, \tag{23}$$

$$subject \quad to \quad \omega_k^T \omega_k = 1, k = 1, \cdots, d \tag{24}$$

The small sample size problem is avoided by the formation $S_b - S_\omega$ instead of $S_\omega^{-1} S_b$ of the LDA.

2.6. Knn Classification

The classification performance of the 250-dimensional accumulated natural vector is also required in this paper. The K-nearest neighbor is a simple classification method and has been developed successfully in real applications [21]. The choice of the *k*-value has a huge impact on the final classification results. In practical applications, cross-checking is usually required to select the optimal *k*-value [22]. In particular, we achieved good results when we chose $k = 1$ in our work.

Suppose the training set contains N samples $\{x_1, x_2, \cdots, x_N\}$, which belong to t classes $\{A_1, A_2, \cdots, A_t\}$. Define the Euclidean distance from the unclassified sample x to the training set sample x_i as $d(x, x_i)$, if:

$$d(x, x_k) = min_{i=1,2,\cdots,N}(d(x, x_i)), x_k \in A_j$$

Then, the nearest neighbor classification decision is: $x \in A_j$.

3. Results and Discussion

3.1. Convex Hull Analysis of Bacterial Families

Bacteria are almost everywhere on earth and play an important role in many research fields [23]. The number of bacteria-related protein sequences is enormous and still rising. As of May 2020, the Reference Sequence Database (RefSeq) on NCBI had a total of 140 million bacterial protein sequences. Therefore, we selected 117 bacterial families from 13 different phyla and 150 protein sequences per these families for a total of $117 \times 150 = 17,550$ protein sequences. Table S1 provides the complete list of the 117 bacterial families.

For each protein sequence, we calculated the 250-dimensional accumulated natural vector. Thus, each protein sequence corresponded to a point in a 250-dimensional space, and 117 different bacteria families corresponded to 117 finite point sets in a 250-dimensional space. We considered the convex hull of these 117 finite point sets and verified whether they intersect in pairs. There were $\binom{117}{2} = 6786$ pairs of convex hulls. No intersection was observed using the method in Section 2.2. We were not surprised because the convex hull method had similar results when applied to other datasets in previous work [15]. We applied the linear discriminant analysis method to visualize the results. For example, Figure 1 shows that the Acetobacteraceae family disjoints the Acidiferrobacteraceae family.

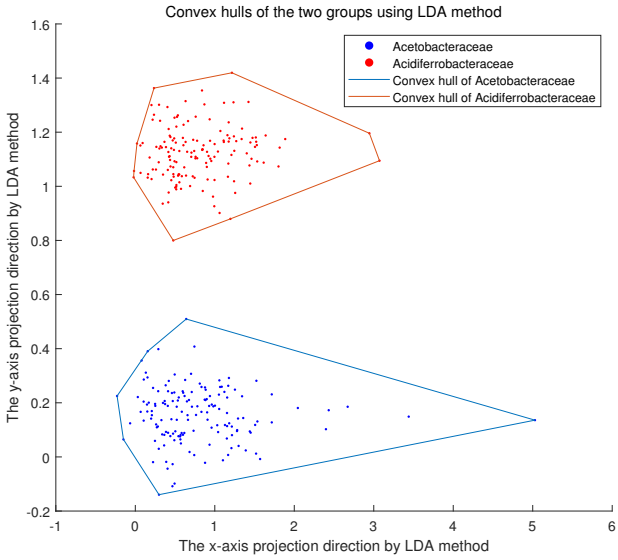

Figure 1. Convex hulls of bacterial family Acetobacteraceae and bacterial family Acidiferrobacteraceae after dimension reduction by LDA(Linear Discriminant Analysis) method.

This result shows that our proposed 250-dimensional accumulated natural vector is very effective when applied to protein sequences. The disjointness property reveals that our method accurately clusters protein sequences from the same family. It is worth noting that the advantage of using the convex hull method is that even if two points are closer together, it does not prevent them from being in different convex hulls. We can suppose that the convex hulls of the different bacterial families are pairwise disjoint. For a new, unclassified protein sequence, we can calculate its 250-dimensional accumulated natural vector in the same way and then analyze in which convex hull the point is located. Subsequently, this new protein is classified into the corresponding family. This is the most intuitive application of our method.

3.2. Classification of Protein Enzyme Classes

We performed a classification analysis on protein enzyme classes using the accumulated natural vector. Four datasets were selected from UniProt. According to UniProt, each dataset is composed of seven enzyme classes: Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, Ligases, and Translocases. We used the one-nearest neighbor classification method mentioned in Section 2.6 to divide these protein sequences into the seven enzyme classes. All protein sequences appeared in the training set and the test set. By predicting the enzyme classes in the test set and comparing these with the training set, we calculated the accuracy of our classification method.

E. coli (*Escherichia coli*) is a type of bacteria closely related to human life [24]. The *E. coli* dataset contained 12284 protein sequences. A total of 11788 protein sequences are classified accurately, and the total accuracy is 11,782/12,284(95.9%). The detailed results of this classification are shown in Table 3.

Table 3. Classification results of the *E. coli* dataset.

Enzyme Class	Total Number	Correct Number	Accuracy
Oxidoreductases	1829	1720	0.940404
Transferases	4054	3905	0.963246
Hydrolases	2783	2632	0.945742
Lyases	1387	1354	0.976208
Isomerases	823	788	0.957473
Ligases	906	893	0.985651
Translocases	502	490	0.976096
Total	12,284	11,782	0.959134

This method was also applied to the other three bacterial family datasets. The total accuracies were 6652/6890(96.5%), 3940/4017(98.1%), and 4511/4655(96.7%), respectively. Tables S2–S4 show the detailed results of classification.

Using the one-nearest neighbor classification method, one sequence is incorporated into its nearest convex hull, which represents one bacterial family. The high accuracy rate results indicate that our method combined with the one-nearest neighbor algorithm performed robustly in the classification of protein enzyme classes. Although the accuracy cannot reach 100%, it is still a good result. We must admit that our dataset does not guarantee that every point with the nearest distance is put into the correct convex hull, which is equivalent to saying that two points that are close together can be in different convex hulls. This non-ideal situation is possible but at a low frequency. This is also another piece of evidence to support the notion that accumulated natural vectors provide a good representation of protein space.

3.3. Convex Hull Analysis of Virus Families

3.3.1. Intersection of Virus Families in a 250-Dimensional Space

We chose 73 virus families and downloaded all of their reviewed protein sequences. Detailed information about the dataset is shown in Table S5. Then, we obtained the accumulated natural vector of each protein sequence in a 250-dimensional space. These vectors of different families constituted the different vector sets. We wanted to determine whether the convex hulls of these vectors intersected in 250 dimensions. Here, we introduce the Baltimore virus taxonomy, which divides viruses into seven categories based on differences in the gene expressions of different viruses.

Our expected result was that all virus families under the same Baltimore virus category would be disjoint.

For one Baltimore classification, we tested the intersection of the convex hulls under each Baltimore classification. The test method follows the theorem in Section 2.2; the input was the two "point set" (vector set), and the output was the intersection of the convex hull pairs.

The intersection results of the 73 virus families are shown in Table 4. A total of 738 pairs of convex hull pairs were counted. Among these, 727 pairs were disjoint in 250 dimensions, and 11 pairs intersected in 250 dimensions. The disjoint convex hull pairs accounted for 0.9851 of all convex hull pairs.

Table 4. Baltimore virus taxonomy.

Virus Classification	Samples	Number of Virus Families
dsDNA virus	Herelleviridae	23
ssDNA virus	Microviridae	7
dsRNA virus	Totiviridae	8
ssRNA(+) virus	Alphatetraviridae	30
ssRNA(-) virus	Bornaviridae	1
ssRNA-RT virus	Metaviridae	1
dsDNA-RT virus	Caulimoviridae	2

Among the seven Baltimore taxonomies, several convex hulls intersect under the dsDNA Baltimore taxonomy, and all convex hulls disjoint under the other six Baltimore taxonomies. There are 11 intersecting dsDNA virus family pairs in the 250-dimensional space, as shown in Table S6.

Then, we used the distances of the convex hulls corresponding to these 11 virus family pairs to perform a cluster analysis. The convex hull distances calculated here were the Euclidean distance between the two centers of the two convex hulls defined in Section 2.3. The result of our cluster analysis is in Figure 2.

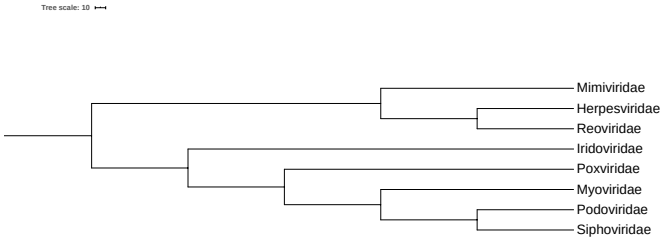

Figure 2. Phylogenetic tree of eight virus families by NJ method: *Reoviridae*, *Herpesviridae*, *Mimiviridae*, *Iridoviridae*, *Poxviridae*, *Myoviridae*, *Siphoviridae*, *Podoviridae*.

From Figure 2, we can see that the three nearest virus families are *Podoviridae, Siphoviridae and Myoviridae*, which all belong to *Caudovirales* and are parasitic in bacteria. Their convex hulls have closer distances, so we deduce that these three families have closer evolutionary distances.

Similarly, we performed a cluster analysis of all virus families in the dataset by calculating the distance matrix of their convex hulls. The result is that all virus families are divided into three clusters. Cluster 2 includes *Togaviridae, Tobaniviridae, Secoviridae, Potyviridae, Picornaviridae and Hypoviridae*. Cluster 3 includes *Iflaviridae and Flaviviridae*. To visualize the result of the cluster analysis, we represent each virus family by the mean vector of its corresponding convex hull, and we project these points onto a 2-dimensional plane. The result is shown in Figure 3.

As shown in Figure 3, we can see that all virus families are divided into three clusters in a 2-dimensional space, and the virus families in each cluster remain consistent with those in a 250-dimensional space.

Above all, most of the convex hull pairs of the 73 virus families are disjoint in a 250-dimensional space. Those virus families whose convex hulls intersect with those of the other families are all from the dsDNA Baltimore taxonomy. Although these virus families have intersecting convex hulls with each other, their evolution relationships can also be reflected by the convex hull distances.

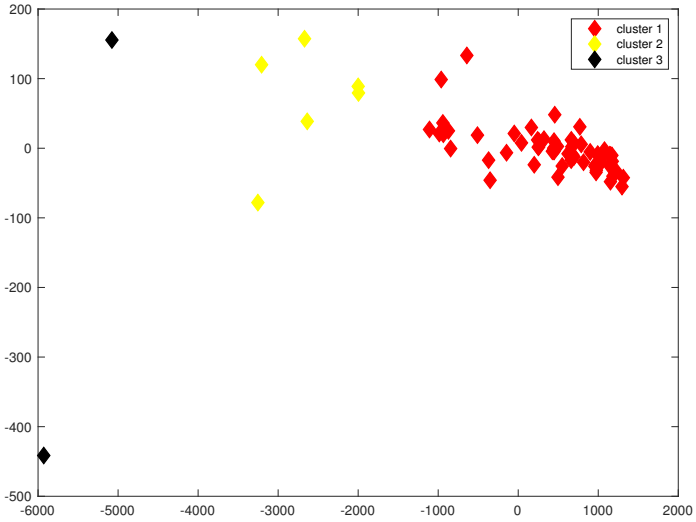

Figure 3. Cluster analysis of all virus families by shortest distance method: red, yellow and black points represent different virus families in three clusters. *x*-axis and *y*-axis are the two coordinate axes of the two-dimensional plane. Values on the axes are position coordinates.

3.3.2. Intersection of Virus Families in a 2-Dimensional Space

Visualization of the results of classification is very important. Convex hulls in a 2-dimensional space are much more intuitive than those in 250-dimensional space. Thus, we project the convex hulls constructed above onto 2-dimensional space to observe their intersection. The method used is the maximum margin criterion, instead of the traditional method of linear discriminant analysis [25], to avoid the small sample size problem [22]. A part of the results of the classification is shown in Table 5, and the complete result is shown in Table S7. Compared to our newly proposed 250-dimensional natural vector of protein sequences, the intersection result of a 60-dimensional natural vector [26] is bad in a 2-dimensional space. All convex hulls intersect in a 2-dimensional space. The detailed results are shown in Table S8. This illustrates that the introduction of covariance between the amino acids is an improvement for the intersection analysis.

Table 5. Partial results of the convex hull intersection: 'Percentage' in this Table is the percentage of disjoint convex hull pairs of all convex hull pairs within one virus family of non-intersection.

Virus Family	Percentage	Virus Family	Percentage
Adenoviridae	0.6667	Hepeviridae	0.9861
Alloherpesviridae	0.8472	Herelleviridae	0.8472
Alphaflexiviridae	0.8056	Herpesviridae	0.1389
Alphatetraviridae	1.0000	Hypoviridae	1.0000
Ampullaviridae	0.8472	Inoviridae	0.8056
Anelloviridae	0.9861	Iridoviridae	0.5333
Arteriviridae	0.9583	Kitaviridae	0.9861
Ascoviridae	0.9444	Lavidaviridae	0.9583
Astroviridae	1.0000	Leviviridae	0.9444

The percentage of non-intersecting convex hull pairs is nearly 0.85 for most of the virus families, except for *Adenoviridae*, *Herpesviridae*, *Iridoviridae*, *Mimiviridae*, *Myoviridae*, *Podoviridae*, *Poxviridae*, *Reoviridae*, and *Siphoviridae*.

We dropped the above nine families whose convex hulls intersect with the convex hulls of other families in a 250-dimensional space, and then the percentage of non-intersecting convex hull pairs is 0.9588.

In conclusion, most of the convex hulls of the virus families disjoint both in a 250-dimensional space and a 2-dimensional space. The percentage of non-intersection of the virus families may be improved by incorporating other elements in the accumulated natural vector;

for example, we can incorporate higher-order moments of the 20 amino acids. Correspondingly, the addition of more elements will add to the dimensions of the vectors and make calculations more complicated.

4. Conclusions

In this paper, the 250-dimensional accumulated natural vector method is proposed by describing the distribution of 20 amino acids within a protein sequence. Protein sequences with similar properties correspond to closer points, so proteins in the same family tend to cluster together. Our proposed method makes it easy to classify the protein sequences since it avoids the high computational complexity associated with sequence alignment and takes advantage of mathematical concepts only. Its applications to real datasets suggest that the accumulated natural vector method is a powerful tool for the classification of protein sequences. However, there is still a lot of room for improvement of this novel method by incorporating other elements.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13101744/s1. Table S1: The bacterial families' names in this paper; Table S2: The classification result of *Mycobacteriaceae* dataset; Table S3: The classification result of *Xanthomonadaceae* dataset; Table S4: The classification result of *Vibrionaceae* dataset; Table S5: The virus families' names and the number of their protein sequences in this paper, Table S6: 11 intersecting dsDNA virus family pairs in a 250-dimensional space, Table S7: The results of convex hull intersection in a 2-dimensional space: Percentage of non-intersection is the percentage of disjoint convex hull pairs of all convex hull pairs under one virus family, Table S8. The results of convex hull intersection in 2-dimension space based on 60-dimensional natural vector: Percentage of non-intersection is the percentage of disjoint convex hull pairs of all convex hull pairs under one virus family. All convex hulls intersect in 2-dimension space based on 60-dimensional natural vector. So some families are omitted.

Author Contributions: Conceptualization, S.S.-T.Y.; Methodology, L.Z. and M.G.; software, L.Z. and M.G.; writing–original draft preparation, L.Z. and M.G.; writing–review and editing, L.Z. and M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (NSFC) grant (91746119), the Tsinghua University Spring Breeze Fund (2020Z99CFY044), the Tsinghua University Start-Up Fund, and the Tsinghua University Education Foundation Fund (042202008).

Data Availability Statement: All datasets used in this study may be found here: https://github.com/Gmcen20/ANV250 (accessed on 19 September 2022) and http://www.uniprot.org/ (accessed on 19 September 2022).

Acknowledgments: Stephen Shing-Toung Yau is grateful to the National Center for Theoretical Sciences (NCTS) for providing an excellent research environment while part of this research was completed.

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

References

1. Mount, D.W. *Bioinformatics: Sequence and Genome Analysis*, 2nd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, USA, 2004.
2. Lodish, H.; Berk, A.; Kaiser, C.A.; Krieger, M. *Molecular Cell Biology*; W.H. Freeman and Company: New York, NY, USA, 2004.
3. Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Res.* **2000**, *28*, 304–305. [CrossRef] [PubMed]
4. Nelson, D.L.; Cox, M.M. *Lehninger Principles of Biochemistry*; W.H. Freeman and Company: New York, NY, USA, 2008.
5. Ng, P.C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **2003**, *31*, 3812–3814. [CrossRef] [PubMed]
6. Black, D.L. Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome biology. *Cell* **2000**, *103*, 367–370. [CrossRef]
7. Wu, C.; Gao, R.; De Marinis, Y.; Zhang, Y. A novel model for protein sequence similarity analysis based on spectral radius. *J. Theor. Biol.* **2018**, *446*, 61–70. [CrossRef]
8. Yao, Y.H.; Dai, Q.; Li, L.; Nan, X.Y.; He, P.A.; Zhang, Y.Z. Similarity/dissimilarity studies of protein sequences based on a new 2D graphical representation. *J. Comput. Chem.* **2010**, *31*, 1045–1052. [CrossRef]

9. Pham, T.D.; Zuegg, J. A probabilistic measure for alignment-free sequence comparison. *Bioinformatics* **2004**, *20*, 3455–3461. [CrossRef]
10. Schwende, I.; Pham, T.D. Pattern recognition and probabilistic measures in alignment-free sequence analysis. *Brief. Bioinform.* **2014**, *15*, 354–368. [CrossRef]
11. Sims, G.E.; Jun, S.R.; Wu, G.A.; Kim, S.H. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 2677–2682. [CrossRef]
12. Steinegger, M.; Söding, J. Clustering huge protein sequence sets in linear time. *Nat. Commun.* **2018**, *9*, 2542. [CrossRef]
13. Yau, S.S.T.; Yu, C.; He, R. A protein map and its application. *DNA Cell Biol.* **2008**, *27*, 241–250. [CrossRef]
14. Zhang, Y.; Wen, J.; Yau, S.S.T. Phylogenetic analysis of protein sequences based on a novel k-mer natural vector method. *Genomics* **2019**, *111*, 1298–1305. [CrossRef] [PubMed]
15. Zhao, X.; Tian, K.; He, R.L.; Yau, S.S.T. Convex hull principle for classification and phylogeny of eukaryotic proteins. *Genomics* **2019**, *111*, 1777–1784. [CrossRef] [PubMed]
16. Dong, R.; He, L.; He, R.L.; Yau, S.S.T. A Novel Approach to Clustering Genome Sequences Using Inter-nucleotide Covariance. *Front. Genet.* **2019**, *10*, 234. [CrossRef] [PubMed]
17. Chou, K.C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.* **2011**, *273*, 236–247. [CrossRef] [PubMed]
18. Tian, K.; Zhao, X.; Yau, S.S.T. Convex hull analysis of evolutionary and phylogenetic relationships between biological groups. *J. Theor. Biol.* **2018**, *456*, 34–40. [CrossRef] [PubMed]
19. Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: Cambridge, MA, USA, 1990.
20. Chen, L.F.; Liao, H.Y.M.; Ko, M.T.; Lin, J.C.; Yu, G.J. New LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognit.* **2000**, *33*, 1713–1726. [CrossRef]
21. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
22. Zhang, S.; Li, X.; Zong, M.; Zhu, X.; Wang, R. Efficient knn classification with different numbers of nearest neighbors. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 1774–1785. [CrossRef]
23. Alexander, C.; Rietschel, E.T. Invited review: Bacterial lipopolysaccharides and innate immunity. *J. Endotoxin Res.* **2001**, *7*, 167–202. [CrossRef]
24. Palmer, B.R.; Marinus, M.G. The dam and dcm strains of Escherichia coli—A review. *Gene* **1994**, *143*, 1–12. [CrossRef]
25. Li, M.; Yuan, B. 2D-LDA: A statistical linear discriminant analysis for image matrix. *Pattern Recognit. Lett.* **2005**, *26*, 527–532. [CrossRef]
26. Wang, Y.; Tian, K.; Yau, S.S.T. Protein Sequence Classification Using Natural Vector and Convex Hull Method. *J. Comput. Biol.* **2019**, *26*, 315–321. [CrossRef] [PubMed]