Separating Phages From Other Virus Families and Classifying the Different Phage Families By GI-Clusters

Xingang Jia (hanqh15@163.com)
Southeast University
Qiuhong Han
Nanjing Forestry University
Zuhong Lu
Southeast University

Research Article

Keywords: MG-Euclidean, Icc-cluster, F-feature, GI-feature, GI-cluster

Posted Date: December 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1130357/v1

METHODOLOGY

Separating phages from other virus families and classifying the different phage families by GI-clusters

Xingang Jia[1*], Qiuhong Han[2] and Zuhong Lu[3]

*Correspondence:
hanqh15@163.com
[1]School of Mathematics, Southeast University , 210096 Nanjing, PR China
Full list of author information is available at the end of the article

Abstract

Background: Phages are the most abundant biological entities, but the commonly used clustering techniques are difficult to separate them from other virus families and classify the different phage families together.

Results: This work uses GI-clusters to separate phages from other virus families and classify the different phage families, where GI-clusters are constructed by GI-features, GI-features are constructed by the togetherness with F-features, training data, MG-Euclidean and Icc-cluster algorithms, F-features are the frequencies of multiple-nucleotides that are generated from genomes of viruses, MG-Euclidean algorithm is able to put the nearest neighbors in the same mini-groups, and Icc-cluster algorithm put the distant samples to the different mini-clusters. For these viruses that the maximum element of their GI-features are in the same locations, they are put to the same GI-clusters, where the families of viruses in test data are identified by GI-clusters, and the families of GI-clusters are defined by viruses of training data.

Conclusions: From analysis of 4 data sets that are constructed by the different family viruses, we demonstrate that GI-clusters are able to separate phages from other virus families, correctly classify the different phage families, and correctly predict the families of these unknown phages also.

Keywords: MG-Euclidean; Icc-cluster; F-feature; GI-feature; GI-cluster

Background

Phages were defined as viruses that infected bacteria, where they were the most abundant microbial entities, and they represented the largest reservoir of unexplored genetic information[1, 2, 3]. With the rapid development of high-throughput biotechnologies, we were able to obtain phage genomes in the public databases[4]. However, for the commonly used clustering techniques with these features that were generated from phage genomes, they were difficult to classify the different phage families[5]. Moreover, because of the lack of corresponding biological and experimental data, clustering techniques with these genomes into the ICTV(the International Committee on Taxonomy of Viruses) scheme were difficult also[5, 6, 7]. Furthermore, some approaches of phage classifications used the singly selected marker molecules to define sequence alignment and similarities, but these approaches were restricted to closely related phage taxa only[8, 9], such as these comparative sequence analysis[10, 11, 12, 13, 14].

To search appropriate methods that were able to separate phages from other virus families and classify the different phage families, we constructed 4 data sets to verify our methods, where Data-1 mixed phages and other 9 different virus families, Data-2 contained 6 different phage families, Data-3 owned 9 different virus families that deleted phages of Data-1, Data-4 contained 5 different Ebolavirus families, Data-3 and Data-4 mainly illustrated the complexity of phages. Here, we used t-SNE maps to select efficient features, where t-SNE was able to map nearest neighbor samples onto adjacent points on the plane[15, 16]. That is, if t-SNE projections of the different virus families distributed in different regions, the used features were reliable to define the similarity of viruses. In fact, we firstly used t-SNE to verify the reliability of *F*-features, where *F*-features were the frequencies of multiple-nucleotides that were generated from genomes of viruses, 15 types of *F*-features were designed in this study, and *F*-features had successfully applied to the classification of siRNAs that their two types of samples had no clear boundary[17]. But for t-SNE map of any type of *F*-features, results showed that the different phage families mixed together on almost all of regions. Then, *G*-features were constructed by the comprehensive application of these 15 types of *F*-features, training data and *MG*-Euclidean algorithm, where *MG*-Euclidean algorithm was able to put the nearest neighbors in the same mini-groups[15], and viruses of training data were used to define the families of mini-groups. However, t-SNE map of *G*-features showed that the different phage families still had slight mixture on their boundaries. Furthermore, *I*-features of phages were constructed by the comprehensive application of these 15 types of *F*-features, training data and *Icc*-cluster algorithm, where *Icc*-cluster was able to put the distant samples were in the different mini-clusters[16], the number of viruses of training data was selected as the number of mini-clusters, and viruses of training data were used to define the families of mini-clusters. However, for t-SNE map of any data set that were generated from *I*-features, its one region mixed some viruses that came from many families, but any family only had few viruses, where these mixed viruses came from test data, and their *I*-features were zero vectors. Importantly, for these viruses in the mixed region of *I*-features, almost all of them were not in the mixed boundaries of *G*-features.

To comprehensively use the information of *G*-features and *I*-features, *GI*-features of viruses were constructed by the sum of *M*-features and *I*-features. Moreover, for viruses of any data set, t-SNE map of their *GI*-features showed that their different families had clear boundaries. Importantly, for viruses of the different families, the maximum elements of their *GI*-features were in different locations. Thus, for these viruses that the maximum element of their *GI*-features were in the same locations, they were put to the same *GI*-clusters, and the families of viruses in test data were predicted by these *GI*-clusters also. For any data set, results showed that *GI*-clusters were able to correctly classify the different virus families. That is, *GI*-clusters were able to separate phages of from other

virus families, correctly classify the different phage families, and correctly predict the families of these unknown phages also. We hoped GI-clusters were able to help the researchers to distinguish different phage families.

Materials and Methods

Data and Data Source

Data-1

Here, we collect 432 complete genomes that contain 143 Phages, 78 Alphaviruses, 11 Arteriviruses, 62 Ebolaviruses, 17 Flaviviruses, 10 Hiv, 40 Marburgviruses, 8 Pestiviruses, 58 Rubiviruses and 5 SARS, where these 432 genomes are used to construct Data-1, these genomes are downloaded from https://www.ncbi.nlm.nih.gov/, and they are defined as the first to tenth families in the above order. Here, Ebolaviruses and Marburgviruses are filoviruses, Alphaviruses, Arteriviruses and Rubiviruses are togaviridae, Flaviviruses and Pestiviruses are flaviviridae, SARS are coronavirus, and Hiv are retroviridae. Moreover, for viruses of each family, 80% of them are put to construct training-1, and other ones are put to test-1, where the details of Data-1, training-1 and test-1 are summarized in Additional file 1.

Data-2

For 143 complete genomes of phages, their 6 families are over 5 genomes that contain 12 Enterobacteria, 14 Mycobacterium, 13 Prochlorococcus, 7 Pseudomonas, 18 Synechococcus and 19 Vibrio, where these 83 phages are used to construct Data-2, and they are defined as the first to sixth families in the above order. Here, for phages of each families, 20% of them are put to test-2, and other ones are put to training-2, where the details of Data-2, training-2 and test-2 are summarized in Additional file 1.

Data-3

Here, Data-3, training-3 and test-3 are that Data-1, training-1 and test-1 delete phages, respectively. That is, Data-3, training-3 and test-3 contain 9 virus families, where the details of Data-3, training-3 and test-3 are summarized in Additional file 1.

Data-4

Data-4 contains 62 complete genomes of Ebolaviruses that come from 5 virus families(B, R, S, T and Z), where 20% of Ebolaviruses are put to test-4, and other ones are put to training-4, where the details of training-4 and test-4 are summarized in Additional file 1.

Methods

The biological clusters of viruses

For viruses of data-s($s = 1, 2, 3, 4$), training-s and test-s, these ones that belong to the l-th family are put to $D_{s,l}$-cluster, $Tr_{s,l}$-cluster and $Te_{s,l}$-cluster respectively, where data-1, data-2, data-3 and data-4 have 10, 6, 9 and 5 virus families, respectively.

F_k-features

For the i-th virus of data-s, we use $X(i)$ to represent its complete genome, and $F_k(i)$ to represent its F_k-feature, where

$$\begin{cases} F_1(i) = \{f_1^1(i), f_1^2(i), f_1^3(i), f_1^4(i)\} = \{f_A(i), f_C(i), f_G(i), f_T(i)\} \\ F_2(i) = \{f_2^1(i), f_2^2(i), \cdots, f_2^{16}(i)\} = \{f_{AA}(i), f_{AC}(i), \cdots, f_{TT}(i)\} \\ F_3(i) = \{f_3^1(i), \cdots, f_3^{64}(i)\} = \{f_{AAA}(i), f_{AAC}(i), \cdots, f_{TTT}(i)\} \\ F_4(i) = \{f_4^1(i), \cdots, f_4^{256}(i)\} = \{f_{AAAA}(i), \cdots, f_{TTTT}(i)\} \end{cases}, \quad (1)$$

and these $F_k(i)(k = 5, 6, \cdots, 15)$ are constructed by two or more $F_1(i)$, $F_2(i)$, $F_3(i)$ and $F_4(i)$. That is, $F_1(i)$, $F_2(i)$, $F_3(i)$ and $F_4(i)$ are the frequency of mononucleotide, dinucleotide, trinucleotide and quadnucleotide of $X(i)$ respectively, and other F_k-features($k = 5, 6, \cdots, 15$) are their various combinations.

Moreover, we use $Y(j)$ to represent the j-th virus genome of training-s, and $Z(n)$ to represent the n-th virus genome of test-s.

MG-Euclidean and Icc-cluster algorithms

MG-Euclidean algorithm does not directly divide viruses into clusters, but put the nearest neighbor viruses to the same mini-groups[15]. That is, when a virus belongs to a mini-group, its nearest neighbors are in the mini-group also. In this study, viruses are put to mini-groups by *MG*-Euclidean with F_k-features, where the similarity of the viruses are defined by Euclidean distance of their F_k-feature.

For *Icc*-cluster algorithm, it puts the distant samples in the different clusters, and it has great ability to remove the effect of the clustering numbers even if clustering number is relative large or small compared to the optimal one[16].

G-features and G-clusters

For any type of F_k-features, *MG*-Euclidean with them divides viruses of data-*s* into mini-groups, where $G_{s,k}^p$-group is used to denote the p-th mini-group, all $G_{s,k}^p$-groups are specified as m categories, and m is the number of the virus families. That is, $G_{s,k}^p$-group is defined as $G_{s,k}^{p,t}$-group if

$$\begin{cases} G_{s,k}^p(t) &= \max\{G_{s,k}^p(1), G_{s,k}^p(2), \cdots, G_{s,k}^p(m)\} \\ G_{s,k}^p(l) &= \dfrac{\mathrm{N}\ \{G_{s,k}^p\text{-group} \cap Tr_{s,l}\text{-cluster}\}}{\mathrm{N}\ \{Tr_{s,l}\text{-cluster}\}} \end{cases}, \quad (2)$$

where $N\{G_{s,k}^p\text{-group} \bigcap Tr_{s,l}\text{-cluster}\}$ is the virus number of the intersection of $G_{k,p}$-group and $Tr_{s,l}$-cluster, and $N\{Tr_{s,l}\text{-cluster}\}$ is the virus number of $Tr_{s,l}$-cluster.

Here, these $G_{s,k}^{p,t}$-groups are used to construct m $G_{s,k}^l$-clusters, where

$$G_{s,k}^l\text{-cluster} = \bigcup_{t=l} G_{s,k}^{p,t}\text{-group}. \quad (3)$$

Then, $G_{s,k}$-features of viruses are constructed by $G_{s,k}^l$-clusters, where $G_{s,k}(i)$ is used to denote $G_{s,k}$-feature of $X(i)$,

$$\begin{cases} G_{s,k}(i) = \{G_{s,k}^1(i), G_{s,k}^2(i), \cdots, G_{s,k}^m(i)\} \\ G_{s,k}^l(i) = \dfrac{N\{G_{s,k}^l\text{-cluster} \cap Tr_{s,l}\text{-cluster}\}}{N\{Tr_{s,l}\text{-cluster}\}} \end{cases}, \quad (4)$$

$N\{G^l_{s,k}\text{-cluster} \bigcap Tr_{s,l}\text{-cluster}\}$ is the virus number of the intersection of $G^l_{s,k}$-cluster and $Tr_{s,l}$-cluster, and $N\{Tr_{s,l}\text{-cluster}\}$ is the virus number of $Tr_{s,l}$-cluster.

And then, G_s-features of viruses are constructed by $G_{s,k}$-features, where $G_s(i)$ is used to denote G_s-feature of $X(i)$, and

$$\begin{cases} G_s(i) = \{G^1_s(i), G^2_s(i), \cdots, G^m_s(i)\} \\ G^l_s(i) = \sum_{k=1}^{15} G^l_{s,k}(i) \end{cases}. \tag{5}$$

Here, $X(i)$ is put to $G_{s,l}$-cluster if

$$G^l_s(i) = \max\{G^1_s(i), \cdots, G^m_s(i)\}, \tag{6}$$

where we use $G_{s,l}$-cluster to denote the l-th G-cluster of data-s.

I-features and I-clusters

For any type of F_k-features, *Icc*-cluster algorithm with them divided viruses of data-s into mini-clusters, where the clustering number is the number of viruses in training data, $I^q_{s,k}$-cluster is used to denote the q-th mini-cluster, and all $I^q_{s,k}$-clusters are specified as m categories. That is, $I^q_{s,k}$-cluster is defined as $I^{q,t}_{s,k}$-cluster if

$$\begin{cases} I^q_{s,k}(t) &= \max\{I^q_{s,k}(1), I^q_{s,k}(2), \cdots, I^q_{s,k}(m)\} \\ I^q_{s,k}(l) &= \frac{N\{I^q_{s,k}\text{-cluster} \bigcap Tr_{s,l}\text{-cluster}\}}{NTr_{s,l}\text{-cluster}} \end{cases}, \tag{7}$$

where $N\{I^q_{s,k}\text{-cluster} \bigcap Tr_{s,l}\text{-cluster}\}$ is the virus number of the intersection of $I^q_{s,k}$-cluster and $Tr_{s,l}$-cluster, and $N\{Tr_{s,l}\text{-cluster}\}$ is the virus number of $Tr_{s,l}$-cluster.

Here, these $I^{q,t}_{s,k}$-clusters are used to construct m $I^l_{s,k}$-clusters, where

$$I^l_{s,k} = \bigcup_{t=l} I^{q,t}_{s,k}. \tag{8}$$

Then, $I_{s,k}$-features of viruses are constructed by $I^l_{s,k}$-clusters, where $I_{s,k}(i)$ is used to denote $I_{s,k}$-feature of $X(i)$,

$$\begin{cases} I_{s,k}(i) = \{I^1_{s,k}(i), I^2_{s,k}(i), \cdots, I^m_{s,k}(i)\} \\ I^l_{s,k}(i) = \frac{N\{I^l_{s,k}\text{-cluster} \bigcap Tr_{s,l}\text{-cluster}\}}{N\{Tr_{s,l}\text{-cluster}\}} \end{cases}, \tag{9}$$

$N\{I^l_{s,k}\text{-cluster} \bigcap Tr_{s,l}\text{-cluster}\}$ is the virus number of the intersection of $I^l_{s,k}$-cluster and $Tr_{s,l}$-cluster, and $N\{Tr_{s,l}\text{-cluster}\}$ is the virus number of $Tr_{s,l}$-cluster.

And then, I_s-features of viruses are constructed by $I_{s,k}$-features, where $I_s(i)$ is used to denote I_s-feature of $X(i)$, and

$$\begin{cases} I_s(i) = \{I^1_s(i), I^2_s(i), \cdots, I^m_s(i)\} \\ I^l_s(i) = \sum_{l=1}^{15} I^l_{s,k}(i) \end{cases}. \tag{10}$$

Here, $X(i)$ is put to $I_{s,l}$-cluster if

$$I_s^l(i) = \max\{I_s^1(i), \cdots, I_s^m(i)\}, \tag{11}$$

where we use $I_{s,l}$-cluster to denote the l-th I-cluster of data-s.

GI-features and GI-clusters
For $X(i)$ of data-s, its GI_s-feature is the sum of its G_s-feature and I_s-feature. That is,

$$\begin{array}{rcl} GI_s(i) & = & G_s(i) + I_s(i) = \{GI_s^1(i), \cdots, GI_s^m(i)\} \\ & = & \{G_s^1(i) + I_s^1(i), \cdots, G_s^m(i) + I_s^m(i)\}. \end{array} \tag{12}$$

Here, for $X(i)$ of data-s, it is put to $GI_{s,l}$-cluster if

$$GI_s^l(i) = \max\{GI_s^1(i), \cdots, GI_s^m(i)\}, \tag{13}$$

where we use $GI_{s,l}$-cluster to denote its l-th GI-cluster of data-s.

0.0.1 Predicting the types of viruses of test data
For $Z(n)$ of test-s, it is predicted as the l-th family if it belongs to $GI_{s,l}$-cluster.

Results

Results of F_k-features of phages
Here, viruses of $D_{s,l}$-clusters($s = 1, 2, 3, 4$) were mapped on t-SNE maps(Fig 1) by their F_4-features, where these viruses were marked by their families. Fig 1 (a) showed that phages of Data-1 spread over four regions, but these regions did not contain other family viruses. For instance, some phages of Data-1 projected among Alphaviruses, Arteriviruses, and Flaviviruses, but these different family viruses had clear boundary.

For t-SNE map of 6 phage families of Data-2, results showed that most of the same family phages were located in the same regions, but some different family phages mixed on their boundaries(Fig 1 (b)). Furthermore, we also used t-SNE map of other F_k-features to display Data-2, but these maps were not able to separate these 6 family phages also. That is, for some phages, their nearest neighbors came from other family ones when the similarity were defined by F_k-features.

However, Fig 1 (c) and (d) showed that t-SNE maps gave $D_{s,l}$-clusters($s = 3, 4$) good separation, where viruses of $D_{3,l}$-cluster($l = 1, \cdots, 9$) were the same as $D_{1,l+1}$-cluster. In fact, most of F_k-features were able to distinguish these 9 family viruses of Data-3 and 5 family Ebolaviruses of Data-4 also. Importantly, $D_{s,l}$-clusters($s = 3, 4$) had clear boundaries.

Figure 1 The t-SNE maps of virus of data-s, where t-SNE maps were generated from F_4-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 $D_{1,l}$-clusters. **(b)** The t-SNE maps of 6 $D_{2,l}$-clusters. **(c)** The t-SNE maps of 9 $D_{3,l}$-clusters. **(d)** The t-SNE maps of 4 $D_{4,l}$-clusters.

Results of G-features of phages

Here, for viruses of data-s($s = 1, 2, 3, 4$), their $D_{s,l}$-clusters were mapped on t-SNE maps(Fig 2) by their G-features, respectively, where these viruses were marked by their families. Fig 2 (a) showed that projections of 10 $D_{1,l}$-clusters of data-1 had clear boundaries, and only 2 phages were projected in regions of Alphaviruses and Arteriviruses, respectively.

For 6 family phages of Data-2, t-SNE map of G-features(Fig 2 (b)) showed that only Enterobacteria and Mycobacterium mapped on together, but other four families spread over two regions. Moreover, six regions of t-SNE map contained different families ones, but these mixed regions mainly contained the same family ones. That is, for the same family phages, most of them were able to search the nearest neighbors in their families.

Moreover, Fig 2 (c) showed that 9 $D_{3,l}$-clusters were able to be separated by their G-features. But for one Ebolavirus(FJ217162) of $D_{4,4}$-cluster, it was projected in regions of $D_{4,1}$-cluster. The reason was that $D_{4,4}$-cluster only contained the single Ebolavirus, it did not search the nearest neighbors in $D_{4,4}$-cluster.

Figure 2 The t-SNE maps of viruses of data-s, where t-SNE maps were generated from G-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 $D_{1,l}$-clusters. **(b)** The t-SNE maps of 6 $D_{2,l}$-clusters. **(c)** The t-SNE maps of 9 $D_{3,l}$-clusters. **(d)** The t-SNE maps of 4 $D_{4,l}$-clusters.

Results of I-features

Here, for viruses of data-s($s = 1, 2, 3, 4$), their $D_{s,l}$-clusters were mapped on t-SNE maps(Fig 3) by their I-features, respectively, where these viruses were marked by their families. For any data-s, Fig 3 showed that its one region mixed some viruses that came many families, but any family only had few ones, where these mixed viruses came from test data, and their I-features were zero vectors. In fact, since the clustering number of Icc-cluster algorithm was equal to the number of viruses in training data, its many mini-clusters only contained one virus. Thus, for viruses in these mini-clusters that only contained themselves, if they came from test-s, their I-features were zero vectors.

Figure 3 The t-SNE maps of viruses of data-s, where t-SNE maps were generated from I-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 $D_{1,l}$-clusters. **(b)** The t-SNE maps of 6 $D_{2,l}$-clusters. **(c)** The t-SNE maps of 9 $D_{3,l}$-clusters. **(d)** The t-SNE maps of 4 $D_{4,l}$-clusters.

Results of GI-features

Here, for viruses of data-s($s = 1, 2, 3, 4$), their $D_{s,l}$-clusters were mapped on t-SNE maps(Fig 4) by their GI-features respectively, where these viruses were marked by their families. For the same family phages, Fig 4 (b) showed that they spread over two or more regions, but any region hardly contained other family ones.

Moreover, Fig 4 showed that t-SNE maps gave $D_{s,l}$-clusters($s = 1, 3, 4$) good separation also. That is, GI-features were able to separate phages from other family viruses, distinguish 9 family viruses of Data-3 and 5 family Ebolaviruses of Data-4 also.

Figure 4 The t-SNE maps of viruses of data-s, where t-SNE maps were generated from GI-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 $D_{1,l}$-clusters. **(b)** The t-SNE maps of 6 $D_{2,l}$-clusters. **(c)** The t-SNE maps of 9 $D_{3,l}$-clusters. **(d)** The t-SNE maps of 4 $D_{4,l}$-clusters.

Comparison of G-features, I-features and GI-features of phages

Here, for phages in 6 $D_{2,l}$-clusters, the curve shapes of their G-features, I-features and GI-features were shown in Fig 5. For phages of any $D_{2,l}$-cluster, Fig 5 showed that the profiles of their GI-feature were hardly difference. Importantly, for phages of different $D_{2,l}$-cluster, the position of the maximum value of their GI-features were different. Thus, GI-clusters were able to group the same family phages of Data-2 together.

Moreover, for phages in the same $D_{2,l}$-clusters, the curve shape of their G-features were relatively chaotic, and I-features of few phages were zero vectors.

Figure 5 The profile plots of G-features, I-features and GI-features of 6 $D_{2,l}$-clusters, where the X-axis represented the positions of the feature components, the Y-axis represented the value of the feature components. (a_1), (b_1) and (c_1) The profiles of G-features, I-features and GI-features of $D_{2,1}$-cluster. **(a_2), (b_2) and (c_2)** The profiles of G-features, I-features and GI-features of $D_{2,2}$-cluster. **(a_3), (b_3) and (c_3)** The profiles of G-features, I-features and GI-features of $D_{2,3}$-cluster. **(a_4), (b_4) and (c_4)** The profiles of G-features, I-features and GI-features of $D_{2,4}$-cluster. **(a_5), (b_5) and (c_5)** The profiles of G-features, I-features and GI-features of $D_{2,5}$-cluster. **(a_6), (b_6) and (c_6)** The profiles of G-features, I-features and GI-features of $D_{2,6}$-cluster.

Results of G-clusters

Here, for viruses of data-s($s = 1,2,3,4$), their G-clusters were constructed by Eq.(6), where the statistical results of these G-clusters were summarized in Table 1. For phages of Data-2, Table 1 showed that 2 Enterobacteria, 2 Mycobacterium, 3 Prochlorococcus, 2 Pseudomonas, 1 Synechococcus and 3 Vibrio were misjudged by $G_{2,l}$-clusters, respectively. That is, 15.7%(13/83) phages of Data-2 were misjudged by $G_{2,l}$-clusters. Moreover, for 432 viruses of Data-1, only 2 phages were misjudged as Alphavirus and Flavivirus by G-clusters, respectively.

Furthermore, all viruses of Data-3 were correctly distinguished by G-clusters, and only 1 Ebolavirus of Data-4 were misjudged by G-clusters. In fact, the misidentified Ebolavirus came from $D_{4,4}$-cluster that only contained itself. Since mini-groups that were generated from MG-Euclidean algorithm contained two samples at least, G-clusters contained two samples at least. Thus, the misidentified Ebolavirus was caused by the imperfection of MG-Euclidean algorithm.

Results of I-clusters

Here, for viruses of data-s($s = 1,2,3,4$), their I-clusters were constructed by Eq.(11), where the statistical results of these G-clusters were summarized in Table 2. For I-clusters of all data sets, Table 2 showed that all of their sensitivity were 100%, but most of their specificity were less than 1. For instance, Table 2 showed that 1 Enterobacteria, 1 Mycobacterium, 1 Prochlorococcus, 1 Pseudomonas, 1 Synechococcus and 4 Vibrio of Data-2 were incorrectly distinguished

by $I_{2,l}$-clusters, respectively. In fact, these incorrectly distinguished viruses came from $Te_{s,l}$-clusters, and their I-feature were zero vectors.

The reliability of GI-clusters

Here, for viruses of data-s($s = 1, 2, 3, 4$), their GI-clusters were constructed by Eq.(13), where the statistical results of these GI-clusters were summarized in Table 3. For GI-clusters of all data sets, Table 3 showed that only 1 Synechococcus of Data-2 were misjudged as Vibrio.

Predicting the phages of test data

Here, for viruses of test-s($s = 1, 2, 3, 4$), their biological types(or subtypes) were predicted by GI-clusters, where these predicted results were summarized in Table 3. For viruses of of all test sets, Table 3 showed that only 1 Synechococcus of test-2 were incorrectly predicted as Vibrio. These results demonstrated that GI-clusters were able to separate phages from other virus families, and grouped the same phage families together also.

Moreover, for viruses of test-s($s = 1, 2, 3, 4$), their families were predicted by G-clusters and I-clusters also, where these distinguishing results were summarized in Table 1 and Table 2, respectively. Table 1 showed that G-clusters were able to separate almost all of phages from other virus families, but did not group the same phage families together. For instance, Table 1 showed that 1 Prochlorococcus, 1 Pseudomonas, 1 Synechococcus and 1 Vibrio of test-2 were misjudged by $G_{2,l}$-clusters, respectively. Furthermore, Table 2 showed that I-clusters were not able to separate phages from other family viruses, and did not group the same family phages together also. In details, Table 2 showed that 2 Phages, 1 Ebolaviruses, 3 Flaviviruses, 2 Hiv, 1 Marburgviruses, and 1 Pestiviruses of test-1 were not distinguished by $I_{1,l}$-clusters, and also 1 Enterobacteria, 1 Mycobacterium, 1 Prochlorococcus, 1 Pseudomonas, 1 Synechococcus and 4 Vibrio of test-2 were did not distinguished by $I_{2,l}$-clusters.

Discussion

In theory, for each sample, MG-Euclidean algorithm firstly searches its nearest neighbor, and then puts it and its nearest neighbor to the same mini-group. Moreover, MG-Euclidean algorithm construct too many mini-groups that is far beyond the number of sample families. To merge the same families of mini-groups together, samples of training data is used to define the families of mini-groups. Furthermore, if the number of the same family viruses is relative small, some mini-groups may contain these nearest neighbor that come from the different families. The reason is that the some samples do not search their nearest neighbors in themselves families.

To compensate the shortcomings of G-clusters, I-clusters are constructed, where I-clusters are able to assure that the samples of training data are correctly distinguished. The reason is that the number of mini-clusters is the number of viruses in training data. Moreover, for these samples of test data that are misjudged by I-clusters, their mini-clusters do not contain samples of training data, so their I-features were zero vectors. To make full use of the togetherness with G-clusters

and I-clusters, GI-features are constructed by the sum of G-features and I-features. For each sample, results show that it is able to be correctly distinguished by GI-clusters if it is able to be put correct I-cluster or G-cluster.

Conclusion

In this study, we use GI-clusters to separate phages from other family viruses, group the same family phages together, and correctly predict the families of unknown phages. We hoped GI-clusters are able to help the researchers to distinguish the families of phages.

Abbreviations

$D_{s,l}$-cluster: it contains these viruses that belong to the l-th family of Data-s.
$Tr_{s,l}$-cluster: it contains these viruses that belong to the l-th family of training-s.
$Te_{s,l}$-cluster: it contains these viruses that belong to the l-th family of test-s.
F_k-features: $F_1X(i)$, $F_2X(i)$, $F_3X(i)$ and $F_4X(i)$ are the frequency of mononucleotide, dinucleotide, trinucleotide and quadnucleotide respectively, and other F_k-features($(k > 4)$) are their various combinations.
$G^p_{s,k}$-group: the p-th mini-group of Data-s that is generated from MG-Euclidean with F_k-feature.
$G^{p,t}_{s,k}$-group: $G^p_{s,k}$-group are re-defined by Eq.(2).
$G^l_{s,k}$-cluster: it is the union of these $G^{p,t}_{s,k}$-groups that their t- parameters are equal to l(Eq.(3)).
$G_{s,k}$-feature: it is defined by $G^l_{s,k}$-clusters(Eq.(4)).
G_s-feature: it is the sum of all $G_{s,k}$-features(Eq.(5)).
$G_{s,l}$-cluster: it is the l-th G-cluster of data-s that is defined by Eq.(6).
$I^q_{s,k}$-cluster: it is the q-th mini-cluster of data-s that is generated from Icc-cluster algorithm with F_k-feature.
$I^{q,t}_{s,k}$-cluster: it is $I^q_{s,k}$-cluster are re-defined by Eq.(7).
$I^l_{s,k}$-cluster: it is the union of these $I^{q,t}_{s,k}$-clusters that their t- parameters are equal to l(Eq.(8)).
$I_{s,k}$-feature: it is defined by $I^l_{s,k}$-clusters(Eq.(9)).
I_s-feature: it is the sum of all $I_{s,k}$-features(Eq.(10)).
$I_{s,l}$-cluster: it is the l-th I-cluster of data-s(Eq.(11)).
GI_s-feature: it is the sum of G_s-feature and I_s-feature(Eq.(12)).
$GI_{s,l}$-cluster: it is generated from GI_s-features(Eq.(13)).

Declarations

Acknowledgements
This work rests almost entirely on open data. Contributors were gratefully acknowledged. Moreover, This work is supported by Major Program of National Natural Science Foundation of China(2016YFA0501600).

Funding
This work was supported by Major Program of National Natural Science Foundation of China(2016YFA0501600).

Availability of data and materials
The data sets were collected from the NCBI database. The more detailed report on data set was included in the article, in "Materials and Methods" section.

Author's contributions
XJ analyzed and discussed the model, and wrote the manuscript. QH performed a portion of the model. ZL supervised the study. All co-authors actively commented and improved the manuscript, as well as finally read and approved the final manuscript.

Competing interests
The authors declared that they had no competing interests.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declared that they had no competing interests.

Additional Files

Additional file
Additional file 1: the details of Data-s, training-s and test-s.

Author details
[1]School of Mathematics, Southeast University , 210096 Nanjing, PR China. [2]Department of Mathematics, Nanjing Forestry University, 210037 Nanjing, PR China. [3]State Key Laboratory of Bioelectronics, School of Biological Science and Medical Engineering, Southeast University, 210096 Nanjing, PR China.

References

1. Shapiro, J.W.; Putonti, C. Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution. MBio 2018, 9, 1-14.
2. Chow, C.-E.T.; Suttle, C.A. Biogeography of Viruses in the Sea. Annu. Rev. Virol. 2015, 2, 41-66.
3. Classifying the Unclassified: A Phage Classification Method. Viruses 2019, 11(2)
4. Roux, S.; Solonenko, N.E.; Dang, V.T.; Poulos, B.T.; Schwenck, S.M.; Goldsmith, D.B.; Coleman, M.L.; Breitbart, M.; Sullivan, M.B. Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ 2016, 4, e2777.
5. Adriaenssens, E.M.; Rodney Brister, J. How to name and classify your phage: An informal guide. Viruses 2017, 9, 70.
6. Krupovic, M.; Dutilh, B.E.; Adriaenssens, E.M.; Wittmann, J.; Vogensen, F.K.; Sullivan, M.B.; Rumnieks, J.; Prangishvili, D.; Lavigne, R.; Kropinski, A.M.; et al. Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee. Arch. Virol. 2016, 161, 1095-1099.
7. Lefkowitz, E.J.; Dempsey, D.M.; Hendrickson, R.C.; Orton, R.J.; Siddell, S.G.; Smith, D.B. Virus taxonomy: The database of the International Committee on Taxonomy of Viruses (ICTV). Nucleic Acids Res. 2018, 46, D708-D717.
8. Simmonds, P.; Adams, M.J.; Benk, M.; Breitbart, M.; Brister, J.R.; Carstens, E.B.; Davison, A.J.; Delwart, E.; Gorbalenya, A.E.; Harrach, B.; et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat. Rev. Microbiol. 2017, 15, 161-168.
9. Meier-Kolthoff, J.P.; G?ker, M. VICTOR: Genome-based phylogeny and classification of prokaryotic viruses. Bioinformatics 2017, 33, 3396-3404.
10. Shapiro, J.W.; Putonti, C. Gene Co-occurrence Networks Reflect Bacteriophage Ecology and Evolution. MBio 2018, 9, 1-14.
11. Roux, S.; Hallam, S.J.; Woyke, T.; Sullivan, M.B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. Elife 2015, 4, 1-20.
12. Lima-Mendez, G.; Van Helden, J.; Toussaint, A.; Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. Mol. Biol. Evol. 2008.
13. Iranzo, J.; Krupovic, M.; Koonin, E.V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. MBio 2016, 7, 1-21.
14. Deschavanne, P.; DuBow, M.S.; Regeard, C. The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virol. J. 2010, 7, 1-12.
15. Jia X, Han Q, Lu Z: Analyzing the similarity of samples and genes by MG-PCC algorithm, t-SNE-SS and t-SNE-SG maps. BMC Bioinformatics 2018, 19(1):512.
16. Jia X, Liu Y, Han Q, Lu Z: Multiple-cumulative probabilities used to cluster and visualize transcriptomes. *FEBS Open Bio* 2017, 7(12):2008-2020.
17. He F, Han Y, Gong J, Song J, Wang H, Li Y: Predicting siRNA efficacy based on multiple selective siRNA representations and their combination at score level. Sci Rep 2017, 7:44836.

Tables

Table 1 The distinguishing results of all $G_{s,l}$-clusters.

	Types(or Subtypes)	N1	N2	N3	$Se^{'}$	$Sp^{'}$	N4	N5	N6	Se	Sp
Data-1											
	Phages	143	141	141	100	98.6	28	27	27	100	96.4
	Alphaviruses	78	79	78	98.7	100	16	16	16	100	100
	Arteriviruses	11	11	11	100	100	2	2	2	100	100
	Ebolaviruses	62	62	62	100	100	12	12	12	100	100
	Flaviviruses	17	18	17	94.4	100	4	5	4	80	100
	Hiv	10	10	10	100	100	2	2	2	100	100
	Marburgviruses	40	40	40	100	100	8	8	8	100	100
	Pestiviruses	8	8	8	100	100	1	1	1	100	100
	Rubiviruses	58	58	58	100	100	12	12	12	100	100
	SARS	5	5	5	100	100	1	1	1	100	100
Data-2											
	Enterobacteria	12	13	10	76.9	83.3	2	3	2	66.7	100
	Mycobacterium	14	12	12	100	85.7	3	3	3	100	100
	Pseudoalteromonas	13	13	10	76.9	76.9	2	2	2	100	100
	Pseudomonas	7	5	5	100	71.4	2	1	1	100	50
	Synechococcus	18	20	17	85	94.4	3	2	2	100	66.7
	Vibrio	19	20	16	80	84.2	4	5	4	80	100
Data-3											
	Alphaviruses	78	78	78	100	100	16	16	16	100	100
	Arteriviruses	11	11	11	100	100	2	2	2	100	100
	Ebolaviruses	62	62	62	100	100	12	12	12	100	100
	Flaviviruses	17	17	17	100	100	4	4	4	100	100
	Hiv	10	10	10	100	100	2	2	2	100	100
	Marburgviruses	40	40	40	100	100	8	8	8	100	100
	Pestiviruses	8	8	8	100	100	1	1	1	100	100
	Rubiviruses	58	58	58	100	100	12	12	12	100	100
	SARS	5	5	5	100	100	1	1	1	100	100
Data-4											
	B	6	7	6	100	85.7	1	1	1	100	100
	R	8	8	8	100	100	1	1	1	100	100
	S	11	11	11	100	100	3	3	3	100	100
	T	1	0	0		0	0	0	0		
	Z	36	36	36	100	100	7	7	7	100	100

N1: The virus number of $D_{s,l}$-cluster. N2: The virus number of $G_{s,l}$-cluster. N3: The virus number of $D_{s,l}$-cluster $\cap$ $G_{s,l}$-cluster. N4: The virus number of $Te_{s,l}$-cluster. N5: The virus number of test-s $\cap$ $G_{s,l}$-cluster. N6: The virus number of $Te_{s,l}$-cluster $\cap$ $G_{s,l}$-cluster. $Se^{'}$(sensitivity): N3/N2. $Sp^{'}$(specificity): N3/N1. Se(sensitivity): N6/N5. Sp(specificity): N6/N4.

Table 2 The distinguishing results of all $I_{s,l}$-clusters.

	Types(or Subtypes)	N1	N2	N3	$Se^{'}$	$Sp^{'}$	N4	N5	N6	Se	Sp
Data-1											
	Phages	143	141	141	100	98.6	28	26	26	100	92.9
	Alphaviruses	78	78	78	100	100	16	16	16	100	100
	Arteriviruses	11	11	11	100	100	2	2	2	100	100
	Ebolaviruses	62	61	61	100	98.4	12	11	11	100	91.7
	Flaviviruses	17	14	14	100	82.4	4	1	1	100	25
	Hiv	10	8	8	100	80	2	0	0		0
	Marburgviruses	40	39	39	100	97.5	8	7	7	100	87.5
	Pestiviruses	8	7	7	100	87.5	1	0	0		0
	Rubiviruses	58	58	58	100	100	12	12	12	100	100
	SARS	5	5	5	100	100	1	1	1	100	100
Data-2											
	Enterobacteria	12	11	11	100	91.7	2	1	1	100	50
	Mycobacterium	14	13	13	100	92.9	3	2	2	100	66.7
	Pseudoalteromonas	13	12	12	100	92.3	2	1	1	100	50
	Pseudomonas	7	6	6	100	85.71	2	1	1	100	50
	Synechococcus	18	17	17	100	94.4	3	2	2	100	66.7
	Vibrio	19	15	15	100	78.9	4	0	0		0
Data-3											
	Alphaviruses	78	78	78	100	100	16	16	16	100	100
	Arteriviruses	11	11	11	100	100	2	2	2	100	100
	Ebolaviruses	62	62	61	100	98.4	12	11	11	100	91.7
	Flaviviruses	17	17	14	100	82.4	4	1	1	100	25
	Hiv	10	10	8	100	80	2	0	0		0
	Marburgviruses	40	40	39	100	97.5	8	7	7	100	87.5
	Pestiviruses	8	8	7	100	87.5	1	0	0		0
	Rubiviruses	58	58	58	100	100	12	12	12	100	100
	SARS	5	5	5	100	100	1	1	1	100	100
Data-4											
	B	6	6	6	100	100	1	1	1	100	100
	R	8	7	7	100	87.5	1	0	0		0
	S	11	11	11	100	100	3	3	3	100	100
	T	1	1	1	100	100	0	0	0		
	Z	36	35	35	100	97.2	7	6	6	100	85.7

N1, N2, N3, N4, N5, N6, $Se^{'}$, $Sp^{'}$, Se and Sp had define in Table 1.

Table 3 The distinguishing results of all $GI_{s,l}$-clusters.

	Types(or Subtypes)	N1	N2	N3	Se'	Sp'	N4	N5	N6	Se	Sp
Data-1											
	Phages	143	143	143	100	100	28	28	28	100	100
	Alphaviruses	78	78	78	100	100	16	16	16	100	100
	Arteriviruses	11	11	11	100	100	2	2	2	100	100
	Ebolaviruses	62	62	62	100	100	12	12	12	100	100
	Flaviviruses	17	17	17	100	100	4	4	4	100	100
	Hiv	10	10	10	100	100	2	2	2	100	100
	Marburgviruses	40	40	40	100	100	8	8	8	100	100
	Pestiviruses	8	8	8	100	100	1	1	1	100	100
	Rubiviruses	58	58	58	100	100	12	12	12	100	100
	SARS	5	5	5	100	100	1	1	1	100	100
Data-2											
	Enterobacteria	12	12	12	100	100	2	2	2	100	100
	Mycobacterium	14	14	14	100	100	3	3	3	100	100
	Pseudoalteromonas	13	13	13	100	100	2	2	2	100	100
	Pseudomonas	7	7	7	100	100	2	2	2	100	100
	Synechococcus	18	17	17	100	94.4	3	2	2	100	66.7
	Vibrio	19	20	19	95	100	4	5	4	80	100
Data-3											
	Alphaviruses	78	78	78	100	100	16	16	16	100	100
	Arteriviruses	11	11	11	100	100	2	2	2	100	100
	Ebolaviruses	62	62	62	100	100	12	12	12	100	100
	Flaviviruses	17	17	17	100	100	4	4	4	100	100
	Hiv	10	10	10	100	100	2	2	2	100	100
	Marburgviruses	40	40	40	100	100	8	8	8	100	100
	Pestiviruses	8	8	8	100	100	1	1	1	100	100
	Rubiviruses	58	58	58	100	100	12	12	12	100	100
	SARS	5	5	5	100	100	1	1	1	100	100
Data-4											
	B	6	6	6	100	100	1	1	1	100	100
	R	8	8	8	100	100	1	1	1	100	100
	S	11	11	11	100	100	3	3	3	100	100
	T	1	1	1	100	100	0	0	0		
	Z	36	36	36	100	100	7	7	7	100	100

N1, N2, N3, N4, N5, N6, Se', Sp', Se and Sp had define in Table 1.

Figures

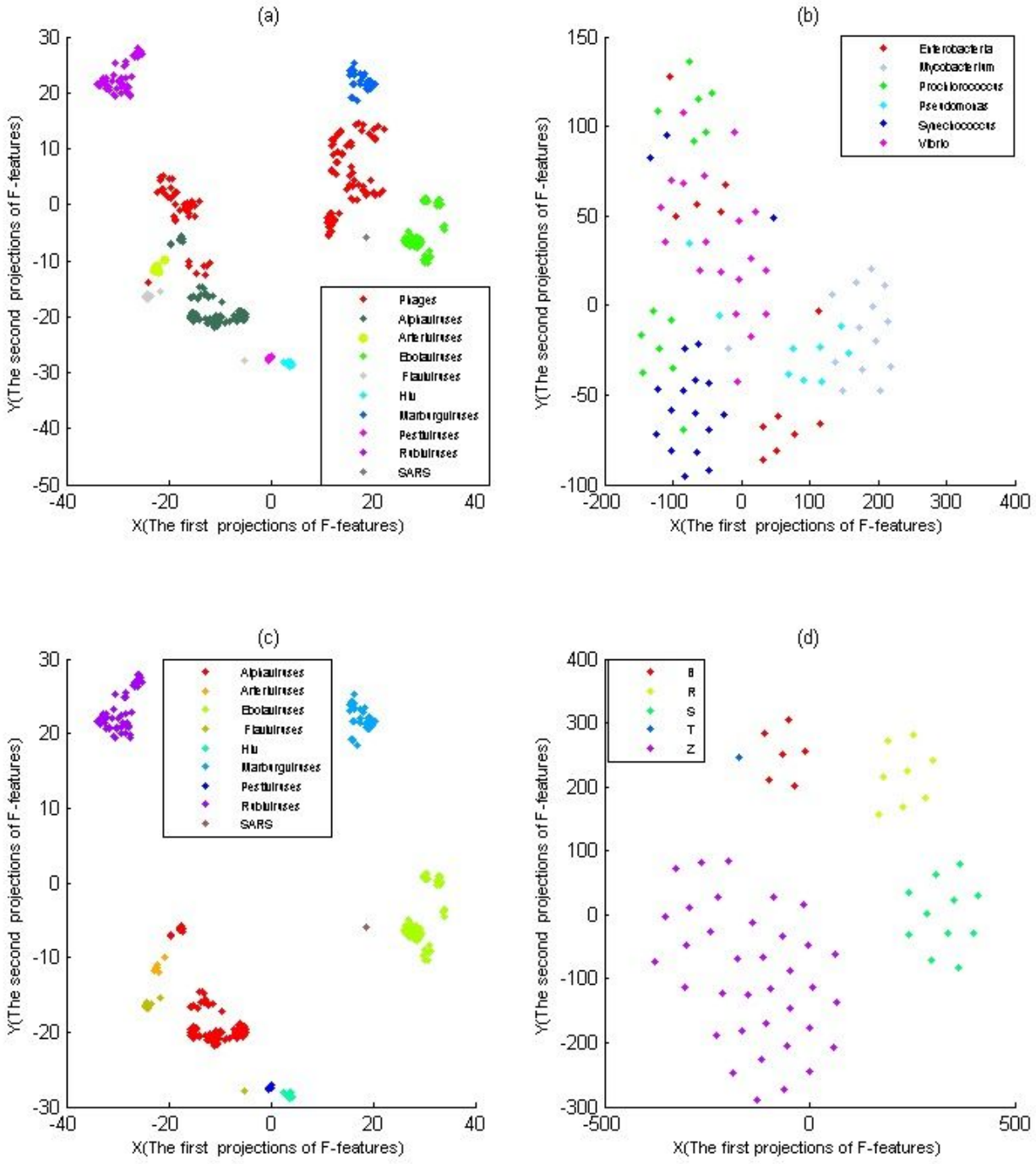

Figure 1

The t-SNE maps of virus of data-s, where t-SNE maps were generated from F4-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 D1,l-clusters. (b) The t-SNE maps of 6 D2,l-clusters. (c) The t-SNE maps of 9 D3,l-clusters. (d) The t-SNE maps of 4 D4,l-clusters.

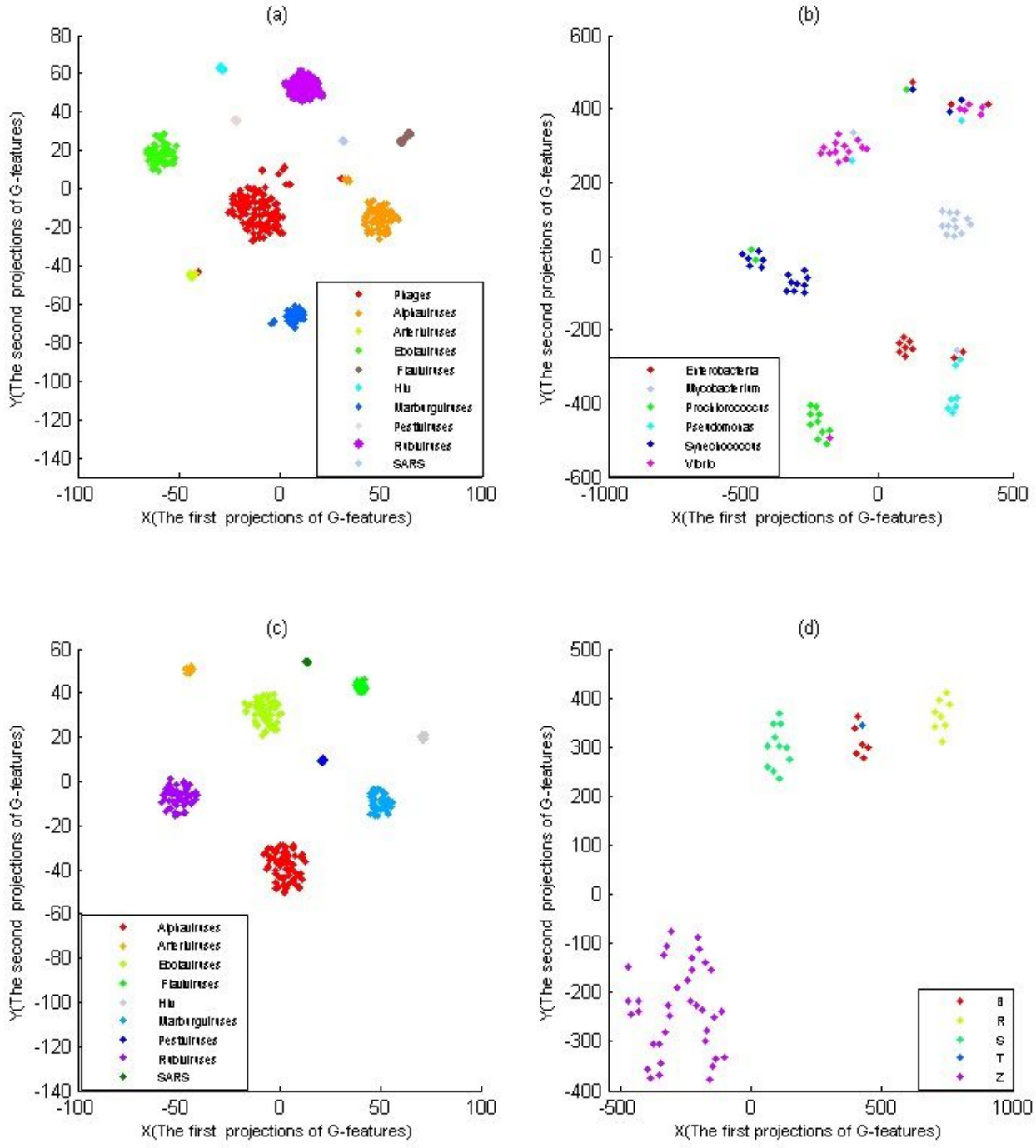

Figure 2

The t-SNE maps of viruses of data-s, where t-SNE maps were generated from G-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 D1,l-clusters. (b) The t-SNE maps of 6 D2,l-clusters. (c) The t-SNE maps of 9 D3,l-clusters. (d) The t-SNE maps of 4 D4,l-clusters.

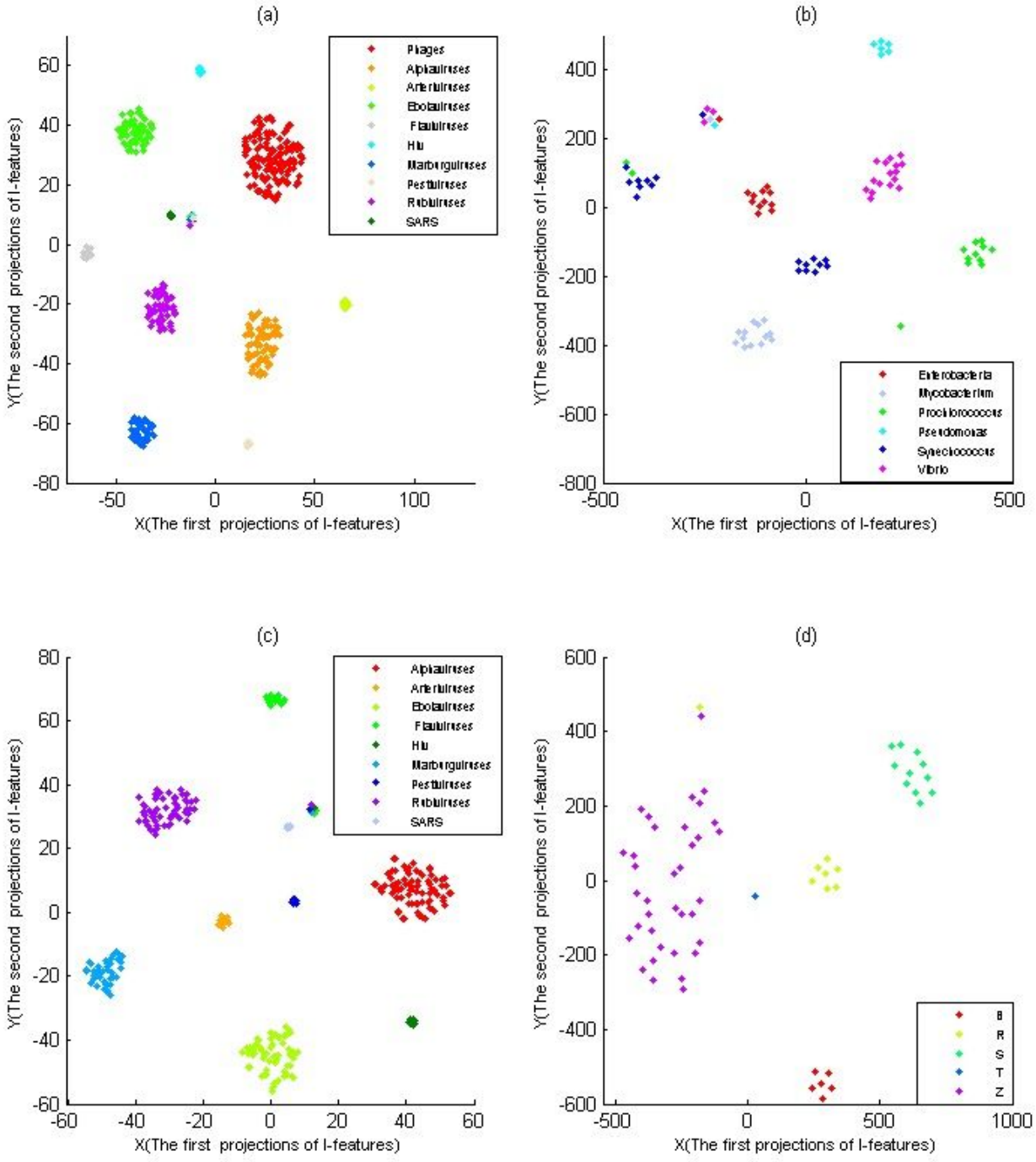

Figure 3

The t-SNE maps of viruses of data-s, where t-SNE maps were generated from l-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 D1,l-clusters. (b) The t-SNE maps of 6 D2,l-clusters. (c) The t-SNE maps of 9 D3,l-clusters. (d) The t-SNE maps of 4 D4,l-clusters.

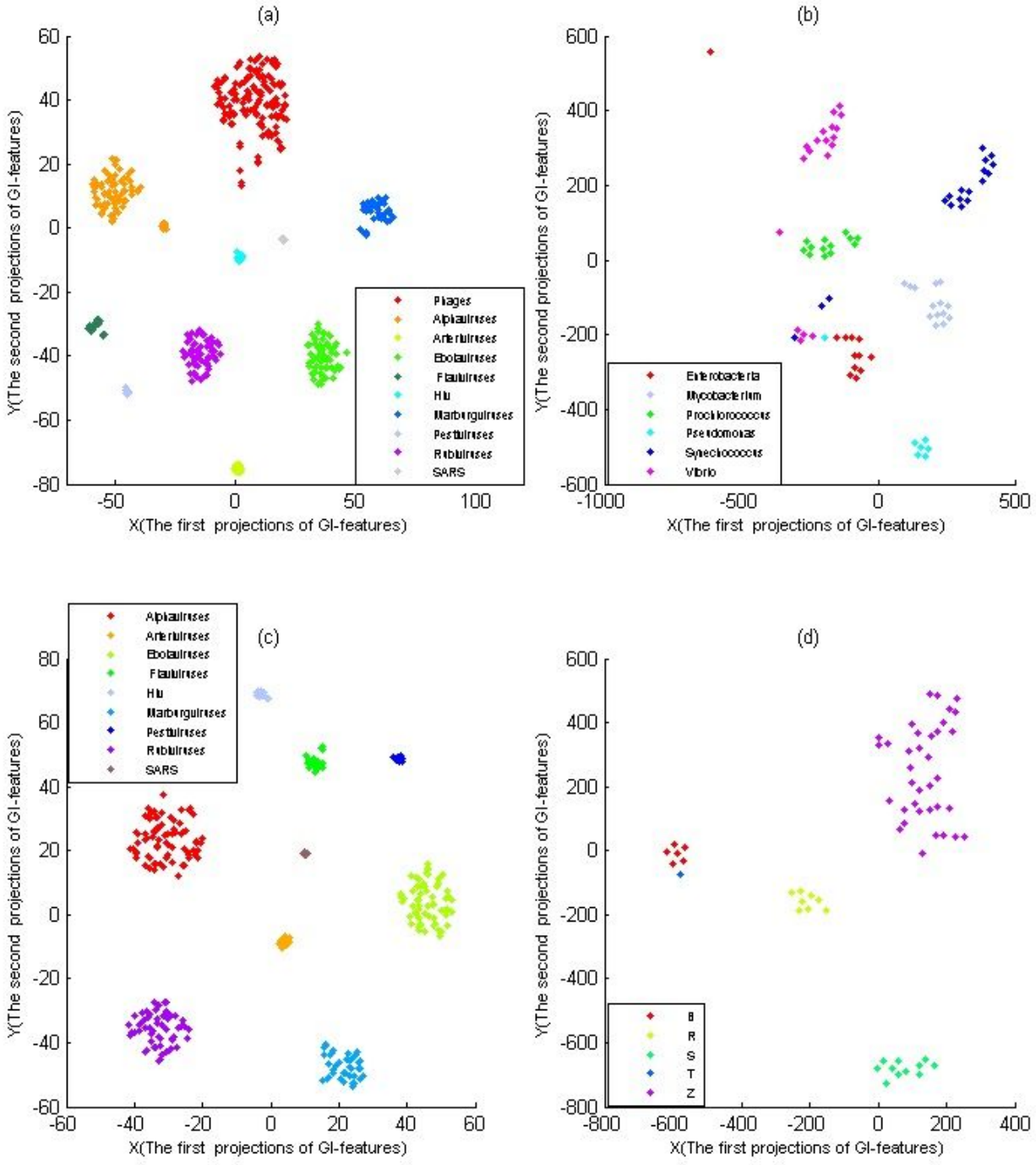

Figure 4

The t-SNE maps of viruses of data-s, where t-SNE maps were generated from GI-features, and viruses were colored according to their families. (a) The t-SNE maps of 10 D1,l-clusters. (b) The t-SNE maps of 6 D2,l-clusters. (c) The t-SNE maps of 9 D3,l-clusters. (d) The t-SNE maps of 4 D4,l-clusters.

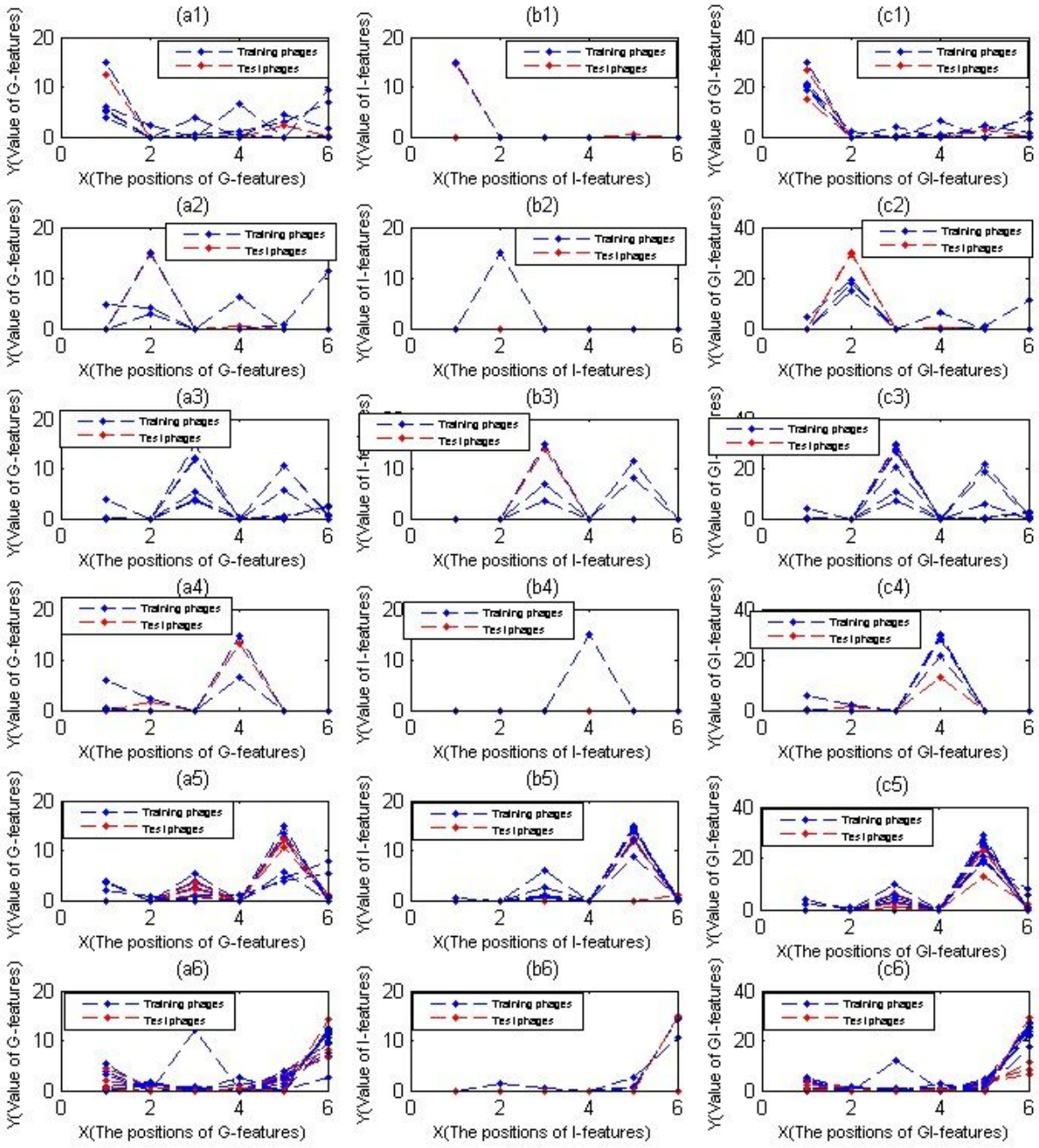

Figure 5

The prole plots of G-features, I-features and GI-features of 6 D2,I-clusters, where the X-axis represented the positions of the feature components, the Y-axis represented the value of the feature components. (a1), (b1) and (c1) The proles of G-features, I-features and GI-features of D2,1-cluster. (a2), (b2) and (c2) The proles of G-features, I-features and GI-features of D2,2-cluster. (a3), (b3) and (c3) The proles of G-features, I-features and GI-features of D2,3-cluster. (a4), (b4) and (c4) The proles of G-features, I-features and GI-

features of D2,4-cluster. (a5), (b5) and (c5) The proles of G-features, I-features and GI-features of D2,5-cluster. (a6), (b6) and (c6) The proles of G-features, I-features and GI-features of D2,6-cluster.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xlsx