Hindawi Publishing Corporation
BioMed Research International
Volume 2017, Article ID 5404180, 7 pages
http://dx.doi.org/10.1155/2017/5404180

Research Article

Construction of Multilevel Structure for Avian Influenza Virus System Based on Granular Computing

Yang Li, Qi-Hao Liang, Meng-Meng Sun, Xu-Qing Tang, and Ping Zhu

School of Science, Jiangnan University, Wuxi 214122, China

Correspondence should be addressed to Ping Zhu; zhuping@jiangnan.edu.cn

Received 11 September 2016; Revised 1 December 2016; Accepted 14 December 2016; Published 16 January 2017

Academic Editor: Hao-Teng Chang

Exploring the genetic structure of influenza viruses attracts the attention in the field of molecular ecology and medical genetics, whose epidemics cause morbidity and mortality worldwide. The rapid variations in RNA strand and changes of protein structure of the virus result in low-accuracy subtyping identification and make it difficult to develop effective drugs and vaccine. This paper constructs the evolutionary structure of avian influenza virus system considering both hemagglutinin and neuraminidase protein fragments. An optimization model was established to determine the rational granularity of the virus system for exploring the intrinsic relationship among the subtypes based on the fuzzy hierarchical evaluation index. Thus, an algorithm was presented to extract the rational structure. Furthermore, to reduce the systematic and computational complexity, the granular signatures of virus system were identified based on the coarse-grained idea and then its performance was evaluated through a designed classifier. The results showed that the obtained virus signatures could approximate and reflect the whole avian influenza virus system, indicating that the proposed method could identify the effective virus signatures. Once a new molecular virus is detected, it is efficient to identify the homologous virus hierarchically.

1. Introduction

Exploring the genetic structure of biological population attracts the focus in the field of population biology, molecular ecology, and medical genetics [1]. Influenza A virus is a negative-strand RNA virus, which encodes the 8 structural proteins and 2 nonstructural proteins. In the past several decades, some subtypes of influenza viruses have been identified to infect humans, whose epidemics cause morbidity and mortality worldwide [2, 3]. Subtyping identification of a virus is typically based on viral hemagglutinin (HA) and neuraminidase (NA) fragments among the 10 encoded proteins [4, 5]. So far, dozens of subtypes, combination of the 16 HA and 9 NA types, make up the whole viral system and it was verified that different labeled viruses descend from the same ancestor according to microscopic structural features and genome organization analysis [6]. Evolutionary forces, treated as the most important molecular mechanisms, such as natural selection acting upon rapidly mutating viral populations could shape the genetic structure of influenza viruses in different hosts, geographic regions, and periods of time with genetic mutation [7]. In addition, influenza viruses are equipped with antigenic changes, known as antigenic shifts among different subtypes of influenza viruses, which results in structural changes to escape the immunity [8]. It is of crucial importance to identify the subtypes and analyze the evolutionary relationships for developing antiviral drugs and vaccines. Thus, accessing the viral genomes in a timely fashion and developing effective analyzing methods are urgently needed.

The dramatic progress in sequencing technologies provides unprecedented prospects for the exploration of virus homologous and mutation trajectory in space and time. Understanding the evolution of influenza viruses has benefited from phylogenetic reconstructions of the hemagglutinin protein [9]. In an alternative approach, Lapedes and Farber [10] applied a technique called multidimensional scaling to study antigenic evolution of influenza. Plotkin et al. [8] clustered hemagglutinin protein sequences using the single-linkage clustering algorithm and found that influenza viruses group into several clusters. Upon the dimensional projection technique to characterize hemagglutination inhibition (HI)

data, a low-dimensional clustering method that can detect the clusters containing an incipient dominant strain was presented by He and Deem [11]. However, those works just focused on the one fragment, especially HA protein, to explore the evolutional relationships. And large volume of data poses some daunting challenges for exploring the structure of the complex system and the intrinsic relationship. Therefore, there is a need for less computationally intensive methods.

In recent years, the granular computing (GrC) theory has become a hotspot in the field of artificial intelligence and machine learning, which comes from the idea that people solve the problems from different levels and views [12]. Clustering technique is an effective way to generate granules of complex system. Y. Y. Yao and J. T. Yao accomplished a series of research work for applying the theory to data mining and some other fields [13]. Hartmann et al. [14] proposed supervised hierarchical clustering in fuzzy model identification by using hierarchical tree construction. Tang et al. [15, 16] introduced the granular space to describe the hierarchical structural information by using the algebraic topology based on the fuzzy quotient space theory [12]. He also studied the hierarchical clustering structure and analyzed the fuzzy equivalence (or proximity) relation based on the fuzzy granular space. Constructing the hierarchical structure of complex system and extracting the essential information among the granules on different granularities are the goals.

In this paper, our aim is to explore the evolutional relationships of the avian influenza viruses in the same subtype and among the subtypes considering both HA and NA fragments in the virus system. Moreover, the complex virus system should be reduced for further exploration, faced with thousands of samples in the dataset. Jointing the two protein sequences, the feature vectors are extracted from HA and NA proteins, respectively, for labeling the specific virus. Furthermore, the granular signatures in the viral granules are identified based on the obtained features to reduce the systematic and computational complexity and then its performance will be evaluated. This will provide the supports for the rationality of subtype identification. Once a new molecular virus is detected, it could be analyzed with obtained viral signatures and then the prevention and treatment measures can follow what were applied in the viral signature.

2. Materials and Methods

2.1. Materials. The influenza virus dataset was downloaded from the NCBI Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/) [17]. The influenza virus contains eight linear negative-strand RNA fragments, which encode 10 viral proteins, that is, PB1, PB2, PA, HA, NP, NA, M1, M2, NS1, and NS2, among which most are structural proteins except NS1 and NS2. Notably, HA and NA fragments play the direct and important roles in the viral subtyping identification and the functions [18]. It has been verified that 8 subgroups of avian influenza virus (H5N1, H5N2, H7N2, H7N3, H7N7, H9N2, H10N7, and H7N9) could infect people, which occurred from 1902 to 2015 around the world. The avian influenza viruses are labeled with unambiguous symbols such as the host, outbreak time, and detection sites. Removing some vague and uncompleted viruses, there are 8274 influenza viruses which reserve HA and NA protein fragments simultaneously (13143 HA protein fragments and 9401 NA protein fragments), compositing the whole avian virus protein system, denoted as Ω. According to the physicochemical property [19], amino acids are divided into four types, namely, the polar and hydrophilic (*pq*), polar and hydrophobic (*pr*), nonpolar and hydrophilic (*sq*), and nonpolar and hydrophobic (*sr*). Considering the adjacency statistical information, the 16-dimension feature vector is extracted by calculating the frequency from one protein sequence. Therefore, 32-dimension feature vector is extracted to represent a virus molecule.

2.2. The Optimization Model for Extracting the Hierarchical Structure. A relation R on a universe X is a fuzzy proximity (FP) relation if it satisfies the reflexivity and symmetry [16, 20]. Furthermore, if R is an FP relation on the universe X and satisfies the separable condition ($\forall x, y \in X$, $R(x, y) = 1 \leftrightarrow x = y$), then R is called a separable FP relation (or SFP relation).

In [16], the granular space of FP (or SFP) relations on the universe X was introduced, and then their properties were explored. Let R be an FP (or SFP) relation on a finite universe $X = \{x_1, x_2, \ldots, x_n\}$, where X is a dataset of K-dimension space. For any $\lambda \in [0, 1]$, we define a relation $R_\lambda : (x, y) \in R_\lambda \leftrightarrow R(x, y) \geq \lambda$, where R_λ is a crisp proximity relation that satisfies the reflexivity and symmetry. Then, the equivalent classes of the transitive closure $\mathrm{tr}(R_\lambda)$ can be marked by $[x]_\lambda$, which is derived by R_λ, and then $X(\lambda) = \{[x]_\lambda \mid x \in X\}$ is a granularity corresponding to λ. The set $\{X(\lambda) \mid \lambda \in [0, 1]\}$ represents a fuzzy granular space on X, which is an ordered set, and satisfies that the bigger the threshold λ is, the finer the granularity is, denoted by $\aleph_{\mathrm{TR}}(X)$ [16].

The granularity derived by λ is marked as $X(\lambda) = \{a_1, a_2, \ldots, a_{c_\lambda}\}$, where $a_i = \{x_{i1}, x_{i2}, \ldots, x_{iJ_i}\}$ satisfying the conditions that $|a_i| = J_i$ ($|\cdot|$ stands for the number of the elements in a set) and $\sum_{i=1}^{c_\lambda} J_i = n$. Some properties are explored, such as $\overline{a_i} = \sum_{k=1}^{J_i} x_{ik}/J_i$ ($i = 1, 2, \ldots, c_\lambda$) is the center of granule a_i and the center of X is $\overline{a} = \sum_{i=1}^{c_\lambda} \sum_{k=1}^{J_i} x_{ik}/n$. From the perspective of statistical theory, two indexes are introduced to measure the deviations within the classes and among the classes on the granulation $X(\lambda)$ [18, 20], defined, respectively, as follows:

$$S_{\mathrm{among}}(X(\lambda)) = \frac{\sum_{i=1}^{c_\lambda} J_i \left\| \overline{a_i} - \overline{a} \right\|_2^2}{n},$$
$$S_{\mathrm{within}}(X(\lambda)) = \frac{\sum_{i=1}^{c_\lambda} \sum_{k=1}^{J_i} \left\| x_{ik} - \overline{a} \right\|_2^2}{n}, \quad (1)$$

where $\|\cdot\|_2$ stand for the 2-norm number in K-dimension space.

By analyzing the variance within and among the classes in statistics [21], $S_{\mathrm{among}}(X(\lambda))$ is monotone increasing, with the granularity changing from the coarse to the fine, while

$S_{\text{within}}(X(\lambda))$ is gradually decreasing. Notably, the total deviation ($S(X(\lambda)) = S_{\text{among}}(X(\lambda)) + S_{\text{within}}(X(\lambda))$) is always constant $S(X(\lambda)) = \sum_{i=1}^{n} \|x_i - \overline{a}\|_2^2/n$. Additionally, $S_{\text{among}}(X(0)) = S_{\text{within}}(X(1)) = 0$ and $S_{\text{among}}(X(1)) = S_{\text{within}}(X(0)) = \sum_{i=1}^{n} \|x_i - \overline{a}\|_2^2/n$. Therefore, a fuzzy hierarchical evaluation index (FHEI) based on the fuzzy granular space is proposed as follows:

$$\text{FHEI}(X(\lambda)) = \left| S_{\text{among}}(X(\lambda)) - S_{\text{within}}(X(\lambda)) \right|. \quad (2)$$

We establish an optimization model to determine the reasonable granulation in the granular space with the minimal objective; that is, FHEI($X(\lambda)$) reaches the minimum. There exists only one $\lambda = \lambda_0$ to meet the optimization model, marked as Model (2):

$$X(\lambda_0) = \arg \min_{X(\lambda) \in \aleph_{\text{TR}}(X)} \{\text{FHEI}(X(\lambda))\}. \quad (3)$$

Remark 1. Model (2) is a global optimization model without constraints on the hierarchical structure of the finite universe X. Compared with [18], their model for determining the optimal hierarchical clustering has the restriction $S_{\text{among}}(X(\lambda)) > S_{\text{within}}(X(\lambda))$.

Given an FP relation (or SFP relation) R on the finite set $X = \{x_1, x_2, \ldots, x_n\}$ and $D = \{R(x, y) \mid x, y \in X\} = \{r_0, r_1, \ldots, r_N\}$, satisfying $1 = r_0 > r_1 > \cdots > r_N$, an algorithm is presented to detect the optimized hierarchical clustering and construct the hierarchy of complex system based on the fuzzy granular space [16].

Algorithm A.

Input: an FP relation (or SFP relation).

Output: the optimized hierarchical structure and the corresponding threshold.

Step 1

$$X(r_i) = C = \{a_1, a_2, \ldots, a_{c_1}\},$$
$$S_0 \Longleftarrow \left| S_{\text{among}}(X(r_i)) - S_{\text{within}}(X(r_i)) \right| \quad (4)$$
$$i = 0.$$

Step 2

$$i \Longleftarrow i + 1. \quad (5)$$

Step 3

$$A \Longleftarrow C. \quad (6)$$

Step 4

$$B \Longleftarrow \varnothing,$$
$$C \Longleftarrow \varnothing. \quad (7)$$

Step 5. For any $a_j \in A$, $B \Leftarrow B \cup a_j$, $A \Leftarrow A \setminus a_j$.

Step 6. For $\forall a_k \in A$, if $\exists x_j \in a_j$, $y_k \in a_k$ satisfying $R(x_j, x_k) \geq r_i$, $B \Leftarrow B \cup a_k$, $A \Leftarrow A \setminus a_k$.

Step 7

$$C \Longleftarrow \{B\} \cup C. \quad (8)$$

Step 8. If $A = \varnothing$, $X(r_i) = C$; otherwise, go to Step 5.

Step 9. If $X(r_i) \neq X(r_{i-1})$, $S_1 \Leftarrow |S_{\text{among}}(X(r_i)) - S_{\text{within}}(X(r_i))|$; otherwise, go to Step 2.

Step 10. If $S_0 > S_1$, $S_0 \Leftarrow S_1$, go to Step 2.

Step 11. Output r_{i-1}, $X(r_{i-1})$ and S_0.

The computational complexity of Algorithm A is $O(n^2)$. The concrete problems are decomposed hierarchically, which is consistent with the core idea of GrC. Given an FP (or SFP) relation on the finite set X, the optimization clustering structure constructed by Algorithm A is its first level structure. Furthermore, its second level structure is obtained if Algorithm A is repeatedly applied to all the equivalent classes in its first level structure. Therefore, Algorithm A can be used to construct multilevel structure in practical application.

2.3. Identification of Granular Signature. Once the optimal granularity of the complex system is determined, it is of crucial importance to construct information granules for abstracting original samples. Generally, the granules are obtained according to the principle: the samples with the same features assemble in one granule. And the average of all samples in one class or the center of the class is efficacious to represent the core information. Suppose that a multilevel structure (or granularity) $X^* = \{a_1, a_2, \ldots, a_J\}$ is constructed, where $J = |X^*|$. To reduce the complexity of the system, feature viruses (or signature viruses) could be extracted to approximately represent the equivalent class. According to the nearest-to-center principle, an objective function to select the signature is established, and it is formulated as follows:

$$p_i = \arg \max_{1 \leq k \leq J_i} \{R(x_{ik}, \overline{a_i})\}, \quad (9)$$

where p_i is the signature item of the granule a_i and $P = \{p_1, p_2, \ldots, p_J\}$ is a signature set of the granularity X^*. In some way, the signature set P can be used to represent approximately the complex system X.

2.4. Validation of Granular Signature Set. To evaluate the performance of selected signature set P, a classifier is designed for classifying the rest of the samples of the corresponding classes according to the principle of maximum similarity, marked as Model (3). Given a virus q_j ($\in X \setminus P$), the classifier is designed:

$$L_j = \arg \max_i \{R(q_j, p_i)\}, \quad (10)$$

Table 1: The 8 subtypes of avian influenza virus.

Subtype	Number	Subtype	Number	Subtype	Number	Subtype	Number
H5N1	306	H5N2	127	H7N3	70	H9N2	199
H7N9	24	H7N2	40	H7N7	68	H10N7	75

where $p_i \in P$, $i = 1, 2, \ldots, J$, and L_j is the class the virus q_j belongs to.

Model (3) states that the signature viruses are treated as the classifying targets and the other samples in $X \setminus P$ are assigned to $|P|$ classes. All samples in $X \setminus P$ are divided into $|P|$ classes according to Model (3), marked as b_k, $k = 1, 2, \ldots, |P|$. The accuracy ratio r is introduced to measure the efficiency of signature set for constructing the multilevel structure X^*. It is defined as

$$r = \frac{\sum_{k=1}^{|P|} |a_k \cap b_k|}{|X \setminus P|}. \tag{11}$$

In formula (11), the overlapped ratio r is proposed, which measures the rationality of the obtained signature to represent the whole virus system. And the bigger the value r is, the better the result is.

3. Results and Analysis

In this section, we apply the proposed model to the avian influenza virus system for constructing the evolutionary structure, which contains 8274 viral HA and NA protein fragments simultaneously within 8 subtypes, listed in Table 1.

Based on the feature vectors extracted from the viral HA and NA proteins, the 32-dimension vector $x_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_{32}})$ labels the specific virus x_i. Furthermore, the similarity between viruses x_i and x_j is measured:

$$R(x_i, x_j) = \frac{(x_i, x_j)}{\sqrt{(x_i, x_i) \cdot (x_j, x_j)}}, \tag{12}$$

where $(x_i, x_j) = \sum_{i=1}^{32} x_{ik} \cdot x_{jk}$ stands for the inner product in 32-dimension space. Obviously, R is an SFP relation.

The virus dataset has redundant information as many viruses are labeled with the same host, the same occurrence time, and the same outbreak sites, which could pose the obstacle to explore the intrinsic relationship and difference among the subtypes. Thus, those with the same host, the same occurrence time, and the same location combine as one new point (a representative virus), which is the preliminary system simplification, and then a unique virus database Ω^* is obtained. The FHEI is applied to virus system Ω^* containing 909 avian influenza viruses, to obtain the reasonable partition and evolutionary structure.

On the basis of the virus database Ω^*, the viral granular space (evolutionary structure) is constructed by using Algorithm A. On the first level, 3 equivalent classes were finally determined to partition the whole system, and the corresponding signature viruses are obtained, shown in Table 2. For the virus granules on the first level, class A1 contains the most viruses (about 93.5%), and granule A3 arises, containing an isolated virus (A/American green-winged teal/Washington/1595750/2014(H5N1)). Therefore, it is necessary to construct the second level of virus system. For each virus granule on the first level, Algorithm A is used repeatedly, which is to refine the granules to get the detailed evolutional structure. 14 equivalent subclasses are identified, denoted as b_k^* ($k = 1, 2, \ldots, 14$), and the virus signatures are extracted, shown in Table 3. From Tables 2 and 3, we construct the two-level feature structure of the whole virus system by using the signature viruses on first level and second level structure.

Table 2: Three signature viruses of the first level structure.

Number	Virus number	Virus signature
A1	850	A/Pekin duck/Singapore/F59/04/98(H5N2)
A2	58	A/chicken/Tunisia/145/2012(H9N2)
A3	1	A/American green-winged teal/Washington/1595750/2014(H5N1)
Sum	909	

The virus signature could be used to approximate the whole system for they are selected from the classes as the granule information. Moreover, the classifier, designed based on the principle of maximum similarity, is applied to validate the performance of virus signature. The accuracy rate of the signature virus set P^* on the second level structure is 76.57% by comparison, indicating that the second level structure of viruses system constructed by our model is effective.

Remark 2. Evaluating the performance of virus signature, the error rate is still 23.43%, which might be caused by the approximation process since all signature viruses are selected according to the nearest-to-center principle and they are not just on the center of each subclass, respectively. From the perspective of approximation, the signature set contains the most information of virus system according to the accuracy rate 76.57%. Therefore, the signature virus set P^* containing 14 viruses can be used to approximate the whole system containing 909 viruses.

The phylogenetic tree of the signature virus set P^* can be constructed by applying the hierarchical clustering algorithm [16], shown in Figure 1. According to Remark 2, it can also be treated as the core structure of whole influenza viruses system, which helps us understand the evolutionary history and the mechanism of evolution [22].

Table 3: The virus classes on the second level structure.

Number	Virus number	First level	Virus signature
B1	1	A1	A/chicken/Cambodia/LC/2006(H5N1)
B2	1	A1	A/dog/Shandong/JT01/2009(H5N2)
B3	177	A1	A/chicken/Israel/184/2009(H9N2)
B4	665	A1	A/duck/Taiwan/DV1236/2009(H5N2)
B5	1	A1	A/blue-winged teal/LA/AI13-1225/2013(H7N7)
B6	2	A1	A/duck/Korea/A349/2009(H7N2)
B7	2	A1	A/chicken/Abbottabad/NARC-2419/2005(H7N3)
B8	1	A1	A/mallard/Netherlands/22/2010(H10N7)
B9	12	A2	A/swine/Hong Kong/2106/98(H9N2)
B10	35	A2	A/chicken/Italy/330/1997(H5N2)
B11	1	A2	A/chicken/Queensland/1995(H7N3)
B12	9	A2	A/chicken/Iran/261/01(H9N2)
B13	1	A2	A/oystercatcher/Peru/MM152/2008(H10N7)
B14	1	A3	A/American green-winged teal/Washington/195750/2014(H5N1)

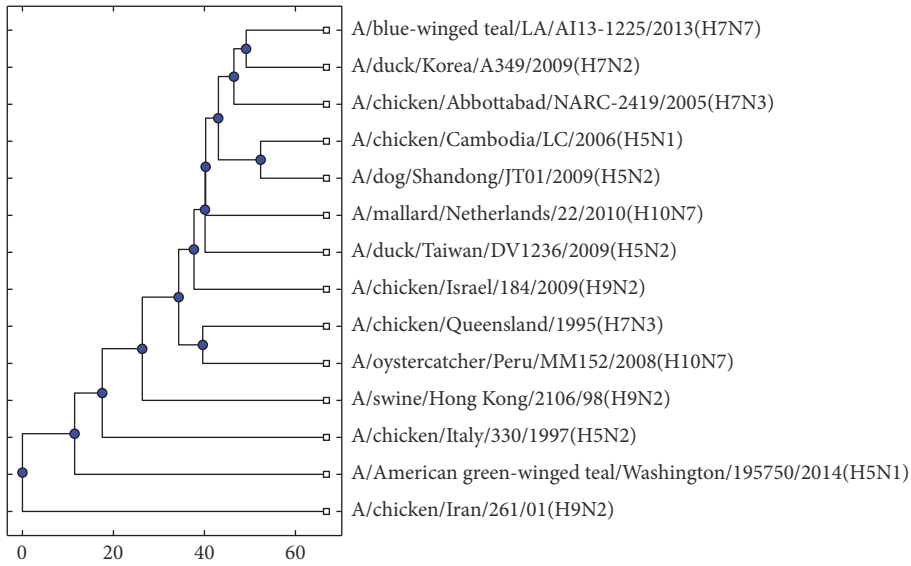

Figure 1: The phylogenetic tree of signature viruses on second level structure.

Among the 8 virus subtypes, 7 viruses are identified as the signature viruses except H7N9, for H7N9 viruses account for the minority of whole system (Figure 1). Exploring the intrinsic relation, it is obvious that H7N9 belongs to class B4, elucidating that the variation of H7N9 is not significant [23] and can be viewed as a new member in the family of viruses. Based on the coarse-grained idea, one signature virus represents the corresponding class. However, some isolated points are detected, such as A/chicken/Cambodia/LC/2006(H5N1), A/dog/Shandong/JT01/2009(H5N2), and A/chicken/Queensland/1995(H7N3), which might be caused by the big change to virus RNA strain.

From the hierarchical structure of the feature viruses, A/blue-winged teal/LA/AI13-1225/2013(H7N7), A/duck/Korea/A349/2009(H7N2), and A/chicken/Abbottabad/NARC-2419/2005(H7N3) have similar evolution relationship (connect closely) for they equip the same HA type (H7). Besides, A/chicken/Cambodia/LC/2006(H5N1) and A/dog/Shandong/JT01/2009(H5N2) have the consistent conclusions. However, A/chicken/Italy/330/1997(H5N2) is far from them, which could be due to the fact that the outbreak time plays an important role in sequence mutation. If just considering the HA and NA proteins, the subtypes, such as H9N2 [24], should be redefined. Comparing the two-level

structure and hierarchical structure of virus signature, the intrinsic relationship among A1 on the first level structure is consistent with that in the hierarchical structure, while class A2 has the dispersed structure where the feature viruses in different subtypes scatter in chaos, which indicates that constructing the second level structure is meaningful.

4. Conclusions

The rapid variation of influenza viruses results in low-accuracy subtyping identification and makes it difficult to develop effective drugs. This article explored the homology of avian influenza virus system and identified the subtypes according to HA and NA protein fragments, which might provide the support for developing antiviral drugs and vaccines according to different subtypes. Phylogenetic reconstructions serve understanding the evolution of influenza viruses. However, the large amounts of virus dataset pose an obstacle for analyzing the evolutionary relationship and identifying the correct subtypes to predict the biological functions. Granular computing theory was applied to determine the partition of virus system based on the constructed granular space. A method and the corresponding algorithm were proposed for detecting the rational granularity. With the proposed algorithm applied repeatedly, a multilevel structure of whole system was constructed. To reduce the computational complexity, some key viruses were selected to approximate the whole system based on the coarse-grained idea. According to the nearest-center principle, virus signatures were identified and constructed the granular signature set of a multilevel structure of complex system. By designing a classifier, the performance of virus signatures was evaluated and the result showed that the virus signatures could reflect the most properties of virus system. Furthermore, hierarchical structure of virus signature was constructed by using hierarchical clustering algorithm. Both of the two structures have some consistent intrinsic relationship among the virus systems and between the different subtypes. Some viruses were detected as isolated points in the structure thought equipped with the same labels, which might be caused by the rapid variations in the RNA strands. The virus signatures have the potential use in new virus subtyping comparison and functional prediction.

Disclosure

The work was previously presented in "The 10th International Conference on Systems Biology (ISB 2016, held in Weihai, China, August 19–22, 2016)."

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The work was supported by National Natural Science Foundation of China (Grant nos. 11371174, 11271163) and Colleges and Universities in Jiangsu Province Plans to Graduates Research and Innovation (Grant no. KYLX15_1188).

References

[1] T. Jombart, S. Devillard, and F. Balloux, "Discriminant analysis of principal components: a new method for the analysis of genetically structured populations," *BMC Genetics*, vol. 11, no. 1, article 94, 2010.

[2] F. G. Hayden, "Prevention and treatment of influenza in immunocompromised patients," *The American Journal of Medicine*, vol. 102, no. 3, pp. 55–60, 1997.

[3] N. J. Cox and K. Subbarao, "Global epidemiology of influenza: past and present," *Annual Review of Medicine*, vol. 51, pp. 407–421, 2000.

[4] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch, "Predicting the evolution of human influenza A," *Science*, vol. 286, no. 5446, pp. 1921–1925, 1999.

[5] J. Xu, C. T. Davis, M. C. Christman et al., "Evolutionary history and phylodynamics of influenza A and B neuraminidase (NA) genes inferred from large-scale sequence analyses," *PLoS ONE*, vol. 7, no. 7, Article ID e38665, 2012.

[6] R. G. Webster, W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka, "Evolution and ecology of influenza A viruses," *Microbiological Reviews*, vol. 56, no. 1, pp. 152–179, 1992.

[7] E. Ghedin, N. A. Sengamalay, M. Shumway et al., "Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution," *Nature*, vol. 437, no. 7062, pp. 1162–1166, 2005.

[8] J. B. Plotkin, J. Dushoff, and S. A. Levin, "Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 9, pp. 6263–6268, 2002.

[9] C. A. Russell, T. C. Jones, I. G. Barr et al., "The global circulation of seasonal influenza A (H3N2) viruses," *Science*, vol. 320, no. 5874, pp. 340–346, 2008.

[10] A. Lapedes and R. Farber, "The geometry of shape space: application to influenza," *Journal of Theoretical Biology*, vol. 212, no. 1, pp. 57–69, 2001.

[11] J. He and M. W. Deem, "Low-dimensional clustering detects incipient dominant influenza strain clusters," *Protein Engineering, Design and Selection*, vol. 23, no. 12, pp. 935–946, 2010.

[12] B. Zhang and L. Zhang, *Theory and Applications of Problem Solving*, Elsevier Science Inc, Amsterdam, Netherlands, 1992.

[13] Y. Y. Yao and J. T. Yao, "Granular computing as a basis for consistent classification problems," in *Proceedings of the Workshop on Foundations of Data Mining (PAKDD '02)*, pp. 101–102, 2002.

[14] B. Hartmann, O. Bänfer, O. Nelles, A. Sodja, L. Teslić, and I. Škrjanc, "Supervised hierarchical clustering in fuzzy model identification," *IEEE Transactions on Fuzzy Systems*, vol. 19, no. 6, pp. 1163–1176, 2011.

[15] X.-Q. Tang, P. Zhu, and J.-X. Cheng, "The structural clustering and analysis of metric based on granular space," *Pattern Recognition*, vol. 43, no. 11, pp. 3768–3786, 2010.

[16] X.-Q. Tang and P. Zhu, "Hierarchical clustering problems and analysis of fuzzy proximity relation on granular space," *IEEE Transactions on Fuzzy Systems*, vol. 21, no. 5, pp. 814–824, 2013.

[17] Y. Bao, P. Bolotov, D. Dernovoy et al., "The influenza virus resource at the National Center for Biotechnology Information," *Journal of Virology*, vol. 82, no. 2, pp. 596–601, 2008.

[18] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, Amsterdam, Netherlands, 2011.

[19] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *Journal of Molecular Biology*, vol. 238, no. 1, pp. 54–61, 1994.

[20] B. De Baets and E. Kerre, "Fuzzy relations and applications," *Advances in Electronics and Electron Physics*, vol. 89, pp. 255–324, 1994.

[21] M. R. Anderberg, *Cluster Analysis for Applications: Probability and Mathematical Statistics: A Series of Monographs and Textbooks*, Academic Press, Cambridge, Mass, USA, 2014.

[22] Z.-W. Chen and X.-Q. Li, "Whole-genome phylogeny based on protein domain information," *China Journal of Bioinformatics*, vol. 1, article 10, 2012.

[23] R. Gao, B. Cao, Y. Hu et al., "Human infection with a novel avian-origin influenza A (H7N9) virus," *The New England Journal of Medicine* , vol. 368, no. 20, pp. 1888–1897, 2013.

[24] M. Peiris, K. Y. Yuen, C. W. Leung et al., "Human infection with influenza H9N2," *The Lancet*, vol. 354, no. 9182, pp. 916–917, 1999.