REDUCED AMINO ACID ALPHABET-BASED ENCODING AND ITS IMPACT ON MODELING INFLUENZA ANTIGENIC EVOLUTION

Forghani M.[a],

Firstkov A. L.[a],

AlyanNezhadi M. M.[b],

Danilenko D. M.[c],

Komissarov A. B.[c]

[a] N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences (IMM UB RAS), Yekaterinburg, Russia

[b] Mazandaran University of Science and Technology, Iran.

[c] Smorodintsev Research Institute of Influenza, Saint Petersburg, Russia

КОДИРОВАНИЕ С ПОМОЩЬЮ СОКРАЩЁННОГО АМИНОКИСЛОТНОГО АЛФАВИТА И ЕГО ВЛИЯНИЕ НА МОДЕЛИРОВАНИЕ АНТИГЕННОЙ ЭВОЛЮЦИИ ГРИППА

Форгани М.[1],

Фирстков А. Л.[1],

Аляннеджади М. М.[2],

Даниленко Д. М.[3],

Комиссаров А. Б.[3]

[1] Федеральное государственное бюджетное учреждение науки Институт математики и механики им. Н. Н. Красовского Уральского отделения Российской академии наук (ИММ УрО РАН), Екатеринбург, Россия

[2]Университет науки и технологии Мазандарана, Иран.

[3]ФГБУ Научно-исследовательский институт гриппа имени А.А. Смородинцева, Минздрава России.

ABSTRACT. Currently, vaccination is one of the most efficient ways to control and prevent influenza infection. Vaccine production largely relies on the results of laboratory assays, including hemagglutination inhibition and microneutralization assays, which are time-consuming and laborious. Viruses can escape from the immune response that results in the need to revise and update vaccines biannually. The hemagglutination inhibition assay can measure how effectively antibodies against a reference strain bind and block an antigen of the test strain. Various computer-aided models have been developed to optimize candidate vaccine strain selection. A general problem in modeling of antigenic evolution is the representation of genetic sequences for input into the research model. Our motivation stems from the well-known problem of encoding genetic information for modeling antigenic evolution. This paper introduces a two-fold encoding approach based on reduced amino acid alphabet and amino acid index databases called AAindex. We propose to apply a simplified amino acid alphabet in modeling of antigenic evolution. A simplified alphabet, also called a sub-alphabet or reduced amino acid alphabet, implies to use the 20 amino acids being clustered and divided into amino acid groups. The proposed encoding allows to redefine mutations termed for amino acid groups located in reduced alphabets. We investigated 40 reduced amino acid sets and their performance in modeling antigenic evolution. The experimental results indicate that the proposed reduced amino acid alphabets can achieve the performance of the standard alphabet in its accuracy. Moreover, these alphabets provide deeper insight into various aspects of the relationship between mutation and antigenic variation. By checking identified high-impact sites in the Influenza Research Database, we found that not only antigenic sites have a significant influence on antigenicity, but also other amino acids located in close proximity. The results indicate that all selected non-antigenic sites are related to immune responses. According to the Influenza Research Database, these have been experimentally determined to be T-cell epitopes, B-cell epitopes, and MHC-binding epitopes of different classes. This highlighted a caveat: while simulating antigenic

evolution, the model should consider not only the genetic information on antigenic sites, but also that of neighboring positions, as they may indirectly impact antigenicity. Additionally, our findings indicate that structural and charge characteristics are the most beneficial in modeling antigenic evolution, which is in agreement with previous studies.

Keywords: AAindex, antigenic evolution, hemagglutinin, influenza, modeling, reduced amino acid alphabet

РЕЗЮМЕ. В настоящее время, вакцинация является одним из наиболее эффективных способов контроля и профилактики гриппозной инфекции. Производство вакцин в основном зависит от результатов лабораторных анализов, включая анализ реакции торможения гемагглютинации и микронейтрализации, которые требуют много времени и труда. Вирусы могут избегать иммунного ответа, что приводит к необходимости пересмотра и обновления вакцин два раза в год. Анализ реакции торможения гемагглютинации позволяет измерить, насколько эффективно антитела против эталонного штамма связывают и блокируют антиген испытуемого штамма. Для оптимизации выбора вакцинного штамма-кандидата были разработаны различные компьютерные модели. Одна из общих проблем в моделировании антигенной эволюции является представление генетических последовательностей для ввода в исследовательскую модель. Наша мотивация связана с хорошо известной проблемой кодирования генетической информации для моделирования антигенной эволюции. В данной работе представлен двухэтапный подход к кодированию, основанный на сокращенных аминокислотных алфавитах и базах данных аминокислотных индексов под названием AAindex. Мы предлагаем использовать упрощенные аминокислотные алфавиты для моделирования антигенной эволюции.

Упрощённый алфавит, также называемый субалфавитом или сокращённым аминокислотным алфавитом, это алфавит, в котором 20 аминокислот разделены на группы. Предложенное кодирование позволяет переопределить мутации в терминах групп аминокислот, расположенных в сокращенном алфавите. Мы исследовали 40 сокращённых алфавитов и их эффективность при моделировании антигенной эволюции. Результаты экспериментов показывают, что предложенные сокращенные аминокислотные алфавиты могут достичь показателей стандартного алфавита по точности. Более того, эти алфавиты позволяют лучше понять взаимосвязь между мутациями и антигенными изменениями с различных точек зрения. Проверив полученные высокоэффективные сайты в исследовательской базе данных гриппа (Influenza Research Database), мы обнаружили, что не только антигенные сайты оказывают значительное влияние на антигенность, но и их соседние аминокислоты. Результаты показывают, что все выбранные неантигенные участки связаны с иммунным ответом. Согласно исследовательской базы данных гриппа, экспериментально установлено, что это эпитопы Т-клеток, эпитопы В-клеток и MHC-связывающие эпитопы различных классов. Это подчёркивает значимость того, что: при моделировании антигенной эволюции модель должна учитывать не только генетическую информацию антигенных участков, но и генетическую информацию соседних позиций, поскольку они могут косвенно влиять на антигенность. Кроме того, наши результаты показывают, что, в соответствии с предыдущими исследованиями, структурные и зарядовые характеристики аминокислот являются наиболее значимыми при моделировании антигенной эволюции.

Ключевые слова: AAindex, антигенная эволюция, гемагглютинин, грипп, моделирование, сокращённый аминокислотный алфавит

INTRODUCTION. Influenza is a contagious respiratory infection that affects 5%-15% of the population worldwide annually, resulting in 3-5 million cases of severe illness and 250,000 to 500,000 deaths [36]. Influenza epidemics influence public health and involve severe economic consequences, making it the subject of various economic studies [4]. The World Health Organization (WHO) continuously monitors viral pathogens, especially those that can become epidemics or pandemics, and decides on strategies to combat them. Given the special status of influenza, the WHO created the Global Influenza Surveillance and Response System, the primary function of which is to monitor the evolution of the influenza virus and to provide recommendations for the annual vaccine's composition for the Northern and Southern Hemispheres.

Influenza viruses are part of the *Orthomyxoviridae* family. According to antigenic characteristics of their nuclear proteins, they are grouped into four types: IVA (A); IVB (B); IVC (C); and IVD (D). Among them, types A and B are associated with influenza outbreaks. Type C appears to evolve slowly and leads to less severe and less significant health consequences. Type D is an influenza C-like virus that is observed in non-human hosts, e.g., cattle and swine [30]. Type A is further classified according to the combination of hemagglutinin (HA) and neuraminidase (NA), the two main surface antigens of influenza that play a key role in infectivity and immune responses. HA has 18 variants (H1-H18), while the NA protein can be one of 11 variants (N1-N11). Hence, the virus can theoretically be any of 198 different subtypes; this provides an ability to infect a broad spectrum of various hosts [37]. Despite this diversity, humans are infected with only a limited number of influenza A subtypes (i.e., H1N1, H2N2, H3N2), with H1N1 and H3N2 being currently relevant. Thus, we consider them in this paper. Other zoonotic subtypes represent only sporadic infections and are out of the scope of this study.

Influenza A viruses are capable of enormous genetic variation, both through continuous, gradual mutation and by reassortment of gene segments between

viruses, resulting in emerging novel antigenic variants. Epidemics are the result of gradual evolutionary changes called antigenic drift, which leads to the generation of new strains from existing ones through mutation. In addition to antigenic drift, the influenza virus can be altered by antigenic shift. It is an abrupt significant change in influenza viruses resulting in the emergence of new HA and/or NA. It is the process by which at least two subtypes combine into a new subtype that has a mixture of surface antigens of two or more strains [35].

The only effective method to control influenza is vaccination, eliciting protective neutralizing antibodies and memory T-cell responses. Since HA antigen abundance on the viral surface is approximately four-fold greater than NA [31], it is the primary component in vaccine compositions. This is the reason why we consider only HA protein sequences in this paper.

The influenza vaccine requires an update if the vaccine composition strains are antigenically distinct from currently circulating viruses. A gold-standard and widespread laboratory procedure called hemagglutination inhibition (HI) assay is used to assess the measure of antigenic similarity between strains. The HI assay can measure how effectively antibodies against a reference strain bind and block an antigen of the test strain. High HI titers indicate a high degree of antigenic similarity between strains [16]. The main conclusion of HI assay analysis is determining antigenic distance (i.e., similarity between reference and test antigens), which further can be presented in terms of a binary variable called antigenic variant. Currently, there are two widely used definitions of antigenic distance [18, 27]:

$$d_1(i,j) = log_2(\frac{M}{H_{i,j}}) \qquad (1)$$

$$d_2(i,j) = \sqrt{\frac{H_{i,i} \times H_{j,j}}{H_{i,j} \times H_{j,i}}} \qquad (2)$$

where $H_{i,j}$ is the obtained HI titer for antiserum of (reference) strain *j* against the antigen of (test) strain *i*, and *M* is the maximum titer observed for antiserum *j* against any antigen in the HI table. The antigenic variant is determined by applying the threshold to the obtained antigenic distance. The pair of test and reference viruses whose antigenic distance meets the threshold are designated as antigenic variants; otherwise, they are only antigenically similar.

The HI assay is a labor-intensive and time-consuming procedure, while vaccine development is under time pressure. Over the past decade, various computer-aided approaches have been developed to speed up the process of strain selection and to increase the quality of vaccine production. Klingten *et al*. [16] has provided a comprehensive review of antigenic evolution prediction associated with vaccine production. They classified the approaches into phylogenetic and population genetics-based methods, statistical methods, epidemiological models, and other methods based on information and graph theories. The approaches employ different data types serving as model inputs, e.g., viral sequence, HI assay data, protein structure, physicochemical properties, etc. A critical step in antigenic variant modeling is describing the biological significance of a mutation between test and reference viruses and linking it to antigenicity.

Unfortunately, the exact roles and how they affect biological properties within evolution are not yet fully understood for many such changes. Generally, it is known that evolution is influenced by several biological properties, especially the volume and hydrophobicity of amino acids [32]. Studies on amino acid property changes provide fundamental information about the evolution of specific proteins. Earlier studies indicated that HA antigen is positively charged, while on the contrary, the glycan receptors of the host cell are negatively charged. Thus, changes in electrostatic charge due to mutation can play a significant role in receptor specificity, enhancing or diminishing the receptor binding affinity and avidity [2, 17]. Moreover, Huang *et al*. [14] recently showed that charged amino acid mutations impact

influenza virus evolution and are beneficial in vaccine research. Accordingly, mutation can be considered a multidimensional event, wherein each dimension represents an amino acid attribute.

Several techniques reflect the biological characteristics of mutation in numerical domains, among which application of the AAindex database [15] is the most popular. The AAindex database is a comprehensive collection of biological, physical, and chemical amino acid properties collected from scientific papers and accessed through *www.genome.jp*. The database mainly consists of three sections: AAindex1; AAindex2; and AAindex3. AAindex1 includes various amino acid indices, each of which can be represented as a numerical vector of 20 numbers representing 20 standard amino acids. AAindex2 contains different amino acid mutation matrices, while AAindex3 consists of statistical protein contact potentials. The AAindex database (ver.9.2) currently covers 566, 94, and 47 records for AAindex1, AAindex2, and AAindex3, respectively.

As mentioned, the AAindex database has been employed for encoding protein sequence in various studies. Here, we mention some of the more relevant studies in which the AAindex database was used for exploring genetic and antigenic evolution. Yao *et al*. [39] proposed an algorithm called joint random forest regression to predict antigenic variants. They compared 95 amino acid matrices, including AAindex2, to assess the relationship between genetic and antigenic evolution by amino acid attributes at different protein sites. Their results indicated that structural features are more significant to the antigenicity of the influenza virus. Wang *et al* [34] suggested an approach based on matrix completion for predicting antigenic evolution. They studied the impact of 65 amino acid substitution matrices taken from the AAindex database to predict antigenic evolution. Their results suggested that the “homologous structure derived matrix (called HSDM) for alignment of distantly related sequences” outperformed others in terms of RMSE.

Moreover, Qui *et al*. [24] developed a structure-based antigenicity scoring model. Their model engages antigenically dominant positions according to structural context, including correlation with local amino acid attribute changes, to analyze antigenicity. They demonstrated that incorporating the structural context of protein can enhance antigenic evolution prediction. Additionally, Forghani and Khachay [10] carried out a principal component analysis on AAindex1 and introduced 11 indices that explained 91% of the total variation in the database. The new indices are further used to encode HA protein sequence and create an input tensor fed into a convolutional neural network. Their model achieves a mean absolute error of 0.935 antigenic units for yearly, non-anticipating prediction of antigenic distance for subtype H1N1 (2001-2009). Cui *et al*. [6] suggested modeling influenza virus antigenicity by selecting the most significant sites, clustering the AAindex1 based on mutual information, and encoding the sites by the representative from clusters to form the feature vector. The feature vector is further given to a classifier to discretize antigenic variant classes. Recently, we performed a preliminary analysis to study the impact of amino acid encoding on modeling the antigenic evolution of the influenza virus [11]. Apart from Cui *et al*.'s work, our work introduces an early-stage mutation encoding by applying reduced amino acid alphabets.

The current paper addresses one of the fundamental challenges in bioinformatics: deciding how to represent input genetic information for modeling more efficiently and meaningfully. In response to this problem, we employed simplified amino acid alphabets. A simplified alphabet, also called a sub-alphabet or reduced amino acid alphabet (RAAA), is an alphabet in which the 20 amino acids are clustered and divided into amino acids groups. RAAA construction is a problem that belongs to the set partitioning problem, which is out of this paper's scope. Previous studies have shown that RAAAs have been successfully applied in various domains, including: protein annotation and description; protein functionality

prediction [21, 41]; protein folding assessment; sequence classification [19]; consensus sequence search; and genetic pattern identification [5].

A reduced amino acid set simplifies protein system complexity, providing a better insight into structural similarities across protein sequences [42]. We considered different definitions of similarity via RAAAs to reconstruct the relationship between genotype and phenotype. A RAAA represents genetic information on a coarse scale, which may highlight attributes that drive antigenic evolution of the influenza virus.

In our approach, encoding is conducted in two steps. In addition to the standard amino acid alphabet, the first step employs a RAAA to represent the mutation in different structural, biological, and physicochemical contexts. Further, the second step encodes the alphabetical information of the encoded genetic sequence into a numerical one, which enables its use in various types of mathematical modeling. Preliminary results indicate that some RAAA-based models outperform models based on the standard amino acid alphabet in terms of accuracy.

In this paper, we take a step forward and perform a comprehensive analysis to further refine result accuracy. The contributions of this paper are three-fold:

1. We propose a novel encoding method using reduced amino acid alphabets, which helps to clarify the genetic/antigenic relationship.
2. Relative to similar previous studies [6, 11], we improve the approach at several levels:
    2.1. Increasing the resolution of thresholds.
    2.2. Clustering by several methods and comparing their results to find the optimal number of clusters.
    2.3. Selecting the closest index to the center of a cluster as its representative.
    2.4. Applying five well-known classification algorithms.

2.5. Optimizing of classifier hyperparameters through a comprehensive grid search.

3. Relying on experimental results, we found that incorporating structural and charge properties can enhance modeling quality, which is in agreement with previous studies.

The rest of the paper is organized as follows. Section II describes the general computational pipeline, data preparation, and all necessary algorithms for primary and secondary encoding. Experimental setup and its outcomes are presented in Section III. This section also covers interpretation and discussion of the obtained results. Finally, the conclusion is given in Section IV.

II. METHODOLOGY

As mentioned earlier, our experimental design was inspired by a published methodology [6]. However, we propose some modifications and enhancements to improve modeling quality. Our approach is mainly divided into five steps: encoding genetic sequences by RAAA; selecting the most relevant sites; clustering the AAindex1 data set based on selected sites; encoding the selected sites by a representative from each cluster; and modeling antigenic variants by a classifier. The general schema of our pipeline is shown in Figure 1.

2.1. Data Preparation

Our approach relies on three database types, each of which requires specific preparation in order to be used in the computational pipeline.

2.1.1. Simplified Amino Acid Alphabets

Apart from the standard amino acid alphabet with 20 letters, there are various RAAAs, in which the number of letters is less than 20. Typically, an RAAA is obtained by grouping the 20 amino acids. There are several strategies to perform

this, some of which have been described [29]. For example, the set of 20 amino acids can be divided into three groups based on Van Der Waals volume by setting three ranges (0-2.78, 2.95-4.00, 4.43-8.08), resulting in three partitions: GASCTPD; NVEQIL; and MHKFRYW. This permits new interpretation of the mutation from a different point of view, such as change in hydrophobicity. In total, 40 published RAAAs were collected [8, 29, 38] and are presented in Table 1.

2.1.2. HI Assay Database

Typically, an HI assay database record includes three fields of information: test virus identifier; reference antiserum identifier; and HI titer. Sometimes additional metadata, such as experiment date, may be appended. HI assay results can be presented in four forms: raw HI titer; standardized HI titer; antigenic distance; and antigenic variant. At this point, we only used the antigenic variant obtained via the antigenic distance threshold. We employed Eq. (1) with threshold 4 for calculating the antigenic variant. Duplicated entries were averaged in terms of titer. Therefore, each test/reference virus combination is unique within the database. Here, we considered two subtypes in the influenza vaccine, H1N1 and H3N2. The HI assay database was taken from references [13, 34]. The final obtained database had 7,449 H3N2 and 3,747 H1N1 entries. There were 506 viruses for the H1N1 subtype (506 test against 44 references) and 772 for H3N2 (666 test against 130 references).

2.1.3. AAindex1 Database

The latest version of AAindex1, ver. 9.2, consists of 566 entries. A typical database entry includes a vector of 20 numbers, each of which is assigned to a standard amino acid. Since the range of numbers in vectors varies within the database, we individually scaled each vector into the unit interval $[0,1]$. After removing vectors with missing values, 553 remaining entries were used for analysis.

2.2. Encoding of HA Sequences

Here, we use RAAAs to take into account the impact of the mutation on antigenic evolution from different physicochemical (amino acid) perspectives. The

first step of encoding the sequence by RAAA is selecting an arbitrary amino acid from each group in the alphabet as a group representative. Further, we replace all members of the group with its representative in the protein sequence. This step does not influence data if the standard amino acid alphabet is chosen since this alphabet has 20 groups, not less.

2.3. Selection of High Impact Sites

The model's input is a feature vector produced from encoded relevant sites in the genetic sequence. The model utilizes these sites for reconstructing the relationship between genetic and antigenic evolution in the feature space. Therefore, it is necessary to measure the relevance of site mutations according to the antigenic variation. Cui *et al*. [6] proposed measurement by introducing the below score for the site's antigenic significance:

$$S_i = |\Phi_i| \times E_i \qquad (3)$$

where i is the index of the site in the sequence, S_i is the significance score, and E_i is Shannon's entropy of site i in the whole database as computed by the following formula:

$$E_i = -\sum_{j=1}^{20} P_{i,j} log P_{i,j} \qquad (4)$$

where $P_{i,j}$ is the probability of amino acid j occurrence at position i. Φ_i is a coefficient expressed with the following formula:

$$\Phi_i = \frac{(N_{11} \times N_{00} - N_{10} \times N_{01})}{\sqrt{N_{X1} \times N_{X0} \times N_{1Y} \times N_{0Y}}} \qquad (5)$$

where N_{mn} ($m, n \in \{0,1\}$) is the number of HI entries with $X=m$ and $Y=n$. The variable X represents the occurrence of mutation at site i (0 or 1 for conserved or mutated cases, respectively). The variable Y expresses the antigenic relationship between the test-reference pair of viruses in HI entries. If the test and reference are antigenically

similar, the Y variable value is zero. Otherwise, they are variants, and it takes the value of one. $N_{X,n}$ denotes the number of entries with $Y=n$, whereas X can take any value from {0,1}. Similarly, $N_{m,Y}$ represents the number of entries with $X=m$, while Y has a value from {0,1}. Note that all variables in Eq. (5) are calculated only for site i. In the case of a conserved site, the significance score is set to zero.

The application of Eq. (3) can be extended to sequences encoded by RAAAs. Encoding genetic sequences by such an alphabet notably changes the entropy and Φ values and, accordingly, the significance score. The significance score for all sites obtained by applying a RAAA is further scaled into the unit interval *[0,1]*. This allows us to compare the significance of a specific site considering different alphabets. The final high-impact sites are determined by setting a threshold on the results of the scaled significance score. The threshold value is selected from the set {0.2,0.3,0.4,0.5,0.6,0.7,0.8}. It's worth noting that a site is selected if its scaled significance score is higher than the target threshold. Obviously, decreasing the threshold leads to an increase in the number of selected (high-impact) sites.

2.4. Clustering the AAindex1 Database

The AAindex1 database is used to perform the second stage of encoding. We select some entries from AAindex1 (called representatives) that are further used to encode the genetic information of obtained selected sites in the previous step. It is known that there is a high correlation between AAindex1 entries. Therefore, we cluster them and choose a representative from each cluster to diminish the number and correlation of final features. Clustering should be performed so that the objects of a cluster have almost the same encoding impact on antigenicity modeling.

2.4.1. Computing Mutual Information

To cluster the AAindex1 database, we create a feature vector for each entry by a similar scenario as described [6] with a modification for RAAAs. In the suggested method, the feature vector characterizes the AAindex1 entry by mutual

information (MI). The MI value expresses not only the significance of genetic information but also the impact of encoding for a selected site individually. Note that the size of the feature vector for clustering AAindex1 is the same as the size of selected sites. Indeed, each element of the feature vector is the measure of mutual dependency between the changes in a selected site, encoded by an AAindex1 entry, and antigenic variants within the HI database.

The number of amino acids in the RAAA is less than in the standard alphabet. Thus, a question arises on how encoding is carried out using AAindex1 regarding a RAAA. In order to solve this issue, we define a new database, called pseudo-AAindex1, derived from the original AAindex1 database. The procedure of generating pseudo-AAindex1 is described in Figure 2.

As previously stated, each amino acid group has a representative, which replaces all amino acids of the group in protein sequences. In order to assign a value to the representative, we compute the average of AAindex1 values for the amino acids within the group. This allows each amino acid to participate and have its own effect through the representative. Thus, a pseudo-AAindex1 is created for each RAAA, making it possible to calculate the mutual information in RAAA encoding. For simplicity, we hereafter refer to both the original AAindex1 and the pseudo-AAindex1 simply as AAindex1.

2.4.2. Determining the of optimal number of cluster

When considering an alphabet, we create a feature vector for each AAindex1 entry, the size of which depends on the number of determined high-impact sites. Before clustering AAindex1, it is necessary to determine the optimal number of clusters. Indeed, this number affects the final feature vector, which is used for antigenic variant modeling. For this purpose, we conduct a comprehensive search for the optimal number by employing three algorithms: K-means; agglomerative clustering with different linkage criteria; and spectral clustering.

First, we determine the number of unique feature vectors. Clustering is not required if the number is less than a threshold (e.g., five). When the number of unique vectors is more than the threshold, we cluster the set of vectors, while the number of clusters starts from two and increases up to ten.

Generally, six clustering variants are applied, including: K-means; agglomerative clustering with four different criteria (ward, average, complete/maximum, and single/minimum); and spectral clustering. The obtained clustering from each algorithm is individually evaluated by four scores, including Silhouette, Calinski-Harabasz, Davies-Bouldin, and the sum of squared distances of objects to their closest cluster. Further, the results are plotted and manually checked to decide the optimal number of clusters for AAindex1 associated with an alphabet.

2.4.3. Clustering

Generally speaking, the aim of clustering is to decrease correlation between AAindex1 entries. This also leads to diminishing the number of features, which are used in the final classification. To cluster the AAindex1 database, we apply the associated clustering algorithm by which the optimal number of clusters was determined from the previous subsection. Next, we select a representative from each cluster. The representative of a cluster is the closest object to its center. The representative is further employed to encode the information of high-impact sites for the classification.

2.5. Classification of Antigenic Variants

We use the obtained cluster representatives to apply the secondary encoding. This is carried out by replacing an amino acid group representative in the selected sites with its numerical value from the cluster's representative. Then, we individually calculate the differences between the test and the reference strains for each HI assay database entry by subtracting their encoded selected sites (or feature vectors). If we denote the number of high-impact sites and number of clusters

representatives with N and M, respectively, then the final feature vector has the size of $N \times M$.

Before performing the final classification, the last step is to determine the best classifier. To decrease the effect of the classifier on the results, we consider five different classifiers, including random forest, multilayer perceptron, logistic regression, support vector, and Gaussian naïve Bayes. Each classifier has its own parameters optimized through grid search (parameter list in Table II).

Grid search is carried out by cross-validation with different parameter combinations. Note that the Gaussian naïve Bayes classifier has no parameters for grid search, but it assumes that features are independent. Thus, we perform principal component analysis on the feature matrix to decrease the dependency. A threshold on the percentage of variance explained by the selected components was set as a parameter for Gaussian naïve Bayes.

By comparing grid search results, we were able to choose the best classifier with high performance in terms of accuracy. Note that the selection of optimal classifier depends on the results of three procedures:

- Encoding by the alphabet (primary encoding)
- Selection of high-impact sites
- Clustering the AAindex1 database and choosing representatives for secondary encoding

Among these procedures, the first has the most decisive influence on classification results. In fact, it changes the amino acid space globally, resulting in different representations of genetic variation, as well as different relationships between genotype and phenotype.

III. Results & Discussion

Considering all parameters, we ran 224,147 fits (41 alphabets × 7 thresholds × 781 5-fold cross-validations) for each subtype (H3N2 and H1N1) in the

experimental data to obtain the best classifiers. Knowing the best classifier for each triple-combination case (subtype, alphabet, threshold), we performed a 10-fold cross-validation by applying its best classifier. A comprehensive report of the results, including the evaluation criteria, is publicly available at: *github.com/viroinformatics/Simplified_Alphabets*.

The maximum accuracy achieved by each threshold is presented in Table III. Since the length of the feature vector is decreased by increasing the threshold, this also leads to accuracy reduction. From Table III, it is observed that threshold 0.4 seems to be a good choice for modeling the antigenic variants. Compared with previous studies [10, 11], our results indicate a high degree of accuracy, especially for H3N2, which suggests potential application in the field of public health.

As expected, some RAAAs achieved the same accuracy as the standard amino acid alphabet. Table IV presents the alphabets with the highest performance for different thresholds and subtypes. In the case of subtype H1N1 with thresholds 0.3 and 0.5, there are alphabets, the accuracy of which are slightly less (about 0.01) than the standard and Prlic-SDM12-2000 alphabets, but are not added to the table. Since prediction accuracy significantly drops from threshold 0.5 to threshold 0.8 Table III), we did not consider their results in Table IV. Interestingly, the Risler-88 and Li-2003 alphabets are observed in the list of each subtype.

Moreover, the Cannata-2002 alphabet seems to be more informative for subtype H3N2 rather than for H1N1. In some cases, e.g., subtype H1N1 with threshold 0.4 and subtype H3N2 with threshold 0.3, the feature vector obtained from RAAAs is shorter in length than that obtained from the standard alphabet, while their accuracy is the same. This indicates that the amino acid space represented by the standard alphabet has redundant dimensions to express genetic variation of antigenic variants. Next, we briefly discuss each of the alphabets from Table IV.

Stephenson & Freeland analyzed 34 different RAAAs [29] and classified them into five classes based on how grouping was carried out. The classes are chemistry, sequence alignment, structural alignment, contact potential, and protein

blocks. Of the alphabets in Table IV, four are based on sequence alignment methods, whereas two rely on structural alignment. Only one alphabet (Zou-2009) was created by protein blocks. The complete classification of RAAAs presented in Table I is based on published work [29].

Similarity between amino acids can be defined from various viewpoints, e.g., hydrophobic residues (I, V) and aromatic residues (F, W, Y). The main idea of constructing a RAAA is to use amino acid properties to define similarity, with placing of similar amino acids in a group. For example, the RAAA presented by Li *et al*. [20] was obtained from amino acids substitutions by scoring similarities that may be beneficial in recognition of protein folds. Their results imply that at least ten amino acid types are required to characterize protein complexity.

Cannata *et al*. [5] presented a method to produce RAAAs by scoring different amino acid compositions using a branch and bound algorithm and substitution matrix. Their alphabet belongs to the 'alignment-based methods' class of sequences. Furthermore, Lenckowski *et al*. [19] suggested an alphabet generated using a genetic algorithm and strategy based on global sequence alignment. Their results indicate that the proposed alphabet outperformed the standard amino acid set and other RAAAs in the sequence classification task. Andersen and Brunak's RAAA [1] includes 13 letters; it is also constructed based on sequence alignment. In contrast, RAAAs proposed by Prlic *et al*. [23] and Risler *et al*. [25] are both derived by substitution frequency of structural alignments.

Zou *et al*. [42] applied reduced amino acid alphabets to predict defensin family and subfamily. They clustered amino acids by the protein blocks (PBs) method [7, 9], in which the distribution of amino acids in PBs was used to generate clusters of equivalent amino acids with respect to local structure. Indeed, this kind of alphabet can be considered a structural alphabet. Their results indicate that use of such alphabets can improve prediction accuracy with defensin family and subfamily. Surprisingly, no alphabet based on attributes of individual amino acids attained a

high level of performance. Taken together, the high-performing RAAAs emphasize the role of structural features in antigenic evolution modeling.

By checking the high-impact sites in the Influenza Research Database (IRD) [40], we found that not only antigenic sites have a significant influence on antigenicity, but also other amino acids located in close proximity. The results indicate that all selected non-antigenic sites are related to immune responses. According to IRD, these have been experimentally determined to be T-cell epitopes, B-cell epitopes, and MHC-binding epitopes of different classes. This highlighted a caveat: In modeling of antigenic evolution, the model should consider not only the genetic information of antigenic sites, but also that of neighboring positions, as they may indirectly impact antigenicity. Note that feature vector construction relies on high-impact sites, but the evolutionary history showed that even one amino acid substitution can change the antigenic cluster of the influenza virus [28]. Such a substitution may present a low impact through the mutual information score. We believe a desirable model must take into account the effects of both high and low impact sites. The visualizations of selected high-impact sites for H1N1 (threshold 0.3), and H3N2 (threshold 0.4), are presented in Figure 3. These are cases with high accuracy and shorter feature vector length.

Various AAindex1 entries were designated as representatives during all experiments with optimized classifiers. The top ten entries and their frequencies are listed in Table V. The complete list of AAindex1 entries and their frequencies is available (*github.com/viroinformatics/Simplified_Alphabets*).

It can be seen that the majority of AAindex1 attributes used in model construction are associated with charge properties. This emphasizes that antigenicity notably depends on protein conformation, which cannot be fully reflected in a one-dimensional representation of protein as a sequence. However, the model can capture some attributes information by encoding the genetic sequence using physicochemical properties presented in the AAindex1 database.

Table V also indicates that nine out of the ten most frequently AAindex1 entries are common in both subtypes. The last AAindex1 entry in the list of each subtype is different. To better understand the characteristics of entries in Table V, we computed the Pearson correlation coefficient and visualized it in Figure 4. It is observed that the majority of entries in Table V are not correlated, with two exceptions: FAUJ880111/KLEP840101 and FAUJ880105/CHAM830103.

Although the uncommon entries between H1N1 and H3N2 are different, it can be seen that they are correlated. In addition to the correlation matrix, we used principal component analysis (PCA) to identify the main components of 11 distinct AAindex1 entries in Table V and their expression in terms of explained variance. Figure 5 indicates that the first six components describe more than 90% of the explained variance. Seven and nine components represent 95% and 99% of the explained variance, respectively.

We also considered the performance of classifiers for antigenic variant modeling. Among five classifiers, random forest and multilayer perceptron outperformed others, in terms of accuracy, for both the H1N1 and H3N2 subtypes. The Gaussian naive Bayes classifier gave the worst results, so it may not be suitable for this kind of modeling.

In summary, the outstanding ability of our approach is based on redefining the mutation by RAAA and amino acid attributes used for encoding through a two-fold procedure. The primary encoding plays the main role with high priority, whereas secondary encoding has a supplementary role. From one point of view, the primary encoding determines the high-impact sites, while the secondary encoding gives the numerical interpretation to the genetic information of selected sites. From another point of view, the primary encoding interprets the mutation, and the secondary encoding reconstructs the specific relationship between genetic and antigenic differences (for the test and reference strains).

The proposed two-fold encoding approach revealed some aspects of mutations related to the antigenicity. Our findings indicate that encoding associated

with structural or charge properties of the protein dramatically impacts the performance of the antigenic model. This is in agreement with recent studies done by other researchers [14, 39]. In addition, RAAA encoding can lead to a smaller feature space dimension, while performance is maintained or improved. So far, this approach was applied only to seasonal human influenza strains. However, there are no theoretical limitations that would prevent further testing as a universal computational model for predicting antigenicity in other influenza subtypes, such as zoonotic H5, H7, H9, or other relevant influenza A subtypes that cause sporadic human infection.

III. Conclusion

Determining the degree of antigenic similarity between influenza virus strains is crucial in choosing candidate vaccine strains and subsequent timely vaccine production. Currently, the degree is measured via HI assay, a widespread laboratory procedure. Although HI assay is the gold standard method, it suffers from several shortcomings. Therefore, it has been suggested to employ computer-aided models as auxiliary tools to assess preliminary information about viral antigenicity prior to HI assay.

A notable problem in modeling antigenic evolution is the representation of genetic information to better express the relationship between genetic and antigenic variations. This paper proposes a two-fold encoding approach to genetic information using both a reduced amino acid alphabet (RAAA) and an amino acid index database. By applying a RAAA, we redefine the mutation as changes between amino acid groups of the alphabet, while the output sequence of the primary encoding is still alphabetical. The secondary encoding uses representatives from the AAindex1 database to convert the alphabetical sequence of the primary encoding into the numerical. The experimental results indicate that models built using RAAA encoding are able to achieve the same accuracy as models using the standard amino acid alphabet. The RAAA-based approach, however, features reduced computational complexity and associated cost.

Moreover, the suggested encoding can reveal the amino acid attributes which drive antigenic evolution. In agreement with previous studies, we find that structural and charge characteristics are the most beneficial in modeling antigenic evolution. Although the results obtained by our approach are desirable and promising, they are achieved by taking into account only high-impact sites. It is known that even one substitution can change the antigenic cluster, so we believe that further incorporating the role of low-impact sites into the model may enhance its accuracy and prediction potential; this will be addressed in future studies. Additionally, the model can be improved by: introducing new reduced amino acid alphabets; employing more significant and descriptive criteria for selecting key sites; and incorporating neighboring amino acid effects into the model.

Computational approaches for predicting antigenic properties from genetic sequence are also quite relevant for highly virulent influenza viruses. Laboratory testing of these pathogens requires high biosafety certification levels, and such analysis is not only time-consuming and labor-intensive, but also costly. Unlike current laboratory approaches, computational prediction of antigenic properties from viral sequence has the potential to enable rapid, large-scale antigenic characterization of influenza viruses. It is worth mentioning that application of our approach is not limited to modeling of antigenic evolution. It can be used in modeling any phenotype that is based on protein sequence, such as interactions with monoclonal antibodies.

IV. Funding

This study was funded by the Russian Foundation for Basic Research (RFBR), project number 19-31-60025.

V. Acknowledgments

514 Our work was performed using the 'Uran' supercomputer (IMM UB RAS).
515 The authors would like to thank Edward S. Ramsay for his valuable improvements
516 on the manuscript.

10.15789/2220-7619-RAA-1968

TABLES

Table I. The list of alternative and standard amino acid alphabets employed to encode the protein sequences in our experiments. The alphabets are borrowed from [8, 29, 38]. The color of RAAA determines the method of its obtaining. The classification of alphabets is borrowed from [29].

Таблица I. Список исследованных альтернативных и стандартных аминокислотных алфавитов, использованных для кодирования белковых последовательностей. Алфавиты заимствованы из [8, 29, 38]. Цвет RAAA определяет метод его получения. Классификация алфавитов заимствована из [29].

Name of alphabet Название алфавита	Groups группы
Standard Стандартный	A, C, D, E, Q, F, Y, G, H, I, V, K, R, L, M, N, P, S, T, W
Hydrophobic Гидрофобный	RKEDQN, GASTPHY, CVLIMFW
V Объем Ван-дер-Ваальса	GASCTPD, NVEQIL, MHKFRYW
Polarity Полярность	LIFWCMVY, PATGS, HQRKNED
Polarizability Поляризуемость	GASDT, CPNVEQIL, KMHFRYW
Mahler 1966	DE, KRH, QN, ST, P, CM, WYF, GALIV
Lehninger 1970	DE, KRH, NQSTGCY, PAWFMLIV
Dickerson 1983	DENKRQH, STGPACWY, FMLIV
Taylor 1986	DE, N, KRH, Q, T, SGAC, P, YWF, M, LIV
Weathers 2004	DENRH, KQST, GPACM, WYFLIV
SE-B(14)	A, C, D, EQ, FY, G, H, IV, KR, LM, N, P, ST, W
SE-B(8)	AST, C, DHN, EKQR, FWY, G, ILMV, P
Risler 1988	D, E, N, KRQ, S, T, G, P, H, A, C W, YF, ML, IV
Riddle 1997	DE, NKRQS, THA, GP, CWYFMLIV
Mirny 1999	DE, KR, NQST, GP, HWYF, ACMLIV
Prlic SDM12 2000	D, N, EKR, QST, G, P, H, A, C, W, YF, MLIV
Prlic SDM17 2000	D, EK, N, R, Q, S, T, G, P, H, A, C, W, Y, F, M, LIV
Melo 2005	DENKRQSTP, GA, H, C, WYFMLIV
Robson 1976	DKR, EA, GP, STNQ, H, C, WY, FMLIV
Solis(G) 2000	D, N, S, T, G, P, H, C, Y, EKRQAWFMLIV
Solis(D) 2000	DNS, EKRQ, TH, GP, AM, C, W, F, YL, IV
Rogov 2001	DNSTA, EKRQ, G, P, H, C, W, M, YFLIV
Etchebest 2007	DN, EKRQ, SH, TC, G, P, WYF, AML, IV
Solis GBMR4 2009	DENKRQSTA, G, P, HCWYFMLIV
Zuo 2009	DN, E, KRQ, SH, T, G, P, A, C, WYF, M, L, IV
Dayhoff 1978	DENQ, KRH, STGPA, C, WYF, MLIV
Murphy 2000	DENQ, KR, ST, G, P, H, A, C, WYF, MLIV
Cannata 2002	D, E, N, KR, Q, ST, G, P, H, A, C, W, Y, F, ML, IV
Fan 2003	DEQ, KR, STA, G, P, NH, C, WYF, ML, IV
Li 2003	DE, KRQ, ST, G, P, NH, AC, WYF, ML, IV
Edgar Se-B 2003	DN, EQ, KR, STA, G, P, HW, C, YF, MLIV
Edgar Se-V 2003	DEN, KRQ, STA, G, P, H, C, W, YF, MLIV

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

Kosiol 2003	DENKRQSTGPHA, C, W, YF, MLIV
Anderson 2004	D, E, KRQ, NS, T, G, P, H, A, C, WYF, ML, IV
Lenckowski 2007	DSHFM, ERQL, KPAC, NTWY, GIV
Crippen 1990	ENRSGHY, DKQTPW, AV, CFMLI
Maiorov 1992	DENQ, KR, G, P, AV, STHWY, CFMLI
Thomas 1996	DE, KR, QSTNGPH, C, AWYFMLIV
Wang 1999	DE, NKRQS, GP, THA, CWYFMLIV
Ceiplak 2001	DENRQSTG, K, HA, CWYMV, FLI
Liu 2002	DE, KR, NQSTGPHY, ACW, FMLIV

Amino acid physicochemical attributes

Физико-химические свойства аминокислот

Substitution frequency -- Structural alignment

Частота замещения -- структурное выравнивание

Spatial frequency -- Protein blocks

Пространственная частота -- Белковые блоки

Substitution frequency -- Sequence alignment

Частота замены -- Выравнивание последовательности

Spatial frequency -- Contact potential

Пространственная частота -- Контактный потенциал

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

Table II. Parameters used in the grid search. Parameter names are based on the machine learning package Scikit-learn [22].

Таблица 2. Параметры, используемые при поиске по сетке. Имена параметров основаны на пакете машинного обучения Scikit-learn [22].

Method **Метод**	**Parameters** **Параметры**	**Total cases in gird search** **Общее число случаев при поиске по сетке**
Random forest алгоритм случайного леса	Criterion, n_estimator, min_samples_split	Fitting 5-fold cross- validation for each of 40 candidates, totaling 200 fits фитинг 5-кратной перекрестной проверки для каждого из 40 кандидатов, всего 200 фитингов
Logistic regression Логистическая регрессия	Solver, penalty, max_iter	Fitting 5 folds for each of 66 candidates, total 330 fits фитинг 5 кратностей для каждого из 66 кандидатов, всего 330 фитингов
Multilayer perceptron Многослойный персептрон	Solver, learning_rate, activation, max_iter, learning_rate_init, hidden_layer_sizes	Fitting 5 folds for each of 648 candidates, total 3240 fits Подгонка 5 кратностей для каждого из 648 кандидатов, всего 3240 фитингов
SVM	Kernel, gamma, C, degree	Fitting 5 folds for each of 24 candidates, total 120fits фитинг 5 кратностей для каждого из 24 кандидатов, всего 120 фитингов
Gaussian naïve bayes Гауссовский наивный байесовский классификатор	PCA_threshold	Fitting 5 fold for each of 3 candidates, total 15 fits фитинг 5 кратностей для каждого из 3 кандидатов, всего 15 фитингов

Table III. The maximum accuracy was obtained by 10-fold cross-validation using different thresholds for scaled significance score. Note that increasing the threshold may lead to shorter feature vector length and consequent reduced accuracy.

Таблица 3. Максимальная точность была получена путем 10-кратной перекрестной проверки с использованием различных пороговых значений для масштабируемой оценки значимости. Увеличение порога может привести к уменьшению длины вектора признаков и, как следствие, к снижению точности.

Threshold Порог Subtype Подтип	0.2	0.3	0.4	0.5	0.6	0.7	0.8
H1N1	0.88	0.88	0.87	0.84	0.77	0.77	0.77
H3N2	0.92	0.92	0.92	0.9	0.88	0.85	0.83

Table IV. High performance alphabets in our experiments by virus type and site selection threshold. The amino acid groups for each alphabet are presented in Table I.

Таблица 4. Высокоэффективные алфавиты в экспериментах по типу вируса и порогу выбора сайта. Аминокислотные группы для каждого алфавита представлены в таблице I.

Threshold **Порог**	**H1N1**	**H3N2**
0.2	Standard, Risler-88, Li-2003, Anderson-2004 Стандартный, Risler-88, Li-2003, Anderson-2004	Standard, Risler-88, Cannata-2002, Zuo-2009 Стандартный, Risler-88, Cannata-2002, Zuo-2009
0.3	Standard* Стандартный *	Standard, Cannata-2002 Стандартный, Cannata-2002
0.4	Standard, Lenckowski-2007 Стандартный, Lenckowski-2007	Standard, Cannata-2002 Стандартный, Cannata-2002
0.5	Prlic-SDM12-2000*	Risler-88, Li-2003, Zuo-2009

* There are other alphabets, not listed, whose accuracy was slightly less than the alphabets shown in the table.

* Существуют и другие алфавиты, не указанные в списке, точность которых была несколько меньше, чем у алфавитов, приведенных в таблице.

Table V. List of the top ten most frequent AAindex1 entries in the experiments with optimized classifiers.

Таблица V. Список первых десяти наиболее частых записей AAindex1 в экспериментах с оптимизированными классификаторами.

ID Идентификатор	Description Описание	Freq. Частота
H1N1		
ANDN920101	alpha-CH chemical shifts химические сдвиги альфа-CH (Andersen *et al*., 1992)	147
CHAM830104	The number of atoms in the side chain labelled 2+1 Количество атомов в боковой цепи, помеченной 2+1 (Charton-	106
KLEP840101	Net charge Результирующий заряд (Klein *et*	97
CHAM830103	The number of atoms in the side chain labelled 1+1 Количество атомов в боковой цепи, помеченной 1+1 (Charton-Charton, 1983)	92
FAUJ880111	Positive charge Положительный заряд (Fauchere *et al*., 1988)	83
CHAM830107	A parameter of charge transfer capability Параметр способности переноса заряда (Charton-Charton, 1983)	83
VENT840101	Bitterness Горечь (Venanzi, 1984)	74
FAUJ880112	Negative charge Отрицательный заряд (Fauchere *et al*., 1988)	70
FAUJ880105	STERIMOL minimum width of the side chain STERIMOL минимальная ширина боковой цепи (Fauchere *et al*., 1988)	47
CHAM830105	The number of atoms in the side chain labelled 3+1 Количество атомов в боковой цепи, помеченной как 3+1 (Charton-Charton, 1983)	38
H3N2		
VENT840101	Bitterness Горечь (Venanzi, 1984)	119
CHAM830103	The number of atoms in the side chain labelled 1+1 Количество атомов в боковой цепи, помеченной 1+1 (Charton-Charton, 1983)	117
FAUJ880111	Positive charge Положительный заряд (Fauchere *et al*., 1988)	101
ANDN920101	alpha-CH chemical shifts химические сдвиги альфа-CH (Andersen *et al*., 1992)	101
KLEP840101	Net charge Результирующий заряд (Klein *et al*., 1984)	88
FAUJ880112	Negative charge Отрицательный заряд (Fauchere *et al*., 1988)	87
CHAM830107	A parameter of charge transfer capability Параметр способности переноса заряда (Charton-Charton, 1983)	66
CHAM830104	The number of atoms in the side chain labelled 2+1 Количество атомов в боковой цепи, помеченной 2+1 (Charton-Charton, 1983)	60

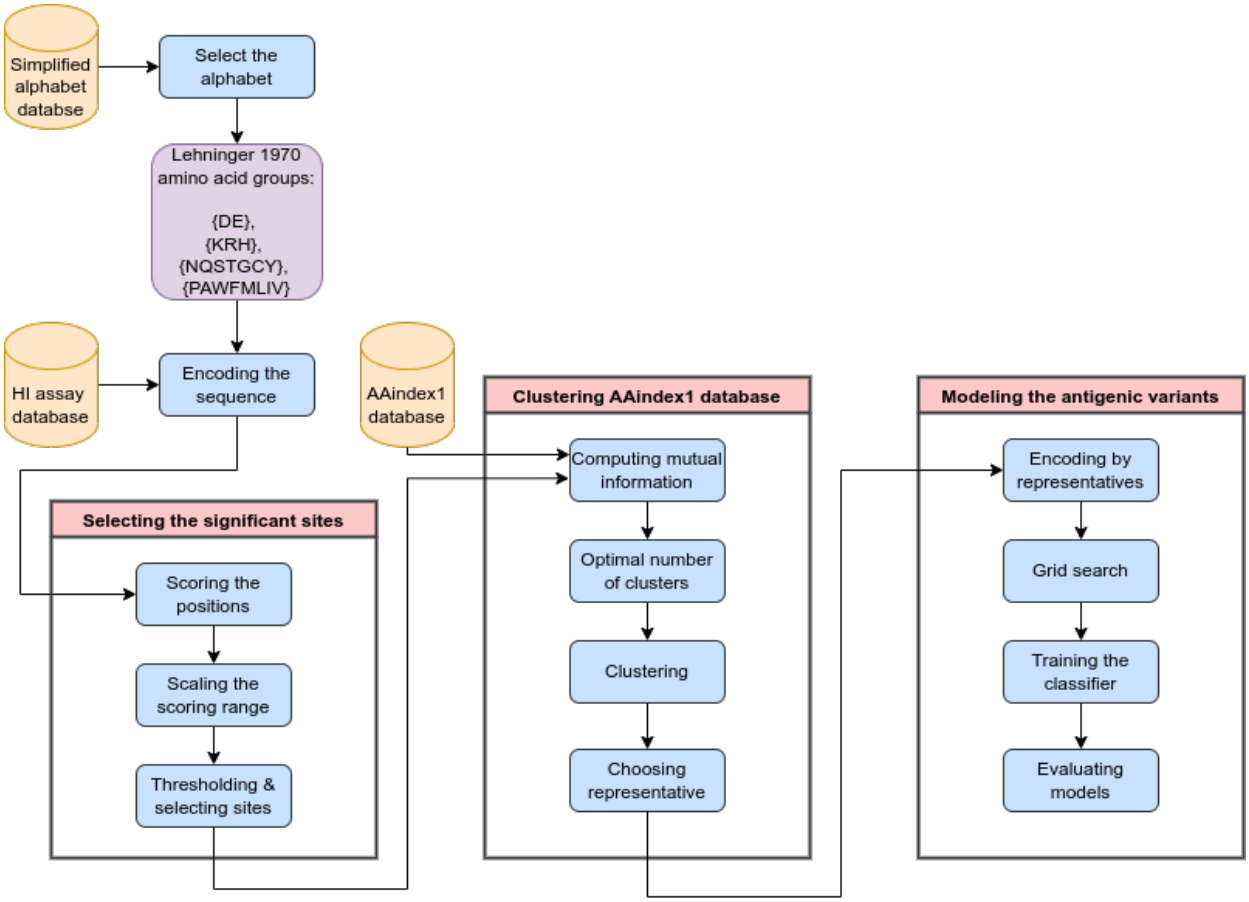
FAUJ880105	STERIMOL minimum width of the side chain STERIMOL минимальная ширина боковой цепи (Fauchere *et al*., 1988)	59
FAUJ880109	Number of hydrogen bond donors Количество доноров водородных связей (Fauchere *et al*., 1988)	58

10.15789/2220-7619-RAA-1968

FIGURES

Figure 1. General scheme of the computational pipeline. It consists of five parts: encoding HA sequences by a reduced amino acid alphabet; selecting significant sites; clustering the AAindex1 database using mutual information of selected sites; encoding the sites by a representative from each cluster; and finally training the classifier.

Рис. 1. Общая схема вычислительного конвейера. Он состоит из пяти частей: кодирование последовательностей НА сокращенным аминокислотным алфавитом; выбор значимых участков; кластеризация базы данных AAindex1 с использованием взаимной информации выбранных сайтов; кодирование сайтов представителем от каждого кластера; и, наконец, обучение классификатора.



ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

Figure 2. Generation of the pseudo-AAindex1 database from the hydrophobicity index. The pseudo database is created based on the selected RAAA. Note that the value assigned to each group is the average of the group's amino acid values in the scaled AAindex1 vector.

Рисунок 2. Создание базы данных псевдо-AAindex1 из индекса гидрофобности. Псевдобаза данных создается на основе выбранного RAAA. Обратите внимание, что значение, присвоенное каждой группе, представляет собой среднее значение величин аминокислот группы в масштабированном векторе AAindex1.

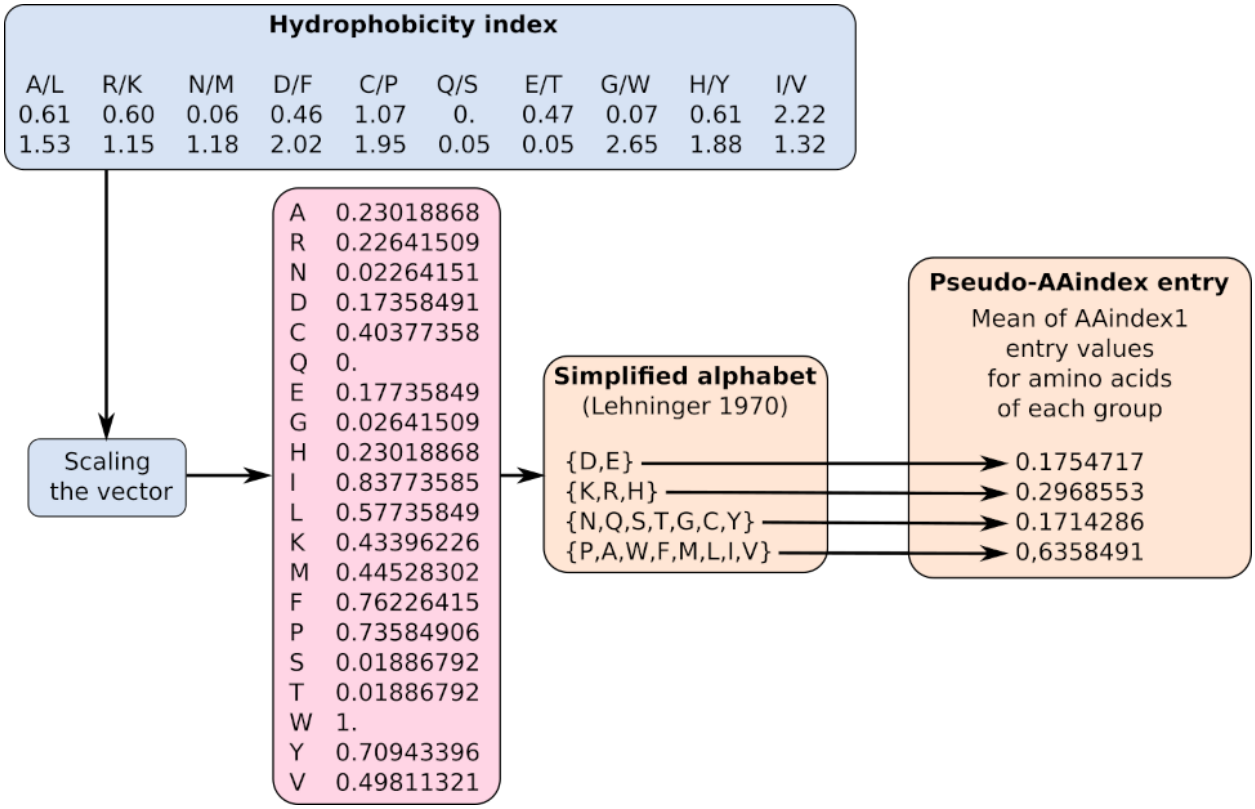
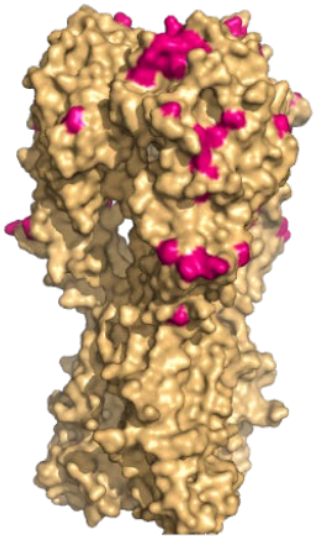
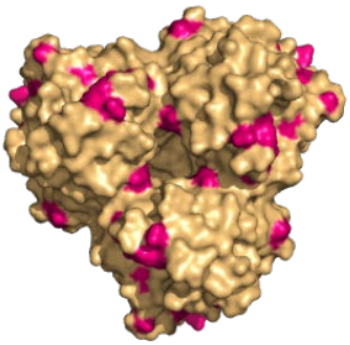
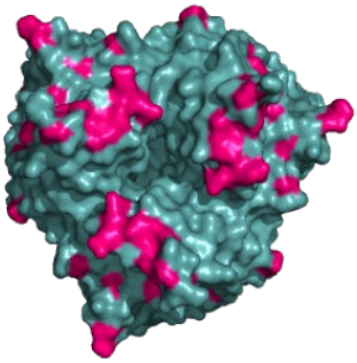


Figure 3. Visualization of high-impact sites on the surface of hemagglutinin protein by PyMOL [26]. Top – H1 protein (PDB ID: 1RUY [3, 12]). Bottom – H3 protein (PDB ID: 5THF [3, 33]). Note that the highlighted sites include not only the antigenic sites but also those experimentally determined as T-cell epitopes, B-cell epitopes, as well as MHC-binding epitopes of different classes.

Рисунок 3. Визуализация участков сильного воздействия на поверхности белка гемагглютинина с помощью PyMOL [26]. Вверху – белок H1 (PDB ID: 1RUY [3, 12]). Внизу – белок H3 (PDB ID: 5THF [3, 33]). Обратите внимание, что выделенные сайты включают не только антигенные сайты, но и те, которые экспериментально определены как Т-клеточные эпитопы, В-клеточные эпитопы, а также MHC-связывающие эпитопы разных классов.

10.15789/2220-7619-RAA-1968



ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

Figure 4. Correlation matrix of 11 unique AAindex1 entries from Table V. Note that the majority of indices have low correlation.

Рисунок 4. Матрица корреляции 11 уникальных записей AAindex1 из таблицы V. Обратите внимание, что большинство индексов имеют низкую корреляцию.

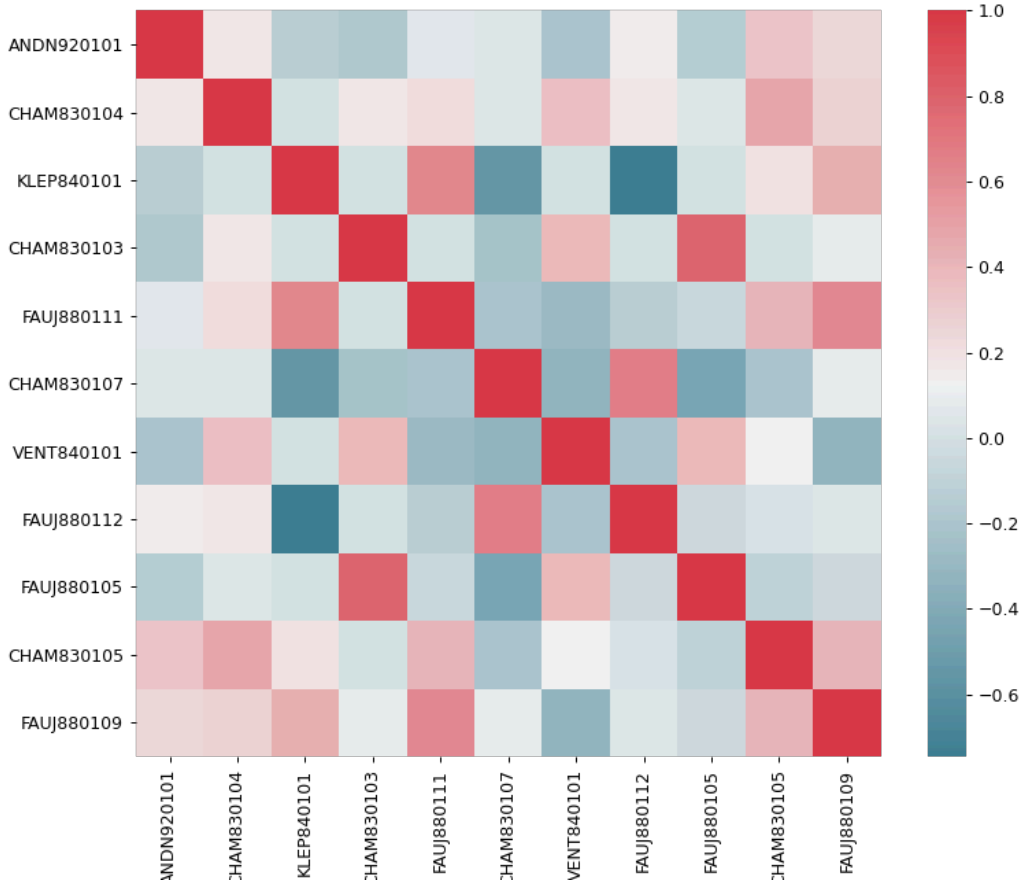
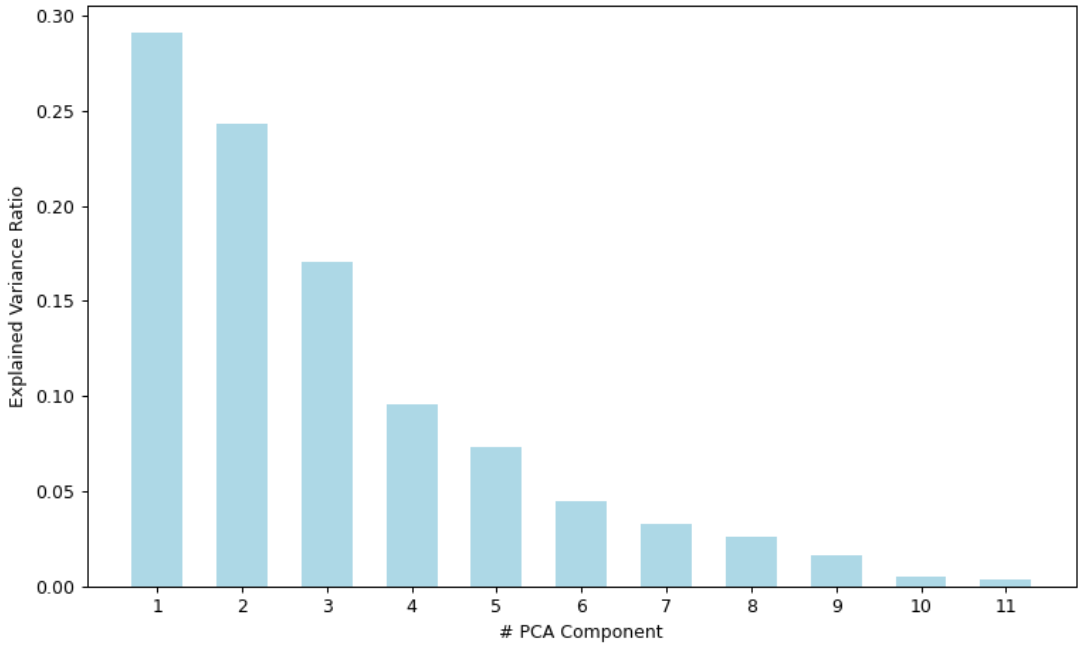


Figure 5. Explained variance ratios for PCA analysis components. Analysis was performed for 11 unique AAindex1 indices from Table V. The result shows that the first six components represent more than 90% of the identified variance.

Рисунок 5. Выявленные коэффициенты дисперсии для компонентов анализа PCA. Анализ был проведен для 11 уникальных индексов AAindex1 из таблицы V. Результат показывает, что первые шесть компонентов представляют более 90% выявленной дисперсии.



TITLE PAGE_METADATA

CORRESPONDING AUTHOR: Forghani Majid – Ph.D. in Physics and Mathematics, researcher (N.N. Krasovskii Institute of Mathematics and Mechanics), senior Researcher (Ural Federal University), 16 S.Kovalevskaya Str., Yekaterinburg, 620108 Russia, +7 (906) 811 9201, forghani@imm.uran.ru

Форгани Маджид –кандидат физико-математических наук, научный сотрудник отдела математического программирования, Федеральное государственное бюджетное учреждение науки Институт математики и механики им. Н. Н. Красовского Уральского отделения Российской академии наук (ИММ УрО РАН), старший научный сотрудник Уральского Федерального Университета (УрФУ)
620108, Россия, г. Екатеринбург, ул. Софьи Ковалевской, д. 16, +7 (906) 811 9201, forghani@imm.uran.ru

Фирстков А.Л – математик первой категории отдела математического программирования, Федеральное государственное бюджетное учреждение науки Институт математики и механики им. Н. Н. Красовского Уральского отделения Российской академии наук (ИММ УрО РАН)

Даниленко Д.М. – кандидат биологических наук**,** заместитель директора по научной работе, заведующий отделом этиологии и эпидемиологии (ФГБУ Научно-исследовательский институт гриппа имени А.А. Смородинцева, Минздрава России)

Комиссаров А.Б. – заведующий лабораторией молекулярной вирусологии (ФГБУ Научно-исследовательский институт гриппа имени А.А. Смородинцева, Минздрава России)

10.15789/2220-7619-RAA-1968

Running head/Краткое название: Вложение CAA и антигенная эволюция/RAAA encoding & antigenic evolution

Keywords: AAindex, antigenic evolution, hemagglutinin, influenza, modeling, reduced amino acid alphabet

Ключевые слова: AAindex, антигенная эволюция, гемагглютинин, грипп, моделирование, сокращённый аминокислотный алфавит

Original article

20 pages, 5 table, 5 figure

31.05.2022

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

REFERENCES

Порядковый номер ссылки	Авторы, название публикации и источника, где она опубликована, выходные данные	ФИО, название публикации и источника на английском	Полный интернет-адрес (URL) цитируемой статьи и/или
1	Andersen, C. A., & Brunak, S., Representation of protein-sequence information by amino acid subalphabets. AI magazine, 2004, vol. 25, no. 1, pp. 97–97	---	https://doi.org/10.1609/aimag.v25i1.1750 [10.1609/aimag.v25i1.1750]
2	Arinaminpathy, N., & Grenfell, B. Dynamics of glycoprotein charge in the evolutionary history of human influenza. PloS one, 2010, vol. 5, no. 12, pp. e15674.	---	https://doi.org/10.1371/journal.pone.0015674 [10.1371/journal.pone.0015674]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

3	Berman H. M. et al. The protein data bank. Nucleic acids research. Vol. 28, no. 1, pp. 235–242, 2000.	---	https://doi.org/10.1093/nar/28.1.235 [10.1093/nar/28.1.235]
4	Burns, A., Van der Mensbrugghe, D., & Timmer, H. Evaluating the economic consequences of avian influenza, in Plastics, World Bank Washington, DC, 2006.	---	https://www.academia.edu/download/72419367/474170WP0Evalu101PUBLIC10Box334133B.pdf
5	Cannata, N., Toppo, S., Romualdi, C., & Valle, G. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. Bioinformatics, vol. 18, no. 8, pp. 1102–1108, 2002.	---	https://doi.org/10.1093/bioinformatics/18.8.1102 [10.1093/bioinformatics/18.8.1102]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

6	Cui, H., Wei, X., Huang, Y., Hu, B., Fang, Y., & Wang, J. Using multiple linear regression and physicochemical changes of amino acid mutations to predict antigenic variants of influenza A/H3N2 viruses. Bio-medical materials and engineering, vol. 24, no. 6, pp. 3729–3735, 2014.	---	https://doi.org/10.3233/BME-141201 [10.3233/BME-141201]
7	de Brevern, A. G. New assessment of a structural alphabet. In silico biology, vol. 5, no. 3, pp. 283–289, 2005.	---	https://content.iospress.com/articles/in-silico-biology/isb00186
8	Edgar, R. C. Local homology recognition and distance measures in linear time using compressed amino acid alphabets. Nucleic acids research, vol. 32, no. 1, pp. 380–385, 2004.	---	https://doi.org/10.1093/nar/gkh180 [10.1093/nar/gkh180]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

9	Etchebest, C., Benros, C., Bornot, A., Camproux, A. C., & De Brevern, A. G. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. European Biophysics Journal, vol. 36, no. 8, pp. 1059–1069, 2007.	---	https://doi.org/10.1007/s00249-007-0188-5 [10.1007/s00249-007-0188-5]
10	Forghani, M., & Khachay, M. Convolutional Neural Network Based Approach to in Silico Non-Anticipating Prediction of Antigenic Distance for Influenza Virus. Viruses, vol. 12, no. 9, pp. 1019, 2020.	---	https://doi.org/10.3390/v12091019 [10.3390/v12091019]
11	Forghani, M., Khachay, M., & AlyanNezhadi, M. M. The Impact of Amino Acid Encoding on the Prediction of Antigenic Variants. In 2020 6th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), pp. 1–5, 2020.	---	https://doi.org/10.1109/ICSPIS51611.2020.9349560 [10.1109/ICSPIS51611.2020.9349560]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

12	Gamblin, S. J., et al. The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science, vol. 303, no. 5665, pp. 1838–1842, 2004.	---	https://doi.org/10.1126/science.1093155 [10.1126/science.1093155]
13	Gregory, V., et al. Human former seasonal Influenza A (H1N1) haemagglutination inhibition data 1977-2009 from the WHO Collaborating Centre for Reference and Research on Influenza, London, UK.University of Glasgow, 2016.	---	http://dx.doi.org/10.5525/gla.researchdata.289 [10.5525/gla.researchdata.289]
14	Huang, Z. Z., Yu, L., Huang, P., Liang, L. J., & Guo, Q. Charged amino acid variability related to N-glyco-sylation and epitopes in A/H3N2 influenza: Hem-agglutinin and neuraminidase. PloS one, vol. 12, no. 7, pp. e0178231, 2017.	---	https://doi.org/10.1371/journal.pone.0178231 [10.1371/journal.pone.0178231]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

15	Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., & Kanehisa, M. AAindex: amino acid index database, progress report 2008. Nucleic acids research, vol. 36, no. suppl 1, pp. D202–D205, 2007.	---	https://doi.org/10.1093/nar/gkm998 [10.1093/nar/gkm998]
16	Klingen, T. R., Reimering, S., Guzmán, C. A., & McHardy, A. C. In silico vaccine strain prediction for human influenza viruses. Trends in microbiology, vol. 26, no. 2, pp. 119–131, 2018.	---	https://doi.org/10.1016/j.tim.2017.09.001 [10.1016/j.tim.2017.09.001]
17	Kobayashi, Y., & Suzuki, Y. Compensatory evolution of net-charge in influenza A virus hemagglutinin. PloS one, vol. 7, no. 7, pp. E40422, 2012.	---	https://doi.org/10.1371/journal.pone.0040422 [10.1371/journal.pone.0040422]

10.15789/2220-7619-RAA-1968

18	Lee, M. S., & Chen, J. S. E. Predicting antigenic variants of influenza A/H3N2 viruses. Emerging infectious diseases, vol. 10, no. 8, pp. 1385, 2004.	---	https://doi.org/10.3201%2Feid1008.040107 [10.3201%2Feid1008.040107]
19	Lenckowski, J., & Walczak, K. Simplifying amino acid alphabets using a genetic algorithm and sequence alignment. In European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, pp. 122–131, 2007.	---	https://doi.org/10.1007/978-3-540-71783-6_12 [10.1007/978-3-540-71783-6_12]
20	Li, T., Fan, K., Wang, J., & Wang, W. Reduction of protein sequence complexity by residue grouping. Protein Engineering, vol. 16, no. 5, pp. 323–330, 2003.	---	https://doi.org/10.1093/protein/gzg044 [10.1093/protein/gzg044]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

21	Nanni, L., & Lumini, A. A genetic approach for building different alphabets for peptide and protein classification. BMC bioinformatics, vol. 9, no. 1, pp. 1–10, 2008.	---	https://doi.org/10.1186/1471-2105-9-45 [10.1186/1471-2105-9-45]
22	Pedregosa, F., et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.	---	https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com
23	Prlić, A., Domingues, F. S., & Sippl, M. J. Structure-derived substitution matrices for alignment of distantly related sequences. Protein Engineering, vol. 13, no. 8, pp. 545–550, 2000.	---	https://doi.org/10.1093/protein/13.8.545 [10.1093/protein/13.8.545]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

24	Qiu, J., Qiu, T., Yang, Y., Wu, D., & Cao, Z. Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2. Scientific reports, vol. 6, no. 1, pp. 1–9, 2016.	---	https://doi.org/10.1038/srep31156 [10.1038/srep31156]
25	Risler, J. L., Delorme, M. O., Delacroix, H., & Henaut, A. Amino acid substitutions in structurally related proteins a pattern recognition approach: Determination of a new and efficient scoring matrix. Journal of molecular biology, vol. 204, no. 4, pp. 1019–1029, 1988.	---	https://doi.org/10.1016/0022-2836(88)90058-7 [10.1016/0022-2836(88)90058-7]
26	Schrödinger, L. L. C. The PyMOL molecular graphics system, version 1.8, 2015.	---	https://pymol.org/2/

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

27	Smith, D. J., Forrest, S., Ackley, D. H., & Perelson, A. S. Variable efficacy of repeated annual influenza vaccination. Proceedings of the National Academy of Sciences, vol. 96, no. 24, pp. 14001–14006, 1999.	---	https://doi.org/10.1073/pnas.96.24.14001 [10.1073/pnas.96.24.14001]
28	Smith, D. J., Lapedes, A. S., De Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D., & Fouchier, R. A. apping the antigenic and genetic evolution of influenza virus. Science, vol. 305, no. 5682, pp. 371–376, 2004.	---	https://doi.org/10.1126/science.1097211 [10.1126/science.1097211]
29	Stephenson, J. D., & Freeland, S. J. Unearthing the root of amino acid similarity. Journal of molecular evolution, vol. 77, no. 4, pp. 159–169, 2013.	---	https://doi.org/10.1007/s00239-013-9565-0 [10.1007/s00239-013-9565-0]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

30	Su, S., Fu, X., Li, G., Kerlin, F., & Veit, M. Novel Influenza D virus: Epidemiology, pathology, evolution and biological characteristics. Virulence, vol. 8, no. 8, pp. 1580–1591, 2017.	---	https://doi.org/10.1080/21505594.2017.1365216 [10.1080/21505594.2017.1365216]
31	Sylte, M. J., & Suarez, D. L. Influenza neuraminidase as a vaccine antigen. Vaccines for Pandemic Influenza, pp. 227–241, 2009.	---	https://doi.org/10.1007/978-3-540-92165-3_12 [10.1007/978-3-540-92165-3_12]
32	Tomii, K., & Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Engineering, Design and Selection, vol. 9, no. 1, pp. 27–36, 1996.	---	https://doi.org/10.1093/protein/9.1.27 [10.1093/protein/9.1.27]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

33	Tzarum N. et al. The 150-loop restricts the host specificity of human H10N8 influenza virus. Cell reports. vol. 19, no. 2, pp. 235–245, 2017.	---	https://doi.org/10.1016/j.celrep.2017.03.054 [10.1016/j.celrep.2017.03.054
34	Wang, P., Zhu, W., Liao, B., Cai, L., Peng, L., & Yang, J. Predicting influenza antigenicity by matrix completion with antigen and antiserum similarity. Frontiers in microbiology, vol. 9, pp. 2500, 2018.	---	https://doi.org/10.3389/fmicb.2018.02500 [10.3389/fmicb.2018.02500]
35	Wikramaratna, P. S., Sandeman, M., Recker, M., & Gupta, S. The antigenic evolution of influenza: drift or thrift?. Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 368, no. 1614, pp. 20120200, 2013.	---	https://doi.org/10.1098/rstb.2012.0200 [10.1098/rstb.2012.0200]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

36	World Health Organization, Influenza fact sheet: Overview, Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire, vol. 78, no. 11, pp. 77–80, 2003.	---	https://apps.who.int/iris/handle/10665/232113
37	Yang, H., Carney, P. J., Chang, J. C., Guo, Z., Villanueva, J. M., & Stevens, J. Structure and receptor binding preferences of recombinant human A (H3N2) virus hemagglutinins. Virology, vol. 477, pp. 18–31, 2015.	---	https://doi.org/10.1016/j.virol.2014.12.024 [10.1016/j.virol.2014.12.024]
38	Yang, X. Y., Shi, X. H., Meng, X., Li, X. L., Lin, K., Qian, Z. L., ... & Cai, Y. D. Classification of transcription factors using protein primary structure. Protein and peptide letters, vol. 17, no. 7, pp. 899–908, 2010.	---	https://doi.org/10.2174/092986610791306670 [10.2174/092986610791306670]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

10.15789/2220-7619-RAA-1968

39	Yao, Y., Li, X., Liao, B., Huang, L., He, P., Wang, F., ... & Yang, J. Predicting influenza antigenicity from Hemagglutintin sequence data based on a joint random forest method. Scientific reports, vol. 7, no. 1, pp. 1–10, 2017.	---	https://doi.org/10.1038/s41598-017-01699-z [10.1038/s41598-017-01699-z]
40	Zhang, Y., Aevermann, B. D., Anderson, T. K., Burke, D. F., Dauphin, G., Gu, Z., ... & Scheuermann, R. H. Influenza Research Database: An integrated bioinformatics resource for influenza virus research. Nucleic acids research, vol. 45, no. D1, pp. D466–D474, 2017.	---	https://doi.org/10.1093/nar/gkw857 [10.1093/nar/gkw857]
41	Zhang, Z. H., Wang, Z. H., Zhang, Z. R., & Wang, Y. X. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS letters, vol. 580, no. 26, pp. 6169–6174, 2006.	---	https://doi.org/10.1016/j.febslet.2006.10.017 [10.1016/j.febslet.2006.10.017]

ISSN 2220-7619 (Print)
ISSN 2313-7398 (Online)

42	Zuo, Y. C., & Li, Q. Z. Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. Peptides, vol. 30, no. 10, pp. 1788–1793, 2009.	---	https://doi.org/10.1016/j.peptides.2009.06.032 [10.1016/j.peptides.2009.06.032]