MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **80** (2018) 295-310

ISSN 0340 - 6253

An Application of the 2D-Dynamic Representation of DNA/RNA Sequences to the Prediction of Influenza A Virus Subtypes

Damian Panas[a], Piotr Wąż[b], Dorota Bielińska–Wąż[a], Ashesh Nandy[c], Subhash C. Basak[d]

[a] *Department of Radiological Informatics and Statistics, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland*

[b] *Department of Nuclear Medicine, Medical University of Gdańsk, Tuwima 15, 80-210 Gdańsk, Poland*

[c] *Centre for Interdisciplinary Research and Education, 404B Jodhpur Park, Kolkata 700068, India*

[d] *University of Minnesota Duluth-Natural Resources Research Institute and Department of Chemistry and Biochemistry, University of Minnesota Duluth, 5013 Miller Trunk Highway, Duluth, MN 55811, USA*

(Received February 5, 2018)

Abstract

A new theoretical method for the virus identification has been proposed. *The 2D-Dynamic Representation of DNA/RNA Sequences* has been applied to the prediction of influenza A virus subtypes. We have shown that the method can be successfully combined with novel supervised machine learning algorithms, such as C5.0. The descriptors of the 2D-Dynamic Representation of DNA/RNA Sequences have been evaluated. High mean accuracy of predicting the subtype of the influenza A virus has been obtained (over 90% of correct predictions). As a consequence, the combination of the machine learning algorithms and the 2D-Dynamic Representation of DNA/RNA Sequences has been shown to constitute a simple and accurate tool for the classification of unidentified virus strains.

1 Introduction

Influenza is a deadly viral disease that has claimed millions of lives since the epidemic in 1918, the Spanish Flu, when over 20 million people succumbed to the illness [1]. The flu virion is characterized by a segmented genome that permits new subtypes to form when two or more varieties of the flu infect a single host cell. This characteristic has enabled various subtypes to evolve and create pandemics over the last 100 years, the most recent manifestation being the Swine Flu epidemic of 2009 that spread rapidly across the world from Mexico where it was first reported to Singapore and beyond [2].

One feature that characterizes this virus to create epidemics every few years is its ability to mutate to new strains through reassortments of its constituent genetic material. Apparently, it transpires, the viral genome is composed of several genes grouped together in individual segments. In a host cell during replication, it would be possible, even at a low probability level, to reassemble individual virions that combine segments from two or more viral elements as available and remain viable. New viruses could presumably, in some instance, give rise to more intense ailments to humans not used to the novel strains and cause epidemics [3].

Given that there are many possibilities of combinations among the two important genes or RNA sequences coding for surface proteins, hemagglutinin and neuraminidase, of the influenza virion, it is surprising to note that only a few combinations appear to be functional in nature [4]. Several authors have opined on various aspects of this phenomenon [4–9]. In an earlier paper of our group it was postulated that a coupling existed between the hemagglutinin and neuraminidase constituents that precluded certain combinations [4].

Here, in the present paper, we carry this idea forward through a 2D-Dynamic Representation of DNA/RNA Sequences using a supervised machine learning model and a set of trainee sequences to predict, with over 90% accuracy, the subtypes of the remaining influenza strains. We hope that this study may provide a guide for the future work on influenza surveillance to control, if not to avoid, the upcoming epidemics. The presented approach is also a methodological contribution – it constitutes a novel virus identification tool.

2 Materials and Methods

A DNA/RNA sequence is a succession of letters (A – adenine, C – cytosine, G – guanine, T/U – thymine/uracil) that indicate the order of bases within a DNA/RNA chain. A class of methods in bioinformatics called *Graphical Representations* allows for both graphical and numerical similarity/dissimilarity analysis of such objects. Many examples of such approaches with different applications may be found in the literature [10–27] (for reviews see [28–30]). Within these methods one can create a large number of different types of values which characterize numerically graphs representing the sequences. After [28] we refer to these quantities as *descriptors*. The aim of these methods is the creation of both graphs and descriptors in a unique way. An important requirement which should be fulfilled by a new method is the lack of degeneracy[1].

A popular technique used for the construction of graphs representing the DNA/RNA sequences are walks in either two-dimensional space [31–33] or in three-dimensional space [34–37]. The methods utilizing a space of a given dimension may differ from each other by the way of assigning basis vectors to particular bases, by some details of the construction of the graphs, or by the kind of descriptors representing the graphs. Each base is represented by different basis vector in two or three-dimensional space. Starting from the origin of the coordinate system a shift indicated by the basis vector representing the first base in the sequence is performed. The end of this vector is the starting point for the next shift, as indicated by the basis vector representing the second base in the sequence, and so on. As a consequence, a two- or three-dimensional curve representing the sequence is obtained.

In the present work we apply the method called by us *the 2D-Dynamic Representation of DNA/RNA Sequences* which is based on shifts in a two-dimensional space [38–42]. Within this approach, the sequence is represented by the 2D-dynamic graph which may be interpreted as a set of material points in a two-dimensional space. After a unit shift the point with the unit mass is localized. If the ends of the vectors during the shifts overlap then the mass of this point increases accordingly. Descriptors proposed as a representation of the 2D-dynamic graphs take into account different masses of the points of the graphs. In this way, the so-called repetitive walks, i.e., shifts along the same trace can be taken into account both graphically and numerically.

[1] A degeneracy (nonuniqueness) exists if several different sequences are represented by the same graph or by the same descriptors.

In the present work we consider the following descriptors of the 2D-dynamic graphs [38, 41, 43]:

- Coordinates $(\mu_x,\ \mu_y)$ of the centers of mass of the graphs.

$$\mu_\gamma = \frac{1}{N}\sum_{i=1}^{p} m_i\gamma_i, \quad \gamma = x, y, \quad N = \sum_{i=1}^{p} m_i, \tag{1}$$

  where x_i, y_i are the coordinates of mass m_i in the Cartesian coordinate system for which $(0, 0)$ is the origin of all the sequences and N is the length of the sequence (equal to the total mass of the graph) and p is the number of the material points in the graph.

- Principal moments of inertia (I_{11}, I_{22}) of the graphs.

  The moment of inertia tensor is defined by the matrix

$$\hat{I} = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix} \tag{2}$$

  with elements

$$I_{xy} = I_{yx} = -\sum_{i=1}^{p} m_i x_i^\mu y_i^\mu, \tag{3}$$

$$I_{xx} = \sum_{i=1}^{p} m_i (y_i^\mu)^2, \tag{4}$$

$$I_{yy} = \sum_{i=1}^{p} m_i (x_i^\mu)^2, \tag{5}$$

  where x_i^μ, y_i^μ denote the coordinates of mass m_i in the Cartesian coordinate system with the origin at the center of mass of the graph. Principal moments of inertia are equal to the solutions $I = I_{11}, I_{22}$ of equation

$$\begin{vmatrix} I_{xx} - I & I_{xy} \\ I_{xy} & I_{yy} - I \end{vmatrix} = 0. \tag{6}$$

- Other descriptors.

  We also consider graph radius

$$g_R = \sqrt{\mu_x^2 + \mu_y^2} \tag{7}$$

  and four descriptors $D_k^\gamma = D_1^x, D_2^x, D_1^y, D_2^y$,

$$D_k^\gamma = \frac{\mu_\gamma}{I_{kk}}, \quad k = 1, 2; \quad \gamma = x, y, \tag{8}$$

  related to a relation between coordinates of the center o mass and moments of inertia.

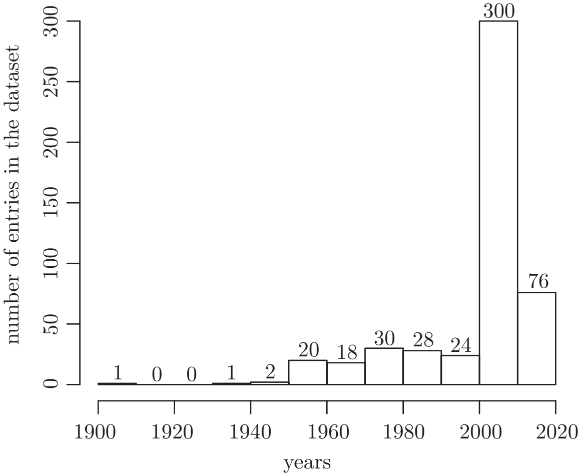

Figure 1. Histogram of the collection years. Most strains in the dataset were collected between 2000 and 2010.

In order to precisely evaluate these parameters using machine learning, it is recommended to operate on relatively large, diverse yet proportionally distributed dataset. To this end, 20 of the most prevalent subtypes of the influenza A virus were employed: H1N1, H1N2, H2N2, H2N3, H3N2, H3N6, H3N8, H4N6, H5N1, H5N2, H6N1, H6N2, H6N6, H7N2, H7N3, H7N7, H7N9, H9N2, H10N7, H11N9 [4]. For each subtype 25 strains ($\omega = 500$ in total), with full coding RNA sequences, were prepared. Only 4th and 6th segment of the genome was considered. The segments code viral proteins, accordingly hemagglutinin (HA) and neuraminidase (NA), which directly characterize the subtype of the influenza A virus.

A collection of the data on the studied viruses are presented in Figure 1. The oldest strain was collected in 1902[2]. The newest ones are from 2017[3]. The mean collection year is 2000. The largest number of entries (50) correspond to 2005.

[2] A/chicken/Brescia/1902(H7N7).

[3] A/duck/NC/91347/01(H1N2) and A/Alaska/20/2017(H1N1).

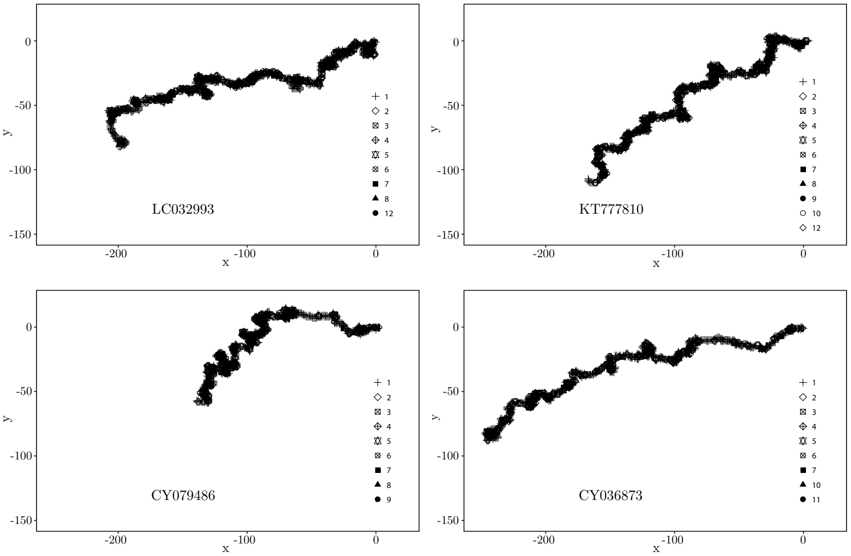

Figure 2. 2D-dynamic graphs describing some selected sequences.

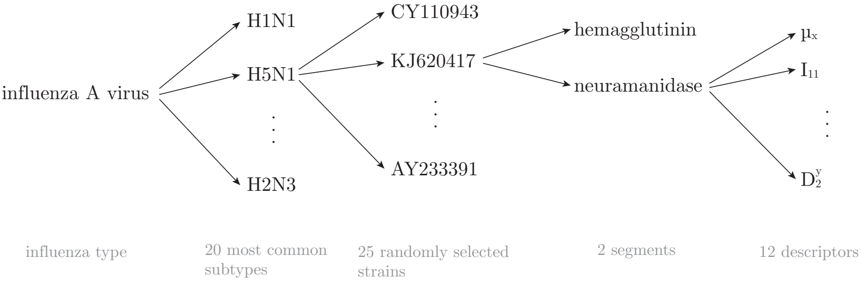

Figure 3. Dataset organization.

Examples of the 2D-dynamic graphs describing several sequences are shown in Figure 2. After the graphs were constructed, the corresponding parameters (the descriptors of the 2D-dynamic graphs) have been calculated. The general structure of the dataset is pictured in Figure 3.

For every segment the appropriate 2D-dynamic graph was built. Each segment of the influenza's RNA genome has, in essence, been transformed into a set of 12 parameters: μ_x, μ_y, g_R, I_{xx}, I_{yy}, $I_{xy} \equiv I_{yx}$, I_{11}, I_{22}, D_x^1, D_y^1, D_x^2, D_y^2. We denote the set as $\mathcal{D}$. Every influenza strain is therefore described by 26 parameters – half of them correspond to hemagglutinin and the other half to neuraminidase.

In order to determine the predictive abilities of particular parameters, a supervised machine learning algorithm has been applied. Using the C5.0 algorithm a decision tree, based on the 80% of the randomly selected observations, has been built. The decision tree has then been tested on the remaining 20% observations of the dataset.

We define *accuracy* as a following fraction:

$$\Theta_{\mathcal{S}}^{\mathcal{F}} = \frac{s}{a}, \tag{9}$$

where s denotes the number of correctly classified subtypes, a is the number of all classified subtypes (in this study a is always equal to 100), $\mathcal{F}$ corresponds to the collection of the selected descriptors[4] and $\mathcal{S} = \{\mathrm{HA}, \mathrm{NA}, \mathrm{HN}\}$ represents the segment coding a particular protein (hemagglutinin, neuraminidase, or both of them).

Because of the randomness of the dataset division, the accuracy of predictions is a random variable as well. For every set of parameters the mean accuracy (defined below) as well as its 95% confidence interval (95% ci) have been calculated. In all the calculations, 95% ci were calculated using the adjusted bootstrap percentile method with 10 000 replicates. In order to obtain such results, the process of splitting the data, creating decision tree, testing the tree, and calculating the accuracy of the predicted outcome had to be repeated n times (Figure 4).

The mean accuracy $\bar{\Theta}_{\mathcal{S}}^{\mathcal{F}}$, which evaluates predictive capabilities of the descriptors, can be, therefore, written as

$$\bar{\Theta}_{\mathcal{S}}^{\mathcal{F}} = \frac{1}{n}\sum_{i=1}^{n} \Theta_{\mathcal{S},i}^{\mathcal{F}}, \tag{10}$$

where $\Theta_{\mathcal{S},i}^{\mathcal{F}}$ is the accuracy of the tree constructed from $\mathcal{F}$ descriptors of the strain S in the i-th iteration.

[4] One should remember that we have not tested every possible combination of the descriptors, which would be $2^{12} - 1 = 4095$, but only some particular sets: $\mathcal{F} = \{\mu_x, \mu_y, g_R, I_{xx}, I_{yy}, I_{xy}, I_{11}, I_{22}, D_x^1, D_y^1, D_x^2, D_y^2, I, \hat{I}, D_k^\gamma, all\}$, where $all = \{\mu_x, \mu_y, g_R, I_{xx}, I_{yy}, I_{xy}, I_{11}, I_{22}, D_x^1, D_y^1, D_x^2, D_y^2\}$.

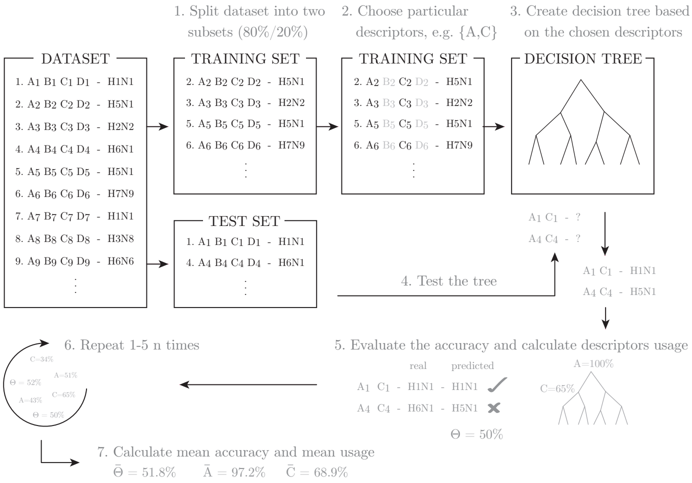

Figure 4. A flowchart of the method.

In order to optimize the computation time and the accuracy of predictions, the number of repetitions (n) and boosting trials have been adjusted. The results of the adjustment are shown in Figure 5 and in Figure 6. As one can see, the mean accuracy is roughly constant when the computation time increases due to the increasing number of repetitions (Figure 5). Taking a big value of n is therefore unnecessary since it would not noticeably improve the quality of the predictions, but would extend the computation time. On the other hand, a very small n, seen as the statistical population size, may affect the bootstrap method. It is assumed that for the considered application $n = 30$ is a reasonable estimate of the optimal value.

The accuracy of the decision tree has been improved by an adjustment of the number of boosting trials. The idea of the boosting trial itself is to create several classifiers instead of only one. Later, in an iterative process, these classifiers converge to the final classifier. On the other hand, the number of boosting trials distinctly affects the computation time. From Figure 6 one may conclude that the improvement of the results is, in our case, insignificant, yet computation time changes considerably. The number of boosting trials equal to 10 seems to be a reasonable choice.

Since one of the goals of this study is to evaluate every descriptor alike, the winnowing of the descriptors (which, for the sake of simplicity, dismisses some descriptors from the decision tree) has not been considered. Other advanced parameters of C5.0, such us fuzzy

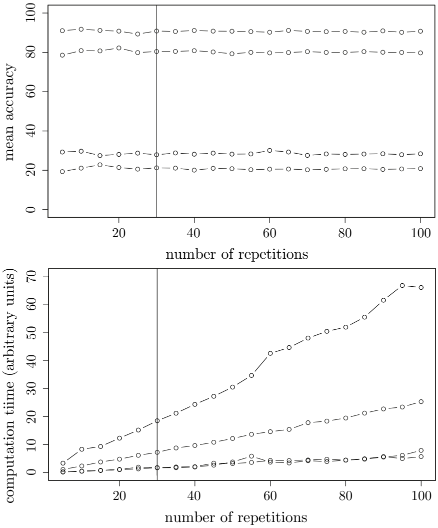

Figure 5. The influence of the number of repetitions on the quality of the results. Top panel: the accuracy versus the number of repetitions; bottom panel: the computation time versus the number of repetitions. The vertical line indicates the estimated optimum value of n.

threshold and global pruning, are of some secondary importance and have been set by default (either tuned off or tuned on).

Another way of evaluating descriptors using the C5.0 algorithm is by calculating their percentage usage in the process of classification. For example, the descriptor in the first split (node) of the decision tree takes part in classifying all observations. Therefore its usage is equal to 100%. Contrary, the terminal nodes, that cover only a minority of set samples, have vastly smaller importance, that may be even close to 0%. The average usage of every descriptor from n decision trees, based on the set of all parameters, has also been calculated. In general, one may define the average usage as

$$\bar{\Psi}^{\mathcal{F},\mathcal{D}}_{\mathcal{S},\hat{\mathcal{S}}} = \frac{1}{n}\sum_{i=1}^{n} \Psi^{\mathcal{F},\mathcal{D}}_{\mathcal{S},\hat{\mathcal{S}},i} \qquad \text{for} \quad \Psi^{\mathcal{F},\mathcal{D}}_{\mathcal{S},\hat{\mathcal{S}},i} = \frac{q}{\omega - a}, \tag{11}$$

where $\Psi^{\mathcal{F},\mathcal{D}}_{\mathcal{S},\hat{\mathcal{S}},i}$ corresponds to the usage of the descriptor $\mathcal{D}$ and segment $\hat{\mathcal{S}} = \{\text{HA}, \text{NA}\}$ for the tree constructed from set $\mathcal{F}$ of segment $\mathcal{S}$ in the i-th iteration. Value q represents number of observations that take part in classification by descriptor $\mathcal{D}$ and segment $\hat{\mathcal{S}}$ in the tree, and $\omega - a$ stands for the number of entries in the training set (in our case $\omega - a$ is always equal to 400).

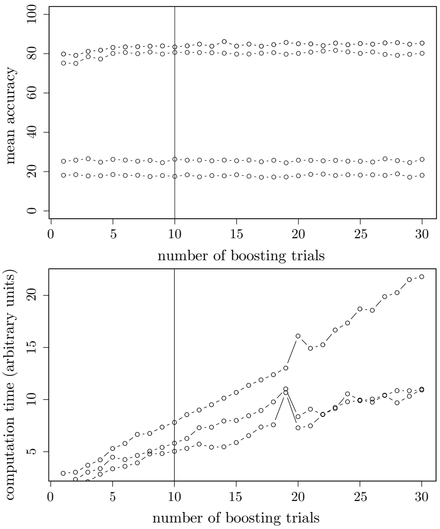

Figure 6. Influence of the boosting trials. Top panel: the dependence of the accuracy on the number of boosting trials; bottom panel: the dependence of the computation time on the number of boosting trails. The vertical line indicates the estimated optimum value of n.

Since within the scope of our interest is comparing descriptors on the same decision tree, the set of all possible parameters can be significantly narrowed down. For our purpose the equation (11) can be rewritten to

$$\bar{\Psi}^{all,\mathcal{D}}_{\mathrm{HN},\hat{\mathcal{S}}} = \frac{1}{n}\sum_{i=1}^{n}\Psi^{all,\mathcal{D}}_{\mathrm{HN},\hat{\mathcal{S}},i},$$

which from now will be denoted as

$$\bar{\Psi}^{\mathcal{D}}_{\hat{\mathcal{S}}} = \frac{1}{n}\sum_{i=1}^{n}\Psi^{\mathcal{D}}_{\hat{\mathcal{S}},i}.$$

3 Results and discussion

From Figure 7 and Table 1 one may conclude that using the present method one may obtain over 90% accuracy of the predictions ($\bar{\Theta}^{all}_{\mathrm{HN}} = 91.89\%$). Therefore, the descriptors of the 2D-dynamic graphs, especially their whole set, carry a lot of information about the represented sequence. It is worth mentioning that some smaller sets of parameters may give satisfactory results as well. For example, the accuracy is over 80% both for $\hat{I}$ and for D_k^{γ} ($\bar{\Theta}^{\hat{I}}_{\mathrm{HN}} = 81.338\%$, $\bar{\Theta}^{D_k^{\gamma}}_{\mathrm{HN}} = 84.412\%$), and 76.036% for I. Most accurate decision trees with only one parameter from each protein are based on the D_y^1, I_{xx} and μ_y ($\bar{\Theta}_{\mathrm{HN}}^{D_y^1} = 63.654\%$, $\bar{\Theta}_{\mathrm{HN}}^{I_{xx}} = 61.936\%$ and $\bar{\Theta}_{\mathrm{HN}}^{\mu_y} = 61.249\%$). As yet, it is not clear why for some parameters the mean accuracy significantly differs among proteins. For example, the accuracy for hemagglutinin's descriptor D_y^1 is 41.1% and for neuraminidase's only 18.7%.

The mean usage of descriptors is pictured in Figure 8 and Table 2. As one can see, the most important descriptor is nauraminidaze's I_{xx}. Its average usage $\bar{\Psi}_{\mathrm{NA}}^{I_{xx}} = 100\%$ means that for every decision tree this descriptor occurs in the very first node. Neuraminidaze's D_y^1 and μ_y take part in almost every classification as well ($\bar{\Psi}_{\mathrm{NA}}^{D_y^1} = 96.847\%$, $\bar{\Psi}_{\mathrm{NA}}^{\mu_y} = 98.388\%$).

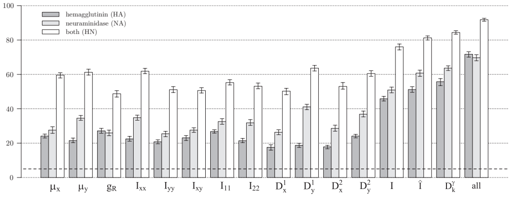

Figure 7. Mean accuracy. Dashed line represents level of completely random predictions.

Table 1. The mean accuracy of the predicted subtypes

$\mathcal{F}$	$\bar{\Theta}_{\mathrm{HA}}^{\mathcal{F}}$	95% ci	$\bar{\Theta}_{\mathrm{NA}}^{\mathcal{F}}$	95% ci	$\bar{\Theta}_{\mathrm{HN}}^{\mathcal{F}}$	95% ci
μ_x	24.113	23.014–25.166	27.580	25.698–29.521	59.571	57.925–61.959
μ_y	21.507	20.277–22.940	34.536	33.267–36.087	61.249	59.637–63.046
g_R	27.093	25.733–28.621	25.928	24.303–27.570	48.713	46.829–50.556
I_{xx}	22.412	21.121–23.911	34.807	33.311–36.291	61.936	60.412–63.527
I_{yy}	20.738	19.676–21.908	25.434	23.716–26.954	51.112	49.537–52.933
I_{xy}	23.186	21.625–24.412	27.605	26.365–28.935	50.622	49.217–53.243
I_{11}	26.736	25.826–27.822	32.698	31.002–34.232	55.255	53.954–56.932
I_{22}	21.498	20.323–22.718	31.832	30.334–33.737	53.128	30.385–33.745
D_x^1	17.648	15.908–19.003	26.305	24.973–27.848	50.338	48.249–51.908
D_y^1	18.732	17.543–19.987	41.112	39.532–42.674	63.654	62.009–65.232
D_x^2	17.865	16.712–18.741	28.658	27.011–30.546	53.034	51.436–55.328
D_y^2	24.170	23.148–25.158	36.998	35.329–38.792	60.522	58.943–62.273
I	45.821	44.553–47.243	50.951	49.337–52.669	76.039	74.230–77.743
$\hat{I}$	51.235	49.642–52.903	60.843	58.932–62.505	81.338	80.038–82.480
D_k^γ	55.771	53.521–57.639	63.742	62.395–65.043	84.412	83.465–85.479
all	71.734	70.107–73.254	69.708	67.953–71.532	91.890	91.028–92.744

Table 2. The mean usage of descriptors

	HA		NA	
$\mathcal{D}$	$\bar{\Psi}^{\mathcal{D}}_{\mathrm{HA}}$	95% ci	$\bar{\Psi}^{\mathcal{D}}_{\mathrm{NA}}$	95% ci
μ_x	70.938	67.799–74.707	62.706	57.764–67.665
μ_y	72.030	66.993–77.680	98.388	97.007–99.344
g_R	34.800	29.798–40.333	67.303	62.921–71.009
I_{xx}	85.553	81.693–89.030	100.000	99.798–100.000
I_{yy}	73.042	69.785–76.393	31.282	25.266-40.038
I_{xy}	78.248	72.438–83.844	68.259	62.688–75.081
I_{11}	26.778	21.590–33.939	50.781	45.659–58.121
I_{22}	48.167	44.037–52.159	75.701	64.952–84.514
D^1_x	63.637	59.196–67.706	25.713	22.072–29.785
D^1_y	48.680	40.261–55.859	96.847	88.677–100.000
D^2_x	37.842	33.513–42.062	63.158	58.061–68.156
D^2_y	70.068	64.367–75.095	81.832	76.214–86.296

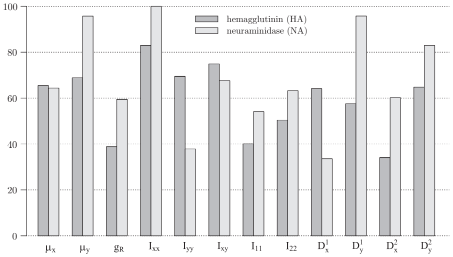

Figure 8. The mean usage of descriptors.

4 Conclusions

Combining novel machine learning algorithms, such as C5.0, with the 2D-Dynamic Representation of the DNA/RNA Sequences can set up simple yet accurate virus characterization tool. In the present study we have shown that such combination conserves its promising properties even for viruses with diverse and complex lineages, i.e., influenza. One can therefore use, or even extend, this approach by applying the 2D-Dynamic Representation of the DNA/RNA Sequences, along with machine learning algorithms, to characterize unknown virus strains.

References

[1] T. M. Tumpey, C. F. Basler, P. V. Aguilar, H. Zeng, A. Solórzano, D. E. Swayne, N. J. Cox, J. M. Katz, J. K. Taubenberger, P. Palese, A. García–Sastre, Characterization of the reconstructed 1918 Spanish influenza pandemic virus, *Science* **310** (2005) 77–80.

[2] CDC 2010. https://www.cdc.gov/h1n1flu/cdcresponse.htm (Accessed 23 Jan 2018).

[3] A. Nandy, S. C. Basak, Viral epidemics and vaccine preparedness, *J. Mol. Pathol. Epidemiol.* **2** (2017) 1–5.

[4] A. Nandy, T. Sarkar, S. C. Basak, P. Nandy, S. Das, Characteristics of influenza HA-NA interdependence determined through a graphical technique, *Curr. Comput. Aided Drug Design* **10** (2014) 285–302.

[5] A. De, T. Sarkar, A. Nandy, Bioinformatics studies of influenza A hemagglutinin sequence data indicate recombination-like events leading to segment exchanges, *BMC Res Notes* **9** (2016) 222.

[6] Y. Zhang, X. Lin, G. Wang, J. Zhou, J. Lu, H. Zhao, F. Zhang, J. Wu, C. Xu, N. Du, Z. Li, Y. Zhang, X. Wang, S. Bi, Y. Shu, H. Zhou, W. Tan, X. Wu, Z. Chen, Y. Wang, Neuraminidase and hemagglutinin matching patterns of a highly pathogenic avian and two pandemic H1N1 influenza A Virus, *PLoS One* **5** (2010) #e9167.

[7] R. Wagner, M. Matrosovich, H. D. Klenk, Functional balance between haemagglutinin and neuraminidase in influenza virus infections, *Rev. Med. Virol.* **12** (2002) 159–166.

[8] S. J. Gamblin, J. J. Skehel, Influenza hemagglutinin and neuraminidase membrane glycoproteins, *J. Biol. Chem.* **285** (2010) 28403–28409.

[9] W. Hu, The interaction between the 2009 H1N1 influenza a hemagglutinin and neuraminidase: mutations, co-mutations and the NA stalk motif, *J. Biomed. Sci. Eng.* **3** (2010) 1–12.

[10] M. Randić, M. Vračko, N. Lerš, D. Plavšić, Novel 2-D graphical representation of DNA sequences and their numerical characterization, *Chem. Phys. Lett.* **368** (2003) 1–6.

[11] J. Zupan, M. Randić, Algorithm for coding DNA sequences into ”spectrum-like” and ”zigzag” representations, *J. Chem. Inf. Model.* **45** (2005) 309–313.

[12] J. Song, K. Tang, A new 2-D graphical representation of DNA sequences and their numerical characterization, *J. Biochem. Bioph. Meth.* **63** (2005) 228–239.

[13] M. Randić, J. Zupan, D. Vikić–Topić, D. Plavšić, A novel unexpected use of a graphical representation of DNA: Graphical alignment of DNA sequences, *Chem. Phys. Lett.* **431** (2006) 375–379.

[14] M. Randić, Spectrum–like graphical representation of DNA based on codons, *Acta Chim. Slov.* **53** (2006) 477–485.

[15] Q. Dai, X. Liu, T. Wang, A novel graphical representation of DNA sequences and its application, *J. Mol. Graph. Model.* **25** (2006) 340–344.

[16] B. Liao, W. Zhu, Y. Liu, 3D graphical representation of DNA sequence without degeneracy and its applications in constructing phylogenic tree, *MATCH Commun. Math. Comput. Chem.* **56** (2006) 209–216.

[17] B. Liao, X. Xiang, W. Zhu, Coronavirus phylogeny based on 2D graphical representation of DNA sequence, *J. Comput. Chem.* **27** (2006) 1196–1202.

[18] Z. Cao, B. Liao, R. Li, A group of 3D graphical representation of DNA sequences based on dual nucleotides, *Int. J. Quant. Chem.* **108** (2008) 1485–1490.

[19] G. Huang, B. Liao, Y. Li, Z. Liu, H-L curve: A novel 2D graphical representation for DNA sequences, *Chem. Phys. Lett.* **462** (2008) 129–132.

[20] M. Randić, D. Plavšić, Novel spectral representation of RNA secondary structure without loss of information, *Chem. Phys. Lett.* **476** (2009) 277–280.

[21] Z. Liu, B. Liao, W. Zhu, G. Huang, A 2D graphical representation of DNA sequence based on dual nucleotides and its application, *Int. J. Quant. Chem.* **109** (2009) 948–958.

[22] D. Bielińska–Waż, Four–component spectral representation of DNA sequences, *J. Math. Chem.* **47** (2010) 41–51.

[23] Z. Zhang, T. Song, X. Zeng, Y. Niu, Y. Jiang, L. Pan, Y. Ye, ColorSquare: A colorful square visualization of DNA sequences, *MATCH Commun. Math. Comput. Chem.* **68** (2012) 621–637.

[24] B. Liao, Q. Xiang, L. Cai, Z. Cao, A new graphical coding of DNA sequence and its similarity calculation, *Physica A* **392** (2013) 4663–4667.

[25] N. Jafarzadeh, A. Iranmanesh, C-curve: a novel 3D graphical representation of DNA sequence based on codons, *Math. Biosci.* **241** (2013) 217–224.

[26] V. Aram, A. Iranmanesh, Z. Majid, Spider representation of DNA sequences, *J. Comput. Theor. Nanosci.* **11** (2014) 418–420.

[27] D. Bielińska–Waż, P. Waż, Spectral–dynamic representation of DNA sequences, *J. Biomed. Inform.* **72** (2017) 1–7.

[28] A. Nandy, M. Harle, S. C. Basak, Mathematical descriptors of DNA sequences: development and application, *Arkivoc* **ix** (2006) 211–238.

[29] D. Bielińska–Waż, Graphical and numerical representations of DNA sequences: Statistical aspects of similarity, *J. Math. Chem.* **49** (2011) 2345–2407.

[30] M. Randić, M. Novič, D. Plavšić, Milestones in graphical bioinformatics, *Int. J. Quant. Chem.* **113** (2013) 2413–2446.

[31] M.A. Gates, Simpler DNA sequence representations, *Nature* **316** (1985) 219–219.

[32] A. Nandy, A new graphical representation and analysis of DNA sequence structure. I: Methodology and application to globin genes, *Curr. Sci.* **66** (1994) 309–314.

[33] P.M. Leong, S. Morgenthaler, Random walk and gap plots of DNA sequences, *Comput. Appl. Biosci.* **11** (1995) 503–507.

[34] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.* **258** (1983) 1318–1327.

[35] M. Randić, M. Vračko, A, Nandy, S. C. Basak, On 3-D graphical representation of DNA primary sequences and their numerical characterization, *J. Chem. Inf. Comput. Sci.* **40** (2000) 1235–1244.

[36] Y. Yang, Y. Zhang, M. Jia, C. Li, L. Meng, Non-degenerate graphical representation of DNA sequences and its applications to phylogenetic analysis, *Comb. Chem. High Throughput Screen.* **16** (2013) 585–589.

[37] P. Waż, D. Bielińska–Waż, 3D-dynamic representation of DNA sequences, *J. Mol. Model.* **20** (2014) 2141.

[38] D. Bielińska–Waż, T. Clark, P. Waż, W. Nowak, A. Nandy, 2D-dynamic representation of DNA sequences, *Chem. Phys. Lett.* **442** (2007) 140–144.

[39] D. Bielińska–Waż, W. Nowak, P. Waż, A. Nandy, T. Clark, Distribution moments of 2D-graphs as descriptors of DNA sequences, *Chem. Phys. Lett.* **443** (2007) 408–413.

[40] D. Bielińska–Waż, P. Waż, T. Clark, Similarity studies of DNA sequences using genetic methods, *Chem. Phys. Lett.* **445** (2007) 68–73.

[41] P. Waż, D. Bielińska–Waż, A. Nandy, Descriptors of 2D-dynamic graphs as a classification tool of DNA sequences, *J. Math. Chem.* **52** (2014) 132–140.

[42] D. Panas, P. Waż, D. Bielińska–Waż, A. Nandy, S. C. Basak, 2D-Dynamic representation of DNA/RNA sequences as a characterization tool of the zika virus genome, *MATCH Commun. Math. Comput. Chem.* **77** (2017) 321–332.

[43] A. Nandy, S. Dey, S.C. Basak, Bielińska–Waż, P. Waż, Characterizing the zika virus genome – a bioinformatics study, *Curr. Comput. Aided Drug Des.* **12** (2016) 87–97.