International Journal of Peptide Research and Therapeutics (2022) 28:35
https://doi.org/10.1007/s10989-021-10345-2

VirVACPRED: A Web Server for Prediction of Protective Viral Antigens

Jesús Herrera-Bravo[1,2] · Jorge G. Farías[3] · Fernanda Parraguez Contreras[3] · Lisandra Herrera-Belén[3] · Juan-Alejandro Norambuena[3,4] · Jorge F. Beltrán[3]

Accepted: 7 December 2021

Abstract
Viral antigens are key in the development of vaccines that prevent or eradicate infections caused by these pathogens. Bioinformatics tools are modern alternatives that facilitate the discovery of viral antigens, reducing the costs of experimental assays. We developed a bioinformatics tool called VirVACPRED, which is highly efficient in predicting viral antigens. In this study, we obtained a model based on the gradient boosting classifier, which showed high performance during the training, leave-one-out cross-validation (accuracy = 0.7402, sensitivity = 0.7319, precision = 0.7503, F1 = 0.7251, kappa = 0.4774, Matthews correlation coefficient = 0.4981) and testing (accuracy = 0.8889, sensitivity = 1.0, precision = 0.8276, F1 = 0.9057, kappa = 0.7734, Matthews correlation coefficient = 0.7941). VirVACPRED is a robust tool that can be of great help in the search and proposal of new viral antigens, which can be considered in the development of future vaccines against infections caused by viruses.

Keywords Protective antigen · Vaccine · Bioinformatics · Virus · Machine learning · Server

Introduction

There is a considerable number of antiviral drugs against many viruses, and some of them do not eliminate infections but simply alter the clinical course of the disease (Huang et al. 2004). Antiviral vaccines have been the most successful alternatives in the prevention of epidemics, and it is for this reason that is necessary to exploit new technologies that identify critical antigens in order to induce a potent immune response (Graham 2013). Vaccination has allowed combating various infectious diseases mediated by viruses, like influenza, smallpox, varicella, diphtheria, polio, hepatitis, rotavirus, papillomavirus, among others (Graham 2013; Soria-Guerra et al. 2015). A vaccine is a molecular agent that induces specific protective immunity that triggers an enhanced adaptive immune response to reinfection by pathogens through the enhancement of immune memory (Pollard and Bijker 2021). Conventional vaccines are composed of attenuated or killed pathogens and they can take up to 15 years to develop. While it is true that these vaccines have saved many lives, they can also have adverse effects that could compromise the life of the patient (Bogdanos et al. 2001; Jarząb et al. 2013; Olson et al. 2001). The main component of vaccines are molecules called antigens, which are foreign to the immune system, and in turn, can have the ability to induce an immune response (Lahariya 2016).

Protective antigens are capable of inducing protection against a disease caused by an infectious agent after they are evaluated by means of an immunization scheme in an animal model. This approach to vaccine development includes several steps such as pathogen culturing, purifying the components (candidate antigens), and evaluating immunogenicity in an animal model ("An overview of biotechnology in vaccine development" 2020). Recombinant DNA and sequencing technology have led to a new concept within the field of vaccine development, where antigens capable of stimulating a specific

✉ Jorge F. Beltrán
beltran.lissabet.jf@gmail.com

1 Departamento de Ciencias Básicas, Facultad de Ciencias, Universidad Santo Tomas, Santiago, Chile

2 Center of Molecular Biology and Pharmacogenetics, Scientific and Technological Bioresource Nucleus, Universidad de La Frontera, Temuco, Chile

3 Department of Chemical Engineering, Faculty of Engineering and Science, Universidad de La Frontera, Ave. Francisco Salazar, 01145, Temuco, Chile

4 Program on Natural Resources Sciences, Universidad de La Frontera, Avenida Francisco Salazar, 01145, P.O. Box 54-D, 4780000 Temuco, Chile

Published online: 17 December 2021

immune response are identified (Brusic and Petrovsky 2005; Soria-Guerra et al. 2015; Tomar and De 2014, 2010). In recent years, RNA vaccines have been attracting increasing attention due to their ability to induce a safe and long-lasting immune response using in vivo models (Pardi et al. 2018; Zhang et al. 2019). RNA vaccines differ from traditional ones in that they do not administer live attenuated agents or fragments of it, eliminating the risk of causing the disease that is intended to be prevented. For the development of RNA vaccines, it is necessary to find the DNA sequences that encode essential antigens of the infectious agent and then transcribe them to obtain the corresponding RNA, which will be used as a vaccine (Brisse et al. 2020; Tombácz et al. 2021; Verbeke et al. 2019). However, like the traditional approach to vaccine development, the identification of candidate antigenic molecules is necessary.

The field of bioinformatics has allowed the acceleration and discovery of new vaccine candidates, through the large-scale prediction of different molecules that constitute potential protective antigens. Currently, there are many bioinformatic tools that predict antigenicity from a protein sequence, which is usually divided into small peptides called epitopes, which have the ability to induce an immune response mediated by T lymphocytes (Soria-Guerra et al. 2015). However, the tools that allow predicting whether a protein is antigenic or not are very scarce, with Vaxijen v2.0 being a widely cited tool and the only one of its kind to date (Doytchinova and Flower 2007). The Vaxijen v2.0 approach is extremely interesting and useful since it allows predicting antigenic proteins from various sources such as bacteria, viruses, and tumor cells. However, the viral antigen prediction model has not been updated for years. In consequence, taking into account the concept of Vaxijen v2.0, the main objective of the present work, was to develop an updated immunoinformatic tool for the robust and reliable prediction of viral antigens.

Materials and Methods

Dataset

The dataset used in this work was extracted from the publication of Vaxijen v2.0. This dataset is composed of 100 sequences of viral antigens referenced in the literature and 100 sequences identified as non-antigens (Doytchinova and Flower 2007), for a total of 200 sequences. This dataset was divided into training and testing datasets in a relationship of 80% and 20%, respectively (Fig. 1).

Antigen Feature Computation

To calculate the characteristics of the antigens, we use our script called AIDApy (Herrera-Bravo et al. 2021). AIDApy allows the calculation of 544 physicochemical and biochemical properties derived from the AAindex database. The AAindex database contains numerical indices that indicate different physicochemical and biological characteristics of amino acids and amino acid pairings (Kawashima et al. 2008). As mentioned above, in this study, all the indices contained in this database were calculated for the antigens and non-antigens by selecting the equation number (4) as shown below:

$a.a_n$: Total number of any amino acid of the 20 natural ones (1)

$a.a[AAindex]_n$: numerical value of each amino acid of the 20 natural ones in one of the 544 AAindex (2)

$$AAindexa.a_n = \frac{a.a_n \times a.a[AAindex]_n}{Sequence\ Length} \quad (3)$$

$$AAindex\ a.a_{antigen} = \frac{\sum AAindexa.a_n}{20} \quad (4)$$

Feature Selection

Usually, machine learning models that include many variables show low performance, for this reason reducing the dimensionality of the variables is a procedure that helps solve this problem (Mladenić 2006). For this reason, after calculating all AAindex characteristics (a total of 544), the best ten predictors were filtered and selected. For this purpose, the information gain function (Quinlan 1986) contained in the Orange3 3.28.0 library and written in Python 3 was used (Demšar et al. 2013).

Training, Cross-Validation, and Testing

The training, leave-one-out cross-validation (LOOCV) and testing, were carried out with the use of the open source PyCaret 2.3.1 (https://pypi.org/project/pycaret/) and scikit-learn 0.24.2 (https://pypi.org/project/scikit-learn/) libraries. PyCaret allows evaluation of several machine learning algorithms in an efficient and fast way, abstracting the functionalities of the popular Scikit-learn library on which it is based. A total of 16 machine learning algorithms were evaluated as shown below: random forest classifier (RF), extra trees classifier (ETC), quadratic discriminant analysis (QDA), light gradient boosting machine (LGBC), gradient boosting classifier (GBC), naive Bayes classifier (NBC), linear discriminant

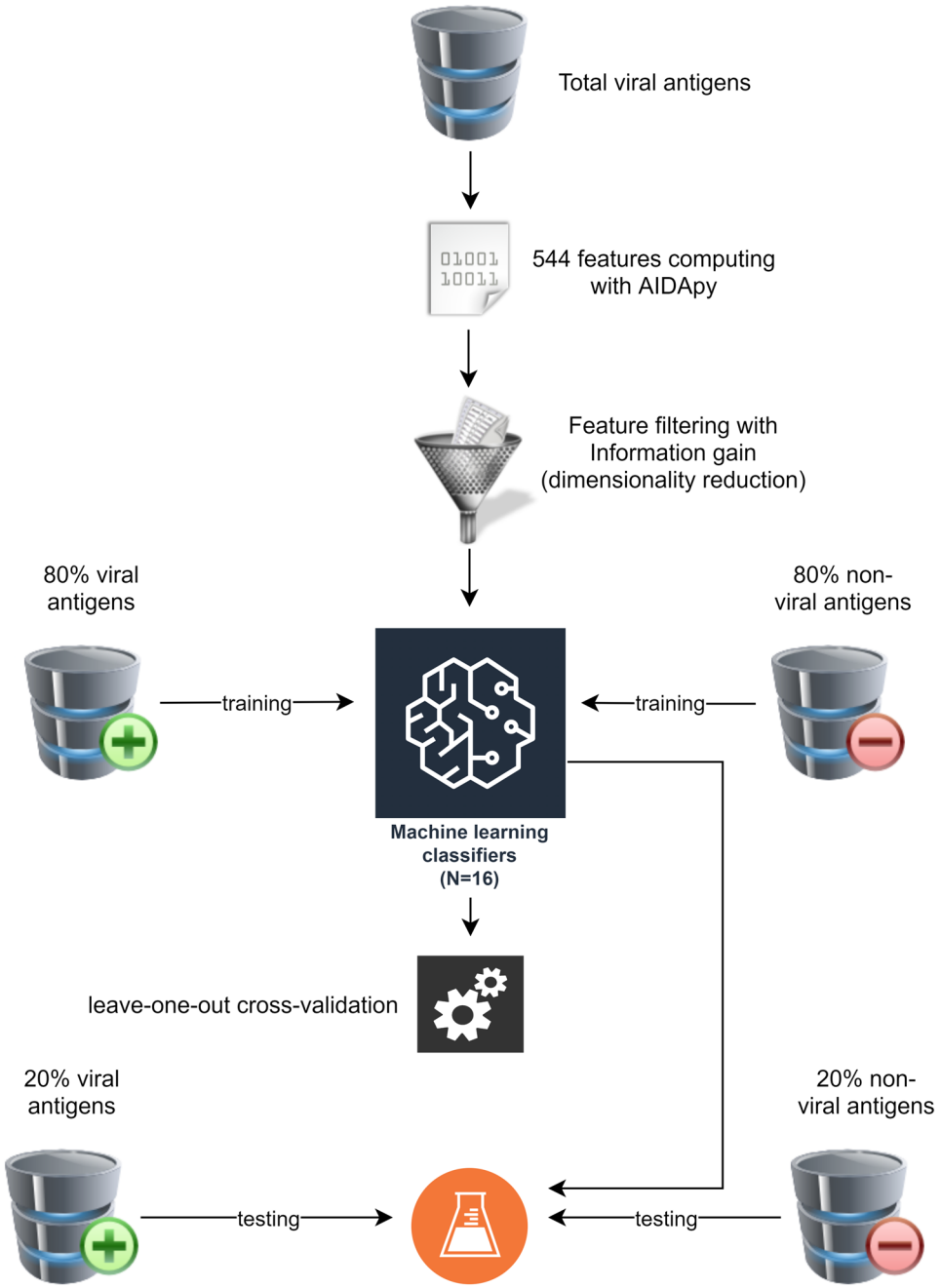

Fig. 1 The architecture used for the generation of the predictive models of protective viral antigens

analysis (LDA), ada boost classifier (ABC), K neighbors classifier (KNN), decision tree classifier (DTC), SVM-linear kernel (SVM-LK), logistic regression (LR), SVM-radial kernel (SVM-RK), Gaussian process classifier (GPC), MLP classifier (MLPC), and Ridge classifier (RC). For this study the PyCaret and scikit-learn library default parameter of all classifiers were used. The selection of the best classifier against the training, LOOCV, and testing phases was made based on the following metrics:

$$Sensitivity(TPR) = TP/(TP + FN) \quad (5)$$

$$Accuracy(ACC) = TP + TN/(TP + FP + FN + TN) \quad (6)$$

$$Precision(PVV) = TP/(TP + FP) \quad (7)$$

$$F1 = 2TP/(2TP + FP + FN) \quad (8)$$

$$MCC = (TP)(TN) - (FP)(FN)/\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)} \quad (9)$$

$$Kappa = p_0 - p_e/1 - p_e \quad (10)$$

All of the performance measures shown above have a range from zero to one (0–1). Models with measurements close to one are considered more reliable.

Results

The results of the information analysis allowed identifying the best predictors AAindex as shown below: AURR980113 (score: 0.207), FINA770101 (score: 0.191), QIAN880116 (score: 0.190), QIAN880102 (score: 0.183), KOEP990101 (score: 0.179), QIAN880133 (score: 0.174),

Table 1 Training performance measurements obtained during the LOOCV using 16 machine learning algorithms

Algorithms	ACC	AUC	TPR	PVV	F1	Kappa	MCC
RF	0.7471	0.8365	0.7611	0.7504	0.7437	0.4938	0.5086
ETC	0.7467	0.8485	0.7486	0.7507	0.7393	0.4927	0.5041
QDA	0.7412	0.8110	0.8278	0.7048	0.7543	0.4824	0.5010
LGBM	0.7405	0.8265	0.7333	0.7575	0.7311	0.4792	0.4962
GBC	0.7402	0.8045	0.7319	0.7503	0.7251	0.4774	0.4981
NBC	0.7173	0.8265	0.8750	0.6715	0.7542	0.4386	0.4710
LDA	0.7075	0.7827	0.7264	0.7012	0.7057	0.4142	0.4246
ABC	0.7069	0.7665	0.7375	0.7023	0.7028	0.4142	0.4339
KNN	0.7052	0.7792	0.7778	0.6821	0.7194	0.4104	0.4294
DTC	0.6775	0.6750	0.6389	0.7004	0.6560	0.3506	0.3582
SVM-LK	0.5144	0.0000	0.1222	0.2000	0.1048	0.0111	0.0243
LR	0.5088	0.7415	0.0000	0.0000	0.0000	0.0000	0.0000
SVM-RK	0.5088	0.2573	0.0000	0.0000	0.0000	0.0000	0.0000
GPC	0.5088	0.7427	0.0000	0.0000	0.0000	0.0000	0.0000
MLPC	0.5088	0.7591	0.0000	0.0000	0.0000	0.0000	0.0000
RC	0.5033	0.0000	0.0000	0.0000	0.0000	-0.0111	-0.0243

AUC area under the curve

Table 2 Performance measurements obtained during the testing phase (independent dataset) with the 16 machine learning algorithms assessed

Algorithms	ACC	AUC	TPR	PVV	F1	Kappa	MCC
RF	0.8667	0.9266	1.0	0.8000	0.8889	0.7273	0.7559
ETC	0.8444	0.9266	0.9583	0.7931	0.8679	0.6828	0.7010
QDA	0.7556	0.8492	0.8333	0.7407	0.7843	0.5045	0.5092
LGBM	0.8667	0.9315	0.9167	0.8462	0.8800	0.7305	0.7335
GBC	0.8889	0.9008	1.0	0.8276	0.9057	0.7734	0.7941
NBC	0.8000	0.8373	0.9583	0.7419	0.8364	0.5897	0.6222
LDA	0.8000	0.8611	0.8333	0.8000	0.8163	0.5970	0.5976
ABC	0.8667	0.8502	1.0	0.8000	0.8889	0.7273	0.7559
KNN	0.8222	0.8621	0.9167	0.7857	0.8462	0.6386	0.6492
DTC	0.8667	0.8661	0.8750	0.8750	0.8750	0.7321	0.7321
SVM-LK	0.5333	0.5000	1.0	0.5333	0.6957	0.0000	0.0000
LR	0.4667	0.7778	0.0000	0.0000	0.0000	0.0000	0.0000
SVM-RK	0.4667	0.2222	0.0000	0.0000	0.0000	0.0000	0.0000
GPC	0.4667	0.7778	0.0000	0.0000	0.0000	0.0000	0.0000
MLPC	0.4667	0.7560	0.0000	0.0000	0.0000	0.0000	0.0000
RC	0.4667	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

AUC area under the curve

LEVM780103 (score: 0.172), QIAN880113 (score: 0.170), SUEM840101 (score: 0.168), and RICJ880117 (score: 0.166), which were used for training with all the machine learning algorithms mentioned above. The LOOCV of the 16 evaluated algorithms allowed obtaining models with marked differences in performance measures, where the RFC, ETC, QDA, LGBM, GBC, NBC, LDA, ABC, KNN, and DTC algorithms, showed the best performance measures (Table 1).

On the other hand, an excellent performance was observed on the independent dataset (testing), where these measures increased considerably, which is indicative of robust prediction models (Table 2). It is important to highlight that the GBC algorithm presented the best performance measures during the testing phase, which allows its selection for the construction of a tool for the prediction of viral antigens (Table 2).

Taking into account the aforementioned aspects, we developed a web application called VirVACPRED, which includes the predictive model based on the gradient boosting classifier. This application was developed with the Python 3.9 programming language and the Flask framework, both open sources. VirVACPRED has a friendly and robust interface for the reliable and fast prediction of viral antigens, which is available at https://virvacpred.herokuapp.com/. VirVACPRED returns probability scores in the range of 0 and 1, where probability scores $>=0.5$ indicate that the input sequence is a viral antigen.

Discussion

During the past decade, viruses have emerged or re-emerged that have suddenly become major threats to humanity and the global economy, which was a concern regarding their epidemic transmission (Afrough et al. 2019; Trovato et al. 2020). Zoonoses such as Lassa fever, dengue fever, Middle East respiratory syndrome (MERS), swine flu, Ebola and Marburg hemorrhagic fevers, yellow fever, severe acute respiratory syndrome (SARS), West Nile fever, Zika, Chikungunya vector-borne diseases, and recently the coronavirus disease 2019 (COVID-19), are examples of the damage that viruses can cause in the world population (Trovato et al. 2020). In this sense, the development of innovative and technological platforms that allow the discovery of new drugs to prevent and combat viral infections is essential.

Bioinformatics has emerged as a powerful tool for solving different problems within the biological sciences, including the field of immunology (Soria-Guerra et al. 2015). Currently, there are dissimilar bioinformatics tools focused on the prediction of small linear peptides presented in the context of MHC. However, tools aimed at predicting the antigenicity of a complete protein are scarce. The prediction of the antigenicity of a protein is extremely important, considering that 90% of the epitopes processed by B cells are conformational and only 10% linear (Benjamin 1995; Huang and Honda 2006). Taking into account the aforementioned aspects, the need for tools that predict the global antigenicity of a protein is an important factor to take into consideration. Vaxijen v2.0 is a widely cited tool, it allows evaluating global antigenicity from an input amino acid sequence (Doytchinova and Flower 2007). However, since the development of Vaxijen v2.0 to date, there have been important advances in the field of machine learning, which could be used to improve the predictive capacity of viral antigens using the same approach as Vaxijen v2.0.

The results of the information gain analysis showed that the ten best predictive AAindexes are related to characteristics of secondary protein structures such as helix, beta-sheet, alpha-helix, coil, beta-turn, and helix-coil. In this sense, we suggest that future tools focused on predicting antigenicity take these structural properties into account. In fact, it has been reported that the secondary structure of viral antigens is key to the development of an immune response mediated by T lymphocytes (Gairin and Oldstone 1993).

In this work, the GBC presented the best performance measures in the classification of viral antigens during the training and testing phase. This classifier has been successfully used in the development of predictive models in the area of bioinformatics, such as the prediction of submitochondrial localization (Yu et al. 2020), DNA-binding residue (Deng et al. 2018), gene-expression data analysis (Blagus and Lusa 2015), prediction of the interaction between target and ligand (Xuan et al. 2019), diagnostic classification of cancers (Ma et al. 2020), and prediction of RNA-protein interactions (Jain et al. 2018), among others. However, other classifiers such as ETC, QDA, LBGM, GBC, NBC, LDC, ABC, KNN, DTC, and RF, also presented good performance measures in both phases. It is important to highlight that the performance measures obtained with GBC, even outperforming the RF classifier, the latter very popular and widely used in the field of bioinformatics (Beltrán Lissabet et al. 2019a, b; Jorge Félix Beltrán Lissabet et al. 2019a, b; Boulesteix et al. 2012; Herrera-Bravo et al. 2021). For this reason, as mentioned above, the GBC was selected to develop the VirVACPRED tool.

In this work, we make a comparison of VirVACPRED with the performance measures reported by Vaxijen v2.0. In this comparison, it was observed that both tools present a similar performance during training. However, VirVACPRED presented a better performance over the independent dataset (Table 3 and Fig. 2), due to the high-performance measures obtained, demonstrating its high efficiency in the prediction of viral antigens.

Table 3 Comparison of the Vaxijen v2.0 and VirVACPRED performance measures

Tool	Phase	ACC	AUC	TPR	PVV	F1	Kappa	MCC
VirVACPRED	LOOCV	0.7402	0.8045	0.7319	0.7503	0.7251	0.4774	0.4981
Vaxijen v2.0 Doytchinova and Flower (2007)	LOOCV	0.73	0.810	0.74	0.71	–	–	–
VirVACPRED	Testing	0.8889	0.9008	1.0	0.8276	0.9057	0.7734	0.7941
Vaxijen v2.0 Doytchinova and Flower (2007)	Testing	0.70	0.743	0.84	–	–	–	–

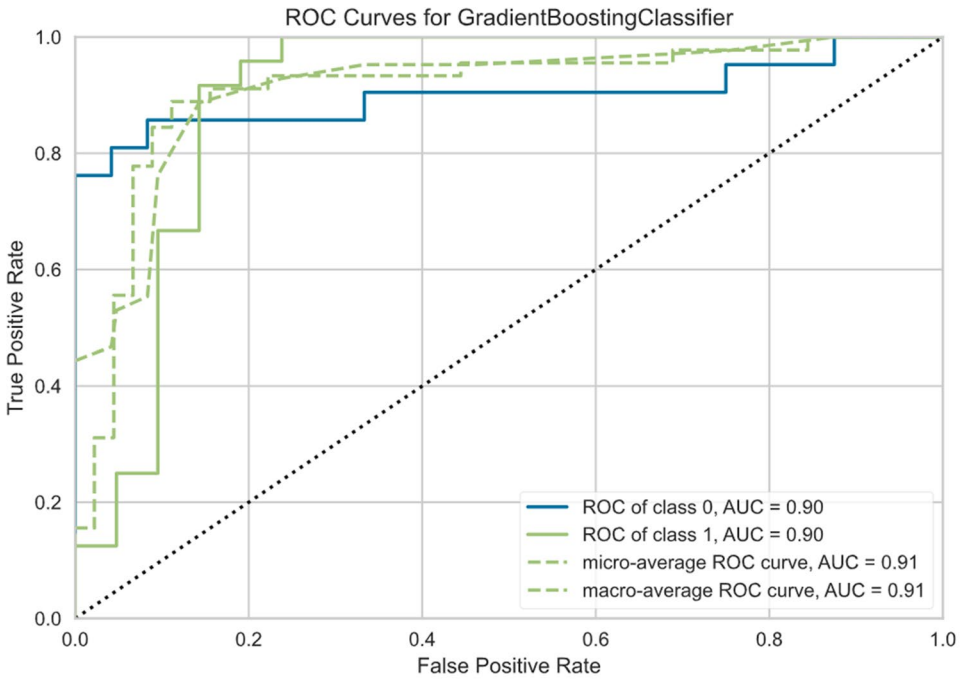

Fig. 2 Receiver operating characteristic curves of the gradient boosting classifier on the independent dataset. This classifier shows an AUC value of 0.90 on the unseen data (testing data), which is an indicative of a good model for prediction of the viral antigen and non-viral antigen classes represented by zero and one, respectively

As mentioned above, the datasets used to train and test VirVACPRED consisted of antigenic and non-antigenic protein sequences in monomeric states (primary sequence), obtained from different virus species (Doytchinova and Flower 2007). Consequently, we recommend that users make predictions using the viral primary sequences as input. VirVACPRED is a tool that has a friendly interface, which unlike Vaxijen v2.0, can process multiple protein sequences in FASTA format. We believe that VirVACPRED can be very useful in the discovery of new protective viral antigens, which could be considered in the formulation of future vaccines to prevent future epidemics. The tool is freely available at https://virvacpred.herokuapp.com/. This tool has a simple user interface for amino acid sequence processing (Fig. 3).

Fig. 3 User interface of the VirVACPRED tool for prediction of protective viral antigens. **A** Input and **B** result interfaces

Conclusions

The discovery of viral antigens plays a key role in the development of vaccines that allow the prevention of viral infections. Vaxijen v2.0 and VirVACPRED are the only tools of their kind, which allow predicting the global antigenicity of a protein. VirVACPRED is an updated tool that allows predicting viral antigens with high efficiency according to the performance measures obtained in the training and testing phases. The present server is limited to processing no more than 1000 protein sequences per prediction. We believe that VirVACPRED can be of great help in the discovery of new viral antigens, which will allow the development of future vaccines that prevent the risk of infections caused by viruses.

References

Afrough B, Dowall S, Hewson R (2019) Emerging viruses and current strategies for vaccine intervention. Clin Exp Immunol. https://doi.org/10.1111/cei.13295

Beltrán Lissabet JF, Belén LH, Farias JG (2019) AntiVPP 1.0: a portable tool for prediction of antiviral peptides. Comput Biol Med 107:127–130. https://doi.org/10.1016/j.compbiomed.2019.02.011

Beltrán Lissabet JF, Herrera Belén L, Farias JG (2019) TTAgP 1.0: a computational tool for the specific prediction of tumor T cell antigens. Comput Biol Chem 83:107103. https://doi.org/10.1016/j.compbiolchem.2019.107103

Benjamin DC (1995) B-cell epitopes: fact and fiction. Advances in experimental medicine and biology. Springer, Boston, pp 95–108

Blagus R, Lusa L (2015) Boosting for high-dimensional two-class prediction. BMC Bioinform. https://doi.org/10.1186/s12859-015-0723-9

Bogdanos DP, Choudhuri K, Vergani D (2001) Molecular mimicry and autoimmune liver disease: virtuous intentions, malign consequences. Liver. https://doi.org/10.1034/j.1600-0676.2001.021004225.x

Boulesteix AL, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Min Knowl Discov 2:493–507. https://doi.org/10.1002/widm.1072

Brisse M, Vrba SM, Kirk N, Liang Y, Ly H (2020) Emerging concepts and technologies in vaccine development. Front Immunol. https://doi.org/10.3389/fimmu.2020.583077

Brusic V, Petrovsky N (2005) Immunoinformatics and its relevance to understanding human immune disease. Expert Rev Clin Immunol 1:145–157. https://doi.org/10.1586/1744666x.1.1.145

Demšar J, Erjavec A, Hočevar T, Milutinovič M, Možina M, Toplak M, Umek L, Zbontar J, Zupan B (2013) Orange: data mining toolbox in python Tomaž Curk Matija Polajnar Laň Zagar. J Mach Learn Res 14:2349–2353

Deng L, Pan J, Xu X, Yang W, Liu C, Liu H (2018) PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine. BMC Bioinform. https://doi.org/10.1186/s12859-018-2527-1

Doytchinova IA, Flower DR (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinform. https://doi.org/10.1186/1471-2105-8-4

Gairin JE, Oldstone MB (1993) Virus and cytotoxic T lymphocytes: crucial role of viral peptide secondary structure in major histocompatibility complex class I interactions. J Virol 67:2903–2907. https://doi.org/10.1128/jvi.67.5.2903-2907.1993

Graham BS (2013) Advances in antiviral vaccine development. Immunol Rev. https://doi.org/10.1111/imr.12098

Herrera-Bravo J, Herrera Belén L, Farias JG, Beltrán JF (2021) TAP 1.0: a robust immunoinformatic tool for the prediction of tumor T-cell antigens based on AAindex properties. Comput Biol Chem 91:1052. https://doi.org/10.1016/j.compbiolchem.2021.107452

Huang J, Honda W (2006) CED: a conformational epitope database. BMC Immunol. https://doi.org/10.1186/1471-2172-7-7

Huang DB, Wu JJ, Tyring SK (2004) A review of licensed viral vaccines, some of their safety concerns, and the advances in the development of investigational viral vaccines. J Infect. https://doi.org/10.1016/j.jinf.2004.05.018

Jain DS, Gupte SR, Aduri R (2018) A data driven model for predicting RNA-protein interactions based on gradient boosting machine. Sci Rep. https://doi.org/10.1038/s41598-018-27814-2

Jarząb A, Skowicki M, Witkowska D (2013) Subunit vaccines—antigens, carriers, conjugation methods and the role of adjuvants. Postepy Hig Med Dosw 67:1128–1143. https://doi.org/10.5604/17322693.1077807

Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. https://doi.org/10.1093/nar/gkm998

Lahariya C (2016) Vaccine epidemiology: a review. J Fam Med Prim Care 5:7. https://doi.org/10.4103/2249-4863.184616

Levine MM, Levine MM, Dougan G, Kaper JB, Good MF, Liu MA, Nabel GJ, Rappuoli R, Nataro JP (2004) An overview of biotechnology in vaccine development. New generation vaccines. CRC Press, Boca Raton, pp 38–51

Ma B, Meng F, Yan G, Yan H, Chai B, Song F (2020) Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2020.103761

Mladenić D (2006) Feature selection for dimensionality reduction. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Springer, Berlin, pp 84–102

Olson JK, Croxford JL, Calenoff MA, Dal Canto MC, Miller SD (2001) A virus-induced molecular mimicry model of multiple sclerosis. J Clin Invest 108:311–318. https://doi.org/10.1172/jci13032

Pardi N, Hogan MJ, Porter FW, Weissman D (2018) mRNA vaccines-a new era in vaccinology. Nat Rev Drug Discov. https://doi.org/10.1038/nrd.2017.243

Pollard AJ, Bijker EM (2021) A guide to vaccinology: from basic principles to new developments. Nat Rev Immunol. https://doi.org/10.1038/s41577-020-00479-7

Quinlan JR (1986) Induction of decision trees. Mach Learn. https://doi.org/10.1023/A:1022643204877

Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S (2015) An overview of bioinformatics tools for epitope prediction: implications on vaccine development. J Biomed Inform. https://doi.org/10.1016/j.jbi.2014.11.003

Tomar N, De RK (2010) Immunoinformatics: an integrated scenario. Immunology. https://doi.org/10.1111/j.1365-2567.2010.03330.x

Tomar N, De RK (2014) Immunoinformatics: a brief review. Methods Mol Biol. https://doi.org/10.1007/978-1-4939-1115-8_3

Tombácz I, Weissman D, Pardi N (2021) Vaccination with messenger RNA: a promising alternative to DNA vaccination. Methods in molecular biology. Springer, New York, pp 13–31

Trovato M, Sartorius R, D'Apice L, Manco R, De Berardinis P (2020) Viral emerging diseases: challenges in developing vaccination strategies. Front Immunol. https://doi.org/10.3389/fimmu.2020.02130

Verbeke R, Lentacker I, De Smedt SC, Dewitte H (2019) Three decades of messenger RNA vaccine development. Nano Today. https://doi.org/10.1016/j.nantod.2019.100766

Xuan P, Sun C, Zhang T, Ye Y, Shen T, Dong Y (2019) Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. Front Genet. https://doi.org/10.3389/fgene.2019.00459

Yu B, Qiu W, Chen C, Ma A, Jiang J, Zhou H, Ma Q (2020) SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. Bioinformatics 36:1074–1081. https://doi.org/10.1093/bioinformatics/btz734

Zhang C, Maruggi G, Shan H, Li J (2019) Advances in mRNA vaccines for infectious diseases. Front Immunol. https://doi.org/10.3389/fimmu.2019.00594