Computational Biology and Chemistry 91 (2021) 107452

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

TAP 1.0: A robust immunoinformatic tool for the prediction of tumor T-cell antigens based on AAindex properties

Jesús Herrera-Bravo [a,b], Lisandra Herrera Belén [c], Jorge G. Farias [c], Jorge F. Beltrán [c,*]

[a] Departamento de Ciencias Básicas, Facultad de Ciencias, Universidad Santo Tomas, Chile
[b] Center of Molecular Biology and Pharmacogenetics, Scientific and Technological Bioresource Nucleus, Universidad de La Frontera, Chile
[c] Universidad de La Frontera, Department of Chemical Engineering, Faculty of Engineering and Science, Ave. Francisco Salazar 01145, Temuco, Chile

ARTICLE INFO

Keywords:
Tumor
Antigen
Prediction
Machine learning
T-cell
Peptide

ABSTRACT

Immunotherapy is a research area with great potential in drug discovery for cancer treatment. Because of the capacity of tumor antigens to activate the immune response and promote the destruction of tumor cells, they are considered excellent immunotherapeutic drugs. In this work, we evaluated fifteen machine learning algorithms for the classification of tumor antigens. For this purpose, we build robust datasets, carefully selected from the TANTIGEN and IEDB databases. The feature computation of all antigens in this study was performed by developing a script written in Python 3.8, which allowed the calculation of 544 physicochemical and biochemical properties extracted from the AAindex database. All classifiers were subjected to the training, 10-fold cross-validation, and testing on an independent dataset. The results of this study showed that the quadratic discriminant classifier presented the best performance measures over the independent dataset, accuracy = 0.7384, AUC = 0.817, recall = 0.676, precision = 0.7857, F1 = 0.713, kappa = 0.4764, and Matthews correlation coefficient = 0.4834, outperforming common machine learning classifiers used in the bioinformatics area. We believe that our prediction model could be of great importance in the field of cancer immunotherapy for the search of potential tumor antigens. Taking all aspects mentioned before, we developed an immunoinformatic tool called TAP 1.0 with a friendly interface for tumor antigens prediction, available at https://tapredictor.herokuapp.com/.

1. Introduction

Tumor antigens (TAs) have great potential to be considered excellent immunotherapeutic molecules (Olsen et al., 2014). Tumor immunotherapy is a field of research showing rapid growth by exploiting T-cells tumor-associated antigens to induce rejection of tumors (Ilyas and Yang, 2015). The goal of tumor immunotherapy is to activate the adaptive immune response to destroy tumors or prevent their occurrence, where the key effector are cytotoxic T lymphocyte, which recognizes and destroy tumor cells (Geiger and Sun, 2016; Hodi et al., 2010; Robert et al., 2011; Topalian et al., 2012). TAs are classified into two groups, tumor-specific antigens, which are only present on tumor cells, and tumor-associated antigens that are overexpressed on tumor cells (Boon et al., 1994; Olsen et al., 2017). Currently, a wide variety of immunoinformatics tools have been developed for the prediction of immunogenic epitopes, both in the context of MHC class I (MHC-I) and II (MHC-II), for example, NetMHC 4.0 (Andreatta and Nielsen, 2016; Nielsen et al., 2003), NetMHCIIpan 4.0 (Reynisson et al., 2020), MixMHC2pred (Racle et al., 2019), among other summarized in an excellent review by Soria-Guerra et al. (Soria-Guerra et al., 2015). However, none of these tools was developed specifically for the prediction of antitumor epitopes (tumor antigens).

To date, the two most recent immunoinformatic tools for predicting antigenic peptides presented by MHC-I molecules with antitumor activity are TTAgP 1.0 (Beltrán Lissabet et al., 2019) and iTTCA-Hybrid (Charoenkwan et al., 2020). However, it is important to highlight some drawbacks of both tools. TTAgP 1.0 is a tool based on the random forest algorithm, and it is currently under license by the University of La Frontera, and part of the dataset used to train the algorithm is private. In addition, for the execution of this software, a series of steps must be followed that for the common user not related to the area of computing can become tedious. Other drawbacks of this tool are its limited peptide processing capacity since it can only process one peptide at a time. Finally, this software lacks a score function for ranking predictions, which is very useful in the filtering process against a high number of peptides.

* Corresponding author.
E-mail address: j.beltran07@ufromail.cl (J.F. Beltrán).

https://doi.org/10.1016/j.compbiolchem.2021.107452
Received 30 November 2020; Received in revised form 4 February 2021; Accepted 4 February 2021
Available online 8 February 2021

On the other hand, iTTCA-Hybrid is based on the support vector machine and random forest algorithms for the prediction of tumor antigens. Like TTAgP 1.0, this tool has several details to take into account. The iTTCA-Hybrid authors propose that the prediction model of this program presents better performance than the one reported by TTAgP 1.0, making a comparison on its own negative dataset and probably very different from the one used by TTAgP 1.0, which is not correct. On the other hand, it is not clear whether the selection of the peptides that make up the negative dataset includes peptides presented by MHC-I and/or MHC-II molecules, which is an important detail to take into account because the positive dataset selected it is based on peptides expressed in the context of MHC class I, and a mixture of peptides (MHC-I and MHC-II) could affect the reliability of the predictive model.

In terms of peptide processing capacity, iTTCA-Hybrid has advantages over TTAgP 1.0, because it can process several sequences at the same time in FASTA format. However, in practice, this iTTCA-Hybrid processing approach is also not very useful, because when predictions of immunogenic antigens presented by MHC-I and MHC-II molecules are made, these are carried out from of complete protein sequences (Andreatta and Nielsen, 2016; Reynisson et al., 2020), which are divided into small fragments known as *k*-mers (Manekar and Sathe, 2018).

In general, the peptides presented by MHC-I molecules are in the range of 8 to 11-mers (Jojic et al., 2006; Soria-Guerra et al., 2015; Wieczorek et al., 2017; Zhao and Sher, 2018) and to a lesser extent longer. (Zhao and Sher, 2018). In this sense, the current immunogenic peptide prediction programs presented by MHC-I molecules divide the input protein sequences into small peptides in the range of 9 to 14-mers, thus, allowing the complete evaluation of immunogenic sites in a protein under study. (Soria-Guerra et al., 2015). Both TTAgP 1.0 and iTTCA-Hybrid lack this functionality. Taking into account all the aforementioned aspects, the present work aims to develop an immunoinformatics tool for the prediction of tumor antigens presented by MHC-I molecules, which allows us to carry out the predictions from protein sequences with good reliability.

2. Materials and methods

2.1. Datasets

For this study the TANTIGEN database (http://cvc.dfci.harvard.edu/tadb/), was selected. This database contains human tumor antigens presented by MHC-I and MHC-II molecules that elicit *in vivo* and *in vitro* T-cell response (Olsen et al., 2017). A total of 592 unique antigens were selected as the positive dataset, which was divided into a training dataset (80 % = 442 tumor antigen) and an independent dataset (20 % = 111 tumor antigen). For the construction of the negative dataset, the IEDB database was selected (https://www.iedb.org/) (Vita et al., 2019), and the selection of the non-tumor antigens was performed using a rigorous approach, establishing different search parameters within this database as shown below: (1) antigens validated by *in vitro* and *in vivo* assays, (2) antigens presented by MHC-I molecules, (3) antigens with a length between 9-mer and 14-mer, (4) *Homo sapiens* as antigen source and host, and (5) antigen discard with associated terms such as 'tumor', 'cancer', 'carcinoma', and 'metastasis'. The resulting negative dataset consisted of 592 unique antigens, which were divided into a training dataset (80 % = 442 non-tumor antigen) and an independent dataset (20 % = 111 non-tumor antigen) (see datasets in: https://github.com/jfbldevs/AIDApy). The diagram of the models training, cross-validations, and testing are shown in Fig. 1.

2.2. Peptide features

The features of all antigens were calculated from 544 properties extracted from the AAindex database. This database is composed of amino acid indices and amino acid mutation matrices. In the amino acid index, a group of 20 numerical values represents various

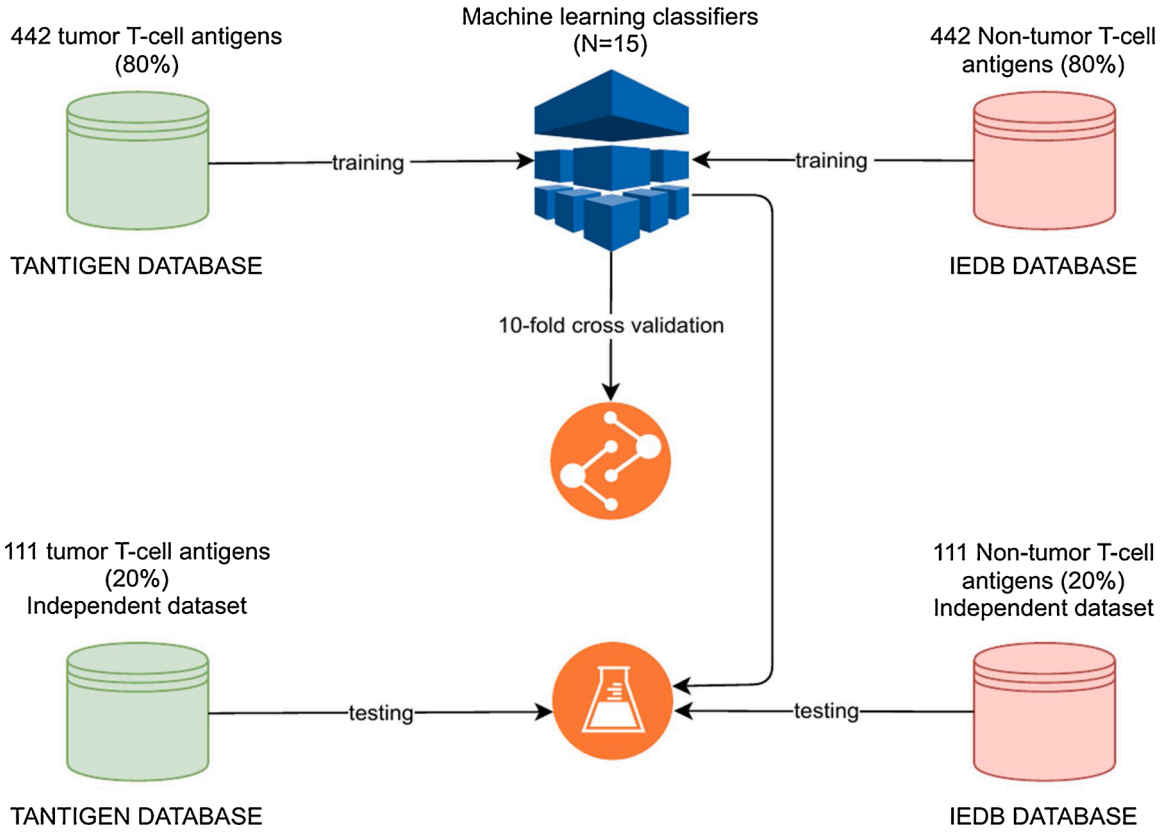

Fig. 1. Architecture of the training, cross-validation and testing phases.

physicochemical and biochemical properties of amino acids (Kawashima and Kanehisa, 2000). For this work, 544 features were extracted from the AAindex database using a script written in Python 3.8 (https://www.python.org/).

2.3. AAindex feature computation

In the $AAindex_{property}$ calculation, aa_i is the number of a specific amino acid of any of the 20 natural amino acids in an antigen, $AAindex_i$ is the numerical value of each amino acid of any of the 20 natural amino acids available in the AAindex database.

$$\text{AAindexproperty} = \sum_{i=1}^{20} (\text{aa}i * \text{AAindex}i) \Big/ N$$

2.4. Feature selection

The information gain (Quinlan, 1986) was used to select the feature more relevant. The number of features was reduced from 544 to 10, which were selected for the training and testing phases. The information gain was computed using the Orange3 3.26.0 library (Demšar et al., 2013) (https://pypi.org/project/Orange3/).

2.5. Training and testing

In this work, fifteen machine learning algorithms were evaluated as shown below: Logistic Regression (LR), k-nearest neighbors (KNN), Naives Bayes (NB), Decision Tree (DT), SVM – Radial Kernel (RBFSVM), Gaussian Process Classifier (GPC), Multi-Level Perceptron (MLP), Random Forest Classifier (RF), Quadratic Discriminant Analysis (QDA), Ada Boost Classifier (ADA), Gradient Boosting Classifier (GBC), Linear Discriminant Analysis (LDA), Extra Trees Classifier (ET), Extreme Gradient Boosting (XGBOOST), and Light Gradient Boosting (LIGHTGBM). The training, cross-validation 10-fold (10-fold CV) and testing, were carried out with the use of the PyCaret 2.0 library (https://pypi.org/project/pycaret/2.0/). The receiver operating characteristic (ROC) is used to evaluate the performance of binary classifiers, by determining the area under the curve (AUC), where values in the range of 0.5–1.0 are indicative of models with good performance (Obuchowski et al., 2005). Moreover, the measures shown below were also used to evaluate the performance of the classifiers:

Accuracy = (TP + TN) / (TP + FP + FN + TN)

F1 = 2 * (Precision * Sensitivity) / (Precision + Sensitivity)

Recall = TP / (TP + FN)

Precision = TP / (TP + FP)

MCC = TP * TN – FP * FN / sqrt((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))

Kappa = p_0 - p_e / 1 - p_e

TP = true positives, TN = true negatives, FP = false positives and FN = false negatives. The Matthews Correlation Coefficient (MCC) ranges from −1 to 1, a higher MCC means a better prediction (Boughorbel et al., 2017).

3. Results

The selection of the ten best predictive features were made by means of the information gain analysis, and the following AAindex properties were obtained: GRAR740102(score = 0.132), WOEC73010(score = 0.127), KIDA85010(score = 0.126), QIAN88018(score = 0.123), FAUJ830101 (score = 0.121), PONP800103(score = 0.118), LAWE840101(score = 0.112), MIYS990105 (score = 0.112), EISD860103(score = 0.111), and ARGP820102(score = 0.106), which were selected for the training of the fifteen classification algorithms included in this study.

As mentioned above, both the training and the testing phase of all algorithms were carried out with the PyCaret 2.0 library. For this purpose, the split function of this library was used to divide the total dataset, into a training dataset (80 %) and an independent dataset (20 %). During the 10-fold CV of the fifteen machine learning algorithms, it was observed that in general all algorithms showed a good performance in tumor antigens classification (Table 1).

It is important to highlight that the QDA classifier presented the best balance in its performance measures, even outperforming common classifiers in the bioinformatics area such as the MLP, RF, and RBFSVM (Table 1). However, when all the models were evaluated on the independent dataset, it was observed that most of the algorithms tend to overfit, with the exception of the QDA classifier, which showed an increase in its performance measures, indicating correct learning (Table 2 and Fig. 2).

4. Discussion

The field of immunotherapy has promoted a new area of research within bioinformatics called immunoinformatic, whose objective is the development and application of *in silico* tools that allow the reduction of the high costs of *in vitro* assays and rapid decision-making in the discovery of immunotherapeutic drugs (Backert and Kohlbacher, 2015; Raoufi et al., 2020). In this work, 544 chemical-physical properties extracted from the AAindex database were calculated for a total of 1184 tumor and non-tumor antigens. For this calculation, we developed a script written in Python 3.8 called $AIDA^{py}$, which is available at https://github.com/jfbldevs/AIDApy.

Of the top ten best features determined using the information gain analysis, it was observed that 60 % of these features (KIDA850101, FAUJ830101, PONP800103, LAWE840101, EISD860103, and ARGP820102) are related to hydrophobicity, which is in correspondence with what is reported in the literature regarding the importance of this property as a key feature in antigens presented by MHC-I molecules (Chowell et al., 2015; Huang et al., 2011), as is the case of the antigens present in all our datasets. In fact, in a study carried out by Chowell et al., an artificial hydrophobicity-based neural network was developed for the prediction of immunogenic epitopes presented by MHC-I molecules with excellent results, thus, highlighting the importance of this property for antigen classifications (Chowell et al., 2015).

One of the main problems in health informatics, bioinformatics, and computational biology projects are the problems related to model overfitting, which lead to bad practices in the development and selection of predictive models (Chicco, 2017). A model is considered to be overfitted when the performance measures obtained during the training phase are superior to the testing phase (Hawkins, 2004). In this study, it was observed that most of the classifiers evaluated in our work tend to the overfitting when their performance was assessed against the independent dataset (testing phase) (Table 2), in consequence, we highlight the importance of taking this aspect into account when working with this type of dataset in future studies. However, it was observed that the QDA classifier, in addition to presenting the best performance measures during training, did not show a tendency to the overfitting, since it outperformed all the performance measures evaluated on the independent dataset, showing its excellent potential in the classification of tumor antigens. The importance of the QDA classifier has been present in many investigations in the area of bioinformatics such as the classification of antimicrobial peptides (Chen and Luo, 2009; Feng et al., 2019), prediction of toxins (Lin and Li, 2007), DNA sequence motif recognition (Zhang, 2000), determination of the expression genes level (Arevalillo and Navarro, 2011), among others. The results of our study show that this classifier can also be used in the classification of tumor antigens.

Table 1
Performance measures obtained during the 10-fold cross-validation.

Classifier	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
QDA	0.7171	0.7804	0.6755	0.7386	0.7047	0.4343	0.4369
LDA	0.7011	0.7646	0.6689	0.7158	0.6905	0.4023	0.4041
LR	0.6969	0.7617	0.6816	0.7056	0.6921	0.3938	0.3956
NB	0.6948	0.7593	0.6184	0.7371	0.6697	0.3896	0.3974
GBC	0.6833	0.7567	0.6246	0.7113	0.6626	0.3665	0.3713
ET	0.6789	0.7554	0.6098	0.7147	0.6549	0.3579	0.3648
ADA	0.6748	0.7432	0.6269	0.6940	0.6568	0.3498	0.3525
RF	0.6738	0.7350	0.5909	0.7110	0.6427	0.3477	0.3546
LIGHTGBM	0.6727	0.7467	0.6328	0.6881	0.6588	0.3453	0.3468
KNN	0.6674	0.7131	0.6222	0.6868	0.6513	0.3347	0.3377
XGBOOST	0.6642	0.7325	0.6392	0.6741	0.6548	0.3282	0.3297
DT	0.6093	0.6088	0.5907	0.6114	0.5995	0.2186	0.2194
MLP	0.7064	0.7612	0.6500	0.7357	0.6885	0.4129	0.4174
RBFSVM	0.6947	0.7550	0.6012	0.7452	0.6642	0.3895	0.3985
GPC	0.6979	0.7629	0.6203	0.7399	0.6732	0.3958	0.4028

AUC: *area under the curve*.

Table 2
Performance measures obtained on the independent dataset.

Classifier	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
QDA	0.7384	0.817	0.676	0.7857	0.713	0.4764	0.4834
LDA	0.73	0.7832	0.6356	0.7812	0.7009	0.4595	0.4676
LR	0.7215	0.7683	0.6695	0.7453	0.7054	0.4428	0.4451
NB	0.6962	0.756	0.5763	0.7556	0.6538	0.3918	0.4032
GBC	0.7004	0.7646	0.5932	0.7527	0.6635	0.4003	0.4095
ET	0.6835	0.758	0.5932	0.7216	0.6512	0.3666	0.3725
ADA	0.7004	0.7659	0.6017	0.7474	0.6667	0.4003	0.4081
RF	0.6498	0.7214	0.5424	0.6882	0.6066	0.2989	0.3058
LIGHTGBM	0.6667	0.7513	0.5678	0.7053	0.6291	0.3328	0.3392
KNN	0.6667	0.7146	0.6102	0.6857	0.6457	0.333	0.335
XGBOOST	0.6962	0.777	0.6356	0.7212	0.6757	0.3921	0.3949
DT	0.6287	0.6286	0.6186	0.6293	0.6239	0.2573	0.2574
MLP	0.7046	0.7796	0.5678	0.7791	0.6569	0.4086	0.4244
RBFSVM	0.6835	0.7768	0.5339	0.759	0.6269	0.3663	0.3834
GPC	0.7089	0.7686	0.5932	0.7692	0.6699	0.4172	0.4284

AUC: *area under the curve*.

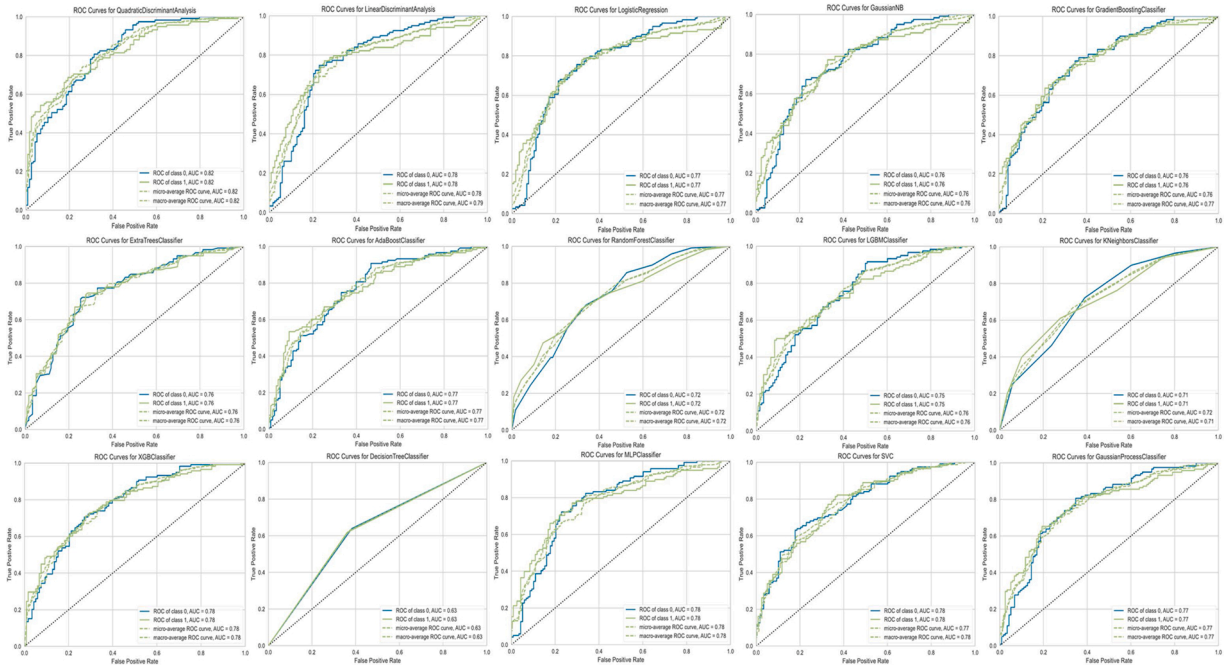

Fig. 2. ROCs curves obtained during the assessment of the fifteen classifiers over the independent dataset.

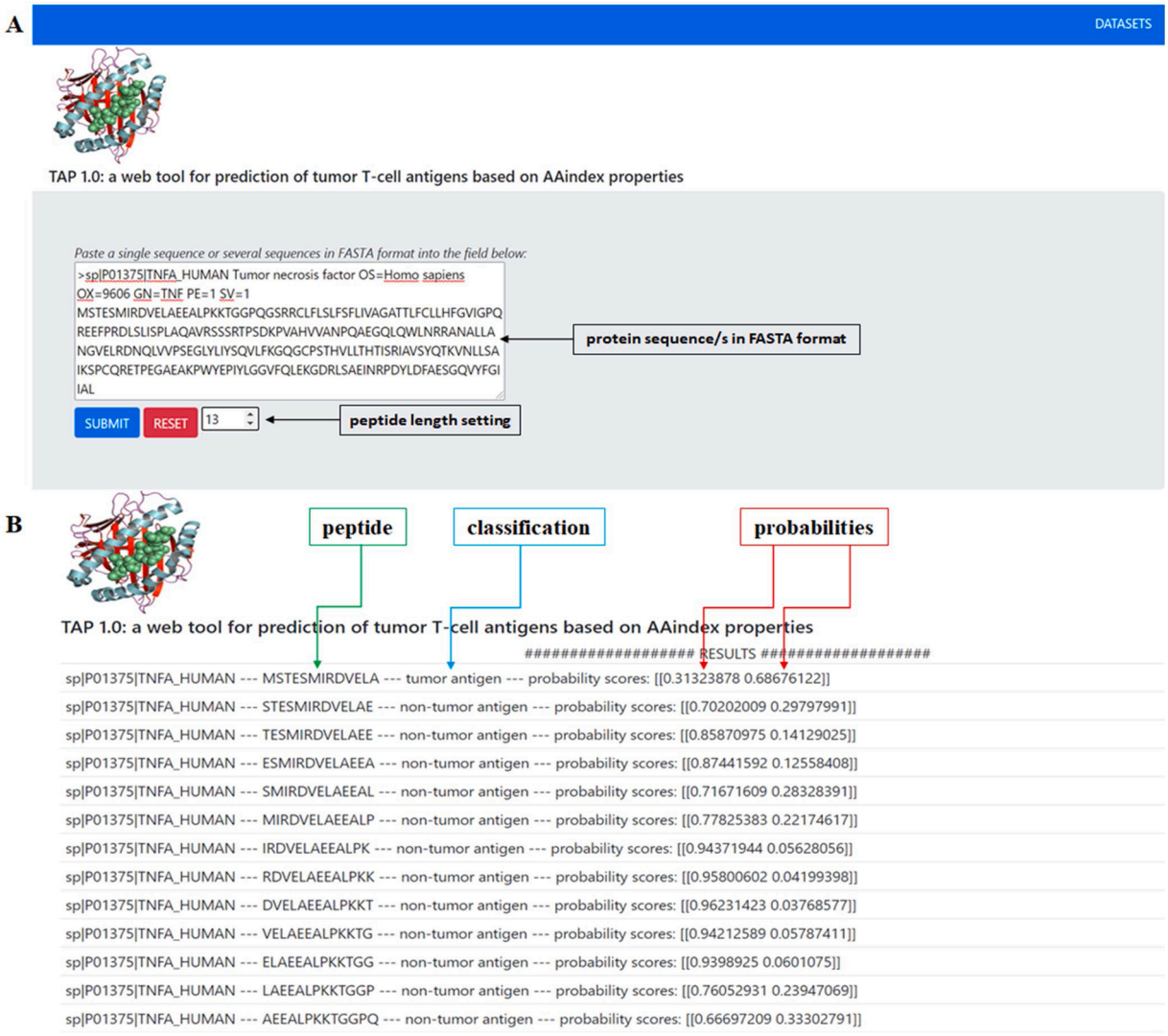

Fig. 3. TAP 1.0 user interface for tumor antigen predictions. (A) Interface for data entry (amino acid sequence/s in FASTA format). (B) Results interface.

In this work, we developed a web tool called Tumor Antigen Predictor (TAP 1.0) (Fig. 3), which constitutes a more practical alternative to our first tool, TTAgP 1.0 (Beltrán Lissabet et al., 2019) and iTTCA-Hybrid, the latter developed by Charoenkwan et al. (Charoenkwan et al., 2020). As mentioned above, the advantages of TAP 1.0 are the use of a robust negative dataset, carefully selected from the IEDB database, and its ability to scan entire protein sequences for searching tumor antigens. In this work, a comparison with the TTAgP 1.0 and iTTCA-Hybrid programs was not made, because both tools were trained over a negative dataset different from the one used by TAP 1.0, and this comparison would not be correct. However, in this work, we propose and highlight the importance of using a robust negative dataset like the one used in this study for the development of future prediction models with this objective.

TAP 1.0 has a user-friendly interface, which allows the search of tumor antigens with different lengths given to one or more protein sequences, this tool is currently available at: https://tapredictor.herokuapp.com/. We believe that TAP 1.0 can be a robust alternative in the field of immunotherapy for searching tumor antigens with potential use in the discovery of immunotherapeutic drugs against cancer.

5. Conclusions

Tumor antigens are excellent drugs in the field of cancer immunotherapeutic. TAP 1.0 is a quadratic discriminant analysis-based tool with excellent performance for the prediction of tumor antigens presented by MHC-I. This tool has a friendly interface that allows the search of tumor antigens from protein sequences of any length. We believe that TAP 1.0 could have a significant impact on the discovery of new tumor antigens for use in tumor immunotherapy.

CRediT authorship contribution statement

Jesús Herrera-Bravo: Conceptualization, Investigation, Writing - review & editing. **Lisandra Herrera Belén:** Investigation, Writing - review & editing. **Jorge G. Farias:** Investigation, Writing - review & editing. **Jorge F. Beltrán:** Investigation, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors report no declarations of interest.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.compbiolchem.2021.107452.

References

Andreatta, M., Nielsen, M., 2016. Gapped sequence alignment using artificial neural networks: application to the MHC class i system. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv639.

Arevalillo, J.M., Navarro, H., 2011. A new method for identifying bivariate differential expression in high dimensional microarray data using quadratic discriminant analysis. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-12-S12-S6.

Backert, L., Kohlbacher, O., 2015. Immunoinformatics and epitope prediction in the age of genomic medicine. Genome Med. https://doi.org/10.1186/s13073-015-0245-0.

Beltrán Lissabet, J.F., Herrera Belén, L., Farias, J.G., 2019. TTAgP 1.0: a computational tool for the specific prediction of tumor T cell antigens. Comput. Biol. Chem. https://doi.org/10.1016/j.compbiolchem.2019.107103.

Boon, T., Cerottini, J.C., Van Den Eynde, B., Van Der Bruggen, P., Van Pel, A., 1994. Tumor antigens recognized by T lymphocytes. Annu. Rev. Immunol. https://doi.org/10.1146/annurev.iy.12.040194.002005.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One. https://doi.org/10.1371/journal.pone.0177678.

Charoenkwan, P., Nantasenamat, C., Hasan, M.M., Shoombuatong, W., 2020. iTTCA-Hybrid: improved and robust identification of tumor T cell antigens by utilizing hybrid feature representation. Anal. Biochem. https://doi.org/10.1016/j.ab.2020.113747.

Chen, W., Luo, L., 2009. Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. J. Microbiol. Methods. https://doi.org/10.1016/j.mimet.2009.03.013.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData Min. https://doi.org/10.1186/s13040-017-0155-3.

Chowell, D., Krishna, S., Becker, P.D., Cocita, C., Shu, J., Tan, X., Greenberg, P.D., Klavinskis, L.S., Blattman, J.N., Anderson, K.S., 2015. TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. Proc. Natl. Acad. Sci. U. S. A. https://doi.org/10.1073/pnas.1500973112.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B., 2013. Orange: data mining toolbox in python. J. Mach. Learn. Res.

Feng, P., Wang, Z., Yu, X., 2019. Predicting antimicrobial peptides by using increment of diversity with quadratic discriminant analysis method. IEEE/ACM Trans. Comput. Biol. Bioinform. https://doi.org/10.1109/TCBB.2017.2669302.

Geiger, T.L., Sun, J.C., 2016. Development and matura. In: Geiger, T.L., Sun, J.C. (Eds.), Development and Maturation of Natural Killer Cells. Current Opinion in Immunology, 39, pp. 82–89. https://doi.org/10.1016/j.coi.2016.01.007.

Hawkins, D.M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. https://doi.org/10.1021/ci0342472.

Hodi, F.S., O'Day, S.J., McDermott, D.F., Weber, R.W., Sosman, J.A., Haanen, J.B., Gonzalez, R., Robert, C., Schadendorf, D., Hassel, J.C., Akerley, W., Van Den Eertwegh, A.J.M., Lutzky, J., Lorigan, P., Vaubel, J.M., Linette, G.P., Hogg, D., Ottensmeier, C.H., Lebbé, C., Peschel, C., Quirt, I., Clark, J.I., Wolchok, J.D., Weber, J.S., Tian, J., Yellin, M.J., Nichol, G.M., Hoos, A., Urba, W.J., 2010. Improved survival with ipilimumab in patients with metastatic melanoma. N. Engl. J. Med. https://doi.org/10.1056/NEJMoa1003466.

Huang, L., Kuhls, M.C., Eisenlohr, L.C., 2011. Hydrophobicity as a driver of MHC class i antigen processing. EMBO J. https://doi.org/10.1038/emboj.2011.62.

Ilyas, S., Yang, J.C., 2015. Landscape of tumor antigens in t cell immunotherapy. J. Immunol. https://doi.org/10.4049/jimmunol.1501657.

Jojic, N., Reyes-Gomez, M., Heckerman, D., Kadie, C., Schueler-Furman, O., 2006. Learning MHC I - peptide binding. Bioinformatics. https://doi.org/10.1093/bioinformatics/btl255.

Kawashima, S., Kanehisa, M., 2000. AAindex: amino acid index database. Nucleic Acids Res. https://doi.org/10.1093/nar/28.1.374.

Lin, H., Li, Q.Z., 2007. Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem. Biophys. Res. Commun. https://doi.org/10.1016/j.bbrc.2007.01.011.

Manekar, S.C., Sathe, S.R., 2018. A benchmark study of k-mer counting methods for high-throughput sequencing. Gigascience. https://doi.org/10.1093/gigascience/giy125.

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S.L., Lamberth, K., Buus, S., Brunak, S., Lund, O., 2003. Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. Protein Sci. https://doi.org/10.1110/ps.0239403.

Obuchowski, N.A., Blackmore, C.C., Karlik, S., Reinhold, C., 2005. Fundamentals of clinical research for radiologists. Am. J. Roentgenol. https://doi.org/10.2214/ajr.184.2.01840364.

Olsen, L.R., Campos, B., Winther, O., Sgroi, D.C., Karger, B.L., Brusic, V., 2014. Tumor antigens as proteogenomic biomarkers in invasive ductal carcinomas. BMC Med. Genomics. https://doi.org/10.1186/1755-8794-7-S3-S2.

Olsen, L.R., Tongchusak, S., Lin, H., Reinherz, E.L., Brusic, V., Zhang, G.L., 2017. TANTIGEN: a comprehensive database of tumor T cell antigens. Cancer Immunol. Immunother. https://doi.org/10.1007/s00262-017-1978-y.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. https://doi.org/10.1023/A:1022643204877.

Racle, J., Michaux, J., Rockinger, G.A., Arnaud, M., Bobisse, S., Chong, C., Guillaume, P., Coukos, G., Harari, A., Jandus, C., Bassani-Sternberg, M., Gfeller, D., 2019. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. Nat. Biotechnol. https://doi.org/10.1038/s41587-019-0289-6.

Raoufi, E., Hemmati, M., Eftekhari, S., Khaksaran, K., Mahmodi, Z., Farajollahi, M.M., Mohsenzadegan, M., 2020. Epitope prediction by novel immunoinformatics approach: a state-of-the-art review. Int. J. Pept. Res. Ther. https://doi.org/10.1007/s10989-019-09918-z.

Reynisson, B., Barra, C., Kaabinejadian, S., Hildebrand, W.H., Peters, B., Peters, B., Nielsen, M., Nielsen, M., 2020. Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. J. Proteome Res. https://doi.org/10.1021/acs.jproteome.9b00874.

Robert, C., Thomas, L., Bondarenko, I., O'Day, S., Weber, J., Garbe, C., Lebbe, C., Baurain, J.F., Testori, A., Grob, J.J., Davidson, N., Richards, J., Maio, M., Hauschild, A., Miller, W.H., Gascon, P., Lotem, M., Harmankaya, K., Ibrahim, R., Francis, S., Chen, T.T., Humphrey, R., Hoos, A., Wolchok, J.D., 2011. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. N. Engl. J. Med. https://doi.org/10.1056/NEJMoa1104621.

Soria-Guerra, R.E., Nieto-Gomez, R., Govea-Alonso, D.O., 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.

Topalian, S.L., Hodi, F.S., Brahmer, J.R., Gettinger, S.N., Smith, D.C., McDermott, D.F., Powderly, J.D., Carvajal, R.D., Sosman, J.A., Atkins, M.B., Leming, P.D., Spigel, D.R., Antonia, S.J., Horn, L., Drake, C.G., Pardoll, D.M., Chen, L., Sharfman, W.H., Anders, R.A., Taube, J.M., McMiller, T.L., Xu, H., Korman, A.J., Jure-Kunkel, M., Agrawal, S., McDonald, D., Kollia, G.D., Gupta, A., Wigginton, J.M., Sznol, M., 2012. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N. Engl. J. Med. https://doi.org/10.1056/NEJMoa1200690.

Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D. K., Sette, A., Peters, B., 2019. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. https://doi.org/10.1093/nar/gky1006.

Wieczorek, M., Abualrous, E.T., Sticht, J., Álvaro-Benito, M., Stolzenberg, S., Noé, F., Freund, C., 2017. Major histocompatibility complex (MHC) class I and MHC class II proteins: conformational plasticity in antigen presentation. Front. Immunol. https://doi.org/10.3389/fimmu.2017.00292.

Zhang, M.Q., 2000. Discriminant analysis and its application in DNA sequence motif recognition. Brief. Bioinform. https://doi.org/10.1093/bib/1.4.331.

Zhao, W., Sher, X., 2018. Systematically benchmarking peptide-MHC binding predictors: from synthetic to naturally processed epitopes. PLoS Comput. Biol. https://doi.org/10.1371/journal.pcbi.1006457.