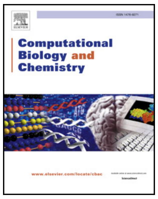
Contents lists available at ScienceDirect

Computational Biology and Chemistry

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



Research Article

TTAgP 1.0: A computational tool for the specific prediction of tumor T cell antigens

Jorge Félix Beltrán Lissabet, Lisandra Herrera Belén, Jorge G. Farias*

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

ARTICLE INFO

Keywords:
Immunoinformatics
Peptide
Prediction
Cancer
Machine learning
MHC

ABSTRACT

Nowadays, cancer is considered a global pandemic and millions of people die every year because this disease remains a challenge for the world scientific community. Even with the efforts made to combat it, there is a growing need to discover and design new drugs and vaccines. Among these alternatives, antitumor peptides are a promising therapeutic solution to reduce the incidence of deaths caused by cancer. In the present study, we developed TTAgP, an accurate bioinformatic tool that uses the random forest algorithm for antitumor peptide predictions, which are presented in the context of MHC class I. The predictive model of TTAgP was trained and validated based on several features of 922 peptides. During the model validation we achieved sensitivity = 0.89, specificity = 0.92, accuracy = 0.90 and the Matthews correlation coefficient = 0.79 performance measures, which are indicative of a robust model. TTAgP is a fast, accurate and intuitive software focused on the prediction of tumor T cell antigens.

1. Introduction

When the Greek physician Hippocrates first used the terms *carcinoma* and *carcinos* to refer to ulcer-forming tumors and non-ulcer forming tumors (Doytchinova and Flower, 2018), he could not even imagine that this group of diseases would become the second cause of death worldwide. According to World Health Organization data, a total of 9.6 million deaths due to cancer were reported in 2018 (Bray et al., 2018).

Many tumors are immunogenic because they have antigenic determinants, as peptides and proteins, able to induce an adaptive type of immune response mediated by T cells (Blankenstein et al., 2012). However, because spontaneous T cell responses to tumors are not entirely efficient in eliminating them, researchers are trying to develop immunotherapies, such as anti-tumor vaccines (Kumai et al., 2017; Obara et al., 2018; Pol et al., 2015; Wada et al., 2016), to enable the immune system of cancer patients to destroy tumors. Tumor-associated peptide vaccines, once inside the bloodstream, bind with MHC (Major Histocompatibility Complex) molecules expressed in the antigen-presenting cells. This complex of the antigenic peptide bound to the MHC migrates to the cell surface and is recognized by the T cell receptors, triggering their activation (Aranda et al., 2013; Klausen et al., 2018; Reche et al., 2015). Lately, effective antitumor immunity in humans has been associated with the presence of T cells recognizing cancer neoantigens (Schumacher and Schreiber, 2015; Wirth and Kühnel, 2017). A very interesting fact is that neoantigens are not present in normal cells; therefore, they are therapeutically specific targets and minimize the risk of autoimmunity and immune tolerance (Hacohen et al., 2013; Saini et al., 2017).

Bioinformatic and immunoinformatic approaches focusing on the immune system have played a crucial role in peptide vaccine discovery (Kazi et al., 2018; Knutson et al., 2001; Roberts et al., 2006; Soria-Guerra et al., 2015). To date, there are only two tools to predict human tumor antigens: VaxiJen v2.0 and TIminer. VaxiJen v2.0 is the first tool developed for the prediction of human tumor antigens that uses physicochemical properties for proteins through an auto cross-covariance transformation model (Doytchinova and Flower, 2007). By contrast, TIminer is an integrative tool that types human leukocyte antigens, predicts neoantigens, characterizes immune infiltrates and predicts tumor immunogencity. Its prediction model of human tumor antigens is based mainly on the NetMHCpan program, which uses an artificial neural network for predictions (Tappeiner et al., 2017).

Cancer cells express tumor antigens that are recognized and killed by T cells in the MHC class I and II context. For this reason, T cells play a key role in the rejection of tumors and have been used effectively in the field of immunotherapeutic cancer (Restifo et al., 2012; Gill and June, 2015). In this study, we developed a portable immunoinformatics tool for the prediction of human tumor antigens recognized by T cells in

* Corresponding author.
E-mail address: jorge.farias@ufrontera.cl (J.G. Farias).

https://doi.org/10.1016/j.compbiolchem.2019.107103
Received 18 February 2019; Received in revised form 20 June 2019; Accepted 10 August 2019
Available online 12 August 2019

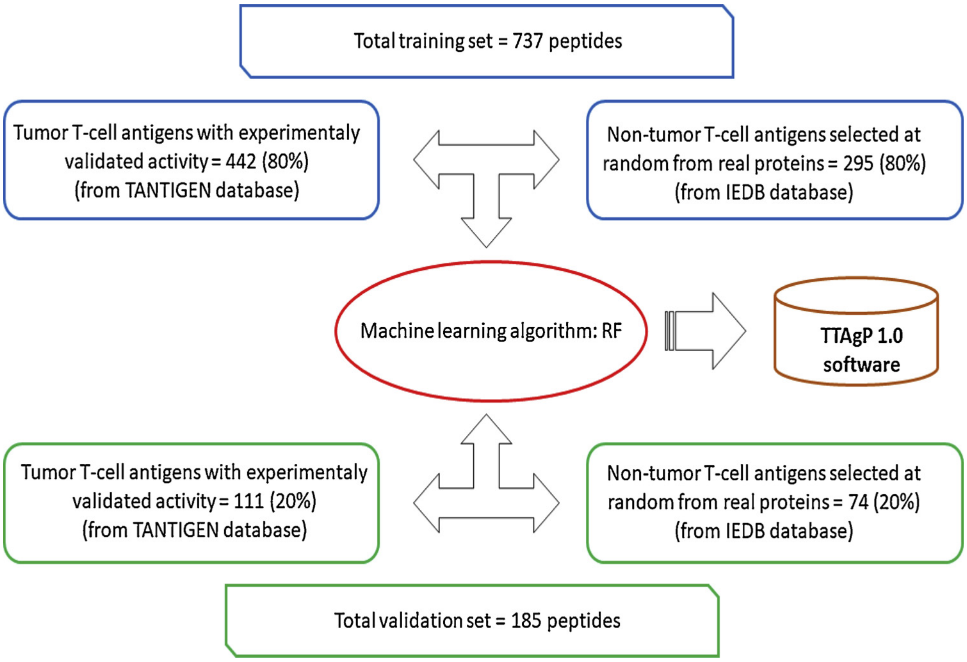

Total training set = 737 peptides
Tumor T-cell antigens with experimentaly validated activity = 442 (80%) (from TANTIGEN database)
Non-tumor T-cell antigens selected at random from real proteins = 295 (80%) (from IEDB database)
Machine learning algorithm: RF
TTAgP 1.0 software
Tumor T-cell antigens with experimentaly validated activity = 111 (20%) (from TANTIGEN database)
Non-tumor T-cell antigens selected at random from real proteins = 74 (20%) (from IEDB database)
Total validation set = 185 peptides

Fig. 1. The architecture of the model training and validation. A total dataset of 737 and 185 peptides for training and validation were selected, respectively.

Table 1
Comparison among current programs for human tumor antigen predictions.

Program	Performance measurements			
	TPR	SPC	ACC	MCC
TTAgP 1.0	0.89	0.92	0.90	0.79
VaxiJen v2.0	0.4306	0.3514	0.4037	−0.2066
TIminer	0.863	0.94	0.843	0.74

TPR: sensitivity, **SPC:** specificity, **ACC:** accuracy, **MCC:** Matthews correlation coefficient.

the MHC class I context, based on a predictive model built with the random forest algorithm.

2. Material and methods

2.1. Datasets

For the collection of antigens, we selected the TANTIGEN database (http://cvc.dfci.harvard.edu/tadb/), which contains human tumor antigens reported to elicit either *in vivo* or *in vitro* T cell response (Olsen et al., 2017). For this study, a total of 686 MHC class I peptides was selected, which were reduced to 553 unique peptides when identical peptides were considered one since the TANTIGEN repository contains promiscuous peptides, which are reported to bind to different alleles. For the training and validation of our model, the dataset was divided as follows: 442 (80%) for training and 111 (20%) for validation. The non-tumor T cell antigens were randomly selected from the IEDB database (https://www.iedb.org/) (Vita et al., 2018), which were divided as follows: 295 (80%) for training and 74 (20%) for validation. The architecture of the model training and validation is shown in Fig. 1.

2.2. Peptide features

For this study, the following features were assessed: net charge (Klein, et al., 1984), number of hydrogen bond donors (FAUCHÈRE, et al., 1988), and hydropathy index (Kyte and Doolittle, 1982), composition of charged (DEKHR), aliphatic (ILV), aromatic (FHWY), polar (DERKQN), neutral (AGHPSTY), hydrophobic (CVLIMFW), positively charged (HKR), negatively charged (DE), tiny (ACDGST), small (EHILKMNPQV), large (FRWY) residues and the relative frequency of all 20 natural amino acids. All features were calculated by using the Python 3.7 programming language available at https://www.python.org/.

2.3. Relative frequency (Rfre) of all 20 natural amino acids

Rfre [a.a] = Xi/N where Rfre [a.a] is the relative frequency of a type i natural amino acid. N is the total number of natural amino acids in the peptide (peptide length).

2.4. Amino acid composition of peptides (PEP [comp])

Ex: PEP [positively charged] = Rfre [H] + Rfre [K] + Rfre [R] where PEP [comp] is the sum of all Rfre [a.a] in a peptide.

2.5. Training and validation

The random forest algorithm (RF) was evaluated for model prediction building. The model training was performed using the Python 3.7 programming language and the Anaconda 3 package (available at https://www.anaconda.com). The 'score' function (accuracy) of the Anaconda 3 package was implemented to choose models with scores > 0.95 for subsequent validations. The 'score' function measures the accuracy of probabilistic predictions and ranges from 0 to 1. For model validations the following equations were used:

$$\text{Sensitivity (TPR)} = TP/(\ TP + FN)$$

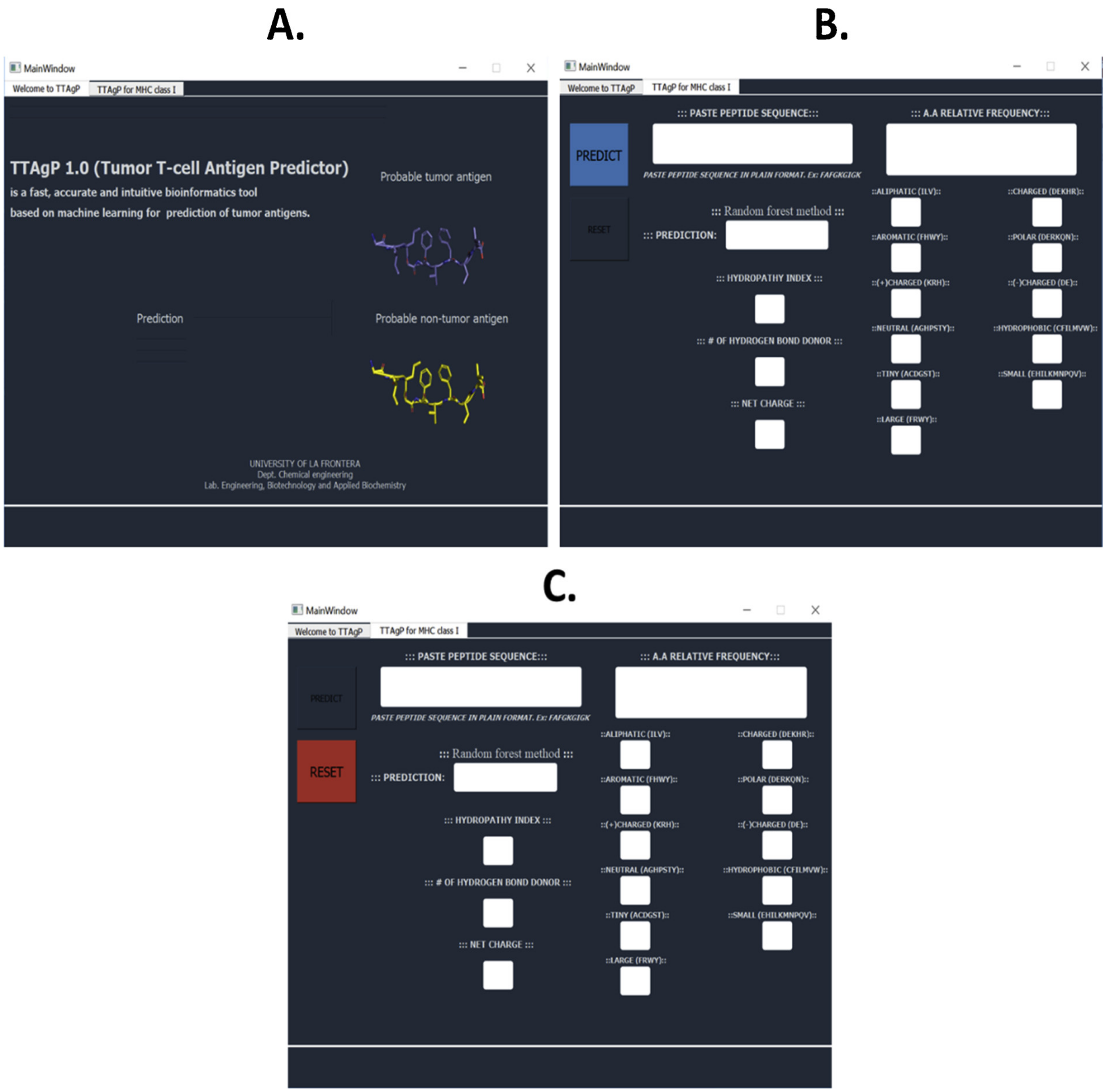
$$\text{Specificity (SPC)} = TN/(\ TN + FP)$$

$$\text{Accuracy (ACC)} = TP + TN/(\ TP + FP + FN + TN)$$

where TP represents the true positives, TN the true negatives, FP the false positives and FN the false negatives. For model validation, in addition to the equations mentioned above, the Matthews correlation coefficient (MCC) was calculated:

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

MCC is used to evaluate the performance of the predictor. Its value



TTAgP 1.0 (Tumor T-cell Antigen Predictor)

is a fast, accurate and intuitive bioinformatics tool

based on machine learning for prediction of tumor antigens.

UNIVERSITY OF LA FRONTERA
Dept. Chemical engineering
Lab. Engineering, Biotechnology and Applied Biochemistry

PASTE PEPTIDE SEQUENCE IN PLAIN FORMAT. Ex: FAFGKGIGK

Fig. 2. TTAgP interface. (A) Initial interface. (B) Bottom PREDICT (blue) to perform predictions. (C) Bottom RESET (red) to clean the fields.

ranges from -1 to 1 and a higher MCC means a better prediction (Boughorbel et al., 2017).

2.6. Software

For the development of our software, called TTAgP (Tumor T cell Antigen Predictor), the programming language Python 3.7 and the framework PyQt5 were used. Qt is a set of cross-platform C++ libraries that implement high-level APIs for accessing many aspects of modern desktop and mobile systems. PyQt5 is a comprehensive set of Python bindings for Qtv5 (https://pypi.org/project/PyQt5/). TTAgP has a friendly and intuitive interface, which, in addition predicting tumor T cell antigens, can also be used to calculate different physicochemical characteristics of peptides. The software as well as the instructions to run it are available at https://github.com/bio-coding/TTAgP.

3. Results and discussion

Immune-based approaches have garnered a lot of interest in the efficient treatment of tumors. Targeted immunotherapy against cancer include a triad of key approaches: monoclonal antibodies, immune checkpoint inhibitors and vaccines(Riley et al., 2019; Seliger, 2019). The immune response has two arms: the humoral, or antibody-mediated, and the cellular, mediated primarily by T cells (Doytchinova and Flower, 2018). Given that CD8+ and CD4 + T cells play a significant role in tumor rejection (Knocke et al., 2016), the strategies that are most frequently used to search for human tumor antigens are based on the use of immunoinformatics tools that predict peptides presented by the MHC class I and II complexes. Examples of these are: NetMHC4.0 (Andreatta and Nielsen, 2015), NetMHCpan4.0 (Jurtz et al., 2017), IEDB consensus (Zhang et al., 2008) and SYFPEITHI (Rammensee et al., 1999), among others. The training and validation of these tools, however, involve the mixing of immunogenic epitopes derived from different sources (bacteria, viruses and others) (Soria-Guerra et al., 2015), which could lead to erroneous results when making specific predictions like those about human tumor antigens.

The bioinformatics arena is progressively relying on machine learning algorithms to conduct predictive analytics and gain greater insights into the complex biological processes of the human body

(Olson et al., 2017). Among all the machine learning algorithms, the random forest algorithm is one of the best in the field of bioinformatics due to its high predictive capacity (Boulesteix et al., 2012; Couronné et al., 2018; Qi, 2012), and it has been widely used in the search for peptides and proteins with different biological activities (Kandaswamy et al., 2011; Chen et al., 2015; Chang and Yang, 2013; Zhang et al., 2011; Bhadra et al., 2018). Based on the relevance of this algorithm, we proceeded to evaluate it in the prediction of tumor antigens by taking several of the predictive characteristics most used in different studies on predictions of antimicrobial (Bhadra et al., 2018; Meher et al., 2017) and antiviral activities (Chang and Yang, 2013; Thakur et al., 2012; Lissabet et al., 2019).

During the training phase, we obtained a prediction model with a score = 0.9946. This model showed the highest performance measures during the testing phase: TPR = 0.89, SPC = 0.92, ACC = 0.90 and MCC = 0.79, which are indicative of an excellent model.

In this work, we compared the performance measures of TTAgP with the VaxiJen and TIminer programs using our validation dataset (Table 1).

In this comparison, TTAgP was found to have a higher overall performance than the other programs, demonstrating its high predictive capacity for human tumor antigens. All this proves the quality of TTAgP as a tool for the study and discovery of new human tumor antigens. TIminer also showed high-performance measures; however, it only outperforms TTAgP in terms of SPC. On the other hand, VaxiJen v2.0 showed a low performance with respect to the evaluated programs.

Based on our RF model, we developed an intuitive software for making predictions of tumor antigens called TTAgP (Fig. 2).

TTAgP returns two types of results: 'Probable: [True]' for positive cases and 'Probable: [False]' for negative cases. Additionally, it calculates and returns all the features used by the program for the prediction of tumor antigens. It is valid to highlight that TTAgP is the first immunoinformatics tool based on machine learning that uses a database of purely tumor antigens with experimentally verified *in vitro* and *in vivo* activity for the training and validation of its predictive model.

We believe that TTAgP could be a very useful alternative in the search for and development of new human tumor antigens, given that it is a tool that has a high predictive capacity and a friendly interface for any user.

4. Conclusion

TTAgP is an immunoinformatics tool based on the random forest algorithm for predicting human tumor antigens of MHC class I molecules. This tool presents an excellent balance in its performance measures, making it a useful alternative in the search for human tumor antigens. TTAgP is characterized by its portability, easy handling and predictive power. We believe that this tool can have an impact on the development of drugs in immunotherapy against cancer.

Notes

TTAgP 1.0 is protected by copyright. This software is free for academic users. For commercial purposes contact: jorge.farias@ufrontera.cl

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Acknowledgments

This work was supported by the following projects: DI12-PEO1DIUFRO and DIE14-0001 DIUFRO of the Universidad de La Frontera, Chile.

References

Doytchinova, I.A., Flower, D.R., 2018. BMC Immunol. 19, 11.
Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. CA Cancer J. Clin. 68, 394–424.
Blankenstein, T., Coulie, P.G., Gilboa, E., Jaffee, E.M., 2012. Nat. Rev. Cancer 12, 307.
Kumai, T., Kobayashi, H., Harabuchi, Y., Celis, E., 2017. Curr. Opin. Immunol. 45, 1–7.
Obara, W., Kanehira, M., Katagiri, T., Kato, R., Kato, Y., Takata, R., 2018. Cancer Sci. 109, 550–559.
Pol, J., Bloy, N., Buqué, A., Eggermont, A., Cremer, I., Sautes-Fridman, C., Galon, J., Tartour, E., Zitvogel, L., Kroemer, G., 2015. Oncoimmunology 4, e974411.
Wada, S., Yada, E., Ohtake, J., Fujimoto, Y., Uchiyama, H., Yoshida, S., Sasada, T., 2016. Immunotherapy 8, 1321–1333.
Aranda, F., Vacchelli, E., Eggermont, A., Galon, J., Sautès-Fridman, C., Tartour, E., Zitvogel, L., Kroemer, G., Galluzzi, L., 2013. Oncoimmunology 2, e26621.
Klausen, U., Holmberg, S., Holmström, M.O., Jørgensen, N.G.D., Grauslund, J.H., Svane, I.M., Andersen, M.H., 2018. Front. Immunol. 9, 2264.
Reche, P., Flower, D.R., Fridkis-Hareli, M., Hoshino, Y., 2015. J. Immunol. Res. 2015.
Schumacher, T.N., Schreiber, R.D., 2015. Science 348, 69–74.
Wirth, T.C., Kühnel, F., 2017. Front. Immunol. 8, 1848.
Hacohen, N., Fritsch, E.F., Carter, T.A., Lander, E.S., Wu, C.J., 2013. Cancer Immunol. Res. 1, 11–15.
Saini, S.K., Rekers, N., Hadrup, S.R., 2017. Ann. Oncol. 28, xii3–xii10.
Kazi, A., Chuah, C., Majeed, A.B.A., Leow, C.H., Lim, B.H., Leow, C.Y., 2018. Pathog. Glob. Health 112, 123–131.
Knutson, K.L., Schiffman, K., Disis, M.L., 2001. J. Clin. Invest. 107, 477–484.
Roberts, J.D., Niedzwiecki, D., Carson, W.E., Chapman, P.B., Gajewski, T.F., Ernstoff, M.S., Hodi, F.S., Shea, C., Leong, S.P., Johnson, J., 2006. J. Immunother. 29, 95–101.
Soria-Guerra, R.E., Nieto-Gomez, R., Govea-Alonso, D.O., Rosales-Mendoza, S., 2015. J. Biomed. Inform. 53, 405–414.
Doytchinova, I.A., Flower, D.R., 2007. BMC Bioinformatics 8, 4.
Tappeiner, E., Finotello, F., Charoentong, P., Mayer, C., Rieder, D., Trajanoski, Z., 2017. Bioinformatics 33, 3140–3141.
Restifo, N.P., Dudley, M.E., Rosenberg, S.A., 2012. Nat. Rev. Immunol. 12, 269.
Gill, S., June, C.H., 2015. Immunol. Rev. 263, 68–89.
Olsen, L.R., Tongchusak, S., Lin, H., Reinherz, E.L., Brusic, V., Zhang, G.L., Immunology, Cancer, 2017. Immunotherapy 66 (6), 731–735.
Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., Peters, B., 2018. Nucleic Acids Res. 47 (D1), D339–D343.
Boughorbel, S., Jarray, F., El-Anbari, M., 2017. PLoS One 12, e0177678.
Riley, R.S., June, C.H., Langer, R., Mitchell, M.J., 2019. Nat. Rev. Drug Discov. 1.
Seliger, B., 2019. Front. Immunol. 10, 999.
Knocke, S., Fleischmann-Mundt, B., Saborowski, M., Manns, M.P., Kühnel, F., Wirth, T.C., Woller, N., 2016. Cell Rep. 17, 2234–2246.
Andreatta, M., Nielsen, M., 2015. Bioinformatics 32, 511–517.
Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., Nielsen, M., 2017. J. Immunol. ji1700893.
Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P.E., Bui, H.-H., Buus, S., Frankild, S., Greenbaum, J., 2008. Nucleic Acids Res. 36, W513–W518.
Rammensee, H.-G., Bachmann, J., Emmerich, N.P.N., Bachor, O.A., Stevanović, S., 1999. Immunogenetics 50, 213–219.
Olson, R.S., La Cava, W., Mustahsan, Z., Varik, A., Moore, J.H., 2017. arXiv preprint arXiv:1708.05070.
Boulesteix, A.L., Janitza, S., Kruppa, J., König, I.R., 2012. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2, 493–507.
Couronné, R., Probst, P., Boulesteix, A.-L., 2018. BMC Bioinformatics 19, 270.
Qi, Y., 2012. Random forest for bioinformatics. Ensemble Machine Learning. Springer, pp. 307–323.
Kandaswamy, K.K., Chou, K.-C., Martinetz, T., Möller, S., Suganthan, P., Sridharan, S., Pugalenthi, G., 2011. J. Theor. Biol. 270, 56–62.
Chen, L., Chu, C., Huang, T., Kong, X., Cai, Y.-D., 2015. Amino Acids 47, 1485–1493.
Chang, K.Y., Yang, J.-R., 2013. PLoS One 8, e70166.
Zhang, W., Xiong, Y., Zhao, M., Zou, H., Ye, X., Liu, J., 2011. BMC Bioinformatics 12, 341.
Bhadra, P., Yan, J., Li, J., Fong, S., Siu, S.W., 2018. Sci. Rep. 8, 1697.
Meher, P.K., Sahu, T.K., Saini, V., Rao, A.R., 2017. Sci. Rep. 7, 42362.
Thakur, N., Qureshi, A., Kumar, M., 2012. Nucleic Acids Res. 40, W199–W204.
Lissabet, J.F.B., Belén, L.H., Farias, J.G., 2019. Comput. Biol. Med. 107, 127–130.