proceedings

MDPI

Proceedings

Prediction of Peptide Vascularization Inhibitory Activity in Tumor Tissue as a Possible Target for Cancer Treatment †

Jose Liñares-Blanco and Carlos Fernandez-Lozano *

Department of Computer Science, Faculty of Computer Science, University of A Coruña, CITIC, A Coruña 15071, Spain; j.linares@udc.es

* Correspondence: carlos.fernandez@udc.es; Tel.: +34-881-01-6013

† Presented at the 2nd XoveTIC conference, A Coruña, 5–6 September 2019.

Published: 31 July 2019

Abstract: The prediction of metabolic activities in silico form is crucial to be able to address all research possibilities without exceeding the experimental costs. In particular, for cancer research, the prediction of certain activities can be of great help in the discovery of different treatments. In this work it has been proposed to predict, through Machine Learning, the anti-angiogenic activity of peptides is currently being used in cancer treatment and is giving hopeful results. From a list of peptide sequences, three types of molecular descriptors were obtained (AAC, DC and TC) that offered the possibility of training different ML algorithms. After a Feature Selection process, different models were obtained with a predictive value that surpassed the current state of the art. These results shown that ML is useful for the classification and prediction of the activity of new peptides, making experimental screening cheaper and faster.

Dataset: https://doi.org/10.6084/m9.figshare.6016994

Dataset License: CC0

Keywords: machine learning; feature selection; activity prediction; peptides; cancer; screening

1. Introduction

The prediction of metabolic activities in-silico is crucial to address all research possibilities without exceeding the experimental costs. Specifically, in cancer research, there are endless opportunities that may be tested for possible treatment. Among them, one method that is having hopeful results in this field is tumor treatment with anti-angiogenic peptides. Attacking the tumor by destabilizing its micro-environment is crucial to prevent its development. The vast majority of treatments of this type are based on peptides, mainly due to their low toxicity and design possibilities.

Being able to predict the anti-angiogenic activity of these peptides from their amino acid sequence opens the doors to the discovery of new natural peptides or designed in the laboratory with this activity. In this sense, experimental researchers would be able to carry out a previous filter at the time of experimentally validating the treatments and focus only on those that had a previous significance, by means of in-silico techniques.

In this work, Machine Learning techniques were able to predict with high precision if a peptide has anti-angiogenic activity, only from its amino acid sequence. Our work was published before in detail as open access in the Scientific Reports journal [1]. The list of sequences was obtained from the article [2].

2. Results

Previous studies shown that peptides with anti-angiogenic activity have a common structure, presenting mostly folds of beta anti-parallel sheets, with a high incidence of hydrophobic and cationic residues. On the other hand, the composition of these peptides is not fully defined, although it has been observed that these peptides are more prone to present certain amino acid residues in their sequences.

2.1. Baseline Algorithms without Feature Selection

Firstly, a comparative experiment was carried out under the same conditions in order to observe which dataset-algorithm pair achieves the best performance. Four algorithms (SVM [3], RF [4], Glmnet [5] and k-NN [6]) were trained with the three descriptors (AAC, DC and TC) [7] and the union of AAC-DC and AAC-TC. The results obtained in this phase of the experiment do not improve those reported in the literature (0.809 in Accuracy) [2], but they do indicate a certain trend in the data.

2.2. Feature Selection

We ranked the features accordiing with their p-values after performing a Kruskal test of each variable with the dependent variable (Anti-angiogenic and Non Anti-angiogenic). We explored the size of different subgroups, (5, 10 and 15) for AAC, (25, 50, 75, 100) for DC, (75, 100, 125, 150) for TC and (50, 100, 150, 200) for the combination of the three datasets.

In this case, several models exceed the values marked in the literature (red line), all of them with RF and Glmnet algorithms. The best performance was achieved with the Glmnet algorithm trained with the dataset of the union of the three descriptors, using the 200 most significant variables after the action of the statistical test. Furthermore, as can be seen in Figure 1b,d, all the models show great stability, which indicates that they have obtained homogeneous results in all the repetitions.

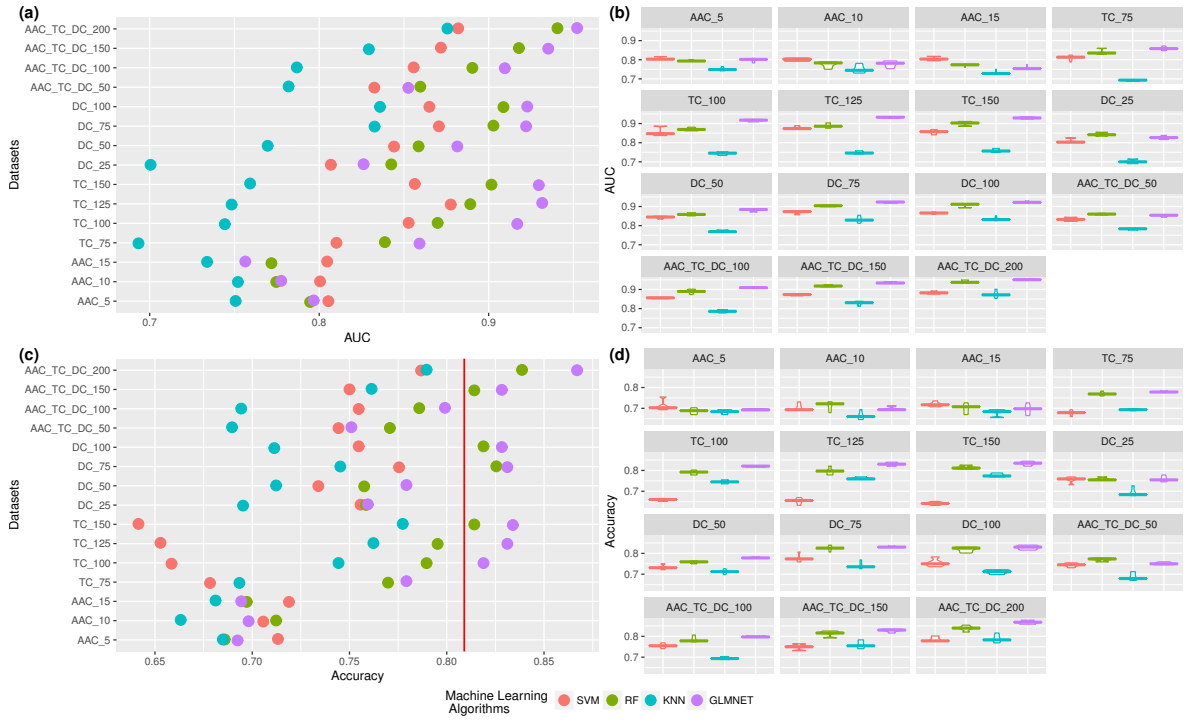

Figure 1. Results obtained in the feature selection process. (**a**) Performance using AUC, (**b**) boxplot using AUC, (**c**) performance using accuracy, and (**d**) boxplot using accuracy. Best previously published value in the literature by Ramaprasad et al. [2] (red line).

3. Discussion

The results obtained in this work reflect the importance of the use of new data analysis technologies to support experimental research. This work reports a set of models that have surpassed previous state of the art models for this type of problem (p-value $= 1.8665 \times 10^{-9}$). A high proportion of amino acids within the sequence such as Alanine, Valine and Cysteine are important for classifying peptides. In addition, it is observed in the higher positions, as di-peptide sequences (SP, VD, ID and CK) and tripeptides (LSL, DIT and PDL) provide significant information to this model.

4. Materials and Methods

Machine Learning Models

The following algorithms were implemented: Support Vector Machines (SVM) [3], Random Forest (RF) [4], k-Nearest Neighbors (k-NN) [6] and Generalized linear model (Glmnet) [5]. A Nested Cross Validation was used for training the models. In other words, there were two validation phases. Firstly, a holdout was used for the selection of the best hyperparameters (2/3 for training and 1/3 for testing) and secondly, a 10-fold CV was used for the validation of the model (we ran 5 times this CV process).

Author Contributions: Conceptualization, C.F.-L.; methodology, C.F.-L.; software, J.L.B. and C.F.-L.; formal analysis, J.L.B. and C.F.-L.; Writing—Original Draft preparation, J.L.B.; Writing—Review and Editing, J.L.B. and C.F.-L.; supervision, C.F-L.

Funding: This research received no external funding.

Acknowledgments: This work is supported by the "Collaborative Project in Genomic Data Integration (CICLOGEN)" PI17/01826 funded by the Carlos III Health Institute from the Spanish National plan for Scientific and Technical Research and Innovation 2013–2016 and the European Regional Development Funds (FEDER)—"A way to build Europe". This project was also supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. ED431G/01, ED431D 2017/16), the "Galician Network for Colorectal Cancer Research" (Ref. ED431D 2017/23), and the Spanish Ministry of Economy and Competitiveness via funding of the unique installation BIOCAI (UNLC08-1E-002, UNLC13-13-3503) and the European Regional Development Funds (FEDER) by the European Union and the "Juan de la Cierva" fellowship program supported by the Spanish Ministry of Economy and Competitiveness (Carlos Fernandez-Lozano, Ref. FJCI- 2015-26071).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liñares Blanco, J.; Porto-Pazos, A.B.; Pazos, A.; Fernandez-Lozano, C. Prediction of high anti-angiogenic activity peptides in silico using a generalized linear model and feature selection. *Sci. Rep.* **2018**, *8*, 1–11.
2. Ramaprasad, A.S.E.; Singh, S.; Venkatesan, S. AntiAngioPred: a server for prediction of anti-angiogenic peptides. *PLoS ONE* **2015**, *10*, e0136990.
3. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
4. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
5. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1.
6. Cover, T.M.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27.
7. Bhasin, M.; Raghava, G.P. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* **2004**, *279*, 23262–23266.