UDC 004.4

Application of neural networks to the analysis of the resistance of the human immunodeficiency virus to HIV reverse transcriptase inhibitors

Anastasia V. Demidova[*], **Olga A. Tarasova**[†]

[*] *Department of Applied Probability and Informatics*
Peoples' Friendship University of Russia
Miklukho-Maklaya str. 6, Moscow, 117198, Russia
[†] *Institute of Biomedical Chemistry*
Pogodinskaya str. 10, Moscow, 119121, Russia

Email: demidova_av@rudn.university, olga.a.tarasova@gmail.com

AIDS and the opportunistic infections and some other complications, associated with this syndrome lead to more than one million of human deaths per year. Human Immunodeficiency Virus is a cause of AIDS. The drugs, targeted proteins of HIV, can only lead to decrease of HIV copies in the human organism but still do not eliminate HIV from an organism. The main cause of antiretroviral drug therapy failure is HIV resistance to main drug classes: inhibitors of HIV structural proteins, protease and reverse transcriptase. One of the approaches to the classification of HIV variants into the resistant and susceptible ones is the use of machine learning methods. The aim of this work is the classification of the HIV variants into susceptible and resistant based on the nucleotide sequence of the HIV protease using neural networks. In our work, we used two topologies of neural networks: multilayer perceptron and convolutional neural network. Neural Networks were built using Python Tensor Flow and Keras libraries, where optimization of the neural networks can be performed. The training and test sets include experimental data on the nucleotide sequence of HIV protease and their resistance. Sensitivity, specificity, balanced accuracy were used as the main parameters, reflected the quality of classification. Those parameters were calculated for the test set, collected in the later period comparing to the sequences of the data set.

Key words and phrases: neural networks, HIV/AIDS, inhibitors, resistance.

In: K. E. Samouylov, L. A. Sevastianov, D. S. Kulyabov (eds.): Selected Papers of the IX Conference "Information and Telecommunication Technologies and Mathematical Modeling of High-Tech Systems", Moscow, Russia, 19-Apr-2019, published at http://ceur-ws.org

1. Introduction

Human immunodeficiency virus type 1 (HIV-1) is a retrovirus that causes the Acquired Immunodeficiency Syndrome (AIDS) leading to the failure of immune system and resulting in death. There are more than 36 HIV-infected people globally and more than one million of them live in Russian Federation. Protocols for HIV/AIDS treatment typically include use various combinations of antiretroviral drugs (Highly Active Antiretroviral Treatment (HAART)). There are two main classes on antiretroviral drugs included in HAART: inhibitors of structural HIV proteins, protease and reverse transcriptase, encoded by single HIV gene named pol. Marketed antiretroviral drugs are incapable of eliminating of virus from human organism. Mutations in the genes of HIV cause its resistance to the drugs. HIV characterizes by a high velocity of mutations occurrence, therefore it is able to develop resistance in short terms. HIV resistance results in loss of drugs effects and necessity to use different drug combinations to prevent viral replication. HIV resistance to the main antiretroviral classes of drugs is typically estimated using two types of experimental tests: phenotypic tests and genotypic ones. A genotypic test produces nucleotide sequences of the pol gene, while a phenotypic test represents the data on the genotype of HIV resistant variant together with the data on its resistance. The results of phenotypic and genotypic tests can be used to develop on their basis the computational method aimed at predicting HIV resistance to the particular antiretroviral drugs. There have been developed many of them and an accuracy of prediction is over 90% (for some of them is more than 95%), for details, please, see review by [1–4].

There are several machine learning approaches for the prediction of HIV drug resistance [1, 2, 5–11]. There were demonstrated using these approaches that an accuracy of prediction of HIV resistance to a certain drug may vary depending on the type of descriptor chosen, a drug, to which the resistance should be predicted. Application of neural networks have become widely used in biology and chemistry for a past decade [12], [13]. The representation of a nucleotide sequence of HIV reverse transcriptase and protease as a set of nucleotide fragments can be used for successful prediction of HIV drug resistance, as was shown in a study by [2]. In the current work, we demonstrate the application of the neural networks to the classification of HIV protease sequences into resistant and susceptible groups based on the descriptors generated from the nucleotide sequences, as we previously developed.

2. The description of the data

Nucleotide sequences of HIV protease were use as the training and test sets with the data on their resistance produced using Phenosense test system for phenotypic tests. In our study we used data on the HIV resistance to six marketed inhibitors of HIV protease:

- fosamprenavir (FPV);
- azatanavir (ATV);
- indinavir (IDV);
- lopinavir (LPV);
- nelfinavir (NFV).

The data include sequences of the pol collected using samples of the patients and available from HIV Stanford drug resistance database. A detailed description of the data processing is given in the study by [2]. In our earlier study we tried to simulate a prospective validation using the set of the sequences collected from the patients, examined no later than in January, 2006; sequences, collected later were used a the test set. Nucleotide sequences were divided into the short sequences (descriptors), where each short sequence was represented by a central nucleotide, eleven nucleotides before the central one and twelve nucleotides after it. For each sequence, we generated the set of descriptors. Then we collected all of them in a list and sorted by their frequency in the whole set of nucleotide sequences of protease. We used 500 descriptors and generated a set of binary descriptors from them, where each value of the descriptor was “1” if the particular short sequence in the nucleotide sequence from the training or test set or “0” if it is not in the while nucleotide sequence.

3. The architecture of the neural networks

In the current study we used two types of artificial neural networks: convolutional neural networks (CNN) and multilayer perceptron (MLP). The neural network were built using Python as the programming language and its libraries: Keras and TensorFlow. These tools include several options to optimize the parameters for the neural networks and to estimate an accuracy of the classification. Multilayer network included an input layer, five hidden layers and an output layer. Rectifier Linear Unit (ReLU) was used as a function of activation in hidden layers and a sigmoid function was used in the output layer. A dropout option was used to prevent an overtraining [14].

The architecture of the neural network, which is based on the convolution operation, was first developed in the late 1990s by Lekun et al. [15]. Today convolutional neural networks are considered to be the best for solving image recognition problems. In this work we use the sequences of binary descriptors as training data that take the value «0» or «1». Thus, this data can be assigned in the same way as images in recognition tasks. This view allows one to apply the apparatus of convolutional neural networks.

Convolutional neural network includes the input layer, the host matrix of size 22x22, two convolutional layers with feature map size 44x44 and 88x88, pooling layer (MaxPooling), two hidden fully connected layers with 352 and 151 neurons and an output layer. The convolution kernel in all layers is 3. ReLU is used as the activation function in hidden layers, and sigmoid activation function is used on the output layer.

To assess the quality of neural network for each of the 6 drugs were calculated indicators such as [16]:

– Sensitivity, also known as recall, reflects the proportion of positive results that are correctly identified by the classifier.
– Specificity — reflects the proportion of negative results that are correctly identified by the classifier.
– Balanced accuracy is the proportion of true results (both positive and true negative) among the total number of considered cases, i.e. the probability that the class will be predicted correctly.
– The precision of the classification of positive results is the proportion of positive results that are correctly identified by the classifier among the total number of considered cases.

To obtain an assessment of the classifier quality, a cross-validation with a 10-fold division into training and test samples was used. In tables 1 and 2 the results of the classifier are presented.

Table 1

Assessment of the neural network quality MLP

Drug	Sensitivity	Specificity	Precision	Balanced accuracy
FPV	0.655	0.945	0.45	0.917
ATV	0.838	0.63	0.787	0.757
IDV	0.902	0.882	0.866	0.892
LPV	0.824	0.857	0.741	0.847
NFV	0.879	0.856	0.89	0.868
SQV	0.795	0.87	0.812	0.839

Table 2

Assessment of the neural network quality CNN

Drug	Sensitivity	Specificity	Precision	Balanced accuracy
FPV	0.951	0.965	0.966	0.958
ATV	0.827	0.825	0.823	0.815
IDV	0.886	0.867	0.857	0.872
LPV	0.871	0.897	0.902	0.882
NFV	0.9	0.844	0.832	0.864
SQV	0.847	0.878	0.885	0.859

Low values of accuracy may be associated with a small amount of training sample, as well as with the peculiarities of biological data on HIV resistance, which are characterized by a certain incompleteness and heterogeneity [4,17,18].

The analysis of the results showed that the classifier constructed with the help of convolutional neural network for the majority of metrics and preparations give better results than multilayer perceptron. In addition, convolutional neural network use almost two times less parameters – 1 153 126 MLP parameters against 578 444 CNN parameters, which improvs convergence and reducs computation time.

4. Conclusion

The paper demonstrates the application of neural networks of different topologies to the problem of classification of human immunodeficiency virus resistance to HIV protease inhibitors. In the future we plan to improve the performance of the classifier through the use of other topologies of neural networks, the selection of the optimal number of layers, maps, features, sizes of the convolution kernel and other hyperparameters.

Acknowledgments

The publication has been prepared with the support of the "RUDN University Program 5-100".

References

1. M. Riemenschneider, R. Senge, U. Neumann, E. Hüllermeier, D. Heider, Exploiting hiv-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification, BioData Min 9 (2016) 10–10, 26933450[pmid]. doi:10.1186/s13040-016-0089-1.
URL https://www.ncbi.nlm.nih.gov/pubmed/26933450
2. O. Tarasova, N. Biziukova, D. Filimonov, V. Poroikov, A computational approach for the prediction of hiv resistance based on amino acid and nucleotide descriptors, Molecules 23 (11) (2018) 2751, 30355996[pmid]. doi:10.3390/molecules23112751.
URL https://www.ncbi.nlm.nih.gov/pubmed/30355996
3. M. J. Dapp, R. H. Heineman, L. M. Mansky, Interrelationship between hiv-1 fitness and mutation rate, J Mol Biol 425 (1) (2013) 41–53, 23084856[pmid]. doi:10.1016/j.jmb.2012.10.009.
URL https://www.ncbi.nlm.nih.gov/pubmed/23084856
4. O. Tarasova, V. Poroikov, Hiv resistance prediction to reverse transcriptase inhibitors: Focus on Open Data, Molecules 23 (4) (2018) 956, 29671808[pmid]. doi:10.3390/

molecules23040956.
URL https://www.ncbi.nlm.nih.gov/pubmed/29671808

5. N. Beerenwinkel, B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, J. Selbig, Diversity and complexity of hiv-1 drug resistance: A bioinformatics approach to predicting phenotype from genotype, Proceedings of the National Academy of Sciences 99 (12) (2002) 8271–8276. arXiv:https://www.pnas.org/content/99/12/8271.full.pdf, doi:10.1073/pnas.112177799.
URL https://www.pnas.org/content/99/12/8271
6. N. Beerenwinkel, M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, H. Walter, Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes, Nucleic Acids Res 31 (13) (2003) 3850–3855, 12824435[pmid].
URL https://www.ncbi.nlm.nih.gov/pubmed/12824435
7. S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, R. W. Shafer, Genotypic predictors of human immunodeficiency virus type 1 drug resistance, Proceedings of the National Academy of Sciences 103 (46) (2006) 17355–17360. arXiv:https://www.pnas.org/content/103/46/17355.full.pdf, doi:10.1073/pnas.0607274103.
URL https://www.pnas.org/content/103/46/17355
8. D. Heider, J. Verheyen, D. Hoffmann, Machine learning on normalized protein sequences, BMC Res Notes 4 (2011) 94–94, 21453485[pmid]. doi:10.1186/1756-0500-4-94.
URL https://www.ncbi.nlm.nih.gov/pubmed/21453485
9. G. J. P. van Westen, A. Hendriks, J. K. Wegner, A. P. Ijzerman, H. W. T. van Vlijmen, A. Bender, Significantly improved hiv inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data, PLoS Comput Biol 9 (2) (2013) e1002899–e1002899, 23436985[pmid]. doi:10.1371/journal.pcbi.1002899.
URL https://www.ncbi.nlm.nih.gov/pubmed/23436985
10. O. Tarasova, D. Filimonov, P. V.V., Computational prediction of human immunodeficiency resistance to reverse transcriptase inhibitors, Biomeditsinskaya khimiya 63 (5) (2017) 457–460. doi:10.18097/PBMC20176305457.
11. O. Tarasova, D. Filimonov, V. Poroikov, Pass-based approach to predict hiv-1 reverse transcriptase resistance, Journal of Bioinformatics and Computational Biology 15 (02) (2017) 1650040, pMID: 28033735. doi:10.1142/S0219720016500402.
12. I. I. Baskin, D. Winkler, I. V. Tetko, A renaissance of neural networks in drug discovery, Expert Opinion on Drug Discovery 11 (8) (2016) 785–795, pMID: 27295548. doi:10.1080/17460441.2016.1201262.
URL https://doi.org/10.1080/17460441.2016.1201262
13. D. Dana, S. V. Gadhiya, L. G. St Surin, D. Li, F. Naaz, Q. Ali, L. Paka, M. A. Yamin, M. Narayan, I. D. Goldberg, P. Narayan, Deep learning in drug discovery and medicine; scratching the surface, Molecules 23 (9) (2018) 2384, 30231499[pmid]. doi:10.3390/molecules23092384.
14. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.
URL http://jmlr.org/papers/v15/srivastava14a.html
15. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324. doi:10.1109/5.726791.
16. D. M. W. Powers, Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation, Journal of Machine Learning Technologies 2 (1) (2011) 37–63.
17. O. A. Tarasova, A. F. Urusova, D. A. Filimonov, M. C. Nicklaus, A. V. Zakharov, V. V. Poroikov, Qsar modeling using large-scale databases: Case Study for HIV-1 Reverse Transcriptase Inhibitors, Journal of Chemical Information and Modeling 55 (7) (2015) 1388–1399. doi:10.1021/acs.jcim.5b00019.
URL https://doi.org/10.1021/acs.jcim.5b00019

18. D. Fourches, E. Muratov, A. Tropsha, Curation of chemogenomics data, Nature Chemical Biology 11 (2015) 535, correspondence.
URL https://doi.org/10.1038/nchembio.1881