Maj Institute of Pharmacology
Polish Academy of Sciences

Kraków, 2024-02-14

Review of the doctoral thesis of Mr. Oleksandr Myronov, M.Sc. entitled

In-silico Modeling of Antigen Recognition During Immune Response by Analyzing the Sequential and Structural Peptide-HLA-TCR Data Using Machine Learning

prepared under the supervision of prof. Dariusz Plewczyński

Although various artificial intelligence (AI) methods, especially deep learning (DL), are increasingly being used in many areas of our lives, it was only when OpenAI released its ChatGPT in November 2022 that everyone realized the level of sophistication and possibilities of the new technology. Developing new DL-based models requires, at first, access to large and diverse datasets, data preparation for training, application of effective learning algorithms, robust validation, and testing. The best if models are integrated into practical contexts, considering real conditions and our needs, and if the model could be understandable. It seems that there is no more critical yet challenging field of research fitting DL models' potential than exploring various aspects of human health.

The doctoral thesis of Olexandr Myronov, M.Sc. fits precisely into this area, addressing a particularly crucial issue of human immunity, specifically, predicting antigen recognition by the human adaptive immune system.

The 156-page dissertation is in English and generally follows a standard format. It begins with a short introduction section containing (i) the assumptions and objectives of the research, (ii) discussing the basics of the biochemical mechanism of antigen fragments (peptides) presentation on a cell surface by Human Leukocyte Antigen (HLA) and their recognition by specialized T-cell receptors (TCR), and (iii) a brief overview of DL methods used. The Materials and Methods section, on the one hand, extends the last part of the introduction by discussing the basics of machine learning and, on the other hand, presents the methods used directly in the research part. Appropriate to the dissertation topic, special emphasis is placed on DL approaches used for sequence classification and include convolutional neural

12 Smętna Street
31-343 Krakow
Poland

e-mail: ifpan@if-pan.krakow.pl
www.if-pan.krakow.pl

phone: +48 12 662 32 20
+48 12 662 32 05
+48 12 662 32 49
fax: +48 12 637 45 00

networks (CNNs), recurrent neural networks (RNNs) and more recent the Transformer architecture. These methods have demonstrated remarkable effectiveness across various domains, from image classification, bioinformatics to natural language processing, text generation and beyond. It is worth noting the consistent, hierarchical introduction and explanation of relevant terms and abbreviations, the mathematical description layer comprising 77 equations, and the graphical illustrations, particularly those with special notions showing layer architectures, also used consequently in the rest of the dissertation. Next, a brief characterization and sources of datasets used for model training and validation are provided, i.e., peptide-HLA presentation, peptide-HLA immunogenicity, and peptide-TCR binding sets.

The results section first presents the development of a general peptide-HLA presentation model using CNN. The same architecture was used for the viral immunogenicity model, whereas the peptide-HLA cancer immunogenicity model was based on RNN. The models were thoroughly validated, and their performance was better when compared to several commonly used tools described in the literature. They were also implemented to the Ardigen software – ArdImmune Rank and ArdImmune Vax, and used for a design of potential COVID-19 vaccine. The developed models were also applied to explore the relationships between immunogenicity and selected immune escape mechanisms. The section is based on the analysis of vast data on patients with five types of cancer collected in The Cancer Genome Atlas and The Cancer Immunome Atlas.

The third and last part of the thesis deals with predicting peptide binding to TCR receptors, i.e., the event that triggers an immune response. Due to several orders of magnitude higher numbers of both presented human antigens and TCR sequences than available experimental data, which are moreover limited only to positive examples, the goal was highly challenging. The Author first elaborated a method for generating negative decoys and then created a model called BERTrand based on transformer architecture with extensive unsupervised pre-training. From the carefully reviewed benchmark methods only those that could compete in predicting unseen earlier peptides were selected. Many model development and evaluation issues were comprehensively explained, illustrated, and discussed. Although BERTrand performance on the external test set was moderate, it significantly outperformed

other methods, thus should be the first choice tool in predicting peptide-TCR binding. It has to be stressed that BERTrand software is freely available on GitHub.

Finally, in conclusions, in addition to highlighting the accomplishments, the author points to encountered problems and proposes solutions facilitating the progress of computational immunology. The 98-item bibliography is relevant and up-to-date.

Generally, I found Mr. Myronov's dissertation extremely valuable. The topic of computational immunology is highly relevant, and the elaborated state-of-the-art in silico tools for the prediction of antigen recognition by the immune system were made commercially or publically available. Most of the results presented have already been published in two publications, one of which Ph.D. Candidate authored as the first and corresponding author. The results were also presented at prestigious international conferences. It should also be stressed that the project of the thesis has been granted under the Industrial PhD Program financed by the Polish Ministry of Science and Higher Education. Most of the research was done at Ardigen company in Kraków.

Mr. Myronov has demonstrated proficiency in biology, bioinformatics, research design, data collection and curation, advanced DL methods, programming, comprehensive analysis of results, their presentation, discussion, drawing conclusions, problem-solving, and scientific article writing.

I fully agree with the Author that the issue of experimentally confirmed negative examples in training datasets for ML-model construction is very serious. It is in line with my experience in the case of available affinity data of small molecule ligands of G-protein coupled receptors since the publication of negative results, i.e., inactive compounds, is practically impossible. I appreciate the Author's efforts in generating the most relevant sets of negative decoys.

The following aspects of the dissertation are, in my opinion, also noteworthy:

- a mature and versatile approach to problem-solving
- a logical structure with detailed chapter divisions
- recognition of problems and critical comments
- frequent cross-references that integrate the material and enhance the readability

- skillful descriptions of the Author's personal contributions
- writing relevant passages in the first person
- meticulous and informative figures and schemes

Below there are some minor remarks and questions:

p11 – “The enzyme called TAP cuts the old proteins into short chains...” TAP does not cut proteins

p13 – “Inside the endometrial reticulum, the proteins are cleaved by the proteasome...” – proteasome is not inside the ER

p16 – “It is estimated that an average person experiences a potential cancer event twice a week...” what is the source of this statement?

P72, Figure 22 – dataset for the peptide-HLA presentation model – it seems that data on more HLA types was available in the original paper by Sarkizova et al. 2019 (Figure 1C).

P80, Figure 27 – unusual leave-one-family-out cross-validation results for the Pneumoviridae family should be commented on.

There is no list of abbreviations and I really missed it. There are some abbreviations that are not explained under the first usage (e.g. TCGA, APP, TME), and some are not explained (e.g. GWAS, MHC).

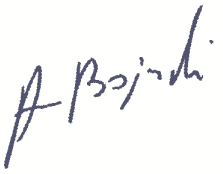
While the dissertation is well illustrated, and most figures are well-thought-out and informative, figure captions are generally minimalistic (e.g., Figure 2), making referring to the accompanying text necessary. The same figures in publications have much more elaborated captions.

Are there any plans to publish the results of section 3.8 Immune escape mechanisms analysis? Shouldn't more cancer types be analyzed to analyze relationships between immunogenicity and immune escape mechanisms?

In summary, I have no doubts that the reviewed doctoral dissertation of Olexandr Myronov, M.Sc., meets the appropriate legal requirements, and I strongly recommend the

acceptance of this thesis by the Scientific Council of the Discipline Information and Communication Technology of the Warsaw University of Technology.

Due to the high scientific quality and commercial applications of the developed methods, I hereby recommend the thesis for distinctions.



Podpisany elektronicznie przez
Andrzej Jacek Bojarski
14.02.2024
16:29:29 +01'00'

Prof. Andrzej J. Bojarski
Head of the Department of Medicinal Chemistry
Maj Institute of Pharmacology Polish Academy of Sciences
12 Smetna Street, 31-343 Kraków, Poland