Modelowanie in silico rozpoznania antygenu w odpowiedzi immunologicznej na podstawie analizy danych sekwencyjnych oraz kompleksów strukturalnych peptyd-HLA-TCR przy pomocy metod uczenia maszynowego

Streszczenie

Przedstawiono wybrane problemy z dziedziny immunologii obliczeniowej – rozpoznawanie antygenu (peptydu) przedstawionego na powierzchni komórki przez białko głównego układu zgodności tkankowej (ang. Major histocompatibility complex, MHC), przez receptor limfocytu T (TCR). Skupiłem się na 3 problemach – przewidywania prezentacji peptyd-MHC, immunogenności peptyd-MHC oraz wiązania peptyd-TCR. Zebrano dane dotyczące prezentacji peptydów pochodzących z eksperymentów spektrometrii masowej; immunogenności peptydów w kontekście MHC z eksperymentów badających rozpoznanie antygenu oraz wiązania peptyd-TCR pochodzące z eksperymentów sekwencjonowania TCR za pomocą metod nowej generacji (ang. Next Generation Sequencing, NGS). Dane były użyte do stworzenia 4 modeli uczenia maszynowego: 1) model prawdopodobieństwa prezentacji oraz 2) model wirusowej immunogenności na podstawie splotowych sieci neuronowych; 3) model prawdopodobieństwa immunogenności dla nowotworowych peptydów na podstawie rekurencyjnych sieci neuronowych oraz 4) model wiązania peptyd-TCR na podstawie architektury Transformer z uwagą. Wszystkie modele oceniono na niezależnych zbiorach testowych, osiągając bardzo dobre wyniki. Modele 1 oraz 2 zastosowano do badania podstawowej immunogenności peptydów wirusa SARS-nCov-2, zaproponowano hipotetyczny skład szczepionki T limfocytowej. Modele 1 oraz 3 zastosowano do badania mechanizmów ucieczki nowotworu przed systemem odpownościowym, stwierdzono również związek pomiędzy ich wystąpieniem a immunogennością. Model 4 pod nazwą BERTrand zawiera innowacyjne rozwiązania dotyczące generowania sztucznych negatywnych przykładów oraz stworzenia modelu językowego. BERTrand wykazał się bardzo dobrymi wynikami w przewidywaniu wiązania peptyd-TCR dla nieznanych peptydów, czyli osiągnął generalizację modelu uczenia maszynowego. BERTrand został udostępniony w postaci oprogramowania w języku Python, zawiera jeden z największych zbiorów dla trenowania i walidacji modeli oraz przyjazny interfejs użytkownika dla inferencji oraz trenowania własnych modeli.

Słowa kluczowe: uczenie maszynowe, sieci neuronowe, immunologia obliczeniowa, peptyd, MHC, TCR

Warszawa, 13.06.23

Aleksandr Mironov

In-silico modeling of antigen recognition during immune response by analyzing the sequential and structural peptide-HLA-TCR data using machine learning

Summary

In my PhD thesis I addressed several problems from computational immunology – the recognition of a peptide, presented on the surface of the cell by the Major Histocompatibility Complex (MHC), by a T-cell receptor (TCR). I focused on three critical problems in the domain – peptide-MHC presentation prediction, peptide-MHC immunogenicity prediction and peptide-TCR binding prediction. The data from public sources was collected and curated from peptide presentation mass spectrometry experiments, peptide-MHC immunogenicity assays and TCR next generation sequencing (NGS). The data was used to create four machine learning models: 1) peptide-MHC presentation prediction model and 2) peptide-MHC viral immunogenicity model using convolutional neural networks (CNNs); 3) peptide-MHC cancer immunogenicity model based on recurrent neural networks (RNNs); 4) peptide-TCR binding model using the Transformer architecture and the attention mechanism. All the models were evaluated using independent test sets and achieved very good results. Models 1 and 2 were applied to basic research of the immunogenicity of SARS-nCov-2 peptides. A hypothetical T-cell vaccine formulation was suggested. Models 1 and 3 were applied to the immune escape mechanisms study. The connection between immune evasion and immunogenicity was uncovered. Model 4, named BERTrand, contains novel solutions of the negative decoys observation problem, as well as language modeling. BERTrand achieved good results in prediction peptide-TCR binding for peptides unseen during model training, thus achieving generalization. BERTrand was published as a Python package, it contains one of the biggest datasets for model training and evaluation, as well as a user-friendly API for inference and training user's own models.

Keywords: machine learning, neural networks, computational immunology, peptide, MHC, TCR

Warszawa, 13.06.23

Aleksandr Mityanov