طراحی پپتیدهای ضدباکتریایی با استفاده از روش‌های یادگیری ماشین جمعی

فاطمه ابراهیمی ترکی، سمیه دباغ صادق‌پور، محبوبه ضرابی*

آزمایشگاه زیست‌شناسی محاسباتی، گروه بیوتکنولوژی، دانشکده علوم زیستی، دانشگاه الزهرا تهران

*m.zarrabi@alzahra.ac.ir

مقاومت‌های آنتی‌بیوتیکی چالش بزرگی است. پپتیدهای ضدمیکروبی غشای سلول را هدف قرار می‌دهند و غالباً، یک هدف خاص پروتئینی ندارند، احتمال ایجاد مقاومت در برابر اینها توسط باکتری‌ها، پائین است.آنالیزهای آماری و الگوریتم‌های یادگیری ماشین اخیراً مورد توجه قرار گرفته‌اند. تکنیک‌های یادگیری تجمیعی در یادگیری ماشین از تلفیق چند مدل برای ارائه‌ی مدل بهینه به منظور پیشگویی و طبقه‌بندی داده‌ها استفاده می کند. از الگوریتم‌های پرکاربرد در این زمینه می‌توان به الگوریتم‌های Bagging، Adaboost و RandomForest با تخمین‌گرهای متعدد، اشاره کرد. در این پژوهش به منظور پیشگویی پپتیدهایی با عملکرد اختصاصی ضد باکتریایی، داده‌ها از DRAMP 2.0 استخراج شد.روش‌های EDA با استفاده از کتابخانه‌های seaborn، numpy و pandas پایتون انجام شد. ۵۵۴ توالی با عملکرد ضدباکتری و ۶۲۶ توالی فاقد این عملکرد فراهم شدند. توصیف کننده‌ها برمبنای ویژگی‌های بیوفیزیکی پپتیدها ازجمله طول توالی، وزن مولکولی، بار، چگالی بار، pI، ضریب ناپایداری، آروماتیسیته، ضریب آلیفاتیک، ضریب Boman و میزان آب‌گریزی تعیین شدند. مدلسازی توسط الگوریتم SVM با توابع Kernel خطی، چند جمله‌ای با درجه ۵ و RBF با میزان گامای ۳ و الگوریتم‌های RandomForest با تعداد درخت ۱۰۰و۱۰۰۰ Bagging Classifier و Adaboost با تعداد تخمین‌گرهای ۱۰۰ و ۱۰۰۰ انجام شد. صحت و دقت مدل ساخته شده با استفاده از الگوریتم RandomForest با تعداد تخمین‌گر ۱۰۰۰، ۸۷٪ و ۹۰٪ و بهینه‌ترین حالت در مقایسه با روش‌های دیگر بود. میانگین صحت و دقت برای روش‌های SVM با Kernelهای اشاره شده، Bagging و Adaboost به ترتیب برابر بود با ۷۸٪، ۸۷٪ و ۸۶٪. برای داده‌ها و مشخصه‌های این پژوهش، رویکرد تجمیعی به دلیل نحوه‌ی استفاده از داده‌های train در حالت کلی در مقایسه با روش SVM نتیجه بهتری داشت.داده‌ها به صورت تصادفی تقسیم‌بندی شده و چندین بار برای یادگیری مدل استفاده شدند.با وجود پیشرفت روش‌های محاسباتی نیاز به روش‌های آزمایشگاهی به منظور ارزیابی‌های دقیق‌تر وجود دارد که این مرحله از گام‌های آتی این پژوهش است.

کلمات کلیدی: مقاومت آنتی‌بیوتیکی، EDA، الگوریتم SVM، Random Forest، پپتید

De novo design of Antibacterial peptides by ensemble machine learning methods

Fatemeh Ebrahimi Tarki, Somayyeh Dabbagh Sadeghpour, Mahboobeh Zarrabi*

Computational biology laboratory, Biotechnology department, Faculty of biological science, Alzahra University

*m.zarrabi@alzahra.ac.ir

Antibiotic resistance is a great challenge. Since Antimicrobial peptides directly act on the microbial membrane and normally didn’t have any specific protein targets, it is less likely, bacteria arise resistance against these molecules. Recently statistical analysis and machine learning algorithms have been considered. Ensemble learning techniques, in machine learning, are a combination of several models that are used to provide an optimal model for predicting or classifying data. The most widely used algorithms are Bagging, Adaboost and RandomForest with several estimators. In this study, to predict peptides with specific antibacterial effects, the data has been gathered from the DRAMP2.0, EDA were performed with the Seaborn, Numpy, and Pandas packages in Python. 554 peptides with antibacterial function and 626 without it were provided. Descriptors have been defined based on biophysical features like length, Molecular weight, Charge, Charge density, pI, Instability index, Aromaticity, Aliphatic index, Boman index, and Hydrophobic ratio. Modeling was performed using an SVM algorithm with linear, polynomial (degree=5) and RDF (gamma=3) kernel functons, RandomForest algorithm, Bagging classifier and Adaboost with 100 and 1000 estimators. The accuracy and precision of the model made using the RandomForest algorithm with 1000 estimators was

87% and 90% and this model was the most optimal compared to other methods. The average of accuracy and precision for SVM method with mentioned kernels,Bagging and Adaboost was 78%,87% and 86%, respectively. For the data and features of this study, the ensemble technique had better results than the SVM method due to the way the train data is used, the data is randomly segmented and used several times to learn the model. Despite the advancement of computational methods in drug design and therapeutic peptides, there is still a need for laboratory methods for more accurate evaluations, which is one of the next steps in this research.

Keywords: Antibiotic resistance, EDA, SVM algorithms, Random Forest, Peptide