AI and Biological Data Fusion: Theoretical Model Construction and Validation Boundary Exploration for Disease Target Prediction

Yichi Chen
Nouvelle Academy of Shenzhen, 518001, China
CYC18320978684@outlook.com

Abstract

This article focuses on the application of the integration of AI and biological data in the field of disease target prediction. Firstly, its significant importance was expounded, and it was pointed out that this integration can enhance the accuracy of prediction and accelerate drug research and development, etc. Then, the key elements for the construction of theoretical models for disease target prediction were discussed in detail, including data preprocessing, feature selection and extraction, and model algorithm selection, etc. Subsequently, the verification boundaries of the theoretical model were discussed, covering aspects such as verification methods, verification metrics, and the assessment of the model's generalization ability. Finally, summarize the full text, emphasize the potential and challenges of the integration of AI and biological data in the prediction of disease targets, and look forward to the future development direction.

Keywords: AI; Biological data fusion; Prediction of disease targets; Theoretical model construction; Verify the boundary.

1. Introduction

Disease target prediction is a core link in biomedical research, which is of vital significance for understanding the occurrence and development mechanisms of diseases, developing effective therapeutic drugs, and achieving personalized medicine [1]. Traditional methods for predicting disease targets mainly rely on biological experiments and prior knowledge, such as gene knockout experiments, protein-protein interaction studies, and drug screening. However, these methods often have problems such as low efficiency, high cost and limited accuracy. For instance, screening drug targets through experimental methods requires a large number of cell experiments and animal experiments, which not only consume a great deal of time and funds, but also, due to the complexity of biological systems, the experimental results may have problems such as false positives and false negatives [2].

As biotechnology has advanced at a swift pace and high-throughput technologies like genomics, transcriptomics, proteomics, and metabolomics have been extensively put into use, a vast quantity of biological data has been produced [3]. This biological data reflect the physiological and pathological states of organisms from different levels, providing rich information resources for the prediction of disease targets. For instance, gene expression data can reveal the changes in gene expression during the occurrence of diseases, protein interaction data can demonstrate the functional associations between proteins, and metabolomics data can reflect the changes in the metabolic state of organisms, etc. However, biological data possesses characteristics such as multi-source heterogeneity, high dimensionality, and high noise. How to extract valuable information from these complex data for disease target prediction is a major challenge currently faced [4].

Meanwhile, artificial intelligence (AI) technology has demonstrated powerful capabilities in data processing, pattern recognition, and predictive modeling. AI technology can automatically learn complex patterns and rules in data and efficiently analyze and process large-scale data [5]. Machine learning techniques, including Support Vector Machine (SVM), Random Forest (RF), and neural networks, have attained significant accomplishments across numerous domains. Deep learning, a subset of machine learning, excels in automatic feature extraction and nonlinear modeling, and has witnessed groundbreaking advancements in areas like image recognition, speech recognition, and natural language processing [6]. The integration of AI with biological data provides new ideas and methods for disease target prediction, which is expected to break through the limitations of traditional methods and achieve more accurate and efficient disease target prediction.

In recent years, the integration of AI and biological data has achieved some significant research results in the field of disease target prediction. For instance, some studies have utilized deep learning algorithms to analyze gene expression

data and successfully predicted potential targets for various diseases [7]. However, there are still some issues in this field that need to be further addressed at present, such as how to select appropriate AI algorithms and model structures, how to handle multi-source heterogeneous biological data, and how to evaluate the performance and reliability of models. Therefore, it is of great theoretical and practical significance to deeply explore the application of the integration of AI and biological data in the prediction of disease targets, construct effective theoretical models and clarify their verification boundaries.

2. The Significance of AI and Biological Data Integration in Disease Target Prediction

2.1 Enhance Prediction Accuracy

Biological data have the characteristics of multi-source heterogeneity. Different types of data reflect the physiological and pathological states of organisms from different perspectives. For instance, gene expression data can reveal the expression changes of genes during the occurrence of diseases, and protein interaction data can demonstrate the functional associations between proteins. AI technology can integrate these multi-source heterogeneous biological data, explore potential correlations and complex patterns among the data, and thereby identify disease targets more comprehensively and accurately. Through machine learning or deep learning algorithms, models can learn the subtle features and patterns in the data, improving the accuracy of predicting disease targets.

2.2 Accelerate the Drug Research and Development Process

Drug development is a long and costly process. Traditional drug development methods require a significant amount of time and funds for target screening, drug design, and clinical trials, among other steps. The integration of AI and biological data can accelerate the early stage of drug development, that is, the discovery and validation of disease targets. By rapidly and accurately predicting disease targets, a clear direction can be provided for drug development, reducing the time and cost of blind screening. Meanwhile, AI technology can also simulate and predict the mechanism of action of drugs, helping to optimize drug design and increase the success rate of drug research and development.

2.3 Promote the Development of Personalized Medicine

Everyone's genome and biological characteristics are different, so the pathogenesis of diseases and treatment responses also vary from person to person. The integration of AI and biological data can combine an individual's multi-omics data (such as genomic, transcriptomic, proteomic, etc.) and clinical information to achieve personalized prediction of disease targets. Based on individual characteristics, personalized treatment plans are formulated for patients to enhance the pertinence and effectiveness of treatment, reduce the occurrence of adverse reactions, and promote the development of personalized medicine.

2.4 Reveal the Complex Mechanisms of Diseases

The occurrence and development of diseases is a complex biological process involving the interaction of multiple genes, proteins and signaling pathways. Traditional biological research methods often can only focus on the local mechanisms of diseases and are difficult to comprehensively reveal the complex networks of diseases. The integration of AI and biological data can consolidate multi-level biological data, construct disease-related biological network models, and understand the occurrence and development mechanisms of diseases at the system level. By analyzing the key nodes and pathways in biological networks, new disease targets and therapeutic targets can be discovered, providing new ideas and methods for disease treatment.

3. Key Elements for Constructing Theoretical Models of Disease Target Prediction

3.1 Data Preprocessing

Biological data comes from a wide range of sources and varies in quality. There may be problems such as missing values, outliers, and duplicate values. Data cleaning is an important step in data preprocessing, aiming to remove noise and errors from the data and improve data quality. For missing values, methods such as deleting samples containing missing values, filling in the mean or media can be adopted for handling. Outliers can be identified and corrected through statistical methods or visual analysis. For duplicate values, directly delete the duplicate records.

Biological data from different sources have different dimensions and distribution characteristics. Direct fusion and analysis may lead to some features having an excessive impact on the model, while the influence of others is ignored.

Data normalization is capable of transforming data with diverse characteristics into a uniform scale range. Widely used normalization techniques encompass Z-score normalization and Min-Max normalization, among others. Z-score normalization adjusts the data to conform to a distribution with an average of 0 and a standard deviation of 1. Conversely, Min-Max normalization linearly projects the data onto the range [10, 1].

Biological data have the characteristics of multi-source isomerism, including gene expression data, protein interaction data, metabolomics data, etc. Data integration is the process of combining data from different data sources to form a unified data set for subsequent analysis and modeling. Data integration needs to take into account issues such as data consistency, integrity and redundancy. Methods such as database technology, data warehouse technology or data fusion algorithms can be adopted to achieve data integration.

3.2 Feature Selection and Extraction

3.2.1 Feature Selection

Biological data usually have the characteristic of high dimensionality and contain a large number of features, but many of these features may be irrelevant or redundant to disease target prediction. The purpose of feature selection is to screen out the features most relevant to disease target prediction from the original features, reduce the feature dimension, and improve the training efficiency and prediction performance of the model. Common feature selection methods include filtering, packaging and embedding, etc. The filtering method screens features based on their statistical properties (such as variance, correlation, etc.). The packaging method selects the optimal feature subset by constantly trying different feature subsets and evaluating the performance of the model. The embedding method combines the feature selection process with the model training process and automatically selects important features during the model training process.

3.2.2 Feature Extraction

Feature transformation is the procedure of converting initial features into novel feature representations, aiming to more effectively uncover the inherent structure and patterns embedded within the data. In contrast to feature selection, which directly eliminates original features, feature transformation creates new features by applying linear or nonlinear mappings. Widely recognized feature transformation techniques encompass linear approaches like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA), along with nonlinear methods found in deep learning, such as Autoencoders and Convolutional Neural Networks (CNNs). Deep learning approaches possess the capability to autonomously learn intricate feature representations from data and exhibit extensive potential for application in the feature extraction of biological datasets.

3.3 Model Algorithm Selection

3.3.1 Machine Learning Algorithms

Machine learning techniques are extensively utilized for predicting disease targets. Among the frequently employed machine learning algorithms are Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB), among others. SVM operates by identifying an optimal hyperplane to distinguish between samples of different classes, demonstrating strong generalization capabilities. RF, as an ensemble learning method, improves model accuracy and robustness by building multiple decision trees and aggregating their predictions through voting. LR is particularly suited for binary classification tasks, providing the likelihood of a sample belonging to a specific category. NB relies on Bayes' theorem and assumes feature independence, offering a straightforward and computationally efficient classification approach.

3.3.2 Deep Learning Algorithms

Deep learning methodologies exhibit remarkable proficiency in automatically extracting features and performing nonlinear modeling, making them particularly adept at managing extensive and high-dimensional biological datasets. Widely adopted deep learning algorithms encompass deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), along with their derivatives (such as Long Short-Term Memory networks (LSTMs) and Gated Recurrent Units (GRUs), among others. DNN is a multi-layer neural network structure that can learn complex nonlinear relationships in data. CNN is suitable for processing data with local correlations, such as gene sequence data, image data, etc. RNNS and their variants are suitable for processing sequence data, such as time series gene expression data, protein sequence data, etc.

4. Discussion on the Validation Boundary of Theoretical Models for Disease Target Prediction

4.1 Verification Method

4.1.1 Cross-validation

Cross-validation stands as a prevalent technique for model validation. This approach partitions the dataset into k distinct subsets, systematically utilizing k-1 subsets for training and the remaining subset for validation in each iteration. It conducts such experiments and computes the average performance across these iterations as the ultimate assessment of the model's efficacy. Notable cross-validation strategies encompass K-fold cross-validation and leave-one-out cross-validation, among others. K-fold cross-validation evenly splits the dataset into k parts, thereby optimizing the utilization of the dataset for comprehensive model evaluation. Leave-one-out cross-validation represents a specific instance of K-fold cross-validation, where k is set equal to the total number of samples, leaving only one sample as the validation set in each round, a method particularly apt for datasets with limited samples.

4.1.2 Independent Test Set Verification

Independent test set validation is to validate a model using a test set that is completely independent of the training set. The independent test set should have a data distribution similar to that of the training set, but the samples cannot be duplicated with those of the training set. Verification through an independent test set can more accurately assess the performance of the model in practical applications and avoid overestimation of model performance due to data leakage.

4.2 Verification Indicators

4.2.1 Classify Task Indicators

When predicting disease targets, the task is typically reframed as a classification problem, focusing on identifying whether a specific molecule serves as a disease target. Widely adopted classification metrics encompass accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic curve (ROC-AUC), among others. Accuracy denotes the ratio of correctly predicted samples to the total number of samples. Precision measures the proportion of samples correctly predicted as positive that are indeed positive. Recall indicates the ratio of actual positive samples that are accurately predicted as positive. The F1 score represents the harmonic means of precision and recall, offering a balanced evaluation of both metrics. ROC-AUC reflects the model's capacity to differentiate between positive and negative classes across various thresholds, with values ranging from 0 to 1; the closer it is to 1, the superior the model's performance.

4.2.2 Regression Task Indicators

When reframing the disease target prediction challenge as a regression problem, where the goal is to forecast continuous variables like the activity or binding affinity of disease targets, frequently employed regression metrics consist of mean squared error (MSE), mean absolute error (MAE), and the coefficient of determination (R^2), among others. MSE and MAE serve to quantify the extent of the discrepancy between predicted and actual values. R^2, on the other hand, gauges how well the model fits the data, with its values spanning from 0 to 1; the closer it approaches 1, the more effectively the model fits the data.

4.3 Evaluation of Model Generalization Ability

4.3.1 Definition of Generalization Ability

The generalization ability of a model refers to its predictive performance on unseen data, that is, whether the model can apply the knowledge and rules learned in the training set to new data. A model with good generalization ability can achieve good predictive performance on different datasets, not just on the training set.

4.3.2 Evaluation Methods

Apart from the cross-validation and independent test set validation approaches discussed earlier, the model's generalization capability can also be assessed by examining how its performance varies across distinct data subsets. For example, the dataset can be partitioned based on various characteristics (such as disease categories, sample origins, etc.), and the model's performance can be evaluated separately for each subset. If the model demonstrates relatively consistent performance across all subsets, it suggests that the model possesses strong generalization ability. On the contrary, if the

model's performance drops notably on certain subsets, this indicates that the model may be suffering from overfitting and has limited generalization ability.

5. Conclusion

The integration of AI and biological data brings new opportunities and challenges to the prediction of disease targets. By integrating multi-source heterogeneous biological data, AI technology can enhance the accuracy of disease target prediction, accelerate the drug development process, promote the development of personalized medicine, and reveal the complex mechanisms of diseases. In the process of constructing theoretical models for disease target prediction, key elements such as data preprocessing, feature selection and extraction, and model algorithm selection have a significant impact on the performance of the model. At the same time, to ensure the validity and reliability of the model, it is necessary to conduct strict validation of the model, clarify the validation boundaries, and adopt appropriate validation methods and indicators to evaluate the performance and generalization ability of the model.

However, the integration of AI and biological data still faces some problems in the field of disease target prediction, such as the quality and annotation of biological data, the interpretability of models, and the challenges of multimodal data fusion. Future research should be dedicated to addressing these issues and further enhancing the accuracy and reliability of disease target prediction. With the continuous development of AI technology and biotechnology, it is believed that the integration of AI and biological data will play a greater role in the field of disease target prediction, bringing new breakthroughs to biomedical research and clinical practice.

REFERENCES

[1] Peeriga, R., Manubolu, K., Shaik, A., Arige, S., Mohammed, J., & Kumar, V. L. V. Target Identification and Validation. In Innovations in Drug Discovery (pp. 38-53). Routledge.
[2] Liu, B., Li, S., & Hu, J. (2004). Technological advances in high-throughput screening. American Journal of Pharmacogenomics, 4(4), 263-276.
[3] Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. Bioinformatics and biology insights, 14, 1177932219899051.
[4] Vahabi, N., & Michailidis, G. (2022). Unsupervised multi-omics data integration methods: a comprehensive review. Frontiers in genetics, 13, 854752.
[5] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.
[6] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.
[7] Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., ... & Shi, T. (2018). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Frontiers in genetics, 9, 477.