TheSainsburyLaboratory
TSL

Effector Proteins Prediction Using Deep Learning

Research Project Plan

Introduction

In order to gain entry into the plant interior, colonize diverse tissues, and cause disease, pathogens must be able to disable plant defense responses. A critical component required for pathogenesis is the secretion of pathogen proteins, called effectors, which modulate plant immunity and facilitate infection [1]. Additionaly, when effector proteins are translocated into host cells, they manipulate the host defence systems [2]. Since the effectors play important roles in pathogen–host interactions, identifying them is crucial to understand how these effector proteins manipulate the host cell, and more generally to understand the pathogenic mechanisms of these effector [3]. Classical strategies for the identification of effectors is using the experimental approaches including (for fungal effectors) map-based cloning, analysis of fungal secretomes during infection, identification of hypersensitive response(HR)-inducing pathogen genes, mutagenesis and screening of expressed sequence tag (EST) libraries [4] and (for bacterial effectors) detection of common promoter elements [5]. However, since implementation of these approaches is often expensive and time-consuming, therefore computational tools need to be developed to tackle this issue.

Goal

Developing a computational tools to predict effector proteins based on amino acids sequence using deep learning methods.

Hypothesis

Using deep learning to train the amino acids sequence data of effector proteins will outperform traditional machine learning approaches and result in both higher accuracy and efficiency.

Rationale

The hypothesis is based on both the widely use of deep learning over past few years in bioinformatics, computational biology and medical informatics community [3] and its robust performance as well as its superiority in dealing with biological sequence data, such as DNA, RNA, and protein sequence data [6]. Moreover, the most attractive aspect of deep learning methods is their ability to perform these tasks without time-intensive feature engineering as any standard machine learning technique require more feature engineering skills [7].

TheSainsburyLaboratory
TSL

Research Plan

In order to achieve the objective stated previously, the following research plan needs to be done:

1. Understand the dataset of effector proteins' amino acid sequences (stastical summary) and analyze what type of classification will be binary or multiclass classification problems (define the problem).
2. Reprocess the data:
    - Split the data into training, development, and testing (the ratio will depend on the number of data we have).
    - Encode the sequence data (can be done in various ways).
    - Balance the data (if they are imbalance).
    - Shuffle the data.
    - Flatten the data (if necessary).
3. Develop deep learning models and use those models to train and evaluate the data.
4. Once the base models found, then use the hyperparamaters scan to get the best hyperparameters setting that give the best accuracy values.
5. Analyze and visualize the results.
6. Make the report.

TheSainsburyLaboratory
TSL

References

[1] Tania Y. Toruño, Ioannis Stergiopoulos and G. Coaker. Plant-Pathogen Effectors: Cellular Probes Interfering with Plant Defenses in Spatial and Temporal Manners:419–441, 2017. DOI: 10.1146/annurev-phyto-080615-100204.

[2] Z. Esna et al. Using an optimal set of features with a machine learning-based approach to predict effector proteins for Legionella pneumophila:1–12, 2019.

[3] L. Xue et al. Sequence analysis DeepT3 : deep convolutional neural networks accurately identify Gram-negative bacterial type III secreted effectors using the N-terminal sequence. (November):1–7, 2018. DOI: 10.1093/bioinformatics/bty931.

[4] D. G. O. Saunders et al. Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi. 7(1), 2012. DOI: 10.1371/journal.pone.0029847.

[5] J. E. Mcdermott et al. MINIREVIEW Computational Prediction of Type III and IV Secreted Effectors in Gram-Negative Bacteria. 79(1):23–32, 2011. DOI: 10.1128/IAI.00537-10.

[6] Y. Li et al. Deep learning in bioinformatics : introduction , application , and perspective in big data era, 2019.

[7] L. Deng and D. Yu. Deep Learning: Methods and Applications:3–4, 2013. DOI: 10.1561/2000000039. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DeepLearning-NowPublishing-Vol7-SIG-039.pdf.

[8] S. Kamoun. GOHREP - How to plan and manage a research project, 2013.