Human Protein Atlas challenge : Finding individual human cell differences in microscope images

Rohan Bhukar[1,2]

[1] BMED-6517 : Machine Learning in Biosciences, Georgia Institute of Technology
[2] Bioinformatics, School of Biological Sciences, EBB, Georgia Institute of Technology, GA-30332

rbhukar3@gatech.edu
https://bioinformatics.gatech.edu/graduate-students

Abstract- Proteins have important roles in almost all cellular processes. Different proteins interact at a specific location to perform a task, the exact outcome of that task depends on which proteins are present. Differences in subcellular distributions of one protein can give rise to functional heterogeneity between cells. Not only finding differences, but figuring out how and why they occur, is important for understanding how cells function, how diseases develop. Present machine learning models for classifying protein localization patterns in microscope images summarize the entire population of cells. Single-cell studies require precise classification models. The specific aim of this project is to develop models capable of segmenting and classifying each individual cell with accurate labels. With recent advancements in deep-learning algorithm development and single-cell revolution, models that can precisely classify patterns in each individual cell in image, will fundamentally improve our understanding of how cells function and how diseases develop.

Index Terms- Machine Learning, Image Classification, Protein localization, Deep Learning, Single-cell analysis

I. Motivation

More than 7.674 billion humans exist on this earth, and each of us is made up of trillions of cells. Like every individual is unique including identical twins, scientists observe differences between the genetically identical cells in our bodies. Differences in the location of proteins can give rise to cellular heterogeneity. Proteins are critical to almost all cellular processes, with many coming together at a specific location to perform a role, resulting in an outcome depending on protein composition at location. Different subcellular distributions of one protein can give rise to great functional heterogeneity between cells. Finding such differences, and figuring out how and why they occur, is important for understanding how cells function, how diseases develop, and ultimately how to develop better treatments for those diseases.

II. Problem Statement

Study of a single cell enables the discovery of mechanisms that are difficult to observe with multi-cellular research. This is a weakly supervised multi-label classification problem and a code competition. Solving single-cell image classification challenge will help to characterize single-cell heterogeneity in the dataset provided, by generating more accurate annotations of the subcellular localizations for thousands of human proteins in individual cells. This may accelerate the growing understanding of how human cells function and how diseases develop, both of which are critical to the research community.

III. Data

The Human Protein Atlas is an initiative based in Sweden aimed at mapping proteins in all human cells, tissues, and organs. The data in the Human Protein Atlas database is freely accessible. :

1) The data page on Kaggle provides a set of full size original images (a mix of 1728x1728, 2048x2048 and 3072x3072 PNG files) in train.zip and test.zip.
2) The image level labels from train.csv and the filenames for the test set from sample_submission.csv.
3) Public HPA images available to download in the notebook.
4) The 16-bit version of the training images are available here.
5) Each sample consists of four files, with every file representing a different filter on the subcellular protein patterns represented by the sample. The format is [filename]_[filter color].png for the PNG files. Colors are red for microtubule channels, blue for nuclei channels, yellow for Endoplasmic Reticulum (ER) channels, and green for protein of interest, as mentioned.

IV. Method

Deep learning is a part of the machine learning methods, based upon artificial neural networks with representation learning. A convolutional neural network (CNN) is a class of deep neural networks which is commonly applied to analyzing visual image data. Google has developed a faster bread of CNN architecture called NFNet (normalizer free network) that can be trained in larger batch sizes and stronger data augmentations. NFNet and other DNN architectures will be explored to generate a better classification model.

V. Expected outcome

Better image classifier model which predicts protein organelle localization labels for each cell in the image. There are in total 19 different labels present in the dataset, which will be used in classifying 17 different cell types of highly different morphology. The predicted labels of each individual cell within those images is the expected outcome.

VI. Timeline

1) **January 26, 2021 - Start Date**
2) **May 4, 2021 - Entry deadline.**
3) **May 4, 2021 - Team Merger deadline.**
4) **May 11, 2021 - Final submission deadline.**

VII. References

https://www.kaggle.com/c/hpa-single-cell-image-classification