PROTEIN STUDY IN CELL IMAGES USING PYTHON AND ML

Name: Debaraj Roy
Registration number: 19BCE1789
Name of organization : VIT Chennai
Location: Chennai, India
Email:
debaraj.roy2019@vitstudent.ac.in

Name : Amitesh Roy
Registration number: 19BCE1507
Name of organization : VIT Chennai
Location: Chennai, India
Email:
amitesh.roy2019@vitstudent.ac.in

Name: Mithesh Raja A
Registration number: 19BCE1799
Name of organization : VIT Chennai
Location: Chennai, India
Email:
mithesh.raja2019@vitstudent.ac.in

Abstract—**Proteins are the mode of expression of the genetic information. They perform a variety of duties in the cells such as they act as the structural components of cells, enzymes, hormones, pigments, storage proteins and some toxins in the cells. The aim of this project is to develop a model capable of classifying mixed patterns of proteins in microscopic images. This model can then be taken further to identify a protein's location from a high throughput image.**

Keywords—Protein, Cell, Organelles, Endoplasmic reticulum,

Channels, etc.

I. Introduction

Cells are the building blocks of any living being. And in today's time it has become very important to differentiate the cell organelles. For this we came up with a CBIR based protein synthesis in cell images using python and deep learning.

II. Motivation

Proteins are thought of as the "doers" in the human call, executing many functions that together enable life. Historically, classification of proteins has been limited to single patterns in one or a few cell types but in order to fully understand the complexity of the human cell, models must classify mixed patterns across a range of different human cells. Pinpointing sub cellular protein localisation from microscopy images is easy to the trained eye. However it is very challenging to automate. Once we're able to find the localisation Non of these protein localisations, it can be integrated with a smart-microscopy system to identifying a protein's location from a high throughout image. Therefore, the need is greater than ever for automation biomedical image analysis to accelerate the understanding of human cells and disease.

III. Literature Survey

The Cell Atlas in the Human Protein Atlas (HPA) program has built a dataset including millions of high-resolution immunofluorescent microscopy images of human cells which detail the subcellular distribution pattern of proteins. We can learn a lot by recognising this distribution pattern which is a crucial step towards tackling different diseases at a cellular level. UnNl now a lot of models have been proposed towards tackling this classification problem but not all of them look towards the class imbalance which is caused due to the rare cell structures that exist.[1] takes this class imbalance into consideration and uses transfer learning but the accuracy of this approach is not that high

In [2] when we want to recognise the patterns in the sub-cellular proteins it uses the approach of a hardware system that implements software dynamic programming Algorithms and uses BLSOUM Matrices (Block substitution matrices are used in sequence alignment of proteins and are used to align sequences of proteins and match them according to the) Although this approach uses Dynamic Programming the computation is much slower and given k sequences of length N the alignment algorithm takes ø(K2N2) and it doesn't work for images as the data needs to be present in the database and the feature extraction of the protein should be performed first before performing the alignment of the different types of the patterns in the proteins.

Although the paper states using VHDL for implementation this alignment and creating a hardware system this is rather consuming and acts as an extra machine which is not needed for our purpose one more con of using the database is that in case there is some abnormality in the given image of a specific protein it becomes difficult to identifying the given protein and that is why we use deep learning so that even when something doesn't exist in the database our model should be able to perform image classification for the given protein. Analysis of the human genome results in identification of many potential drug targets. Target-based drug discovery is a commonly used technique as it reduces cost of laboratory experiments. Based on recent studies, most durable proteins are classified as functional proteins. Enzymes and GPCRs are the most significant target proteins and more than half of drug targets are categorised under only two protein families: 'receptors and ion channels'.Here,the training dataset includes both positive (proteins which can interact with drugs) and negative(can't be considered as drug targets). 1611 proteins approved as drug targets are stored in the Drug Bank database. 1224 are considered as positive sample sets.

Here we evaluated SVM, Neural Network (NN), k-Nearest Neighbourhood (kNN), Naïve Bayes, RF, and Decision Tree (DT) algorithms to determine which was best.

Analysis of the human genome results in identification of many potential drug targets. Target-based drug discovery is a commonly used technique as it reduces cost of laboratory experiments. Based on recent studies, most durable proteins are classified as functional proteins. Enzymes and GPCRs are the most significant target proteins and more than half of drug targets recategorised underonlytwoproteinfamilies:'receptors and ion channels'. Here, the training data set includes both positive (proteins which can interact with drugs) and negative (can't be considered as drug targets). 1611 proteins approved as drug targets are stored in the DrugBank database. 1224 are considered as positive sample sets.

Here we evaluated SVM, Neural Network (NN), k-Nearest Neighbourhood (kNN), Naïve Bayes, RF, and Decision Tree (DT) algorithms to determine which was best.

In [4], a systematic attempt has been made to predict viral-host Protein- Protein interaction (PPI). Three well known machine learning methods, SVM, Naive Bayes, Random Forest were used. The specificity of Naive Bayes was highest (99.52%) as compared to the other two.

Unknown targets of hepatitis B, hepatitis E PPI were predicted.

The proposed method can predict large scale interspecies viral-human PPIs. The nature and function of unknown viral proteins (HBV and HEV), interacting partners of host protein, were identified using an optimised SVM model. The aim of

[5] is to identify the four levels of protein structure characterisation: primary, secondary, terNary and quaternary structures using artificial neural networks (ANNs).

The two ANNS classify protein sequences into 698 UniProt families (AUC=99.99%) and 983 Gene Ontology classes (AUC=99.45%).

5. Then goes on to discuss various methods of applying neural networks for protein classification; they are frequently deployed for classification and nearly real time image and sound processing. Today's neural networks are mostly deep neural networks, meaning that they have a much larger number of layers than earlier variations, resulting in a vastly increased learning capacity.

5. Then goes on to discuss various methods of applying neural networks for protein classification; they are frequently deployed for classification and nearly real time image and sound processing. Today's neural networks are mostly deep neural networks, meaning that they have a much larger number of layers than earlier variations, resulting in a vastly increased learning capacity. If one

the neural network can be trained to perform its classification task.specifies the non- linear functions of the neutrons, and the Architecture of the network, then the neural network can be trained to perform its classification task. This learning capability is the most appealing property of neural nets.

They then go on to train their model using a dataset downloaded from the UniProt protein database. Two models were trained on this dataset: one for Gene Ontology functional classification and one for UniProt family classification.

There was a logical relation among the attributes (functions/families) in both cases, describable using a directed acyclic graph (DAG), where each edge signifies an implication: for each edge A ! B, if an entry belongs to the class (has attributes) A, then it will also belong to the class B.

The results obtained were highly accurate, outperformed the existing solutions and have attained a near 100% of accuracy in multi label, multi family classification. Simplified the network architecture. Unlisted a deeper neural networks with more parameters have a much larger capacity for learning good representations but can only distinguished two ways

6. Focuses on utilising a deep convolutional neural network (DeepLoc) to analyse yeast cell images. They do this because existing computational pipelines for quantitative analysis of high content microscopy data rely on traditional machine learning approaches that fail to accurately classify more than a single dataset without substantial tuning and training, requiring

IV. PROPOSED METHODOLOGY

For this project we have used the InceptionResnetV2 architecture. The InceptionResnetV2 network is considerably deeper than the InceptionV3 network and is a perfect fit for this task with the ability to provide higher accuracy.

The main deep learning frameworks unlisted in this project are Keras and Tensorflow.

The start is to import all the desired libraries and frameworks. Second, we tend to produce a perform referred to as create model, that utilises the

pertained model origin ResnetV2.

We then stack a dense 2nd convolution neural network to the present

pertained model. The activation perform unlisted within the network is ReLu

To stop this model from over fitting to the data, we tend to use the regularisation technique drop out.

To stop covariate shift, batch standardisation is used. It normalises the activation of every layer.

So our entire model can have the subsequent design during this sequence:

1. Input Layer

2. Batch standardisaTIon Layer

3. Inception Resnet_v2 Model

4. 2D convolutional Neural Network

v) Flatten Layer to scale the values into a tensor vi) Dropout Layer1 vii) Dense Layer one - totally connected Layer viii) Dropout Layer2 ix) Dense Layer two - totally connected Layer.

x) The ultimate activation in perform is that the sigmoid function for the final classification.

We retrain this model with the binary cross entropy loss function.

Since this is often a classification task, we've additionally created use of the binary cross entropy loss Perform to scale back the loss whereas training.

We predict the legion's categories for the new set of images. During this sense, the scores are the chances created by the sigmoid function.

Finally, all those categories with a expected score worth from the sigmoid of bigger than 0.2 are thought of because the classes for that exact image.

We calculate the F1 score to envision the accuracy of the model and create numerous plots to search out the relation between the coaching and therefore the validation loss, accuracy and F1 score.

V. Proposed Methodology

Dataset: [*https://github.com/sequencer55/CBIR_J_COMPONENT_19BCE1789/tree/main/Dataset%20of%20cell*](https://github.com/sequencer55/CBIR_J_COMPONENT_19BCE1789/tree/main/Dataset%20of%20cell)

VI. Hardware and software requirements

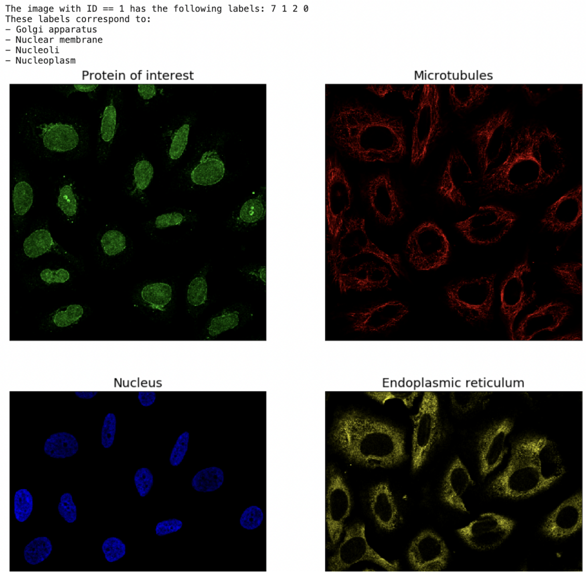
For hardware:

8GB Ram 256GB minimum storage and i5 6th gen processor

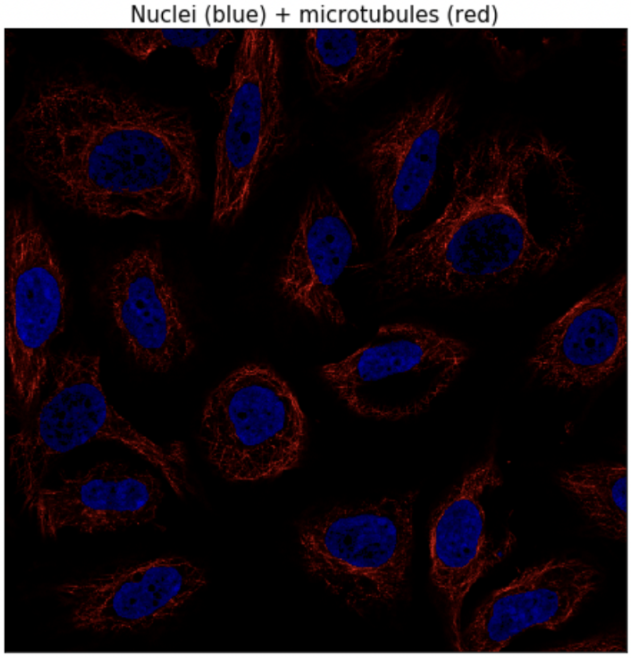
For software:

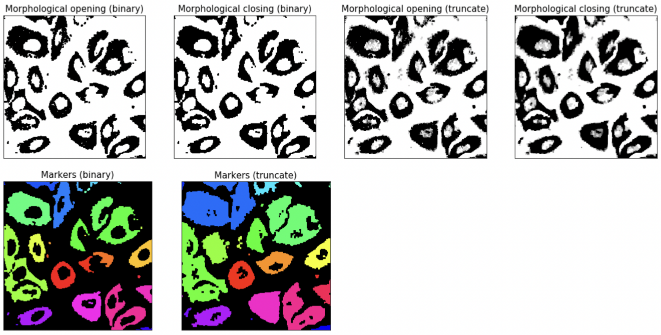
windows 10 home or mac OS

VIII. Results obtained:

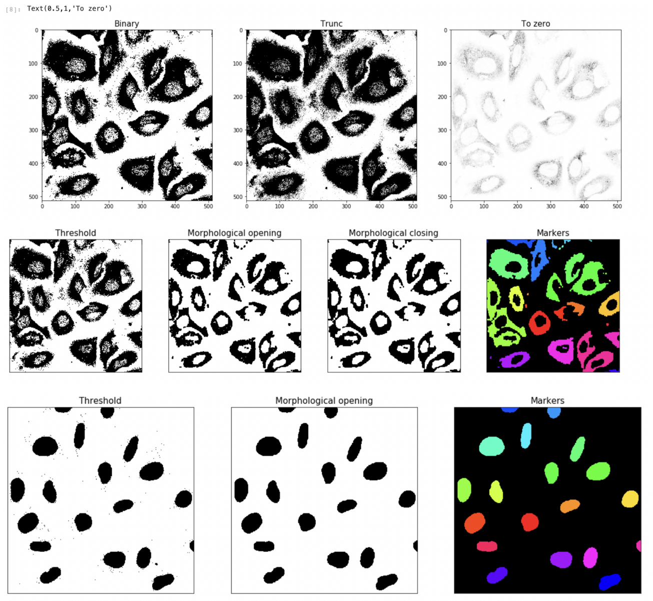


```
The image with ID == 1 has the following labels: 7 1 2 0
These labels correspond to:
- Golgi apparatus
- Nuclear membrane
- Nucleoli
- Nucleoplasm
```

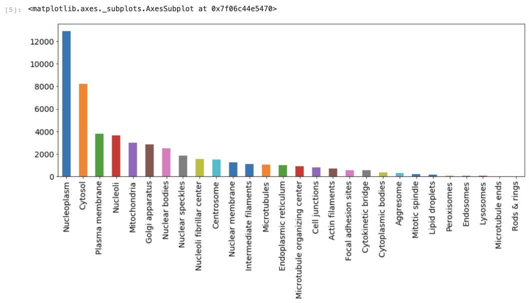




Morphological opening (binary)
Morphological closing (binary)
Morphological opening (truncate)
Morphological closing (truncate)
Markers (binary)
Markers (truncate)



```
[8]: Text(0.5,1,'To zero')
```



```
[5]: <matplotlib.axes._subplots.AxesSubplot at 0x7f06c44e5470>
```

IX. Conclusion

Today's neural networks are mostly deep neural networks, meaning that they have a much larger number of layers than earlier variations, resulting in a vastly increased learning capacity. If one the neural network can be trained to perform its classification task.specifies the non- linear functions of the neutrons, and the Architecture of the network, then the neural network can be trained to perform its classification task. This learning capability is the most appealing property of neural nets. They then go on to train their model using a dataset downloaded from the UniProt protein database. Two models were trained on this dataset: one for Gene Ontology functional classification and one for UniProt family classification. There was a logical relation among the attributes (functions/families) in both cases, describable using a directed acyclic graph (DAG),where each edge signifies an implication: for each edge A ! B, if an entry belongs to the class (has attributes) A, then it will also belong to the class B.

The results obtained were highly accurate, outperformed the existing solutions and have attained a near 100% of accuracy in multi label, multi family classification. Simplified the network architecture.

Unlisted a deeper neural networks with more parameters have a much larger capacity for learning good representations t

The results achieved by our model, as we can see our initial accuracy for the first epoch itself is 90.4 percent which helps us realise our transfer learning approach has been successful on running a total of 15 Epochs we conclude with an accuracy of 94.13 percent on the train set.

Acknowledgment

We would like to thank respected Dr.Geetha mam who has helped us throughout the journey and gave us all the assistance possible. It has been a truly knowledgeable journey.

Under her excellent guidance we have truly learnt a lot and we hope to work under her again.

References

1. Classification of Subcellular Protein Patterns in Human Cells with Transfer

Learning

Authors: Li, Zongyao, Togo, Ren, Ogawa, Takahiro, Haseyama, Miki

Source: 2019 IEEE 1st Global Conference on Life Sciences and Technologies (LifeTech)

Life Sciences and Technologies (LifeTech), 2019 IEEE 1st Global Conference on.

:273-274 Mar, 2019

Publisher Information: IEEE

Publication Year: 2019

Link: http://ieeexplore.ieee.org.egateway.vit.ac.in/document/8884002/

2. BioCircuit - A Hardware Based Methodology for Protein Recognition

Authors: Gajda, Dominik ,Pulka, Andrzej

Source: 2018 International Conference on Signals and Electronic Systems (ICSES) Signals

and Electronic Systems (ICSES), 2018 International Conference on. :289-294 Sep, 2018

Link: http://ieeexplore.ieee.org.egateway.vit.ac.in/document/8507340

3.DrugMiner[LM1]: comparative analysis of machine-learning algorithms for

prediction of potential druggable proteins

Authors:Ali Akbar Jamali , Reza Ferdousi , Saeid Razzaghi , Jiuyong Li , Reza

Safdari , and Esmaeil Ebrahimie

Publisher Info: Drug Discovery Today (2016), http://dx.doi.org/10.1016/

j.drudis.2016.01.007

Publishing year: 2016

4. Prediction of Interactions between Viral and Host Proteins Using Supervised Machine

Learning Methods

Authors: Ranjan Kumar Barman , Sudipto Saha, Santasabuj Das

Source: https://doi.org/10.1371/journal.pone.0112034

Publication year: 2014

5. Near Perfect Protein Multi-Label Classification with Deep Neural Networks

Authors: Balázs Szalkai, Vince

Grolmusz Source: https://arxiv.org/abs/1703.10663

Publication year: 2017

6. Automated analysis of high content microscopy data with deep learning

Authors: Oren Z Kraus ,Ben T Grys, Jimmy Ba, Yolanda Chong, Brendan J

Frey, Charles Boone Brenda J Andrews

Source: https://www.embopress.org/doi/full/10.15252/msb.2017751

Publication year: 2017

Reference papers

https://github.com/sequencer55/CBIR_J_COMPONENT_19BCE1789/tree/main/Reference%20Papers