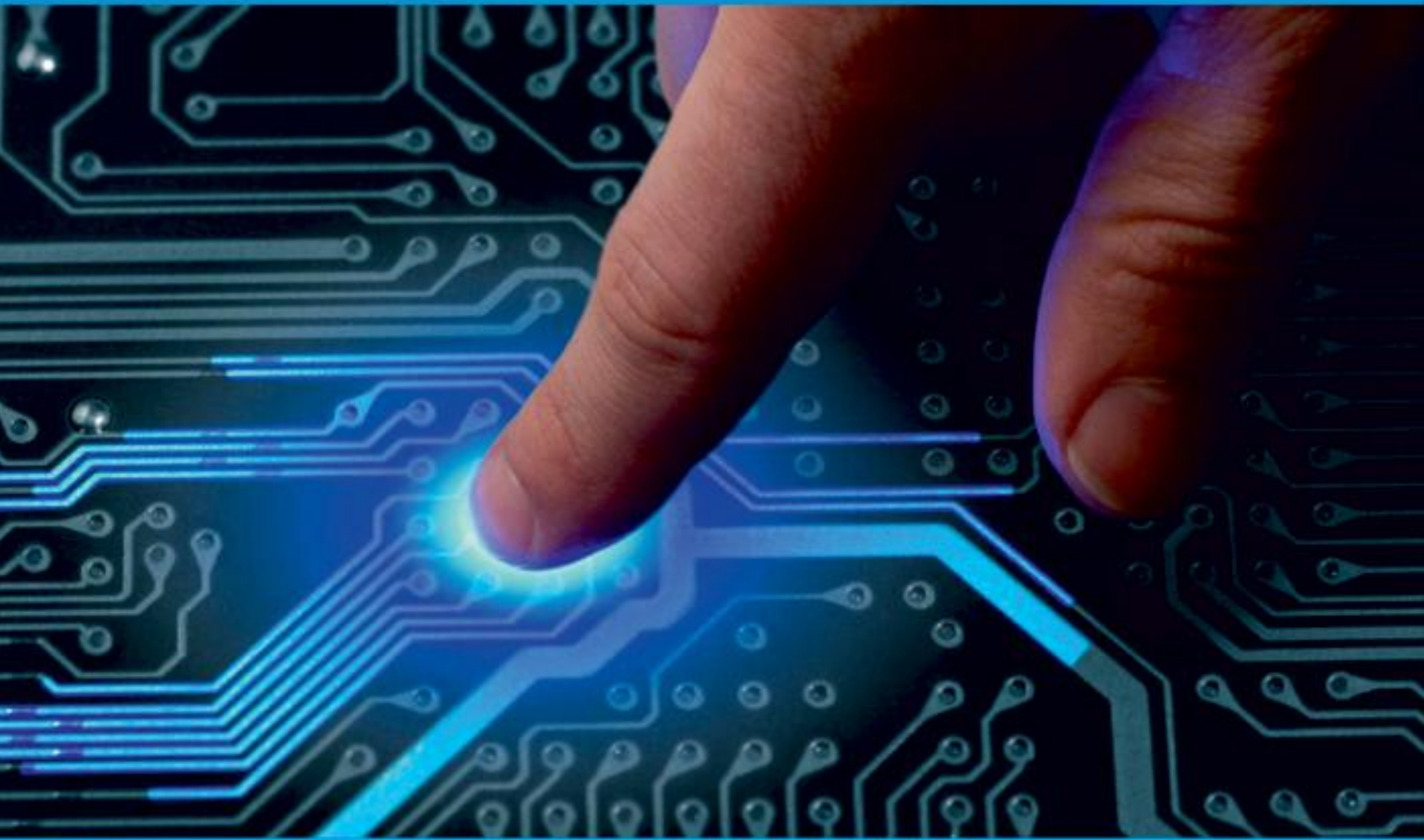
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 6, June 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0906048 |

Implementation of Protein Function Detection and Classification using CNN Algorithm

Keerthana M M

Assistant Professor, Dept. of CSE., ATME College of Engineering, Mysore, Karnataka, India

ABSTRACT: Proteomics research has become one of the most important topics in the fields of life science and natural science. The project determined that proteins participate in life activities mainly in the form of complexes. At present, research on protein–protein interaction networks (PPINs) have mainly focused on detecting protein complexes or function modules. This problem has been transformed into a recognizable dense subgraph problem in a PPIN diagram. The situation in PPIN research in recent years is introduced in this study, including commonly used databases, traditional detection algorithms, recent solutions, and the application of the swarm intelligence algorithms in this field. We then propose a detection scheme based on particle swarm optimization (PSO) and gene ontology knowledge. This scheme combines PSO and biological gene ontology knowledge to identify complexes from PPINs. Simultaneously, network topology knowledge improves the detection
accuracy of the protein module.

KEYWORDS: Protein–Protein Interaction Networks(PPIN), Protein Function Module(PFM), Particle Swarm Optimization(PSO), Machine Learning(ML).

I. INTRODUCTION

Man has entered a post-genomic era with the completion of the Human Genome Project. At present, proteomics has become one of the most important research topics in the fields of life science and natural science. Proteomics involves the systematic study of the characteristics of proteins. It aims to provide a detailed description of the structures, functions, and regulations of biological systems under healthy and diseased conditions. The importance of proteins, as one of the indispensable materials in life, is self-evident. Proteins are involved in nearly every process of life activities, such as the replication and transport of genetic materials, the control of gene expressions, signal transduction, metabolism, and energy storage, among others. All these activities are dependent on protein functions. However, the project found that a single protein alone does not embody protein function. Different proteins form protein complexes through various interactions. These complexes satisfy the demands of diverse functions in life processes. Consequently, research on and analysis of protein–protein interaction (PPI) or PPI network (PPIN) have provided the basis to understand the organization, process, and function of cells in life activities. Moreover, research on PPI can promote the study of the mechanisms of various diseases and new drug development processes. The increasing amount of protein interaction data contribute to PPIN formation. PPIN is similar to other networks because it also has scale-free and small-world features. At present, research on PPINs can identify complexes from a large number of protein data by adopting the clustering of PPINs as the main approach. Authoritative international publications and conferences, such as Nature, Science, Proceedings of the National Academy of Sciences, and Nucleic Acids Research, among others, consider research on PPIN an important subject.

II. RELATED WORK

[1] Decision Tree Based Approaches for Detecting Protein Complex in Protein Interaction Network (PPI) via Link and Sequence Analysis.

Author: Aisha Sikandar, Waqas Anwar, Usama IjazBajwa , Xuan Wang. Year of Conference : 2018

A network of modular protein complexes inside a cell coordinates many biological processes and is known as PPI network.A PPI network can be modeled as a graph, in which edges represent interactions among proteins, and sub graphs represent protein complexes.Previous methods for protein complex mining from PPI network mainly focused on few topological features like density and degree statistics based on the assumption that proteins inside a complex are highly interactive with each other and thus form dense subgraphs.While this assumption is true for some complexes, it doesn‟t hold for many others.The important biological information within the protein amino acid sequences, which estimates the interacting property among two proteins for performing a specific biological function isn‟t considered in most of the previous studies.There is a need for algorithms that consider both topological and biological features for correctly identifying protein complexes having varying topological structures and biological patterns inside a PPI network.In this study, we present an algorithm for detecting protein complexes from interaction graphs. By using graph topological patterns and biological properties as features, we model each complex sub graph by decision tree learners.node.

[2] Protein quantitative based on simulated annealingalgorithm
Author: MingyuShao , Yi Yang , Jihong Guan† and Shuigeng Zhou. Year of Conference : 2017
Proteomics is a hot pot topic in current, its development experienced from proteins qualitative research to quantitative research.Label-free quantification method is the most widely used protein quantification method. But during the process of label-free quantification, lots of Mass spectrometry (MS) data do not be used so that cause the waste of data resource.This paper, a quantitative algorithm of protein based on simulated annealing algorithm was proposed to improve the efficiency of the use of MS spectral.This paper is about using all possible peptide ions to extract quantitative information from MS spectral.Verified by Experimental data set about that this protein quantification algorithm can increase the coverage of proteins on the condition of ensured accuracy.This paper focused on the development of proteomics and the analysis and excavation of mass spectrum data, its study results can be widely applied in protein qualification area.

[3]**Single-particle electron microscopy in the study of membrane proteinstructure**
Author: Rita De Zorzi, Wei Mi1, Maofu Liao1, and Thomas Walz. Year of Conference: 2015
Single-particle electron microscopy (EM) provides the great advantage that protein structure can be studied without the need to grow crystals. However, due to technical limitations, this approach played only a minor role in the study of membrane protein structure.This situation has recently changed dramatically with the introduction of direct electron detection device cameras, which allow images of unprecedented quality tobe recorded, also making software algorithms, such as three-dimensional classification and structure refinement, much more powerful.The enhanced potential of single-particle EM was impressively demonstrated by delivering the first long-sought atomic model of a member of the bio medically important transient receptor potential channel family.Structures of several more membrane proteins followed in short order. This review recounts the history of single-particle EM in the study of membrane proteins, describes the technical advances that now allow this approach to generate atomic models of membrane proteins and provides a brief overview of some of the membrane protein structures that have been studied by single-particle EM to date.

[4] **A Comparison Study on Protein-protein Interaction Network Models.**
Author: Mingyu Shao, Yi Yang ,Jihong Guan† and Shuigeng Zhou . Year of Conference : 2015
This paper presents a comprehensive comparison study on the performances of major existing models over two PPI datasets, by comparing the global and local statistical properties of the original PPI networks and the model-reproduced ones.Our experimental results show that the DD model has best fitting ability while iSite model and STICKY model also fit well with the PPI datasets over most statistical properties.By comparing the statistical properties between the original PPI networks and the model-reproduced networks, we find that the DD model fits best with the PPI data while iSite model and STICKY model also fit well with the PPI datasets over most statistical properties.By analysing the embedded mechanisms of these models, we speculate that PPI networks exhibit "degree-weighted" behaviour and evolve by gene duplication and divergence.

III. Proposed Algorithm

3.1 CNN ALGORITHM

•In deep learning, a convolutional neural network (CNN, or ConvNet) is a class of deep neural networks, most commonly applied to analyzing visual imagery.
•They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics.
•A convolutional neural network consists of an input and an output layer, as well as multiple hidden layers. The hidden layers of a CNN typically consist of a series of convolutional layers that convolve with a multiplication or other dot product.
•The activation function is commonly a RELU layer, and is subsequently followed by additional convolutions such as pooling layers, fully connected layers and normalization layers, referred to as hidden layers because their inputs and outputs are masked by the activation function and final convolution.

•Though the layers are colloquially referred to as convolutions, this is only by convention. Each neuron in a neural network computes an output value by applying a specific function to the input values coming from the receptive field in the previous layer.
•The function that is applied to the input values is determined by a vector of weights and a bias (typically real numbers). Learning, in a neural network, progresses by making iterative adjustments to these biases and weights. The vector of weights and the bias are called filters and represent particular features of the input (e.g., a particular shape).

3.2 MODULES

A.*KERAS*

Keras is one of the leading high-level neural networks APIs. It is written in Python and supports multiple back-end neural network computation engines. The Model is the core Keras data structure. There are two main types of models available in Keras: the Sequential model, and the Model class used with the functional API.
Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML. Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.

- pip install keras

B.*TENSORFLOW*

TensorFlow is an open-source library for machine learning applications. TensorFlow applications can be written in a few languages: Python, Go, Java and C. Tensorflow supports both cpus and gpus. Goggle has even produced its own specialized hardware for computing in cloud,called Tensor processing unit.
Tensorflow is developed by the Google Brain team for internal Google use. It is released under the Apache License 2.0 on November 9, 2015. Tensorflow is Google Brain's second-generation system.1st Version of tensorflow was released on February 11, 2017.While the reference implementation runs on single devices, Tensorflow can run onmultiple CPU‟s and GPU (with optional CUDA and SYCL extensions for general-purpose computing on graphics processing units). TensorFlow is available on various platforms such as64-bit Linux, macOS, Windows, and mobile computing platforms including Android and iOS. The architecture of tensorflow allows the easy deployment of computation across a variety of platforms (CPU‟s, GPU‟s, TPU‟s), and from desktops - clusters of servers to mobile and edge devices.
Tensorflow computations are expressed as stateful dataflow graphs. The name Tensorflow derives from operations that such neural networks perform on multidimensional data arrays, which are referred to as tensors.

- pip install tensorflow –command

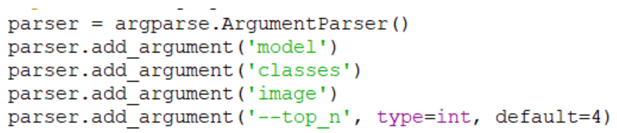
C. *XCEPTION MODEL*

Xception is a convolutional neural network that is 71 layers deep. Xception was proposed by none other than François Chollet himself, the creator and chief maintainer of the Keras library. The Xception architecture has 36 convolutional layers forming the feature extraction base of the network. In our experimental evaluation, we will exclusively investigate image classification.
Max pooling is a sample-based discretization process. The objective is to down-sample an input representation (image, hidden-layer output matrix, etc.), reducing its dimensionality and allowing for assumptions to be made about features contained in the sub-regions binned.
Softmax function, a wonderful activation function that turns numbers aka logits into probabilities that sum to one. Softmax function outputs a vector that represents the probability distributions of a list of potential outcomes.
Average pooling involves calculating the average for each patch of the feature map.

3.3 Steps Involved in Deployment of Code

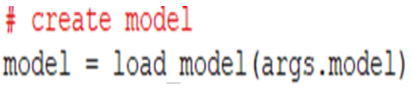
1. Get the Model & Required Paths via CMD.
2. Load the Model with weights.
3. Load the Labels.
4. Load the Input / Test Image.
5. Predict the Test Image.

3.3.1 Code Snippet of getting the Path Links Via CMD



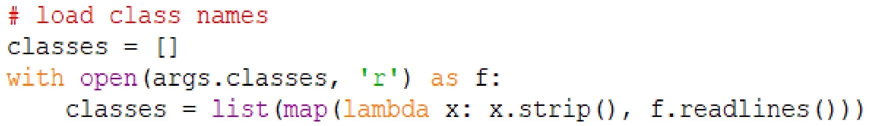
```
parser = argparse.ArgumentParser()
parser.add_argument('model')
parser.add_argument('classes')
parser.add_argument('image')
parser.add_argument('--top_n', type=int, default=4)
```

3.3.2 Code Snippet of Loading the Model



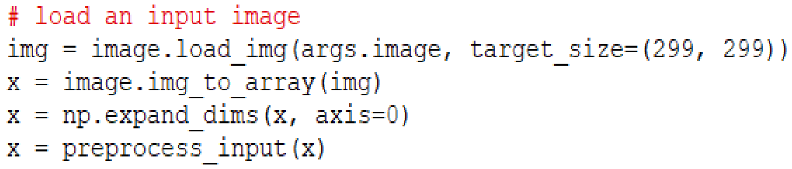
```
# create model
model = load_model(args.model)
```

3.3.3Code Snippet of Loading the Class Labels



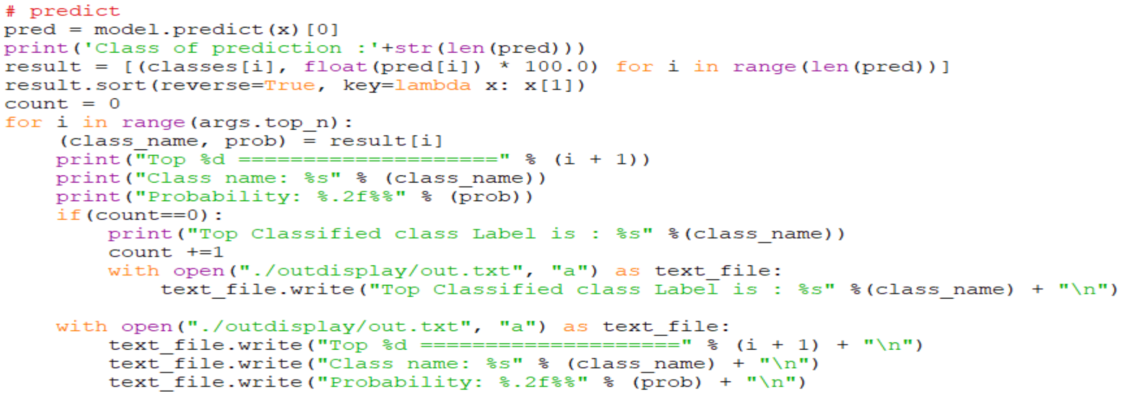
```
# load class names
classes = []
with open(args.classes, 'r') as f:
    classes = list(map(lambda x: x.strip(), f.readlines()))
```

3.3.4 Code Snippet of Loading the Test Image



```
# load an input image
img = image.load_img(args.image, target_size=(299, 299))
x = image.img_to_array(img)
x = np.expand_dims(x, axis=0)
x = preprocess_input(x)
```

3.3.5 Code Snippet of Predicting & Printing The Result

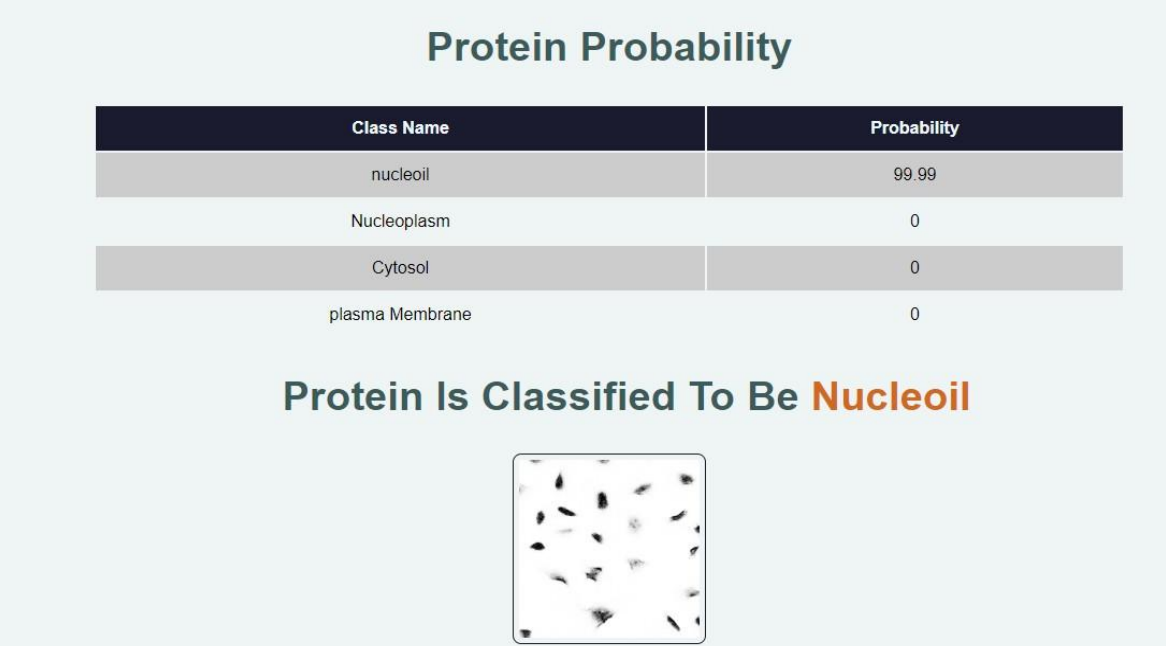


```
# predict
pred = model.predict(x)[0]
print('Class of prediction :'+str(len(pred)))
result = [(classes[i], float(pred[i]) * 100.0) for i in range(len(pred))]
result.sort(reverse=True, key=lambda x: x[1])
count = 0
for i in range(args.top_n):
    (class_name, prob) = result[i]
    print("Top %d ====================" % (i + 1))
    print("Class name: %s" % (class_name))
    print("Probability: %.2f%%" % (prob))
    if(count==0):
        print("Top Classified class Label is : %s" %(class_name))
        count +=1
        with open("./outdisplay/out.txt", "a") as text_file:
            text_file.write("Top Classified class Label is : %s" %(class_name) + "\n")

    with open("./outdisplay/out.txt", "a") as text_file:
        text_file.write("Top %d ====================" % (i + 1) + "\n")
        text_file.write("Class name: %s" % (class_name) + "\n")
        text_file.write("Probability: %.2f%%" % (prob) + "\n")
```

IV. Simulation Results

Load the testing data on to the training model,here, test image is compared with the training data set, the probability of nucleoli is higher than the other proteins, and hence the output predicted here is nucleoli, similarly the probability of cytosol is higher than the presence of other proteins, hence protein predicted as cytosol.

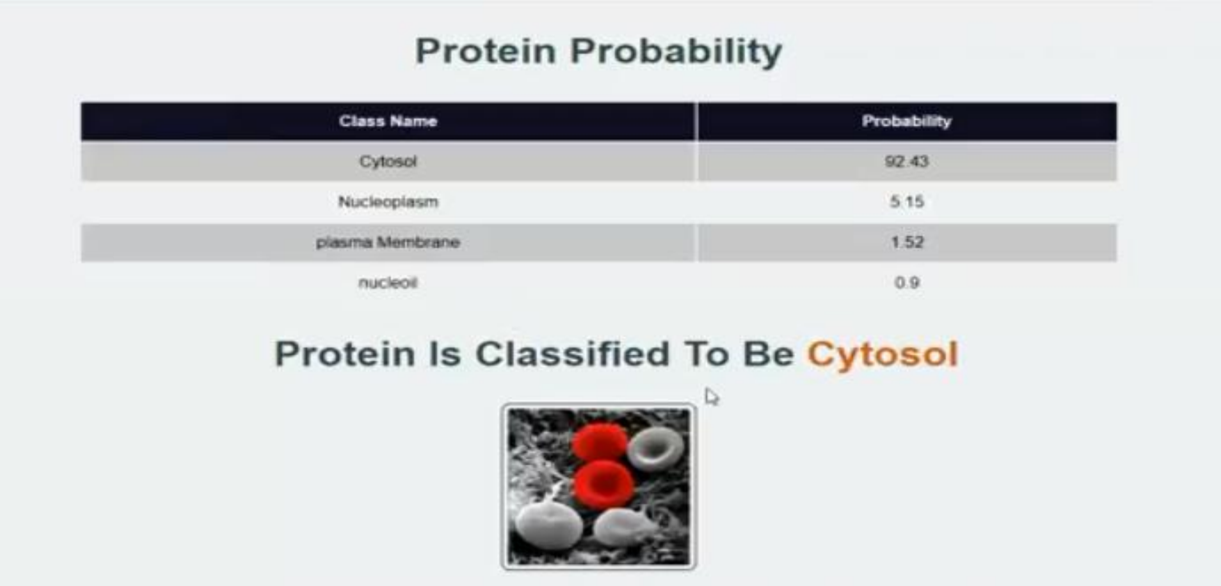


Protein Probability

Class Name	Probability
nucleoil	99.99
Nucleoplasm	0
Cytosol	0
plasma Membrane	0

Protein Is Classified To Be Nucleoil

Fig. 1: Predicted protein is Nucleoil



Protein Probability

Class Name	Probability
Cytosol	92.43
Nucleoplasm	5.15
plasma Membrane	1.52
nucleoil	0.9

Protein Is Classified To Be Cytosol

Fig. 2: Predicted protein is cytosol

V. Conclusion and Future Work

In this model, the identification of protein were proposed, that is complete network of proteins are identified in a tissue image. In this study we presented a CNN based approach for discovering complexes from PPI network. In CNN network there are three layers input, hidden and output layer. The classification and complete process will be done in the hidden layer. After processing we will get the output through the output layer. Thus CNN algorithm has made it possible to use this approach to identify the proteins. Microscopic tissue image is used to classify and identify the protein present in the image.Further , the model can be enhanced to analyze and compare the result to identify the protein in the tissue. Also other protein data will be collected and the existing model will be enhanced using more feature for identifying more proteins and identifying the amount of particular protein present in the tissue.

REFERENCES

1. Aisha Sikandar, Waqas Anwar, Usama IjazBajwa , Xuan Wang “Decision Tree Based Approaches for Detecting Protein Complex in Protein Interaction Network (PPI) via Link and Sequence Analysis”, IEEE 2018.
2. Mingyu Shao , Yi Yang , Jihong Guan† and Shuigeng Zhou “Protein quantitative based on simulated annealing algorithm” Published on 2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications (ISPA/IUCC)
3. Rita De Zorzi, Wei Mi1, Maofu Liao1, and Thomas Walz “Single-particle electron microscopy in the study of membrane protein structure” . Year of Conference : 2015,PMID: 26470917, PMCID: PMC4749050, DOI: 10.1093/jmicro/dfv058
4. Mingyu Shao, Yi Yang ,Jihong Guan and Shuigeng Zhou “ Comparison Study on Protein-protein Interaction Network Models” Conference: Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference.

BIOGRAPHY

Mrs.Keerthana M M is an assistant professor in Dept. of computer science and engineering in ATME College of Engineering Mysuru, Karnataka, India. She received her master degree in computer science and engineering from VKIT, Bangalore affiliated to VTU University, Belagavi. She has 6 years teaching experience and her field of interest is cloud computing, IOT and Machine learning.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

Scan to save the contact details

www.ijircce.com