Convolutional Neural Networks with Image Representation of Amino Acid Sequences for Protein Function Prediction

Samia Tasnim Sara
012172010

Department of Computer Science and Engineering
United International University

A thesis submitted for the degree of
MSc in Computer Science & Engineering

December 15, 2021

Abstract

Proteins are one of the most important molecules that govern the cellular processes in organisms. Various functions of the proteins are of paramount importance to understand the basics of life. Several supervised learning approaches are applied to this field to predict the functionality of proteins. In this thesis, we propose a convolutional neural network based approach protconv to predict the functionality of proteins by converting the amino-acid sequences to a two dimensional image. We have used a protein embedding technique using transfer learning to generate the feature vector. Feature vector is then converted into a square sized single channel image to be fed into a convolutional network. The neural network architecture used here is a combination of convolutional filters and average pooling layers followed by dense fully connected layers to predict a binary function. We have performed experiments on standard benchmark datasets taken from two very important protein function prediction task: proinflammatory cytokines and anticancer peptides. Our experiments shows that the proposed method, ProtConv achieves state-of-the-art performances on both of the datasets.

Acknowledgements

This work would have not been possible without the input and support of many people. I would like to express my gratitude to everyone who contributed to it in some way or other.

First of all, I would like to thank my supervisor Swakkhar Shatabda, Associate Professor and Undergraduate Program Coordinator United International University for the continuous support of my research. Without his guidance it was impossible for me to complete the research and writing of this thesis.

Besides my supervisor, my sincere gratitude goes to Prof. Dr. Mohammad Nurul Huda, Professor and Director - Master of Science in Computer Science and Engineering United International University, Dr. Dewan Md. Farid, Associate Professor, United International University & Rubaiya Rahtin Khan, Assistant Professor, United International University for critical reading of the manuscript.

Last but not the least, I owe to my family including my parents for their unconditional love and immense emotional support.

Published Paper

Work relating to the research presented in this thesis has been published/ submitted by the author in the following peer-reviewed journals and conferences:

Samia Tasnim Sara, Md Mehedi Hasan, Ahsan Ahmad, and Swakkhar Shatabda. Convolutional neural networks with image representation of amino acid sequences for protein function prediction. Computational Biology and Chemistry, 92:107494, 2021

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is one of the most lethal diseases, accounting for millions of deaths globally each year. Traditional chemotherapy is the primary cancer treatment technique at the moment. It has bad impact on good cells also. Anticancer peptides offer a promising new avenue for cancer treatment, as well as a number of appealing advantages. In this chapter, we introduce our project. In Section 1.1, the project overview is presented. In Section 1.2, we present the motivation of solving the problem using our method.

1.1 Problem Statement

In the case of protein function prediction, the tasks are often prediction their various properties like DNA binding [8], anticancer [9], secondary or tertiary structures [10], different types of post-translational modifications [6], subcellular localizations [5], etc. Different representations are used in the vast literature for the vectorization of the protein sequences with help of sequence based, physico-chemical, evolutionary and structural features.Recent developments in deep learning has enabled the researchers to apply knowledge from other domains into this area as well. However, one of the major bottlenecks here is the lack of a large dataset required to avoid overfitting. However, a few approaches have been their to use transfer learning from pre-trained models using generic prediction tasks [11, 12]. In a recent work, it has been found that DNA sequences converted into two dimensional images could be fed into convolutional networks [13]. Such conversion strengthens the classification methods by enabling use of computer vision based techniques. However, it also depends on the transformation technique applied to represent the basic sequence.

1.2 Motivation

Systems biology is a quickly growing field driven by the omics age and fresh technological developments that have enhanced information accuracy that can be established. Concentrated on easy single cell organisms like bacteria contributes to their tractability and implies the quickly maturing science of systems microbiology.
In the other hand, peptides play an dominant role in cancer formation, Unlike other therapies, peptides show superiority due to their specificity. Peptides which have anti tumor activity are called anticancer peptides.

All blood cells and other cells that aid the body's immunological and inflammatory responses are affected by cytokines. They also aid anti-cancer activities by transmitting signals that cause aberrant cells to expire while normal cells live longer.
Inflammatory peptides are molecules that plays a very important role in signaling the immune system against the pathogens. Regulation of inflammatory cytokines are often key to treat inflammatory diseases [1]. Computational methods have been recently proposed to reduce the cost and time needed in laboratory methods to identify pro-inflammatory peptides. ProInflam was proposed in [2] by Gupta et al. using Support Vector Machine Classifier and sequence based features. Manavalan et al. proposed PIP-EL [3]. In their work, they have also utilized several sequence based features.

Proteins are the most important and adaptable macromolecules in life, and understanding their roles is important for developing novel medications, improved crops, and even synthetic biochemicals. However, there hasn't been much research into protein function prediction, which could open up a whole new field of biomedicine. On the other hand computational biology is focused with finding answers to problems that have arisen as a result of bioinformatics research. It is important in scientific study, such as the simulation of protein folding, mobility, and interaction to see how proteins interact with one another. [4]

1.3 Objectives

Based on our observations of other thesis work, we believe that the accuracy of the other predictors can be improved, and that better tools for scientists can be developed. Also we didn't find enough work on Proinflammatory Cytokines, So we also want to if we can use this in future work.

So here is the objectives we want to cover in this thesis work:

1. To develop a new methodology for protein function prediction.
2. To test and validate the method with two different data set and open an area for future research.

1.4 Methodology

We will use transfer learning in our model as main method. Transfer learning has several advantages, but the major ones are that it saves training time, improves neural network performance (in most cases), and does not require a large amount of data. We will discuss the details in the following chapter.
In our thesis we will use convolutional neural network to test and validate the data. The important benefit of CNN over its predecessors is that it automatically discovers essential features without any user intercession. The brief discussion of details will be in the following chapter.
We chose two protein function prediction tasks: pro-inflammatory peptide prediction and anticancer peptide prediction. Both topics are described in the literature as binary classification problems, and a number of supervised learning methods have been used to solve them. In the following chapter, we briefly describe the problems and related datasets, as well as provide a glimpse of the state-of-the-art literature for both problems.

1.5 Project Outcome

In this paper, we present ProtConv, a convolutional neural network based protein function prediction method. We propose to convert the vector representation of the protein or peptide sequence into a two dimensional image with a single channel which is fed into the convolutional neural network. The convolutional neural network architecture that we use in ProtConv is a simple one inspired from Lenet-5 [5]. We have used TAPE [6] embeddings and pre-trained models to convert the protein or peptide sequence to a vector representation. The framework is simple and yet extendable. Any feature representation and network architecture can replace these proposed ones. We have tested the performance of the proposed method on two very important prediction tasks: identification of proinflammatory peptides and anticancer peptides. Both of the problems ProtConv were trained using standard benchmark datasets and the performance was evaluated on the independent test

sets. On both of these prediction tasks, ProtConv achieves improved results than the state-ofthe-art methods and thus shows the effectiveness of the proposed method.

1.6 Organization of the report

The report is divided into 5 chapters. We are now in Chapter 1 which contains Declaration, Certificate, Abstract, Acknowledgments, the Table of Contents and Introduction. Other chapters are designed as follows:

Chapter 2 provides brief idea about background works related to the work that inspired me to work with this topic.

Chapter 3 describes the environment we would like to explore, the datasets, feature description. Finally the CNN architecture that gives the best result on the dataset we prepare.

Chapter 4 shows the result we got from our experiments and finally the result we want to stop for the time being.

Chapter 5 presents the conclusions, summaries the thesis contributions, and discusses the future works.

Chapter 2

Background

In recent years, there has been a great deal of interest in the use of therapeutic peptides in cancer therapy. The current research addresses the creation of computational models for predicting and discovering novel anticancer peptides. In this chapter, we will explore the approaches and processes that have been used in recent years.

2.1 Biological Preliminaries

Here are some biological preliminaries which will be needed to understand the chapter.

2.1.1 DNA

DNA stands for deoxyribonucleic acid, which is the transporter of genetic instructions within a cell. It calls "Hereditary Materials" means the information passed on to the next generation. Nucleotides are the basic building blocks of DNA. Phosphate, sugar, and nitrogen bases are all present in each nucleotide. Adenine (A), thymine (T), guanine (G), and cytosine (C) are four forms of nitrogen bases (C). DNA is considered as the "blue print" of the cells. All living things have DNA in their cells. Children inherit their DNA from their parents. This is why children have similar skin, hair, and eye colors to their parents. A person's DNA is made up of a mixture of their parent's DNA. A molecule of DNA contains two chains that are folded up each other which a bit like a ladder that has been twisted several times [7][8].

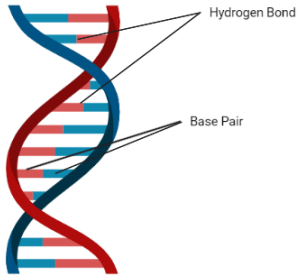

Figure 2.1: Structure of DNA

2.1.2 RNA

Ribonucleic acid or RNA is one of the three main biological macro molecules that are important for all known life forms (along with DNA and proteins) [9]. Adenine, cytosine, uracil, and guanine are the four nitrogenous bases of RNA. The pyrimidine uracil is structurally identical to the pyrimidine thymine, which is also found in DNA. Uracil(like thymine) can base-pair with adenine. RNA is used by the cell for a variety of purposes, one of which is messenger RNA, or mRNA. The major task of mRNA is to carry the genetic code require for the formation of proteins from the nucleus to the ribosome. Another form of RNA is tRNA or transfer RNA which carries amino acid to the translation site. RNA can also act as enzymes [10][11].

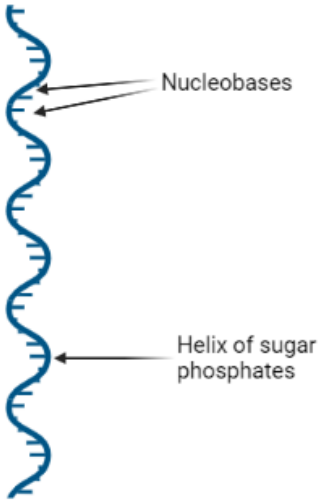

Figure 2.2: Structure of RNA

2.1.3 Gene

The physical and functional unit of heredity is the gene. DNA is used to make genes. Some genes serve as blueprints for creating protein-like molecules. Many genes, on the other hand, don't produce proteins. Genes range in size from a few hundred to over 2 million DNA bases in humans. Humans have between 20,000 and 25,000 genes, according to the Human Genome Project, an international research project aimed at determining the sequence of the human genome and identifying the genes it contains.
Each individual is born with two copies of each gene, one from each parent. The majority of genes are identical in all humans, but a small number of genes (less than 1% of the total) vary slightly. Alleles are variants of the same gene with minor variations in DNA base sequence. These minor variations contribute to the individuality of each person's physical features.

2.1.4 Amino-Acid

Amino acid, any of a class of organic compounds composed of a basic amino group (-NH2), an acidic carboxyl group (-COOH), and a unique organic R group (or side chain) for each amino acid. The phrase amino acid refers to the carboxylic acid (α-amino).

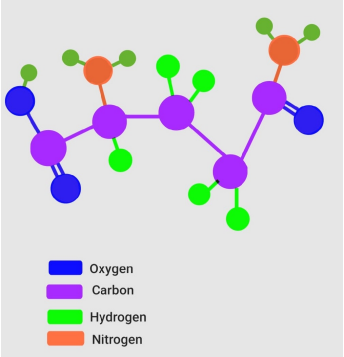

Figure 2.3: Peptide Bond of Amino Acid

Each molecule has a core carbon (C) atom, referred to as the α-carbon, to which an amino and a carboxyl group are linked. A hydrogen (H) atom and the R group usually satisfy the remaining two bonds of the α-carbon atom. [12]

Here are some essential amino acids used in protein formation.

Abbreviation	**Name**	**Side-Chain**	**pKa**
A ala	Alanine	Hydrophobic	
C cys	Cysteine	Hydrophobic	8.5
D asp	Aspartic Acid	Negative	4.4
E glu	Glutamic Acid	Negative	4.4
F phe	Phenylalanine	Hydrophobic	
G gly	Glycine	Hydrophobic	
H his	Histidine	Positive	6.5
I ile	Isoleucine	Hydrophobic	
K lys	Lysine	Positive	10.0
L leu	Leucine	Hydrophobic	
M met	Methionine	Hydrophobic	
N asn	Asparagine	Polar	
P pro	Proline	Hydrophobic	
Q gln	Glutamine	Polar	
R arg	Arginine	Positive	12.0
S ser	Serine	Polar	
T thr	Threonine	Polar	
W trp	Tryptophan	Hydrophobic	
Y tyr	Tyrosine	Polar	10.0
V val	Valine	Hydrophobic	

Table 2.1: Important amino acid list.

2.1.5 Protein and Peptides

Protein and peptides are elementary portion of cells that carry out important biological tasks. Peptides are tied-up chain of amino acids which are held together by peptide bonds. Functionally, proteins and peptides are very similar. The basic distinguishing facts are size and structure. Peptides are smaller than proteins [13]. Traditionally, proteins are consist of 50 or more amino acid, whereas peptides are defined as molecules that consist of between 2 and 50 amino acids [14].

2.1.6 Central Dogma of Biology

The central dogma is the process by which the instruction in DNA are converted into a function product. It describes the two-step process, transcription and translation. DNA contains the information needed to make all of our proteins and RNA is a messenger that carries the information to the ribosomes. The ribosomes serve as factories in the cell where the information is translated from a code into the functional product. How it is works we can understand that from figure 2.5.

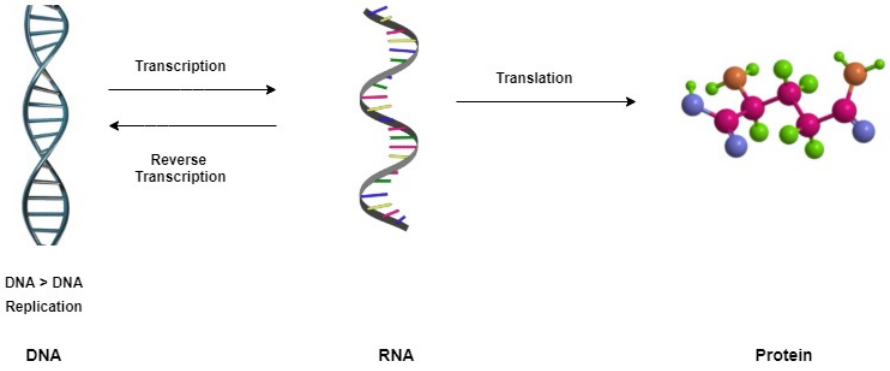

Figure 2.4: Central Dogma

2.1.7 Anticancer Peptides

The antimicrobial peptides which have anti tumor activity are called anticancer peptides (ACPs). Anticancer peptides (ACPs) are generally short peptides with a length of 10-50 amino acids, and have been widely explored over the years as one of the most effective treatments for cancer. ACPs can reduce tumor cell proliferation or migration, as well as the creation of tumor blood vessels, and are less effective to control drug resistance. ACPs are the most promising anti-cancer agent due to the aforementioned benefits [15]. ACPs show a wide range of cytotoxicity to different

cancer cells but not to normal cells; it is believed that the cancer-selective toxicity of ACPs is closely linked to the electrostatic communication of ACPs with negatively charged plasma membrane components of cancer cells [16].

2.1.8 Inflammatory Cytokines

An inflammatory cytokine or proinflammatory cytokine is a type of signaling molecule that is secreted from immune cells like helper T cells and macrophages, and certain other cell types that promote inflammation [17].
Inflammatory cytokines play a key role in the production of atherosclerotic plaque, causing effects all along the atherosclerotic vessel. Importantly, independent of the risk factor, e.g., diabetes, hypertension, or obesity, the development of atherosclerotic lesions is defined by a disruption in normal endothelial cell activity [18].

2.2 Machine Learning Preliminaries

In this section we have discuss some machine learning approaches which generally used to build various model to find protein function.

2.2.1 Supervised Learning

Supervised learning is a sort of machine learning in which machines are trained using well-labeled training data and then predict the output based on that data. There are several advanced machine learning approaches, such as Support-vector machines, Linear regression, Logistic regression, Linear discriminant analysis, Naive Bayes, Decision trees, Neural networks etc. We have used some of them which we have describe below:

Support Vector Machine

Most of the identification or prediction tasks are modelled as supervised learning problem. Support Vector Machine (SVM) is an algorithm for supervised machine learning that can be used for supervised challenges with classification or regression. SVM models have been developed using amino acid composition, dipeptide composition and binary profiles as input features on both practical datasets and balanced datasets [19].
The use of therapeutic peptides in cancer treatment has gained significant attention in recent years. Present study describes the creation of computational models for the prediction and detection of novel anticancer peptides.[20] Two methods were

built in the present study to predict these peptides using the support vector machine (SVM) as a strong algorithm for machine learning. Classifiers were used on the basis of the definition of the pseudo-amino acid composition of Chou (PseAAC) and the local alignment kernel.The results show that 89.7 percent and 92.68 percent, respectively, are the precision and specificity of the local alignment kernel-based process. PseAAC-based system accuracy and specificity are 83.82% and 85.36% , respectively. Out of 22 peptides of the p24 protein, 4 peptides are anticancer and 18 are non-anticancer through computational review. It is evident in the Ames test results that anticancer peptides are not mutagenic. The results therefore show that the mentioned methods of computation are useful for identifying potential anticancer peptides that are worthy of further experimental validation.[21]

Locally Deep Support Vector Machine

Locally deep support vector machine is a well-researched class of supervised learning techniques. This unique implementation is appropriate based on either continuous or categorical variables, for the prediction of two possible effects. To construct a model based on the Support Vector Machine Algorithm, the Two-Class Support Vector Machine is used. To predict two potential outcomes that depend on continuous or categorical predictor variables, the classifier initialized by this module is useful [22]. This model is a supervised method of learning and thus involves a dataset that includes a column that has been labelled. By offering the model and the tagged dataset as an input to Train Model or Tune Model Hyper parameters, you can train the model. For the new input cases, the trained model will then be used to predict values.

Decision Forest Tree

We also tested a Decision Forest technique, which mixes numerous Decision Tree models. A unique set of descriptors is used to create each Decision Tree model. When similar predictive performance measures are integrated using the Decision Forest approach, quality is consistently and considerably enhanced in both the training and testing processes when compared to the individual models [23].

Artificial Neural Network

Another machine learning algorithm that can be supervised or unsupervised is the neural network or artificial neuron network (ANN). Two Class Neural Network is a supervised version. This method's work flow and structure are inspired by bio-

logical neural networks. Two Class Neural Networks are made up of two or more interconnected layers. The first layer is known as the input layer, and the last layer is known as the output layer. We can add multiple hidden layers between these two layers.
The value of the output layer is determined at each node of the hidden layers in order to compute the output. The weighted total of the values from the previous layer is used to calculate t his value. To determine the weighted total, an activation function is also used.

Features

However, we implemented Transfer learning in CNN to predict protein function; traditionally, protein function prediction was performed via feature extraction and feature selection. G-Gap and N-Gram are two of the most commonly used feature extraction techniques.
G-Gap
An significant concept for understanding generalization is the generalization gap. The difference between the performance of a model on training data and its performance on unseen data from the same distribution is commonly known as g-gap [24].
N-Gram
An n-gram is a contiguous sequence of n items from a given sample of text or speech in the fields of computational linguistics and probability. According to the application, the components may be phonemes, syllables, letters, words or base pairs. Usually, n-grams are gathered from a corpus of text or speech. When the things are words, shingles can also be called n-grams [25].

2.2.2 Convolutional Neural Network

A convolutional neural network is a subset of deep neural networks in deep learning, most widely applied to visual imagery analysis. Convonutional neural networks are very similar to traditional neural networks. They are made up of neurons which have weights and prejudices that are learn able. Each neuron receives such inputs, conducts a dot product, and with a non-linearity optionally follows it [26].
A convolution is the simple process of applying a filter to an input to produce an activation. When the same filter is applied to an input multiple times, a feature map is created, showing the locations and strength of a detected feature in an input, such as an image [27] [5] [1].
The capability of convolutional neural networks to acquire a large number of filters

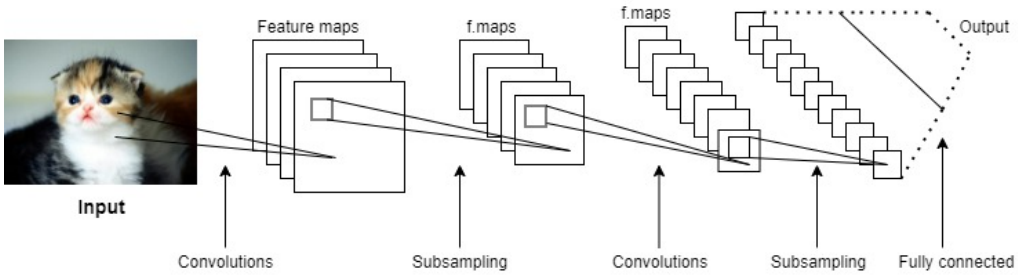

Figure 2.5: Convolutional Neural Network with Image Classification

in parallel unique to a training dataset under the conditions of a particular predictive modeling problem, such as image classification, is their unique feature. As a result, highly specific features appear on input images that can be found anywhere [28].
We also used Kernal in CNN. The kernel in a convolutional neural network is nothing more than a filter that extracts features from images. The kernel is a matrix that passes over the input data, operates a dot product with a sub-region of the input data, and outputs a matrix of dot products. The stride value pushes the kernel on the input data. If the stride value is 2, the kernel pushes the input matrix by two columns of pixels. In a nutshell, the kernel is used to extract high-level features from an image, such as edges [29] [30].
Convolutional neural networks (CNNs) consist of alternating convolutional layers and pooling layers.Another part of a CNN is the pooling layer.Its aim is to gradually shrink the representation's spatial size in order to reduce the number of parameters and computations in the network. Each function map is handled separately by the pooling layer [31] [32].
Pooling can be categorized into two kinds: maximum and average. A window passes over the input matrix in max pooling and generates a matrix with the windows' maximum values.Average pooling is similar to max pooling, but instead of using the maximum value, it uses the average value. The stride value defines how the window moves. If the stride value is 2, the window in the matrix shifts 2 columns to the right after each operation. In a nutshell, the pooling technique reduces the amount of computing power needed to analyze the data [33] [34] [35].

2.2.3 Transfer Learning

Transfer learning is an approach to machine learning. In the field of transfer learning, knowledge of an already trained machine learning model is applied to a separate but related problem [36]. The general idea is to use the information that the model has learned from a task with a lot of accessible, labeled training data in a new task

that does not have much data. It uses pre-training and fine-tuning to learn how to customize an existing model [22].
In machine learning, neural networks typically aim to detect edges in the first layer, forms in the middle layer, and task-specific properties in the latter layers. The early and middle layers are utilised in transfer learning, and only the latter layers are retrained. It aids in utilizing the labeled data from the task on which it was initially trained.

2.3 Literature Review

Our literature review is divided into three areas. To begin, we'll go through the machine learning method we utilized and where it was previously employed.Then we'll go over the two data sets we used: anticancer peptides and proinflammatory cytokines.

2.3.1 Transfer Learning

Stevo Bozinovski and Ante Fulgosi published a paper in 1976 that addressed transfer learning in neural network training directly. A mathematical and geometrical model of transfer learning is presented in this paper [37]. A paper on the use of transfer learning to train a neural network on a dataset of images representing letters on computer terminals was published in 1981. Both positive and negative transfer learning were demonstrated in the lab. Lorien Pratt proposed the discriminability-based transfer (DBT) method in a paper on transfer in machine learning released in 1993 [38].
Machine learning and data mining have various applications in biological domains, where we are trying to develop predictive models using labeled training data. In practice, however, high-quality labeled data is limited, and labeling fresh data is expensive. Transfer and multitask learning are appealing alternatives because they allow usable knowledge to be extracted and transferred from data in auxiliary domains, which helps to overcome the target domain's shortage of data problem [39] [40] [41].

2.3.2 Anticancer Peptides

Previously Atul Tyagi, Pallavi Kapoor, in silico model [20], have gained the ability to develop in silico methodologies for the prediction and design of ACPs. Support vector machine (SVM)-based models based on peptide properties such as amino acid

composition, dipeptide composition, and binary profile pattern have been created. Models that distinguish ACPs from AMPs have also been created. With MCC and AUC values of 0.83 and 0.94, respectively, a binary profile-based SVM model utilizing the NT10 dataset achieved maximum accuracy of 91.44 percent. To aid the scientific community, AntiCP, a user-friendly website, has been built for the first time to anticipate and construct highly efficacious ACPs [42] [43]. We introduced ACP-DL, a deep learning long short-term memory (LSTM) neural network model for predicting anticancer peptides.Peptide sequences are converted by a k-mer sparse matrix of the reduced amino acid alphabet, which is a 2D matrix, and practically complete sequence order and amino acid component information are preserved [44]. also compared the purposed ACP-DL with existing state-of-the-art machine-learning models, for example SVM,Random Forest (RF), and Naive Bayes (NB), The 5-fold cross-validation. The implementation of these three machine-learning models comes from Scikit-learn. They have used ACP740 and ACP240 and got 89.4% and 90.0% accuracy respectively [27] [45] [46] [47].

2.3.3 Inflammatory Cytokines

Osteoarthritis (OA) is associated with cartilage degeneration, subchondral bone remodeling, and synovial membrane inflammation, yet the origin and pathogenesis of this devastating disease are unknown. Secreted inflammatory chemicals, such as proinflammatory cytokines, are key mediators of the disrupted processes implicated in OA pathogenesis [48] [49]. Aside from IL-1 and TNF, numerous additional cytokines, including IL-6, IL-15, IL-17, IL-18, IL-21, leukemia inhibitory factor, and IL-8 (a chemokine), have been demonstrated to be involved in OA and may be therapeutically targeted [50] [51]. In another study, Serum samples were taken from 56 patients who had been diagnosed with rheumatoid arthritis for at least six months. Patients with psoriatic arthritis (PsA) and ankylosing spondylitis (n = 21) as well as healthy people (n = 19) had their cytokine profiles assessed. The Kruskal–Wallis test with Dunn's multiple comparison adjustment, linear correlation tests, significance analysis of microarrays (SAM), and hierarchical clustering software were used to analyze the data [52] [53] [54].

2.4 Summary

The use of therapeutic peptides in treatment of cancer has received a lot of attention in recent years. The current paper highlights the creation of computational models for the prediction and discovery of new anticancer peptides.Previously they have

used some algorithms as like Amino Acid Composition, Dipeptide Composition, Binary Profile and N-terminal Approach and get 91.44% accuracy in the result[20]. There are some other papers where they have used some more complex algorithms along with the previous one as like , Overlapping property, Twenty-One-Bit Features, Composition-Transition-Distribution, G-gap dipeptide composition, Adaptive skip dipeptide composition, 10-fold cross-validation etc [55] [56]. to get the result and they have got 91% and 89% accuracy respectively. They got a decent result, but it was a long and time-consuming process to use. In ProtConv, we employed Tape Embedding and Transfer Learning, and our proposed model is quite simple. Any feature representation and network architecture can be used instead of the ones we proposed. In the performance test, ProtConv outperformed state-of-the-art approaches.

Chapter 3

Materials and Methods

We go over the methodology and outcomes in detail in this chapter. Also, there's the data set from a prior paper.

3.1 Overview

A system diagram of ProtConv is given in figure 3.1. Protein sequences from a dataset is converted into a vector representation using transfer learning from pre-trained models. After that, each of these vectors are fed into a padding and re-shaping module to convert them into two dimensional square images. After that all the training images are passed into a training model where a convolutional neural network (CNN) is learned. The output is a trained CNN model which is then used to validate the results on a independent test set of sent for production. Rest of this chapter presents the necessary details of each of these steps.

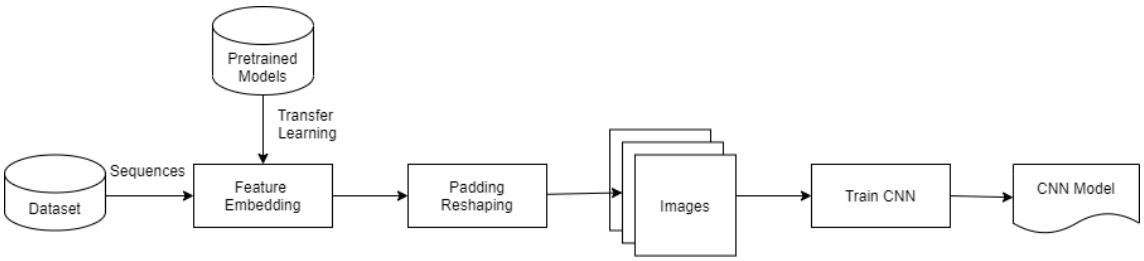

Figure 3.1: Methodology of our project.

3.2 Description of Datasets

We have used two protein function prediction task. Proinflamatory Cytokines and Anticancer peptide. The brief discussion of these datasets are given below :

3.2.1 Benchmark Dataset

To evaluate the performance of ProtConv we have selected two protein function prediction task: proinflammatory and anticancer peptide prediction. Both of the problems are presented in the literature as binary classification problem and a number of supervised learning algorithms have been deployed to address the problems. In this section, we briefly introduce the problems and the related datasets with a glimpse of the state-of-the-art literature of the related problems. A summary of the datasets are given in Table 3.1

Problem	type	Positive Sample	Negative Sample	Total
Proinflamatory Cytokins	train	607	1098	1705
	test	134	156	290
Anticancer Peptides	train	250	250	500
	test	82	82	164

Table 3.1: Summary of the datasets used in this paper.

3.2.2 Proinflammatory peptides

Inflammatory peptides are molecules that plays a very important role in signaling the immune system against the pathogens. Regulation of inflammatory cytokines are often key to treat inflammatory diseases [1]. Computational methods have been recently proposed to reduce the cost and time needed in laboratory methods to identify proinflammatory peptides. ProInflam was proposed in [2] by Gupta et al. using Support Vector Machine Classifier and sequence based features. Manavalan et al. proposed PIP-EL [3]. In their work, they also utilized several sequence based features. The dataset that we use in this paper is curated from The immune epitope database (IEDB) [[57], [58]]. It has been observed that the benchmark datasets used in the previous studies might have some redundancies. To reduce that, CD-HIT [56] algorithm was run to remove the sequences upto 60% similarity. The resulting dataset was then split into 80%-20% ratio intro train and test datasets. These datasets were originally curated and made available in http: //kurata14.bio.kyutech.ac.jp/ProInFuse/. We are using the dataset with their permission. The final train dataset has 607 positive and 1098 negative 80 samples and the test set contains 134 positive and 156 negative instances. All the peptide sequences have lengths in the range between 5 to 25.

3.2.3 Anticancer peptides

The second prediction task is to identify anticancer peptides. Anticancer peptides have recently gained attention due to their increased specificity and less toxic nature [59]. These are short sequences as well ranging from 10 to 50 amino acids. Computational methods are in use to predict anticancer peptides given their sequences [[55], [20], [44], [60], [14], [21]]. AntiCP was proposed by Tyagi et al. [20] using Support Vector Machines and residue composition based features. Hajisharifi et al. [21] used pseudo-amino acid composition (PseAAC) and local alignment kernel support vector machines. Later works, also employed several sequences based and evolutionary features and traditional classification models. In a recent work, Yi et al. used long short term memory based deep neural networks and proposed ACP-DL to predict anticancer peptides. The datasets that we have chosen for anticancer peptides is well used in the literature. This was first proposed in [60] and curated from CancerPPD database [42] . Both of the training and test datasets are well balanced and difficult benchmarks for the prediction task[**?**].

3.3 Feature Description

Several feature representation techniques have been proposed in the literature for generic protein attribute prediction and particularly for the two tasks that are selected in this paper. Among them are sequence based features [61], gapped k-mers [62], pseudo-amino acid composition [63] , evolutionary features [64] , structure based features [[65] , [43]], etc. Also note that a number of feature generators are available based on these such as PyFeat [19] , iFeature [66], iLearn [15], PseAAC-Builder [67]. However, recent trends shows the use of deep learning engineered features to represent proteins as vectors. A number of sequence based, attention models and transofrmer models have been applied in the field of genomics and proteomics inspired from the models in Natural Language Processing domain. Among them are, ProtVec [68], Progen [69], UDSMProt [70], Bertology [26] , etc. In this paper, we have used transfer learning [22] and used the embeddings of TAPE [6]. Three different architectures: long short term memory (LSTM) [28], dilated residual network (ResNet) [46] and transformer [71] and were trained in [6] to learn protein embeddings in a semi-supervised fashion for several protein tasks to obtain generalization. The pre-trained embedding of the model provides a (32, 768) shaped tensor which is then averaged to get a vector representation of length 768. This vector is then fed to subsequent padding and reshaping phases to be converted into a 28×28 image. Sample protein images are shown in Figure 3.2.

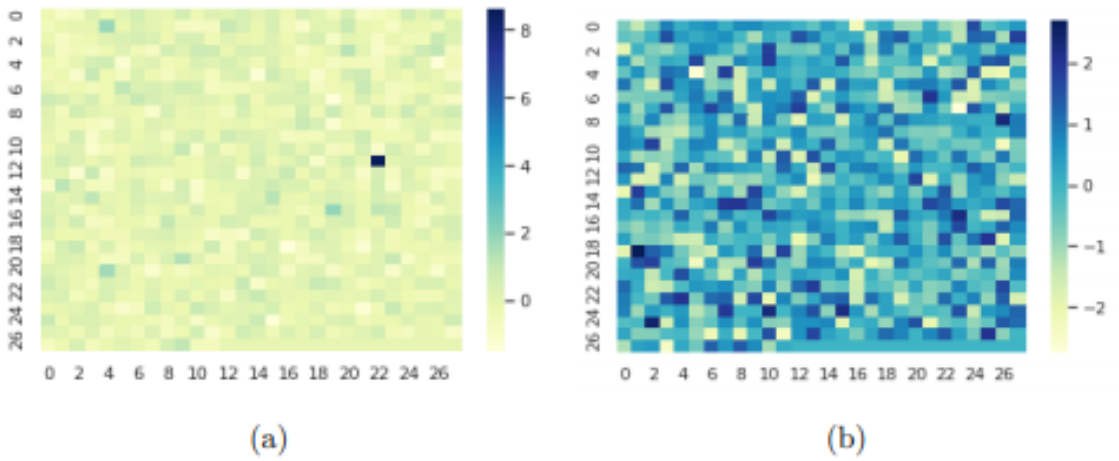

Figure 3.2: A sample protein image from proinflammatory peptide dataset shown using heat map, (a) before normalization and (b) after normalization.

3.4 CNN Architecture

The convolutional neural network architecture used in this paper is inspired from LeNet-5 [5]. Convolutional neural networks have been proven effective in the computer vision domain as they incorporate ideas that allows noise or distortion due to scaling or shifting using subsampling, local receiptive fields and shared weights. The architecture of the Convolutional neural network that is used in ProtConv is given in figure 3.3.

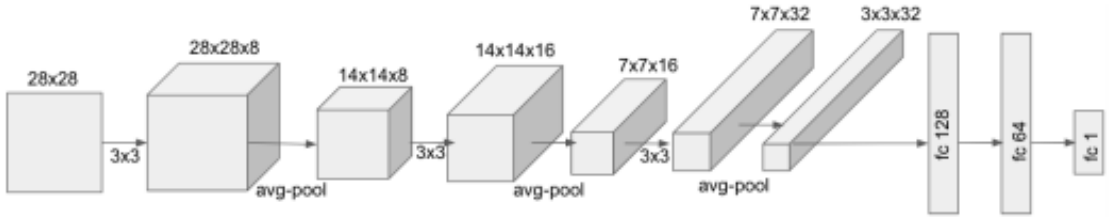

Figure 3.3: ProtConv Convolutional Neural Network Architecture.

It takes an two dimensional image of shape 28×28 as input to the network. It goes through six consecutive layers of covolutional and sub-sampling layers. All convolutional layers use filters of size 3×3. Paddings are used to keep the dimensions same which was not a feature in original LeNet-5 architecture. Also note, we have added another pair of subsequent layers of convolutional and subsampling layers. We have used average sampling layer to subsample as done in the original LeNet-5 paper. In our experiments, we have seen that in comparison to average pooling, max pooling layers quickly overfits the data.

In the three layers of convolutional neural networks, the number of filters used

were 8,16 and 32. In each case, rectified linear unit (RelU) [38] was used as activation function. After these layers, the output is flattened and fed to subsequent fully connected layers. First two layers are comprised of and 64 relU units followed by a single sigmoid unit layer for binary classification. Total number of trainable parameters in this network was 51,201. We kept the neural network architecture simple to avoid overfitting but also capable of discriminating the input image vectors to solve important prediction tasks. We have used binary cross entropy function as a loss function in the output layer of the network.

$$L = -\frac{1}{m}\sum_{i=1}^{m}[y(i)log(\hat{y}(i)) + (1 - y(i))log(1 - \hat{y}(i))] \tag{3.1}$$

To learn the model, ADAM optimization algorithm [13] was used.

However, experiments showed that RMSProp [72] achieves almost similar results to that of ADAM algorithm. Please note that, one alternative to this architecture was to use 1D convolutions used traditionally on single dimensional data. Protein sequences are often converted to a single dimensional vector. However, the methodology, we used here tries to capture the spatial properties of the features in a two dimensional space.

3.5 Performance Evaluation

For both of the problems, we have used extra validation set with removed redundancy from the training set to test the performance of our proposed method. Since the size of the training sets are not that large, the validation set were kept at least 20% of the training set to avoid any over fitting and loss of generalization. Four standard metrics have been used: Sensitivity (Sn), Specificity (Sp), Accuracy (Acc) and Mathew's Correlation Coefficient (MCC). All these metrics are suitable and applicable for binary prediction or classification tasks. In a binary prediction task, true negatives (TN) are samples that are negative in class and also predicted as negative; true positives (TP) are positive instances that are predicted correctly as positive; false positive (FP) are negative instances mistakenly predicted as negative and false positives (FP) are negative instances incorrectly predicted positive by the predictor. Based on these, the metrics used in this paper are formulated as in the following equations

$$S_n = \frac{TP}{TP + FN} \tag{3.2}$$

$$S_p = \frac{TN}{TN + FP} \tag{3.3}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.5}$$

Among these metrics Sn, Sp and Acc have the values in the range [0,1] where, 0 is the worst classifier and 1 is the best classifier. In case of MCC, it has a best classifier with value +1 and 0 is a random classifier and -1 denotes the worst classifier. The range is in [-1,+1].

3.6 Summary

When it comes to biological elements, it is obvious difficult to forecast a situation. To do this assignment, we began by gathering a good dataset. Then we extracted features from the dataset and identified the features that were most connected with the target concerned level. Then we implemented some classification algorithms and discovered that Support Vector Machine produces the best results.

Chapter 4

Experimental Analysis

We have performed our experiments in Google colaboratory environment. The models were all built on Tensor Flow version 1.x using keras library to build convolutional neural network. We used a batch size of 64 for both of the prediction tasks. In each case, the independent test set was used only to validate the results. Since we are using deep neural networks, it is often found that such networks over estimates and shows high bias. To avoid that, we have used regularization and early stopping.

4.1 Environment Setup

Initially, we used Azure Machine Learning Studio. We employed support vector machines, two-class locally deep SVMs, K-mers, n-grams, cross fold validation, and other techniques. Because our data set was so limited, we relied on feature extraction and feature selection. The process took too long, which was inconvenient for us, and the outcome was unsatisfactory. Following that, we employed the Keras library, but the results were unsatisfactory. Following that, we employed ProtConv, a protein function prediction approach based on a convolutional neural network.We proposed converting the vector representation of a protein or peptide sequence into a two-dimensional image using a single channel and feeding it into a convolutional neural network. In ProtConv, we adopt a simple convlutional neural network architecture inspired by the Lenet-5 [5]. To transform a protein or peptide sequence to a vector representation, we used TAPE [6] embeddings and pre-trained models. The framework is simple, but it can be extended. These proposed feature representations and network architectures can be replaced by any other feature representation and network design.

4.2 Results

We kept our models simple and kept same for both of the prediction tasks. However, in the case of proinflammatory peptide prediction task, number of epochs were 50 compared to 30 in case of the anticancer peptide prediction task.

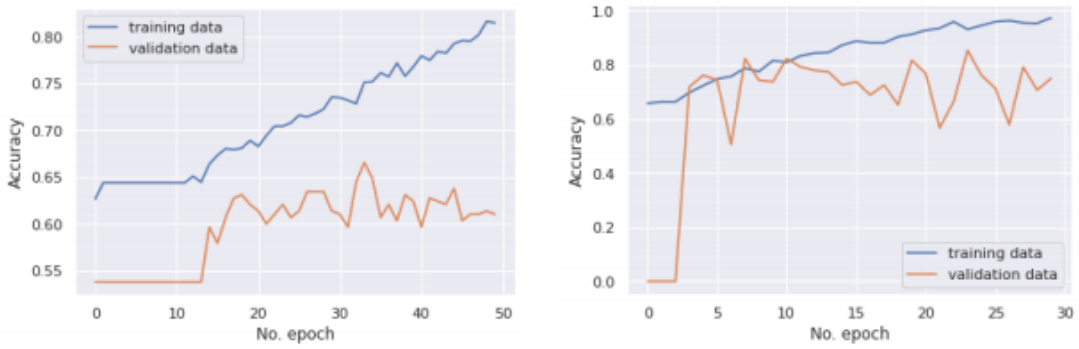

Figure 4.1: Plot of accuracy vs number of epochs for training and validation set for (a) proinflammatory peptides dataset and (b) anticancer peptides dataset.

4.2.1 Convergence Analysis

In each eopch, 500 iterations were made. The history of convergence of both of the prediction tasks are shown in Figure 4.1. Note that for both of the problems training set accuracy increases. In case of proinflammatory peptide prediction task the training set accuracy reaches upto 82% and for the anticancer peptide prediction task, it reaches upto 97%. However, if we apply early stopping, the best training set accuracy for these two problems is around 73% and 93.25% respectively.

4.2.2 Comparison to Other Methods

To compare the performance of ProtConv, we have taken state-of-the-art methods for both of the tasks and compare our results with those. In case of proinflammatory peptide prediction task, we have used ProInflamm [2] and PIP-EL [3]. Results on the independent test set in terms of sensitivity, specificity, accuracy and MCC are given in Table 4.1 Bold faced values in the table are showin the better values. We could see that ProtConv is producing better or improved metrics in terms of sensitivity, accuracy and MCC. However, the specificity of PIP-EL [3] is better than ProtConv.

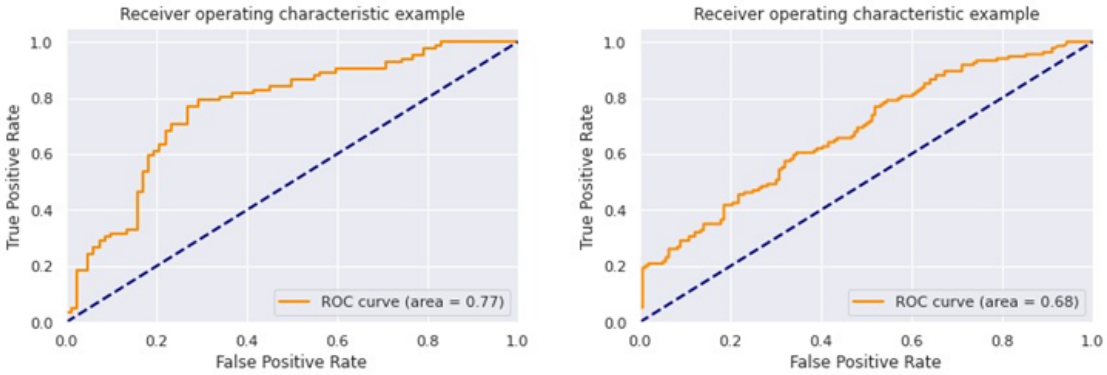

Figure 4.2: ROC curve of anticancer peptides (left) and proinflamatory cytokins (right) repectively.

Method	**Sn**	**Sp**	**Acc**	**MCC**
ProInflam	0.666	0.596	0.628	0.264
PIP-EL	0.542	0.741	0.649	0.299
ProtConv	0.686	0.647	0.665	0.333

Table 4.1: Results Comparison on Independent dataset of Proinflamatory cytokins prediction problem.

In case of anticancer peptides task, we have used ACP-DL [55] for comparison. We have not used other methods since this is a latest method providing state-of-the-art results using deep neural networks. In the case of proinflammatory peptide prediction task, we could not find any deep learning or deep neural network based approaches in the literature. The results for anticancer prediction task are given in Table 4.2

Method	**Sn**	**Sp**	**Acc**	**MCC**
ACP-DL	0.890	0.804	0.847	0.670
ProtConv	0.890	0.817	0.853	0.709

Table 4.2: Results Comparison on Independent dataset of Anticancer peptide prediction problem.

Here again, bold faced values are showing the better values. We could notice that performance of ProtConv is significantly improved or similar to the performances of ACP-DL.

4.2.3 ROC Analysis

One of the important metric for performance analysis of classifiers is the area under receiver operating characteristic which is independent of the threshold chosen by the classifiers for decision making. In Figure 4.2, we show the receiver operating characteristic curves of the both of the problems to show the nature of the classifiers on each of the datasets.

4.2.4 Other Classifiers

Previously we have used some traditional machine learning methods such as Support Vector Machine, decision tree, logistic regression, bayes point and neural network. Unfortunately we didn't get the satisfactory result. So we have used Tape in our data set. Here is the comparison of the results with traditional methods.

Methods	**Accuracy**
Decision Tree	80%
Logistic Regression	85%
Bayes point	87%
Neural Network	85%
ProtConv	89%

Table 4.3: Results Comparison between traditional machine learning methods with ProtConv

Here, we have got the best result for ProtConv compared to other traditional machine learning model.

4.3 Discussion

In this paper, we have presented ProtConv a convolutional neural network based approach for protein fucntion prediction task. Here, the protein sequence is first converted into a two dimensional image from a vector representation. Note that, the vector representation can be from any sources. ProtConv uses TAPE based encodings from a pre-trained model. However, it is possible to incorporate any feature representation technique for the vector represention. Note that, after the conversion into image, it is also possible to apply sophisticated methods like wavelet transform or Hilbert transforms to improve the representation. However, the purpose of this paper is to keep the settings simplest to show the hypothesis is working. We see that for both of this prediction tasks the proposed model is providing satisfactory

results.

Note that, there are scope of improvement in the methodology proposed here. Using evolutionary information helps to improve the performance of the prediction tasks. However, attention models have shown to achieve similar performances to evolutionary approaches in recent times [26]. We have used the TAPE encoding without any kind of fine-tuning. In case of deep learning, there exists a large number of hyper parameters and all of these are subject to optimization. Thus, there are a lot of scope of improvement of our proposed model. Our goal here was to show that the hypothesis of converting the vector representation of a protein sequence into two dimensional image and using a convolutional neural network architecture can solve prediction tasks with satisfactory results. We have demonstrated the effectiveness on two standard prediction tasks. We believe it is possible to extend this method beyond these initial setup and apply it to other prediction tasks on proteins and other sequence types.

Chapter 5

Conclusion

This chapter presents an overall summary of the work followed by the limitations and possible future work from this work.

5.1 Summary

In this thesis, we have presented ProtConv a convolutional neural network based approach for protein fucntion prediction task. Here, the protein sequence is first converted into a two dimensional image from a vector representation. Note that, the vector representation can be from any sources. ProtConv uses TAPE based encodings from a pre-trained model. However, it is possible to incorporate any feature representation technique for the vector represention. Note that, after the conversion into image, it is also possible to apply sophisticated methods like wavelet transform or Hilbert transforms to improve the representation. However, the purpose of this paper is to keep the settings simplest to show the hypothesis is working. We see that for both of this prediction tasks the proposed model is providing satisfactory results. Note that, there are scope of improvement in the methodology proposed here. Using evolutionary information helps to improve the performance of the prediction tasks. However, attention models have shown to achieve similar performances to evolutionary approaches in recent times [26]. We have used the TAPE encoding without any kind of fine-tuning. In case of deep learning, there exists a large number of hyper parameters and all of these are subject to optimization. Thus, there are a lot of scope of improvement of our proposed model.

5.2 Limitation

The lack of a large dataset is a key obstacle in this case. However, There have been a number approaches to integrating transfer learning from pre-trained models with generic prediction challenges. Our experimens were limited to using only two problems and a simple CNN architecture. The hypothesis required more exploration on other datasets and various types of architectures to be generalized. In addition to this exploratory and explanatory studies could have been performed on the methodology to show the pros and cons of this feature representation technqiue.

5.3 Future Work

There are scope of improvement in the methodology proposed here. The use of evolutionary information aids in the improvement of prediction task performance. In recent years, however, attention models have proved to perform similarly to evolutionary techniques [26]. We didn't do any fine-tuning with the TAPE encoding. There are a vast number of hyper parameters in deep learning, all of which are subject to optimization. As a result, our proposed model has a lot of scope for improvement. The purpose of this study was to demonstrate that the hypothesis of transforming a protein sequence's vector representation into a two-dimensional image and using a convolutional neural network architecture to solve prediction challenges successfully. On two standard prediction tasks, we have shown that it is effective. We believe that this method can be extended beyond these initial setups and used to do various protein and sequence type prediction tasks.

References

[1] Roberto Scarpioni, Marco Ricardi, and Vittorio Albertazzi. Secondary amyloidosis in autoinflammatory diseases and the role of inflammation in renal damage, 2016.

[2] Sudheer Gupta, Midhun K Madhu, Ashok K Sharma, and Vineet K Sharma. Proinflam: a webserver for the prediction of proinflammatory antigenicity of peptides and proteins, 2016.

[3] Balachandran Manavalan, Tae Hwan Shin, Myeong Ok Kim, and Gwang Lee. Pip-el: A new ensemble learning method for improved proinflammatory peptide predictions, 2018.

[4] Jean-Michel Claverie. From bioinformatics to computational biology. *Genome research*, 10(9):1277–1279, 2000.

[5] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition, 1998.

[6] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape, 2019.

[7] Scott O Rogers and Arnold J Bendich. Extraction of dna from plant tissues, 1989.

[8] Nadrian C Seeman. Dna in a material world, 2003.

[9] Matthew G Seetin and David H Mathews. Rna structure prediction: an overview of methods, 2012.

[10] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years, 2019.

[11] Steven P Gygi, Yvan Rochon, B Robert Franza, and Ruedi Aebersold. Correlation between protein and mrna abundance in yeast, 1999.

[12] Wikipedia. "Amino acid — Wikipedia, the free encyclopedia", 2021.

[13] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines, 2010.

[14] Lei Xu, Guangmin Liang, Longjie Wang, and Changrui Liao. A novel hybrid sequence-based model for identifying anticancer peptides, 2018.

[15] Zhen Chen, Pei Zhao, Fuyi Li, Tatiana T Marquez-Lago, André Leier, Jerico Revote, Yan Zhu, David R Powell, Tatsuya Akutsu, Geoffrey I Webb, et al. ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of dna, rna and protein sequence data, 2020.

[16] Jamie S Mader and David W Hoskin. Cationic antimicrobial peptides as novel cytotoxic agents for cancer treatment, 2006.

[17] Anthony Cerami. Inflammatory cytokines, 1992.

[18] Cuihua Zhang. The role of inflammatory cytokines in endothelial dysfunction. *Basic research in cardiology*, 103(5):398–406, 2008.

[19] Rafsanjani Muhammod, Sajid Ahmed, Dewan Md Farid, Swakkhar Shatabda, Alok Sharma, and Abdollah Dehzangi. Pyfeat: a python-based effective feature generation tool for dna, rna and protein sequences, 2019.

[20] Atul Tyagi, Pallavi Kapoor, Rahul Kumar, Kumardeep Chaudhary, Ankur Gautam, and GPS Raghava. In silico models for designing and discovering novel anticancer peptides, 2013.

[21] Zohre Hajisharifi, Moien Piryaiee, Majid Mohammad Beigi, Mandana Behbahani, and Hassan Mohabatkar. Predicting anticancer peptides with chou's pseudo amino acid composition and investigating their mutagenicity via ames test, 2014.

[22] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning, 2012.

[23] Weida Tong, Huixiao Hong, Hong Fang, Qian Xie, and Roger Perkins. Decision forest: combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences*, 43(2):525–531, 2003.

[24] Hao Lin, Wei-Xin Liu, Jiao He, Xin-Hui Liu, Hui Ding, and Wei Chen. Predicting cancerlectins by the optimal g-gap dipeptides, 2015.

[25] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language, 1992.

[26] Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models, 2020.

[27] Monowar Md Anjum, Ibrahim Asadullah Tahmid, and M Sohel Rahman. Cnn model with hilbert curve representation of dna sequence for enhancer prediction, 2019.

[28] Jürgen Schmidhuber and Sepp Hochreiter. Long short-term memory, 1997.

[29] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning, 2018.

[30] Youngmin Cho. Kernel methods for deep learning, 2012.

[31] Meng Joo Er, Yong Zhang, Ning Wang, and Mahardhika Pratama. Attention pooling-based convolutional neural network for sentence modelling, 2016.

[32] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences, 2014.

[33] Dingjun Yu, Hanli Wang, Peiqiu Chen, and Zhihua Wei. Mixed pooling for convolutional neural networks, 2014.

[34] Phil Kim. Convolutional neural network, 2017.

[35] Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection, 2015.

[36] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning, 2009.

[37] Stevo Bozinovski. Reminder of the first paper on transfer learning in neural networks, 1976, 2020.

[38] Lorien Y Pratt et al. Discriminability-based transfer between neural networks, 1993.

[39] Qian Xu and Qiang Yang. A survey of transfer and multitask learning in bioinformatics. *Journal of Computing Science and Engineering*, 5(3):257–268, 2011.

[40] John M Giorgi and Gary D Bader. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094, 2018.

[41] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[42] Atul Tyagi, Abhishek Tuknait, Priya Anand, Sudheer Gupta, Minakshi Sharma, Deepika Mathur, Anshika Joshi, Sandeep Singh, Ankur Gautam, and Gajendra PS Raghava. Cancerppd: a database of anticancer peptides and proteins, 2015.

[43] Swakkhar Shatabda, Sanjay Saha, Alok Sharma, and Abdollah Dehzangi. iphloc-es: identification of bacteriophage protein locations using evolutionary and structural features, 2017.

[44] Balachandran Manavalan, Shaherin Basith, Tae Hwan Shin, Sun Choi, Myeong Ok Kim, and Gwang Lee. Mlacp: machine-learning-based prediction of anticancer peptides, 2017.

[45] Sucheta Chauhan and Shandar Ahmad. Enabling full-length evolutionary profiles based deep convolutional neural network for predicting dna-binding proteins from sequence, 2020.

[46] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks, 2017.

[47] Yasen Jiao and Pufeng Du. Performance measures in evaluating machine learning based bioinformatics predictors for classifications, 2016.

[48] Robert R McLean. Proinflammatory cytokines and osteoporosis, 2009.

[49] CK Wong, CY Ho, FWS Ko, CHS Chan, ASS Ho, DSC Hui, and CWK Lam. Proinflammatory cytokines (il-17, il-6, il-18 and il-12) and th cytokines (ifn-γ, il-4, il-10 and il-13) in patients with allergic asthma, 2001.

[50] Mohit Kapoor, Johanne Martel-Pelletier, Daniel Lajeunesse, Jean-Pierre Pelletier, and Hassan Fahmi. Role of proinflammatory cytokines in the pathophysiology of osteoarthritis, 2011.

[51] Xavier Ayral, EH Pickering, TG Woodworth, N Mackillop, and M Dougados. Synovitis: a potential predictive factor of structural progression of medial tibiofemoral knee osteoarthritis–results of a 1 year longitudinal arthroscopic study in 422 patients, 2005.

[52] Wolfgang Hueber, Beren H Tomooka, Xiaoyan Zhao, Brian A Kidd, Jan W Drijfhout, James F Fries, Walther J Van Venrooij, Allan L Metzger, Mark C Genovese, and William H Robinson. Proteomic analysis of secreted proteins in early rheumatoid arthritis: anti-citrulline autoreactivity is associated with up regulation of proinflammatory cytokines, 2007.

[53] Kyung-Joo Cho, Chang-Hyun Yun, Lester Packer, and An-Sik Chunga. Inhibition mechanisms of bioflavonoids extracted from the bark of pinus maritima on the expression of proinflammatory cytokines, 2001.

[54] Anders Johannisson, Robert Jonasson, Johanna Dernfalk, and Marianne Jensen-Waern. Simultaneous detection of porcine proinflammatory cytokines using multiplex flow cytometry by the xmap™ technology, 2006.

[55] Hai-Cheng Yi, Zhu-Hong You, Xi Zhou, Li Cheng, Xiao Li, Tong-Hai Jiang, and Zhan-Heng Chen. Acp-dl: a deep learning long short-term memory model to predict anticancer peptides using high-efficiency feature representation, 2019.

[56] Limin Fu, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. Cd-hit: accelerated for clustering the next-generation sequencing data, 2012.

[57] Ward Fleri, Kerrie Vaughan, Nima Salimi, Randi Vita, Bjoern Peters, and Alessandro Sette. The immune epitope database: how data are entered and retrieved, 2017.

[58] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The immune epitope database (iedb): 2018 update, 2019.

[59] Yuan Qin, Zuo D Qin, Jing Chen, Che G Cai, Ling Li, Lu Y Feng, Zheng Wang, Gregory J Duns, Nong Y He, Zhe S Chen, et al. From antimicrobial to anticancer peptides: the transformation of peptides, 2019.

[60] Wei Chen, Hui Ding, Pengmian Feng, Hao Lin, and Kuo-Chen Chou. iacp: a sequence-based tool for identifying anticancer peptides, 2016.

[61] Sheikh Adilina, Dewan Md Farid, and Swakkhar Shatabda. Effective dna binding protein prediction by using key features via chou's general pseaac, 2019.

[62] Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A Beer. Enhanced regulatory sequence prediction using gapped k-mer features, 2014.

[63] M Saifur Rahman, Swakkhar Shatabda, Sanjay Saha, Mohammad Kaykobad, and M Sohel Rahman. Dpp-pseaac: A dna-binding protein prediction model using chou's general pseaac, 2018.

[64] Shahana Yasmin Chowdhury, Swakkhar Shatabda, and Abdollah Dehzangi. idnaprot-es: Identification of dna-binding proteins using evolutionary and structural features, 2017.

[65] Md Mofijul Islam, Sanjay Saha, Md Mahmudur Rahman, Swakkhar Shatabda, Dewan Md Farid, and Abdollah Dehzangi. iprotgly-ss: Identifying protein glycation sites using sequence and structure based features, 2018.

[66] Zhen Chen, Pei Zhao, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Yanan Wang, Geoffrey I Webb, A Ian Smith, Roger J Daly, Kuo-Chen Chou, et al. ifeature: a python package and web server for features extraction and selection from protein and peptide sequences, 2018.

[67] Pufeng Du, Xin Wang, Chao Xu, and Yang Gao. Pseaac-builder: A cross-platform stand-alone program for generating various special chou's pseudo-amino acid compositions, 2012.

[68] Ehsaneddin Asgari and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics, 2015.

[69] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation, 2020.

[70] Nils Strodthoff, Patrick Wagner, Markus Wenzel, and Wojciech Samek. Udsmprot: universal deep sequence models for protein classification, 2020.

[71] N Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[72] Kingma Da. A method for stochastic optimization, 2014.