PROTEIN DRUG INTERACTION FROM THEIR SEQUENCE

Using the SMILES and SEQUENCE information

Anup Adhikari

Supervisor
Dr. Surendra Shrestha

This Thesis is carried out as a part of the education at the Tribhuwan University and is therefore approved as a part of this education. However, this does not imply that the University answers for the methods that are used or the conclusions that are drawn.

Tribhuwan University, 2019
Institute of Engineering
Pulchowk Campus
Department of Electronics and Computer Engineering

Abstract

Protein and Drug interactions are long debated terms in the field of computational bioinformatics. Finding them based on molecular fingerprints and protein sequences alone is itself challenging as the process involving the true interaction depends on pathways, molecular properties, chaperones and more. Moreover the structural properties in the case of protein has different dimensions among which the efficient representation exists in the form of primary and secondary information. In this work the representation of drugs in fingerprints and proteins in sequence are used to generate features. The major components of feature vectors used in this work that bring the better prediction are PSSM-DT, Embedding and RPT vectors. These features are transformed to create suitable feature sets for training a deep learning algorithm using state of art technique. We use KIBA score to quantify the interaction to discriminate the similarly interacting proteins and drugs.

Acknowledgement

I would like to express the deepest appreciation to my supervisor and Head of Department of Electronics and Computer Engineering, Pulchowk Campus Dr. Surendra Shresthafor his guidance throughout the period of this work. His invaluable support, understanding and expertise have been very important in completing this work. It was a great honor for me to pursue my thesis under his supervision.

I pay my sincere gratitude Dr. Aman Shakya, to MSCSKE Coordinator for his supervision and help during this research work.

I am highly grateful to Prof. Dr. Shashidhar Ram Joshi, Prof. Dr. Subarna Shakya, Dr. Sanjeeb Prasad Pandey, Dr. Dibakar Raj Pant and Dr. Basanta Joshi for their encouragement and guidance.

I would like to express my heartily gratitude towards the Institute of Engineering, Pulchowk Campusalong with all my respected teachers, my friends, my family for giving me continuous support for their invaluable help.

Anup Adhikari

073 MSCS 652

Institute of Engineering

Contents

1 Introduction **7**

1.1 Background . . . 7

1.2 Statement of Problem . . . 8

1.3 Objectives . . . 8

1.4 Scope of Work . . . 8

1.5 Organization of Report . . . 8

2 Theorical Background **9**

2.1 No Free Lunch Algorithm . . . 9

2.2 Stacking Generalization . . . 9

2.3 Literature Review . . . 9

3 Methodology **10**

3.1 System Block Diagram . . . 10

3.2 Dataset . . . 10

3.2.1 KEGG . . . 10

3.2.2 UniProt and CHEMBL . . . 11

3.2.3 PSI-BLAST . . . 12

3.2.4 Residue feature . . . 13

3.3 Deep Learning Model . . . 13

3.3.1 Components description used from Tensorflow (Keras) . . . 14

4 Experiments and Results **17**

4.1 Experiments . . . 17

5 Conclusion **18**

5.1 First Section . . . 18

A R2RSRV **20**

Bibliography **22**

List of Figures

3.1 System Block Diagram 10

3.2 Data Distribution of KIBA-interaction scores, Drug Sequences and Protein Sequences . . . 11

3.3 Deep Learning Model 14

3.4 Dense Layer 15

3.5 Dropout Layer 15

3.6 Pooling Layer 15

3.7 Convolutional Neural Network 16

3.8 Long Short Term Memory 16

List of Tables

3.1 Inputs Used in the Deep Learning Network 13

A.1 R2RSRV Matrix 21

Chapter 1

Introduction

1.1 Background

Finding the interaction of drugs and proteins based simply on primary structure information of drugs and proteins is one of the many challenges faced in drug-synthesis process.

With the advent of new machine learning techniques and along with the rise of deep-learning techniques, we are closer to create a good prediction of analogy. However, the chemical properties of drugs and the targets complicate the situation as they react differently with slight change in protein sequence. Moreover, the complexes tend to behave similarly even when the protein sequences are distantly related, one of the results of tertiary structures that the proteins are form of.

The deep learning methods are quite good at predicting the molecular behaviour of the drug. However they present no good means when predicting the behaviour of proteins. The major fallback being that the simple encoding techniques don't incorporate the proteins behaviour related to hydrophobicity, acidity, secondary and tertiary structures information.

The Stacked Generalized Prediction on the other hand works by basing the prediction guesses based on a number of prediction functions. Here, we use the sequence information of proteins to calculate the predictions on different feature transformation techniques and generalize those predictions using a stack of dense layers. The Dataset we used scores the interaction of proteins and drugs based on Kb scores. We use 52498 drugs from CHEMBL and 254 proteins from UniProt to get an interaction of 180244, by removing the unrecognized interactions. The interactions are based on KIBA score, collected from KEGG(Kyoto Encyclopedia of Genes and Genomes) dataset [1].

1.2 Statement of Problem

1.3 Objectives

1.4 Scope of Work

1.5 Organization of Report

Chapter 2

Theorical Background

2.1 No Free Lunch Algorithm

2.2 Stacking Generalization

2.3 Literature Review

Chapter 3

Methodology

3.1 System Block Diagram

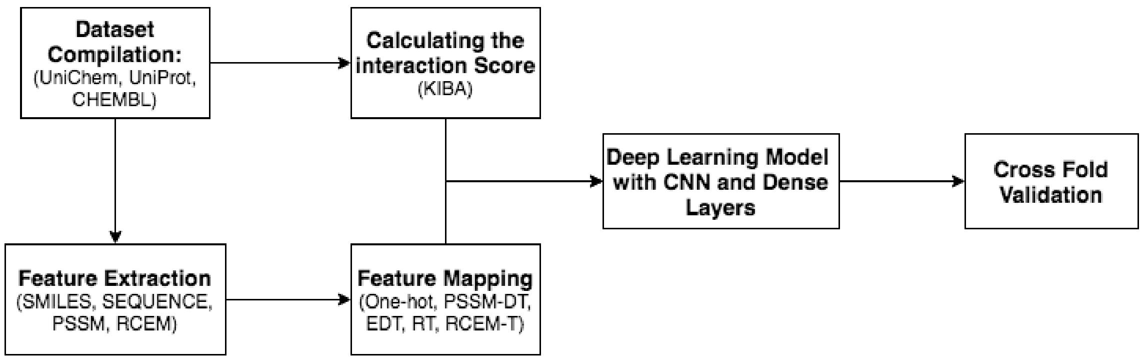


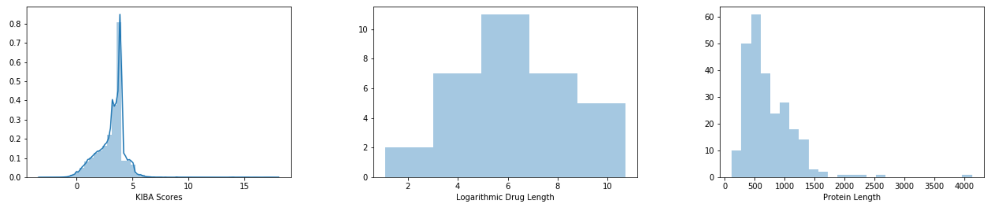
Figure 3.1: System Block Diagram

The figure 3.1 shows the various components used to form the prediction system. The idea is basic in that protein interaction depends on the structural and chemical properties. The structural components are fulfilled and

3.2 Dataset

3.2.1 KEGG

It is a community-driven database which holds large-scale molecular datasets generated by genome sequencing and high-throughput experimental techniuqe.[1] We use KEGG DRUG dataset for finding the interaction set between DRUG and PROTEIN. The interaction score is based on Equation 3.1:



(a) KIBA Scores (b) Logarithmic One Hot Encodings of Drug Sequence (c) One Hot Encodings of Protein Sequence

Figure 3.2: Data Distribution of KIBA-interaction scores, Drug Sequences and Protein Sequences

$$KIBA = \begin{cases} K_i.adj & if\ IC_{50}\ and\ K_i\ are\ present \\ K_b.adj & if IC_{50}\ and\ K_d\ are\ present \\ \frac{K_i.adj\ K_b.adj}{2} & if\ IC_{50}\ , K_i\ and\ K_d\ are\ present \end{cases} \tag{3.1}$$

where L_d and L_i are parameters defining weights of IC_{50} in model adjustments for K_i and K_b

For a kinase inhibitor drug−target interaction, we consider the medians of three major bioactivity types IC_{50}, K_i, K_d where IC_{50} [2] is the concentration at which the inhibitor causes a 50% inhibition of enzymatic activity and K_i is defined by

$$Ki = \frac{IC_{50}}{1+[S]K_m} \tag{3.2}$$

where, [S] is the experimental substrate concentration and K_m is the concentration of the substrate.

$$K_i.adj = \frac{IC_{50}}{1+L_i(IC_{50}/K_i)} \tag{3.3}$$

$$K_d.adj = \frac{IC_{50}}{1+L_d(IC_{50}/K_d)} \tag{3.4}$$

All the bioactivity types are available from CHEMBL.[3] We thus have 254 proteins and 52498 drugs. Based on interaction data available, we remove the unknown values and get a total of 180244 interaction KIBA score values in the range of -3.09 to 17.8. With the standard deviation of 1.22, we try to predict the best KIBA score of drug and protein based on the sequence information alone.

3.2.2 UniProt and CHEMBL

UniProt

The sequence related information of protein is referenced using UniProt Identifier and protein sequence (FASTA) is called using the api from UniProt. [4]

CHEMBL

The molecular fingerprints related to drugs are referenced usning CHEMBL Identifier and the drug sequence is called from CHEMBL database. [3]

3.2.3 PSI-BLAST

It relates with multiple sequence alignments from a family of protein sequences[5]. This helps us to create a PSSM 3.5 matrix referred to as secondary protein structure. The improvement in drug-contact prediction can be thought for amino acid composition being tuned with the scoring system. For this study, the PSSM profile of every protein sequence is obtained by executing iteration of PSI-BLAST against [5, KEGG] protein. PSSM profile is a matrix of L*20 dimensions where, 20 referring to standard type of amino acids and L being the length of the protein. The larger positive scores represent conserved positions, which in turn implies critical functional residues that are required to perform various intermolecular interactions.[5, PSSM]

$$PSSM = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,20} \\ P_{2,1} & P_{2,2} & \dots & P_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ P_{2,1} & P_{2,2} & \dots & P_{2,20} \end{bmatrix} \tag{3.5}$$

PSSM-DT

Two forms of PSSM distance transformation techniques are used to transform the PSSM information into fixed dimensional vectors [6]. The PSSM-DT (PSSM-Distance Transformation) can transform the PSSM information into uniform numeric representation by approximately measuring the occurance probabilities of any pairs of amino acid. It results in two types of feauture matrices: PSSM-SDT and PSSM-DDT defined by:

$$PSSM - SDT(i,lg) = \sum_{i=1}^{L-lg} S_{i,j} \times \frac{S_{i,j+lg}}{L-lg} \tag{3.6}$$

lg = distance of separation between same amino acid sequence

$$PSSM - DDT(i_1,i_2,lg) = \sum_{j=1}^{L-lg} S_{i_1,j} \times \frac{S_{i_2,j+lg}}{L-lg} \tag{3.7}$$

i_1 and i_2 refer to tow different types of amino acids

Thus we have (380 Eqn: 3.7+20 Eqn: 3.6 = 400) x lg matrix which will be used as protein-specific vector in this work.

Evolutionary Distance Transformation Matrix

The mutational information of protein can be more informative than the sequence information itself[7]. Evolutionary difference formula(EDF) is used to represent mutation difference between adjacent residues. Secondly, the PSSM is converted into 20 x 20 matrix (ED-PSSM). This extracts the non co-occurence probability for two amino acids separated by a certain distance *d* in the protein from the PSSM profile. For example, d=1 implies that the two amino acids are consecutive; d=2 implies that there is one amino acid between the two. Then the EDT feature vector computed from ED-PSSM can be represented as (3.8):

$$P = [\partial_1, \partial_2, \dots, \partial_\Omega] \tag{3.8}$$

where Ω is an integer that represents the dimension of the vector whose value is 400.. The non-co-occurance probability of two amino acids separated by distance *d* can be computed as:

$$f(A_x,A_y) = \sum_{d=1}^{D} \frac{1}{L-d} \sum_{i=1}^{L-d} (P_{i,x} - P_{i+d,y})^2 \tag{3.9}$$

where $P_{i,x}$ and $P_{i+d,y}$ are the elements in the PSSM profile; A_x and A_y represent any of the the 20 different amino acids in the protein sequence. Finally we spread the $f(A_x,A_y)$ in equation 3.8 as: $\partial_1 = f(A_1,A_2)$, $\partial_{400} = f(A_20,A_20)$

3.2.4 Residue feature

The Statistical Residue Vector Space R2RSRV [8] plays an important role in Residue Residue Interaction and thus creates a basis for structural stability of the protein sequence itself. Though related more to the tertiary structure of protein sequence itself, we regard it to create a correlated sequence information where two proteins are related distantly by sequence but highly related with functional characteristic of protein. Table A shows the table used in this work. It is a 20 x 20 matrix whose rows and columns represent 20 standard amino acids.

Residue Probing Transformation(RPT) feature

RPT as proposed by [9, Jeong et al.], and implemented by [10, Pujan et al.], emphasizes domains with similar conservation rates by grouping domain families based on their conservation score in the PSSM profile.

$$RPT = \begin{bmatrix} H_{1,1} & H_{1,2} & \dots & H_{1,20} \\ H_{2,1} & H_{2,2} & \dots & H_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ H_{2,1} & H_{2,2} & \dots & H_{2,20} \end{bmatrix} \tag{3.10}$$

The RPT matrix (Equation 3.10) is then tranformed into feature vector of 400 dimensions, as shown in Equation 3.11.

$$V = [f_{s_{1,1}}, f_{s_{1,2}}, \dots, f_{s_{i,j}}, \dots, f_{s_{20,20}}] \tag{3.11}$$

where,

$$f_{s_{i,j}} = \frac{s_{i,j}}{L}(i, j = 1, 2, \dots, 20) \tag{3.12}$$

3.3 Deep Learning Model

The Features thus formed are then subjected to deep learning model using keras library in python. We use the Embedding feature provided by keras as other features for both drug fingerprint and protein sequence. The implemented model is represented by Fig. 3.3. The input layers are described in Table 3.3.

S.No.	Input Layer Name	Used Feature Vector	Type
1	input_1	One Hot Encoding	Drug
2	input_2	One Hot Encoding	Protein
3	input_3	Evolutionary Distance Transformation	Protein
4	input_4	PSSM-DT	Protein
5	input_5	RDT	Protein

Table 3.1: Inputs Used in the Deep Learning Network

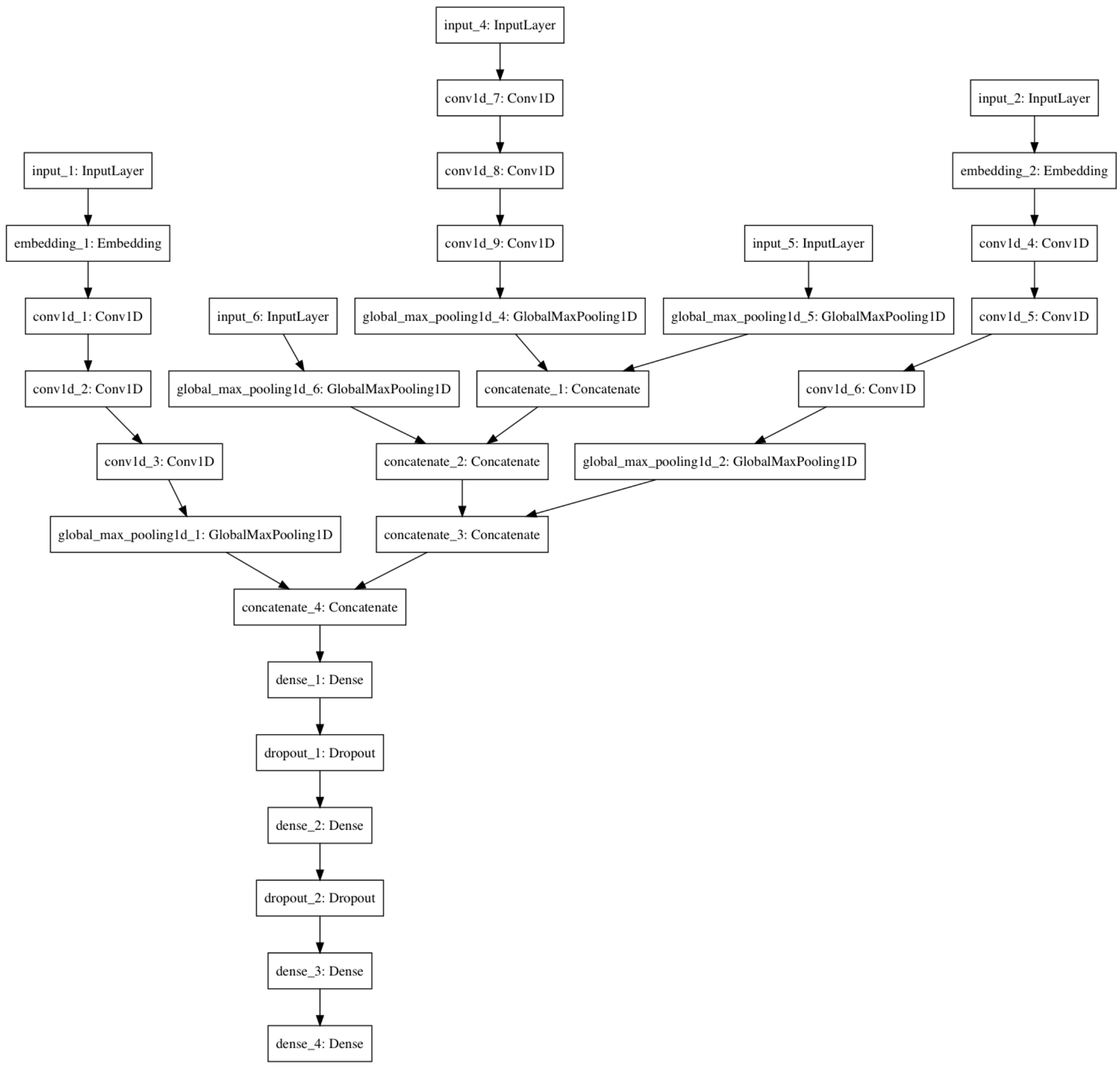


Figure 3.3: Deep Learning Model

3.3.1 Components description used from Tensorflow (Keras)

Embedding Layer

The one-hot encodings of the drugs and protein sequences are inputs to this layer. It turns positive integers (indexes) into dense vectors of fixed size. eg. [[4], [20]] -> [[0.25, 0.1], [0.6, -0.2]].

Dense Layer

It is a neural layer which fully connects the input layer to output layer. It can be seen from Figure 3.4.

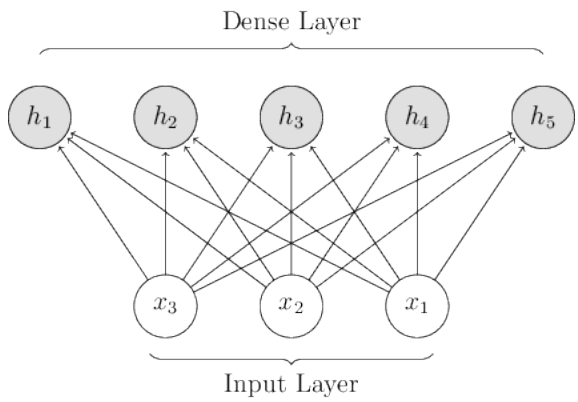


Figure 3.4: Dense Layer

Dropout Layer

It is undesirable when every component of the input layer makes a significant changes to the output layer. To reduce the effect of unimportant features we use dropout layer. Thus the backpropagation network tries to ignore the noise features and minimizes the unrealizable prediction of the learning problem. This can be expressed diagrammatically in the Figure 3.5.

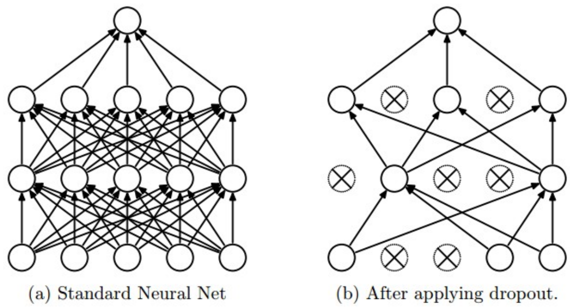


Figure 3.5: Dropout Layer

Global Max Pooling Layer

We use this to sample the learned parameters from the grid of 3 dimensions returned by Convolution Layer. It gets reduced to 1 dimension by taking the highest values from the window size(corresponding to shape of 1st dimensional element).

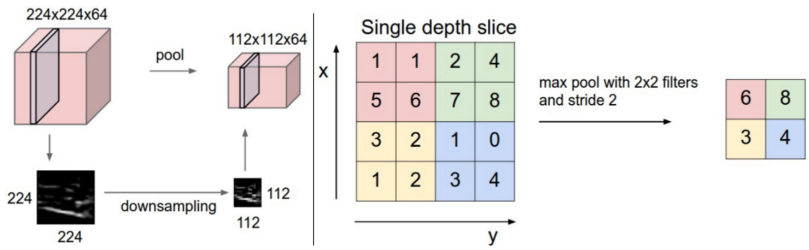


Figure 3.6: Pooling Layer

Concatenation Layer

It is used to simply join two vectors so that we create a feature set comprising of multiple features whose positional index indicates the feature set being manipulated.

Convolution Neural Network

To learn the local patterns in the input vector, we use CNN. While Dense Layers and LSTM learn the global patterns, CNN is used to understand the local patterns. It does so by increasing the depth layer, which in turn is designed to learn different patterns as shown in Figure 3.7.

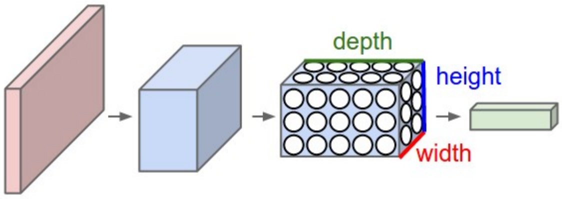


Figure 3.7: Convolutional Neural Network

LSTM

As the RNN often suffers from vanishing gradient problem, we use a LSTM Layer to learn the global pattern of the featuresets resulting after concatenation of different stacked layers outputs. The LSTM architecture can be seen in figure below:

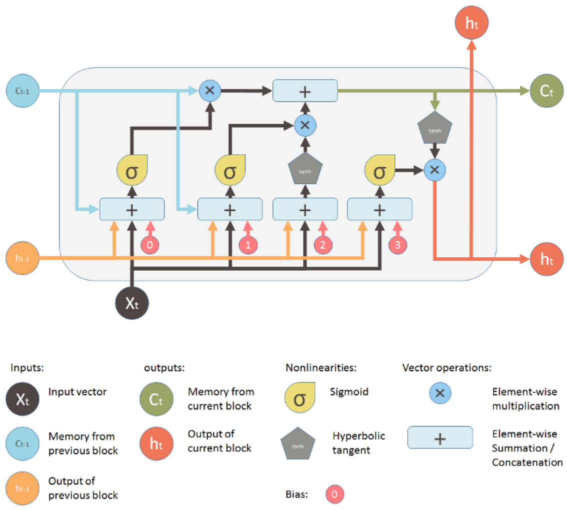


Figure 3.8: Long Short Term Memory

Chapter 4

Experiments and Results

4.1 Experiments

The simple technique of encoding the sequence information of drugs and proteins to identify if a drug will interact or not has a major pitfall in that while drugs encoding information can be used to make drug related predictions, the protein encodings live i

Chapter 5

Conclusion

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

5.1 First Section

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Appendix A

R2RSRV

	I	V	L	F	C	M	A	G	T	S	W	Y	P	H	E	Q	D	N	K	R
I	5.21	2.42	0.88	1.71	-1.59	1.13	0.95	0.48	-1.05	-3.20	0.65	1.44	-0.82	-1.54	-0.94	-0.62	-1.66	-3.14	-2.23	-2.14
V	2.42	9.46	1.33	0.49	-0.32	0.54	1.55	-2.12	-0.91	-1.80	-2.88	-1.05	-0.81	-1.32	-0.29	-0.58	-2.39	-3.69	0.66	-1.42
L	0.88	1.33	9.90	1.08	-0.42	2.17	2.41	-2.29	-3.40	-2.32	0.48	-0.77	-2.28	1.67	-0.77	-0.08	-3.49	-2.16	-2.10	0.19
F	1.71	0.49	1.05	6.11	0.55	0.89	0.52	-2.00	-1.10	-2.09	-0.11	1.14	0.83	-1.33	-1.79	0.42	-3.62	-0.96	-1.71	-1.33
C	-1.59	-32	-0.42	0.55	15.35	-1.35	-0.21	0.59	-1.52	1.53	-1.07	-1.16	0.28	0.95	-0.52	-1.47	-1.95	-2.23	-1.80	-0.84
M	1.13	0.54	2.17	0.89	-1.35	5.40	-0.28	0.44	-2.15	-1.50	-0.71	-0.33	-0.31	0.19	0.01	0.27	-3.38	-1.74	-0.72	-1.51
A	0.95	1.55	2.41	0.52	-0.21	-0.28	7.08	-2.04	-1.04	-0.61	-1.15	-1.22	-1.58	0.11	-0.53	-0.82	-1.06	0.17	-1.11	-2.74
G	0.48	-2.12	-2.29	-2.00	0.59	0.44	-2.04	5.65	1.67	-1.32	-0.82	0.27	-0.60	0.75	-2.24	1.68	0.70	-1.01	1.72	1.22
T	-1.05	-0.91	-3.40	-1.10	-1.52	-2.15	-1.04	1.67	4.42	1.23	0.59	-1.36	-0.04	-1.48	-0.06	-2.61	4.66	0.02	0.29	-0.74
S	-3.20	-1.80	-2.32	-2.09	1.53	-1.50	-0.61	-1.32	1.23	6.22	-1.10	-1.40	-0.79	-2.66	2.14	-0.08	4.57	0.95	0.11	-0.38
W	0.65	-2.88	0.48	-0.11	-1.07	-0.71	-1.15	-0.82	0.59	-1.10	1.08	-0.45	5.88	0.15	-2.84	-2.84	-1.98	-1.35	-0.27	4.08
Y	1.44	-1.05	-0.77	1.14	-1.16	-0.33	-1.22	0.27	-1.36	-1.40	-0.45	6.40	0.21	1.11	0.75	-2.73	-3.07	-0.45	0.87	-0.33
P	-0.82	-0.81	-2.28	0.83	0.28	-0.31	-1.58	-0.60	-0.04	-0.79	5.88	0.21	1.73	-1.13	0.66	0.82	-2.51	1.37	0.14	-0.40
H	-1.54	-1.32	1.67	-1.33	0.95	0.19	0.11	0.75	-1.48	-2.66	0.15	1.11	-1.13	5.03	-2.22	0.32	3.11	-1.46	-1.90	-0.06
E	-0.94	-0.29	-0.77	-1.79	-0.52	0.01	-0.53	-2.24	-0.06	2.14	-2.84	0.75	0.66	-2.22	2.59	-1.98	-4.29	0.07	3.52	3.45
Q	-0.62	-0.58	-0.08	0.42	-1.47	0.27	-0.82	1.68	-2.61	-0.08	-2.84	-2.73	0.82	0.32	-1.98	3.44	0.79	0.92	-0.67	0.24
D	-1.66	-2.39	-3.49	-3.62	-1.95	-3.38	-1.06	0.70	4.66	4.57	-1.98	-3.07	-2.51	3.11	-4.29	0.79	1.69	3.85	0.86	2.73
N	-3.14	-3.69	-2.16	-0.96	-2.23	-1.74	0.17	-1.01	0.02	0.95	-1.35	-0.45	1.37	-1.46	0.07	0.92	3.85	7.91	-0.63	-0.43
K	-2.23	0.66	-2.10	-1.71	-1.80	-0.72	-1.11	1.72	0.29	0.11	-0.27	0.87	0.14	-1.90	3.52	-0.67	0.86	-0.63	2.61	-3.54
R	-2.14	-1.42	0.19	-1.33	-0.84	-1.51	-2.74	1.22	-0.74	-0.38	4.08	-0.33	-0.40	-0.06	3.45	0.24	2.73	-0.43	-3.54	0.73

Table A.1: R2RSRV Matrix

Bibliography

[1] M. Kanehisa. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, jan 2000.

[2] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis. *Journal of Chemical Information and Modeling*, 54(3):735–743, mar 2014.

[3] Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, jan 2017.

[4] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 46(5):2699–2699, mar 2018.

[5] A. A. Schaffer. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Research*, 29(14):2994–3005, jul 2001.

[6] Ruifeng Xu, Jiyun Zhou, Hongpeng Wang, Yulan He, Xiaolong Wang, and Bin Liu. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Systems Biology*, 9(1):1–12, 2015.

[7] Lichao Zhang, Xiqiang Zhao, and Liang Kong. Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition. *Journal of Theoretical Biology*, 355:105–110, aug 2014.

[8] Andrew K.C. Wong, Ho Yin Sze-To, and Gary L. Johanning. Pattern to Knowledge: Deep Knowledge-Directed Machine Learning for Residue-Residue Interaction Prediction. *Scientific Reports*, 8(1):1–14, 2018.

[9] Jong Cheol Jeong, Xiaotong Lin, and Xue Wen Chen. On position-specific scoring matrix for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):308–315, 2011.

[10] Avdesh Mishra, Pujan Pokhrel, and Md Tamjidul Hoque. Thesis – StackDPPred: a stacking based prediction of DNA-binding protein from sequence. *Bioinformatics*, 35(3):433–441, feb 2019.