Protein-Protein Interaction Classification Using Jordan Recurrent Neural Network

Dilpreet Kaur
Department of Computer Science and Engineering
PEC University of Technology
Chandigarh, India
dilpreet.kaur88@gmail.com

Dr. Shailendra Singh
Associate Professor, Department of Computer Science and Engineering
PEC University of Technology
Chandigarh, India
sscse@in.com

Abstract- **Proteins form a very important part of a living cell. The biological functions are carried out by the proteins within the cell by interacting with other proteins in other cells. This is called protein-protein interaction. Protein-Protein Interactions are very important in understanding the diseases and finding their cause. It can also provide the basis for new therapeutic approaches. A number of classifiers have been developed till date to classify protein-protein interactions namely SVM, SVM-KNN, Back-propagation Neural Network (BPNN). In this work Jordan Recurrent Neural Network (JRNN) is used to classify the protein-protein interactions. The classifier developed for this work uses amino acid composition of proteins as input to classify the percentage of interacting and non-interacting proteins. The results obtained were best at the threshold value of zero. The classifier gives an accuracy of 97.25% which is 8.7% more than BPNN. The overall accuracy of JRNN for threshold ranging from -1 to +1 with a difference of 0.1 comes out to be 80.1%.**

Keywords: Protein-Protein Interaction; Jordan Recurrent Neural Network (JRNN); Amino acid composition; Back-Propagation Neural Network

I. INTRODUCTION

Bioinformatics is a conceptualization of biology in terms of molecules i.e. in sense of physical-chemistry and then applying informatics techniques, derived from math, computer science and statistics, to understand and organize the information associated with these molecules on a large scale. Bioinformatics is more of a tool than a discipline, the tools for analysis of Biological Data. The primary goal of bioinformatics is focus on developing and applying computationally intensive techniques (e.g. pattern recognition, data mining, machine learning algorithms and visualization) to increase the understanding of biological processes. The bioinformatics is extremely broad and is rapidly changing, particularly in recent years. The current scope of bioinformatics is mainly at bimolecular level particularly on macromolecules such as DNA, RNA. Proteomics one of the fields of bioinformatics deals with the study of proteins especially its structure and function. Proteins work in collaboration with other proteins so the main goal of proteomics is predict the proteins that interact [14]. Prediction and classification of interacting and non-interacting proteins is helpful in improving the understanding of diseases. Protein-Protein Interaction has become a very important research area now a days. Protein-Protein Interactions occur when proteins bind together to carry out some biological function. The most important molecular process in a cell such as DNA replication is carried out by large number of protein components organizes by their protein-protein interactions [15]. Protein interactions are studied in the aspect of biochemistry, quantum chemistry, molecular dynamics, chemical biology, signal transduction and other metabolic or genetic/epigenetic networks. Most of the biological functions are performed due to the protein-protein interactions. For example, signals from the exterior of a cell are mediated to the inside of that cell by protein–protein interactions of the signaling molecules. This process, called signal transduction, plays a fundamental role in many biological processes and in many diseases.

The classification of interacting and non-interacting proteins has been done using various classifiers till date namely SVM [1], SVM-KNN [2], BPNN [3] but no classifier gave better accuracy. In this work a classifier is developed using Jordan Recurrent Neural Network (JRNN). The JRNN classifier takes amino acid composition of proteins as input. Amino acid composition has been calculated for different purposes in [4] [5]. In [4] the local composition or composition profile of patters is calculated by the authors i.e. a pattern is represented by the amino acid composition.

This paper is divided into different sections that include the material and method that are used to develop the classifier, results given by the classifier and at the end the conclusion and future scope of the work is discussed.

II. MATERIALS AND METHODS

Protein–protein interaction prediction is a field combining bioinformatics and structural biology in an attempt to identify and catalog physical interactions between pairs or groups of proteins. This section describes various materials and methods used to develop the classifier for protein-protein interaction classification. This section tells about the data material used, amino acid calculation of proteins, and explanation of Jordan recurrent neural network. The phases involved in the development of the classifier are shown in Figure 1.

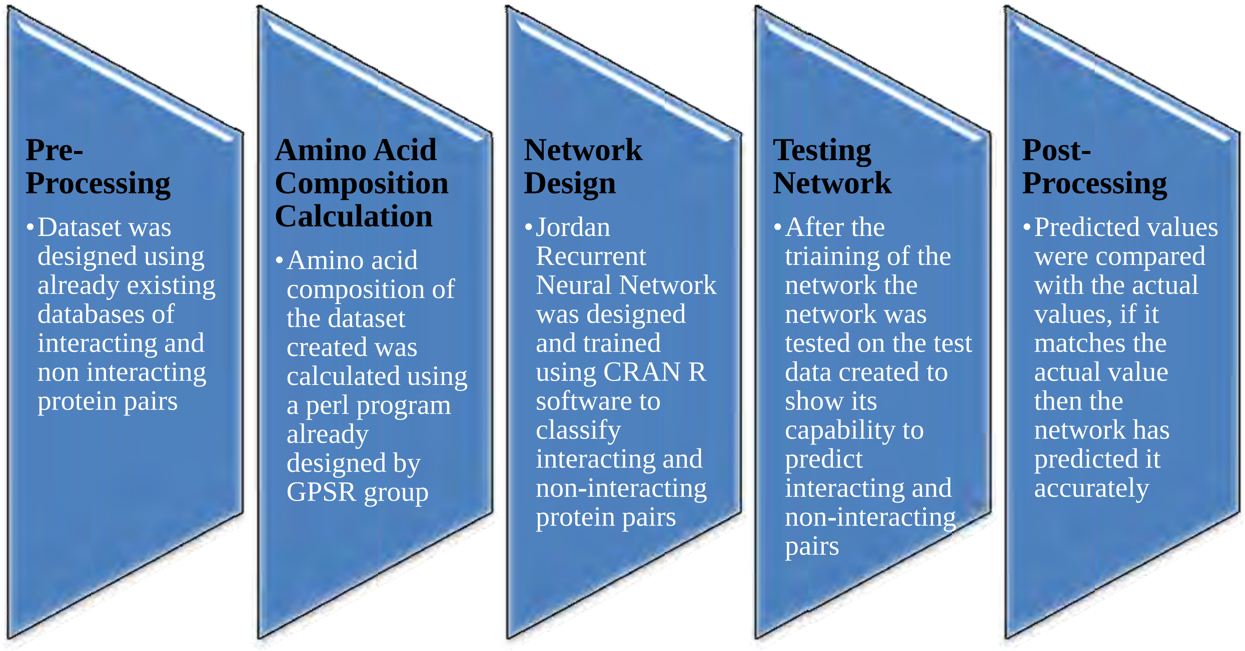
Pre-Processing
• Dataset was designed using already existing databases of interacting and non interacting protein pairs
Amino Acid Composition Calculation
• Amino acid composition of the dataset created was calculated using a perl program already designed by GPSR group
Network Design
• Jordan Recurrent Neural Network was designed and trained using CRAN R software to classify interacting and non-interacting protein pairs
Testing Network
• After the traiining of the network the network was tested on the test data created to show its capability to predict interacting and non-interacting pairs
Post-Processing
• Predicted values were compared with the actual values, if it matches the actual value then the network has predicted it accurately

Figure 1. Phases of Classifier Development

A. Dataset of Interacting and Non-Interacting Proteins

Proteins perform their functions within a cell by interacting with each other and by passing signals to other proteins. A number of protein databases has been developed in past years by various researchers. The major protein databases developed includes UniProt [6], SwissProt [7], PDB [8], HPRD [9] and Pfam [10]. In this work a dataset is developed from already existing databases namely Pfam [10], 3DID [11], Negatome [12], DSSP [13]. The dataset developed have equal number of interacting and non-interacting proteins, in which the positive patterns were randomly picked from the pool of interacting proteins. Positive patterns contain interacting residues in its center while negative patterns contain non-interacting residues in its center. This dataset is used because machine-learning techniques are more efficient in learning when negative and positives patterns are equal. The dataset developed includes 753 positive patterns and 656 negative patterns.

B. Amino Acid Composition Calculation

The most typical sequential representation for a protein sample is its entire amino acid (AA) sequence, which can contain its most complete information. This is an obvious advantage of the sequential model [16]. However, this kind of approach failed to work when a query protein did not have significant homology to the attribute-known proteins. Thus, various discrete models were proposed.

The simplest discrete model is using the amino acid composition (AAC) to represent protein samples, as formulated as follows. Given a protein sequence P with L amino acid residues, I [16].

(1)

where R_1 represents the 1st residue of the protein P, R_2 represents the 1st residue of the protein P and so forth according to the amino acid composition (AAC) model, the protein P of Eq.1 [16] can be expressed by

(2)

where f_i (i=1,2,3,........,20) are the normalized occurrence frequencies of the 20 native amino acids in P and T the transposing operator. Accordingly, the amino acid composition of a protein can be easily derived once the protein sequencing information is known.

In this work the sequence is represented by a vector a vector of dimension 21 as used in [4] which represents twenty natural amino acids and one dummy amino acid ‘‘X’’. Amino acid composition of a pattern was computed using the following formula [4] [5]:

$$comp(i) = \frac{R(i)}{N} \qquad (3)$$

where comp(i) is the fraction of residue or composition of residue of type i. Ri and N are number of residues of type i and total the number of residue in protein i (length of protein) respectively.

C. Jordan Recurrent Neural Network

The Jordan Neural Network is a simple recurrent network (SRN) developed by Michael I. Jordan [18] in 1986. The context layer holds the previous output from the output layer and then echos that value back to the hidden layer's input. The hidden layer then always receives input from the previous iteration's output layer [17]. Jordan neural networks are generally trained using genetic, simulated annealing, or one of the propagation techniques. Jordan neural networks are typically used for prediction. The architecture of Jordan Recurrent Neural Network is shown in Fig. 2.

In this work a Jordan Recurrent Neural Network based classifier is designed using RSNNS [19] package of CRAN R [20]. The network used five-fold cross validation to train and test the input data. The neural network used JE_BP learning function, which is a standard back-propagation training function, to train the network.

III. RESULTS

The results of the Jordan Recurrent Neural Network classifier are shown in Table I. There are a total of 1379 protein pairs that are taken out of which 753 are interacting protein pairs and 656 are non-interacting protein pairs.

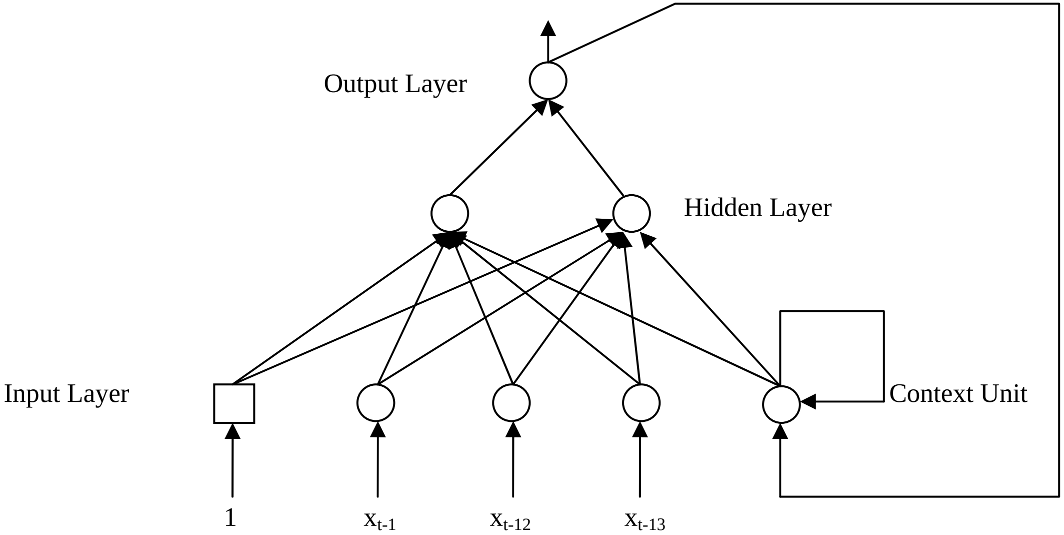

Figure 2. Jordan Recurrent Neural Network

From the confusion matrix shown in table I the sensitivity of Jordan recurrent neural network classifier is found to be 95.09% and the specificity is 99.84%. These values show that Jordan recurrent neural network classifier can differentiate between interacting and non-interacting protein pair with high probability. The positive predictive value (PPV) and negative predictive value (NPV) are calculated to be 99.86% and 94.41% respectively. The high values of PPV indicate that Jordan recurrent neural network classifier can correctly identify interacting protein pairs.

TABLE I CONFUSION MARTIX FOR JORDAN RECURRENT NEURAL NETWORK CLASSIFIER

	Positive	Negative	
Positive	TP 717	FP (Type I Error) 1	PPV= 99.86%
Negative	FN (Type II Error) 37	TN 625	NNV= 94.41%
	Sensitivity= 95.09%	Specificity= 99.84%	

A. Comparison with Back-Propagation Neural Network

The comparison of Jordan neural network classification model is done with Back-propagation Neural Network [3] is done on the basis of sensitivity, specificity and accuracy. Table II shows the specificity, sensitivity and accuracy values for JRNN and BPNN.

TABLE II SPECIFICITY, SENSITIVITY AND ACCURACY VALUES OF BPNN & JRNN

Classifier/ Parameter	Specificity	Sensitivity	Accuracy
BPNN	86.0	91.1	88.5
JRNN	99.84	95.9	97.25

The Specificity comparison of Jordan neural network classification model with Back-propagation Neural Network is shown in Figure 3.

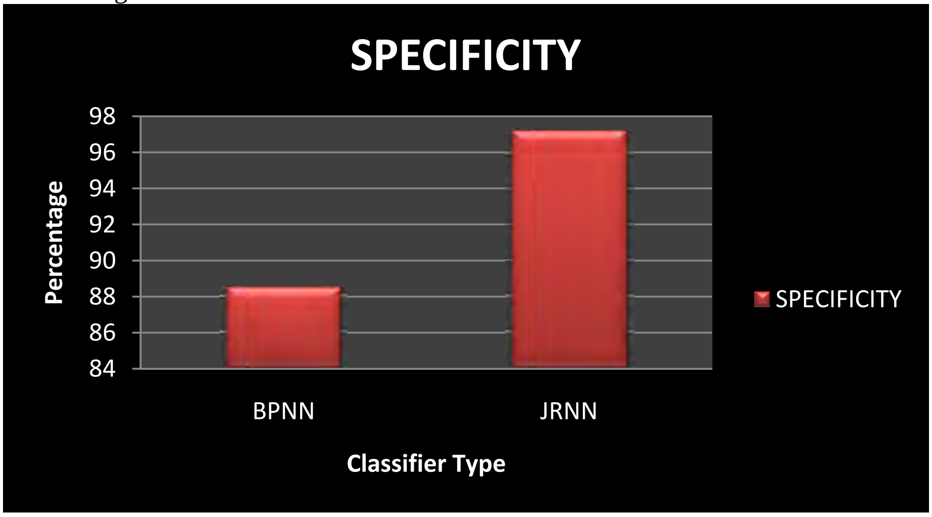

Figure 3. Specificity Comparison of BPNN and JRNN

The Sensitivity comparison of Jordan neural network classification model with Back-propagation Neural Network is shown in Figure 4. The value of sensitivity gives the percentage of interacting proteins classified as interacting.

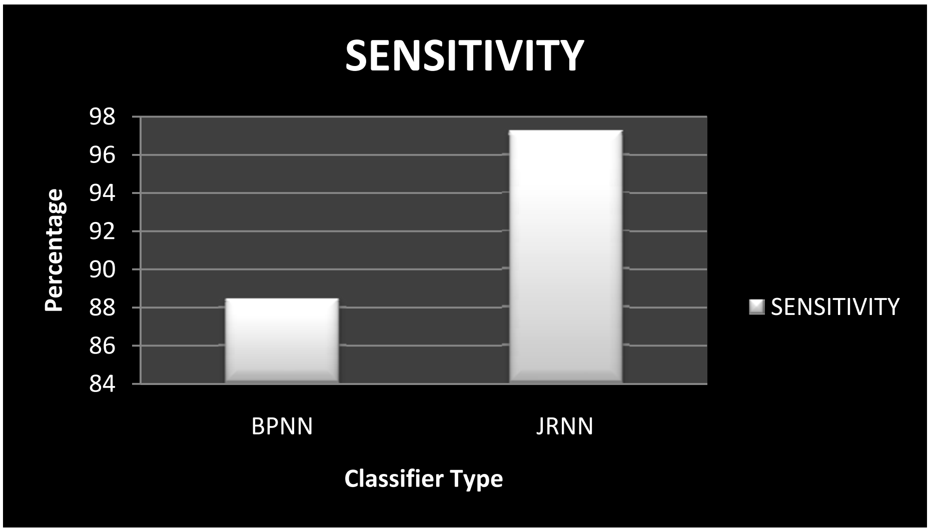

Figure 4. Sensitivity Comparison of BPNN and JNNCM

The Accuracy comparison of Jordan neural network classification model with Back-propagation Neural Network is shown in Fig. 5.

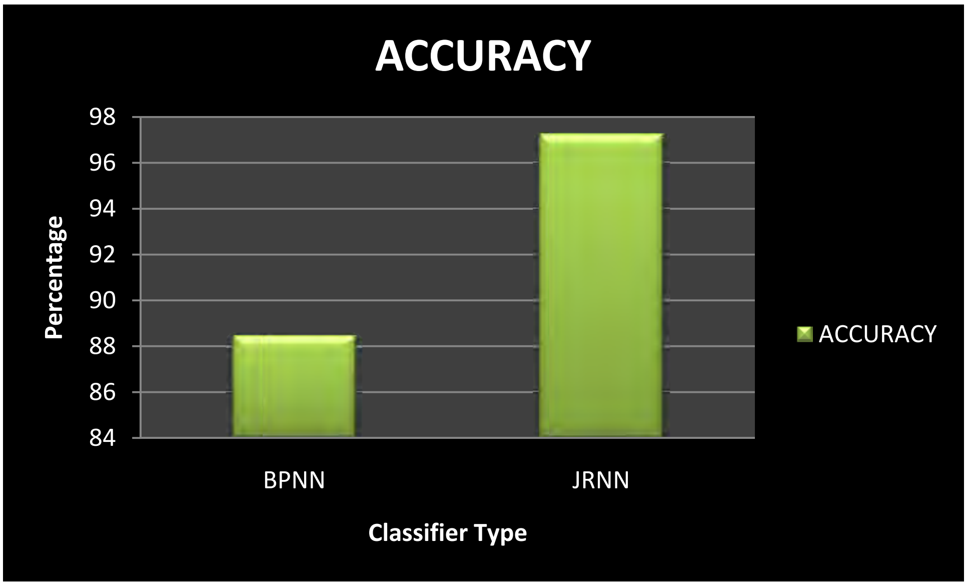

Figure 5. Accuracy Comparison of BPNN and JNNCM

B. Discussion

The Jordan neural network classification model used the amino acid composition of protein pairs as input to predict and classify interacting and non-interacting protein pairs. The accuracy of Jordan neural network classification model has increased by 8.7%. The accuracy improvement has helped to better classify interacting and non-interacting protein pairs. Jordan neural network classification model can classify protein pairs as interacting and non-interacting protein pairs with an accuracy of 97.25% i.e. Jordan neural network classification model can correctly identify up to 97.25% of protein pairs as pairs with and without interactions.

The analysis, interpretation and comparison of JRNN with various techniques for the classification of interacting and non-interacting protein pairs prove that Jordan recurrent neural network classifier is a better method for classification among interacting and non-interacting protein pairs.

IV. CONCLUSION AND FUTURE SCOPE

Proteomics is the large-scale study of proteins, particularly their structures and functions. Proteins are vital parts of living organisms, as they are the main components of the physiological metabolic pathways of cells. A number of techniques have been developed for the identification and classification of protein-protein interactions. The techniques developed in past years are still far from perfect. The Jordan neural network classification model tries to overcome this problem. The Jordan Neural Network takes amino acid composition of protein pairs to classify them interacting and non-interacting. On comparing, Jordan neural network classification model is found to have higher accuracy (97.25%) as compared to BP neural network (88.55). The percentage improvement is 8.7%.

Jordan neural network classification model outperforms the other methods for protein-protein interaction classification. Jordan neural network classification model proves to be better model with higher accuracy along with improved specificity and sensitivity than the various existing techniques.

A. Future Scope

Jordan recurrent neural network classifier the input given had almost equal positive and negative patterns. It gives the output which shows very good results nearly equal to perfect. In this model the input can be changed i.e. the input file can be altered having more negative patterns and less positive patterns as compared to the negative patterns to get better results than the results given by Jordan neural network classification model with input file having equal negative and positive patterns.

The Jordan Neural Network can also use other parameters related to proteins to predict and classify protein-protein interactions. These parameters include the six physiochemical properties of proteins namely assessable residues, buried residues, hydrophobicity, molecular weight, polarity and average area buried as used in [3].

REFERENCES

[1] Lishuang Li, Linmei Jing et. al., “Protein-Protein Interaction Extraction from Biomedical Literatures Based on Modified SVM-KNN”, IEEE International Conference on Natural Language Processing and knowledge Engineering, pp. 1-7, 2009.

[2] Hong-Wei Liu, “Protein-Protein Interaction Detection by SVM from Sequence Information”, The Third International Symposium on Optimization and Systems Biology, pp. 198-206, 2009.

[3] Zhiqiang Ma, Chunguang Zhou et. al., “Predicting Protein-Protein Interactions Based on BP Neural Network” IEEE Conference on Bioinformatics and Biomedicine Workshops, pp. 3-7, 2007.

[4] Agarwal S, Singh H et. al. “Identification of Mannose Interacting Residues Using Local Composition”, PLoS ONE, 2011.

[5] Gajendra PS Raghava, Joon H Han, “Correlation and prediction of gene expression level from amino acid and dipeptide composition of its protein”, BMC Bioinformatics, 2005.

[6] Cathy H. Wu, Rolf Apweiler, Amos Bairoch et. al., “The Universal Protein Resource (UniProt): an expanding universe of protein information”, Nucleic Acids Research,vol. 34, pp. 187–191, 2006.

[7] Amos Bairoch, Rolf Apweiler, “The SWISS-PROT protein sequence data bank and its supplement TrEMBL”, Nucleic Acids Research, vol. 25, no. 1, pp. 31–36, 1997.

[8] Helen M. Berman, John Westbrook et. al., “The Protein Data Bank”, Nucleic Acid Research, vol. 28, no.1, pp. 235-242, 2000.

[9] Suraj Peri, J. Daniel Navarro, Troels Z. Kristiansen, Ramars Amanchy et. al., “Human protein reference database as a discovery resource for proteomics”, Nucleic Acids Research, vol. 32, pp. 497-501, 2004.

[10] Robert D. Finn, John Tate et. al., “The Pfam protein families database”, Nucleic Acids Research, vol. 36, pp. 281–288, 2008.

[11] Amelie Stein, Robert B. Russell and Patrick Aloy, “3did: interacting protein domains of known three-dimensional structure”, *Nucleic Acids Research,* vol. 33, pp. 413–417, 2005.

[12] Pawel Smialowski, Philipp Page et. al., “The Negatome database: a reference set of non-interacting protein pairs”, Nucleic Acids Research, pp. 1–5, 2009.

[13] Kabsch W, Sander C, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features”, Biopolymers, pp. 2577-2637, 1983.

[14] http://en.wikipedia.org/wiki/Proteomics

[15] http://en.wikipedia.org/wiki/Protein%E2%80%93proteininteraction

[16] http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition

[17] http://www.heatonresearch.com/wiki/Jordan_Neural_Network

[18] Jordan, M.I., “Serial Order: A parallel Distributed Processing Approach”, Tech. rep. Report, pp. 86-104, 1986.

[19] Christoph Bergmeir, José M. Benítez, “Neural Networks in R using the Stuttgart Neural Network Simulator”, Repository CRAN, 2012.

[20] W. N. Venables, D. M. Smith, “R: A Programming Environment for Data Analysis and Graphics”, Version 2.15.0, 2012.