Recurrent Neural Network based Classification of Protein-Protein Interactions

Dilpreet Kaur
ME Final Year Student
PEC University of Technology
Chandigarh

Shailendra Singh
Phd, Associate Professor
PEC University of Technology
Chandigarh

ABSTRACT
Proteomics is an attempt to describe or explain biological state and qualitative and quantitative changes of protein content of cells and extracellular biological materials under different conditions to further understand biological processes. Protein-Protein interaction prediction and classification is a very important task. Prediction and classification of protein-protein interactions can help in improving the understanding of diseases and can provide the basis for new therapeutic approaches. In this work a model is proposed to classify protein-protein interactions. Jordan Recurrent Neural Network is used to classify the protein-protein interactions. The model developed gives 97.25% of accuracy which is 8.7% more than Back-Propagation Neural Network.

Keywords
Protein-Protein Interactions, Jordan Recurrent Neural Network, Back-Propagation (BP) Neural Network, SVM, SVM-KNN, Amino Acid Composition

1. INTRODUCTION
Bioinformatics is the science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, structural biology, software engineering, data mining, image processing, modeling and simulation, signal processing, discrete mathematics, control and system theory, circuit theory and statistics [18]. Bioinformatics generates new knowledge as well as the computational tools to create that knowledge. Proteomics is a field of bioinformatics that deals with the study of proteins, particularly their structures and functions [19]. Proteomics is an attempt to describe or explain biological state and qualitative and quantitative changes of protein content of cells and extracellular biological materials under different conditions to further understand biological processes. Proteins function in collaboration with other proteins so it is the main goal of proteomics to identify the proteins that interact [19]. Protein-Protein interaction prediction and classification is a very important task. Prediction and classification of protein-protein interactions can help in improving the understanding of diseases and can provide the basis for new therapeutic approaches [20].

Protein–protein interactions occur when two or more proteins bind together, often to carry out their biological function. A large number of protein components organized by their protein-protein interactions helps in carrying out the most important molecular process in the cell such as DNA replication [20]. 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.

A number of classifiers has been developed till date for the classification of protein-protein interactions namely SVM [5], SVM-KNN [4], BP Neural Network [3] but no classifier gave better accuracy. In this work, a new model for the prediction and classification of protein-protein interactions is presented. In Jordan Neural Network Classification Model (JNNCM) amino acid composition is used as an input for classification. Amino acid composition gives the percentage occurrence of a particular amino acid in a protein sequence. Amino acid composition has been used in [1] [14] for different purposes. In [1] authors used local composition or composition profile of patterns (CPP). It means that they represented a pattern by amino acid composition (AAC).

This paper is divided into different sections that include Materials and Methods that are used to develop the classification model, results given by the classification model. At the end the conclusion and future scope of the model is discussed.

2. MATERIALS AND METHODS
In Protein–protein interaction prediction field bioinformatics and structural biology is combined to identify and catalog physical interactions between pairs or groups of proteins [17]. In this section prediction and classification of protein-protein interactions is described. The phases involved in developing the method are shown in Fig 1.

2.1 Dataset of Interacting Proteins
Proteins perform their function by interacting with each other and by transmitting signals to other proteins. In past years, a number of protein and protein interaction databases have been developed by researchers to conduct further experimental work. 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 available datasets namely Pfam [10], 3DID [11], Negatome [12], DSSP [13].The dataset developed have equal number of

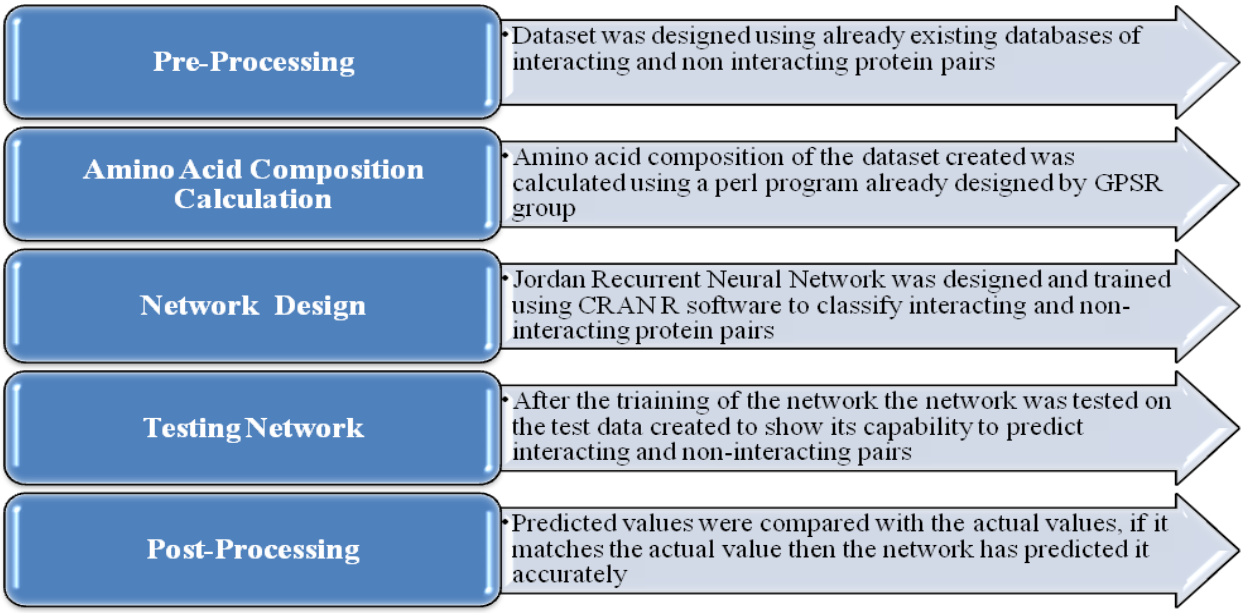

Fig 1: Phases of Classification Model

positive and negative patterns, where positive patterns were randomly picked up from the pool of positive patterns. 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.

2.2 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 [21]. 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 [21].

$$P=[R_1 R_2 R_3 R_4 R_5 R_6 R_7 \ldots . R_L] \quad (1)$$

where R_1 represents the 1st residue of the protein P, R_2 represents the 2nd residue of the protein P and so forth, according to the amino acid composition (AAC) model, the protein **P** of Eq.1 [21] can be expressed by

$$P=[f_1 \quad f_2 \quad \ldots\ldots \quad f_{20}]^T \quad (2)$$

where f_u (u= 1, 2, ……, 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 was be represented by a vector of dimension 21 as used in [1], which represents twenty natural amino acids and one dummy amino acid ‘‘X’’. Amino acid composition of a pattern was computed using the following formula [1] [14]:

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

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.

2.3 Jordan Recurrent Neural Network

The Jordan Neural Network is a simple recurrent network (SRN) developed by Michael I. Jordan [2] 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 [22]. 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 is designed using RSNNS [15] package of CRAN R [16]. 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.

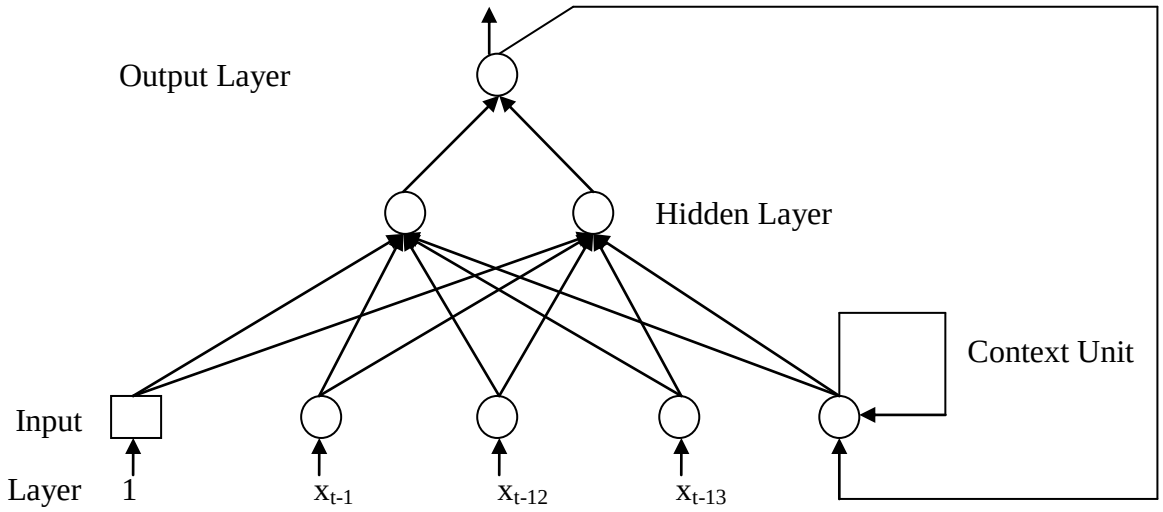

Fig2: Jordan Recurrent Neural Network

3. RESULTS

The results of the prediction and classification of interacting and non-interacting protein pairs using Jordan neural network classification model is shown in Table 1. 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 pair

Table 1. Confusion Matrix for Jordan Neural Network Classification Model

	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%	

From the confusion matrix shown in table 1 the sensitivity of Jordan neural network classification model is found to be 95.09% and the specificity is 99.84%. These values show that Jordan neural network classification model 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 neural network classification model can correctly identify interacting protein pairs.

3.1 Comparison with Known Classifiers

The results obtained are then compared with other known classifiers that have been used in past few years to classify interacting and non-interacting proteins. Jordan neural network classification model is compared with SVM [5] and SVM-KNN [4] on the basis of precision, recall and F-score. The comparison with Back-Propagation neural network [3] based classifier is done on the basis of specificity, sensitivity and accuracy.

3.1.1 Comparison with SVM and SVM-KNN

Table 2 shows the Recall, Precision and F-score values of Jordan neural network classification model, SVM and SVM-KNN.

Table 2. Recall, Precision and F-Score values of SVM, SVM-KNN and JNNCM

Classifier/ Parameter	RECALL	PRECISION	F-SCORE
SVM	84.2	63.9	72.7
SVM-KNN	87.5	70.5	82.4
JNNCM	97.9	99.86	95

The Recall comparison of Jordan neural network classification model with SVM [5] and SNM-KNN [4] is shown in Fig. 3.

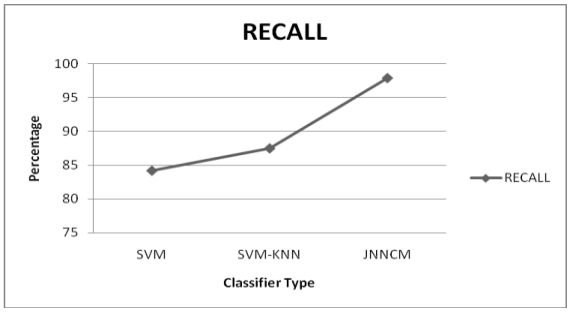

Fig 3: Recall Comparison of PPI Classifiers

The Precision comparison of Jordan neural network classification model with SVM [5] and SNM-KNN [4] is shown in Fig. 4.

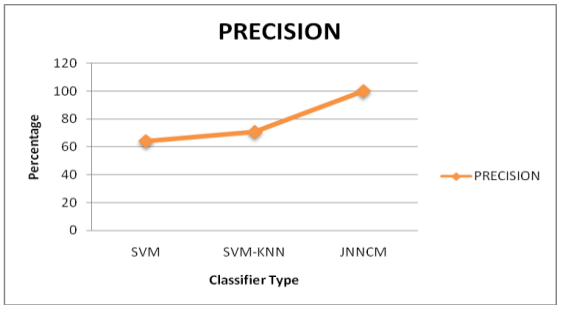

Fig 4: Precision Comparison of PPI Classifiers

The F-Score comparison of Jordan neural network classification model with SVM [5] and SNM-KNN [4] is shown in Fig. 5.

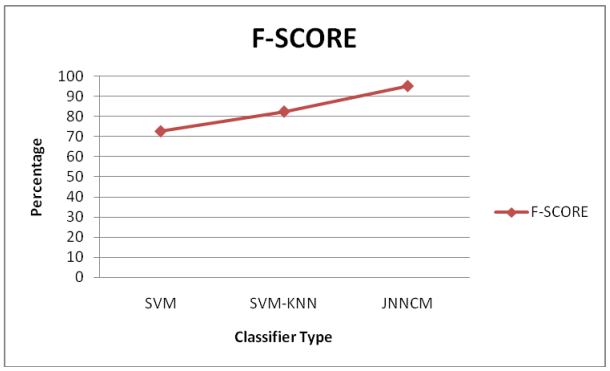

Fig 5: F-Score Comparison of PPI Classifiers

3.1.2 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 3 shows the specificity, sensitivity and accuracy values for JNNCM and BPNN.

Table 3. Specificity, Sensitivity and Accuracy Values of BPNN and JNNCM

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

The Specificity comparison of Jordan neural network classification model with Back-propagation Neural Network is shown in Fig. 6.

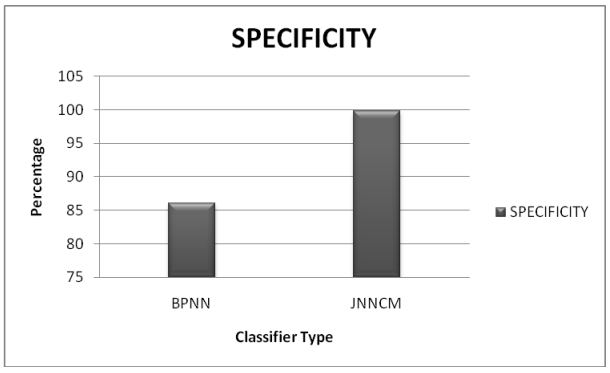

Fig 6: Specificity Comparison of BPNN and JNNCM

The Sensitivity comparison of Jordan neural network classification model with Back-propagation Neural Network is shown in Fig. 7. The value of sensitivity gives the percentage of interacting proteins classified as interacting.

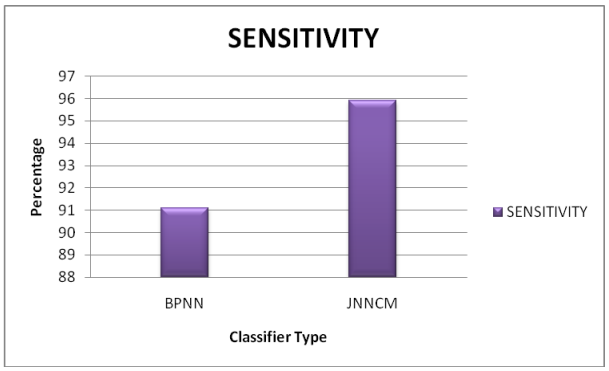

Fig 7: Sensitivity Comparison of BPNN and JNNCM

The Accuracy comparison of Jordan neural network classification model with Back-propagation Neural Network is shown in Fig. 8.

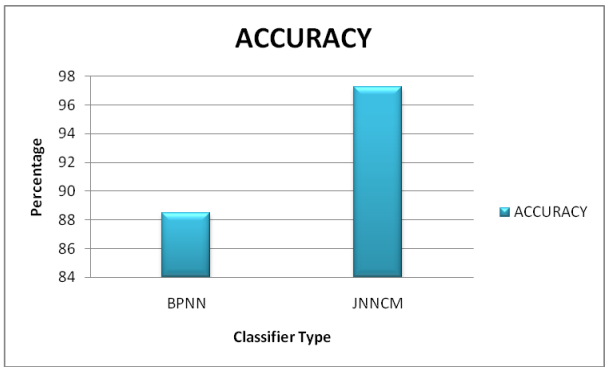

Fig 8: Accuracy Comparison of BPNN and JNNCM

3.2 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 JNNCM with various techniques for the classification of interacting and non-interacting protein pairs prove that Jordan neural network classification model (JNNCM) is a better method for classification among interacting and non-interacting protein pairs.

4. 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.

4.1 Future Scope

Jordan neural network classification model 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].

5. REFERENCES

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

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

[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] 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.

[5] 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.

[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] 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.

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

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

[17] http://en.wikipedia.org/ wiki /Protein%E2%80%93protein_interaction_prediction

[18] http://en.wikipedia.org/wiki/Bioinformatics

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

[20] http://en.wikipedia.org/wiki/Protein%E2%80%93protein interaction

[21] http://en.wikipedia.org/wiki/Pseudo_amino_acid_compo -sition

[22] http://www.heatonresearch.com/wiki/Jordan_Neural_Net -work