A Pattern Recognition Based Approach for Prediction of Protein Drug Interactions Using Neural Networks

Attia Anis[1], Muhammad Abuzar Fahiem[1], Muhammad Hassan Rasheed[2]

[1.] Department of Computer Science, Lahore College for Women University, Lahore,54000 Pakistan.
[2.] Department of Computer Science, FAST National University of Computer and Emerging Science, Lahore, 54000 Pakistan.
attiasultana22@gmail.com

Abstract: The prediction of protein-drug interaction is a key area in drug discovery and it is considered to be a complex task. Drug discovery is a crucial task therefore there is a deep motivation towards developing new methods for identifying protein-drug interaction efficiently. The study aimed is to predict that how strong binding orientation exists between protein-drug interactions, and these interaction results further used for drug discovery. The research was carried out from various datasets of protein-drug interaction gathered from different databases and all these datasets are considered as an unstructured data. This research is focuses on drug-target interaction networks in human beings involving protein families such as, Enzymes, Ion channels, G-protein-coupled receptors (GPCRs) Nuclear Receptors, Alpha and Beta. The comprehensive pattern recognition technique is used for predicting new protein-drug interactions. Implementation is done using neural network pattern recognition tool. Neural Network is considered to be a most efficient method for the datasets related to bioinformatics and neuron sciences. The satisfactory result of our research is 93%. The success of our approach is evident from the result. The study recommended that the proposed technique should be implemented in health department for efficient drug discovery results.

[Attia Anis, Muhammad Abuzar Fahiem, Muhammad Hassan Rasheed. **A Pattern Recognition Based Approach for Prediction of Protein Drug Interactions Using Neural Networks.** *Researcher* 2013;5(12):69-74]. (ISSN: 1553-9865). http://www.sciencepub.net/researcher. 8

Keywords: Drug Discovery; Protein-drug interaction; Pattern Recognition; Neural Network.

1. Introduction

Bioinformatics is playing a vital role in the field of computer science and biology. For development of drug the extensive work has done on protein-Drug interaction. The bioinformatics field has evolved because of the most demanding task involves the analysis and clarification of different category of data related to nucleotide, protein domains and their structures and their interactions, amino acid and biological sciences. Computational biology is referred to as analyzing, interpreting and computing data related to bioinformatics significant sub-disciplines within biology science and computational science is to introduce, develop and implement new tools that enable efficient management, access and use of various type of material. Biological reaction consists on interaction of different molecules, receptors, and enzymes. A computational based prediction predicts that how strong binding orientation exists between them and whether this drug will be favorable for human body or not. Drug interaction within human body is one of the main reasons for the antagonistic side effects. Drug Discovery is a major and necessary task for pharmacists and in the field bioinformatics. Biological Sequences consists on DNA, RNA and twenty amino acids. The sequence of amino acids in protein is defined by a sequence of gene, sequence of gene is encoded in the genetic code and the genetic code specifies 20 standard amino acids [A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,P]. The protein is long chain assembled from twenty different of amino acids. Proteins are not rigid molecules they are divided in three different classes. Globular, fibrous and membrane are protein classes. Protein is the chief actor of cell. It binds the other protein and arranges their diverse set of functions. Protein interacts with protein for performing specific function. Protein interacts with smaller molecule or ligands for functional process. Ligands can be a drug or a medicine. Protein also has binding orientation with drug in the case of disease, decode in human genome and help in drug discovery. Most important drug transport protein in human plasma is Human serum Albumin (HAS). Mostly drugs in body interacts with HAS. Protein-drug interactions is a chemical substance, exists in bound and unbound states, such that,

Protein + Drug -> Protein-Drug Complex.

Pattern recognition has achieved using linear and quadratic discriminations K-nearest neighbor algorithm, neural network, template matching. Existing approaches are Kernel based(Yong-Cui – Wang et al., 2011).Statistical and systematic method based (Kevin Bleakley and Yoshihiro Yamanishi, 2009). (Wang et al., 2010).(Yoshihiro Yamanishi et

al., 2008), (Jacob L and vert J.-P, 2008). Network based (Xing Chen and Gui-Ying Yan, 2010). (Q.L. Li, and L.H. Lai, 2007). (Yildirim MA et al., 2007). (Leonid Chindelevitch et al., 2010). Biological and chemical features and information based (Zhang J et al., 2010).(Jean-Loup Faulonet al., 2008).Based on webbased tool(Kang L et al., 2006). (Zhu et al., 2011). Screening based (Yao L, Rzhetsky A 2008). Based on docking and scoring methods (Andrew R et al., 2006). (Dariusz Plewczynskiet al., 2010). (Thomas Lengauer and Ralf Zimmer 2000). Based on 3D structure of Protein (Alasdair T.R. Laurie and Richard M. Jackson 2006) based on Do Novel approach(YanayOfran 2005). Our approach is targeted the prediction of Protein-Drug Interaction based on Pattern Recognition approach. For this purpose we used classification of protein etc,Enzymes, Ion Channels, Nuclear Receptor, GPCRs, Alpha and Beta to accomplish this task.

2. Materials and Methods.

We have taken Protein-Drug interaction datasets from different sources like KEGG (21), DRUGBANK (22), PUBMED (23), PUBCHEM (24), and ZINC (25). Whenever we will get Drug-target interaction information it will be considered as an unstructured data.Unstructured data is considered to be as a poor data. It is very tricky to extract "training datasets" from unstructured data. Necessary information in tabulated form is shown in Table 1.Protein- drug interaction data include that how many drugs interact with protein, protein families. In this research weare considering families and two alpha and beta subunits.An algorithm is designed for extracting data according to need and unstructured to structure conversion (normalization) is performed on these tables for further used. Prediction is based on these structured datasets. Algorithm is shown in Figure. 1. Data According to Specified Table Entries has depicted in Table 2.

```
Id, name, P1, E, G, R, I, A, B,
C, S,T, Loc
    for (1 to m)
for (1 to 12)
c [ i, j ] = 0
d [m, j] = id
d[ m, j+1] = name
fori (3 to 12) → col
    for  j (1 to m ) → row
    do  if  "Protein interact with
    drug "
P1 = = id
p (j, i ) = P1
Switch (P1, S, T){
```

→

```
Case 1: P1 = = E
v ( j,i +1 ) = "Yes"
if "enzymes has subcellular location"
v(j, i+8)=Loc
else break;
Case 2: P1 = = G
v ( j,i +2 ) = "Yes"
if "enzymes has subcellular location"
v(j, i+8)=Loc
else break
Case 3: P1 = = R
v ( j,i +3 ) = "Yes"
if "enzy
mes has subcellular location"
v(j, i+8)=Loc
```

→

```
Case 4: P1 = = Ion
v ( j,i +4 ) = "Yes"
if "enzymes has subcellular location"
v(j, i+8)=Loc
else break;
Case 5: P1 = = A
v ( j,i +5 ) = "Yes"
if "enzymes has subcellular location"
v(j, i+8)=Loc
if A=T
v(j, i+9)= T;  else break;
Case 6: P1 = = B
v ( j,i +6 ) = "Yes"
if "enzymes has subcellular location"
v(j, i+8)=Loc;  else break;}
End Switch
```

Figure. 1. Design Algorithm for inserting data in table

Table. 1. Specified Table Entries

Families

Drug Name	Drug Code	Protein Unique ID	Enzymes	GPCRs	Receptors	Ion Channels	Alpha	Beta	Protein sequence	Sub Cellular localization	Type of protein

Table 2: Data According to Specified Table Entries

Families

Drug Name	Drug Code	Protein Unique ID	Enzymes	GPCRs	Receptors	Ion Channels	Alpha	Beta	Protein sequence	Subcellular localization	Type of protein
Cocaine	DB0097	P038A5	Yes		Yes		Yes		FASTA	Brain Catecholaminergic neurons	5
		P03868	Yes					Yes			
		P035A7		Yes			Yes				10

Table 3: Unstructured to Structured Conversion

Drug Code	Drug Name	Protein Sequence
DB0097	Cocain	**FASTA**

Table 4: Unstructured to Structured Conversion

Classes

Drug Code	Enzymes	GPRCs	Ion channels	Receptor	Alpha	Beta
DB0097		Yes	Yes	Yes		Yes

Table 5: Unstructured to Structured Conversion

Drug Code	Protein id	Types of protein	Sub Cellular location
DB0097	P03868		Cytoskeleton
DB0097	P03A67	7	

2.1. Unstructured to Structure Conversion.

There are two ways for normalizing the proposed table i.e.

- One drug can interact with multiple proteins.
- One protein can interact with multiple drugs.

But after overall literature study it is concluded that multiple protein can interact with multiple drugs, so there is many to many relation exist between protein-drug interactions. Further the tables have normalized up to three normal forms. Structured form of data is shown in Table 3 to Table 5.

2.2. Prediction.

After getting the structured data the pattern recognition approach give the impression to demonstrate the several features that could help to predict drug-target interaction. Pattern recognition based prediction is usually categorized according to the sort of learning procedure is used to generate output. If it is assumed that "Supervised" learning method is used then there is need a "training set". There is "pattern", "features" and "datasets" are required that can be expressed as an input. Pattern Recognition algorithm based on neural work proved that it can handled uncertainties in data.

2.3. Formulation of Protein Drug Interaction Data as Pattern Recognition.

So after getting datasets, these datasets are tabulated according to our prediction need and unstructured to structured conversion is performed for further use. The main elements and related data has extracted from structured data i.e

2.3.1. Pattern Feature.

Drug-id Protein-id Class Sub-Cellular Localization
DB00045 O60603 Receptor Cell wall

2.4. Pattern Recognition Algorithm.

Different statistical, structural, probabilistic and neural network has used for pattern recognition in the field of bio-informatics. Here the neural network has applied for pattern recognition the drug-target interaction. Multilayer Perception is forward neural network that maps sets of input for predicting accurate output. A neural network consists of three or more than three layers these are input, hidden, and output. Neural Network can be used to find patterns in data in complex relationships between inputs and outputs.

Data Sets consist on pattern included drug targeted information. Data sets are divided in three kinds of samples training, validation and testing. In the case of training data the data sets are used to adjust neural network according to its error.

Receiver Operating Characteristics (ROC) (Figure 2) will be plot under the Neural Network in the Pattern recognition tool. ROC curves are used to graphical representation of performance. For plotting the fraction of true positive out of positives and false positives out of negatives these ROC curves are used. A perfect test would show the points in the upper left corner. The ROC curve false positive and false negative rate can find the true result. Abnormal

values in Roc curve are larger > 4 and normal values are smaller < 4.It may always not true in every case.

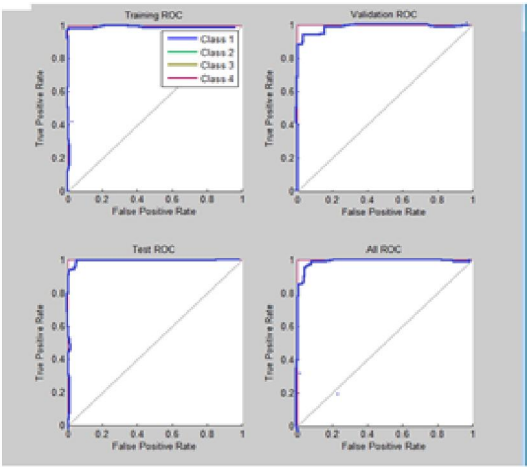

Figure.2. ROC (Receiver Operating Curve)

2.5. Discussion and Results.

We are using datasets Enzymes, Ion Channels, GPCRs, Receptors, Alpha and Beta. Implementation is done using neural network pattern recognition tool. Neural network is considered to be a most efficient datasets related to bioinformatics and neuron sciences. Neural network aim is to build a biological neural system for developing the understanding that how biological system works and some new predictions are possible. Data Sets are consisting on pattern included drug targeted information. Through a very comprehensive technique required data is obtained from different resources. Data sets are divided in three kinds of samples training, validation and testing. In the case of training data the data sets are used to adjust neural network according to its error. They are used to measure network generalization and when its stop to improving the training is halt. They have no effect on training and an independent network performance measurement during and after training are achieved. Validation and testing performance is shown in Figure 3.Validation and test performance can be plot Blue lines represents train data. A green line represents validation data. A red line represents testing of datasets. These lines are overlapping and have shown best interrelation between all elements. Overall accuracy of our system is 93%.The success of our approach is evident from the result is shown in Figure 4. Our approach is quite modest comparatively than others. Comparative study of proposed approach with existing approaches is shown in Table 6.

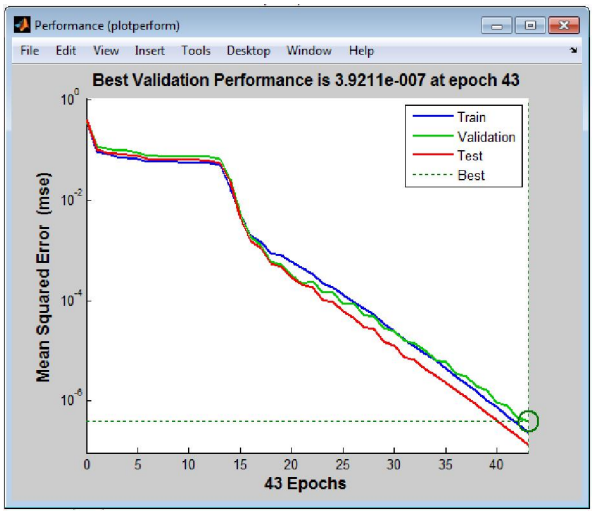

Figure.3. Validation and Testing Performance

Target Matrix

Output Matrix					
	231 **85.7%**	**11** **1.6%**	**45** **4.2%**	**33** **3.3%**	**93** **4.1%**
	14 4.1%	**13** **1.1%**	**23** **4.1%**	**41** **5.1%**	**65** **6.1%**
	11 5.2%	**21** **3.4%**	**62** **1.4%**	**21** **1.5%**	**67** **3.4%**
	13 2.1%	**41** **2.1%**	**34** **3.1%**	**145** **4%**	**91.2** **6.1%**
	92.3 3.1%	**93.4** **6.1%**	**79** **5.1%**	**87** **2.1%**	**93%** **5.1%**

Figure 4: Results

Table 6: Comparison of Proposed Approach with Existing Approaches

Author	Existing Approaches	Classification/Structure of protein	% Accuracy
[1]	Kernel Function (SVM)	Enzymes, Ion channels,GPCRs,Receptor	91
[2]	Statistical Method	Enzymes, Ion channels,GPCRs,Receptor	89
[5]	NRWRH method	Chemical Structure Enzymes, Ion Channels,GPCRs,Receptor	82
[6]	Drug-Cocktail Network	Drug Combinations with all close protein	85
[7]	Nearest Neighbor Algorithm	Enzymes, Ion channels,GPCRs,Receptor	86
[11]	SVM (Kernel Function)	Enzymes, Ion channels,GPCRs	87
[13]	Graphical Network (Statistical Method)	Ion Channels,GPCRs,Receptor	85
[18]	Docking And Scoring	3D Structure	81
[19]	SBDD and Virtual Screening	3D Structure	79
[25]	Do Novel Approach	3D Structure	87
Proposed Approach	Pattern Recognition based prediction	Enzymes, Ion channels,GPCRs,Receptor,Alpha and Beta	93

References

1. Yong-Cui –Wang, Chun-Hua Zhang, Nai-yang Deng, Yong –Wang, "kernel based data fusion improves the Drug Protein interaction prediction", Northwest institute of plateau biology, Information school Renmin university of china, college of science chine agriculture university China. 2011;35(8):352-362.
2. Kevin Bleakley and Yoshihiro Yamanishi "Supervised prediction of drug-Target interactions using bipartite locals methods" 2009; 25(8): 2397-2403.
3. Wang, Yong-Cui; Yang, Zhi-Xia; Wang, Yong; Deng, Nai-Yang "Computationally Probing Drug-Protein Interactions via Support Vector Machine", 2010; 7(5): 370-378.
4. Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge1, Wataru Honda and Minoru Kanehisa "Prediction of drug–target interaction networks from the integration of chemical and genomic spaces" 2008; 24(13): i232-i240.
5. Xing Chen, Gui-Ying Yan, "NRWRH for Drug Target Prediction", The Fourth International Conference on Computational Systems Biology Suzhou, China, September 9–11, 2010: 219–226.
6. Q.L. Li, and L.H. Lai, "Prediction of potential drug targets based on Simple sequence properties", 2007; 8: 353-364.
7. He Z, Zhang J, Shi X-H, Hu L-L, Kong X, et al. "Predicting Drug-Target Interaction Networks Based on Functional Groups and Biological Features", PLoS ONE 2010;5(3):e9603.doi:10.1371/journal.pone.0009603.
8. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, Luo X, Zhu W, Chen K, et al. TarFisDock: "A web server for identifying drug targets with docking approach", Nucleic Acids Res. 2006; 34(Web Server issue):W219–W224.
9. Jean-Loup Faulon, MilindMisra, Shawn Martin, Ken Sale and RajatSapra "Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor" Bioinformatics, 2008; 24(2): 225-233.
10. Yoshihiro Yamanishi, Masaaki Kotera, Minoru Kanehisa, and Susumu Goto "Drug-target interaction prediction from chemical, genomic and Pharmacological data in an integrated framework" Bioinformatics,2010; 24(12): i246-i245.
11. Jacob L, Vert J.-P, "Protein-ligand interaction prediction: an improved chemo genomics approach", 2008; 24:2149-2156.
12. Yildirim MA, Goh KI, Cusick ME, Barabsi AL, Vidal M: "Drug-target Network",Nat Biotechnology 2007; 25(10):1119-1126.
13. Yao L, Rzhetsky A: "Quantitative systems-level determinants of human genes targeted by successful drugs", Genome Res 2008;18(2):206-13.
14. Leonid Chindelevitch Chung-Shou Liao Bonnie Berger "Department of Mathematics and Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139",Pacific Symposium on Biocomputing 2010;15:123-132.
15. Andrew R. Leach, Brian K. Shoichet, and Catherine E. Peishoff "Prediction of Protein-Ligand Interactions. Docking and Scoring:

Successes and Gaps", Journal of medicinal chemistry, copyright 2006 by American chemical society.49 (20): 2006.
16. Alasdair T.R. Laurie and Richard M. Jackson "Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening" Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, UK. Current Protein and Peptide Science, 2006, 7, 395-406.
17. Alasdair T.R. Laurie and Richard M. Jackson "Methods for the Prediction of Protein-Ligand Binding Sites for Structure-Based Drug Design and Virtual Ligand Screening" Institute of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, LS2 9JT, UK. Current Protein and Peptide Science, 2006; 7: 395-406.
18. Zhu, Yuyin Sun, SashikiranChalla, Ying Ding, Michael S Lajiness and David J Wild. "Semantic inference using chemo genomics data for drug discovery" BMC Bioinformatics 2011.
19. Thomas Lengauer and Ralf Zimmer "Protein structure prediction methods for drug design". Bioinformatics 2000; 1(3):275-288.
20. Yanay Ofran, Marco Punta, Reinhard Schneider and BurkhardRost "Beyond annotation transfer by homology: novel protein –function prediction methods to assist drug discovery" DDT. 2000; 10(21).
21. http://www.genome.jp/kegg/
22. http://www.drugbank.ca/
23. http://pubchem.ncbi.nlm.nih.gov/
24. .http://www.ncbi.nlm.nih.gov/pubmed
25. http://zinc.docking.org/

1/12/2013