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

Attia Anis¹, Muhammad Abuzar Fahiem¹, Muhammad Hassan Rasheed²

¹Department of Computer Science, Lahore College for Women University, Lahore, 54000 Pakistan.
²Department 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.

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 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


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 we are 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.

```
for (1 to m)     
for (1 to 12)   
c[i, j] = 0     
d[m, j] = id   
d[m, j+1] = name
for (3 to 12) \rightarrow col
   for (j (1 to m) \rightarrow 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 "enzymes has subcellular location"
               \(v(j, i + 8)\) = Loc
            else break;
            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
            else break;
            Case 6: P1 = = B
               v(j, i + 6) = "Yes"
               if "enzymes has subcellular location"
               \(v(j, i + 8)\) = Loc
               if A=T
               v(j, i+9) = T; else break;
            Case 7: P1 = = C
               v(j, i + 7) = "Yes"
               if "enzymes has subcellular location"
               \(v(j, i + 8)\) = Loc
            else break;
            Case 8: P1 = = T
               v(j, i + 8) = "Yes"
               if "enzymes has subcellular location"
               \(v(j, i + 8)\) = Loc
            else break;
            Case 9: P1 = = S
               v(j, i + 9) = "Yes"
               if "enzymes has subcellular location"
               \(v(j, i + 8)\) = Loc
            else break;
            Case 10: P1 = = Loc
               v(j, i + 10) = "Yes"
               if "enzymes has subcellular location"
               \(v(j, i + 8)\) = Loc
            else break;
            Case 11: P1 = = T
               v(j, i + 11) = "Yes"
               if "enzymes has subcellular location"
               \(v(j, i + 8)\) = Loc
            else break;
            Case 12: P1 = = T
               v(j, i + 12) = "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

<table>
<thead>
<tr>
<th>Families</th>
<th>Drug Name</th>
<th>Drug Code</th>
<th>Protein Unique ID</th>
<th>Enzymes</th>
<th>GPCRs</th>
<th>Receptors</th>
<th>Ion Channels</th>
<th>Alpha</th>
<th>Beta</th>
<th>Protein sequence</th>
<th>Sub Cellular localization</th>
<th>Type of Protein</th>
</tr>
</thead>
</table>
Table 2: Data According to Specified Table Entries

<table>
<thead>
<tr>
<th>Families</th>
<th>Drug Name</th>
<th>Drug Code</th>
<th>Protein Unique ID</th>
<th>Enzymes</th>
<th>GPCRs</th>
<th>Receptors</th>
<th>Ion Channels</th>
<th>Alpha</th>
<th>Beta</th>
<th>Protein sequence</th>
<th>Subcellular localization</th>
<th>Type of protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cocaine</td>
<td>DB0097</td>
<td>P038A5</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>FASTA</td>
<td>Brain</td>
<td>Catecholaminergic neurons</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P03868</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>P035A7</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3: Unstructured to Structured Conversion

<table>
<thead>
<tr>
<th>Drug Code</th>
<th>Drug Name</th>
<th>Protein Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB0097</td>
<td>Cocain</td>
<td>FASTA</td>
</tr>
</tbody>
</table>

Table 4: Unstructured to Structured Conversion

Classes

<table>
<thead>
<tr>
<th>Drug Code</th>
<th>Enzymes</th>
<th>GPCRs</th>
<th>Ion channels</th>
<th>Receptors</th>
<th>Alpha</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB0097</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Unstructured to Structured Conversion

<table>
<thead>
<tr>
<th>Drug Code</th>
<th>Protein id</th>
<th>Types of protein</th>
<th>Sub Cellular location</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB0097</td>
<td>P03868</td>
<td></td>
<td>Cytoskeleton</td>
</tr>
<tr>
<td>DB0097</td>
<td>P03A67</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Target Matrix</th>
<th>Output Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>231 85.7%</td>
<td>11 1.6%</td>
</tr>
<tr>
<td>14 4.1%</td>
<td>13 1.1%</td>
</tr>
<tr>
<td>11 5.2%</td>
<td>21 3.4%</td>
</tr>
<tr>
<td>13 2.1%</td>
<td>41 2.1%</td>
</tr>
<tr>
<td>92.3 3.1%</td>
<td>93.4 6.1%</td>
</tr>
</tbody>
</table>

Figure 4: Results
Table 6: Comparison of Proposed Approach with Existing Approaches

<table>
<thead>
<tr>
<th>Author</th>
<th>Existing Approaches</th>
<th>Classification/Structure of protein</th>
<th>% Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Kernel Function (SVM)</td>
<td>Enzymes, Ion channels, GPCRs, Receptor</td>
<td>91</td>
</tr>
<tr>
<td>[2]</td>
<td>Statistical Method</td>
<td>Enzymes, Ion channels, GPCRs, Receptor</td>
<td>89</td>
</tr>
<tr>
<td>[5]</td>
<td>NRWRH method</td>
<td>Chemical Structure</td>
<td>82</td>
</tr>
<tr>
<td>[6]</td>
<td>Drug-Cocktail Network</td>
<td>Drug Combinations with all close protein</td>
<td>85</td>
</tr>
<tr>
<td>[7]</td>
<td>Nearest Neighbor Algorithm</td>
<td>Enzymes, Ion channels, GPCRs, Receptor</td>
<td>86</td>
</tr>
<tr>
<td>[8]</td>
<td>SVM (Kernel Function)</td>
<td>Enzymes, Ion channels, GPCRs</td>
<td>87</td>
</tr>
<tr>
<td>[18]</td>
<td>Docking And Scoring</td>
<td>3D Structure</td>
<td>81</td>
</tr>
<tr>
<td>[19]</td>
<td>SBDD and Virtual Screening</td>
<td>3D Structure</td>
<td>79</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td>Pattern Recognition based prediction</td>
<td>Enzymes, Ion channels, GPCRs, Receptor, Alpha and Beta</td>
<td>93</td>
</tr>
</tbody>
</table>

References


15. Andrew R. Leach, Brian K. Shoichet, and Catherine E. Peishoff “Prediction of Protein-Ligand Interactions. Docking and Scoring:


1/12/2013