International Journal of Engineering & Technology, 7 (1.9) (2018) 145-148

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

A hybrid approach for hot spot prediction and deep representation of hematological protein – drug interactions

Bipin Nair B. J [1] *, Lijo Joy [1]

[1] *Department of Computer Science Amrita School of Arts and Science, Mysore Campus Amrita Vishwa Vidyapeetham,India*
**Corresponding author E-mail: bipin.bj.nair@gmail.com*

Abstract

In our research work we will collect the data of drugs as well as protein regarding hematic diseases, then applying feature extraction as well as classification, predict hot spot and non-hot spot then we are predicting the hot region using prediction algorithm. Parallelly from the hematological drug we are extracting the feature using molecular finger print then classifying using a classifier and applying deep learning concept to reduce the dimensionality then finally using machine learning algorithm predicting which drug will interact with the help of a hybrid approach.

Keywords: *Deep Learning; Drug Prediction; Hot Spots; Hot Region.*

1. Introduction

Bioinformatics is a field of merging, creating technical information and programming equipment for biological information. It is the scientific concept of scientific terminology, in the physical and chemical sense, after the application of the technology obtained from the command, such as connected mathematics, software engineering, and measurement to understand and form the data associated with these molecules in large scale.
Cheminformatics is a new field of data innovation, focusing, storing, testing and controlling material information. The conspicuous mixed information usually contains data on small atomic recipes, structures, properties, spectra, and exercises. Chemical informatics has initially become a means of helping the drug's revelation and progress, but chemical informatics now has an indisputable key part in many areas of science, science and natural chemistry.
In this paper, we are discussing the combination of the area under Bioinformatics and Cheminformatics to predict the promising drug and hot region for blood coagulation protein. We are introducing the machine learning technique to predict the drug and hot region for hematological protein. In our work deep learning is used to reduce the dimensionality of drugs. Most biological processes include many proteins that communicate with each other. It has been found late that the specific accumulation of protein-protein communication, which is called the problem domain, is more helpful to other aspects of limiting preferences. Problem area deposits have a different and lively attribute that makes them try, but the important focus is to adjust the protein-protein. The remedial operator program in collaboration with the problem area deposit eventually became an important procedure for disrupting the poor protein-protein link. The use of organic technology to determine which accumulation is the problem area can be expensive and cumbersome. The further development of the machine learning approach to deal with the expected problem areas incorporates a number of hot spots and demonstrates a predictable victory.

Recent technologies are feature-based strategies, energy-based strategies and similarity-based approaches. Similarity methods are used to predict the interaction of their compound with sequence similarity of protein. Vector-based approaches are considered as a more advanced strategy to address drug and protein topographies more directly. Energy-based methods use a simplified knowledge-based model to calculate the bound free energy to predict hotspot and non-hotspots in PPIs. These lack of proficiency, high expectations of machine learning technology is increasingly used for protein-protein association. Although machine learning techniques have strengthened the expected implementation, there are still quite a few issues that need clarification.
In this paper, SVM and AdaBoost classifier is used to accumulate the prediction model of the problem area in the protein-protein interface, and the drug target interaction is expected. These have been shown to be a better performance predictor of classification problems, and the clustering algorithm used to predict the hot region

2. Literature survey

Researches are reported in the area of machine learning in bioinformatics for the prediction of protein hotspots and drug, some of the related the work are summarized here.
Buza [1] was proposed a Machine learning technique are used to predict drug target interaction using Bipartite Local Model (BLM), which gives the promising prediction of drug-target interaction. The local model as Hubness-aware & ECkNN. Xiaoli et al [2] was suggested the Hot-spots is binding free energy which contribute portion of interface residues. And hot-spot molecules are known as hot regions. support vector machine based on ensemble learning system used to predict the hotspot and hot regions in protein-protein interactions. Sathien [3] proposed Cytochrome P450 enzymes (CYP450) is commonly used alterations in drug metabolism at the time of Drug-Drug interaction. The Enzyme

action intersection method provided a new technology that can be used to predict drug-drug interactions. Tian [4] was proposed DL-CPI method with the combination of neural network and features of chemical & protein is used to predict new CPI, also it improve the performance of CPIs. The SVM is used to predict ligand-protein interactions with a reasonable accuracy. Cheng [5] proposed that the Drug-Drug interaction predicted by HNAI framework. The calculation is done by drug genomic, chemical, phenotypic and therapeutic similarities, followed by construction of a DDI network. Huang et al [6] was proposed an ELM, is a powerful learning calculation joined with Chous pseudo amino corrosive creation (PseAAC) is a standout amongst the most generally utilized element extractors for proteins. Which is utilized to take care of entangled issue with more than 20 discrete elements without totally losing the sequence-order. You [7] was proposed that the sequence-based method, which combines the ELM with the new representation, can accurately predict the PPI using self-covariance. The AC descriptors used for the interaction between residues of a protein sequence and the ELM at a distance is an accurate and rapid method of learning. Laarhoven [8] was recommended that WNN-GIP be administered in combination with a new compound or target. Computation is done through a regularized least-squares algorithm that combines core products. Wang [9] Proposed drug relocation is a strategy used to predict the interaction of a single drug with multiple targets. The new DTIs are identified by the RBM and the Contrast Divergence (CD), and their corresponding type of interaction is inferred. Mei [10] introduced a improved version of BLM that is, BLM-NII for incorporating the ability to learn neighbors into the original BLM approach And it is used to classification and prediction of new drug target interaction. Gönen [11] A method for predicting the interaction of drug targets is proposed using a novel Bayesian formula that combines dimensionality reduction, matrix factorization and binary classification, respectively. The interaction network uses only the chemical similarity between the drug compounds and the genomic similarity between the target proteins. Van [12] was proposed that the interaction profile be made by using a machine learning method with a binary vector to describe the presence or absence of an interaction with each goal. Interactions can be effectively used for accurate prediction of drug target interactions. In order to test the predictive performance, a simple regularized least-squares algorithm was used, combining the product of a kernel function. Danger [13] suggested a Semi-automated de-curation gives the proper arrangement of drug target interaction article. The de-curation is done by text mining tools and various machine learning technique. Namboori [14] the possibility of proposing a drug is a qualitative analysis to check whether a given molecule is a drug. Mathematical models have been developed using machine learning algorithms to predict whether a given molecule is a proteomics drug. PreADMET and SVM were used to determine the drug similarity of proteomic targets. Bleakley [15] the proposed chemical and genomic data are used as a new monitoring method to predict the interaction of unknown drug targets. Predict compound-protein interactions using bivariate classification by participating biochemical and genomic space into one uniform space, ie, pharmacological space and a nuclear-based approach. Sugaya [16] introduced the supervised learning methods are applied to predict the ability of drug in single drug target proteins using machine learning technique. SVM predicts the drug ability of PPIs. To calculating the interaction by combine the machine learning technique and SVM. Darnell [17] was proposed the combination of knowledge based and learning technique are used to predict the PPI hotspots KFADE and KCON is the two methods and the combination KFC provide the accurate prediction. Chan [18] was suggested from the deep learning to predict the large-scale DTIs. feature extraction and SVM is used to large scale prediction by reducing the dimensionality. Likhitha [19] A new method combining Levenshtein distance algorithm and STR analysis is proposed, in which the DNA barcodes are used to identify the smallest number of mismatches between sequences. And it is used to find and group the standard genome of the species from DNA nucleotide sequences. Bipin [20] was proposed an approach to identify a specific gene by providing relevant protein sequences and identifying exact locations of splice junctions in DNA sequences. They focused the factor for highly essential for the clotting of blood i.e. Factor IX. Longest Common Continuous Subsequence (LCCS) method is used to get the promising results with an accuracy of 96.8% in detecting proper splice junctions for the given gene.

Multi-scale enhancement hypothesis is used to remove drugs and proteins. They propose a reduction in the depth of learning dimension. In addition, stacked automatic encoder can create information layer by layer. More importantly, it re-creates proxy highlights from the hidden layers of the stack and assembles the SVM as the last classifier. Together with AdaBoost (known as BoostSVM), SVM predicts problems with storage and off-site accumulation of problems in cooperation. At this point, LCSDA is additionally received to distinguish hot areas with high accuracy.

3. Problem formulation

Recently similarity-based method and vector-based technique are used to find the drug interaction. The similarity-based approach used to predict the interaction of their compound and sequence similarity of the protein. Vector-based techniques are viewed as further developed procedures that contain drug and protein feature clearly.

The energy-based approach uses a knowledge-based simplified model to calculate the free energy to predict hotspot and non-spot residues in PPIs. These are inefficient, time-consuming and costly. So we proposed a various machine learning technique to find promising drug and hot spots.

4. Problem definition

In this paper, we introduce a hybrid approach throw the various machine learning technique to predict hot spots as well as drugs for blood coagulation protein.

a) Hot spots prediction

i) Feature Extraction

In the studying of PPIs, feature selection is considering the attribute selection. In this paper, we focusing on the blood clotting and collect the protein according to it from protein data bank and generate 3d structure of each clotting protein.

ii) Classification

AdaBoost classifier and SVM have a common point, that gives a promising classification over protein-protein interaction. In this paper, the combination of advantages of the algorithm gives the appropriate prediction of hot spots and non-hotspot residues in the hematological protein. Clustering algorithm used to predict the hot region.

b) Drug prediction

i) Classification

The classification and feature extraction are focused on quality of the data. We examine the large-scale collection of drugs under blood diseases. Classify the dataset and extract the drugs according to bleeding disorders to predict the promising drug for blood coagulation.

ii) Dimensionality reduction

After extracting the hematological drugs, we apply the deep learning technique to reduce the dimensionality of drugs. It is used to reduce the high dimensional features to low dimension. Also used to reduce the large-scale data into the small-scale that is reduce the number of the data in to a small unit to predict the promising drug for the bleeding disorders.

iii) Fingerprinting

The molecular fingerprints used for encoding the structure of a molecule. The system for encoding the structure of drugs for the computational process. In this paper, we provide the fingerprinting to check the similarity of the hematological drug for the accurate prediction.

5. Proposed architecture

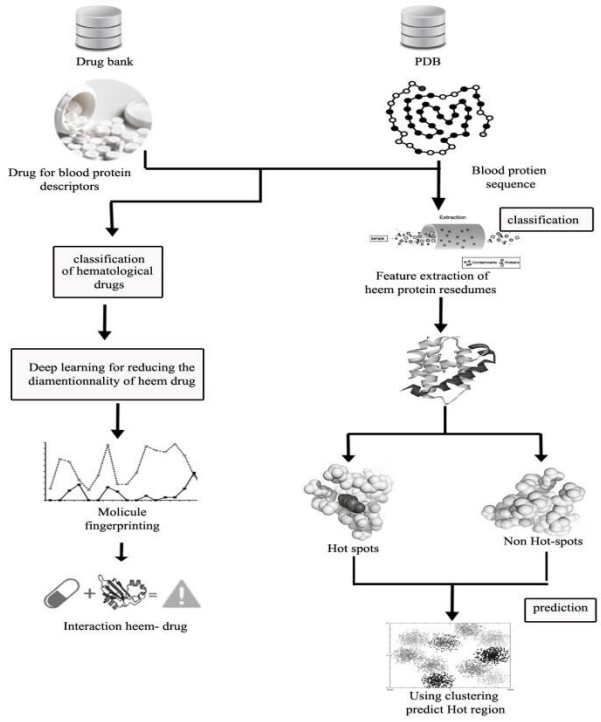

Fig. 1: Proposed Flow.

6. Algorithm

In this paper, we uses a combination of classification algorithms such as SVM and AdaBoost to classify the hot spots and non-hot spots residues.
Algorithm I
Step1: Training set S = {(s,yl), ... , (sn,yn)}'
Step2: Begin
Step3: Initialize weight
step4: for (t = 1... T) samples weight
step5: SVM train the data
step6: hx: Rd {0, 2}
step7: hx applied to the samples
step8: Update weight Dv+1(j) =Dt (j) exp ((Xj))/Zt
step9: End for
step10: Combine

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Step11: end
From the above classification, we identify the hot region using k-mean clustering algorithm
Algorithm II
Step1: begin
Step2: Specify the desired number of clusters K : k=3
Step3: k represents the centroid of the cluster
Step4: calculate the distance using Euclidean

$$d_{x,y} = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$$

Step5: select the least distance
Step6: calculate mean
Step7: pointe are assigned to the cluster
Step8: end

The molecular fingerprinting is used to encode the molecules to computer system and it provides a similarity between the drugs

Algorithm III
Input: SMILES of two drug
Output: similarity in percentage
Begin
For each smiles of two drugs
M1=smile of a drug
M2= smile of another drug
Fp1=fingerprint (m1)
Fp2=fingerprint (m2)
Similarity =fing_similarity (fp1, fp2)
Print similarity
End for
End

7. Dataset

We are collecting the protein from the PDB and drugs from Drug bank. The protein structure data come from the protein database and we filter the appropriate dataset to classify the hot spots and non-hotspots residues.
We are focusing on the fibrinogen from the dataset and calculate a dataset with residue no, weight etc. to classify the hot spots and non-hotspots residues.

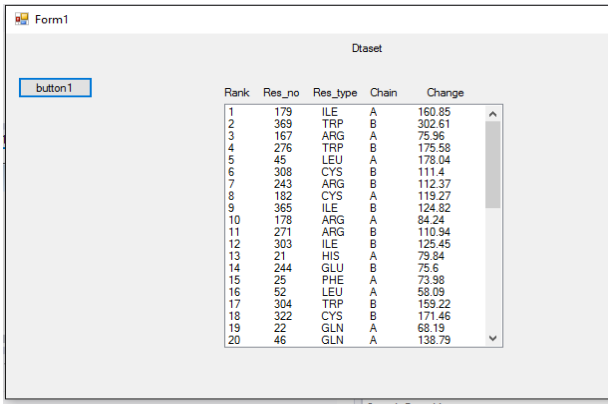

Fig. 2: Data Set for Hotspot Classification.

We are collecting the large-scale data of drug from drug bank and the fingerprinting are used to predict the similarity between drugs. Apply the fingerprinting and the deep learning to reduce the dataset and accurate prediction of blood coagulation drug. Drug smiles dataset is also used for the fingerprinting.

8. Experiment result

In the experimental result, we generate the 3D structure of each fibrinogen protein with the PDB files. From the extraction of the hematological protein, we extract the blood clotting protein to classify hotspots and non-hotspots.

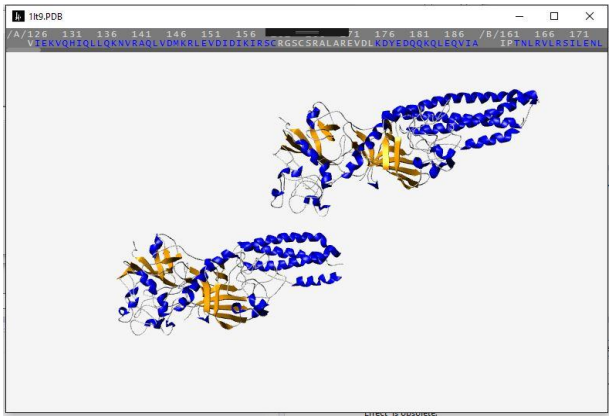

Fig. 3: 3D Structure of Protein.

The AdaBoost classifier is used to classify the hot spot and non-hotspot residues using the above example dataset. The k-mean clustering algorithm used for accurate prediction of the hot region.

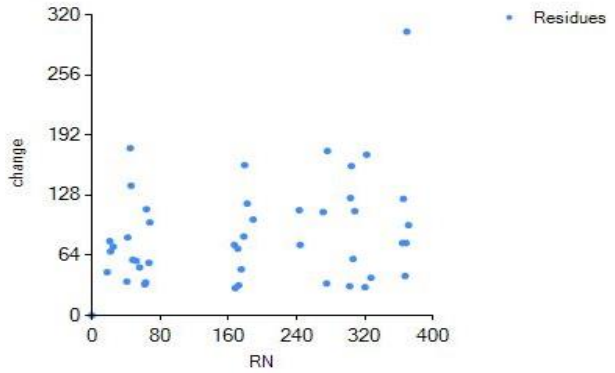

Fig. 4: Example of Hotspot Residues.

The above Fig showed the example of the residues of hotspots. The advantages of the k-mean algorithm used to predict the hot region by mean nodes.

To predict the promising drug for the coagulation protein, we use the deep learning and fingerprinting methods. Deep learning is used to reduce the number of large-scale data of the hematological protein. The Table-I shown the similarity between two drugs. Fingerprinting used to check the similarity. The result gives Heparin and Warfarin promising drug for blood coagulation with less similarity.

Table 1: Similarity between Two Drugs

Drugs	Drugs	Similarity (%)
Cyklokapron	Heparin	46.09375
Cyklokapron	Tranexamic Acid	100.0
Cyklokapron	Warfarin	46.09375
Heparin	Tranexamic Acid	46.09375
Heparin	Warfarin	33.6039975016
Tranexamic Acid	Warfarin	46.09375

9. Conclusion

A hybrid method is been proposed in this paper, to predict both the hot region and drug according to blood coagulation. To predict the hot region, classify hotspots and non-hotspots residues using AdaBoost and SVM classification. in this paper, we generate the 3d structure of the coagulation protein for classifying hotspots and non-hotspots as well as visualization. The k-mean algorithm is used to predict the hot region.

We propose the promising approach for predicting the drugs using deep learning technique. Deep learning technique is used to reduce the dimensionality of drugs. Here deep learning reduces the number of large-scale data. We propose the molecular fingerprinting to encode the chemical of the drug for our computing method and predict the similarity between two drugs. Also predict the prominant drug for blood clotting.

References

[1] Buza, K. (2016, May). Drug-target interaction prediction with hubness-aware machine learning. In *Applied Computational Intelligence and Informatics (SACI), 2016 IEEE 11th International Symposium on* (pp. 437-440). IEEE. https://doi.org/10.1109/SACI.2016.7507416.

[2] Lin, X., & Zhang, X. (2016, December). Prediction and analysis of hot region in protein-protein interactions. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on* (pp. 1598-1603). IEEE. https://doi.org/10.1109/BIBM.2016.7822758.

[3] Hunta, S., Aunsri, N., & Yooyativong, T. (2015, June). Drug-Drug Interactions prediction from enzyme action crossing through machine learning approaches. In *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015 12th International Conference on* (pp. 1-4). IEEE. https://doi.org/10.1109/ECTICon.2015.7207126.

[4] Tian, K., Shao, M., Wang, Y., Guan, J., & Zhou, S. (2016). Boosting compound-protein interaction prediction by deep learning. *Methods*, *110*, 64-72. https://doi.org/10.1016/j.ymeth.2016.06.024.

[5] Cheng, F., & Zhao, Z. (2014). Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association*, *21*(e2), e278-e286. https://doi.org/10.1136/amiajnl-2013-002512.

[6] Huang, Q. Y., You, Z. H., Li, S., & Zhu, Z. (2014, July). Using Chou's amphiphilic Pseudo-Amino Acid Composition and Extreme Learning Machine for prediction of Protein-protein interactions. In *Neural Networks (IJCNN), 2014 International Joint Conference on* (pp. 2952-2956). IEEE. https://doi.org/10.1109/IJCNN.2014.6889476.

[7] You, Z. H., Li, L., Ji, Z., Li, M., & Guo, S. (2013, April). Prediction of protein-protein interactions from amino acid sequences using extreme learning machine combined with auto covariance descriptor. In *Memetic Computing (MC), 2013 IEEE Workshop on* (pp. 80-85). IEEE. https://doi.org/10.1109/MC.2013.6608211.

[8] Van Laarhoven, T., & Marchiori, E. (2013). Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PloS one*, *8*(6), e66952. https://doi.org/10.1371/journal.pone.0066952.

[9] Wang, Y., & Zeng, J. (2013). Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics*, *29*(13), i126-i134. https://doi.org/10.1093/bioinformatics/btt234.

[10] Mei, J. P., Kwoh, C. K., Yang, P., Li, X. L., & Zheng, J. (2012). Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, *29*(2), 238-245. https://doi.org/10.1093/bioinformatics/bts670.

[11] Gönen, M. (2012). Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics*, *28*(18), 2304-2310. https://doi.org/10.1093/bioinformatics/bts360.

[12] Van Laarhoven, T., Nabuurs, S. B., & Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, *27*(21), 3036-3043. https://doi.org/10.1093/bioinformatics/btr500.

[13] Danger, R., Segura-Bedmar, I., Martínez, P., & Rosso, P. (2010). A comparison of machine learning techniques for detection of drug target articles. *Journal of biomedical informatics*, *43*(6), 902-913. https://doi.org/10.1016/j.jbi.2010.07.010.

[14] Namboori, P. K. (2009). Machine Learning Approaches to Determine the "Drug-Likeness" of the Proteomic Targets.

[15] Bleakley, K., & Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, *25*(18), 2397-2403. https://doi.org/10.1093/bioinformatics/btp433.

[16] Sugaya, N., & Ikeda, K. (2009). Assessing the druggability of protein-protein interactions by a supervised machine-learning method. *BMC bioinformatics*, *10*(1), 263. https://doi.org/10.1186/1471-2105-10-263.

[17] Darnell, S. J., Page, D., & Mitchell, J. C. (2007). An automated decision-tree approach to predicting protein interaction hot spots. *Proteins: Structure, Function, and Bioinformatics*, *68*(4), 813-823. https://doi.org/10.1002/prot.21474.

[18] Chan, K. C., & You, Z. H. (2016, July). Large-scale prediction of drug-target interactions from deep representations. In *Neural Networks (IJCNN), 2016 International Joint Conference on* (pp. 1236-1243). IEEE.

[19] Likhitha, C. P., Ninitha, P., & Kanchana, V. (2016). DNA Barcoding: A Novel Approach for Identifying an Individual Using Extended Levenshtein Distance Algorithm and STR analysis. *International Journal of Electrical and Computer Engineering*, *6*(3), 1133.

[20] Nair, B. B., Khamarudheen, K. S., & Ranjitha, H. S. (2016). An Approach for Identifying The Presence of Factor Ix Gene In Dna Sequences Using Position Vector Ann. *Journal of Theoretical And Applied Information Technology*, *87*(3), 396.