ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

LSSAT (Ligand Structure Similarity Analysis Tool): An automated program to generate large scale analysis of targets, ligands and networks.

Hitesh Kumar Jaiswal[1#], Jaisri Jagannadham[1#], Amit Kumar[1], and Kamal Rawal[1,2,3#*].

[1] Jaypee Institute of Information Technology, A-10 sector 62 Noida, Uttar Pradesh, India.
[2] DataPlex Systems, New Delhi, India.
[3] Rawal Labs, New Delhi, India.

ABSTRACT

Identification of off-targets can play important role in network based therapeutics and side effect predictions of a given drug. In this paper, we present a program which can compute ligand similarity based upon binary finger printing using tanimoto coefficient. We have extracted 6512 drugs of drug bank and computed tanimoto coefficient of 6512 x 6512 drugs to find similar drugs among these sets. Out of this dataset, we selected 23 drugs indicated/involved in clinical condition known as hypertension. Thereafter, we selected 46 protein targets involved in hypertension using literature survey. Subsequently, we docked 23 drugs against the 46 protein targets to build binding profiles based upon docking energies using docking pipeline (developed in-house). Based upon our experiments, we propose that similar ligands (i.e. drugs having higher tanimoto coefficient) tend to share similar binding profile against the set of protein targets.

Keywords: Tanimoto coefficient, Perl, Hypertension, Protein targets, Drugs, Binding affinity.

Corresponding author

INTRODUCTION

Drug discovery is an expensive and time consuming process [1]. Many a times a promising drug gets rejected by FDA or other regulatory agencies during clinical trials after many years of investments and hard work. Therefore, it is imperative for pharmaceutical industry to select those molecules for clinical research which has high efficacy and minimal side effects. To improve on these possibilities, computational techniques such as virtual screening (VS) play an important role during drug discovery and development process [2].

Drug databases such as Drug Bank and PubChem provides datasets for VS. Virtual screening process can be divided into two broad categories: ligand based and structure based VS (Figure 1) [3]. The ligand based method uses methods such as similarity searching [4], pharmacophore mapping [5] and machine learning methods [6]. Structure based systems involves docking where a drug candidate binds to target protein and score is computed to identify ligands with relatively better binding affinity [7].

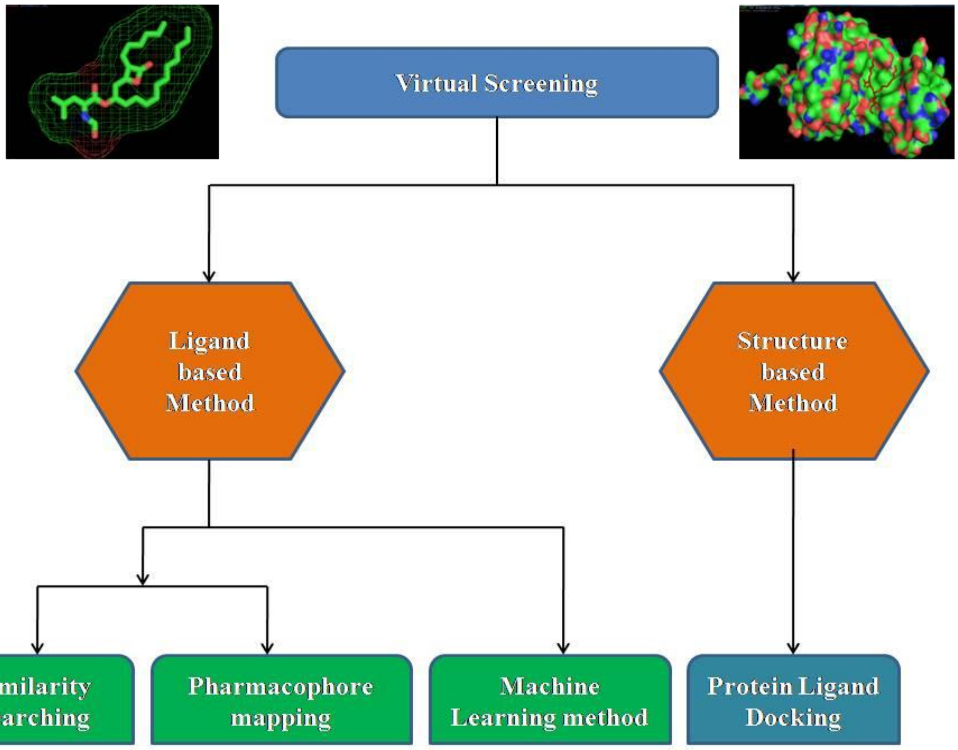

Figure 1 shows Virtual Screening Process

Similarity search approach broadly employs substructure searching. Pharmachophore based approaches takes into account steric and electronic features to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response [5]. On the other hand, machine learning approaches are based on learning to capture structure and functional properties from experimentally verified compounds. This information is used to predict the properties of new compound and the techniques is popularly known as Quantitative –structure activity relationship (QSAR) and Structure–activity relationship (SAR) [6] . In contrast with structure based methods, ligand based screening methods do not take target protein structure into account. These methods work on an assumption that ligands sharing same topological properties will have same biological activity.

Though number of methods has been developed in ligand based VS domain [2-7] still there is a plenty of scope to introduce new methods which can help during drug discovery process. In the present work, we built a system which automates Open Babel to generate drug similarity coefficient (Tanimoto coefficient: Tc) matrix for various drug profiles. This program played a crucial role in our network based therapeutics and side effects prediction system [8].

MATERIALS AND METHODS

Database

DrugBank (http://www.drugbank.ca) and PubChem (http://pubchem.ncbi.nlm.nih.gov/) are well known repository for drug related information [9]. DrugBank contains 8261 drug entries which includes 2021 FDA approved drug, 233 FDA approved protein/peptide drug, 94 nutraceuticals, 6000 experimental drugs and about 4338 non redundant proteins [10]. In this study, we have used ligand information from Drug Bank.

Software

Open Babel is an open collaborative project for analyzing chemical data. It provides several inbuilt features to search, convert, analyze, or store data from molecular modeling, chemistry, solid-state materials, biochemistry, or related areas. These includes file format support, fingerprints and fast searching, bond and atom characterization, canonical representation of molecules, coordinate generation in 2D and 3D space, stereochemistry, force fields [11]. We wrote Perl scripts to automate windows based Open Babel - 2.3.1 (32 bit) code. This script was used to calculate ligand similarity between two molecules. The Open Babel executable program was downloaded from the following link http://openbabel.org/wiki/Get_Open_Babel.

Ligand similarity analysis

Here, we introduce a ligand similarity method to produce binary finger printing based upon molecular properties or features. We label a feature as '0' if feature/property is absent from the molecular profile and '1' if the feature is present. Tanimoto coefficient is a well studied measure to find similarity between two molecules using their fingerprints. For instance, if we need to find similarity between two drug molecules (i.e. X and Y) then their tanimoto coefficient (Tc) can be defined as

$$Tc = N_{XY} / (N_X + N_Y – N_{XY})$$

Where N_X is number of features present in molecule X, N_Y represent features present in Y, N_{XY} account for common features present in X and Y (Figure 2). Further, we use set theory and label different numbers of properties for molecule X and Y as set A and set B respectively. Now, let us assume that set A contains five features labeled as {A, B, C, D, E} and set B contains 6 features {I, H, G, F, E, D}.

Set A = {A, B, C, D, E} Eq (1)
Set B = {I, H, G, F, E, D} Eq (2)
$N_A = 5$
$N_B = 6$
$N_{AB} = 2$
Tc = 2/(5+6-2) = 2/9 = 0.22

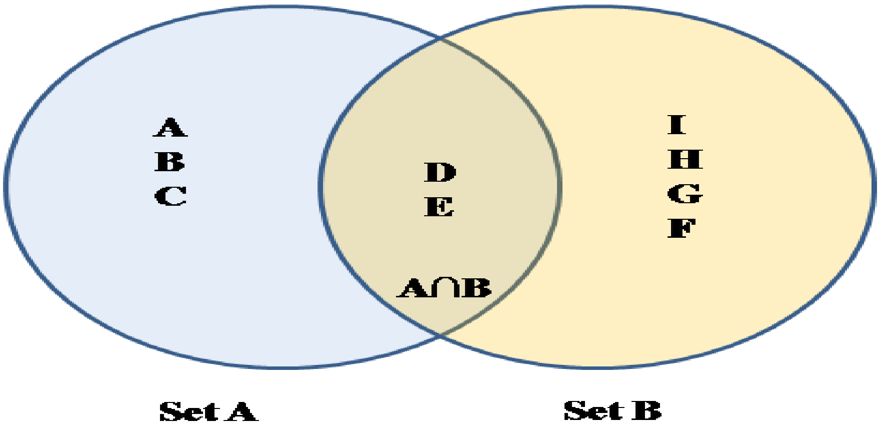

Figure 2: Venn diagram set A (for molecule in X) and set B (for Y)

Programming Language

Perl is a de facto scripting language for bioinformatics community due to ease of coding and flexibility to build variety of computational pipelines and automation [12-13]. In present work, we have used Perl for mining relevant information from DrugBank structure-data file (sdf) to parse relevant information such as DrugBank ID, Generic_Name, etc to create an input file for subsequent processing (Figure 3). The obtained information was used by Open Babel software to compute ligand similatity (namely tanimoto coefficient) and to construct similarity matrices using Perl scripts.

RESULTS

DrugBank contains information on sequence, structure and pathway of target molecule along with chemical, pharmacological and pharmaceutical data of experimental and withdrawn drugs. The structural information is provided in Structure Data Format (SDF) file. We have retrieved SDF files for different drugs and used Perl scripts to extract information such as DrugBank ID, generic name, chemical formula and drug SMILES string (Simplified Molecular-Input Line-Entry System) (See Figure 3 and 4). We have also coded for clustering algorithm to cluster data on drugs and their properties. As an application, we extracted drugs involved in hypertension (https://tinyurl.com/yaqnj7dt: Supplementary 1, 2, 3 & 4). A total of 23 drug molecules were identified: twelve (12) out of these decreases blood pressure (labeled as: decreases hypertension), whereas ten are found to increase blood pressure (causes hypertension). One drug out of this group known as anti-obesity drug (orlistat) was used as additional example due to our earlier studies on obesity network [8].

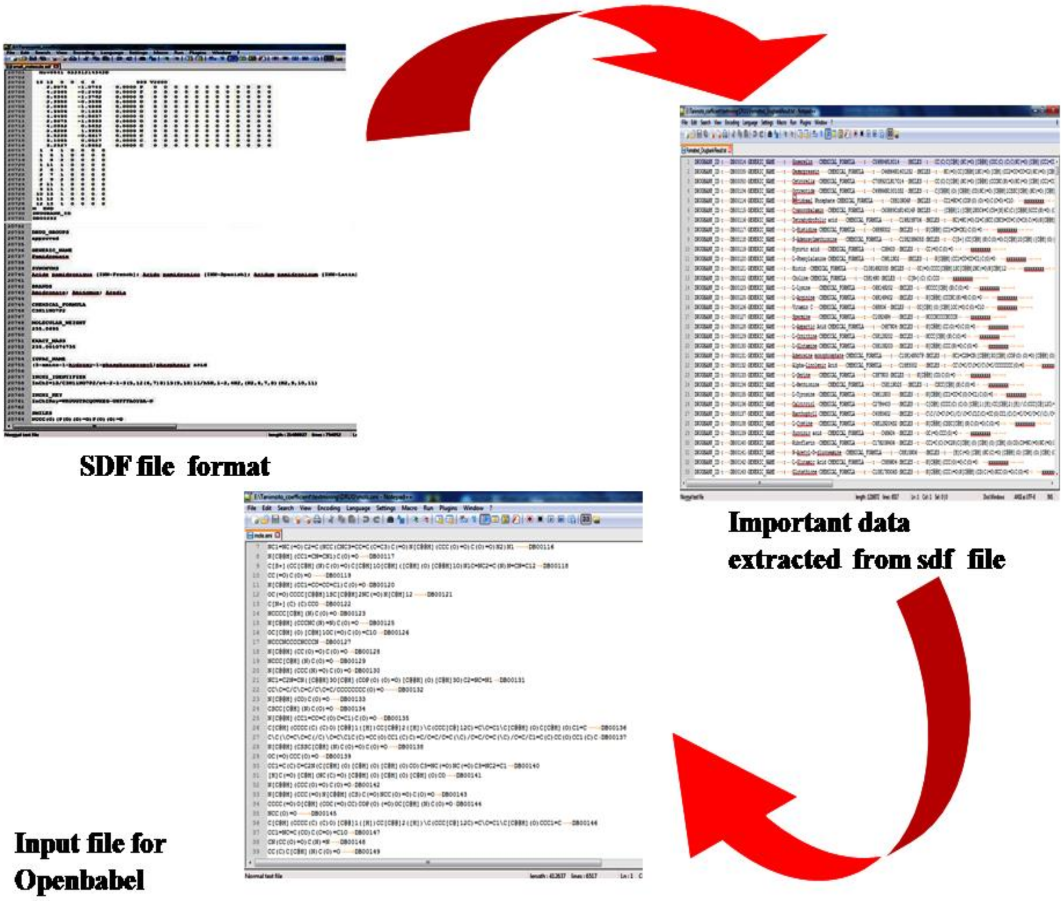

Figure 3 shows Preprocessing of sdf file to smi file an Open Babel input

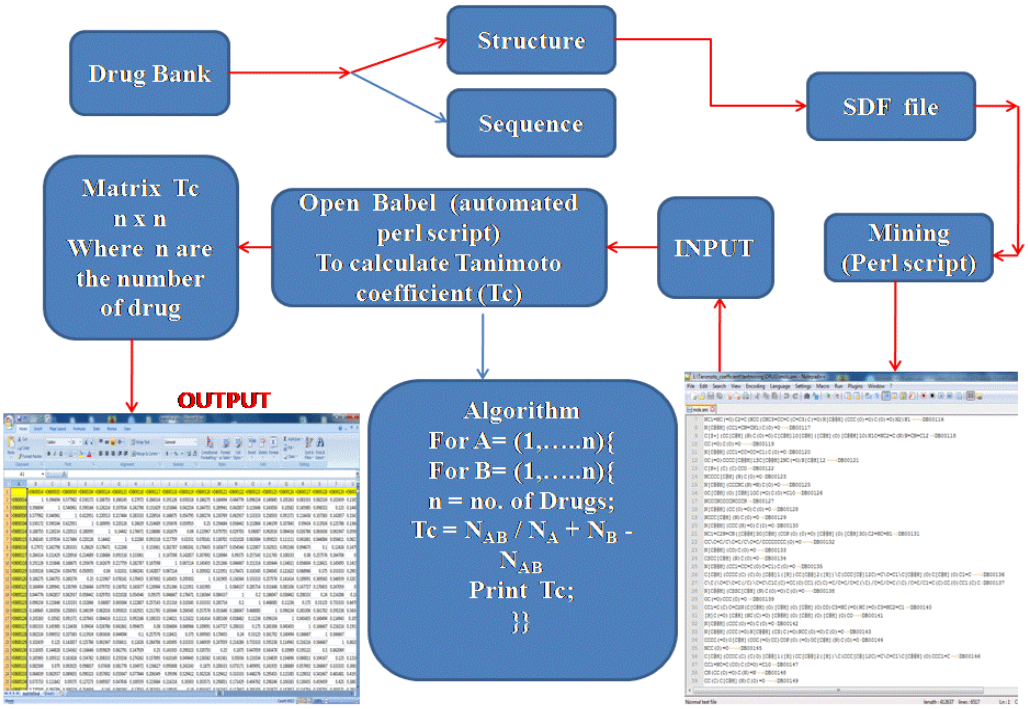

Figure 4 Process flow diagram of LSSAT

Each ligand in each set was compared to each ligand in every other set. Overall, 23 versus 23 set comparisons were made and a matrix of 23x23 ligands is generated (https://tinyurl.com/yaqnj7dt: Supplementary 5). Tanimoto coefficients (Tc) of chemical similarity were calculated for each pair of ligands. This ligand similarity matrix is subjected to hierarchical clustering and heat map plot is generated (Figure 5). Principle component analysis algorithm was used for clustering data based on Tc values as shown in Figure 6. Tc (Tanimoto coefficient) >=40% are listed in Table 1.

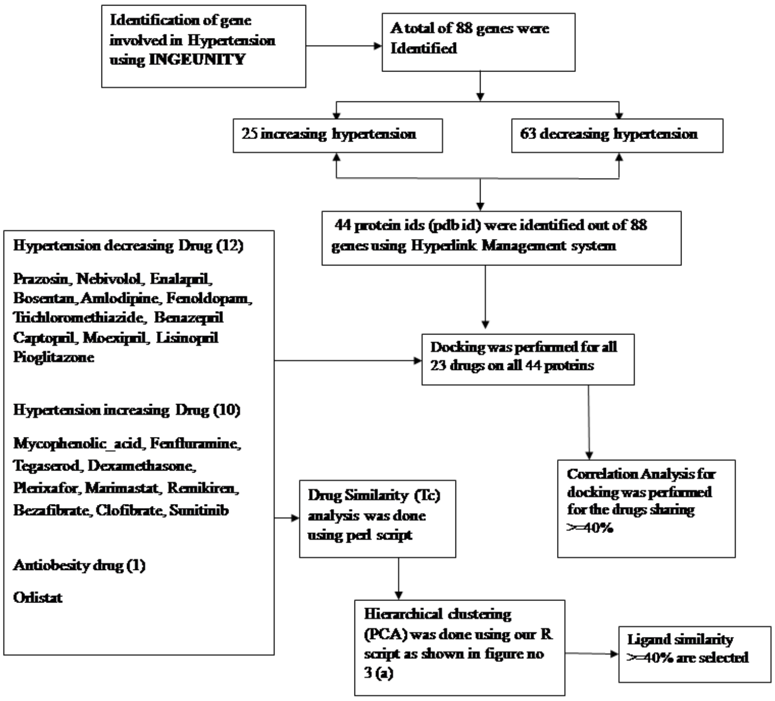
Identification of gene involved in Hypertension using INGEUNITY
A total of 88 genes were Identified
25 increasing hypertension
63 decreasing hypertension
44 protein ids (pdb id) were identified out of 88 genes using Hyperlink Management system
Hypertension decreasing Drug (12)
Prazosin, Nebivolol, Enalapril, Bosentan, Amlodipine, Fenoldopam, Trichloromethiazide, Benazepril Captopril, Moexipril, Lisinopril Pioglitazone
Hypertension increasing Drug (10)
Mycophenolic_acid, Fenfluramine, Tegaserod, Dexamethasone, Plerixafor, Marimastat, Remikiren, Bezafibrate, Clofibrate, Sunitinib
Antiobesity drug (1)
Orlistat
Docking was performed for all 23 drugs on all 44 proteins
Correlation Analysis for docking was performed for the drugs sharing >=40%
Drug Similarity (Tc) analysis was done using perl script
Hierarchical clustering (PCA) was done using our R script as shown in figure no 3 (a)
Ligand similarity >=40% are selected

Figure 5 Flow chart of ligand based similarity analysis

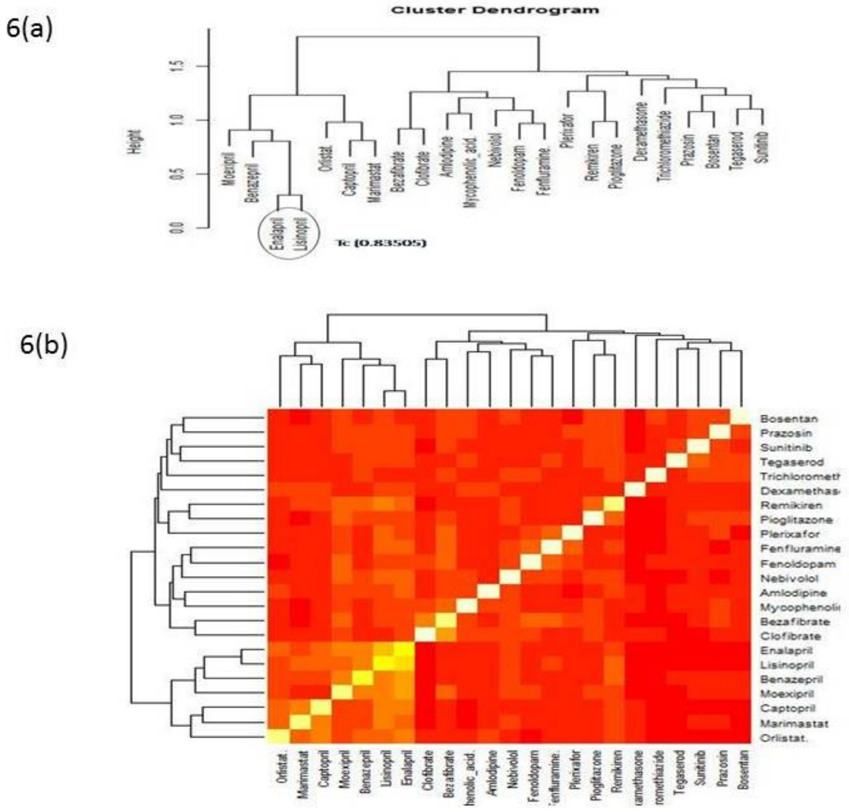

Figure 6 (a) Clustering (b) Heat map plot of the 23 drugs based on their similarity scores

Table 1: Drug-Drug Similarity score and docking energy when compared with subset of proteome (23 drug targets).

Drug 1	Drug 2	Tc	Percentage Similarity (%)	Correlation coefficient (R)
Lisinopril	Enalapril	0.83505	83%	0.97006
Lisinopril	Moexipril	0.505376	50%	0.60243
Lisinopril	Benazepril	0.52381	52%	0.65877
Enalapril	Moexipril	0.55	55%	0.69308
Enalapril	Benazepril	0.61111	61%	0.7495
Benazepril	Moexipril	0.462687	46%	0.53508
Orlistat	Lisinopril	0.401869	40%	0.27562
Orlistat	Enalapril	0.4905	49%	0.32077

We extracted around 44 target proteins involved in hypertension. Thereafter, 23 ligands were docked against these targets (https://tinyurl.com/yaqnj7dt: Supplementary 6 & 7). A correlation/regression analysis was performed between binding energies and those drugs having more than Tc value >=40% (0.4). Drugs with high similarity (Tc) shows good correlation with their binding energies (Table 1 and Figure 7; https://tinyurl.com/yaqnj7dt: Supplementary 8 and 9). We found significant correlation between binding energies and Tc values when compared with control (P<0.05). These data suggest that structurally similar drugs show similar binding affinities towards their targets. (Table 2 and Figure 8).

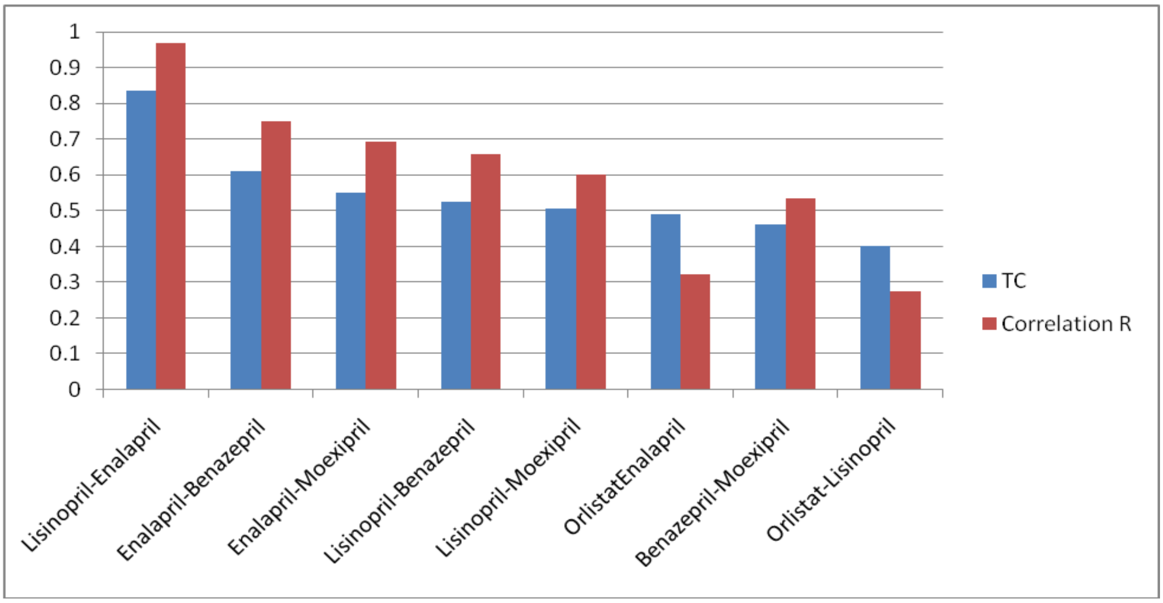

Figure 7 Correlation coefficient of binding energies of drugs w.r.t. Tc values

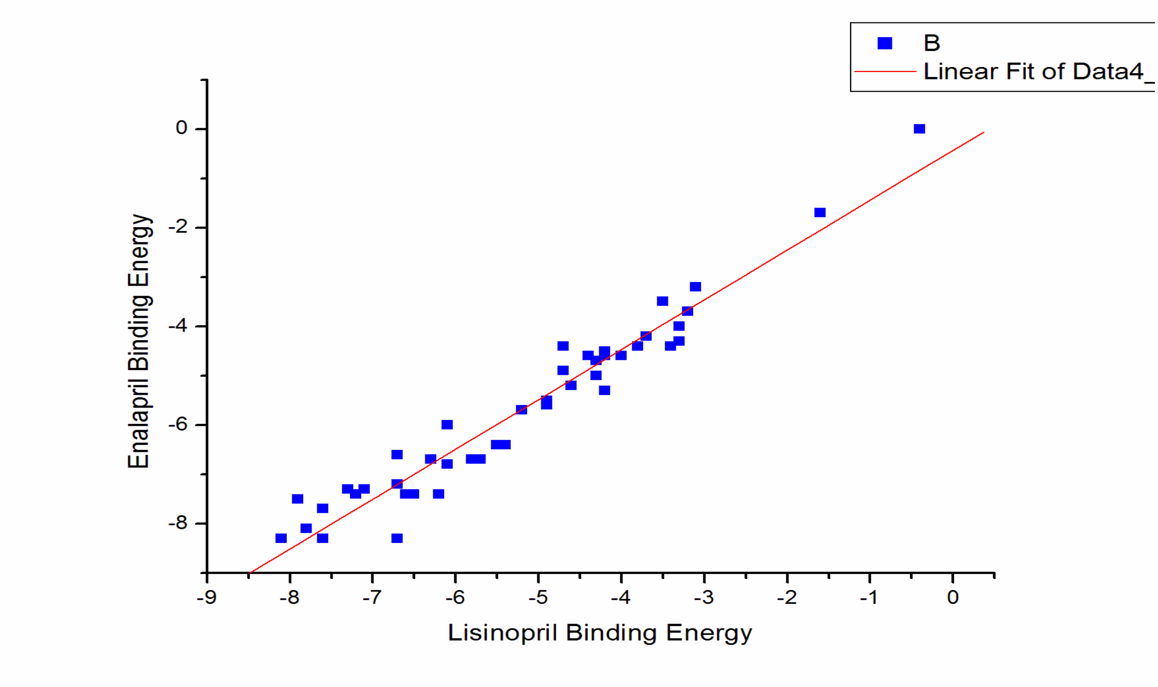

Figure 8 Correlation graph Lisinopril v/s Enalapril

CONCLUSION

In summary, we have shown that protein targets may be quantitatively related by their ligands. The ligand similarity plays a vital role in drug designing and drug re-positioning efforts. This tool can help pharmaceutical companies and researchers to screen large database such as DrugBank and PubChem to identify those ligand molecules that shares similar structure and property. This tool is able to calculate ligand similarity based upon their Tc value of drugs as well as perform large scale docking analysis to correlate two measures. We have shown utility of this tool in clinical conditions hypertension and obesity [8]. We also believe that this tool will play important role in network based therapeutics and side effect predictions.

Table 2: Binding energies of two highly similar drugs: Lisinopril and Enalapril with 83% similarity TC (0.8350)

Protein	Lisinopril (Binding energies Kcal/Mol)	Enalapril (Binding energies Kcal/Mol)
1B68	-3.4	-4.4
1BDA	-5.7	-6.7
1BIL	-6.7	-6.6
1C9H	-5.8	-6.7
1D2Q	-3.2	-3.7
1DB1	-7.6	-7.7
1GCZ	-6.2	-7.4
1HKN	-4.2	-4.5
1I7I	-7.3	-7.3
1IAP	-3.8	-4.4
1M9J	-8.1	-8.3
1MJV	-4.4	-4.6
1MMP	-6.1	-6.8
1N3U	-4.2	-4.6
1O86	-7.8	-8.1
1R4L	-7.9	-7.5
1SG1	-4.9	-5.5
1UVZ	-4.6	-5.2
1YK0	-5.4	-6.4
1YXJ	-3.3	-4.3
1Z3S	-3.5	-3.5
2A4Z	-6.7	-7.2
2AF0	-4.3	-4.7
2BXS	-7.6	-8.3
2D86	-4.2	-5.3
2GBT	-4.7	-4.9
2ILT	-7.1	-7.3
2NMP	-0.4	0
2QFA	-5.5	-6.4
2YGG	-3.7	-4.2
2Z8C	-4.3	-5
ZNN	-6.6	-7.4
3A7E	-5.2	-5.7
3B4V	-3.3	-4
3E7G	-6.1	-6
3EQM	-6.5	-7.4
3G2F	-4.9	-5.6
3H7W	-1.6	-1.7
3HNG	-6.7	-8.3
3J0A	-4.7	-4.4
3ODU	-7.2	-7.4
3UX0	-6.3	-6.7
4AL1	-4	-4.6
4JYO	-3.1	-3.2

REFERENCES

[1] Mullard A. Breakthrough programme turns two. Nat Rev Drug Discov. 2014; 13; 873-875.

[2] Reddy AS, Pati SP, Kumar PP, Pradeep HN, Sastry GN, Virtual Screening in Drug Discovery – A Computational Perspective, Current Protein and Peptide Science, 2007; 8; 329-351.

[3] McInnes C, Virtual screening strategies in drug discovery, Curr Opin Chem Biol., 2007; 5: 494-502.

[4] Willett P, Chemical Similarity Searching, J. Chem. Inf. Comput. Sci., 1998; 38; 983–996.

[5] Sun H, Pharmacophore-Based Virtual Screening, Curr Med Chem, 2008; 15; 1018-1024.

[6] Reynolds CR, Amini AC, Muggleton SH, Sternberg MJE, Assessment of a Rule-Based Virtual Screening Technology (INDDEx) on a Benchmark Data Set, J. Phys. Chem. B, 2012; 116; 6732–6739.

[7] Cavasotto CN, Orry AJ, Ligand docking and structure-based virtual screening in drug discovery, Curr Top Med Chem., 2007; 7; 1006-1014.

[8] Jagannadham J, Jaiswal HK, Rawal K, Comprehensive map of molecules implicated in obesity, Plos One, 2016, 2, e0146759.

[9] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M, DrugBank: a knowledgebase for drugs, drug actions and drug targets, Nucleic Acids Res., 2008; 36; D901–D906.

[10] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS, DrugBank 4.0: shedding new light on drug metabolism, Nucleic Acids Res. ; 2014; 42; D1091-D1097.

[11] Boyle NMO, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR, Open Babel: An open chemical toolbox, Journal of Cheminformatics, 2011; 3; 33.

[12] Rawal K, Ramaswamy R, Genome wide analysis of mobile genetic elements insertion sites, Nucl. Acids Res., 2011, 39, 6864-6878.

[13] Mandal P, Rawal K, Ramaswamy R, Bhattacharya A, Bhattacharya S. Identification of Insertion hot spots for non-LTR retrotransposons: Computational and Biochemical application to Entamoeba histolytica, Nucl. Acids Res.; 2006; 34; 5752-5763.