Project- Design of Potential Drugs Molecules for Coronavirus SARS Using Machine Learning

B. Shadrack Jabes

May 27, 2020

1 Introduction

The aim of this project is to predict the biological activity of drug molecules.

2 Dataset

SARS-CoV 3C-like protease (SARS-CoV $3CL^{pro}$), is a receptor, which is a part of the replicase polyproteins, cleaves a functional polypeptide and consequently leads to the maturation of SARS-CoV. Because of its functional importance in the SARS-CoV replication cycle, SARS-CoV $3CL^{pro}$ is considered a potential target to develop novel anti-SARS drugs[1].

28 compounds reported in[2] as novel inhibitors against SARS. SARS-CoV $3CL^{pro}$ is considered a potential target to develop novel anti-SARS drugs[1]. The initial dataset consists of 28 chemical compounds that share the same molecular backbone but different functional groups[2] Biological activity or IC_{50} (μM) is the concentration of the drug compounds leading to 50% inhibitory effect. The logarithm transformation of this parameter has been used as biological end points (log IC_{50}) and also gives a normal distribution. I used a web-based application known as pre-ADMET for predicting ADME data and building drug-like library. Using this *in-silico* method I retrived 955 diverse molecular descriptors, which includes 60 constitutional descriptors, 61 electrostatic descriptors, 16 geometrical descriptors, 130 physicochemical descriptor, and 688 topological descriptors. The other structural features of the molecules like polar surface area, *logp* including 14 other descriptors were retrived from ADME-tox server *(http://bioserv.rpbs.jussieu.fr/rpbs/html /an/t0_home.html)*. Here m corresponds to the number of drug molecules, which is 28. The biological activity data y collected from the literature is a 28 dimensional vector. The features n for each drug is 969. Total features consist of (28x969) dimensional vector.

2.1 Dimensional reduction: principal component analysis

I used principal component analysis to reduce the dimentionality of the descriptors for each molecule. Finally 5 descriptors for each of the 28 molecules was kept. I first compute the covariance matrix of the data. Then use Octaves *svd* function to compute the eigen vectors. These eigen vectors correspond to the principal components of variation in the dataset provided. The covariance $\sum$ is given as follows:

$$\sum = \frac{1}{m} X^T X$$

Here, X is the feature matrix with m molecules in rows and n features. Thus the covariance is a nxn dimensional matrix. To compute the principal components I use the following Octave command

$$[U, S, V] = svd(\sum);$$

Here U contains the principal components and S contains eigen values as a diagonal matrix. I choose the top K values in S and the corresponding eigen vectors U. The projected data can be usd instead of the high dimensional original dataset. Now each drug molecule will have K components instead of n. But the information about the drug molecules is retained because I only choose the eigen vector that corresponds to high eigen values. The K features of the drug molecules are shown in table (1)

Table 1: Descriptors and its class that are used to predict the biological activity of the ligands using neural network

S.No	class	descriptors
1	topological	chi 3 path
2	topological	$I\_edge\_adj\_deg\_mag$
3	topological	valence charge index 0
4	topological	ATS Moreau-Broto 1 vdW radius
5	constitutional	No. of rigid bonds

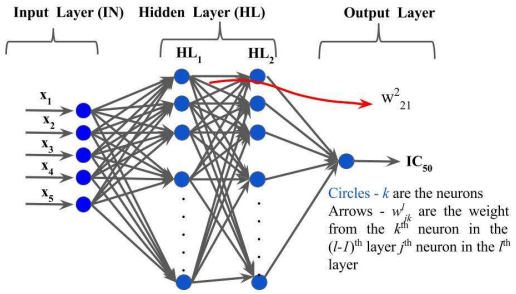

Figure 1: The neural network model used in this project consists of four layers. The input layer consist of 5 principal components. Then two hidden layers and the final layer is the output layer which predicts the biological activity of the drug molecule. Here, W represents the weights. The error δ is computed based on comparison between the biological activity predicted and the literature values y. Out of the 28 ligands considered in this study, 18 ligands are treated as training and the rest are treated as test data set. The 5 features of the drugs are classified as: (ATS Moreau-Broto 1 vdW radius, no. of rigid bonds, valence charge index 0, I_edge_adj_deg_mag and chi 3 path)

3 Artificial neural network model

The details about implementation of this algorithm is described elsewhere in another project *Project : Automated Handwritten Digit Recognition*. Model refinement was based on the consideration of statistical parameters such as correlation coefficient (r^2), cross validated correlation coefficient (q^2), MSE is the mean squared error. Cross-validation and the other statistical analyses were performed in MATLAB environment. The algorithm used for training neural network model was back propagation method. Using the same random seed, following parameters were fixed: max training epoch = 100, learning rate= 0.02, output learning rate= 0.045, initial weights and biases are taken randomly, number of neurons in the hidden layer = 12. Here I show only the results. The model used in this project is shown in figure (1). The mean squared error (MSE) is shown in table (3) and the predicted and actual values of the biological activity for the training and test cases are shown in table (2)

Table 2: Observed and the calculated $logIC_{50}$ values using ANN for the 28 inhibitors. The molecules shown in color are the test data set.

S.No	Observed $logIC_{50}$	Calculated $logIC_{50}$	residuals
1	0.477121	0.4908	-0.01368
2	1.00000	1.0089	-0.0089
3	1.04139	1.0491	-0.00771
4	1.07918	1.0908	-0.01162
5	1.14613	1.1529	-0.00677
6	1.17609	1.1829	-0.00681
7	1.17609	1.1826	-0.00651
8	1.17609	1.1787	-0.00261
9	1.47712	1.4796	-0.00248
10	1.60206	1.6035	-0.00144
11	1.60206	1.6035	-0.00144
12	1.65321	1.6512	0.00201
13	1.77815	1.7796	-0.00145
14	1.77815	1.7733	0.00485
15	2.00000	1.9968	0.0032
16	2.30103	2.298	0.00303
17	2.30103	2.2962	0.00483
18	2.30103	2.2989	0.00213
19	2.30103	2.2974	0.00363
20	2.30103	2.2983	0.00273
21	2.39794	2.3928	0.00514
22	2.47712	2.4693	0.00782
23	2.47712	2.4726	0.00452
24	2.477127	2.4738	0.00332
25	2.54407	2.541	0.00307
26	2.60206	2.5977	0.00436
27	2.69897	2.6955	0.00347
28	3.00000	2.9883	0.0117

Table 3: Statistical parameters of correlated data set using ANN method. Here r^2 is the correlation coefficient, q^2 is the cross validated correlation coefficient and MSE is the mean squared error.

S.No	r^2	q^2	MSE	No. of descriptors
1	0.7956	0.7955	0.0111	1
2	0.9543	0.9542	0.0036	2
3	0.9738	0.9735	0.0017	3
4	0.9868	0.9864	0.0008	4
5	0.9999	0.9999	0.0001	5

4 Conclusion

New molecules can be now designed based on the 5 features (shown in table 1) and its respective biological activity can be known. A potential drug candidate should show less biological activity.

References

[1] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, and R. Hilgenfeld. Coronavirus Main Proteinase ($3CL^{pro}$) Structure: Basis for Design of Anti-SARS Drugs. *Science*, 300:1763–1767, 2003.

[2] K.C. Tsai, S.-Y. Chen, P.-H. Liang, I.-L. Lu, N. Mahindroo, H.-P. Hsieh, Y.-S. Chao, and et. al. Discovery of a Novel Family of SARS-CoV Protease Inhibitors by Virtual Screening and 3D-QSAR Studies. *J. Med. Chem.*, 49:3485–3495, 2006.