Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604
Vol 2, Issue 2, June-2011, pp 161-166 http://www.bipublication.com

2D-QSAR ANALYSIS OF DIHYDROFOLATE REDUCTASE (DHFR) INHIBITORS WITH ACTIVITY IN *TOXOPLASMA GONDII* AND *LACTOBACILLUS CASEI*

Saumya K. Patel, S. Prasanth Kumar, Himanshu A. Pandya,

Yogesh T. Jasrai and Mehul I. Patni

Bioinformatics Laboratory Applied Botany Center, Gujarat University, Ahmedabad -380009, Gujarat, India.

ABSTRACT:

Methotrexate (MTX), an inhibitor of Dihydrofolate reductase (DHFR), is a well known drug given in the treatment of rheumatoid arthritis (RA). Due to its potential neurotoxicity, the patient has to discontinue the chemotherapy. In the present study, DHFR inhibitors which were structurally similar to MTX and had reported biological activity in model organisms such as *Toxoplasma gondii* and *Lactobacillus casei* was considered. A 2D-QSAR was modeled based on certain topological and constitutional descriptors along with its biological activity and found best 5 inhibitory molecules. *in vitro* validation of this inhibitors will be an alternative for effective drug development against RA.

Keywords: *Methotrexate, Dihydrofolate reductase, Rheumatoid arthritis, 2D-QSAR, Descriptors*

[I] INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory disease of unknown etiology, principally affecting smaller synovial joints in a symmetrical fashion, frequently leading to joint destruction. Methotrexate (4-amino-N10-methylpteroyl glutamic acid; MTX) has been largely used for this disease, which has been shown to be effective in 46 to 65% of the cases [1] and its neurotoxicity lead to discontinuation of treatment in RA varies from 10 to 30% [2]. Pharmacogenomic studies showed that gene polymorphism of HLA-12B may be correlated to some extent with the effectiveness and toxicity of this drug in RA patients. MTX is a folate analog and acts as a competitive inhibitor of the enzyme dihydrofolate reductase (DHFR). MTX with high affinity binds and inactivates DHFR, resulting in the depletion of metabolically active intracellular folates with subsequent inhibition of the synthesis of thymidylate and inosinic acid. Inhibition of DHFR causes termination of the synthesis of purine metabolites which are important for cell proliferation [3].

Quantitative structure-activity relationships (QSAR) derive models which describe the structural dependence of biological activity either by physicochemical parameters, by variables encoding different structural features, or by three-dimensional molecular property profile of the compounds. A compound with a biological profile is an initial point in drug designing strategy. It requires optimization of both the activity profile and structural features to become a lead molecule for testing in clinical trials. If a particular feature (e.g. steric) is contributing for the biological activity, optimization is performed through combinatorial explosion. Combinatorial explosion, the possibilities of substituents in the defined chemical space is greater than the thousands of magnitude. QSAR attempts to find the biological activity (prediction) of molecules with known structural properties and without the knowledge of activity profile. This is usually performed by molecules having similar structural properties and known activity spectrum. Now, this molecules will be trained through a multi-step multiple linear regression (MLR) analysis and a model will be generated. Testing is carried out by providing molecules with descriptors which was used in developing the model. Subsequently, the biological activity will be predicted by fitting a best line in the regression model and it is validated by certain statistical measures such as regression coefficient, correlation coefficient, F static, residuals, etc. A QSAR generally takes the form of a linear equation:

Biological Activity = Constant + $(C_1 * P_1)$ + $(C_2 * P_2)$ + $(C_3 * P_3)$ + ….... $(C_n * P_n)$

where the parameters P_1 through P_n are computed for each molecule in the series and the coefficients C_1 through C_n are calculated by fitting variations in the parameters and the biological activity [4].

[II] MATERIALS AND METHODS

2.1 Ligand preparation

31 unique DHFR inhibitors with reported activity in *Toxoplasma gondii* and *Lactobacillus casei* were taken from BindingDB [5]. The biological activity in terms of IC50 (also known as half maximal inhibitory concentration, defined as the concentration of an inhibitor required for 50% inhibition of its target) was used in the study and the respective chemical structure was retrieved from Pubchem in Structure Data Format (SDF) [6]. Conversion of SDF to Protein Data Bank (PDB) format was carried out using Swiss PDB viewer [7]. The ligand dataset was subjected to chemical atomic corrections and energy minimization using HyperChem v8.0.7 [8]. Molecular mechanics (MM+) force field with bond, angle, torsion, non-bonded, electrostatic and hydrogen-bonded atoms as components without any cutoff was utilized.

2.2 Calculation of molecular descriptors

The molecular descriptors were calculated for the ligand dataset using QSAR properties utility of HyperChem v8.0.7. Topological (surface area) and constitutional descriptors (LogP, molecular weight, polarizability and refractivity) were computed.

2.3 Calculation of pIC50

Reported IC50 values were manually converted into pIC50 (predicted IC50) using the formula given below. The term ‘pIC50’ is a scale for expressing IC50 value exponentially which normalizes the actual activity using negative logarithmic function which is considered as a prediction. Hence, the term ‘predicted’ is used.

pIC50 = -log IC50

2.4 Training and test data set for developing QSAR model

The ligand dataset comprised of 31 molecules were categorized into training and test data set. 16 molecules with predicted biological activity (pIC50) and 5 molecular descriptors were randomly selected for training the model. The remaining 15 molecules with computed descriptors were tested in the generated model without specifying its respective pIC50. It gave the predicted biological activity which was then compared with manually computed pIC50 to analyze the prediction behavior and accuracy of the generated model.

[III] RESULTS AND DISCUSSION

The computed molecular descriptors, reported and predicted biological activities (IC50 and pIC50) of the ligand dataset were shown in Table 1. The model was trained with training set using MLR analysis which generated a 2D-QSAR equation,

Biological activity = 1.580354080040E + 001 + 9.200251214572E-003*(surface area) + - 4.649129538003E-001*(logP) + 2.030973405986E-002*(molecular weight) + 1.017789824217E-001*(polarizability) + - 2.206177356776E-001*(refractivity)

In the above equation, biological activity was taken as dependent variable and all the descriptors were considered as independent variables which influenced the regression line of fitness.

The model was generated based upon the MLR equation with training data.

S.No	CID	Activity		Topological	Constitutional			
		IC50 (nm)	pIC50**	Surface Area (Å^2)	LogP	Molecular Weight (amu)	Polarizability (Å^3)	Refractivity (Å^3)
1	54369*	1.5	8.82	430.08	-1.07	325.37	35.47	95.88
2	72440	0.01	11	575.36	-1.45	454.45	45.06	120.24
3	126941*	6.0	8.22	572.69	-1.45	454.45	45.06	120.24
4	329367*	3.8	8.42	514.93	-0.26	439.43	43.71	115.74
5	425380	0.00125	11.90	588.46	-2.15	440.42	43.08	117.01
6	444617*	6.3	8.20	444.76	-1.91	340.38	36.82	100.97
7	447021*	3.9	8.41	431.94	-2.25	339.40	37.53	105.03
8	447815*	0.58	9.24	486.57	-3.00	384.44	41.12	111.51
9	457754	3.1	8.51	525.15	-2.69	422.44	44.13	123.35
10	462103*	0.84	9.07	350.29	-0.39	280.33	31.87	90.10
11	462123	4.7	8.33	458.40	-2.25	339.40	37.53	105.03
12	462125*	5.4	8.27	464.39	-3.24	369.42	40.00	111.40
13	462129	3.9	8.41	421.53	-2.25	339.40	37.53	105.03
14	462577	2.7	8.57	526.66	-2.66	398.40	42.96	116.26
15	472907*	1.7	8.77	402.70	-0.25	328.80	35.64	97.10
16	476035	0.88	9.05	442.17	-2.69	326.36	34.98	97.78
17	476043	2.2	8.66	339.85	-0.71	266.31	30.04	85.03
18	5479796*	1.8	8.74	506.46	-3.05	408.42	42.29	118.05
19	5479797	3.7	8.43	498.68	-3.92	407.43	43.01	119.87
20	5479798	4.3	8.37	557.59	-3.67	421.46	44.84	124.77
21	5481387*	5.8	8.24	534.61	-4.29	443.48	42.22	125.31
22	11979618*	6.2	8.21	700.54	-3.96	537.53	53.63	146.14
23	13942163	0.0023	11.64	623.22	-2.95	458.47	45.53	121.39
24	22708989*	2.5	8.60	607.28	0.10	441.49	45.51	119.90
25	25099170	3.7	8.43	446.62	-1.91	353.42	39.36	109.77
26	25132195	3.6	8.44	396.90	0.80	293.37	34.42	97.02
27	44281312*	2.2	8.66	544.18	-2.14	452.47	46.48	124.25
28	44285361*	0.005	11.30	595.07	-3.01	444.45	43.69	116.52
29	44342970	4.1	8.39	390.22	0.33	333.44	39.15	108.24
30	44388480*	6.3	8.20	372.80	-0.02	294.36	33.71	92.38
31	103918140	1.3	8.89	291.29	3.78	432.27	36.01	98.43

Table: 1. Molecular descriptors and predicted biological activity of the ligand dataset. Legends: CID- chemical identifiers *Training set for model generation, **Computed using formula.

Correlation among descriptors was initially analyzed to identify the descriptors producing partial positive impact on the overall model. The following were the high correlating descriptor pairs with correlation coefficient shown in bracket: molecular weight with surface area (0.97), polarizability with surface area (0.97), polarizability with molecular weight (0.98), refractivity with surface area (0.94), refractivity with molecular weight (0.97) and refractivity with polarizability (0.97). Hence, it was distinguished that surface area, molecular weight, polarizability and refractivity were good descriptors to generate the model. However, logP descriptor was identified as the worst descriptor due to the non-involvement in correlation with other descriptors.

To study the contribution of individual descriptor in biological activity, correlation was performed with individual descriptor and its respective pIC50 values. Descriptors such as LogP and refractivity was found to be negatively correlated (-0.11 and -0.06), respectively. It was also found that polarizability made no contribution for biological activity as its correlation coefficient was found to be 0.00. Surface area (0.12) and molecular weight (0.04) with contribution percentage of 1.48 % and 0.17 % was expressed as better descriptor for QSAR model. The model was statistically validated by regression coefficient which was found to be 57.91 %. It means that only ~58 % of the descriptors features (including negative impact conferred by certain descriptors) were contributing to biological activity. F-statistics, a measure to describe the level of heterozygosity in a population, was found to be 2.75. It may be simply due to the similar chemical architecture of the ligand dataset as it was embodied with analogues, which greatly eliminated the molecular diversity.

S.No	CID	Actual pIC50*	Calculated pIC50**	S.No	CID	Actual pIC50*	Calculated pIC50**
Training dataset				Test dataset			
1	54369	8.82	9.32	17	72440	11	9.06
2	126941	8.22	9.04	18	425380	11.90	9.73
3	329367	8.42	8.50	19	457754	8.51	7.74
4	444617	8.20	9.17	**20**	**462123**	**8.33**	**8.61**
5	447021	8.41	8.36	**21**	**462129**	**8.41**	**8.27**
6	447815	9.24	9.07	**22**	**462577**	**8.57**	**8.70**
7	462103	9.07	8.27	23	476035	9.05	9.74
8	462125	8.27	8.58	24	476043	8.66	8.97
9	472907	8.77	8.51	**25**	**5479797**	**8.43**	**8.42**
10	5479796	8.74	8.44	**26**	**5479798**	**8.37**	**8.24**
11	5481387	8.24	8.38	27	13942163	11.64	10.07
12	11979618	8.21	8.22	28	25099170	8.43	7.77
13	22708989	8.60	8.49	29	25132195	8.44	7.14
14	44281312	8.66	8.31	30	44342970	8.39	6.12
15	44285361	11.30	10.44	31	103918140	8.89	7.46
16	44388480	8.20	8.27				

Table: 2. Actual and calculated pIC50 values of training and test dataset. Legends: CID- chemical identifiers, *Computed using formula, **Predicted using the generated model. Best (fitted) molecules were highlighted in bold text.

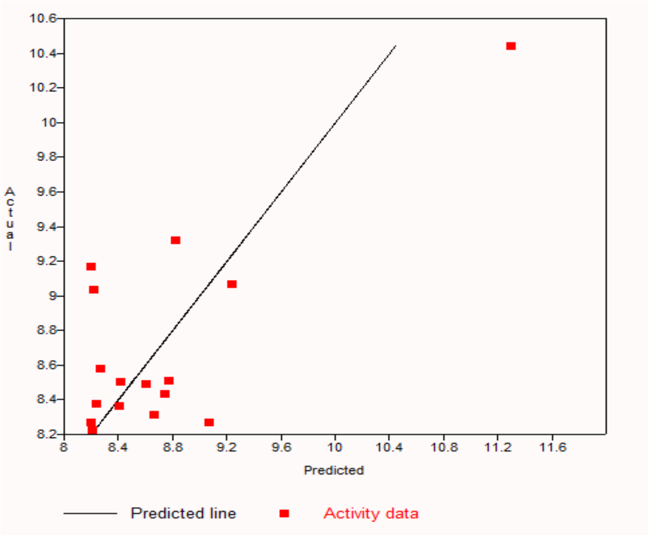


Fig: 1. Scatter plot of actual and calculated (predicted) pIC50 values of the training dataset.

The scatter plot of actual versus predicted pIC50 value was shown in Fig. 1. It showed that 3 molecules were occupied the regression line and 4 molecules were found to be around the line (Table 1). Thus, scatter plot analysis exhibited 7 molecules were very close to the model trained upon the 16 molecules. In other words, the pIC50 (computed by formula) and the calculated pIC50 values (result of generated model) were very close to each other. Test data set was given as input with the only specification of molecular descriptors. The model predicted the biological activity of the test set which is termed as ‘calculated pIC50 value’. To understand the prediction accuracy of the model, the calculated pIC50 and actual pIC50 (computed by formula) values were graphically analyzed with connected line graph (Fig. 2). Compounds with chemical identifiers (CID) viz. 5479797, 5479798, 462129, 462577 and 462123 were found to be best molecules studied from the 2D QSAR analysis (Table 2). Molecular similarity among the best molecules were analyzed graphically when superimposed (Fig. 3).

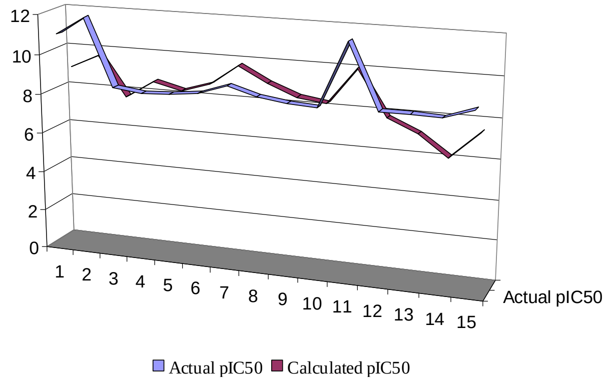
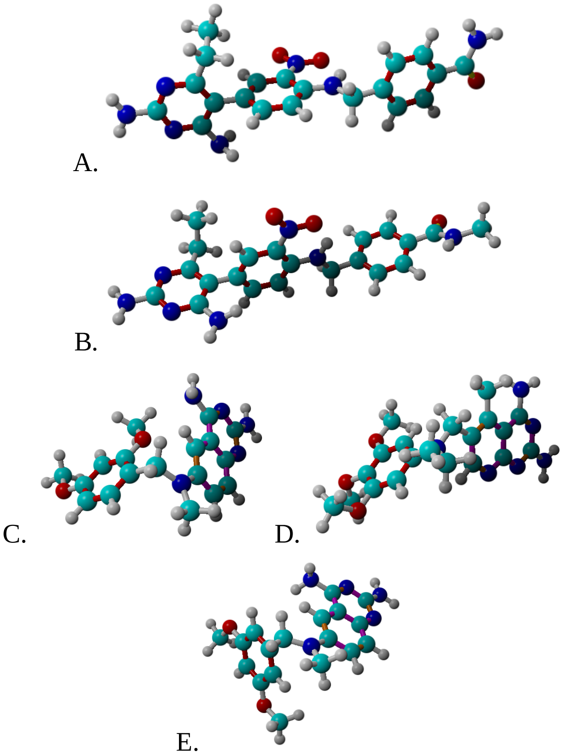


Fig: 2. Line connected graph of actual and calculated (predicted) pIC50 values of the test dataset. Overlapping regions indicate the best molecule fitted over the regression equation.



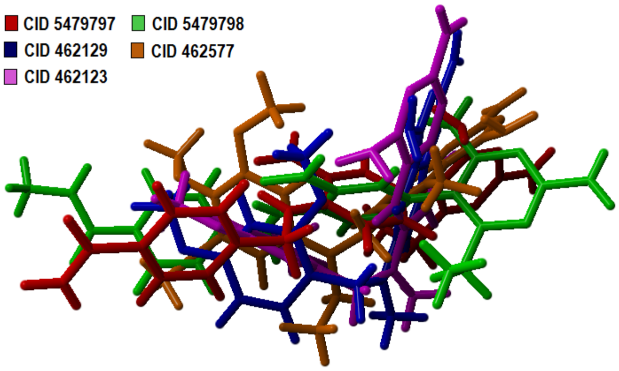


Fig: 3. 5 Best molecule (fitted over the regression equation) generated by the model whose actual and calculated pIC50 values were very close. (A) CID 5479797,(B) CID 5479798, (C) CID 462129, (D) CID 462577, (E) CID 462123 and (F) Superimposed view of all the best 5 molecules.

[IV] CONCLUSION

2D QSAR analysis of DHFR inhibitors with activity reported in *Toxoplasma gondii* and *Lactobacillus casei* was carried out. Randomly selected training molecules generated a model provided a MLR equation, by which the activity of test set, was conducted. Correlation studies among the descriptors and with the activity revealed LogP had a negative impact on the biological activity. However, descriptors such as surface area, molecular weight, polarizability and refractivity contributed very less towards the activity of the molecules. Hence, it is clear that similar chemical structures (analogues) did not optimize the activity of the molecule and stresses the requirement of additional inhibitors reported elsewhere with maximum molecular diversity which can reproduce a better model for structure-activity relationships. 5 best molecules (best fitted on the regression line) was identified based on the actual pIC50 and calculated pIC50 values whose highest and lowest residuals (difference between actual and predicted pIC50 values) was found to be 0.28 and 0.01. It was also noted that the first 2 and the next 3 best molecules were conformers to each other.

ACKNOWLEDGEMENT

The authors deeply acknowledges Gujarat University for providing Bioinformatics Infracstructure.

REFERENCES

[1]. Bathon, J.M., Martin, R.W., Fleischmann, R.M., Tesser, J.R., Schiff, M.H., Keystone, E.C., Genovese, M.C., Wasko, M.C., Moreland, L.W., Weaver, A.L., Markenson, J. and Finck, B.K., 2000. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. N. Engl. J. Med., 343: 1586–1593.

[2]. Alarcon, G.S., Tracy, I.C. and Blackburn, W.D., 1989. Methotrexate in rheumatoid arthritis: toxic effects as the major factor in limiting long-term treatment. Arthritis Rheum., 32: 671–676.

[3]. Oewierkot, J. and Szechiñski, J., 2006. Methotrexate in rheumatoid arthritis. Pharmacological Reports, 58: 473-492.

[4]. Recent Advances in QSAR Studies: Methods and Applications (Challenges and Advances in Computational Chemistry and Physics), 2004. Editors: Tomasz Puzyn, Jerzy Leszczynski and Mark T. Cronin. *Springer Verslag*, 1st Ed., ISBN: 1402097824

[5]. Liu, T., Lin, Y., Wen, X., Jorrisen, R.N. and Gilson, M.K., 2007. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Research, 35: pp.198-201. www.bindingdb.org/ Last accessed on 18.05.2011

[6]. PubChem. Free database of chemical structures of small organic molecules and information on their biological activities. National Centre for Biotechnology Information (NCBI). www.pubchem.ncbi.nih.gov/. Last accessed on 18.05.2011

[7]. Guex,N., Diemand, A. and Peitsch, M.C., 1999. Protein modelling for all. TiBS, 24:364-367. http://www.expasy.org/spdbv/

[8]. HyperChem^{TM} v8.0.7, Hypercube, Inc., USA.