Iranian Journal of Pharmaceutical Research (2017), 16 (3): 981-998
Received: June 2015
Accepted: October 2015

Original Article

A Comparative QSAR Analysis, Molecular Docking and PLIF Studies of Some N-arylphenyl-2,2- Dichloroacetamide Analogues as Anticancer Agents

Masood Fereidoonnezhad[a,b], Zeinab Faghih[b], Ayyub Mojaddami[a,b], Zahra Rezaei[b] and Amirhossein Sakhteman[b,c*]

[a]Department of Medicinal Chemistry, School of Pharmacy, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. [b]Department of Medicinal Chemistry and Pharmaceutical Sciences Research Centre, School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran. [c]Medicinal and Natural Products Chemistry Research Center, Shiraz University of Medical Sciences, Shiraz, Iran.

Abstract

Dichloroacetate (DCA) is a simple and small anticancer drug that arouses the activity of the enzyme pyruvate dehydrogenase (PDH) through inhibition of the enzyme pyruvate dehydrogenase kinases (PDK1-4). DCA can selectively promote mitochondria-regulated apoptosis, depolarizing the hyperpolarized inner mitochondrial membrane potential to normal levels, inhibit tumor growth and reduce proliferation by shifting the glucose metabolism in cancer cells from anaerobic to aerobic glycolysis. In this study, a series of DCA analogues were applied to quantitative structure–activity relationship (QSAR) analysis. A collection of chemometrics methods such as multiple linear regression (MLR), factor analysis–based multiple linear regression (FA-MLR), principal component regression (PCR), and partial least squared combined with genetic algorithm for variable selection (GA-PLS) were applied to make relations between structural characteristics and cytotoxic activities of a variety of DCA analogues. The best multiple linear regression equation was obtained from genetic algorithms partial least squares, which predict 90% of variances. Based on the resulted model, an *in silico*-screening study was also conducted and new potent lead compounds based on new structural patterns were designed. Molecular docking as well as protein ligand interaction fingerprints (PLIF) studies of these compounds were also investigated and encouraging results were acquired. There was a good correlation between QSAR and docking results.

Keywords: DCA; QSAR; *in silico* screening; Descriptor analysis; Docking; PLIF studies.

Introduction

There has been a great detonation in the number of potential molecular targets that can be investigated for cancer treatment. Some metabolic pathways that play a great role in tumor growth are being explored as novel targets for anticancer drug development (1, 2). Mitochondria are essential for the continuation of life in higher eukaryotic cells, including cancer cells. Several common characteristics of recognized tumor cells directly or indirectly depend on mitochondrial deregulation (3). Meanwhile, they control programmed cell death (apoptosis). Extensive investigation has

* Corresponding author:
E-mail: *asakhteman@sums.ac.ir

been focused on the progression of strategies designed in order to selectively induce apoptosis in cancer cells (1, 4). Pyruvate dehydrogenase complex (PDC) is one of the major regulators of mitochondrial function. PDC is a complex of three enzymes that convert pyruvate into acetyl-CoA by pyruvate decarboxylation. PDC via production of reactive oxygen species (ROS) and followed by oxidative damage, can induce apoptosis. The activity of PDC is regulated by reversible phosphorylation of three serine residues on the E1α subunit. PDH kinases (PDK) phosphorylate these sites. There are four known isoforms of PDKs that are distributed in a different manner in the tissues. Their expressions are regulated by factors like hypoxia, starvation and employment of glucose and fatty acids in various tissues. It should be noted that the role of PDK (1–4) is inactivation of PDC (1, 5).

It was discovered that dichloroacetate (DCA) acts as a pyruvate dehydrogenase activator through stimulating PDC activity. DCA is a lactate-lowering drug, which has been in use for many years to treat various diseases such as lactic acidosis, inborn errors in mitochondrial function (6, 7). In 2007 it was discovered that the drug DCA induced the death of human lung, breast and brain cancer cells that were embedded into rats, while being non-toxic to healthy cells (8). DCA prevent cell growth of a large range of tumor cells like lung, breast, glioblastoma (8), endometrial (9), prostate (10), pediatric (11), pancreatic (12), cervical (13) and colorectal (14) cancer cells by promoting mitochondria-regulated apoptosis and decreasing proliferation. Nevertheless, it exerted no obvious toxicities on the normal cells.

Molecular modelling studies such as quantitative structure activity relationship (QSAR) and molecular docking have a great importance in the field of medicinal chemistry. There are different variable selection methods available for QSAR studies such as multiple linear regression (MLR), principal component or factor analysis (PCA/FA), genetic algorithm, and so on (15). Recently structure-based design of some PDK2 inhibitors from molecular docking studies has been reported and some compounds were introduced as the potent inhibitors of PDK2 (16, 17).

Here, in this paper, QSAR studies of a series of N-arylphenyl-2, 2-dichloroacetamide analogues with cytotoxic activity on human non-small cell lung cancer cell line (A 549). which recently designed and synthesized by Li *et al*. (18) have been explored. Among different QSAR models, the best multiple linear regression equation was obtained from GA-PLS models, which was a linear seven-parameter model. Thereafter, a virtual screening study was employed to determine novel biologically active patterns by insertion, deletion and substitution of different substitutes on the primary molecules. The results of this study led to the identification of novel structures, which are potent anticancer agents according to the QSAR model. It also should be mentioned that molecular docking as well as PLIF studies of these compound were also carried out and the promising results were obtained.

Experimental

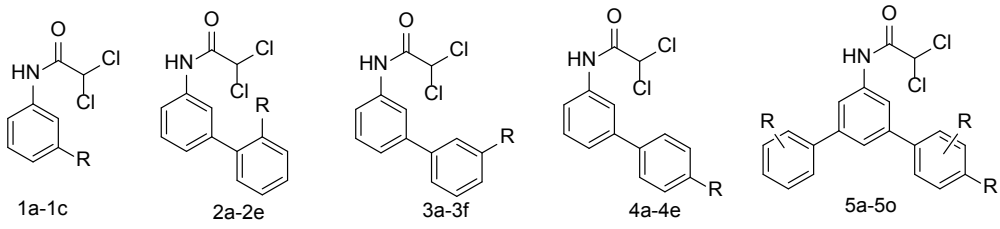
Data set

The biological data used in this paper are cytotoxic activity of a series of N-arylphenyl-2, 2-dichloroacetamide analogues on human non-small cell lung cancer cell line (A 549). which were designed, synthesized and evaluated for their anticancer activity by Li *et al.* (18). The structural features and biological activity of these compounds are listed in Table 1 The biological data were converted to logarithmic scale (pIC_{50}) and then used for subsequent QSAR analysis as dependent variables.

Molecular descriptors

The two dimensional structures of the ligands were drawn using ACD chemsketch software. Then the ligands were subjected to minimization procedures by means of an in house TCL script using Hyperchem (Version 8, Hypercube Inc., Gainesville, FL, USA). Each ligand was optimized with different minimization methods such as commonly used molecular mechanics method (MM+) and then quantum based semi-emprical method (AM1) using Hyperchem package. The Z-matrices of the structures were constructed by the software and then transferred to the Gaussian 98 program (19). HyperChem,

Table 1. Chemical structure of the N-arylphenyl-2, 2-dichloroacetamideanalogues used in this study and their docking binding energy, experimental and cross-validated predicted activity (by GA-PLS) for cytotoxic activity.



Name	R	Exp. pIC_{50}	Pred. pIC_{50} [a]	Binding Energy (kcal/mol)	Leverage
1a	Ph	5.05061	4.8485	-6.36	0.01161
1b	Furan-3-yl	5.071604	4.9098	-6.43	0.00281
1c	Thiophen-2-yl	4.322484	4.4432	-5.89	0.01231
2a	Me	5.006123	4.8475	-6.12	0.01121
2b	F	4.879097	4.8434	-6.01	0.01127
2c	Cl	4.646661	4.9101	-5.93	0.00278
2d	OMe	4.881735	4.8335	-5.85	0.01073
2e	CF3	5.124939	4.9422	-6.59	0.01051
3a	Me	5.060481	4.8621	-6.39	0.01123
3b	F	5.020907	4.8436	-6.51	0.01153
3c	Cl	4.668978	4.9230	-6.01	0.00170
3d	OMe	5.365523	4.9751	-7.05	0.0064
3e	CF_3	4.732828	4.9418	-6.07	0.00200
3f	$NHCOCHCl_2$	4.320118	4.3482	-5.86	0.01161
4a	Me	5.080922	4.9111	-6.29	0.01012
4b	F	5.105684	4.9632	-6.57	0.01159
4c	Cl	4.440812	4.5109	-5.98	0.00186
4d	OMe	5.069051	4.9034	-6.45	0.01208
4e	CF_3	4.627456	4.7315	-5.92	0.01117
5a	2-OMe	4.59346	4.4290	-5.95	0.00283
5b	3-OMe	5.026872	5.1020	-6.32	0.01401
5c	4-OMe	4.682354	4.6284	-5.99	0.02014
5d	2-OEt	5.137869	5.0302	-6.73	0.00294
5e	3-OEt	5.054039	5.0937	-6.39	0.01232
5f	4-OEt	5.094204	5.1284	-6.3	0.02014
5g	2-SMe	5.186419	4.9985	-6.71	0.01132
5h	3-SMe	4.798603	4.8185	-6.02	0.01195
5i	4-SMe	4.651695	4.7485	-5.93	0.01201
5j	4-OiPr	5.417937	5.3289	-6.77	0.02025
5k	3,4-diO-CH_2	5.761954	5.5788	-7.89	0.29556
5l	4-(Tetrahydro-2H-pyran-2-yl)oxy	5.560667	5.6437	-7.27	0.36181
5m	4-iPr	4.759201	4.8005	-6.05	0.01161
5n	3-F-4-OMe	5.709965	5.7285	-7.74	0.02016
5o	3-F-5-OMe	5.04624	5.0865	-6.4	0.01092

[a]Cross-validated prediction by

Gaussian 98 and Dragon softwares (20) were used for calculation of molecular descriptors. Highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies and molecular dipole moment were calculated by Gaussian98. Quantum chemical indices of hardness ($\eta = 0.5$ (HOMO+LUMO)); softness ($S = 1 / \eta$); electronegativity (χ = -0.5 (HOMO-LUMO)); and electrophilicity ($\omega = \chi^2/2\eta$) were calculated according to the equations proposed by Thanikaivelan *et al.* (21). Some chemical parameters including molar volume (V). molecular surface area (SA), hydrophobicity (logP), hydration energy (HE) and molecular polarizability were calculated using Hyperchem software. Dragon calculated different topological, geometrical, charge, empirical and constitutional descriptors for each molecule. 2D autocorrelations, aromaticity indices, atom-centered fragments and functional groups were also calculated by dragon.

In the case of docking procedure, each ligand was optimized with different minimization MM^+ then AM1 using HyperChem 8. The output structures were thereafter converted to PDBQT using MGL tools 1.5.6 (22). The three dimensional crystal structure of pyruvate dehydrogenase kinase 2 (PDB ID: 2BU8) was retrieved from protein data bank (23). Co-crystal ligand molecules were excluded from the structures and the PDBs were checked in terms of missing atom types by modeller 9.12 (24). An *in house* application (MODELFACE) was used for generation of python script and running modeller software. Subsequently, the enzymes were converted to PDBQT and gasteiger partial charges were added using MGLTOOLS 1.5.6.

Model development

Four different regression methods were conducted for constructing QSAR equations: 1) simple multiple linear regression with stepwise variable selection (MLR) 2) factor analysis as the data preprocessing step for variable selection (FA-MLR), 3) principal component regression analysis (PCRA), and 4) genetic algorithm–partial least squares (GA-PLS). These methods are well substantiated in the QSAR studies, and therefore, these methods are described briefly (25).

Stepwise regression is a semi-automated process of building a model by successively adding or removing variables based solely on the t-statistics of their estimated coefficients. In stepwise regression (26), a multiple-term linear equation was constructed step by step. The basic procedures include (i) recognizing a primary model, (ii) iteratively ‹stepping′, that is, repetitively changing the model at the prior step by adding or removing a predictor variable in accordance with the ‹stepping criteria′ (in our case, probability of F = 0.05 for inclusion; probability of F = 0.1 for leaving out for the forward selection method), and (iii) terminating the search when stepping is no longer possible given the stepping criteria, or when a known maximum number of steps have been obtained. Particularly, at each step, for determining which one will contribute most to the equation, all variables are reviewed for evaluation (26). The variable will then be applied in the model, and the process starts again. A limitation of the stepwise regression search approach is that it assumes there is a single ‹best′ subset of X variables and search for identifying it. There is often no unique ‹best′ subset, and whole possible regression models with a similar number of X variables as in the stepwise regression solution should be fitted subsequently to explore whether some other subsets of X variables might be better (27). Here in this study, MLR with stepwise selection and elimination of variables was applied for developing QSAR models using SPSS software (version 21; SPSS Inc., IBM, Chicago, IL, USA). Using MATLAB 2015 software (version 8.5; Math work Inc., Natick, MA, USA), the resulted models were validated by leave-one-out cross-validation procedure to check their prediction ability and robustness.

In FA-MLR method, although classical approach of multiple regression technique was applied as the final statistical tool for developing QSAR relation, factor analysis (FA) (15, 26) was used as the data-preprocessing step to identify the important predictor variables contributing to the response variable and to avoid collinearities among them. In a typical factor analysis procedure, standardizing of the data matrix followed by constructing a correlation matrix is done. An eigenvalue problem is then solved

and the factor pattern can be acquired from the corresponding eigenvectors (characteristic vector). The principal objectives of factor analysis (FA) are to display multidimensional data in a space of lower dimensionality with minimum loss of information (explaining >95% of the variance of the data matrix) and to extract the basic features behind the data with ultimate goal of interpretation or prediction. Factor analysis was done on the data set containing biological activity and all descriptor variables, which were to be considered. The factors were extracted by principal component method and then rotated by (VARIMAX) rotation (28).

Along with FA-MLR, PCRA was also tried for the data set. In this method (15, 26), factor scores that obtained from FA are used as the predictor variables. PCRA has a benefit that collinearities among X variables are not a disturbing factor and that the number of variables included in the analysis may exceed the number of observations (29). While the main purpose of FA-MLR is to identify relevant descriptors, in PCRA model all descriptors are supposed to be important.

Genetic algorithms (GA) generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

Partial least square (PLS) is a generalization of regression, that can handle data with forcefully correlated and numerous X variables (30). It gives reduced solution, which is statistically more robust and reliable than MLR. The linear PLS model finds ‹new variables′ (latent variables or X scores) that are linear combination of the original variables. To avoid overfitting, a strict test for the significance of each consecutive PLS component is necessary and then stopping when the components are non-significant. Cross-validation is a practical and credible method of testing this significance (31). Application of PLS thus allows the construction of larger QSAR equations while still avoiding over fitting and eliminating most variables. Usually PLS is applied in combination with cross-validation to obtain the optimum number of components (26, 32, 33). In the GA-PLS procedure, in addition to the best set of descriptor, the optimum number of concealed variable must be determined. Here, for each subset of descriptors (i.e., for each chromosome of the GA), a PLS model was developed separately and therefore the number of latent variables was optimized. The PLS regression method was applied the NIPALS-based algorithm existed in the chemometrics toolbox of MATLAB software. Leave-one-out cross-validation procedure was used to obtain the optimum number of factors based on the Haaland and Thomas F-ration criterion (26, 34). The MATLAB PLS toolbox developed by eigenvector company was used for PLS and GA modeling. All calculations were run on a core i7 personal computer (CPU at 6 MB) with Windows 7 operating system.

Model validation

Statistical parameters including correlation coefficient (R^2), standard error of regression (SE), and variance ratio (F) at specified degrees of freedom were used for validating the goodness-of-fit of the resulted QSAR models. The generated QSAR equations were also validated by leave-one-out cross-validation correlation coefficient (Q^2), root mean square error of cross-validation (RMScv) and cross validation cross validation (Cvcv). According to Tropsha *et al.* (35) the predictive ability of a QSAR model should be tested on an external set of data that has not been taken into account during the process of developing the model. Therefore, as it was shown in table 1, an external test set composed of randomly selected 7 molecules (for example 1a, 2a, 2e, 3c, 4d, 5 h. and 5m) were applied to determine the overall prediction ability of the resulted models. It should be emphasized that we carried out each QSAR model with more than 3 test set and the best equation was considered as the best model.

Applicability domain

One of the great uses of a QSAR model is based on its precise prediction ability for new compounds. A model validation is just within its training domain, and new compounds must be appraised as belonging to the domain before the model is applied. The applicability domain is appraised by the leverage values for each compound. A Williams's plot (the plot of standardized residuals versus leverage values (h)) can then be used for an immediate and

simple graphical detection of both the response outliers (Y outliers) and structurally influential chemicals (X outliers) in our model. In this graph, the applicability domain is established inside a squared area within ±x (standard deviations) and a leverage threshold h*. The threshold h* is generally fixed at 3(k + 1)/n (k is the number of model parameters and n is the number of training set compounds), whereas x = 2 or 3. Prediction must be considered unreliable for compounds with a high leverage value (h > h*). From the other point of view, when the leverage value of a compound is lower than the threshold value, the probability of agreement between observed and predicted values is as high as that for the training set compounds (36, 37).

Docking procedure

The docking simulations were carried out by means of an *in house* batch script (DOCKFACE) for automatic running of AutoDock 4.2 (38) in a parallel mode using all system resources. In all experiments Genetic algoritm search method was used to find the best pose of each ligand in the active site of the target enzyme. Random orientations of the conformations were generated after translating the center of the ligand to a specified position within the receptor active site, and making a series of rotamers. This process was recursively repeated until the desired number of low-energy orientations was obtained. No attempt was made to minimize the ligand-receptor complex (rigid docking). For Lamarckian GA method; 27,000 maximum generations; 2500000 maximum No. of evaluations, 150population size, mutation rate of 0.02; and a crossover rate of 0.8 were used. A grid box of 50×50×50 points in x, y, and z direction with a grid spacing of 0.375 Å was built. No. of points in x, y and z was 50, 40 and 81 respectively.

Protein ligand interaction fingerprint (PLIF)

In order to perform PLIF studies on docking results, the poses of docking were extracted from dlg files using an *in house* vb.net application (pre Aupos SOM) (39). The resulted pdbqts and the receptor were converted to mol2 using Open Babel 2.3.1. The resulted mol2 files were submitted to Aupos SOM 2.1 web server (40-42). Two training phases with 1000 iterations were set in the self-organizing map settings of Aupos SOM conf files. Other parameters of the software were remained as default. The output files were subjected to Dendroscope 3.2.10 for visualization of the results (43, 44). The PLIF parameters were set as default of the AuPos SOM v2.1 Web Application.

Results and Discussion

In this study, we executed a detailed QSAR study using a combination of chemical, electronic and substituent constant, to explore structural parameters affecting cytotoxic activity of novel N-arylphenyl-2,2- dichloroacetamide analogues. Among the different chemometrics tools available for modeling the relationship between the biological activity and molecular descriptors, four methods (i.e. stepwise MLR, FA-MLR, PCRA, and GA-PLS) were applied and compared here. A comparison between stepwise FA-MLR and MLR will indicate which variable selection method (stepwise or FA) is well suited for MLR, whereas a comparison between FA-MLR and PCRA reveals for modeling of the studied biological activities, using original descriptors selected based on factor loading or using the factor scores calculated based on all calculated descriptors results in more suitable model. Eventually, GA-PLS, which is assumed to produce the most useful model, was employed, and its results were compared with the other employed models.

MLR modeling

Firstly, separate stepwise selection-based MLR analyses were performed using different types of descriptors, and then, a MLR equation was obtained utilizing the pool of all calculated descriptors. As it was shown in Table 2, statistical parameters such as correlation coefficient (R^2). correlation coefficient (R^2_p) of test set, standard error of regression (SE), and variance ratio (F) at specified degrees of freedom, leave-one-out cross-validation correlation coefficient (Q^2), cross validation cross validation (Cvcv) and root mean square error of cross-validation (RMScv) were used for validating the goodness-of-fit of the resulted QSAR equations. Equation 1

Table 2. The results of different QSAR models with different type of dependant variables

Model	Eq.no.	MLR Equation	n[a]	R^2_c	Q^2	Rmscv	Cvcv	F	SE	R^2_p
MLR	1	pIC50 = 0.010G(O..O) (±0.003) - 0.376nPhX (±0.058) + 0.265DipY(±0.072) -1.574GATS7v (±0.258) + 1.076MATS2e (±0.362)+ 0.205nROR (±0.063) + 0.997MATS7e(±0..401)+7.562 (±0.488)	27	0.917	0.76	0.159	2.78	25.0	0.12	0.70
FA-MLR	2	pIC50 = 2.152MATS7v(±0.537) + 0.230DipY(±0.083) + 0.244nROR (±0.048) + 0.020Ss (±0.003) +3.538 (±0.204)	27	0.895	0.81	0.197	3.29	19.8	0.22	0.69
PCRA	3	pIC50 = 0.240 FAC1 (±0.048) + 0.139 FAC2 (±0.048) + 0.114 FAC4 (±0.048) + 0.117 FAC7 (±0.048) + 0.103 FAC9 (±0.048) + 4.969 (±0.047)	27	0.906	0.87	0.168	3.24	19.8	0.27	0.71
GA-PLS	4	pIC50 = -20.126X3A (±7.555)+3.685MATS7v (±0.391)+ 2.655MATS5p (±0.471) + 0.319DipY(±0.053) + 0.230H-048 (±0.036) -1.084MATS6e (±0.304) -0.637 ASP (±0.234)+8.553 (±1.397)	27	0.943	0.82	0.148	2.99	31.7	0.09	0.87

[a]Number of molecules of training set used to derive the QSAR modelT

was selected as the best equation in the MLR model. The selected variables demonstrate that quantum (DipY), geometrical (G (O..O)), 2D autocorrelations (MATS2e, MATS7e, GATS7v), and functional (nPhX, nROR) descriptors affect the cytotoxic activity of the studied compounds.

A small difference between the conventional and cross-validate correlation coefficients of the different MLR equations reveals that none of the models are over fitted, which can be partially attributed to absence of collinearity between the variables in one hand and using of no extra variables in the other hand. The correlation coefficient (r^2) matrix for the descriptors used in MLR equation 1, shows that no significant correlation exists between pairs of descriptors (Table 3).

FA-MLR and PCRA

It was discovered that five factors could explain the data matrix to the extent of 96.3%, from the factor analysis on the data matrix

Table 3. Correlation coefficient (R^2) matrix for descriptors represented in multiple linear regression eqn 1.

	MATS2e	MATS7e	GATS7v	DipY	nROR	nPhX	G(O..O)
MATS2e	1	-0.160	0.217	0.315	-0.192	-0.207	-0.294
MATS7e		1	0.130	0.013	0.003	-0.048	0.205
GATS7v			1	0.088	0.209	0.090	0.075
DipY				1	-0.233	-0.088	-0.218
nROR					1	-0.134	0.227
nPhX						1	-0.159
G(O..O)							1

Table 4. Factor loadings of some significant descriptors after VARIMAX rotation.

Descriptor	factor1	factor2	Factor4	Factor7	Factor9	Communalities	PIC50
pIC50	0.596	0.476	0.389	0.218	-0.169	0.939	
Mp	-0.578	-0.397	-0.210	0.082	-0.049	0.993	-0.021
G(O..O)	0.679	0.445	-0.463	-0.022	-0.023	0.988	0.345
qpos	0.821	0.515	-0.175	0.083	-0.069	0.984	0.234
H-048	0.310	0.371	-0.552	-0.288	-0.128	0.956	0.561
H-052	0.544	-0.019	0.165	0.306	-0.016	0.932	0.508
nROR	0.259	0.364	-0.608	-0.270	-0.067	0.956	0.676
nPhX	-0.159	0.282	-0.063	0.611	0.342	0.970	0.441
X3A	-0.706	-0.386	-0.059	0.198	0.094	0.979	0.398
X3AV	-0.078	-0.757	0.026	-0.125	0.162	0.965	-0.243
lop	-0.482	-0.180	0.415	0.409	-0.243	0.983	-0.129
ATS8p	0.054	-0.527	-0.180	0.387	0.157	0.981	0.256
GATS6m	-0.581	-0.431	-0.133	0.052	0.001	0.940	0.436
GATS8m	-0.590	-0.537	-0.028	0.166	0.122	0.894	0.450
GATS5e	-0.605	0.355	0.037	-0.132	0.079	0.871	0.164
GATS8e	-0.817	-0.100	0.113	0.055	0.027	0.830	0.237
GATS4p	0.120	0.059	0.417	-0.698	-0.151	0.938	0.461
GATS7p	-0.371	0.156	-0.127	-0.118	-0.013	0.903	0.219
GATS4v	0.035	0.086	-0.111	-0.050	0.603	0.965	0.065
MATS5p	-0.145	0.209	-0.616	-0.124	0.086	0.881	-0.432
MATS7v	-0.815	-0.218	-0.018	0.313	0.115	0.944	0.712
MATS4m	-0.288	0.114	-0.008	0.053	-0.008	0.993	0.349
MATS6m	-0.352	-0.088	0.264	0.102	-0.454	0.816	0.415
MATS6e	0.576	-0.346	0.086	-0.040	0.116	0.954	-0.291
MATS7e	0.637	0.047	0.038	-0.098	-0.198	0.858	-0.123
ASP	-0.907	-0.012	-0.017	0.019	-0.072	0.971	-0.293
Ss	0.708	0.586	0.232	0.122	0.032	0.997	0.608
G(N..F)	-0.310	0.721	0.488	-0.046	0.258	0.992	0.341
MAXDP	0.551	0.435	0.145	0.134	.264	0.954	0.326
DipY	-0.027	-0.210	0.357	-0.144	-0.693	0.774	0.632

consisting of the pIC50 and calculated molecular descriptors. Table 4 shows that the biological activity is highly loaded with factors 2 and especially 1. The highest loading values for factor 2 are associated with X3AV, and G (N.F) descriptors whereas Ss, ASP, qpos, G(O.O), MATS7v, MATS7e, GATS5e and GATS8earethe highly loaded descriptors of factor 1. Table 4 revealed that, factors 1 and 2 are moderately loaded with cytotoxicity activity. Interestingly, the former possessed the highest loadings from geometrical (G(O..O), ASP), constitutional

(Ss), 2D autocorrelations (MATS7v, MATS7e, GATS5e, GATS8e) and charge (qpos) descriptors whereas the latter is containing the information from topological (X3Av) and geometrical (G (N.F)) descriptors. As it was shown in equation 6, the highly loaded descriptors of factors 1, 2, 4, 7and 9 (instead of applying the pool of all calculated descriptors) can be considered as the source of molecular descriptors for QSAR model building. So, the probability of obtaining chance models is decreased (45).

The subsequent FA-MLR equation using highly loaded descriptors is shown in Table 2, Eq.2.

PCRA

When factor scores were used as the predictor parameters in a multiple regression equation (instead of their highly loaded descriptors), a predictive QSAR model with factor scores number 1, 2, 4, 7and 9 as input variable was obtained (Eq. 3). This equation shows statistical quantities similar to those obtained by FA-MLR method (Table 2). However, it shows slightly higher calibration and lower cross-validation statistics with respect to Eq 2. This shows a sign of overfitting since the factors considered in Eq. 3 have information from irrelevant descriptors too. Considering this information in modeling may apparently increase the model variances (i.e., R^2) but they are not useful for prediction. On the other hand, the advantage of the QSAR model obtained by PCRA is that the factors appeared in the MLR equation 3 are orthogonal. The regression coefficients calculated for such variables are more stable and thus are closer to the real values. In addition, from the factor scores used, significance of the original variables for modeling the activity can be obtained. Factor score 1 indicates the importance of geometrical (G(O..O), ASP). constitutional (Ss), 2D autocorrelations (MATS7v, MATS7e, GATS5e, GATS8e) and charge (qpos) descriptors. The factor score 2 indicates importance of topological (X3Av) and geometrical (G (N.F)) descriptors, and factor score 4 signifies the importance of functional (nROR) and 2D autocorrelations (MATS5p) descriptors. The factor score 7reveals the importance of the 2D autocorrelations parameters (GATS4p) and functional (nPhX) descriptors. The factor score 9 signifies the importance of quantum (DipY) and 2D autocorrelations (GATS4v) descriptors.

GA-PLS

In PLS analysis, having decomposed the descriptors data matrix to orthogonal matrices, then the scores are constrained to have inner relationship with the dependent variables. Hence similar to PCRA, the multicollinearity problem in the descriptors is omitted by PLS analysis. Genetic algorithm was applied to find the more useful set of descriptors in PLS modeling. So, many different GA-PLS runs were done using different initial set of populations. The results of this model are summarized in Table 2.

As it is shown in Table 2 Eq 4, a combination of quantum (DipY)2 .D autocorrelations (MATS7v, MATS5p, MATS6e), atom- centered fragments (H-048), geometrical (ASP) and topological (X3A) descriptors have been selected by GA-PLS to account for the cytotoxic activity of N-arylphenyl-2, 2-dichloroacetamide analogues. The resulted GA-PLS model possessed very high statistical quality parameters (i.e., R^2 = 0.94and Q^2 = 0.82). The predictive ability of the model was measured by application to 7 external test set molecules. The squared correlation coefficient for prediction was 0.87, and standard error of prediction was 0.099.

Table 2 shows that none of the proposed QSAR models were obtained by chance and the GA-PLS model because of its greatest statistical parameters is the best predictive model.

The brief description of the descriptors used by QSAR models are summarized in Table 5.

In silico screening

In silico research in medicine is thought to have the potential to speed the rate of discovery, predicting and identifying new biologically active compounds while reducing the need for expensive lab work and clinical trials. One way to attaint his is by generating and screening drug candidates more effectively. On the other hand, the *in silico* procedure minimizes the time and cost associated with identifying new leads (46, 47).

A virtual screening was applied by deletion, insertion and substitution of different substitutes

Table 5. Definitions of molecular descriptors present in the models.

No.	Descriptors	Brief description
1	ATS8p	Broto-Moreau autocorrelation of a topological structure - lag 8 / weighted by atomic polarizabilities
2	MATS7v	Moran autocorrelation - lag 7 / weighted by atomic van der Waals volumes
3	MATS4m	Moran autocorrelation - lag 4 / weighted by atomic masses
4	MATS6m	Moran autocorrelation - lag 6 / weighted by atomic masses
5	MATS5p	Moran autocorrelation - lag 5 / weighted by atomic polarizabilities
6	MATS6e	Moran autocorrelation - lag 6 / weighted by atomic Sanderson electronegativities
7	MATS7e	Moran autocorrelation - lag 7 / weighted by atomic Sanderson electronegativities
8	GATS4v	Geary autocorrelation - lag 4 / weighted by atomic van der Waals volumes
9	GATS7v	Geary autocorrelation - lag 7 / weighted by atomic van der Waals volumes
10	GATS6m	Geary autocorrelation - lag 6 / weighted by atomic masses
11	GATS4p	Geary autocorrelation - lag 4 / weighted by atomic polarizabilities
12	GATS7p	Geary autocorrelation - lag 7 / weighted by atomic polarizabilities
13	GATS8e	Moran autocorrelation - lag 8 / weighted by atomic Sanderson electronegativities
14	X3A	average connectivity index chi-3
15	X3AV	average valence connectivity index chi-3
16	H-048	H attached to C2(sp3) / C1(sp2) / C0(sp)
17	H-052	H attached to C0(sp3) with 1X attached to next C
18	G(O..O)	sum of geometrical distances between O..O
19	Lop	Lopping centric index
20	nPhX	number of X-C on aromatic ring
21	nROR	number of ethers (aliphatic)
22	ASP	Asphericity
23	DMY(DipY)	Molecular dipole moment at Y-direction

on the parent molecules and the effects of the structural modifications on the biological activity were investigated. Then, the domain application of QSAR model was determined to apply the model for screening new compounds. The applicability domain (AD) of QSAR model was used to verify the prediction reliability, to identify the troublesome compounds and to predict the compounds with accep table activity that falls within this domain.

The important descriptors selected by GA-PLS model (because of its greatest statistical parameters compared to the others it was chosen as the best model) could be used for designing new active compounds. Analyzing the model applicability domain (AD) in the Williams plot (Figure 1) of the GA-PLS model based on the whole data set, appeared that none of the compounds were identified as an obvious outlier for the cytotoxic activity if the limit of normal values for the Y outliers (response outliers) was set as 2.5 times of the standard deviation units. As it is cleared, none of the compounds have leverage (h) values greater than the threshold leverages (h*). The warning leverage (h*), was found to be 0.89 for the developed QSAR model. The compounds that had a standardized residual more than three times of the standard deviation units were considered to be outliers. For both the training set and prediction set,

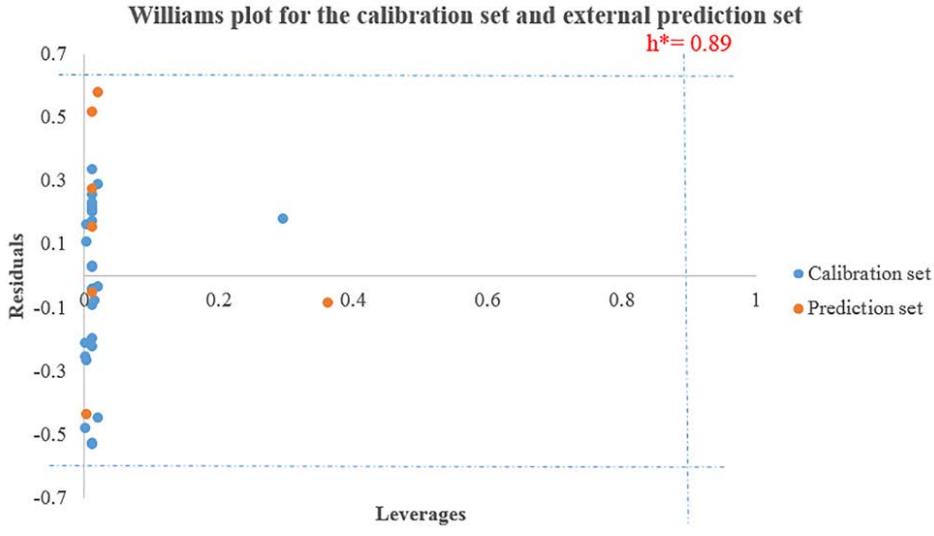


Figure 1. Williams plot for the training set and external prediction set for cytotoxic activity of N-arylphenyl-2,2-dichloroacetamide analogues.

Next, the *in silico* screening was used to the design of new compounds with potential cytotoxic activity according to the developed QSAR model and was validated by the developed GA-PLS model. So, the compounds in Tables 1 with IC_{50} <9.0μm were selected as template due to their good cytotoxic activity. Then, the in *silico* screen was applied by substituting different bioisosteric groups (O, S) in the place of-NH group; the results of this investigation are summarized in Table 6.

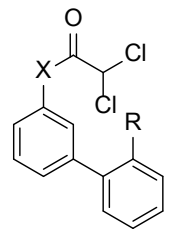
the presented model matches the high quality parameters with good fitting power and the capability of assessing external data. Moreover, almost all of the compounds were within the applicability domain of the proposed model and were evaluated accurately. While chemicals with a leverage value higher than h* were considered to be influential or high leverage chemicals (26, 34).

Next, the *in silico* screening was used to the design of new compounds with potential cytotoxic activity according to the developed QSAR model and was validated by the developed GA-PLS model. So, the compounds in Tables 1 with IC_{50} <9.0μm were selected as template due to their good cytotoxic activity. Then, the *in silico* screen was applied by substituting different bioisosteric groups (O, S) in the place of-NH group; the results of this investigation are summarized in Table 6.

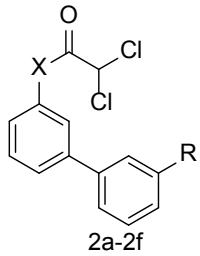
The model tolerated various bioisosteric changes of NH groups by sulfur and oxygen groups. Since all of the studied derivatives were within the applicability domain. Among different designated molecules, the compound 4c, 4g, 4i, 4j, 4k, 4m showed the best activity (pIC50 >5.25). Thus, in order to clarify the relation between the activities of the compounds with different functional groups, this compound was chosen for more structural modification. As it was shown in table 9, some esteric and thioesteric derivatives of this class of anticancer compounds have a good potentially for becoming anticancer agent. Finally, this result confirms the reliability of the models and it shows that with the aim of the QSAR model and use of *in silico* screening, it is possible to identify new synthetic compounds for drug discovery.

The proposed QSAR models have all conditions to be considered as predictive models. Firstly, all have correlation coefficient of cross-validation (Q^2) larger than 0.5 and of prediction (r^2) higher than 0.6. Thus, according to great statistics, GA-PLS can be considered as the most predictive model. According to the cross-validation results all models have Q^2> 0.7 and can be considered predictive models. To have a consideration on the cross-validated prediction results, the predicted activity data are plotted against the experimental activities in Figure 2. As it was mentioned in the article, the least scattering of data was obtained from GA-PLS.

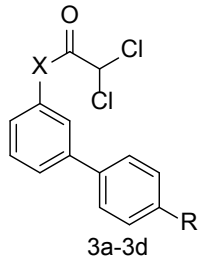
Table 6. Structural modification of N-arylphenyl-2, 2-dichloroacetamide analogues and their predicted activities and docking binding energy.



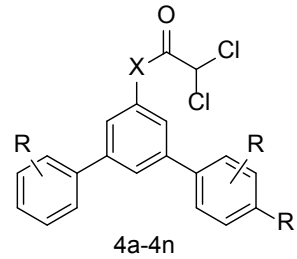
1a-1d



2a-2f



3a-3d



4a-4n

Name	R	X	pIC_{50} pred	leverage	Docking Binding Energy (kcal/mol)
11a	H	O	4.82834	0.010839	-5.93
11b	Furan-3-yl	O	4.97586	0.003205	-5.82
11c	H	S	4.74324	0.010721	-5.76
11d	Furan-3-yl	S	4.90973	0.002659	-6.05
12a	Me	O	4.74095	0.010605	-5.91
12b	F	O	4.97864	0.010118	-6.06
12c	CF3	O	5.14431	0.010435	-6.12
12d	Me	S	4.66754	0.012149	-5.89
12e	F	S	4.86423	0.022139	-5.91
12f	CF3	S	5.04231	0.002139	-6.01
13a	F	O	4.81908	0.010553	-6.15
13b	Cl	O	4.54678	0.010299	-5.79
13c	F	S	4.74390	0.011123	-5.58
13d	Cl	S	4.60043	0.010075	-5.83
14a	2-OEt	O	5.10686	0.013994	-6.01
14b	2-OEt	S	4.94502	0.004364	-6.26
14c	4-OEt	O	5.28557	0.069692	-6.53
14d	4-OEt	S	5.13035	0.019082	-6.24
14e	2-SMe	O	4.95452	0.010975	-5.76
14f	2-SMe	S	4.83654	0.010499	-5.69
14g	4-OiPr	O	5.28557	0.069692	-6.37
14h	4-OiPr	S	5.13026	0.019061	-6.24
14i	3,4-diO-CH2	O	5.87894	0.078891	-7.14
14j	3,4-diO-CH2	S	5.58176	0.218303	-6.97
14k	4-(Tetrahydro-2H-pyran-2-yl)oxy	O	5.97755	0.027392	-6.92
14l	4-(Tetrahydro-2H-pyran-2-yl)oxy	S	5.64426	0.025294	-7.02
14m	3-F-4-OMe	O	5.28666	0.070138	-6.62
14n	3-F-4-OMe	S	5.13081	0.019188	-6.09

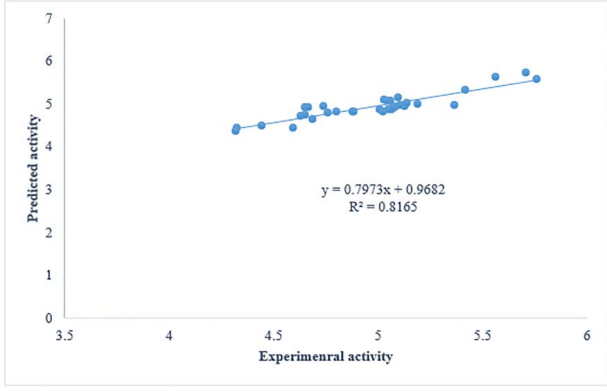


Figure 2. Plots of cross-validated predicted values of activity by GA-PLS against the experimental values

Docking Studies

Dockingisfrequentlyusedtopredictthebinding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule. Hence docking plays a great role in the rational design of drugs. DCA stimulates the activity of the enzyme PDH through inhibition of the enzyme PDKs. The crystal structure of PDK2 in complex with DCA has been acquired, and it shows that DCA indwells the pyruvate binding site in the N-terminal regulatory (R) domain (1).

Here, in this study docking studies were carried out on the compounds in Table 1 and 6 to find their binding site, binding modes and the best direction on the base of their binding energy. The docking simulations were carried out by means of an *in house* batch script (DOCKFACE) for automatic running of AutoDock 4.2 in a parallel mode using all system resources. Having completed the docking process, then the protein–ligand complex was analyzed to investigate the type of interactions. Top ranked binding energies (kcal/moL) in AutoDock dlg output file were considered as response in each run. Docking results were supported almost by high cluster populations. The conformation with the lowest binding energy was considered as the best docking result in each case.

As it was shown in figure 3 there is a good relationship between experimental pIC_{50} and docking binding energy. Hence, our docking protocol can discriminate between the ligand (active) and decoys (non-active). The validated

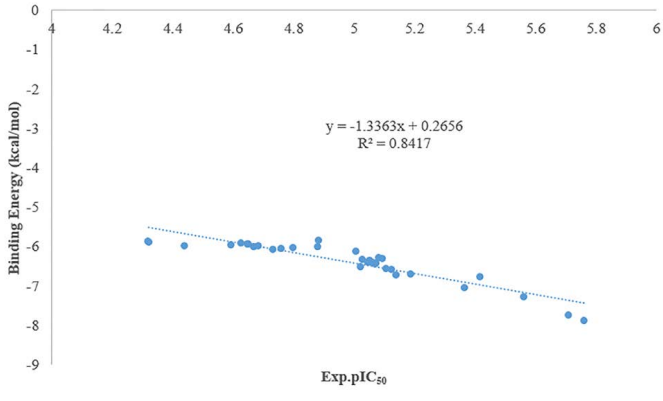


Figure 3. Plots of experimental pIC_{50} values versus docking binding energy.

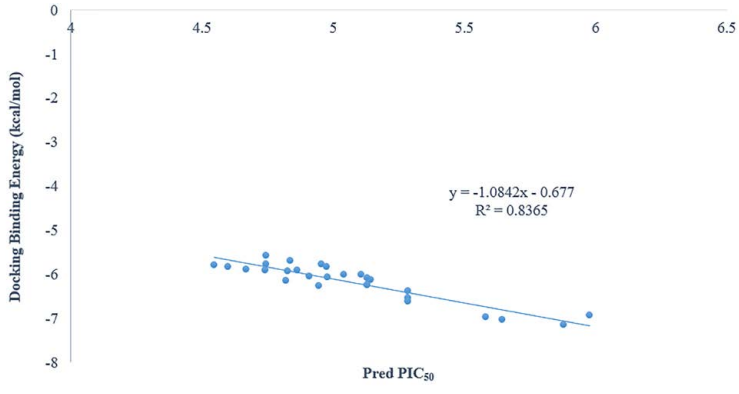


Figure 4. Plots of predicted pIC_{50} values versus docking binding energy.

docking procedure was also applied to our designed ligands. Figure 4, shows that this correlation was also existed between predicted pIC_{50} of QSAR studies and docking binding energy. Compounds 14i-m based on the best docking binding energy can be considered as good candidates for synthesis.

The results for each ligand were compared to its corresponding co-crystal ligand. Hydrogen bindings between docked potent agents such as 3g and the PDK receptor (2BU8) was analyzed using Autodock tools program (ADT, Version 1.5.6). ligplotv.4.5.3 (48) and Ligand Scout 3.12 (49). As it was shown in figure 5, a hydrogen bond acceptor interaction exists between oxygens of carboxyl group of co-crystal ligand (DCA) and Arg 154, Tyr 80 in receptor (figure 5A). Meanwhile, a hydrogen bond acceptor interaction existed between oxygen of methoxy group of 4d with Arg158, in receptor. There is also exists an arene-cation interaction between the phenyl group that bearing amide substituent with Arg158 and an arene-cation interaction between the phenyl group that bearing methoxy group in the receptor with Arg154 (Figure 5B).

Protein ligand interaction fingerprint (PLIF) studies could be used as a more reliable analysis technique (40). This method makes it possible to study the effect of different starting states of the structures on generated poses as well as their corresponding vector of contacts towards receptor during docking procedure. For this purpose, the docking of all 34 compounds of QSAR study as well as our designated compounds were carried out, then all generated poses of the ligands were subjected to Aupos SOM 2.1 to calculate their contact vectors within the receptor binding cavity. In this procedure, the contacts between the structures and the protein comprise of hydrophobic, hydrogen bonding and coulombic interactions. The resulted vectors of contacts are then analyzed using self-organizing map as implemented in Aupos SOM software. The output of self-organizing map is a classification pattern for ligands. For visualization of the results, the output files were subjected to Dendroscope 3.2.10. To the best of our knowledge, ligands in the same subgroup may show a similar behavior. As it was shown in figure 6, designated ligands such as 14g, 14h, 14k and 14l are clustered in the 5b (the best compound due to its greatest IC_{50}), 5e, 5f and 5i-o subgroup. Meanwhile, compounds 2e, 3a-d, 11d, 12c-f, 13a-c are clustered in the same subgroup. So these compounds may have a similar behavior as theirs and can be good candidates for synthesis.

Conclusion

In this study, four different QSAR modeling methods, MLR, FA-MLR, PCR and GA-PLS as well as FWA were used in the construction of a QSAR model for cytotoxic activity of N-arylphenyl-2, 2-dichloroacetamide analogues and the resulting models were compared. As it was shown in the article, having performed GA before the calibration, a regression model with

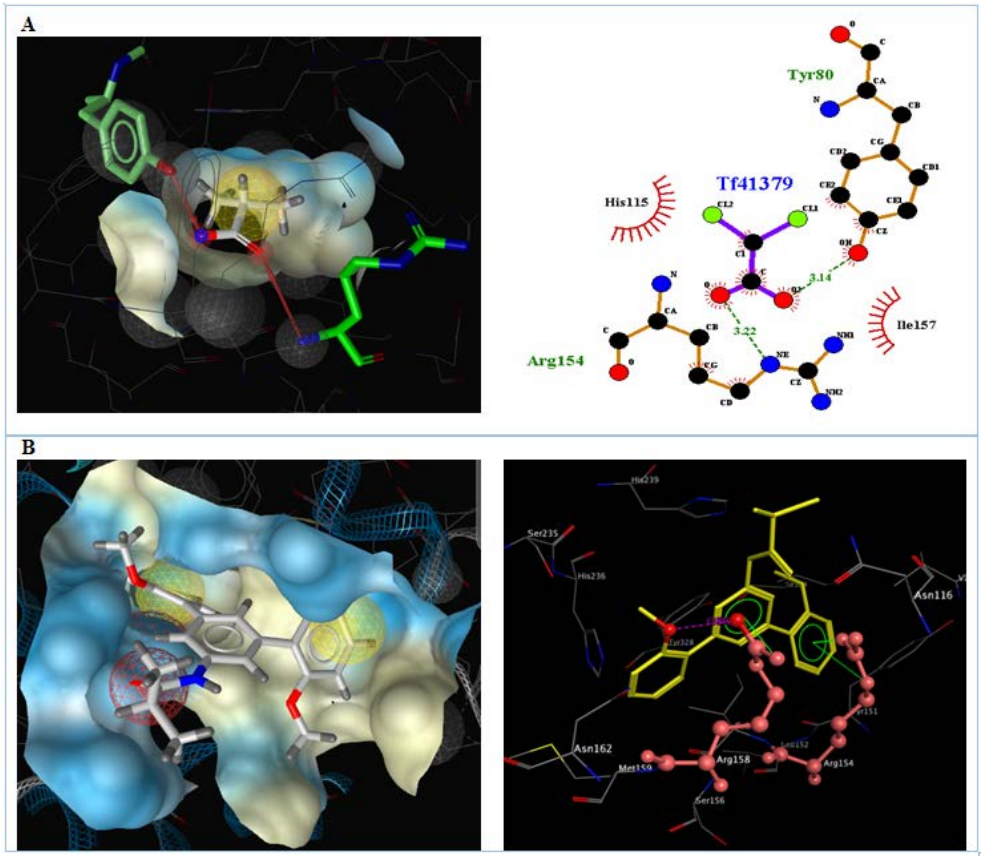


Figure 5. Interactions of A) DCA and B) compound 4d with the residues in the binding site of PDK (2BU8) receptor.

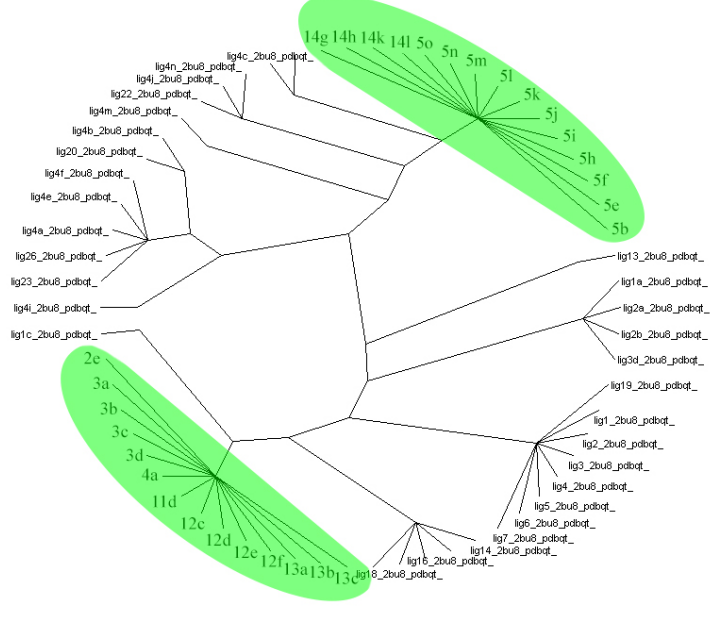


Figure 6. AuposSOM results for poses of docking.

enhanced predictive power would be obtained. The reliability, accuracy and predictability of the proposed models were evaluated by various criteria, including cross-validation, the root mean square error of prediction (RMSEP), root mean square error of cross-validation (RMSECV), validation through and Y-randomization. It was also shown that the proposed model is a useful aid for reduction of the time and cost of synthesis and biological evaluation of DCA analogues. Moreover, the results confirm that among the applied models, the GA-PLS is superior for prediction of the pIC50 of DCA analogues. The statistical parameters of the four different chemometrics methods used in this study are represented in Table 2. Some models represent high goodness of fit (measured by R^2), whereas that obtained by GA-PLS is significantly better than that of the other models. To the best of our knowledge, GA-PLS is the best choice for the prediction purpose of QSAR study, and for descriptive purpose it should be better to use MLR method. The cross-validation statistics reported in Table 2 suggest the higher prediction ability of the GA-PLS model. This can be ascribed to the exploit of a large number of descriptors by GA-PLS in compared to the MLR. The study suggests the importance of dipole moment in y-direction (DMY), 2D autocorrelations and a sphericity (ASP) of molecules for DCA analogues cytotoxic activity. It is clearly understood that 2D autocorrelation descriptors such as MATS7v, MATS6e, MATS5p, geometrical descriptors such as ASP, atom- centered fragments likeH-048, topological descriptors like X3A and quantum chemical parameter (DMY) are important structural parameters that significantly influence the cytotoxic activity. The 2D autocorrelation descriptors depict the topological structure of the compounds, but are more complicated in nature with respect to the classical topological descriptors. The calculation of these descriptors includes the summations of different autocorrelation functions corresponding to different structural lags and leads to different autocorrelation vectors corresponding to the lengths of the substructural fragments. As a result, these descriptors address the topology of the structure or parts thereof in association with a specific physicochemical property. According to the developed QSAR model, *in silico* screening was applied and new compounds such as 4c, 4g, 4i, 4j, 4k, and 4m with potential cytotoxic activity were suggested for synthesis.

There was a good correlation between docking binding energy and experimental pIC_{50}. The molecular docking study revealed that there is an arene-arene interaction between phenyl group the phenyl group that bearing amide substituent with Arg158 and an arene-cation interaction between the phenyl group that bearing substituents with Arg154 in the receptor. As it was shown in figure 4, based on the substituent on phenyl group, a hydrogen bond acceptor interaction also existed with the substituent and Arg158 in receptor. The docking results were also subjected to PLIF studies and compounds 11d, 12c-f, 13a-c, 14g, 14h, 14k and 14l are introduced as a good candidates for synthesis.

Acknowledgment

The authors would like to thank department of medicinal chemistry at school of pharmacy, Shiraz University of Medical Sciences for its kind contribution in providing the needed facilities for this work.

References

(1) Papandreou I, Goliasova T and Denko NC. Anticancer drugs that target metabolism: Is dichloroacetate the new paradigm? *In.t J. Cancer*. (2011) 128: 1001-8.

(2) Stockwin LH, Yu SX, Borgel S, Hancock C, Wolfe TL, Phillips LR, Hollingshead MG. and Newton DL. Sodium dichloroacetate selectively targets cells with defects in the mitochondrial ETC. *Int. J. Cancer*. (2010) 127: 2510-9.

(3) Galluzzi L, Morselli E, Kepp O, Vitale I, Rigoni A, Vacchelli E, Michaud M, Zischka H, Castedo M and Kroemer G. Mitochondrial gateways to cancer. *Mol. Aspects. Med*. (2010) 31: 1-20.

(4) Hans HK, Taeho K, Euiyong K, Ji Kyoung P, Seok-Ju P, Hyun J and Han JK. The Mitochondrial Warburg Effect: A Cancer Enigma. *IBC*. (2009) 1:7.

(5) Kankotia S and Stacpoole PW. Dichloroacetate and cancer: New home for an orphan drug? *Biochim. Biophys. Acta*. (2014) 1846: 617-29.

(6) Stacpoole PW, Wright EC, Baumgartner TG, Bersin RM, Buchalter S, Curry SH, Duncan CA, Harman EM, Henderson GN, Jenkinson S and Lachin JM. A controlled clinical trial of dichloroacetate for treatment of lactic acidosis in adults. *N Engl. J. Med.* (1992) 327: 1564-9.

(7) Abdelmalak M, Lew A, Ramezani R, Shroads AL, Coats BS, Langaee T, Shankar MN, Neiberger RE, Subramony SH and Stacpoole PW. Long-term safety of dichloroacetate in congenital lactic acidosis. *Mol. Gen. Metab*. (2013) 109: 139-43.

(8) Bonnet S, Archer SL, Allalunis-Turner J, Haromy A, Beaulieu C, Thompson R, Lee CT, Lopaschuk GD, Puttagunta L, Bonnet S and Harry GA. A Mitochondria-K+ Channel Axis Is Suppressed in Cancer and Its Normalization Promotes Apoptosis and Inhibits Cancer Growth. *Cancer. Cell* (2007) 11: 37-51.

(9) Wong JY, Huggins GS, Debidda M, Munshi NC and De Vivo I. Dichloroacetate induces apoptosis in endometrial cancer cells. *Gynecol. Oncol*. (2008) 109: 394-402.

(10) Cao W, Yacoub S, Shiverick KT, Namiki K, Sakai Y, Porvasnik S, Urbanek C and Rosser CJ. Dichloroacetate (DCA) sensitizes both wild-type and over expressing Bcl-2 prostate cancer cells *in-vitro* to radiation. *The*

(11) *Prostate*. (2008) 68: 1223-31.
Heshe D, Hoogestraat S, Brauckmann C, Karst U, Boos J and Lanvers-Kaminsky C. Dichloroacetate metabolically targeted therapy defeats cytotoxicity of standard anticancer drugs. *Cancer Chemother. Pharmacol*. (2011) 67: 647-55.

(12) Chen Y, Cairns R, Papandreou I, Koong A and Denko NC. Oxygen consumption can regulate the growth of tumors, a new perspective on the Warburg effect. *PloS One*. (2009) 4: e7033.

(13) Anderson KM, Jajeh J, Guinan P and Rubenstein M. *In-vitro* effects of dichloroacetate and CO2 on hypoxic HeLa cells. *Anticancer Res*. (2009) 29: 4579-88.

(14) Shahrzad S, Lacombe K, Adamcic U, Minhas K and Coomber BL. Sodium dichloroacetate (DCA) reduces apoptosis in colorectal tumor hypoxia. *Cancer Lett*. (2010) 297: 75-83.

(15) Khoshneviszadeh M, Edraki N, Miri R and Hemmateenejad B. Exploring QSAR for substituted 2-sulfonyl-phenyl-indol derivatives as potent and selective COX-2 inhibitors using different chemometrics tools. *Chem. Biol. Drug Des*. (2008) 72: 564-74.

(16) Kakkar R. Structure-based design of PDHK2 inhibitors from docking studies. *Int. Res. J Pharm*. (2011) 1: 51-9.

(17) Subramanian K and Ramaian AS. Development of a less toxic dichloroacetate analogue by docking and descriptor analysis. *Bioinformation* (2010) 5: 73-6.

(18) Li T, Yang Y, Cheng C, Tiwari AK, Sodani K, Zhao Y, Abraham I and Chen ZS. Design, synthesis and biological evaluation of N-arylphenyl-2,2-dichloroacetamide analogues as anti-cancer agents. *Bioorg. Med. Chem. Lett*. (2012) 22: 7268-71.

(19) Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JJA, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H and Hada M. Gaussian 09. Wallingford, CT, USA: Gaussian, Inc.; 2009.

(20) Mauri A, Consonni V, Pavan M and Todeschini R. DRAGON Software: An Easy Approach to Molecular Descriptor Calculations. *MATCH Commun. Math. Comput. Chem*. (2006) 56: 237-48.

(21) Thanikaivelan P, Subramanian V, Raghava Rao J and Unni Nair B. Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chem. Phys. Lett.* (2000) 323: 59-70.

(22) Morris GM, Huey R and Olson AJ. Using AutoDock for Ligand-Receptor Docking. *Curr. Protoc. Bioinformatics* (2008) 8: 8-14.

(23) Protein Data Bank, http://www.rcsb.org/pdb/home/home.do (2015).

(24) Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U and Sali A. Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinformatics*: John Wiley & Sons, Inc.; (2002).

(25) Sagrado S and Cronin MTD. Application of the modelling power approach to variable subset selection for GA-PLS QSAR models. *Anal. Chim. Acta*. (2008) 609: 169-74.

(26) Khoshneviszadeh M, Edraki N, Miri R, Foroumadi A and Hemmateenejad B. QSAR study of 4-aryl-4H-chromenes as a new series of apoptosis inducers using different chemometric tools. *Chem. Biol. Drug Des*. (2012) 79: 442-58.

(27) Leonard JT and Roy K. QSAR by LFER model of HIV protease inhibitor mannitol derivatives using FA-MLR, PCRA, and PLS techniques. *Bioorg. Med. Chem*. (2006) 14: 1039-46.

(28) S Ramos LS, Beebe KR, Carey WP, Sanchez ME, Erickson BC, Wilson BE, Wangen LE and Kowalski BR. Chemometrics. *Anal. Chem*. (1986) 58: 294R-315R.

(29) Cho SJ and Hermsmeier MA. Genetic Algorithm Guided Selection: Variable Selection and Subset Selection. *J. Chem. Inf. Model*. (2002) 42: 927-36.

(30) Leardi R. Genetic algorithms in chemometrics and chemistry. *J. Chemom*. (2001)15: 559-69.

(31) Fan Y, Shi LM, Kohn KW, Pommier Y and Weinstein JN. Quantitative structure-antitumor activity relationships of camptothecin analogues: Cluster analysis and genetic algorithm-based studies. *J. Med. Chem*. (2001) 44: 3254-63.

(32) Edraki N, Hemmateenejad B, Miri R and Khoshneviszade M. QSAR study of phenoxypyrimidine derivatives as potent inhibitors of p38 kinase using different chemometric tools. *Chem. Biol. Drug Des*. (2007) 70: 530-9.

(33) Miri R, Javidnia K, Mirkhani H, Hemmateenejad B, Sepeher Z, Zalpour M, Behzad T, Khoshneviszadeh M, Edraki N and Mehdipour AR. Synthesis, QSAR and calcium channel modulator activity of new hexahydroquinoline derivatives containing nitroimidazole. *Chem. Biol. Drug Des*. (2007) 70: 329-

36.
(34) Asadollahi T, Dadfarnia S, Shabani AMH, Ghasemi JB and Sarkhosh M. QSAR models for CXCR2 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the PLS linear regression method and design of the new compounds using *in silico* virtual screening. *Molecules* (2011) 16: 1928-55.
(35) Tropsha A, Gramatica P and Gombar VK. The Importance of being Earnest: Validation is the Absolute Essential for successful application and interpretation of QSPR models. *QSAR Comb. Sci*. (2003) 22: 69-77.
(36) Roy K, Kar S and Das RN. Chapter 7 - Validation of QSAR Models. Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment. Academic press (2015) 231-89.
(37) Weaver S and Gleeson MP. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics. Modell*. (2008) 26: 1315-26.
(38) Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, White SWand Lee RE. Validation of molecular docking programs for virtual screening against dihydropteroate synthase. *J. Chem Inf. Model*. (2009) 49: 444-60.
(39) Sakhteman A. PreAuposSOM, https://www.biomedicale.univ-paris5.fr/aupossom/ (2015).
(40) Mantsyzov AB, Bouvier G, Evrard-Todeschi N and Bertho G. Contact-based ligand-clustering approach for the identification of active compounds in virtual screening. *Adv. Appl. Bioinform. Chem*. (2012) 5: 61-79.
(41) Bouvier G, Evrard-Todeschi N, Girault JP and Bertho G. Automatic clustering ofdocking poses in virtual screening process using self-organizing map. *Bioinformatics* (2010) 26: 53-60.
(42) Faghih Z, Fereidoonnezhad M, Tabaei SMH, Rezaei Z and Zolghadr AR. The binding of small carbazole derivative (P7C3) to protofibrils of the Alzheimer's disease and β-secretase: Molecular dynamics simulation studies. *Chem. Phys*. (2015) 459: 31-9.
(43) Huson DH and Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol*. (2012) 61: 1061-7.
(44) Huson DH, Richter DC, Rausch C, Dezulian T, Franz M and Rupp R. Dendroscope: An interactive viewer for large phylogenetic trees. *BMC. Bioinformatics* (2007) 8: 460.
(45) Hemmateenejad B. Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based PCR. *J. Chemom*. (2004) 18: 475-85.
(46) Bauch C, Bevan S, Woodhouse H, Dilworth C and Walker P. Predicting *in-vivo* phospholipidosis-inducing potential of drugs by a combined high content screening and *in silico* modelling approach. *Toxicol. In-Vitro* (2015) 29: 621-30.
(47) Murakami Y, Hayakawa M, Yano Y, Tanahashi T, Enomoto M, Tamori A, Kawada N, Iwadate M and Umeyama H. Discovering novel direct acting antiviral agents for HBV using in silico screening. *Biochem. Biophys. Res. Commun*. (2015) 456: 20-8.
(48) Wallace AC, Laskowski RA and Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*. (1995) 8: 127-34.
(49) Wolber G and Langer T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model*. (2005) 45: 160-9.

This article is available online at http://www.ijpr.ir