4D-QSAR Analyses for EGFR Inhibitors Based on CDDA-OPS-GA Method

Biying Cai, Tiansheng Zhao, Daogang Qin and Guogang Tu*
School of Pharmaceutical Science, Nanchang University, 330006, China.
tugg110@163.com*

(Received on 18th September 2023, accepted in revised form 23rd August 2024)

Summary: Epidermal growth factor receptor (EGFR) is a preferred target for treating cancer. Compared to 3D-QSAR, 4D-QSAR has the feature of conformational flexibility and free alignment for individual ligands. In present studies, the 4D-QSAR of 131 analogs of 4-anilino quinazoline for EGFR inhibitors was built. The GROMACS package was employed to yield the conformational ensemble profile. The field descriptors of Coulomb and Lennard−Jones potentials were calculated by LQTA-QSAR (Laboratory of Theoretical and Applied Chemometrics, QSAR). The filter descriptors and variable selection is very important, which was performed by means of comparative distribution detection algorithm (CDDA), ordered predictors selection (OPS) and genetic algorithm (GA) method. Best 4D-QSAR model yielded satisfactory statistics ($R^2 = 0.71$), good performance in internal ($Q^2_{Loo} = 0.60$) and external prediction ($R^2_{pred} = 0.69$, $k = 0.97$, $k' = 1.01$). The 4D-QSAR was shown to be robust ($Q^2_{LMO} = 0.59$) and was not built by chance ($R^2_{YS} = 0.17$, $Q^2_{YS} = -0.25$). The model has a good potential for rational design new EGFR inhibitors.

Keywords: 4D-QSAR, comparative distribution detection algorithm (CDDA), ordered predictors selection (OPS), genetic algorithm (GA), EGFR Inhibitors

Introduction

Epidermal growth factor receptor (EGFR) is the prototype of receptor tyrosine kinases (TKs) family [1]. As a transmembrane glycoprotein, several signal transduction cascades are initiated, and lead to DNA synthesis and cell proliferation when EGFR is activated [2]. So EGFR plays an important role in the regulation of several cellular functions such as survival, cell growth, differentiation, proliferation, and apoptosis [3]. The mutation or amplification of EGFR was found in solid tumors, such as lung cancer, glioma, breast cancer, and ovarian cancer. Currently, EGFR is a potential target for cancer therapy [4-7]. The EGFR inhibitor, lapatinib, is approved for the treatment of breast cancer by the FDA [8]. Other EGFR inhibitors, such as lomustine, temozolomide, gefitinib, and erlotinib, have also been approved by the FDA for the treatment of glioma [9]. However, in many cancer types as breast cancer, hepatocellular carcinoma, pancreatic cancer, non-small cell lung cancer, colorectal carcinoma, glioblastoma, and melanoma, there are significant resistance to the used EGFR inhibitors [10]. All these findings make requires to design and synthesize new potent EGFR inhibitors.

Before the synthesis, it is necessary to develop a prediction method for biological activities. Quantitative structure-activity relationship (QSAR) is an important part for modern drug design, including risk assessment, drug discovery and predictive toxicology [11, 12]. Initially, two-dimensional quantitative structure-activity relationship (2D-QSAR) and three-dimensional quantitative structure activity relationship (3D-QSAR) were extensively explored in medicinal chemistry study. However, the major constraint of 3D-QSAR is its dependency and sensitivity to conformations and alignments of compounds [13], because only one conformation, not a conformational ensemble profile, is considered for each compound [14]. To overcome inherent constraint of 3D-QSAR, the four-dimensional quantitative structure activity relationship (4D-QSAR) was originally developed which includes the freedom of alignment and the conformational flexibility to build 3D-QSAR by performing molecular state ensemble averaging, i.e., the fourth "dimension" [15].

The LQTA-QSAR (*Laboratório de Quimiometria Teórica e Aplicada*), a 4D-QSAR approach, often calculates a large number of descriptors (variable), frequently several thousands

*To whom all correspondence should be addressed.

[16]. Hence, the variable selection is very important when generating 4D-QSAR model. The comparative distribution detection algorithm (CDDA) can classify descriptors according to distribution profile, which compares individual distributions of a descriptor with dependent variables, and computes dissimilarity statistics [17]. CDDA enables numerical inspection of bivariate scatter plots and helps in filtering (selection) of descriptors which is suitable to establish 4D-QSAR model. The ordered predictors selection (OPS) is able to get an informative vector that contains information about the location of the best response variables for prediction. The OPS was shown to avoid overfitting and chance correlation, and be robust for QSAR model [18]. The genetic algorithm (GA), a stochastic method, enables to solve optimization problems of fitness criteria, which applies different genetic functions and evolution hypothesis of Darwin, i.e. mutation and crossover [19].

In present work, we aimed to build a 4D-QSAR model of the EGFR inhibitors by means of CDDA-OPS-GA method for descriptors selection. The regression methods used were multiple linear regression (MLR). It is the first to report the 4D-QSAR model of 4-anilino quinazoline derivatives as EGFR inhibitors.

Experimental

Data set

All inhibitors of EGFR were taken from literature [20-23]. In order to provide numerically larger data values, the biological activities expressed as IC_{50} values in units of molarity were transformed to pIC_{50} ($-\log IC_{50}$) which were used as dependent variables. All of the compounds were divided into the training set of 105 compounds and the test set of 26 compounds taking into account both the distribution of dependent variables and the structural diversity. The training set was used to construct 4D-QSAR model, and the test set was used to evaluate the predictive quality. The chemical structures and IC_{50} values of the data set are presented in Fig. 1 and Tab. S1.

4D-QSAR study

The 3D structures of all of the compounds were built by means of Ghemical program [24]. The structures were optimized with the ffG43a1 force field. Then partial atomic charges of AM1-BCC method were computed with AMBER ff03 atom types using UCSF Chimera [25]. The topology files of compounds were obtained by the topobuild program. In order to obtain conformational ensemble profile (CEP), the molecular dynamics (MD) simulations of all compounds were performed by the GROMACS software (version 4.5.4) [26]. All compounds were put in the dodecahedron box which filled with water molecules. Long-range electrostatics and van der Waals interaction energies were computed by means of Particle Mesh Ewald method with a cut off radius of 10 Å [27]. System temperature was controlled by Berendsen thermostat coupling, and pressure was kept by Parrinelloe Rahman coupling [28]. System was optimized by the steepest descent and conjugated gradient method. Using the script of LQTAgrid software [29], the stepwise heating method was run which included heating the system at 50 K, 100 K, 200 K and 350 K for 20 ps in 1 fs step size. The system was then backed to 300 K for 500 ps. The trajectory file was recorded every 10 ps simulation time.

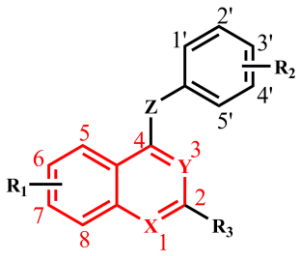


Fig. 1: The structure of data set and red atoms used for alignment.

The compound 47 was chosen as the reference of alignment due to the most active compound among all compounds. All conformations generated in MD simulations at 300 K were superimposed to the reference using the index number of common atoms. The atoms, which were selected for alignment, are shown in Fig. 1. During the alignment, the initial conformer generated at 20 ps was selected, then other trajectories, which were generated up to 100 ps times with 2 ps increment, were subjected to alignment using the least squares method to compute the minimum root mean- square of the distances (RMSD). The aligned CEPs of the most active compound 47 (reference) and alignment with CEPs of least active compound 126 are shown in Fig. 2.

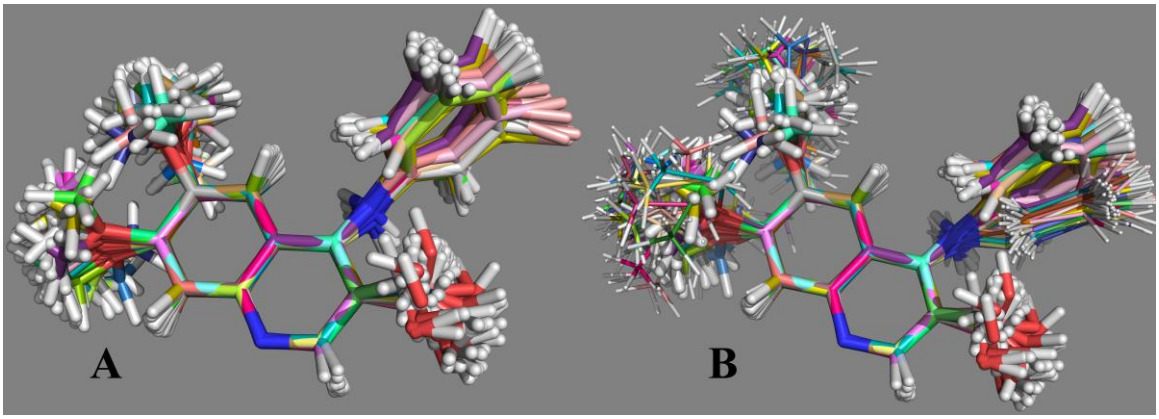


Fig. 2: The aligned CEPs generated in MD simulations. A): aligned CEPs of the least active compound 126 (represented by licorice), B): alignment of the most active compound 47 (represented by line) with compound 126.

The grid box was defined as 20 × 19 × 17 Å which was large enough to accommodate all conformers. The aligned molecules were submitted to the LQTAgrid program to calculate the energy descriptors of intermolecular interaction every grid point of a 1 Å grid cell lattice. The NH_3^+ probe was selected to simulated N-terminus moiety of protein. The Coulomb interaction descriptors (C descriptors) and Lennard-Jones potential descriptors (LJ descriptors) were calculated. The dimension of the descriptor matrix was 131 × 15120, where each row is a compound and each column is a descriptor.

$$LJ'_{x,y,z} = \begin{cases} LJ_{x,y,z} & LJ_{x,y,z} \le 30 \\ 30 + \log_{10}(LJ_{x,y,z} - 29) & LJ_{x,y,z} > 30 \end{cases} \quad (1)$$

$$C'_{x,y,z} = \begin{cases} C_{x,y,z} & |C_{x,y,z}| \le 30 \\ -(30 + \log_{10}(-C_{x,y,z} - 29)) & C_{x,y,z} < -30 \\ 30 + \log_{10}(C_{x,y,z} - 29) & C_{x,y,z} > 30 \end{cases} \quad (2)$$

Fig. 3: The equation for truncating both LJ and C descriptors.

Descriptor selection and model construction

First, it is necessary to truncate both LJ and C descriptors, in order to avoid large values with high orders of magnitude, and to keep information in the region close to the compounds [30]. When the distance between the atoms of compound and probe is close to zero, interaction energy generates a large value which do not benefit to the model. Based on equation (1) and (2) (Fig. 3), if the absolute value of interaction energy was more than 30 kcal/mol, the logarithmic value of residual was added to 30 kcal/mol.

Second, if the variance of descriptors is below of 0.01, then the descriptors are excluded. Because the descriptors are far from compounds, and contains very little information.

Third, the pearson correlation coefficients between descriptors and dependent variables were calculated (r vector) using correlation coefficient cut-off according to the equation (3) [17], where $Z_{0.99}$ is the number of standard deviations equal to 2.33, extending from the mean of normal distribution ($\mu = 0$) required to contain 99% of the area, and σ is the standard deviation of r_{rand}. When the absolute value of $|r|_{cut\text{-}off}$ was lower than 0.3, the descriptors were excluded. This method can eliminate most of noise.

$$|r|_{cut\text{-}off} = Z_{0.99}\sigma \quad (3)$$

Forth, the CDDA was performed to exclude descriptors whose distribution is inconsistent with dependent variables [31]. The descriptors were sorted in descending order according to their absolute value of correlation coefficients. The hyperparameter m (0.05 - 1, step 0.01) was applied to adjust the number of descriptors which were used to build 4D-QSAR model.

Fifth, the OPS method attaches importance to each descriptor based on a vector; then the matrix of descriptors is rearranged according to their relevancy. The most relevant/important descriptors are

represented by the first column of the matrix [31]. Successive partial least squares (PLS) regressions are performed by increasing the descriptors number in order to select the set that build the best latent variables (LVs) for correlation with the endpoints [32]. The process is repeated in an iterative manner

Finally, GA in QSRINS package [33], which is a software for the development and validation of multiple linear regression (MLR) model, was applied to choose the most appropriate descriptors for model. The GA performs its optimization make use of variation and selection via the evaluation of the fitness function. GA is a stochastic technique well suited to the problem of variable selection and optimization, and is proved to be effective as a variable selection method.

Model evaluation

The internal validation of 4D-QSAR model was performed to establish robustness and internal stability. For internal validation, leave-one-out cross-validation (Q^2_{LOO}) is the most preferred technique in which each compound of the training set was removed once from the dataset, and the biological activity of removed compound was predicted from the model. Leave-many-out cross-validation (Q^2_{LMO}) method was also used which carried out for 30% of data out of training each run [34]. In order to check the chance correlation, Y-Scrambling testing was performed in which the dependent variable vector, Y-vector, is randomly shuffled many times, then a new QSAR model is built making use of the original independent variable matrix and the R^2_{YS} and Q^2_{YS} values are calculated each time.

The external validation was used to evaluate the predictive accuracy. The model equation, built using the training set compounds, was applied to test set compounds, and the biological activity of test set compounds was computed. The predictive accuracy is checked in terms of the RMSE, MAE, S, CCC, PRESS, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, Δr^2_m, $\overline{r}^2_m$ and Golbraikh-Tropsha statistics which calculates the slopes (i.e. k and k′) of the regression lines of the external validation [35, 36].

Results and Discussion

Analysis of MD trajectories

Following successful simulation, the MD trajectories were investigated for dynamic behavior. The last trajectories, obtained for the 500 ps simulations, were analyzed. For some less active compounds (compounds 73, 110, 125, 126, 130 and 131) and more active compounds (compounds 12, 36, 47, 48, 61 and 67), the root mean square deviations (RMSD) values, which were calculated from the reference trajectories obtained at the end of simulations of 350 K with the trajectories obtained during simulations of 300 K, stayed within 0.20 nm range (Fig. 4). The RMSD fluctuation of compound 125 is greater than that of other compound, due to compound 125 containing the flexible ester group at position 3. The conformations of target compounds did not drastically change during the MD simulations. This indicates that an equilibrium state of target compounds was reached characterized by the RMSD profile.

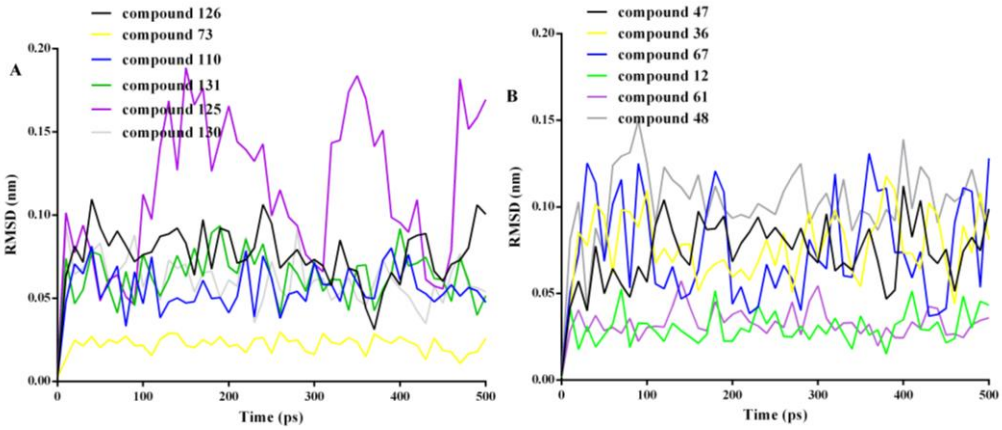


Fig. 4: Graphs showing RMSD value. A): RMSD value of less active compounds. B): RMSD value of more active compounds.

4D-QSAR model

The 4D-QSAR model with 18 variables was shown to be the best model by the leave-one-out cross-validation, which resulted in $R^2 = 0.71$ and $Q^2_{LOO} = 0.60$ and $R^2_{Pred} = 0.69$. The details of the 4D-QSAR model statistics are shown in Tab. 1. R^2 = correlation coefficient, R^2_{ad} = adjusted correlation coefficient, $RMSE_{tr}$ = root mean square error of training set, RMSEcv = root mean square error of cross-validation, $RMSE_{ext}$ = root mean square error external, MAE_{tr} = mean absolute error of training set, MAE_{cv} = mean absolute error of cross-validation, MAE_{ext} = mean absolute error external, CCC_{tr} = concordance correlation coefficient of training set, CCC_{ext} = concordance correlation coefficient external, S = standard error of estimate, F = fischer ratio, Q^2_{loo} = square of the cross-validated correlation coefficient from leave-one-out, Q^2_{LMO} = square of the cross-validated correlation coefficient from leave many-out, $\overline{r}^2_m$ = average r^2_m, Δr^2_m = delta r^2_m, Q^2_{F1}, Q^2_{F2}, Q^2_{F3} = predictive squared correlation coefficient, k′, k = slope of regression line, $PRESS_{cv}$ = predictive residual sum of squares from cross-validation, $PRESS_{ext}$ = predictive residual sum of squares external, $R^2_{YS} = R^2$ of Y-Scrambling, $Q^2_{YS} = Q^2$ of Y-Scrambling, R^2_{Pred} = correlation coefficient of test set. It can be concluded that the model presents good predictive power, as the model satisfied the external and internal validation criteria: $R^2 > 0.6$, Q^2_{LMO} and $Q^2_{LOO} > 0.5$, CCC > 0.85, Q^2_{F1}, Q^2_{F2} and $Q^2_{F3} > 0.6$, $\overline{r}^2_m > 0.5$, $\Delta r^2_m < 0.2$ and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ [37]. The difference between R^2 and Q^2_{LOO} is 0.11 units, less than 0.2, indicating the absence of over-fitting [38]. The Q^2_{LMO} (0.59) is close to Q^2_{LOO} (0.60), the model is considered robust. The averages R^2_{YS} (0.17) and Q^2_{YS} (-0.25) are smaller than the values of the original model, so the 4D-QSAR was not built by chance. In addition, other validation parameters were within acceptable limits.

Table-1: The statistical parameters of 4D-QSAR model.

Statistical parameter	Model with 18 variables	Statistical parameter	Model with 18 variables
R^2	0.71	MAE_{cv}	0.84
R^2_{adj}	0.65	$PRESS_{cv}$	13.82
$RMSE_{tr}$	0.87	Q^2_{LMO}	0.59
MAE_{tr}	0.69	R^2_{YS}	0.17
CCC_{tr}	0.89	Q^2_{YS}	-0.25
S	0.96	$RMSE_{ext}$	0.84
F	11.84	MAE_{ext}	0.70
Q^2_{LOO}	0.60	$PRESS_{ext}$	8.47
$RMSE_{cv}$	1.04	R^2_{Pred}	0.69
$\overline{r}^2_m$	0.57	CCC_{ext}	0.86
Q^2_{F1}	0.67	Δr^2_m	0.16
Q^2_{F3}	0.73	Q^2_{F2}	0.67
k′	1.01	k	0.97

The model equation with 18 variables is given in Equation 4. In terms of least squares curve fitting method, the values of the regression coefficient were calculated. The variable with positive values is favorably contributing to the model, whereas the variable with negative coefficients is inversely contributing to the model. The plot for experimental pIC_{50} against predicted pIC_{50} is shown in Fig. 5.

$$pIC_{50} = 0.0167*(C1) - 0.0009*(C2) - 0.0114*(C3) - 0.0076*(C4) - 0.0051*(C5) + 0.0160*(LJ1) + 0.0125*(LJ2) + 0.0024*(LJ3) + 0.2405*(LJ4) + 0.0037*(LJ5) + 0.0147*(LJ6) + 1.4731*(LJ7) - 0.0113*(LJ8) - 0.0239*(LJ9) - 0.0096*(LJ10) - 0.0146*(LJ11) - 0.0241*(LJ12) - 0.0157*(LJ13) + 8.7873 \quad (4)$$

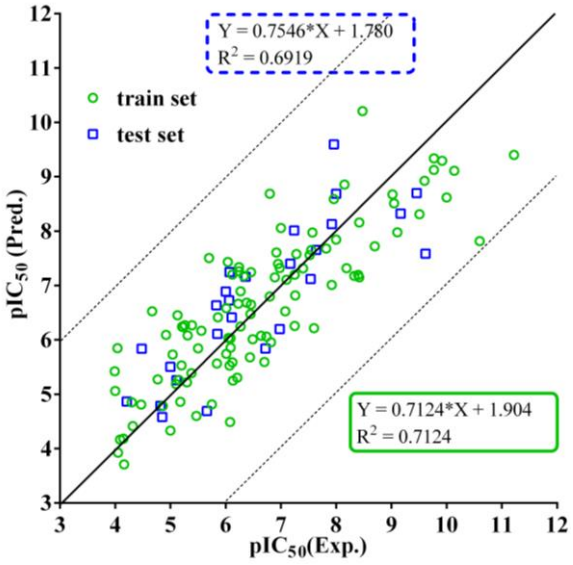


Fig. 5: Predicted pIC_{50} versus experimental pIC_{50} for training set and test set compounds.

The descriptors are designated as C1: X24_21_15_NH_3^+_C, C2: X21_27_11_NH_3^+_C, C3: X25_29_10_NH_3^+_C, C4: X29_21_17_NH_3^+_C, C5: X29_24_14_NH_3^+_C, LJ1: X21_22_16_NH_3^+_LJ, LJ2: X27_26_17_NH_3^+_LJ, LJ3: X28_18_15_NH_3^+_LJ, LJ4: X28_27_6_NH_3^+_LJ, LJ5: X29_18_13_NH_3^+_LJ, LJ6: X32_22_16_NH_3^+_LJ, LJ7: X34_29_10_NH_3^+_LJ, LJ8: X25_21_14_NH_3^+_LJ, LJ9: X28_29_9_NH_3^+_LJ, LJ10: X29_24_16_NH_3^+_LJ, LJ11: X29_26_13_NH_3^+_LJ,

LJ12: X31_21_19_NH_3^+_LJ, LJ13: X32_21_11_NH_3^+_LJ. Each descriptor denotes the specific interaction energy at the specific grid point. The C1 represents the Coulomb descriptor at the grid point 24, 21 and 15 along the X, Y and Z axes respectively. Similarly, the LJ1 represents the Lennard–Jones descriptor at the grid point 21, 22 and 16 along the X, Y and Z axes respectively. Compare with the R^2 and Q^2 of the original model, R^2_{YS} and Q^2_{YS} was lowest in the Y-scrambling test which implies no chance correlation in the model (Fig. 6).

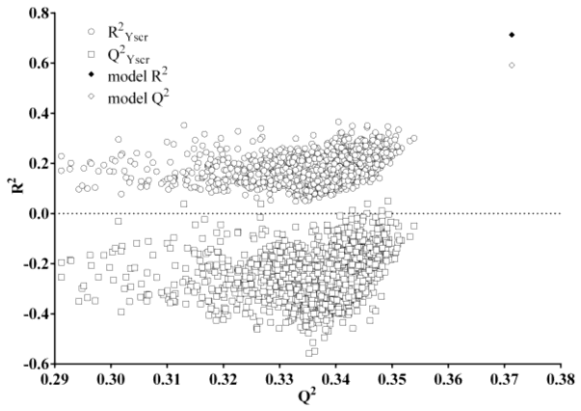


Fig. 6: Plot of Y-scrambled models compared with the original model.

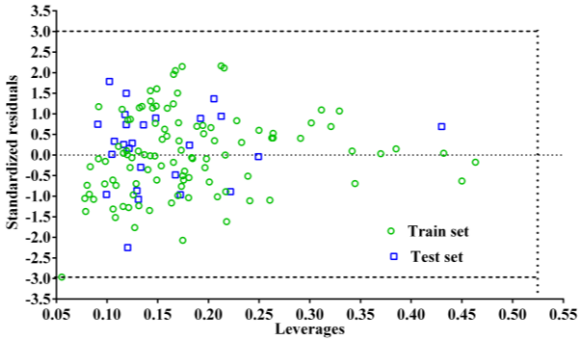


Fig. 7: Plot of leverages against standardized residuals. Dashed lines and dotted lines represent ±3.0 standardized residual and warning leverage (h* = 0.543).

Model applicability domain

The leverage method was used to verify the chemical applicability domain (AD) and the robustness of model. The leverage, h^*, for each molecule was calculated by this method. The warning leverage is generally fixed at $3LV/m$, where LV is the latent variable and m is the number of training set compounds. It can be seen from the AD analysis results presented in Fig. 7 that there are no outliers for all compounds. More importantly, the test compounds which were not applied to build model are predicted with similar accuracy of the training compounds.

Contour maps of 4D-QSAR

Graphical representations of the 4D-QSAR model are shown in Fig. 8. Blue regions are the electrostatic descriptors corresponding to negative regression coefficients and red regions are the electrostatic descriptors related to positive regression coefficients. Likewise, Yellow regions and green regions denote steric descriptors with negative and positive regression coefficients, respectively. The descriptors LJ7 and LJ11 at 3 position (the numbers of compounds as shown in Fig. 1) indicate sterically favorable and unfavorable region respectively for biological activity (such as compound 125). The descriptors LJ4 and LJ9 near the R_3 position also show sterically favorable and unfavorable region respectively (such as compound 52). The blue descriptor C3 near 1 position describes electron-withdrawing substituent is preferred, like compound 130, and 131. The descriptors C2 and LJ1 are related to R_1 group, which suggests bulky group at 6 position (such as compound 65 contain butoxyl group) and electron-withdrawing group at 7 position (such as compounds 15 – 19 containing nitro group) is preferred. The R_2 group at phenyl ring describes structural information related to conformational flexibility of substituent group. The descriptors C4 and LJ12 show small group with negative-charge increase the compound's bioactivity. The descriptors LJ5 and LJ3 near R_2 position of compound 111 containing isopropyl group indicate that bulky group is favorable. The descriptors LJ8 and C1 describe that a small group with positive-charge is preferred, such as compound 103. The descriptors LJ10 and C5 prefer small groups with negative-charge like compound 110. The descriptors LJ6 and LJ13 show sterically favorable and unfavorable region respectively, such as compounds 106 and 108, the former contains the 2'-Cl substituent and the latter contains the 2'-SMe substituent. The coefficient of LJ7 is the largest in the model equation, so the green descriptor LJ7 can be mostly connected with the activity.

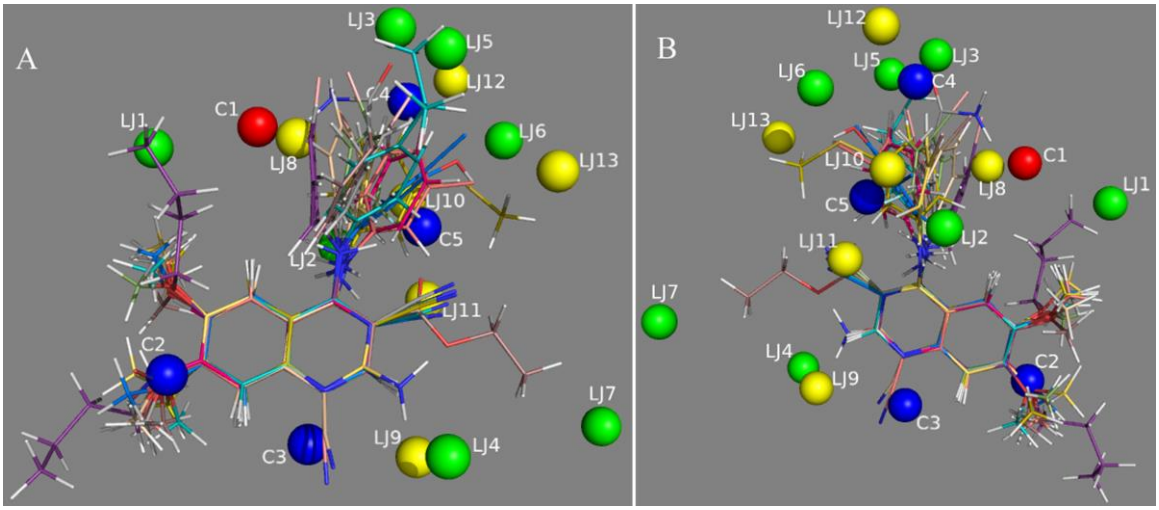


Fig. 8: Different perspectives of the steric and electrostatic contour maps of 4D-QSAR model.

Propose of new compounds

Based on these observations, 3 new compounds were designed, and the 4D-QSAR model generated was used to predict biological activity for each compound (shown in Tab. 2). We first consider -$O(CH_2)_2CH(CH_3)_2$ group as the substituent R_1, because its terminal CH_3 group just interacts with the green LJ1. Meanwhile, we also consider -$(CH_2)_3CH(CH_3)_2$ group as the substituent R_4, because its terminal CH_3 group can interact with the green LJ7. For R_5, the -SO_3H group was introduced, because -SO_3H has the steric bulk contribution and the electrostatic contribution and fall into the area of LJ3, LJ5 and C4. For R_2 and R_3, electron-withdrawing groups of -OH and -CH_2NO_2 were introduced to interact with blue C2 and C3 (Fig. 9).

From Table-2, we can see that the predicted biological activities are all higher than that of compound 47 (pIC_{50} = 11.22). Such results further suggest that this 4D-QSAR model has a strong predictive ability and can be prospectively used in structural modification or molecular design

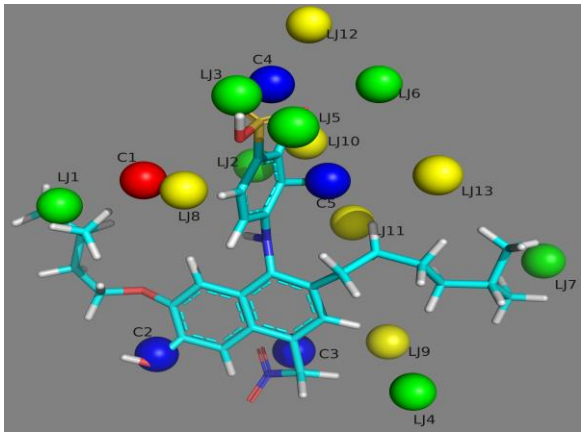
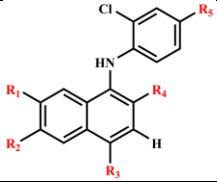


Fig. 9: The steric and electrostatic contour maps of new compound D3 with 4D-QSAR model.

Table-2: Potential new compounds with *in silico* biological activity.



Comp	R_1	R_2	R_3	R_4	R_5	Predicted pIC_{50}
D1	-$O(CH_2)_2CH(CH_3)_2$	-OH	-CH_2OH	-CH_2OH	SO_3H	11.69
D2	-$O(CH_2)_2CH(CH_3)_2$	-OH	-CH_2NO_2	-CH_2OH	SO_3H	11.82
D3	-$O(CH_2)_2CH(CH_3)_2$	-OH	-CH_2NO_2	-$(CH_2)_3CH(CH_3)_2$	SO_3H	12.62

Conclusion

In this paper, the CEPs were constructed by using GROMACS dynamics simulation and interaction energy descriptors, i.e. Lennard–Jones interaction energy descriptors and Coulomb interaction energy descriptors were calculated using LQTAgrid program. Through the combination of CDDA-OPS-GA method for filtration and selection descriptors, this 4D-QSAR model had achieved satisfactory results. It could be concluded that large group with electron-withdrawing increases the compound's bioactivity for R_1 group, the R_2 group describes conformational flexibility of substituent group, the bulky group is preferred for R_3 group, the electron-withdrawing group and large group is preferred for 1 position and 3 position of quinazoline ring.

Acknowledgments

The project was supported by the Undergraduate Innovation and Entrepreneurship Foundation (2021CX242) and the Jiangxi Province Science Foundation (20171BAB205104).

References

1. Y. He, B. S. Harrington, J. D. Hooper, New crossroads for potential therapeutic intervention in cancer - intersections between CDCP1, EGFR family members and downstream signaling pathways, *Oncoscience*, **3**(1), 5 (2016).
2. R. Wang, X. Wang, J. Q. Wu, B. Ni, L. B. Wen, L. Huang, Y. Liao, G. Z. Tong, C. Ding, X. Mao, Efficient porcine reproductive and respiratory syndrome virus entry in MARC-145 cells requires EGFR-PI3K-AKT-LIMK1-COFILIN signaling pathway, *Virus Res.*, **225**, 23 (2016).
3. C. Yewale, D. Baradia, I. Vhora, S. Patil, A. Misra, Epidermal growth factor receptor targeting in cancer: a review of trends and strategies, *Biomaterials*, **34**(34), 8690 (2013).
4. A. Cho, J. Hur, Y. W. Moon, S. R. Hong, Y. J. Suh, Y. J. Kim, D. J. Im, Y. J. Hong, H. J. Lee, Y. J. Kim, H. S. Shim, J. S. Lee, J. H. Kim, B. W. Choi, Correlation between EGFR gene mutation, cytologic tumor markers, 18F-FDG uptake in non-small cell lung cancer, *BMC Cancer*, **16**, 1471 (2016).
5. K. Wang, D. Li, L. Sun, High levels of EGFR expression in tumor stroma are associated with aggressive clinical features in epithelial ovarian cancer, *Onco Targets Ther.*, **9**, 377 (2016).
6. F. Imamura, J. Uchida, Y. Kukita, T. Kumagai, K. Nishino, T. Inoue, M. Kimura, S. Oba, K. Kato, Monitoring of treatment responses and clonal evolution of tumor cells by circulating tumor DNA of heterogeneous mutant EGFR genes in lung cancer, *Lung Cancer*, **94**, 68 (2016).
7. X. B. Holdman, T. Welte, K. Rajapakshe, A. Pond, C. Coarfa, Q. Mo, S. Huang, S. G. Hilsenbeck, D. P. Edwards, X. Zhang, J. M. Rosen, Upregulation of EGFR signaling is correlated with tumor stroma remodeling and tumor recurrence in FGFR1-driven breast cancer, *Breast Cancer Res.*, **17**, 1465 (2015).
8. K. Oda, Y. Matsuoka, A. Funahashi, H. Kitano, A comprehensive pathway map of epidermal growth factor receptor signaling, *Mol. Syst. Biol.*, **1**(1), 1 (2005).
9. N. Minkovsky, A. Berezov, BIBW-2992, a dual receptor tyrosine kinase inhibitor for the treatment of solid tumors, *Curr. Opin. Investig. Drugs.*, **9**(12), 1336 (2008).
10. M. Burotto, V. L. Chiou, J. M. Lee, E. C. Kohn, The MAPK pathway across different malignancies: a new perspective, *Cancer*, **120**(22), 3446 (2014).
11. G. F. Yang, X. Q. Huang, Development of quantitative structure-activity relationships and its application in rational drug design, *Curr. Pharm. Des.*, **12**(35), 4601 (2006).
12. A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, QSAR modeling: where have you been? where are you going to?, *J. Med. Chem.*, **57**(12), 4977 (2014).
13. J. Shim, A. D. Mackerell Jr, Computational ligand-based rational design: role of conformational sampling and force fields in model development, *Medchemcomm*, **2**(5), 356 (2011).
14. J. Ghasemi, M. Salahinejad, M. Rofouei, Review of the quantitative structure–activity relationship modelling methods on estimation of formation constants of macrocyclic compounds with different guest molecules, *Supramol. Chem.*, **23**(9), 614 (2011).
15. A. Hopfinger, S. Wang, J. S. Tokarski, B. Jin, M. Albuquerque, P. J. Madhav, C. Duraiswami, Construction of 3D-QSAR models using the 4D-QSAR analysis formalism, *J. Am. Chem. Soc.*, **119**(43), 10509 (1997).
16. J. P. Martins, E. G. Barbosa, K. F. Pasqualoto, M. M. Ferreira, LQTA-QSAR: a new 4D-QSAR methodology, *J. Chem. Inf. Model.*, **49**(6), 1428 (2009).

17. E. G. Barbosa, M. M. Ferreira, Digital filters for molecular interaction field descriptors, *Mol. Inform.*, **31**(1), 75 (2012).
18. R. F. Teofilo, J. P. A. Martins, M. M. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, *J. Chemom.*, **23**(1), 32 (2009).
19. B. Hemmateenejad, R. Miri, M. Akhond, M. Shamsipur, QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. an application of genetic algorithm for variable selection in MLR and PLS methods, *Chemometrics Intellig. Lab. Syst.*, **64**(1), 91 (2002).
20. A. J. Bridges, H. R. Zhou, D. R. Cody, G. W. Rewcastle, A. McMichael, H. D. Showalter, D. W. Fry, A. J. Kraker, W. A. Denny, Tyrosine kinase inhibitors. 8. an unusually steep structure-activity relationship for analogues of 4-(3-bromoanilino)-6,7-dimethoxyquinazoline (PD 153035), a potent inhibitor of the epidermal growth factor receptor, *J. Med. Chem.*, **39**(1), 267 (1996).
21. S. Li, C. Guo, H. Zhao, Y. Tang, M. Lan, Synthesis and biological evaluation of 4-[3-chloro-4-(3-fluorobenzyloxy)anilino]-6-(3-substituted-phenoxy)pyrimidines as dual EGFR/ErbB-2 kinase inhibitors, *Bioorg. Med. Chem.*, **20**(2), 877 (2012).
22. N. Suzuki, T. Shiota, F. Watanabe, N. Haga, T. Murashi, T. Ohara, K. Matsuo, N. Oomori, H. Yari, K. Dohi, M. Inoue, M. Iguchi, J. Sentou, T. Wada, Synthesis and evaluation of novel pyrimidine-based dual EGFR/Her-2 inhibitors, *Bioorg. Med. Chem. Lett.*, **21**(6), 1601 (2011).
23. A. G. Waterson, K. G. Petrov, K. R. Hornberger, R. D. Hubbard, D. M. Sammond, S. C. Smith, H. D. Dickson, T. R. Caferro, K. W. Hinkle, K. L. Stevens, S. H. Dickerson, D. W. Rusnak, G. M. Spehar, E. R. Wood, R. J. Griffin, D. E. Uehling, Synthesis and evaluation of aniline headgroups for alkynyl thienopyrimidine dual EGFR/ErbB-2 kinase inhibitors, *Bioorg. Med. Chem. Lett.*, **19**(5), 1332 (2009).
24. T. Hassinen, M. Peräkylä, New energy terms for reduced protein models implemented in an off-lattice force field, *J. Comput. Chem.*, **22**(12), 1229 (2001).
25. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, UCSF Chimera: a visualization system for exploratory research and analysis, *J. Comput. Chem.*, **25**(13), 1605 (2004).
26. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, H. J. Berendsen, GROMACS: fast, flexible, and free, *J. Comput. Chem.*, **26**(16), 1701 (2005).
27. T. Darden, D. York, L. Pedersen, Particle mesh ewald: an $N\cdot\log(N)$ method for ewald sums in large systems, *J. Chem. Phys.*, **98**(12), 10089 (1993).
28. H. J. Berendsen, J. V. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak, Molecular dynamics with coupling to an external bath, *J. Chem. Phys.*, **81**(8), 3684 (1984).
29. R. Patil, S. Sawant, Molecular dynamics guided receptor independent 4D QSAR studies of substituted coumarins as anticancer agents, *Curr. Comput. Aided Drug Des.*, **11**(1), 39 (2015).
30. W. Ma, Y. Wang, D. Chu, H. Yan, 4D-QSAR and MIA-QSAR study on the bruton's tyrosine kinase (Btk) inhibitors, *J. Mol. Graph. Model.*, **92**, 357 (2019).
31. L. J. de Campos, E. B. de Melo, Modeling structure–activity relationships of prodiginines with antimalarial activity using GA/MLR and OPS/PLS, *J. Mol. Graph. Model.*, **54**, 19 (2014).
32. E. B. de Melo, J. P. A. Martins, E. H. Miranda, M. M. C. Ferreira, A best comprehension about the toxicity of phenylsulfonyl carboxylates in vibrio fischeri using quantitative structure activity/property relationship methods, *J. Hazard. Mater.*, **304**, 233 (2016).
33. P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, QSARINS: a new software for the development, analysis, and validation of QSAR MLR models, *J. Comput. Chem.*, **34**(24), 2121 (2013).
34. R. B. Patil, E. G. Barbosa, J. N. Sangshetti, S. D. Sawant, V. P. Zambre, LQTA-R: a new 3D-QSAR methodology applied to a set of DGAT1 inhibitors, *Computat. Biol. Chem.*, **74**, 123 (2018).
A. Golbraikh, A. Tropsha, Beware of q^2!, *J. Mol. Graph. Model.*, **20**(4), 269 (2002).
35. A. Golbraikh, A. Tropsha, Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection, *Mol. Divers.*, **5**(4), 231 (2002).
36. A. Balupuri, P. K. Balasubramanian, S. J. Cho, 3D-QSAR, docking, molecular dynamics simulation and free energy calculation studies of some pyrimidine derivatives as novel JAK3 inhibitors, *Arab. J. Chem.*, **13**(1), 1052 (2020).
37. R. Kiralj, M. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, *J. Braz. Chem. Soc.*, **20**(4), 770 (2009).