MOLECULAR DESCRIPTORS USED IN QSAR

Molecular Descriptors used in QSAR

Molecular descriptors can be defined as a numerical representation of chemical information encoded within a molecular structure via mathematical procedure. This mathematical representation has to be invariant to the molecule's size and number of atoms to allow model building with statistical methods.

The information content of structure descriptors depends on two major factors:

(1) The molecular representation of compounds

(2) The algorithm which is used for the calculation of the descriptor.

The three major types of parameters initially suggested are,

(1) Hydrophobic

(2) Electronic

(3) Steric

Lead Molecules

A lead compound in drug discovery is a chemical compound that has pharmacological or biological activity likely to be therapeutically useful, but may still have suboptimal structure that requires modification to fit better to the target. Its chemical structure is used as a starting point for chemical modifications in order to improve potency, selectivity, or pharmacokinetic parameters. Furthermore, newly invented pharmacologically active moieties may have poor druglikeness and may require chemical modification to become drug-like enough to be tested biologically or clinically.

Discovering lead compounds

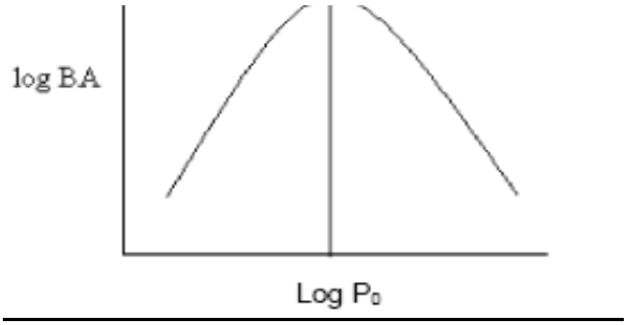
A lead compound may arise from a variety of different sources. Lead compounds are found by characterizing natural products, employing combinatorial chemistry, or by molecular modeling as in rational drug design. Lead compounds are often tested by high-throughput screenings (active compounds are designated as "hits") which can screen compounds for their ability to inhibit (antagonist) or stimulate (agonist) a receptor of interest as well as determine their selectivity for them.

Linear and non linear modeled equations

Hansch Analysis

In 1969, Corwin Hansch extends the concept of linear free energy relationships (LFER) to describe the effectiveness of a biologically active molecule. It is one of the most promising approaches to the quantification of the interaction of drug molecules with biological system. It is also known as linear free energy (LFER) or extra thermodynamic method which assumes additive effect of various substituents in electronic, steric, hydrophobic, and dispersion data in the non-covalent interaction of a drug and biomacro molecules. This method relates the biological activity within a homologous series of compounds to a set of theoretical molecular parameters which describe essential properties of the drug molecules. Hansch proposed that the action of a drug as depending on two processes.

1. Journey from point of entry in the body to the site of action which involves passage of series of membranes and therefore it is related to partition coefficient
log P (lipophilic) and can be explained by random walk theory.



Interaction with the receptor site which in turn depends on,

a) Bulk of substituent groups (steric)

b) Electron density on attachment group (electronic)

He suggested linear and non-linear dependence of biological activity on different parameters.

$\log (1/C) = a(\log P) + b\,\sigma + cES + d$ .....linear

$\log (1/C) = a(\log P)2 + b(\log P) + c\,\sigma + dES + e$ .....nonlinear

Where a-e are constants determined for a particular biological activity by multiple regression analysis. Log P, σ, ES etc, are independent variables whose values are obtained directly from experiment or from tabulations. Other parameters than those shown may also be included. If there are 'n' independent variables to be considered, then there are 2n-1 combinations of these variables that may be used to best explain the tabulated data.

PHYSICOCHEMICAL PARAMETERS AND MOLECULAR DESCRIPTORS

Type	Descriptors
Hydrophobic Parameters	Partition coefficient ; log P
	Hansch's substitution constant; π
	Hydrophobic fragmental constant; f, f'
	Distribution coefficient; log D
	Apparent log P
	Capacity factor in HPLC; log k' , log k'_w
	Solubility parameter; log S
Electronic Parameters	Hammett constant; σ, σ^+, σ^-
	Taft's inductive (polar) constant; σ*
	Swain and Lupton field parameter
	Ionization constant; pK_a , ΔpK_a
	Chemical shifts: IR, NMR
Steric Parameters	Taft's steric parameter; Es
	Molar volume; MV
	Van der waals radius
	Van der waals volume
	Molar refractivity; MR
	Parachor
	Sterimol
Quantum chemical descriptors	Atomic net charge; Q^σ, Q^π
	Superdelocalizability
	Energy of highest occupied molecular orbital; E_{HOMO}
	Energy of lowest unoccupied molecular orbital; E_{LUMO}
Spatial Descriptor	Jurs descriptors, Shadow indices, Radius of Gyration, Principle moment of inertia

Classification of descriptors based on the dimensionality of their molecular representation

Molecular representation	Descriptor	Example
0D	Atom count, bond counts, molecular weight, sum of atomic properties	Molecular weight, average molecular weight number of: atoms, hydrogen atoms carbon atoms, hetero-atoms, non-hydrogen atoms, double bonds, triple bonds, aromatic bonds, rotatable bonds, rings, 3-membered ring, 4-membered ring, 5-membered ring, 6-membered ring
1D	Fragments counts	Number of: primary C, secondary C, tertiary C, quaternary C, secondary carbon in ring, tertiary carbon in ring, quaternary carbon in ring, unsubstituted aromatic carbon, substituted carbon, number of H-bond donar atoms, number of H-bond acceptor atoms, unsaturation index, hydrophilic factor, molecular refractivity
2D	Topological descriptors	Zagreb index, Wiener index, Balaban J index, connectivity indices chi (χ), kappa (K) shape indices
3D	Geometrical descriptors	Radius of gyration, E-state topological parameters, 3D Wiener index, 3D Balaban index

Molecular descriptors are final products of mathematical procedures transforming chemical information encoded within a molecular structure to a numerical representative. Dimensionality of molecular descriptors can identify QSAR model type as described below:

0D QSAR- These are descriptors derived from molecular formula e.g., molecular weight, number and type of atoms etc.

1D QSAR- A substructure list representation of a molecule can be considered as a one-dimensional (1D) molecular representation and consists of a list of molecular fragments (e.g. functional groups, rings, bonds, substituents etc.).

2D QSAR- A molecular graph contains topological or two-dimensional (2D) information. It describes how the atoms are bonded in a molecule, both the type of bonding and the interaction of particular atoms (e.g. total path count, molecular connectivity indices etc.).

3D QSAR- These are calculated starting from a geometrical or 3D representation of a molecule. These descriptors include molecular surface, molecular volume and other

geometrical properties. There are different types of 3D descriptors e.g., electronic, steric, shape etc.

4D QSAR- Four-dimensional information is described in this type of models, and the fourth dimension is an ensemble of conformation of each ligand.

5D-QSAR – Five-dimensional information is described in this type of models, and the fifth dimension is the possibility to represent an ensemble of up to six different induced-fit models.

The descriptors are fall into 4 classes: Topological, Geometrical, Electronic and Hybrid. Topological descriptors in chemistry are graph invariants generated by applying the theorems of graph theory. Examples of topological descriptors are: atom counts, ring counts, molecular weight, weighted paths, molecular connectivity indices, substructure counts, molecular distance edge descriptors, kappa indices, electro-topological state indices, and some other invariants.
Aspects of the structures related to the electrons are encoded by calculating electronic descriptors. Examples of electronic descriptors are: partial atomic charges, HOMO or LUMO energies, dipole moment. Geometric descriptors are used to encode the 3-D aspects of the molecular structure such as moments of inertia, solvent accessible surface area, length-to-breadth ratios, shadow areas, gravitational index. A class of hybrid descriptors called charged partial surface area descriptors encode the propensity of compounds to engage in polar interactions. The set of cpsa descriptors is based on the partial atomic charges and the partial surface area of each atom. The two attributes lists are mixed and a set of approximately 25 cpsa descriptors can be generated by matching the two mixed lists with different weighting schemes. Examples of cpsa descriptors can include: fractional positive surface area, charged weighted negative surface area.

• QSAR models validation.
Validation process aims to provide a model which is statistically reliable with selected descriptors as a consequence of the cause-effect and not only of pure numerical relationship obtained by chance. However, non-statistical validations such as verification of the model in terms of the known mechanism of action or other chemical knowledge are necessary; it is not acceptable to rely on statistics only in validation process. Actually, this is somehow a hard procedure for cases where no mechanism of action is known or where data sets are small. Validation methods are needed to establish the predictiveness of a model. There are two types of validation methods: Internal and external. Internal methods depend on training datasets like Q2 (squared correlation coefficient), R2 (coefficient of determination or the coefficient of multiple determination for multiple regression), chi-squared (X2), and root-mean squared error (RMSE). The major disadvantage of this approach is the lack of predictability of the model when it is applied to a new data set. However, external methods depend on the testing set and it is considered as best validation method. It was reported that, in general, there is no relationship between internal and external predictivity: high internal predictivity may result in low external predictivity and vice versa. In many cases, comparable models are obtained

where some models show comparatively better internal validation parameters and some other models show comparatively superior external validation parameters. This
may create a problem in selecting the final model. Therefore, it is must to develop some good validation techniques to overcome the entire above-mentioned disputes.

B. QSAR in Drug design

QSAR is involved in drug discovery and designing to identify chemical structures with good inhibitory effects on specific targets and with low toxicity levels [25- 41]. The implementation of QSAR in designing different types of drugs as antimicrobial, and antitumor compounds by numerous works is a strong evidence of its efficiency in drug designing. Previous research in this field has been undertaken by different researchers. Researchers investigated QSAR study on a series of 8-substituted xanthines as adenosine antagonists have been carried out. The chemical structure was described with parameters effect the receptors affinity. IN [26], two multilayer feed forward neural networks and docking studies were developed to investigate the hypothetical binding mode of the target compounds. Two 3D-QSAR models for a series of non-purine xanthine oxidase inhibitors were designed to study different factors affect the oxidase inhibitors. QSAR model of xanthine oxidase inhibitory flavylium salts was implemented to predict the inhibitory potency of anthocyanidins as a function of their molecular properties. A three-dimensional QSAR study has been implemented to study epothilones – tubulin depolymerization inhibitors. QSAR models is established for the toxicity of polycylic aromatic hydrocarbons (PAHs). Four dimensional QSAR models is used to study a set of 18 structurally diverse antifolates including pyrimethamine, cycloguanil, methotrexate, aminopterin and trimethoprim, and 13 pyrrolo [2,3-d] pyrimidines. The utility of Topological polar surface area (TPSA) was demonstrated in 2D QSAR for 14 sets of diverse pharmacological activity data. QSAR of Hydrazones of N-Amino-N‘-hydroxyguanidine as Electron Acceptors for Xanthine Oxidase was built. Antiviral QSAR models are implemented to predict by the first time an mt-QSAR model for 500 drugs tested in the literature against 40 viral species. The Markov Chain theory is used to calculate new multi-target entropy that fits a QSAR
model.

III. RESULTS & DISCUSSION

A. QSAR Implementing in Drug Designing Results.

It is very important to validate the model’s performance to conclude whether the results satisfy researcher’s expectations or not. R2 and Q2 are two statistical measures used for this purpose [35-41]. Values of R2 and Q2 obtained from different previous researches are listed in Table I. R2 (called as coefficient of determination or the coefficient of multiple determination for multiple regression.) is a statistical measure of how close the data are to the fitted regression line; High R-squared indicates that the model has a good fit. According to previous research, R2 should be ≥ 0.6 to consider the model fits well. As it is shown in Table I, all QSAR models developed have a higher R2 value than 0.6. Q2 is squared correlation coefficient and it is used as a criterion of both robustness and predictive ability of the model. It can be considered as an indicator of the high predictive power of the QSAR model. However, high Q2 value is not enough to conclude that the model has acceptable predictive ability; models should be tested for their ability to predict the activity of

compounds of an external test set also. It was proven that for good predictability R2 – Q2 value should not be larger than 0.3. R2 – Q2 values are calculated for researches and added in Table I. As it can be noticed from Table 1, only in the R2 – Q2 value exceeds 0.3, while in all other works the values are very small (lower than 0.3), which indicates a good predictability of the constructed models in these works. However, QSAR predictability and robustness levels cannot be proved by R2 and Q2 values only; more parameters should be involved to obtain a strong conclusion, as: chisquared ($\chi 2$), root-mean squared error (RMSE), correlation coefficient R between the predicted and observed activities, slopes k and k' of the regression lines through the origin. Also, the chemical space of training and test sets has to be discussed and studied; real outliers, with respect to character and structure similarities, have to be found and removed. Only a small number of reported QSAR studies were implementing numerous different validation characteristics in their QSAR validation processes.

TABLE 1. R^2 AND Q^2 VALUES OBTAINED BY APPLYING QSAR MODELS IN DIFFERENT WORKS. $R^2 – Q^2$ VALUES ARE CALCULATED FOR EACH WORK.

Reference	Paper Title	Descriptors	R^2 value	Q^2 value	$R^2 – Q^2$
[35]	More Effective DPP4 Inhibitors as Antidiabetics Based on Sitagliptin Applied QSAR and Clinical Methods	Hydrophobicity, counts of rotatable bonds, hydrogen bond donor and acceptor atoms, and topological polar surface area.	0.85	0.77	0.08
[36]	Molecular modelling studies of 3,5-dipyridyl-1,2,4-triazole derivatives as xanthine oxidoreductase inhibitors using 3D-QSAR, Topomer CoMFA, molecular docking and molecular dynamic simulations	Steric, electrostatic, and hydrophobic fields.	0.988	0.578	0.41
[37]	Prediction of caspase-3 inhibitory activity of 1,3-dioxo-4-methyl-2,3-dihydro-1h-pyrrolo[3,4-c] quinolines: QSAR study	HOMO, LUMO energies	0.955	0.885	0.07
[38]	Predictive QSAR modeling on tetrahydropyrimidine-2-one derivatives as HIV-1 protease enzyme inhibitors	Radial Distribution Function (RDF)	0.824	0.773	0.05
[39]	Development of an in Silico Model of DPPH• Free Radical Scavenging Capacity: Prediction of Antioxidant Activity of Coumarin Type Compounds	van der Waals volume	0.713	0.654	0.06
[40]	Comparative Molecular Field Analysis (CoMFA) of a Series of Selective Adenosine Receptor A2A Antagonists	Electrostatic and steric field	0.970	0.840	0.13
[41]	QSAR and docking studies on xanthone derivatives for anticancer activity targeting DNA topoisomerase II α	Dielectric energy, group count, LogP, shape index basic (order 3), solvent-accessible surface area	0.840	Not detected	/

Some applications of QSAR study in drug design are described in table 1. QSAR study was a predictive tool for investigations antidiabetic drugs based on sitagliptin as potential antioxidant agents. Hydrophobicity, counts of rotatable bonds, hydrogen bond donor and acceptor atoms, and topological polar surface area were used as descriptors in this research. Based on the established QSAR equations, new sitagliptin derivatives with possibly improved pharmacological effect as DPP4 inhibitors are proposed to investigate. Also, by using QSAR study can be predict Antioxidant Activity of Coumarin Type Compounds. In this study the best correlation between activity and structure has shown van der Waals volume, that was used as molecular descriptor. In silico methods can be good predictive tool for evaluation inhibitory activity of molecules. In this investigations, QSAR studies often combine with other methods as docking studies and neural network. For predict 3,5-dipyridyl-1,2,4-triazole derivatives as xanthine oxidoreductase inhibitors, QSAR study was used. The results suggested that the steric, electrostatic, and hydrophobic fields played an important role in the models. A QSAR study was performed on a series of 1,3-dioxo-4-methyl-2,3-dihydro-1H-pyrrolo[3,4-c] quinolones in pursuit of better caspase-3 inhibitors.

The study reveals that when increasing the conformational minimum energy while decreasing the lowest unoccupied molecular orbital energy (LUMO) and highest occupied molecular orbital energy (HOMO), the biological activity can be increased. On the basis of a selected QSAR model, a new series of 1,3-dioxo-4- methyl-2,3-dihydro-1H-pyrrolo[3,4-c]quinolines compounds, calculated their caspases inhibitory activity and found that the designed compounds were more potent than the existing compounds.
QSAR model was carried out to predict HIV-1 protease receptors inhibitors activity. In this study Radial Distribution Function (RDF) was used as molecular descriptor that has shown the best correlation with HIV-1 protease inhibition. The QSAR model also indicates that the descriptors (RDF010u, RDF010m, TPSA (NO), F04[C–N]) play an important role in enzyme binding. The CoMFA approach to studies of 3D-QSAR for series of compounds has proven to be a valuable technique for building predictive model. In this study electrostatic and steric field were used as descriptors. A QSAR model was developed to explore the anticancer compounds from xanthone derivatives by the multiple linear regression method. A high activity–descriptors relationship accuracy are obtained referred by regression coefficient and a high activity prediction accuracy. Molecular descriptors: dielectric energy, group count (hydroxyl), LogP (the logarithm of the partition coefficient between n-octanol and water), shape index basic, and the solvent-accessible surface area – were found to correlate with anticancer activity.

IV. CONCLUSION

In all described articles QSAR study were good prediction tool for investigation drug activity or binding mode on specific receptors. Descriptors that have shown the best correlation in this investigation gives information about important functional groups in the structures of tested compounds. According to this, by changing some groups in the structure of drugs, we can increase their pharmacological activity or physicochemical properties. In general, the experimental determinations are very expensive and the QSPR studies allow a reduction of this cost. It is basically used to study the biological activities with various properties associated with the structures, which is helpful to explain how structural features in a drug molecule influence the biological activities. QSPR/QSAR methods can be used to build models that can predict properties or activities for organic compounds. However, an effective way to encode the structures with calculated molecular structure descriptors are required for accurate models development. The descriptors incorporated in models development can provide an opportunity to focus on specific features account for the property or activity of interest in the compounds. QSAR should not replace experimental values, but it is useful predictive tool and might be usable if no data were available.

REFERENCES

[1] Kapetanovic IM. Drug Discovery and Development - Present and Future. InTech. 2016; DOI: 10.5772/1179.

[2] Badnjevic A, Beganovic E, Music O. Facts about solution based and cartridge based devices for blood gas analyses. IEEE 18th International Conference on System, Signals and Image Processing. pp:1-5, 16-18 June 2011, Sarajevo, Bosnia and Herzegovina.

[3] Badnjevic A, Gurbeta L, Boskovic D, Dzemic Z. Medical devices in legal metrology. IEEE 4th Mediterranean Conference on Embedded Computing (MECO). pp: 365-367, 14 – 18 June 2015, Budva, Monténégro

[4] Badnjevic A, Gurbeta L, Boskovic D, Dzemic Z. Measurement in medicine – Past, present, future. Folia Medica Facultatis Medicinae Universitatis Saraeviensis Journal, 2015; 50(1): 43-46

[5] Boskovic D, Badnjevic A. Opportunities and Challenges in Biomedical Engineering Education for Growing Economies. IEEE 4th Mediterranean Conference on Embedded Computing (MECO), pp: 407-410, 14 – 18 June 2015, Budva, Monténégro

[6] Badnjevic A, Gurbeta L. Development and Perspectives of Biomedical Engineering in South East European Countries. IEEE 39th International convention on information and communication technology, electronics and microelectronics (MIPRO), 30. May to 03. June 2016. Opatija, Croatia

[7] Badnjevic A, Beganovic E, Gvozdenovic V, Sehic G. Automated Closed Loop Controller of Insipred Oxygen System for Improved Mechanical Ventilation in Newborns. IEEE 34th International convention on information and communication technology, electronics and microelectronics (MIPRO), pp: 145-149, 23.-27. May 2011. Opatija, Croatia

[8] Kapetanovic IM. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. Chem-Biol. Interact. 2008; 171(2), 165-176.

[9] Aparoy P, Reddy K, Reddanna P. Structure and Ligand Based Drug Design Strategies in the Development of Novel 5-LOX Inhibitors, Current Medicinal Chemistry. 2012; 19(19), ISSN 3763-3778.

[10] Santos-Filho OA, Hopfinger AJ, Cherkasov A, de Alencastro RB. The receptor-dependent QSAR paradigm: an overview of the current state of the art. Med. Chem. (Shāriqah (United Arab Emirates)) 2009; 5, 359–366.

[11] Kubinyi H. QSAR: Hansch Analysis and Related Approaches. In Methods and Principles in Medicinal Chemistry; Mannhold R, Kroogsgard-Larsen P, Timmerman H, Eds.; Wiley-VCH: Weinheim, Germany, 1993; 1, 240.

[12] Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in Drug Design - A Review. Current Topics in Medicinal Chemistry. 2010; 10, 95-115.