Vol. 5, No. 3 March 2014 ISSN 2079-8407
Journal of Emerging Trends in Computing and Information Sciences

http://www.cisjournal.org

The Predictive Learning Role in Drug Design

[1] Mohammed E. El-Telbany
[1] Assoc. Prof., Department of computers and Systems, Electronics Research Institute, Egypt

[1] telbany@eri.sci.eg

ABSTRACT

QSAR (quantitative structure-activity relationship) modeling is one of the well developed areas in drug development through computational chemistry. Similar molecules with just a slight variation in their structure can have quit different biological activity. This kind of relationship between molecular structure and change in biological activity is center of focus for QSAR Modeling. Predictions of property and/or activity of interest have the potential to save time, money and minimize the use of expensive experimental designs, such as, for example, animal testing. This paper, presents a survey of the machine learning algorithms' roles in the field of QSAR modeling and their impact on modern drug design processes.

Keywords: *QSAR, drug design, machine learning, prediction.*

1. INTRODUCTION

In recent years, with products of human genome project helping to reveal many new disease targets to which drug treatments might be aimed, all the major pharmaceutical companies have invested heavily in the routine ultra-High Throughput Screening (uHTS) of vast numbers of 'drug-like' molecules guided by chemo informatics investigations [1, 2]. The development of a new drug is still a challenging, time-consuming and cost-intensive process and due to the enormous expense of failures of candidate drugs late in their development, uHTS in vitro assays now cover liabilities such as possible side effects [3] as well as therapeutic properties, computational methods can be used to assist and speed up the drug design process. It is obvious that the drug discovery and development process would greatly benefit from faster and cheaper procedures to identify chemical compounds with desired biological properties and to optimize their structure in order to obtain effective drugs. Several major bottlenecks in drug discovery may be addressed with computer-assisted drug design methods, such as quantitative structure–activity relationships (QSAR) models [4]. Drug design and discovery is a systematic, serial process of identification and modification of chemical structure to achieve desired results against biological targets associated with a particular disease. The design approach to drug discovery starts with scientists understanding the genetic and molecular base of a disease and using that information to select a specific therapeutic target by using hundreds of thousands of compounds that are screened against the targets to identify those compounds that hit the targets using high throughput screening (HTS). By analyzing the structure of the selected compounds and identifying common active substructures, novel compounds containing those substructures are synthesized to significantly lower the number of lead compounds. Finally, the leads that identified are further refined to comply with pharmacokinetic constraints such as absorption and bioavailability, and to increase their potency and efficacy, while decreasing side effects and toxicity, usually termed absorption, distribution, metabolism, elimination, toxicity (ADMET) properties [5].

Recently, machine learning algorithms play an increasingly important role in drug design, screening and identification of candidate molecules and studies on quantitative structure-activity relationships (QSAR), which can discriminate between sets of chemicals that are active(inactive) towards a certain biological receptor [6, 7, 8].The machine learning field [9, 10, 11, 12] is versatile methods or algorithms such as decision trees, lazy learning, k-nearest neighbors, Bayesian methods, Gaussian processes, artificial neural networks, artificial immune systems, particle-swarm optimization, artificial bee optimization, cuckoo search, support vector machines, and kernel algorithms for a variety of tasks in drug design. Among them are:

- **Virtual screening:** In virtual screening machine learning techniques are applied to rank or filter compounds with respect to different properties. Especially in legend based virtual screening, where no structural information about the target is available, machine learning methods enhance similarity search—even if only very few reference compounds are given [13]. Additionally, Machine learning methods may be applied to create diverse compound libraries that can serve as input for virtual screening (library design) [14].

- **Quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR):** QSAR and QSPR models are statistical models used to infer dependencies between chemical structures and their biological activity or physicochemical properties [15].

- **Prediction of protein structure, function and interaction:** Machine learning methods have found extensive applications in biochemical tasks like protein structure prediction, protein function prediction and characterization of protein-protein interaction.

The machine learning algorithms are becoming more popular in analysis of structure–activity data, establishment of QSAR, area of research. We will restrict ourselves solely to machine learning applications on QSAR applications [16]. The machine learning algorithms extract information from experimental data by computational and statistical methods and generate a set of rules, functions or procedures that allow them to predict the properties of novel objects that are not included in the learning set. Formally, a learning algorithm is tasked with selecting a hypothesis that best supports the data. Considering the hypothesis to be a function f mapping from the data space X to the response space Y; i.e., $f: X \rightarrow Y$. The learner selects the best hypothesis f^* from a space of all possible hypotheses F by minimize errors when predicting value for new data, or if our model includes a cost function over errors, to minimize the total cost of errors. As shown in Figure 1, the QSAR modeling is heavily dependent on the selection of molecular descriptors; if the association of the descriptors selected to biological property is strong the QSAR model can identify valid relations between molecular features and biological property/activity. Thus, uninformative or redundant molecular descriptors should be removed using some feature selection methods during (Filters) or before (wrappers) the learning process. Subsequently, for tuning and validation of the predicatively of learned QSAR model, one of the validation strategy can be applied likes cross-validation, leave-one-out or the full data set is divided into a training set and a testing set prior to learning. The QSAR models based on machine learning algorithms are applied during the drug development cycles to optimize the biological activity, target selectivity, and other physico-chemical and biological properties of selected chemical compounds. Machine learning models are also used to eliminate chemical compounds that have undesirable effects, such as mutagens, carcinogens, teratogens, or other toxic compounds. During the learning process, machine learning algorithms are used to build models that describe the empirical relationship between the structure and property of interest. The optimal model is obtained by searching for the optimal modeling parameters and feature subset simultaneously. This finalized model built from the optimal parameters will then undergo validation with a testing set to ensure that the model is appropriate and useful. This paper gives a survey to the machine learning algorithms that have been commonly used in constructing QSAR models. The rest of the paper is organized as follows. Section 2 briefly introduces the QSAR models. Section 3 is a survey to the related methods used in QSAR modeling. Section 4 describes an evaluation of QSAR modeling and prediction results. Section 5, describes a comparison among presented algorithms with insights into the benefits of learning algorithms. Finally Section 6 presents the findings and conclusions.

2. QSAR Models

QSAR models are in essence a mathematical function that relates features and descriptors generated from small molecule structures to some experimental determined activity or property [17]. The structure-activity study can indicate which features of a given molecule correlate with its activity, thus making it possible to synthesize new and more potent compounds with enhanced biological activities. QSAR analysis is based on the assumption that the behavior of compounds is correlated to the characteristics of their structure. In general, a QSAR model is represented as follows:

$$activity = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \qquad (1)$$

Where the parameters X_i are a set of measured (or computed) properties of the compounds and $\beta_0$through β_i are the calculated coefficients of the QSAR model. The computational techniques should be used to detect the functional group in compounds in order to refine the discovered drug. This can be done using QSAR that consists of computing every possible number that can describe a molecule then doing an enormous curve fit to find out which aspects of the molecule correlate well with the drug activity or side effect severity. This information can then be used to suggest new chemical modifications for synthesis and testing. Typical molecular parameters that are used to account for electronic properties, hydrophobicity, steric effects, and topology can be determined empirically through experimentation or theoretically via computational chemistry. The QSAR has typically been used for drug design and discovery and development and has gained wide applicability for correlating molecular information with not only biological activities but also with other physicochemical properties, which has therefore been termed quantitative structure-property relationship (QSPR). QSPR models are used often to model and predict ADMET properties. QSAR and QSPR are very similar in that much of the same computational approaches are used in their development and optimization. The major differences arise from the activities/ properties they are designed to predict.

3. FROM STATISTICAL TO LEARNING ALGORITHMS

In the past decades there are several statistical methods that can be applied to QSAR studies such as multiple linear regression and partial least squares. Recently, there is also a growing interest in the application of machine learning algorithms in the field of QSAR, as well as other molecular modeling approaches have been recognized as important tools in drug design [8]. Machine learning comprises a set of algorithms that enable computers to learn. The concept of learning usually builds on two different approaches: inductive and deductive learning. In unsupervised learning one typically tries to uncover hidden regularities or to detect anomalies in the data. In supervised learning, there is a label associated with each example. It is supposed to be the answer to a question about the example. If the label is discrete, then the task is called classification problem – otherwise, for real valued labels we speak of a regression problem. These set of algorithms have been successfully applied to QSAR and QSPR modeling and classification by predicting experimental activities based on descriptors or features.

The shift of using machine learning algorithms comes from the fact that the drug design is very complex and requires the use of hybrid techniques [8, 7, 18].

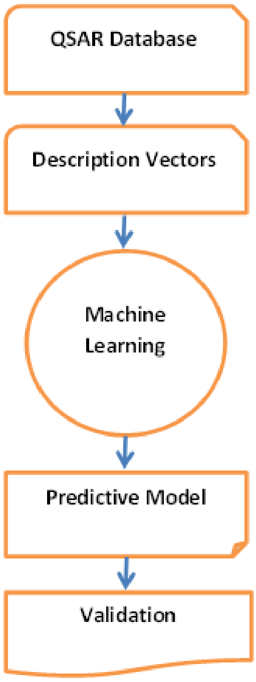

Fig 1: General Steps of Developing QSAR Models

3.1 Statistical Algorithms

- **Multiple Linear Regressions:** Traditionally, QSAR models may be constructed through linear regression analysis of the biological activity data against a number of features or 'descriptors' of chemical structure. The linear regression is one of the most fundamental and common modeling method for regression QSAR. The linear regression is attracted for its simplicity and ease of interpretation as the model assumes a linear relationship between the compound's property, $\hat{y}$, and its feature vector, denoted$\mathbf{X}$, which is usually the molecular descriptors. Thus, with the notion of $\mathbf{X}$, the property of an unknown compound can be predicted by the fitted model [19, 20].

$$\hat{y} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \qquad (2)$$

The establishment of a linear model is the parameterization of $\boldsymbol{\beta}$ based on training data. Once $\boldsymbol{\beta}$ is solved, the prediction of $\mathbf{y}$, $\hat{y}$, can be easily obtained from $\mathbf{X}$. The size of the coefficients may reveal the degree of influence of the corresponding molecular descriptors on the target property. In addition, a positive coefficient suggests that the corresponding molecular descriptor contributes positively to the target property, while a negative coefficient suggests negative contribution. The linear regression has been successful techniques used to construct QSAR models. However, even with moderate numbers of features this technique can result in over-fitting. In order to avoid over-fitting, linear regression is often used in combination with principal component analysis (PCA) or other feature selection methods like genetic algorithm. Moreover, it is assumed that the underlying relationship is linear and that any deviation from linearity will be distributed normally (a parameter assumption). It also assume that the drug properties are real no and they are independent each other, so that the effect of one variable is the other variables. In many QSAR problems it is desirable to learn relationships that are non-linear.

- **Partial Least Square:** To overcome this problem the Partial least squares (PLS) is more appropriate when the number of features greatly exceed the number of samples and when features are highly collinear [21]. PLS projection to latent structures is a robust multivariate generalized regression method using projections to summarize multitudes of potentially collinear variables [22]. PLS regression technique is especially useful in quite in common case where the number of descriptors is greater than the no of compounds (data points) and/or there exist other factors leading to correlations between variables [23]. PLS first projects both the predictor $\mathbf{X}$ and response $\mathbf{Y}$ variables onto one or more new axes with "outer relations" (factor) yielding scores that contain most of the information in the observed variables.PLS leads to stable, correct and highly predictive models even for correlated descriptors [24]. The number of latent factors used in PLS is an important consideration for QSAR modeling, and it is usually obtained through the use of cross-validation methods like n-fold cross validation and leave-one-out methods, where a portion of the samples is used as training set, while the other portion is set aside as testing set to validate the model that was built from the training set.

3.2 Learning Algorithms

Recently, learning algorithms were found to be efficient in constructing QSAR models. The advantage of using a non-linear method compared to a linear method such as linear regression is that more complex and non-linear QSAR models can be derived, which in turn can better reflect the possible relationship between the features of the molecule and its activity.

- **Artificial Neural Networks:** One of these techniques is the artificial neural networks which have been developed as generalizations of mathematical models of biological nervous systems. In a simplified mathematical model of the neuron, the effects of the synapses are represented by weights that modulate the effect of the associated input signals, and the nonlinear characteristic exhibited by neurons is represented by a transfer function, which is usually the sigmoid, Gaussian function etc. The neuron impulse is then computed as the weighted sum of the input signals, transformed by the transfer function. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. The learning situations in neural networks may be classified into three distinct sorts. These are supervised learning, unsupervised learning and reinforcement learning [12]. Feed-forward neural networks has been applied to QSAR modeling where feed-forward neural networks capture the relation between some feature of the molecules and the property that must be predicted using supervised learning algorithms [25]. The feature selection process is needed to be performed in order to select which ones to include or exclude from the model as input. Each input or feature then needs to be weighted with respect to maximizing predictive accuracy on the output decision over the training examples. Self-organizing maps (SOM) are also applied for QSAR to determining if the structurally related compounds will have similar properties determining similarity is a complex task. The SOM are applied for QSAR to produce low dimensional representations of the training sample while preserving the topological properties of the input space [26].

- **Decision Trees**: Another method is decision trees(DT) learning can provide an informative model, through which predictive rules are induced to solve classification/regression problems [27]. The method uses a process called recursive partitioning. In their simplest form, e.g. C4.5 [28], each attribute of the data is examined in turned and ranked according to its ability to partition the remaining data. This algorithm performs a greedy search using some loss function (usually referred to as an impurity function) to find the best axis-parallel split while partitioning the data space as it grows. The data are propagated along the branches of the tree until sufficient attributes have been chosen to correctly classify them. The trained classifier has a tree-like structure. They are popular in QSAR domain for their ease of interpretability. The tree effectively combines the training process with descriptor selection, limiting the complexity of the model to be analyzed. The DT is used to predicting QSAR of pyrimidines [29].

- **Support Vector Machines**: Recently, major contributions to the machine learning field are being achieved through *kernel methods*[12, 30, 31]. One of the kernel machines is the support vector machines (SVM) are a set of related supervised learning methods that can used for classification and regression. SVM represent the input descriptors/features as vectors that are projected onto higher-dimensional space. A special property of SVM is that they simultaneously minimize the empirical classification error and minimize geometric margin. SVM was created to address challenging problems in QSAR analysis. The goal of QSAR analysis is to predict the bioactivity of molecules. Each molecule has many potential descriptors that may be highly correlated with each other or irrelevant to the target bioactivity [29]. The bioactivity is known for only a few molecules. These issues make model validation challenging and over fitting easy. The results of the SVMs are somewhat unstable small changes in the training and validation data or on model parameter may produce rather different sets of nonzero weight attributes. An optimal hyper-plane is then constructed separating the actives and in-actives. The hyper-plane is used to predict the activity of new compounds that are tested.

- **Evolutionary Algorithms**: Evolutionary algorithms are other techniques which are adaptive search methods, which may be used to solve search and optimization problems, based on the genetic processes of biological organisms [32,33,34]. Over many generations, natural populations evolve according to the principles of natural selection and 'survival of the fittest'. By mimicking this process, evolutionary algorithms are able to 'evolve' solutions to real world problems, if they have been suitably encoded. Usually grouped under the term evolutionary algorithms or evolutionary computation, we find the domains of genetic algorithms, evolution strategies, evolutionary programming, genetic programming and learning classifier systems. Genetic algorithms have been used in feature selection for QSAR with a range of learning algorithms, e.g. Artificial Neural Networks [18, 35,36].D. Turner *et al*.[37] improved the predictive value of a QSAR model by variable selection using a GA. Fitness is evaluated by a PLS (partial least squares)cross validation. Also, a genetic algorithm is used to optimize ANN connection weights [36].Reference[38] applied Bayesian-regularized genetic neural networks (BRGNNs) and GA-optimized SVM (GA-SVM) to QSAR modeling in drug design. References [42, 43] applied e*volutionary strategies algorithms* for *de-novo* drug (or ligand) design which is an attempts to generate ligands from scratch based only on information about the pharmaceutical target site or known ligands as shown in Figure 2.

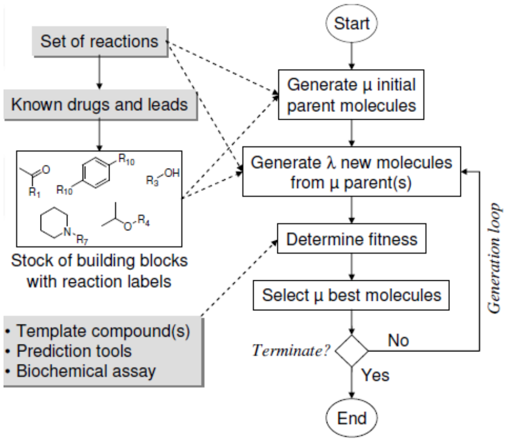

Fig 2: De novo Fragment Assembly using EA.

- **Ensemble Methods**: To improve the performance of single predictor or classifiers, ensemble or meta approaches have been developed [39, 40]. Ensembles achieve predictive improvement by reducing the variance in their members and leaving the bias unaltered. By taking advantage of the bias and variance trade-off the ensemble can obtain a lower prediction error than any single member. Bagging, boost and stacking are different ensemble strategies. In bagging different data is used for each classifier. Boosting weighting the data and subsequent iterations involve improving the existing model by focusing on the instances poorly predicted in the previous iteration. Bagging and boosting are concerned with using the same classifiers for the whole ensemble, whereas stacking can mix any number of different classifiers together. In stacking instead a selection of classifiers can be used in order to benefit from the different learning schemes. The overall classification reflects the combined predictions of classifiers

4. QSAR MODELS VALIDATION

The validation of a quantitative structure-activity relationship is probably the most important step of all. The validation estimates the reliability and accuracy of predictions before the model is put into practice. Poor predictions misguide the direction of drug development and turn downstream efforts meaningless. To verify model quality in regression tasks, predictions are made on the testing set in order to check the agreement between the theoretical values and experimental values by calculating root-mean square error of prediction (RMSE).

$$RMSE = \sqrt[2]{\frac{\sum_{j=1}^{m}(\hat{y}_j - y_j)}{m}} \qquad (3)$$

Where, $\hat{y}_j$, values of the predicted values, and y_j, values of the actual values. However, it is necessary to get a large number of testing compounds in order to draw statistically convincing conclusion. There is another measure of model quality which is the R^2 value. The R^2 is also known as the Pearson coefficient and ranges from -1 to +1. Good models are characterized by high value of R^2 and low values of *RMSE*. However, it is well known that R^2 is not always a good indicator of model quality and in many cases can be misleading. There are four methods for validation of QSAR models [41] namely: internal validation or cross-validation; validation by dividing the data set into training and test compounds; true external validation by application of model on external data and data randomization or Y-scrambling. Cross-validation is probably the most popular technique for estimating generalization error of QSAR models. One of the most widely used techniques is k-fold cross-validation. In k-fold cross validation, training data are partitioned into disjoint k folds of the same size. The average accuracy of k-run is called k-fold cross-validation accuracy (See Figure 3).

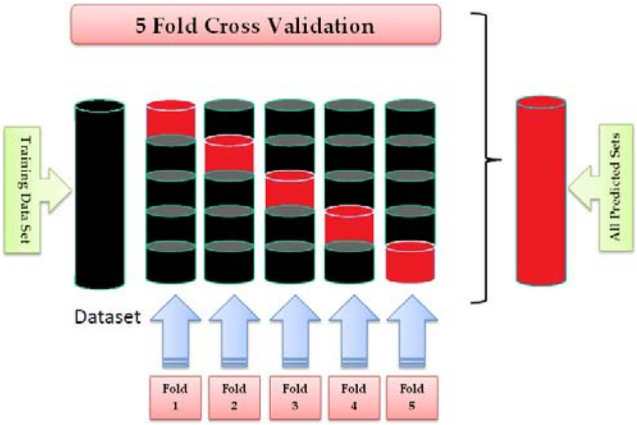

Fig 3: 5-fold cross-validation.

It could be argued that leave-one-out (LOO) cross-validation is the optimal method as it gives the most available data for training. Generalization error of QSAR models measures their average performance when the models are generalized to the entire possible chemistry space. However, there are different aspects of validation of QSAR models that need attention include methods of selection of training set compounds, setting training set size and impact of variable selection methods for training set models for determining the quality of prediction.

5. TRADITIONAL AND LAERNING ALGORITHMS: A COMPARSION

Comparing the main characteristics of the modeling algorithms for QSAR data is presented in a comparative way. We consider the traditional techniques and learning algorithms that have been discussed earlier.

The desired characteristics of thesis algorithms depend on the particular problem under consideration. However, the evaluation of learning algorithms is done along a set of evaluation criteria. These include, *scalability*, *parameter setting requirements*, *dealing with noise and outliers*, and *sensitivity to control parameters*. It is clear that all machine learning techniques have advantages and disadvantages. As the research in drug design depends on a series of factors each method could be efficient for specific problems. The most widely used MLP suffers with some limitations such as architecture and parameter setting, slow convergence of the back-propagation algorithm (with the risk to get stuck in a local minimum), poor ability of generalization, and lack of robustness due to random initialization of the weights. The main disadvantages of SVM include the high complexity of the model and the long computing time if a significant prediction power is required. The advantages of SOM are related to their nondeterministic characteristics, as well as to the robustness of these techniques to outliers. It is important to note that various machine learning techniques address the problem of high dimensionality of data in QSAR studies and Genetic Algorithm (GA) is one of the most applied techniques to solve this problem. This approach is based on natural selection, mutation, evolution and genetic crossover [8].

6. CONCLUSIONS

The traditional methods underlying QSAR are well-established techniques continue to be used, providing successful results. These techniques include linear methods such as partial least squares. However, as discussed previously, there is great interest in developing new and improved QSAR models in order to improve the efficiency and productivity of drug design and development. Because of the great complexities, scarce and “*noisy data*,” as well as overwhelming numbers of parameters involved, researchers have borrowed heavily from the field of machine learning. In this paper we have reviewed different algorithms of statistical and machine learning to the development of predictive QSAR models. There is a great opportunity for the development of novel approaches and methodologies that will increase the likelihood of survival of drug candidates through the development process. Most of the studies that used machine learning algorithms have proven to be of practical value for approximating non-linear separable data, especially for predict or classifying QSAR data. The goal of future research is to continuously improve modeling for QSAR data using natural inspired machine learning algorithms [33] such as particle swarm optimization and cuckoo search algorithm.

REFERENCES

[1] C. Lipinski, Lead- and drug-like compounds: the rule-of-five revolution. Drug Discovery Today: Technologies 1(4):337-341, 2004.

[2] P. Lesson, A. Davis, J. Steele, Drug-like properties: guiding principles for design–or chemical prejudice? Drug Discovery Today: Technologies, 1(3):189-195, 2004.

[3] A. Li, Preclinical in vitro screening assays for drug-like properties. Drug Discovery Today: Technologies, 2(2):179-185, 2005.

[4] Hansch, A Quantitative Approach to Biochemical Structure-Activity Relationships. Acct. Chem. Res. 2: 232-239, 1969.

[5] M. Segall Why is it Still Drug Discovery?, European Biopharmaceutical Review Spring '08 issue, Samedan Ltd., 2008.

[6] W. Duch, K. Swaminathan and J. Meller, Artificial Intelligence Approaches for Rational Drug Design and Discovery, Current Pharmaceutical Design, 13, 2007.

[7] L. Chin Yee and Y. Chun Wei, Current Modeling Methods Used in QSAR/QSPR, in Statistical Modelling of Molecular Descriptors in QSAR/QSPR, 1st,Edited by M. Dehmer, K. Varmuza, and D. Bonchev, Wiley-VCH Verlag GmbH & Co, 2012.

[8] J. Gertrudesa, V. Maltarollob, R. Silvaa, P. Oliveiraa, K. Honórioa and A. da Silva. Machine Learning Techniques and Drug Design, Current Medicinal Chemistry, 19, 4289-4297, 2012.

[9] E. Burke and G.Kendall, Search Methodologies Introductory Tutorials in Optimization and Decision Support Techniques, 2nd (ed.) Springer, 2014.

[10] M. Mohri, A. Rostamizadeh, and A.Talwalkar, Foundations of Machine Learning, MIT Press, 2012.

[11] S. Marsland Machine Learning: An Algorithmic Perspective, (Chapman & Hall/CRC, 2009.

[12] C. Bishop, Pattern Recognition and Machine Learning, Springer, 2nd, 2006.

[13] J. Hert, P. Willett, D. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer. New methods for Ligand-Based virtual screening:? use of data fusion and machine learning to enhance the effectiveness of similarity searching. Journal of Chemical Information and Modeling, 46(2):462–470, 2006.

[14] G. Schneider and K. Baringhaus. Molecular design: concepts and applications. Wiley-VCH, Feb. 2008.

[15] H. Kubiny, QSAR: Hansch Analysis and Related Approaches, VCH, 2003.

[16] J.Malik, H.Soni, S. Singhai, H.Pandey, QSAR - Application in Drug Design, International Journal of Pharmaceutical Research & Allied Sciences, Volume 2, issue 1,1-13, 2013.

[17] D. Livingstone. Data Analysis for Chemists {Applications to QSAR and Chemical product Design. Oxford University Press, 1995.

[18] A, Dudek, T., Arodzb and J. Galvez, Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review, Combinatorial Chemistry & High Throughput Screening, 9, 213-228, 2006.

[19] E. Papa, J., Dearden and P.Gramatica , Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors. Chemosphere. 67(2): 351-358, 2007.

[20] E. Ibezim et al., Computer-Aided Linear Modeling Employing Qsar for Drug Discovery, African Journal of Basic & Applied Sciences 1 (3-4): 76-82, 2009.

[21] I. Skoglund Algorithms for a Partially Regularized Least Squares Problem, Linköpings universitet, 2007.

[22] H.Waterbeemd et al., Glossary of Terms Used in Computational Drug Design (IUPAC) Recommendations 1997. Annu. Rep. Med. Chem. 33: 397-409, 2008.

[23] A. Khlebnikov,I. Schepetkin, N. Domina, L. Kirpotina and Q.Quinn, Improved Quantitative Structure-Activity Relationship Models to Predict Antioxidant Activity of Flavonoids in Chemical, Enzymatic, and Cellular Systems. Bioorg. Med. Chem.15 (4): 1749–1770, 2007.

[24] R. Gieleciak and J. Polanski, Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pKa Values. J. Chem. Inf. Model. 47: 547–556, 2007.

[25] A. Kustrin, R. Beresford, M. Pauzi and A.Yusof , ANN modeling of the penetration across a polydimethylsiloxane membrane from theoretically derived molecular descriptors. J. Pharm. Biomed. Anal. 26(2): 241-254, 2001.

[26] F. Cheng and V. Sutariya, Applications of Artificial Neural Network Modeling in Drug Discovery, Clin Exp Pharmacol, 2:3, 2012.

[27] L., Breiman, J., Friedman, R., Olshen, P., Stone. Classification and Regression Trees. Wadsworth, Belmont, CA., 1984.

[28] J. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos, CA., 1992.

[29] R. Burbidge, M. Trotter, B. Buxton, and S. Holden, Drug design by machine learning: support vector machines for pharmaceutical data analysis, Computers and Chemistry 26, 5–14, 2001.

[30] J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis Cambridge University Press, 2004.

[31] Steinwart and A. Christmann, Support Vector Machines, Springer, 2008.

[32] D. Fogel, Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. IEEE Press, Piscataway, NJ, 2nd, 1999.

[33] L. de Castro, Fundamentals of Natural Computing: An Overview, Physics of Life Reviews (4) 1–36, 2007.

[34] D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, MA., 1989.

[35] L. Terfloth and J. Gasteige, Neural networks and genetic algorithms in drug design, DDT Vol. 6, No. 15 (Suppl.),2001.

[36] J. Liu and L. Zhou, QSAR modeling for thiolactomycin analogues using genetic algorithm optimized artificial neural networks, Molecular Simulation, Vol. 33, No. 8, 629–638, 2007.

[37] D. Turner and P. Willett, Evaluation of the EVA descriptor for QSAR studies: 3. The use of a genetic algorithm to search for models with enhanced predictive properties (EVA_GA). J. Comput.-Aided Mol. Design 14, 1–21, 2000.

[38] M. Fernandez, · J. Caballero, L. Fernandez, and A. Sarai, Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM), Mol Divers, 15:269–289, 2011.

[39] Z. Zhou, Ensemble Methods Foundations and Algorithms, CRC Press, 2012.

[40] L. Kuncheva, Combining Pattern Classifiers Methods and Algorithms, John Wiley & Sons, 2004.

[41] S. Wold. and L. Eriksson, Statistical Validation of QSAR Results. in Waterbeemd, Han van de. Chemometric Methods in Molecular Design. Weinheim: VCH. pp: 309-318, 1995.

[42] V. Gillet. De Novo Molecular Design, in Evolutionary Algorithms in Molecular Design, by David E. Clark (Eds.), WILEY-VCH Verlag, 2000.

[43] C. Nicolaou, C. Kannas, and E. Loizidou. Multi-Objective Optimization Methods in De Novo Drug Design, in Mini-Reviews in Medicinal Chemistry, 2012.

AUTHOR PROFILES

Mohammed El-Telbany, Ph.D. was born in Dammitta, Egypt, in 1968. He received the B.S. degree in computer engineering and science from the University of Minufia in 1991 and the M.Sc. and Ph.D. degree in Computer Engineering, from Electronics and Communication Department, Cairo University, Faculty of Engineering, in 1997 and 2003 respectively. He has been an associative professor at the Electronics Research Institute. He has also worked at the ESA at European Space Research Institute (ESRIN), 1998-1999, Frascati, Italy, at the Faculty of Engineering, Al-Ahliyya Amman University, Jordan, 2004-2005, College of Computer Sciences, King Khalid University, KSA, 2005-2008, College of Computer Sciences, Taif University, KSA and College of Computer Sciences, Islamic University, KSA. He has been involved in the field of autonomous mobile robots and machine leaning. His previous research includes work on Evolutionary Computation and Forecasting. Current research includes work in robotics and machine learning, data mining, bioinformatics, chemo informatics and swarm intelligence.