International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307, Volume-4, Issue-ICCIN-2K14, March 2014

Validation of SOMFA using Data Mining Technique

Payal Pahwa, Manju Papreja

Abstract- In drug design, the investigation of properties of chemical compounds is the most important task. For determining the properties, the analysis of the existing data set is essential. Instead of describing individual molecules, in drug design, methods are used to characterize complete sets of chemical compounds and their relationship. Data mining analyzes large amount of data to obtain useful information leading to understanding of relationships within chemical compounds to extract "hidden" information for decision making. This paper describes the various data mining techniques used in Cheminformatics to analyze chemical data sets for molecular patterns, for extraction of relevant information and the production of reliable secondary information for drug discovery. Data mining techniques also helpful for construction of statistical model (QSAR Model) that is useful to testing of validation and reliability of Cheminformatics tools. Validation is a crucial element for design of any tool. The reliability of any tool/model depends on how well the tool can predict the activity of compounds outside the training set and reproduces the biological activity of compounds included in the model. In this paper we validate SOMFA tool by using QSAR model on dataset taken from literature.

Keywords (QSAR Model), QSAR, SOMFA.

I. INTRODUCTION

Cheminformatics is the use of computer and informational techniques, applied to a range of problems in the field of chemistry [1]. It is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information [2].It is a rapidly growing field, with a huge application potential. Chemistry has produced an enormous amount of data and this data avalanche is rapidly increasing. More than 45 million chemical compounds are known and this number is increasing by several millions each year [3]. Novel techniques such as combinatorial chemistry and high-throughput screening generate huge amounts of data. All this data and information can only be managed and made accessible by storing them in proper databases. That is only possible through Cheminformatics [4].It is process of gathering and systematic use of chemical information, and the use of these data to predict the behavior of unknown dataset [5]. In Chemistry the investigation of molecular structures and of their properties is one of most important task. So the task of Data Mining in chemical context is to evaluate "hidden" information, correlations and other systematic relationships between chemical compounds in a set of chemical data. Data mining is also applied on various computational methods that allow us to gain insights into the biological actions of chemicals by analyzing large amounts of data. The tools or techniques of data mining that are used in Cheminformatics include visualization, classical QSAR or Statistical analysis, clustering, Decision tree and Neural Networks.

II. DATA VISUALIZATION

According to Friedman (2008) “the main goal of data visualization is to communicate information clearly and effectively through graphical means”. Data visualization is the creation and study of the visual representation of data and information extracted for decision making[6].

Data visualization technique in Cheminformatics is used to visualize chemical structures, and organize chemical structures in visualization by structural similarity. This requires specialized extensions to visualization software or by working with structural descriptors such as fingerprints.

III. CLASSICAL QSAR OR REGRESSION ANALYSIS

Quantitative structure– activity relationship analysis is regression or classification based analysis mainly used in the chemical and biological sciences. The QSAR regression models relate a set of "predictor" variables (X) to the potency of the response variable (Y), while classification QSAR models relate the predictor variables to a categorical value of the response variable [7]. QSAR model summarizes relationship between chemical structures and biological activity in a data-set of chemicals and predict the activities of new chemicals. A QSAR model uses following mathematical formula

Activity=f(physiochemical properties and /or structural properties) + error

The error includes difference between actual and predicted values.

QSAR modeling produces predictive models derived from application of statistical tools used for correlating biological activity or properties in chemicals with molecular structure and/or properties. The QSAR modeling should lead to statistically robust and predictive models that will be capable of making accurate and reliable predictions of the modeled response of new compounds [8].

IV. CLUSTERING

Cluster analysis aims to divide a group of objects into clusters so that the objects within n a cluster are similar but objects taken from different clusters are dissimilar. Once a set of molecules has been clustered then a representative subset can be chosen simply by selecting one compound from each cluster [9]. Most clustering analysis methods are non-overlapping that is each object belongs to just one cluster. While in overlapping methods, an object can be

Manuscript received March 2014

Payal Pahwa, Bhagwan Parshuram Institute of Technology, Rohini (New Delhi), India.
Manju Papreja, GVM Institute of Technology and Management, Sonipat (Haryana), India.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

present in more than one cluster. The Key steps for cluster based compound selection are as follows:

i. Generate descriptors for each compound in the data set.
ii. Calculate similarity or distance between all compounds in the data set.
iii. Use cluster algorithm to group the compounds within the dataset.
iv. Select a representative subset by selecting one or more compounds from each cluster.

Also, clustering cannot tolerate the heterogeneity of the data. This makes one turn to partitioning approaches [10].

V. NEURAL NETWORKS

Neural networks are powerful data mining tools with a wide range of applications in drug design. Sometimes the statistical methods are unsuccessful to solve chemical problems, then artificial neural networks can be used for analyzing non-linear and complex relationships between descriptors [11]. The neural networks are self-adaptive auto-associative systems, i.e. they learn by processing a set of training data about the relationships within this data set. The important tasks for neural networks in Data Mining are:

i. Classification: i.e. assigning data to predefined categories
ii. Modeling: Describing complex relationships between data by mathematical functions;
iii. Auto-Association: extrapolation and prediction of new data using already learned relationships.

VI. DECISION TREE

Decision trees, also known as partitioning algorithms are non-parametric approaches. It is difficult for regression or parametric classification approaches to work on heterogeneous types of data. The excessively large number of descriptors can make clustering computation infeasible [12]. Decision trees are introduced to solve these problems. One of the most popular decision tree techniques is recursive partitioning (RP). It has been reported that RP algorithms can partition on data sets with over 100,000 compounds and 2,000,000 descriptors, in less than an hour [13-14]. RP algorithms can also be used to build multivariable regression models. One of the disadvantages of the decision tree approach is similar to a problem with the clustering algorithm approach, namely: it suggests too many solutions.

Comparison of Cheminformatics Datamining techniques

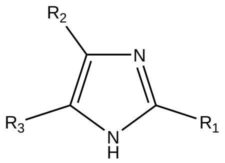
Method	Remarks
Regression or QSAR	Regression methods are the most traditional approaches for pattern recognition. These methods assume the variables are continuous and the curve shapes are pre-defined. For multidimensional data, curve patterns are not known and trying all possible curves is very time consuming. In these cases, genetic algorithms may be applied to partially solve the problem of identifying curve patterns.
Decision tree classification	This approach is applied when there are a great number of descriptors and, the descriptors have various value types and ranges.
Hierarchical clustering	This approach assumes the objects have hierarchical characters. The methods require similarity or distance matrices. The approach may produce multiple answers for users to explain or with which to experiment.
Nonhierarchical clustering	The approach assumes the objects have nonhierarchical characters, and the number of clusters is known prior the computation. The method requires similarity or distance matrices. The approach may produce multiple answers for users to explain or with which to experiment.
Neural Networks	Neural networks are model-free mapping devices that are capable of capturing complex nonlinear relationships in the underlying data that are often missed by conventional QSAR approaches. However, neural networks are known to be unstable, in the sense that minor changes in the training data and/or training parameters can have serious consequences in the generalization ability of the resulting models

VII. VALIDATION OF SOMFA BY QSAR-MODEL

With increase use of Cheminformatics, now days, data mining techniques are used for the development of relationships between physicochemical properties of chemical substances and their biological activities to obtain a reliable statistical model(QSAR model) for prediction of the activities of new chemical entities. Self-organizing molecular field analysis (SOMFA) is a new field based tool for drug design discovered by Robinson and Co-workers (2000). Quantitative structure activity relationship (3D-QSAR) model is used here to test the validity of SOMFA tool on dataset taken from literature. It uses intrinsic molecular properties, such as the molecular shape and electrostatic potential, which are used to develop the QSAR models (15).

The main objective of our present 3D-QSAR study is to get a validated correlation between the structural features of glucagon receptor inhibitors and triarylimidaozle activities. For developing a statistically validated reliable model for, a dataset of 27 compounds was taken from the literature and used for 3D-QSAR study using SOMFA.

Table 1: Structure of triarylimidazole derivatives



No.	R_1	R_2	R_3
1	(4-Br)Ph	(4-F)Ph	4-pyridyl
2	(3-Br)Ph	(4-F)Ph	4-pyridyl
3	(4-Cl)Ph	(4-F)Ph	4-pyridyl
4	(4-F)Ph	(4-F)Ph	4-pyridyl
5	(4-I)Ph	(4-F)Ph	4-pyridyl
6	(4-Me)Ph	(4-F)Ph	4-pyridyl
7	(4-iPrPh	(4-F)Ph	4-pyridyl
8	(4-Ph)Ph	(4-F)Ph	4-pyridyl
9	(4-NH_2)Ph	(4-F)Ph	4-pyridyl
10	(4-OMe)Ph	(4-F)Ph	4-pyridyl
11	(4-CNPh	(4-F)Ph	4-pyridyl

12	(4-COOMe) Ph	(4-F)Ph	4-pyridyl
13	(4-SMe)Ph	(4-F)Ph	4-pyridyl
14	(4-Br)Ph	Ph	4-pyridyl
25	(4-Cl)Ph	(4-F)Ph	3-Me(4-pyridyl)
16	(4-Cl)Ph	(4-Cl)Ph	4-pyridyl
17	(4-Cl)Ph	(4-I)Ph	4-pyridyl
18	(4-Cl)Ph	(4-Ph)Ph	4-pyridyl
19	(4-Cl)Ph	(4-t-Bu)Ph	4-pyridyl
20	(4-Cl)Ph	(4-n-Bu)Ph	4-pyridyl
21	(4-Cl)Ph	(3-Ph)Ph	4-pyridyl
22	(4-Cl)Ph	(2-OPh)Ph	4-pyridyl
23	(4-Cl)Ph	(3-OPh)Ph	4-pyridyl
24	(4-Cl)Ph	(4-OPh)Ph	4-pyridyl
25	(4-Cl)Ph	(2O-n-Bu)Ph	4-pyridyl
26	(4-Cl)Ph	(2,4-(O-n-Pr)$_2$)Ph	4-pyridyl
27	(4-Cl)Ph	(2,4-(O-n-Bu)$_2$)Ph	4-pyridyl

Table 2: Actual and Predicted activities for Training and Test set molecules from SOMFA model:

No	Actual Activity (pIC$_{50}$)	Predicted Activity	Residual Activity
1	6.568	6.050	0.491
2^T	5.853	6.045	-0.156
3	6.398	6.067	0.282
4	5.699	6.042	-0.371
5^T	6.292	6.050	0.197
6	5.886	6.016	-0.187
7	6.155	5.944	0.216
8	5.000	5.901	-0.946
9	5.699	6.077	-0.361
10^T	4.886	5.953	-1.094
11	5.097	5.949	-0.914
12	5.06	5.724	-0.758
13	6.31	5.907	0.352
14	6.107	6.420	-0.027
15^T	5.959	6.181	-0.489
16	6.721	6.126	0.464
17	6.886	6.277	0.644
18	6.854	6.588	0.236
19	6.886	6.716	0.292
20^T	7.131	6.665	0.605
21	7.215	6.735	0.551
22	**8.187**	**8.242**	**-0.065**
23	7.886	7.391	0.22
24	7.569	7.428	0.072
25	8.071	7.183	1.107
26	7.886	8.684	-0.700
27	8.187	8.861	-0.718

T- Test Set Molecules

Table 3: Statistical results of 3D-QSAR studies

Parameter	**Resolution (0.5 Å)**
q^2	0.6911
r^2	0.7197
S	0.5541
F	51.3441
r^2_{pred}	0.6731
Contributions	*Shape 52% Electrostatic 48%*

q^2: cross-validated correlation coefficient by leave one out method; r^2: conventional correlation coefficient; S: standard error of estimate; F: Fisher Test value; r^2_{pred}: Correlation coefficient f or prediction (test) set

The r^2_{cv} (q^2) can take up values in the range from 1, suggesting a perfect model, to less than 0 where errors of prediction are greater than the error from assigning each compound mean activity of the model. Fischer Statistics (F-Test) is another useful parameter to check the statistical reliability of a model. The larger is the value of F, the greater the probability that the QSAR models will be statistically significant. The statistical results, cross-validated r2 cv and non cross-validated r2, F-Test value showed a satisfied predictive ability (r2 pred) obtained from SOMFA that validates the tools.

VIII. CONCLUSION

In summary, we have developed a predictive QSAR model using data mining techniques to validate SOMFA tool on Glucagon Receptor Inhibitors and Triarylimidaozle activities evidenced by statistical measures. The statistical results, cross-validated $r2_{cv}$ and non cross-validated r_2, F-test value showed a satisfied predictive ability (r_{2pred}) from SOMFA indicating usefulness of this tool for the design of drug candidate.

REFERENCES

[1] V. Umashankar, S. Gurunathan, “Chemoinformatics and its Applications” General, Applied and Systems Toxicology, John Wiley & Sons, 2009.’

[2] N. Prakash, D. A. Gareja, “Cheminformatics” J Proteomics Bioinform vol. 3, pp. 249-252, 2010

[3] S. A. Bhalerao, D. R. Verma, R. L. D’souza, N. C. Teli, V. S. Didwana, “Chemoinformatics: The Application of Informatics Methods to Solve Chemical Problems” Research Journal of Pharmaceutical, Biological and Chemical Sciences, vol.4, pp. 475-499, 2013.

[4] V. S. Velingkar, G. Pokharna, N. S. Kolhe, “Chemoinformatics: A Novel tool in drug discovery” Int. J. Curr. Pharm. Res, vol. 3, pp. 71-75, 2007.

[5] Available online: http://www.rsc.org/images/valgillet_tcm18-46919.pdf

[6] Diana Burley, “Information isualization As A Knowledge Integration Tool”, Journal of Knowledge Management Practice, Vol. 11, 2010.

[7] Available online: http://en.wikipedia.org/wiki/Quantitative_structure-activity_relationship

[8] E.C.Ibezim, P.R .Duchowicz, N.E. Ibezim, “Computer-Aided Linear Modeling Employing Qsar for Drug Discovery”, African Journal of Basic & Applied Sciences vol.1(3-4), pp.76-82,2002

[9] Lior Rokach , Oded Maimon , “CLUSTERING METHODS”, data mining and knowledge discovery handbook,available online :http://www.ise.bgu.ac.il/faculty/liorr/hbchap15.pdf

[10] Jun X., Hagler.A, “Chemoinformatics and Drug Discovery”, Molecules vol.7, pp. 566-600,2002.

[11] Markus C. H., Gasteiger.J, “Data Mining in Chemistry”, Proceed. Chem. Inf. Conf., H. Collier (Ed.), Infonortics Ltd. Calne, UK, pp.1-6,1997

[12] Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. “Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning”, J. Chem. Inf.Comput. Sci., 1999, 39, 1017-1026.

[13] Rusinko, A., III; Young, S. S.; Drewry, D. H.; Gerritz, S. W. “Optimization of Focused Chemical Libraries Using Recursive Partitioning”, Comb. Chem. High T. Scr., vol.5, pp. 125-133,2002.

[14] W. L. Chen, “Chemoinformatics: past, present, and future. Journal of Chemical Information Model” vol. 46, pp. 2230-2255, 2006.