# Mutagenicity analysis based on Rough Set Theory and Formal Concept Analysis

Mostafa A. Salama, Mohamed Mostafa M. Fouad, Nashwa El-Bendary, Aboul Ella Hassanien

**Abstract** Most of the current Machine Learning applications in cheminformatics are black box applications. Support vector machine and neural networks are the most used classification techniques in prediction of the mutagenic activity of compounds. The problem of these techniques is that the rules/reasons of prediction are unknown. The rules could show the most important features/descrpitors of the compounds and the relations among them. This article proposes a model for generating the rules that governs prediction through the rough set theory. These rules, which based on two levels of selection for the highly discriminating power features, are visualized by lattice generated using the formal concept analysis approach. That is, better understanding of the reasons that leads to the mutagenic activity can be obtained. The resulted lattice shows that lipophilicity, number of nitrogen atoms, and electronegativity are the most important parameters in mutagenicity detection. Moreover, experimental results are compared against previous researches for validating the proposed model.

Mostafa A. Salama
British University in Egypt (BUE), Cairo - Egypt,
Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net
e-mail: mostafa.salama@gmail.com

Mohamed Mostafa M. Fouad
Arab Academy for Science,Technology, and Maritime Transport, Cairo - Egypt,
Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net
e-mail: mohamed_mostafa@aast.edu

Nashwa El-Bendary
Arab Academy for Science,Technology, and Maritime Transport, Cairo - Egypt,
Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net
e-mail: nashwa.elbendary@ieee.org

Aboul Ella Hassanien
Information Technology Dept., Faculty of Computers and Information, Cairo University,
Cairo - Egypt,
Scientific Research Group in Egypt (SRGE), http://www.egyptscience.net
e-mail: aboitcairo@gmail.com

# 1 Introduction

Studying the role of chemical compounds in the biological systems has been further strengthened. The prediction of the effect these compounds on humans, and animals is one of the benefits that could be resulted from this study. An approach is proposed to find the common patterns, similarities, between compounds having the same reactions, but this solution is an NP-complete problem [1]. Chemists provides a set of properties, descriptors, that describes the structure of the chemical compounds like the molecular type, atomic type, and bond type [2]. These descriptors are categorized as constitutional, topological, geometrical, charge related, semi-empirical, and thermodynamic [3]. Several approaches have been proposed to do a matching between these descriptors and the activity of the corresponding chemical compounds. One of the problems that faced these approaches is the high number of molecular descriptors. The main concern of these approaches is the emerging and evolving of a huge number of molecular descriptors. Therefore, most of them are selecting a subset of these descriptors. The selection process reduces the computational complexity needed for too many descriptors, and also removes the redundancy of information. A main step in cheminformatics is to reduce the number of descriptors in order to apply the prediction of the activity of the compounds. The next step is to apply a classification technique that uses a set of compounds of known activity as a training set, then predict the activity of new compounds of un-known activity.
In order to apply an accurate feature selection of the descriptors and an accurate analysis of the mapping between the descriptors and the right decision, a new model should be proposed that differs from the classical mode.

This proposed model should handle all the features together when applying the feature selection technique, this to make use of the dependency among features. On the other hand, this model should clearly interprets the relationship among these features in a visual illustration for chemists. In this paper other machine learning techniques like descriptor/feature selection, rule generation and visualization techniques are applied. The descriptor/feature selection technique shows the most important descriptors that discriminate between the different activities of the drugs, for example, the difference between the mutagen and non-mutagen activity of drugs. Feature selection techniques select the features whose distribution correlates the distribution of the class labels. The rule generation techniques, shows the rules that governs the prediction of the aspects of the drugs. Finally the visualization technique shows the relation between the different descriptors and each other and between these descriptors and the predicted aspects.

The model proposed in this paper applies two phases of feature selection, the first phase uses a classical ChiMerge feature selection technique, while the second phase applies the rough set technique. This will compel a highly filtration of the unneeded descriptors and on the other hand, the rules will describe the relation among these important selected descriptors. The rest of this paper is organized as follows. Section 2 presents a machine learning based model for visualization. Section 3 presents experimental results, Section 4 presents analysis and discussion. Conclusion and future work are discussed in section 5.

## 2 Visualization Model

Machine learning and data visualization has a great contribution in knowledge discovery. The number of descriptors can be considered in cheminformatics is too high, it can reach 6122 descriptor like those extracted by PowerMV software [4] as shall be discussed later within the context. The corresponding expression of descriptors in machine learning is called features or attributes, and the effect (activity) of each chemical compound is called the target class. An example of this effect is whether this chemical compound is mutagenic or non-mutagenic. Set of compounds with a known target class are used as a training set, while other set of compounds will be test whether the prediction of the target class are correct or not. The prediction here will be refereed as the classification of the compounds in the training set.

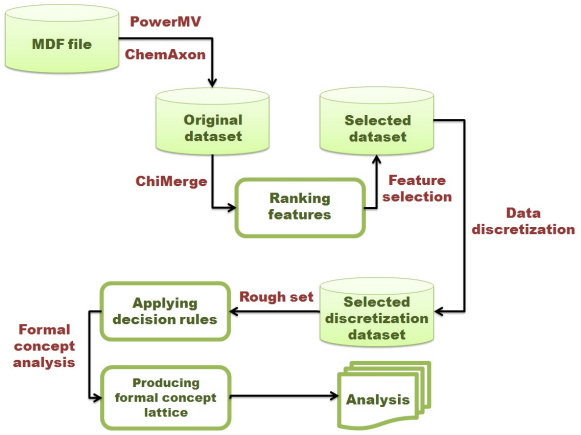


**Fig. 1** Model of visualization of Mutagenicity in chemical compounds

Figure 1 shows the proposed model that illustrates the steps required for rule generation and visualization. First of all, the huge number of descriptors prevents an accurate analysis and prediction of the target class. This problem is named as the curse of dimensionality [5], where the data set contains high number of features (words dimensions). Therefore a forward feature selection technique is applied, where the features are ranked according to its importance in the discrimination between the different target classes. The ChiMerge technique is the most applicable ranking techniques, as it is capable of handling continuous data set [6]. Then these features are ordered, and tested in a series of tests, first test will be applied on data set of the highest feature, then applied on a data set of the highest two features, and so on until this data will include all the ranked features. The testing of the set of features selected is applied using any classification technique like Naieve Bayesian Tree (NBT) classifier. Only the set features that leads to the local and global maximum success in the tests will be selected [7]. This will shrink the original descriptors data set into a set of high level features those required to predict the target class. This

shrinking process reduces the processing time required for knowledge learning process as well as enhancing the accuracy of the prediction of the target class of new compounds. The second step is to state the relation among these attributes through the extraction of rules governing this relation using the rough set theory. In order to apply rough set on the reduced data set, a discretization method is applied. The discretization method involves the creation of cuts dependent on the target class, in order to simplify the rule extraction and class prediction. Finally, the rules are extracted according to the rough set technique [8] to describe the range of the values of some/all features that indicates the corresponding predicted class. The third step is to visualize the relation between attributes. One of the visualization techniques applied is the formal concept analysis to generate a formal concept lattice [9]. The formal concept analysis is usually applied on the original highly dimensional data set. This leads to a high dense, unclear and non-informative formal concept lattice. Here, the formal concept analysis is applied on the generated rules from the rough set technique [12]. For each class, each extracted rule contains a set of attributes' range that leads this class. A list of all attributes' ranges is prepared, and for each rule, the attribute range that included in this rule is marked by one, otherwise it is marked by zero. This will generates a binary data for each class, the data is provided to the formal concept analysis technique for the generation of the formal concept lattice. The visualization of the rough set rules provides a simple and more descriptive formal concept lattice. Finally this formed lattice can be provided to the domain expert for analysis. This simplifies his task through the visualization of the relation among attributes for each target class.

## 3 Experimental results

The Bursi Mutagenicity Dataset [10] were used in this paper. It consists of 2401 mutagenic compound and of 1936 non-mutagenic compound. In order to extract the descriptors of these compounds, two software are used the PowerMV software and the ChemAxon software [11]. PowerMV calculates a total of 6122 descriptors classified as 546 atom pair descriptors, 4662 Carhart descriptors, 735 fragment pair descriptors, 147 pharmacophore fingerprints, 24 Weighted Burden Number descriptor and 8 properties descriptors. ChemAxon also have a various kinds of descriptors, including structural and topological analyzer descriptors. Due to the huge number of descriptors extracted by PowerMV, a feature ranking and selection techniques are applied to select the most important features. This applied on only 200 compounds, divided equally between the two class. After applying ChiMerge technique, the WBN_EN_H_0.75 $WEN$, WBN_LP_H_0.75 $WLP$ descriptors shows the highest ranks. When a test is applied on the 200 compounds data set, and based on these two descriptors only, the classification accuracy is 78.0% success. When the third ranked descriptor is applied the classification accuracy result is decreased. Therefore, the only two selected descriptors are the $WEN$ and $WLP$, descriptors. These two descriptors are from the 24 Weighted Burden number-continuous descriptors.

They refer to the electronegativity and the atomic lipophilicity property respectively. Then, the rest of the descriptors are extracted from the ChemAxon software. The number of extracted descriptors from the ChemAxon software are 64 descriptor. The used descriptors are as shown in Table 1.

**Table 1** The extracted descriptors for the mutagenic data set

| Descr. | Description | Descr. | Description | Descr. | Description | Descr. | Description |
|---|---|---|---|---|---|---|---|
| *MW* | Molecular Mass | *MEM* | Exact Mass | *nBT* | #Bonds | *nAT* | #Atoms |
| *nN* | #N Atoms | *nS* | #S Atoms | *IMD* | Molecular Dim. | *nH* | #H Atoms |
| *nP* | #P Atoms | *HRC* | #Heterogenous | *nO* | #Oxygen Atoms | *nC* | #Carbon Atoms |
| *nBR* | #Br Atoms | *RC* | #Rings | *RBC* | #bonds in Rings | *nF* | #F Atoms |
| *ALAC* | #Aliphatic Atom | *TC* | Total Charge | *BL* | Avg. bond len. | *RAC* | #atoms in Rings |
| *ALBC* | #Aliphatic Bond | *nDB* | #double Bonds | *RBN* | #Rotatable Bond | *nCL* | #Cl Atoms |
| *ALRC* | #Aliphatic Ring | *ARAC* | #Aromatic Atom | *nAB* | #aromatic Bonds | *ARBC* | #Aromatic Bond |
| *ARRC* | #Aromatic Ring | *ASAC* | #Asymmetric | *wP* | wiener Polarity | *wI* | wiener Index |
| *CRC* | #Carbo Ring | *rI* | randic Index | *hwI* | hyperWiener Ind. | *nTB* | #trible Bonds |
| *sI* | szeged Index | *CBC* | #Chain Bond | *cN* | #cyclomatic | *CAR* | #Carbo-aromatic |
| *CAC* | #Chain Atom | *CCC* | #Chiral Center | *pI* | platt Index | *CAL* | #Carbo-aliphatic |
| *hI* | harary Index | *DB* | Double Precision | *fC* | #fragment | *bI* | balaban Index |
| *FAL* | #Fused Aliphatic | *FC* | #Fragment | *FAR* | #Fused Aromatic | *FRC* | #Fused |

The rest of the descriptors describe the number of different types of bonds like the Number of AROMATIC Bonds Between "c and o" *nAbBco*, the Number Of Single Bonds Between "c and c" *n*1*bBcc*, and the Number Of double Bonds Between "c and o" *n*2*bBco*.

After the combination of these 64 descriptors to the selected descriptors from PowerMV descriptors, the data set resulted is now 66 features data set. Again the feature selection technique is applied on this data set. Figure 2 shows the classification accuracy of increasing set of features according to the ascending ranked features. The feature numbers are corresponding to the order of the maintained features. The selected descriptors from these set of features from both software are only 22 descriptors. When applying different classification techniques on the selected 22

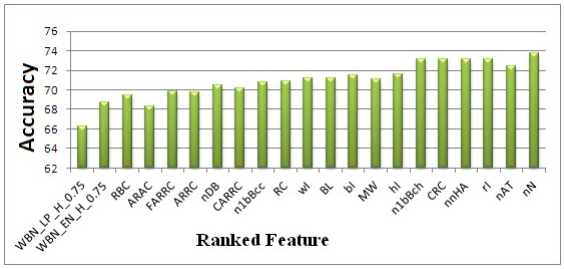


**Fig. 2** Forward feature selection technique selectees 22 features out of 66 features

features data set, the resulted classification accuracy varies according to the used

technique. A 10-fold cross validation method is used to have an average results that indicate the accuracy of the used classifier. Also, as discussed before, the input data set is nearly balanced among the two target classes. Naive Bayesian Tree classifier is used in the forward feature selection technique, the resulted classification accuracy of the selected data set is 75.78%. The resulted classification accuracy of other classifiers are as follows, Support Vector Machine shows 62.02%, Decision Tree shows 76.45% accuracy, the Lazy Classifier, Ib1, shows 77.15%, the Random Forest Classifier shows 77.28 % and finally Multi-Layer Perceptron Classifier shows 73.18% accuracy. The previous tests are performed using the WEKA software [13]. Then, rough set theory is applied on the 22 descriptors data set. The generated rules from the rough set theory is dependent only on 12 features. Those features are the most discriminating ones for compound classification whether it is mutagenic or not. These features are $WLP$, $WEN$, $CAR$, $nN$, $nAT$, $hl$, $bI$, $wl$, $nDB$, $rl$, $n1bBch$, $CRC$. The number of the generated rules for both mutagen and non-mutagen classes are more than 1000 rule. Each rule is applied to a number of compounds that varies according the covering of this rule. The following statement is an example of the generated rules:
(WLP=”(3.205, 3.318)”)AND(bI=”(1.407, 1.541)”)
$\Rightarrow$ (class = mutagenicity ) {covered by 71 compound}

Where this rule consists of two subrules, the first subrule stated that the value of the $WLP$ descriptor lies between the two values 3.205 and 3.318. The second subrule stated that the value of the bI descriptor lies between the values 1.407 and 1.541. If these two subrules are achieved, then the class of this compound is mutagenicity. This rule covered 71 compound and not leads to any compound whose class is non-mutagenicity. In order to visualize these rules a smaller subset of these rules is used, these selection of these rules is based on the highest number of compounds covered by each rule. Only 20 rules are selected for each class to simplify the visualization of relation and dependencies between descriptors. The visualization applied is based on the formal concept analysis which accepts only binary input. To convert this subset of rules into binary input to the formal concept analysis, the ranges of the 12 features will be tested either achieved or not for each rule. For example in the previously maintained rule, the range 3.205 and 3.318 of feature $WLP$ is stated to 1 and the range (1.407 and 1.541 is stated to 1, while the rest of feature ranges will be assigned 0. The result of the formal concept analysis is a formal concept lattice. Two lattices are generated, Figure 3 shows the relationship between descriptors of mutagenicity class whereas Figure 4 illustrates the non-mutagenicity class' descriptors.

## 4 Observations from Visually Generated Lattices

Figures 4 and 3 show the relation among these descriptors, the degree of effectiveness of each one on the others, and the range of values for each one. $WLP$ and $nN$

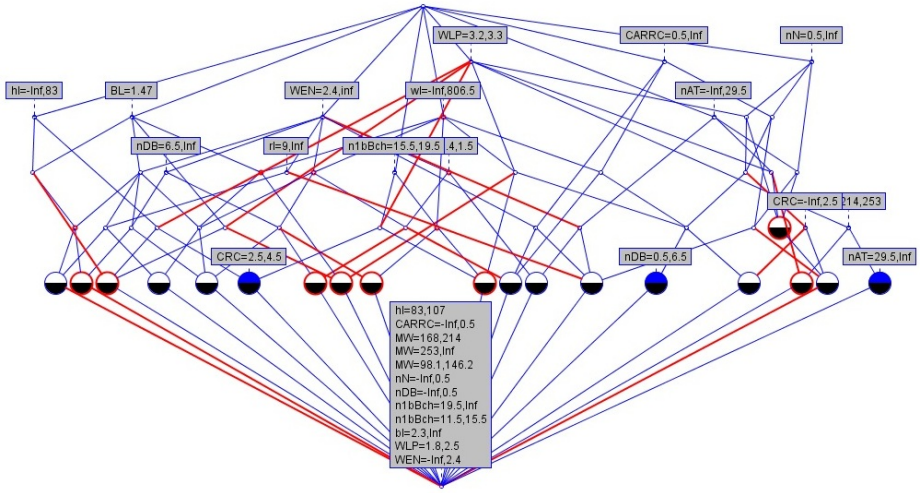


**Fig. 3** Formal concept lattice for mutagenic data

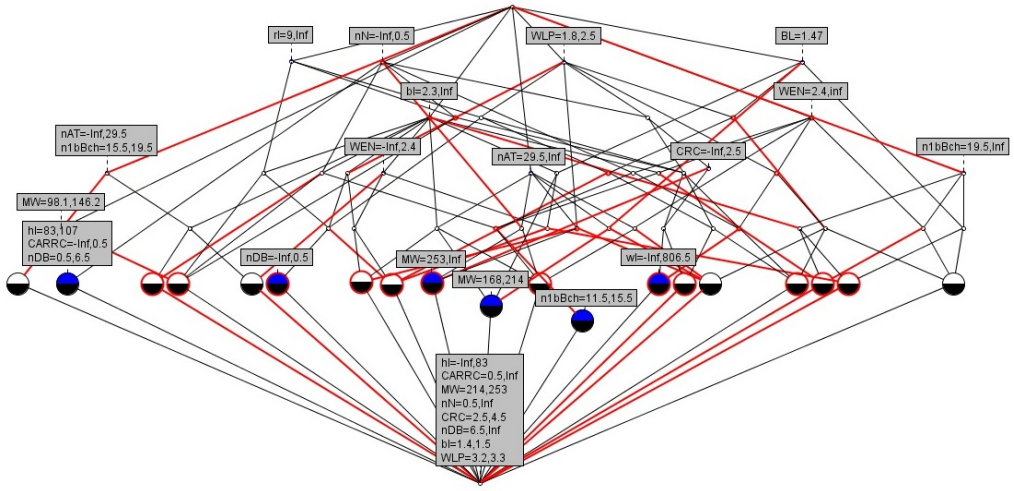


**Fig. 4** Formal concept lattice for non-mutagenic data

descriptors appears as the highest important descriptors for the discrimination between the two mutagen and non-mutagen classes. The two range of values of these two descriptors are different from each other in the two classes. This shows a connection between atomic lipophilicity and the number of nitrogen, and a connection between these two descriptors and the mutagenicity of the chemical compounds. It appears that research, since 1960s, shows an importance of lipophilicity for biological potency. It was thought that lipophilicity is an important descriptor in the QSAR research and required especially in the drug design [14]. On the other hand

it appears that a good correlation exists between the resistance reversal activity and the intrinsic basicity of the nitrogen atom at the tricyclic ring system, frontier orbital energies, and lipophilicity [15]. This provides a proof that the connections between these two descriptors in the generated lattices is correct. Also, it gives an evidence that the influence of these two descriptors on the mutagenicity predication, as appeared in the generated formal concept lattice is correct. On the level of the values, Figures 4 and 3 show that as the values of the $WLP$ and $nN$ descriptors increases, as the compound becomes more mutagenic. This is because the ranges of $WLP$ and $nN$ descriptors in the case of mutagenic compounds are [3.2, 3.3] and [0.5, $infinity$] respectively. Another descriptors that show a high discrimination between the two classes is $WEN$. The range of values of the $WEN$ descriptor in the case mutagenic class is [2.4, $infinity$]. While in the case of non-mutagenic compounds, the range of values of $WEN$ descriptor is [-$infinity$, 2.4]. This means that as the value of the electronegativity of the compound $WEN$ increases, as the possibility of the compound to be mutagenic increases. A study in 1995 shows that more electronegative compounds are more reactive to DNA and more capable of inducing mutation than other compounds [16]. Again, this is another proof that the generated lattices are correct. Also, there is a high correlation between the electronegativity and the number of atoms $nAT$ which appears in the 12 selected features. The generated lattices show that in case of mutagenic class, this descriptor $nAT$ range is [-$infinity$, 29.5], while in case of non-mutagenic class objects, the range is [29.5, $infinity$]. This means that as the number of atoms in the compound decreases as the tendency of the compound to become non-mutagenic increases, and this is corresponding to its correlation to the electronegativity descriptor. Resistance reversal activity and intrinsic basicity of the nitrogen atom at the tricyclic ring system has a relation to the lipophilicity [15], while lipophilicity has a direct effect on the mutagenicity. A research in 2005 shows that the nitrogen-substituted position in the chrysene molecule, has a direct effect on the mutagenic activity. The ratio of participation of the metabolic activation enzyme isoforms of cytochrome is influenced by the differences in these positions [17].

## 5 Conclusions

The massive data generated from cheminformatics descriptors deteriorates the quality of the results' analysis. The proposed model applies different layers of filtration of the generated descriptors to come out with a small set of the highest effective descriptors. After applying feature selection and rule generations, the resulted set of descriptors includes the lipophilicity, the number of nitrogen atoms and electronegativity. These descriptors are the most important parameters required to detect mutagenicity. The generated lattice by the formal concept analysis shows a set of facts that were proved separately in previous medical researches. This proves the quality and preciseness of the proposed model.

## References


1. N. Brown, Chemoinformatics : an introduction for computer scientists, ACM COMPUTING SURVEYS 41(2) (2009) Article 8.
2. J. Xu and A. Hagler, Chemoinformatics and Drug Discovery, Molecules 7 (2002) 566-600.
3. A.R. Katritzky, L. Pacureanu, D. Dobchev, and M. Karelson, QSPR Study of Critical Micelle Concentration of Anionic Surfactants Using Computational Molecular Descriptors, Journal of Chemical Information and Modeling 47(3) (2007) 782-793.
4. K. Liu, J. Feng, and S.S. Young, PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation, Journal of Chemical Information and Modeling 45(2) (2005) 515-22.
5. M.A. Salama, N. El-Bendary, A.E. Hassanien, K. Revett, A.A. Fahmy, Interval based attribute evaluation algorithm, In: Proc. The Federated Conference on Computer Science and Information Systems, FedCSIS 2011, Szczecin, Poland, pp. 153-156.
6. F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, Naieve Bayesian based on Chi Square to Categorize Arabic Data, In: Proc. The 11th International Business Information Management Association Conference, IBIMA, on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, 2009, pp. 930-935.
7. H.F. Eid, M.A. Salama, A.E. Hassanien, and T.-H. Kim, Bi-Layer Behavioral-Based Feature Selection Approach for Network Intrusion Classification, In: T.-H. Kim, H. Adeli, W.-C. Fang, L.J. Garca-Villalba, K.P. Arnett, M.K. Khan (Eds.), Proc. Security Technology - International Conference, SecTech 2011, Part of the Future Generation Information Technology Conference, FGIT'11, Jeju Island, Korea, 2011, pp. 195-203.
8. H. Al-Qaheri, A.E. Hassanien, A. Abraham, A Generic Scheme for Generating Prediction Rules Using Rough Sets, In: A. Abraham, R. Falcan, R. Bello (Eds.), Springer, Rough Set Theory: A True Landmark in Data Analysis, Studies in Computational Intelligence 174 (2009) 163-186.
9. S. Motameny, B. Versmold, and R. Schmutzler, Formal Concept Analysis for the Identification of Combinatorial Biomarkers in Breast Cancer, Lecture Notes in Computer Science 4933 (2008) 229-240.
10. J. Kazius, R. McGuire, R. Bursi, Derivation and Validation of Toxicophores for Mutagenicity Prediction. J. Med. Chem. (2005), 48(1), 312-320.
11. ChemAxon Software: available on: http://www.chemaxon.com/, [Last accessed: January 2013].
12. S.O. Kuznetsov, Machine Learning and Formal Concept Analysis, Lecture Notes in Computer Science 2961 (2004) 3901-3901.
13. WEKA: Waikato Environment for Knowledge Analysis, version 3.5.9, available on: http://www.cs.waikato.ac.nz/ml/weka/, [Last accessed: January 2013].
14. Q. Du, P.G. Mezey, and K.C. Chou, Heuristic Molecular Lipophilicity Potential (HMLP): A 2D-QSAR Study to LADH of Molecular Family Pyrazole and Derivatives, Journal of Computational Chemistry 26(5) (2005) 461-470.
15. A.K. Bhattacharjee, D.E. Kyle, J.L. Vennerstrom, and W.K. Milhous, A 3D QSAR pharmacophore model and quantum chemical structure–activity analysis of chloroquine(CQ)-resistance reversal, Journal of Chemical Information and Computer Sciences 42(5) (2002) 1212-1220.
16. H.S. Rosenkranz and G. Klopman, Relationships between electronegativity and genotoxicity, Mutation Research 328(2) (1995) 215-227.
17. Yamada K, Hakura A, Kato TA, Mizutani T, Saeki K., Nitrogen-substitution effects on the mutagenicity and cytochrome P450 isoform-selectivity of chrysene analogs, Mutat Res. 2005 Sep 5;586(1):87-95.