PJHRD Philippine Journal of Health Research and Development

Molecular Descriptors for Drugs: A Discriminant Analysis

Liza T. Billones*, Alex C. Gonzaga, Junie B. Billones

*Corresponding author's email address: ltbillones@up.edu.ph

Department of Physical Sciences and Mathematics, College of Arts and Sciences, University of the Philippines Manila, Padre Faura, Ermita, Manila 1000 Philippines

RESEARCH ARTICLE

Abstract

Background: The biological activity of a compound is assumed to be encoded in its chemical composition and geometric structure, from which physico-chemical, electrotopological, and graph theory-derived properties can be determined.
Objective: This study aimed to identify the molecular descriptors derived from Dragon® 6 software that can discriminate compounds as drug or nondrug.
Methodology: Over 4000 molecular properties were obtained for approximately 2000 known drugs and 2000 nondrugs on which Linear Discriminant Analysis was performed.
Results: Compounds can be discriminated between drug and nondrug with 81% accuracy using only two molecular descriptors, the information index HVcpx and the topological index MDDD.
Conclusion: A "Rule of Three" (HVcpx ≤ 3 and MDDD ≥ 30) seems to confer druglikeness in compounds. This rule can be used as additional filter in high throughput screening of compounds in any drug discovery research.

Keywords: *Dragon®descriptors, discriminant analysis, druglikeness, topological, information index, drug discovery*

Introduction

The properties of small molecules have been analyzed in efforts to find the essential factors required to produce good lead compounds in drug discovery [1,2]. An illustrious development in medicinal chemistry is the seminal paper of Lipinski on Rule of Five (RO5) which characterizes most orally bioavailable drug candidates [3,4]. The original RO5 covers orally active compounds and specifies critical range of values for four simple physico-chemical parameters (i.e. molecular weight, MW ≤ 500; octanol-water partition coefficient, log P ≤ 5, hydrogen-bond donors, HBD ≤ 5, hydrogen-bond acceptors, HBA ≤ 10) satisfied by 90% of orally active drugs that have reached the phase II clinical stage. The RO5 has been modified by Veber and co-workers, who discovered that the optimum number of rotatable bonds (NROT) is 7 and that the NROT must not exceed 10 for a compound to display good oral bioavailability [5]. Clark and Pickett also demonstrated that polar surface area (PSA) is another key property [6]. They proposed that the PSA should not exceed 140 Å2 to avoid the problem of low oral bioavailability.

Moreover, several accounts are confronting the issues facing the compounds that are identified by screening of small molecule libraries. One novel alternative approach that is gaining wider acceptance is 'fragment-based' discovery [7,8,9]. In fragment-based lead discovery, small chemical fragments are allowed to weakly bind to the biological target and then are allowed to grow or joined together to produce a lead with higher affinity. The hits identified in this method generally obey a 'Rule of Three', which could be utilized in the construction of fragment libraries for lead generation. Congreve and co-workers carried out an analysis of a diverse set of fragment hits that were identified against a range of targets and found that the hits seemed to follow a 'Rule of Three' in which molecular weight was ≤ 300, the number of hydrogen bond donors was ≤ 3, the number of hydrogen bond acceptors was ≤ 3 and ClogP was ≤ 3, number of rotatable bonds, NROT (≤ 3) and polar surface area, PSA (≤ 60). These findings indicate that a 'Rule of Three' could be used to speed up the screening of fragments for efficient lead discovery.

Aside from the abovementioned simple parameters associated with oral bioavailability, an enormous number of physico-chemical, electrotopological and graph theory-

derived molecular descriptors [11] can be generated from the composition and structure of a compound. A cheminformatics software DRAGON® [12-13], for example, is capable of generating over 4000 molecular descriptors per molecule. The descriptors can be classified as constitutional (0D) properties; 1D descriptors (e.g. functional groups, atom centered fragments, information and properties descriptors; 2D descriptors (e.g. topological, molecular walk counts, Burden eigenvalues, eigenvalue-based indices, topological charge indices, connectivity, edge adjacency and 2D autocorrelation descriptors); and 3D descriptors namely, charge, Randic molecular profiles, geometry, RDF, 3D-MoRSE, WHIM, and GETAWAY descriptors [11].

Recently, the utility of these molecular descriptors in predicting the inhibitory activity of curcumin analogues as anti-proliferative agents of human prostate cancer cell Line (PC-3) [14], dihydroquinazoline derivatives of Retro-2cycl against Shiga toxin [15], and dihydrothiophenones against dihydroorotate dehydrogenase of malaria parasite [16] has been demonstrated. Moreover, by performing cluster analysis, key molecular descriptors that allow segregation of anti-inflammatory drugs into COX-2 selective and nonselective inhibitors have been identified [17].

The identification of key molecular properties that confer druglikeness has been a long-standing goal in drug discovery. The activity of many drugs, for instance, has been strongly correlated with properties that promote oral bioavailability such as those featured in the Lipinski Rule of Five (Ro5). However, meeting the requirements of RO5 does not guarantee druglikeness. The limited scope of RO5 and its variants has prompted the search for key molecular properties that could discriminate known drugs from nondrugs. Specifically, this work examined two sets of compounds - approved drugs from DrugBank and nondrugs that were randomly selected from the Enamine database of synthetic compounds. The 3D structure of each compound was optimized, and the molecular descriptors for each compound were calculated using Dragon 6 software. Subsequently, discriminant analysis was performed on these compounds in order to identify the crucial molecular properties that discriminate between drugs from nondrugs.

Methodology

All computational works were performed in a personal computer running on Microsoft Windows 7 Professional 64-bit Operating System using a 3.50-GHz Intel Core i7-4770K processor with 8.00-GB random access memory. The structures of approved drugs were retrieved from the DrugBank database (http://drugbank.ca) while the set of nondrugs was obtained through random selection of entries in the Enamine HTS Collection database (https://enamine.net). The structures were drawn using MarvinSketch (https://chemaxon.com) and saved in .mol format. The generation of the 3D structures of the compounds was performed using the CHARMm force field in BIOVIA Discovery Studio (DS) (http://3dsbiovia.com). Each structure was saved as standard database format (.sdf) file.

The molecular descriptors were calculated using the Dragon 6 software (https://chm.kode-solutions.net), which calculates over 4000 descriptors per molecule. The data set was cleaned by deleting rows (compounds - drugs or nondrugs) and columns (variables or descriptors) of the data file that are duplicate compounds or with at least one NAN (i.e. Not A Number) entries, descriptors with invariant values or with mostly zero values. Thus, the original number of approved drugs of 1887 was reduced to 1792 and correspondingly the 2000 nondrugs reduced to 1792, by random selection, making a total of 3584 compounds; and from 4888, the number of variables was also reduced to 245.Linear Discriminant Analysis using SPSS (https://www.ibm.com) was performed on a data set consisting of 3584 rows (compounds) and 245 columns (descriptors) *vide infra*.

Results and Discussion

The development of bioactive molecules into therapeutic agents is an important step in drug discovery. It is in the lead drug development stage where most bioactive molecules get eliminated due to problems primarily associated with bioavailability. Unfortunately, this phase in drug discovery is typically reached only when a substantial amount of efforts and resources has already been expended in discovering new leads and their variants. In order to prevent the high attrition rate at the later stages of drug discovery, few screens that can discriminate potentially nondruglike compounds in the clinical phase have been introduced even at the hit discovery stage so that only those bioactive hits with no serious ADMET (absorption, distribution, metabolism, excretion, toxicity) issues are forwarded for further development.

Linear Discriminant Analysis (LDA) was performed in order to classify compounds as drug or nondrug and identify the essential properties of compounds that are associated with druglikeness. “Group” referring to drug group, as dependent variable, was assigned a value of “1” for drugs and “0” for nondrugs. Initially, LDA was carried out using

backward stepping method involving all the 245 descriptors. This generated a discriminant function (DF) consisting of 42 descriptors that correctly classifies a compound as drug or nondrug with 92% overall accuracy. LDA was done 42 times more with these descriptors involving 1 descriptor at a time and the 5 descriptors with the least prediction accuracy were identified and subsequently deleted. The prediction accuracy of the resulting 37-predictor DF remained at 92%. With the 37 remaining descriptors, another round of 37 DA runs was done with 36 descriptors (*i.e.* 37 – 1) removing one descriptor every run. This round identified 8 descriptors with the least contribution in the classification of drugs. Removing these from the DF reduced the prediction accuracy by only 0.2%. With the remaining 29 descriptors succeeding several rounds were done which allowed the removal of 12 more descriptors with only roughly 2% reduction in accuracy, now with only 17 descriptors the accuracy of prediction was still 90.3%. More rounds were done and more descriptors with the least contributions in prediction were subsequently removed from the DF. With only six descriptors left in the DF namely, nHet, NNRS, ON0, ON0V, MDDD, and HVcpx, the overall accuracy of prediction was still high at 85% (Table 1), correctly classifying 93% of nondrugs and 77% of drugs (Figure 1).

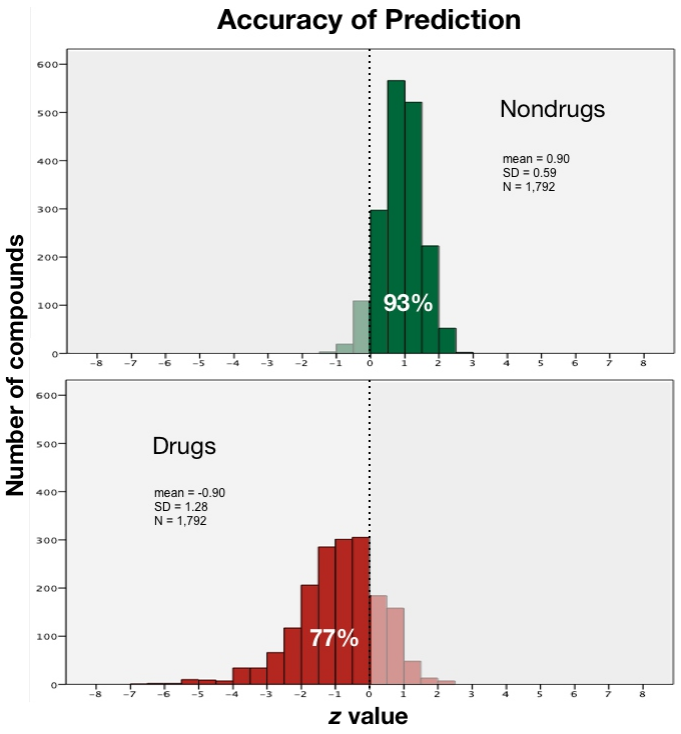

Figure 1. *Prediction Accuracy of Equation 1 for Nondrugs (above) and Drugs (below)*

Table 1. *Results of the Classification of Compounds by Discriminant Analysis of a Six-Predictor Discriminant Function*

Classification Results[a,c]					
			Predicted Group Membership		Total
		Group	0	1	
Original	Count	0	1661	131	1792
		1	410	1382	1792
	%	0	92.7	7.3	100.0
		1	22.9	77.1	100.0
Cross-validated[b]	Count	0	1661	131	1792
		1	414	1378	1792
	%	0	92.7	7.3	100.0
		1	23.1	76.9	100.0

a. 84.9% of original grouped cases correctly classified.
b. Cross-validation is done only for those cases in the analysis. In cross-validation, each case is classified by the functions derived from all cases other than that case.
c. 84.8% of cross-validated grouped cases correctly classified.

The discriminant function, z, involving 6 predictors, is given in Equation 1.

$$z = 0.70\,nHet + 1.52\,NNRS - 0.72\,ON0 + 0.70\,ON0V - 0.08\,MDDD + 3.18\,HVcpx - 9.09$$

(Equation 1)

where: *nHet* – number of heteroatoms
NNRS – normalized number of ring systems
ON0 – overall modified Zagreb index of order 0
ON0V – overall modified Zagreb index of order 0 by valence vertex degrees
MDDD – mean distance degree deviation
HVcpx – graph vertex complexity index

In order to predict whether a compound will likely be a drug or not, one only needs to determine the value of these predictors for a compound and use the above DF to calculate the z value. If z is negative, the compound is classified as a drug, albeit this prediction is only true for roughly 8 out of 10 compounds. The DF shows that small values of *nHet*, *NNRS*, *ON0V*, and *HVcpx*, and large values of *ON0*, and *MDDD* will make *z* value small that will result in classifying a compound as a drug.

The *nHet* descriptor is a constitutional index that describes the number of heteroatoms (*i.e.* not C or H) in a molecule [18]. For example, in the drug paracetamol or acetaminophen with a chemical formula of $C_8H_9NO_2$, the *nHet* value is 3 due to one N and two O atoms present in the molecule. The heteroatoms, which often define the functional groups in the molecule, are usually the hot spots of chemical reactivity. *NNRS* is a ring descriptor that stands for normalized number of ring systems [18]. The ring portions in the molecule provide structural rigidity, which can be useful for anchorage or precise disposition of functional groups during interaction with a biological target.

On0 and *ON0V* are Zagreb-originated topological indices based on a graph generated from the structure of the molecule by replacing atoms with vertices and bonds with edges [19]. The original Zagreb indices M_1 and M_2 are simply the summation of the square of vertex degrees and the summation of the product of vertex degrees connected by an edge, respectively [20]. The drawback with the Zagreb indices is that they place more weight on the inner vertices and edges of a graph, contrary to chemical intuition which puts more importance on outer vertices and edges because they are associated with a larger part of the molecular surface. This defines the molecular size, volume, and shape, properties that are expected to make greater contribution to the physical, chemical, and biological properties of the molecule [20]. The amended Zagreb indices, called modified Zagreb indices ${}^mM_1$ and ${}^mM_2$, use the inverse values of the vertex degrees [20]. ^{m}M2 is identical to the first-order overall index ${}^1ON$ descriptor introduced by Bonchev [21]. The *ON0* descriptor is the overall modified Zagreb index of order 0, whereas *ON0V* is the overall modified Zagreb index of order 0 by valence vertex degrees, a Zagreb index calculated using valence vertex degree δ^{ω}, in place of simple vertex degree. Zagreb indices have been utilized in modeling the boiling point of alkanes [20]. Since the boiling point of a compound depends primarily on the strength of the intermolecular interaction, the Zagreb indices and their successors can be related with relevant properties such as polarizability, solubility, octanol-water partition coefficient (logP), etc.,

The mean distance degree deviation (*MDDD* or Δσ) is defined as:

$$\Delta\sigma = \frac{1}{A}\sum_{i=1}^{A}|\sigma_i - \bar{\sigma}|$$

(Equation 2)

where *A* is the number of atoms, σ_i is the vertex distance degree (*i.e.* the sum of topological distances d_{ij}, the sum of the distances of all the other atoms *j* to atom *i*), and σ is the mean vertex distance degree [19]. Terminal vertices are observed to have high values of vertex distance degree compared to central vertices. In addition, the vertex distance degree is small if the vertex is near a branching site compared to a terminal vertex that is far away from it [19]. Lastly, HVcpx is an information index called graph vertex complexity index. It is defined by the equation [22,23]:

$$HVcpx = \frac{1}{A}\sum_{i=1}^{A} v_i^c$$

(Equation 3)

where v_i^c is the vertex complexity for vertex *i* given by:

$$v_i^c = -\sum_{j=0}^{\sigma(v_i)} \frac{k_j^i}{A} log \frac{k_j^i}{A}$$

(Equation 4)

In Equation 4, $\sigma(v_i)$ is the eccentricity of vertex *i*, that is, the maximum graph distance between vertex *i* and any other vertex in the graph. *A* is the number of atoms or vertices, and each vertex *i* has *j k*-neighbors that are k steps away. *HVcpx* apparently encodes the steric aspect of the molecule.

In the stepwise removal of descriptors from the DF, the reductions in prediction accuracy were noted. With only 3% reduction in accuracy, *NNRS* turned out to have the

smallest contribution to prediction, followed by *nHet*, *ON0V*, *ON0*, *MDDD*, and *HVcpx*, in order of increasing contribution. The removal of either *MDDD* or *HVcpx* drastically reduced the prediction accuracy by around 10%. Then an LDA with only these two predictors was done and it showed a prediction accuracy of 81% (Table 2).

The DF with these two predictors, *HVcpx* and *MDDD*, is given in Equation 5.

$$z = 4.58\,\mathbf{HVcpx} - 0.12\,\mathbf{MDDD} - 12.09 \quad \text{(Equation 5)}$$

This correctly classifies 91% of nondrugs and 71% of drugs. In particular, Equation 5 correctly classified 1273 of the 1792 drugs and 1635 of the 1792 nondrugs (Figure 2) in the data set. The prediction accuracy of the DF is 66% with *HVcpx* alone as predictor, and 48% with *MDDD* alone. Interestingly, *MDDD* is one of the topological indices, which were proven effective in discriminating between drugs and nondrugs in a study employing artificial neural network [24].

The *HVcpx* values exhibit normal distributions with skewness values of -0.776 and -0.111 for the nondrugs and drugs, respectively; and 96% of the values from each group fall within two standard deviations from their corresponding means, 3.328 ± 2(0.214) and 3.101 ± 2(0.476) (Figure 3 and Figure 4). Except for only 4%, the *HVcpx* values of nondrugs

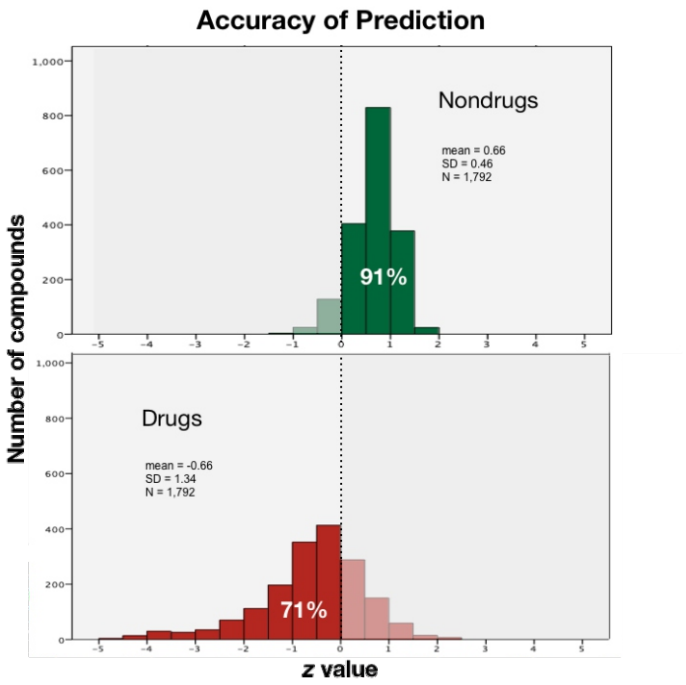

Figure 2. *Prediction Accuracy of Equation 2 for Nondrugs (above) and Drugs (below)*

are at least 2.90, and most values are greater than 3.25. Thus, it is quite safe to classify a compound with an *HVcpx* value of less than 3.00 as a drug. In the data set, only 147 out of 1792 (or 8.2%) nondrugs have *HVcpx* values below 3.00

Table 2. *Results of Classification of Compounds by Discriminant Analysis of a Two-Predictor Discriminant Function*

Classification Results[a,c]					
			Predicted Group Membership		Total
		Group	0	1	
Original	Count	0	1635	157	1792
		1	519	1273	1792
	%	0	91.2	8.8	100.0
		1	29.0	71.0	100.0
Cross-validated[b]	Count	0	1635	157	1792
		1	519	1273	1792
	%	0	91.2	8.8	100.0
		1	29.0	71.0	100.0

a. 81.1% of original grouped cases correctly classified.
b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
c. 81.1% of cross-validated grouped cases correctly classified.

and misclassified as drugs. For the drugs group, 757 or 42.2% have *HVcpx* values below 3.00. Thus, using *HVcpx* ≤ 3 as a criterion for druglikeness, two-thirds (2402 out of 3584) of the compounds are correctly classified. However, an *HVcpx* value above 3.00 may not necessarily mean that a compound is a nondrug as more than half of the drugs have *HVcpx* values above 3.00.

The *MDDD* values are also normally distributed with skewness of 0.06 and 2.11 for nondrugs and drugs, respectively; and around 96% of the values from each group fall within two standard deviations from their corresponding means, 20.35 ± 2(5.64) and 22.22 ± 2(17.62), with the drugs

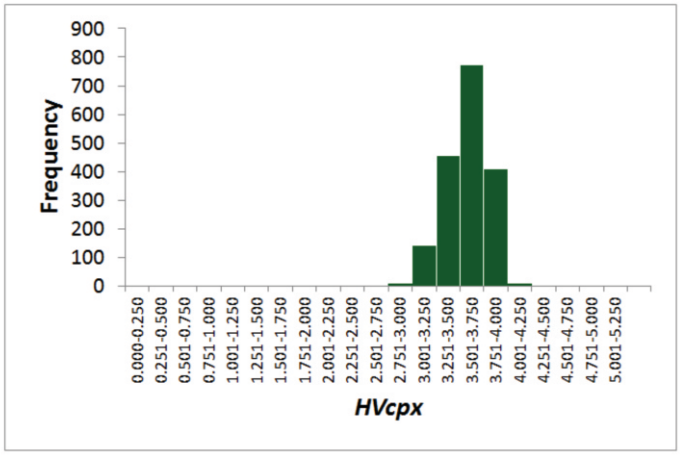

Figure 3. *Frequency Distribution of HVcpx Values of Nondrugs (Mean = 3.328, SD = 0.214)*

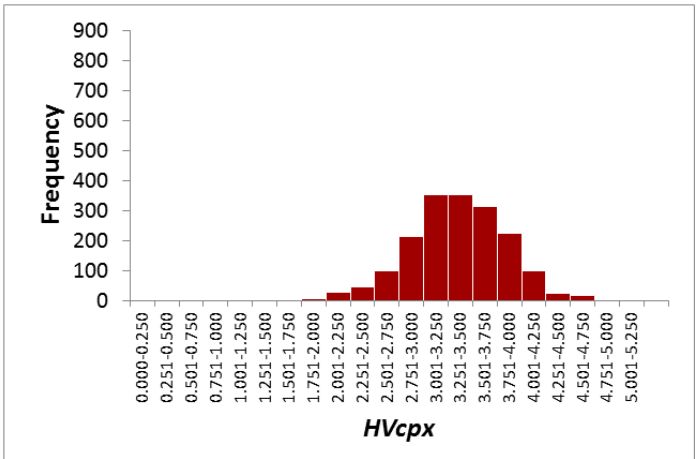

Figure 4. Frequency Distribution of HVcpx Values of Drugs (Mean = 3.101, SD = 0.476)

group having the higher mean. As a criterion for druglikeness, "*MDDD* ≥ 30" would misclassify only 4% of nondrugs but would correctly classify 23% of drugs. Thus, it can be said that if a compound has an *MDDD* value above 30, it is highly likely to be a drug because only very few nondrugs have those *MDDD* values.

Finally, these results show that the known drugs, when compared to representative synthetic organic compounds from Enamine HTS Collection, somewhat obey a "Rule of Three": *HVcpx* ≤ 3 and *MDDD* ≥ 30.

Conclusion

Using Discriminant Analysis, this study revealed six properties (*i.e. nHet*, *NNRS*, *ON0*, *ON0V*, *MDDD*, and *HVcpx*) that can correctly classify 85% of compounds as drug or nondrug. Together, *HVcpx* and *MDDD* can discriminate 81% of the compounds; the predictor *HVcpx* alone correctly classifies two-thirds of the compounds. The juxtaposition in this study of approved drugs in DrugBank and synthetic compounds in Enamine HTS Collection squeezed out a kind of "Rule of Three" (*HVcpx* ≤ 3, *MDDD* ≥ 30) criteria for druglikeness. These may be used as filters in the hit discovery phase to help minimize the costly high attrition rate of candidates at the later stage of the drug discovery process.

Acknowledgment

This work was fully supported by the Office of the Vice President of Academic Affairs, University of the Philippines System through the Enhanced Creative Work and Research Grant (ECWRG 2016-1-006).

References

1. Hann M, *et al.* (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. Journal of Chemical Information and Computer Scientists, 41: 856-864.
2. Oprea TI. (2001) Is there a difference between leads and drugs? A historical perspective. Journal of Chemical Information and Computer Scientists, 41: 1308-1315.
3. Lipinski CA, *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews, 46: 3-26.
4. Owens J. (2003) Chris Lipinski discusses life and chemistry after the Rule of Five. Drug Discovery Today, 8: 12-16.
5. Veber DF, *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. Journal of Medicinal Chemistry, 45: 2615-2623.
6. Clark DE, Picket SD. (2000) Computational methods for the prediction of drug-likeness. Drug Discovery Today, 5: 49-58.

7. Carr R, Jhoti H. (2002) Structure-based screening of low-affinity compounds. Drug Discovery Today, 7: 522–527.
8. Erlanson DA, *et al.* (2000) Site-directed ligand discovery. Proceedings of the National Academy of Sciences of the United States of America, 97: 9367 – 9372.
9. Vetter D. (2002) Chemical microarrays, fragment diversity, label-free imaging by plasmon resonance - a chemical genomics approach. Journal of Cellular Biochemistry, 39: 79 – 84.
10. Congreve M, Carr R, Murray C, Jhoti H. (2003) A 'rule of three' for fragment-based lead discovery?. Drug Discovery Today, 8(19): 876 – 877.
11. Todeschini R, Consonni V. (2009) In: Molecular Descriptors for Chemoinformatics, WILEY-VCH, Weinheim (Germany).
12. Tetko IV, *et al.* (2005) Virtual computational chemistry laboratory - design and description, Journal of Computer-Aided Molecular Design, 19: 453-463.
13. Tetko IV. (2005) Computing chemistry on the web, Drug Discovery Today, 10: 1497-500.
14. Reyes AMM, Billones JB. (2013) Quantitative structure-activity relationship study of curcumin analogues as anti-proliferative agents of human prostate cancer cell line (PC-3), Kimika, 24(1):8– 17.
15. Billones LT, Billones JB. (2013) Multiple Linear Regression Model of Shiga Toxin Inhibitory Activity of Dihydroquinazoline Derivatives of Retro-2cycl". Philippine Science Letters, 6(2): 231 – 240.
16. Billones LT, Billones JB. (2014) A Univariate analysis of molecular properties and inhibitory activity of dihydrothiophenones against dihydroorotate dehydrogenase of malaria parasite, Journal of Chemical and Pharmaceutical Research, 6(8):209-217.
17. Vios VSL, Billones JB. (2015) Cluster and multi-linear regression analyses guided identification of molecular descriptors that account for cyclooxygenase activities, Journal of Chemical and Pharmaceutical Research, 2015, 7(8): 735-742.
18. Talete. List of Molecular Descriptors Calculated by Dragon, n.d.
19. Todeschini R, Consonni V. (2000) Handbook of Molecular Descriptors, Wiley-VCH, Weinheim.
20. Nikolic S, Kovacevic G, Milicevic A, Trinajstic N. (2003) The Zagreb Indices 30 Years After, Croatia Chemica Acta, 76(2): 113-124.
21. Bonchev D. (2001) Journal of Molecular Graphics and Modelling, 2001, 20: 65-75.
22. Shojaie A, Sedaghat N. (2017) How Different Are Estimated Genetic Networks of Cancer Subtypes? In: Big and Complex Data Analysis, Ahmed SE (Ed.), Springer International Publishing AG, Switzerland.
23. Mekenyan O, Basak S. (1994) Topologicak Indices and Chemical Reactivity. In: Graph Theoretical Approaches to Chemical Reactivity, Bonchev D, Mekenyan O (Eds.), Springer Science+Business Media, Dordrecht.
24. Murcia-Soler M, *et al.* (2003)Drugs and Nondrugs: An Effective Discrimination with Topological Methods and Artificial Neural Networks, Journal of Chemical Information and Computer Scientists, 43:1688-1702.