MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **78** (2017) 5-16

ISSN 0340 - 6253

The Past within the Future. Graph Eigenvalues as Powerful Discriminant Variables in Drug Design

Jorge Galvez[1]*, Riccardo Zanni[1,2], Maria Galvez-Llompart[1,2], Ramón García-Domenech[1]

[1] *Molecular Topology and Drug Design Unit, Department of Physical Chemistry, University of Valencia, Spain*
jorge.galvez@uv.es , ramon.garcia@uv.es

[2] *Department of Microbiology, University of Malaga, Spain*
riccardo.zanni@uv.es , maria.galvez@uv.es

(Received November 29, 2016)

Abstract

The past within the future is the emblematic title of the present paper to describe the actual situation of molecular topology (MT) and graph theory in drug design. Graph theory can be considered as a classical and well-known discipline, which however cannot be considered out of date because its role in chemistry has become more and more important during the last years. Graph theory is the theoretical support of molecular topology, a paradigm capable to depict molecules from a topological viewpoint. Thanks to a pure mathematical representation of chemical structures, molecular topology (MT) uses topological descriptors to operate biological, pharmacological, physical and chemical properties. The present paper illustrates another interesting application of MT, providing discrimination between drug and non-drug molecules based on simple graph differences. This original formalism may help in the difficult task of finding out new *hit* and *lead* compounds.

1 Introduction

A little bit of history: The Riemann zeta function

The Riemann zeta function, ζ(s), named both by and for Bernhard Riemann [1], was first studied in the 18th century by Leonhard Euler, who defined it as a sum over all the natural numbers as:

$$\zeta(s) = \sum_n \frac{1}{n^s}$$

where the power *s* is variable.

Euler demonstrated that the sum is finite for any value of *s* above 1. For example, for S=2, ζ(s) is equal to $\pi^2/6$:

$$\zeta(2) = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \ldots = \frac{\pi^2}{6}$$

Euler also proved that the summatory is equal to a product formula including one term for each prime number, what suggests some connection between the zeta function and the distribution of primes among the integers.

Later on, in 1859, Riemann showed that the zeta function applies not only if *s* is a real number greater than 1 but for any number either real or complex (excepting the numbers whose real part is equal to 1) [2].

The crossing points, where ζ(*s*)=0, are called zeros of the zeta function. Riemann studied the distribution of the zeros and thereby established a hypothesis, which represents a milestone in modern mathematics.

Many years later, Montgomery [3] realized that pair-correlation function of the zeta zeros was essential to understand the statistics of the fluctuations, so that the distribution of the zeros matches that of the eigenvalues of a random Hermitian matrix, which is a large Hermitian matrix whose entries are random variables. It also fits the energy levels in heavy nuclei (for example U-238) and other patterns such as galaxies layout across the universe, genes distribution along a strand of chromatin or prime numbers location among the integers. Figure 1 exhibits the distribution patterns for different systems. It can be seen that although the distances between the levels vary significantly from one case to another, on average their values are identical.

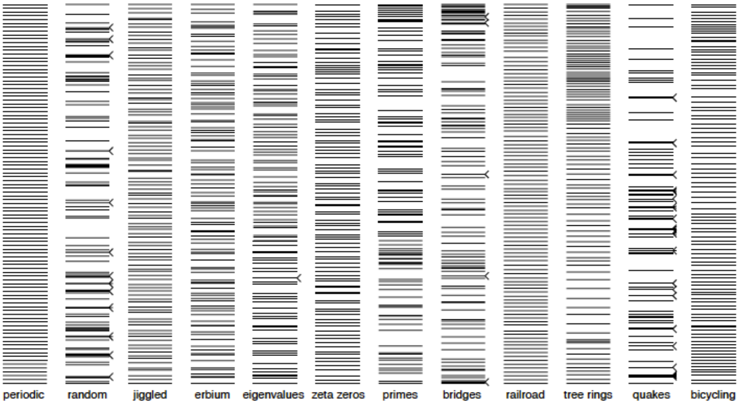
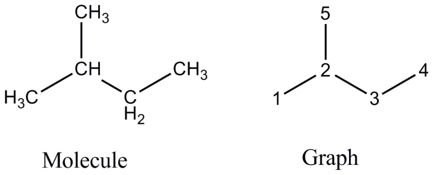


Figure 1. Some interesting common patterns for a random Hermitian matrix [1].

Molecular topology and the Riemann zeta function

Given the outlined above, we may wonder if it possible to establish some parallelism between Figure 1's patterns and molecular topology.

Molecular topology (MT) is a discipline based on graph theory that allows representing every molecular structure as with a particular topology [4,5]. One of the key advantages of this formalism is that the graph-molecule can be transformed into a matrix, called topological or adjacency matrix. Starting by the matrix, different sets of descriptors called topological indices (TIs) can be calculated [5]. TIs are an algebraic representation of the graph and they can be related to several properties of molecules [6]. Figure 2 shows the representation of isopentane as a graph, along with its topological matrix.



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{array}{c} \delta \\ \begin{bmatrix} 1 \\ 3 \\ 2 \\ 1 \\ 1 \end{bmatrix} \end{array}$$

Adjacency matrix

Figure 2. The chemical graph and adjacency matrix of isopentane.

The atoms are represented by vertices and the bonds by edges. To build up the topological matrix each one of the vertices (atoms) in the graph is randomly numbered. This way it is possible to assign any ordinal integer number to any vertex in the graph [6]. The mathematicians call TIs also as *graph invariants*, because they remain unaltered under isomorphism and their value must not depend on the order of numbering the vertices. As can be seen in Figure 2, the adjacency matrix is arranged so that its a*ij* entries take value 1 if vertex *i* is linked to vertex *j* and 0 otherwise.

Once calculated, the TIs can be correlated with many physical, chemical or biological properties of the graph associated chemical compounds [6].

Now, considering that **every adjacency matrix *A* is also a Hermitian one** (every symmetric matrix is a particular case of Hermitian) we analyse here the possibility to discover any type of pattern in the distribution of matrix *A*'s eigenvalues. Considering that *A* matrix cannot be a random one, if such pattern exists it could be disclosed averaging the eigenvalues of a large number of compounds.

This way, the randomness related to the big range of the matrix is supplied by a large number of smaller matrices what makes the outcome statistically significant. Following this method and after confirming the existence of a particular pattern, it is applied to the distinction between two categories of molecules: Drugs and Non-drugs. Concretely, the first 15 eigenvalues from the edge adjacency matrix weighted by edge degrees of different molecular simple graphs were calculated. Altogether 4600 molecules (2300 drugs and 2300 non-drugs) were included in the study.

2 Material and methods

The mathematical and statistical methods used to achieve the objective are described.

Dataset

One half of the 4600 molecules dataset were drugs selected from the Comprehensive Medicinal Chemistry (CMC) database [7] while the other half, compounds not showing known pharmacological activity, were selected from the Sigma-Aldrich database [8]. The dataset for the different pharmacological classes analyzed independently was collected from the Prous databank [9], while the inactive compounds were selected from literature.

Calculation of the topological descriptors. The Eigenvalues (EEig)

Each compound was characterized by a set of the 15 first eigenvalues of the edge adjacency matrix weighted by edge degrees. Eigenvalues are a special set of scalars associated with a linear system of equations (i.e., a matrix equation) that are sometimes known as characteristic roots, characteristic values [10], proper values, or latent roots [11]. All indices were calculated with Dragon software version 5.4 [12].

Modeling techniques

Linear discriminant analysis (LDA) is used to distinguish between drugs and non- drugs from the eigenvalues distribution. It is a statistical method to find the best linear combination of variables (eigenvalues in our case) that better distinguish between two or more categories or objects (in our case drug-molecules and nondrug-molecules). The software used for the statistical study was BMDP [13].

3 Results and discussion

Distribution of the eigenvalues

Figure 3 shows the plot of the differences between each eigenvalue and its nearest neighbor, versus the mean value of the two neighboring eigenvalues for the first 15 eigenvalues of the edge adjacency matrix weighted by edge degrees (EEigx). A mixed set of 4600 molecules, half of them drugs and the other half non-drugs, was included in the analysis.

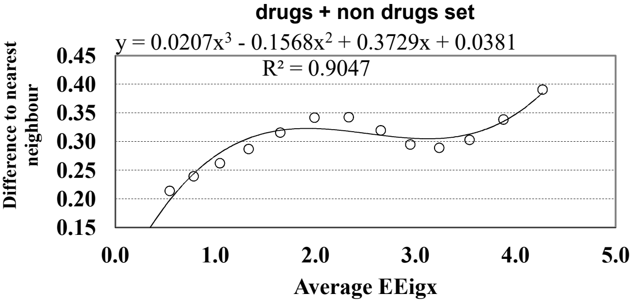


Figure 3. Distribution of the average values of EEigx for the whole dataset.

A simple observation of the determination coefficient (R^2= 0.905) highlights a pretty good fitting to a three-degree polynomial, although for the first six eigenvalues, the dependence is clearly linear. The following step consisted in representing the distribution of the drug group (about 2300 compounds).

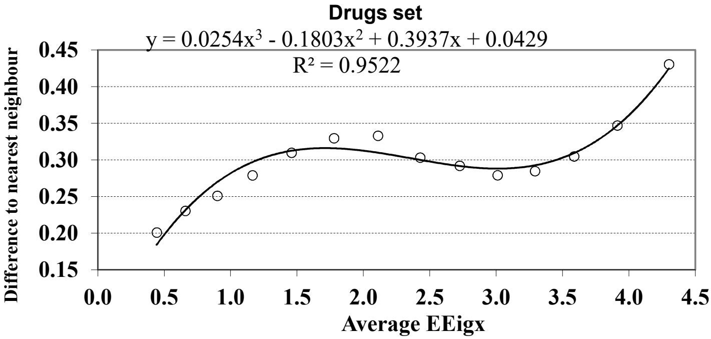


Figure 4. Distribution of the average values of EEigx for the “drug” dataset.

As it can be seen in Figure 4, the fitting improves significantly with respect to the mixed set. In fact, R^2 rises up to 0.952, what means that the drugs fulfill the pattern much better than the non-drugs. Finally, the same graphic representation was done for the set of non-drugs (Fig. 5). It is noteworthy that the regression is worse (R^2=0.8604) and that the shape of the curve is broken down, particularly the first straight frame corresponding to the first six eigenvalues.

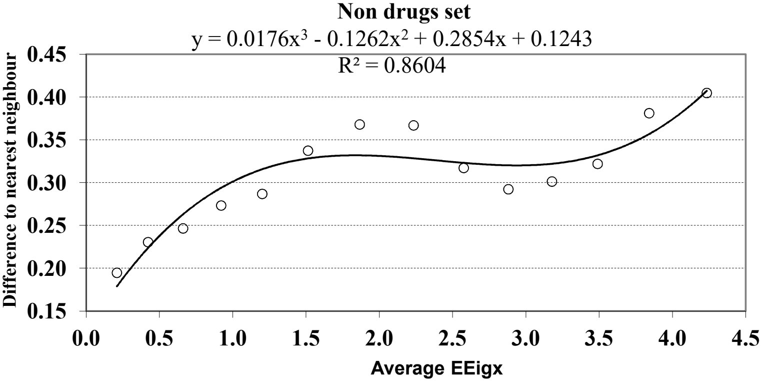


Figure 5. Distribution of the average values of EEigx for the "non drugs" dataset.

These results clearly demonstrate that there are different patterns of distribution for drugs and non-drugs, what can be used to discriminate between them.

Finally, the difference of ordinates between drugs and non-drugs was plotted against the ordinal of the corresponding eigenvalue. As can be seen in Fig. 6, EEig09x is the eigenvalue showing the largest difference.

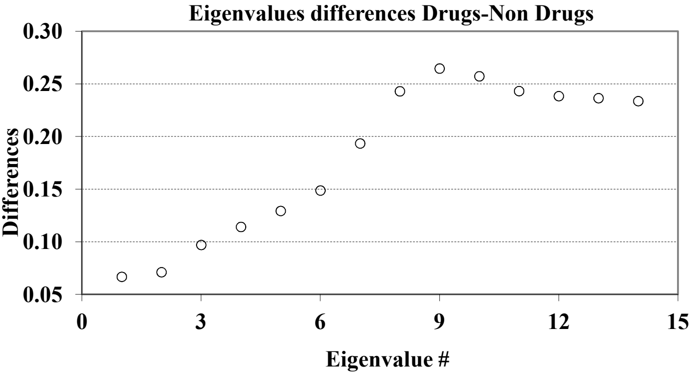


Figure 6.- Eigenvalues differences drugs – non drugs versus eigenvalue #.

Therefore, it is to be expected that this eigenvalue, i.e. EEig09x, is to be the best discriminant eigenvalue to distinguish drugs from non-drugs. This was confirmed using LDA, as can be appreciated in Table 1. In the Table are also included the results for the discrimination of particular types of drugs (anti-Alzheimer, analgesics, antibacterials and antineoplastics).

Table 1. Classification matrix obtained through the selected $DF_{1\text{-}6}$ for the *training set*.

Set	**A**	**I**	**TP**	**FP**	**TN**	**FN**	**Sensitivity (%)**	**Specificity (%)**
	Drugs + non drugs set			**DF_1 = 1.57EEig09x -3.19**				
Training	1855	2287	1423	807	1480	432	76.7	64.7
	Drugs + nondrugs set			**DF_2 = 1.29EEig11x -1.83**				
Training	1855	2330	1246	684	1646	609	67.2	70.6
	Anti-Alzheimer			**DF_3 = 0.501EEig11x -0.866**				
Training	520	523	351	228	295	169	67.5	56.4
	Anti-bacterial			**DF_4 = 1.52EEig09x -3.11**				
Training	350	375	268	135	240	82	76.5	64.0
	Analgesic			**DF_5 = -0.44EEig11x +0.70**				
Training	238	436	132	149	287	106	55.5	65.8
	Anti-neoplastic			**DF_6= -0.37EEig07x +1.01**				
Training	464	629	195	173	456	269	42.0	72.5

A: number of active compound; I: number of inactive compounds; TP: true positive (active); FP: false positive (active); TN: true negative (inactive); FN: false negative (inactive).
Sensitivity (%) =100xTP/(TP+FN), Specificity (%) =100xTN/(TN+FP)

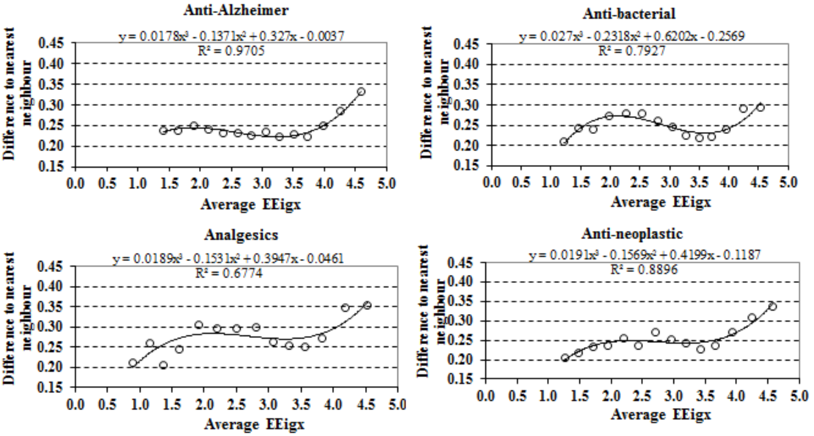


Figure 7. Distribution of the average values of EEigx for four pharmacological groups of molecules. All the data was collected from the CMC and Prous databases.

Taking into account the results depicted in Table 1 and Figure 7, it is noteworthy that the eigenvalues of the edge adjacency matrix perform very well not only to distinguish between drugs and non-drugs but also within the four classes of drugs analyzed here. Results not disclosed here point towards the following of the same pattern for other classes of drugs. The results were noteworthy, because the same pattern observed for the entire drugs group is reproduced in the particular classes of drugs analyzed. The only exception was the group of analgesics, which gave a worse correlation (R^2= 0.67), although still statistically significant.

Moreover, DF_1 allows an overall accuracy of 70% (76.7% sensitivity and 64.7% specificity) as illustrated in Table 1, whereas DF_2, which depends on EEig11x, reached the maximum specificity.

Hence, both DFs can be used jointly or alternatively depending on our needs regarding sensitivity and specificity.

Anyway, the stepwise application of both discriminant equations should yield a good record in selecting novel scaffolds with drug-like characteristics, because we have a good sensitivity plus a good specificity.

Regarding specific classes of drugs seems like analgesics were the less accurate, while anti-Alzheimer were the best. Antineoplastics differ from the others because the descriptor which most contributed was EEig07x (see DF_6 equation in Table 1). The most interesting aspect of this equation lies in its ability to recognize the inactive compounds. With a 72.5% of correct classification, the specificity of the model is very high, as compared to the others.

Altogether it is outstanding how structural descriptors based on simple graphs, i.e. no chemical or biological information was provided, were capable to correctly classify above

70% of the molecules either as drugs or as non-drugs. Moreover, the actual specificity is probably greater than 70% because not all the compounds in Aldrich database are non-drugs and it is to expect that a slight percentage of them are actually drugs.

Application to frameworks

Based on these interesting results, discriminant function was used to distinguish between frameworks showing drug and non-drug profiles. Frameworks are defined as a set of two or more un-substituted (clean) cycles linked by bonds [14]. In other words: a framework is the underlying rings and connectors [14].

The analysis of frameworks has proved to be a good strategy for the structure-activity study of drugs [15,16], among other reasons because the same framework can be common to many different molecules. Hence, it is possible to study different levels (graph / graph-node / graph-node-bond), using identifiers able to differ one level from another [17]. An example of these three levels is shown in figure 8.

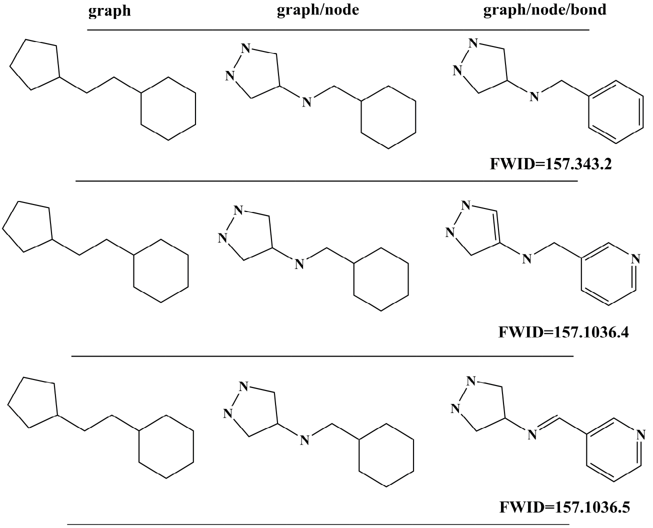
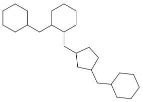
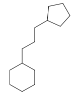
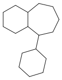
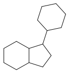
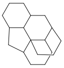
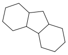
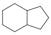
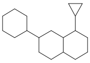
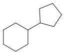
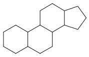
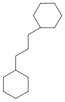
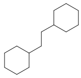
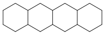


Figure 8.- Framework identifier (FWID) for three frameworks (157 shows the graph level, 343 or 1036 the graph/node level and 2, 4 or 5 the graph/node/bond level).

Since our objective is to demonstrate the efficacy of molecular depiction using simple graphs, the equation DF_1 was employed to select two sets of frameworks: one for drugs and another one for non-drugs. Table 2 shows the results for a small representative group of drug and non-drug frameworks.

Table 2. Frameworks corresponding to non-drugs and drugs. Most of drugs exhibit positive values of EEig09x usually above 0.5.

Framework			Framework	
Drugs	Classification	EEig09x	Non drugs	EEig09x
	Aminoglycosides	2.296		0.226
	Benzodiazepines	1.069		0.407
	Morphine	1.379		0.295
	Penicillins	0.0		-0.879
	Quinolones	2.0		-0.414
	Steroids	1.481		0.382
	Sulfamides	0.147		
	Tetracyclines	1.347		

As expected, there is a notable difference between the two groups. Thus, the majority of drugs show values of the discriminant index EEig09x above 0.5, while most of non-drugs show values below it. So, 0.5 is the threshold to distinguish both classes. As can be seen in Table 2, the frameworks of important pharmacological groups, such as benzodiazepines, quinolones, steroids, etc. show high EEig09x values, ranging from 1 to 2. A significant exception is that of penicillins (Eeig09x=0). We can conclude that EEig09x works quite well as discriminant between drugs and non-drugs.

Briefly, in summary, the present results highlight how simple graphs associated with pharmacological active molecules are structurally different from those corresponding to molecules without pharmacological activity. However, it remains unclear if this difference is intrinsic to the pharmacological activity or to the fact that drugs exhibit particular structures, which are more likely to be found than those of non-drugs. This might be because of the bias of chemists in pharmaceutical industry to synthesize molecules that have already shown pharmacological activity. In other words, consolidated drugs tend to be more synthesized in search of derivatives, establishing a process of *rich gets richer* [17] and a structural difference between drugs and non-drugs. We believe that although this factor may play an important role, it is not the reason for this difference, because many authors (including ourselves) have demonstrated that the molecular structures of drugs differ intrinsically from those of non-drugs [18]. Under this view, it is noteworthy that this intrinsic structural difference can be disclosed in a straightforward manner by using simple graphs.

4 Summary and conclusions

The molecular simple graphs of organic compounds show a well-defined pattern of distribution of the eigenvalues from the edge adjacency matrix weighted by edge degrees. Such a pattern fits to a third degree polynomial, with a better fitting for drugs than for non-drugs. The most discriminant eigenvalue (EEig09x), enables a discrimination of above 70% between a large set of 2300 compounds from CMC database (all drugs) and another set of 2300 compounds from Sigma-Aldrich database (the very most of them non-drugs). This discriminating capability may also be extended to molecular frameworks so broaden the usefulness of the formalism for the design and discovery of new drugs. The distribution of the eigenvalues of the edge adjacency matrix weighted by edge degrees for molecular simple graphs is parallel to what observed by Montgomery in 1972 for the distribution of the eigenvalues of random hermitic matrices as far as the edge adjacency are particular case of Hermitian matrices. Moreover, the polynomial fiiting is better for drugs than for non-drugs, which makes the eigenvalues themselves the best discriminant variables between both types of molecules. Eigenvalues such as Eeig09x and Eeig11x, play a key role and are applicable to particular groups of drugs, such as analgesics, antibacterials, anti-Alzheimer, antineoplastics and others.

These results are an important conceptual contribution about the relevance of molecular skeleton described in terms of simple graphs for the detection of drug activity. The role of the eigenvalues as discriminant variables is emphasized. Altogether, the present results imply that we may setup novel mathematical tools for drug design and discovery.

Acknowledgments: Authors acknowledge the MINECO (Spanish Ministery of Economy, Industry and Competitivity) Project: “DESARROLLO DE NUEVAS HERRAMIENTAS PARA EL CONTROL DE OIDIOS” (AGL2016-76216-C2-2-R.).

References

[1] B. Hayes, The spectrum of Riemannium, *Am. Sci.* **91** (2003) 296–300.

[2] A. Laurinčikas, The Riemann zeta-function: Approximation of analytic functions, in: S. V. Rogosin, A. A. Koroleva (Eds.), *Advances in Applied Analysis*, Springer, Basel, 2012, pp. 95–114.

[3] H. L. Montgomery, The pair correlation of the zeta function, *Proc. Sympos. Pure Math.* **24** (1973) 181–193.

[4] L. Pogliani, Phase diagrams and physicochemical graphs. How did it start, *MATCH Commun. Math. Comput. Chem.* **49** (2003) 141–152.

[5] R. Zanni, M. Galvez-Llompart, R. Garcia-Domenech, J. Galvez, Latest advances in molecular topology applications for drug discovery, *Exp. Op. Drug Disc.* **10** (2015) 945–957.

[6] J. Galvez, M. Galvez-Llompart, R. Zanni, R. Garcia-Domenech, Molecular topology – dissimilar similarities, *Drug Discov. Today Tech.* **10** (2013) e475–e481.

[7] Biovia. (2016). Comprehensive Medicinal Chemistry (CMC). Retrieved from http://accelrys.com/products/collaborative-science/databases/bioactivity-databases/comprehensive-medicinal-chemistry.html

[8] Sigma-Aldrich database. (2016). Retrieved from http://www.sigmaaldrich.com/chemistry/chemistry-services/aldrich-market-select.html

[9] CIPSLINE, Prous Science, Barcelona, Spain (2016). Retrieved from www.prous.com

[10] K. H. Hoffman, R. Kunze, *Linear Algebra*, Pearson, London, 1971.

[11] M. Marcus, H. Minc, *Introduction to Linear Algebra*, Macmillan, New York, 1965.

[12] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Dragon for Windows (Software for Molecular Descriptor Calculations), version 5.4. Talete srl: Milan, Italy 2006.

[13] W. J. Dixon, BMDP statistical software manual: to accompany the 1990 software release, Univ. California Pr., San Francisco, 1990.

[14] G. W. Bemis, M. A. Murcko, The properties of known drugs. 1. Molecular frameworks, *J. Med. Chem.* **39** (1996) 2887–2893.

[15] H. Chen, Y. Yang, O. Engkvist, Molecular topology analysis of the differences between drugs, clinical candidate compounds, and bioactive molecules, *J. Chem. Inf. Model.* **50** (2010) 2141–2150.

[16] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework, *Bioinformatics* **26** (2010) i246–i254.

[17] A. H. Lipkus, Q. Yuan, K. A. Lucas, S. A. Funk, W. F. Bartelt, R. J. Schenck, A. J. Trippe, Structural diversity of organic chemistry. A scaffold analysis of the CAS registry, *J. Org. Chem.* **73** (2008) 4443–4451.

[18] J. Galvez, M. Galvez-Llompart, R. Garcia-Domenech, Molecular topology as a novel approach for drug discovery, *Exp. Op. Drug Disc.* **7** (2012) 133–153.