Chemometric Study on Molecules with Anticancer Properties

João Elias Vidueira Ferreira[1], Antonio Florêncio de Figueiredo[2],
Jardel Pinto Barbosa[3] and José Ciríaco Pinheiro[3]
[1]*Universidade do Estado do Pará*
[2]*Instituto Federal de Educação, Ciência e Tecnologia do Pará*
[3]*Laboratório de Química Teórica e Computacional, Universidade Federal do Pará*
Brasil

1. Introduction

Cancer is a class of diseases characterized by uncontrolled growth of abnormal cells of an organism. All over the world millions of people die every year owing to one of the different types of cancer. Unfortunately cancer chemotherapy finds a serious limitation since treatment with drugs is followed by drug resistance in the tumorous cells and side effects (Efferth, 2005). So researches have been directed to make chemotherapy treatment more efficient.

In the late years literature has reported the research on natural products as a good strategy to discover new chemotherapy agents. One of the plants that have shown anticancer properties is *Artemisia annua L.* (*qinghao*). It has the active ingredient artemisinin, which is used as antimalarial. Artemisinin and derivatives have excellent efficacy against multidrug-resistant strains of *P. falciparum* and they are very well tolerated (Price et al., 1998). Recently the sensibility to artemisinin has been evaluated in some tumorous cells. Studies suggest that artemisinin is more toxic to cancerous cells than to normal cells, so giving a new perspective in cancer therapy (Lai et al, 2009).

However ... This book is on chemometrics and what has chemometrics to do with cancer chemotherapy? Well... understanding how these two different areas can be related to one another is the purpose of this chapter. You just must keep on reading this chapter and you will see the many ways chemometrics can be employed to investigate the "behavior" molecules exhibit considering anticancer activity and to make predictions about drugs that were not tested yet. The potential application of chemometrics to analytical data arising from problems in biology and medicine is enormous and, in fact, the applications of chemometrics have diversified substantially over the last few years (Brereton, 2007; 2009). At the end of the chapter you will note that, as in many areas of research, chemometrics plays an important role in medicinal chemistry, fortunately.

Firstly it is necessary to remember that producing a drug is something that takes time and money, so the process must be rationalized! However, in the past, drugs were discovered by synthesizing a lot of molecules, rather without rigorous criteria, and testing experimentally all of them to evaluate their capacity of cure of the disease or at least to control it. But in process of time this methodology became more and more inadequate, for the more new compounds are studied the less a new compound may be discovered to be potent against a disease. It has long been desired to design active structures on the basis of logic and calculations, not relying on chance or trial-and-error (Fujita, 1995).

Nowadays, in science, there is a basic assumpion that molecular properties and structural characteristics are closely connected to biological functions of the compounds. It is often assumed that compounds with similar properties and structures also display similar biological responses. Chemical structure encodes a large amount of information explaining why a certain molecule is active, toxic or insoluble (Rajarshi, 2008). Thus to understand the mechanism of action of a drug it is necessary to interpret the role played by its molecular and structural properties.

In the last decades, much scientific research has focused on how to capture and convert the information encoded in a molecular structure into one or more numbers used to establish quantitative relationships between structures and properties, biological activities or other experimental properties (Puzyn et al., 2010). Quantitative structure-activity relationship (QSAR) studies have been of great value in medicinal chemistry. Statistical tools can be used for the prediction of the biological activities of new compounds based only on the knowledge of their chemical structures, i.e., not depending on experimental data, which are unknown. Such a strategy gives very useful information for the understanding of the mechanisms of the action of drugs and proposals for syntheses, in this way rationalizing drug discovery. QSAR is alive and well (Doweyko, 2008), that is, QSAR has been used with success and so it is still of relevance today.

Moreover advances in computation brought software that made possible to get many different types of information (descriptors) about the molecules. Consequently data gathered through experiments and computers can produce a huge matrix whose elements are information related to molecules. But it seems that analyzing all them will require infinite patience!

What to do?

Chemometics has the solution!

That is true because chemometrics is the art of extracting chemically relevant information from data produced in chemical experiments (Wold, 1995). Most people only think of statistics when faced with a lot of quantitative information to process (Bruns et al., 2006). In this text we show a common and efficient methodology used in medicinal chemistry to rationalize the process of producing a new drug by employing chemometric methods. It is presented a molecular modeling and a chemometric study of 25 artemisinins, which involves artemisinin and derivatives (training set, Fig. 1) with different degrees of cytotoxicities against human hepatocellular carcinoma HepG2 (Liu et al, 2005), since among the malignant tumors in the liver, the hepatocellular carcinoma is very commom. Literature has showed the application of the methodology here described to investigate biological properties (antimalarial and anticancer) of artemisinin and derivatives (Barbosa et al., 2011); (Cardoso et al., 2008); (Pinheiro et al., 2003).

2. Methodology

Any chemometric study requires data. In this study data are obtained from molecular descriptors calculated through computation. The start point is the molecular modeling step, which consists on the construction of the structures and the complete optimization of their geometries through a quantum chemistry approach implemented in computer. This is necessary to represent molecules as real as possible and thus to compute their molecular descriptors. The B3LYP/6-31G** method (Levine, 1991) as implemented in the Gaussian 98 program was employed (Frisch et al., 1998), considering this strategy is suitable for optimizing well all structures since a good description of the geometrical parameters of artemisinin is achieved.

The 25 compounds investigated include artemisinin, amides, esters, alcohols, ketones, and five-membered ring derivatives. All compounds have been associated to their in vitro bioactivity against a human hepatocellular carcinoma cell line, HepG2, and were labeled previously into two classes according with their activities: (-) less active (those with $IC_{50} \geqslant 97$ μM) and (+) more active (those with $IC_{50} < 97$ μM) derivatives. The criteria for choosing this value of IC_{50} are rather subjective. Nevertheless it is convenient to say that 97 μM is the IC_{50} for artemisinin and the higher IC_{50} the less active is the compound.

After molecular modeling, 1700 descriptors (independent variables) were computed for each molecule in the training set. They represent different source of chemical information (features) regarding the molecules and include geometric, electronic, quantum-chemical, physical-chemical, topological descriptors and others. They are assumed to be important to understand molecular characteristics such as bioactivity against cancer. In fact one of the purposes of a research like this is to find which descriptors of the molecules are better related to the disease under study, in this example cancer. The software used to compute these descriptors were e-Dragon (Virtual Computational Laboratory , 2010), a product from the Virtual Computational Laboratory and Gaussian 98 (Frisch et al., 1998).

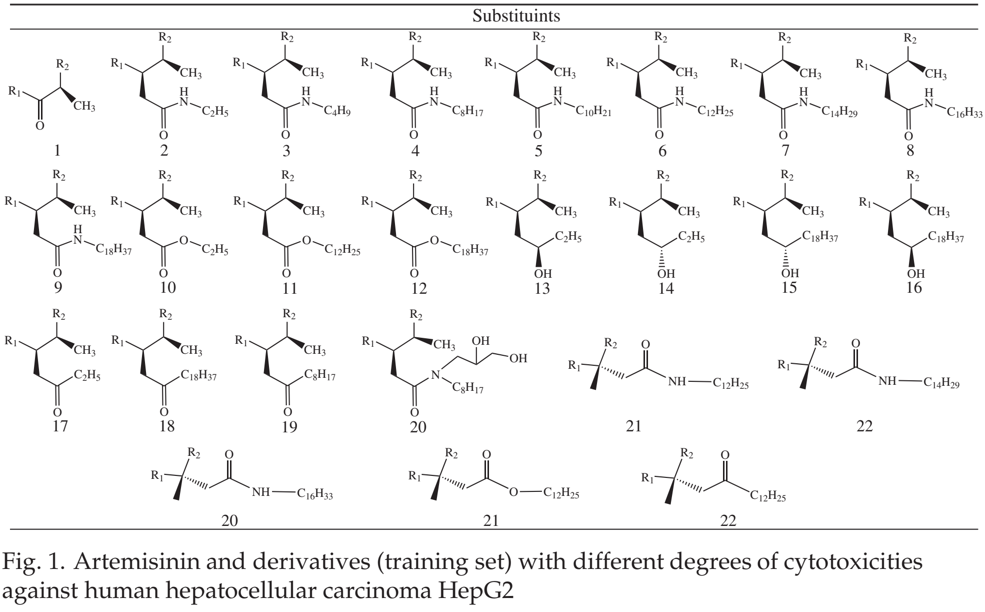


Fig. 1. Artemisinin and derivatives (training set) with different degrees of cytotoxicities against human hepatocellular carcinoma HepG2

However, a crucial point to be considered in any data analysis is preprocessing. The original data matrix usually does not have optimal value distribution for the analysis (for example it has different units and variances in variables), which requires some pretreatment prior to multivariate analysis. In general, the autoscale preprocessing, which results in scaled variables with zero mean and unit variance, is used (Ferreira, 2002). Then, all variables were auto-scaled as a preprocessing so that they could be standardized and this way could have the same importance regarding the scale.

Then the next step consists on application of multivariate statistical methods to find key features involving molecules, descriptors and anticancer activity. The methods include principal component analysis (PCA), hiererchical cluster analysis (HCA), K-nearest neighbor method (KNN), soft independent modeling of class analogy method (SIMCA) and stepwise discriminant analysis (SDA). The analyses were performed on a data matrix with dimension 25 lines (molecules) x 1700 columns (descriptors), not shown for convenience. For a further study of the methodology applied there are standard books available such as (Varmuza & Filzmoser, 2009) and (Manly, 2004).

2.1 PCA

Suppose that in your study, like in the example exhibited in this chapter, you have a large set of data, certainly it will not be a simple task to analyze so many variables and extract useful information from them. It would be a "revolution" in your research if you could confidently interpret all data in a simpler way. Fortunately, with the aid of PCA technique, this "revolution" can happen. Through PCA you can reduce the total number of variables to a smaller set while maintaining as much of the original information as is possible. No matter your area of research this is a great advantage.

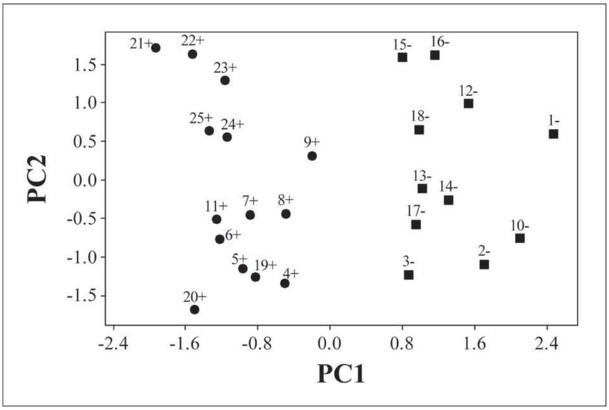


Fig. 2. Plot of PC1-PC2 scores for artemisinin and derivates (training set) with activity against human hepatocellular carcinoma HepG2. More active compounds displayed on the left side (plus sign) while less active ones on the right side (minus sign)

Now considering our data matrix, PCA was employed looking for a small group of descriptors so that they alone were responsible for classifying all 25 samples into two distinct classes: more active and less active. Besides it is desirable to choose uncorrelated descriptors that could be easier to interpret and analyze, trying to associate them to cytotoxicities against human hepatocellular carcinoma HepG2.

Furthermore, given the large quantity of multivariate data available, it was necessary to reduce the number of variables. Thus, if two any descriptors had a high Pearson correlation coefficient (r > 0.8), one of the two was randomly excluded from the matrix, since theoretically they describe the same property to be modeled (biological response). Therefore it is sufficient to use only one of them as an independent variable in a predictive model (Ferreira, 2002). Moreover those descriptors that showed the same values for most of the samples were eliminated too.

Compound	*IC5*	*Mor29m*	*O1*	*MlogP*	**Activity**
1	4.862	-0.305	-0.246	2.845	97
2	5.253	-0.308	-0.200	2.630	>100
3	5.389	-0.372	-0.202	3.080	>100
4	5.628	-0.445	-0.194	4.845	9.5
5	5.684	-0.474	-0.205	5.250	2.8
6	5.624	-0.525	-0.214	5.644	1.2
7	5.501	-0.514	-0.211	6.027	0.46
8	5.364	-0.518	-0.191	6.400	0.79
9	5.225	-0.501	-0.210	6.765	4.2
10	5.217	-0.236	-0.205	3.036	>100
11	5.597	-0.526	-0.218	6.050	0.72
12	5.197	-0.179	-0.225	7.171	>100
13	5.253	-0.364	-0.246	3.141	>100
14	5.253	-0.322	-0.237	3.141	>100
15	5.159	-0.294	-0.259	7.095	>100
16	5.159	-0.232	-0.258	7.095	>100
17	5.180	-0.443	-0.219	2.996	>100
18	5.168	-0.307	-0.209	7.131	>100
19	5.624	-0.485	-0.186	5.644	1.8
20	5.856	-0.518	-0.218	3.941	3.5
21	5.543	-0.562	-0.344	5.449	1.3
22	5.419	-0.560	-0.320	5.837	0.77
23	5.280	-0.591	-0.281	6.215	0.74
24	5.516	-0.498	-0.269	5.855	3.7
25	5.488	-0.545	-0.273	5.815	0.47
Mean	5.378	-0.425	-0.234	5.164	
Stardard Deviation	0.225	0.121	0.040	1.570	

Table 1. Values of the four descriptors selected through PCA for compounds from the training set

After this step, PCA was performed in order to continue reducing the dimensionality of the data, find descriptors that could be useful in characterizing the behavior of the compounds acting against cancer and look for natural clustering in the data and outlier samples. While processing PCA, several attempts to obtain a good classification of the compounds are made. At each attempt, one or more variables are removed, PCA is run and the score and loading plots are analyzed.

The score plot gives information about the compounds (similarities and differences). The loading plot gives information about the variables (how they are connected to each other and which are the best to describe the variance in the original data). Depending on the results displayed by the plots, variables remain removed or included in the data matrix. If a removal of a variable contributes to separate compounds showed by the score plot into two classes (more and less active), then in the next attempt PCA is run without this variable. But if no improvement is achieved, then the variable removed is inserted in the data matrix, another variable is selected to be removed and PCA is run again. The loadings plot gives good clues on which variables must be excluded. Variables that are very close to one another indicate they are correlated and, as stated before, only one of them needs to remain.

This methodology comprises part of the art of variable selection: patience and intuition are the fundamental tools here. It is not necessary to mention that the more you know about the system you are investigating (samples and variables and how they are connected), the more you can have success in the process of finding variables that really are important to your investigation. Variable selection does not occur like magic, at least, not always!

The descriptors selected in PCA were *IC5*, *Mor29m*, *O1* and *MlogP*, which represent four distinct types of interactions related to the molecules, especially between the molecules and the biological receptor. These descriptors are classified as steric (*IC5*), 3D-morse (*Mor29m*), electronic (*O1*) and molecular (*MlogP*). The main properties of a drug that appear to influence its activity are its lipophilicity, the electronic effects within the molecule and the size and shape of the molecule (steric effects) (Gareth, 2003).

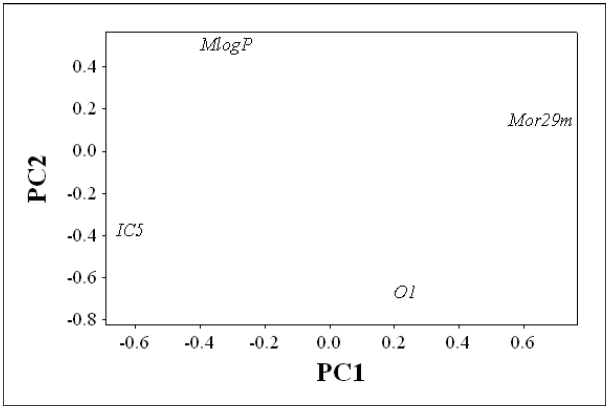


Fig. 3. Plot of the PC1-PC2 loadings for the four descriptors selected through PCA

The PCA results show the score plot (Fig. 2) relative to the first and second principal components. In PC1, there is a distinct separation of the compounds into two classes. More active compounds are on the left side, while less active are on the right side. They were chosen among all data set (1700 descriptors) and they are assumed to be very important to investigate anticancer mechanism involving artemisinins. Table 1 displays the values computed for these four descriptors. This step was crucial since a matrix with 1700 columns was reduced to only 4 columns. No doubt it is more appropriate to deal with a smaller matrix. The first three principal components, PC1, PC2 and PC3 explained 43.6%, 28.7% and 20.9% of the total variance, respectively. The Pearson correlation coefficient between the variables is in general low (less than 0.25, in absolute values); exception occurs between *Mor29m* and *IC5*, which is -0.65).

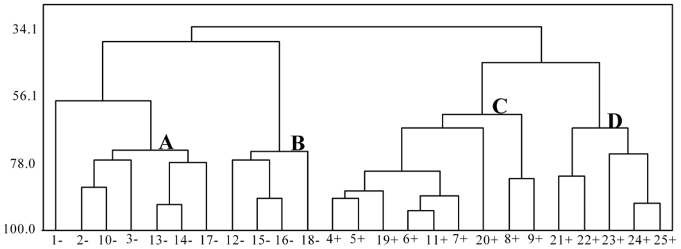


Fig. 4. HCA dendogram for artemisinin and derivatives (training set) with biological activity against human hepatocellular carcinoma HepG2. Plus sign for more active compounds while minus sign for less active ones

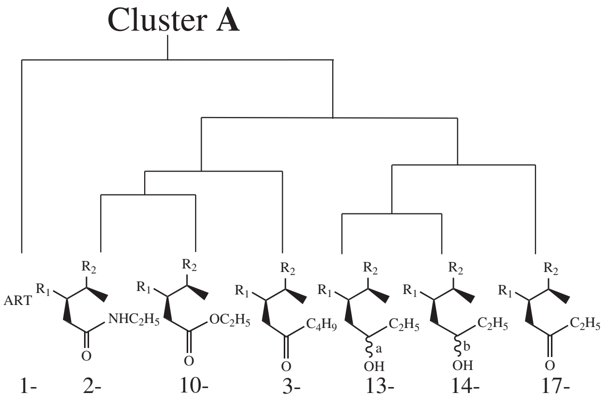


Fig. 5. Cluster **A**

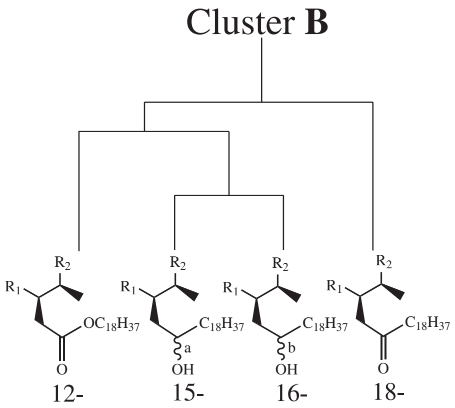


Fig. 6. Cluster **B**

The loading plot relative to the first and second principal components can be seen in Fig. 3. PC1 and PC2 are expressed in Equations 1 and 2, respectively, as a function of the four selected descriptors. They represent quantitative variables that provide the overall predictive ability

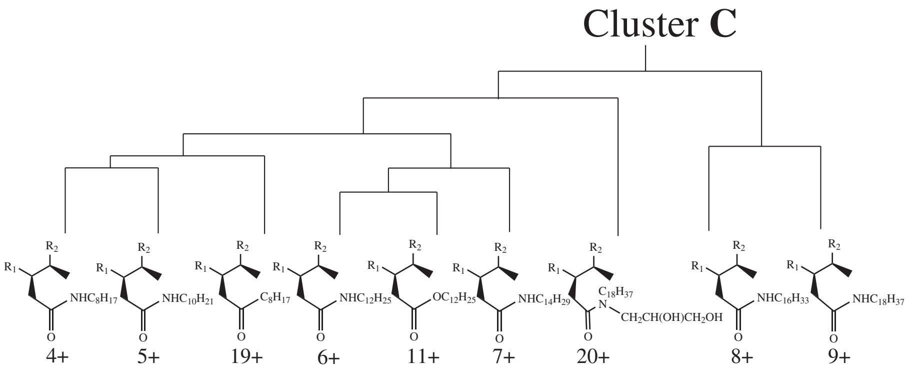


Fig. 7. Cluster **C**

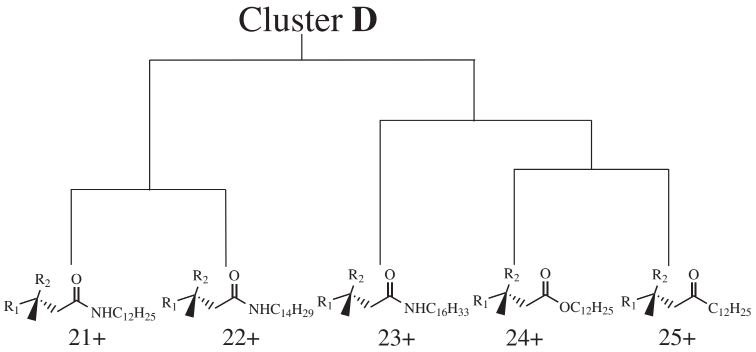


Fig. 8. Cluster **D**

of the different sets of molecular descriptors selected. In Equation 1 the loadings of *IC5* and *MlogP* are negative whereas they are positive for *Mor29m* and *O1*. Among all of them *IC5* and *Mor29m* are the most important to PC1 due to the magnitude of their coefficients (-0.613 and 0.687, respectively) in comparison to *O1* and *MlogP* (0.234 and -0.313, respectively). For a compound to be more active against cancer, it must generally be connected to negative values for PC1, that is, it must present high values for *IC5* and *MlogP*, but more negative values for *Mor29m* and *O1*.

$$PC1 = -0.613IC5 + 0.687Mor29m + 0.234O1 - 0.313MlogP \quad (1)$$

$$PC2 = -0.445IC5 + 0.081Mor29m - 0.743O1 + 0.493MlogP \quad (2)$$

2.2 HCA

Considering the necessity of grouping molecules of similar kind into respective categories (more and less active ones), HCA is suitable for this purpose since it is possible to visualize the disposition of molecules with respect to their similarities and so make suppositions of how they may act against the disease. When performing HCA many approaches are available. Each one differs basically by the way samples are grouped.

Compound	K1	K2	K3	K4	K5	K6
1	-	-	-	-	-	-
2	-	-	-	-	-	-
3	-	-	-	-	-	-
4	+	+	+	+	+	+
5	+	+	+	+	+	+
6	+	+	+	+	+	+
7	+	+	+	+	+	+
8	+	+	+	+	+	+
9	+	+	+	+	+	+
10	-	-	-	-	-	-
11	+	+	+	+	+	+
12	-	-	-	-	-	-
13	-	-	-	-	-	-
14	-	-	-	-	-	-
15	-	-	-	-	-	-
16	-	-	-	-	-	-
17	-	-	-	-	-	-
18	-	-	-	-	-	-
19	+	+	+	+	+	+
20	+	+	+	+	+	+
21	+	+	+	+	+	+
22	+	+	+	+	+	+
23	+	+	+	+	+	+
24	+	+	+	+	+	+
25	+	+	+	+	+	+

Table 2. Classification of compounds from the training set according to KNN method

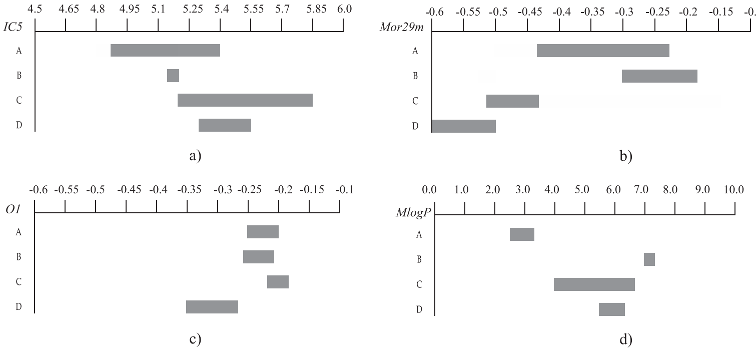


Fig. 9. Variations in descriptors: a) Variations in *IC5* for each cluster; b) Variations in *Mor29m* for each cluster; c) Variations in *O1* for each cluster; d) Variations in *MlogP* for each cluster

In this work, classification through HCA was based on the Euclidean distance and the average group method. This method established links between samples/cluster. The distance between two clusters was computed as the distance between the average values (the mean vector or centroids) of the two clusters. The descriptors employed in HCA were the same selected in

PCA, that is, *IC5*, *Mor29m*, *O1* and *MlogP*. The representation of clustering results is shown by the dendogram in Fig. 4, which depicts the similarity of samples. The branches on the bottom of the dendogram represent single samples. The length of the branches linking two clusters is related to their similarity. Long branches are related to low similarity while short branches mean high similarity. On the scale of similarity, a value of 100 is assigned to identical samples and a value of 0 to the most dissimilar samples. For a better interpretation of the dendogram, the clusters are also analyzed alone (Figs. 5, 6, 7 and 8), and variations in descriptors in each cluster are presented in Fig. 9. The scale above each figure is associated to the property considered and the letters indicate the cluster in the dendogram. It is easily recognized that descriptors in clusters in general have different pattern of variations, a characteristic supported by the fact that clusters have different groups of molecules.

Group or class	Number of Compounds	Compounds wrongly classified					
		K1	K2	K3	K4	K5	K6
Less active	11	0	0	0	0	0	0
More active	14	0	0	0	0	0	0
%Correct information	25	100	100	100	100	100	100

Table 3. Classification matrix obtained by using KNN

The dendogram shows compounds classified into two different classes according to their activities with no sample incorrectly classified. Less active compounds are on the left side and are divided into clusters **A** (Fig. 5) and **B** (Fig. 6). In cluster **A** substituints have either C_2H_5 (**2**, **10**, **13**, **14** and **17**) or C_4H_9 (**3**). Here the lowest values for *IC5* (Fig. 9a) and *MlogP* are found (Fig. 9d). In cluster **B** (**12**, **15**, **16** and **18**) all substituints have $C_{18}H_{37}$ and are present the highest values for *MlogP* (Fig. 9d). Considering more active samples, right side of the figure, in cluster **C** (Fig. 7) compounds have amide group (exception is **11**, ester, and **19**, ketone) and attached to this group there is an alkyl chain of 8 to 18 carbon atoms. Here the descriptor *IC5* displays the highest values (Fig. 9a). In Cluster **D** (Fig. 8) substituints have an alkyl chain of 12 to 16 carbon atoms and the six-membered ring molecules with oxygen O_{11} are replaced by five-membered ring molecules. Compounds display the lowest values for *Mor29m* (Fig. 9b) and *O1* (Fig. 9c).

Besides these two methods of classification (PCA and HCA), others (KNN, SIMCA and SDA) were applied to data. They are important to construct reliable models useful to classify new compounds (test set) regarding their ability to face cancer. This is certainly the ultimate purpose of many researches in planning a new drug.

2.3 KNN

This method categorizes an unknown object based on its proximity to samples already placed in categories. After built the model, compounds from the test set are classified and their classes predicted taking into account the multivariate distance of the compound with respect to K samples in the training set. The model built for KNN in this example employs leave one out method, has 6 (six) as a maximum k value and autoscaled data. Table 2 shows classification for each sample at each k value. Column number corresponds to k setting so that the first column of this matrix holds the class for each training set sample when only one neighbor (the nearest) is polled whereas the last column holds the class for the samples when the kmax nearest neighbors are polled. Tables 2 and 3 summarizes the results for KNN analysis. All 6-nearest neighbors classified samples correctly.

2.4 SIMCA

The SIMCA method develops principal component models for each training set category. The main goal is the reliable classification of new samples. When a prediction is made in SIMCA, new samples insufficiently close to the PC space of a class are considered non-members. Table 4 shows classification for compounds from the training set. Here sample **9** was classified incorrectly since its activity is 4.2 (more active) but it is classified by SIMCA as less active.

	Compound																								
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Class	-	-	-	+	+	+	+	+	-	-	+	-	-	-	-	-	-	-	+	+	+	+	+	+	+

Table 4. Classification of compounds from the training set according to SIMCA method

Probably the reason for this misclassification lies in the fact that compound **9** may not be "well grouped" into one of the two classes. In fact when you analyze Fig. 2 you note that **9** is the compound classified as more active that is closer to compounds classified as less active.

Group or Class	Number of Compounds	True group	
		More active	Less active
Less active	11	0	11
More active	14	14	0
Total	25		
%Correct information		100	100

Table 5. Classification matrix obtained by using SDA

2.5 SDA

SDA is also a multivariate method that attempts to maximize the probability of correct allocation. The main objectives of SDA are to separate objects from distinct populations and to allocate new objects into populations previously defined.

The discrimination functions for less active and more active classes are, respectively, Equations 3 and 4, given below:

$$Y_{LESS} = -5.728 - 2.825MlogP - 0.682O1 - 3.243IC5 + 7.745Mor29m \tag{3}$$

$$Y_{MORE} = -3.536 + 2.220MlogP + 0.536O1 + 2.548IC5 - 6.086Mor29m \tag{4}$$

The way the method is used is based on the following steps:

(a) Initially, for each molecule, the values for descriptors (*IC5*, *Mor29m*, *O1* and *MlogP*) are computed;

(b) The values from (a) are inserted in the two discrimination functions (Equation 3 and Equation 4). However, since these equations were obtained from autoscaled values from Table 1 (training set), it is necessary that values from Table 7 (test set) are autoscaled before inserted into the equations;

(c) The two values computed from (b) are compared. In case the value calculated from Equation 3 is higher than that from Equation 4, then the molecule is classified as less active. Otherwise, the molecule is classified as more active.

Group or Class	Number of Compounds	True group	
		More active	Less active
Less active	11	0	11
More active	14	14	0
Total	25		
%Correct information		100	100

Table 6. Classification matrix obtained by using SDA with Cross Validation

Through SDA all compounds of the training set were classified as presented in Table 5. The classification error rate was 0% resulting in a satisfactory separation between more and less active compounds.

The reliability of the model is determined by carrying out a cross-validation test, which uses the leave-one-out technique. In this procedure, one compound is omitted of the data set and the classification functions are built based on the remaining compounds. Afterwards, the omitted compound is classified according to the classification functions generated. In the next step, the omitted compound is included and a new compound is removed, and the procedure goes on until the last compound is removed. The obtained results with the cross-validation methodology are summarized in Table 6. Since the total of correct information was 100%, the model can be believed as being a good model.

Compound	*IC5*	*Mor29m*	*O1*	*MlogP*
26	5.371	-0.437	-0.238	2.461
27	5.526	-0.544	-0.249	2.496
28	5.402	-0.516	-0.241	1.649
29	5.336	-0.481	-0.239	2.461
30	5.572	-0.553	-0.238	2.305
31	5.464	-0.411	-0.226	3.117
32	5.584	-0.323	-0.244	3.328
33	5.282	-0.496	-0.226	3.225
34	5.483	-0.570	-0.345	2.090
35	5.583	-0.667	-0.262	2.922

Table 7. Values of the four descriptors for the compounds from the test set

2.6 Classification of unknown compounds

The models built from compounds from the training set through PCA, HCA, KNN, SIMCA and SDA now can be used to classify others compounds (test set, Fig. 10) whose anticancer activities are unknown. So ten compounds were proposed here to verify if they must be classified as less active or more active against a human hepatocellular carcinoma cell line, HepG2. In fact, they were not selected from any literature, so it is supposed that they have not been tested against this carcinoma. These compounds were selected so that they have substituitions at the same positions as those for the training set (R1 and R2) and the same type of atoms. It is important to keep the main characteristics of the compounds that generated the models. This way good predictions can be achieved. The classification of the test set was based on the four descriptors used in the models: *IC5*, *Mor29m*, *O1* and *MlogP*, according to Table 7.

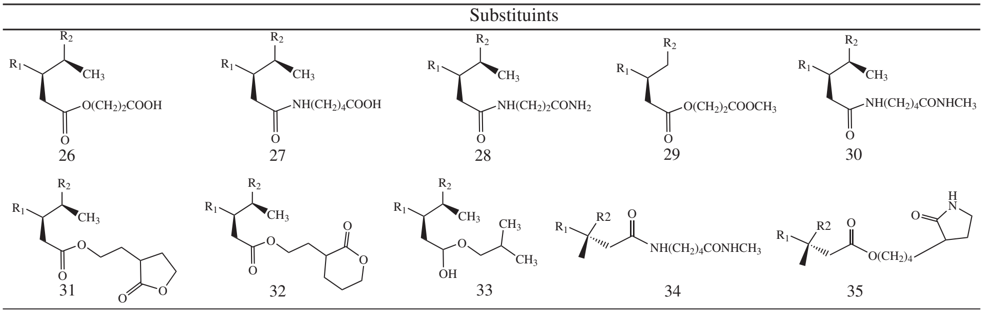


Fig. 10. Compounds from the test set which must be classified as either less active or more active

Compound	PCA	HCA	KNN	SIMCA	SDA
26	-	-	-	-	-
27	+	+	+	+	+
28	-	-	-	-	-
29	-	-	-	-	-
30	+	+	+	+	+
31	-	-	-	-	-
32	-	-	-	-	-
33	-	-	-	-	-
34	+	+	+	+	+
35	+	+	+	+	+

Table 8. Predicted classification for unknown compounds from the test set through different methods. Minus sign (-) for a compound classified as less active while plus sign (+) for a compound classified as more active

The result presented in Table 8 reveal that all samples (test set) receive the same classification by the four methods. Compounds **26**, **28**, **29**, **31**, **32** and **33** were classified as less active while compounds **27**, **30**, **34** and **35** were classified as more active. If you look for an explanation for such a pattern you will note that **26** and **27** present carboxylic acid group at the end of the chain, but only **27** is classified as more active. So it is possible that the change of an ester group by an amide group causes increase in activity. However when two amide groups are considered as occurs in **28** and **30** more carbon atoms in substituent means more active. Now comparing **26**, **29**, **31** and **32**, all of them have ester group associated with another different group and they all are classified as less active. The presence of the second group seams not to modify activity too much. The same effect is found in **34** and **35**, both more active.

3. Conclusion

All multivariate statistical methods (PCA, HCA, KNN, SIMCA and SDA) classified the 25 compounds from the training set into two distinct classes: more active and less active according to their degree of anticancer HepG2 activity. This classification was based on *IC5*, *Mor29m*, *O1* and *MlogP* descriptors. They represent four distinct classes of interactions related to the molecules, especially between the molecules and the biological receptor: steric (*IC5*), 3D-morse (*Mor29m*), electronic (*O1*) and molecular (*MlogP*).

A test set with ten molecules with unknown anticancer activity has its molecules classified, according to their biological response, into more active or less active compound. The results reveal in which classes they are grouped. In general molecules classified as more active must be seen as more efficient in cancer treatment than those classified as less active. Then the developed studies with PCA, HCA, KNN, SIMCA and SDA can provide valuable insight into the experimental process of syntheses and biological evaluation of the new artemisinin derivatives with activity against cancer HepG2. Without chemometrics no model and, consequently, no classification could be possible unless you are a prophet!

The interfacioal location of chemometrics, falling between measurements on the one side and statistical and computational theory and methods on the other, poses a challenge to the new practioner (Brow et al., 2009). The future of chemometrics lies in the development of innovative solutions to interesting problems. Some of the most exciting opportunities for innovation and new developments in the field of chemometrics lie at the interface between chemical and biological sciences. These opportunities are made possible by the exciting new scientific advances and discoveries of the past decade (Gemperline, 2006).

Finally, after reading this chapter you certainly must have noticed that chemometrics is a useful tool in medicinal chemistry, mainly when the great diversity of data is taken into account, because a lot of conclusions can be achieved. A study like this one here presented, where different methods are employed, is one of the examples of how chemometrics is important in drug design. Thus applications of statistics in chemical data analysis looking for the discovery of more efficacious drugs against diseases must continue and will certainly help researches.

4. References

Barbosa, J.; Ferreira, J.; Figueiredo, A.; Almeida, R.; Silva, O.; Carvalho, J.; Cristino, M.; Ciriaco-Pinheiro, J.; Vieira, J. & Serra, R. (2011). Molecular Modeling and Chemometric Study of Anticancer Derivatives of Artemisini. *Journal of the Serbian Chemical Society*, Vol. 76, No. 9, (September 2011), pp. 1263-1282, ISSN 0352-5139

Brereton, R. (2009). *Chemometrics for Pattern Recognition*, John Wiley & Sons, Ltd, ISBN 978-0-470-74646-2, West Sussex,England

Brereton, R. (2007). *Applied Chemometrics for Scientists*, John Wiley & Sons, Ltd, ISBN 978-0-470-01686-2, West Sussex, England

Brown, S.; Tauler, R. & Walczak, B. (Ed(s)) (2009). *Compreensive Chemometrics: Chemical and Biochemical Data Analysis*, Vol. 1, Elsevier, ISBN 978-0-444-52702-8, Amsterdan, The Netherlands

Bruns, R.; Scarminio, I. & Barrros Neto, B. (2006) *Statistical Design - Chemometrics*, Elsevier, ISBN 978-0-444-52181-1, Amsterdan, The Netherlands

Cardoso, F.; Figueiredo, A.; Lobato, M.; Miranda, R.; Almeida, R. & Pinheiro, J. (2008). A Study on Antimalarial Artemisinin Derivatives Using MEP Maps and Multivariate QSAR. *Journal of Molecular Modeling*, Vol. 14, No. 1, (January 2008), pp. 39-48, ISSN 0948-5023

Doweyko, A. (2008). QSAR: Dead or Alive? *Journal of Computer-Aided Molecular Design*, Vol. 22, No. 2, (February 2008), pp. 81-89, ISSN 1573-4951

Efferth, T. (2005). Mechanistic Perspectives for 1,2,4-trioxanes in Anti-cancer Therapy. *Drug Resistance. Updat*, Vol. 8, No.1-2, (February 2005), pp. 85-97, ISSN 1368-7646

Ferreira, M. (2002). Multivariate QSAR. *Journal of the Brazilian Chemical Society*, Vol.13, No. 6, (November/December 2002), pp. 742-753, ISSN 1678-4790

Fujita, T. (1995). *QSAR and Drug Design: New Developments and Applications*, Elsevier, ISBN 0-444-88615-X, Amsterdan, The Netherlands

Gareth, T. (2003). *Fundamental of Medicinal Chemistry*, John Wiley & Sons, Ltd, ISBN 0-470-84307-1, West Sussex, England

Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, Jr J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K.N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K. J.; Foresman, B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E. S. & Pople, J. A. (1998) *Gaussian, Inc.*, Gaussian 98 Revision A.7, Pittsburgh PA

Gemperline, P. (2006). *Practical Guide to Chemometrics* (2nd), CRC Press, ISBN 1-57444-783-1, Florida, USA

Varmuza, K. & Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, ISBN 9781420059472, Florida, USA

Lai, H.; Nakasi, I.; Lacoste, E.; Singh, N. & Sasaki (2009). T. Artemisinin-Transferrin Conjugate Retards Growth of Breast Tumors in the Rat. *Anticancer Research*, Vol. 29, No. 10, (October 2009), pp. 3807-3810, ISSN 1791-7530

Levine, I. (1991). *Quantum Chemistry* (4th), Prentice Hall, ISBN 0-205-12770-3, New Jersey, USA

Liu,Y.; Wong, V.; Ko, B.; Wong, M. & Che, C. (2005). Synthesis and Cytotoxicity Studies of Artemisinin Derivatives Containing Lipophilic Alkyl Carbon Chains. *Organic Letters*, Vol. 7, No. 8, (March 2005), pp. 1561-1564. ISSN 1523-7052

Pinheiro, J.; Kiralj, R.; & Ferreira, M. (2003). Artemisinin Derivatives with Antimalarial Activity against Plasmodium falciparum Designed with the Aid of Quantum Chemical and Partial Least Squares Methods. *QSAR & Combinatorial Science*, Vol. 22, No. 8, (November 2003), pp. 830-842, ISSN 1611-0218

Manly, B. (2004). *Multivariate Statistical Methods: A Primer* (3), Chapman and Hall/CRC, ISBN 9781584884149, London, England

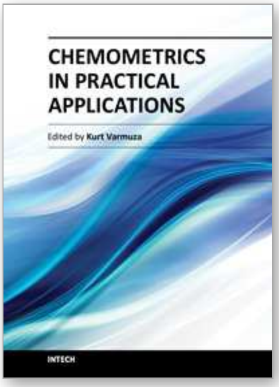
Price, R.; van Vugt, M.; Nosten, F.; Luxemburger, C.; Brockman, A.; Phaipun, L.; Chongsuphajaisiddhi, T. & White, N. (1998). Artesunate versus Artemether for the Treatment of Recrudescent Multidrug-resistant Falciparum Malaria. *The American Journal of Tropical Medicine and Hygiene*, Vol. 59, No. 6, (December 1998), pp. 883-888, ISSN 0002-9637

Puzyn, T.; Leszczynski, J. & Cronin, M. (Ed(s)). (2010). *Recent Advances in QSAR Studies: Methods and Applications*, Springer, ISBN 978-1-4020-9783-6, New York, USA

Rajarshi, G. (2008). On the interpretation and interpretability of quantitative structure-activity relationship models. *Journal of Computer-Aided Molecular Design*, Vol. 22, No. 12, (December 2008), pp. 857-871, ISSN 1573-4951

Wold, S. (1995). *Chemometrics, what do we mean with it, and what do we want from it? Chemometrics and Intelligent Laboratory Systems*, Vol. 30, No. 1, (November 1995), pp. 109-115, ISSN 0169-7439

Virtual Computational Laboratory, VCCLAB In: e-Dragon, 13.05.2010, Available from http://www.vcclab.org



Chemometrics in Practical Applications
Edited by Dr. Kurt Varmuza

ISBN 978-953-51-0438-4
Hard cover, 326 pages
Publisher InTech
Published online 23, March, 2012
Published in print edition March, 2012

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821