Sesquiterpene Lactones with Anti-Hepatitis C Virus Activity Using Molecular Descriptors

LETTERS IN DRUG DESIGN & DISCOVERY, SHARJAH, v. 9, n. 9, pp. 881-890, NOV, 2012
http://www.producao.usp.br/handle/BDPI/41019

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo
Sesquiterpene Lactones with Anti-Hepatitis C Virus Activity Using Molecular Descriptors

Michelle Rossini¹, Marcus Tullius Scotti², Mauro Vicentini Correia¹, Harold Hilarion Fokoue¹, Luciana Scotti*³, Francisco Jaime Bezerra Mendonça Júnior⁴, Marcelo Sobral da Silva³ and Vicente de Paulo Emerenciano¹

¹Departamento de Química, Institute of Chemistry, University of São Paulo, Brazil; ²Center of Applied Sciences and Education, Federal University of Paraíba; ³LTF Laboratory of Pharmaceutical Technology, Departamento de Farmácia, Federal University of Paraíba, Departamento de Ciências Biológicas, Universidade Estadual da Paraíba, Brazil; ⁴Laboratório de Síntese e Vetorização de Moléculas, Departamento de Ciências Biológicas, Universidade Estadual da Paraíba, Rua Horácio Trajano de Oliveira s/n, Cristo Redentor, João Pessoa 58070-450, PB, Brazil.

Received April 25, 2011; Revised August 13, 2012: Accepted August 15, 2012

Abstract: Hepatitis C is a worldwide public health problem. The available therapies are limited by their partial effectiveness and with meaningful side-effects. Sesquiterpene lactones (SLs) are a group of natural products with a wide variety of chemical structures and biological activities associated. There are few studies about the influence of the molecular structure of SLs for the anti-hepatitis C virus activity. In the present work, SLs are investigated in a subgenomic RNA replicon assay system and were analyzed using multiple linear regression along with self-organizing maps with DRAGON descriptors in order to identify the structural requirements for their biological activity and to predict the inhibitory potency of SLs. Characteristics such as stereochemistry and electronic effects demonstrated to be important for their anti-HCV activity, and the SOM produced a clear separation between active and inactive compounds. Therefore, it is possible to use this map as a filter for virtual screening to predict the anti-HCV activity of SLs.

Keywords: Anti-hepatitis C virus, artificial neural network, descriptors, QSAR, self-organizing maps, sesquiterpene lactones

INTRODUCTION

The HCV genome heterogeneity makes prevention and treatment of hepatitis C very difficult. First of all, the diversity in the development of vaccines is hindered by the vaccine antigens which come from multiple serotypes which are necessary for the development of global protection. Furthermore, the different genotypes vary in their responsiveness to the combined therapy interferon/ribavirin, which is the current standard of care. It is also worthwhile to mention that this associated therapy is limited by its partial efficacy and meaningful side effects [1,2].

SLs are a representative group of natural products (NPs).This wide variety of chemical structures is matched by a diversity of biological activities [3-6]. There are few studies involving the influence of the molecular structure of SLs related to their anti-HCV activity [7-14].

In order to understand the relationship among physical-chemical parameters and biological activities, MLR are used to select the most relevant molecular descriptors that codify the molecular structure, by different selection strategies, like artificial neural network (ANN). ANN are defined as computational models with structures derived from the simplified concept of the brain in which a number of nodes, called neurons, are interconnected in a network-like structure. Some ANNs studies about SLs have already been developed targeting other biological activities. The most used ANNs architecture for pattern recognition and classification is the Self-organizing maps (SOM). SOM is a powerful visualization tool able to reduce dimensions of projections and displays similarities among objects, which has already been successfully used in several applications in chemical database [15-19].

METHODS AND MATERIALS

Dataset

This study analyzes nineteen compounds (Fig. 1) previously published by Hwang et al. 2006 [20], in which the anti-HCV activity of seventeen SLs (represented by four skeletons: eleven germacronolides, four eudesmanolides, one guaianolide and one pseudo guaianolide) and two lactones were investigated in a sub genomic RNA replicon assay system. Numbers, names, skeletons and in vitro potential of inhibition of these compounds are shown in Table 1. The potential of inhibition in vitro was quantified in EC50 (drug concentration producing a 50% decrease in virus replication). Molecules were classified as actives (twelve molecules with EC50 ≤ 10μM) and as inactives (seven molecules with EC50 > 10μM).

All structure data were extracted from the SISTEMAT X database which is the newest version of SISTEMAT [21]
The chemical database (using MySQL) that includes the occurrence data (ca. 28,000 entries) with a Java interface, which can draw compounds and associate their respective biological activities and botanical occurrences. After that, they were codified in 3D using .mol (mdl) files.

Molecular modeling computations were performed on Hyperchem 8.0 for Windows [22]. The molecules were subjected to geometry optimization and conformational analysis. The semi-empirical quantum chemical method used was the Austin model 1 and the root mean square gradient

Fig. (1). Molecular structures of SLs (1-8), exo-methylene lactones (9,10) and parthenolide analogues (11-19) analyzed for the anti-HCV activity.
value of 0.001 kcal/mol as termination condition. The semi-empirical AM1 method is suitable to give a satisfactory conformation of small organic compounds but as a semi-empirical molecular orbital method it has limited accuracy. The major drawback of semi-empirical methods is in the parametrization process, as conformation flexibility [8, 23-25].

The molecules were saved as .mol (mdl) files for computing various molecular descriptors using DRAGON Professional v. 5.4 [26].

### Molecular Descriptors

Three dimensional descriptors provided by DRAGON were calculated, totaling 680 descriptors: 2D autocorrelations, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, GETAWAY, WHIM, among others [27-32]. For each block of descriptors, the constant variables were excluded, as well as those that presented only a different value of the series. For the remaining descriptors, pairwise correlation (r < 0.99) analysis was performed to exclude those highly correlated. Thus, were used 336 3D descriptors in our final calculations.
SOM inactive and active compounds. how many structures in each neuron, numbers 0 and 1 represent the and active compounds, respectively. Numbers in brackets show selection in MLR. The black and gray colors represent the inactive and active compounds. Fig. (2).

3D-MoRSE (103 descriptors) descriptors are related to the stereochemistry of the compounds3D-MoRSE descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transformation used in electron diffraction studies for preparing theoretical scattering curves. GETAWAY (108 descriptors) descriptors have been proposed as chemical structure descriptors derived from a new representation of molecular structure. WHIM (57 descriptors) are geometrical descriptors based on statistical indices calculated on the projections of the atoms along principal axes in order to capture relevant molecular 3D information regarding molecular size, shape, symmetry and atom distribution. Geometrical Descriptors (20 descriptors) are defined in several different ways but always derived from the three-dimensional structure of the molecule.

RDF (48 descriptors) are based on a radial distribution function which can be interpreted as the probability distribution of finding an atom in a spherical volume of radius.

SOM

All DRAGON descriptors selected previously by pairwise correlation analysis were analyzed with SOMs in Matlab 6.5 and SOM Toolbox 2.0 [34-36]: Matlab is a powerful and easy program to use with scientific computing language and is the choice for most scientific simulation and data analysis. SOM toolbox is a set of Matlab functions that can be used for creation, visualization and analysis of self-organizing maps. The literature shows that the determination of the size of the SOM is an empiric process [16,17]. Initially, a heuristic formula of \( m = 5(n)^{0.2} \) is used for total number of map units, where \( n \) is the number of samples. The ratio of side lengths is based on the two biggest eigenvalues of the covariance matrix of the given data. As a result, SOM with different dimensions were analyzed for the anti-HCV activity.

The descriptors of the training set were selected using the Batch-training algorithm, where the dataset is presented to the network before doing any adjustment. In each training step, the dataset is partitioned according to the regions of the map weight vectors and it was divided in order to reduce the intrinsic error in the dependent variable, dividing the compounds into two sets: active and inactive. Then, the correct prediction of these groups and the total correct prediction of compounds are evaluated. This step will be performed again using only the selected descriptors by multiple linear regression (MLR). The dataset contained only 19 compounds which did not allow the use of a test set that is not critical for an unsupervised and robust method as SOM.

**Calculation of MLR**

MobyDigs program [37] was used for the calculation of MLR models by using Gas [38] with \( pEC_{50} = -\log EC_{50} \) (EC50 provided by the antiviral activity of the twelve molecules with specific activity in the subgenomic HCV replicon system) as dependent variable, and all DRAGON descriptors as independent variables.

As it is necessary to select the most statistically significant models, the search for the best ones are usually performed using the ordinary least squares regression under the GA approach, that is: by the variable subset selection-genetic algorithm method. In the GA terminology, a population is characterized by a set of candidate variables (the genetic heritage of the population) and is constituted by individuals, it means: models made of one or more population variables [39].

The inactive molecules were not considered in this approach because they don’t have a definite EC50 value. Therefore, we will have only 12 compounds to perform the MLR and it is not possible to use an external test set that is a suitable analysis of the predictive performance. The validation will be performed by using the cross-validation (leave-one-out) method.

**Self-organizing Maps Generated by Selected Descriptors in MLR**

The self-organizing maps were generated by using only the selected descriptors by MLR. Each contribution of descriptor for the classification of SLs according to their anti-HCV activity [40].

The selected descriptors (present in the equation generated by the MLR) were reanalyzed in Kohonen RNNs with the 19 compounds, after data pre-treatment, in Matlab 6.5 [34] and SOM Toolbox 2.0.[35].

**CONCLUSION**

After pre-treatment, where constant variables, as well as those that presented only a different value of the series and the highly correlated were excluded, remaining 336 3D descriptors. Table 2 summarizes the results of different SOM trainings before the descriptors selection in MLR. Fig. (2) shows the best Kohonen map (using 3D-MoRSE block of descriptors) obtained for the 19 compounds at the SOM analysis done before the descriptors selection in MLR. The
map with dimension 4 x 3 presents 89.5% of match for all analyzed compounds, 100% for the inactive compounds and 83.3% for the active ones.

After the linear regression using 336 3D descriptors and GA to select them, the best model showed the descriptors MOR11u, G1s and HATS3v (equation 1) with the highest value for Qcv, and it was able to explain 94% of variance for the anti-HCV activity. The self-organizing maps were generated using only the selected descriptors by MLR.

\[
\text{EC50} = +0.10(\pm0.06) \text{Mor11u} -22.12(\pm4.90) \text{G1s} +8.95(\pm2.17) \text{HATS3v} +8.74(\pm0.76) \quad (n=12; r^2=0.94; s=0.05; F=39.78; Qcv^2=0.88; SPRESS=0.06) \tag{1}
\]

Using equation 1 to calculate the error, which is the difference between the calculated activity values and the experimental activity values (pEC50 calculated – pEC50 experimental) for the anti-HCV compounds (Table 3), and plotting these data on a graphic (Fig. 3), the adjustment of a straight line of the points used for the calibration of the model demonstrates the significant correlation degree previously verified.

The block of descriptors present in the most robust model were: 3D MoRSE (Mor11u), WHIM (G1s) and GETAWAY (HATS3v). The value of the coefficient of internal prediction

---

**Table 3. Experimental values of pEC50, values calculated by Equation 1 and errors for the anti-HCV compounds.**

<table>
<thead>
<tr>
<th>Compound*</th>
<th>pEC50 experimental</th>
<th>pEC50 calculated</th>
<th>Error (Calculated-Experimental)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.66</td>
<td>5.66</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>5.57</td>
<td>5.53</td>
<td>-0.04</td>
</tr>
<tr>
<td>3</td>
<td>5.51</td>
<td>5.59</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>5.90</td>
<td>5.89</td>
<td>-0.02</td>
</tr>
<tr>
<td>5</td>
<td>5.69</td>
<td>5.65</td>
<td>-0.04</td>
</tr>
<tr>
<td>12</td>
<td>5.65</td>
<td>5.66</td>
<td>0.01</td>
</tr>
<tr>
<td>13</td>
<td>5.60</td>
<td>5.56</td>
<td>-0.04</td>
</tr>
<tr>
<td>14</td>
<td>5.55</td>
<td>5.60</td>
<td>0.05</td>
</tr>
<tr>
<td>15</td>
<td>5.79</td>
<td>5.81</td>
<td>0.02</td>
</tr>
<tr>
<td>16</td>
<td>5.69</td>
<td>5.64</td>
<td>-0.05</td>
</tr>
<tr>
<td>17</td>
<td>5.27</td>
<td>5.27</td>
<td>0.00</td>
</tr>
<tr>
<td>18</td>
<td>5.65</td>
<td>5.67</td>
<td>0.02</td>
</tr>
</tbody>
</table>

* Compounds which molecular structures are in Fig. (1).
Q_{cv}^2 is significant (0.88) and the F value 39.78, with 95% of trust with 3 and 8 grades of freedom.

The G1s descriptors were calculated by using the first component of the molecule (the axis of higher length). The higher the number of symmetric atoms according to the electrotopology, higher is the G1 value.

The HATS3v descriptor is a kind of GETAWAY descriptor, and codifies the information that is related to the influence of the molecular shape. The value of this descriptor will be higher, when the molecules that have higher van der Waals.

After obtaining and interpreting equation 1, the SOM were generated to relate the properties codified in its descriptors to the biological activity.

Fig. (4) shows the best Kohonen map obtained for the 19 compounds at the SOM analysis done after the selection of the descriptors in the MLR. The map with dimension size 4 x 3 presents 100% of match for all analyzed compounds. It can be seen, that there is a high similarity between the maps obtained with G1s and HATS3v descriptors. Both maps present similar weights, with the lowest values in the northeastern and the higher values in the southeastern of the maps. The observation for SOMs was generated with the same dimensions (4 x 3) associating only two descriptors at the same time: Mor11u with HATS3v (Fig. 5) and MOR11u with G1s (Fig. 6).

The maps obtained with Mor11u and HATS3v descriptors (Fig. 5) with Mor11u and G1s descriptors (Fig. 6) presented 100% and 89.5% of match, respectively, for all
analyzed compounds. 85.7% for the inactive compounds and 91.7% for the active ones. This proved that the maps had the ability to distinguish active molecules from the inactive ones considering only two descriptors. As the combined analysis of the Mor11u and HATS3v descriptors (Fig. 5) showed the same maximum capacity that had been observed for the analysis with three descriptors, they were the pair of descriptors of choice.

Anti-HCV activity (-log EC<sub>50</sub>) and Mor11u and HATS3v descriptors values for each compound are reported on Table 4.

In general, it can be stated that the most representative descriptors are those that have two principal features: the highest weight in the predominantly active and inactive regions, and a considerably difference between the highest and lowest values of descriptors.

Some considerations could be made after analyzing the structures of the compounds (Fig. 1), dividing them in active and inactive sets and then comparing them according to the chemical groups.

First of all, it is important to note that all active compounds analyzed by equation 1 have the α-metilene-γ-lactone moiety (1-5) or are substituted analogues that, somehow, recover this portion (11-18).

In the active set of the analyzed molecules (EC<sub>50</sub> ≤ 10μM), there are germacronolides (1, 2, 12-18), guaianolide (3), pseudo-guaianolide (4) and eudesmanolide (5); however, the majority of eudesmanolides (6, 7, 8) and the lactones (9, 10) are inactive.

The descriptors Mor11u and HATS3v confirm the positive influence of the oxygen atoms and the stereochemistry of the α-metilene-γ-lactone moiety for the anti-HCV activity.

These observations allow us to explain the high values for the activity of the compounds 4 and 5, which respectively are, the first and the third most active compounds. Both have the configuration 7β - 8β in the fusion of the α-metilene-γ-lactone ring with the terpenoid skeleton. This fact also helps to explain the fact that compound 5 is the only active eudesmanolide among the four eudesmanolides analyzed.

Nevertheless, it is observed that the HATS3v descriptor does not permit to explain exactly the activity of the parthenolide (1) analogues (12-18), which have amino substituent groups in the 13position. These groups collaborate to acquire a linear periphery, instead of a spherical one, that would be much more influent to HATS3v. This way, it is possible to partially justify why compound 17, even having a substituent with bulky atoms (nitrogen and oxygen) and with topological distance 3 between them, doesn’t have a high biological activity. Additionally, compounds 1(active) and 6 (inactive) present pretty similar values for HATS3v, once compound 6 has more oxygen atoms that collaborate for its high value of HATS3v, as well as 9 and 10. However 6 is a tricycle, making the molecule more rigid and it will also makes the influence of the oxygen atoms is comparable to the influence of the lower quantity of this atom in 1, that is more spherical, and so, promotes a higher influence of the oxygen atoms and increasing the HATS3v value.

Although the descriptor G1s is not present in the SOM map chosen to be analyzed in this work, it appears in equation 1 and its interpretation shows that molecules less symmetric have a higher anti-HCV activity than the symmetric compounds. This evidence is valid to explain the low activity of compound 3, once its configuration makes the compound to be more spherical.

The descriptors selected in equation 1 bring out the importance of characteristics as stereochemistry and electronic effects caused by the presence of atoms with high
volume (as oxygen and nitrogen) with topological distance 3 between them and located in the periphery of the molecule.

The Kohonen network produced a clear separation of the active and inactive compounds on the map with only two of the three initially selected descriptors. Furthermore, the SOM had a match of 100%. Therefore, it is possible to use this map as a filter for virtual screening databases with the aim to predict the anti-HCV activity of unknown SLs.

The difficulty to explain the biological activity of analogues (12-18) with the analyzed descriptors corroborate with the results observed previously by Hwang et al. [20] who have already demonstrated that the same analogues were not recovered from the media where the anti-HCV activity was analyzed.

In addition to the importance of the spatial arrangement provided by the terpenoid skeleton fused with an α-methylene-γ-lactone already stated, it has also been verified that when there is the 7β - 8γ configuration in the fusion of the α-methylene-γ-lactone ring with the terpenoid skeleton, the anti-HCV activity increases, at least in eudesmanolides and pseudoguaianolides derivatives. Some structural modifications in these SLs skeletons, as including amino substituent groups in position 13 are suitable according to this study.

Although the lack of a more accurate value of EC50 to discriminate the inactivity of some compounds (which does not allow the use in regression analysis) and the difficult interpretation of the selected descriptors (that are not fast and directly interpretable as the sigma of Hammet, for instance, but allows a good interpretation), the present study achieved its objective: it combined different and complementary methodologies to identify more structural requirements for the anti-HCV activity and to predict the inhibitory potency of unknown SLs.

This preliminary study examined the quantitative relationship between 19 SLs and theirs activities anti-hepatitis C virus, using MLR and SOM. Other research projects will be carried out using the same methodology, based on the highest number of natural products: chalcones, coumarins, flavonoids, terpenes and SLs.

**CONFLICT OF INTEREST**

The authors confirm that this article content has no conflicts of interest.

**ACKNOWLEDGEMENTS**

This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).
The authors thank the financial support of CNPq and Inct-IF.

**ABBREVIATIONS SECTION**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D-Morse</td>
<td>Molecule Representation of Structure based on Electron diffraction</td>
</tr>
<tr>
<td>ANN</td>
<td>artificial neural network</td>
</tr>
<tr>
<td>EC50</td>
<td>half maximal effective concentration</td>
</tr>
<tr>
<td>F</td>
<td>F-test value</td>
</tr>
<tr>
<td>GETAWAY</td>
<td>GGeometry, Topology, and Atom-Weights Assembly</td>
</tr>
<tr>
<td>HCV</td>
<td>hepatitis C virus</td>
</tr>
<tr>
<td>MLR</td>
<td>multiple linear regressions</td>
</tr>
<tr>
<td>NPs</td>
<td>Natural products</td>
</tr>
<tr>
<td>pEC50</td>
<td>negative logarithm of EC50</td>
</tr>
<tr>
<td>Qc^2</td>
<td>Cross validated squared correlation coefficient</td>
</tr>
<tr>
<td>QSAR</td>
<td>Quantitative structure–activity relationship</td>
</tr>
<tr>
<td>r^2</td>
<td>Coefficient of determination</td>
</tr>
<tr>
<td>RDF</td>
<td>Radial Function Distribution</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SLs</td>
<td>Sesquiterpene lactones</td>
</tr>
<tr>
<td>SOMs</td>
<td>self-organizing maps</td>
</tr>
<tr>
<td>SPRESS</td>
<td>standard deviation of sum of square of difference between predicted and observed values</td>
</tr>
<tr>
<td>WHIM</td>
<td>Weighted Holistic Invariant Molecular</td>
</tr>
<tr>
<td>μM</td>
<td>micromole</td>
</tr>
</tbody>
</table>

**REFERENCES**


