Computer-Aided Drug Design Methodologies Toward the Design of Anti-Hepatitis C Agents

Alejandro Speck-Planche, M. Natália D. S. Cordeiro

Abstract: Hepatitis C constitutes an infectious disease that causes severe damages to the liver, and is caused by hepatitis C virus. There is no vaccine against this type of disease and the number of people infected continues to grow worldwide. The anti-viral therapy which is currently used is a mixture of interferon alpha-2a with ribavirin, but approximately half of the patients do not respond to therapy. Therefore, it is necessary to search for new compounds with anti-hepatitis C activity. Computer-aided drug design methodologies have been vital in the discovery of candidates to drugs. This review is dedicated to the role of computer-aided drug design methodologies for the development of new anti-hepatitis C agents. In addition, we introduce a QSAR model based on substructural approaches in order to model the anti-hepatitis C activity in vivo.

Keywords: Anti-HCV agents, QSAR, 3D-QSAR, structure-based drug design, linear discriminant analysis, fragments.

INTRODUCTION

Hepatitis C (HC) is an infectious disease that affects the liver. It is caused by the hepatitis C virus (HCV), the only known member of the hepacivirus genus in the family Flaviviridae [1]. Although this infection is often asymptomatic, once established, chronic infection can progress to scarring of the liver (fibrosis), and advanced scarring (cirrhosis) which is generally apparent after many years. In some cases, those with cirrhosis will go on to develop liver failure or other complications of cirrhosis, including liver cancer, and in other cases, life threatening esophageal and gastric varices. Currently HCV infection is thought to be transmitted by parenteral and nosocomial routes and it has been estimated that nearly 200 million people worldwide are infected with HCV, being in addition 2-3 million infected per year. HC has thus the potential to be the next pandemic [2, 3]. Spontaneous viral clearance rates are highly variable and between 10-60% of people infected with HCV clear the virus from their bodies during the acute phase as shown by normalization in liver enzymes (alanine transaminase (ALT) & aspartate transaminase (AST)), and plasma HCV-RNA clearance (this is known as spontaneous viral clearance). However, persistent infections are common and most patients develop chronic hepatitis C [4]. The prevalence of HC is higher in Asian and African countries, where the health conditions are extremely unfavorable. The worst case is Egypt, country which is believed that HC is linked to a now-discontinued mass-treatment campaign for schistosomiasis, which is endemic in that country [5]. An important factor in HCV is its ability to cause co-infection with the human immunodeficiency virus (HIV) [6], constituting that an important cause of mortality in immunocompromised patients.

 Nowadays, it is very difficult to treat HC. The current treatment is a combination of pegylated interferon-alpha-2a [7] (can be used also pegylated interferon-alpha-2b) and the anti-viral drug ribavirin for a period of 24 or 48 weeks, depending on hepatitis C virus genotype [8-11]. Interferon alpha-2a is a protein which is released by lymphocytes in response to the presence of pathogens —such as viruses, bacteria, or parasites— or tumor cells (Fig. 1). It is one of the proteins which allows communication between cells to trigger the protective defenses of the immune system that eradicate pathogens or tumors. In its pegylated form (40 kDa; commercial name Pegasyx) is an anti-viral drug discovered at the pharmaceutical company F. Hoffmann-La Roche; it has a dual mode of action - both anti-viral and on the immune system. The addition of polyethylene glycol to the interferon, through a process known as pegylation, enhances the half-life of the interferon when compared to its native form.
Fig. (2). Structure of ribavirin.

On the other hand, ribavirin is a nucleoside anti-metabolite used as an anti-viral agent which blocks nucleic acid synthesis and is used against both RNA and DNA viruses (Fig. 2). This drug inhibits the activity of the enzyme RNA dependent RNA polymerase, due to its resemblance to building blocks of the RNA molecules. The oral form is used in the treatment of HC, in combination with interferon drugs.

Anyway this therapy is generally recommended for treatment of patients with proven HCV chronic infection and persistently abnormal liver function tests. However, approximately 50% of patients do not respond to anti-viral therapy. In general terms, we can say that there is no vaccine against HC. For this reason, the search of an effective anti-viral therapy against HCV, constitutes a challenge for the scientific community, in order to discover in a more rational and efficient way, compounds with anti-HC activity. In this sense, computer-aided drug design (CADD) methodologies constitute a vital branch of Chemoinformatics [12-14] and Bioinformatics [15-18], which have been determinant for the better understanding of infinite chemical, biochemical and biological processes in order to perform a rational drug design in less time and with an appreciable resource savings. That is because they are focused on the use of several computational approaches that permit the processing of biological data which will be related with the chemical structure to different levels of complexity and diversity. This review is focused on the current state of the CADD methodologies toward the design of new anti-HCV agents. Herein, we confirm the great importance of CADD by introducing a model based on substructural approaches for the design of new anti-viral agents with potent HC activity in vivo.

**CADD METHODOLOGIES FOR THE DISCOVERY OF DRUGS AND TARGETS**

Nowadays, CADD methodologies for the development of new anti-HCV agents and in general, to any class of drugs, can be divided into two great categories. The first category is related to methodologies which are focused on the knowledge of the three-dimensional structure of the biological receptor obtained through experimental methods such as X-ray crystallography or NMR spectroscopy. If the structure of the receptor is not available, a homology model of the biological receptor can be created. This first category of methodologies is often known as structure-based drug design. The second category comprises methodologies which do not take into consideration the three-dimensional structure of the receptor in explicitly. These are based on the search of molecular patterns from the structure of the molecule which show a high biological activity because of the interaction with the receptor. The name usually accepted for this kind of methodologies is ligand-based drug design. Mainly the second category but (more recently) the first include the type of models known as Quantitative Structure-Activity Relationship (QSAR) models. In all these cases we can find models that use the structural parameters of molecular systems like drugs, protein structure, RNA secondary structure, protein-protein interaction networks (PINs), genes network as input to predict the properties of such system (output). That is why, the editor González-Díaz has extended the discussion to different collective of authors editing special issues on CADD techniques (including QSAR and others), which have been published on journals such as the Current Topics in Medicinal Chemistry [19-28], Current Proteomics [29-36], Current Drug Metabolism [37-45], Current Pharmaceutical Design [46-55], and Current Bioinformatics [56-65].

**Structure-Based Drug Design**

In general terms, structure-based drug design is strongly related with Bioinformatics, which is concerned with the application of statistics and computer science to the field of molecular biology, and it has been determinant for a better understanding of biochemical and biological processes. Techniques such as molecular docking, homology modeling, structure-based virtual screening and many others comprise the select group of molecular modeling techniques (MMT) in structure-based drug design.

In the field of anti-viral therapy against HCV, several works have been reported using MMT [66-74]. In general terms, almost all the works have been focused in the discovery of potent anti-HCV agents through the inhibition of NS5B and NS3 proteins. NS5B is the RNA dependent RNA polymerase of Hepatitis C virus, and like other RNA dependent RNA polymerases, is error prone. This viral RNA replicase displays approximately a million times lower fidelity than a replicative prokaryotic or eukaryotic DNA polymerase. This is partly due to the fact that NS5B contains no exonuclease or proofreading domain. In NS5B, two divalent cations coordinated by carboxyl groups (as seen in DNA polymerases) catalyze the polymerization of monomers of RNA triphosphates to extend a primer strand, which may have initiated de novo. In the case of NS5B, the residues that coordinate divalent cations (Mg$^{2+}$ or Mn$^{2+}$ in vitro) are the three active site aspartates (220, 318 and 319). Another target which has been studied is the NS3 serine protease. This HCV-encoded serine protease is essential for viral replication and, hence, is an attractive target for HCV-specific anti-viral therapy.

An important work was reported by combining QSAR tools with ligand based (LB) and structure based (SB) alignment procedures for in silico screening of new HCV-NS5B polymerase allosteric inhibitors [67]. Herein, the combination of a complete computational procedure together with biological studies led to the identification of novel molecular scaffolds, hitherto untested toward NS5B polymerase. Structure based 3D-QSAR models were generated employing NS5B non-nucleoside inhibitors (NNIs), whose bound conformations were readily available from the protein database (PDB). These were grouped into two training sets of structurally diverse NS5B NNIs, based on their binding to the enzyme thumb (15 NNIs) or palm (10 NNIs) domains. Thus, LB and SB alignments were rigorously investigated to assess the reliability on the correct molecular alignment for unknown binding mode of the modeled compounds. Computational approaches applied were able to reproduce with...
minimal errors the experimental binding conformations of 24 experimental NS5B allosteric inhibitors (Fig. 3). Eighty-one (thumb) and 223 (palm) modeled compounds taken from literature were LB and SB aligned and used as external validation sets for the development of 3-D QSAR models. Low error of prediction proved the 3-D QSARs to be useful scoring functions for the in silico screening procedure. Finally, the virtual screening of the NCI Diversity Set led to the selection for enzymatic assays of 20 top scoring molecules for each final model. Among the 40 selected molecules, preliminary data yielded four derivatives exhibiting IC$_{50}$ values ranging between 45 and 75 μM.

On the other hand, a promising work was reported for the identification of new NS5B inhibitors by pharmacophore-guided virtual screening [68]. This work was focused also on the search of new NS5B inhibitors using pharmacophore-guided virtual screening [70]. Towards that goal, the aforementioned methodology was applied to identify novel HCV-NS5B inhibitors. The pharmacophore model generated from this previous analysis of the binding modes, as well as structure-based 3D-QSAR studies of aryl diketo acid analogues, were used. In this pharmacophore-guided virtual screening study, among 37 447 compounds from the LeadQuest chemical library [75], 40 compounds were selected as new candidates of HCV-NS5B inhibitors, and their biological activities been evaluated. Especially, T29 was chosen for further development (Fig. 5). This fact was because this compound resulted to be highly active against HCV-NS5B.

In the field of investigations for the development of NS3 serine protease inhibitors, a work based on the prediction of binding for a kind of non-peptic HCV-NS3 serine protease inhibitors from plants was reported [74]. Molecular doc-king (MDock) combined with the Molecular Mechanics and Poisson-Boltzmann Surface Area (MM-PBSA) technique
were applied to predict the binding mode of polyphenol inhibitors in the binding pocket of the HCV NS3 serine protease for which the ligand-protein crystal structure is unavailable (Fig. 6). The most favorable geometry of three candidates from MDock studies had a binding free energy about 3 and 6 kcal/mol more favorable in one candidate than in the other two and was identified as the correct binding mode. For such mode, the correlation of the calculated and experimental binding affinities of all five polyphenol compounds was satisfactory as indicated by the determination coefficient ($R^2 = 0.92$). The most favorable binding mode suggested that two galloyl residues at 3 and 4 positions of the glucopyranose ring of the inhibitors interact with SER139, GLY137, ALA157, and ASP81 by hydrogen bond interactions and with ALA156 and HIE57 by hydrophobic interaction and are essential for the activities of the studied inhibitors.

**Ligand-Based Drug Design**

In the context of the methodologies which rely on the knowledge of the structure of the receptor, Quantitative Structure-Activity Relationships techniques (including 3D-QSAR approaches) have played a determinant role. These methodologies have been strongly supported by chemometric methods such as multiple linear regression (MLR) [13, 76, 77], linear discriminant analysis (LDA) [77, 78], artificial neural networks (ANNs), partial least squares (PLS) [76-78] and many other techniques. On the other hand, in the last 10 years, complex network theory (CNT) has provided a deeper rapprochement between ligand-based and structure-based drug design, and leading to a better understanding of several biological phenomena to different levels of chemical diversity and complexity [79-81].

Some works have been published by application of QSAR techniques for the discovery of new anti-HCV compounds [66, 67, 69, 71, 82-87]. Many of these works have employed 3D-QSAR as the methodology to model the anti-HCV activity (Table 2). The aim has been to find the most active conformations of the molecules which can provide more information about the molecule-receptor interactions.

Table 1. Structure-Based Drug Design Methodologies Applied to the Discovery of NS5B Inhibitors

<table>
<thead>
<tr>
<th>Chemical Family</th>
<th>Target</th>
<th>Methodology</th>
<th>Authors</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>benzothiadiazine derivatives</td>
<td>NS5B</td>
<td>MDock and MD</td>
<td>Wang et al.</td>
<td>[66]</td>
</tr>
<tr>
<td>carboxylic acid derivatives</td>
<td>NS5B</td>
<td>PBVS and MDock</td>
<td>Musmucu et al.</td>
<td>[67]</td>
</tr>
<tr>
<td>HDBC</td>
<td>NS5B</td>
<td>PBVS and MDock</td>
<td>Ryu et al.</td>
<td>[68]</td>
</tr>
<tr>
<td>benzinimazole derivatives</td>
<td>NS5B</td>
<td>MDock</td>
<td>Patel et al.</td>
<td>[69]</td>
</tr>
<tr>
<td>HDBC</td>
<td>NS5B</td>
<td>PBVS and MDock</td>
<td>Kim et al.</td>
<td>[70]</td>
</tr>
<tr>
<td>diketo acid derivatives</td>
<td>NS5B</td>
<td>PBVS and MDock</td>
<td>Di Santo et al.</td>
<td>[72]</td>
</tr>
<tr>
<td>hexapeptides</td>
<td>NS3</td>
<td>MDock</td>
<td>Frecer et al.</td>
<td>[73]</td>
</tr>
<tr>
<td>polyphenol derivatives</td>
<td>NS3</td>
<td>MDock and MM-PBSA</td>
<td>Xudong et al.</td>
<td>[74]</td>
</tr>
</tbody>
</table>

HDBC: Heterogeneous Database of Compounds; NS5B: RNA dependent RNA polymerase of hepatitis C virus; NS3: serine protease of hepatitis C virus; MDock: Molecular Docking; MD: Molecular Dynamics; PBVS: Pharmacophore Based-Virtual Screening; MM-PBSA: Molecular Mechanics and Poisson-Boltzmann Surface Area.

Ligand-Based Drug Design

Fig. (5). Structure of T29, the most potent HCV-NS5B inhibitor.

Fig. (6). Structure of a polyphenol extracted from *Rhodiola kirilowii*, the most potent NS3 serine protease inhibitor.

Table 1. Structure-Based Drug Design Methodologies Applied to the Discovery of NS5B Inhibitors
conformations employing the atom fit alignment method. Second, receptor-based 3D-QSAR models were derived from the predicted binding conformations obtained by docking all NS5B inhibitors at the allosteric binding site of NS5B (PDB ID: 2dxs). The predictive ability of the models was validated using a structurally diversified test set of 22 compounds that had not been included in the preliminary training set of 45 compounds. The predictive $R^2$ values for the ligand-based CoMFA and CoMSIA models were 0.734 and 0.800, respectively, while the corresponding predictive $R^2$ values for the receptor-based CoMFA and CoMSIA models were 0.538 and 0.639, respectively. Tryptophan derivatives showed greater potency than the tyrosine derivatives (Fig. 7), and that was interpreted in terms of the CoMFA steric and electrostatic contour maps. The CoMSIA results revealed that for a NS5B inhibitor to have appreciable inhibitory activity it required hydrogen bond donor and acceptor groups at the 5-position of the indole ring and a R substituent at the chiral carbon, respectively. Interpretation of the CoMFA and CoMSIA contour maps in context of the topology of the allosteric binding site of NS5B provided insight into NS5B-inhibitor interactions. Taken together, these 3D-QSAR models were found to accurately predict the HCV-NS5B polymerase inhibitory activity of structurally diverse test set compounds and to yield reliable clues for further optimization of the benzimidazole derivatives in the data set.

Another promising work was based on the performance of CoMFA and CoMSIA towards thiazolone derivatives as hepatitis C virus NS5B polymerase allosteric inhibitors [85]. With this aim, 3D-QSAR models for a series of thiazolone derivatives as novel inhibitors bound to the allosteric site of HCV-NS5B polymerase were developed (Fig. 8). Two different conformations of the template molecule and the combinations of different CoMSIA fields were considered to build predictive CoMFA and CoMSIA models. The CoMFA and CoMSIA models with best predictive ability were obtained by the use of a template conformation from X-ray crystal structures. The information obtained from the CoMFA and CoMSIA 3D-contour maps enabled the interpretation of their structure-activity relationship and was also used to the design of several new inhibitors with improved activity.

Finally, an ANNs analysis was performed for predicting the antigenic activity for a major conformational epitope in the hepatitis C virus NS3 protein [86]. This fact was motivated by the insufficient knowledge of general principles for accurate quantitative inference of biological properties from
sequences, and is a major obstacle in the rationale design of proteins with predetermined activities. Due to this deficiency, protein engineering frequently relies on the use of computational approaches focused on the identification of Structure-Activity Relationships (SAR) for each specific task. In that work, a computational model was developed to define SAR for a major conformational antigenic epitope of the HCV non-structural protein 3 (NS3) in order to facilitate a rationale design of HCV antigens with improved diagnostically relevant properties. In this sense, one ANNs model that connected changes in the antigenic properties and structure of HCV-NS3 recombinant proteins representing all 6 HCV genotypes was developed. The ANNs performed quantitative predictions of the enzyme immunoassay (EIA) Signal/Cutoff (S/Co) profiles from sequence information alone with 89.8% accuracy. Amino acid positions and physicochemical factors strongly associated with the HCV NS3 antigenic properties were identified. The positions most significantly contributing to the model were mapped on the NS3 3D-structure. The location of these positions validated the major associations found by the ANNs model between antigenicity and structure of the HCV-NS3 proteins.

**QSAR FRAGMENT-BASED APPROACH TOWARD THE DESIGN OF ANTI-HCV AGENTS**

All the computational approaches and methodologies which have been discussed until now, have been determinant for the search and design of novel anti-HCV agents. However, a disadvantage remains, that is to say, a compound can be extremely active against any of the targets that nowadays are constantly studied, but that is a necessary and sufficient condition to ensure that the same compound which was a very potent receptor inhibitor, will be active in cells already infected by HCV? In an attempt to overcome this problem we introduce here a QSAR model based on a fragment-based approach.

**Methods**

**Atom-Centered Fragments**

Atom-centered fragments have demonstrated to be very useful descriptors, and have been employed in some QSAR studies, providing useful information about the hydrophobic and dispersive interactions that are involved in both transport and distribution of the drugs through the membrane and drug-receptor interactions [88-91]. These descriptors are defined as the number of fragments which contain specific atom types in a molecule and they are calculated from the molecular composition and atom connectivities. Thus, each type of atom in the molecule is described in terms of its neighboring atoms. Hydrogen and halogen atoms are classified by the hybridization and oxidation state of the carbon atom to which they are bound and for hydrogen atoms, heteroatoms attached to a carbon atom in α-position are further considered. Carbon atoms are classified by their hybridization state and depending on whether their neighbors are carbon or heteroatoms.

**Functional Group Counts**

These are other type of descriptors that express certain fragmental features. These are simple molecular descriptors defined as the number of specific functional groups in a molecule and are also calculated from the molecular composition and atom connectivities. The functional groups defined by these descriptors are those traditionally used in Organic Chemistry.

**Spectral moments of the bond adjacency matrix**

The approach which encloses the calculation of the spectral moments of the bond adjacency matrix is known as the TOPS-MODE (TOPOlogical Substructural MOlecular DEsign) approach, and it has been applied on the modeling of some physicochemical properties of organic compounds [92-98]. Also, these descriptor have been extended to QSAR studies [99-106], and the analysis of toxicological properties [107-114]. In order to perform the calculation of spectral moments, the molecular structure is codified by means of the edge adjacency matrix \( \mathbf{E} \) (known also as the bond adjacency matrix \( \mathbf{B} \)) [115]. The \( \mathbf{E} \) matrix is a square symmetric matrix of order \( m \) (\( m \) being the number of chemical bonds in the molecular graph) whose elements \( e_{ij} \) are equal to 1 if the bonds \( i \) and \( j \) are adjacent, that is if they are incident to a common atom, or 0 otherwise. In order to codify information of heteroatoms, the TOPS-MODE approach uses weighted matrices \( \mathbf{E}(wij) \), instead of \( \mathbf{E} \). The weights \( wij \) are chemically meaningful quantities such as bond distances, bond dipoles, bond polarizabilities or mathematical expressions involving atomic weights. For this reason, the spectral moments of the bond adjacency matrix can be used as molecular fingerprints in QSAR studies and molecular design [116-118].

By mathematical definition, the term spectral moment must be understood as the sum of the main diagonal elements \( (e_{ii})\) of the natural powers of the weighted \( \mathbf{E} \) matrix. Then, the spectral moments of order \( k \) (\( \mu_k \)) are defined as:

\[
\mu_k = Tr(\mathbf{E}^k) = \sum_i (e_{ii})^k
\]

where Tr means the trace of the matrix, that is the sum of the diagonal entries \( (e_{ii}) \) of the k-th power of the weighted \( \mathbf{E} \) matrix. In a similar way, local spectral moments are defined as the sum of the diagonal entries of different powers of the edge matrix corresponding to a given molecular fragment:

\[
\mu_k(f) = \sum_{i|f} (e_{ii})
\]

where \( f \) is the corresponding fragment for which the spectral moment is to be defined and the sum is carried out over all bonds that form the fragment \( f \).

**Selection of the Dataset: Calculation of the Descriptors and Development of the Model**

The dataset comprised 1958 compounds (see supplementary material file 1 available upon request to corresponding authors; Supp. Inf. 1) and 937 of them were considered as active against the HCV [119]. The condition to consider a compound as active was that \( EC_{50} \leq 2.5 \mu M \), being \( EC_{50} \), the half maximal effective concentration. The inactive group was formed by 1021 compounds, many of them were compounds with \( EC_{50} > 2.5 \mu M \) and also, drugs which belong to different therapeutic categories [120]. This dataset was divided into training and prediction series. The training series
contained 1469 compounds: 703 active and 766 inactive, while the prediction series was formed by 489 compounds: 234 active and 255 inactive. The atom-centered fragment and functional group count descriptors were calculated using DRAGON program (version 5.3) [121]. The spectral moments of the weighted edge adjacency matrix, were calculated using Modeslab software (version 1.5) [122]. In this case, the spectral moments were weighted by the dipole moments, molar refractivities and Abraham molar refractivities. As the modeling technique, we have elected the linear discriminant analysis (LDA), which has been used extensively in QSAR studies [123-131], in order to find a classification model (eq. 3), which best describes the anti-HCV activity ($A_{\text{HCV}}$), as a linear combination of the predictor $X$-variables (molecular descriptors $D_k$), with the coefficients $a_k$. Such coefficients are optimized by means of LDA, specifically the LDA technique implemented in the STATISTICA software (version 6.0) [132], using only the training set compounds.

$$A_{\text{HCV}} = a_0 + a_1 D_1 + a_2 D_2 + \ldots + a_k D_k$$  \hspace{1cm} (3)

In developing the models, $A_{\text{HCV}}$ values of +1 and -1 were assigned to active and inactive compounds, respectively, but a posteriori probabilities were then used instead to assert the classification of compounds. In particular, when the a posteriori probabilities were then used instead to assert the assigned to active and inactive compounds, respectively, but having as few descriptors as possible, was chosen.

The statistical quality of the model was estimated by examining several statistical indices, such as the Wilks’ lambda ($\lambda$), the squared of the Mahalanobis’s distance ($D^2$), the Fisher ratio ($F$), the corresponding $p$-level, as well as the percentage of good classifications. The Wilks $\lambda$ statistic is a multivariate measure of the group differences over several variables, and can take values from zero (perfect discrimination) to one (no discrimination). The $D^2$ statistic is also a measure of the separation between the active and inactive groups, it shows if the model displays an appropriate discriminatory power for differentiating those groups.

Additionally, other statistical indices were used to confirm the quality of the model, and to validate it, such as Sensitivity ($Sens$) – the ability for classifying active cases, Specificity ($Spec$) – the ability for classifying inactive cases, Accuracy ($Acc$) – the overall predictivity, the positive predictive value (PPV), and the negative predictive value (NPV). These indices were determined according to the following equations:

$$Sens = \frac{TP}{C+} \times 100$$  \hspace{1cm} (4)

$$Spec = \frac{TN}{C-} \times 100$$  \hspace{1cm} (5)

$$Acc = \frac{TP + TN}{(C+) + (C-)} \times 100$$  \hspace{1cm} (6)

$$PPV = \frac{TP}{TP + FP} \times 100$$  \hspace{1cm} (7)

$$NPV = \frac{TN}{TN + FN} \times 100$$  \hspace{1cm} (8)

where $TP$ means the cases (compounds) classified correctly by the model as active, $C+$ the total active compounds, $TN$ means the cases classified correctly by the model as inactive, $C-$ represents the total inactive compounds, and $FP$ and $FN$ stand for the false positives and false negatives, respectively.

Finally, we also evaluated the predictive ability of our final discriminant model by using an external set of compounds not used in the model setup. It should be emphasized that, validation of the final model with compounds, which are not part of the training set, is a crucial but necessary step to ensure generalization, and also of great relevance to future QSAR studies.

**Discriminant Model**

The best classification model derived from the training set, by combining the LDA and FS techniques along with the $D_i$ structural representation, is given below together with the statistical parameters of the LDA:

$$A_{\text{HCV}} = + 6.219 \times 10^{-10} \mu_{\text{Ab}} - 3.1 \times 10^{-2} \mu_{\text{Ab}}^{(2)} + 3.722 \times 10^{-3} \mu_{\text{Ab}}^{(3)} + 2.448(S-108) + 0.612(\text{ArOH}) - 0.427(F-084) + 0.429(\text{ROR})$$  \hspace{1cm} (9)

and

$$N=1469 \ \lambda=0.523 \ \ D^2=3.645 \ \ F(14,1454)=94.586 \ \ p<0.001$$

This model includes three types of descriptors, i.e. three spectral moments of the edge adjacency matrix, six functional group counts and five atom-centered fragments. Thus, $\mu_{\text{Ab}}^{(\text{Ab}-\text{R}^2)}$ represents the spectral moment of order 15, weighted by the Abraham molar refractivity, $\mu_{\text{Ab}}^{(\text{Dip})}$ the spectral moment of order 2, weighted by the dipole moments and $\mu_{\text{Ab}}^{(\text{MBR})}$ represents the spectral moment of order 2, weighted by the molar refractivity. In the case of functional groups counts, descriptor $\text{ArOH}$ means the number of phenol fragments, $\text{ROR}$ is the number of aliphatic ether residues, $\text{Pyrr}$ represents the presence of pyrrole rings, $\text{RsR}$ indicates the number of aliphatic thioether residues, $\text{Isotz}$ takes into consideration the presence or absence of isothiazole rings and $\text{Pymd}$ represents the presence or absence of pyrimidine fragments. Regarding atom-centered fragments, descriptor $S-108$ indicates the number of fragments containing a carbon-sulfur double bond, $F-084$ means the number of fragments which contain fluorine atom attached to a $\text{C(sp)}_2$ atom which is attached at the same time with a electronegative atom (only O, N, S, P, Se, and halogens are considered electronegative atoms) by a simple bond, $O-060$ represents the number of fragments in which an oxygen atom defines aliphatic-aromatic, aromatic, or aromatic cyclic ethers, or when the oxygen atom is attached with two carbon atoms, being one of them attached to an electronegative atom by a double bond, while $C-039$ takes into account the number of fragments where a $\text{C(sp)}_2$ atom is attached to an aromatic ring, with an aliphatic group and with an electronegative atom by a double bond. Finally, $C-042$ indicates the presence or absence of cyclic fragments in which a $\text{C(sp)}_2$ atom is attached...
Furthermore, the large $F$ index, and small $p$ value are indicative of the model’s statistical significance. In addition, the values of the Wilks statistic and of the Mahalanobis distance show that the model displays an adequate discriminatory power for differentiating active from inactive groups. The latter is also confirmed by the classification results; the model has a sensitivity of 83.21% and a specificity of 85.38% in the training series, for an accuracy of 84.34%. Additionally, the positive and the negative predictive values were 83.93% and 84.71%, respectively. This means that, if the QSAR model predicts a compound that resulted to be anti-HCV agent, the probability of this compound to be really active is 83.93%. Similarly, if the model predicts a compound that resulted inactive as anti-HCV agents, the probability of the compound to be really inactive is 84.71%. We have also examined all the compounds, searching for misclassified cases because they can be outliers and have influence in the quality of a model, by thus checking the Mahalanobis’s distance of each molecule with respect to the two centroids of both groups (active and inactive). Generally, in the case of abnormal values, the compounds should be excluded from the model. Even though they were misclassified compounds, the deletion of them did not improve the model.

In order to validate our model, we took into consideration the sensitivity, the specificity, the accuracy and positive and negative predictive values (all these statistical indices in the prediction series). The sensitivity of the model in the prediction series was 83.33% and the specificity was 83.52%, for an accuracy of 83.42%. Positive and negative predictive values were 82.27% and 84.52% respectively. The names or codes, and the probabilities of activity for each compound (expressed as percentages) are recorded in a supplementary material file available upon request to the corresponding authors (Supp. Inf. 2).

**ROC Curve**

![ROC Curve](image)

The sensitivity and the specificity can describe adequately the quality of a model. However, these two statistical indices have disadvantages. The most important one is that they cannot provide information about how many times the probabilities indicate that a compound, observation or case will be predicted more as positive (active) than negative (inactive), and this is very important since it confirms together with the positive predictive value if a given case is active. However, that information can be provided by a Receiver-Operating Characteristic (ROC) analysis. ROC is a classic methodology from signal detection theory [133]. The ROC curve is created by plotting the true-positive rate against false-positive rate, or sensitivity against $(1− specificity)$. The ROC curve going along the diagonal from bottom left to upper right represents pure-chance performance. The areas under the ROC curves were 0.95 and 0.91 for the training and prediction series, respectively (Fig. 9). These areas can be interpreted as follows: in the case of the training series, that value of area (0.95), means that a randomly selected compound or case from the active group will has a larger value of probability than a randomly selected compound or case from the inactive group, 95% of the times. A similar conclusion can be inferred from the value of the area under the ROC curve in the prediction series. Altogether, this proves that our model is not a random classifier because the areas under the ROC curves are different and statistically significant from those obtained by random classifiers (area = 0.5).

**Structural Interpretation of the Descriptors**

In the model represented by eq. 9, three descriptors are based on spectral moments. They take into consideration different physicochemical properties. The molecular accessibility in regions with different size (codified by $\mu_1^{(Ab-R2)}$ and $\mu_2^{(MR)}$) is very important because its increment means that the molecule will have several regions which will be able to interact properly with the biological receptor, causing the enzymatic inhibition and the decrease in the proliferation of the virus. Also, these descriptors are related in some way, with the increasing of the hydrophobicity which will provide the pass of the molecule through the membrane. On the other hand, the diminution of the bond polarity (codified by $\mu_2^{(Dip)}$), will improve the activity of the molecule against HCV, because of the increment in van der Waals interactions between the molecule and the corresponding receptor.

The atom-centered fragments and the functional group count descriptors have an easy interpretation because they
indicate certain types of group of atoms that form fragments and/or functional groups. In this sense the information provided by these descriptors depends on the structure of the fragment or functional group and for this reason, it will exist a strong relationship with the reactivity (acidity, nucleophilic and electrophilic characteristic, ability to be hydrolyzed) or with specific physicochemical properties that will not be codified by the spectral moments. Although it is not possible to determine exactly which kind of property has more influence in the fragment codified by atom-centered fragments and the functional group count descriptors, the signs of the corresponding coefficients of these descriptors in the equation will provide an idea about the desirability of the different fragments, i.e. if the fragment will be favorable or unfavorable for the development of anti-HCV activity. The most important advantage of this model is the possibility of computing the quantitative contributions of any fragment to the anti-HCV activity. In this sense, we selected some fragments which are present in the molecules (Fig. 10) and we were able to calculate the above mentioned quantitative contributions of those fragments to the activity under study.

The calculation of fragment contributions provides useful information about the molecular patterns that can be determinant for the development of anti-HCV activity, and which are those with negative influence in the potency of the compound to be used as anti-HCV agent (Table 3). So, it is possible to design new molecules from those fragments with positive contributions and such molecules, in principle, should be very active as HCV inhibitors.

CONCLUSIONS

Computational-aided drug design methodologies have been essential for the discovery and development of anti-HCV agents. However, it is necessary to extend the existing techniques of drug design to the discovery of new targets for anti-viral therapy against HCV, for example, by application of molecular modeling techniques such as homology modeling. More efforts should be made for the design of new molecular entities with potential in vivo activity. Our model which resorts to a substructural approach, and is based on a large heterogeneous database of compounds, is an attempt to overcome this problem. Also, we consider that powerful and promising concepts like complex network theory should be considered for a better understanding of the biological and biochemical processes, including the study of mechanism of resistance in HCV. We consider that the future perspectives in the design of new and potent anti-HCV agents should take more into consideration the following aspects:

To develop and/or apply new approaches based in QSAR models to combine strategies used graph-theoretical descriptors in order to predict easily, rapidly and rationally, the anti-HCV activity in vivo of several compounds.

To extend and apply the structure-based drug design methodologies like homology modeling in order to find more relationships at a biomolecular level between HCV and another viruses.

To apply approaches related to complex networks that will permit to develop new strategies for the control of hepatitis C.

ACKNOWLEDGEMENTS

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) for financial support (grant Nº Pest-C/EQB/LA0006/2011).

REFERENCES


Table 3. Quantitative Contributions of Different Fragments for the Anti-HCV Activity

<table>
<thead>
<tr>
<th>ID</th>
<th>Contrib.</th>
<th>ID</th>
<th>Contrib.</th>
<th>ID</th>
<th>Contrib.</th>
<th>ID</th>
<th>Contrib.</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>0.040</td>
<td>F6</td>
<td>-2.376</td>
<td>F11</td>
<td>-2.022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2</td>
<td>-0.137</td>
<td>F7</td>
<td>-0.095</td>
<td>F12</td>
<td>-0.808</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F3</td>
<td>-0.038</td>
<td>F8</td>
<td>-1.243</td>
<td>F13</td>
<td>0.389</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F4</td>
<td>-0.445</td>
<td>F9</td>
<td>0.396</td>
<td>F14</td>
<td>0.290</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F5</td>
<td>0.358</td>
<td>F10</td>
<td>1.056</td>
<td>F15</td>
<td>-0.301</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


In ChEMBL Database (2010). Available at: http://www.ebi.ac.uk/chembl/


