

QSAR Studies on the Antiviral Compounds of Natural Origin

B. Hemmateenejad^{a,b,*}, K. Javidnia^a, M. Nematollahi^a and M. Elyasi^a

^aMedicinal & Natural Products Chemistry Research Center, Shiraz University of Medical Sciences, Shiraz, Iran

^bDepartment of Chemistry, Shiraz University, Shiraz, Iran

(Received 19 February 2008, Accepted 3 July 2008)

Two data sets of natural antiviral agents including 107 anti-HIV1 and 18 anti-polio molecules were collected and subjected to quantitative structure-activity relationship (QSAR) analyses. A wide variety of molecular descriptors belonging to various structural properties were calculated for each molecule. Multiple linear regression (MLR) based on stepwise variable selection was employed to find the most convenient quantitative models. For each antiviral data set different QSAR models were established in two steps. Firstly, for each type of molecular descriptors separate QSAR analysis was performed, and then a new QSAR model was calculated using the selected descriptors in the first phase. For both types of antiviral data sets significant QSAR models were obtained. The atom-centered fragment descriptors represented the highest impact on the anti-HIV1 activity whereas for anti-polio agents, radial distribution function and three-dimensional MoRSE descriptors showed the most significant influences. Cross-validation and a separate prediction set were used to evaluate the stability and prediction ability of the models. It was found the discovered QSAR models for anti-HIV1 and anti-polio agents could reproduce about 80% and 90% of variances in the antiviral activity data with root mean square error of prediction of 0.421 and 0.171, respectively.

Keywords: QSAR, Antiviral, HIV, Polio, Natural products, Theoretical descriptors

INTRODUCTION

Viruses produce a wide variety of clinical illnesses. A virus consists of either DNA or RNA wrapped within a protein nucleocapside [1]. The nucleocapside may be covered by an envelope composed of glycoproteins and lipids. Viral genes can code for only a limited number of proteins and viruses possess no metabolic machinery [2-4]. They are entirely dependent upon host cells for protein synthesis and replication and are therefore obligate intracellular parasites [5].

The human immune deficiency virus type-1 (HIV-1) pandemic has grown to become one of the greatest infectious disease threats to human health and social stability that the

world has ever encountered [6,7]. Nearly 40 million persons are living with HIV-1 infection and more than 21 million people have already died from HIV-induced disease. Although effective anti-retroviral therapy has slowed the epidemic in some industrialized countries, worldwide there are still an estimated 15,000 new HIV infections occurring daily. In addition to the vast personal suffering, the loss of young adult parents, caretakers, and wage-earners, HIV has created an unprecedented strain on the social and economic infrastructure of many developing countries, particularly in Sub-Saharan Africa [7-9]. These facts make it imperative that the epidemic be controlled as rapidly as possible through prevention of new infections. Although education and available public health approaches should be vigorously pursued, development of a preventive vaccine is the best hope of controlling the HIV

*Corresponding author. E-mail: hemmatb@sums.ac.ir

epidemic.

Polio (also called poliomyelitis) is a contagious, historically devastating disease that was virtually eliminated from the western hemisphere in the second half of the 20th century [10-12]. Although polio has plagued humans since ancient times, its most extensive outbreak occurred in the first half of the 1900s before the vaccination, created by Jonas Salk, became widely available in 1955. At the height of the polio epidemic in 1952, nearly 60,000 cases with more than 3,000 deaths were reported in the United States alone. However, with widespread vaccination, wild-type polio, or polio occurring through natural infection, was eliminated from the United States by 1979 and the Western hemisphere by 1991.

Therefore, design and discovery of antiviral agents are in the frontier of world health research. An effective antiviral drug must interfere with virus-coded molecular processes without affecting any cellular metabolic process. The main drawback associated with antiviral agents is that many drugs that have an inhibitory effect on virus replication; they may also inhibit other molecular processes in both infected and uninfected tissues. As a progressive field in drug design and discovery, attentions have been directed toward using natural compounds isolated from plants for treatment of a wide variety of diseases [13]. Plants represent a large and untapped potential source of antiviral agents and a large number of compounds of varied chemical structures isolated from medicinal plants have been shown to possess antiviral activity. There is an excellent review article regarding the antiviral agents isolated from plants [14]. In addition to the interesting biological activity of the natural products, they can be used as the starting point for design and discovery of more potent compounds [15].

Currently, molecular modeling and computational chemistry are the inseparable parts in drug design and discovery and no one can talk about drug design without having a bit knowledge about computational methods [16-19]. Computational methods result in saving in time and money for discovering a new drug in all steps of drug production. Among the different computational methods existed in drug design, Quantitative structure-activity relationships (QSARs) have found the major popularity [20]. QSAR, as one of the most important areas in chemometrics, gives information that is useful for drug design and medicinal chemistry [21-24].

QSAR models are mathematical equations relating biological activity of series of molecules to their chemical structure. QSAR models are particularly useful for screening chemical databases and virtual libraries before the synthesis of chemicals, for setting testing priorities, for reducing reliance on animal testing and, in conclusion, for the timely assessment of the health and environmental risks of chemicals.

The official birth date of QSAR is considered to be 1962, when Hansch *et al.* developed quantitative relationships between biological activity and the octanol-water partition coefficient [25]. It is assumed that the sum of substituent effects on the steric, electronic and hydrophobic interaction of the compounds with their receptor determines the biological activity. Another criterion in constructing the QSAR models is finding one or more molecular descriptors that represent variation in the structural property of the molecules by a number [26,27]. The derived relationships between molecular descriptors and activity are used to estimate the property of other molecules and/or finding the parameters affecting the biological activity. Multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) regression are the mostly used modeling methods in QSARs [28-30]. MLR yields models that are simpler and easier to interpret than PCR and PLS, because these methods perform regression on latent variables that do not have physical meaning.

Molecular descriptors play a fundamental role in developing QSAR models and finding a set of molecular descriptors affect the biological activity of interest is the essential part of modeling procedure in QSAR analyses. Molecular descriptors can be collected from different sources such as substituent constants, physicochemical properties, quantum chemical calculations and the theoretical structural parameters derived from one or two dimensional molecular structures [31]. Theoretical descriptors such as constitutional descriptors and topological indices have found the major popularity in QSAR studies for many reasons such as a) their calculation are simple and fast, b) they do not need information about three-dimensional structure of molecules, c) they are exact number without uncertainty and d) they represent high correlation with many biological activities [31-35].

There are some literature report on the QSAR analyses of

the natural compounds with different biological activity [36-39]. However, to the best of our knowledge, no article is available regarding the QSAR study of the natural antiviral agents. In this article, we developed QSAR models for antiviral activity (*i.e.*, anti-polio and anti-HIV1) of some chemical compounds isolated from plants. A data set of 125 molecules with natural resource (107 molecules with anti-HIV1 activity and 18 molecules with anti-polio activity) was collected and a large number of theoretical descriptors were calculated for each molecule. Multiple linear regression was used to develop QSAR equations, and the resulted models were validated by cross-validation as well as external validation.

EXPERIMENTAL

Data Set

The biological data used in this study are the antiviral activity (as IC_{50}) of a series of anti-HIV and anti-Polio compounds. The data were collected from the review article of Perez [14, and references therein], in which a large number of antiviral natural products isolated from plants is reported. However, we confined our data set to compounds whose anti-HIV1 or anti-polio activities were reported by similar research groups or similar methods. We could obtain a data set size of 107 anti-HIV1 agents and 18 anti-polio agents. The structural features and biological activity of these compounds are listed in Schemes 1 and 2, and Tables 1 and 2, respectively. The antiviral activity was taken as $-\log IC_{50}$ or pIC_{50} . In this case, molecule with larger pIC_{50} having higher antiviral activity.

Descriptor Generation

All structures were generated with the HyperChem (Hypercube Inc. version 7) program and optimized by the AM1 semi-empirical method of the software. Since the calculated values of the electronic features of the molecules will be influenced by the conformation used, in the current research we made attempt to use the most stable conformations. To avoid the local stable conformations of the compounds, geometry optimization was run many times with different starting points for each molecule, and conformation with the lowest energy was considered for calculation of the electronic properties. Some electronic descriptors such as

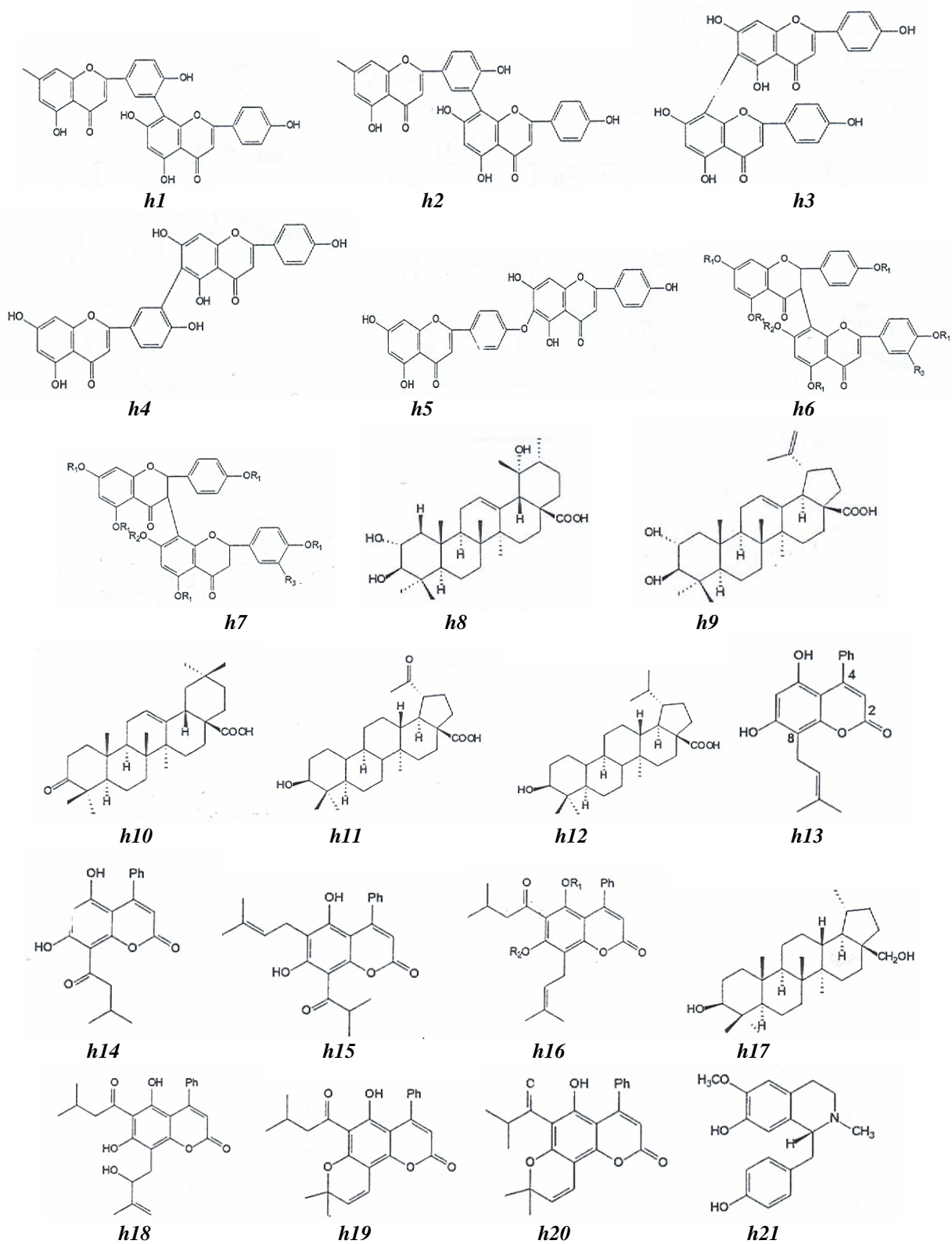
dipole moments and orbital energies of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) were calculated by the HyperChem software. Constitutional descriptors and topological indices were calculated utilizing Dragon software created by the Milano QSAR and Chemometrics Research Group (www.disat.unimib.it/chm/). These descriptors are calculated using two-dimensional representation of the molecules and therefore geometry optimization is not essential for calculating these types of descriptors. In addition, Dragon calculates a large number of descriptors from the optimized three-dimensional structure of the molecules.

Data Preprocessing

In the case of each antiviral activity, the calculated descriptors were collected in a data matrix (**D**) with dimension of $(n \times m)$, where n and m are being the number of molecules in each data set and the number of calculated descriptors for each molecule, respectively. Firstly, the descriptors were checked for constant or near constant values and those detected were removed from the original data matrix. Then, the correlation of descriptors with each others and with the activity data was determined. Among the collinear descriptors detected ($r > 0.9$), one of them that had the highest correlation with activity was retained and the rest were omitted.

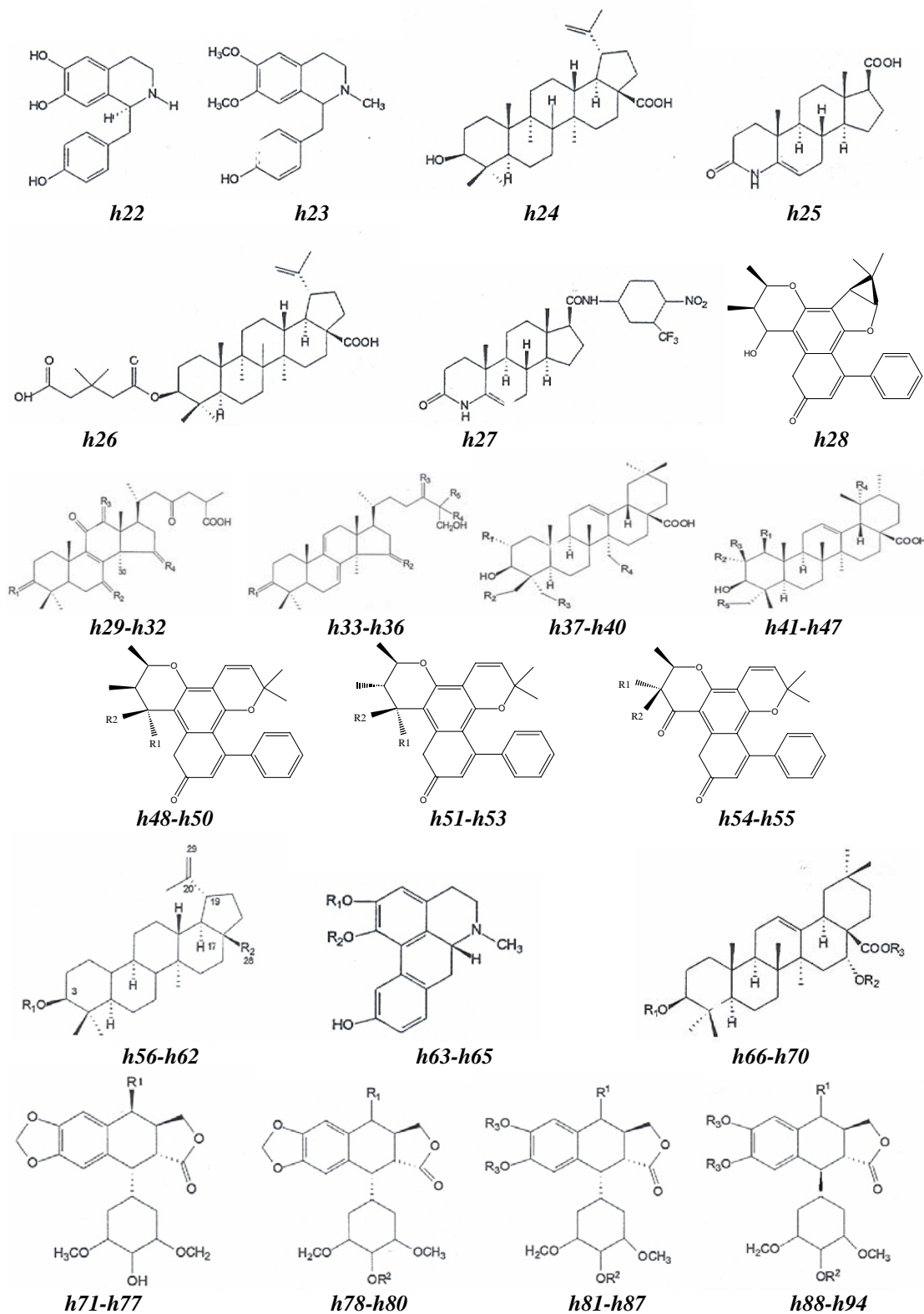
Model Development and Validation

QSAR models were obtained by multiple linear regression (MLR) analysis. The stepwise selection of variables, a combination of forward selection and backward elimination procedure, was used to select the most relevant subset of descriptors. Regression analyses were performed by SPSS software (SPSS Inc., Version 11.5). In the case of each regression problem, SPSS produces many models and ranked them based on standard error of calibration (Se) and coefficient of multiple determinations (R^2), where some models have large number of input variables and thus they are over-fitted. To hinder obtaining over-fitted models, the generated multilinear QSAR models by SPSS were validated by cross-validation for prediction ability and generalization. A balance between the high cross-validation correlation coefficient (R^2_{CV}) and low number of descriptors were used as the criterion for model selection.

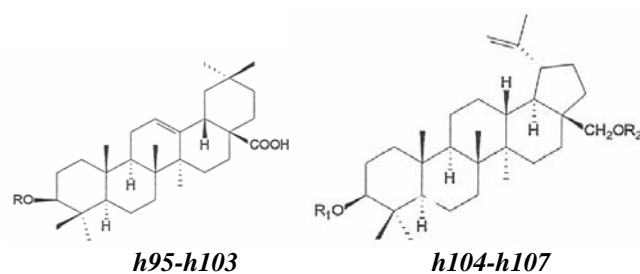


Scheme 1

QSAR Studies on the Antiviral Compounds of Natural Origin



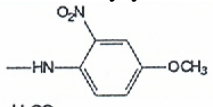
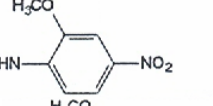
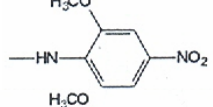
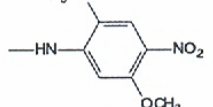
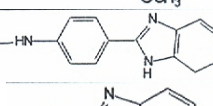
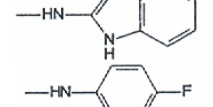
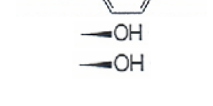
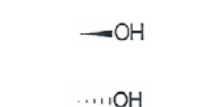
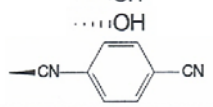
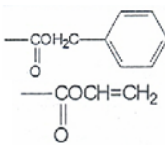
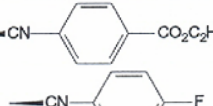
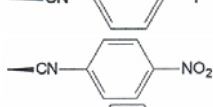
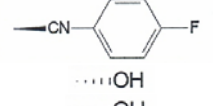
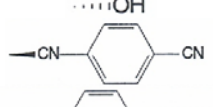
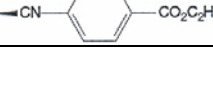


Scheme 1



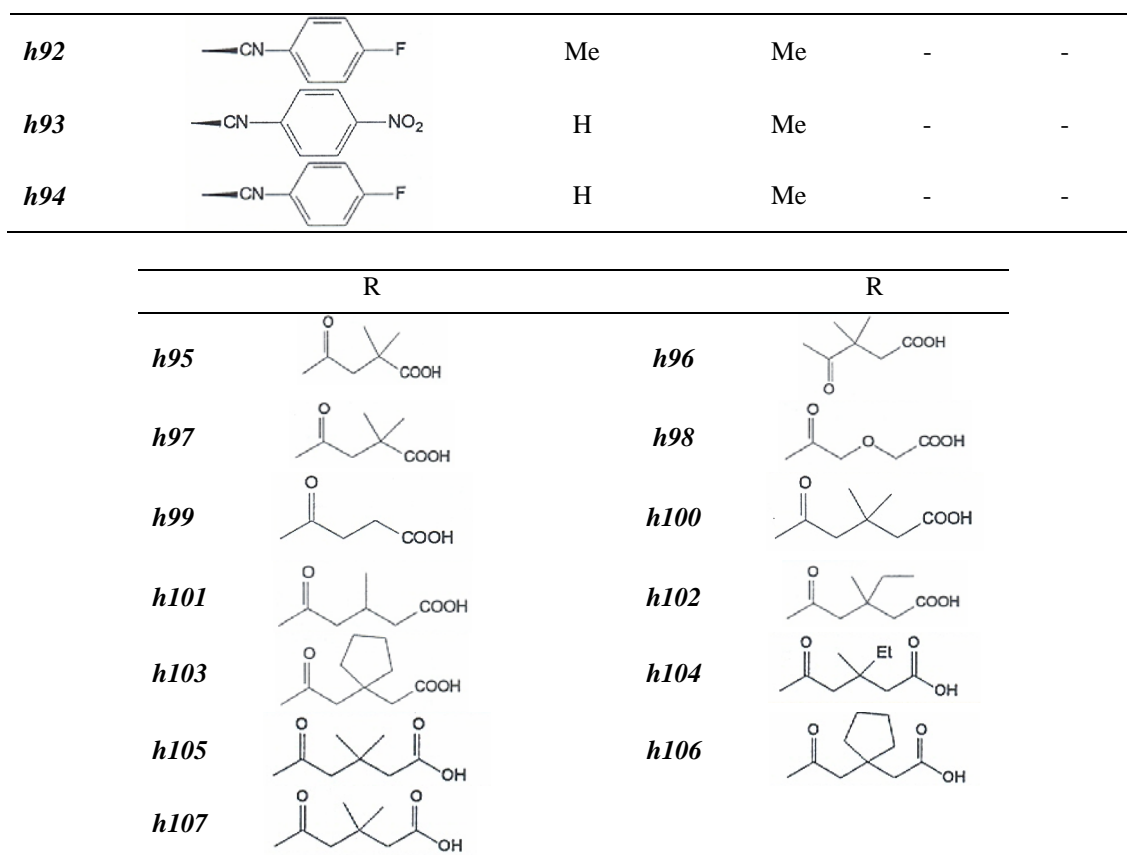
| | R1 | R2 | R3 | R4 | R5 |
|-----------------|--|---------------------------------------|-----------------|-----------------|--------------------|
| <i>h29</i> | | O | | | - |
| <i>h30</i> | | | | O | - |
| <i>h31</i> | O | O | | O | - |
| <i>h32</i> | | | O | O | - |
| <i>h33</i> | | H ₂ | | OH | CH ₃ |
| <i>h34</i> | O | | CH ₃ | CH ₃ | CH ₂ OH |
| <i>h35</i> | O | H ₂ | CH ₃ | CH ₃ | CH ₂ OH |
| <i>h36</i> | O | H ₂ | | OH | CH ₃ |
| <i>h137</i> | H | H | H | H | - |
| <i>h38</i> | OH | H | OH | H | - |
| <i>h39</i> | OH | OH | H | H | - |
| <i>h40</i> | H | H | H | OH | - |
| <i>h41</i> | H | H | H | OH | H |
| <i>h42</i> | H | H | H | H | H |
| <i>h43</i> | H | OH | H | H | H |
| <i>h44</i> | H | OH | H | H | OH |
| <i>h45</i> | H | OH | OH | OH | H |
| <i>h46</i> | H | OH | OH | OH | H |
| <i>h47</i> | OH | OH | OH | OH | H |
| <i>h48, h51</i> | H | OH | - | - | - |
| <i>h49, h52</i> | OH | H | - | - | - |
| <i>h50, h53</i> | H | OAc | - | - | - |
| <i>h54</i> | Me | H | - | - | - |
| <i>h455</i> | H | Me | - | - | - |
| <i>h56</i> | H | COOH | - | - | - |
| <i>h57</i> | H | CH ₂ OH | - | - | - |
| <i>h58</i> | H | COOCH ₂ COOCH ₃ | - | - | - |
| <i>h59</i> | Ac | COOH | - | - | - |
| <i>h60</i> | COC ₆ H ₅ | COOH | - | - | - |
| <i>h61</i> | COCH=CHCH ₃ | COOH | - | - | - |
| <i>h62</i> | COCH ₂ CH ₂ COOH | COOH | - | - | - |
| <i>h63</i> | Me | OH | - | - | - |
| <i>h64</i> | Me | Me | - | - | - |
| <i>h65</i> | H | Me | - | - | - |

Scheme 1

QSAR Studies on the Antiviral Compounds of Natural Origin

| | | | | | |
|-------------|---|---|----|---|---|
| h66 | H | H | H | - | - |
| h67 | H | H | Me | - | - |
| h68 | Ac | H | Me | - | - |
| h69 | Ac | Ac | H | - | - |
| h70 | Butyryl | H | Me | - | - |
| h71 |  | - | - | - | - |
| h72 |  | - | - | - | - |
| h73 |  | - | - | - | - |
| h74 |  | - | - | - | - |
| h75 |  | - | - | - | - |
| h876 |  | - | - | - | - |
| h77 |  | - | - | - | - |
| h78 |  | H | - | - | - |
| h79 |  |  | - | - | - |
| h80 |  | - | - | - | - |
| h81 |  | Me | H | - | - |
| h82 |  | Me | Me | - | - |
| h83 |  | H | Me | - | - |
| h84 |  | H | Me | - | - |
| h85 |  | Me | Me | - | - |
| h86 |  | H | Me | - | - |
| h87 | | H | Me | - | - |
| h88 | | Me | H | - | - |
| h89 | | Me | Me | - | - |
| h90 | | H | Me | - | - |
| h91 | | H | Me | - | - |

Scheme 1



Scheme 1

The overall prediction abilities of the final models were accessed by using prediction set containing about 25% of the original molecules. To do so, the data sets of each antiviral activity were classified to calibration and prediction sets, randomly. The model coefficients were calculated using calibration data and then used to calculate the antiviral activity of the molecules in the prediction set. The data splitting was run seven times and the root mean square errors of predictions were averaged.

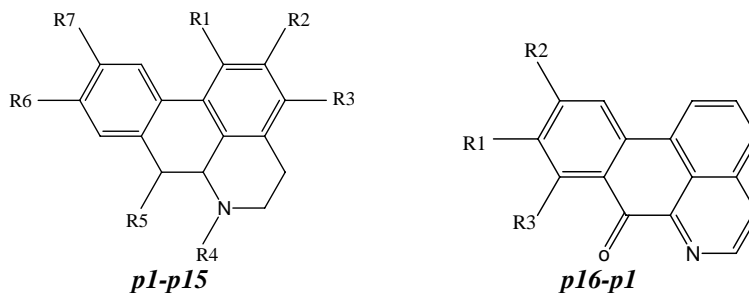
RESULTS AND DISCUSSION

As it is shown in Schemes 1 and 2, a data set size of 107 molecules was used for QSAR analysis of the anti-HIV1 activity and a size of 18 molecules were used for anti-Polio activity. The anti-HIV1 agents used in this study have a very large diverse molecular structure belonging to different

molecular family. As it is seen from Table 1, the anti-HIV1 activity of the studied compounds is varied between 3.495 (**h35**) and 7.569 (**h79**) in pIC_{50} unit. On the other hand, in the case of anti-Polio compounds, not only the number of molecules is very smaller but also the molecules share similar structural backbone. The pIC_{50} for this type of molecules (Table 2) is varied between 3.301 (**p17**) and 5.046 (**p13**).

To investigate the effects of molecular structure on the antiviral activity of the studied natural compounds, a large number of molecular descriptors belonging to wide variety of structural features were considered. At the first, separate QSAR models were obtained using the pools of different type of molecular descriptors. This helped us to identify the molecular descriptors of each group that represented higher impact on the antiviral activity of interest. Then, the selected descriptors of different types were used to develop a final QSAR model for each data set. The results are discussed in

QSAR Studies on the Antiviral Compounds of Natural Origin



| | R1 | R2 | R3 | R4 | R5 | R6 | R7 |
|------------|----------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| <i>p1</i> | OCH ₃ | OCH ₃ | H | CH ₃ | H | OCH ₃ | OCH ₃ |
| <i>p2</i> | OCH ₃ | OCH ₃ | H | COH ₃ | H | OCH ₃ | OCH ₃ |
| <i>p3</i> | OCH ₃ | OCH ₃ | H | H | H | OH | OCH ₃ |
| <i>p4</i> | OCH ₃ | OCH ₃ | H | CH ₃ | H | OH | OCH ₃ |
| <i>p5</i> | OCH ₃ | OCH ₃ | H | H | H | OH | OCH ₃ |
| <i>p6</i> | OCH ₃ | OCH ₃ | H | CH ₃ | H | OH | OCH ₃ |
| <i>p7</i> | OH | OCH ₃ | H | CH ₃ | H | OH | OCH ₃ |
| <i>p8</i> | -OCH ₂ O- | | H | H | H | OH | OCH ₃ |
| <i>p9</i> | -OCH ₂ O- | | H | CH ₃ | H | OH | OCH ₃ |
| <i>p10</i> | OCH ₃ | OCH ₃ | H | CH ₃ | H | H | H |
| <i>p11</i> | -OCH ₂ O- | | OCH ₃ | CH ₃ | OH | H | H |
| <i>p12</i> | -OCH ₂ O- | | H | CH ₃ | OH | OCH ₃ | H |
| <i>p13</i> | -OCH ₂ O- | | H | CH ₃ | OCH ₃ | OCH ₃ | H |
| <i>p14</i> | -OCH ₂ O- | | H | CH ₃ | OCH ₃ | H | H |
| <i>p15</i> | -OCH ₂ O- | | H | CH ₃ | OCH ₃ | H | H |
| <i>p16</i> | -OCH ₂ O- | | H | - | - | - | - |
| <i>p17</i> | OCH ₃ | OCH ₃ | H | - | - | - | - |
| <i>p18</i> | -OCH ₂ O- | | OCH ₃ | - | - | - | - |

Scheme 2

below.

QSAR Models for Anti-HIV1 Agents

The resulted QSAR models derived from the pools of different types of molecular descriptors is represented in Table 1 for the natural anti-HIV agents. It is clearly observed that no accurate model has been obtained from none of the descriptor types. This can be attributed to the fact that the anti-HIV activity of the diverse set of natural compounds can not be related to the single structural feature of the molecules and therefore, a combination of different structural features should be considered in order to find desirable QSAR model. However, significant QSAR models have been obtained from the pool of some descriptor types such as 3D MoRSE, Atom-centered fragments, BCUT and Functional group descriptors, among which the Atom-centered fragments descriptors

represented the most significant QSAR model. These descriptors are comprehensively described in [31]. The cross-validated correlation coefficients (R^2_{CV}) of the QSAR models derived from these types of descriptors are higher than 0.50, which means that these models could explain more than 50% of the variances in the anti-HIV activity of the studied natural compounds.

To obtain QSAR model containing different structural features of the studied molecules, the selected descriptors appeared in Table 3 were collected and QSAR models were generated from the pool of these descriptors. Among the different QSAR models proposed by SPSS software, that model represented the highest cross-validated correlation coefficient with lower number of input descriptors is represented in Table 4. In this table, the selected descriptors along with their regression coefficient and the corresponding

Table 1. Ant-HIV1 Activity of the Selected Natural Compounds

| <i>ID</i> | <i>pIC₅₀</i> | <i>ID</i> | <i>pIC₅₀</i> | <i>ID</i> | <i>pIC₅₀</i> | <i>ID</i> | <i>pIC₅₀</i> |
|-------------------|-------------------------|-------------------|-------------------------|-------------------|-------------------------|--------------------|-------------------------|
| <i>h1</i> | 4.924 | <i>h28</i> | 4.046 | <i>h55</i> | 4.886 | <i>h82</i> | 5.564 |
| <i>h2</i> | 4.000 | <i>h29</i> | 3.721 | <i>h56</i> | 4.089 | <i>h83</i> | 6.533 |
| <i>h3</i> | 4.187 | <i>h30</i> | 3.770 | <i>h57</i> | 4.347 | <i>h84</i> | 6.488 |
| <i>h4</i> | 4.208 | <i>h31</i> | 3.745 | <i>h58</i> | 4.585 | <i>h85</i> | 4.885 |
| <i>h5</i> | 3.936 | <i>h32</i> | 3.699 | <i>h59</i> | 4.699 | <i>h86</i> | 5.347 |
| <i>h6</i> | 3.627 | <i>h33</i> | 3.638 | <i>h60</i> | 4.824 | <i>h87</i> | 4.360 |
| <i>h7</i> | 3.770 | <i>h34</i> | 3.770 | <i>h61</i> | 4.319 | <i>h88</i> | 4.849 |
| <i>h8</i> | 4.115 | <i>h35</i> | 3.495 | <i>h62</i> | 4.796 | <i>h89</i> | 4.735 |
| <i>h9</i> | 5.071 | <i>h36</i> | 3.658 | <i>h63</i> | 4.131 | <i>h90</i> | 4.688 |
| <i>h10</i> | 4.523 | <i>h37</i> | 4.279 | <i>h64</i> | 5.310 | <i>h91</i> | 4.975 |
| <i>h11</i> | 4.951 | <i>h38</i> | 4.747 | <i>h65</i> | 5.119 | <i>h92</i> | 4.728 |
| <i>h12</i> | 5.161 | <i>h39</i> | 4.094 | <i>h66</i> | 4.886 | <i>h93</i> | 4.666 |
| <i>h13</i> | 4.496 | <i>h40</i> | 4.237 | <i>h67</i> | 4.569 | <i>h94</i> | 4.550 |
| <i>h14</i> | 4.654 | <i>h41</i> | 4.301 | <i>h68</i> | 6.370 | <i>h95</i> | 4.719 |
| <i>h15</i> | 3.907 | <i>h42</i> | 4.836 | <i>h69</i> | 6.425 | <i>h96</i> | 4.780 |
| <i>h16</i> | 4.493 | <i>h43</i> | 4.759 | <i>h70</i> | 6.539 | <i>h97</i> | 4.625 |
| <i>h17</i> | 4.548 | <i>h44</i> | 4.724 | <i>h71</i> | 4.625 | <i>h98</i> | 4.415 |
| <i>h18</i> | 5.315 | <i>h45</i> | 4.500 | <i>h72</i> | 4.757 | <i>h99</i> | 4.936 |
| <i>h19</i> | 4.133 | <i>h46</i> | 4.064 | <i>h73</i> | 6.569 | <i>h100</i> | 4.593 |
| <i>h20</i> | 5.248 | <i>h47</i> | 4.076 | <i>h74</i> | 6.678 | <i>h101</i> | 4.328 |
| <i>h21</i> | 4.578 | <i>h48</i> | 4.959 | <i>h75</i> | 5.514 | <i>h102</i> | 5.144 |
| <i>h22</i> | 4.031 | <i>h49</i> | 4.337 | <i>h76</i> | 6.380 | <i>h103</i> | 5.124 |
| <i>h23</i> | 3.927 | <i>h50</i> | 7.420 | <i>h77</i> | 6.382 | <i>h104</i> | 4.405 |
| <i>h24</i> | 4.535 | <i>h51</i> | 6.143 | <i>h78</i> | 5.520 | <i>h105</i> | 4.515 |
| <i>h25</i> | 5.319 | <i>h52</i> | 5.000 | <i>h79</i> | 7.569 | <i>h106</i> | 4.172 |
| <i>h26</i> | 4.496 | <i>h53</i> | 4.000 | <i>h80</i> | 6.399 | <i>h107</i> | 3.620 |
| <i>h27</i> | 4.851 | <i>h54</i> | 4.155 | <i>h81</i> | 5.695 | | |

Table 2. Anti-Polio Activity of the Selected Natural Compounds

| <i>ID</i> | <i>pIC₅₀</i> | <i>ID</i> | <i>pIC₅₀</i> | <i>ID</i> | <i>pIC₅₀</i> | <i>ID</i> | <i>pIC₅₀</i> |
|------------------|-------------------------|-------------------|-------------------------|-------------------|-------------------------|-------------------|-------------------------|
| <i>p1</i> | 3.848 | <i>p6</i> | 3.602 | <i>p11</i> | 3.971 | <i>p15</i> | 4.165 |
| <i>p2</i> | 3.466 | <i>p7</i> | 3.664 | <i>p12</i> | 5.155 | <i>p16</i> | 4.319 |
| <i>p3</i> | 3.783 | <i>p8</i> | 3.971 | <i>p13</i> | 5.046 | <i>p17</i> | 3.301 |
| <i>p4</i> | 3.602 | <i>p9</i> | 4.328 | <i>p14</i> | 4.602 | <i>p18</i> | 4.553 |
| <i>p5</i> | 4.022 | <i>p10</i> | 3.362 | | | | |

QSAR Studies on the Antiviral Compounds of Natural Origin

Table 3. The Resulted QSAR Models for Anti-HIV Agents Derived from the Pools of Different Types of Molecular Descriptors

| <i>Descriptor groups</i> | <i>descriptors</i> | R^2 | <i>Se</i> | R^2_{CV} |
|--------------------------|--|-------|-----------|------------|
| 2D Autocorrelation | MATS8e, MATS1p, GATS6m | 0.512 | 0.600 | 0.483 |
| 3D MoRSE | Mor06u, Mor11v, Mor07e, Mor11e, Mor29e, Mor11p | 0.644 | 0.520 | 0.593 |
| Atom-centered fragments | C-006, C-011, C-012, C-014, C-024, C-026, O-029, N-070 | 0.713 | 0.471 | 0.647 |
| BCUT | BEHv2, BELv8, BELe8, BEHp6, BELp6 | 0.553 | 0.579 | 0.511 |
| Charge | qpos, Qpos, Qmean, TE2 | 0.460 | 0.634 | 0.332 |
| Constitutional | Me, Ms, nDB, nN, nR12 | 0.532 | 0.593 | 0.384 |
| Functional | nCt, nCaH, nNHRPh, nOHPh, nROR, nPhX | 0.654 | 0.512 | 0.612 |
| Galvez | JGI3, JGI4, JGI6, JGI7 | 0.365 | 0.687 | 0.303 |
| Geometrical | L/Bw, G(N..O) | 0.401 | 0.661 | 0.336 |
| Getway | H2e, H3e | 0.520 | 0.592 | 0.449 |
| Mol walk | SRW05, SRW07, SRW09 | 0.429 | 0.649 | 0.400 |
| RDF | RDF030m, RDF035v, RDF090v, RDF010e, RDF020e, RDF090p | 0.577 | 0.5664 | 0.553 |
| Topological | J, Jhetp, HVcpx, IC2, BIC5, D/Dr12, T(N..O) | 0.590 | 0.5604 | 0.521 |
| WHIM | P1m, Km | 0.258 | 0.7357 | 0.188 |

Table 4. The Resultant QSAR Model for Anti-HIV1 Agents Derived from the Pool of Molecular Descriptors Appeared in Table 3

| <i>Variable</i> | <i>Coefficient</i> | <i>Standard error of coefficient</i> | <i>Definition</i> |
|-----------------|--------------------|--------------------------------------|---|
| Intercept | 4983 | 126 | - |
| O-059 | 0.078 | 0.019 | Number of C _{aliphatic} -O-C _{aliphatic} fragment in the molecule |
| C-012 | -0.295 | 0.068 | Number of CR ₂ X ₂ fragment in molecule |
| logP | 0.004 | 0.001 | Logarithm of octanol-water partition coefficient or lipophilicity index |
| T (N..O) | 0.009 | 0.003 | Topological distances between O and N atoms |
| C-014 | 1.072 | 0.311 | Number of CX ₄ fragment in molecule |
| D/Dr12 | -0.014 | 0.003 | Distance/detour ring index of order 12, it is a topological index |
| C-011 | -0.421 | 0.019 | Number of CR ₃ X fragment in molecule |
| MATS1p | -3.084 | 0.422 | Mora autocorrelation-lag 1 weighted by atomic polarizability, it is a 2D autocorrelation descriptor |

Table 5. Statistical Parameters for the QSAR Models of Anti-HIV1 (Table 4) and Anti-polio (Table 7) Agents

| | N^a | K^b | N_D^c | F^d | F_{max}^e | R^2 | Se | R_{cv}^2 |
|------------|-------|-------|---------|-------|-------------|-------|-------|------------|
| Anti-HIV | 107 | 10 | 63 | 98.3 | 78.3 | 0.868 | 0.389 | 0.753 |
| Anti-Polio | 18 | 3 | 38 | 41.4 | 18.9 | 0.984 | 0.142 | 0.895 |

^aN is the number of molecules. ^bK is number of selected descriptors. ^cN_D is number of original descriptors, from which the most convenient QSAR model was obtained. ^dF is conventional Fisher variance ratio of model ^eF_{max} is the Livingstone-Salt variance ratio.

standard error of coefficients are given. Obviously, for all the variables appeared in Table 4, the standard errors of coefficients are much lower than the coefficient itself, which indicates the statistical significance of the selected descriptors in the resulted QSAR model for the anti-HIV activity of the studied compounds.

The statistical parameters of this 8-parametric equation are represented in Table 5. In this Table, N, K and N_D are the number of molecules used in model development, number of selected descriptors and the number of original descriptors used to obtain QSAR equation, respectively. The calibration statistics calculated are coefficient of multiple determination (R^2), standard error of calibration (Se) and the Fisher variance ratio (F). The R^2 value is 0.868, which means that the resultant model can explain about 86% of variances in the anti-HIV activity data. In addition, the F-ratio statistics is larger than the critical value with the significance level lower than 0.001. On the other hand, Livingstone and Salt stated that since MLR models suffer from selection bias, the significance of these models can not be judged by conventional statistics and proposed an F_{max} function considering the number of original variables used to derive the MLR models too [40]. This value was calculated for the system under study using the online version of the F_{max} calculator (<http://www.cmd.port.ac.uk/cmd/fmax.shtml>). As it is shown in Table 5, the calculated F-value of the model is also larger than the critical F_{max} value. This is another indication that the resulted model developed for the anti-HIV activity of the selected natural compounds is statistically valid and does not suffer from selection bias.

Moreover, the high value of the cross-validated correlation coefficients ($R_{cv}^2 = 0.753$) indicate the predictivity of the proposed QSAR model. The plot of cross-validated predicted values of anti-HIV1 activity against the experimental values

are represented in Fig. 1 A, which shows the scattering of data around a straight line with slope and intercept close to one and zero, respectively.

The overall prediction ability of the resulted model was established by using prediction set samples. To do so, 25 molecules, out of 107 molecules, were randomly selected as prediction set and the rest were chosen as calibration samples. Model coefficients were calculated using calibration samples and then they used to predict the activity of the prediction samples. To investigate the effect data splitting on the model performances, data splitting into calibration and prediction sets was repeated seven times. The resulted correlation coefficients of the prediction sets are shown in Fig. 2A. Obviously, the resulted correlation coefficients are higher than 0.80, which shows the ability of the resulted QSAR model to predict 80% of the anti-HIV1 activity data. The average root mean square error of prediction for seven analyses was 0.421 which is a tiny error with respect to the average of the anti-HIV1 activity data (*i.e.*, 4.80).

The majority of the variables appeared in Table 4 are atom-centered fragment descriptors (*i.e.*, O-059, C-012, C-014, C-011). The rest are two topological indices (*i.e.*, T(N..O) and D/Dr1), the lipophilicity index (logP) and a 2D autocorrelation descriptor (MATS1p). Interestingly, among the QSAR models obtained by using separate groups of descriptors, represented the most significant QSAR model was obtained from the atom-centered fragment descriptors. This explains the significance of the atom-centered fragment descriptor in modeling the anti-HIV activity of the studied natural products.

QSAR Models for Anti-polio Agents

The number of molecules used in the anti-polio data set is

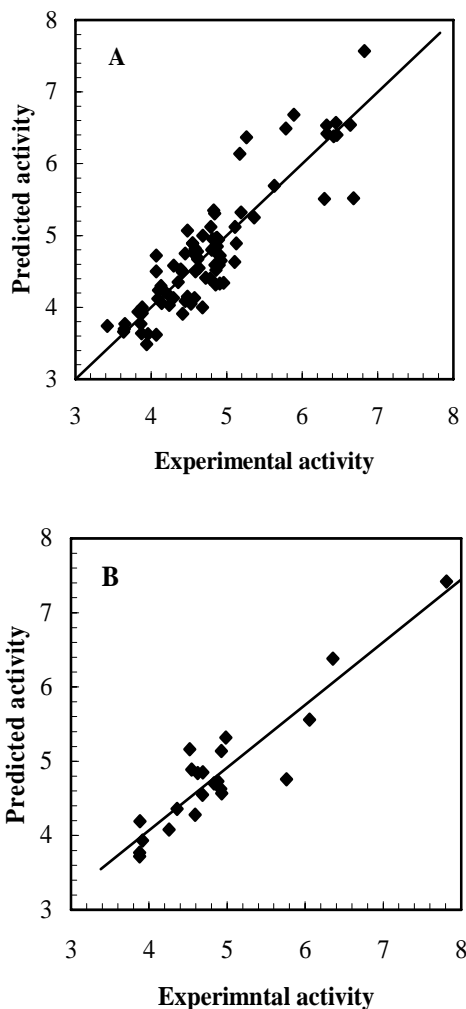


Fig. 1. Plot of predicted activity against the experimental activity of the anti-HIV1 data set: (A) cross-validation and (B) prediction set.

very smaller than the data set size of the anti-HIV1 agents. Therefore, the QSAR models that will be discussed can not be generalized in the same manner as those found for anti-HIV1 data set. However, this QSAR analysis gives introductory information about the structure-activity relationships of the natural anti-polio agents. The resulted QSAR models derived from the pools of different types of descriptors are listed in Table 6. First of all, the resulted models represent higher statistical quality with respect to the models found for anti-HIV1 agents (R^2 and R^2_{CV} are much higher). This is not unexpected for the smaller number of molecules in the anti-

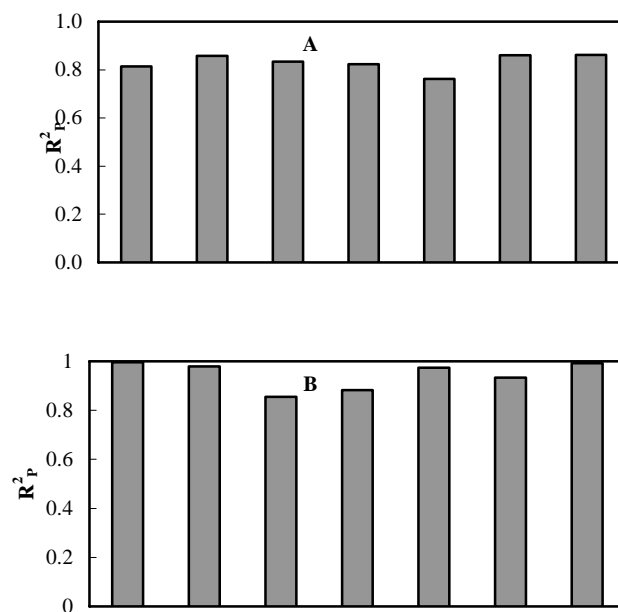


Fig. 2. Correlation coefficient of prediction for seven replicate data splitting: (A) anti-HIV1 (B) anti-polio.

polio data set. As it is observed from the data shown in Table 2, no significant QSAR models are obtained from some types of descriptors such as aromaticity indices, chemical descriptors and Galvez charge topological indices. On the other hands, the most significant QSAR models are provided from the pool of radial distribution function (RDF), BCUT and 3D-MoRSE descriptors. In comparison with the QSAR models found anti-HIV1 compounds, BCUT and 3D-MoRSE descriptors represented significant impact on both types of antiviral activities. However, the atom-centered fragments that represented the most significant QSAR model for anti-HIV1 activity is replaced by RDF descriptor in anti-polio data set.

The descriptors appeared in Table 6 were collected in descriptor data matrix and used as source of molecular structural parameters to obtain a QSAR model containing various structural features of the studied molecules. The most convenient QSAR model obtained by MLR analysis based on stepwise variable selection and cross-validation test is reported in Table 7. In this table, the selected descriptors along with their regression coefficient and the corresponding standard error of coefficients are given. As it is observed, the standard errors of coefficients are much lower than the coefficient

Table 6. The Resulted QSAR Models for Anti-polio Agents Derived from the Pools of Different Types of Molecular Descriptors

| Descriptor groups | Descriptors | R ² | Se | R ² _{CV} |
|-------------------|---|----------------|--------|------------------------------|
| 2D | MATS2e, MATS3e, GATS3e | 0.823 | 0.2507 | 0.744 |
| 3D | Mor28m, Mor03p, Mor31p | 0.894 | 0.1942 | 0.786 |
| Aromaticity | HOMT | 0.366 | 0.4445 | 0.322 |
| Atom-centered | C-007 | 0.621 | 0.3437 | 0.559 |
| BCUT | BEHm8, BEHp8, BELp4 | 0.838 | 0.2398 | 0.791 |
| Chem | Surface area | 0.274 | 0.4757 | 0.271 |
| Constitutional | nCIR | 0.621 | 0.3437 | 0.603 |
| Functional | nROR | 0.634 | 0.3377 | 0.597 |
| Galvez | JGI2 | 0.282 | 0.4729 | 0.224 |
| Geometrical | DELS, G2, SEig | 0.731 | 0.3093 | 0.700 |
| Getway | REIG, R2m ⁺ , R3m ⁺ | 0.806 | 0.2631 | 0.767 |
| Mol walk | SRW05 | 0.621 | 0.3437 | 0.610 |
| RDF | RDF035m, RDF040m, RDF050e, RDF070p, RDF110p | 0.938 | 0.1603 | 0.807 |
| Topological | D/Dr05 | 0.655 | 0.3277 | 0.611 |
| WHIM | G1u, E2m, Av, Vu | 0.807 | 0.2721 | 0.779 |

Table 7. The Resultant QSAR Model for Anti-polio Agents Derived from the Pool of Molecular Descriptors Appeared in Table 6

| Variable | Coefficient | Standard error of coefficient | Definition |
|-----------|-------------|-------------------------------|--|
| Intercept | 1.860 | 0.214 | - |
| Mor28m | -3.118 | 0.366 | 3D-MoRSE-signal 28/weighted by atomic masses |
| RDF035m | -1.072 | 0.118 | Radial distribution function 3.5/weighted by atomic masses |
| Mor03p | 0.844 | 0.191 | 3D-MoRSE-signal 3/weighted by atomic polarizabilities |

itself, which indicates the statistical significance of the selected descriptors. A comparison between the selected descriptors in Table 6 and 7 reveals that only two types of molecular descriptors, which represented the most significant QSAR models (*i.e.*, 3D-MoRSE and RDF descriptors), are chosen for the final QSAR model of the anti-polio agents.

The statistical parameters of the resultant three-parametric QSAR model of the anti-polio agents are given in Table 5. The statistical significances of this model are established not only by the high correlation coefficient and low standard error of calibration but also by conventional F and F_{max} criterion

(Livingstone and Salt, 2005). In addition, the resultant model represents very high correlation coefficient for cross-validation, which confirms its stability and accuracy. The overall prediction ability of the resulted QSAR model was examined in the same manner as performed for anti-HIV1 data set. Among the 18 studied natural anti-polio molecules, 5 compounds were randomly selected as prediction set, and the rest were employed to calculate model coefficients. The resulted correlation coefficients of prediction for seven repeated data splitting are shown in Fig. 2B. Obviously, the correlation coefficients of prediction are higher than 0.87. The

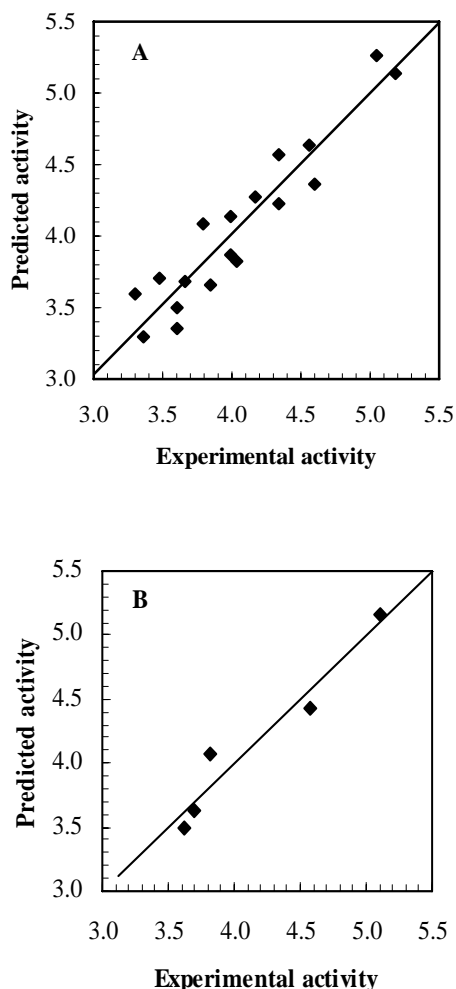


Fig. 3. Plot of predicted activity against the experimental activity of the anti-polio data set: (A) cross-validation and (B) prediction set.

average root mean square error of prediction for seven randomly selected prediction sets is 0.174 in pIC_{50} unit, which is a very small error in comparison with the average of pIC_{50} for selected anti-polio compounds (*i.e.*, 4.04). The plot of cross-validated activity of the anti-polio agents, obtained from the QSAR model of Table 7, against the experimental values are shown in Fig. 3A. Also, the plot of predicted activity of an external prediction set with medium quality against the corresponding experimental activity is given in Fig. 3B. Both plots represent a uniform distribution around a straight line with respective slope and intercept very close to one and zero. This is again another confirmation of model accuracy.

ACKNOWLEDGEMENTS

Financial support of this project by the Research Council of Shiraz University of Medical Sciences is acknowledged.

REFERENCES

- [1] S.P. Whelan, J.N. Barr, G.W. Wertz, *Curr. Top. Microbiol. Immunol.* 283 (2004) 61.
- [2] T. Das, M. Mathur, A.K. Gupta, G.M. Janssen, A.K. Banerjee, *Proc. Natl. Acad. Sci. USA* 95 (1998) 1449.
- [3] X. Shen, P.S. Masters, *Proc. Natl. Acad. Sci. USA* 98 (2001) 2717.
- [4] J.H. Strauss, E.G. Strauss, *Viral RNA replication. With a Little Help from the Host*, *Science* 283 (1999) 802.
- [5] J.H. Connor, M.O. McKenzie, G.D. Parks, D.S. Lyles, *Virology* 362 (2007) 109.
- [6] P.J. Hotez, D.H. Molyneux, A. Fenwick, E. Ottesen, S.E. Sachs, J.D. Sachs, *Plos Med.* 3 (2006) 576.
- [7] E. Blignaut, *Socio-Med. Asp. Aids/HIV* 19 (2007) 532.
- [8] C.T. Fang, P.C. Hsiung, C.F. Yu, M.Y. Chen, J.D. Wang, *Qual. Life Res.* 11 (2002) 753.
- [9] S. Yallop, A. Lowth, M.H. Fitzgerald, J. Reid, A. Morelli, *Cult. Health Sex.* 4 (2002) 431.
- [10] J.K. Andrus, K. Banerjee, B.P. Hull, J.C. Smith, I. Mochny, *J. Infec. Diseases* 175 (1997) S89.
- [11] S. Bonu, M. Rani, O. Razum, *Health Policy* 70 (2004) 327.
- [12] C. Kapp, *World Health Organization Tackles Polio in the Congo*, *Lancet* 353 (1999) 1949.
- [13] A. DerMarderosian, J.A. Beutler, *Review of Natural Products*, 3rd ed., *Facts & Comparisons*, a Wolters Kluwer Company, Missouri, 2006.
- [14] R.M.G. Perez, *Pharm. Biol.* 41 (2003) 107.
- [15] J.S. Driscoll, V.E. Marquez, *Stem Cells* 12 (1994) 7.
- [16] N. Castagnoli, L. Kier, *Chem. Biodivers.* 2 (2005) 409.
- [17] T. Kiss, P. Erdi, *Biosystems* 86 (2006) 46.
- [18] A.R. Ortiz, P. Gomez-Puertas, A. Leo-Macias, P. Lopez-Romero, E. Lopez-Vinas, A. Morreale, M. Murcia, K. Wang, *Curr. Top. Med. Chem.* 6 (2006) 41.
- [19] P.H. Reggio, *AAPS J.* 8 (2006) E322.
- [20] T. Fujita, H. Timmerman, *QSAR and Drug Design, New Developments and Applications*, 1st ed., Elsevier,

- Amsterdam, 1995.
- [21] C. Hansch, D. Hoekman, H. Gao, *Chem. Rev.* 96 (1996) 1045.
- [22] B. Hemmateenejad, *J. Chemometr.* 18 (2004) 475.
- [23] V. Lozitsky, V. Kuzmin, A. Artemenko, R. Lozytska, A. Fedchuk, Y. Boschenko, T. Gridina, L. Shitikova, L. Mudrik, J.J. Vanden Eynde, E. Muratov, D. Kryzhanovsky, *Antivir. Res.* 57 (2003) A83.
- [24] M. Seierstad, D.K. Agrafiotis, *Chem. Biol. Drug Des.* 67 (2006) 284.
- [25] C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir, *Nature* 194 (1962) 178.
- [26] B. Hemmateenejad, M. Sancholi, *J. Chemometr.* 21 (2007) 96.
- [27] L. Kier, L. Hall, *Molecular Structure Description*, Academic Press, New York, 1999.
- [28] L. Eriksson, E. Johansson, *Chemom. Intell. Lab. Syst.* 34 (1996) 1.
- [29] M.M.C. Ferreira, *J. Brazil. Chem. Soc.* 13 (2002) 742.
- [30] B. Hemmateenejad, R. Miri, M. Jafarpour, M. Tabarзад, A. Foroumadi, *QSAR Comb. Sci.* 25 (2006) 56.
- [31] R. Todeschini, V. Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH Verlag, Weinheim, 2000.
- [32] A.T. Balaban, A. Beteringhe, T. Constantinescu, P.A. Filip, O. Ivanciuc, *J. Chem. Inf. Model.* 47 (2007) 716.
- [33] B. Hemmateenejad, R. Miri, N. Edraki, M. Khoshneviszadeh, A. Shafiee, *J. Iran. Chem. Soc.* 4 (2007) 182.
- [34] M. Shamsipur, R. Ghavami, B. Hemmateenejad, H. Sharghi, *QSAR Comb. Sci.* 23 (2004) 734.
- [35] S. Vilar, E. Estrada, E. Uriarte, L. Santana, Y. Gutierrez, *J. Chem. Inf. Model.* 45 (2005) 502.
- [36] I.R.A. Menezes, J.C.D. Lopes, C.A. Montanari, G. Oliva, F. Pavao, M.S. Castilho, P.C. Vieira, M.T. Pupo, *J. Comput.-Aided Mol. Des.* 17 (2003) 277.
- [37] G. Ramirez-Galicia, R. Garduno-Juarez, B. Hemmateenejad, O. Deeb, S. Estrada-Soto, *Chem. Biol. Drug Des.* 70 (2007) 143.
- [38] F.L. Stahura, J.W. Godden, L. Xue, J. Bajorath, *J. Chem. Inf. Comput. Sci.* 40 (2000) 1245.
- [39] A.D. Wright, R. De Nys, C.K. Angerhofer, J.M. Pezzuto, M. Gurrath, *J. Nat. Prod.* 69 (2006) 1180.
- [40] D.J. Livingstone, D.W. Salt, *J. Med. Chem.* 48 (2005) 661.