Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 25 (2006) 37–45
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Evaluation of a neural networks QSAR method based on ligand representation using substituent descriptors Application to HIV-1 protease inhibitors

Adina-Luminiţa Milac [a], Speranţa Avram [b], Andrei-José Petrescu [a,*]

[a] *Institute of Biochemistry, Splaiul Independenţei 296, Sector 6, Bucharest, Romania*
[b] *Faculty of Biology, Department of Biophysics and Physiology, Splaiul Independenţei 91–95, Bucharest, Romania*

Received 1 April 2005; received in revised form 17 June 2005; accepted 29 September 2005
Available online 1 December 2005

Abstract

We present here a neural networks method designed to predict biological activity based on a local representation of the ligand. The compounds of the series are represented by a vector mapping for each of four substituent properties: volume, log *P*, dipole moment and a simple 'steric' parameter relating to its shape. This ligand representation was tested using neural networks on a set of 42 cyclic-urea derivatives, inhibiting HIV-1 protease. The leave-one-out cross-validation using all descriptors in the input gave a correlation factor between prediction and experiment of 0.76 for the overall set and 0.88 when three outliers were left out. To rank the significance of the four descriptors, we further tested all combinations of two and three parameters for each substituent, using two disjunctive testing sets of five inhibitors. In these sets, vectors with extreme descriptor values were used either in the training or the testing set (sets A and B, respectively). The method is a very good interpolator (set A, $95 \pm 2\%$ accuracy) but a less effective extrapolator (set B, $85 \pm 2\%$ accuracy). Generally, the combinations including the 'steric' parameter predict better than average, while those containing the volume are less effective. The best prediction, $98.8 \pm 1.2\%$, was obtained when log *P*, the dipole and the steric parameter were used on set A. At the opposite end, the lowest ranked descriptor set was obtained when replacing log *P* with the volume, giving $92.3 \pm 6.7\%$ accuracy over the set A.

Keywords: Neural networks; QSAR; Compound library; Molecular descriptor; Biological activity prediction; HIV-1 protease inhibitors

1. Introduction

Neural networks (NN) are able to create internal models for complex input–output relationships based on learning from examples and therefore are useful in prediction.

In protein science NN were successfully used to predict secondary structure [1–3] and transmembrane segments [4], the structural class [5–8] and family [9,10], motifs such as co- and post-translational modifications [11–13], antigenic segments [14], signal sequence [15] or intracellular localization [16,17].

The NN techniques are also suited for quantitative structure–activity relationship (QSAR) applications because here a set of compounds with known activities is available for training. In contrast to simple QSAR methods based on regression analysis, where one has to priorly assume an input–output relation (e.g. linear or quadratic function), NN do not require any prior model of how input and output are connected and have the unique ability to adapt to highly complex non-linear relations [18–20]. Consequently, the essential features of NN: non-linearity, adaptivity, independence of any statistical and modelling assumptions, fault tolerance, universality and real time operation make them particularly suitable for pharmacokinetic applications, especially where extremely complex and unfamiliar responses are studied [21].

Recent examples include prediction of biological targets for chemical compounds using probabilistic NN and atom type descriptors, with 90% accuracy [22], prediction of drug resistance of HIV-1 protease mutants based on the number of drug–protein contacts, using Kohonen NN [23], selection of focussed drug libraries using feed-forward NN and 3D BCUT descriptors [24], prediction of toxicity of chemicals to aquatic species [25]. The current state-of-the-art in this field has been recently reviewed [26].

* Corresponding author. Tel.: +40 21 223 90 69; fax: +40 21 223 90 68.
E-mail address: Andrei.Petrescu@biochim.ro (A.-J. Petrescu).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.09.014

The first step in designing a NN is data pre-processing, which mainly consists in encoding the input information into an object representation so that this could be processed by the NN. This is a crucial step as the NN performance critically depends on how information is presented to the NN. An ideal encoding scheme should extract maximal information from the input data and satisfy the basic coding assumption that similar items are represented by close vectors [27]. In QSAR-like NN methods, the compounds are usually encoded by molecular descriptors—physico-chemical parameters that may be either experimental (e.g. refractive index, octanol/water partition coefficient or spectral data) or theoretical (e.g. molecular volume, weight, charge, electronic, lipophilic and steric properties).

Such a variety of parameters could generate large descriptor sets that may result either in redundancy of information—when descriptors are correlated, or chance correlations—when the dataset contains more descriptors than compounds [28]. Choosing a set of descriptors which is small enough to avoid redundancy and chance correlation, but large enough to allow an accurate representation of the ligand is therefore very important.

In this work we aim at evaluating various representations of a ligand set, focussed on substituents properties, that may be used as input in a feed-forward NN for QSAR-like applications.

The ligand set model chosen to test the method consists of 42 cyclic-urea derivatives with known inhibition constant against HIV-1 protease (PR). This system is also of practical interest and lot of data are available in the literature.

Human immunodeficiency virus type 1 (HIV-1) proteins are translated as part of the larger polyprotein precursor whose proteolytic processing during virus assembly and maturation is performed by PR [29–31]. PR is therefore an essential enzyme for HIV-1 life-cycle and a very attractive target for new antiviral drugs. The enzyme is a homodimer of 99 amino acids per chain and belong to eukaryotic aspartic protease family. The dimer has one active site region, situated at the interface between the two monomers, with one catalytic triad (Asp-Thr-Gly) from each monomer. The β-sheet configurations, which include the triplet active site Asp 25/125-Thr 26/126-Gly 27/127 are present in the major part of the enzyme (amino acids 1–85/101–185), whereas the α-helix domain is represented by the amino acids 86–99 [32,33]. PR has a high mutagenesis rate, thus being able to develop strong resistance to inhibitors [34–37]. This represents a serious problem for the anti-HIV-1 therapy. Taking into account the possible changes of the inhibitors structures, fast and precise techniques predicting the biological activity for new inhibitors are needed. In the last years the computational techniques as: molecular energy calculation [38,39], molecular docking techniques [40,41], molecular dynamics simulations [42–46] or QSAR procedures [47–54] have been useful tools for the study of the PR mutants and their inhibitors. This NN method is therefore also complementary to previous studies on interaction of HIV-1 PR inhibitors with the target enzyme.

2. Methodology

2.1. Molecular modelling of HIV-1 PR inhibitors

The set of 42 HIV-1 PR inhibitors, symmetric (benzyl, isopropyl, 4-hydroxybenzyl) cyclic-urea derivatives, was compiled from literature [55]. The criteria used for selection were: (i) the level of inhibition constants $K_i < 0.11$ nM and (ii) the variety of substituents to cyclic urea, covering as many as possible classes, e.g. methoxybenzyl, aminobenzyl, isobuthyl and hydroxybenzyl. This resulted in a highly diverse set (Table 1) in which most of the compounds have high activity. HIV-1 PR inhibitors were modelled in InsightII, starting from the cyclic-urea derivative DMP323 [56–58] complexed with HIV-1 PR (PDB code 1qbs [59]). The common cyclic urea was kept unchanged and specific substituents were added in R_1 and R_2 positions (Fig. 1A). The minimum potential energy calculations for all inhibitors were performed in Insight/Discover running conjugate-gradient method, convergence = 0.01. Electric charges of the HIV-1 PR inhibitors were loaded from InsightII dictionary applying Potentials within Force Field module.

2.2. Inhibitor parameters calculation

Each molecule is described by a vector whose elements are parameters measuring physical factors that we considered important for protein–inhibitor interaction: size (volume V), hydrophobicity (water/octanol partition coefficient $\log P$), charge (dipole moment Δ) and shape (steric factor σ). We introduced also a steric factor to account for the orientation of structural units relative to a benzene cycle contained in R_2. This was defined as shown in Fig. 1B.

Except the steric factor which was computed only for R_2, all other parameters were computed both for R_1 and for R_2.

Molecular volume was computed with Tinker [60,61]. The hydrophobic coefficient ($\log P$) was calculated considering the Crippen incremental value [62] using Schrodinger software. This was also used to compute the dipole moment starting form R_1/R_2 partial charges.

2.3. Neural networks

The multi-layered feed-forward NN was trained with Levenberg–Marquardt algorithm [63,64]. Due to the non-linear input–output dependency the transfer function was chosen sigmoid. All units were fully interconnected and the input-to-output information flow was feed-forward (no feed-back connections). The number of neurons in the input layer was set equal to the number of dimensions of the input vectors, while the number of neurons in the output layer was set equal to 1, i.e. the number of parameters to be predicted in this case. Based on the finding that two hidden layers with non-linear neurons are required to approximate arbitrary functions [65], we used NN having two hidden layers, each with the number of neurons taking seven