Bioorganic & Medicinal Chemistry Letters 21 (2011) 6728–6731

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

One-class classification as a novel method of ligand-based virtual screening: The case of glycogen synthase kinase 3β inhibitors

Pavel V. Karpov [a], Dmitry I. Osolodkin [a], Igor I. Baskin [a,b], Vladimir A. Palyulin [a,c,*], Nikolay S. Zefirov [a,c]

[a] Department of Chemistry, Moscow State University, Leninskie Gory 1/3, Moscow 119991, Russia
[b] Laboratoire d'Infochimie, UMR 7177 CNRS, Université de Strasbourg 4, rue B. Pascal, Strasbourg 67000, France
[c] Institute of Physiologically Active Compounds, RAS, Severny proezd 1, Chernogolovka, Moscow Region 142432, Russia

ARTICLE INFO

Article history:
Received 22 August 2011
Revised 13 September 2011
Accepted 14 September 2011
Available online 21 September 2011

Keywords:
One-class classification
Virtual screening
Glycogen synthase kinase inhibitors
Neural networks
Auto-encoders

ABSTRACT

A virtual screening system based on one-class classification with molecular fingerprints as descriptors is developed and tested on a series of 1226 inhibitors and 209 noninhibitors of glycogen synthase kinase 3β (GSK-3β). The suggested system outperforms the ones based on pharmacophore hypothesis and molecular docking in a retrospective study. However, in a prospective study it should not be used as a sole classifier. The system is exceptionally useful for the identification of new scaffolds among the virtual screening results obtained with other methods.

© 2011 Elsevier Ltd. All rights reserved.

The main task of a virtual screening (VS) is to discriminate putative active compounds from inactive ones. The main requirement of major classification methods is to use both classes of active and inactive compounds during the ligand-based model construction, allowing one to find a hyperplane in a feature space that would separate active samples from inactive ones. The problem is that data for the inactive samples should be collected in the same conditions as for the active ones, but usually information on inactive compounds is not available, and researchers are obliged to create decoy datasets by themselves using their own rules (e.g., Refs. 1,2), that leads to several disadvantages. First, certain decoys may be really active. Second, the performance of a model is influenced by the quality of data and chemical diversity of the training dataset. Therefore, usage of standard classification procedures is methodologically incorrect when the decoy set is not rigorously defined.

The simplest and fastest VS method is the similarity search when a reference molecule with the desired biological activity is selected and all compounds from a database are ranked in ascending order of similarity to the reference molecule.[3] This procedure generally does not require building a model and using the negative samples, but recently it has been shown that the similarity search works only when the structure/activity surface is smooth enough and there are no 'activity cliffs'.[4] This condition is difficult to meet because such information is unattainable a priori, so the new methods of similarity search have to be proposed.

All these problems could be successfully resolved using the one-class classification (OCC) method,[5–7] the main idea of which is to construct the model based on the active compounds exclusively. In this Letter, we demonstrate the application of OCC to the VS of glycogen synthase kinase 3β (GSK-3β) inhibitors (for a recent review of this target see Ref. 8).

The reconstruction methods of the OCC approach include the auto-encoder neural networks, self-organizing maps (SOM) and principal component analysis (PCA). Auto-encoder (replicator, bottleneck or sand-glass) networks[9] are feed forward neural networks which have at least one hidden layer with the number of neurons many times smaller than in the other layers (Fig. 1). This layer

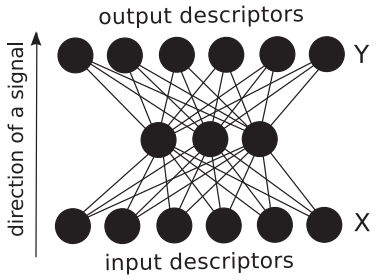

Figure 1. Typical scheme of an auto-encoder neural network.

* Corresponding author. Tel.: +7 495 939 39 69; fax: +7 495 939 02 90.
E-mail addresses: vap@org.chem.msu.su, vap@qsar.chem.msu.ru (V.A. Palyulin).

0960-894X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2011.09.051

acting as a compressing element attempts to reconstruct the input data to the output layer. The number of neurons in this layer can be roughly correlated with the number of features that are responsible for the biological activity.

Biological activity of an organic compound is defined by the constellation of its pharmacophoric features, encoded in the chemical structure, but it is very difficult to guess a priori which features are responsible for a certain activity. Moreover, one structure can possess several biological activities and have different features related to them. One way to predict the probability for a molecule to be biologically active is to compare features of the ligand with those for the target. If they are similar, the compound can be considered as potentially active. The auto-encoder network does actually the same: it tries to reconstruct the target features. In other words, auto-encoders work like the nonlinear PCA and sometimes similarly to pseudo-receptor models, implicitly reconstructing the binding cavity from the ligand features. PCA acts as a linear method, but auto-encoder networks perform the nonlinear transformation in data.

The second layer in Figure 1 is comprised of three neurons which filter only desired features. The reconstructed descriptors (**Y**) at the output layer should be the same as the input descriptors (**X**), thus the size of the output layer is also the same as the size of the input layer. The network can be divided into two parts: the first part encodes input information into the vector of features, whereas the second part acts as a decompressor which reconstructs this vector into the output one. The well trained network produces the output very similar to the input for those samples that possess the same characteristics that were in the training data, and for such samples the reconstruction error defined by the dissimilarity between the input and the output vectors would be minimal. If the reconstruction error is relatively high then the sample was

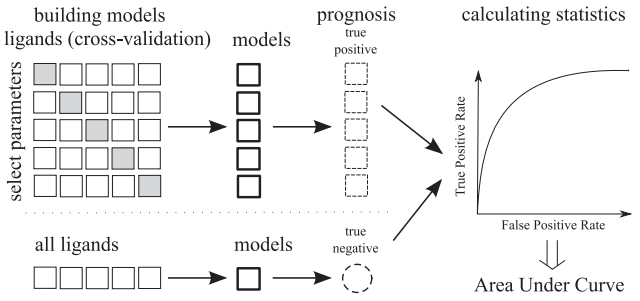

Figure 2. The workflow for choosing parameters of OCC models.

Table 1
Criteria of ZINC pre-filtering

Parameter	Minimum	Maximum
Molecular weight	150	500
Number of heavy atoms	10	30
Number of cycles	1	4
Number of hydrogen bond donors	1	4
Number of hydrogen bond acceptors	1	6
Number of rotatable bonds	0	10
Number of rigid bonds	0	25
Heteroatoms count	2	12
Calculated log*P*	−5	4
Formal charge	−2	+2

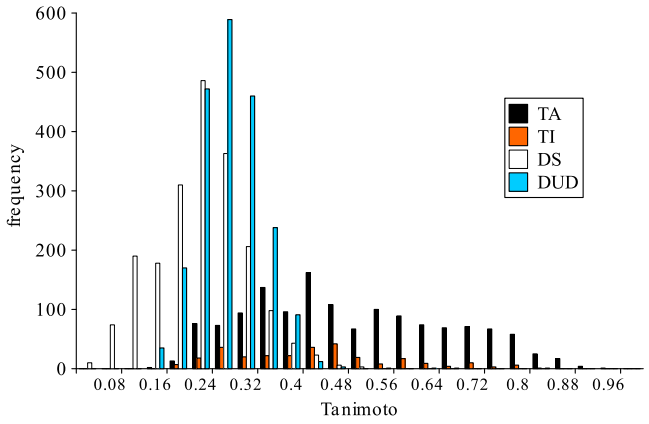

Figure 3. Distribution of scores for the retrospective VS. The [0,1] range was uniformly divided into 26 bins. TA, true actives; TI, true inactives; DS, NCI Diversity Set; DUD, Directory of Useful Decoys; see also Ref. 13.

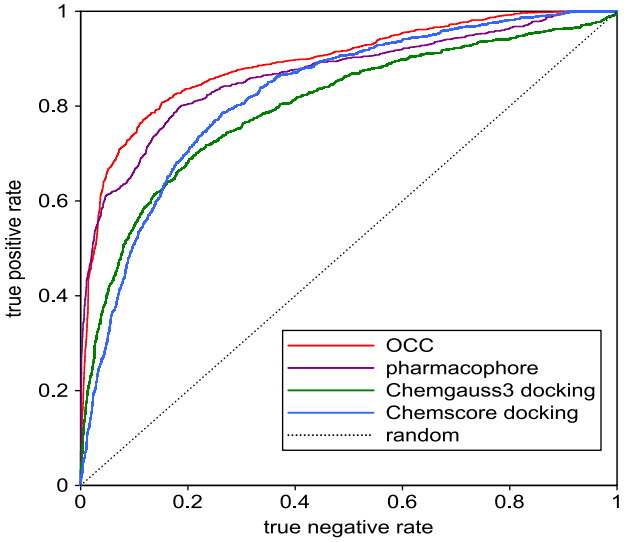

Figure 4. ROC curves for the different VS approaches.

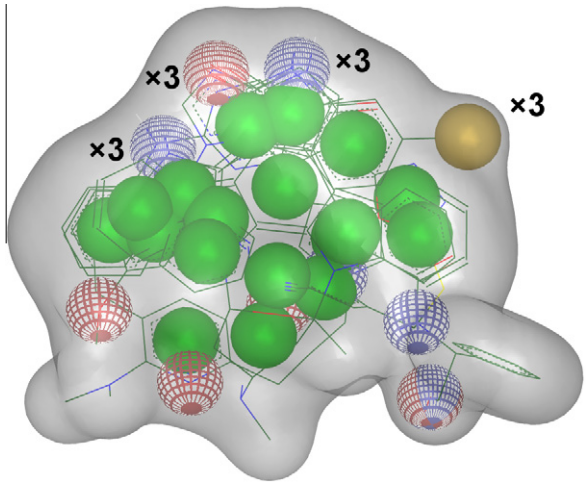

Figure 5. Pharmacophore model for GSK-3 inhibitors[19] is built by ROCS based on the co-crystal structures of hinge-aligned GSK-3-inhibitor complexes (PDB IDs 1Q3D,[20] 1UV5,[21] 2OW3,[22] 3F88,[23] and 3I4B[24]). The model was chosen from the series of the best ones owing to presence of hydrophobic feature (yellow ball) and maximal allowed molecular volume (grey cloud). Weight tripling for the features corresponding to the interaction of the inhibitor with the hinge (two hydrogen bond donors (blue balls), hydrogen bond acceptor (red balls)) and gatekeeper (hydrophobic) led to maximal *AUC* increase. Green balls correspond to the presence of ring features in the ligands.

Download English Version:

https://daneshyari.com/en/article/1370244

Download Persian Version:

https://daneshyari.com/article/1370244

Daneshyari.com