Mol Divers (2011) 15:947–955
DOI 10.1007/s11030-011-9325-2

FULL-LENGTH PAPER

Theoretical study of GSK-3α: neural networks QSAR studies for the design of new inhibitors using 2D descriptors

Isela García · Yagamare Fall · Xerardo García-Mera · Francisco Prado-Prado

Received: 25 February 2011 / Accepted: 20 June 2011 / Published online: 7 July 2011

Abstract Glycogen synthase kinase-3 (GSK-3) targets encompass proteins implicated in AD and neurological disorders. The functions of GSK-3 and its implication in various human diseases have triggered an active search for potent and selective GSK-3 inhibitors. In this sense, QSAR could play an important role in studying these GSK-3 inhibitors. For this reason, we developed QSAR models for GSK-3α, linear discriminant analysis (LDA), and artificial neural networks (ANNs) from nearly 50,000 cases with more than 700 different GSK-3α inhibitors obtained from ChEMBL database server; in total we used more than 20,000 different molecules to develop the QSAR models. The model correctly classified 237 out of 275 active compounds (86.2%) and 14,870 out of 15,970 non-active compounds (93.2%) in the training series. The overall training performance was 93.0%. Validation of the model was carried out using an external predicting series. In these series, the model classified correctly 458 out of 549 (83.4%) compounds and 29,637 out of 31,927 non-active compounds (83.4%). The overall predictability performance was 92.7%. In this study, we propose three types of non-linear ANN as alternative to already existing models, such as LDA. Linear neural network: LNN: 236:236-1-1:1 which had an overall training performance of 96% proved to be the best model. In addition, we did a study of the different fragments of the molecules of the database to see which fragments had more influence in the activity. This can help design new inhibitors of GSK-3α. This study reports the attempts to calculate, within a unified framework probabilities of GSK-3α inhibitors against different molecules found in the literature.

Keywords GSK-3α · QSAR · Artificial neural network · Linear neural network · Fragment contribution · Linear discriminant analysis

Introduction

Alzheimer's disease (AD) [1] is a serious and degenerative disorder that causes a gradual loss of neurons. In spite of the efforts realized by the big pharmaceutical companies of the world, the origin of this pathology is still not very clear. Glycogen synthase kinase-3 (GSK-3) is a serine–threonine kinase encoded by two isoforms in mammals, termed GSK-3α and GSK-3β. GSK-3 targets encompass proteins implicated in AD, neurological disorders, in the *Wnt* and insulin signaling pathway, glycogen and protein synthesis, regulation of transcription factors [2], embryonic development, cell proliferation and adhesion, tumorigenesis, apoptosis [3], circadian rhythm, etc. GSK-3β knock-out mice die in utero [4], whereas GSK-3α knock-out mice are viable and display improved glucose tolerance in response to glucose load and elevated hepatic glycogen storage and insulin sensitivity [5]. The functions of GSK-3 and its implication in various human diseases have triggered an active search for potent and selective GSK-3 inhibitors [6] in the last years. In this sense, quantitative structure–activity relationships (QSAR) could play an important role in studying these GSK-3 inhibitors (GSKI-3α); QSARs can be used as predictive tools

Electronic supplementary material The online version of this article (doi:10.1007/s11030-011-9325-2) contains supplementary material, which is available to authorized users.

I. García (✉) · Y. Fall
Faculty of Chemistry, Department of Organic Chemistry, University of Vigo, Vigo, Spain
e-mail: iselapintos@yahoo.es

X. García-Mera · F. Prado-Prado
Department of Organic Chemistry, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain

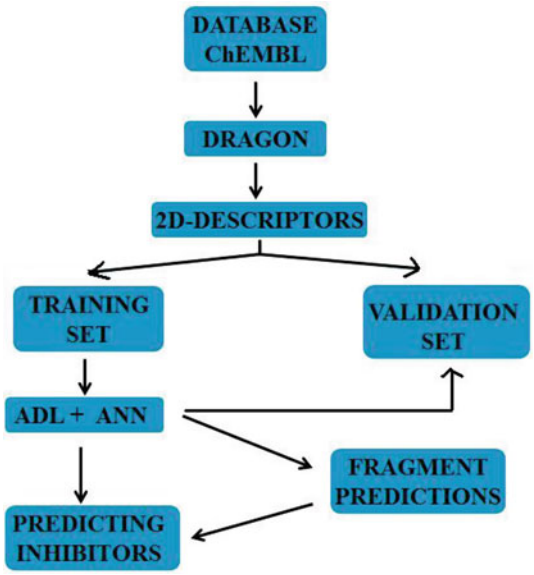

Fig. 1 Graphical approaches to study biological problems

for the development of molecules. Computer-aided drug design techniques based on QSAR could play an important role in drug discovery programs. The QSAR approach involves the development of models that relate the structure of drugs to their biological activity against different targets. Using graphical approaches to study biological problems can provide an intuitive picture or useful insights for helping analyzing complicated mechanisms in these systems, as demonstrated by many previous studies on a series of important biological topics, such as enzyme-catalyzed reactions [7], protein folding kinetics and folding rates, inhibition of HIV-1 reverse transcriptase, inhibition kinetics of processive nucleic acid polymerases and nucleases, drug metabolism systems, analysis of DNA sequence [8], interaction between amphiphilic helices in proteins, protein sequence evolution [9], etc. (Fig. 1).

The classical QSAR methods use physicochemical descriptors, whereas 3D-QSAR: CoMFA, CoMSIA, etc. use properties fields. In principle, there are currently more than 5,000 molecular descriptors that may be generalized and used to solve the problem outlined above [10]. Some of these indices are compiled in the Dragon software. Dragon provides 1,664 molecular descriptors divided into several families: 0D (constitutional descriptors), 1D (e.g., functional group counts), 2D (e.g., topological descriptors and connectivity indices), and 3D (e.g., GETAWAY, WHIM, RDF, and 3D-MoRSE descriptors) [11]. Numerous different molecular descriptors have been reported to encode chemical structures in QSAR studies. Furthermore, there are multiple chemometric approaches that can, in principle, be selected for this step. Multiple linear regression (MLR), linear discriminant analysis (LDA) [12], partial least squares (PLS), and different kinds of artificial neural networks (ANN) can be used to relate molecular structure (represented by molecular descriptors) to biological properties. The ANNs are particularly useful in QSAR studies in which the linear models fit poorly due to high data complexity [12], as an example we can cite the study of Prado-Prado et al. in which four types of non-linear ANN were developed to calculate within an unified framework probabilities of antiparasitic action of drugs against different parasite species [13].

There are several kinds of ANN and these include multilayer perceptron (MLP), radial basis functions (RBF), and PNNs; the latter is a variant of RBF systems. In particular, PNN is a type of neural network that uses a kernel-based approximation to form an estimate of the probability density functions of classes in a classification problem [14].

In this article, we developed QSAR models for GSK-3α, LDA, and ANNs from more than 48,000 cases. This database contains 700 different molecules inhibitors of GSK-3α (active against GSK-3α). To develop a different QSAR model, we used in total more than 20,000 different molecules which non-active or had not interaction with GSK-3α. All the tested compounds were compiled from ChEMBL database http://www.ebi.ac.uk/chembldb/index.php/target/browser/classification [15].

First, we used 237 molecular descriptors calculated with Dragon software [11]. Next, we developed LDA and ANNs models to find new compounds that present inhibitor action against GSK-3α and hence might be used as new treatments for neurological pathologies such as Parkinson's disease (PD) or AD. Finally, to check if our QSAR models are consistent, we established a Molecular Fragments study; we used different fragments from the known molecules of the database to see which fragments had more influence in the activity, and which fragments interacted more with the GSK-3α protein. The design of new inhibitors of this enzyme is very important for the study of neurodegenerative diseases [16].

This article is the first reported study that attempts to calculate framework probabilities of GSK-3α inhibitors against different molecules using a tool server ChEMBL database.

According to a recent comprehensive review [17], to establish a really useful statistical model for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the model; (ii) formulate the samples with an effective descriptor that can truly reflect their intrinsic correlation with the attribute to be investigated; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation (CV) tests to objectively evaluate the anticipated accuracy of the classifier; and (v) establish a user-friendly web-server for the classifier that is accessible to the public. Given below is a description of how to deal with these steps.

Methods

Linear classifier

A database from ChEMBL database [15] containing assayed GSK-3α inhibitors was used (Table S1 in Supplementary Material). To remove the homologous sequences from the benchmark dataset, a cutoff threshold of 25% was imposed [18] to exclude from the benchmark datasets, those proteins with equal sequence identity or 25% greater than any other in a same subset. However, in this study we did not use such a stringent criterion because the currently available data do not allow us to do so. Otherwise, the number of proteins for some subsets would be too few to have statistical significance.

We used the Dragon software 4.0 [11] to calculate the 2D molecular descriptors. This software has 1,664 descriptors classified as 0D, 1D, 2D, and 3D descriptors depending on whether they are computed from the chemical formula, substructure list representation, molecular graph, or geometrical representation of the molecule, respectively [19]. In this study, we calculated all 2D-descriptors that exist in the Dragon software: 2D autocorrelations, Burden eigenvalues, topological charge indices, eigenvalue-based indices, functional group counts, atoms-centered fragments, charge descriptors, and molecular properties. The QSAR model was constructed with the multivariate regression technique, the LDA, employing the Forward stepwise method for the selection of variables. All statistical analyses and data exploration were carried out in STATISTICA 6.0 [20]. In the actual study, the independent data test is used by splitting the data randomly in a training series used for a model construction and a CV one. In statistical prediction, the following three CV methods are often used to examine a classifier for its effectiveness in practical application: independent (or external) dataset test, subsampling test, and jackknife test. However, as elucidated in and demonstrated by Eqs.28–32 of [17], among the three CV methods, the jackknife test is deemed the least arbitrary that can always yield an unique result for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators to examine the accuracy of various predictors or classifiers (see e.g., [4–14,21–24]).

The general formula of the QSAR classification function is the following:

$$\text{GSKI} - 3\alpha_{\text{score}} = \sum W_m \cdot^m 2D_i + W_0, \tag{1}$$

where GSKI-3α_{score} is the continuous and dimensionless score value for the GSKI-3α/non-GSKI-3α classification that gives relatively higher values to molecules with more probability to act as GSKI-3α, ^m2D_i are the 2Ds of type m, W_m is the coefficient (weights) of these indices in the QSAR model, and W_0 is the independent term.

The reported statistical parameters of the QSAR model are the following: N, χ^2, and P-level as well as sensitivity, specificity, and accuracy for both training and CV [20]. N is the number of molecules used to train the model, λ is Wilks statistic parameter, χ^2 is Chi-square, and P-level is the probability of error.

Nonlinear classifiers

We processed our data with different ANNs using the STATISTICA 6.0 software [20] looking for a better model to predict activity against GSK-3α. Three types of ANNs were used, namely, Radial basis function (RBF: has a hidden layer of radial units, each actually modeling a Gaussian response surface) [25], Multi layers perceptron (MLP: the units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output, and the units are arranged in a layered feedforward topology), and linear neural network (LNN: where the fitted function is a hyperplane). The profile of a ANN is: Ni:I–H1–H2–O:No. It means that we have inputs variables (Ni), neurons in the input layer (I), neurons in the first hidden layer (H1), in the second hidden layer (H2), neuron in the output layer (O), and output variable (No).

We used a very simple type of ANN called linear neural network (LNN) to fit this discriminant function. The model deals with the classification of a compound set with or without affinity for different receptors. A dummy variable affinity class (AC) was used as input to codify the affinity. This variable indicates either high (AC = 1) or low (AC = 0) affinity of the drug for the receptor. $\text{S(DTP)}_{\text{pred}}$ or DTP affinity predicted score is the output of the model and it is a continuous dimensionless score that sorts compounds from low to high affinity for the target coinciding DTPs with higher values of $\text{S(DTP)}_{\text{pred}}$ and nDTPs with lowest values. In Eq. 2, b represents the coefficients of the LNN classification function, determined by the ANN module of the STATISTICA 6.0 software package [20]. We used Forward Stepwise algorithm for a variable selection.

Let be $^k\chi(G)$ drugs molecular descriptors and $^k\xi(R)$ receptor or drug target descriptors for different drugs (d) with different receptor; we can attempt to develop a simple linear classifier of mt-QSAR type with the general formula:

$$\text{S (DTP)}_{\text{pred}} = \sum_{k=0}^{5} b\,(G_k) \cdot {}^k\chi\,(G) + \sum_{k=0}^{5} b\,(R_k) \cdot {}^k\xi\,(R) + b. \tag{2}$$

We assessed the quality of models with different statistical parameters such as specificity (see Eq. 2), sensitivity (see Eq. 3), accuracy (see Eq. 4), and receiver operating

characteristic (ROC) curve which is a graphical plot of the sensitivity, or true positives, versus (1-specificity), or false positives,

$$\text{Specificity} = \frac{\text{NTN}}{\text{NTN} + \text{NFP}} \tag{3}$$

$$\text{Sensitivity} = \frac{\text{NTP}}{\text{NTP} + \text{NFN}} \tag{4}$$

$$\text{Accuracy} = \frac{\text{NTP} + \text{NTN}}{\text{NTP} + \text{FN} + \text{FP} + \text{TN}}, \tag{5}$$

where NTN means number of true negatives, NFP is number of false positives, NTP is number of true positives, NFN is number of false negatives, FN is false negatives, FP is false positives, and TN is true negatives.

Study of molecular fragments

In this study, we calculated contributions of different molecular fragments for activity against 15 different molecules inhibitors of GSK-3α. In so doing, we gave the following steps [13,26]:

- First, we calculated the specie-dependent atomic descriptors included in the QSAR equation for selected molecular fragments using DRAGON software [11].
- Second, we calculated the contribution scores of each fragment against 15 species of molecules studied by substituting the atomic descriptors into the QSAR equation using the Microsoft Excel application.
- Third, the contributions of each molecular fragment were standardized dividing each value by the sum of all contributions for each molecule. These molecular fragment contributions can indicate the potential relation between molecular fragments with the activity against GSK-3α, each separated fragment or each fragment inside a molecule.
- Fourth, contributions for each atom of the drug were scaled into a percentage value.
- Fifth, scaled atom contributions were grouped into different molecular fragments.
- Sixth, molecular fragment contributions to the biological activity were back-projected onto the molecular structure for obtaining a color-scaled biological structure–activity.

Data set

We developed QSAR models for GSK-3α, LDA, and ANNs from more than 48,000 cases. This database contains 700 different molecules inhibitors of GSK-3α (active against GSK-3α); in total, we used more than 20,000 different molecules non-active or not have interaction with GSK-3α to develop different QSAR models. All the compounds were compiled from ChEMBL database http://www.ebi.ac.uk/chembldb/index.php/target/browser/classification [15]. This is a database of bioactive drug-like small molecules, which contains 2D structures, calculated properties (e.g., log$P$, molecular weight, Lipinski parameters, etc.), and abstracted bioactivities (e.g., binding constants, pharmacology, and ADMET data). ChEMBL normalizes the bioactivities into an uniform set of end-points and units where possible, and also tags the links between a molecular target and a published assay with a set of varying confidence levels. The cutoff for selected active or non-active compounds was determinated by original papers with the biological experiments (IC_{50}, selectivity, % inhibition, etc.), for all units recopiled in the ChEMBL dataset. The data is abstracted and curated from the primary scientific literature, and covers a significant fraction of the structure–activity relationship (SAR) and discovery of modern drugs. The codes and activity for all compounds as well as the references used to collect them are depicted in Table S1 of the Supplementary Material.

Results and discussion

LDA

In this article, we obtained a LDA study, Eq. 6, and we can observe that 18 variables entry inside equation:

$$\begin{aligned}\text{GSKI} - 3\alpha_{\text{score}} = &-34.3 \cdot \text{ATS1}m + 19.3 \cdot \text{ATS3}m \\ &- 14.3 \cdot \text{ATS4}e - 2.9 \cdot \text{ATS6}e \\ &+ 37.0 \cdot \text{ATS1}p + 39.1 \cdot \text{MATS3}v \\ &- 7.3 \cdot \text{MATS3}e - 43.3 \cdot \text{MATS3}p \\ &- 3.7 \cdot \text{GATS5}m - 4.0 \cdot \text{GATS2}e \\ &- 46.5 \cdot \text{BEL}m3 - 34.6 \cdot \text{BEL}m4 \\ &+ 6.8 \cdot \text{BEL}e8 + 41.0 \cdot \text{BEL}p3 \\ &+ 36.0 \cdot \text{BEL}p4 + 1.1 \cdot \text{GGI3} \\ &- 227.2 \cdot \text{JGI4} + 163.4 \cdot \text{VE}m2 \\ &+ 40.7 \\ N = 48{,}721 \quad \chi^2 = 3127.3 \quad &P\text{-level} < 0.001 \end{aligned} \tag{6}$$

The nomenclature used in the descriptors of the equation is the same as establishing the Dragon software, where N is the number of cases, χ^2 is the Chi-square and P is the level of error. The model correctly classified 237 out of 275 active compounds (86.2%) and 14,870 out of 15,970 non-active compounds (93.2%) in the training series. The overall training performance was 93.0%. Validation of the model was carried out using an external predicting series. In this series, the model classified correctly 458 out of 549 (83.4%) active compounds and 29,637 out of 31,927 non-active compounds (92.8%). The overall predictability performance was 92.7% (see Table 1).

Table 1 Comparison of LDA and different ANNs classification models

Model profile	Train			Stat. Par.	Validation		
	Active	Non-active	%		%	Active	Non-active
LDA	237	38	86.2	Sn	83.4	458	91
	1100	14870	93.2	Sp	92.8	2290	29637
			93.0	Ac	92.7		
LNN	258	17	93.8	Sn	93.3	513	37
236:236-1:1	860	15625	94.8	Sp	94.4	1834	31146
			96.2	Ac	95.6		
MLP	97	178	35.3	Sn	34.6	190	360
22:22-27-1:1	11038	5447	33.0	Sp	33.3	22004	10976
			33.1	Ac	33.3		
RBF	210	65	76.4	Sn	72.4	398	152
98:98-740-1:1	4713	11772	71.4	Sp	71.4	9441	23539
			71.5	Ac	71.4		

ANN models

The ANN models are non-linear models useful to predict the biological activity of a large datasets of molecules. This technique is an alternative to linear methods such as LDA [27]. Figure 2 depicts the networks maps for some of the ANN models. In general, at least one ANN of every types tested was statically significant. However, one must note that the profiles of each network indicate that these are highly non-linear and complicated models [28,29].

In Fig. 2, we depict the ROC curve [30] for LNN tested. Notably, the ROC curve can also be represented equivalently by plotting the fraction of true positives (sensitivity) versus the fraction of false positives (1-specificity). The vitality of this type of procedures developing ANN-QSAR models has been demonstrated before [31]; see, for instance, the study of Fernandez and Caballero [32]. The same is true about the ANNs tested in this study, we illustrated a ROC curve for the best ANN model, in this case was a LNN (236:236-1:1) which results of ROC curve values (AUC) were with an area higher than 0.98. It indicates that this model gives statistically significant results and clearly different from those obtained with a random classifier (area $= 0.5$) [33]. To show how important this result is, we compared this model with other model used to address the same problem. We processed our data with ANNs looking for a better model [27].

The network found was LNN and it showed training performance higher than 96%. The summary of results is shown in Table 1. After direct inspection of the results reported in Table 1 for ANN methods, we can conclude that a complex ANN method is a good method to predict the activity. We compared different types of networks to obtain a better model; Table 1 shows the classification matrix of the different networks. LNN 236:236-1:1 was taken as the main network because it presented a wider range of variables, 236 inputs in the first layer and 236 neurons in second layer, and two sets of cases (training and validation). Regarding other tested networks, MLP 22:22-27-1:1 and RBF 98:98-740-1:1 presented low accuracy and PNN 237:237-16760-2-2:1 had a very low percentage of DTPs leading to possible errors in the model although, its accuracy was very good, see Table 1. We depict the ROC curve for LNN 236:236-1:1 to show how reliable was the developed network model, see Fig. 3.

Study of molecular fragments (F)

One application of QSAR is the calculation of the contribution of different molecular fragments to the desired activity [34,35]. In this sense, one important application of QSAR models is the calculation of molecular fragments contribution to activities or action against different drug targets [26]. Before, to get the different QSAR models we calculated the contribution of different molecular fragments from the better model obtained in this study. The LDA model was better than ANN models obtained, because the LDA model presented 18 variables to obtain one result (active or non-active GSK-3α), while the best LNN model presented 236 variables to obtain the same results. In spite of percentage of good classification the LNN model presents the highest classified percentage, but LNN model needs more variables than LDA model to predict one result. This LDA model developed in this study is a single equation that uses few parameters to predict the inhibitory action of a new compound against GSK-3α. For this reason, to obtain the best and the most reliable results, we calculated fragments of the molecules using the LDA method. As a result, we selected different molecular fragments against 15 molecules of database; we selected these

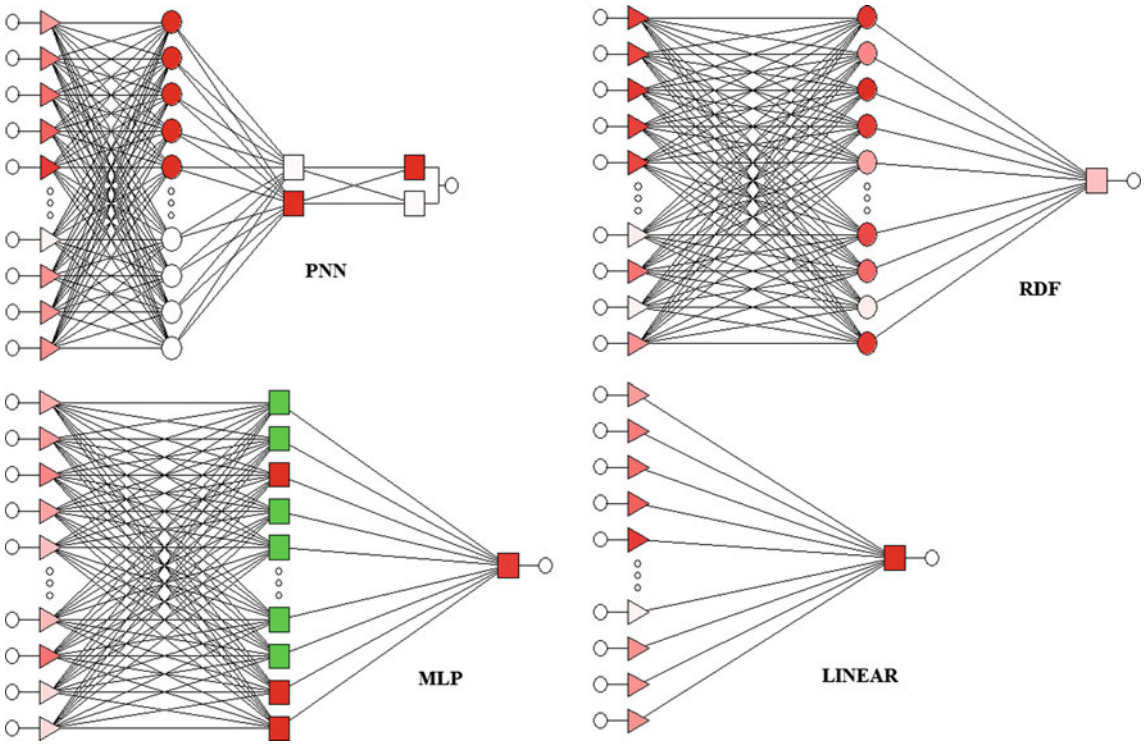

Fig. 2 Networks maps for some of the ANN models

molecules at random. In Fig. 4, we observed contribution of all fragments for the 15 molecules selected at random and because they are known molecules as potent inhibitors of GSK-3; whose results in our model are as follows: green color indicates major contribution to the activity, yellow color indicates medium contribution, and red color indicates minor contribution to the biological activity against GSK-3α. An example is the Dasatinib see Louise N. Johnson [36], where a pair of hydrogen bonds is formed in the hinge region of the ATP-binding site (i.e., between the 3-nitrogen of the aminothiazole ring of dasatinib and the amide nitrogen of Met318 and between the 2-amino hydrogen of dasatinib with the carbonyl oxygen of Met318). A hydrogen bond is also formed between the side chain hydroxyl oxygen of Thr315 and the amide nitrogen of dasatinib.

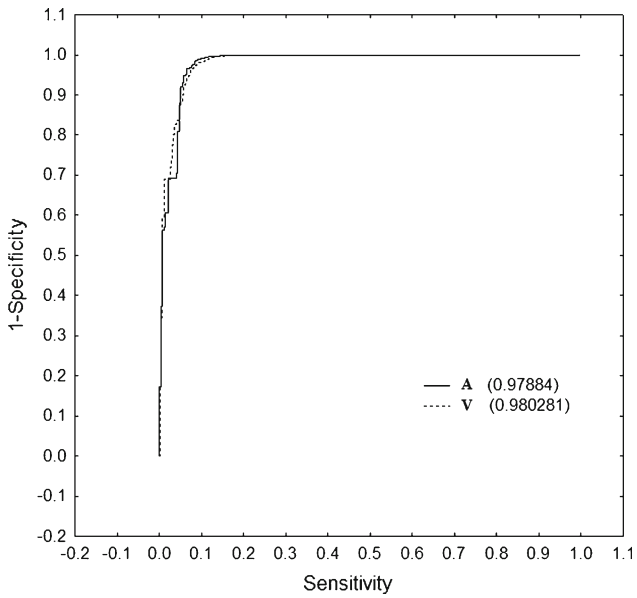

Fig. 3 ROC curve for LNN 236:236-1:1 tested

Domain of applicability of the model

QSAR studies can assess the accuracy of predictions in different ways. The simple ones try to distinguish reliable versus non-reliable predictions. They usually assume that the accuracy of prediction of molecules, which are inside a space of descriptors covered by the training set, is similar to the estimated accuracy of the model. These methods include: descriptor boxes, leverage-based approaches that evaluate the probability distribution of predictions and empirical approaches based on the "distance to model" concept [37]. The interest in QSAR has steadily increased in recent decades. It is generally acknowledged that these empirical relationships are valid only within the same domain for which they were developed. However, model validation is occasionally neglected and the application domain poorly defined [38].

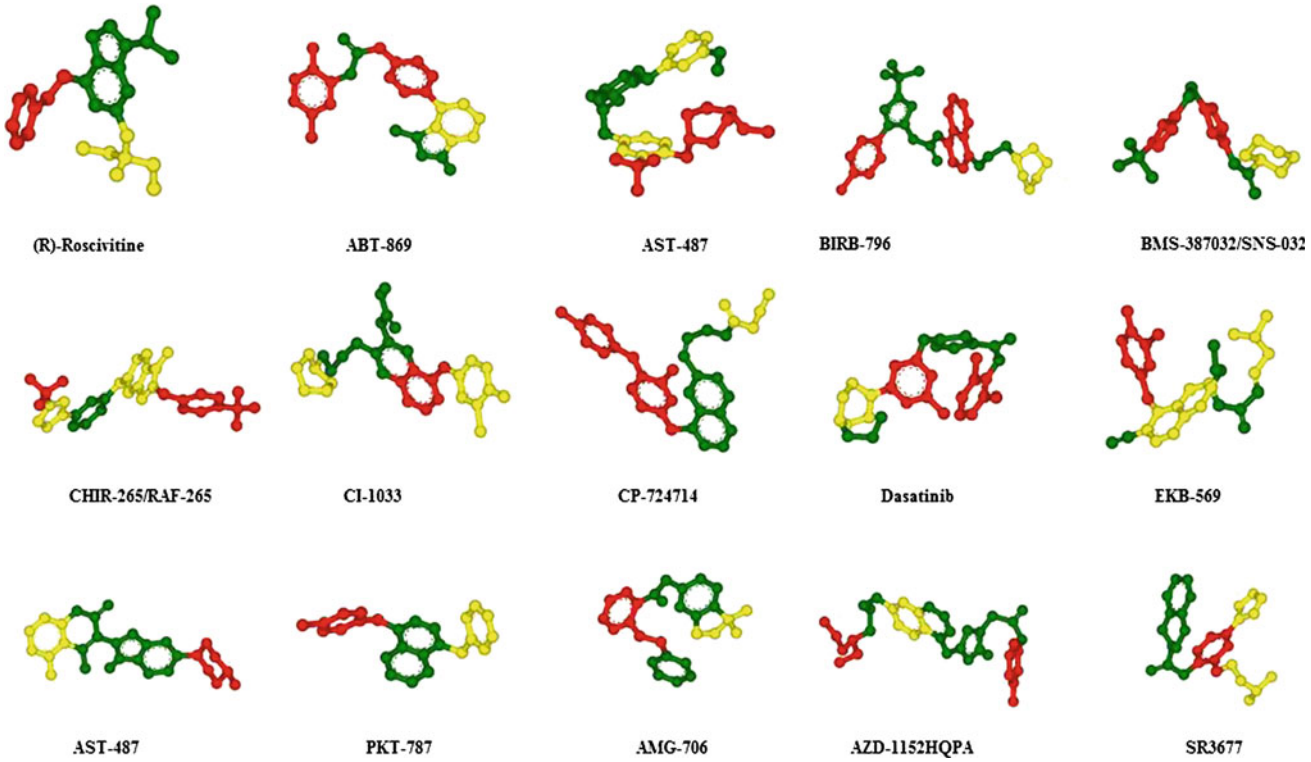

Fig. 4 Contribution of all fragments for the molecules. (Color figure online)

The purpose of this section is the method leverage-based to outline how validation and domain definition determines in which situation it is correct to use the model [39]. The domain of applicability can be characterized in various ways as it is defined by the descriptors used in the model and the studied response. Here a leverage approach was employed to verify the prediction reliability [39,40]. The leverage (h) is calculated by $hi = xi(XT\,X) - 1 \times Ti(i = 1, \ldots, m)$, where xi is the descriptor row-vector of the query compound i, m is the number of query compounds, and X is the $n \times k$ matrix of the training set (k is the number of model descriptors and n is the number of training set samples). The limit of normal values for X outliers ($h*$) is set as $3(k+1)/n$ and a leverage greater than $h*$. For the training set this means the chemical is highly influential in determining the model, while for the test set it means the prediction is the result of substantial extrapolation of the model and as such could be unreliable [41]. To examine this in further detail, a double ordinate Cartesian plot of deleted-residuals (first ordinate), standard residuals (second ordinate), and leverages (abscissa) defined the domain of applicability of the model as a squared area within ± 2 band for residuals and a leverage threshold of $h = 0.0011$. As can be noted in Fig. 5, almost all cases used in training and validation lie within this area. Some sequences have leverage higher than the threshold, but show leave-one out (LOO) residuals, deleted-residuals, and standard residuals within the limits. In short, no apparent outliers

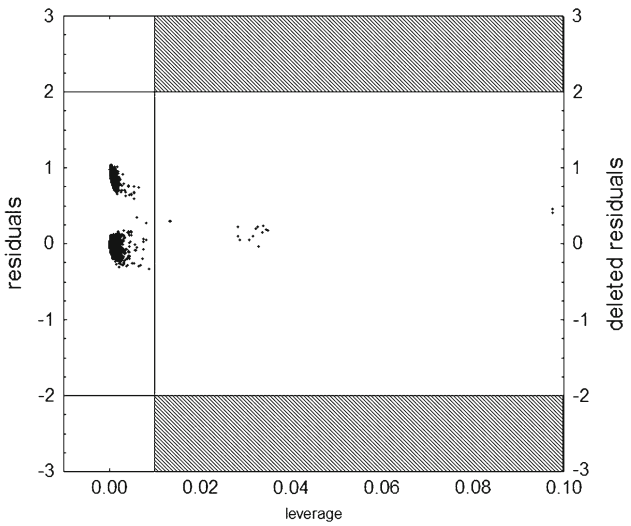

Fig. 5 Domain of applicability

were detected and the model can be used with high accuracy in this applicability domain [42–44].

Conclusion

The functions of GSK-3 and its implication in various human diseases have triggered an active search for potent and selective GSK-3 inhibitors. Nowadays theoretical studies such as QSAR models have become a very useful tool in this

context to substantially reduce time and resources consuming experiments. In this study, we developed a new LDA model using the Dragon descriptors and a large database with about 20,000 different drugs obtained from the ChemBL server. We conclude that a large database gives a much more precise model; the use of tools such as ChembL database allows us to develop models with large data bases and get more reliable results. To improve the model, we developed non-linear models and compared them with LDA. We proposed non-linear models, and for the first time, ANN models based on Dragon Descriptors series of GSK-3α, we concluded that they are alternative methods to study the activity of different families of molecules compared with other methods found in the literature. The use of tools such as fragment contributions can help us design the best GSK-3α inhibitors to be synthesized later in the laboratory, thus avoiding random synthesis of many molecules with few possibilities of being active.

Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful predictors or classifiers [45], we shall make efforts in our future study to provide a web-server for the method presented in this article.

Acknowledgments Prado-Prado F. thanks sponsorships for research position at the University of Santiago de Compostela from Angeles Alvariño, Xunta de Galicia. All authors acknowledge the Project 07CSA008203PR. We are grateful to the Xunta de Galicia (INCITE08 PXIB314255PR) for partial financial support.

References

1. Olson RE (2000) Secretase inhibitors as therapeutics for Alzheimer's disease. Annu Rep Med Chem 35:31–40. doi:10.1016/S0065-7743(00)35005-9
2. Troussard AA, Tan C, Yoganathan TN, Dedhar S (1999) Cell–extracellular matrix interactions stimulate the AP-1 transcription factor in an integrin-linked kinase- and glycogen synthase kinase 3-dependent manner. Mol Cell Biol 19:7420–7427. doi:0270-7306/99/$04.0010
3. Turenne GA, Price BD (2001) Glycogen synthase kinase3 beta phosphorylates serine 33 of p53 and activates p53's transcriptional activity. BMC Cell Biol 2:12–21. doi:10.1186/1471-2121-2-12
4. Hoeflich KP, Luo J, Rubie EA, Tasao MS, Jin O, Woodgett JR (2000) Requirement for glycogen synthase kinase-3beta in cell survival and NF-kappaB activation. Nature 406:86–90. doi:10.1038/35017574
5. MacAulay K, Doble BW, Patel S, Hansotia T, Sinclair EM, Drucker DJ et al (2007) Glycogen synthase kinase 3alpha-specific regulation of murine hepatic glycogen metabolism. Cell Metab 6:329–337. doi:10.1016/j.cmet.2007.08.013
6. Droucheau E, Primot A, Thomas V, Mattei D, Knockaert M, Richardson C et al (2004) *Plasmodium falciparum* glycogen synthase kinase-3: molecular model, expression, intracellular localisation and selective inhibitors. Biochim Biophys Acta 1697:181–196. doi:10.1016/j.bbapap.2003.11.023
7. Andraos J (2008) Kinetic plasticity and the determination of products ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Can J Chem 86:342–357. doi:10.1139/VO8-020
8. Xie G, Mo Z (2011) Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. J Theor Biol 269:123–130. doi:10.1016/j.jtbi.2010.10.018
9. Wu ZC, Xiao X, Chou KC (2010) 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol 267:29–34. doi:10.1016/j.jtbi.2010.08.007
10. Todeschini R, Consonni V (2002) Handbook of molecular descriptors. Wiley-VCH, New York
11. Talete srl (ed) DRAGON for Windows (Software for Molecular Descriptor Calculations)
12. Prado-Prado FJ, Borges F, Perez-Montoto LG, Gonzalez-Diaz H (2009) Multi-target spectral moment: QSAR for antifungal drugs vs. different fungi species. Eur J Med Chem 44:4051–4056. doi:10.1016/j.ejmech.2009.04.040
13. Prado-Prado FJ, García-Mera X, González-Díaz H (2010) Multi-target spectral moment QSAR versus ANN for antiparasitic drugs against different parasite species. Bioorg Med Chem 18:2225–2231. doi:10.1016/j.bmc.2010.01.068
14. Mosier PD, Jurs PC (2002) QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. J Chem Inf Comput Sci 42:1460–1470. doi:10.1021/ci020039i
15. War WA (2009) ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). J Comp Aided Mol Des 23:195–208. doi:10.1007/s10822-009-9260-9
16. Prado-Prado FJ, Ubeira FM, Borges F, Gonzalez-Diaz H (2010) Unified QSAR & network-based computational chemistry approach to antimicrobials. II. Multiple distance and triadic census analysis of antiparasitic drugs complex networks. J Comput Chem 31:164–173. doi:10.1002/jcc.21292
17. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. J Theor Biol 273:236–247. doi:10.1016/j.jtbi.2010.12.024
18. Chou KC, Shen HB (2010) Plant-mPLoc: a top-down strategy to augment the power for predicting plant protein subcellular localization. PLoS One 5:e11335
19. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
20. Hill T, Lewicki P (2002) STATISTICS. Tulsa:StatSoft
21. Chou KC, Shen HB (2010) Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Sci 2:1090–1103. doi:10.4236/ns.2010.210136
22. Kandaswamy KK, Chou KC, Martinetz T, Moller S, Suganthan PN, Sridharan S et al. (2011) AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol 270:56–62. doi:10.1016/j.jtbi.2010.10.037
23. Lin H, Ding H (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 269:64–69. doi:10.1016/j.jtbi.2010.10.019
24. Zakeri P, Moshiri B, Sadeghi M (2011) Prediction of protein submitochondria locations based on data fusion of various features of sequences. J Theor Biol 269:208–216. doi:10.1016/j.jtbi.2010.10.026
25. Melagraki G, Afantitis A, Sarimveis H, Igglessi-Markopoulou O, Alexandridis A (2006) A novel RBF neural network training methodology to predict toxicity to *Vibriofischeri*. Mol Divers 10:213–221. doi:10.1007/s11030-005-9008-y
26. Prado-Prado FJ, Gonzalez-Diaz H, de la Vega OM, Ubeira FM, Chou KC (2008) Unified QSAR approach to antimicrobials. Part 3: first multi-tasking QSAR model for input-coded prediction,

structural back-projection, and complex networks clustering of antiprotozoal compounds. Bioorg Med Chem 16:5871–5880. doi:10.1016/j.bmc.2008.04.068
27. Roy K, Mandal AS (2008) Development of linear and nonlinear predictive QSAR models and their external validation using molecular similarity principle for anti-HIV indolyl aryl sulfones. J Enzyme Inhib Med Chem 23:980–995. doi:10.1080/14756360701811379
28. Patra JC, Singh O (2009) Artificial neural networks-based approach to design ARIs using QSAR for diabetes mellitus. J Comp Chem 30:2494–2508. doi:10.1002/jcc.21240
29. Roy K, Mandal AS (2009) Predictive QSAR modeling of CCR5 antagonist piperidine derivatives using chemometric tools. J Enzyme Inhib Med Chem 24:205–223. doi:10.1080/14756360802051297
30. Gonzalez-Diaz H, Bonet I, Teran C, De Clerck E, Bello R, Garcia MM et al. (2007) ANN-QSAR model for selection of anticancer leads from structurally heterogeneous series of compounds. Eur J Med Chem 42:580–585. doi:10.1016/j.ejmech.2006.11.016
31. Aguero-Chapin G, Varona-Santos J, de la Riva GA, Antunes A, Gonzalez-Villa T, Uriarte E et al. (2009) Alignment-free prediction of polygalacturonases with pseudofolding topological indices: experimental isolation from *Coffeaarabica* and prediction of a new sequence. J Proteome Res 8:2122–2128. doi:10.1021/pr800867y
32. Fernandez M, Caballero J, Tundidor-Camba A (2006) Linear and nonlinear QSAR study of *N*-hydroxy-2-[(phenylsulfonyl)amino]acetamide derivatives as matrix metalloproteinase inhibitors. Bioorg Med Chem 14:4137–4150. doi:10.1016/j.bmc.2006.01.072
33. Prado-Prado FJ, Uriarte E, Borges F, Gonzalez-Diaz H (2009) Multi-target spectral moments for QSAR and complex networks study of antibacterial drugs. Eur J Med Chem 44:4516–4521. doi:10.1016/j.ejmech.2009.06.018
34. Estrada E, Molina E (2006) Automatic extraction of structural alerts for predicting chromosome aberrations of organic compounds. J Mol Graph Model 25:275–288. doi:10.1016/j.jmgm.2006.01.002
35. Estrada E, Uriarte E, Molina E, Simon-Manso Y, Milne GW (2006) An integrated in silico analysis of drug-binding to human serum albumin. J Chem Inf Model 46:2709–2724. doi:10.1021/ci600274f
36. Johnson LN (2009) Protein kinase inhibitors: contributions from structure to clinical compounds. Quart Rev Biophys 42:1–40. doi:10.1017/S0033583508004745
37. Sushko I, Novotarskyi S, Körner R, Pandey AK, Cherkasov A, Li J et al (2010) Applicability domains for classification problems: benchmarking of distance to models for Ames mutagenicity set. J Chem Inf Model 50:2094–2111. doi:10.1021/ci100253r
38. Oberg T (2004) A QSAR for baseline toxicity: validation, domain of application, and prediction. Chem Res Toxicol 17:1630–1637. doi:10.1021/tx0498253
39. Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701. doi:10.1002/qsar.200610151
40. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification and regression-based QSARs. Environ Health Perspect 111:1361–1375. doi:10.1289/ehp.5758
41. Li J, Gramatica P (2009) The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. Mol Divers doi:10.1007/s11030-009-9212-2
42. Gramatica P, Giani E, Papa E (2006) Statistical external validation and consensus modeling: a QSPR case study for K(oc) prediction. J Mol Graph Model 25:755–766. doi:10.1016/j.jmgm.2006.06.005
43. Liu H, Papa E, Gramatica P (2006) QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. Chem Res Toxicol 19:1540–1548. doi:10.1021/tx0601509
44. Papa E, Villa F, Gramatica P (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephalespromelas* (fathead minnow). J Chem Inf Model 45:1256–1266. doi:10.1021/ci0502121
45. Chou KC, Shen HB (2009) Recent advances in developing web-servers for predicting protein attributes. Nat Sci 2:63–92. doi:10.4236/ns.2009.12011