Implementation of PLS discriminant analysis to rank indirubin derivatives against decoys

Luminita Crisan¹, Liliana Pacureanu¹*, Sorin Avram¹, Alina Bora¹, Ludovic Kurunczi²

¹Institute of Chemistry of Romanian Academy, Department of Computational Chemistry, 300223 Timisoara, Romania
²University of Medicine and Pharmacy "Victor Babes", Department of Chemical Physics, Faculty of Pharmacy, 300041 Timisoara, Romania

Received 1 March 2013; Accepted 21 May 2013

Abstract: Partial Least Squares Discriminant Analysis (PLS-DA) is employed to obtain novel combinations of energetic terms present in classical scoring functions, which exceed and compensates the “traditional” consensus scheme. These novel scoring functions were involved to rank the database of indirubin inhibitors of glycogen synthase kinase-3β and cyclin dependent kinase-2 decoys from Directory of Useful Decoys. The ability of docking-scoring algorithm to prioritize the actives is assessed by means of several metrics. The best classification function includes donor component of Chemgauss2, steric contribution from Chemgauss3 and rotatable bond term of ScreenScore and provide significant improvement of enrichment factor at 5% of database.

Keywords: Indirubin • Glycogen synthase kinase-3 • PLSDA-DOCET • Docking • Scoring

1. Introduction

Recently, the interest for glycogen synthase kinase-3 (GSK-3) inhibitors increased significantly due to their promising pharmacological potential [1-3]. The elucidation of pharmacological role and mechanism of action represents the mainsprings of the continuous attention devoted to these compounds. Indirubins are bis-indole derivatives, which showed inhibitory activity upon GSK-3β. It has been shown that indirubin displays several pharmacological roles such as: (i) apoptosis induction in leukemic cells [4-6]; (ii) inhibition of the interferon-gamma production [7]; (iii) inhibition of the rabbit muscle glycogen phosphorylase β and aryl hydrocarbon receptor [8-10]. Since indirubin and its analogs inhibit cyclin dependent kinases (CDKs) and GSK-3β by ATP - competitive mechanism, this holds strong promises for the treatment of cancer and other diseases [11], conferring considerable relevance to the investigation of indirubin derivatives.

Molecular docking is an important tool for drug discovery. A large number of docking programs has been developed in the last years based on a variety of algorithms [12-14]. The use of docking algorithms in partnership with scoring functions to prioritize potentially active molecules from chemical libraries is now one of the most frequent prototypes of virtual screening. In the current study we used FRED (Fast Rigid Exhaustive Docking) developed by Open Eye Scientific Software because of its rapidity - only few seconds per ligand, which is significantly superior to most docking programs which usually allocate minutes per ligand [15-21]. The current report is concerned with a specific application of the protocol developed by our group, called Partial Least Squares Discriminant Analysis - Docking Optimized Combined Energetic Terms (PLSDA-DOCET) [22]. The PLSDA-DOCET algorithm [22] consists of two steps: (i) to reduce the data and to find the energetic terms, ETs, (the ETs of seven scoring functions available in FRED) that best discriminate the actives from the decoys by applying...
PLS-DA method: (ii) to identify the most successful combination showing the best early enrichment, the best overall discriminative power and both characteristics simultaneously. The aim of the current paper is to propose a novel, validated scoring function which exhibit superior performance for early enrichment of active compounds involving the ATP binding site of GSK-3β as target. The validation procedure uses a set of indirubins with selectivity towards GSK-3. The obtained scoring function can be used in virtual screening experiments to identify new GSK-3 inhibitors, with a certain degree of similarity against indirubin derivatives.

2. Computational details

2.1. Dataset preparation and docking procedure

A variety of indirubin derivatives were identified as powerful inhibitors of GSK3β [5]. The current data set consists of 109 indirubin derivatives [23-27] presented in Table 1 (see the corresponding SMILES codes in the Supplementary Information). In this work the indirubin derivatives are considered active, as they produce IC50 values below 200 μM in a GSK-3 inhibition assay (in Table 1 the domain for these values is between 200 - 0.0013 μM). These indirubin derivatives were assembled with a decoy set of 1778 molecules designed for CDK2 downloaded from DUD (Directory of Useful Decoys), which includes ionization states and tautomers, a total of 2070 distinct species [28,29]. The tautomers and ionizations states of indirubin derivatives were generated with LigPrep module from the Schrödinger package [30], in the pH range of 7.4±2.0. We used two criteria to define this combined (GSK-3β, CDK2) dataset. First, we have considered the high identity/similarity (52.9%/85%) of the amino acid binding sites between GSK-3β and CDK2 [28]. Second, the similarity test (based on MACCS fingerprints and Tanimoto coefficient) applied to indirubin derivatives and CDK-2 actives downloaded from DUD database, demonstrated structural similarity. Hence, 35.8% of indirubin derivatives show a Tanimoto similarity coefficient higher than 0.75 with at least one active from DUD database.

The CDK-2 decoys set has been considered inactive, even if they have not been experimentally tested against GSK-3β. It is worth mentioning that the molecule ZINC03814440, an compound included in CDK-2 actives from DUD, is an indirubin derivative with experimentally determined biological activity on both kinases GSK-3β (IC50 = 280 nM) and CDK-2 (IC50 = 35 nM) [28], and most part of CDK-2 active data set from DUD is purine-like inhibitors, a class of compounds that are active also against GSK-3 [31]. This fact suggests the opportunity to consider the CDK-2 decoys instead of specially designed decoys against GSK-3. The conformational sampling of the indirubin derivatives and the decoy set has been carried out with the Omega 2.4.3 module from OpenEye package [32-36]. The crystal structure of GSK-3β (PDB entry: 1Q41) in complex with (Z)-1H,1'H-[2,3']biindolyldene-3,2'-dione-3-oxime (IXM) has been downloaded from Brookhaven Protein Database (PDB http://www.rcsb.org/pdb). The receptor was prepared for docking [37] using the FRED RECEPTOR which generated an active site box of 7222 Å3 and inner contour of 63 Å3 and outer contour of 2009 Å3. We used the docking software FRED [15-21] to perform rigid protein - ligand docking on the pre-generated multiconformer database of molecules. The FRED docking consists roughly of two steps: shape fitting and optimization. During shape fitting, the ligand was placed into a 0.5 Å resolution grid box that includes all active-site atoms (including hydrogens) using a smooth Gaussian potential [21]. Further, all seven empirical scoring functions available in FRED were employed to score the docking poses: Shapegauss (SG) [21], Piecewise Linear Potential (PLP) [38], Chemgauss2 (CG2) [15], Chemgauss3 (CG3) [15], Chemscore (CS) [21,39], OEChemScore (OECS) [15], and ScreenScore (SS) [40,38]. Only the best scored tautomer and/or ionization state was retained for each molecule to perform PLS-DA DOCET consensus scoring evaluation.

2.2. PLS-DA method

The PLS-DA is a frequently used classification method that is based on the PLS approach, where the dependent variable is categorical [41]. The PLS-DA can reliably analyze a large number of independent variables, managing also the multicollinearity problem [42]. The procedure implemented in PLS-DA rotates the projection of the principal components in the latent variables space focusing on class separation [43]. The binary Y-variable was defined as follows: -1 for the decoy set and 1 for the indirubin derivatives. The aim is to find a discriminant plane in the principal component space that separates the active with respect to the inactive compounds.

The energetic component outputs of all scoring functions [15] were submitted to the SIMCA P 9.0 package [44] as independent variables to perform initially a Principal Component Analysis (PCA) [45], followed by the PLS-DA analysis. The PCA score plots were examined to distinguish object division into clusters and how much of the variance introduced in the X matrix (independent variables matrix) was explained.
Table 1. The template and experimental biological activity (IC$_{50}$) of indirubin derivatives.

| No | X | W | Y | Y' | Z | Z' | Q | IC$_{50}$ (μM) |
|----|---|---|---|----|---|----|---|----------------|
| 1  | H | =O | H | H | H | H | H | 1.000          |
| 2  | H | =NOH | H | H | H | H | H | 0.022          |
| 3  | H | =NOAc | H | H | H | H | H | 0.200          |
| 4  | H | =NOCH$_3$ | H | H | H | H | H | 0.150          |
| 5  | H | =O | H | H | Br | Br | H | 4.500          |
| 6  | H | =NOH | H | H | Br | Br | H | 0.120          |
| 7  | H | =O | H | H | Br | H | H | 22.00          |
| 8  | H | =NOH | H | H | H | Br | H | 0.340          |
| 9  | H | =O | H | H | Br | H | H | 0.045          |
| 10 | H | =NOH | H | H | Br | H | H | 0.005          |
| 11 | H | =NOAc | H | H | Br | H | H | 0.010          |
| 12 | H | =NOCH$_3$ | H | H | Br | H | H | 0.030          |
| 13 | H | =O | H | H | Cl | H | H | 0.140          |
| 14 | H | =NOH | H | H | Cl | H | H | 0.020          |
| 15 | H | =NOAc | H | H | Cl | H | H | 0.017          |
| 16 | H | =O | H | H | I  | H | H | 0.055          |
| 17 | H | =NOH | H | H | I  | H | H | 0.010          |
| 18 | H | =NOAc | H | H | I  | H | H | 0.013          |
| 19 | H | =O | H | H | CH=CH$_2$ | H | H | 0.240          |
| 20 | H | =NOH | H | H | CH=CH$_2$ | H | H | 0.060          |
| 21 | H | =NOAc | H | H | CH=CH$_2$ | H | H | 0.065          |
| 22 | H | =O | H | H | F  | H | H | 0.650          |
| 23 | H | =NOH | H | H | F  | H | H | 0.130          |
| 24 | H | =NOAc | H | H | F  | H | H | 0.090          |
| 25 | H | =O | CH$_3$ | H | Br | H | H | 0.025          |
| 26 | H | =NOH | CH$_3$ | H | Br | H | H | 0.006          |
| 27 | H | =NOAc | CH$_3$ | H | Br | H | H | 0.007          |
| 28 | H | =O | Cl | H | Cl | H | H | 0.030          |
| 29 | H | =NOH | Cl | H | Cl | H | H | 0.004          |
| 30 | H | =NOAc | Cl | H | Cl | H | H | 0.004          |
| 31 | H | =O | NO$_2$ | H | Br | H | H | 0.100          |
| 32 | H | =NOH | NO$_2$ | H | Br | H | H | 0.007          |
| 33 | H | =NOAc | NO$_2$ | H | Br | H | H | 0.006          |
| 34 | H | =NOH | I  | H | H | H | H | 0.009          |
| 35 | H | =O | SO$_2$-NH-C$_2$H$_5$-OH | H | H | H | H | 0.033          |
| 36 | H | =O | SO$_2$NH$_2$ | H | H | H | H | 0.040          |
| 37 | H | =O | NO$_2$ | H | H | H | H | 0.042          |
Continued Table 1. The template and experimental biological activity (IC₅₀) of indirubin derivatives.

| No | X     | W  | Y   | Y'  | Z   | Z'  | Q   | IC₅₀(μM) |
|----|-------|----|-----|-----|-----|-----|-----|---------|
| 38 | H     | -O | Cl  | H   | H   | H   | H   | 0.050   |
| 39 | H     | -O | Br  | H   | H   | H   | H   | 0.055   |
| 40 | H     | -O | CH₃ | H   | H   | H   | H   | 0.062   |
| 41 | H     | -O | I   | H   | H   | H   | H   | 0.068   |
| 42 | H     | -O | F   | H   | H   | H   | H   | 0.078   |
| 43 | H     | =NOH| SO₂H| H   | H   | H   | H   | 0.080   |
| 44 | H     | -O | SO₂-NHCH₃| H | H   | H   | H   | 0.110   |
| 45 | H     | -O | SO₂-N(CH₃)₂| H | H   | H   | H   | 0.180   |
| 46 | H     | -O | Br  | Br  | H   | H   | H   | 0.250   |
| 47 | H     | -O | SO₂H| H   | H   | H   | H   | 0.280   |
| 48 | H     | -O | H   | Br  | H   | H   | H   | 0.350   |
| 49 | H     | -O | SO₂-N(C₂H₄OH)₂| H | H   | H   | H   | 0.400   |
| 50 | H     | -O | SO₂H| Br  | H   | H   | H   | 4.000   |
| 51 | Phenyl| -O | H   | H   | H   | H   | H   | 200.000 |
| 52 | H     | -O | H   | H   | H   | F   |     | 0.400   |
| 53 | H     | =NOH| H   | H   | H   | F   |     | 0.270   |
| 54 | H     | =NOCH₃| H   | H   | H   | F   |     | 0.440   |
| 55 | H     | =NOCOCH₃| H   | H   | H   | F   |     | 0.330   |
| 56 | H     | =NOH| H   | H   | H   | Cl  |     | 21.000  |
| 57 | H     | =NOH| H   | H   | H   | Br  |     | 32.000  |
| 58 | H     | =NOH| H   | H   | H   | I   |     | 16.000  |
| 59 | CH₃   | =NOH| H   | H   | H   | I   |     | 30.000  |
| 60 | H     | =NOCH₂CH₂Br| H   | H   | H   | Br  |     | 100.000 |
| 61 | H     | NOCH₂CH₂-N  | H   | H   | H   | H   | Br  | 7.000  |
| 62 | H     | NOCH₂CH₂-N  | H   | H   | H   | H   | Br  | 3.000  |
| 63 | H     | NOCH₂CH₂-N  | H   | H   | H   | H   | Br  | 0.570  |
| 64 | H     | NOCH₂CH₂-N  | H   | H   | H   | H   | Br  | 9.000  |
| 65 | H     | NOCH₂CH₂-N  | H   | H   | H   | H   | Br  | 11.000 |
Continued | Table 1. The template and experimental biological activity (IC\textsubscript{50}) of indirubin derivatives.

| No | X | W | Y | Y' | Z | Z' | Q | IC\textsubscript{50}(μM) |
|----|---|---|---|---|---|---|---|--------------------------|
| 66 | H | NOCH\textsubscript{2}CH\textsubscript{2}N | NH | H | H | H | H | Br | 5.000 |
| 67 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3} | H | H | H | H | H | Br | 8.000 |
| 68 | H | =NOH | H | H | H | H | Br | 0.016 |
| 69 | H | =NOH | NO\textsubscript{2} | H | H | H | H | 0.0021 |
| 70 | CH\textsubscript{3} | =NOH | NO\textsubscript{2} | H | H | H | H | 0.530 |
| 71 | H | =NOH | NO\textsubscript{2} | Br | H | H | H | 0.055 |
| 72 | CH\textsubscript{3} | =NOH | NO\textsubscript{2} | Br | H | H | H | 9.300 |
| 73 | H | =O | NH\textsubscript{2} | H | H | H | H | 0.080 |
| 74 | H | =O | NHAc | H | H | H | H | 0.0075 |
| 75 | H | =O | NH\textsubscript{2} | Br | H | H | H | 0.440 |
| 76 | H | =O | NHAc | Br | H | H | H | 0.073 |
| 77 | H | =NOH | NH\textsubscript{2} | H | H | H | H | 0.360 |
| 78 | H | =NOH | NHAc | H | H | H | H | 0.350 |
| 79 | H | =NOH | NH\textsubscript{2} | Br | H | H | H | 6.600 |
| 80 | H | =NOH | NHAc | Br | H | H | H | 40.000 |
| 81 | H | =NOH | F | H | H | H | H | 1.300 |
| 82 | H | =NOH | F | Br | H | H | H | 15.000 |
| 83 | H | =NOH | Br | Br | H | H | H | 24.000 |
| 84 | H | =NOCH\textsubscript{2}CH\textsubscript{2}Br | H | H | Br | H | H | 0.14 |
| 85 | H | =NOCH\textsubscript{2}CH\textsubscript{2}OH | H | H | Br | H | H | 0.03 |
| 86 | H | =NOCH\textsubscript{2}CH\textsubscript{2}OH | H | H | Br | H | H | 0.034 |
| 87 | H | =NOCON(CH\textsubscript{2}CH\textsubscript{3})\textsubscript{3} | H | H | Br | H | H | 0.03 |
| 88 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3} | H | H | Br | H | H | 0.033 |
| 89 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3}HCl | H | H | Br | H | H | 0.029 |
| 90 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3} | H | H | Br | H | H | 0.035 |
| 91 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3}HCl | H | H | Br | H | H | 0.027 |
| 92 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3}OH | H | H | Br | H | H | 0.040 |
| 93 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3}OHHCl | H | H | Br | H | H | 0.041 |
| 94 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3}OHCH\textsubscript{2}OH | H | H | Br | H | H | 0.067 |
| 95 | H | =NOCH\textsubscript{2}CH\textsubscript{2}N(CH\textsubscript{2})\textsubscript{3}OHCH\textsubscript{2}OHHCl | H | H | Br | H | H | 0.023 |
| 96 | H | NOCH\textsubscript{2}CH\textsubscript{2}N | H | H | Br | H | H | 0.026 |
| 97 | H | NOCH\textsubscript{2}CH\textsubscript{2}N | H | H | Br | H | H | 0.054 |
Continued Table 1. The template and experimental biological activity (IC<sub>50</sub>) of indirubin derivatives.

![Chemical structure](image)

| No  | X | W      | Y     | Y'    | Z | Z' | Q  | IC<sub>50</sub>(μM) |
|-----|---|--------|-------|-------|---|----|----|--------------------|
| 98  | H | NOCH₂CH₂-N | H     | H     | Br| H  | H  | 0.060              |
| 99  | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.110              |
| 100 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.0033             |
| 101 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.0013             |
| 102 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.0070             |
| 103 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.0050             |
| 104 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.0050             |
| 105 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.0042             |
| 106 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.0110             |
| 107 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.02000            |
| 108 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.014              |
| 109 | H | NOCH₂CH₂-N  | H     | H     | Br| H  | H  | 0.033              |

* IC<sub>50</sub> values, calculated from the dose-response curves, are reported in μM

The statistical significance of the estimated predictive power and the absence of data overfit for the final model were certified by Y-permutation using 999 randomizations [46]. The dependent variable values are randomly interchanged and novel PLS-DA models based on the same independent variable matrix were built [47]. If the PLS-DA models obtained in this manner tend to exhibit minimal R<sup>2</sup> and Q<sup>2</sup> values, the chance correlation bias is excluded. True positive rate (TPR) and false positive rate (FPR) are defined as follows:
Implementation of PLS discriminant analysis to rank indirubin derivatives against decoys

TPR = \frac{TP}{TP + FN}
FPR = \frac{FP}{FP + TN}

where $TP$ denotes true positives, $FN$ is the false negative, $FP$ represents false positives, and $TN$ is true negatives. The TPR is also known as sensitivity and FPR as one minus specificity [48]. Enrichment give information about the improvement of the hit rate obtained by a VS protocol in comparison to random selection and is dependent on the ratio of the active molecules in the database [49].

$$EF = \frac{TP}{n}$$

where $N$ denotes the number of molecules in database, $A$ is the total number of active compounds and $n$ is the number of molecules in the % top.

3. Results and discussion

3.1. Docking results

In order to evaluate the performance of the docking method, the receiver operating characteristic (ROC) curve, area under the curve (AUC) and enrichment factor (EF) were used [48,50]. An ROC curve (Fig. 1) describes the evolution of the true positive rate (TPR, y-axis) plotted versus the false positive rate (FPR, x-axis) during the ranking of database (the thresholds were considered at every single compound detection) [50]. The position of ROC curve with respect to the diagonal line (random selection) provides information about the ability of the scoring function to prioritize the true positives with respect to true negatives, whereas the area under the curve (AUC) gives a quantitative measure of the global performance of the method employed [50]. Generally speaking, the effectiveness of the method to discriminate the actives from decoys is directly proportional with the AUC [50]. The AUC allows overall evaluation of virtual screening algorithm, but enrichment factor is essential for the assessment of early enrichment, since only the top ranked compounds are experimentally tested. Along these lines, the enrichment at 2%, 5%, 10% and 25% of the database were calculated (Table 2). The values of the AUC parameter are situated between 0.679 for Shapegauss and 0.965 in the case of Chemgauss3, where only two scoring functions Shapegauss and Chemgauss2 performance is below 0.8.

These three criteria (AUC, ROC, enrichment factor) designated Chemgauss3 as the best single scoring function to rank the indirubin derivatives with respect to the decoy set (Fig. 1, Tables 2, 3). More negative mean values of interaction energies registered for active compounds were observed for most of the scoring functions and energetic terms (see Table 2). In order to obtain a maximum discrimination among these two classes (indirubin derivatives versus decoys) a PLS-DA analysis was performed.

3.2. PLS-DA results

The PLS discriminant analysis was involved to improve the ranking of the actives with respect to decoys (see Table 2). The PLS discriminant analysis provides a reliable solution to the problem of classification, whilst the output parameters can offer a suggestive graphical representation of class separation [51].

In the first step of the PLS-DA analysis a PCA model was constructed for the whole $X$ matrix: $N = 1887$ rows (109 indirubin derivatives and 1778 decoys), and $K = 34$ columns (32 energetic terms for the $X$ matrix, and 2 columns for dependent variable matrix $Y$). The PCA on the $X$ block was carried out using variable autoscaling algorithm that multiply the variables by suitable weights to
provide them unit variance (i.e., the same importance) [51,52]. The PCA score plot for the first two principal components accounting for 51.6% of the X matrix information content, generated from the energetic terms, is represented in Fig. 2. As can be seen in Fig. 2, the actives (squares, in black) and the inactive compounds (triangles, in red) are separated well enough using these two latent variables alone. Subsequently, the initial PLS-DA model was constructed starting from the same X matrix. In the PLS-DA method the latent variables are constructed to describe the maximum covariance between the X and Y block matrices, allowing the separation of the two classes [53,54]. In Fig. 3 one can observe the clustering of actives and decoys in the space of the first two PLS latent variables, indicating the opportunity of the PLS-DA method for classification.

**Table 2.** The minimum, maximum and mean scores for indirubin derivatives and decoys.*

| No. | Energetic Terms | Score for indirubin derivatives | Score for decoy sets |
|-----|-----------------|----------------------------------|----------------------|
|     |                 | min | max   | mean  | min | max   | mean  |
| 1   | SG              | -632.481 | -415.854 | -501.110 | -559.467 | -345.350 | -465.275 |
| 2   | PLP             | -76.550 | -48.129 | -60.735 | -73.136 | 4.506   | -48.905 |
| 3   | CG2             | -79.991 | -49.926 | -60.854 | -72.560 | -40.531 | -54.688 |
| 4   | CG3             | -109.506 | -77.952 | -92.213 | -106.288 | -24.120 | -73.077 |
| 5   | CS              | -30.617 | -14.753 | -22.974 | -30.906 | 3.099   | -14.995 |
| 6   | OECS            | -51.681 | -33.827 | -40.250 | -46.541 | -16.361 | -34.532 |
| 7   | SS              | -183.920 | -126.829 | -151.994 | -175.125 | -28.783 | -121.417 |
| 8   | CG2_Steric      | -70.256 | -45.373 | -54.9408 | -63.564 | -39.007 | -51.7726 |
| 9   | CG2_Acc/Metal   | -5.325 | 0.000 | -2.35886 | -10.687 | 0.175   | -1.85281 |
| 10  | CG2_Donor       | -11.383 | -0.260 | -3.59441 | -10.409 | 0.326   | -1.37068 |
| 11  | CG2_Aromatic    | -0.298 | 0.047 | -0.14438 | -0.795 | 1.069   | -0.09133 |
| 12  | CG3_Steric      | -108.503 | -79.201 | -93.5504 | -109.337 | -16.933 | -77.4889 |
| 13  | CG3_Desolvation | 4.798 | 18.300 | 11.1748 | 3.640 | 24.944 | 12.5745 |
| 14  | CG3_Acc         | -10.432 | 0.000 | -3.47337 | -25.938 | 0.000   | -4.82461 |
| 15  | CG3_Donor       | -25.499 | 0.000 | -8.21608 | -29.387 | 0.000   | -4.75558 |
| 16  | CS_RB           | 0.000 | 5.990 | 0.84633 | 0.000 | 9.882 | 3.34481 |
| 17  | CS_LiPO         | -29.401 | -17.388 | -22.6266 | -31.387 | -9.286 | -19.949 |
| 18  | CS_HB           | -11.139 | 0.000 | -3.82236 | -13.630 | 0.000   | -3.20109 |
| 19  | CS_Clash        | 0.000 | 8.037 | 1.92316 | 0.000 | 23.633 | 4.23369 |
| 20  | OECS_Lipo       | -46.473 | -30.184 | -35.9393 | -49.871 | -20.547 | -34.0417 |
| 21  | OECS_HB         | -15.199 | 0.000 | -6.66043 | -16.278 | 0.000   | -5.19218 |
| 22  | OECS_Clash      | 0.000 | 8.037 | 1.92316 | 0.000 | 23.633 | 4.23369 |
| 23  | PLP_HB          | -15.410 | -0.280 | -7.60053 | -17.422 | 23.757 | -4.38957 |
| 24  | PLP_NP          | -66.838 | -38.983 | -53.7788 | -62.559 | -11.960 | -44.3397 |
| 25  | PLP_Sulphur     | 0.000 | 0.000 | 0 | 0.000 | 9.882 | 3.34481 |
| 26  | SS_RB           | 0.000 | 14.400 | 3.43486 | 1.600 | 14.000 | 8.458043 |
| 27  | SS_Lipo         | -82.847 | -48.901 | -66.0379 | -78.460 | -27.763 | -56.7085 |
| 28  | SS_Ambig        | -56.110 | -19.817 | -32.0554 | -59.223 | -18.167 | -36.3397 |
| 29  | SS_Clash        | 3.120 | 22.310 | 8.45914 | 3.008 | 46.756 | 15.2837 |
| 30  | SS_PLP          | -67.278 | -37.846 | -53.6853 | -64.307 | -4.003 | -49.4375 |
| 31  | SS_HB           | -16.938 | 0.000 | -7.2841 | -18.553 | 0.000   | -4.94375 |
| 32  | SS_Aromatic     | -10.368 | -3.647 | -6.89561 | -12.231 | 0.000   | -5.09123 |

* SG-Shapegauss; PLP-Pairwise Linear Potential; CG2-Chemgauss2; CG3-Chemgauss3; CS-Chemscore; OECS-OpenEye Chemscore; SS-ScreenScore.
Implementation of PLS discriminant analysis to rank indirubin derivatives against decoys

In order to eliminate the overfitting, six subsequent different PLS-DA models were constructed, retaining in each step only the variables with PLS coefficient significantly different from zero (t-Student test) [52]. Since the data set is imbalanced - the number of inactives (1778) is much higher than the number of actives (109) - the statistical results for the final model are not spectacular - the cumulative sum of squares (SS) of all the X values explained by all extracted components $R^2_X(CUM) = 0.502$, the cumulative SS of all the Y's and X's explained by all extracted components $R^2_{Y(CUM)} = 0.416$, and the fraction of the total variation of Y values that can be predicted for all extracted principal components $Q^2_{Y(CUM)} = 0.412$. Nevertheless, these results are acceptable, bearing in mind the very low difference between the $R^2$ and $Q^2$ values (0.004). This last fact indicates a robust, predictable model. The absence of overfit in the final PLS-DA model was assessed by the Y-permutation procedure using 999 randomizations [46]. The plot displayed in Fig. 4 demonstrates that the Y-intercept values corresponding to the $R^2$ and $Q^2$ lines do not exceed 0.3, and 0.05 respectively (in fact they are −0.00311, and −0.0148 respectively). These results indicate a valid model, without overfit. In this valid model, nine out of thirty two energetic terms were selected as significant variables for the discrimination between actives and inactives. These variables are CG2_Donor (Chemgauss2 contributions from donors on the ligand interacting with acceptors on the protein), CG3_Steric (Chemgauss3 steric interactions), CS_Lipo (Chemscore

Figure 2. PCA score plots for the first two principal components t1,t2 - actives (black squares) and inactives (red triangles).

Figure 3. PLS-DA score plots, for the first two latent variables - actives (black squares) and inactives (red triangles).
interaction between lipophilic atoms), OECS_Lipo (OEChemScor interaction between lipophilic atoms), SS_RB (ScreenScore rotatable bond penalty), SS_Ambig (ScreenScore ambiguous interactions), SS_HB (ScreenScore hydrogen bonds), SS_Clash (Screencore penalty for clashes with the protein) and SG_SH (Shapegauss shape term) [21]. For seven energetic terms excluding OECS_Lipo and SS_Ambig that shows VIP values below 0.8, all the possible combinations were constructed, their sums were calculated, and the first ten best combinations were selected (Table 3). The sums of these equally weighted energetic terms represent the new PLS-DA “mixed” scoring functions.

The performance of CG3 scoring function in addition to AUC (0.965) and enrichment factor at 2% of database (29.358%) is noteworthy. However, these results were surpassed by the AUC values of the resulted PLS-DA combinations of the energetic components (C1-C10), exhibiting AUC values greater than 0.982 and enrichment factors higher than for single scoring functions, exceeding 30% (maximum 33.945%) at 2% of database (see Table 3). At 5% of the database the improvement registered by C3 is higher 75.229%, while for the individual scoring function the maximal enrichment factor is 54.128% in the case of Chemscore. The best enrichment factor at 10% of the database is provided by C2 (98.165%) with respect to single scoring function CG3 (82.567%). The improvement of early enrichment is crucial for virtual screening purposes since only the top ranked compounds are tested experimentally.

Analyzing the ten combinations one can observe that two energetic terms are repeatedly included: CG3 steric interactions and SS rotatable bond (CG3_Steric and SS_RB). In combination with Chemgauss2 donor (CG2_Donor), the energetic terms above offer the best separation of indirubin derivatives and decoys (see the AUC values (Table 3) and the ROC curve (Fig. 5)).
The CG3_Steric takes into account the contacts between protein heavy atoms and heavy atoms of the ligand, and includes a correction term [15,21]. The steric component accounts also for the potential energy described by two effects: (i) the van der Waals interactions when the ligand is fitted in the binding site, (ii) protein desolvation energy owing to water transfer from the binding site into solvent, disregarding the hydrogen bonds that may take place between water and the binding site [15,21]. The CG2_Donor term registers the hydrogen bonding interaction energy that takes place between the ligand and protein [15,21]. The SS_RB component from ScreenScore represents a penalty term proportional to the number of rotatable bonds in the ligand [40]. The SS_RB is an important term in our situation since a significant number of compounds display a considerable number of flexible bonds in the decoys (≤11) and ligands (≤9). The selected energetic terms refer to the main interactions that occur in the binding site such as steric, donor and one important penalty term - rotatable bond. Hence, the combination of these terms can constitute a novel scoring function whose ability to rank the database; especially the early enrichment is superior to that of standard scoring functions for the studied receptor.

### 4. Conclusions

A promising workflow in terms of speed and performance for structure-based virtual screening using the rigid docking engine FRED followed by PLS-DA analysis of the energetic terms corresponding to the scoring functions has been applied to indirubin derivatives and CDK-2 DUD decoys, targeting the GSK3β receptor. A new “mixed” scoring function was built. It assembles the energy terms from different scoring functions that illustrate the particular interactions in the GSK-3 binding site. The results obtained demonstrated better evaluation metrics than those obtained employing ‘single’ scoring functions available “de facto” in the OpenEye package. The present study helped us to identify the optimal modus operandi to acquire the highest early enrichment of actives in the top 2% to 25% of the database for seven classical and ten “assembled” scoring functions. The AUC of 0.965 and enrichment factor of 49.541% at 5% of the database for the classical CG3 scoring function represents a good performance for indirubin derivatives. However, the comparison of enrichment factors and AUC provided by the best standard scoring function and PLS-DA outcome showed a better performance of the PLS-DA scoring (AUC = 0.991, EF=75.229% at 5% of the database). In extended virtual screening experiments our mixed scoring function C1 can furnish with very good chances valuable new GSK-3 inhibitors.

### Acknowledgements

We thank OpenEye for providing academic license, Dr. Erik Johansson (Umetrics, Sweden) for kindly providing the SIMCA-P 9.0 package and to Dr. Ramona Curpan for kindly providing the Schrödinger software. This project was financially supported by the Project No. 1.2 of the Institute of Chemistry of Romanian Academy, Timisoara.
References

[1] L. Meijer, N. Guyard, L.A. Skaltsounis, G. Eisenbrand (Eds.), Indirubin, the Red Shade of Indigo (Life in Progress), (Station Biologique, Roscoff, 2006)
[2] A.R. Katritzky, L.M. Pacureau, D.A. Dobchev, D.C. Faro, P.R. Duchowicz, M. Karelson, Bioorg. Med. Chem. 14, 4987 (2006)
[3] A.M.F. Oliveira-Campos, A.M. Salaheldin, F.A. Almeida Paz, L.M. Rodrigues, Arkivoc xi, 121 (2011)
[4] G.Y. Wu, F.D. Fang, Acta Acad. Med. Sinica 2, 83 (1980)
[5] K. Suzuki, R. Adachi, A. Hirayama, H. Watanabe, S. Otani, Y. Watanabe, T. Kasahara, Br. Haematol. 130, 681 (2005)
[6] G. Sethi, K.S. Ahn, S.K. Sandur, X. Lin, M.M. Chaturvedi, B.B. Aggarwal, J. Biol. Chem. 281, 34225 (2006)
[7] T. Kunikata, T. Tatefuji, H. Aga, K. Iwaki, M. Ikeda, M. Kurimoto, Eur. J. Pharmacol. 410, 93 (2000)
[8] K. Suzuki, R. Adachi, A. Hirayama, H. Watanabe, S. Otani, Y. Watanabe, S. Leclerc, M. Garnier, R. Hoessel, D. Marko, J.A. Bibb, G.L. Snyder, P. Greengard, J. Biernat, Y.Z. Wu, E.M. Mandelkow, G. Eisenbrand, J. Biol. Chem. 276, 251 (2001)
[9] M.N. Kosmopoulou, D.D. Leonidas, E.D. Chryssina, N. Bischler, G. Eisenbrand, C.E. Sakarellos, R. Paupit, N.G. Oikonomakos, N. G. Eur. J. Biochem. 271, 2280 (2004)
[10] J. Adachi, Y. Mori, S. Matsui, H. Takigami, J. Fujino, H. Kitagawa, C.A. Miller III, T. Kato, K. Saeki, T. Matsuda, J. Biol. Chem. 276, 31475 (2001)
[11] S.E. Nikoulinia, T.P. Ciardaldi, S. Mudaliar, L. Carter, K. Johnson, R.R. Henry, Diabetes 51, 681 (2005)
[12] I. Muegge, M. Rarey, Small molecule docking and scoring (Wiley-VCH, New-York, 2001) Vol. 17, pp.1-60
[13] R.D. Taylor, P.J. Jewsbury, J.W. Essex, J. Comput. Aided Mol. Des. 16, 151 (2002)
[14] I. Halperin, B. Ma, H. Wolfson, R. Nussinov, Proteins: Struct., Funct., Genet. 47, 409 (2002)
[15] FRED, version 2.2.5 OpenEye (Scientific Software, Inc., Santa Fe, NM, 2009)
[16] M. McGann, J. Comp. Aid. Mol. Design, 26, 897 (2012)
[17] M. McGann, J. Chem. Inf. Model. 51, 578 (2011)
[18] S.L. Swann, S.P. Brown, S.W. Muchmore, H. Patel, P. Merta, J. Locklear, P.J.. Hajduk, J. Med. Chem. 54, 1223 (2011)
[19] A.J.S. Knox, T. Price, M. Pawlak, G. Golfs, C.T. Flood, D. Fayne, D.C. Williams, M.J. Meegan, D.G. Lloyd, J. Med. Chem. 52, 2177 (2009)
[20] G.B. McGaughey, R.P. Sheridan, C.I. Bayly, J.C. Cuberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.F. Truchon, W.D. Cornell, J. Chem. Inf. Model. 47, 1504 (2007)
[21] M.R. McGann, H.R. Almond, A. Nicholls, J.A.Grant, F.K. Brown, Biopolymers 68, 76 (2003)
[22] S. Avram, L.M. Pacureauu, E. Seclaman, A. Bora, L. Kurunczi, J. Chem. Inf. Model. 51, 3169 (2011)
[23] S. Leclerc, M. Garnier, R.Hoessel, D. Marko, J.A. Bibb, G.L. Snyder, P. Greengard, J. Biernat, Y.Z. Wu, E.M. Mandelkow, G. Eisenbrand, J. Biol. Chem. 276, 251 (2001)
[24] P. Polychronopoulou, P. Magiatis, A.L. Skaltsounis, V. Myriantopoulos, E. Mikros, A. Tarricone, A. Musachio, S.M. Roe, L. Pearl, M. Leost, P. Greengard, L. Meijer, J. Med. Chem. 47, 935 (2004)
[25] A. Beauchard, Y. Ferandion, S. Fre’re, O. Lozach, M. Blairvacq, L. Meijer, V. Thie’ry, T. Besson, Bioorg. Med. Chem. 14, 6434 (2006)
[26] Y. Ferandion, K. Bettayeb, M. Kritsanida, O. Lozach, P. Polychronopoulou, P. Magiatis, A.L. Skaltsounis, L. Meijer, J. Med. Chem. 49, 4638 (2006)
[27] K. Vougogiannopoulou, Y. Ferandion, K. Bettayeb, V. Myriantopoulos, O. Lozach, Y. Fan, C.H. Johnson, P. Magiatis, A.L. Skaltsounis, E. Mikros, L. Meijer, J. Med. Chem. 51, 6421 (2008)
[28] N. Huang, B.K. Shoichet, J.J. Irwin, J. Med. Chem. 49, 6789 (2006)
[29] A. Jahn, G. Hinselmann, N. Fechner, A. Zell, J. Cheminf. 1, 14 (2009)
[30] Schrodinger Suite 2010, LigPrep, version 2.4 (Schrödinger, LLC, New York, NY, 2010)
[31] H. Eldar-Finkelman, A. Martinez, Front. Mol. Neurosci. 4, 32 (2011)
[32] OpenEye Scientific Software (OpenEye Scientific Software, Inc., USA, 2009) http://www.eyesopen.com/
[33] P.C.D. Hawkins, A. Nicholls, J. Chem. Inf. Model. 52, 2919 (2012)
[34] P.C.D. Hawkins, A.G. Skillman, G.L. Warren, B.A. Ellingson, M.T Stahl, J. Chem. Inf. Model. 50, 572 (2010)
[35] E. Perola, P.S. Charifson, J. Med. Chem. 47, 2499 (2004)
[36] J. Bostrom, J.R. Greenwood, J. Gottfries, J. Mol. Graphics Model. 21, 449 (2003)
[37] L. Crisan, L. Pacureauu, A. Bora, S. Avram, L. Kurunczi, Z. Simon, Cent. Eur. J. Chem. 11, 63 (2013)
[38] D.K. Gehlhaar, G.M. Verkhivker, P.A. Rejto, C.J. Sherman, D.B. Fogel, L.J. Fogel, S. T. Freer, Chem. Biol. 2, 317 (1995)
[39] M.D. Eldridge, C.W. Murray, T.R. Auton, G.V. Paolini, R.P. Mee, J. Comput.-Aided Mol. Des. 11, 425 (1997)
[40] M. Stahl, M. Rarey, J. Med. Chem. 44, 1035 (2001)
[41] M. Barker, W. Rayens, J. Chemometrics 17, 166 (2003)
[42] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58, 109 (2001)
[43] M. Sjöström, S. Wold, B. Söderström, PLS Discriminant Plots (Proceedings of PARC in Practice) (Elsevier Science Publishers B.V., Amsterdam, North-Holland, 1985)
[44] SIMCA P, version 9.0 (Umetrics AB: Umea, Sweden, 2001)
[45] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, B. Walczak, Chemom. Intell. Lab. Syst. 85, 203 (2007)
[46] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi and Megivariate Data Analysis, Principles and Applications (Umetrics AB: Umea, Sweden, 2001) 92-97, 489-491
[47] P.P. Roy, S. Paul, I. Mitra, K. Roy, Molecules, 14, 1660 (2009)
[48] J.A. Hanley, B.J. McNeil, Radiology. 143, 29 (1982)
[49] J. Kirchmair, P. Markt, R. Distinto, G. Wolber, T. Langer, J. Comput. Aided. Mol. Des. 22, 213 (2008)
[50] N. Triballeau, F. Acher, I. Brabet, J.P. Pin, H.O. Bertrand, J. Med. Chem. 48, 2534 (2005)
[51] S. Wold, Chemom. Intell. Lab. Syst. 2, 37 (1987)
[52] (a) S. Wold, E. Johansson, M. Cocchi, 3D QSAR in Drug Design Theory, Methods, and Applications (ESCOM, Leiden, Holland, 1993) 523–550; (b) SIMCA P, version 9.0. (Umetrics AB: Umea, Sweden, 2001) http://www.umetrics.com
[53] L. Stahle, S. Wold, J. Chemom. 1, 185 (1987)
[54] S. Wold, M. Sjostrom, L. Eriksson, In: P.v.R. Schleyer, N.L. Allinger, T. Clark, J. Gasteiger, P.A. Kollman, H. F., Schaefer, III, P.R. Schreiner, (Eds.) PLS in chemistry: The Encyclopedia of Computational Chemistry (Wiley, Chichester, UK, 1999) 2006-2020