PLS (Partial Least Square) Study for GSK-3 (Glycogen synthase kinase-3) Inhibition by Indirubin Derivatives

LUMINITA CRISAN [1], ALINA BORA[1]*, LILIANA PACUREANU[1], SORIN AVRAM[1], LUDOVIC KURUNCZI[2]
[a]Romanian Academy, Institute of Chemistry Timisoara, 24 Mihai Viteazul , 300223, Timisoara, Roumania
[b]University of Medicine and Pharmacy "Victor Babes" Timisoara, Faculty of Pharmacy, 2 E. Murgu, 300041 Timisoara, Roumania

GSK-3 (Glycogen synthase kinase-3) is a serine/threonine protein kinase involved in various physiological processes and its inhibitors have been evaluated as promising drug candidates for severe pathologies. In this paper a PLS regression analysis was used, to explore the relationship between field of descriptors (as independent variable) and biological activities (pIC50 as dependent variables). The main purpose of this work is to develop a robust QSAR model for predicting inhibitory activity of the 109 indirubin derivatives and better understanding of structural features of these compounds and their relation with the inhibitory activity using the Dragon descriptors. The statistical qualities of the final model have been assessed by several parameters such as: the cumulative sum of squares of correlation coefficient R^2_Y(CUM)=0.872, the cumulative fraction of the total variation of the Y values that can be predicted for all the extracted principal components by cross validation Q^2(CUM) = 0.804,R^2 supplied by the Golbraikh-Tropsha external validation method and Y-randomization test.

Keywords: Projections in Latent Structures (PLS), glycogen synthase kinase-3β (GSK-3β), Dragon descriptors, indirubin derivatives, validation, randomization.

Protein kinases are involved in numerous physiological processes and in many diseases. They are one of the largest known families of enzymes characterized by having a well-conserved ATP binding pocket [1].

Glycogen synthase kinase-3 (GSK-3) is a serine/threonine protein kinase found in two closely related isforms, GSK-3α and GSK-3β, which share high homology (ca. 90%) in the catalytic domain. GSK-3β plays a critical role in glucose homeostasis, CNS function (via tau proteins and b-catenin), and cancer (via angiogenesis, apoptosis, and tumorigenesis) [2]. Several publications have emerged describing structurally diverse molecules that inhibit GSK-3 [3]. Many compounds have also been reported to stimulate glycogen synthesis in vitro as well as lower plasma glucose in diabetic animals [4].

Different classes of ATP-competitive GSK-3β inhibitors have been described, such as indirubins, maleimide derivatives, paullone, staurosporine, aloisines, and the marine sponge hymenialdisine. Up to now 32 X-ray cocrystallographic structures of GSK-3 with the above mentioned inhibitors were deposited in the Brookhaven Protein Data Bank (PDB). They provide a considerable amount of information regarding the interaction between inhibitors and GSK-3. These findings revealed the flexibility of the ATP binding pocket that will help the design of novel, selective GSK-3 inhibitors [5].

Regarding the high therapeutical potential of targeting GSK-3 in different pathologies, the search for its inhibitors is a successful area in both academic centers and pharmaceutical companies. The bis-indole indirubin is an active component of Danggui Longhui Wan, a histhoric Chinese medicine formula used against chronic diseases such as leukemias and Alzheimer. However, indirubin shows pharmacokinetic properties. In order to improve pharmacological properties and to reduce toxicity, two indirubin analoges, such as 5-chloro-indirubin and indirubin-3'- monoxime have been synthesized. Indirubins have been demonstrated to be powerful inhibitors (IC_{50} =5–50 nM) of GSK-3β , while other indigoids are inactive. Testing a series of indoles and bis-indoles against GSK-3β, CDK1/cyclin B, and CDK5/p25 shows that only indirubins inhibit these kinases [6]. Numerous structure-activity relationship studies suggests that indirubins bind to GSK-3β's ATP binding pocket in a similar manner to their binding to CDKs, information ustained by the existing crystallographic data [7].

In the current study, we have applied PLS technique to a series of 109 indirubin derivatives with known biological activity. The relevant PLS model for this series was constructed in the following steps: (1) generation of ligand molecular structures; (2) geometry optimization on the molecular structures using semiempirical AM1 method; (3) calculation of various structural descriptors using the DRAGON software; (4) generation of PLS models; (5) the evaluation of the developed QSAR models by validation, predictability and robustness. Until now, in the literature a complete study concerning PLS approaches for all the 109 indirubin derivatives with known biological activity was not published. Such a study could guide the design of more convenient inhibitors and contribute to a better understanding of GSK-3β inhibition.

Our goal is to develop a robust model that meaningfully suggests a structure activity relationship of indirubin derivatives in terms of their affinity to GSK-3β receptor.

Experimental part

Methods and materials

Dataset

For this study 109 indirubin derivatives, that were collected from the literature [8- 12], and their biological activities were selected and listed in table 1. The biological activity for indirubin derivatives has been expressed as negative logarithm of experimental half-maximal effective concentrations (molar units, M) pIC50. The general template of indirubin analogues built with the help of Symyx Draw 3.3 [13] program is depicted in figure1.

* *email: alina.bora@gmail.com; Tel.: +40256491818*

Structural calculations

First of all, 109 investigated molecules were pre-optimized by means of the Molecular Mechanics Force Field (MM+) included in HyperChem version 7.52 [14] package. The resulted minimized structures were further refined using the semiempirical AM1 Hamiltonian. We choose a RMS gradient norm limit of 0.01kcal/A for the geometry optimization. In order to have the "real" spatial orientation of the subtituents of the bis-indol moiety, a conformational search for the flexible lateral chains of this rigid skeleton has been carried out with Conformational search module also available in HyperChem 7.52 and only the low energy conformations were retained (maximum 1 kcal/mol above the lowest).

Molecular descriptors calculation

A set of 1668 molecular descriptors of different kinds was used to describe the chemical diversity of the 109 compounds. In our study, the whole set of descriptors resulted from DRAGON [15] was used to build the X-matrix suitable for PLS analysis as following: 35 constitutional descriptors, 91 topological descriptors, 46 walk and path counts, 30 connectivity indices, 47 information indices, 96 2D autocorrelations, 106 edge adjacency indices, 64 BCUT descriptors, 17 topological charge indices, 44 eigenvalue-based indices, 41 Randic molecular profiles, 50 geometrical descriptors, 150 RDF descriptors, 160 3D-MoRSE descriptors, 99 WHIM descriptors, 195 GETAWAY descriptors, 31 functional group counts, 36 atom-centered fragments, 14 charge descriptors, 23 molecular properties, 134 2D binary fingerprints, 159 2D frequency fingerprints. The complete list of these molecular descriptors, and their meaning, is provided with literature references by the DRAGON package [15]; the calculation procedure is explained in detail, with related literature references, in the Molecular Descriptors for Chemoinformatics [16].

PLS method

PLS is a regression method that works with two matrices, **X** (e.g., chemical descriptors) and **Y** (e.g., biological responses), and has two objectives, namely approximate well X and Y, and to model the relationship between them [17]. The QSAR matrix was submitted to the SIMCA P 9.0 package [18] to perform initially PCA (Principal Component Analysis) [19], and afterwards a PLS analysis. In order to interpret the results of PLS, the equation in latent variables was transformed as function of the original X_{ij} (i =1, 2,..., N; j=1, 2,..., K) variables, resulting eq. (1) (with $\hat{Y}_i$ the calculated dependent variable, i.e. the calculated pKi value, and b_j the PLS coefficients):

$$\hat{Y}_i = b_0 + b_1X_{i1} + b_2X_{i2} + ... + b_jX_{ij} + ... + b_KX_{iK} \quad (1)$$

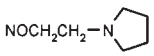
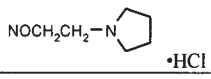
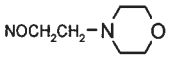
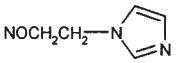
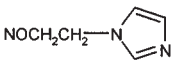
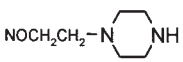
This method is preferable for large data sets. For the present study the dataset was split into training and test set using Golbraikh - Tropsha algorithm based on a sphere-exclusion algorithm [20-22], (the compounds from the test set were not used in the model development phase). The first two statistical parameters that provide a measure of the quality and validity for the final PLS model are the following: the correlation coefficient, R^2, and the cross-validated correlation coefficient, Q^2 (PLS results).

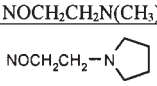
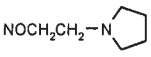
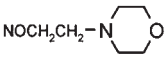
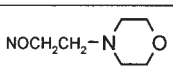
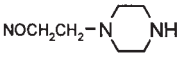
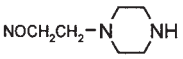
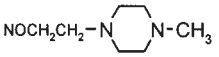
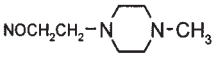
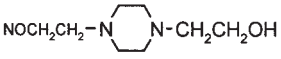
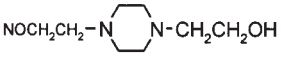
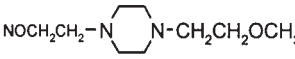
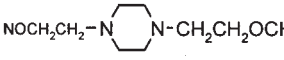
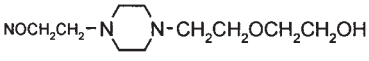
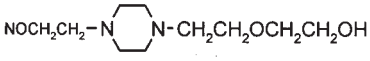
Model validity

To certify the predictive ability of our PLS model, the following satisfied conditions for the test set were considered: (i) cross validation correlation coefficient $Q^2>0.5$; (ii) the correlation coefficient R between the predicted and observed activities $R^2 > 0.6$; (iii) concerning the coefficients of determination for predicted versus observed activities R^2_0, and observed versus predicted

Table1
THE STRUCTURE AND BIOLOGICAL ACTIVITY OF INDIRUBIN ANALOGUES

	No	pIC50	R1/R5'/R6'/R7/ R5/R6	R3'
1	001	6	H/H/H/H/ H/H	O
2	002	7.657	H/H/H/H/ H/H	NOH
3	003	6.698	H/H/H/H/ H/H	NOAc
4	004	6.823	H/H/H/H/ H/H	$NOCH_3$
5	005	5.346	H/H/Br/H/ H/Br	O
6	006	6.921	H/H/Br/H/ H/Br	NOH
7	007	4.657	H/H/Br/H/ H/H	O
8*	008	6.468	H/H/Br/H/ H/H	NOH
9	009	7.347	H/H/H/H/ H/Br	O
10	010	8.301	H/H/H/H/ H/Br	NOH
11	011	8	H/H/H/H/ H/Br	NOAc
12	012	7.522	H/H/H/H/ H/Br	$NOCH_3$
13	015	6.854	H/H/H/H/ H/Cl	O
14*	016	7.699	H/H/H/H/ H/Cl	NOH
15	017	7.769	H/H/H/H/ H/Cl	NOAc
16	018	7.259	H/H/H/H/ H/I	O
17	019	8	H/H/H/H/ H/I	NOH
18	020	7.886	H/H/H/H/ H/I	NOAc
19	021	6.619	H/H/H/H/ H/ $CH=CH_2$	O
20	022	7.222	H/H/H/H/ H/ $CH=CH_2$	NOH
21	023	7.187	H/H/H/H/ H/ $CH=CH_2$	NOAc
22	024	6.187	H/H/H/H/ H/F	O
23	025	6.886	H/H/H/H/ H/F	NOH
24	026	7.046	H/H/H/H/ H/F	NOAc

25	029	7.602	H/H/H/H/ CH_3/Br	O
26	030	8.222	H/H/H/H/ CH_3/Br	NOH
27*	031	8.155	H/H/H/H/ CH_3/Br	NOAc
28	032	7.523	H/H/H/H/ Cl/Cl	O
29	033	8.398	H/H/H/H/ Cl/Cl	NOH
30	034	8.398	H/H/H/H/ Cl/Cl	NOAc
31	035	7	H/H/H/H/ NO_2/Br	O
32	036	8.155	H/H/H/H/ NO_2/Br	NOH
33	037	8.222	H/H/H/H/ NO_2/Br	NOH
34	039	8.046	H/H/H/H/ I/H	NOH
35	040	7.482	H/H/H/H/ SO_2-NH-C_2H_4-OH/H	O
36	041	7.398	H/H/H/H/ SO_2NH_2/H	O
37	042	7.377	H/H/H/H/ NO_2/H	O
38*	043	7.301	H/H/H/H/ Cl/H	O
39*	044	7.259	H/H/H/H/ Br/H	O
40	045	7.207	H/H/H/H/ CH_3/H	O
41	046	7.1675	H/H/H/H/ I/H	O
42	047	7.108	H/H/H/H/ F/H	O
43	048	7.097	H/H/H/H/ SO_3H/H	NOH
44	049	6.958	H/H/H/H/ SO_2-$NHCH_3$/H	O
45	050	6.745	H/H/H/H/ SO_3-$N(CH_3)_2$/H	O
46	051	6.602	H/Br/H/H/ Br/H	O
47	052	6.553	H/H/H/H/ SO_3H/H	O
48	053	6.456	H/Br/H/H/ H/H	O
49*	054	6.398	H/H/H/H/ SO_2-$N(C_2H_4OH)_2$/H	O
50	055	5.398	H/Br/H/H/ SO_3H/H	O
51	061	3.699	Ph/ H/H/H/ H/H	O
52	080	6.398	H/H/H/F/ H/H	O
53	081	6.569	H/H/H/F/ H/H	NOH
54	082	6.356	H/H/H/F/ H/H	$NOCH_3$
55	083	6.481	H/H/H/F/ H/H	$NOCOCH_3$
56	089	4.678	H/H/H/Cl/ H/H	NOH
57*	097	4.495	H/H/H/Br/ H/H	NOH
58	105	4.796	H/H/H/I/ H/H	NOH
59	109	4.523	CH_3/ H/H/I/ H/H	NOH
60	112	4	H/H/H/Br/ H/H	$NOCH_2CH_2Br$
61	114	5.155	H/H/H/Br/ H/H	
62	115	5.523	H/H/H/Br/ H/H	 •HCl
63	118	6.244	H/H/H/Br/ H/H	
64	122	5.046	H/H/H/Br/ H/H	
65	123	4.959	H/H/H/Br/ H/H	 •HCl
66*	126	5.301	H/H/H/Br/ H/H	
67*	128	5.097	H/H/H/Br/ H/H	$NOCH_2CH_2N(CH_3)_2$
68	138	7.796	H/H/H/Br/ Br/H	NOH
69	142	8.678	H/H/H/H/ NO_2/H	NOH
70	143	6.276	CH_3/ H/H/Br/ NO_2/H	NOH
71	144	7.259	H/Br/H/H/ NO_2/H	NOH

72	145	5.031	CH_3/Br/H/H/ NO_2/H	NOH
73	146	7.097	H/H/H/H/ NH_2/H	O
74*	147	8.125	H/H/H/H/ NH-Ac/H	O
75	148	6.356	H/Br/H/H/ NH_2/H	O
76	149	7.136	H/Br/H/H/ NH-Ac/H	O
77*	150	6.444	H/H/H/H/ NH_2/H	NOH
78	151	6.456	H/H/H/H/ NH-Ac/H	NOH
79	152	5.180	H/Br/H/H/ NH_2/H	NOH
80	153	4.398	H/Br/H/H/ NH-Ac/H	NOH
81*	158	5.887	H/H/H/H/ F/H	NOH
82	159	4.824	H/Br/H/H/ F/H	NOH
83	161	4.619	H/Br/H/H/ Br/H	NOH
84	164	6.854	H/H/H/H/ H/Br	$NOCH_2CH_2Br$
85	165	7.523	H/H/H/H/ H/Br	$NOCH_2CH_2OH$
86	166	7.468	H/H/H/H/ H/Br	$NOCH_2CH(OH)CH_2OH$
87	167	7.523	H/H/H/H/ H/Br	$NOCON(CH_2CH_3)_2$
88	168	7.485	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_3)_2$
89	169	7.537	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_3)_2$•HCl
90	170	7.456	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_2CH_3)_2$
91	171	7.568	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_2CH_3)_2$•HCl
92	172	7.398	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_2CH_2OH)_2$
93	173	7.387	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_2CH_2OH)_2$•HCl
94*	174	7.174	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_3)CH_2CH(OH)CH_2OH$
95*	175	7.638	H/H/H/H/ H/Br	$NOCH_2CH_2N(CH_3)CH_2CH(OH)CH_2OH$•HCl
96*	176	7.585	H/H/H/H/ H/Br	
97	177	7.267	H/H/H/H/ H/Br	 •HCl
98	178	7.222	H/H/H/H/ H/Br	
99	179	6.958	H/H/H/H/ H/Br	 •HCl
100	180	8.481	H/H/H/H/ H/Br	
101	181	8.886	H/H/H/H/ H/Br	 •2HCl
102	182	8.155	H/H/H/H/ H/Br	
103	183	8.301	H/H/H/H/ H/Br	 •2HCl
104	184	8.301	H/H/H/H/ H/Br	
105	185	8.376	H/H/H/H/ H/Br	 •2HCl
106*	186	7.958	H/H/H/H/ H/Br	
107*	187	7.699	H/H/H/H/ H/Br	 •2HCl
108	188	7.853	H/H/H/H/ H/Br	
109	189	7.481	H/H/H/H/ H/Br	 •2HCl

* Compounds used in the test set; these data will be used later in the model validation phase

activities R'^2_0 through the origin $(R^2 - R_0^2)/R^2 < 0.1$ or $(R^2 - R_0'^2)/R^2 < 0.1$ and $|R_0^2 - R_0'^2| < 0.3$; (iv) slopes k and k' of the regression lines through the origin $(0.85 \le k \le 1.15$ or $0.85 \le k \le 1.15$ or $0.85 \le k' \le 1.15$ [20-22].

Y-randomization test is a technique that shows the robustness of a QSAR model. The Y-values (dependent variable) are randomly interchanged and a QSAR model is built using descriptor matrix. The obtained QSAR models (after ten randomizations) must provide the minimal R^2 and Q^2 values. In order to avoid chance correlation, R^2_p parameter [23] was calculated. This R^2_p parameter penalizes the model R^2 for the difference between squared mean correlation coefficient (R^2_r) of randomized models and squared correlation coefficient (R^2) of the non-randomized model (eq.2).

$$R_p^2 = R^2 \cdot \sqrt{R^2 - R_r^2} \qquad (2)$$

For a statistically adequate model that exclude chance correlation, the parameter R^2_p should take values higher than 0.5 [23].

Results and discussions

In order to correlate the activity values with calculated descriptors, the PLS calculations with the SIMCA P09 package were performed. The training and test set construction phase classified 92 molecules in the first category and 17 in the second. A PCA model (M1) was built for the X matrix (N=92 rows/compounds, and K= 1668 columns/descriptors) and 45 principal components resulted. From the 45 principal components resulted, the first three components already explain 57% of the information content of the QSAR matrix. The first PLS model was constructed using the same X matrix. The difference between R^2_Y(CUM) and Q^2(CUM) values demonstrated the overfit of the model. The relatively small values, R^2_Y(CUM) = 0.677, and Q^2(CUM) = 0.364, also suggest the need for enhancement of the model quality. These improvements were obtained using the normal probability plot for Y standardized residuals as outlier criterion (standard deviation higher than 2, as the statistical significance level adopted is 0.05 for all the diagnostic tests used in the PLS procedure) and gradually eliminating the outliers. On the other hand the overfit was exceeded by excluding the noise variables in the model (coefficient values insignificantly different from 0).

Twelve compounds (table2), were gradually eliminated as outliers, the coefficients were reclaimed and 10 intermediary models were obtained. In so doing, we obtained two models namely M11 (N= 80 and K= 87) and M12 (N= 80 and K= 60), considered as possible final models. A part of the variables contained by model, M11, were not important for describing the interaction between the indirubin derivatives and the GSK-3β. From that reason these variables are considered as noise variables. Further the noise variables have been removed and the final model, M12, is robust and include three principal components (table2). For this model, the small difference between the R^2_Y and Q^2 values and high Q^2(CUM) value does not automatically imply a high predictive power of the model. The predictive abilities of our model are tested in the next step.

Table 2
STATISTICAL CHARACTERISTICS OF THE DEDUCED PLS MODELS*

PLS Model	R^2_X(CUM)	R^2_Y(CUM)	Q^2(CUM)	N	A	K	eliminated compounds
M12	0.539	0.872	0.804	80	3	60	1,7,46,56,58,63,68,76,78,79,80,82

* R^2_X(CUM) and R^2_Y(CUM) are the cumulative sum of squares of all the X and Y values, respectively, explained by all extracted principal components; Q^2(CUM) is the fraction of the total variation of the Y values that can be predicted for all the A extracted principal components in the cross validation procedure (7 rounds) used to establish the number of significant principal components, A.

Table 3
THE b_j COEFFICIENTS FROM EQ. (1) IN DECREASING ORDER OF VIP VALUES FOR MODEL M12

No	Variable ID	b_j	VIP[3]	Descriptor Classes	Descriptor significance
1	F08[O-Br]	-0.079	1.634	2D frequency fingerprints	Frequency of O - Br at topological distance 8
2	B08[O-Br]	-0.067	1.606	2D binary fingerprints	Presence/absence of O - Br at topological distance 8
3	B06[N-Br]	-0.083	1.594	2D binary fingerprints	Presence/absence of N - Br at topological distance 6
4	SRW09	-0.076	1.474	Walk and path counts	self-returning walk count of order 09
5	EEig01r	-0.069	1.384	Edge adjacency indices	Eigenvalue 01 from edge adj. matrix weighted by resonance integrals
6	F04[C-Br]	-0.062	1.230	2D frequency fingerprints	Frequency of C - Br at topological distance 4
7	Mor12v	-0.044	1.197	3D-MoRSE	3D-MoRSE – signal 12 / weighted by atomic van der Waals volumes
8	F07[C-Br]	-0.068	1.147	2D frequency fingerprints	Frequency of C - Br at topological distance 7
9	GATS3p	-0.062	1.132	2D - autocorrelations	Geary autocorrelation - lag 3 / weighted by atomic polarizabilities
10	MATS5e	-0.065	1.125	2D - autocorrelations	Moran autocorrelation-lag 5/weighted by Sanderson electronegativites

11	G1e	-0.057	1.122	WHIM	1st component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities
12	RDF100m	-0.057	1.103	Radial Distribution Function	Radial distribution function-10.0/ weighted by atomic masses
13	Mor12p	-0.039	1.102	3D-MoRSE	3D-MoRSE – signal 27 / weighted by atomic polarizabilities
14	Mor09m	-0.037	1.094	3D-MoRSE	3D-MoRSE – signal 27 / weighted by atomic masses
15	GATS7m	-0.055	1.092	2D - AUTOCORREL ATIONS	Geary autocorrelation - lag 7 / weighted by atomic masses
16	B10[N-N]	0.044	1.085	2D binary fingerprints	Presence/absence of N - N at topological distance 10
17	F10[N-N]	0.044	1.085	2D frequency fingerprints	Frequency of N-N at topological distance 10
18	Mor27m	0.052	1.048	3D-MoRSE	3D-MoRSE – signal 27 / weighted by atomic masses
19	Mor03m	0.057	1.046	3D-MoRSE	3D-MoRSE – signal 3 / weighted by atomic masses
20	Mor27v	0.059	1.040	3D-MoRSE	3D-MoRSE – signal 27 / weighted by atomic van der Waals volumes

In the analysis of the model, M12, we used as criteria the Variables Importance in the Projection (VIP) values and the sign of the coefficients (b_j). The VIP values computed by a module of the SIMCA program reflect the importance of descriptors in the PLS model with respect to Y, i.e. its correlation to all responses, and with respect to X. Variables with higher VIP scores are more relevant in explaining the activity. The X variables (here SIMCA-selected descriptors) with the highest VIP values are associated with the active part of the molecule. But, sometimes variables with high VIP scores cannot be related to the activity, because these variables X, preserving in the model, shows insignificant coefficients. The descriptors with VIP >1 (table 3) are considered the most relevant for the model. The principal components of the model (M12) explains more than 53% of the variance in the experimental activity with a good predictive power.

Interpretation

The binary/frequency *fingerprint* classes are considered structural keys in drug research. The binary fingerprint class is a binary bitstring of 1's (for the presence) and 0's (for the absence) of specific fragments and the use of hash functions allows breaking the fingerprint into smaller strings, easier to handle [25]. The beneficial effect of F10 [N-N] and B10 [N-N] descriptors are consistent with the presence of nitrogen atoms in all indirubin derivatives. These two nitrogen atoms are essential in achieving the hydrogen bonds between ligands and C=O backbone of ASP133 and VAL135 key residues [26], respectively. In the last few decades, a great number of molecular fingerprints have been presented in the literature. Atomic, atom-type and total non-stochastic quadratic indices have shown a great ability to encode chemical information, which can be used for the development of QSARs [27]. The *Molecular walk counts* descriptor classes are: MCW (Molecular walk counts), TWC (Total walk counts) and the SRW (Self-returning walk counts). Self-returning walk counts are atomic and molecular descriptors obtained from a molecular graph whose hydrogen atoms have been removed, based on Graf's path starting and ending at the same point, node, which is considered self-return path [28]. *Edge adjacency indices* descriptors are indices that show a good structural selectivity [29]. The most well known *2D-autocorrelation* descriptors, *Moran* and Geary coefficients were defined in order to reflect the contribution of a considered atomic property to the experimental observations under investigation. The Moran coefficient usually takes a value in the interval [-1, +1]. Positive autocorrelation corresponds to positive values of the coefficient whereas negative autocorrelation produces negative values. The Geary coefficient is a distance type function varying from zero to infinity. Strong autocorrelation produces low values of this index; moreover, positive autocorrelation translates into values between 0, and 1 whereas negative autocorrelation produce values larger than 1; therefore the reference "no correlation" is c = 1 [30, 31]. In the present case, analyzing the sign of regression coefficients of the participating descriptors MATS5e and GATS(3p, 7m) from this class, they contributed negatively to the activity. It suggests that the activity profile of indirubin derivatives and the identified descriptors display inverse proportionality relationship.

3D-Morse (3D-Molecule Representation of Structure based on Electron diffraction) descriptors present an advantage as they code with fixed-length representation of 3D molecular structure, allowing us the comparison of data sets comprising molecules of different size, and

Model	Training set	Test set	R^2	R_0^2	$R_0'^2$	k	k'	$\lvert R_0^2 - R_0'^2 \rvert$	$\frac{R^2_p - R_0^2}{R^2_p}$
M12	80	17	0.647	0.646	0.487	0.986	1.01	0.16	0.0015

[a] Golbraikh and Tropsha criterions [20, 21, 22] are used. The significance of R^2, R^2_{pred}, R_0^2, $R_0'^2$, k, and k' are the same as that in references [20, 21, 22]

Table 4
PREDICTIVE POWER RESULTS FOR THE EXTERNAL TEST SET[a]

Model	1	2	3	4	5	6	7	8	9	10
R^2_Y(CUM)	0.073	0.133	0.111	0.092	0.184	0.135	0.063	0.14	0.077	0.08
Q^2(CUM)	-0.07	-0.1	-0.1	-0.1	-0.1	-0.063	-0.056	-0.1	-0.1	-0.089
R^2_r	0.109									
R^2_p	0.762									

R^2_r, R^2_p see text, references [23]

Table 5
THE RESULTED VALUES OF EXPLAINED VARIATION R^2_Y(CUM) AND Q^2(CUM) FOR Y-RANDOMIZATION TESTS FOR PLS MODEL

number of atoms [32]. Besides, these descriptors 3D-MoRSE descriptors are based on the idea of obtaining information from 3D atomic coordinates by the function employed in electron diffraction studies to construct scattering curves.

WHIM (Weighted Holistic Invariant Molecular Descriptors) descriptors are built in such a way as to capture the relevant molecular 3D information regarding the molecular size, shape, symmetry, and atom distribution with respect to some invariant reference frame [33]. *RDF* (Radial distribution function) descriptors are based on the distances distribution in the geometrical representation of a molecule and constitute a radial distribution function code. The RDF descriptors are interpretable by using simple rules sets, and thus they provide a possibility for conversion of the code back into the corresponding 3D structure. Besides information about interatomic distances in the entire molecule, the RDF descriptors provides further valuable information, e.g. about bond distances, ring types, planar and non-planar systems and atom types [34]. For the case of RDF100m, the sphere radius is of 10.0Å and the atomic masses are employed to distinguish their nature.

Model Validation

External validation is an essential step in the model evaluation. The predictive abilities of our models are tested using the externally predicted R^2 value and the Golbraikh-Tropsha set of criteria (see Model Validity). In case of external validation, predictive capacity of the PLS model was judged by its application for prediction of test set activity values and calculation of predictive R^2 value.

Analyzing the external test set results listed in table 4 we can assume that all the Golbraikh-Tropsha criteria are accomplished. The $R^2 > 0.6$ condition is fulfilled by the M12 model, R_0^2 is close enough to R^2 (0.646 versus 0.647) and the slope values, k and k' are between 0.85 and 1.15.

The robustness of PLS model is confirmed by the close values of the R^2_Y(CUM) and Q^2(CUM) statistical parameters. In addition, the R_p^2 parameter value higher than 0.5 excludes the possibility of chance correlation from our PLS model. These results are listed in table 5.

Conclusions

A PLS approach has been applied for the linear modeling of chemically diverse GSK-3 inhibitors using various molecular descriptors (3D-MoRSE, GETAWAY, RDF, WHIM, 2D-autocorrelation, etc) that encompass different 2D and 3D aspects of the chemical structure. Three-dimensional quantitative structure-activity relationship study on a series of 109 maleimide derivatives as GSK-3 inhibitors yielded stable and statistically significant predictive models as indicated by moderate to high cross-correlation coefficients. The predictive ability of the PLS model was estimated from the prediction of the GSK-3 inhibitory activity of the test set of 17 compounds. The statistically significant model is based on PLS regression using the 2D and 3D descriptors based on the internal (Q^2(CUM)= 0.804) and external (R^2_Y(CUM) = 0.872) predictive power with the R^2_p parameter value higher than 0.5 (0.762).

In PLS a measure of the Xj variable importance for both modeling of X and Y is the VIP value (Variable Importance for the Projection). The descriptors with VIP >1 are the most relevant for a model. This model explains more than 53% of the variance in the experimental activity with a good predictive power.

Selected molecular descriptors captures 3D information (WHIM, 3D-Morse), molecular selectivity (Edge adjacency indices), the geometric representation of the molecules, providing information about interatomic distances, topological distances, types of atoms (RDF, 2D-Fingerprints, 2D-Autocorrelation). In conclusion we can assume that molecular descriptors represent an attractive tool for efficient indirubin design process.

Acknowledgement: This project was supported by Project 1.2 of the Institute of Chemistry Timisoara of the Romanian Academy. We thank Dr. Erik Johansson (Umetrics, Sweden) for kindly providing the SIMCA package. The authors are indebted to Prof. Mircea Mracec for giving access to the Hyperchem software and to Dr. Simona Funar-Timofei for the access to the Dragon software.

References

1.VULPETTI, A., CRIVORI, P., CAMERON,A., BERTRAND, J., BRASCA, M.G., D'ALESSIO,R., PEVARELLO, P., J. Chem. Inf. Model., 45, 2005, p.1282.
2.ZHANG,H.C., YE, H., CONWAY, B.R., DERIAN, C.K., ADDO, M.F., KUO, G.H., HECKER, L.R., CROLL, D.R., LI, J., WESTOVER, L., XU, J.Z., LOOK, R., DEMAREST, K.T., ANDRADE-GORDON, P., DAMIANO B.P., MARYANOFF, B.E., Bioorg. Med. Chem. Lett., 14, 2004, p. 3245.
3.WITHERINGTON, J., BORDAS, V., GARLAND, S.L., HICKEY, D.M.B., IFE, R.J., LIDDLE, J., SAUNDERS, M., SMITH, D.G., WARD, R.W., Bioorg. Med. Chem. Lett., **13** 2003, p. 1577.
4.PEAT, A.J., GARRIDO, D., BOUCHERON, J.A., SCHWEIKER, S.L., DICKERSON, S.H., WILSON, J.R., WANG T.Y., THOMSON, S.A., Bioorg. Med. Chem. Lett., 14, 2004, p. 2127.
5.LESCOT, E., BUREAU, R., SANTOS, J.S.O., ROCHAIS, C., LISOWSKI, V., LANCELOT, J. C., RAULT, S., J. Chem. Inf. Model., 45, 2005, p. 708.
6.MARTINEZ, A., CASTRO, A., DORRONSORO, I., ALONSO, M., Medicinal Research Reviews, 22, No. 4, 2002, p. 373.
7.HOESSEL, R., LECLERC, S., ENDICOTT, J., NOBLE, M., LAWRIE, A., TUNNAH, P., LEOST, M., DAMIENS, E., MARIE, D., MARKO, D., NIEDERBERGER, E., TANG, W., EISENBRAND, G., AND MEIJER, L., Nat. Cell Biol., 1, 1999, p. 60.
8.LECLERC, S., GARNIER, M., HOESSEL, R., MARKO, D., BIBB, J.A., SNYDER, G.L., GREENGARD, P., BIERNAT, J., WU, Y.Z., MANDELKOW,

E.M., EISENBRAND, G., MEIJER, L., The Journal of Biological Chemistry, 276, No. 1(5), 2001, p. 251.
9.POLYCHRONOPOULOS, P., MAGIATIS, P., SKALTSOUNIS, A.L., MYRIANTHOPOULOS, V., MIKROS, E., TARRICONE, A., MUSACCHIO, A., ROE, S.M., PEARL, L., LEOST M., GREENGARD, P., MEIJER, L., J. Med. Chem., 47, 2004, p. 935.
10.BEAUCHARD, A., FERANDIN, Y., FRERE, S., LOZACH, O., BLAIRVACQ, M., MEIJER, L., THIE´RY, V., BESSON, T., Bioorg. Med. Chem., 14, 2006, p. 6434.
11.FERANDIN, Y., BETTAYEB, K., KRITSANIDA, M., LOZACH, O., POLYCHRONOPOULOS, P., MAGIATIS, P., SKALTSOUNIS, A.L., MEIJER, L., J. Med. Chem.**,** 49, 2006, p. 4638.
12.VOUGOGIANNOPOULOU, K., FERANDIN, Y., BETTAYEB, K., MYRIANTHOPOULOS, V., LOZACH, O., FAN, Y., JOHNSON, C. H., MAGIATIS, P., SKALTSOUNIS, A. L., MIKROS, E., MEIJER, L., J. Med. Chem., 51, 2008, p. 6421.
13.*** www.symyx.com
14.HYPERCHEM 7.52 RELEASE FOR WINDOWS; HYPERCUBE, INC., GAINESVILLE, FLORIDA, USA, http://www.hyper.com
15.DRAGON FOR WINDOWS (SOFTWARE FOR MOLECULAR DESCRIPTOR CALCULAIONS), http://www.talete.mi.it
16.TODESCHINI, R., CONSONNI, V., Molecular Descriptors for Chemoinformatics, **Vol. I & II** , Ed. WILEY – VCH, 2009
17.ERIKSSON, L., GOTTFRIES, J., JOHANSSON, E., WOLD, S., Chemom. Intel. Lab. Syst. 73, 2004, p.73.
18.SIMCA P, version 9.0; Umetrics AB: Umea, Sweden. http://www.umetrics.com.
19.DASZYKOWSKI, M., KACZMAREK, K., HEYDEN V., WALCZAK, B., Chemom. Intel .Lab. Syst., 85, 2007, p. 203.
20.GOLBRAIKH, A., TROPSHA, A., J. Comput. Aided Mol. Des., 16, 2002, p. 357.
21.GOLBRAIKH, A., TROPSHA, A., J. Comput. Aided Mol. Des., 120, 2002, p. 269.
22.GOLBRAIKH, A., SHEN, M., XIAO, Z., LEE, K.H., TROPSHA, A., J. Comput. Aided Mol. Des., 17, 2003, p. 241.
23.ROY, P.P., PAUL, S., MITRA I., ROY, K., Molecules, 14, 2009, p. 1660.
24.ERIKSSON, L., JOHANSSON, E., KETTANEH-WOLD, N., WOLD, S., SIMCA P version 9.0; User guide and tutorial. Umetrics AB: Umea, Sweden, 2001, 108, p. 191.
25.SCIABOLA, S., CAROSATI, E., CUCURULL-SANCHEZ, L., BARONI, M., MANNHOLD, R., Bioorg. Med. Chem., 15, 2007, p. 6450.
26. BROOKHAVEN PROTEIN DATA BANK (PDB); http://www.pdb.org/pdb/explore/explore.do?structureId=1q41
27.MONTERO-TORRES, A., VEGA, M.C., MARRERO-PONCE, Y., ROLÓN, M., GÓMEZ-BARRIO, A., ESCARIO, J.A., ARÁN, V.J., MARTÍNEZ-FERNÁNDEZ, A.R., MENESES-MARCEL, A., Bioorg. Med. Chem., 13, 2005, p. 6264.
28.TODESCHINI, R.; CONSONI, V. „Handbook of molecular descriptors", **Vol 11**, 2000, p. 292.
29.ESTRADA, E., RAMÍREZ, A., J. Chem. Inf. Comput. Sci., 36, No. 4, 1996, p. 837.
30.AJLOO, D., SHARIFIAN, A., BEHNIAFAR, H., Bull. Korean Chem. Soc., 29, No.10, 2008, p.2009.
31.SAÍZ-URRA, L., GONZÁLEZ, M.P., TEIJEIRA, M., Bioorg. Med. Chem., 15, 2007, p. 3565.
32.SAIZ-URRA, L., GONZALEZ, M.P., TEIJEIRA, M., Bioorg. Med. Chem., 14, 2006, p. 7347.
33.GONZÁLEZ, M.P., SUÁREZ, P.L., FALL, Y., GÓMEZ, G., Bioorg. Med. Chem. Lett., 15, 2005, p. 5165.
34.GONZÁLEZ, M.P., TERÁN , C., TEIJEIRA , M., HELGUERA, A.M., Eur. J. Med. Chem., 41, 2006, p. 56.

Manuscript received: 20.10.2011