A First QSAR Model for Galectin-3 Glycomimetic Inhibitors Based on 3D Docked Structures

Suzanne Sirois*, Denis Giguère and René Roy*

Département de Chimie, Université du Québec à Montréal (UQÀM), C.P. 8888, Succursale, Centre-Ville, Montréal Québec, Canada H3C 3P8

Abstract: This study presents the first QSAR model for Galectin-3 glycomimetic inhibitors based on docked structures to the carbohydrate recognition domain (CRD). Quantitative numerical methods such as PLS (Partial Least Squares) and ANN (Artificial Neural Networks) have been used and compared on QSAR models to establish correlations between molecular properties and binding affinity values (Kd). Training and validation of QSAR predictive models was performed on a master dataset consisting of 136 compounds. The molecular structures and binding affinities (Kd) (136 compounds) were obtained from the literature. To address the issue of dimensionality reduction, molecular descriptors were selected with PLS contingency approach, ANN, PCA (Principal Component Analysis) and GA (Genetic Algorithms) for the best predictive Galectin-3 binding affinity (Kd). Final sets comprising 56, 31 and 35 descriptors were obtained with PLS, PCA and ANN, respectively. The objective of this prototype QSAR model is to serve as a first guideline for the design of novel and potent Gal-3 selective inhibitors with emphasis on modification at both C-3' and at O-3 positions [1].

Key Words: Galectin-3, 3D-QSAR, glycomimetics, Neural-Network.

INTRODUCTION

Galectins [2] are a family of 14 protein members that constitute important targets for therapeutics development because of their newly identified role in inflammation [3], immunity [4] and cancer [5-8]. The particularity of galectin-3 (Gal-3) as compared to the other members of the galectin family is its monomeric nature in solution. Very few inhibitor design have been made until now [9] for this lectin member. Recently, three high resolution X-ray crystal structures for the human Gal-3 in complex with LacNac the natural ligand: 1A3K, 1KJL and a more potent inhibitor 1KJR [10, 11] have been resolved. These Gal-3/ligand structures are used as templates in the context of a structure-based rational approach for the design of novel classes of inhibitors. In general, galectins share similar carbohydrate recognition domains (CRDs) and affinity for small β-D-galactosides, but show significant differences in binding specificity for more complex glycoconjugates. In Gal-3 CRD, the majority of electrostatic interactions involving hydrogen bonding patterns are made between the galactoside residue, and His158, Asn160 and Arg162 (see Fig. **1**). Direct hydrogen bonding to the protein occurs mostly through galactoside O-3'axial hydroxyl group. Two residues with planar-side chains, Tryp181 and His158, are important to align the sugar correctly in the binding site. In particular, Tryp181 participates in stacking interactions with carbons C3, C4 and C5 of the galactoside residue. Also, Gal-3 X-ray crystal structures resolution revealed that three OH groups of the galactoside residue point towards an extension of the lactose/LacNac (N-acetyllactosamine) binding site [10]. This is thought to be responsible for conferring selectivity for longer oligosaccharides. Taking advantage of this particular feature, chemical modifications of LacNac derivatives were made by Sorme *et al*. [11-14] at the critical C-3'position of the galactoside moiety using a benzamide pattern template. Another class of compound, O-galactosyl aldoximes [15], have also shown moderate potency as compared to the LacNac derivatives. Following this, another class of inhibitors has been investigated based on thio-β-D-galactopyranoside template [16]. Hence, the design of small affinity galectins inhibitors is a new and exciting challenge because of its high potential of applications in improving human health. Carbohydrates are notorious for being poor drugs because of their *in vivo* hydrolysis and their failure to cross membrane due to high polarity. Synthesis of low molecular weight (MW), high affin-

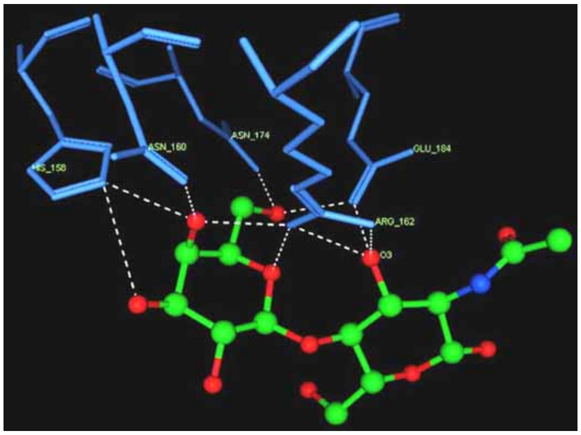


Fig. (1). Hydrogen bond networks between LacNac and Galectin-3.

*Address correspondence to these authors at Département de Chimie, Univer-sité du Québec à Montréal (UQÀM), C.P. 8888, Succursale. Centre-Ville, Montréal Québec, Canada H3C 3P8;
E-mails: sirois.suzanne@uqam.ca; roy.rene@uqam.ca

1573-4064/06 $50.00+.00

ity, and monovalent, and multivalent ligands presents an important challenge. The advantage of non-O-linked monosaccharide derivatives is that they may possess a longer half-life *in vivo,* due to lack of hydrolytically labile glycosidic bonds, and also are less polar which in turn improved cell permeability. Hence, disaccharide glycomimetics with modification of glycosidic bonds is an avenue of intense development. At the molecular level the challenge lies in the understanding of the interactions between the lectin and the saccharide through the complex networks of hydrogen-bonds as well as hydrophobic, salt bridges, and Van der Waals interactions. The first objective is to provide a rationale means to model the various types of interactions responsible for inhibitory activity of newly synthesized glycomimetic compounds against Gal-3. The amount of available molecular structures, 136 published in total to date, is sufficient to enable the development of a first QSAR Gal-3 glycomimetics theoretical model. This prototype QSAR model is currently serving as a first guideline for the design of novel and potent Gal-3 selective inhibitors [1] and new exciting and challenging work into progress.

MATERIALS AND METHODS

A typical QSAR table is composed essentially of rows representing the molecules and columns representing descriptor values. The process of building a predictive model from experimental data can be generalized as follow [17]:

1. Assemble a database of experimental results and molecular structures.
2. Optimize molecular structures in 3D space.
3. Calculate molecular descriptors for each molecule in the training set.
4. Estimate the parameters of a chosen numerical model (PLS, ANN, PCA).
5. Remove outliers from the training set
6. Assess the predictability of the model. If the model is not satisfactory, return to step 4, or add new structures if available.

Galectin-3 Glycomimetics Database: Training Set

Surface plasmon resonance experimental binding affinity Kd values of 136 Gal-3 inhibitors was obtained from litterature [11,15,16]. Each structure was constructed with the Molecular Operating Environment (MOE) [18] and then minimized on LacNac and LacNac-Bek template bound to the Gal-3 three-D structure in order to restraint the domain of conformational flexibilities to those of the complexed ligands (see Fig. **2**). (LacNac-Bek is the name of the ligand given by Sorme *et al.* in the PDB file 1KJR). This rational approach is justifiable in the context that the CRD allows only one specific type of anchorage for the galactoside residue of the carbohydrate based on the observations obtained from various galectins X-ray structures [10,11], PDB number 1KJL and 1KJR.

Molecular Descriptors Calculations

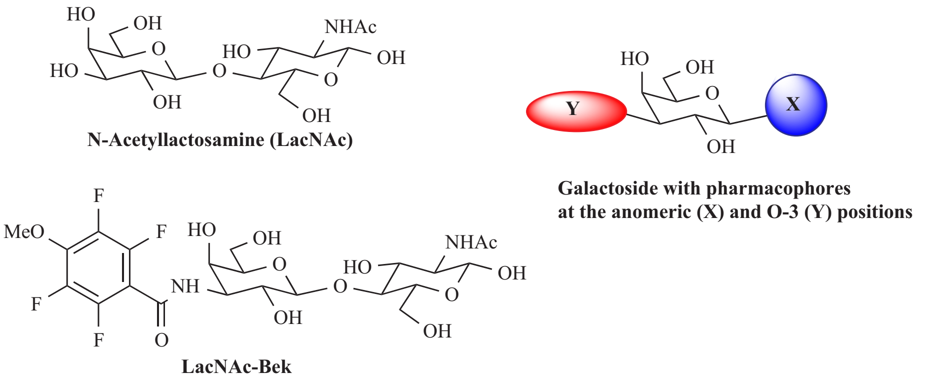
Numerical representation of molecules is described with *n*-vectors of numbers called molecular descriptors. Molecular descriptors, initially more than 155, were generated from the minimized molecular structures in Gal-3 pocket on the basis of 1D, 2D and 3D formulas. The descriptor types used to represent the molecular structure belong to the structural, topological, physical and chemical domains.

Choice of Mathematical Model

PLS: is a linear model in which the experimental result is expressed as a linear combination of the descriptors plus a constant. The parameters, or coefficients, for the model are determined in such a way that the mean squared error between the training sets experimental results and the models results is minimized. Given a matrix $\boldsymbol{X}_{\mathrm{nxp}}$ that contains *n* observations of *p* descriptors, and a matrix $\boldsymbol{Y}_{\mathrm{nxm}}$ that contains *n* observations of *m* dependent variables, the goal of PLS is to allow forming a model that will describe their common structure and allow for the prediction of ***Y*** from ***X*** (for novel observations in ***X***) [10]. Essentially, PLS tries to solve the system:

$$\boldsymbol{Y}_{\mathrm{nxm}} = \boldsymbol{X}_{\mathrm{nxp}}\boldsymbol{B}_{\mathrm{pxm}} + \boldsymbol{E}_{\mathrm{nxm}} \qquad (1)$$

Where n: number of observations, m: number of dependent variables (in our case, m=1 since we are only predicting



N-Acetyllactosamine (LacNAc)

LacNAc-Bek

Galactoside with pharmacophores at the anomeric (X) and O-3 (Y) positions

Fig. (2). Natural ligand LacNAc (1KJL, K_d = 67 μM); a potent analog modified at O-3' LacNAc-Bek (K_d = 0.88 μM); and the structure of a Galactoside scaffold with two pharmacophores X and Y at C-1 and O-3.

K_d), p: number of descriptors, ***E*** is the error residual matrix and ***B*** is the regression coefficient matrix (whose determination leads to the formation of a predictive model).

ANN: it is said that ANN "mimic" brain function [19]. ANN, as any other artificial intelligence method, inherently uses learning. This is achieved through an initial training session at the end of which internal associations - relating patterns of inputs and outputs - are built. Subsequently, ANN can make a prediction based on new inputs or "unknowns". In feed forward neural networks, the neurons are organized in the form of layers. The neurons in a layer get input from the previous layer and feed their output to the next layer. In this kind of networks connections to the neurons in the same or previous layers are not permitted. The last layer of neurons is called the output layer and the layers between the input and output layers are called the hidden layers. The input layer is made up of special input neurons, transmitting only the applied external input to their outputs. In a network if there is only the layer of input nodes and a single layer of neurons constituting the output layer then they are called single layer network. If there are one or more hidden layers, such networks are called multilayer networks (MLP). The MLP [19], in general, is the most popular network structure. It is dependent on iterative training (slow in cases) but it produces networks that are compact and fast in their execution (following training). We tested MLPs with 3 or 4 layers, featuring the hyperbolic tangent as activation function interconnecting these layers. Radial basis function [20] (RBF) networks originate from the regularization theory for solving ill-conditioned problems. They tend to perform slower than MLPs but they train fast. Conversely to the MLPs, the effectiveness of RBF is inversely proportional to the increasing number of input variables. However, inclusion of unnecessary inputs makes them more sensitive which may have an indicative value in unexploited datasets like ours. Generalized regression neural networks [21] (GRNN) train fast (when N<1000 approximately) and perform satisfactorily, yet they execute slowly. GRNNs augment the pros and cons of RBFs thus contrasting MLPs. GRNNs use Bayesian techniques to estimate the expected value of an output variable dependent on a given input.

Selection of Molecular Descriptors

Within the framework of a linear QSAR model (PLS), contingency analysis was performed to assist in the selection of subsets of descriptors from a set of more than 155 descriptors to complement the 38-VSA descriptors [18]. The VSA set of descriptors is a Subdivided Surface Areas type descriptors based on an approximate accessible van der Waals surface area. QSAR-contingency performs a bivariate contingency analysis for each descriptor and the activity or property value [18]. Contingency analysis attempts to measure the degree to which two random variables are dependent.

QSAR Models

Various QSAR models are obtained from suggested descriptors from an initial contingency analysis. For the relative importance of each descriptor, it is obtained from the absolute values of the normalized coefficients divided by the absolute value of the largest normalized coefficient. The quality of the linear-based model is assessed with r^2: correlation coefficient, and RMSE: root mean square error.

Validation of each tested model is made with:

- $PRED a value of the model,
- $RES the difference between the value of the model and the activity field (log (Kd)),
- $Z-SCORE the absolute difference between the value of the model and the activity field (log (Kd)), divided by the square root of the mean square error of the data set.

Cross-Prediction is obtained with:

- $XPRED the value of the model under a leave-one-out cross validation scheme.
- $XRES the difference between the value of the model under a leave-one-out cross validation scheme and the activity field.
- $XZ-SCORE the absolute difference between the value of the model under a leave-one-out cross validation scheme and the activity field (log (Kd)), divided by the square root of the mean square error of the data set.

The two Z-Score fields $Z-SCORE and $XZ-SCORE can be thought of as the number of standard deviations away from the mean and are used for outlier detection.

RESULTS AND DISCUSSION

Data Set Training Compounds

The two most challenging steps for developing a QSAR model are the selection of a set of training compounds having sufficient molecular diversity and the selection of an appropriate set of molecular descriptors describing the activity. A first molecular database of 119 compounds has been built from literature available affinity Kd data. They comprise 59 compounds with C-3'position based on a benzamide pattern template [11], 52 compounds withC-1' based on O-galactosyl aldoximes [15] and 8 compounds with thio-β-D-galactopyranoside template [16]. These compounds were synthesized with the following rationale. Replacement of N-acetyl glucosamine by a non-carbohydrate aglycon to render more drug-like with higher affinity; thioglycosides: to confer stability towards acidic and enzymatic hydrolysis; anomeric oxime ethers: to improve stability against enzymatic hydrolysis: i.e. stable at physiological pH. Then, 17 more compounds were added to the initial database and they comprised 12 triazol-1thio-galactosides [16] and 5 C2-symmetrical thio-galactoside bis-benzamido derivatives [22], for a total of 136 compounds.

Molecular Descriptors Selection

Molecular descriptors selection is intimately related with accuracy, stability and interpretability of a model. There are so many thousands descriptors available that the selection of an appropriate set is nowadays dependent on the type of mathematical model, algorithm used and class of drug target. Within that context Labute [23] has developed a set of descriptors for wide applications. These descriptors are based upon atomic contributions to Van derWaals area, logP (octa-

Table 1. Number of Descriptors and Compounds with Their RMSE and r^2 Values

#descriptors / # compounds	RMSE	r^2 Correlation coefficient	RMSE Cross-Validated	r^2 Cross-Validated
38/119	0.9736	0.91532	1.78298	0.74002
44/119	0.8872	0.92968	1.58980	0.79679
81/119	0.7378	0.95099	2.64724	0.55579
38/129	1.0435	0.897975	1.59498	0.77162
81/129	0.77437	0.94388	1.81190	0.74006
38/136	1.17578	0.881626	1.76659	0.74188
66/136	0.80366	0.94251	1.59166	0.81689
81/136	0.83729	0.96359	1.45851	0.65207
56/136Contingency	0.65408	0.96337	1.03636	0.90939
31/136 GA	0.90278	0.94399	1.07127	0.80876
17/136 PCA	0.89581	0.92098	1.10385	0.96062
35/136 GA and NN	0.8107606	0.9554546	0.3748076	0.9899365
35/136 GA and NN	0.7698428	0.9593227	0.5226707	0.9829417
35/136 GA and NN	0.5657788	0.9787421	0.7049245	0.9375967
35/136 GA and NN	1.099722	0.9268441	0.6359181	0.9626615
35/136 GA and NN	1.152485	0.9056143	0.4901789	0.9878652
35/136 GA and NN	0.7285886	0.9645584	0.8220331	0.9292236
35/136 GA and NN	0.6313888	0.9688487	0.9416063	0.9326992
35/136 GA and NN	0.7631699	0.9597963	0.5527562	0.9760612

nol/water), molar refractivity and partial charge. From these contributions three sets of descriptors have been defined: SlogP_VSA (10) , SMR_VSA (8) and PEOE_VSA (14) which captured hydrophilic and hydrophobic effects, polarisability, and electrostatic interactions, respectively. According to Labute, each of these descriptor sets is derived from the Hansch and Leo descriptors [24] and taken all together they define a 32 dimensional chemistry space. Six more VSA-types descriptors were also added which represents the partial charges Q_VSA-type.

Table **1** shows the validity of the QSAR models based initially on the 38-VSA descriptors set. QSAR PLS-based models produce very good correlation and cross-correlation coefficients but fail for cross-correlation RMSE deviation for a training set of 119 compounds with a value of 1.78298. Adding more compounds to the training dataset improve slightly cross-correlation RMSE deviation: 1.78298 versus 1.76659 for 119 and 136 compounds, respectively. This suggests that linear QSAR model based on VSA descriptors set is insufficient to describe the binding affinity of glycomimetics to their binding sites. Different domains of descriptor types needed to be explored. For this purpose, force-field types were selected after performing contingency analysis on various descriptor domains (see Table **2**). The predictive capacity of the model was improved giving a cross-validated RMSE of 1.45851. Hence, PLS-contingency method was applied to produce the best possible PLS model with RMSE and r^2 correlation coefficient and cross – correlation values of 0.65408, 0.96337, 1.03636 and 0.90939, respectively. PCA, GA, and contingency-based methods were used subsequently to identify other categories of descriptors from a set of 155 descriptors (see Table **2**). The 31 descriptors set obtained with GA produced a PLS model with RMSE and r^2 correlation coefficient and cross– correlation values of 0.90278, 0.94399 1.07127 and 0.80876, respectively. The 17 descriptors set obtained with PCA produced a model with RMSE and r^2 correlation coefficient and cross –correlation values of 0.89581, 0.92098, 1.10385 and 0.96062, respectively. For this ensemble of Gal-3 glycomimetic derivatives, PLS-contin-gency produces a very good model, followed by the GA approach and then by PCA. Regarding the molecular descriptors, force-field charge descriptors based on PEF95SAC force field [25] were amongst the highest ranked. It is important to mention that this force field has been specifically developed to model carbohydrates

Table 2. Sets of Type of Descriptors Evaluated and the Total Number of Descriptors

Descriptor type sets	Number of descriptors
PEOE	19
SLOGP	11
SMR	8
Total	38
FF	18
PM3	7
Vsa	3
Total	66
Q_VSA	5
dipole	4
vdw	2
diameter	1
radius	1
bpol	1
a_IC	1
mr	1
a_hyd	1
rgyr	1
ASA	1
TPSA	1
Total	86
Inductive[49]	50
ASA_	4
Lipinski	2
SMB	5
directional	3
various	5
Total	155

and alcohols. The *ab initio* charges used in the force field are quite similar to those used in most established water potentials. PEF95SAC is based on Consistent Force Field (CFF) optimized potential energy parameters for alcohols and most naturally occurring carbohydrates and has been applied to and tested on β-lactose. Hence, descriptors selection involving charge components is very dependent on the choice of the force field made. Traditionally, QSAR statistical methods - like PLS and Principal Component Analysis (PCA) - are used in drug design and optimization. PLS methods are limited in describing the inherent non-linearity between the descriptors because they are of a reductionism nature and do not take into account inter- correlations.

ANN/GA

One method that has gained popularity recently is the one based on GA and ANN. For this purpose, a combine ANN/GA methodology was applied to both the selection of a reduced minimal set of descriptors from the 56 suggested by the PLS-contingency analysis and to the development of an improved model. GA/ANN suggested a subset of 35 descriptors, which has been evaluated on 10 random split sets of the master dataset of 136 compounds consisting of 110/26 training/validating. To address the non-linear nature of data, we tested different structures of ANN i.e. multi layer perceptron (MLP) with 3 or 4 layers, radial basis function (RBF), and generalized regression neural networks (GRNN). Often a number of variables may carry - to some extent - the same information as other variables. This problem is known as multi-co linearity and the only remedy known is to decrease the dimensionality of the problem in question by selecting only the most significant variables in relation to the predicted outcome. Genetic algorithms and neural networks used in combination to identify the most significant descriptors of the compounds in the training and validating split datasets have demonstrated its strength as compared to linear approach such as PLS described above. As anticipated, combination applications of GA and ANN on the split datasets suggested a ranking of significance for each one of the chemical descriptors (see Table **3**).

This study's objective is to identify and select a set of molecular descriptors for a universal glycomimetic class of Gal-3 inhibitors rather than focusing on a particular subclass at a particular position modification. The notion of relevance to receptor affinity of a collection of descriptors is difficult to quantify and our analysis is based on the ranking and the kind of descriptors suggested by ANN. Suggested descriptors belong to the original 38-VSA types in combination with others and they are describing: **1)** physical properties such as heat of formation (kcal/mol) and total energy calculated using the PM3 Hamiltonian (PM3_HF and PM3-E), log of the octanol/water partition coefficient (log (o/w): logP), polar surface area calculated using group contributions to approximate the polar surface area from connection table information only (TPSA) **2)** surface area descriptors based on force field charges (FF_VSA), volume and shape descriptors such as water accessible surface area calculated using a radius of 1.4 A for the water molecule (ASA), Van der Waals volume (VDW_VOL) **3)** distance matrix descriptors such as molecular diameter and radius **4)** pharmacophore atom type descriptors such as number of hydrophobic atoms (A_HYD) **5)** finally, the sum of hardness of all atoms in molecule (SUM_HARD).

Direct Electrostatic Contributions

Recent work [31] has suggested that the underlying atomic contributions to partial charge, molar refractivity and logP are relevant to receptor affinity. The PEOE_VSA (Partial Equalization of Orbital Electronegativities [26]) are intended to capture direct electrostatic interactions. The PEOE

Table 3. Descriptors Description [18] and Their Relative Ranking According to Their Contribution

Abbreviation	Descriptors Description	Ranking		
		Average	Min	Max
PEOE_VSA_+3	Partial Equalization of Orbital Electronegativities. Sum of v_i where q_i is in the range [0.15,0.20)	2,2	1	3
SLOGP_VSA2	Sum of v_i such that L_i is in (-0.2,0]	3,6	2	7
SLOGP_VSA0	Sum of v_i such that L_i <= -0.4	3,8	1	12
FF_VSA-4	Binned VDW surface area descriptors based on forcefield charges. Total negative 4 vdw surface area	7,2	2	15
DIAMETER	Molecular diameter	7,4	1	21
A_HYD	Number of hydrophobic atoms	8,8	7	10
PEOE_VSA+4	Partial Equalization of Orbital Electronegativities Sum of v_i where q_i is in the range [0.20,0.25).	13	5	23
PM3_HF	The heat of formation (kcal/mol) calculated using the PM3 Hamiltonian [MOPAC]	13,8	5	27
FF_VSA_HYD	Binned VDW surface area descriptors based on forcefield charges. Total hydrophobic vdw surface area	14,6	9	25
PEOE_VSA-3	Partial Equalization of Orbital Electronegativities Sum of v_i where q_i is in the range [-0.20,-0.15)	14,8	10	20
SMR_VSA0	Sum of v_i such that R_i is in [0,0.11]	15	5	27
SMR_VSA3	Sum of v_i such that R_i is in (0.35,0.39]	15,2	3	26
FF_VSA_NEG	Binned VDW surface area descriptors based on forcefield charges. Total negative vdw surface area.	15,4	7	27
ASA	Water accessible surface area calculated using a radius of 1.4 A for the water molecule.	16	8	31
VDW_VOL	Van der Waals volume	16,2	4	35
SMR	Molecular refractivity	17,4	8	27
PEOE_VSA_POL	Partial Equalization of Orbital Electronegativities	18,4	7	34
RADIUS	Molecular radius	18,4	12	28
SUM_HRD	Sum of hardness of all atoms in molecule	18,6	6	33
SMPSSGMM	Sum of all positive Sigma (mol->atom) in molecule	20	7	28
LOGP_O_W	Log of the octanol/water partition coefficient	20	4	29
SMR_VSA1	Sum of v_i such that R_i is in (0.11,0.26]	20,6	9	32
Q_VSA_POL	Total polar van der Waals surface area	20,6	12	26
FF_VSA+1	Binned VDW surface area descriptors based on forcefield charges. Total positive 1 vdw surface area'	21,6	12	30
SLOGP_VSA4	Sum of v_i such that L_i is in (0.1,0.15]	22	15	29
PEOE_VSA_+2	Partial Equalization of Orbital Electronegativities Sum of v_i where q_i is in the range [0.10,0.15)	22,6	13	32
PEOE_VSA_+1	Partial Equalization of Orbital Electronegativities Sum of v_i where q_i is in the range [0.05,0.10)	24,2	14	32
SLOGP_VSA_8	Sum of v_i such that L_i is in (0.30,0.40]	24,4	15	33
FF_VSA_POL	Total polar vdw surface area based on forcefield charges	25,6	14	32
TPSA	Total polar surface area	26,4	17	35
FF_VSA_-3	Total negative 3 vdw surface area	26,6	18	35
PM3_E	The total energy (kcal/mol) calculated using the PM3 Hamiltonian [MOPAC].	26,6	19	34
SLOGP_VSA_9	Sum of v_i such that L_i > 0.40	29	21	34
Q_VSA_POS	Total polar van der Waals positive	29,4	19	35
FF_VSA_+2	Total positive 2 vdw surface area	30,6	24	35

method of calculating charges is an iterative method in which charge is transferred between bonded atoms until equilibrium. Although the notion of partial charges is widely used it is an intellectual concept because it depends on the model and the manners the whole electron distribution in a molecule is divided. What is important to see at this point is the contribution of the charge distribution within the context of glycomimetics. For this, PEOE and FF types are important descriptors, which encapsulate charge distribution, and this is directly related to enthalpic energy contributions.

Hydrophobicity (Lipophilicity)

Hydrophobicity and water solubility are properties which are used as early as ADME screens [27,28] to reject probable development failures early on stage. Study findings suggested that hydrophobicity is an important contributor that controls the binding activity of Gal-3 inhibitors. Non-covalent interactions such as lipophilicity and shape of the molecule account for most of the activity against Gal-3. In drug design the thermodynamic binding of a drug to its therapeutic target includes free energy of binding, enthalpy and entropy. Enthalpy contribution is expressed through the drug-target interaction relative to the solvent. The primary contribution comes from hydrogen bonding and Van der Waals interactions. Entropy contribution is primarily due to hydrophobic interactions caused by an increase in the solvent entropy from burial of hydrophobic groups of the drugs and by release of water molecules upon binding and also from a small loss of conformational degrees of freedom of the candidate molecule. A drug with favorable entropy indicates that the binding is driven by hydrophobic interactions and low hydrogen bound formation. This type of drugs is hydrophobic and poorly water-soluble and is also conformationally restraint. This entails that they lack a potential of adaptability and consequently are highly susceptible to cause drug resistance [29] and side effects [17].

Binding Affinity Prediction

Electrostatic interactions such as hydrogen bonding between the Gal-3 and its natural ligand lactose occur between the O-3 hydroxyl group and surrounding groups Arg 162 and Glu184 (Fig. **3**): and the endocyclic oxygen of galactoside moiety. These interactions create a network of three salt bridges, which are a common feature in carbohydrate-lectin recognitions. Semi-empirical calculations correlate the modifications done at the anomeric position with the charge density on the O-3 oxygen [1].

DISCUSSION

Intuitively, molecular entities should possess specific features, which classify them into drugs or non-drugs [30]. First and foremost, these features have to be specific to their target. Thus, drugs are subdivided into antiviral, antiretroviral, antibiotics, antineoplastics, etc. Secondly, the structural, physical and chemical features of a particular class of compounds should be related to its biological activity as well as the binding interaction, which encompasses molecular recognition between the ligand and the receptor. Characterizing a class of molecule to a specific activity is another major challenge in drug identification and optimization [31]. The pharmaceutical industry routinely uses classical approaches based on: 1) High Throughput Screening (HTS) for identifying new class of compound [32,33] and 2) Linear Quantitative Structure Activity relationship (QSAR) techniques for optimizing lead candidates [24,28,34-36] in drug discovery and development, and to analyze data sets of compounds. It has also been helpful in understanding chemical–biological interactions in the drug-design process. It has also been utilized for the evaluation of ADMET phenomena in many organisms and whole animal studies [24,28,34-36]. Most commonly used QSAR techniques are primarily linear methods that correlate the changes at a specific point on a small molecular structure. This linear approach is limited because non-linear effects within the overall molecule potency is not considered. Furthermore, these methods are of a reductionism nature because interrelations between the various types of descriptors are excluded from the analysis. For a candidate molecule to be considered as an inhibitor for a specific target it has to attain a threshold experimental value in the range of the nanomolar (nM) concentration. However, the structural, chemical and physical characteristics leading to this threshold value have to be defined. Several characteristics can be considered individually as per the classical QSAR approach [17,34,37,38]. However, the synergistic effect of these characteristics is greater than that of their individual contribution. The complexity of these interrelations differentiates a potent inhibitor versus a non-potent one. Unfortunately, recent discoveries are not always leading to the creation of more affordable and safer drugs for patients. Moreover, new therapeutic development processes are becoming increasingly challenging, complex and costly. In particular, we aim to address one of the recent concern on the pipeline

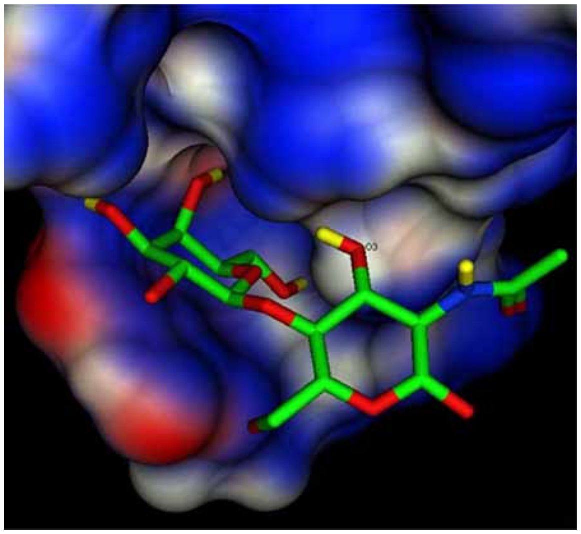


Fig. (3). Connolly surface using a space grid of 0.75 colored by Active Lone Pair showing the CRD pocket and O-3 into a hydrogen-bonding network. The bleu regions are hydrophobic, the red are hydrophilic, while the white represent regions through which hydrogen bonds (hydrogen atoms colored yellow) are likely to form for novel Galectin-3 inhibitors.

problem expressed in the FDA report [39], i.e. the causes of recent slowdown, instead of expected acceleration, in innovative medical therapies reaching patients. The approach that is taken therein find its justification on the Critical Path Initiative (CPI) put in place recently by the FDA which demands a rationale approach for a fast an accurate means of predicting the biological properties of small molecules to be developed since prototyping is both expensive and time consuming [39]. The price tag for the development of a new drug can goes as high as $800 million. It follows that green chemistry is becoming more and more a principle for the twenty first century. Within this green context approach in mind, compounds are design and tested virtually before being synthesized. This rational approach not only permit economy of time and money but more importantly the avoidance of the uses of chemical materials that will need to be wasted eventually or recycled [40]. Structural bioinformatics has been applied to timely derive the 3D structures of some functionally important proteins, helping to understand their action mechanisms and stimulating the course of drug discovery [40-48]. Thus, a rational approach including computational chemistry, structural bioinformatics, and cheminformatics encompasses designing, synthesising and testing, instead of synthesizing, testing and designing. For this purpose hybrid QSAR-ANN models for galectin-3 glycomimetics inhibitors have been developed and presented therein to predict the complex inhibition activity of new molecules that belong to this particular class of compounds.

CONCLUSION

Study findings suggested that PEOE and FF types are important descriptors, which encapsulate charge distribution that is directly, related to enthalpic energy contributions and also hydrophobicity is an important contributor that controls the binding activity of Gal-3 inhibitors. Non-covalent interactions such as lipophilicity and shape of the molecule account for most of the activity against Gal-3. High affinity and selectivity, synthetic accessibility, no chemically reactive group, oral bioavailability, favorable pharmacokinetics, metabolism, elimination pathway, lack of side effects, lack of toxic effects are currently considered in upstream of the drug development process of novel glycomimetics Gal-3 inhibitors with emphasis on modification at both C-3' and at O-3 positions [1] and continuing work in progress. By combining chemistry based enumeration, filtering and structure based evaluation we plan to go rapidly from a chemical hit structure with a known synthetic route pathway to a model of small molecule targeted library into a protein active site. With this approach chemists are able to select a set of targeted compounds to being synthesized and tested further in biological assays. This process is repeated until compounds with a satisfactory ADMET property and activity profiles are obtained. The platform that we are currently developing should allow the organisation of data from various domains and also the possibility to integrate the data from various parts of the platform through the development of hybrid QSAR and artificial intelligence. In the future we plan to develop novel QSAR models from publicly available screens that will serve as benchmarks for various diseases such as SRAS, influenza, cystic fibrosis, and cancer.

ACKNOWLEDGEMENTS

`This work was supported from Natural Sciences and Engineering Research Counsel of Canada (NSERC) and Canadian Research Chair Therapeutic Chemistry to RR.

REFERENCES

[1] Giguère, D.; Sato, S.; St-Pierre, C.; Sirois, S.; Roy, R. *Bioorg. Med. Chem. Lett.,* **2006**, *16*, 1668.

[2] Barondes, S. H.; Castronovo, V.; Cooper, D. N.; Cummings, R. D.; Drickamer, K.; Feizi, T.; Gitt, M. A.; Hirabayashi, J.; Hughes, C.; Kasai, K.; Leffler, H.; Liu, F.-T.; Lotan, R. M.; Monsigny, A.M.; Pillai, S.; Poirer, F.; Raz, A.; Rigby, P.W.J.; Rini, J..M.; Wang, J.L. *Cell,* **1994**, *76*, 597.

[3] Almkvist, J.; Karlsson, A. *Glycoconj. J.,* **2004**, *19*, 575..

[4] Sato, S.; Nieminen, J. *Glycoconj. J.,* **2004**, *19*, 583.

[5] Rabinovich, G. A.; Baum, L. G.; Tinari, N.; Paganelli, R.; Natoli, C.; Liu, F. T.; Iacobelli, S. *Trends Immunol.,* **2002**, *23*, 313.

[6] Nakahara, S.; Oka, N.; Raz, A. *Apoptosis,* **2005**, *10*, 267.

[7] Rabinovich, G. A.; Cumashi, A.; Bianco, G. A.; Ciavardelli, D.; Iurisci, I.; D'Egidio, M.; Piccolo, E.; Tinari, N.; Nifantiev, N.; Iacobelli, S. *Glycobiology,* **2005**.

[8] Hsu, D. K.; Liu, F. T. *Glycoconj. J.,* **2004**, *19*, 507.

[9] Morris, S.; Ahmad, N.; Andre, S.; Kaltner, H.; Gabius, H. J.; Brenowitz, M.; Brewer, F. *Glycobiology,* **2004**, *14*, 293.

[10] Seetharaman, J.; Kanigsberg, A.; Slaaby, R.; Leffler, H.; Barondes, S. H.; Rini, J. M. *J. Biol. Chem.,* **1998**, *273*, 13047.

[11] Sorme, P.; Arnoux, P.; Kahl-Knutsson, B.; Leffler, H.; Rini, J. M.; Nilsson, U. J. *J Am. Chem. Soc.,* **2005**, *127*, 1737.

[12] Sorme, P.; Qian, Y.; Nyholm, P. G.; Leffler, H.; Nilsson, U. J. *Chembiochem.,* **2002**, *3*, 183.

[13] Sorme, P.; Kahl-Knutsson, B.; Huflejt, M.; Nilsson, U. J.; Leffler, H. *Anal. Biochem.,* **2004**, *334*, 36..

[14] Sorme, P.; Kahl-Knutson, B.; Wellmar, U.; Nilsson, U. J.; Leffler, H. *Methods Enzymol.,* **2003**, *362*, 504.

[15] Tejler, J.; Leffler, H.; Nilsson, U. J. *Bioorg. Med. Chem. Lett.,* **2005**, *15*, 2343.

[16] Cumpstey, I.; Carlsson, S.; Leffler, H.; Nilsson, U. J. *Org. Biomol. Chem.,* **2005**, *3*, 1922.

[17] Sirois, S.; Wei, D. Q.; Tsoukas, C. M.; Chou, K.-C.; Hatzakis, G. E. *Med. Chem.,* **2005**, *1*, 173.

[18] MOE, Molecular Operating Environment version 2005.04, CCG Chemical Computing Group, Montréal, QC, Canada H3A 2R7.

[19] Lippmann, R. P.; Shahian, D. M. *Ann. Thorac. Surg.,* **1997**, *63*, 1635.

[20] Bang, S. Y.; Hwang, Y. S. *Neural. Netw,* **1997**, *10*, 1495-1503.

[21] Mosier, P. D.; Jurs, P. C. *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 1460-70.

[22] Cumpstey, I.; Sundin, A.; Leffler, H.; Nilsson, U. J. *Angew. Chem. Int. Ed. Engl.,* **2005**, *44*, 5110.

[23] Labute, P. *J. Mol. Graph. Model.,* **2000**, *18*, 464.

[24] Leo, A.; Hansch, C.; Church, C. *J. Med. Chem.,* **1969**, *12*, 766.

[25] Fabricius, J.; Englesen, S. B.; Rasmussen, K. *J. Carbo. Chem.,* **1997**, *16*, 751.

[26] Gasteiger, M.; Marsili, J. *Tetrahedron,* **1980**, *36*, 3219.

[27] Li, A. P. *Drug Discov. Today,* **2001**, *6*, 357.

[28] Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. *Bioorg. Med. Chem.,* **2004**, *12*, 3391.

[29] Ohtaka, H.; Velazquez-Campoy, A.; Xie, D.; Freire, E. *Protein Sci.,* **2002**, *11*, 1908.

[30] Sirois, S.; Hatzakis, G.; Wei, D.; Du, Q.; Chou, K. C. *Comput. Biol. Chem.,* **2005**, *29*, 55.

[31] Burke, M. D.; Berger, E. M.; Schreiber, S. L. *Science,* **2003**, *302*, 613.

[32] Blanchard, J. E.; Elowe, N. H.; Huitema, C.; Fortin, P. D.; Cechetto, J. D.; Eltis, L. D.; Brown, E. D. *Chem. Biol.,* **2004**, *11*, 1445.

[33] Kevorkov, D.; Makarenkov, V. *J. Biomol. Screen.,* **2005**, *10*, 557.

[34] Verma, R. P.; Hansch, C. *Bioorg. Med. Chem.,* **2005**.

[35] Hansch, C.; Steinmetz, W. E.; Leo, A. J.; Mekapati, S. B.; Kurup, A.; Hoekman, D. *J. Chem. Inf. Comput. Sci*., **2003**, *43*, 120.

[36] Hansch, C.; Hoekman, D.; Leo, A.; Weininger, D.; Selassie, C. D. *Chem. Rev.,* **2002**, *102*, 783.

[37] Verma, R. P.; Mekapati, S. B.; Kurup, A.; Hansch, C. *Bioorg. Med. Chem.,* **2005**, *13*, 5508.

[38] Hansch, C.; Verma, R. P.; Kurup, A.; Mekapati, S. B. *Bioorg. Med. Chem. Lett.,* **2005**, *15*, 2149.

[39] CPI, Critical Path Initiative http://www.fda.gov/oc/initiatives/criticalpath/.

[40] Wei, D. Q.; Zhang, R.; Du, Q.-S.; Gao, W.-N.; Li, Y.; Gao, H.; Wang, S.-Q.; Zhang, X.; Li, A.-X.; Sirois, S.; Chou, K.-C. *Amino Acids,* **2006**, 1-8.

[41] Sirois, S.; Sing, T.; Chou, K. C. *Protein Pept. Sci.,* **2005**, *6*, 413.

[42] Sirois, S.; Wei, D. Q.; Du, Q.; Chou, K. C. *J Chem. Inf. Comput. Sci*., **2004**, *44*, 1111.

[43] Wei, D. Q.; Du, Q. S.; Sun, H.; Chou, K. C. *Biochem. Biophys. Res. Commun.,* **2006**.

[44] Chou, K. C. *Curr. Med .Chem.,* **2004**, *11*, 2105.

[45] Chou, K. C.; Wei, D. Q.; Zhong, W. Z. *Biochem. Biophys. Res. Commun.,* **2003**, *308*, 148-51.

[46] Du, Q.; Wang, S.; Wei, D.; Sirois, S.; Chou, K. C. *Anal. Biochem.,* **2005**, *337*, 262-70.

[47] Du, Q. S.; Wang, S. Q.; Zhu, Y.; Wei, D. Q.; Guo, H.; Sirois, S.; Chou, K. C. *Peptides,* **2004**, *25*, 1857-64.

[48] Chou, K. C. *Biochem. Biophys. Res. Commun.,* **2004**, *319*, 433-8.

[49] Cherkasov, A.; Shi, Z.; Fallahi, M.; Hammond, G. L. *J. Med. Chem.,* **2005**, *48*, 3203-13.

Received: 19 December, 2005 **Revised: 05 May, 2006** **Accepted: 06 May, 2006**