JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

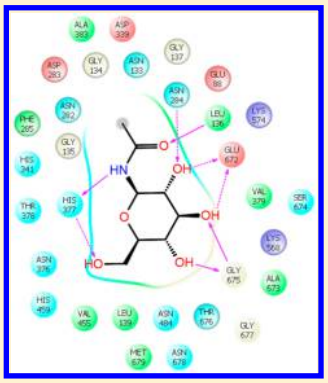
CoRILISA: A Local Similarity Based Receptor Dependent QSAR Method

Vijay M. Khedkar and Evans C. Coutinho*

Department of Pharmaceutical Chemistry, Bombay College of Pharmacy, Kalina, Santacruz (E), Mumbai 400098, India

Supporting Information

ABSTRACT: Molecular similarity methods have played a crucial role in the success of structure-based and computer-assisted drug design. However, with the exception of CoMSIA, the current approaches for estimating molecular similarity yield a global picture thereby providing limited information about the local spatial molecular features responsible for the variation of activity with the 3D structure. Application of molecular similarity measures, each related to the *functional "pieces"* of a ligand−receptor complex, is advantageous over a composite molecular similarity alone and will provide more insights to rationally interpret the activity based on the receptor and ligand structural features. Building on the ideas of our previously published methodologies—CoRIA and LISA, we present here a local molecular similarity based receptor dependent QSAR method termed CoRILISA which is a hybrid of the two approaches. The method improves on previous techniques by inclusion of receptor attributes for the calculation and comparison of similarity between molecules. For validation studies, the CoRILISA methodology was applied on three large and diverse data sets—glycogen phosphorylase *b* (GP*b*), human immunodeficiency virus-1 protease (HIV PR), and cyclin dependent kinase 2 (CDK2) inhibitors. The statistics of the CoRILISA models were benchmarked against the standard CoRIA approach and with other published approaches. The CoRILISA models were found to be significantly better, especially in terms of the predictivity for the test set. CoRILISA is able to identify the thermodynamic properties associated with residues that define the active site and modulate the variation in the activity of the molecules. It is a useful tool in the fragment-based drug discovery approach for ligand activity prediction.



INTRODUCTION

Molecular similarity is a tool for rationalizing the pharmacological activities of potential drug candidates. It is now being extensively used in the selection of analogs of chemicals, for estimation of molecular properties, to search compound libraries for analogous compounds and so on.[1−4] However, so far only a few reports on the relationships between molecular similarities and biological activities have appeared. The interest in quantification of the similarity between molecules arises from the anticipation that molecules with similar structural features will exhibit similar physicochemical properties and thus biological activities. In recent years, the concept of molecular similarity has become progressively more quantified in chemistry with the help of improved statistical techniques. However, the basic principle underlying the similarity-based QSAR was enunciated explicitly by Johnson and Maggiora, who stated that "molecules that are structurally similar likely will have similar properties".[5] Therefore, the activity of a new molecule can be predicted by taking into account the similarity values between the molecule under study and the molecules of a data set whose activities are known. The similarity between a pair of molecules is estimated using either distance or angular information to characterize the level of resemblance between the molecules or based on the overlap of analogous fields on a three-dimensional (3D) grid, using various 3D-molecular similarity indices. Each similarity method is user defined, and its efficacy relies on the set of descriptors used to define the intermolecular similarity as well as on the mathematical function used to quantify similarity. In 1991, Rum et al. used the $N \times N$ similarity matrices (N is the number of compounds) derived from 2D topological descriptors as the input vector to correlate with the biological activity using stepwise regression.[6] Two years later, Good et al. did a systematic investigation to correlate 3D electronic similarities of molecules (using electrostatic potential similarity indices) with their biological activities using neural network or PLS regression.[2,7] Seri-Levy et al. used shape similarity as a single independent variable in QSAR equations and demonstrated higher predictive abilities for their approach than for multivariate analyses.[8] However, they restricted their method to homologous series of compounds. Ghuloum et al. used molecular hash keys based on molecular surface similarity.[9]

Comparative Molecular Field Analysis (CoMFA) is one of the most notable methods of analyzing 3D-molecular similarity where the shape information on the molecular field is indirectly coded as attribute index numbers signifying the value of the field at the sampled grid points. Comparative Molecular Similarity Index Analysis (CoMSIA) on the other hand uses a SEAL-based molecular similarity index derived from steric,

Received: October 23, 2014
Published: December 23, 2014

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

electrostatic, hydrophobic, and hydrogen-bond acceptor and donor fields as 3D descriptors for QSAR analysis. Comparative Molecular Moment Analysis (CoMMA) compares the lower-order moments of the molecular mass and the charge distribution between the molecules.[10] Molecular quantum similarity measure (MQSM) uses the first-order molecular density function of two molecules as a measure of their similarity.[11,12] In the 4D-QSAR method developed by Hopfinger and co-workers, the molecular similarity is measured as a function of conformation, alignment, and atom type and has been used to study chiral and isosteric molecules as well as for the identification of common pharmacophores.[13] However, with the exception of CoMSIA, the current approaches for estimating molecular similarity yield measures that consider the entire molecule (global approach) thereby providing very limited information about the local spatial molecular features responsible for the variation of activity with the 3D structure. Therefore, in order to address the issue of local similarity and its influence on activity, we recently reported a 3D-QSAR formalism termed Local Indices for Similarity Analysis (LISA) wherein the molecular similarity is investigated by dissecting the global molecular similarity into local values.[14] This approach was adopted in the QSAR method, HomoSAR, being applied to peptides.[15] Hopfinger's 4D-QSAR method[13] considers molecular similarity with respect to the whole molecule (global) as well as functional segments of the molecule (local). However, these techniques are limited by the fact that the wealth of valuable information available in the target receptor is completely ignored. Prediction of biological activity of new molecules against a specific target protein is of paramount importance in drug discovery. Therefore, application of molecular similarity measures, each related to the *functional "pieces"* of a ligand−receptor complex will be more advantageous than a single composite measure of ligand molecular similarity alone and will provide greater insight into rationally interpret the activity based on the receptor and ligand structural features. Also the advantages of imperiling ligand−receptor interaction energies to statistical analysis over the classical approach involving molecular mechanics computation of binding free energies are that the noise due to inaccuracies in the potential energy functions and the respective molecular models can be reduced, and mechanistically relevant interaction terms identified.

With this motive, we introduce herein a new *local molecular similarity based receptor dependent* QSAR methodology termed CoRILISA (derived from CoRIA and LISA) by extending the current receptor-independent LISA formalism to regard the 3D-structures of small molecules in complex with their macromolecular targets. The resultant receptor dependent approach employs thermodynamic elements involving the ligand and the individual active site residues to compute the molecular similarity index against a reference molecule in a CoRIA-like approach. These molecular similarity measures are then corelated with the biological activity via regression analysis to identify rational, site-specific alteration to the molecules for improving their activity. The CoRILISA methodology has been examined and validated by three data sets—inhibitors of glycogen phosphorylase *b* (GP*b*), human immunodeficiency virus-1(HIV) protease, and cyclin dependent kinase 2.

METHODOLOGY

Theory. As stated earlier, the fundamental principle underlying any similarity-based QSAR formalism was enunciated by Johnson and Maggiora that structures which display substantial similarity will have considerable similarity in their physical, chemical and biological properties.[5] It thus postulates a direct relation between the structure of a compound and its putative pharmacological activity. In the LISA formalism, molecular similarity was investigated by dissecting the overall molecular similarity into local values at discrete points in 3D space surrounding the molecule. The local similarity at a given point on the 3D grid surrounding the molecule is calculated as the "potential" at that grid point of a molecule compared to a reference molecule in the data set. The local similarity indices calculated at all grid points serve as the independent variables which are then correlated with the biological activity. It has been realized that a wealth of information is closely associated with the presence/absence of certain thermodynamic interactions between the ligand and specific residues in the surrounding active site. Therefore, it can be assumed that extensive decomposition of ligand−receptor interactions allows those components that are predictive of binding free energy and hence biological activity to be detected. However, LISA being a ligand-based approach does not include receptor information in its formalism. Comparative residue interaction analysis (CoRIA) on the other hand is a receptor dependent 3D-QSAR formalism that accounts for the complete thermodynamic events involved in ligand−receptor binding.

CoRILISA is an adoption of the LISA formalism in a receptor-based setting finding its roots in the CoRIA approach to regard the geometry of the target protein in a structural similarity measure. The idea is to break down the global molecular similarity in context of the receptor into local similarity analogous to LISA. However, in contrast to LISA, the CoRILISA approach uses active site residues as the "probes" to calculate the similarity between molecules using the thermodynamic approach of CoRIA. The local similarity of a molecule is calculated as the property (any thermodynamic property such hydrophobic, van der Waals, electrostatic, etc.) of interaction of an active site residue with a given molecule in relation to a reference molecule. Therefore, one of the requisites for the CoRILISA formalism is the selection of a reference molecule with respect to which the Local Similarity Indices (LSIs) for the remaining molecules are calculated. This according to general convention is the most active molecule in the data set. The reference molecule not only embraces the appropriate size and shape complementary to the active site but also possess the essential pharmacophoric features necessary to interact with the crucial active site residues. The primary objective of CoRILISA is to extract not only from the reference molecule but also from the remaining ones in the data set, the crucial features associated with the nature and type of interactions at the level of both the receptor and the ligand that contribute to the biological activity. These can then be utilized for the design of novel compounds and/or to optimize existing leads.

Local Similarity Index As Descriptor. A local similarity index is a kind of QSAR descriptor slightly different from the traditional physicochemical parameters.[14] These indices are derived from numerical integration and normalization of the specific thermodynamic interaction values and represent a local measure of the similarity between a pair of molecules based on their specific interaction with active site residues. In essence, any thermodynamic interaction between the ligand and the active site residue can be used to compute the molecular similarity index. The LSI calculated using the Petke's formula divides the region surrounding the molecules into "equivalent",

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

"favored/similar", and "disfavored/dissimilar" potentials against a reference molecule in the data set.

The LSI between molecules as defined by the residue surrounding the ligand is calculated by adaption of the formula reported by Petke in 1993[16] for calculating 3D-molecular similarity:

$$\mathrm{LSI}_{ri} = \frac{2P_{r_i}^{\mathrm{ref}}P_{r_i}^{\mathrm{test}}}{((P_{r_i}^{\mathrm{ref}})^2 + (P_{r_i}^{\mathrm{test}})^2)} \quad (1)$$

where LSI_{ri} is the local similarity index based on residue r at point i on the receptor; $P_{r_i}^{\mathrm{ref}}$ is any thermodynamic property (P) between the residue r at point i on the receptor and the reference molecule; $P_{r_i}^{\mathrm{test}}$ is any thermodynamic property (P) between the residue r at point i on the receptor and the test molecule. The term $2P_{r_i}^{\mathrm{ref}}P_{r_i}^{\mathrm{test}}$ in eq 1 represents the overlap of the property P of molecules (reference and test), while the denominator $((P_{r_i}^{\mathrm{ref}})^2 + (P_{r_i}^{\mathrm{test}})^2)$ is the normalization factor.

The LSI provides an attractive approach for obtaining quantitative measures of molecular similarity based upon the ligand as well as receptor attributes and has the advantage that its ability to explain activity can be evaluated quantitatively using QSAR techniques. This LSI defines similarity wrt discrete points (i) in the receptor space while the overall molecular similarity can be obtained as the ratio of the sum of similarities over all residues in the receptor to the total number of residues considered, thereby leading to an average similarity over the measured space. A typical data set gives LSI values that range from 1 (completely similar molecules) to −1 (completely dissimilar molecules) because of the normalization factor, with the LSI values of the reference molecule equating to unity at all residues surrounding the ligand. Unlike the classical 3D-QSAR approaches, CoRILISA does not rely on the selection of an artificial probe (like H^+ for electrostatic interactions and CH_3 for steric interactions with some energy cutoff) since every residue in the receptor pocket serves as a probe for the computation of the similarity index.

Elements for Measures of Similarity. The thermodynamics of interaction between each amino acid of the receptor and the ligand are the key elements for the assessment of similarity in the CoRILISA formalism. The thermodynamic interactions that form the basis of similarity comparisons are discussed below.

The **nonbonded interactions** are primarily the van der Waals and Coulombic energies between the ligand and receptor and appear as a consequence of their proximity. These interactions are calculated using the OPLS 2005 force field incorporated in the Schrödinger molecular modeling suite.[17−22] The functional forms of these nonbonded interaction energy terms are expressed as

$$E_{\mathrm{nonbonded}} = \sum_{i>j} f_{ij}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} + \frac{q_i q_j e^2}{4\pi\varepsilon_0 r_{ij}}\right) \quad (2)$$

where $A_{ij} = \sqrt{A_{ii}A_{ij}}$ and $C_{ij} = \sqrt{C_{ii}C_{ij}}$

where i and j are atoms of the protein and the ligand, respectively; A_{ij} and C_{ij} are the repulsive and attractive coefficients for interacting atoms i and j, respectively; r_{ij} is the distance between atoms i and j; q_i and q_j are the charges of atoms i and j, respectively, while ε_o is the dielectric constant.

Another term contributing to the ligand−receptor binding process is the **hydrophobic interaction** between the ligand and receptor. This is a complex process primarily governed by the entropic effects associated with the change in orientation of solvent molecules in the solvation shell wrapping the solute molecules as well as from the bulk form of solvent molecules. The quantified values for the residuewise hydrophobic interactions between the ligand and receptor are obtained in the form of HINT scores using the "HINT" program integrated in Sybyl (v 7.1, Tripos Inc., USA).[23,24] HINT calculates the intermolecular hydrophobic interactions as a double sum over the atoms within each component using the following equation:

$$B = \sum_{i}^{\mathrm{atoms}}\sum_{j}^{\mathrm{atoms}} b_{ij} = \sum\sum = (a_i S_i a_j S_j R_{ij} T_{ij} + r_{ij}) \quad (3)$$

where i and j are atoms of specific amino acid residue in the active site and the ligand; b_{ij} is a microinteraction constant/binding score that describes the specific interaction between atoms i and j in the complex; B is the total interaction between the active site residue and the ligand; a_i = the hydrophobic atom constant for atom i; S_i = the solvent accessible surface area for atom i; R and r = distance functions for the interaction between atoms i and j; T_{ij} = a discriminant function designed to keep the signs of interactions consistent with the HINT convention that favorable interactions are positive and unfavorable interactions are negative.

The contribution of **hydrogen bonding** to drug-receptor interactions lies in determining the specificity. Besides participation in the drug-receptor complexation, hydrogen bonding plays a crucial role in stabilizing the drug in the conformation required for proper binding to the macromolecular target. There is rarely any ligand molecule devoid of functional groups acting as potential hydrogen donor or acceptor and each of this potentially H-bonding group acts along with other intermolecular forces in lowering the energy of the ligand−receptor complex. The overall energy change involved in breaking and forming the hydrogen bonds determines the favorability of the drug-receptor hydrogen bond. The contribution of hydrogen bonds to drug-receptor interactions was calculated using the OPLS 2005 force field incorporated in the Schrödinger molecular modeling suite.

The term "**entropy loss**" accounts for the loss in torsional, vibrational, rotational, and translational free energies upon binding of the ligand to its receptor. When the ligand approaches the receptor, there is a loss of three rotational and three translational degrees of freedom. This loss of entropy, which is a consequence of the abridged conformational flexibility of the ligand upon receptor binding, was measured using Discovery Studio 2.5.[25]

Another parameter constituting the ligand−receptor binding thermodynamics is the energy associated with changes in the conformation of both the protein and the ligand due to their association. However, in many complexes the conformational change associated with the receptor is insignificant as compared to that involved in the ligand. The change in the ligand conformational energy referred to as the **strain energy** is calculated using a molecular mechanics potential function as the energy associated with changes in bond lengths, angles, torsions and nonbonded interactions. The ligands extracted from their complexes are minimized to relieve the "strain", due to constraints imposed by the protein environment, until the energy gradient of 0.001 kcal/mol·Å is reached (Discovery

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

Studio 2.5). The strain energy is computed as a difference between the energy of the ligand in complex with the protein and energy of the ligand after free minimization in vacuo.

In addition to the above contributions, CoRILISA also takes into consideration the **solvation free energy** at physiological conditions. Both the ligand as well as the receptor is solvated before complexation but as the interactions with solvent molecules compete with protein−ligand interactions, the solvent molecules reorganize. The binding free energy contribution by solvation is actually the hydration free energy, which is the difference between the free (e.g., cellular) and the bound state. Because of the relatively high surface area of the receptor compared to the ligand, the net solvation free energy for the receptor (free and bound) is negligible as compared to that of the ligand. The electrostatic contribution to the solvation free energy of the ligand corresponds to the amount of energy required to strip the solvent molecules off the ligand while moving from an aqueous environment to a hydrophobic receptor cavity. It was calculated using the Jaguar module incorporated in the Schrödinger molecular modeling suite.

Construction of CoRILISA Models. A matrix is built with columns representing each of the LSIs calculated from pairwise similarity between the reference and target molecules based on the interaction elements discussed in the previous section and rows representing each molecule in the data set. A final column containing inhibitory activities of the molecules in the data set is then added to the matrix. In vitro measures of biological activity such as K_i, IC_{50}, etc., reflecting ligand binding strength are used as the dependent variables in the QSAR analysis. Majority of ligand−receptor systems have two troublesome features with respect to developing statistical approaches for performing 3D-QSAR analysis. First, the size of the most data sets (number of molecules) is usually small compared to the number of independent variables (LSIs) and second, many of the independent variables are seen to be highly correlated. A methodology for deriving stable QSAR models for such data sets is G/PLS (genetic function approximation in conjunction with partial least squares) developed by Rogers and Hopfinger[13,26−31] which combines the best features of Genetic Function Approximation (GFA) with PLS. The GFA algorithm generates an initial population of individuals followed by a fitness function, a measure of least-squares error termed as "lack of fit" (LOF), which is applied as an estimate of the quality of each individual. Those individuals with the best fitness scores are allowed to mate and propagate their genetic material to offspring through the crossover and/or mutation operators. After repeatedly performing these steps, the average fitness of the individuals in the population increases, as good combination of "genes" (here the LSI descriptors) are identified and spread through the population. Thereafter the best combinations of the LSI descriptors obtained are subjected to PLS for regression analysis. The entire population of equations can then be searched for information on features, patterns and regions wherein different equations predict well. G/PLS is able to systematically separate the "energetic signals" from the "background noise" in order to derive a meaningful correlation between the biological activity and a subset of weighted independent variables. Since these equations have been derived through extensive crossover and/or mutation cycles we feel that G/PLS is able to identify the global minimum (the most statistically significant) solution to the QSAR problem. The ultimate purpose of developing a regression model is not only to predict properties, but also to provide an interpretation of the structural features responsible for the activity or property of interest. The QSAR equations derived from G/PLS regression are able to extract the crucial features associated with the nature and type of interactions at the level of both the receptor and the ligand that contribute to the biological activity.

The data set was divided into training and test sets on the basis of structural, chemical and biological diversity by adopting the similarity search techniques *viz*. D-Optimal design, Tanimoto similarity coefficient and the Euclidian distance matrix criteria defined in Cerius2 (Accelrys Inc., USA).[32] G/PLS as implemented in Cerius2, was used as the chemometric method to derive the QSAR equations. These models were derived with linear terms and the optimal number of components was selected based on the highest cross-validated r^2 (i.e., q^2). The number of generations and the population size was set to 10,000 and 500 respectively with crossover and mutation probabilities of 50% (default settings). The smoothness function which penalizes the equations on their size and thus controls the bias in the scoring factor between equations with different numbers of terms was set to 1.0.

Diagnostic measures adopted to analyze the significance of the resultant QSAR models include correlation coefficient, leave-one-out/leave-five-out cross-validation correlation coefficients, q_{adj}^2, randomization at 99% confidence interval and boot strapping. The predictive power of the selected models was ascertained by predicting the biological activity of compounds (validation set) not included in the training set. These traditionally used validation parameters have been supplemented with two novel parameters r_m^2 and r_p^2 which are more stringent tests of validation.[33] While r_m^2 penalizes a model for large differences between the observed and predicted values of the test set compounds, the term r_p^2 penalizes the model r^2 for large differences between correlation coefficient of the nonrandomized model and the square of the mean correlation coefficient r_r^2 derived from the randomization test. The functional forms of the two parameters are as follows:

$$r_m^2 = r^2(1 - \sqrt{r^2 - r_0^2}) \qquad r_p^2 = r^2\sqrt{r^2 - r_r^2}$$

where r_0^2 is the squared correlation coefficient between the observed and predicted values of the molecules in the test set with intercept set to zero. For an acceptable QSAR model, the values of r_m^2 and r_p^2 should be greater than 0.5.

RESULTS

The scope of the CoRILISA formalism is demonstrated by analyzing three diverse data sets of protein−ligand complexes—inhibitors of glycogen phosphorylase *b* (GP*b*), human immunodeficiency virus-1 protease (HIV PR), and cyclin dependent kinase 2 (CDK2). The wealth of structural data available for these enzymes is a consequence of the fact that they are long-established targets with many structurally diverse inhibitors being synthesized and tested with a good amount of success. There are several literature reports about QSAR studies performed on different series of inhibitors evaluated against these targets, many of these inhibitors are also part of the present data sets. However, a direct comparison of the statistical outcomes obtained in the present investigation is not possible with previous studies since the data sets used in the construction of the models are not exactly the same. Therefore, the outcomes in terms of medicinal chemistry appear to be a better alternative. Also the outcomes of the present approach have been compared with the preceding

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

Table 1. Best QSAR Models Developed by the CoRILISA Methodology and Its Comparison against Other Approaches for the GP*b* Inhibitors Data Set

technique	QSAR equations[a]	r^2	q_{LOO}^2	q_{LSO}^2	q_{adj}^2	r_{pred}^2	r_{bs}^2	q_{rand}^2	r_m^2	r_p^2
	CoRILISA									
1	pK_i = 0.83 + 1.82 Gly:675 Coul + 0.31 His:A571 Coul +1.55 Ile:380 Coul + 1.26 Asn:284 Coul − 0.84 Tyr:573 HINT + 0.69 Asp:339 HINT + 0.46 Leu:136 HINT + 1.50 Glu:A672 Hbond − 0.49 Glu:672 vdW	0.96	0.85	0.74	0.80	0.61	0.94	0.19	0.60	0.84
2	pK_i = 0.70 + 1.95 Gly:675 Coul + 0.28 His:571 Coul + 1.61 Ile:380 Coul + 1.31 Asn:284 Coul + 0.79 Asp:339 HINT + 0.48 Leu:136 HINT + 1.71 Glu:672 Hbond − 0.97 Asn:484 Hbond − 0.56 Glu:672 vdW	0.96	0.81	0.72	0.74	0.66	0.91	0.18	0.60	0.84
3	pK_i = 0.55 + 2.09 Gly:675 Coul + 0.35 His:571 Coul + 1.64 Ile:380 Coul + 1.35 Asn:284 Coul + 1.08 Asn:282 vdW + 1.19 Asp:339 HINT − 0.47 Asn:282 HINT − 0.49 Ser:674 Hbond	0.95	0.86	0.71	0.82	0.62	0.94	0.18	0.56	0.83
	CoRIA									
1	pK_i = 4.01 − 0.78 Ile:570 Coul + 0.39 Asp:339 Coul + 0.38 Ala:383 Coul − 0.40 Asn:484 vdW + 0.46 Gly:134 vdW − 0.82 Ser:674 HINT − 0.99 His:377 Hbond − 0.27 Leu:136 Hbond	0.79	0.36	0.31	0.18	0.20	0.77	0.42	0.17	0.48
2	pK_i = 3.92 − 1.13 Ile:570 Coul + 0.77 Asp:283 Coul − 0.51 Ser:674 HINT − 0.61 His:341 HINT − 0.29 Asn:284 HINT + 0.71 Lys:574 Hbond − 0.51 His:377 Hbond − 0.20 Ala:673 vdW	0.79	0.38	0.33	0.21	0.30	0.76	0.39	0.23	0.49
3	pK_i = 3.87 + 0.42 Thr:378 Coul − 1.16 Asn:282 Coul − 0.56 Leu:136 Coul + 0.27 Gly:675 vdW − 0.71 Glu:572 vdW + 1.02 Ile:570 vdW − 0.65 His:377 Hbond − 0.93 His:341 HINT	0.79	0.37	0.29	0.19	0.24	0.77	0.43	0.17	0.47
	CoMFA									
		0.84	0.41	0.38		0.32	0.89	0.39	0.21	0.56
		0.84	0.51	0.45		0.28	0.89	0.42	0.17	0.54
		0.84	0.42	0.41		0.31	0.89	0.41	0.20	0.55
	CoMSIA									
		0.84	0.39	0.35		0.33	0.86	0.40	0.18	0.55
		0.86	0.47	0.44		0.38	0.80	0.45	0.26	0.55
		0.84	0.37	0.35		0.26	0.86	0.44	0.19	0.53

[a]Coul, vdW, Hbond, and HINT—Coulombic, van der Waals, hydrogen bonding, and hydrophobic interactions, respectively.

J. Chem. Inf. Model. 2015, 55, 194−205 DOI: 10.1021/ci5006367

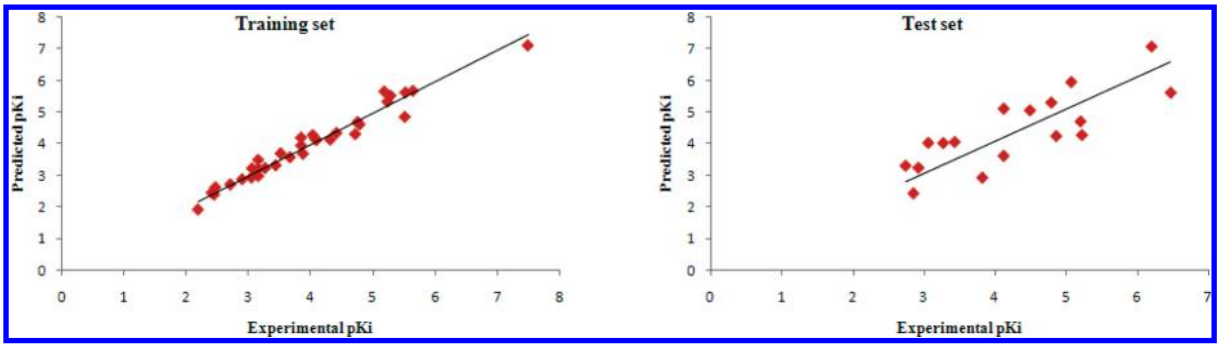


Figure 1. Correlation plots of experimental vs predicted pK_i values for molecules in the training and test sets for the GP*b* inhibitors data set.

ideology i.e. CoRIA alongwith the standard 3D-QSAR techniques-CoMFA and CoMSIA. For the sake of brevity the results have been discussed comprehensively for only the GP*b* data set; its substantiation on the other two data sets, HIV-PR and CDK2 inhibitors, has been conferred only briefly.

Interpretation of CoRILISA Models. Things to be noted while analyzing the CoRILISA models:

1. The sign of the coefficient associated with the LSI descriptor in the QSAR equation will dictate whether an increase (favor) or decrease (disfavor) in the similarity of the target molecule with the reference molecule is required in order to improve its biological activity.
2. The LSI descriptor with a positive coefficient suggests that similarity with the reference molecule should be retained/increased wrt its specific interaction with a particular residue in the receptor. Likewise a negative coefficient for an LSI descriptor indicates that the thermodynamic property of the target molecule should be made "dissimilar" to the reference molecule wrt a specific interaction with a particular residue.
3. In case of electrostatic interactions, a positive coefficient of the LSI descriptor indicates "favored" electrostatic similarity with the reference molecule while a negative coefficient suggests "disfavored" electrostatic similarity with the reference molecule wrt a specific residue around the ligand.
4. Also the more negative the value of the thermodynamic property, the stronger the interaction between the ligand and the particular residue; meanwhile, positive values of these interaction energies imply weaker interaction between the ligand and the receptor residues.
5. In the case of the hydrophobic interactions (HINT Score), positive values signify favorable interactions while negative values indicate unfavorable interactions.

Data Set 1: Glycogen Phosphorylase *b* (GP*b*). Glycogen phosphorylase is an enzyme that catalyzes the degradative phosphorolysis of glycogen to glucose in the liver. Physiologically, the enzyme exists in two interconvertible forms: the dephosphorylated low-activity GP*b* which is transformed by phosphorylation to the high-activity glycogen phosphorylase *a* (GP*a*). Because of its central role in glycogen metabolism, glycogen phosphorylase has been exploited as a target for structure-assisted design of potent inhibitors. An insight into how these inhibitors bind to the glycogen phosphorylase would provide a rational basis for the design of new molecules with an increased affinity and specificity for glycogen phosphorylase. The primary objective for selecting this target is that it has been extensively explored by researchers for evaluating QSAR formalisms. Therefore, a comparison of the QSAR model constructed using the formalism presented herein with the outcome of previous results will assist in evaluating the accuracy, information-content and usefulness of the methodology.

A data set of 50 crystal structures of GP*b* inhibitors was retrieved from the Brookhaven Protein Data Bank (PDB).[34−36] It is worth mentioning that a reasonable model of a complex for which no crystal structure is available can be generated using any docking algorithm and can serve as a good starting point. Every enzyme−inhibitor complex was corrected for any crystallographic errors using the *Protein Preparation Wizard* incorporated in the Schrödinger molecular modeling suite. Structures with missing residues were repaired; correct bond orders and charges were assigned to the atoms; protonation states were assigned consistent with the physiologic pH (7.0). Each enzyme−inhibitor complex was placed in a solvation box with approximately a 10 Å thick shell of TIP3P water molecules. Counter ions were added to maintain electric neutrality in the system. In order to avoid any potential bias resulting from crystal packing forces, the complexes were subsequently minimized with the OPLS 2005 force field until the energy converged to 0.001 kcal/mol. This minimization cycle was used to eliminate any aberrant van der Waals contacts that may exist in the protein structure while preserving the integrity of the X-ray structure.

The experimentally determined inhibition constants (K_i) were converted into pK_i values and are observed to span over a satisfactorily wide range of 5.31 units. The PDB codes for the complexes are summarized in Table 1S together with their experimentally determined inhibition constants. Although the QSAR equations and statistical analysis of the best models are reported in Table 1, other regression equations were also found to be nearly as significant and contain in some cases, different thermodynamic properties as descriptors. The G/PLS regression analysis could yield a statistically significant relationship between the biological activity and the local similarity indices as indicated by the correlation coefficient (r^2) being greater than 0.90. Cross validation using both the leave-one-out (LOO) and leave-five-out (L5O) procedures returned q^2 values more than 0.70 suggesting good internal predictivity for the model. The models were tested for overfitting and chance correlation by y-scrambling (randomization) validation. In this procedure, the dependent variables were randomly shuffled 100 times for the investigated set and the average value for r^2 calculated. Regression models with low r^2 were obtained suggesting that models have not resulted due to a chance correlation or redundancy in the training set. The boot-strap results further advocates the sturdiness of the regression

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

models. For external validation, the G/PLS models obtained using the training set were used to predict the biological activity of molecules (prediction set) not included in the training set. The external validation statistics was found to be acceptable as indicated by the high predictive r^2 values. The residuals [experimental activity (y_{exp}) − predicted activity (y_{pred})] for each compound was found to hardly exceed 1 unit. The correlation plots of experimental vs predicted pK_i values for the molecules in the training as well as test set for the best CoRILISA model (eq 1 in Table 1) are shown in Figure 1. The proposed methodology has been compared with CoRIA and the classical 3D-QSAR (CoMFA and CoMSIA) formalisms using the same training and test sets. It was observed that the QSAR models derived using CoRILISA formalism are significantly better in terms of internal as well as external predictions (Table 1).

An assessment of all the CoRILISA models reveals that almost the same set of amino acids appear consistently in the set of best scoring regression equations, suggesting that these residues play a significant role in the drug-receptor interactions. It was observed that five terms (Gly:675 Coul, Ile:380 Coul, Asn:284 Coul, His:571 Coul) corresponding to Coulombic interactions appear in the equations suggesting importance of the electrostatics over other thermodynamic properties in the inhibiting tendency of the molecules. This also suggests that any conformational change that allows favorable or avoids unfavorable Coulombic interaction of the ligand with residues in the active site of GP*b* should be relatively more important than other interactions. The per residue interaction analysis reveals that an improvement in the activity of a target molecule can be achieved by increasing its electrostatic similarity with respect to the reference molecule in relation to residues—Gly:675, Ile:380, Asn:284, and His:571—owing to the positive coefficients of these descriptors in the equations.

The term "Asn:282 vdW" representing the local similarity index for a molecule in relation to Asn:282 in the terms of van der Waals interaction has a positive coefficient in the equation while the van der Waals interaction energy bears a negative sign at this residue for both the reference as well as the target molecules. Owing to the negative sign for the interaction energies in both the target as well as the reference molecule, the LSI turns out to be positive. Therefore, an increase in the activity can be achieved by increasing the steric similarity of the target molecule with the reference molecule wrt Asn:282. However, the term "Glu:672 vdW" has a negative coefficient in the equations corresponding to the van der Waals interaction energy which also has a negative sign at this residue for both the reference as well as the target molecules. This indicates that rendering the target molecule dissimilar wrt the reference molecule in terms of steric properties at Glu:672 will favor the inhibitory activity. Likewise, on account of the positive coefficient in the QSAR equation, increasing the similarity of the target molecule with the reference molecule in terms of hydrogen bonding properties at residue Glu:672 will favor the activity while reducing the similarity wrt the reference molecule at residue Asn:484 and Ser:674 owing to the negative coefficient will augment the activity.

On the hydrophobic front, the term "Leu:136 HINT" appears with a positive coefficient; correspondingly the sign for the HINT score at Leu:136 for both the target as well as reference molecule is also positive. This indicates that increasing the hydrophobic similarity of the target molecule with respect to the reference molecule will favor the binding process. Also the term "Asp:339 HINT" has a positive coefficient in the equation but as opposed to Leu:136, the HINT score observed for the target as well as reference molecule bears a negative sign (signifying disfavorable hydrophobic interaction). Therefore, owing to the positive coefficient observed for the term "Asp:339 HINT" an increase in the activity for the target molecule can be achieved by rendering it more similar to the reference molecule with respect to hydrophobic property which means that the hydrophobic interaction of the target molecule with Asp:339 needs to be reduced in order to favor the activity. On the other hand, the term "Tyr:573 HINT" appears with a negative coefficient in the equation while the HINT score for both the target as well as reference molecule bears a positive sign at Tyr:573 (LSI at Tyr:573 will be positive), indicating that reducing the hydrophobic similarity of the target molecule with the reference molecule at residue Tyr:573 will improve the activity. Finally, the term "Asn:282 HINT" has a negative coefficient analogous to the corresponding HINT score with a negative sign for reference as well as the target molecule suggesting that rendering the target molecule dissimilar with respect to the reference molecule in terms of the hydrophobic property at Asn:282 will favor the inhibitory activity. Figure 2 provides a depiction of the CoRILISA model with the key residues extracted by the methodology.

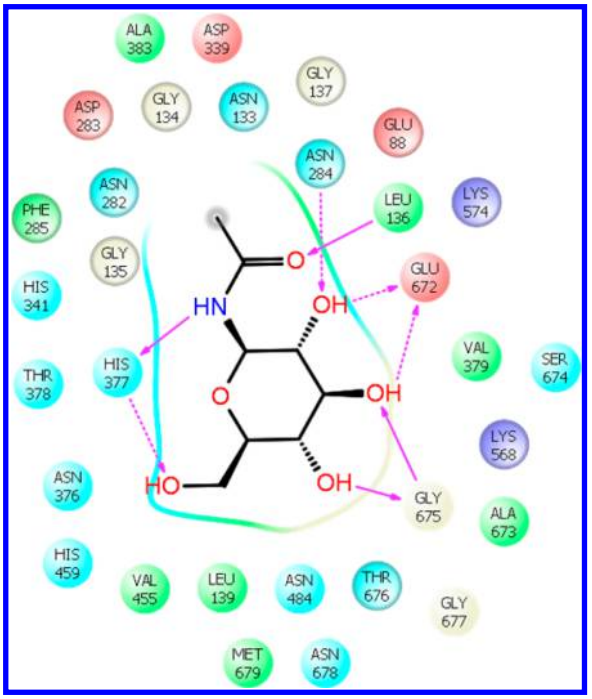


Figure 2. Stereoview of the active site of GP*b* in complex with an inhibitor (PDB: 2prj) showing the key residues extracted by the CoRILISA models.

Comparisons with Other Approaches. CoRIA and classical 3D-QSAR (using CoMFA and CoMSIA; refer to the Supporting Information for details) models were derived using similar training and test sets as adopted for the CoRILISA model and predictions were made for all the compounds. The statistical outcomes of all the approaches are summarized in Table 1. These studies produced a q^2 of 0.36−0.51 as against a q^2 of 0.81−0.86 obtained for the CoRILISA models indicating that the new formalism performs better than the traditional

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

Table 2. Best QSAR Models Developed by the CoRILISA Methodology and Its Comparison against Other Approaches for the HIV-1PR Inhibitors Data Set

technique	QSAR equations[a]	r^2	q_{LOO}^2	q_{LSO}^2	q_{adj}^2	r_{pred}^2	r_{bs}^2	q_{rand}^2	r_m^2	r_p^2
	CoRILISA									
1	pK_i = 0.08 + 0.53 Thr:B26 Coul − 0.74 Asp:B25 Coul + 1.63 Val:B82 vdW + 4.05 Thr:B80 vdW + 3.36 Thr:A80 vdW + 4.00 Ile:A50 vdW − 4.74 Pro:B9 vdW + 1.24 Leu:A90 HINT + 0.78 Met:A46 HINT	0.81	0.74	0.66	0.70	0.69	0.80	0.13	0.58	0.67
2	pK_i = 0.02 + 0.52 Thr:B26 Coul − 0.73 Asp:B25 Coul + 1.64 Val:B82 vdW + 4.05 Thr:B80 vdW + 3.40 Thr:A80 vdW + 4.01 Ile:A50 vdW − 4.75 Pro:B9 vdW + 1.24 Leu:A90 HINT + 0.79 Met:A46 HINT	0.81	0.75	0.64	0.71	0.69	0.80	0.15	0.57	0.66
3	pK_i = 0.03+ 0.54 Thr:B26 Coul − 0.68 Asp:B25 Coul + 1.59 Val:B82 vdW + 4.08 Thr:B80 vdW + 3.53 Thr:A80 vdW + 3.96 Ile:A50 vdW − 4.84 Pro:B9 vdW + 1.26 Leu:A90 HINT + 0.76 Met:A46 HINT	0.81	0.74	0.63	0.69	0.68	0.81	0.18	0.61	0.64
	CoRIA									
1		0.71	0.51	0.39	0.44	0.22	0.69	0.34	0.12	0.43
2		0.71	0.51	0.32	0.42	0.24	0.70	0.37	0.14	0.41
3		0.78	0.52	0.32	0.45	0.25	0.76	0.39	0.13	0.48
	CoMFA									
1		0.85	0.48	0.38		0.28	0.89	0.42	0.20	0.55
2		0.84	0.49	0.36		0.25	0.89	0.43	0.15	0.53
3		0.93	0.49	0.39		0.27	0.95	0.45	0.16	0.64
	CoMSIA									
1		0.87	0.38	0.32		0.24	0.89	0.43	0.15	0.57
2		0.87	0.38	0.33		0.26	0.88	0.45	0.20	0.56
3		0.87	0.39	0.31		0.29	0.89	0.44	0.19	0.57

[a]Coul, vdW, Hbond, and HINT—Coulombic, van der Waals, hydrogen bonding, and hydrophobic interactions, respectively.

QSAR methods. Regarding external validation, when the CoRILISA model is compared with other approaches employing the same external set, significantly better results are obtained as evidenced from an r_{pred}^2 of 0.61−0.66 for CoRILISA against a value of 0.20−0.38 for other methodologies. The validation parameters developed by Roy et al.[33] further advocated the statistical significance of CoRILISA models over the other approaches.

As indicated in the previous section, the GP*b* inhibitors have often been used as a validation set for several QSAR methodologies. Senese et al., using the data set of GP*b* inhibitors, constructed a QSAR model with novel universal 4D fingerprints derived from the 4D-molecular similarity analysis.[37] These are the eigenvalues for a molecule derived from its absolute molecular similarity main distance-dependent matrix (MDDM). The best statistical model obtained in this study had an r^2 of 0.75 with q^2 of 0.68. The results obtained from CoRILISA study are somewhat better achieving an r^2 and q^2 of 0.95−0.96 and 0.81−0.86 respectively. In our recently reported methodology termed LISA, the data set of GP*b* inhibitors was also investigated as a test bed. The statistical outcomes of CoRILISA for the GP*b* data set investigated in this study are significantly better than the LISA models both in terms of internal as well as external validation. LISA is a purely ligand-based approach wherein the influence of the receptor geometry is not considered. Therefore, it is evident that extending the concept of local similarity measures to a receptor-based setting can greatly improve the power of QSAR models with regard to understanding the factors that govern the ligand−receptor binding process vis-à-vis the most active molecule. A similar data set of glucose analog inhibitors of glycogen phosphorylase was studied by Hopfinger et al. to develop a *receptor-independent* (RI) 4D-QSAR analysis wherein the grid cell (spatial) occupancy measures of the atoms of molecules in the training set obtained from the sampling of conformation and alignment spaces were used as descriptors.[38] The best model derived from this study could achieve a q^2 of 0.81−0.83 with an r^2 of 0.86−0.87. Hopfinger et al. extended the RI 4D-QSAR approach to develop the *receptor-dependent* (RD) 4D-QSAR formalism wherein a truncated ligand−receptor model was used to capture ligand−receptor thermodynamics in the functional-region of the protein.[39] It is observed that the models generated using the CoRILISA are comparable in quality, based on statistical measures of fit and test set prediction, to the receptor-dependent 4D-QSAR (q^2 = 0.82, r^2 = 0.85) models. However, in contrast to the 4D-QSAR model, consideration of the complete ligand−receptor structure in the CoRILISA method enables one to capture long-range interactions which may be crucial for binding of the ligand of the enzyme. In terms of statistical outcomes, while there was no significant improvement observed in the RD 4D-QSAR models over the receptor independent counterpart, inclusion of receptor geometry for calculation of the local similarity indices in CoRILISA is significantly able to improve the quality and the predictive ability of the QSAR model as against the model developed using receptor independent LISA (r^2 = 0.83, q^2 = 0.43) method.

Data Set 2: Human Immunodeficiency Virus-1 Protease (HIV-PR). In the pathophysiology of AIDS, HIV protease has proven to be the most promising drug target because of its essential role in viral replication. It is a 99-amino-acid protein encoded by the 5′ portion of the retroviral pol gene which encodes all replicative enzymes. It has been shown that the inactivation of HIV-PR, either by chemical inhibition or certain mutations, leads to the production of immature, noninfectious viral particles. Since the demonstration that HIV-PR plays an essential role in the HIV replication cycle, this enzyme has become one of the primary targets for antiviral drug design and inhibitors of HIV-1 protease are important compounds for establishing highly active antiretroviral therapy for AIDS.

In the present paper, a data set of HIV-1 protease−inhibitor complexes has been studied to test the universal applicability of the CoRILISA formalism. The data set contains 72 inhibitors

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

Table 3. Best QSAR Models Developed by the CoRILISA Methodology and Its Comparison against Other Approaches for the CDK2 Inhibitors Data Set

technique	QSAR equations[a]	r^2	q_{LOO}^2	q_{L5O}^2	q_{adj}^2	r_{pred}^2	r_{bs}^2	q_{rand}^2	r_m^2	r_p^2
	CoRILISA									
1	pIC_{50} 0.18 − 0.25 Asp:145 vdW − 0.71 Gln:131 vdW + 2.90 Thr:14 vdW + 3.85 Ile:10 vdW + 0.38 Thr:14 Coul + 0.52 Gln:85 HINT + 1.47 Phe:82 HINT + 1.08 Val:64 HINT − 0.41 Asp:86 Hbond	0.92	0.81	0.75	0.77	0.63	0.91	0.17	0.59	0.79
2	pIC_{50} 0.09 − 0.62 Gln:131 vdW + 2.79 Thr:14 vdW + 3.84 Ile:10 vdW + 0.67 Gln:85 HINT + 1.91 Phe:82 HINT + 1.03 Val:64 HINT − 0.72 Lys:33 HINT − 0.40 Asp:86 Hbond + 0.34 Thr:14 Coul	0.92	0.84	0.74	0.81	0.67	0.91	0.19	0.60	0.80
3	pIC_{50} 0.13 − 0.73 Gln:131 vdW + 2.73 Thr:14 vdW + 3.88 Ile:10 vdW + 0.37 Thr:14 Coul + 0.65 Gln:85 HINT + 1.55 Phe:82 HINT + 1.0 Val:64 HINT − 0.70 Lys:33 HINT	0.92	0.82	0.73	0.79	0.67	0.90	0.17	0.62	0.80
	CoRIA									
1		0.78	0.40	0.34	0.28	0.30	0.75	0.39	0.28	0.48
2		0.78	0.45	0.40	0.34	0.27	0.71	0.39	0.20	0.48
3		0.78	0.43	0.39	0.32	0.27	0.72	0.37	0.18	0.49
	CoMFA									
1		0.85	0.40	0.37		0.30	0.89	0.38	0.23	0.58
2		0.85	0.43	0.38		0.35	0.80	0.39	0.23	0.57
3		0.84	0.42	0.39		0.28	0.89	0.38	0.24	0.58
	CoMSIA									
1		0.83	0.35	0.33		0.25	0.86	0.37	0.19	0.56
2		0.84	0.39	0.35		0.31	0.86	0.38	0.21	0.56
3		0.85	0.42	0.40		0.30	0.86	0.40	0.21	0.57

[a]Coul, vdW, Hbond, and HINT—Coulombic, van der Waals, hydrogen bonding, and hydrophobic interactions, respectively.

from diverse classes with a wide span of inhibitory potential. The PDB complexes retrieved from the RCSB protein data bank were segregated into a training set for model generation and a test set for model validation on the basis of chemical and biological diversity using the similarity search techniques mentioned in the previous study. The PDB codes for these crystal complexes along with their inhibitory activities are summarized in Table 2S.

The statistical outcomes of the CoRILISA models are compared against CoRIA, CoMFA, and CoMSIA models derived using the same training and test set molecules and are reported in Table 2. The QSAR models generated by the CoRILISA approach are statistically significant producing a correlation coefficients (r^2) of 0.81. Cross validation using both the leave-one-out (LOO) and leave-five-out (L5O) procedures results in statistically acceptable q^2 values. The results obtained from y-randomization and bootstrap analysis further support the sturdiness of the models. Comparison with CoRIA and the classical 3D-QSAR methods shows that the correlation coefficient (r^2) obtained for CoRILISA models appear to be slightly lower than the later approaches. However, the CoRILISA models outperform the CoRIA and 3D-QSAR models in the cross-validation and external predictivity tests. The predictive r^2 for all the CoRILISA models on the test set is more than 0.6 indicating a good predictive power of the models for molecules outside the training set as against an average predictive r^2 of 0.34 obtained by the other approaches for the same test set. Also statistics obtained for the additional validation parameters i.e. r_m^2 and r_p^2 establish the statistical significance of the CoRILISA formalism over the other techniques. The plots of experimental vs predicted pK_i values for the molecules in the training as well as test sets show that the residuals [experimental activity (y_{exp}) − predicted activity (y_{cal})] are less than 1 unit (Figure 1S).

In this study, in addition to building a QSAR model with good predictive power, identification of crucial ligand−receptor interactions that govern the binding process was also the prime objective. Analysis of the best CoRILISA models derived for the HIV-PR inhibitor system shows that an improvement in the activity of the target molecule can be obtained by increasing its steric (expressed as van der Waals interaction energy) similarity at residues Ile:A50, Val:B82, Thr:B80, and Thr:A80 owing to the positive coefficient of these descriptors in the equations and by reducing its steric similarity with residue Pro:B9 (negative coefficient of these descriptors in the QSAR equations) with the reference molecule. The activity of the target molecule can also be augmented by increasing its hydrophobic similarity at Met:A46 and Leu:A90 owing to the positive coefficient for these terms as observed in the equations. Likewise, on account of the negative coefficients in the QSAR equations, rendering the target molecule dissimilar with respect to the reference molecule in terms of electrostatic properties at Asp:B25 will favor the inhibitory activity. On the other increasing the electrostatic similarity with respect to the reference molecule at residue Thr:B26 will augment the activity due to the positive coefficient associated with this term. The CoRILISA models could capture the residues governing the catalytic activity of the HIV-PR enzyme thereby providing a meaningful insight into the ligand−receptor system for this target (Figure 2S).

Data Set 3: Cyclin Dependent Kinase 2 (CDK2). Cyclin dependent kinases (CDKs) are a family of serine/threonine kinases. In concert with cyclins (positive regulators) and the natural inhibitors (CDKI), they serve as the driving force behind the cell cycle regulation and cell proliferation. Cyclin dependent kinase 2, a member of the CDK family encoded by the CDK2 gene, is essential for progression of the cell cycle through G1 to the S phase. Several reports have demonstrated that either over- or underexpression of both the positive regulator cyclin E and the natural inhibitor p27 of CDK2 play a role in the molecular pathology of cancer. This observation makes CDK2 and its regulatory pathways compelling targets for the development of chemical inhibitors that could play an important role in the discovery of a new family of antitumor agents. The availability of the three-dimensional structure of

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

Table 4. Analysis of the Coefficients of the Terms Associated with the Best CoRILISA Equations Obtained for the CDK2 Inhibitors Data Set

sr. no.	terms	sign of the coefficient in QSAR equation	sign of the thermodynamic property: reference molecule	sign of the thermodynamic property: target molecule	inference
1	Ile:10 vdW and Thr:14 vdW	(+)	(−)	(−)	an increase in potency can be achieved by increasing the steric similarity of the target molecule with the reference molecules at Ile:10 and Thr:14
2	Thr:14 Coul	(+)	(−)	(−)	increasing the electrostatic similarity with respect to the reference molecule at Thr:14 will favor binding
3	Gln:131 vdW and Asp:145 vdW	(−)	(−)	(−)	the steric similarity of the target molecule needs to be reduced with respect to the reference molecule at Gln:131 and Asp:145
4	Lys:33 HINT	(−)	(−)	(−)	rendering the target molecule dissimilar with respect to reference molecule at Lys:33 will augment the activity
5	Phe:82 HINT, Gln:85 HINT, and Val:64 HINT	(+)	(+)	(+)	improving the hydrophobic similarity of the target molecule with respect to reference molecule at Phe:82, Gln:85, and Val:64 will improvise the potency
5	Asp:86 Hbond	(−)	(−)	(−)	rendering the query molecule dissimilar with respect to the reference molecule in terms of hydrogen bonding characteristics at Asp:86 will favor the inhibitory activity

CDK2 in complex with diverse classes of inhibitors provides a structural foundation for understanding the mechanisms of activation and inhibition of CDK2. The information prompted us to apply the CoRILISA methodology to this target for further validation of the formalism. To this end a data set of 74 complexes of cyclin dependent kinase 2 with its inhibitors was compiled from the RCSB protein data bank (Table 3S).

The QSAR models derived using G/PLS for this data set show a strong correlation for the training set (r^2 = 0.92, q_{LOO}^2 = 0.81−0.84) with a predictive r^2 of 0.63−0.67 obtained for the test set, demonstrating good external predictability as well (Figure 3S). Other validation parameters such as r_{bs}^2, r_{rand}^2, r_m^2, and r_p^2 further support the statistical significance of the models. The QSAR equations along with the statistical parameters of the models developed using the CoRILISA formalism and its comparison with CoRIA, CoMFA and CoMSIA approaches are summarized in Table 3. A comprehensive analysis of the coefficients associated with the terms of the best CoRILISA equations is provided in Table 4 and Figure 4S.

DISCUSSION AND CONCLUSION

In this paper, we have investigated the similarity between molecules based on the pattern of their interaction with residues in the active site of the receptor. The similarity is expressed by the local similarity index (LSI) against a reference molecule and the LSIs are used as descriptors for deriving the QSAR model.

The main objective of the work presented here was to test the concept and utility of local similarity measures in a receptor based setting and to formalize procedures to utilize this definition of similarity. The formalism was validated by building structure−activity models for three large and diverse data sets of biological importance. Satisfactory correlations were obtained between the local similarity indices and biological activities with good predictions thereby justifying the usefulness of the method. These results are comparable or better than those attained with other QSAR methods. The randomization test and external validation on test sets has confirmed the predictive power of the CoRILISA models. The validation studies performed against three diverse biological systems clearly show that local similarity measures defined in the context of receptor setting can advantageously be used as efficient descriptors for predicting biological activities. Also because of its relative simplicity and absolute generality, the formalism opens an interesting avenue to enrich the traditional QSAR approaches. Moreover, the possibility of identification of individual thermodynamic interactions responsible for the observed biological activity could also be of considerable importance for the rational design of new biologically active candidates.

Finally, the major highlights of the CoRILISA formalism can be summarized as

- The global molecular similarity is broken down into local similarity defined by the receptor surrounding the molecules and this is used as a QSAR descriptor. Systematic investigation of CoRILISA models allows for detection and localization of specific receptor element most likely to be responsible for the observed activity.
- CoRILISA is an expression of similarity encoded as a QSAR formalism in a receptor setting to describe the thermodynamic events involved in ligand−receptor binding and explores both the qualitative as well as the quantitative facets of the ligand−receptor recognition process.
- The approach utilizes the three-dimensional structure of both small molecules as well as their macromolecular targets to extract position-specific information about crucial thermodynamic interactions between the ligand and receptor, which can serve as an aid in the drug design process.

The results are encouraging and we believe that the use of similarity measures in a receptor setting shows promise as a general tool for lead optimization and for designing new drug candidates.

ASSOCIATED CONTENT

Supporting Information

Methodology, Tables 1S−3S, and Figures 1S−4S. This material is available free of charge via the Internet at http://pubs.acs.org.

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

AUTHOR INFORMATION

Corresponding Author

*Telephone: +91-22-26670871. Fax: +91-22-26670816. E-mail: evans@bcpindia.org.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Dean, P. M. Defining molecular similarity and complementarity for drug design. In *Molecular Similarity in Drug Design*, Dean, P. M., Ed.; Blackie Academic & Professional: Glasgow, UK, 1995; Chapter 1, pp 1−23.

(2) Good, A. C. 3D Molecular similarity indices and their application in QSAR studies. In *Molecular Similarity in Drug Design*, Dean, P. M., Ed.; Blackie Academic & Professional: Glasgow, UK, 1995; Chapter 2, pp 24−56.

(3) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225−33.

(4) Maldonado, A. G.; Doucet, J. P.; Petitjean, M.; Fan, B. T. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol. Divers.* **2006**, *10*, 39−79.

(5) Maggiora, G. M.; Johnson, M. A. Introduction to similarity in chemistry. *Concepts and Applications of Molecular Similarity*; J. Wiley & Sons: New York, 1990; pp 1−13.

(6) Rum, G.; Herndon, W. C. Molecular similarity concepts. 5. Analysis of steroid-protein binding constants. *J. Am. Chem. Soc.* **1991**, *113*, 9055−9060.

(7) Good, A. C. The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.* **1992**, *10*, 144−151.

(8) Seri-Levy, A.; West, S.; Richards, W. G. Molecular similarity, quantitative chirality, and QSAR for chiral drugs. *J. Med. Chem.* **1994**, *37*, 1727−32.

(9) Ghuloum, A. M.; Sage, C. R.; Jain, A. N. Molecular hashkeys: a novel method for molecular characterization and its application for predicting important pharmaceutical properties of molecules. *J. Med. Chem.* **1999**, *42*, 1739−48.

(10) Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996**, *39*, 2129−40.

(11) Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: An antitumoral family validation study. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 624−631.

(12) Bultinck, P.; Carbó-Dorca, R. Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 170−177.

(13) Duca, J. S.; Hopfinger, A. J. Estimation of Molecular Similarity Based on 4D-QSAR Analysis: Formalism and Validation. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1367−1387.

(14) Verma, J.; Malde, A.; Khedkar, S.; Iyer, R.; Coutinho, E. Local indices for similarity analysis (LISA)-a 3D-QSAR formalism based on local molecular similarity. *J. Chem. Inf. Model.* **2009**, *49*, 2695−707.

(15) Pissurlenkar, R. R. S.; Coutinho, E. C. HomoSAR: An Integrated Approach Using Homology Modeling and Quantitative Structure-Activity Relationship for Activity Prediction of Peptides. *Scholarly Research Exchange* **2008**, *2008*, 1−12.

(16) Petke, J. D. Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem.* **1993**, *14*, 928−933.

(17) *Glide*, version 5.6, Schrödinger Suite; Schrödinger, LLC: New York, NY, 2010.

(18) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(19) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−9.

(20) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.

(21) Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.

(22) Jorgensen, W. L.; Tirado-Rives, J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665−6670.

(23) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Comput. Aided Mol. Des.* **1991**, *5*, 545−552.

(24) *Sybyl*, version 7.1; Tripos Associates Inc.: St. Louis, MO, USA, 2005.

(25) *Discovery studio*, version 2.5; Accelrys Inc.: San Diego, CA, USA, 2009.

(26) Dunn, W. J., III; Rogers, D. Genetic partial least squares in QSAR. In *Genetic algorithms in molecular modeling*; Devillers, J., Ed.; Academic Press: London, 1996; pp 109−130.

(27) Martin, Y. C. 3D QSAR: Current State, Scope, and Limitations. In *3D QSAR in Drug Design−Recent Advances*; Kubinyi, H.; Folkers, G.; Martin, Y. C., Eds.; Kluwer Academic Publishers: New York, USA, 1998; Vol. 3, pp 3−23.

(28) Verma, J.; Khedkar, V. M.; Coutinho, E. C. 3D-QSAR in drug design–a review. *Curr. Top. Med. Chem.* **2010**, *10*, 95−115.

(29) Datar, P. A.; Khedkar, S. A.; Malde, A. K.; Coutinho, E. C. Comparative residue interaction analysis (CoRIA): a 3D-QSAR approach to explore the binding contributions of active site residues with ligands. *J. Comput. Aided Mol. Des.* **2006**, *20*, 343−60.

(30) Verma, J.; Khedkar, V. M.; Prabhu, A. S.; Khedkar, S. A.; Malde, A. K.; Coutinho, E. C. A comprehensive analysis of the thermodynamic events involved in ligand-receptor binding using CoRIA and its variants. *J. Comput. Aided Mol. Des.* **2008**, *22*, 91−104.

(31) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(32) *Cerius2*, version 4.8; Accelrys, Inc.: San Diego, CA, USA, 1998.

(33) Pratim Roy, P.; Paul, S.; Mitra, I.; Roy, K. On two novel parameters for validation of predictive QSAR models. *Molecules* **2009**, *14*, 1660−701.

(34) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **1977**, *80*, 319−24.

(35) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.* **1978**, *185*, 584−91.

(36) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−42.

(37) Senese, C. L.; Duca, J.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-fingerprints, universal QSAR and QSPR descriptors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1526−39.

(38) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand-receptor binding thermodynamics by free energy force field three-dimensional quantitative structure-activity relationship analysis: applications to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Med. Chem.* **1999**, *42*, 2169−79.

(39) Pan, D.; Tseng, Y.; Hopfinger, A. J. Quantitative structure-based design: formalism and application of receptor-dependent RD-4D-

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205

QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1591−607.

DOI: 10.1021/ci5006367
J. Chem. Inf. Model. 2015, 55, 194−205