**How good are ensembles in improving QSAR models? The case with eCoRIA**

Vijay M. Khedkar, Jose Joseph, Raghuvir Pissurlenkar, Anil Saran and Evans C. Coutinho*

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

Communicated by Ramaswamy H. Sarma

(Received 20 January 2014; accepted 26 March 2014)

A conceptually new idea in quantitative structure–activity relationships (QSAR) which makes use of ensembles from molecular dynamics (MD) trajectories and information retrieved from enzyme–inhibitor binding thermodynamics is presented in this study. This new methodology, termed ensemble comparative residue interaction analysis (eCoRIA), attempts to overcome the current one chemical–one structure–one parameter value dogma in computational chemistry by modeling the biological activity as a function of molecular descriptors derived from an ensemble of conformers of enzyme–inhibitor complexes. The approach is distinctly different from the standard QSAR methodology which uses a single low-energy conformation or the properties averaged over a set of conformers to correlate with the activity. Each conformational ensemble derived from MD simulations is analyzed for the distribution of the non-bonded interaction energies (steric, electrostatic, and hydrophobic) along with solvation, strain, and entropic energy of the inhibitors with the individual amino acid residues in the receptor and these are correlated to the activity through a QSAR model. The scope of the new method is demonstrated with three diverse enzyme–inhibitor data-sets – glycogen phosphorylase b, human immunodeficiency virus-1 protease and cyclin-dependent kinase 2. The QSAR equations derived from the methodology have revealed all the structure activity relationships previously reported for these classes of molecules as well as uncovered some features that were hitherto unknown and may have a hidden role in driving the ligand–receptor-binding process. Impressive improvements in the predictions of affinity have been achieved compared to other QSAR formalisms namely CoMFA, CoMSIA (receptor-independent QSAR techniques), and CoRIA (a receptor-dependent QSAR technique). eCoRIA could provide an understanding of the thermodynamic properties influencing the ligand–receptor binding over a time scale as sampled by the MD simulation. The advantage of analyzing enzyme–inhibitor interaction energies in a statistical domain is that the noise due to inaccuracies in the potential energy functions can be reduced and mechanistically important interaction terms related to protein–ligand binding specificity can be identified which can assist the medicinal chemists in designing new molecules and biologists in studying the influence of position-specific mutations in the receptor on ligand binding.

**Keywords:** ensemble; eCoRIA; receptor-based QSAR; G/PLS

**Introduction**

A quantitative knowledge of the energetics (physical forces) involved in the binding of small organic molecules to a protein is of fundamental importance within the interface of chemistry and biology to link the chemical interactions with biological effects associated with ligand–protein binding. This association is remarkably sensitive to even small measure of the forces which contributes to the overall binding process. A wide variety of computational methods exist for the calculation of such binding affinities; one of these is the free energy perturbation (FEP) (Brandsdal et al., 2003; Lu & Kofke, 2001; Reddy & Erion, 2001) or thermodynamic integration method (Mitchell & McCammon, 1991; Straatsma & Berendsen, 1988) that involves non-physical stepwise conversion between pairs of rather similar ligands in the free and bound states. However, computational demands and issues related to sampling and convergence prevents them from being widely and routinely used in structure-based drug design. Also these techniques are usually “noisy” due to limited conformational sampling, inaccuracies in the modeled structures, and/or inadequacies in current potential energy functions. The quantitative structure-activity relationships (QSAR) approach on the other hand, can balance the different interaction energy contributions in such a way that the “signal” (characterized by a correlation with the biological endpoint) is increased whereas the “noise” (associated with calculation of interaction energies) is reduced. The QSAR model thus establishes a relationship between the differences in activity or binding affinity among members of a series and the differences in the calculated energy contributions. More recently, it was realized that the predictive ability of a molecular mechanics receptor-dependent (RD) QSAR model can be considerably improved if the total interaction energy for each protein–ligand complex is...
partitioned into a number of fragment–based contributions and a subset of the energy variables that account for most of the variance within the series are selected. Considering this fact, a fragment-based approach termed COMparative BINDing Energy (COMBINE) was developed, in which the ligands are divided into fragments chosen according to their spatial location in the protein–binding site (Lozano et al., 2000; Ortiz, Pisabarro, Gago, & Wade, 1995). An improvement of this methodology is Comparative Residue Interaction Analysis (CoRIA) which is based on all elements that define the complete thermodynamic events involved in ligand–receptor association (Datar, Khedkar, Malde, & Coutinho, 2006).

Although these techniques have been successful to a large extent, a conceptual limitation in these approaches is that the methods are based on the current “one chemical–one structure–one parameter value” doctrine. Here, a single conformer characterized by discrete point values of its parameters is used to represent a chemical under study, while all other conformational states are ignored. This assumption runs counter to a growing number of studies indicating that most organic molecules of non-trivial size are not just three-dimensional but are “four-dimensional” because they exist as an ensemble of three-dimensional conformations interchanging over time (or equivalently, consist of a distribution of different microscopic states) at room temperature (Hopfinger et al., 1997; Pissurlenkar, Khedkar, Iyer, & Coutinho, 2011). Their properties and corresponding reactivities depend intimately on this ensemble. Therefore, knowing the “structure” of a molecule requires knowing the structures in this ensemble. The same is true for the corresponding enzyme structure as well. Structural information derived from a single low-energy conformer may be insufficient for a complete description of a chemical and could be the prime reason for failure in QSAR analysis. Ultimately, to predict accurate binding affinities, it is essential to go beyond predicting a single “dominant lowest energy” conformation of the ligand complexed with the protein.

The observation that relatively subtle energy differences between the various conformational states of a molecule can yield significant variation in biological endpoint underscores the necessity to represent the molecular parameter values for each molecule as finite ranges, instead of single point values corresponding to the lowest energy conformer. Hence, the application of traditional logic in the investigation of chemical structures requires the “one chemical–one structure–one parameter value” dogma to be modified to a “one chemical–finite set of structures–range of parameter values” principle. The ability to account for the conformational dynamics associated with enzyme–inhibitor complexes is of specific importance for those methods such as QSAR that quantitatively predict some measurable real-life property of a molecular system. With very few exceptions, researchers (Andrade, Pasqualoto, Ferreira, & Hopfinger, 2010; Araújo et al., 2011; Fontes Ferreira da Cunha, Albuquerque, Cevã Antunes, & Picca de Alencastro, 2005; Hopfinger et al., 1997; Oliveira, Ramalho, & da Cunha, 2009; Pissurlenkar et al., 2011; Vedani & Dobler, 2000; Vedani & Dobler, 2002a, 2002b; Vedani, Briem, Dobler, Dollinger, & McMasters, 2000; Wildman & Crippen, 2002) have tended to neglect the investigation of conformational ensembles in the design of 3D-QSAR techniques. The reason being the basic dogma of traditional 3D-QSAR that only those parameters contribute to good QSAR models which are derived from a specific active conformation generating the molecular interaction field (MIF) (Goodford, 1985) responsible for the ligand–receptor binding. However, it is noteworthy that the success of such MIF-based 3D-QSAR studies does not depend on finding the real receptor-bound/bioactive conformation, but rather on a higher self-consistent molecular alignment. Therefore, almost all the available 3D-QSAR methods require an assumed bioactive conformation and a good molecular alignment as an input (Cramer III, Patterson, & Bunc, 1988; Doweyko, 1991; Gancia, Bravi, Mascagni, & Zaliani, 2000; Klebe, Abraham, & Mietzner, 1994; Pastor, Cruciani, McIlwain, Pickett, & Clementi, 2000; Robinson, Winn, Lyne, & Richards, 1999; Silverman & Platt, 1996), and their ability to predict the real active conformation is consequently very limited. The way out began with the consideration of higher dimensions in QSAR where the effects of conformational flexibility, solvation, and the induced fit were taken into account in a holistic way. Hopfinger et al. (1997) and the group of Vedani (2000) have developed QSAR methods beyond the 3rd dimension by considering the conformational flexibility as the fourth dimension. As an evolution of molecular shape analysis (Hopfinger, 1980), Hopfinger et al. have developed a 4D-QSAR formalism which incorporates ligand conformational flexibility and multiple alignment by ensemble averaging the spatial features i.e. marking the fourth “dimension” of QSAR. Besides this, an exhaustive evaluation of ligand-embedded pharmacophore groupings is also adopted in the QSAR model-building process. Construction of the QSAR model and identification of possible pharmacophore sites is achieved by a three-step statistical analysis, which comprises a genetic algorithm (GFA) optimization followed by backward elimination multidimensional regression and ending with another round of GFA optimization. The methodology has proved useful in the construction of quantitative 3D pharmacophore models for a set of ligands for which the geometry of the corresponding receptor is not known (Receptor Independent: RI-4D-QSAR) as well as for those cases where it is known (Andrade et al., 2010; Duca & Hopfinger, 2001; Hopfinger et al., 1997; Pan, Tseng, & Hopfinger, 2003; Pan, Liu, Senese, Hopfinger, & Tseng, 2004; Senese & Hopfinger, 2003; Venkataraman & Hopfinger, 1999). The most recent evolution in QSAR has
been the incorporation of the induced fit protocol as the 5th dimension allowing for a multiple representation of the topology of the quasi-atomistic receptor surrogate and simultaneous evaluation of various solvation models both implicitly as well as explicitly as the 6th dimension, additionally incorporating contributions from the solvent and entropy into the analysis (Oliveira et al., 2009; Vedani & Dobler, 2000; Vedani & Dobler, 2002a, 2002b; Vedani, Dobler, & Lill, 2005).

A significant amount of structural and activity data is still missing due to the lack of experimental and theoretical studies on the relationship of conformational change and associated variation in parameter values. Hence, an approach where biological activity is modeled as a function of molecular descriptors, derived from specifically selected conformations of the enzyme–inhibitor complex rather than as a property derived from the lowest energy gas phase conformer, has the potential to augment the quality and predictive power of a 3D-QSAR model. In our recently published methodology – ensemble QSAR – the applicability of the conformational ensembles in QSAR has been discussed. However, the approach was tested successfully only on peptide data-sets (ligand-based approach) and its extension into receptor-based setting was thought to be the subject of an ongoing endeavor (Pissurlenkar et al., 2011). With this objective, we address herein the issue of the need for an ensemble of protein–ligand conformers in the QSAR domain by a “dynamic” receptor-based QSAR approach termed ensemble Comparative Residue-Interaction Analysis (eCoRIA), which samples all thermodynamic contributions to ligand–receptor binding from the molecular dynamic (MD) trajectories and uses them as descriptors in the QSAR analysis. The term “ensemble” is used to describe attempts at mimicking the infinite conformational space of a protein–ligand complex by sampling a set of discrete isomers from the conformational trajectory. The method first establishes a collection of reasonable protein–ligand conformations that captures the bioactive conformation as one of diverse, energetically accessible conformations, from which the thermodynamic descriptors are extracted. These thermodynamic properties governing ligand binding are then correlated with the biological activity by a chemometric technique. The eCoRIA formalism has been tested on three diverse enzyme–inhibitor data-sets – glycogen phosphorylase b (GPb), human immunodeficiency virus-1 protease (HIV-PR), and cyclin-dependent kinase (CDK2).

Some features of eCoRIA formalism may overlap with our preceding QSAR techniques, which were adopted directly; some were slightly modified, while others were ignored altogether (Datar et al., 2006; Pissurlenkar et al., 2011; Verma et al., 2008). Therefore, eCoRIA may seem to be an unwieldy mixture of techniques and principles but it is firmly based on our in-house QSAR technique – eQSAR.

Materials and methods

The operational steps involved in the eCoRIA analysis are schematically illustrated in Figure 1.

Theory

A literature survey of current 3D-QSAR techniques indicate, perhaps not too surprisingly, that investigating the thermodynamics of ligand–receptor binding interaction is of primary focus. Nevertheless, when the receptor geometry is available, one can theoretically evaluate all the thermodynamic properties that govern the binding of the ligand to its receptor.

The biological activity is related to the free energy of binding $\Delta G_{\text{bind}}$. The later is equated to the inhibition constant ($K_i$) as:

$$
\Delta G = -RT \ln K_i
$$

where $T$ is the absolute temperature and $R$ the gas constant.

The in vitro measures of biological activity are representative of the relative ligand binding strength ($\Delta G_{\text{bind}}$) and can be used as the dependent variable for constructing QSAR models.

The free energy of binding, $\Delta G_{\text{bind}}$, can be expressed empirically as a sum of the relative free energies of various thermodynamic events as follows:

$$
\Delta G_{\text{bind}} = \Delta G_{\text{solv}} + \Delta G_{\text{inter}} + \Delta G_{\text{conf}} + \Delta G_{\text{hydrophobic}} + \Delta G_{\text{entropy}}
$$

where $\Delta G_{\text{solv}}$ is the change in solvation energy, $\Delta G_{\text{conf}}$ reflects the conformational changes in the ligand as well as receptor, $\Delta G_{\text{inter}}$ represents the specific intermolecular interaction energies i.e. van der Waals and electrostatics, $\Delta G_{\text{hydrophobic}}$ is the hydrophobic interaction between the receptor and ligand, and $\Delta G_{\text{entropy}}$ is change in entropy.

Each one of the terms in Equation (2) can be evaluated with a molecular mechanics forcefield. However, since no force field is exact, the energy terms may be analyzed as descriptors using a QSAR paradigm in order to scale them appropriately.

Workflow

We discuss the sequence of steps in eCoRIA as – modeling molecules and their complexes; generation of protein–ligand ensembles; measurement of thermodynamic components involving ligand–receptor binding; and finally the chemometric analysis to derive the regression equations.
Modeling enzyme–inhibitor complexes

Protein–ligand complexes were sourced from the Protein Data Bank (Bernstein et al., 1977, 1978, 2000). In the absence of an X-ray structure, a model of the ligand obtained by docking into a receptor of known geometry can also serve as a suitable starting point.

Every enzyme–inhibitor complex is first corrected for any crystallographic errors using the Protein Preparation Wizard workflow in the Schrödinger molecular modeling suite (Schrödinger LLC, New York). Structures with missing residues are repaired and correct bond orders and charges assigned to the atoms. The protein is capped with the N-methylamino group at the C-terminal and the

Figure 1. Schematic diagram of the operational steps involved in eCoRIA.
acetyl (ACE) group at the N-terminal to maintain a backbone like structure. Beside the crystallographic water molecules, each ligand–protein complex is solvated with approximately a 5 Å thick shell of TIP3P water molecules. Counterions are added to satisfy the electroneutrality condition in the system. A distance-dependent dielectric is used to account for the electrostatic shielding effects of aqueous solvent. Periodic boundary conditions (NPT ensemble) are applied to the system. The system is then subjected to energy minimization using the OPLS 2005 forcefield, so as to remove aberrant van der Waals contacts while preserving most of the original structure information, using a cycle of steepest descents followed by conjugate gradients until the gradient converges of .01 kcal/mol/Å. The minimized structure is then used in the molecular dynamics simulation (MDS) to generate the ensemble of conformations.

**Generation of conformational ensemble**

The task of generating conformational ensembles can be accomplished by two approaches: deterministic (MD) and stochastic (MC). We have employed MD simulation to generate the conformational ensemble profile (CEP) of the enzyme inhibitor. The geometry of the energy minimized enzyme–inhibitor complex obtained in the first step is used as the starting structure for the MDS. The MD simulation is run with the MD module contained in Desmond (version 2.4, DE Shaw Research, NY, USA). The simulation is run for a period of 5 ns under the NPT (constant number of particles, pressure, and temperature) ensemble. The temperature of the system is maintained at 300 K by coupling to an external bath based on the Berendsen algorithm (Berendsen, Postma, Van Gunsteren, DiNola, & Haak, 1984). An isotropic condition with pressure restrained to 1 bar is achieved with the Berendsen barostat. High-frequency vibrations are eliminated by applying the SHAKE algorithm, constraining all bonds to their equilibrium values. Initial velocities are generated randomly by a Maxwell distribution at 300 K in accordance with the masses assigned to the atoms. The Particle Mesh Ewald (Essmann et al., 1995) method is used for calculating long-range electrostatics. The MD trajectory and corresponding energies are sampled every 5 ps to construct the CEP of each ligand–receptor complex. The MD approach generates ensemble of structures representative of those found in thermal equilibrium.

**Measurement of thermodynamic properties**

As discussed earlier in Equation (2), the complete thermodynamics of enzyme–inhibitor binding leading to a biological response can be described empirically as a linear sum of energy components that describe physical changes in the protein–ligand system caused by binding. An overview of the energy components known to contribute significantly to protein–ligand association and which are adopted in the eCoRiA formalism is as follows.

The specific non-bonded interactions between the ligand and receptor are key elements in the binding thermodynamics. The components of the non-bonded interaction energy – van der Waals ($E_{vdW}$) and Coulombic ($E_{Coul}$) are calculated between the inhibitor and each amino acid residue in the protein environment. To account for the strong distance dependence of the electrostatics we used an off-grid calculation to include long-range electrostatic interactions. These interactions are calculated with the OPLS 2005 (Jorgensen & Tirado-Rives, 1988; Jorgensen, Maxwell, & Tirado-Rives, 1996) force field using the Glide module incorporated in the Schrödinger molecular modeling suite (Friesner et al., 2004, 2006).

The functional form of the non-bonded interaction energies between the ligand and the receptor is expressed as

$$
E_{\text{nonbonded}} = \sum_{i > j} f_j \left( A_{ij} \frac{a_{ij}}{r_{ij}^6} - C_{ij} \frac{q_i q_j e^2}{4\pi\varepsilon_0 r_{ij}} \right) 
$$

(3)

where $A_{ij} = \sqrt{A_i A_j}$ and $C_{ij} = \sqrt{C_i C_j}$, $A_i$ and $C_i$ are the repulsive and attractive coefficients for the 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 $\varepsilon$ is the dielectric constant.

Accounting for hydrogen bonding separately from the electrostatics is crucial in modeling the specificity of drug–receptor interaction. Besides participation in receptor–drug complexation, H-bonding stabilizes the drug in the conformation required for proper binding to the receptor. The free energy of protonation defined as the difference between the total energies of the protonated and neutral forms of the molecule is considered as a good measure of the strength of hydrogen bonds i.e. the higher the energy, the stronger is the bond. The quantified values for the hydrogen bonding interactions between the ligand and receptor were calculated using the Glide module in the Schrödinger molecular modeling package.

Hydrophobic interaction has long been recognized to play an important role in the thermodynamics of ligand–receptor binding. It is an entropy-driven process resulting primarily from the change in the orientation of solvent molecules in the solvation shell wrapping the solute molecules and also from the bulk form of solvent molecules. The HINT (Kellogg, Semus, & Abraham, 1991) program incorporated in the Sybyl suite (version 7.1, Tripos Inc.,
USA) was used to calculate the hydrophobic interactions between all atom pairs in a ligand–protein system. The functional form of the HINT score is:

\[
B = \sum_i \sum_j b_{ij} = \sum_i \sum_j (a_i S_i S_j T_{ij} + r_{ij})
\]  

(4)

where, \(B\) is the total interaction between the molecules, \(i\) and \(j\) are atoms of the protein and ligand, respectively, \(b_{ij}\) is the micro-interaction constant/binding score that represents the attraction/interaction between atoms \(i\) and \(j\) in the protein–ligand complex, \(a_i\) is the hydrophobic atom constant for atom \(i\) based on the CLOGP method of Hansch and Leo, \(S_i\) is the solvent accessible surface area for atom \(i\), \(R\) and \(r\) are the distance functions for the interaction between atoms \(i\) and \(j\), \(T_{ij}\) is a descriptor function, which is essentially a sign change function parameterized to maintain the signs of interactions consistent with the HINT convention that favorable interactions have positive HINT scores, while unfavorable interactions have negative HINT scores.

Another major parameter contributing to the thermodynamics of ligand–protein binding is the conformational entropy. The entropy loss in both the protein and the ligand accounts for the loss of torsional, vibrational, rotational, and translational free energies upon binding. When two molecules bind, there is a loss of three rotational and three translational degrees of freedom due to constraints about single bonds on complex formation. The loss of entropy due to reduced conformational flexibility upon receptor binding was estimated using Discovery Studio 2.5 (version 2.5, Accelrys Inc., San Diego, CA, USA).

eCoRJA also keeps an account of the energetic contribution from changes in the conformation of the ligand upon binding with the protein and vice versa. The conformational change associated with the ligand upon binding to the protein is much more significant compared to that for the protein and is represented as the strain energy. It is computed using a molecular mechanics potential function as the energy associated with changes in bond lengths, angles, torsions, and non-bonded interactions. The ligands extracted from their complexes are minimized using the smart minimizer in Discovery Studio until an energy gradient of .001 kcal/mol\(\AA\) is reached. The difference in energy of the ligands in the complex and the conformation minimized in vacuo is taken as the conformational strain energy.

The last term contributing to the free energy of binding is the solvation energy at physiological conditions (hydration free energy). The ligand–protein interactions compete with water interactions due to the fact that both the protein and the ligand are solvated before complexation. The energy required to strip the solvent molecules off the ligand while moving from an aqueous environment to a hydrophobic protein cavity constitutes the solvation energy. Because of the relatively high surface area of the protein compared to the ligand, the net solvation free energy for the protein is negligible as compared to that of the ligand. For the ligand, the electrostatic contribution to the solvation free energy is calculated using the Jaguar module in the Schrödinger molecular modeling package.

Every term in the energy function described above contributes more or less to the enzyme–inhibitor binding process. Through these energy functions one can explore various details of binding specificity within enzyme–inhibitor systems. More details about these energy functions may be found in our earlier articles (Datar et al., 2006; Verma et al., 2008). It should be noted that in eCoRJA the energy terms are derived from an ensemble of energetically reasonable conformations, a feature unique to eCoRJA and not adopted in earlier techniques.

A final column containing biological activities is then added to the study table. In vitro measures of biological activity, such as \(K_i\) and IC50, are taken to reflect relative ligand-binding strength and used as the dependent variable in the eCoRJA study.

It is hoped that the energy decomposition scheme adopted herein must, at some point, meet the two opposite propensities of the Scylla of considering enough energy terms that the elements responsible for the activity differences can be isolated and the Charybdis of including excessively large number of terms that the signal-to-noise ratio is so low that the subsequent analysis to derive a meaningful statistical model fails (Wade, Ortiz, & Gago, 1998).

**Chemometric analysis**

Static structure analyses by averaging the energy calculations over ensemble of structures have been the norm in understanding protein–ligand interactions. However, it may not reflect the distribution of the properties over the conformational states sampled by the molecule. It is the shape of the energy landscape that is crucial to understand real-time protein conformational fluctuations and dynamics. In the eCoRJA formalism presented here, we have applied moments of distribution as a quantitative measure of the shape of the unseen energy landscape. Best known among these is the first moment, \(M_1\), which describes the mean of the distribution around which central clustering occurs and is calculated as

\[
\bar{x} = 1/N \sum_{j=1}^{N} x_j
\]

(5)

where \(x_j\) is the measured thermodynamic property and \(N\) is the number of samples in the ensemble (1000 in this case).
It is a good approximation of the central tendency for unimodal symmetric distributions. However, it can be misleading in a skewed or multimodal distribution of property. Therefore, it is essential to consider additional measures for quantifying skewed distributions. After having characterized a distribution’s central value, one has to know the spread of the distribution i.e. its “width” or “variability” around that value. The spread of a distribution may be estimated using various parameters, of which variance is the most common one, considered to be the second moment of distribution, $M_2$.

$$\text{variance} \left( x_1 \ldots x_N \right) = \frac{1}{(N - 1)} \sum_{j=1}^{N} (x_j - \bar{x})^2 \quad (6)$$

It is measured as the sum of the squared deviations from the mean divided by the number of samples minus 1 ($N - 1$). The third moment of distribution, $M_3$, is the skewness coefficient, which characterizes the degree of asymmetry of a distribution around its mean. In contrast to the mean which is a dimensional quantity having the same unit as the measured quantities, the skewness is defined in such a way as to make it non-dimensional.

$$\text{skewness} \left( x_1 \ldots x_N \right) = \frac{1}{N} \sum_{j=1}^{N} \left( \frac{x_j - \bar{x}}{\sigma} \right)^3 \quad (7)$$

It is a pure number characterizing only the shape of the distribution for a property over the ensemble. The kurtosis coefficient related to the excess (kurtosis) of the distribution is the fourth moment of distribution, $M_4$.

$$\text{kurtosis} \left( x_1 \ldots x_N \right) = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left( \frac{x_j - \bar{x}}{\sigma} \right)^4 \right\} - 3 \quad (8)$$

It measures the relative flatness of a distribution (as against the normal distribution showing a kurtosis coefficient of zero). A positive kurtosis coefficient signifies a tapering distribution, while a negative kurtosis coefficient indicates a relatively flat distribution of the property over the ensemble of conformations.

The four principle moments of distribution – mean, variance, skewness, and kurtosis are calculated for each free energy component over the ensemble of conformations sampled by the enzyme–inhibitor complex. This approach is distinctly different from selecting a single conformation of an enzyme–inhibitor complex representing the ensemble by a single average value. The methodology explicitly incorporates the features and properties of all conformations of a molecule to represent the distribution of these properties over the conformational states available to a molecule.

As a result of the large number of terms the genetic version of partial least squares (PLS) i.e. G/PLS (Genetic function approximation in conjunction with PLS) (Dunn III & Rogers, 1996; Datar et al., 2006; Khedkar, Malde, & Coutinho, 2007; Pissurlenkar et al., 2011; Verma et al., 2008; Verma, Khedkar, & Coutinho, 2010) is used as the statistical tool to systematically separate the “energetic signals” from the “background noise” in order to obtain a correlation between the activity and a subset of weighted energy terms.

In order to preserve the information contained in the descriptor data-set, the pretreatment of the QSAR table based on correlation matrix was avoided. However, the descriptors (moments of distribution over each property) were scaled according to their mean and standard deviation so that all the columns are placed on the same platform and have equal weight (importance), while deriving the QSAR equations.

To derive statistically significant QSAR models, the data-set was split into training and test sets on the basis of structural, chemical, and biological diversity using similarity search techniques viz. D-Optimal design, Tanimoto similarity coefficient, and the Euclidian distance matrix criteria defined in Cerius2 (version 4.8, Accelrys Inc., San Diego, CA, USA). The selection technique searches for diverse molecules by comparing their chemical nature in 2D space. All QSAR equations were generated with the G/PLS method as implemented in the Cerius2 program. The models were developed with linear terms and the optimal number of components was selected based on the highest cross-validated $r^2$ (i.e. $q^2$) (Stone, 1974; Deep, 2006). The length of the equations was varied from 6 to 10 terms; with a smoothness value of 1.0 (the smoothness function controls the bias in the scoring factor between equations with different number of terms). The number of generations was limited to 10,000 and population size to 500. Crossover and mutation probabilities of 50% (default settings) were used. The models were rigorously validated internally using randomization at 99% confidence interval (Rucker, Rucker, & Meringer, 2007), Leave-One-Out (LOO), Leave-Five-Out (LSO), $q^2_{adj}$ (Romeiro, Albuquerque, de Alencastro, Ravi, & Hopfinger, 2005), and by bootstrap procedures (Richard, Cramer III, Bunce, Patterson, & Frank, 1988). For external validation, the QSAR models were used to predict the biological activity of compounds not included in the training set (validation set). The models were further tested for their statistical significance by calculating validation parameters devised by Roy et al. ($r^2_{adj}$ and $r^2_{p}$) (Pratim Roy, Paul, Mitra, & Roy, 2009). While $r^2_{adj}$ is a measure of the external predictivity of a model, $r^2_{p}$ which is based on the randomization test indicates the statistical significance of the relationship between activity and the descriptors.

Compared with classical molecular mechanics calculations of binding energies, the advantage of calculating ligand–protein interaction energies over an ensemble and then subjecting them to statistical analysis is that the
noise due to inaccuracies in the potential energy functions and molecular models can be reduced and mechanistically important interaction terms can be identified.

The best statistical models derived using the eCoRiA approach were then compared with CoRiA and the standard 3D-QSAR techniques – CoMFA and CoMSIA using the same training and test set molecules.

Results

The objective behind eCoRiA formalism was to develop a technique that accounts for the distribution of thermodynamic properties over an ensemble of enzyme–inhibitor conformers and incorporates their properties in the QSAR framework. The outcome of this technique for three different ligand–receptor systems is discussed comprehensively below. There are several literature reports of QSAR models obtained by various methods on different series of inhibitors acting on these targets. However, as the data-set compiled in this study and those used in the previous reports are not exactly identical, a direct comparison of the statistical outcomes with those obtained by eCoRiA is not fair. However, interpretation of the models in terms of chemistry appears to be a more meaningful alternative.

Data-set 1: GPb

The first application of the eCoRiA formalism is on a data-set of GPb inhibitors, an enzyme that catalyzes the first (or the rate-limiting) step in the phosphorolysis of glycogen to glucose-1-phosphate in the liver. Inhibition of hepatic glycogen phosphorylase has been proposed as a promising strategy for attenuating hyperglycemia in type 2 diabetes. Since glucose production in the liver has been observed to increase in type 2 diabetes, inhibiting the release of glucose from the liver’s glycogen stores appears to be a valid approach. Administration of glycogen phosphorylase inhibitors prevents the conversion of glycogen into glucose and, thus, helps attenuating hyperglycemia in type 2 diabetes by shifting the balance between glycogen breakdown and synthesis in favor of the latter. The enzyme glycogen phosphorylase exists in two interconvertible forms (an inactive b and an active a form); the proportion of each form is regulated by phosphorylation. The less active b (GPb) form is transformed by phosphorylation (glycogen phosphorylase kinase a) to the more active a (GPa) form. The monomer of glycogen phosphorylase in the muscle is a large protein composed of 842 amino acids. The principle reason for selecting this target is that it has been extensively explored by Hopfinger et al. for evaluating their 4D-QSAR formalism. Therefore, a comparison with the outcome of previous results permits an assessment of the accuracy and information content of the eCoRiA models.

A set of 45 enzyme–inhibitor crystal complexes of GPb were compiled from the protein data bank. The experimentally determined inhibition constants were transformed into pKᵢ values and are spread over a satisfactorily large range of 5.09 units. The PDB IDs of these molecules are summarized in Table 1S together with their experimentally determined activities. Although only the optimal eCoRiA equations in terms of statistics are shown in Table 1, other regression equations are also found to be nearly as significant and with slightly different descriptors. Statistically significant models were obtained with correlation coefficient (r²) greater than .90. Cross-validation by the LOO and L5O procedure also produced statistically significant q² values (> .70). The bootstrap values further advocated the robustness of these models. The predictive r² i.e. r² predicted calculated on the test set is also found to be greater than .60, indicating a good predictive power of the models for molecules outside the training set. A value of greater than .50 obtained for r² m and r² also establish the power of predictivity for the models and eliminate the possibility of chance correlation. The residuals [experimental activity (Y(exp)) – predicted activity (Y(pred))] for each compound in the training as well as test set hardly exceed 1 unit. Comparison with the CoRiA, CoMFA, and CoMSIA approaches reveals that the models derived using eCoRiA method are significantly better in terms of r² and predictive capability (r² predicted) (Table 2). The correlation plots of experimental vs. predicted pKᵢ values of the molecules in the training and test sets for eCoRiA model (Equation (1) in Table 2) are shown in Figure 1S.

All 500 equations associated with the best eCoRiA model were analyzed for the frequency with which a particular descriptor appears in the population of equations. The plots of the most frequently appearing descriptors in different equations are shown in Figure 2. The frequency of appearance of different terms is shown on the X-axis, whereas the sign of the coefficients of the terms in the equations are shown on the Y-axis. Terms with positive coefficients in the equations are plotted on the positive Y-axis, whereas those with negative coefficients are on the negative Y-axis. The greater the frequency of appearance for a descriptor in the population of equations, the greater is its role in explaining the variance in the biological activity. A detailed analysis of the models obtained by the eCoRiA approach is described below.

While interpreting the equations derived from the eCoRiA methodology one should remember that

1. More negative the value of the interaction energies, stronger is the interaction between the ligand and the respective amino acid in the target enzyme. Similarly, positive values of these interaction energies signify weaker interaction between the enzyme and the ligand.
Table 1. Best models developed by the eCoRIA and the CoRIA approaches for various combinations of the descriptors.

| Technique | QSAR equations |
|-----------|----------------|
| eCoRIA    |                |
| 1         | $pK_i = 0.81 - 1.04 \text{Tyr:280:vdW} M_3 + 0.57 \text{entropy} M_1 + 2.03 \text{Asn:284:vdW} M_3 - 0.40 \text{Gly:137:vdW} M_3 - 0.93 \text{His:377:Hbond} M_1 + 0.39 \text{Ser:674:vdW} M_2 - 0.29 \text{Gly:135:Coul} M_2$ |
| 2         | $pK_i = 0.44 - 0.86 \text{Tyr:280:vdW} M_3 + 0.50 \text{entropy} M_1 + 1.97 \text{Asn:284:vdW} M_3 - 0.36 \text{Gly:137:vdW} M_3 - 1.13 \text{His:377:Hbond} M_1 + 0.39 \text{Arg:292:Coul} M_3 - 0.38 \text{Gly:135:Coul} M_2$ |
| 3         | $pK_i = 0.62 - 0.79 \text{Tyr:80:vdW} M_3 - 0.35 \text{Gly:137:vdW} M_3 - 1.27 \text{His:377:Hbond} M_1 + 0.45 \text{entropy} M_1 + 1.84 \text{Asn:284:vdW} M_3 + 0.47 \text{Arg:292:Coul} M_3 - 0.36 \text{Gly:135:Coul} M_2$ |
| CoRIA     |                |
| 1         | $pK_i = 2.84 + 1.70 \text{Gly:137:Coul} + 1.52 \text{Gly:134:HINT} - 1.38 \text{Asn:484:Coul} - 0.84 \text{Leu:136:Coul} + 0.35 \text{Asp:283:HINT} - 2.57 \text{Asp:283:Coul} + 0.41 \text{Thr:676:HINT}$ |
| 2         | $pK_i = 2.90 + 2.27 \text{Gly:137:Coul} - 2.11 \text{Asp:283:Coul} - 0.17 \text{Phe:286:vdW} - 0.43 \text{Ser:674:HINT} + 1.23 \text{Gly:134:HINT} - 0.90 \text{Asn:484:Coul} - 0.99 \text{Leu:136:Coul}$ |
| 3         | $pK_i = 3.05 + 1.86 \text{Gly:137:Coul} - 0.58 \text{Ser:674:HINT} - 0.63 \text{Asn:484:Coul} - 1.99 \text{Asp:283:Coul} - 0.96 \text{Leu:136:Coul} + 1.24 \text{Gly:134:HINT} - 0.29 \text{Tyr:573:HINT}$ |

Note: Where $M_1$, $M_2$, $M_3$, $M_4$ are the first, second, third, and fourth moments of distribution respectively, $M_1$ = Mean; $M_2$ = Variance; $M_3$ = Skewness; $M_4$ = Kurtosis; “vdW” refers to the van der Waals interaction, “Coul” refers to the Coulombic interactions, HINT refers to hydrophobic interactions, and Hbond is the hydrogen-bonding interactions.
In case of hydrophobic interactions, positive values imply favorable interactions while unfavorable interactions are associated with a negative coefficient.

It is the sign (+/−) of the coefficient of the equation in the QSAR equation that will eventually guide the need to strengthen/enhance (i.e. shift the interaction energy value towards more negative) or weaken/reduce (i.e. shift the interaction energy value towards more positive) the interaction, in order to favor binding.

An examination of the frequency plot of descriptors appearing in the eCoRIA model (Figure 2) indicates that the Coulombic (/electrostatics), van der Waals (/steric), hydrogen bonding interactions, and entropy are the major forces driving the binding of the ligand with the GPb enzyme. The terms describing steric interaction of the ligand with residues Tyr:280, His:377, and Glu:287 in GPb appear with a negative coefficient in the equations. This indicates that an increase in the biological activity can be achieved by strengthening the steric interaction of the ligand with these residues. On the other hand, the steric interaction with residues Asn:284 and Ser:674 has a positive coefficient in the equations suggesting that this interact needs to be toned down to improve the biological activity. The Coulombic interaction energy of residue Gly:135 has a negative coefficient in the equation, indi-

Table 2. Statistical parameters of eCoRIA, CoRIA, CoMFA, and CoMSIA models derived for the GPb data-set.

| Technique | $r^2$ | $q^2_{LOO}$ | $q^2_{L5O}$ | $q^2_{adj}$ | $r^2_{pred}$ | $r^2_{hs}$ | $r^2_{rand}$ | $r^2_m$ | $r^2_p$ |
|-----------|-------|-------------|-------------|------------|-------------|-----------|------------|--------|--------|
| eCoRIA    | 1     | .93         | .81         | .73        | .75         | .65       | .91        | .19    | .60    |
|           | 2     | .92         | .83         | .74        | .78         | .64       | .91        | .19    | .61    |
|           | 3     | .92         | .81         | .71        | .75         | .61       | .91        | .20    | .58    |
| CoRIA     | 1     | .72         | .41         | .32        | .25         | .36       | .71        | .33    | .27    |
|           | 2     | .71         | .39         | .30        | .22         | .30       | .71        | .33    | .21    |
|           | 3     | .73         | .43         | .33        | .27         | .32       | .72        | .35    | .24    |
| CoMFA     | 1     | .75         | .48         | .40        | .33         | .70       | .34        | .26    | .48    |
|           | 2     | .74         | .42         | .39        | .31         | .71       | .32        | .22    | .47    |
|           | 3     | .71         | .43         | .32        | .31         | .70       | .35        | .28    | .42    |
| CoMSIA    |       | .72         | .41         | .37        | .31         | .70       | .33        | .28    | .44    |
|           |       | .76         | .41         | .36        | .32         | .72       | .35        | .23    | .48    |
|           |       | .73         | .43         | .39        | .30         | .72       | .31        | .27    | .47    |

Note: Where $r^2$: correlation coefficient; $q^2_{LOO}$ and $q^2_{L5O}$ are the cross-validation correlation coefficient by LOO and L5O procedure, respectively; $r^2_{pred}$ is the predictive correlation coefficient; $r^2_{hs}$ and $r^2_{rand}$ are the correlation coefficient obtained by bootstrapping and y-randomization, $r^2_m$ and $r^2_p$ are the validation parameters calculated as per Roy et al.

Figure 2. Frequency plots of descriptors appearing in the 500 equations leading to the eCoRIA model.
cating that an overall negative value of the Coulombic interaction energy of this residue with the ligand will improve binding. On the contrary, the Coulombic interaction energy with residues Arg:292 and Asp:283 have a positive coefficient in the equations suggesting that an overall positive value of the Coulombic interaction energy of these residues with the ligand will favor binding. The term describing the entropy of the system appears with a positive coefficient, which suggests that we need to increase the entropy of the system in order to favor the binding process. It also indicates that the flexibility of the molecules needs to be increased. Also, the hydrogen bonding interaction of the ligand with the residue His:377 should be strengthened to ensure tighter binding due to the negative coefficient of this interaction energy term in the equations. Although the terms describing hydrophobic interaction appear transiently during the evolution, they gradually disappear as the genetic function progresses. However, these terms do appear in the equations when the number of terms in the equation is increased to 10 or more. It is important to mention that no bias was placed on any of the descriptors and all had an equal weightage during development of the eCoRIA equations. A schematic representation of the active site of GPb in complex with an inhibitor (the most active molecules in the data-set-2prj) is shown in Figure 3. Interestingly, the eCoRIA models could successfully pick out the residues involved in the catalytic activity of the enzyme. Such observations have also been described in the literature.

The ultimate goal of eCoRIA is to capture the shape of the energy landscape (i.e. the distribution of variables over the ensemble), which is crucial and may not be observable when a single lowest energy static structure is used for QSAR studies. eCoRIA analysis could successfully capture this distribution of properties as is manifested in the terms that appear in the QSAR equations, e.g. skewness ($M_3$) of the steric interaction of the ligand with Tyr:280 is shown have a strong correlation with the activity. This indicates that the skewness of the steric interaction energy of the ligands with Tyr:280 is one factor for the variation in the biological activity of the ligands. That the asymmetry of the interaction of ligands with a specific residue may be responsible for the variation in their activity is evident from this result. Likewise, the variance ($M_2$) around the mean reflects the

![Figure 3. A diagrammatic representation of the active site of GPb in complex with an inhibitor showing the key residues extracted by the eCoRIA models.](image-url)
distribution of Coulombic interaction of Gly:135 with the ligands and the average ($M_1$) of the hydrogen bonding interaction of His377 are the causes for the variation in biological activity of ligands (Figure 2).

These results indicate that the contribution of each microscopic state to the variation in activity cannot always be weighted by the average value. Thus static structure analyses from ensemble averaged measurements at equilibrium which has been the norm in understanding protein structure and functions needs to be replaced with true ensemble representations to reflect the distribution of the thermodynamic properties over the energy landscape.

Comparison with other 3D-QSAR methodologies

The $e$CoRIA methodology has been compared with our standard CoRIA (receptor-based QSAR), CoMFA, and CoMSIA (receptor-independent QSAR) methods for the same training and test sets. Although these formalisms embody some information contained in the $e$CoRIA methodology, in many ways the $e$CoRIA formalism is significantly different. One important point of difference is that these methodologies rely on static structure analysis, while $e$CoRIA models the biological activity as a function of molecular descriptors derived from an ensemble of discrete microscopic states. The statistics of all these approaches are summarized in Table 2. It is observed that the $e$CoRIA models exhibit significantly better statistics over the other three methodologies as manifested by the basic statistical parameters ($r^2$, $q^2_{LOO}$ and $q^2_{L5O}$). Also, the $e$CoRIA models perform unvaryingly better than other approaches in the external prediction ($r^2_{pred}$). The metrics devised by Roy et al. further establish the statistical significance of $e$CoRIA models over the other approaches. It is noteworthy that though the standard CoRIA models are derived using the same thermodynamic properties incorporated in the $e$CoRIA formalism, an ensemble representation of these properties appears to be a superior mode of describing the activity. The CoMFA and CoMSIA techniques calculate intermolecular interaction energies indirectly by considering probes at each grid point around an aligned static conformation of the molecules, which does not directly reflect the dynamics occurring in the ligand–receptor interactions.

There are few QSAR studies reported for GPb inhibitors in the literature. A data-set of GPb inhibitors, which are analogs of glucose, were investigated by Hopfinger et al. using receptor-independent (RI) 4D-QSAR analysis (Venkatarangan & Hopfinger, 1999). The method uses grid cell occupancy descriptors for defining the thermodynamic averaged spatial pharmacophore. The QSAR model derived using this approach produced a $q^2$ of .81–.83 and an $r^2$ of .86–.87. In comparison to these statistics, the $e$CoRIA models have comparable $q^2$ (.81–.83) but relatively better $r^2$ (.92–.93) values. In another ligand-based QSAR approach termed LISA (Verma, Malde, Khedkar, Iyer, & Coutinho, 2009), a data-set of GPb inhibitors were investigated wherein the global molecular similarity is broken up as local similarity at each grid point surrounding the molecules and is used as a QSAR descriptor. It is observed that the $e$CoRIA statistics surpass those derived by LISA in terms of the internal cross-validation ($q^2$) as well as in the external prediction ($r^2_{pred}$). In yet another report, a RD 4D-QSAR analysis was performed by Hopfinger et al. (Pan et al., 2003) on the same data-set of GPb inhibitors as used in the RI 4D-QSAR analysis. A pruned ligand–receptor model was used to estimate ligand–receptor thermodynamics in the functional region of the protein and explicit water molecules were excluded in this study to simplify the analysis of the ligand-binding process. The statistics of the $e$CoRIA model was found to be comparable to Hopfinger’s RD 4D-QSAR ($q^2 = .82$, $r^2 = .85$).

Data-set 2: human immunodeficiency virus-1 protease (HIV-1 PR)

HIV-1 protease is an obligatory enzyme in the replication of the HIV. It cleaves the polyproteins transcribed from the gag and pol genes into enzymes and structural proteins essential for the assembly and maturation of infectious virions. The finding that inactivation of this viral encoded aspartyl protease produces a progeny of virions that are immature and non-infectious, elicited intense efforts in the development of specific and potent inhibitors targeted against this enzyme, as a novel therapy for AIDS. HIV-1 protease has been the subject of interest for a number of molecular modeling endeavors, which prompted us to test the $e$CoRIA methodology against this target as well. A data-set of 72 enzyme–inhibitor complexes of HIV-1 protease were compiled from the RCSB protein data bank and used as a test bed for the $e$CoRIA methodology. It was then segregated into a training set for model generation and a test set for model validation on the basis of chemical and biological diversity using similarity search techniques. The PDB IDs of these molecules together with their experimentally determined activities are summarized in Table 2S. The QSAR equations developed using $e$CoRIA and those for the standard CoRIA approach are compared in Table 3. Cross-validation using LOO returned $q^2 > .70$, while the L5O procedure produced a slightly lower yet significant value. When the $e$CoRIA equation was used to predict the activity of the test set, the predictive correlation co-efficient was found to be greater than .60 demonstrating good external predictability. The correlation plots of experimental vs. predicted pKi values of the molecules in the training and test sets for the best $e$CoRIA model are shown in Figure 2S. Also an $r^2_{LOO}$ value, close to the
Table 3. The eCoRIA and CoRIA models for the HIV-PR protease-inhibitors data-set.

| Technique | QSAR equation |
|-----------|----------------|
| eCoRIA    |                |
|           | pKᵢ = 0.24 – 0.75 Asp:B25: Coul + 4.82 Pro:B9: vdW + 3.36 Thr:A80: vdW + 3.80 Met:A46: HINT + 1.33 Leu:A90: HINT + 3.92 Thr:B80:vdW + 0.59 Thr:B26: Coul + 1.66 Val:B82:vdW + 0.92 Ile:A50:vdW + 0.81 Met:A46: HINT + 3.33 Thr:A80:vdW + 0.58 Thr:B26: Coul + 0.11 – 0.72 Asp:B25: Coul + 4.78 Pro:B9: vdW + 1.34 Leu:A90: HINT: M₃ + 3.99 Ile:A50:vdW + 4.01 Thr:B80:vdW + 1.62 Val:B82: vdW + 0.81 Met:A46: HINT + 3.33 Thr:A80:vdW + 0.58 Thr:B26: Coul + 1.15 – 4.09 Pro:B9:vdW + 3.13 Thr:B80:vdW + 6.105 Thr:A80:vdW + 0.395 Ala:A28: Coul + 3.42 Gly:A52:vdW + 1.74 Leu:A90: HINT + 0.94 Met:A46: HINT:M₃ + 0.62 Leu:B76: Coule + 0.18 solvation M₃ |
| CoRIA     |                |
|           | pKᵢ = 8.69 + 0.92 Ile:A84: Coul + 0.71 Gly:B78: Coule + 0.91 Leu:A23: Coule + 0.46 Pro:B79: HINT: + 0.44 Pro:B9: HINT: + 0.43 Asn:B83: vdW + 0.54 Asp:B30: HINT: + 0.29 Ile:A50: Coul + 8.77 + 0.42 Leu:A10: HINT: + 0.32 Thr:B74: vdW + 0.67 Asp:B25: HINT: + 0.35 Ile:B84: vdW + 1.26 Gly:B78: vdW + 0.54 Ile:A50: vdW + 0.90 Thr:B80: vdW + 0.30 Asn:B88: Coul + 8.75 – 0.75 Thr:B80: vdW + 0.60 Gly:A51: vdW + 0.55 Asp:B25: HINT: + 0.38 Lys:A45: HINT: + 0.43 Asp:B30: HINT: + 1.18 Gly:B78: vdW + 0.42 Arg:A87: vdW + 0.39 Thr:B74: vdW |

Note: HIV PR being a dimer, A and B represent the two chains of the enzyme.
For the training set, obtained on 100 bootstrap runs, suggests a good internal consistency within the underlying data-set. Overfitting and chance correlation between the activity and descriptors were gauged by employing y-randomization. Poor regression models with low $r^2 (<.20)$ were obtained indicating that the models are stable and the results are not based on chance correlation. The statistical parameters of the best QSAR models are summarized in Table 4. Comparison with the standard CoRIA, CoMFA, and CoMSIA approaches reveal that the QSAR models derived using eCoRIA formalism are significantly better in terms of $r^2$ and predictive capability ($r^2_{pred}$).

In this study, in addition to building a sound QSAR model with good predictive power, identifying the imperative ligand–receptor interactions were also of high interest. To facilitate realization of these objectives, all 500 equations associated with the best QSAR model were examined for the frequency with which a descriptor occurs in the population of equations (Figure 4). The more frequently a descriptor is used in generating new models, the greater is its relative significance in explaining variance in the biological activity. The observations are summarized in the Table 5. A schematic representation of the active site of HIV-PR in complex with an inhibitor is shown in Figure 3S. The eCoRIA models could capture the residues involved in the catalytic activity (e.g. Asp:B25, Pro:B9, Leu:A90, Thr:A80) of the enzyme.

In a computational study by Holloway and co-workers (Holloway et al., 1995), it was found that the intermolecular van der Waals and electrostatic interactions between HIV-1 protease and a series of inhibitors could explain the variance in the observed binding affinities of these ligands. The eCoRIA approach could capture in addition to these properties, the contribution from other thermodynamic properties as well (e.g. hydrophobic interaction). Our previous report on molecular docking and 3D-QSAR studies on a series of cyclic urea analogs with HIV-PR inhibitory activities has some parallel observations to the present eCoRIA model. Thus, crucial interactions with Asp:B25, Pro:B9, Leu:A90, and Thr:A80 residues that influence the biological activity have been captured in both studies (Khedkar et al., 2010). In another report, COMBINE analysis was performed on a data-set of HIV-protease inhibitor complexes, wherein the ligand–receptor interaction energies were broken down into the individual van der Waals and electrostatic contributions (Perez, Pastor, Ortiz, & Gago, 1998). The statistical outcomes of eCoRIA model are found to be comparable with this analysis, where a $q^2$ and $r^2$ of .72 and .90 were obtained for the COMBINE model, the best eCoRIA model produced values of .75 and .82, respectively.

Data-set 3: cyclin-dependent kinase 2

Cyclin-dependent kinase 2 (CDK2) also known as cell division protein kinase 2 is a member of the cyclin-dependent kinase family of Ser/Thr protein kinases, which regulate the cell cycle progression in proliferating eukaryotic cells. Aberrant CDK control and consequent loss of cell cycle checkpoint function have been directly correlated to the molecular pathology of cancer. In view of the strong link between CDKs and cancer, CDKs are being investigated as a potential target for anticancer activity.

Table 4. Statistical analysis of the QSAR models for the eCoRIA, CoRIA, CoMFA, and CoMSIA approaches.

| Technique | $r^2$ | $q^2_{LOO}$ | $q^2_{LO}$ | $q^2_{adj}$ | $r^2_{pred}$ | $r^2_{bs}$ | $r^2_{rand}$ | $r^2_m$ | $r^2_{pred}$ |
|-----------|-------|--------------|------------|-------------|--------------|------------|--------------|--------|--------------|
| eCoRIA    |       |              |            |             |              |            |              |        |              |
| 1         | .82   | .75          | .67        | .71         | .63          | .81        | .11          | .61    | .69          |
| 2         | .82   | .75          | .66        | .71         | .61          | .81        | .12          | .58    | .68          |
| 3         | .82   | .70          | .62        | .65         | .63          | .80        | .18          | .59    | .66          |
| CoRIA     |       |              |            |             |              |            |              |        |              |
| 1         | .78   | .41          | .39        | .32         | .24          | .70        | .30          | .27    | .41          |
| 2         | .77   | .48          | .34        | .40         | .25          | .71        | .47          | .21    | .42          |
| 3         | .79   | .44          | .36        | .35         | .35          | .73        | .46          | .25    | .45          |
| CoMFA     |       |              |            |             |              |            |              |        |              |
| 1         | .84   | .49          | .40        | .19         | .83          | .34        | .17          | .59    | .59          |
| 2         | .84   | .48          | .44        | .22         | .81          | .34        | .21          | .59    | .59          |
| 3         | .85   | .50          | .47        | .27         | .81          | .38        | .22          | .58    | .58          |
| CoMSIA    |       |              |            |             |              |            |              |        |              |
| 1         | .85   | .42          | .40        | .18         | .82          | .32        | .20          | .61    | .61          |
| 2         | .84   | .39          | .40        | .20         | .80          | .35        | .19          | .58    | .58          |
| 3         | .83   | .45          | .42        | .36         | .81          | .32        | .29          | .59    | .59          |
medication. Studies with first generation CDK inhibitors support a role for the use of these agents as both single agents or in combination with cytotoxic drugs. A data set of 81 enzyme–inhibitor complexes of CDK2 were compiled from the protein data bank for testing the eCoRIA approach. The PDB IDs for the molecules are summarized in Table 3S together with their experimentally determined activities. The regression equations and statistical analysis of the eCoRIA models and its comparison with CoRIA, CoMFA, and CoMSIA approaches are provided in Tables 6 and 7, respectively. The correlation plot of experimental vs. predicted pIC50 values of the training and test set molecules for the best eCoRIA model is shown in Figure 4S. Further, the derived eCoRIA models have good predictive ability and provide insight into the features governing the enzyme inhibition.

There are several reports of QSAR studies on CDK2 inhibitors. The CoMFA and CoMSIA models developed by Lan et al. for a series of pyrazolo[4,3-h]quinazoline-3-carboxamides have $q^2$ values of .64 and .50, $r^2$ values of .95 with predictive correlation coefficients of .50 and .46 respectively (Lan, Chen, Xiao, Sun, & Chen, 2010). Samanta et al. have reported a 2D-QSAR study on a data-set of 3-aminopyrazoles analogs to explore the pharmacophoric requirement to exhibit CDK2 inhibitory activity (Samanta et al., 2006). The best models were obtained using topological parameters, indicator variables, and atomic charges have $r^2$ of .88–.91 and $r^2_{pred}$ of
Table 6. The eCoRIA and CoRIA models for the CDK2-inhibitors data-set.

| Technique | QSAR equation |
|-----------|----------------|
| eCoRIA    | pIC_{50} = .50 + .74 Phe:80·Coul M_{1} + .26 Asp:86·HINT M_{3} + .95 Ala:31·Coul M_{1} + 1.56 Val:18·vdW M_{1} + .46 Phe:82·vdW M_{2} + .56 Leu:83·Hbond M_{1} + .59 Lys:89·vdW M_{1} + .88 Gly:147·Coul M_{2} + .32 Leu:298·Coul M_{3} |
|          | pIC_{50} = .12 + .49 Ala:31·Coul M_{2} + .49 Leu:83·Hbond M_{2} + 1.51 Val:18·vdW M_{2} + .25 Asp:86·HINT M_{3} + .52 Phe:82·vdW M_{2} + .82 Phe:80·Coul M_{3} + 1.9 Glu:85·HINT M_{2} + .21 Phe:80·vdW M_{2} + .46 Lys:89·vdW M_{1} |
|          | pIC_{50} = .29 + .23 Asp:86·HINT M_{3} + 1.39 Val:18·vdW M_{2} - .42 Phe:82·Coul M_{1} + .75 Phe:80·Coul M_{3} - .60 Phe:82·vdW M_{2} + .15 Glu:85·HINT M_{3} - .48 Lys:89·vdW M_{1} - .40 Lys:33·vdW M_{1} |
| CoRIA    | pIC_{50} = 6.75 - .42 Val:18·Coul - .57 Asp:145·Hbond - .48 Leu:134·vdW - .38 Gly:11·HINT - .40 Lys:33·vdw - .36281·Thr:14·vdw - .19 Gly:13·Coul - .27 Glu:81·HINT - .22 Phe:82·HINT |
|          | pIC_{50} = 6.94 - .30 Val:18·vdw - .25 Glu:85·HINT - .26 Leu:134·vdw - .49 Gly:11·Coul - 1.29 Lys:33·Coul - 1.85 Glu:12·Coul - 42 Leu:83·Coul - 1.3 Gly:13·Coul - 22 Lys 33·vdw |
|          | pIC_{50} = 6.72 + .43 Gly:11·Coul - .46 Leu:83·Coul - .29 Glu:85·vdw - .73 Gly 147·vdw - .40 Gly 11: vdw - .64 Val:18·Coul + .42 Lys:33·vdw + .31 Asp:145·Hbond + .30 Leu:134: vdw |

Note: Where M_{1}, M_{2}, M_{3}, M_{4} are the moments of distribution; M_{1} = Mean; M_{2} = Variance; M_{3} = Skewness; M_{4} = Kurtosis; Coul, vdW, Hbond and HINT – Coulombic, van der Waals, Hydrogen bonding and Hydrophobic interactions respectively.

Table 7. Statistical analysis of the QSAR models derived using eCoRIA, CoRIA, CoMFA, and CoMSIA approaches.

| Technique | r^2 | q^2_{LOO} | q^2_{LSO} | q^2_{adj} | r^2_{pred} | r^2_{bs} | r^2_{rand} | r_m^2 | r_p^2 |
|-----------|-----|-----------|-----------|-----------|------------|----------|------------|-------|-------|
| eCoRIA    |     |           |           |           |            |          |            |       |       |
| 1         | .92 | .84       | .76       | .81       | .68        | .91      | .19        | .60   | .78   |
| 2         | .91 | .87       | .75       | .84       | .68        | .91      | .21        | .61   | .76   |
| 3         | .91 | .87       | .74       | .84       | .66        | .90      | .20        | .60   | .77   |
| CoRIA     |     |           |           |           |            |          |            |       |       |
| 1         | .69 | .42       | .40       | .32       | .38        | .65      | .41        | .32   | .37   |
| 2         | .70 | .46       | .41       | .37       | .39        | .69      | .37        | .32   | .40   |
| 3         | .72 | .44       | .38       | .35       | .33        | .70      | .42        | .25   | .39   |
| CoMFA     |     |           |           |           |            |          |            |       |       |
| 1         | .71 | .44       | .40       | .31       | .70        | .37      | .29        | .46   | .46   |
| 2         | .72 | .41       | .39       | .33       | .70        | .34      | .30        | .47   | .47   |
| 3         | .71 | .42       | .40       | .31       | .71        | .31      | .28        | .47   | .47   |
| CoMSIA    |     |           |           |           |            |          |            |       |       |
| 1         | .70 | .43       | .41       | .33       | .69        | .38      | .31        | .43   | .43   |
| 2         | .71 | .40       | .36       | .31       | .71        | .33      | .27        | .47   | .47   |
| 3         | .73 | .44       | .41       | .32       | .70        | .34      | .29        | .48   | .48   |

.66–.74. Caballero et al. have reported the use of artificial neural network for modeling the cyclin-dependent kinase inhibition of 1H-pyrazolo[3,4-d] pyrimidine derivatives, which produced models with r^2 of .79–.89 and q^2 of .59–.64 (Fernandez, Tundidor-Camba, & Caballero, 2005). The statistics derived for the eCoRIA models are significantly better than these studies both in terms of internal validation (q^2 > .8, r^2 > .9) as well as external predictivity (r^2_{pred} > .65). Jans et al. have investigated the CDK2 inhibition potential of N2- and O6-substituted guanine derivatives using molecular docking and the hybrid ONIOM method. Using the best ligand poses obtained from docking as an input, a correlation model was derived between the ONIOM-derived quantum chemical descriptor “H-bond interaction energy” and the experimental biological activity, with a r value of .80 (Alzate-Morales, Caballero, Vergara Jague, & Gonzalez Nilo, 2009). They concluded that the residues at the entrance of the binding site (Lys:89 and Asp:86) were the main contributors to the total interaction energy between the protein and the ligand. The eCoRIA models shed more light on these interactions which recommends that increasing the van der Waals interactions of the ligands with Lys:89, while reducing the hydrophobic interaction with Asp:86 will improve binding. In another report, the same authors have used CoMFA to describe the activities of guanine analogs. The best CoMFA model derived with steric and electrostatic fields had q^2
of .80–.84 and $r^2$ of .94–.99 (Alzate-Morales & Caballero, 2009). Fischer et al. have developed a 3D-QSAR model using an active site pressurization (ASP) approach for two CDK2 data-sets containing 65 purines and 91 oxindoles (Mazanetz, Withers, Laughton, & Fischer, 2009; Withers, Mazanetz, Wang, Fischer, & Laughton, 2008). These ASP-QSAR models produced a non-cross-validated correlation coefficient ($r^2$) of .95–.99 and a cross-validated correlation coefficient ($q^2$) between .927 and .929. Comparison with these reported QSAR studies indicates that the eCoRIA models have an equivalent or in some cases better statistics in terms of $r^2$, $q^2$, and $r^2_{\text{pred}}$ values. Also, it is noteworthy that the $r^2_{\text{pred}}$ reported for eCoRIA models is derived from a larger and structurally diverse test set, while the above quoted studies were limited to a single chemical class.

A close appraisal of all eCoRIA models over the three data-sets reveals that some of the interactions shown to be decisive in these studies are distant from the ligand (~10 Å), which indicates that along with the direct interactions, indirect long-range interactions also contribute significantly towards the stability of the ligand–receptor complexes.

### Discussion

A myriad of techniques for analyzing the energetics of enzyme–inhibitor interactions have been reported in the literature and some of them have been able to derive meaningful structure activity relationships for specific systems. However, it is observed that many of these modeling techniques are based on two approximations.

The first involves the approximations necessary to compute the free energies rapidly which usually means ignoring some components of the binding energy such as solvation and configurational entropy (Ajay & Murcko, 1995; Dominy & Brooks, 1999). It has been shown that solvation and desolvation effects inherent to ligand as

| Term and associated residues | Sign of the coefficient in the equation | Suggestions |
|-----------------------------|----------------------------------------|-------------|
| van der Waals interaction energies: Phe:82, Val:18, Lys:89 and Lys:33 | Negative | Increase in the biological activity can be achieved by strengthening the van der Waals interactions of the ligands with these residues. |
| hydrophobic interactions (HINT) Asp:86 | Negative | Reducing the hydrophobic interaction with this residue would favor the inhibition. |
| hydrophobic interactions (HINT) Gln:85 | Positive | Enhancing the hydrophobic interaction will improve the binding with the receptor. |
| Coulombic interaction energy: Leu:298, Phe:82 and Ala:31 | Negative | An overall negative value of the Coulombic interaction energy of these residues with the ligand will favor binding. |
| Coulombic interaction energy: Phe:80, Gly:147 | Positive | The strength of the Coulombic interaction between the ligand and the receptor residues Phe: 80 and Gly: 147 needs to be reduced for tighter ligand binding. |
| hydrogen bonding interactions Leu:83 | Negative | Strengthening the hydrogen bonding interaction with this residue will favor the binding. |
well as receptor can have significant impact on the strength of binding and the corresponding measurement of the resultant biological activity. Also, negligence of explicit water molecules during the MDS can lead to the generation and sampling of artifact states relative to the actual binding mode. The loss of configurational entropy upon binding also contributes significantly to the binding affinity. The restriction of a flexible molecule’s motion upon binding to a protein causes a loss of conformational entropy. Some energy models used in computer-aided molecular design neglect this entropic penalty in the interest of time. However, such a contribution could be especially important in specific systems such as for HIV protease, where binding is associated with a marked reduction in the mobility of the active site flaps. It is possible that other high-affinity protein–ligand systems may undergo similarly large losses of conformational entropy on binding. The eCoRIA methodology presented here tries to account for both these terms along with the typical energy components used in other RD QSAR approaches. Interestingly, the eCoRIA models derived for the GPb data-set could capture conformational entropy as an important thermodynamic property governing the ligand–receptor binding, while the term explaining solvation energy was observed in the eCoRIA equations obtained for the data-set of HIV-PR inhibitors. Other properties like strain energy also appear transiently during the evolution but gradually disappear as the genetic function progress i.e. their frequency of occurrence is too low to be considered as important relative to other interaction terms for ligand binding. One explanation for the absence of these terms in the final equations could be due to the fact that for the data-sets evaluated in the present study, these terms may not be a major determinant in the overall binding process, so that they can be picked up more frequently by the statistical tool. Most of the 3D-QSAR techniques reported till date have only probed and extracted information about a specific interaction involving the ligand (e.g. CoMFA considers only electrostatic and vdW interactions) and this is mostly restricted to its binding site (i.e. receptor pruning), while neglecting critical elements of the ligand-binding process. The receptor pruning approach has been adopted mainly to reduce the time and computational resources, assuming that the enzyme–inhibitor interactions are relatively short range. However, in order to gain a meaningful thermodynamic profile of protein–ligand systems in QSAR, it is necessary to adopt a non-grid approach to incorporate the long-range interactions. Computing the interactions in free space as opposed to using the truncated receptor (as in 4D-QSAR) for the analysis is well able to account for long-range electrostatic interactions. Interestingly, as discussed in the results sections, many of the interactions shown to be decisive in these studies were distant from the ligand, indicating that along with the direct interactions, indirect long-range interactions also contribute significantly towards the stability of the ligand–receptor complexes.

The second type of approximation includes those involved in treating the enzyme–inhibitor complex as a static structure (Dominy & Brooks, 1999; Lightstone, Zheng, & Bruce, 1998). This assumption runs contrary to a growing amount of research indicating that the lowest energy conformation is least likely to represent the binding conformation with solvation and binding interactions capable of compensating for energy differences among the conformers of a molecule. The crystal structure is an average static representation of the dynamic behavior of a ligand–receptor complex and therefore contains limited information on possible conformational changes and corresponding coupled motions occurring in the inhibitor–receptor-binding event. The eCoRIA formalism does away with this approximation by eliminating the assumption of a static ligand–receptor complex and measures the energies over ensembles of structures, which is a major component to the “fourth” dimension in QSAR and hence eCoRIA may be regarded as a 4D-QSAR approach. Analysis of the statistical distribution of each energy component over an ensemble of conformations provides a more robust picture of binding than can be achieved by a single-structure representation. The width of the energy distributions is crucial in correlating with the biological result as evidenced in this study. Moments of distribution computed for the thermodynamic properties gives an account of the spread of these properties over the entire conformational space. This is the first time, as far we know, that a QSAR methodology extracts a feature associated with an ensemble of conformations and packages the dynamic information to establish a correlation equation.

The goal of eCoRIA analysis was to accurately translate the most significant thermodynamic properties of a set of ligand–receptor complexes into a matrix of informative variables to derive a correlation between these properties and the biological activity. The partitioning scheme adopted in eCoRIA analysis allows the categorization of the thermodynamics of binding into many individual energetic contributions and therefore provides a wealth of information that fully account for the biological activity. The data matrix derived from this fragmentation approach could capture the essence of all the structural changes in the molecules and the resulting quantitative model highlights those variables that are crucial for modulating the biological activity.

Conclusion

Although the concept of induced fit in enzyme–inhibitor binding has long been recognized, only a few studies have been published that address the nature and the
energetics of the conformational changes that the small molecules undergo upon binding. In this paper, we have presented a methodology for analysis of protein–ligand association by modifying the current dogma in computational chemistry of “one chemical–one structure–one parameter value” to a “one chemical–finite set of structures–range of descriptors values” principle. The basic idea was derived from the fact that relatively small differences between conformers of a chemical are known to yield significant variation in the properties of the molecule. We have treated the enzyme–inhibitor complex as an ensemble of energetically reasonable conformations and examined the influence of ligand–receptor binding thermodynamics on activity by analyzing the problem as a QSAR calculation. This approach is distinctly different from methods that select a single lowest energy conformation of a molecule or those that average the descriptors over a set of molecular conformations.

Development of mechanistically sound SARs for receptor-mediated events requires one to account for the dynamic behavior of the complex as well as the selection of appropriate descriptors to represent the molecule under investigation. The scan of the potential energy surface by MDSs has validated that this MD-based QSAR approach provides a solution to the uncertainty in deducing the bioactive conformation of small molecules in Q SAR. Moments of distribution computed for the thermodynamic events representing ligand–receptor interactions could provide insight into the influence of the potential energy landscape on the activity that is not captured in a single crystal structure.

Each application of eCoRIA analysis reported here has resulted in QSAR models with good statistics. The Q SAR equations have revealed all structure activity relationships previously reported for these molecules as well as uncovered some features that were hitherto unknown and may have a hidden role in driving the ligand–receptor binding process. The predictive ability of eCoRIA analysis compares very favorably with that of other methods which make more conventional use of molecular mechanics interaction energies, such as COMBINE and FEFF 4D-QSAR (Santos-Filho, Mishra, & Hopfinger, 2001; Tokarski & Hopfinger, 1997).

Decomposition of the interaction energies into van der Waals, electrostatics, hydrogen bonding, hydrophobic, solvation, entropy, and strain energy terms enabled us to examine protein–ligand association in even greater detail. The former interaction energies are associated directly with a specific residue in the protein’s active site and can advantageously be used to extract the residues that are important for binding and also serve as a guide for carrying out mutation studies directed towards understanding ligand binding. This adds to the attempts made by few others to include the influence of conformational entropy, solvation, and strain energy altogether along with the non-bonded interactions on biological activity.

In summary, by combining MD-based conformation analysis with quantitative structure–activity relationship analysis, eCoRIA promises to be a valuable extension to the classical nD-QSAR methodologies, both conceptually and also with respect to reliability in predicting the affinities of new chemical entities.

Funding
The authors are grateful to the Department of Biotechnology (BT/PR11810/BRB/10/690/2009) and the Council of Scientific and Industrial Research (01/23/99/10-EMR-II) for financial support.

Supplementary material
The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2014.909744.

References
Ajay & Mureko, M. A. (1995). Computational methods to predict binding free energy in ligand–receptor complexes. Journal of Medicinal Chemistry, 38, 4953–4967.
Alzate-Morales, J., & Caballero, J. (2009). Computational study of the interactions between guanine derivatives and cyclin-dependent kinase 2 (CDK2) by CoMFA and QM/MM. Journal of Chemical Information and Modeling, 50, 110–122.
Alzate-Morales, J. H., Caballero, J., Vergara Jague, A., & Gonzalez Niño, F. D. (2009). Insights into the structural basis of N2 and O6 substituted guanine derivatives as cyclin-dependent kinase 2 (CDK2) inhibitors: Prediction of the binding modes and potency of the inhibitors by docking and ONIOM calculations. Journal of Chemical Information and Modeling, 49, 886–899.
Andrade, C. H., Pasquado, K. F., Ferreira, E. I., & Hopfinger, A. J. (2010). 4D-QSAR: Perspectives in drug design. Molecules, 15, 3281–3294.
Araújo, J. Q., de Brito, M. A., Hoelz, L. V. B., de Alencastro, R. B., Castro, H. C., Rodrigues, C. R., & Albuquerque, M. G. (2011). Receptor-dependent (RD) 3D-QSAR approach of a series of benzylpiperidine inhibitors of human acetylcholinesterase (HuAChE). European Journal of Medicinal Chemistry, 46, 39–51.
Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinolata, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics, 81, 3684.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. Nucleic Acids Research, 28, 235–242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., … Tassini, M. (1977). The protein data bank. A computer-based archival file for macromolecular structures. European Journal of Biochemistry, 80, 319–324.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr, E. F., Brice, M. D., Rodgers, J. R., … Tassini, M. (1978). The protein data bank: A computer-based archival file for macromolecular structures. Archives of Biochemistry and Biophysics, 185, 584–591.
Brandsdal, B. R. O., Osterberg, F., Almlof, M., Feierberg, I., Luzhkov, V. B., & Aqvist, J. (2003). Free energy calculations and ligand binding. *Advances in Protein Chemistry, 66*, 123–158.

Cramer III, R. D., Patterson, D. E., & Bunce, J. D. (1988). Comparative molecular field analysis (CoMFA). I. Effect of shape on binding of steroids to carrier proteins. *Journal of the American Chemical Society, 110*, 5959–5967.

Datar, P. A., Khedkar, S. A., Malde, A. K., & Coutinho, E. C. (2006). Comparative residue interaction analysis (CoRIA): A 3D-QSAR approach to explore the binding contributions of active site residues with ligands. *Journal of Computer-Aided Molecular Design, 20*, 343–360.

Deep, R. (2006). Regression. In R. Deep (Ed.), *Probability and statistics* (pp. 455–515). London: Academic Press.

Dominy, B. N., & Brooks, C. L. (1999). Methodology for protein-ligand binding studies: Application to a model for drug resistance, the HIV/FIV protease system. *Proteins: Structure, Function, and Bioinformatics, 36*, 318–331.

Doweyko, A. M. (1991). The hypothetical active site lattice – In vitro and in vivo explorations using a three-dimensional QSAR technique. *Journal of Mathematical Chemistry, 7*, 273–285.

Duca, J. S., & Hopfinger, A. J. (2001). Estimation of molecular similarity based on 4D-QSAR analysis: Formalism and validation. *Journal of Chemical Information and Computer Science, 41*, 1367–1387.

Dunn III, W. J., & Rogers, D. (1996). Genetic partial least squares in QSAR. In J. Devillers (Ed.), *QSAR & Combinatorial Science, 24*, 240–245.

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics, 103*, 8577.

Fernandez, M., Tundidor-Camba, A., & Caballero, J. (2005). Modeling of cyclin-dependent kinase inhibition by 1 H-pyrazolo [3, 4-d] pyrimidine derivatives using artificial neural network ensembles. *Journal of Chemical Information and Modeling, 45*, 1884–1895.

Fontes Ferreira da Cunha, E., Albuquerque, M. G., Ceva Antunes, O. A., & Bicca de Alencastro, R. (2005). 4D-QSAR models of HOE/BAY-793 analogues as HIV-1 protease inhibitors. *QSAR & Combinatorial Science, 24*, 240–253.

Friessner, R. A., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., … Shenkin, P. S. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry, 47*, 1739–1749.

Friessner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., … Mainz, D. T. (2006). Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. *Journal of Medicinal Chemistry, 49*, 6177–6196.

Gancia, E., Bravi, G., Mascagni, P., & Zaliani, A. (2000). Global Brandsdal, B. R. O., Osterberg, F., Almlof, M., Feierberg, I., Luzhkov, V. B., & Aqvist, J. (2003). Free energy calculations and ligand binding. *Advances in Protein Chemistry, 66*, 123–158.

Holloway, M. K., Wai, J. M., Halgren, T. A., Fitzgerald, P. M., Vacca, J. P., Dorsey, B. D., … deSolms, S. J. (1995). A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *Journal of Medicinal Chemistry, 38*, 305–317.

Hopfinger, A. J. (1980). A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *Journal of the American Chemical Society, 102*, 7196–7206.

Hopfinger, A. J., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J., & Duraiswami, C. (1997). Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *Journal of the American Chemical Society, 119*, 10509–10524.

Jorgensen, W. L., & Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society, 110*, 1657–1666.

Jorgensen, W. L., Maxwell, D. S., & Tirado-Rives, J. (1996). Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society, 118*, 11225–11236.

Kellogg, G. E., Semus, S. F., & Abraham, D. J. (1991). HINT: A new method of empirical hydrophobic field calculation for CoMFA. *Journal of Computer-Aided Molecular Design, 5*, 545–552.

Khedkar, S. A., Malde, A. K., & Coutinho, E. C. (2007). Design of inhibitors of the MurF enzyme of *Streptococcus pneumoniae* using docking, 3D-QSAR, and de novo design. *Journal of Chemical Information and Modeling, 47*, 1839–1846.

Khedkar, V. M., Ambre, P. K., Verma, J., Shaikh, M. S., Pujari Lenkar, R. R., & Mietzner, T. (1994). Molecular docking and 3D-QSAR studies of HIV-1 protease inhibitors. *Journal of Molecular Modeling, 16*, 1251268.

Klebe, G., Abraham, U., & Mietzner, T. (1994). Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *Journal of Medicinal Chemistry, 37*, 4130–4146.

Lan, P., Chen, W. N., Xiao, G. K., Sun, P. H., & Chen, W. M. (2010). 3D-QSAR and docking studies on pyrazolo [4, 3-h] quinazoline-3-carboxamides as cyclin-dependent kinase 2 (CDK2) inhibitors. *Bioorganic & Medicinal Chemistry Letters, 20*, 6764–6772.

Lightstone, F. C., Zheng, Y.-J., & Bruce, T. C. (1998). Molecular dynamics simulations of ground and transition states for the Sn2 displacement of Cl-from 1, 2-dichloroethane at the active site of Xanthobacter autotrophicus halokalanate dehalogenase. *Journal of the American Chemical Society, 120*, 5611–5621.

Lozano, J. J., Pastor, M., Cruciani, G., Gaedt, K., Centeno, N. B., Gago, F., & Sanz, F. (2000). 3D-QSAR methods on the basis of ligand–receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands. *Journal of Computer-Aided Molecular Design, 14*, 341–353.

Lu, N., & Kokuf, D. A. (2001). Accuracy of free-energy perturbation calculations in molecular simulation. I. Modeling. *The Journal of Chemical Physics, 114*, 7303–7311.

Mitchell, M. P., Withers, I. M., Laughton, C. A., & Fischer, P. M. (2009). A study of CDK2 inhibitors using a novel 3D-QSAR method exploiting receptor flexibility. *QSAR & Combinatorial Science, 28*, 878–884.

Mitchell, M. J., & McCammon, J. A. (1991). Free energy difference calculations by thermodynamic integration: Difficulties in obtaining a precise value. *Journal of Computational Chemistry, 12*, 271–275.
Oliveira, A. A., Ramalho, T. C., & da Cunha, E. F. F. (2009). QSAR study of androstenedione analogs as aromatase inhibitors. Letters in Drug Design & Discovery, 6, 554–562.

Ortiz, A. R., Pinsarbaro, M. T., Gago, F., & Wade, R. C. (1995). Prediction of drug binding affinities by comparative binding energy analysis. Journal of Medicinal Chemistry, 38, 2681–2691.

Pan, D., Tseng, Y., & Hopfinger, A. J. (2003). Quantitative structure-based design: Formalism and application of receptor-dependent RD-DQ SAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. Journal of Chemical Information and Computer Science, 43, 1591–1607.

Pan, D., Liu, J., Senese, C., Hopfinger, A. J., & Tseng, Y. (2004). Characterization of a ligand–receptor binding event using receptor-dependent four-dimensional quantitative structure-activity relationship analysis. Journal of Medicinal Chemistry, 47, 3075–3088.

Pastor, M., Cruciani, G., McClay, I., Pickett, S., & Clementi, S. (2000). GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. Journal of Medicinal Chemistry, 43, 3233–3243.

Pereyra, M., Pastor, M., Ortiz, A. R., & Gago, F. (1998). Comparative binding energy analysis of HIV-1 protease inhibitors: Incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. Journal of Medicinal Chemistry, 41, 836–852.

Pissurlenkar, R. R. S., Khedkar, V. M., Iyer, R. P., & Coutinho, E. C. (2011). Ensemble QSAR: A QSAR method based on conformational ensembles and metric descriptors. Journal of Computational Chemistry, 32, 2204–2218.

Pratim Roy, P., Paul, S., Mitra, I., & Roy, K. (2009). On two novel parameters for validation of predictive QSAR models. Molecules, 14, 1660–1701.

Reddy, M. R., & Erion, M. D. (2001). Free energy calculations in rational drug design. New York, NY: Springer.

Richard, D., Cramer III, R. D., Bunce, J. D., Patterson, D. E., & Frank, I. E. (1988). Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. Quantitative Structure-Activity Relationships, 7, 18–25.

Robinson, D. D., Winn, P. J., Lyne, P. D., & Richards, W. G. (1999). Self-organizing molecular field analysis: A tool for structure-activity studies. Journal of Medicinal Chemistry, 42, 573–583.

Romeiro, N. C., Albuquerque, M. G. O, de Alencastro, R. B., Ravi, M., & Hopfinger, A. J. (2005). Construction of 4D-QSAR models for use in the design of novel p38-MAPK inhibitors. Journal of Computer-Aided Molecular Design, 19, 385–400.

Rucker, C., Rucker, G., & Meringer, M. (2007). y-Randomization and its variants in QSAR/QSAR. Journal of Chemical Information and Modeling, 47, 2345–2357.

Samanta, S., Debnath, B., Basu, A., Gayen, S., Srikanta, K., & Jha, T. (2006). Exploring QSAR on 3-amino pyrazoles as antitumor agents for their inhibitory activity of CDK2/cyclin A. European Journal of Medicinal Chemistry, 41, 1190–1195.

Santos-Filho, O. A., Mishra, R. K., & Hopfinger, A. J. (2001). Free energy force field (FEFF) 3D-QSAR analysis of a set of Plasmodium falciparum dihydrofolate reductase inhibitors. Journal of Computer-Aided Molecular Design, 15, 787–810.

Senese, C. L., & Hopfinger, A. J. (2003). Receptor-independent 4D-QSAR analysis of a set of norstatine derived inhibitors of HIV-1 protease. Journal of Chemical Information and Computer Science, 43, 1297–1307.

Silverman, B. D., & Platt, D. E. (1996). Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. Journal of Medicinal Chemistry, 39, 2129–2140.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 36, 111–147.

Straatsma, T. P., & Berendsen, H. J. C. (1988). Free energy of ionic hydration: Analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. The Journal of Chemical Physics, 89, 5876–5886.

Tokarski, J. S., & Hopfinger, A. J. (1997). Prediction of ligand–receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: Application to a set of peptidomimetic renin inhibitors. Journal of Chemical Information and Computer Science, 37, 792–811.

Vedani, A., Briem, H., Dobler, M., Dollinger, H., & McMasters, D. R. (2000). Multiple-conformation and protonation-state representation in 4D-QSAR: The neurokinin-1 receptor system. Journal of Medicinal Chemistry, 43, 4416–4427.

Vedani, A., & Dobler, M. (2000). Multi-dimensional QSAR in drug research. Predicting binding affinities, toxicity and pharmacokinetic parameters. Progress in Drug Research, 57, 105–135.

Vedani, A., & Dobler, M. (2002a). 5D-QSAR: The key for simulating induced fit? Journal of Medicinal Chemistry, 45, 2139–2149.

Vedani, A., & Dobler, M. (2002b). Quasar. Basel: Biographics Laboratory 3R.

Vedani, A., Dobler, M., & Lill, M. A. (2005). Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. Journal of Medicinal Chemistry, 48, 3700–3703.

Venkataraman, P., & Hopfinger, A. J. (1999). 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. Journal of Medicinal Chemistry, 42, 2169–2179.

Verma, J., Khedkar, V. M., Prabhu, S. A., Khedkar, S. A., Malacle, A. K., & Coutinho, E. C. (2008). A comprehensive analysis of the thermodynamic events involved in lipase–receptor binding using CoMIA and its variants. Journal of Computer-Aided Molecular Design, 22, 91–104.

Verma, J., Khedkar, V. M., & Coutinho, E. C. (2010). 3D-QSAR in drug design–A review. Current Topics in Medicinal Chemistry, 10, 95–115.

Verma, J., Malde, A., Khedkar, S., Iyer, R., & Coutinho, E. (2009). Local indices for similarity analysis (LISA)-a 3D-QSAR formalism based on local molecular similarity. Journal of Chemical Information and Modeling, 49, 2695–2707.

Wade, R. C., Ortiz, A. R., & Gago, F. (1998). Comparative binding energy analysis. Perspectives in Drug Discovery and Design, 9, 19–34.

Wildman, S. A., & Crippen, G. M. (2002). Three-dimensional molecular descriptors and a novel QSAR method. Journal of Molecular Graphics and Modelling, 21, 161–170.

Withers, I. M., Mazanetz, M. P., Wang, H., Fischer, P. M., & Laughton, C. A. (2008). Active site pressurization: A new tool for structure-guided drug design and other studies of protein flexibility. Journal of Chemical Information and Modeling, 48, 1448–1454.