Chemometric Analysis of Ligand Receptor Complementarity: Identifying Complementary Ligands Based on Receptor Information (CoLiBRI)

Scott Oloff,[†,§] Shuxing Zhang,[†] Nagamani Sukumar,[‡] Curt Breneman,[‡] and Alexander Tropsha*[,†]

Laboratory for Molecular Modeling, School of Pharmacy and Department of Pharmacology, School of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599, and Department of Chemistry and Chemical Biology, Rensselaer Polytechnic Institute, Troy, New York 12180

Received February 22, 2005

We have developed a novel structure-based chemoinformatics approach to search for Complimentary Ligands Based on Receptor Information (CoLiBRI). CoLiBRI is based on the representation of both receptor binding sites and their respective ligands in a space of universal chemical descriptors. The binding site atoms involved in the interaction with ligands are identified by the means of a computational geometry technique known as Delaunay tessellation as applied to X-ray characterized ligand−receptor complexes. TAE/RECON multiple chemical descriptors are calculated independently for each ligand as well as for its active site atoms. The representation of both ligands and active sites using chemical descriptors allows the application of well-known *chemometric* techniques in order to correlate chemical similarities between active sites and their respective ligands. We have established a protocol to map patterns of nearest neighbor active site vectors in a multidimensional TAE/RECON space onto those of their complementary ligands and vice versa. This protocol affords the prediction of a virtual complementary ligand vector in the ligand chemical space from the position of a known active site vector. This prediction is followed by chemical similarity calculations between this virtual ligand vector and those calculated for molecules in a chemical database to identify real compounds most similar to the virtual ligand. Consequently, the knowledge of the receptor active site structure affords straightforward and efficient identification of its complementary ligands in large databases of chemical compounds using rapid chemical similarity searches. Conversely, starting from the ligand chemical structure, one may identify possible complementary receptor cavities as well. We have applied the CoLiBRI approach to a data set of 800 X-ray characterized ligand−receptor complexes in the PDBbind database. Using a *k* nearest neighbor (kNN) pattern recognition approach and variable selection, we have shown that knowledge of the active site structure affords identification of its complimentary ligand among the top 1% of a large chemical database in over 90% of all test active sites when a binding site of the same protein family was present in the training set. In the case where test receptors are highly dissimilar and not present among the receptor families in the training set, the prediction accuracy is decreased; however, CoLiBRI was still able to quickly eliminate 75% of the chemical database as improbable ligands. CoLiBRI affords rapid prefiltering of a large chemical database to eliminate compounds that have little chance of binding to a receptor active site.

INTRODUCTION

Computer Aided Drug Design (CADD) is frequently identified with two concurrent approaches: *structure-based* and *ligand-based* modeling. *Structure-based* methodologies employ either an experimentally (X-ray or NMR) determined or predicted three-dimensional structure of a target protein. This structure is used to scan available chemical databases or virtual combinatorial libraries for the identification of potential lead compounds, which are characterized by stereochemical complementarity to the active site and have a high predicted binding constant. The binding constant estimate requires a fast and accurate scoring function. Several successful studies have been reported in the literature using popular docking methods such as DOCK[1−4] and AutoDock.[5,6] Typically, these approaches are capable of identifying a small number of compounds that can fit comfortably into the active site. Despite these success stories, accurate prediction of binding constants still represents a formidable task and is a focus of many ongoing investigations.[7]

Theoretically, the most accurate estimate of the free energy of binding can be obtained using energy based methods. These methods typically employ force fields originally developed for the refinement of experimentally determined molecular structures or for molecular dynamics simulations. Examples include free energy perturbation (FEP)[8,9] or Linear Interaction Energy (LIE) approaches,[10−12] which require significant computational resources. Application of continuum solvation models, instead of explicit solvent, can significantly reduce the cost of these calculations.[13] However, the computational cost of such methods is still too demanding to afford calculations in a high-throughput fashion, and

* Corresponding author e-mail: tropsha@email.unc.edu or alex_tropsha@unc.edu.
† School of Pharmacy, University of North Carolina.
‡ School of Medicine, University of North Carolina.
§ Rensselaer Polytechnic Institute.

10.1021/ci050065r CCC: $33.50
Published on Web 01/11/2006

therefore these calculations can only be practical if applied to either relatively simple systems or small sets of compounds with similar binding modes.

Ligand-based approaches rely on a series of ligands with known binding affinities to build correlations between ligand chemical structure and target properties of interest, such as binding constants or specific biological activities [see ref 14 for a recent review]. The ligand structures are typically represented by multiple chemical descriptors,[15] and statistical data modeling techniques are used to establish quantitative correlations between descriptors and binding affinities. Chemical descriptors and various chemical similarity measures (e.g., Euclidean distances between compounds in multidimensional descriptor space) are at the core of chemometric approaches to the analysis of molecular databases.[16] Such approaches afford rapid chemical similarity calculations and are widely used in database mining or rational library design to discover molecules similar to available compounds that are likely to have similar biological activity.[17] Chemical similarity searches are much more computationally efficient than structure based virtual screening. However, they are more likely to identify false positives that are too bulky or simply not stereochemically complementary to the actual binding site because the binding site information is not typically used as part of the query. Furthermore, chemometric approaches typically identify compounds that are highly similar to the training set compounds, making it difficult to identify novel ligands of a different structural class.

In this paper, we discuss a novel computational drug discovery strategy that combines the strengths of both *structure-based* and *ligand-based* approaches while attempting to surpass their individual shortcomings. To this end, we sought a representation that would allow us to characterize both receptor active sites and their corresponding ligands in the same universal, multidimensional, chemical descriptor space. We reasoned that mapping of both binding pockets and corresponding ligands onto the same multidimensional chemistry space would preserve the complementarity relationships between binding sites and their respective ligands. Thus, we expected that similar binding sites (where similarity is described quantitatively using one of the conventional metrics, such as Manhattan distance in multidimensional descriptor space) would correspond to similar ligands. This would imply that the relative location of a novel binding site in this chemistry space with respect to other binding sites could be used to predict the location of the ligand(s) complementary to this site in the ligand chemistry space. This virtual ligand(s) could then be used as a query in chemical similarity searches to identify putative ligands of the same receptor in available chemical databases.

To implement this strategy, we have used molecular descriptors based on Transferable Atom Equivalents (TAE) developed by Breneman and co-workers.[1,18−20] The major advantage of these descriptors over other descriptor types (discussed briefly in the next section) is that they are derived from the electronic and shape properties of isolated atoms or chemical groups. The additivity principle is used to calculate molecular descriptors by summing up the individual descriptor type values for all atoms in the molecule, using the RECON method. In the case of ligands, this leads to the generation of molecular descriptors, similar to other approaches. The same additivity principle can also be used to derive pseudomolecular descriptors for any group of atoms, e.g., active site fragments, making the TAE descriptors exceptionally well suited for our approach.

In this paper, we report on the application of this chemoinformatics structure based drug discovery strategy, termed CoLiBRI (identification of Complimentary Ligand Based on Receptor Information), to a training set of 800 diverse ligand−receptor complexes comprising the PDBBind data set.[21] We show that the knowledge of the receptor active site affords the highly computationally efficient and accurate identification of its respective ligand(s) within a large compound database. The success of this pilot study indicates that the CoLiBRI method is a rapid prescreening tool to identify the most promising compounds prior to engaging more computationally intensive three-dimensional docking approaches.

COMPUTATIONAL METHODOLOGIES

Data Set. Coordinates for 800 chemically and functionally diverse ligand−receptor complexes were obtained from the PDBbind Database[21] (PDB entry codes for these proteins are listed in the Supporting Information). SYBYL v6.8[22] was used to preprocess the raw macromolecular structures, including elimination of the crystallographic water molecules, removal of salts, and addition of hydrogen atoms.

Chemical Descriptors. Several important considerations went into finding the most capable descriptors in the context of our studies. There are two major classes of traditional chemical descriptors that are derived from either two-dimensional chemical graphs (e.g., molecular connectivity indices, charge descriptors, and others[23−29]) or from three-dimensional molecular models using relative atomic positions in addition to atom properties. A major benefit of 2D compared to 3D chemometric methods is that the former neither requires a conformational search nor structural alignment of molecules. Accordingly, 2D methods are more easily automated and adapted to the task of database searching or virtual screening.[30,31] In fact, 2D descriptors have been shown to be superior to 3D descriptors in database mining.[32] However, most 2D chemical descriptors are typically calculated from only complete molecular graphs. Consequently, they cannot be used to characterize active sites that are composed of fragments or individual atoms of amino acid residues that are involved in specific contacts with ligands. A notable exception is the TAE descriptors, as discussed in the Introduction.

The TAE/RECON method that was developed by Breneman and co-workers is based on the Bader's quantum theory of atoms in molecules (AIM). The TAE method of molecular electron density reconstruction utilizes a library of integrated atomic "basins", as defined by the AIM theory, to rapidly reconstruct representations of molecular electron density distributions and van der Waals electronic surface properties. RECON is capable of rapidly generating 6-31+G* level electron densities and electronic properties of large molecules, proteins or molecular databases, using TAE reconstruction. A library of atomic charge density fragments has been assembled in a form that allows for the rapid retrieval of the fragments, followed by rapid molecular assembly. Additional details of the method are described elsewhere.[1,18−20]

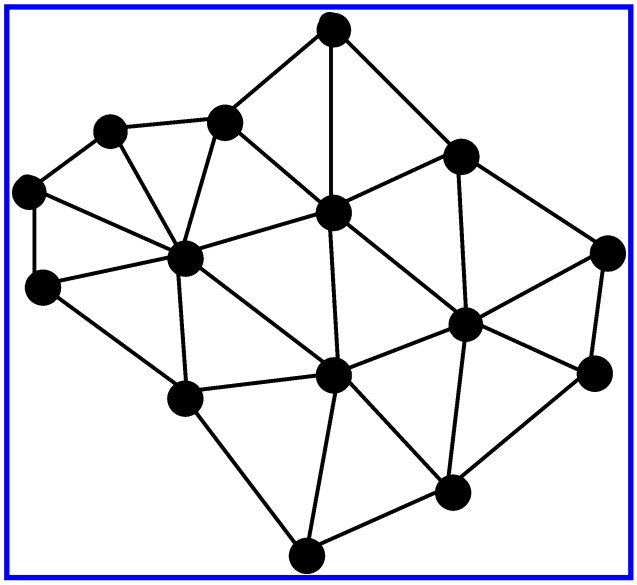

Figure 1. Delaunay tessellations of a collection of random points in 2D.

Calculation of TAE/RECON Descriptors for Ligands and Their Binding Sites. The calculation of TAE/RECON descriptors for the ligands (extracted from their protein complexes) was straightforward. However, similar calculations for the binding sites first required the identification of individual atoms or amino acid fragments involved in specific ligand−receptor interactions. To this end, we have utilized a computational geometry technique known as Delaunay tessellation to isolate the protein atoms that make contacts with bound ligands. Applied to a collection of randomly distributed points, Delaunay tessellation partitions the space occupied by these points into an aggregate of space filling, irregular triangles (in 2D) or tetrahedra (in 3D) with the original points as vertices. Thus, this approach effectively identifies all nearest neighbor triplets (or quadruplets) of vertices. An example of Delaunay tessellation in two dimensions is illustrated in Figure 1.

Protein−ligand complexes are represented by the coordinates of their heavy atoms (i.e., in a hydrogen-depleted form). Delaunay tessellation of this representation uniquely defines all sets of nearest neighbor atom quadruplets, including three types of interfacial quadruplets: three receptor atoms and one ligand atom; two receptor and two ligand atoms; and one receptor and three ligand atoms. Thus, Delaunay tessellation affords an easy way of detecting all receptor atoms that are nearest neighbors of the ligand atoms, which are specified as the active site. The RECON/TAE method is then used as described above to generate a set of descriptors for a pseudomolecule constructed from the active site atoms. An example of a tessellated ligand−receptor complex is shown in Figure 2.

The descriptors generated with TAE/RECON were range scaled prior to distance calculations (eq 1).

$$X_{\mathrm{idNORM}} = \frac{X_{\mathrm{id}} - X_{\mathrm{dMIN}}}{X_{\mathrm{dMAX}} - X_{\mathrm{dMIN}}} \qquad (1)$$

We chose to scale descriptors so that their absolute ranges after scaling are the same. Had we used the real values, the value of the absolute distance between receptors or ligands would be biased by the descriptors that cover the largest ranges.

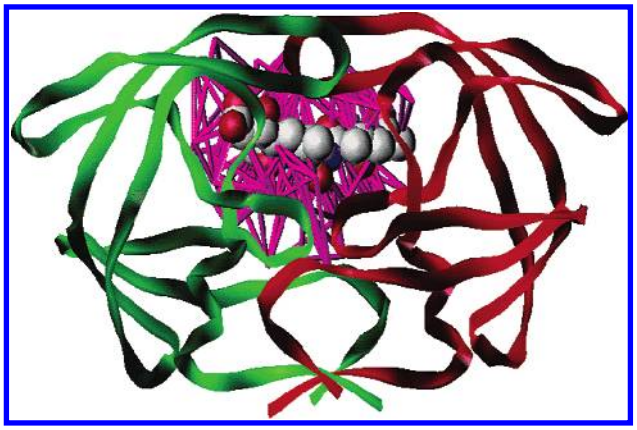

Figure 2. Tessellation of ligand−receptor interface identifies all active site atoms that make contacts with ligand atoms.

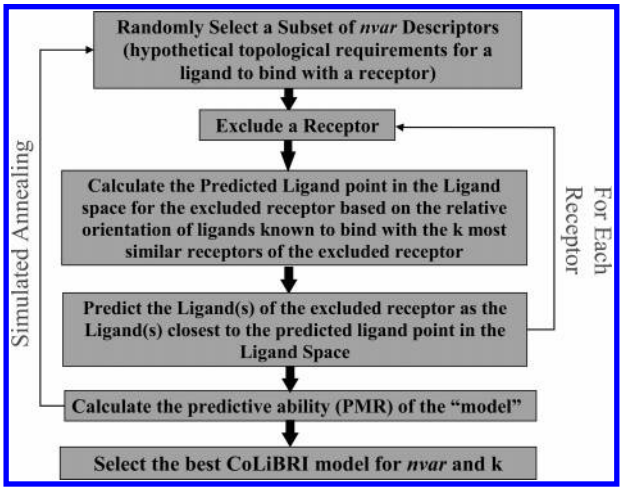

Figure 3. Flowchart of the CoLiBRI method.

CoLiBRI Algorithm. Using the TAE/RECON method, multiple descriptors were generated for both the receptor binding sites and their corresponding ligands so that each chemical entity is represented as a vector in a multidimensional TAE/RECON chemical space. Each dimension of this space corresponds to specific structural features of the ligands and active sites, but not every feature may be important for determining ligand−receptor complementarity. To select the subset of descriptors that best reflect the complementarity between receptors and their respective ligands, we have employed a Leave-One-Out (LOO) approach with variable selection. For each predefined number of variables (nVar), we used stochastic sampling of the descriptors space and simulated annealing to optimize (i) a selection of variables from the original pool of all molecular descriptors that are used to calculate relative similarities between receptors and their respective ligands (i.e., Euclidean distances in nVar-dimensional descriptor space) and (ii) the number of nearest neighbors (k) for each binding site in the receptor chemistry space used to estimate the position of the virtual complementary ligand in the ligand chemistry space. The overall flowchart of the CoLiBRI method is shown in Figure 3 and involves the following steps.

(1) Select a subset of nVar descriptors randomly (nVar is a number between 1 and the total number of available descriptors). nVar is usually set to different values in several

different runs. (2) Perform internal evaluation of this selection by a standard LOO cross-validation procedure and calculate the predictive mean rank (PMR) for the model as described below. (3) Repeat the procedure of generating trial correlations and calculate the corresponding PMR values (steps 1 and 2). The goal is to find the best topological requirements that minimize the PMR value of the CoLiBRI model. This optimization process is driven by a generalized simulated annealing (see below) using PMR as the objective function.

The resulting CoLiBRI model is a series of ligand−receptor complexes mapped into a descriptor subspace with a predefined number of descriptors (nVar), and K is selected based on the most predictive value for the training set. Ligands for a test receptor's binding pocket are predicted by positioning the test receptor pocket in the selected descriptor subspace and find the K most similar receptor pockets from the training set. The known ligands of these K most similar receptor pockets are then used to estimate the position of the test receptor's virtual ligand in the descriptor space using eq 2. All potential ligands are then ranked based on their distance to this predicted virtual ligand point, and the ligand(s) with the smallest distance are considered the most probable hits. Identifying a potential receptor target for a test ligand occurs in the opposite fashion whereby the K most similar training set ligands are found, and the known receptors of those ligands are used to interpolate what receptor target is the most likely candidate.

Cross-Validation and the kNN Principle Applied to the Prediction of Ligands Complementary to the Active Site. The standard leave-one-out cross-validation procedure has been implemented as follows.

(1) Choose a receptor in the training set and select its k nearest neighbors in the TAE/RECON binding site descriptor space. Identify the ligands of the kNN receptors in the ligand space and use their coordinates to predict the coordinates of the chosen receptor's virtual ligand. The coordinates of the virtual ligand are calculated from eq 2 for $k \geq 2$ (different values of k are explored to find the best model as described below)

$$\overrightarrow{X_{\mathrm{ppi}}} = \sum_{k=1}^{K_{\mathrm{Best}}} \left(\frac{\overrightarrow{X_{\mathrm{L}\leftarrow\mathrm{Rk}}}}{K_{\mathrm{Best}} - 1} \cdot \left(1 - \frac{||\overrightarrow{X_{\mathrm{Rk}}} - \overrightarrow{X_{\mathrm{RPredi}}}||}{\sum_{k=1}^{K_{\mathrm{Best}}} ||\overrightarrow{X_{\mathrm{Rk}}} - \overrightarrow{X_{\mathrm{RPredi}}}||} \right) \right) \quad (2)$$

where X_{RPredi} is the chosen receptor i, X_{ppi} is the predicted ligand vector for the receptor i, X_{Rk} is the k nearest receptor, and $X_{\mathrm{L}\leftarrow\mathrm{Rk}}$ is the ligand of the k nearest neighbor receptor. For the case where $K_{\mathrm{Best}} = 1$, then X_{ppj} is simply the position of the nearest receptor's ligand in the ligand space, $X_{\mathrm{L}\leftarrow\mathrm{R1}}$.

(2). Rank known ligands based on their chemical similarity to the virtual ligand. The similarities are evaluated as Euclidean distances (eq 3) using only the subset of descriptors that correspond to the current nVar selection.

$$\mathrm{Dist}_{i,j} = \sqrt{\sum_{d=1}^{\mathrm{nVar}} (X_{id} - X_{jd})^2} \quad (3)$$

(3) Repeat steps 1 and 2 until every receptor in the training set has been eliminated once, and the receptor's virtual ligand and the rank order of all compounds are predicted.

(4) Calculate the PMR for the model using eq 4, where NLR is the number of ligand−receptor complexes in the training set, nVar is the number of descriptors used to build the correlation, X_{jd} and X_{id} are the dth selected descriptor for ligands j and i, and X_{ppi_d} is the dth descriptor of the predicted ligand point.

$$\mathrm{PMR} = \frac{1}{N_{\mathrm{LR}}} \sum_{i=1}^{N_{\mathrm{LR}}} \sum_{j=1}^{N_{\mathrm{LR}}} \left\{ \begin{array}{l} 1 \ \text{if} \ \sum_{d=1}^{\mathrm{nVar}} (X_{jd} - X_{\mathrm{ppi}_d})^2 \leq \sum_{d=1}^{\mathrm{nVar}} (X_{id} - X_{\mathrm{ppi}_d})^2 \ \cap \ i \neq j \\ 0 \ \text{if} \ \sum_{d=1}^{\mathrm{nVar}} (X_{jd} - X_{\mathrm{ppi}_d})^2 > \sum_{d=1}^{\mathrm{nVar}} (X_{id} - X_{\mathrm{ppi}_d})^2 \ \cup \ i = j \end{array} \right\} \quad (4)$$

(5) Repeat steps 1−4 for $k = 3$, 4, 5, etc. Formally, the upper limit of k is the total number of ligand−receptor pairs in the data set minus one; however, the best value has been found empirically to lie between two and five. The k value that leads to the lowest PMR value is chosen as optimal.

Simulated Annealing Based Optimization for Variable Selection. The concept of simulated annealing (SA) is to simulate a physical process called annealing, in which a system is heated to a high temperature and then is gradually lowered to a preset temperature value. During this process, the system samples possible configurations according to the Boltzmann distribution. At equilibrium, low energy states will be mostly populated. The first implementation of the SA procedure was described by ref 33 followed by the development of a more generalized mathematical optimization protocol. The implementation of SA in our studies is as follows. (1) Generate a trial solution to the underlying optimization problem. For example, a CoLiBRI model is built based on a random selection of nVar descriptors. (2) Calculate the value of the fitness function, which characterizes the quality of the trial solution to the underlying problem, i.e., the inverse of the PMR value for a model built using only the selected nVar descriptors ($\mathrm{PMR}_{\mathrm{Current}}$). (3) Perturb the trial solution to obtain a new solution; i.e., change a fraction of the currently used descriptors to other randomly selected descriptors and build a new CoLiBRI model for the new trial set of nVar descriptors. (4) Calculate the new fitness function value for the trial solution as the inverse of the new PMR value ($\mathrm{PMR}_{\mathrm{New}}$). (5) Apply the optimization criteria: if $\mathrm{PMR}_{\mathrm{Current}} < \mathrm{PMR}_{\mathrm{New}}$ the new solution is accepted and used to replace the current trial solution; if $\mathrm{PMR}_{\mathrm{Current}} > \mathrm{PMR}_{\mathrm{New}}$, the new solution is accepted only if the following Metropolis criterion is satisfied (eq 5)

$$\mathrm{rnd} < \mathrm{e}^{-(\mathrm{PMR}^2_{\mathrm{Current}} - \mathrm{PMR}^2_{\mathrm{New}})/T} \quad (5)$$

where rnd is a random number uniformly distributed between 0 and 1 and T is a parameter analogous to the temperature in Boltzmann distribution law. (6) Steps 3−5 are repeated until the termination condition is satisfied. The temperature lowering scheme and the termination condition used in this work have been adapted from ref 34 as implemented by ref 35. Thus, every time a new solution is accepted or when a preset number of successive trial solutions (100 steps) does not lead to a better result, the temperature is lowered by a

preset value, usually by 10% (the default initial temperature is 100). The calculations are terminated when either the current temperature of the simulations is lowered to the value of $T = 10^{-6}$ or the ratio between the current temperature and the temperature corresponding to the best solution found is equal to 10^{-5}. In summary, the CoLiBRI generates both an optimum *k* value and an optimal subset of nVar descriptors, which together produce a model with the best internally predictive power, in terms of PMR. This model can be used to identify correlations between a receptor active site and its ligand(s) such that a predicted point can be found for any receptor, and vice versa, a predicted receptor point can be found for any ligand.

Model Validation. Training and Test Set Selection. The data set of 800 ligand−receptor complexes characterized by TAE/RECON descriptors was divided into the training (data used for model building) and test (data used for model validation) sets using the sphere exclusion method implemented in this laboratory.[36] Only binding site descriptors were used for these calculations. The purpose for this division was to generate a subset of ligands that did not bind any of the receptors in the training set. For the prediction of test set receptor ligands, the entire database of ligands was used. For additional validation, the World Drug Index (WDI) database was added to the list of available test set ligands, and the CoLiBRI models attempted to identify known ligands from the entire database.

Data Shuffling. To ensure that the models used to predict a test set were not based on chance correlations, the training set ligand−receptor associations were randomly shuffled. These data were then used to build CoLiBRI models, which were used for the test set prediction. If no significant model could be built for this randomized data set, it would suggest that the models built using the correct data accurately ligand−receptor complementarity in high-dimensional descriptor space.

RESULTS

Development and Validation of Training Set Models. A diverse training set of 670 receptor binding pockets was selected using the Sphere Exclusion Algorithm,[37] as described above and used by CoLiBRI to build models with the lowest PMR (eq 4). The remaining 130 receptors were used as a test set to evaluate the ability of the optimized model(s) to identify the correct ligand of each test receptor out of the original 800 ligands.

Previous studies from our group in the area of Quantitative Structure Activity Relationship (QSAR) indicated that the most reliable predictions of the test set data are obtained by using the consensus prediction approach.[31] In this approach, multiple variable selection models are built for the training set and used for the prediction of the test set ligands concurrently. To accomplish a consensus prediction, each model ranked all compounds in our ligand database based on the distance of each ligand to a test receptor's virtual complementary ligand. We then re-ranked the ligands based on those that were most similar to the virtual ligand across multiple models. These studies have shown that the inclusion of variable selection improved the mean rank of the test set from 37 to 24 out of 800. Furthermore, by using 100 models for consensus prediction, the mean rank of the test set was improved from 24 to 18.1 out of 800, as shown in Figure 4.

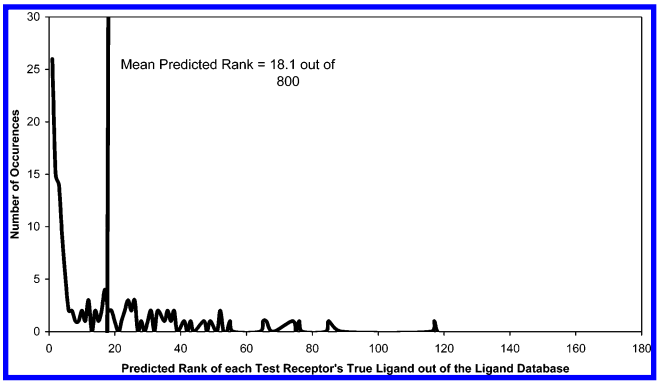

Figure 4. Predictive ability of CoLiBRI to identify ligands of 130 test binding pockets out of the original 800 ligands.

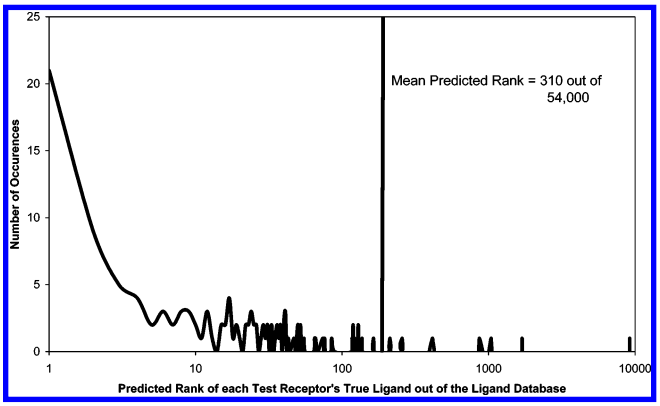

Figure 5. Predictive ability of CoLiBRI to identify ligands of 130 test binding pockets from the WDI and the original 800 ligands.

This increased the CPU time required to predict the test set by more than 2 orders of magnitude. Despite the increased CPU time, the calculations were still completed within 15 min. Since variable selection and consensus modeling vastly improved test set prediction, these methods were used in all subsequent model developments.

Application of CoLiBRI Models to Screening the WDI Database. To simulate the use of CoLiBRI for screening large chemical databases, we added the 800 training set ligands to the WDI data set,[38] which contains ca. 54 000 drugs and drug candidates. Training set CoLiBRI models were used in a consensus manner to predict the correct ligands for each of the 130 test receptors from of the entire combined database. The results illustrated that even when searching a large compound database, CoLiBRI is, on average, able to rank known ligands for a test receptor to within the top 310 ligands out of ca. 54 000, which translates to the top 1% of all compounds, as shown in Figure 5.

The entire screening calculation for 130 test receptors took roughly 4 h on a 2.4 GHz Pentium 4 machine. Figure 5 illustrates that most of the ligands were correctly identified within the top 12 ranked compounds; however, there were two distant outliers that made the average rank much higher. These two outliers (PDB codes 1BM7 and 1G4J) did not contain a receptor−ligand complex from the same family as those in the training set, which could possibly explain the inaccuracy of the predictions. The ligands extracted from 1BM7 and 1G4J, flufenamic acid and 4-(aminosulfonyl)-N-[(2,3,4,5,6-pentafluorophenyl)methyl]benzamide, respectively, also do not appear to be very similar to ligands found within the training data set. This additional dissimilarity may have also played a role in their poor prediction. As discussed

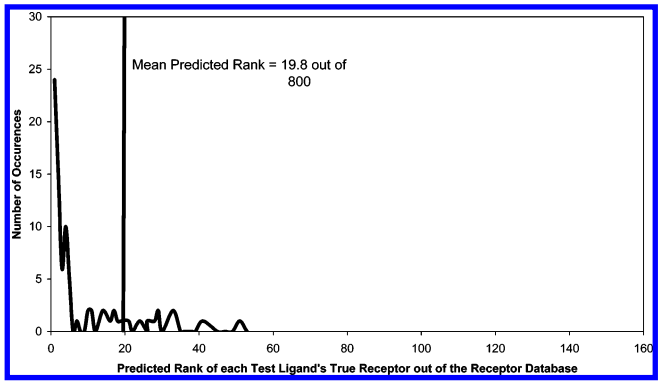

Figure 6. Predictive ability of CoLiBRI to identify receptor binding pockets of 130 test ligands from the library of 800 binding pockets.

in a later section, CoLiBRI appears to perform best when a receptor of the same family as the test set receptor is present in the training set. Otherwise, CoLiBRI is best used as a quick, rough filtering tool that can be used prior to the application of alternative less computationally efficient but perhaps more robust screening methodologies.

Identification of Binding Sites based on Ligand Information. As mentioned above, the CoLiBRI approach could be used to identify ligands based on the binding site descriptors as well as predict binding sites for the ligands. The latter application uses the same formalism (eqs 2−4), except that ligand coordinates are first mapped onto the binding site chemistry space, and then the virtual binding site is used as a query in chemical similarity calculations. We have tested CoLiBRI's ability to identify the correct binding pockets for the same series of 130 test ligands for all 800 pockets. On average, CoLiBRI was able to successfully rank the correct binding pocket to be in the top 20 out of 800, or within the top 2.5%, as shown in Figure 6. As one can see from the graph, most of the predictions actually identified the correct binding pocket within the top 2 or 3 receptors. After careful analysis of the outliers, we discovered that they primarily fell within two classes: either there were multiple receptors of the same family in the data set that a test ligand is known to bind with, and the correct binding site was ranked almost randomly within that top ranked family, or there were only one or no receptors of the same family in the training set.

Identification of Ligands for Novel Receptor Families. As the most rigorous test for CoLiBRI, we explored its ability to predict ligands for receptors that did not have homologous receptor structures in the training set. To this end, we selected a test set of 22 ligand−receptor complexes that consists of receptors from three receptor families: 7 complexes from the peptidase M10A family, 7 complexes from the SRC-Tyrosine kinase family, and 8 complexes from the peptidase S1 family. We then used CoLiBRI to identify the correct ligands from the WDI for each of the test receptors. The results shown in Figure 7 demonstrate that the accuracy of CoLiBRI under this difficult test has decreased significantly; we were only capable of identifying the correct ligand within the top 25% of the database on average, as opposed to the results reported when test receptors were from the same family (Figure 5).

This relative ineffectiveness of CoLiBRI under this test could still be acceptable when one needs to quickly screen a large, multimillion-compound library (25% accuracy for

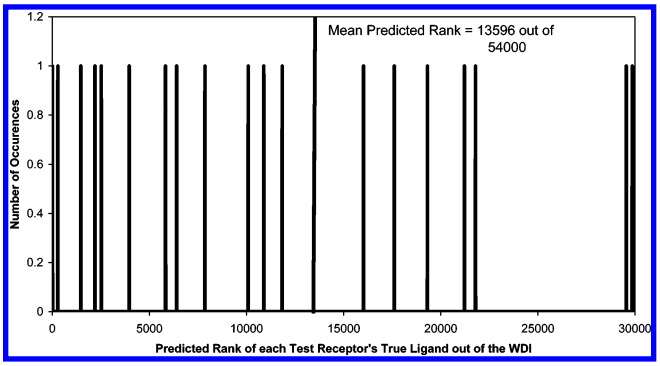

Figure 7. Predictive ability of CoLiBRI to identify ligands of 22 test binding pockets from the WDI when there are no receptors of the same family in the training set.

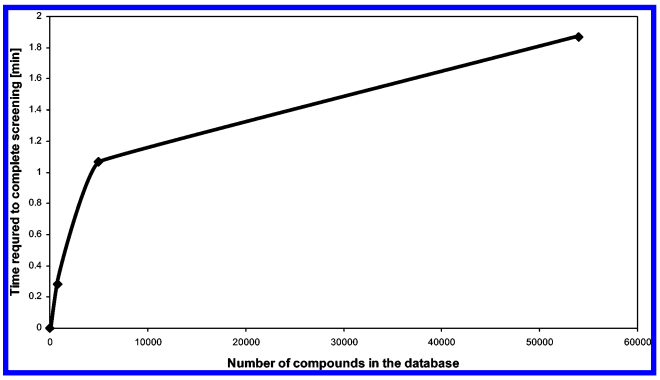

Figure 8. Time required for 100 CoLiBRI models to screen a single test receptor against various compound libraries on a P4 machine using a maximum limit of 5000 sorted hits displayed as output.

correct single hit identification is equivalent to the elimination of 75% of all compounds that are unlikely to be hits). These observations imply that the three test families are significantly different from the protein−ligand complexes in the training set so that the training set models are inapplicable to these test set proteins. In future studies, we shall consider introducing the model applicability domain for the CoLiBRI models, similar to our QSAR studies,[36,39] to serve as a warning that the predictions may be inaccurate due to high dissimilarities between complexes known in the training set and test binding pockets or ligands.

CPU Requirements of CoLiBRI for Efficient Screening of Large Databases. The time required for CoLiBRI to screen the WDI with 100 CoLiBRI models for a single receptor took less than 2 min on a 2.4 GHz Pentium 4 Desktop, as shown in Figure 8. These results illustrate that CoLiBRI may be successfully used as an efficient database filtering tool prior to more thorough and computationally intensive docking studies. For example, in our experience, AutoDock can typically process only about 1000 compounds per day on an SGI octane. The use of CoLiBRI prior to AutoDock will dramatically reduce the computational time required for thorough screening of the database. The reason for the abrupt change in slope in Figure 8 is due to a user-specified limit on the number of potential hits that are sorted in memory and presented as output. For this computational speed assessment test, a limit was set to 5000 ranked ligands presented as output. Thus, CoLiBRI first ranks the first 5000 compounds (or a user-defined number) in the database using all models and then sequentially updates this sorted list as it searches the remainder of the compound database for additional viable hits.

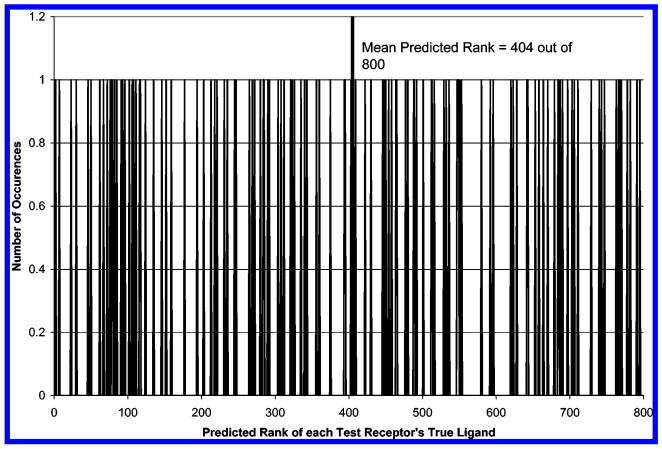

Figure 9. Predictive ability of CoLiBRI to identify ligands of 130 test binding pockets when the training set ligand−receptor associations were shuffled and used for model building.

Validation of CoLiBRI Models through Random Shuffling. To validate that the models were based on an inherently true correlation between properties of a receptor and its respective ligand, the ligands were shuffled such that for any one receptor, a randomly assigned ligand was used in place of the true ligand for model building. A test set of 130 receptors were then used to attempt the identification of the correct binding pockets using the models built on randomly associated ligand−receptor complexes. The PMRs for actual ligands (Figure 9) appear to be completely random suggesting that the models built with real data are based on actual complementarity correlations and are not artifacts of the model building process.

DISCUSSION

The main objective of CoLiBRI is to build chemometric models that reflect complementarity between a receptor's active site and its ligand, such that by knowing either the binding site or the ligand for a given receptor, a researcher could identify its complement from all other possibilities in a virtual database. In contrast with the QSAR approach, which requires a data set of known ligands for a particular receptor in order to build models and search for additional ligands of that receptor, CoLiBRI only requires the structure of the binding pocket. On the other hand, as compared with traditional 3D docking, CoLiBRI training set models are built across the entire available collection of diverse ligand receptor complexes as opposed to using a single receptor of interest for virtual screening. Consequently, CoLiBRI models use information about all other receptors when making prediction of ligands that bind to a particular receptor. In addition, CoLiBRI is significantly more computationally efficient that most 3D docking approaches.

As discussed above (cf. Figure 6), the "inverse" CoLiBRI approach could be also used to identify potential binding sites for a ligand. Binding to alternative sites is a frequent cause of side effects of drugs. The ability to identify potential undesirable interactions before a drug is brought to market is invaluable to the pharmaceutical industry. Foreknowledge of such interactions could send an otherwise effective compounds back through the lead optimization process before immense resources were lost in clinical trials. The structure of the ligand could be modified such that unwanted interactions are removed, while still maintaining its target property, thus leading to a highly useful drug.

To the best of our knowledge, the studies presented in this report are the first attempt to employ chemoinformatics approaches to the analysis of ligand−receptor complementarity. They could be extended in a number of ways. Thus, although this study was done using only TAE descriptors, in principle the descriptors for the receptor binding pockets and ligands could be of different types. This avenue is worth further investigation: while the TAE/RECON method appears to be unique in its ability to generate pseudomolecular descriptors for the collection of active site atoms, a number of other descriptor types are specifically designed for ligands and may better describe their features. Three-dimensional descriptors may also be used for the active site to preserve distance- and orientation-dependent interactions that may occur between a receptor and its ligands. Another important concept that we plan to examine in the future is the implementation of Support Vector Machines[40] as an alternative to kNN. The current approach uses SA variable selection to optimize complementarity. The PMR function could easily be made continuous and applicable to the Support Vector Theory by replacing k with a Gaussian weighting scheme and adding a continuous loss function. This could be done such that all neighbors in the initial starting space would be assigned distance dependent weights, where dissimilar neighbors would be given a weight of zero and closer neighbors would be given a weight higher than zero. Ligands or receptor active sites would be penalized based on their relative distance to their predicted point, which is calculated by orientations in the complementary space. This would allow us to take advantage of the inherent accuracy and generalization constraints of Support Vector Machines that may increase the predictive power of the models. Finally, as briefly mentioned above, a great deal of effort will be placed on defining adequate applicability domains for CoLiBRI models, which should prevent CoLiBRI from making unreliable predictions and therefore improve its accuracy.

CONCLUSIONS

We have developed a novel, predictive approach termed CoLiBRI to the analysis of ligand−receptor complementarity based on the representation of both ligands and their receptor active sites in a universal chemistry space. Unlike traditional docking methodologies that base their prediction of complimentary ligands using active site information alone, CoLiBRI predicts virtual ligands of the receptor based on its relative position in multidimensional chemistry space with respect to other known receptors. This representation affords straightforward and efficient identification of complementary ligands to a receptor from large databases of chemical compounds using rapid chemical similarity searches. Conversely, starting from the ligand chemical structure, one may identify possible complementary receptor cavities as well. This method is also distinct in that it penalizes the model for predicting an incorrect ligand for a receptor binding pocket, rather than optimizing the models by trying to correlate ligand binding to a single receptor. We have demonstrated that the knowledge of the active site structure affords identification of its complimentary ligands among

the top 1% of a chemical database in 90% of cases, where a complex of the same receptor family was present in the training set. In the case where test receptors are highly dissimilar and not present among the receptor families in the training set, the prediction accuracy decreased significantly; however, the method was able to quickly eliminate 75% of the chemical database as improbable ligands. Together, these results suggest that CoLiBRI can be used efficiently as a prescreening tool for traditional docking studies in order to identify a relatively small subset of compounds that are likely to contain actual hits.

Software Availability. The CoLiBRI program is available from the authors upon request. The TAE-RECON descriptors are available from C. Breneman's Web site, http://www.chem.rpi.edu/chemweb/recondoc/WinRecon.html.

ACKNOWLEDGMENT

The authors wish to thank Tripos Assoc. for the software grant. These studies were supported in part by the NIH research grant GM066940.

Supporting Information Available: PDB entry codes for proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Shoichet, B. K.; Kuntz, I. D. Matching Chemistry and Shape in Molecular Docking. *Protein Eng.* **1993**, *6*, 723−732.
(2) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-Based Molecular Design. *Acc. Chem. Res.* **1994**, *27*, 117−123.
(3) Gschwend, D. A.; Good, A. C.; Kuntz, I. D. Molecular docking towards drug discovery. *J. Mol. Recognit.* **1996**, *9*, 175−186.
(4) Oshiro, C. M.; Kuntz, I. D.; Dixon, J. S. Flexible Ligand Docking Using a Genetic Algorithm. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 113−130.
(5) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: Applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1−5.
(6) Osterberg, F.; Morris, G. M.; Sanner, M. F.; Olson, A. J.; Goodsell, D. S. Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **2002**, *46*, 34−40.
(7) Muegge, I.; Rarey, M. Small Molecule Docking and Scoring. In *Reviews in Computational Chemistry*, 17th ed.; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons: 2001.
(8) Kollman, P. Molecular-Dynamics and Free-Energy Perturbation Calculations − What Role Do They Play in Computer-Assisted Molecular Design. *FASEB J.* **1995**, *9*, A1253−A1253.
(9) Kollman, P. Free-Energy Calculations − Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395−2417.
(10) Aqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. Ligand binding affinities from MD simulations. *Acc. Chem. Res.* **2002**, *35*, 358−365.
(11) Aqvist, J.; Medina, C.; Samuelsson, J. E. New Method for Predicting Binding-Affinity in Computer-Aided Drug Design. *Protein Eng.* **1994**, *7*, 385−391.
(12) Chen, X.; Tropsha, A. Calculation of the hydration free energies using an extended linear response method. *J. Comput. Chem.* **1999**, *20*, 749−759.
(13) Zhou, R. H.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M. New linear interaction method for binding affinity calculations using a continuum solvent model. *J. Phys. Chem. B* **2001**, *105*, 10388−10397.
(14) Tropsha, A. Recent Trends in Quantitative Structure−Activity Relationships. In *Burger's Medicinal Chemistry and Drug Discovery*, 6th ed.; Abraham, D., Ed.; John Wiley & Sons: New York, 2003; pp 49−77.
(15) Livingstone, D. J. The characterization of chemical structures using molecular properties. *J. Chem. Inf. Comput. Sci* **2000**, *40*, 195−209.
(16) Willett, P. Chemoinformatics − similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol.* **2000**, *11*, 85−88.
(17) Turner, D. B.; Willett, P. Evaluation of the EVA descriptor for QSAR studies: 3. The use of a genetic algorithm to search for models with enhanced predictive properties (EVA_GA). *J. Comput.-Aided Mol. Des.* **2000**, *14*, 1−21.
(18) Breneman, C. M.; Thompson, T. R.; Rhem, M.; Dung, M. Electron Density Modeling of Large Systems using the Transferable Atom Equivalent Method. *Comput. Chem.* **1995**, *19*, 161−169.
(19) Mazza, C. B.; Sukumar, N.; Breneman, C. M.; Cramer, S. M. Prediction of protein retention in ion-exchange systems using molecular descriptors obtained from crystal structure. *Anal. Chem.* **2001**, *73*, 5457−5461.
(20) Song, M.; Breneman, C. M.; Bi, J.; Sukumar, N.; Bennett, K. P.; Cramer, S.; Tugcu, N. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1347−1357.
(21) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(22) Tripos Inc. Sybyl User's Manual Version 7.8; Tripos, Inc., St. Louis, MO, 2002.
(23) Kier, L. B.; Hall, L. H. Molecular connectivity VII: specific treatment of heteroatoms. *J. Pharm. Sci*. **1976**, *65*, 1806−1809.
(24) Kier, L. B.; Murray, W. J.; Randic, M.; Hall, L. H. Molecular connectivity V: connectivity series concept applied to density. *J. Pharm. Sci.* **1976**, *65*, 1226−1230.
(25) Kier, L. B.; Murray, W. J.; Hall, L. H. Molecular connectivity. 4. Relationships to biological activities. *J. Med. Chem.* **1975**, *18*, 1272−1274.
(26) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. Molecular connectivity. I: Relationship to nonspecific local anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971−1974.
(27) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(28) Murray, W. J.; Kier, L. B.; Hall, L. H. Molecular connectivity. 6. Examination of the parabolic relationship between molecular connectivity and biological activity. *J. Med. Chem.* **1976**, *19*, 573−578.
(29) Murray, W. J.; Hall, L. H.; Kier, L. B. Molecular connectivity. III: Relationship to partition coefficients. *J. Pharm. Sci.* **1975**, *64*, 1978−1981.
(30) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci* **1997**, *37*, 1−9.
(31) Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004**, *47*, 2356−2364.
(32) Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567−597.
(33) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, 1087−1092.
(34) Sun, L.; Xie, Y.; Song, X.; Wang, J.; Yu, R. Cluster Analysis By Simulated Annealing. *Comput. Chem.* **1994**, 103−108.
(35) Zheng, W.; Tropsha, A. Novel variable selection quantitative structure−property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(36) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241−253.
(37) Golbraikh, A.; Tropsha, A. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol. Divers.* **2002**, *5*, 231−243.
(38) Daylight, World Drug Index (WDI), 2004.
(39) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *Quant. Struct. Act. Relat. Comb. Sci.* **2003**, *22*, 69−77.
(40) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer-Verlag: 1995.

CI050065R