GFscore: A General Nonlinear Consensus Scoring Function for High-Throughput Docking

Stéphane Betzi,[†] Karsten Suhre,[‡] Bernard Chétrit,[†] Françoise Guerlesquin,[†] and Xavier Morelli*[,†]

BIP Laboratory, Bioénergétique et Ingénierie des Protéines, CNRS UPR9036 Institute for Structural Biology and Microbiology (IBSM), 31 Chemin Joseph Aiguier 13402 Marseille Cedex 20, France, and IGS Laboratory, Information Génomique et Structurale, CNRS UPR2589 (Structural and Genomic Information Laboratory) Institute for Structural Biology and Microbiology (IBSM), Marseille France, Parc Scientifique de Luminy 163 Avenue de Luminy FR-13288, Marseille Cedex 09, France

Received March 8, 2006

Most of the recent published works in the field of docking and scoring protein/ligand complexes have focused on ranking true positives resulting from a Virtual Library Screening (VLS) through the use of a specified or consensus linear scoring function. In this work, we present a methodology to speed up the High Throughput Screening (HTS) process, by allowing focused screens or for hitlist triaging when a prohibitively large number of hits is identified in the primary screen, where we have extended the principle of consensus scoring in a nonlinear neural network manner. This led us to introduce a nonlinear Generalist scoring Function, GFscore, which was trained to discriminate true positives from false positives in a data set of diverse chemical compounds. This original Generalist scoring Function is a combination of the five scoring functions found in the CScore package from Tripos Inc. GFscore eliminates up to 75% of molecules, with a confidence rate of 90%. The final result is a Hit Enrichment in the list of molecules to investigate during a research campaign for biological active compounds where the remaining 25% of molecules would be sent to in vitro screening experiments. GFscore is therefore a powerful tool for the biologist, saving both time and money.

1. INTRODUCTION

The cost to bring a drug to market has risen from an average of $231 million in 1991 to more than $897 million in 2004 (Tufts Center for the Study of Drug Development, CSDD).[1] Even more, according to a study by Bain & Co. (http://www.bain.com), the cost for a single new drug averages $1.7 billion, almost double the widely accepted $897 million estimate published in May by the Tufts Center for the Study of Drug Development. Part of the problem is that the time to market is increasing tremendously. In 1982, when insulin was approved, it took only 4 years to go through the entire process. Now it takes from 10 to 15 years. Drug companies currently have some 200 million potential drug compounds but less than 0.01% will eventually emerge as a Food and Drug Administration (FDA)-approved treatment.

One of the main solutions came in the mid-1970s when investigators suggested that computational simulations of receptor structures and the chemical forces that govern their interactions would enable 'structure-based' ligand design and discovery.[2,3] Since then, structure-based design has contributed to and even motivated the development of marketed drugs, but a major challenge still remains: i.e., the relationship between in silico screening and the real in vitro screen. How many molecules from the database should be effectively screened in order to get as many true positives as possible? This ratio of the number of hits over the number of tested molecules is called the 'Enrichment Factor', the Holy Grail of the Drug Discovery Process.

High-Throughput Docking (HTD) of chemical databases against protein structures—often abusively called Virtual Library Screening (VLS)—is an emerging and promising step in Computer-Aided Drug Design (CADD).[2] Each compound is sampled in thousands to millions of possible configurations and scored on the basis of its complementarity to the receptor. Of the hundreds of thousands of molecules in the library, only dozens of top-scoring predicted ligands (hits) are subsequently tested for activity in an experimental assay. That is the theory. In per die laboratory, while the docking step is commonly accepted as an efficient tool, the post-processing/rescoring step has impeded progress in this direction. It still remains a current challenge of theoretical chemistry. Indeed, the majority of published scoring functions have been developed in association with docking methods. However, the scoring functions used to rank compounds do not have to be linked to a docking method but, rather, are independent and should be able to estimate the binding affinities of the receptor−ligand complexes generated by any structure-based approaches. Unfortunately, in many cases the scoring functions perform poorly at predicting lower affinity binders.[4] Over the last 5 years many articles have been published on the impact of scoring functions on enrichment,[5] docking/scoring combination,[6] consensus scoring criteria,[7,8] evaluation and assessment of multiple scoring function,[4,9−11] machine learning,[12,13] detailed analysis of scoring function,[14] and even theoretical computer experimentation.[15] All these scientific approaches permit a

* Corresponding author phone: +33 491-164-501; fax: +33 491 164 540; e-mail: morelli@ibsm.cnrs-mrs.fr.

[†] CNRS UPR9036 Institute for Structural Biology and Microbiology (IBSM).

[‡] CNRS UPR2589 (Structural and Genomic Information Laboratory) Institute for Structural Biology and Microbiology (IBSM).

10.1021/ci0600758 CCC: $33.50
Published on Web 06/21/2006

better assessment of the right answer (pose of the ligand) at an affordable price (rank of the true positives).

In this work, we propose an original methodology to parse as many true negatives as possible from proprietary *in house* databases and to send the resulting list of potential hits (focused libraries) to standard High-Throughput Screening (HTS) experiments. This methodology speeds up the drug discovery process by diminishing the time (number of molecules to test in vitro) and also the cost (Enrichment Factor) to discover new hits when the docking mode of inhibitors is unknown. It can also present a great advantage for hitlist triaging when a prohibitively large number of hits is identified in the primary screen, particularly in a FLIPR cell-based assay (Fluorimetric Imaging Plate Reader). An additional advantage of our novel approach is that our consensus scoring function is a ranked-based and not a value-based function. We have used a neural network analysis for learning set evaluation of the scoring results that were extracted from the five scoring functions within CScore (FlexX score, G_score, D_score, ChemScore, and PMF score) following a FlexX-based (TRIPOS, Inc.) HTD. We have comparatively evaluated the ability of our generalist/consensus score with each of the other scoring functions including CScore[7] to eliminate true negatives with a high confidence rate. GFscore (available http://gfscore.cnrs-mrs.fr) eliminated up to 75% of a database with a confidence rate of 90%, while the next best function provided a confidence level of 70% and 62% (G_score or FlexX score respectively), having eliminated the same percentage of the database. We have also calculated the linear version of GFscore (that we are calling here L-GFscore). L-GFscore is able to eliminate 75% of the database with a confidence rate of 80% which is better than each scoring function individually taken but still less accurate than its nonlinear version. We have also analyzed in detail the capacity of these five different scoring functions to discriminate true positives from false positives. Together, these data provide important insight into some basic rules that appear to govern the fitness of each scoring function, supporting the notion that scoring function accuracy may indeed be influenced by the chemical space being targeted.

2. MATERIAL AND METHODS

Preparation of Ligand Databases. The selection of the database was directed by two criteria. First, we needed a short size database to be able to carry out numerous docking experiments. The second important point was the structural diversity of the compounds. Indeed, we aimed at eliminating any bias in the screening conclusions from the chemical compounds distribution. For these reasons we selected the "Diversity" database.

The Diversity database from the National Cancer Institute (NCI) (http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html) was derived from the almost 140 000 compounds available on plates using the program Chem-X (Oxford Molecular Group). Chem-X uses defined centers (hydrogen bond acceptor, hydrogen bond donor, positive charge, aromatic, hydrophobic, acid, and base) and defined distance intervals to create a particular finite set of pharmacophores. The requirements were set as 5 new pharmacophores and, additionally, 5 or fewer rotatable bonds. This procedure resulted in the selection of 1990 compounds. Each compound represents a structural chemical family. Diversity is therefore an interesting database for development studies. Our final Diversity "Drug-Like" database contains 1420 compounds. This database is provided by the NCI in three-dimensional (3D) structures stored in a Structural Data File (SDF), which was then transformed into a Sybyl database for High-Throughput Docking with FlexX. To select only "Drug-Like" compounds and to simulate a real High-Throughput Docking study, we filtered the database using derived-Lipinski rules (details are provided as Supporting Information).

Preparation of the Protein/Ligand Data Set. The FlexX data set contains 200 high-resolution X-ray protein/ligand complexes from the Protein Data Bank (PDB). The description of this data set was discussed by Kramer et al. during a FlexX evaluation.[16] This data set provides all necessary files for FlexX docking: original PDB of the free and bound protein, a Receptor Descriptor File (RDF), and the files of the extracted and minimized ligand in mol2 format. Optimized parameters were selected for these files: charges and ionization states, conformations, and metal atoms. Use of a minimized ligand guarantees a low-energy conformation with suitable bond distances and angles during a docking experiment. Moreover, this new geometry ligand is different from the structural information present in the original PDB. The active sites (receptor files) from the FlexX 200 data set are generally defined using all atoms no farther than 6.5 Å apart from a ligand atom at its crystalline position and without any water molecules. Lysine and arginine residues are protonated. Aspartic and glutamic acid are ionized. Protonation state of hydroxyl groups of serine, threonine, and tyrosine as well as hydrogen positions inside the histidine side chain are defined by the authors. Among this FlexX 200 data set, we only selected 78 complexes with a 2 Å maximum Root Mean Square Deviation (RMSD)[17] between our FlexX docked predictions and the real X-ray ligand structure. Using only these complexes we avoided docking bias in order to focus our work on the question of the scoring function.

Analysis of the Protein/Ligand Data Set. We have analyzed the chemical diversity of this data set using several descriptors. We have investigated each complex to extract all available information about the ligand, the protein, and the interaction between the two partners. Physicochemical descriptors extracted from the Pubchem database (http://pubchem.ncbi.nlm.nih.gov/) were used to define ligands descriptors (Molecular Weight, number of rotatable bonds, and the number of H-bond donors and acceptors). We added to these descriptors an approximation of the ligand polar surface (Topological Polar Surface Area or TPSA). This value is computed with a MOE's function which uses group contributions from connection table information.[18] We also used MOE to calculate the SLogP values.[19] Contacts were analyzed using LPC software[20] available online at "Bioinformatics and Biological Computing Unit" bioportal (Weitzman Institute of Science) (http://bioportal.weizmann.ac.il/oca-bin/lpccsu). LPC calculates the number of nonbonded interactions and solvent accessible surface complementarity between protein and ligand. Protein cavity surface and volume were computed with CASTp[21,22] (http://cast.engr.uic.edu/cast/) using a visual cleft selection.

Docking Software and Procedure. For each of the 78 proteins, our High-Throughput Docking (HTD) experiments consisted of comparing the results obtained for the active compound hidden in the database with those obtained for all the other compounds, which were assumed to be inactive. We docked each compound of the database with each protein of our 78 FlexX data set using the FlexXdocking program as implemented in the 7.0 release of Sybyl Tripos package. FlexX is an incremental construction docking algorithm involving three steps. First, FlexX cuts the ligand into pieces and selects one of these pieces as the base fragment. In the second step, FlexX places the selected base fragment in the active site. The last step is the incremental construction of the whole ligand from the base fragment using the rest of the molecule. Conformational flexibility of the ligand is taken into account by considering both torsion angle flexibility and conformational flexibility of ring systems.[23] FlexX by itself does not treat the receptor flexibility; rather it sees proteins as rigid elements (bodies). Default FlexX parameters were used as supplied in Tripos Sybyl7.0 for carrying out flexible docking with 30 conformations for each molecule, using the place particle option as defined by Rarey et al.[24]

We have applied a common methodology of rescoring described by Wang et al.: "main scoring during docking uses a given scoring function; solutions are re-scored with other adapted scoring functions".[25] In our case, we selected the best pose for each FlexX docking using the FlexX scoring function; we then rescored with the CScore Tripos module.[7] This consensus score (CScore) integrates a number of popular scoring functions to rank ligands affinity to the active site of the receptor. D_score and G_score are two Force-Field derived scoring functions. D_score, derived from DOCK score in the FlexX implementation, uses steric and electrostatic terms based on the AMBER Force-Field;[26] whereas G_score, derived from GOLD score in the FlexX implementation, is a sum of the hydrogen bonding stabilization energy (calculated from van der Waals energy for the ligand and conformers) and a pair wise dispersion potential between ligand and protein to describe the hydrophobic binding energy.[17,27] A knowledge-based scoring function, PMF score, exploits structural information of known complexes and converts it into distance-dependent Helmholtz free energies.[28] Finally, two empirical scoring functions were implemented: ChemScore consists of a term that estimates lipophilic contact energy, a metal−ligand binding contribution, an empirical form for hydrogen bonds, and a penalty for ligand flexibility;[29] whereas FlexX score considers the number of rotatable bonds in the ligand, hydrogen bond interaction, ion pairing, aromatic interactions, and the lipophilic contact energy.[23]

Tripos CScore module uses these five scoring functions to generate the CScore consensus value[7] for the best of the 30 ligand conformations generated as well as an overall consensus CScore for the ranking of the entire database. The 78 CScore scoring results tables of all the HTD experiments were extracted and exported in text files for treatments and analysis.

The Neural Network: GFscore. For each of the 78 proteins from the FlexX data set, the docking of each of the 1420 compounds of the Diversity "Drug-Like" database is described by five scores (see above). For every docking between a given compound and a given protein, the rank of a given score is determined based on the scores of the 1420 other dockings against the same protein. The ranks are then scaled to the range [−1,+1] for convenience. Thus, every individual docking experiment is described by five scaled ranks.

The neural network implementation of the statistical package R (http://www.r-project.org/) was used here. This package allows different kinds of neural network architectures to be used. Based on initial tests (not reported here), we chose the following configuration: one hidden layer with 5 neurons, skip layer connections, and linear output. The five scaled ranks are the input parameters of the neural network; a value of one is output in the case of a correct (reference) docking, while a value of zero is output otherwise. For each protein there are thus 1420 cases that correspond to an output of zero, and only one case that corresponds to an output of one. To train the neural network with as many correct as false dockings, the data of the correct solution are replicated 1420 times in the learning data set. The corresponding five scaled ranks for these reference data sets were slightly perturbed using a Gaussian distribution around the observed scaled ranks to avoid overfitting on the thus highly repeated combinations of input parameters. The standard deviation of this perturbation was varied to find an optimum value. One-third of the data set was set aside as a test set for later cross-validation and was not used in the training of the neural network (27 complexes as positive test set and 38 124 complexes as negative test set). Thus, for the learning set, we used a total of 51 complexes as the positive learn set and 72 012 complexes as the negative learn set. The choice of the best neural network was based on its performance as evaluated on this independent test data set. Repeated neural network fittings (50) were performed while limiting the number of individual learning iterations (100) to avoid overfitting.

3. RESULTS AND DISCUSSION

One of the crucial criteria for the validation of a methodology development in docking and virtual screening is the representativeness of the data selection used to train the system. The use of a representative training data set avoids introducing bias in the conclusions. The two critical points in virtual screening approaches are the training of the ligand database and the selection of the protein/ligand data set.

Representativeness of the Database and Data Set. The first point that we wanted to analyze was the representativeness of this data set in the 'rule of five'/Lipinski-derived chemical space. This kind of analysis has already been published for other databases by Dr. Irvin and Shoichet[30] for the analyses of the ZINC database or by Dr. Rognan et al.[11] for the analysis of a *proprietary* database. Since we wanted to represent the same kind of simple analysis, we thus decided to interpret the same parameters in our work with the Diversity database (i.e, 1D descriptors of the 'rule of five' parameters).

Plotting compounds Topological Polar Surface Area (TPSA) versus Molecular Weight (MW) indicates a linear relationship. Indeed, the tendency curve confirms that the larger the ligand, the more polar is its surface area (Figure 1A). Examination of endpoints also indicates a good disper-

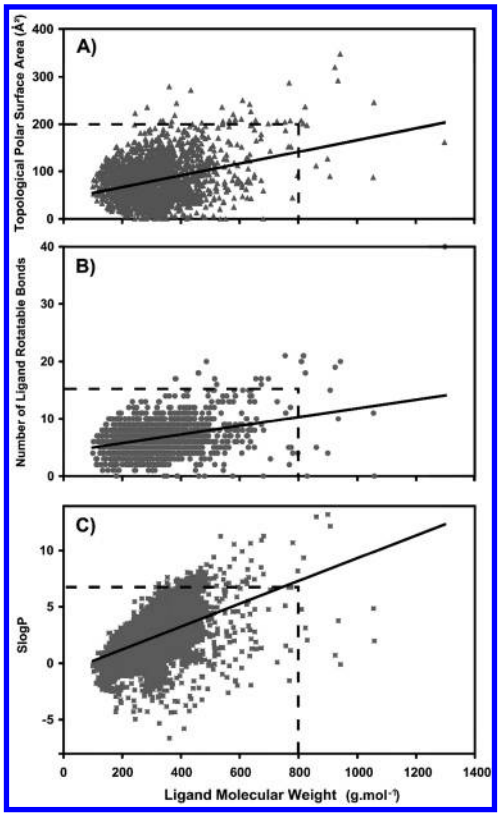

Figure 1. Analysis of the chemical database "Diversity". Fitted curve is displayed in the two graphs. The Drug-Like chemical space is drawn with dashed lines. Ligand molecular weights are represented versus topological polar surface area (Å^2) in (A), versus rotatable bonds in (B) and versus SlogP in (C).

sion in polarity; some compounds outside the tendency are small and hydrophilic, whereas others are large and hydrophobic. The same investigation was performed by plotting the number of Ligand Rotatable Bonds (LRB) versus MW (Figure 1B) and SlogP versus MW (Figure 1C). The tendency curve makes obvious that the degree of freedom is proportional to the ligand size (Figure 1B); it also reveals a good dispersion for ligand flexibility and solubility (Figure 1C). The overall analysis of these data highlights a good diversity in polarity, flexibility, and solubility and demonstrates that we are exploring a good proportion of the 3D "drug-like" chemical space (inside the dashed lines in Figure 1A−C).

The molecular weight of the ligand data set is comprised between 68 g·mol^{-1} and 686 g·mol^{-1} (360 g·mol^{-1} average) with an Accessible Solvent Surface Area from 200 to 900 Å^2. The average number of rotatable bonds is 5, with a maximum of 22 and a minimum of zero. These values clearly show how diverse the ligand data set is. Comparing these ligand data set descriptors to the database descriptors confirms that the ligand data set values are well-distributed along the range of the Diversity database values (Figure 2). This data set is composed of "extreme" representatives in which one can find small- and medium-sized hydrophobic as well as very polar ligands. These are organized in three classes (Figure 2A). Regarding rotatable bonds, conclusions are the same in terms of variability, rigid to semirigid ligands up to very flexible ligands (Figure 2B). These highly flexible ligands concern five complexes: 5P2P, 1PSO, 1APT, 1PPK, and 1AAQ. It is well accepted in the scientific community that FlexX does not perform well with flexible ligands.

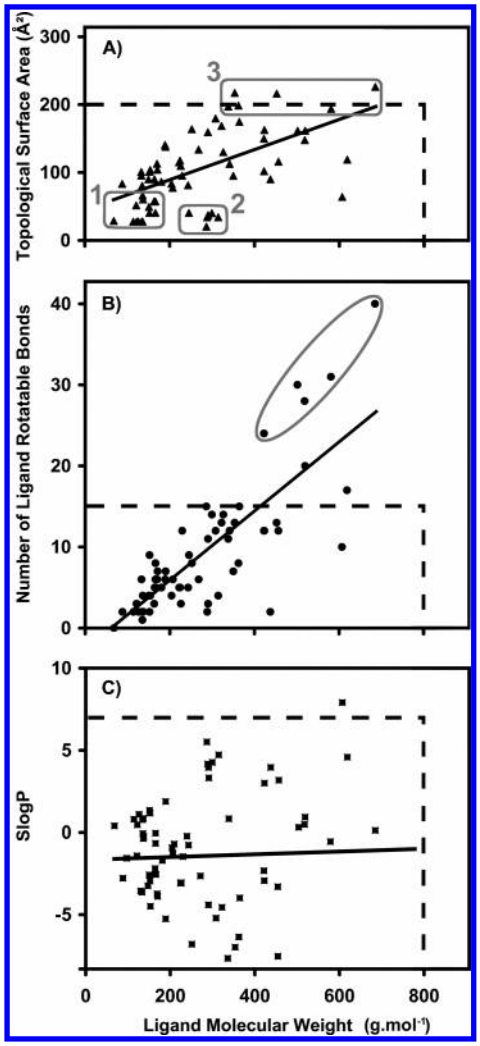

Figure 2. Analysis of the protein/ligand data set (78 high-resolution protein−ligand complexes). A list of the complexes is provided as Supporting Information. The fitted curve is displayed for the three graphs. The Drug-Like chemical space is drawn with dashed lines. (A) Ligand molecular weight is represented as a function of topological polar surface area (Å^2). Small hydrophobic, medium-sized hydrophobic and very polar ligands are enclosed in classes 1−3, respectively. (B) Ligand molecular weight is represented as a function of rotatable bonds. The most five highly flexible ligands (PDB codes 5P2P, 1PSO, 1APT, 1PPK, and 1AAQ) are enclosed in a gray ellipse. (C) Ligand molecular weight is represented as a function of SlogP.

Nevertheless, the analysis of the disparity of the protein/ligand data set allows us to demonstrate that we introduce no bias in our analysis due to a use of a too-specialized data set or a data set that is too far from the training database.

We have also calculated the Fingerprint Database Clustering in MOE (Chemcomp.), using the Jarvis-Patrick clustering method.[31] Using these indices (Tanimoto coefficient of 0.8 for the metric score and bit packed MACCS structural keys as fingerprint), we found 1749 clusters for the Diversity Database and 88 531 clusters for the entire NCI 127K Database. Since our goal here was to compare these two databases, we then decided to express the diversity coefficient by dividing the number of clusters obtained by the number of entries in the database, as proposed by Voigt et al.[32] for the analyses of the NCI database with 9 other Databases. Applying this approach we founded 87.9% for the Diversity database and 69.9% for the NCI 127K Database, confirming

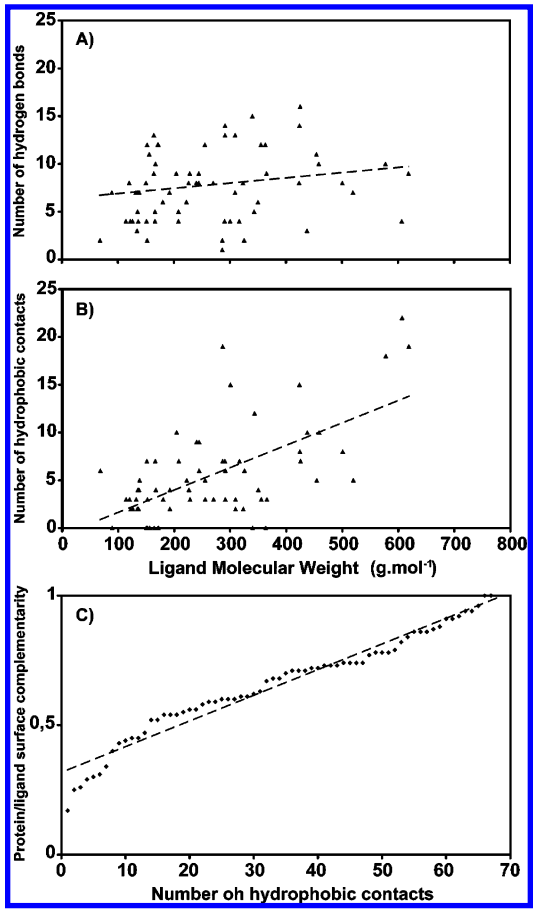

Figure 3. Analyses of nonbonded interactions in the protein/ligand data set. Ligand molecular weights are represented versus number of hydrogen bonds in (A) and versus hydrophobic contacts in (B). (C) Representation of protein/ligand surface complementarity in the protein/ligand data set.

the representativeness of the Diversity Database versus the entire 127K Database.

We also focused on the protein/ligand interaction for the 78 complexes of our data set to ensure that we were covering a large variety of interaction types. We classified them by means of the number of nonbonded contacts between the two partners of each complex. It is expected that the number of nonbonded interactions would increase in proportion to the size of the ligand; indeed large compounds have more heavy polycyclic aromatic groups for hydrophobic contacts and more hydrogen donors and acceptors. Analyzing all contact types, we also found small hydrophobic ligands, small hydrophilic ligands, and at the opposite large hydrophobic or large hydrophilic compounds, thus confirming our expectation (Figure 3A,B).

The last curve, depicting protein/ligand surface complementarity, highlights some very remarkable results (Figure 3C): most ligands fit well in the active site cavity with a 0.7 average complementarity surface, but the complementarity value is distributed between a minimum of 0.2 and a maximum of 1. Consequently, some small ligands bind a large cavity, whereas other ligands fit perfectly in the pocket (ideal case).

Concluding from these results, our protein/ligand data set is composed of different kinds of ligands (small polar ligands, small hydrophobic ligands, large polar ligands, and large hydrophobic ligands), for different kinds of interactions

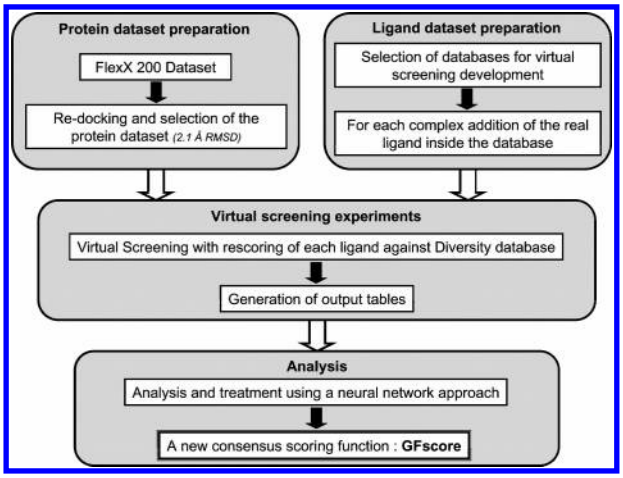

Figure 4. Flowchart of the docking procedure used to develop our Generalist scoring Function, GFscore. We selected a "diverse" database (Diversity) and a set of "diverse" protein−ligand complexes (see Figures 1−3 for more details). We then verified our ability to reproduce the docking of the 3D structure for 200 protein−ligand complexes, the FlexX 200 data set. Among this data set, FlexX was able to generate 78 complexes with an accurate RMSD (<2.1 Å). We then hid each of these 78 ligands with the Diversity database and high throughput docked each of them for each of the 78 targets. The resulting output tables were then used for training the neural network to differentiate true negatives from true positives, GFscore.

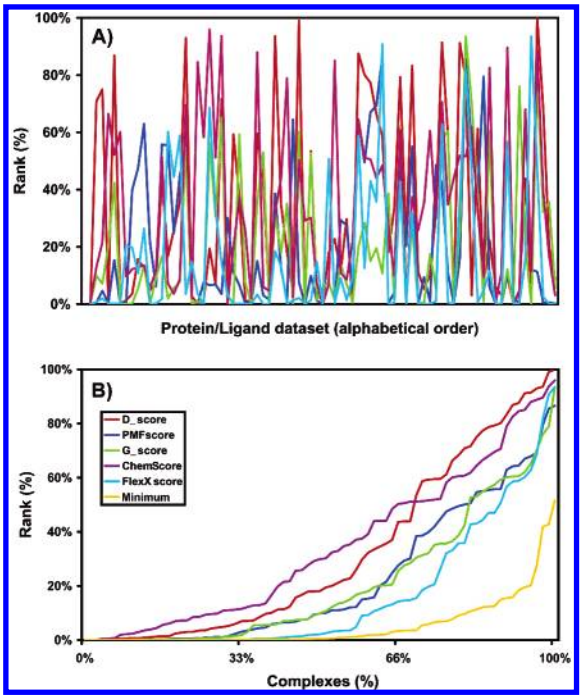

Figure 5. (A) Real ligand rank obtained by each scoring function for the diversity database. (A) The complexes are ranked in alphabetical order, and (B) the complexes are ranked relative to the ability of each scoring function to recognize the true ligand in the database. The curve labeled, 'minimum' represents the best theoretical consensus calculated from the five scoring functions.

(small active sites with large ligands but also large active sites with small ligands) and thus is representative of a diverse chemical space.

The Docking Procedure. In the present article, we aim at defining a nonlinear function through the use of a neural network that would permit an acceleration of the HTS process by defining focused libraries or for hitlist triaging

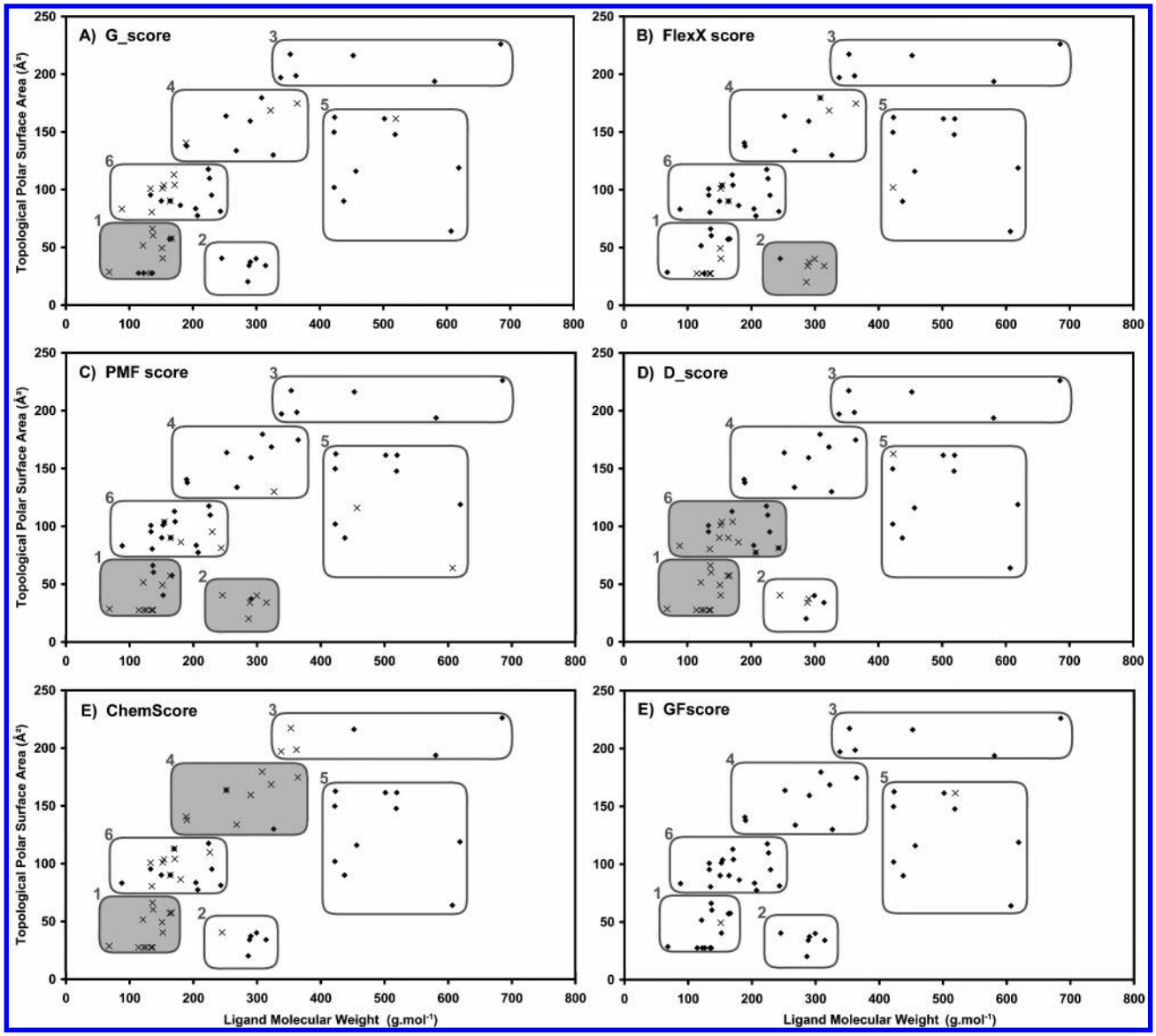

Figure 6. Analysis of the results for the five scoring[31] functions and for GFscore depending on the property of the chemical space. We divided the chemical space of the ligand data set into 6 areas. The ligand is represented by a square (♦) when it is ranked in the TOP25% and a cross (**x**) otherwise. The group corresponding to the chemical space is colored gray when more than 50% of the complexes are not correctly ranked by the scoring function showing a weakness in this chemical space.

when a prohibitively large number of hits is identified in the primary screen. To generate such a function, it is necessary to teach the neural network with accurate and reproducible data. Moreover, these data have to be representative of the complete chemical space. These different constraints have been demonstrated following the flowchart presented in Figure 4. We have selected a "diverse" database and a set of "diverse" protein−ligand complexes (see above for more details and Figures 1−3). We then verified our ability to reproduce the docking of the 3D structure for 200 protein−ligand complexes, the FlexX 200 data set. Among this data set, FlexX was able to generate 78 complexes with an accurate RMSD (<2.1 Å). We then hid each of these 78 ligands within the Diversity database and High Throughput docked each of them for each of the 78 targets. We know that our docking engine (FlexX) is able to correctly predict the pose for the 78 selected complexes. The question now is how to find which scoring function or which combination of scoring functions is most adapted to eliminate a maximum of true negatives, for each target. A plot of ranked versus each complex illustrates that globally D_score and ChemScore perform poorly compared to FlexX score and/or G_score (Figure 5). Figure 5A,B clearly demonstrates that, even if some functions are better for some complexes, they do not work for other complexes. This is the first indication that a linear function may not be sufficient to derive a "Generalist consensus scoring Function". We thus decided to analyze more precisely the different results for each scoring function.

We have classified the 78 complexes into 6 groups (Figure 6). The 24 "extreme" complexes are classified in 3 groups. Group 1 represents small hydrophobic ligands, group 2 medium-sized hydrophobic ligands, and group 3 the very polar ligands. The 54 other "average" complexes are also classified in 3 groups. Group 4 represents small- to average-sized ligands with average polarity, group 5 the average- to large-sized ligands with small to average polarity, and group 6 represents small ligands with average polarity. If the true

ligand is found in the TOP 25% after the docking of the entire database, it is represented as a square. If the function failed to rank the ligand in the TOP 25% after the docking of the entire database, it is represented as a cross. Finally, when a region of the chemical space is not properly predicted for at least 50% of the cases, we highlighted it in gray. The results are represented for the five functions in panel 6A−F. G_score and FlexX score seem to be the "best" functions. However, G_score failed in group 1, the small and hydrophobic ligands, while FlexX score failed in group 2, the medium-sized hydrophobic ligands. The three other functions failed in at least two different groups of the chemical space. PMFscore failed in group 1 and 2, the small- to medium-sized hydrophobic ligands but appeared to be very good in predicting the larger ligands with average polarity or very polar ligands (group 3 and top of group 4 and 5). D_score failed in group 1 and 6, for all the small ligands with hydrophobic or average polarity properties but is the best function for group 3, 4, and 5, the medium-sized to large polar ligands. Finally, the last scoring function in our training case is ChemScore, which clearly failed in group 1 and 4 but was able to predict well group 2 and 5, the medium-sized to large ligands, hydrophobic to partly hydrophilic ligands.

As a general conclusion, this analysis suggests that the 'Achilles' Heel' for each scoring function concerns group 1 (small hydrophobic region). Further work is therefore needed regarding this highly valuable chemical space, thus providing an important challenge for chemists soon. The same kind of results have been recently described in different studies and concluded the necessity to merge these different functions to increase the reliability of the results. Since linear combinations[4,7,8,14] and/or statistical approaches with Principal Component Analysis[33] have already been described, we decided to extend the principle of consensus scoring in a nonlinear neural network manner in order to check if such an approach could lead to substantial improvements.

GFscore. As discussed above, no single scoring function performs perfectly for the entire chemical space. In principle, it should be possible to determine a priori which score is best suited in a given case and then to use this score to decide whether a docking result is likely to represent a hit or not. However, the complexity of the relationship between the performance of a given score and the particularities of a given binding site and docking compound makes formulating explicit rules for such a relationship a daunting task. Therefore, we chose to use a neural network approach. Neural networks have been applied successfully in similar situations in the past[34,35] and represent a special case of supervised learning methods. Briefly, a complex nonlinear function with a reasonable number of free parameters is iteratively fitted to model a set of known input−output parameter relationships. Once trained on this learning data set, a neural network is expected to yield comparable output results when presented with new input data. Its actual performance can be analyzed by its evaluation against a data set that was not used in training the neural network (test data set). While the training of a neural network depends on a number of parameters that can sometimes only be selected on the basis of intelligent guessing at best, the final evaluation against the test data set yields an objective measure of the quality of the network.

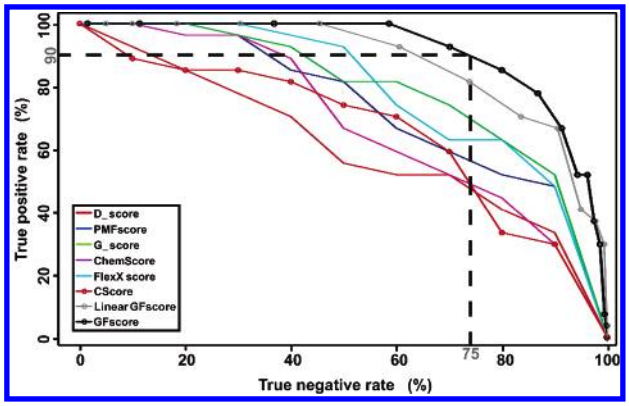

Figure 7. True negative rate versus true positive rate pinpoints GFscore efficiency compared to other scoring functions. GFscore eliminates about 75% of the database with a confidence rate of 90% when the other functions provide confidence rates of 70% for G_score, 62% for FlexX score, 55% for PMF score, 48% for CScore and ChemScore, and 46% for D_score, for the same amount of eliminated molecules. The linear version of GFscore (grey curve) eliminates 75% of the database with a confidence rate of 80% (versus 90% for the nonlinear scoring function) showing the interest of using a nonlinear approach.

To account for the properties of each individual docking situation (i.e. the type of binding cavity), the neural network was not trained using absolute docking scores directly; we instead used score ranks (see methods). For example, a given score may be systematically lower when docking the entire Diversity "Drug-Like" database against one protein as compared to another. However, if the true solution has high ranking scores in both cases, the neural network would still be able to spot such a pattern based on its rank. It is such a combination of neural network learning and the use of ranked scores as input parameters that define the originality of our approach.

Here we have developed GFscore on the basis of the set of 78 complexes and Diversity. The same approach has been developed in the group for the Prestwick database (http://www.prestwickchemical.com), and the results that we obtained are comparable to those presented in this work (unpublished data). The first step was a training of the neural network using the rank-based scores (Figure 7). For the screening, GFscore eliminates about 75% of the database with a confidence rate of 90% when the other functions provide confidence rates of 70% for G_score, 62% for FlexX score, 55% for PMF score, 48% for CScore and ChemScore, and 46% for D_score, for the same amount of eliminated molecules (Figure 7). We have also calculated the linear version of GFscore (that we are calling here L-GFscore). L-GFscore is able to eliminate 75% of the database with a confidence rate of 80%, which is better than each scoring function individually taken but still less accurate than its nonlinear version. Our Internet service (http://gfscore.cnrs-mrs.fr) permits a postprocessing analysis of hundreds of thousands of compounds (and even more), and the user only needs a preliminary docking on the diversity database and on its own database as seeding information.

To analyze the capacity of GFscore to work in a global chemical space for both the ligands and the targets, we have then decided to cluster the protein targets using the EC number (Supporting l Information). The 77 proteins can be classified in 7 families. The first 6 families represent 58 enzymes in general (EC 1.X to 6.X), while the seventh family

represents the "other" type of superfamilies (20 proteins). With GFscore, the average rank that we obtain for the true ligand hidden in the database, for the first six families (enzymes in general) is within the TOP 5% (3.33%), while the individual score for each superfamily (EC1.X to EC6.X) possesses a variability of ±2% around this average score illustrating very stable results in the "enzyme world" in general with no preferences for a particular class. This latter finding clearly supports the idea that GFscore should be a general consensus scoring function used for a general purpose, whatever the target. The score for the seventh family, the "other" cases (other than enzymes), points up an average value within the TOP 7% (6.91%). Obviously the nature of these PDB's is very different compared to the enzymes (surface, cavity depth, charges, etc., ...), and therefore the scores are understandably not as good as in the first case.

Given these data, it now seems obvious to us that GFscore (or any similar approach) should be employed instead of using a specific function, randomly chosen, for rescoring and ranking ligands during a Virtual Library Screening campaign, when no known inhibitor is available. Another important point is that, in our case, D_score and ChemScore give the worst results for ranking results from general library screening, as already noted by Terp et al. in 2001 during the definition of their MultiScore function.[33]

4. CONCLUSIONS

We have shown in this paper that an acceleration of the High Throughput screening process is now possible. Our method was developed using 78 protein/ligand complexes docked on a diverse database, 'Diversity' from NCI. Extending the principle of consensus scoring in a nonlinear neural network manner led us to introduce a nonlinear Generalist scoring Function, GFscore. This method is able to identify true negatives hidden in virtual databases. Verdonk et al. already validated that *rank-by-number* is most effective[36] as proposed in theory by Wang and Wang.[15] Indeed, consensus ranking typically does not perform better than the best of the individual scoring-function-combined, but it does provide a more robust ranking method when the performance of the individual scoring function is unknown. Using our neural network analysis, one can eliminate about 75% of a database without any a priori and send the resulting quarter to real experimental HTS process (focused libraries). In conclusion, it is now possible to develop a model encompassing a variety of different complexes and still be able to obtain an improved score relative to the scores obtained by individual scoring functions. However, if more accurate predictions are desired or when known inhibitors can be used, a more specific model could easily be derived. When the principles of GFscore are used (mainly no known inhibitor or during a primary screening), our method should have a great potential to be included in the process of virtual and real database screening.

ACKNOWLEDGMENT

We thank Drs. Eric Arnoult and J. C. Mozziconacci for their MOE scripts that have been useful to filter out compounds with unwanted chemistry and Dr. Philippe Roche for reading the manuscript. K.S. is also grateful to Prof. J. M. Claverie (head of IGS) for laboratory space and support. This work was partially supported by the French National AIDS Research Agency (ANRS AC14.3), Marseille-Nice Genopole, and the French National Genomic Network (RNG).

Supporting Information Available: PDB code for each of the 78 protein/ligand complexes used in the docking experiments plus a list of the filters used to filter out nondruglike compounds in the diversity database. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) DiMasi, J.; Hansen, R.; Grabowski, H. G. The price of innovation: new estimates of drug development costs. *J. Health Econ.* **2003**, *22*, 151−185.
(2) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.
(3) Cohen, S. S. A strategy for the chemotherapy of infectious disease. *Science* **1977**, *197*, 431−432.
(4) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333−344.
(5) Krovat, E. M.; Langer, T. Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1123−1129.
(6) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(7) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.
(8) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134−1146.
(9) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.
(10) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.
(11) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.
(12) Klon, A. E.; Glick, M.; Davies, J. W. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216−2224.
(13) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 2743−2749.
(14) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(15) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(16) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228−241.
(17) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(18) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(19) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 868−873.
(20) Sobolev, V.; Sorokine, A.; Prilusky, J.; Abola, E. E.; Edelman, M. Automated analysis of interatomic contacts in proteins. *Bioinformatics* **1999**, *15*, 327−332.
(21) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884−1897.
(22) Binkowski, T. A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.* **2003**, *31*, 3352−3355.
(23) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(24) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: placing discrete water molecules during protein−ligand docking predictions. *Proteins* **1999**, *34*, 17−28.
(25) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem*. **2003**, *46*, 2287−2303.
(26) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.
(27) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.
(28) Muegge, I.; Martin, Y. C.; Hajduk, P. J.; Fesik, S. W. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. *J. Med. Chem.* **1999**, *42*, 2498−2503.
(29) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(30) Irwin, J. J.; Shoichet, B. K. ZINC- -a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(31) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbours. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.
(32) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(33) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein−ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333−2343.
(34) Chen, H.; Zhou, H. X. Prediction of interface residues in protein−protein complexes by a consensus neural network method: test against NMR data. *Proteins* **2005**, *61*, 21−35.
(35) Palma, P. N.; Krippahl, L.; Wampler, J. E.; Moura, J. J. BiGGER: a new (soft) docking algorithm for predicting protein interactions. *Proteins* **2000**, *39*, 372−384.
(36) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W. et al. Virtual screening using protein−ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

CI0600758