Ligand−Protein DataBase: Linking Protein−Ligand Complex Structures to Binding Data

Olivier Roche,[†] Ryuichi Kiyama,[‡] and Charles L. Brooks, III*

Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037

Received November 1, 2000

In computational structure-based drug design, the scoring functions are the cornerstones to the success of design/discovery. Many approaches have been explored to improve their reliability and accuracy, leading to three families of scoring functions: force-field-based, knowledge-based, and empirical. The last family is the most widely used in association with docking algorithms because of its speed, even though such empirical scoring functions produce far too many false positives to be fully reliable. In this work, we describe a World Wide Web accessible database that gathers the structural information from known complexes of the PDB with experimental binding data. This database, the Ligand−Protein DataBase (LPDB), is designed to allow the selection of complexes based on various properties of receptors and ligands for the design and parametrization of new scoring functions or to assess and improve existing ones. Moreover, for each complex, a continuum of ligand positions ranging from the crystallographic position to points on the surface of the protein receptor allows an assessment of the energetic behavior of particular scoring functions.

Introduction

The search for new drugs entails the use of a broad range of computational techniques. They are involved in the generation of new databases of compounds[1,2] and in their refinement.[3,4] They are also useful for the selection of the new lead molecules performed by quantitative comparisons of the ligands or by structure-based drug design when the three-dimensional structure is available.[5] For structure-based drug design, many docking/scoring programs exist either to screen very large databases or to optimize lead molecules. If it is assumed that most of these programs perform very well in searching the conformational space in the binding site (the docking part),[6,7] the scoring functions still need improvements to enhance the reliability of discriminating correctly docked from misdocked conformations.[8] Many approaches have been used to improve the accuracy of scoring functions, leading to three families of functions: knowledge-based, force field-based, and empirical. The knowledge-based scoring functions use statistical analysis of three-dimensional complex structures to derive a sum of potentials of mean force between receptor and ligand atoms.[9−12] Force-field-based scoring functions use the classical molecular mechanics force fields, based on physical interactions, to compute the interaction energies (van der Waals and electrostatic) between the receptor and the ligand atoms. Often, they also add empirical terms to take into account the entropy and solvation changes.[13−16] The empirical scoring functions are based on the assumption that the binding free energy can be broken down into different countable contributions such as the number of hydrogen bonds, ionic interactions, apolar contacts, and entropy penalties for fixing rotatable bonds in docking the ligand onto the receptor. Moreover, they suppose that these terms are additive.[17−24]

Empirical scoring functions are the most widely utilized in current drug design/discovery software. However, because of their general lack of reliability, often a combination of scoring functions is used and a "consensus" is sought.[25] The difficult part of empirical scoring function design is the need to determine appropriate weights for each term. These weights are often calculated using multivariable regression methods to fit a training set of receptor−ligand complexes, requiring both high-resolution three-dimensional structures and experimentally known binding constants. Obviously, the larger and less correlated the training set, the better will be the potential for the parameters to yield correlation between energy and accuracy over a broad scope of possible ligands.[26]

The need of such a training set provides the motivation for our efforts to create a World Wide Web accessible database, the Ligand−Protein DataBase (LPDB), which gathers (at present) 195 complexes corresponding to 51 different receptors with both high-resolution structure and known experimental binding affinity. Each complex has been characterized by a set of one-dimensional descriptors, describing the individual features of the protein, the ligands, and the interaction interface. The selected descriptors, extracted from those used in existing empirical scoring functions, allow an accurate description of the complexes and thus an easy selection of targeted subsets that can be used to tune scoring functions. Principal component analysis (PCA) has been performed to understand the possible relationships between the distributions of complexes and to identify outliers and data clusters.[26] The correlation

* Corresponding author: Department of Molecular Biology (TPC6), The Scripps Research Institute, 10550 North Torrey Pines Rd, La Jolla, CA 92037. Tel: (858) 784-8035. Fax: (858) 784-8688. E-mail: brooks@scripps.edu.

[†] Current address: F. Hoffmann-La Roche AG, Pharmaceuticals Division, CH-4070 Basel, Switzerland.

[‡] Current address: Shionogi Research Laboratories, Osaka, Japan.

10.1021/jm000467k CCC: $20.00
Published on Web 09/18/2001

Table 1. PDB Codes for the Training Sets of Six Popular Empirical Scoring Functions

	scoring function					
	ChemScore[17]	Score1[18]	Hammerhead[35]	Autodock3[22]	Validate[21]	Score2[20]
training set	82	45	34	30	51	82
PDB codes	1aaq, 1apt, 1apu, 1hbv 1hpv, 1htf, 1htg, 1hvi, 1hvj, 1hvk, 1hvl, 1hvr, 1lyb 1ppk, 4hvp, 5hvp, 7hvp, 1bra, 1etr, 1ets, 1ett, 1ppc, 1pph, 1tmt, 1tng, 1tnh, 1tni, 1tnj, 1tnk, 1tnl, 3ptb, tmt1, 1cbx, 1mnc, 1tlp, 1tmn, 2tmn, 3cpa, 3tmn, 4tln, 4tmn, 5tln, 5tmn, 6cpa, 6tmn, 7cpa, 8cpa, 1abe, 1abf, 1apb, 1bap, 1dog, 1mfe, 1nsc, 1nsd, 2gbp, 2xis, 5abp, 5cna, 6abp, 7abp, 8abp, 9abp, 1adb, 1dih, 1ebg, 1hsl, 1mbi, 1pgp, 1phf, 1phg, 1rbp, 2cgr, 2cpp, 2ifb, 2tsc, 2ypi, 4dfr, 5cpp, 7dfr, dfr4, tsc2	3ptb, 1dwb, 4cha, 4tmn, 5tmn, 1tlp, 1tmn, 4tln, 1rne, 2er6, 4er2, 4er4, 9hvp, 4phv, 4hvp, 4dfr, 2tsc, 1stp, 1rbp, 2ifb, 2gbp, 1fkf, 2r04, 2phh, 4cna, 1mbi, 4hmg, 2ypi, 3cpa, 6cpa, 2xis, 1ulb + 13 unavailable complexes	7cpa, 1stp, 6cpa, 4tmn, 4dfr, 4phv, 1dwd, 5tmn, 2gbp, 1tlp, 1etr, 1tmn, 1rbp, 1ppc, 5tln, 1pph, 4dfr, 1ett, 1phf, 5cpp, 1xis, 2ifb, 1ulb, 2ypi, 3ptb, 2phh, 2tmn, 3ptb(2), 1dwd, 4tln, 3ptb(3), 4cha, 1dwb, 3ptb(4)	4cna, 3cpa, 6cpa, 2cpp, 4dfr, 1dwb, zer6, 1etr, 1ets, 1ett, 1fkf, 2gbp, 4hmg, 1hvj, 4hvp, 5hvp, 1hvr, 2ifb, 1mbi, 2mcp, 3ptb, 1rbp, 4tln, 1tlp, 1tmn, 4tmn, 5tmn, 1ulb, 2xis, 2ypi	4hvp, 7hvp, 5hvp, 9hvp, 1aaq, 4phv, 1tlp, 1tmn, 2tmn, 3tmn, 4tln, 4tmn, 5tmn, 6tmn, 7tln, 1eed, 2er0, 2er6, 2er7, 2er9, 3er3, 4er1, 4er4, 5er2, 1abe, 1abf, 1abp, 1bap, 9abp, 7abp, 8abp, 1tpa, 2ptc, 1sbn, 2sni, 3sic, 5sic, + 14 unavailable complexes	1add, 1bzm, 1cbx, 1cps, 1ctt, 1ela, 1elc, 1fkf 1hpv, 1hvr 1l82, 1l83, 1l86, 1l87, 1ldm, 1mbi, 1phe, 1phf, 1phg, 1ppc, 1pph, 1pso, 1rbp, 1rne, 1sbp, 1sre, 1stp, 1tlp, 1tmn, 1tnk, 2cpp, 2ctc, 2er6, 2gbp, 2gpb, 2ifb, 2tmn, 2tsc, 2tsc(2), 2xis, 2ypi, 3cpa, 3dfr, 3ptb, 3ptb(2), 3ptb(3), 3ptb(4), 3tpi, 4cha, 4cha(2), 4cha(3), 4cna, 4dfr, 4dfr(2), 4er2, 4er4, 4gr1, 4hvp, 4phv, 4tln, 4tmn, 4tsi, 5cpp, 5tim, 5tln, 5tmn, 5tmn(2), 6acn, 6cpa, 6rsa, 7cpa, 7cpp, 9aat, 1dwb + 7 unavailable complexes
standard error (kcal mol^{-1})	2.07	1.32	1.37	2.18	1.55	1.74

between the selected descriptors and the binding constant is also investigated using partial least squares (PLS).[26] Both linear statistical studies were performed with the default options available in the SIMCA-P8.0 software.[26] Moreover, the LPDB not only provides a large training set of complexes to improve the parametrization of empirical scoring functions but is also designed to assess the energy landscape of a particular scoring function by providing a continuum of ligand positions from the binding site to the receptor surface for each complex. With these sets of "decoys", we plan to assess many of the existing scoring/energy functions for their ability to discriminate docked from misdocked conformations. To date, several popular scoring functions such as AutoDOCK3,[22] DOCK4,[16] and FlexX1.9[23] have been tested. The results will be reported in a future paper.[27]

In this paper, we describe the overall structure and layout of the LPDB. The first section deals with the selection of the complexes, their characterization and parametrization for force field, as well as empirical energy functions. Next, we focus on the selected descriptors used to annotate the complexes and their statistical analysis. We then illustrate the processes used to generate the continuum of "decoy" positions. Finally, we illustrate possible queries that demonstrate the structure of the user interface.

Materials and Methods

Complex Selection. All of the selected complexes are extracted from the Protein DataBank (PDB).[28] The first step of the selection process was to analyze the training sets of the already well-known empirical scoring functions.[17−22] These training sets are summarized in Table 1. After removing all redundancies and complexes for which structural data was unavailable, we were left with 153 complexes.

We then mined the PDB to find new complexes, based on keywords such as "COMPLEX WITH", which yielded approximately 1000 potential complexes. For each hit, we examined the primary citation to assess whether the experimental binding constant was available. Unfortunately, defining automatic approaches to locate binding constant values is not possible since, most of the time, the value is not in the primary citation and the reference paper for the binding constant is too old to be available online. Therefore, this part was very time-consuming and has currently been done for only a small number of the selected complexes. Much still remains to be done; however, at the present time, the Ligand−Protein DataBase consists of 195 complexes divided among 51 different receptors (21 protein classes) and 178 different ligands. As in all of the previous training sets, we avoid covalently bonded complexes.

Parametrization. Our preparation of the suitable complexes for parametrization involved the following steps. For each complex, all of the water molecules were removed and, after a visual inspection, only the cofactors (heme, NADP, NAD, FAD) and structural or catalytic ions (Zn, Mg, Mn, Fe) were kept as a part of the receptor. When there were multiple positions for the ligand, only one was chosen. For multi-subunit structures, we only kept one ligand position.

In the first step of parametrization, we categorized the chemical environment of the ligand atoms using the atom-typing module of the commercial modeling program INSIGHT II, and then all the hydrogen atoms were added.[29] The partial charges were setup using the CHARMm force field[30] together with the typing and charging engine in INSIGHT II. The rest of the complex (receptor, cofactors, ions) was parametrized in the same force field, but using the academic version of CHARMM.[31] We then proceeded to identify whether suitable parameters exist for the force field of interest. Finally, we "regularized" the atomic positions using force-field-based minimization. The minimization consisted of 1000 steps of conjugate gradient (tolgrad 0.5) with a harmonic restraint on the protein backbone atoms. The restraint was incrementally removed during the minimization. The average root-mean-square deviation (RMSD), over all the complexes, between the crystallographic and the minimized forms is 0.34 Å (maximum at 0.66 Å for 1lgr). When we focus on the binding site, the RMSD for the residues within 3.6 Å of the ligand is 0.35 Å and the RMSD of the ligand itself is 0.37 Å. This underlines that the minimization does not change the conformation of the complex, the binding site and the ligand, but only allows for minor readjustment of the atom positions.

Properties. Defining useful and reliable properties to describe the complexes, receptors, and ligands is not a simple task, as numerous descriptors are available. Nevertheless, we isolated common features between many of the popular scoring functions that can lead to both binding constant prediction and specific subset selection.

Complex Features. For the selection of protein–ligand complexes, the resolution of the experimental structure is an important parameter since it reflects the precision of the crystallographic data. Often in training scoring and docking functions, high-resolution structures are used for training and lower-resolution structural data is utilized for validation.[21,32] The interaction surface area between the ligand and the receptor reflects the extent of ligand burial in the receptor. In addition to its use for complex selection, the interaction or buried surface can be part of the scoring function as a measure of desolvation.[10] As with the interaction surface area, the number of hydrogen and ionic bonds can be used in both scoring schemes and as a criteria for the selection of complexes. Finally, the experimental binding constant provides the link between the three-dimensional structure and experimental measurement.

Protein Features. The number of residues and the number of subunits usually characterize the proteins. However, most interesting is to select the proteins with respect to their class, as proteins gathered in a same class share many common features in reactivity and receptor topology. We defined the class based on the EC (Enzyme Commission) number or the name if the protein is not an enzyme. This kind of selection provides another useful means to specifically design/parametrize/assess a scoring function. We also focus on the binding site, defined as the residues that lie around 3.6 Å of the ligand crystallographic position. The formal charge and the number of residues were included as well.

Ligand Features. As for the protein moiety, the LPDB needs descriptors that give an accurate description of the ligand properties. Many descriptors exist to assess the ligand properties; we choose only simple descriptors that have been used in empirical scoring functions.[33,34] The number of rotatable bonds, as well as the number of rings, gives one an insight into the flexibility of the ligand. These properties are often used to assess the loss of conformational entropy due to binding, as in ChemScore,[17] FlexX1.9,[23] AutoDOCK3,[22] and Hammerhead.[35]

The number of donors and acceptors is clearly related to the possibility for the ligand to make hydrogen bonds and ionic interactions with the receptor. We have defined the H-bond donors as the sum of the polar hydrogen atoms on oxygen and nitrogen and the number of acceptors as the nitrogen and oxygen atoms, based on the CHARMm atom types.

The molecular weight of the ligand is well correlated with its size. The logarithm of the molecular weight is used in the Hammerhead scoring function as an approximation for the loss of rotational and translational entropy.[19]

In addition to the use of the previous structural descriptors, we have computed two widely used descriptors for global molecular properties: the calculated logP (ClogP), which is closely correlated with the hydrophobicity of the ligand, and the calculated molar refractivity (CMR), which is related to the binding force between polar portions of the receptor and its ligand. We have chosen Leo's method available from the Daylight package (version 3.05) for the calculation of these properties.[36]

The charge is a difficult parameter to handle, as most of the time the protonation states of the ligand and the residues in the binding site are not available. We have made the assumption that both the protein and the ligand have a protonation state corresponding to their free form at pH 7. However, an important feature of any scoring function is its flexibility, which means its ability to handle the lack of information and even misinformation.

Overall, we have tabulated 16 descriptors, divided into three classes, to annotate aspects of the complexes contained in the LPDB: (i) the resolution, the interaction surface area, and the number of ionic and hydrogen bonds to describe the complex; (ii) the number of basic and acidic residues, the total number of residues, and the formal charge for the protein; and (iii) the number of donors, acceptors, rotatable bonds, rings, and atoms, the molecular weight, ClogP, CMR, and the formal charge for the ligand.

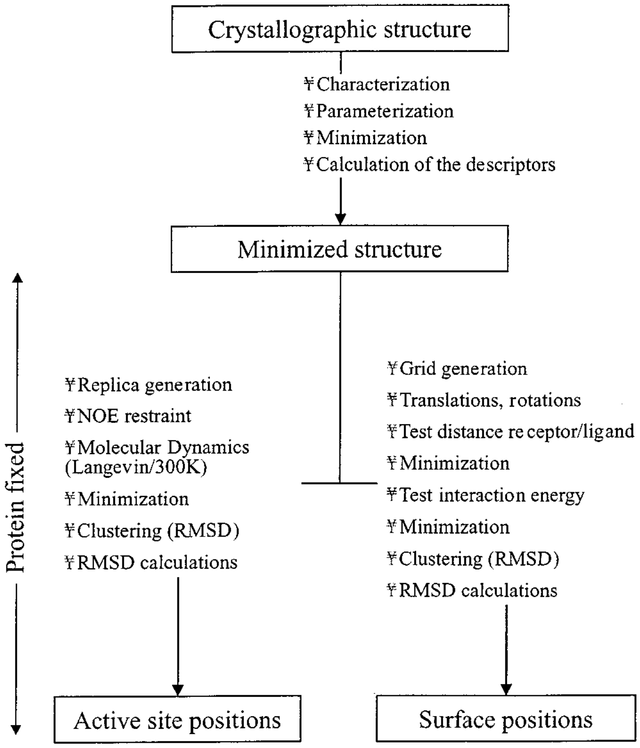

Figure 1. Flowchart describing the generation of a continuum of ligand positions.

Statistical Analysis. To evaluate the relevance of all the chosen descriptors on our data set of 195 complexes, a principal component analysis (PCA) has been performed. This method extracts a small set of orthogonal factors describing the data distribution. It helps to identify correlation between the descriptors and outliers in the data set. Then the projection on latent structures (PLS) method—a multidimensional linear regression technique—was used to see how the chosen descriptors are correlated with the binding constant.

All these descriptors serve our attempt to build a database that can handle very diverse subset selections that can lead to specific scoring function optimization or assessment. However, for an assessment of the energy landscape of such scoring functions, we need to consider not only the "native" ligand–protein complex, but also alternative "decoy" positions of the ligand.

Generation of a Continuum of "Decoy" Positions. For each complex in the LPDB, our goal is to provide a continuum of "decoy" positions that allow assessment of the energetic behavior of a particular scoring function from the binding site to the surface of the receptor (see Figure 1). For many docking programs such as DOCK4 and FlexX1.9, the docking/search process is carried out within a limited volume around a putative receptor site defined by the user. Thus, an assessment of the energy function in the vicinity of the binding site is sufficient. However, the development of genome-scale modeling efforts will provide more and more targets for structure-based drug design with unknown binding site positions. For this application, characterization of putative sites not in the vicinity of identified or known binding pockets is also important. There are already programs, including DOCK4, FlexX1.9, and AutoDOCK3, which can look for the binding sites before docking. In this case, it is useful to generate ligand positions at the surface of the receptor to assess the ability of the program to find the binding site location. As in all the actual docking programs, the protein is fixed for all the calculations.

Distributed Surface Decoys. Starting from the minimized complex structure, we constructed a set of surface distributed decoys by the following procedure. We translated the ligand to points on a spherical shell around the protein. The inner and outer radii of this shell depend of the size of the protein. The outer radius of the shell was defined to be the extreme distance between the geometric center of the protein and the protein surface plus 5 Å. We used the smallest distance as the minimum radius. The grid comprised 10 radial shells between these values with an angular resolution of $\pi/10$, resulting of 1820 points. We removed the points overlapping with the protein. For each remaining grid point, at least three global random rotations were performed to orient the ligand if the ligand had no rotatable bonds. Alternatively, we defined the number of global rotations based on the number of rotatable bonds with a maximum of 15. At the same time, the rotatable bonds adopted random values of −60°, 60°, or 180°. If the closest atom of the ligand was between 1 and 4 Å from the protein, we kept the position and minimized the conformation using the steepest descent method to find a nearby minimum (2000 steps, tolgrad 0.05). We tested the interaction energy of the new position, and if the energy was negative, we performed a second minimization using conjugate gradients (2000 steps, tolgrad 0.1). Finally, we calculated the RMSD of the final position. The average number of resulting ligand positions was 278.5 per complex.

Decoys in the Neighborhood of the Binding Site. Starting from the position of the ligand in the binding site of the minimized structure, we used the REPLICA (multiple copy) method in CHARMM together with molecular dynamics to explore the pathway for release of the ligand. Using this approach, our intent was not to explore all possible positions in the binding site neighborhood, as done in most docking algorithms. Instead we enabled guiding restraints to move the ligand from its original position to a position near the protein surface. We used short molecular dynamics runs (from 200 to 2000 steps) with coupling to a Langevin heat bath (300 K) to enable this sampling. Finally, we minimized the energy using 2000 steps of steepest descent (tolgrad 0.05) to obtain a new final position. We record the positions and RMSD with respect to the "native" complex.

Clustering of Misdocked and Near-Docked Decoys. We used the clustering facility in CHARMM, which is based on a nonhierarchical, K-means clustering algorithm,[37] to identify clusters representing 50 unique positions to be used to evaluate scoring functions. To obtain the desired number of positions, we varied the maximum radius of the clusters. For the surface positions, the clustering was based on the geometric center of the ligand and its RMSD with respect to its "native" position, as the RMSD by itself is not a sufficient discriminator due to the large conformational space represented by the spherical grid. For the positions generated from the binding site, we used only the RMSD, as the conformational space is much smaller. By merging these two sets of 50 positions, we obtained a continuum of ligand positions with reasonable conformations and known RMSD values. These conformations can be used to assess the energy landscape of various scoring functions.

All of the calculations (dynamics and minimization) and the extraction of the properties were performed automatically from the crystallographic structures after their parametrization for the CHARMm force field.[30] We used the programming language Python both as a scripting language for interacting with CHARMM and to compute the descriptors.[38]

Results and Discussion

For each complex, the descriptor values and the set of conformational decoys were integrated using the relational database management system MySQL.[39] As for the generation of the data, we used the programming language Python[38] to handle the database and create a CGI interface between the database and the World Wide Web, allowing for dynamic management of the overall structure. Calculations and database management are performed on a Beowulf Linux cluster of eight dual processors Pentium II 400 MHz PCs.

The Structure of the Queries to LPDB. There are many ways to retrieve information from the LPDB. The CGI interface allows a selection of the complexes based on all of the previously described descriptors (see Materials and Methods). The query interface is divided into three parts.

(i) The first one deals with the properties of the complex: the PDB code, the resolution, the interface surface area, and the experimental value of the binding constant. By selecting a resolution threshold of 2 Å, the LPDB can be divided into two subsets. The first high-resolution subset is composed of 101 complexes with a resolution lower or equal to 2 Å and the other one of 94 complexes.

(ii) The second interface allows selections based on the receptor properties, such as the name of the protein, its class, the number of residues, and the protein's formal charge. As a test case, the selection of the aspartic peptidases returns 53 complexes. However, if you constrain the query to HIV-proteases, only 27 complexes are chosen.

(iii) The last interface deals with the ligand properties. The entries for the query are the name of the ligand, the number of heavy atoms, the number of rotatable bonds, the number of donors and acceptors, the number of rings, the molecular weight, the ClogP, the CMR, and the charge. As an example of a query, we can ask how many of the LPDB ligands match the Lipinski's "rule of five".[40] We have to select a maximum of 10 acceptors, 5 donors, a maximum molecular weight of 500, and ClogP lower than 5.[41] The LPDB extracts 105 complexes with a ligand that matches these criteria.

If we use all the previous selection criteria, we retrieve only one complex: the penicillopepsin complex with a pepstatin analogue (1apu).

The Nature of Compiled Information in the LPDB. For each complex, the LPDB provides a description of the complex, receptor, and ligand properties based on the selected descriptors. It also provides the ligand SMILE code, the identity of the binding site residues (within 3.6 Å of the ligand), a 2D representation of the ligand, and the primary citation of the PDB file.

Furthermore, the user can download the original PDB file, the minimized PDB file, the CHARMm PSF file, and the MMFF mrk file for the complex. The files available for the receptor are the original PDB, the minimized PDB, and the SYBYL mol2 formats.[41] Finally, for the ligand, the LPDB gives the original PDB, the minimized PDB, the CHARMm RTF, and the SYBYL mol2 formats. From these formats, one can easily generate the required files for use in docking programs such as DOCK4, AutoDOCK3, FlexX1.9, and also the molecular modeling programs CHARMM/CHARMm and MMFF.

The 100 positions generated to assess the energetic landscape of an energy function can also be downloaded. Each position can be selected individually based on its RMSD from the minimized protein−ligand complex structure, or on its energy with respect to a particular scoring function. The header of each PDB file contains the RMSD associated with the position.

Table 2. Descriptor Distributions for the Data Set Used in the LPDB

	properties	data set (195 complexes)[a]
complex	binding constant (kcal mol^{-1})	−8.8 (3.5)
	resolution (Å)	2.1 (0.4)
	number of hydrogen and ionic bonds	9.3 (5.5)
	interaction surface area (Å^2)	849.4 (460.9)
receptor	number of residues	323.1 (154.8)
	number of acidic residues	40.2 (35.9)
	number of basic residues	33.4 (21.6)
	formal charge	−4.8 (9.5)
ligand	molecular weight (g mol^{-1})	412.7 (270.6)
	number of atoms	28.9 (19.4)
	number of rotatable bonds	14.9 (13.6)
	number of donors	4.5 (3.6)
	number of acceptors	8.4 (6.1)
	formal charge	−0.2 (1.0)
	number of rings	2.1 (1.6)
	ClogP	0.4 (3.8)
	CMR	10.9 (7.4)

[a] Average (standard deviation).

Statistical Characterization of the LPDB. Even though the current LPDB represents only a relatively small set for statistical characterization (195 complexes), we have explored the correlation between the experimental binding affinity and 16 selected descriptors using linear statistical methods. Table 2 provides on overview of the distribution of each descriptor, illustrating the diversity of our data set. The experimental binding affinity for complexes in the LPDB ranges between −2.03 kcal mol^{-1} (1tnk) and −19.03 kcal mol^{-1} (7cpa). A total of 77% of the binding constants fall in nano- and micromolar range, leading to scoring functions more effective in this area. The LPDB contains 51 different proteins divided into 21 classes. The most important classes are the aspartic proteases (28%), the oxidoreductases (21%), the serine proteases (14%), then the immunoglobulins, the metallo-proteins, the arabinose-binding proteins, the lyases, and the histocompatibility antigens (MHC) (~4.5% each). It is obvious that some families are much more present than others, introducing some bias in the sampling. However, only the availability of new complexes for the missing classes will change the balance. The structural resolution of LPDB complexes lies between 1.25 Å (2wea) and 3.16 Å (1dwb), but ~50% of the complexes have a resolution better than 2 Å, underlying the generally good quality of the structures. The interaction surface areas range from 240 Å^2 (1mbi) to 1905 Å^2 (1vaa), displaying a wide range of buried surface area. We also computed the number of residues within 3.6 Å from the ligand that defined the binding site. This number ranges from 1 (2cpp) to 19 (1hhj, 1vac) with an average of 8.3 residues per binding site. In the LPDB, 44% of the complexes have less than 10 rotatable bonds. Nevertheless, compounds from 0 (imidazole, thiazole) up to 54 rotatable bonds (peptoid ligand in 3er5) are available. The molecular weights range from 69.08 g mol^{-1} (imidazole) to 1269.54 g mol^{-1} (3er5). The molecular weight of 57% of the ligands is lower than 500 g mol^{-1}. Of the ligands, 68% have fewer than five donors and 76% have less than 10 acceptors. Finally, the ClogP of 92% of the ligands is less than 5. All the reference values correspond to the "rule of 5" described by Lipinski et al.[40] We also looked at the number of rings, which falls between 0 and 7 (1hvr), with a maximum at 1 for 35% of the complexes.

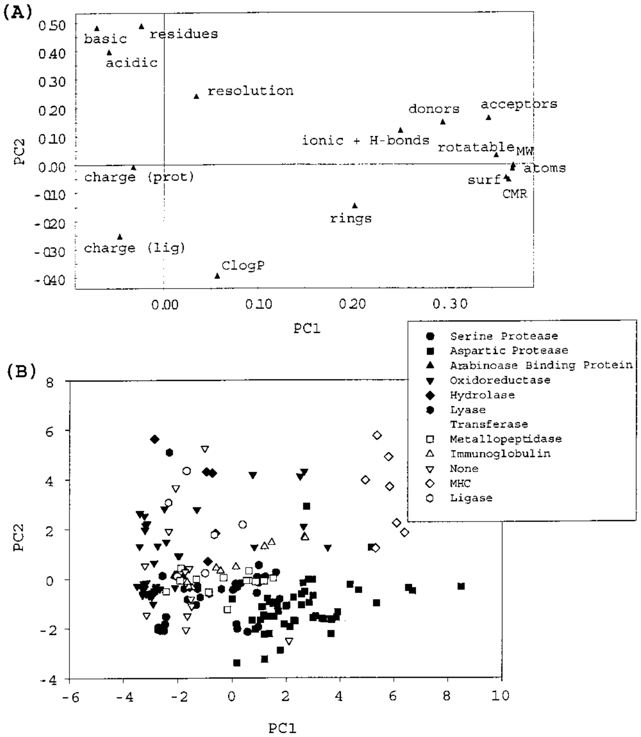

Figure 2. (A) Loading plot of the principal component projection of the 16 variables on the plane formed by the two first principal components. Ionic + H-bonds, resolution, surf refer to the complex properties; residues, basic, acid, charge (prot) to the protein properties; and donors, acceptors, rotatable, rings, MW, ClogP, CMR, atoms to the ligand properties. (B) Score plot of the principal component projection of the 195 complexes on the plane formed by the two first principal components (PC1 and PC2). The complexes are organized by the protein classes. The ellipse represents the projection of the 95% confidence region limit in the two-dimensional score plot.

PCA. From a principal component analysis on the descriptors in our data set, we find that the first principal component explains $R^2 = 44\%$ of the variance. It mainly takes into account descriptors linked to the size of the ligand, e.g., molecular weight, number of atoms, and rotatable bonds of the ligand, and also to the size of the ligand/receptor interface (interaction surface area and number of hydrogen and ionic bonds). This result does not obviously mean that the larger the ligand the better will be the affinity, since the data set is biased toward aspartic proteases, which bind large peptoid ligands with high affinity. The second component, which explains 19% of the variance, is more correlated to the size of the receptor and the lipophilicity of the ligand (ClogP and formal charge). It is obvious from the loading plot (Figure 2a) that some descriptors are highly correlated. However, on the whole they provide a global description of the data set, as displayed in the score plot (Figure 2b). This figure shows that there are no strong outliers in our data set. In this figure, the complexes are gathered in 12 different classes. The clusters mostly correspond to protein classes. Actually, this score plot can be viewed as a local map of the PDB complex space.

PLS. Using this multidimensional linear regression method, we assess the correlation between our descrip-

Table 3. Most Relevant GC Descriptors According to VIP Analysis

variables	VIP
number of hydrogen and ionic bonds	1.50
interaction surface area (Å^2)	1.31
CMR (ligand)	1.31
molecular weight (ligand)	1.29
number of atoms (ligand)	1.28
number of donors (ligand)	1.20
number of rotatable bonds (ligand)	1.16
number of acceptors (ligands)	1.14
ClogP (ligand)	1. 01

tors and the binding constant. The most important descriptors extracted with the variable influence on projection parameter (VIP) included in SIMCA-P are listed in Table 3. As expected, the descriptors that are linked to the interface between the ligand and the receptor, the number of hydrogen and ionic bonds, and the interaction surface area are the most correlated with the binding constant. This result underlines that important descriptors are those related to the interaction area, which play a crucial role in the enthalpic part of the binding energy. Also of note, the molecular refractivity, the molecular weight, the number of donors, or the number of rotatable bonds of the ligand are identified. These descriptors are often used to model the entropic part of the energy. The molecular weight could be correlated to the loss of translational and rotational entropy, the rotatable bonds to the loss of conformational entropy. However, in this work our goal is not to build a new scoring function but to validate our data set. All the obtained results are consistent with the previous approximations made in the classical empirical scoring functions and support a global use of the LPDB for the training and the assessment of scoring functions. Thus, it seems to be reasonable to design scoring terms based on properties that are reflective of the number of possible contacts between the receptor and the ligand as well as the size of the ligand.

Continuum Positions. All of the ligand decoy positions have been generated and minimized in the CHARMm force field, so they should have reasonable conformations. Concerning the surface positions, as the minimization is performed in a vacuum, the electrostatic interactions are emphasized over the lipophilic contacts, favoring the charged interaction sites. As we use molecular dynamics to generate binding site positions, we do not perform a comprehensive sampling of the binding site. Nevertheless, we generate a continuous pathway from the crystallographic position. Unfortunately, in some cases, ligand escape from the binding pocket required significant structural rearrangements in the binding site, not permitted by our protocol. In these instances we have not generated a true continuum of positions. Instead, there is a gap between the surface positions and the binding site positions. Generally, the RMSD ranges from around 0 Å to more than 60 Å. We have already used these positions as a benchmark for the DOCK4, AutoDOCK3, CHARMm, and MMFF energy functions. We will present the results of this work in an upcoming manuscript.[27]

Conclusions

Many specialized databases are currently emerging to exploit the rapid growth of experimental information and underlying the need to make logical connections between the in silico world and the bench. They range from global repository databases, such as GenBank (URL http://www.ncbi.nlm.nih.gov/Web/Genbank/), to very specialized ones, such as the G protein-coupled receptors database (URL http://swift.embl-heidelberg.de/7tm/), from genomic to protein databases, such as the receptor database (URL http://impact.nihs.go.jp/RDB.html); databases are involved in all the domains of the biological research. Focusing on structural databases, in contrast to the very useful database ReLiBase (URL http://rcsb.rutgers.edu:8081/home.html), the Ligand−Protein DataBase not only organizes and analyzes protein−ligand complexes from the PDB but also provides external information about the complexes such as the binding affinity, the ClogP, and the CMR.[42] The LPDB is clearly oriented toward the improvement of empirical scoring functions by providing a large number of complexes with known experimental binding affinity, which may be used to design and parametrize new scoring functions. It may also be used as a benchmark for assessing the behavior of existing and evolving scoring functions, since the LPDB provides a continuum of ligand positions for each complex. Discriminating between docked and misdocked positions is the first purpose of a scoring function; the second one is to give a good estimation of the binding constant. The LPDB provides a means to assess both of these objectives.

It may be noted that our efforts here complement an ongoing project from the NIST group headed by M. Gilson.[43] This effort, named BindingDB (http://www.bindingdb.org) is aimed at providing a database of binding data and experimental conditions but does not archive structural data.

In this first release, only a few descriptors have been used to describe the ligands and receptors. More complicated descriptors can be used to correlate specific complex properties to the experimental binding affinity in order to design more effective/accurate scoring functions, and this will be pursued in continuing development of the LPDB. So far, to our knowledge, no freely accessible tool exists to link the experimental affinity data to the high-resolution structural information. We believe the Ligand−Protein DataBase will be the “missing link”. It will be available in its preliminary version on the Scripps Research Institute (TSRI) server (URL http://lpdb.scripps.edu/). The LPDB will be licensed through TSRI and the license agreement information will be available on the Web site. The LPDB provides a deposit formulary to allow everyone to contribute to its extension. With help from the scientific community, we will try to update the LPDB constantly by adding new complexes, new scoring functions, and new descriptors.

Another goal of the LPDB is to share the results of the scoring function assessments based on the benchmark provided by the continuum of positions for each complex. Such a benchmark tool can be valuable to identify the essential terms to use in new empirical scoring function design in order to decrease the number of false positives and the standard error, which is around 1.2 log K_i (1.64 kcal mol^{-1}). Currently, we are testing several popular scoring functions such as DOCK4, AutoDOCK3, and FlexX1.9.

Acknowledgment. The authors thank Badry Bursulaya for his early involvement in database development, Gisbert Schneider and Hans-Joachim Boehm for their helpful comments, F. Hoffmann-La Roche for providing the SIMCA-P8.0 software, the NIH (GM37554, RR12255), Novartis Pharmaceutical Corporation and the Philippe foundation are acknowledged for financial support. Also, Dr Michael K. Gilson provided a preprint of the manuscript describing the BindingDB (reference 43).

References

(1) Schneider, G.; Lee, M. L.; Stahl, M.; Schneider, P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided. Mol. Des.* **2000**, *14*, 487−494.

(2) Koehler, R. T.; Villar, H. O. Statistical relationships among docking scores for different protein binding sites. *J. Comput. Aided. Mol. Des.* **2000**, *14*, 23−37.

(3) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput. Aided. Mol. Des.* **2000**, *14*, 251−264.

(4) Ajay; Bemis, G. W.; Murcko, M. A. Designing libraries with CNS activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.

(5) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − an overview. *Drug Discovery Today* **1998**, *3*, 148−155.

(6) Dixon, J. S. Evaluation of the CASP2 docking section. *Proteins* **1997**, *Suppl*, 198−204.

(7) Vieth, M.; Hirst, J. D.; Dominy, B. N.; Daigler, H.; Brooks, C. L., III Assessing search strategies for flexible docking. *J. Comput. Chem.* **1998**, *19*, 1623−1631.

(8) Vieth, M.; Hirst, J. D.; Kolinski, A.; Brooks, C. L., III. Assessing energy functions for flexible docking. *J. Comput. Chem.* **1998**, *19*, 1612−1622.

(9) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(10) Wallqvist, A.; Covell, D. G. Docking enzyme−inhibitor complexes using a preference-based free-energy surface. *Proteins* **1996**, *25*, 403−419.

(11) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.* **1995**, *8*, 677−691.

(12) Muegge, I.; Martin, Y. C. A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(13) Liu, M.; Wang, S. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *J. Comput. Aided. Mol. Des.* **1999**, *13*, 435−451.

(14) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.

(15) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M.; Vacca, J. P.; et al. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site [published erratum appears in *J. Med. Chem.* **1996**, *39* (11), 2280]. *J. Med. Chem.* **1995**, *38*, 305−317.

(16) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18*, 1175−1189.

(17) 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.

(18) Bohm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput. Aided. Mol. Des.* **1994**, *8*, 243−256.

(19) Jain, A. N. Scoring noncovalent protein−ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided. Mol. Des.* **1996**, *10*, 427−440.

(20) Bohm, H. J. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aided. Mol. Des.* **1998**, *12*, 309−323.

(21) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; et al. VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands. *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(22) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(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) 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.

(25) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(26) Tame, J. R. Scoring functions: a view from the bench. *J. Comput. Aided. Mol. Des.* **1999**, *13*, 99−108.

(27) Roche, O.; Brooks, C. L., III. Telling right from wrong: An assessment of the discriminatory properties of scoring functions used in docking. Manuscript in preparation.

(28) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; et al. The Protein Data Bank. *Nucleic. Acids. Res.* **2000**, *28*, 235−242.

(29) Molecular Simulations, I. *INSIGHT II*: San Diego, CA.

(30) Momany, F. A.; Rone, R. Validation of the general purpose QUANTA 3.2/CHARMm force field. *J. Comput. Chem.* **1992**, *13*, 888−900.

(31) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. et al. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(32) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(33) Livingston, D. The characterization of the chemical structure using molecular properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(34) Brown, R. D. Descriptors for diversity analysis. *Prescription Drug Discovery Des.* **1997**, *7/8*, 31−49.

(35) Welch, W.; Ruppert, J.; Jain, A. N. Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem. Biol.* **1996**, *3*, 449−462.

(36) Systems, D. C. I. *http://www.daylight.com*: Sante Fe, NM.

(37) Karpen, M. E.; Tobias, D. J.; Brooks, C. L., III. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: analysis of 2.2-ns trajectories of YPGDV. *Biochemistry* **1993**, *32*, 412−420.

(38) Python *http://www.python.org*.

(39) MySQL *http://www.mysql.com*.

(40) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(41) Tripos Associates, I. *SYBYL molecular modeling software*, 6.x ed.; St Louis, MO.

(42) Hendlich, M. Databases for protein−ligand complexes. *Acta. Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 1178−1182.

(43) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: A Web-Accessible Molecular Recognition Database. *Biopolymers/Nucleic Acid Sciences* **2001**, preprint.

JM000467K