SAnDReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions

Article *in* Combinatorial Chemistry & High Throughput Screening · September 2016

DOI: 10.2174/1386207319666160927111347

CITATIONS
8

READS
24,147

7 authors, including:

Mariana Morrone Xavier
Pontifícia Universidade Católica do Rio Grande …
10 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Gabriela Heck
Pontifícia Universidade Católica do Rio Grande …
9 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Maurício Boff de Ávila
Pontifícia Universidade Católica do Rio Grande …
12 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Nayara Maria Bernhardt Levin
Pontifícia Universidade Católica do Rio Grande …
11 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Combinatorial Chemistry & High Throughput Screening, 2016, *19*, 1-12

RESEARCH ARTICLE

SAnDReS a Computational Tool for Statistical Analysis of Docking Results and Development of Scoring Functions

Mariana Morrone Xavier[1], Gabriela Sehnem Heck[1], Mauricio Boff de Avila[1,2], Nayara Maria Bernhardt Levin[1,2], Val Oliveira Pintro[1], Nathália Lemes Carvalho[1], Walter Filgueira de Azevedo Jr.*[,1,2]

[1]*Laboratory of Computational Systems Biology, Faculty of Biosciences, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil;* [2]*Graduate Program in Cellular and Molecular Biology, Faculty of Bioscience, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681. Porto Alegre-RS 90619-900, Brazil*

Walter F. de Azevedo Jr.

Abstract: ***Background***: Docking allows to predict ligand binding to proteins, since the 3D-structure for the target is available. Several docking studies have been carried out to identify potential ligands for drug targets. Many of these studies resulted in the leads that were later developed as drugs.

Objective: Our goal here is to describe the development of an integrated computational tool to assess docking accuracy and build new scoring functions to predict ligand-binding affinity.

***Method*:** We carried out docking simulations using MVD program for a data set available on CSAR 2014 database (coagulation factor Xa) for which ligand-binding information and structures are available. These docking results were analyzed using SAnDReS available at www.sandres.net. Machine learning methods were applied to build new scoring functions and our results were compared with previously published benchmarks.

***Results*:** Our integrated docking strategy generated poses with docking accuracy higher than previously published benchmarks. In addition, the new scoring function developed using SAnDReS shows better performance than well-established scoring functions such the ones available in Autodock, Autodock-Vina, Gold, Glide, and MVD.

***Conclusion*:** The big data generated during docking lacked an integrated computational tool for statistical analysis of the influence of structural parameters on docking and scoring function performance. Here we describe methods to evaluate docking results using SAnDReS, a computational environment for statistical analysis of docking results and development of scoring functions. We believe that SAnDReS is a computational tool with potential to improve accuracy in docking projects.

ARTICLE HISTORY

Received: June 4, 2016
Revised: August 31, 2016
Accepted: September 14, 2016

DOI: 10.2174/1386207319666160927111347

Keywords: Dock, protein, target, drug, machine learning.

1. INTRODUCTION

The data explosion in the number of macromolecules deposited in the Protein Data Bank (PDB) [1-3] opens the possibility to investigate the correlation of these experimentally determined structures with functional information. This is a favorable scenario for application of computational systems biology approaches [4]. Such approaches can be used to develop mathematical models to predict ligand-binding affinity for a target protein. It is also possible to use these three-dimensional structures to study drug targets. The use of structural information makes possible to apply virtual screening (VS) methodology to identify novel hits and guide future development of new drugs. The main method to investigate potential new hits for a target protein is the procedure of protein-ligand docking simulations [5-11].

Protein-ligand docking simulations employ scoring functions to evaluate ligand-binding energy [10]. For validation of scoring functions, it is common to investigate the correlation between the experimental binding affinity with scoring functions. This statistical analysis can be based on squared Pearson's (R^2) or Spearman's (ρ) correlation coefficients [12]. Analysis of scoring function performance can also be carried out using data sets with active and decoy ligands, as proposed in the directory of useful decoys, enhanced (DUD-E) [13].

*Address correspondence to this author at the Faculty of Biosciences - Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre-RS 90619-900, Brazil; Tel/Fax: ++55-51-3353-4529; E-mails: walter.junior@pucrs.br, walter@azevedolab.net
MMX, GSH, MBA, NMBL, and VOP contributed equally to this work and should be considered first authors.

1386-2073/16 $58.00+.00

Furthermore, the richness of structural information opens the possibility to analyze molecular interactions that may be of pivotal importance for the success of protein-ligand docking simulations (re-docking and ensemble-docking simulations). If we focus our analysis on crystallographic structures, we could expect that crystallographic parameters such as resolution, R-free and R-factor [14] would exhibit some correlation with docking results. Moreover, deviations from ideal geometry, such as bond length, bond angle, and torsion angle [15] may also exhibit correlation with docking root-mean-square deviation (RMSD). It was with this in mind, that we developed SAnDReS, to integrate in one computational tool the statistical methods to investigate re-docking, ensemble docking, correlation of docking results with structural parameters, correlation of scoring functions with ligand-binding affinity, and a method to build polynomial scoring functions.

Here we describe the development of the SAnDReS and its application to an ensemble of crystallographic structures, for which binding affinity information is available. In the next sections, we describe the methods to evaluate docking accuracy, and discuss the docking results obtained for structures from Community-Structure Active Resource CSAR database, 2014 Benchmark Exercise for coagulation factor Xa.

2. METHODS

2.1. SAnDReS

SAnDReS is an acronym for Statistical Analysis of Docking Results and Scoring functions. SAnDReS was developed in Python programming language (version 3), using the SciPy (http://scipy.org/), NumPy (http://www.numpy.org/), SciKit-Learn (http://scikit-learn.org/stable/), and Matplotlib (http://matplotlib.org/) libraries. SAnDReS can analyze data generated by any protein-ligand docking program, the only requisite is to have protein-ligand structures in the PDB format, ligands in Structure Data Format (SDF), docking and scoring function data in comma separated values (CSV) format. SAnDReS automatically retrieves binding affinity information from PDB as a CSV file. The binding affinity CSV files bring a summary of experimental binding affinity data available in the PDB. For inhibitory constant (K_i), dissociation constant (K_d), half-maximal inhibitory concentration (IC_{50}), and half maximal effective concentration (EC_{50}) these data are expressed in nM (10^{-9}M). For thermodynamic data, such as Gibbs free energy of binding (ΔG) and enthalpy (ΔH), binding information is expressed in kJ/mol, as well as K_a in 1/M. This binding information was gathered from three other databases: MOAD [16], BindingDB [17] and PDBbind [18]. SAnDReS presents three programs, one is a GUI script to launch a SAnDReS window to manage all analysis that can be carried out. The second program is the SAnDReS main program, which is called by SAnDReS GUI window to run most of its tasks. The third program, called scikit_regression_methods_v1.py is an implementation of machine learning techniques for regression.

2.2. Statistical Analysis and Plots

SAnDReS calculates two correlation coefficients, squared correlation coefficient (R^2) and Spearman's rank correlation coefficient (ρ). R^2 is defined by equation (1)

$$R^2 = 1 - \frac{RSS}{TSS} \quad \text{(Eq. 1)}$$

RSS and TSS are defined by the following relationships:

$$TSS = \sum_{j=1}^{N} \left((y_j - \langle y \rangle) \right)^2$$

$$RSS = \sum_{j=1}^{N} \left(y_j - y_{calc,j} \right)$$

where $y_{calc,j}$ are the values obtained by feeding independent variables into the regression equation, y_j are the experimental observations, for instance $\log(K_i)$, <y> is the mean value for y, and N the number of observations.

The Spearman's rank correlation coefficient is defined by equation (2):

$$\rho = 1 - \frac{6 \sum_{j=1}^{N} d_j}{N(N^2 - 1)} \quad \text{(Eq. 2)}$$

where d_j is the difference in the ranks given to the two variable values for each item of data [12].

Calculation of statistical parameters such the correlation coefficients, p-values, maximum, minimum, median and mean values, F-stat, and standard deviation of docking results are based on SciPy and NumPy libraries. All protein-ligand docking results should be in CSV file to be readable by SAnDReS. Besides the statistical analysis of the correlation between RMSD and scoring functions, SAnDReS can also carry out statistical analysis of scoring functions and binding affinity.

SAnDReS calculates the correlation between experimental binding affinity and predicted values (scoring functions), where the binding information is automatically read from the PDB, as previously explained. SAnDReS can generate high-quality scatter plots for these CSV files, using a plot interface. All scatter plots are generated using Matplotlib library.

2.3. Overall Docking Strategy

There are different approaches to carrying out protein-ligand docking simulation, for instance, a recent study reported the development of a strategy on binding-pose selection and docking selection [19]. In the present work, we consider the selection of the biomolecular system (protein), docking simulations, scoring function development, and validation phases, which are all included in the flowchart shown in Fig. (**1**). This strategy is independent of the program used in the docking simulation. In Fig. (**1**), the grey boxes indicate the functions that can be carried out by SAnDReS, the other steps can be performed by any docking program.

For all structures discussed in this work, we adopted the molecular docking strategy described in the Fig. (**1**). Briefly, all molecular docking simulations were carried out by Molegro Virtual Docker (MVD) [20] and CLC Drug Discovery Workbench (http://www.clcbio.com/products/clc-drug-discovery-workbench/). MVD has shown better docking performance when compared with modern docking programs [20]. All protein and ligand atoms were prepared using default charge values for the programs MVD and CLC Drug Discovery Workbench. For each data set, the highest resolution structure was chosen as the most adequate for re-docking simulations. This structure was then submitted to the 32 docking protocols described in Table **1**. For each protocol, 1000 poses were generated. Besides the MolDock and Plants Scores, we also analyzed docking results using the additional scoring functions implemented in the program MVD, as described in Table **2**. For a full description of these scoring functions see the following references [20-22]. The most promising protocol was selected using as criteria the lowest RMSD and the highest correlation coefficient between the scoring function and docking RMSD.

After selecting the protocol, the rest of the structures in the ensemble were submitted to the same protocol for comparison. This procedure is referred to as ensemble dock, and the results were stored in a CSV file, where each line brings one structure in the ensemble. The ensemble-dock CSV file is used for two types of statistical analysis, one to investigate the correlation between docking RMSD and structural parameters and the second to evaluate the correlation between docking RMSD and scoring functions (ensemble dock).

We also calculate the scoring functions using MVD for each structure in the ensemble, using the crystallographic position of the active ligand. Our goal here is to test the accuracy of scoring functions in predicting binding affinity. We focus on crystallographic position in order to have the most reliable information to test the prediction ability of scoring functions. Next we describe the details of each task of SAnDReS program.

2.4. Download from PDB

SAnDReS has tasks to download structures and related binding affinity information from the PDB [1-3]. To download atomic coordinates in the PDB format, SAnDReS has the GETSTR task, which reads a CSV file (pdbCodes.csv) with the PDB access codes and directly downloads the structures from the PDB site. Since one of the major goals of any molecular docking simulation is to have a reliable model to predict ligand-binding affinity, SAnDReS has a task to download the experimental binding affinity from the PDB. This task is called GETBIND. It reads the same pdbCodes.csv file and downloads the binding affinity information as a CSV file. The structures and binding affinity information of structures discussed in the present work have been downloaded using SAnDReS tasks GETSTR and GETBIND.

2.5. Pre-Docking Analysis

The main goal of pre-docking analysis is to investigate the overall quality of the crystallographic structures in the data set. Pre-docking analysis is able to identify which structure in the data set has the most reliable crystallographic information, or the better overall stereo-chemical quality, using analysis of RMSD deviation from ideal geometry. This information is read from the PDB files in the data set. Prior to statistical analysis of the ensemble of structures, SAnDReS performs data filtering by eliminating repeated ligands in the data set. Pre-docking analysis (STSTRU task) identifies the structure for minimum and maximum values for structural parameters.

After statistical analysis of the ensemble of structures, SAnDReS generates the biological assembly for each structure, if the structure in the asymmetric unit is different from the biological assembly. Our goal here is to have a reliable biomolecular model for docking simulations, for instance, structures where the ligand-binding pockets are at interface between monomers of an oligomeric structure. What SAnDReS does is to read the rotation matrix and translation vector, present in REMARK 350 of the PDB file, and then applies the rotation and translation for each monomer to generate the biological assembly. Let's consider the structure of human purine nucleoside phophorylase (EC 2.4.2.1) as an example, the asymmetric unit content is a monomeric structure. The biological assembly is trimer [23]. Furthermore, the binding pocket is at interface between the monomers. Therefore, any docking project should focus on

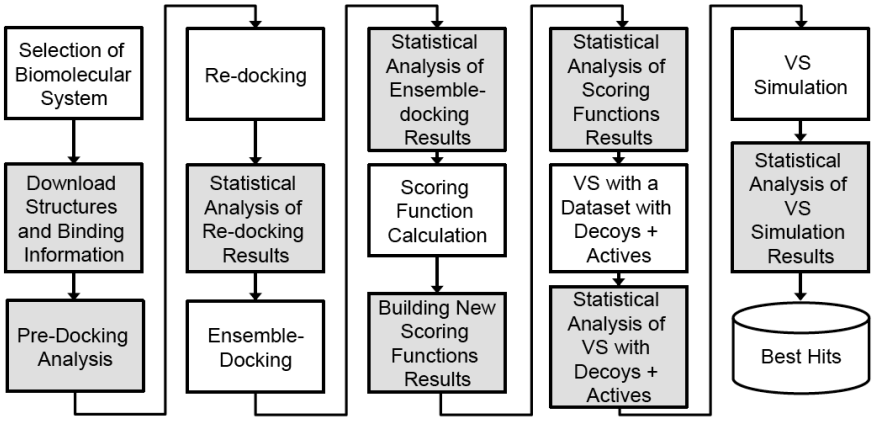

Fig. (1). Protein-ligand docking flowchart.

Table 1. Docking protocols.

Protocol	Scoring Function	Search Algorithm	Displaceable Water?
1	MolDock Score	MolDock Optimizer	Yes
2	MolDock Score	MolDock Optimizer	No
3	MolDock Score	MolDock (Simplex Evolution)(SE)	Yes
4	MolDock Score	MolDock (Simplex Evolution) (SE)	No
5	MolDock Score	Iterated Simplex	Yes
6	MolDock Score	Iterated Simplex	No
7	MolDock Score	Iterated Simplex (Ant Colony Optimization)	Yes
8	MolDock Score	Iterated Simplex (Ant Colony Optimization)	No
9	MolDock Score [GRID]	MolDock Optimizer	Yes
10	MolDock Score [GRID]	MolDock Optimizer	No
11	MolDock Score [GRID]	MolDock (Simplex Evolution) (SE)	Yes
12	MolDock Score [GRID]	MolDock (Simplex Evolution) (SE)	No
13	MolDock Score [GRID]	Iterated Simplex	Yes
14	MolDock Score [GRID]	Iterated Simplex	No
15	MolDock Score [GRID]	Iterated Simplex (Ant Colony Optimization)	Yes
16	MolDock Score [GRID]	Iterated Simplex (Ant Colony Optimization)	No
17	Plants Score	MolDock Optimizer	Yes
18	Plants Score	MolDock Optimizer	No
19	Plants Score	MolDock (Simplex Evolution) (SE)	Yes
20	Plants Score	MolDock (Simplex Evolution) (SE)	No
21	Plants Score	Iterated Simplex	Yes
22	Plants Score	Iterated Simplex	No
23	Plants Score	Iterated Simplex (Ant Colony Optimization)	Yes
24	Plants Score	Iterated Simplex (Ant Colony Optimization)	No
25	Plants Score [GRID]	MolDock Optimizer	Yes
26	Plants Score [GRID]	MolDock Optimizer	No
27	Plants Score [GRID]	MolDock (Simplex Evolution) (SE)	Yes
28	Plants Score [GRID]	MolDock (Simplex Evolution) (SE)	No
29	Plants Score [GRID]	Iterated Simplex	Yes
30	Plants Score [GRID]	Iterated Simplex	No
31	Plants Score [GRID]	Iterated Simplex (Ant Colony Optimization)	Yes
32	Plants Score [GRID]	Iterated Simplex (Ant Colony Optimization)	No

Standard protocols available for MVD and CLC Drug Discovery Workbench were applied.
GRID option means a faster scoring function calculation, without considering hydrogen-bond angles.

the biological assembly, to have a more reliable system to carry out protein-ligand docking simulations. If there is no REMARK 350 matrix and vector in the PDB file, no biological assembly will be generated, and the asymmetric unit content will be used for docking simulations.

In the next step of the Pre-docking analysis, SAnDReS reads each PDB entry in the ensemble of structures, and checks if there are water molecules close to the active ligand. SAnDReS tests all molecules inside a virtual sphere centered at the active ligand. After finishing finding the water molecules, SAnDReS writes all PDB files, for which water molecules were found inside the virtual sphere.

The last step of the pre-docking analysis is to generate a merged SDF file with all ligands in the data set. This file is intended to be used to create a data set with actives and decoys, to be described later.

2.6. Analysis of Re-Docking Results

In this analysis, SAnDReS evaluates the correlation between scoring functions and docking RMSD for each structure, which is performed by running STRMSD task. In addition the docking RMSD information, obtained from any

Table 2. List of all scoring functions used in this study.

Scoring Function	Description
MolDock Score	Protein ligand scoring function
Plants Score	Protein ligand scoring function
Re-rank Score	Protein ligand scoring function
Interaction Score	Total interaction energy between the pose and the target molecule(s)
Co-factor Score	Interaction energy between the pose and the co-factor(s)
Protein Score	Interaction energy between the pose and the protein
Water Score	Interaction energy between the pose and the water molecules
Internal Score	Internal energy of the pose
Electro Score	Short-range electrostatic protein-ligand interactions (r<4.5Å)
Electro Long Score	Long-range electrostatic protein-ligand interactions (r>4.5A)
H-Bond Score	Hydrogen bonding energy
LE1 Score	Ligand Efficiency 1: MolDock Score divided by Heavy Atoms count
LE3 Score	Ligand Efficiency 3: Re-rank Score divided by Heavy Atoms count
Docking Score	Score evaluated before post-processing (either Plants or MolDock). Only used for re-docking.
Displaced Water Score	Energy contributions from non-displaced and displaced water interactions (for odd number protocols in Table **1**).

docking program, SAnDReS uses this data to evaluate docking accuracy [24]. The equation to calculate docking accuracy (*DA1(a,b)*) has been implemented in the program SAnDReS as follows,

$$DA1(a, b) = f_a + 0.5(f_a - f_b) \quad \text{(Eq. 3)}$$

where f_a is the fraction poses for which the docking RMSD is less than a and f_b is the fraction poses for which the docking RMSD is less than b, where $a<b$. More recently, it has been proposed the use of an extended docking accuracy *DA2(a,b,c)* [19], which is also implemented in the program SAnDReS as follows:

$$DA2(a, b, c) = DA1(a, b) + 0.25(f_c - f_b) \quad \text{(Eq. 4)}$$

where f_c is the fraction poses for which the docking RMSD is less than c, where $a<b<c$, and *DA1* has been previously defined in the equation (3). In the current version of SAnDReS, the values for *a, b,* and *c* are 2.0, 3.0, and 4.0 Å, respectively. The main goal in the pre-docking is not only to use the docking RMSD as criterion to select the protocol for docking simulations, but also the correlation between docking RMSD and scoring functions.

2.7. Analysis of Structural Parameters

We can analyze the correlation of docking (ensemble docking) results against structural parameters (derived from the crystallographic information stored in the PDB files of the data set), this is done by running STDOCK task. SAnDReS investigates the correlation between docking RMSD and over one hundred structural parameters (supplementary material 1). The main goal of STDOCK task is to investigate correlation of crystallographic parameters, such as R-factor and R-free with the docking results. It is noteworthy, that since SAnDReS parses structural parameters directly from the information stored in the PDB files, the data is noisy due to the lack of standardization in crystallographic refinement programs [25-27]. For instance, a structural parameter such as phase error (obtained from maximum likelihood method) is calculated in the structure refined using the program PHENIX [27], but it is not calculated in another crystallographic refinement program such as TNT-Buster [26]. Therefore, some of the structural parameters may be missing in certain structures of the data set. Besides structural parameters directly read from the PDB files, SAnDReS uses this information to calculate mean values for B-factor and occupancy. In addition, SAnDReS calculates modified Matthews coefficients for protein content inside a virtual sphere centered at the active ligand. The idea is to investigate the correlation of modified Matthew coefficient with docking RMSD. Modified Matthews (V_M^*) coefficient is calculated using equation (5),

$$V_M^* = \frac{V_{sphere}}{MW} \quad \text{(Eq. 5)}$$

where V_{sphere} is the volume of a virtual sphere centered at active ligand and MW is the molecular weight the protein inside the virtual sphere. The volume is expressed in $Å^3$ and MW in Daltons. The original Matthews coefficient is used to evaluate solvent content in protein crystals and estimate the number of protein molecules in the unit cell [28]. It is determined dividing the volume of unit cell by its protein content. Besides the modified Matthews coefficient, SAnDReS also calculates the mean values for occupancy and B-factors inside each virtual sphere centered at the active ligand. The goal here is to evaluate the correlation of these structural parameters close to the active ligand, which may reveal some influence on docking results.

Molecular docking simulation relies heavily on structural information derived from crystallographic structures, so it is natural to open the possibility to investigate the influence of structural parameters in the docking results.

2.8. Analysis of Ensemble Docking

Here we intend to analyze the correlation between docking RMSD and scoring functions. Differently from the re-docking analysis, here we have the docking RMSDs for an ensemble of structures. This analysis can be performed on two types of data sets. The first type is one where all structures are of the same protein. The second type occurs when we have several different proteins in the same data set.

2.9. Scoring Function

The goal in this step is to test the performance of scoring functions in predicting binding affinity. To carry out the

analysis of correlation between the binding affinity and the scoring functions, SAnDReS runs STSCOR task.

2.10. Polynomial Scoring Functions

SAnDReS allows using the scoring functions as templates to build new polynomial scoring functions, where each term in the polynomial equation is a scoring function. This opens the possibility to build new regression models, based on the scoring functions used in the previous analysis. We could also use descriptors in this analysis, but our discussion here is focused on scoring functions. The current version of SAnDReS (1.0.1) builds polynomial scoring functions up to three independent variables, where each independent variable is a scoring function, or a mixed term involving two different scoring functions or the squared scoring function. What SAnDReS does is to find the coefficients (weights) for the polynomial equation indicated below using regression analysis,

$$\begin{aligned} score = {} & \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \\ & \omega_4 x_1 x_2 + \omega_5 x_1 x_3 + \omega_6 x_2 x_3 + \\ & \omega_7 x_1^2 + \omega_8 x_2^2 + \omega_9 x_3^2 \end{aligned} \quad \text{(Eq. 6)}$$

where *score* is the scoring function value, ω_0 is the regression constant, and other ω's are the weights for each independent variable in the equation. SAnDReS uses machine learning methods from SciKit-Learn library to carry out regression analysis [29]. Since we have 9 terms in the equation (6), we can have up to 511 different polynomial equations (2^9 -1), we don't consider the equation *score* = ω_0.

SAnDReS generates 9-bit binary strings (bit strings) to build the polynomial equations, being the first 000000001 and the last 111111111. When we have "1" means that the term will be included in the polynomial equation and "0" means that the equivalent term will be omitted from the equation. For instance, the bit string 100000001 represents polynomial equation *score* = ω_0 + ω_1 x_1 + ω_9 x_3^2. A simplified version of the polynomial scoring function method has been previously described for the program Polscore [30]. For binding information K_i, K_d, IC_{50}, EC_{50}, and K_a, it is used the log of the value, for instance log(IC_{50}). For thermodynamics functions (ΔG and ΔH), it is used the values in kcal/mol. The percentage of the data to be used in the training set is determined by the user. The default value for the percentage of ligands in the training set is approximately 70%, as suggested by Cichero *et al* 2010 [31].

The training set will be used to build the regression model and the test set will be used to evaluate the predictive ability of the regression model. After including new polynomial equations for evaluating ligand-binding affinity, new round of scoring function analysis can be carried out.

In the present work, we used the MolDock Score ($E_{MolDock\ Score}$) implemented in the program MVD to calculate the protein-ligand interaction energy defined as follows:

$$E_{MolDock\ Score} = E_{intra} + E_{inter} \quad \text{(Eq. 7)}$$

where the term E_{inter} is intermolecular energy, for the protein-ligand structure. This term is computed as follows,

$$E_{inter} = \sum_{i=1}^{N1}\sum_{j=1}^{N2}\left(E_{PLP}\left(r_{ij}\right) + 332.0\frac{q_i q_j}{4r_{ij}^2}\right) \quad \text{(Eq. 8)}$$

In the above equation, N1 and N2, indicate the numbers of atoms in the ligand and protein, respectively. The component E_{PLP} indicates the piecewise linear potential described elsewhere [20] and r_{ij} accounts for interatomic distance. The second term is an electrostatic potential, where q_i and q_j represent the punctual electric charges for ligand and protein atoms, respectively. The intramolecular term of equation (7) (E_{intra}) is calculated as:

$$E_{intra} = \sum_{i=1}^{N1}\sum_{j=1}^{N1}\left(E_{PLP}\left(r_{ij}\right) + E_{clash}\right) + \sum_{\theta\ \in\ rotatable\ \ bonds} A\left[1-\cos\left(m.\theta-\theta_0\right)\right] \quad \text{(Eq. 9)}$$

N1 and r_{ij} are the same as defined for equation (8), but for equation (9), we have a double summation for N1 non-hydrogen atoms in the ligand, which are more than 2 bonds apart and computes the E_{PLP} and E_{clash} for each atom pair. E_{clash} is a fixed penalty term of 1000 given to intra-atomic distance <2 Å. Furthermore, we have also a term for torsion energy, which is determined for torsional angles found in the ligand. The terms m, θ_0 and A are described by Thomsen and Christensen, 2006 [20]. The term θ is the dihedral angle. We also analyzed docking results using the additional scoring functions implemented in the program MVD, as described in Table **2**. There scoring functions were used as terms (x_j) in the equation (6).

2.11. Decoys and Actives

SAnDReS is able to generate user-defined data sets which are composed of decoys + actives. This data set is partially based on DUD-E data [13]. SAnDReS can merge in one SDF file, the actives being studied in the data set (actives for which there are crystallographic structures) and part of the decoys available in the DUD-E data set. The decoys are gathered from the DUD-E database. SAnDReS generates a new merged file with actives and decoys. The decoys are randomly selected from the DUD-E file specified by the user. In addition, the user can define the percentage of the actives and decoys in the data set.

This SAnDReS-generated data set can be used to test the ability of a protein-ligand docking program to find active ligands embedded in a data set composed of decoys and actives. To do so, we run a small VS, using the best docking protocol, previously selected for the program MVD, and rank the ligands using MVD scoring functions and polynomial scoring functions. In this analysis, SAnDReS calculates enrichment factors, and generates receiver operating characteristics (ROC) curves for evaluation of scoring function performance. SAnDReS generates ROC curves based on the data generated in a VS simulation focused on a data set composed of decoy and active ligands. The ROC curve is a standard method of machine learning research and it is valuable graphical tool for analysis of docking results, since it shows the overall performance of a binary classifier system, specifically discriminating between actives and decoys. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Plots are generated using Matplotlib library and the area under the curve for ROC curves are calculated using SciKit-learn library [29].

SAnDReS uses decoy and active results to determine enrichment factor [32], as defined below,

$$EF = \frac{H_a / H_t}{a / N} \quad \text{(Eq. 10)}$$

where H_a is the number of active ligands in the n top-ranked compounds (H_t) of a total database of N compounds of which a indicates the number of actives. It is expected $EF>>1$ for successful VS simulations. SAnDReS calculates EF for top-ranked compounds for 1 %, 2 %, 5 %, 10 % and 20 % of N, named *EF1, EF2, EF5, EF10*, and *EF20*, respectively.

2.12. Data Set

The structures were obtained from Community-Structure Active Resource CSAR database (2014 Benchmark Exercise for coagulation factor Xa (EC 3.4.21.6)) (http://csardock.org/). We filtered the data set to include only structures for which K_i information was available in the PDB and water molecules were inside a sphere centered at active ligand with radius of 15.0 Å. We ended up with an ensemble of 82 structures out 125 available at CSAR database. This data set will be referred to as Xa data set. The PDB access codes for all structures in the Xa data set are shown in the supplementary material 2.

3. RESULTS AND DISCUSSION

3.1. Analysis of Pre-Docking and Re-Docking Results

Using the X-ray crystallographic resolution as a selection criterion, SAnDReS identified the PDB access code 2JKH [33] as a best structure for re-docking simulation. It is noteworthy that, a lower numerical value of crystallographic resolution means better overall X-ray crystallographic data. The structure 2JKH was employed for re-docking simulations, using the 32 docking protocols listed in the Table **1**.

If we consider the lowest docking RMSD, the best protocol is the number 23. This protocol uses as search engine Iterated Simplex (Ant Colony Optimization) algorithm [20]. To rank results the program MVD makes possible to apply all scoring functions available in Table **2**. For protocol 23, low values of docking RMSDs (< 1.0 Å) were observed for most of the scoring functions (Plants, MolDock, Re-rank, Interaction, Protein, LE1, LE3, and Docking Scores), which indicates that we could use any of these scoring functions to rank our docking results obtained with protocol 23. The highest Spearman's rank correlation coefficient (ρ = 0.923), was obtained for Plants Score function, with p-value < 0.001. Protocol 23 was used to carry out docking simulation for the rest of the entries in the ensemble of crystallographic structures (ensemble docking).

3.2. Statistical Analysis of Structural Parameters

Linking structural parameters and model quality is a common procedure in macromolecular X-ray crystallography [34-36]. On the other hand, linking structural parameters and docking results is rare [15]. Analysis of the correlation between docking RMSD and structural parameters is intended to identify key features that may influence docking results. We considered structural parameters statistically significant if the p-value1 < 0.05 (p-value1 is the p-value for Spearman's correlation). Only the top-ranked structural parameter is discussed here.

The lowest p-value was observed for solvent content parameter (ρ= -0.247 and p-value1 < 0.029) (supplementary material 3). The scatter plot for solvent content *vs* docking RMSD is shown Fig. (**2**). Since the value of ρ= -0.247, it indicates only a weak correlation between the solvent content and docking RMSD. It is tempting to speculate, that this correlation may be attributable in part to the fact that higher solvent content contributed positively to the ligand diffusion in the preparation of protein crystals. It was observed in crystal soaking experiments, that the ligand-binding process is facilitated in high-solvent protein crystals, which allows high occupation of the binding pocket by the ligand [37].

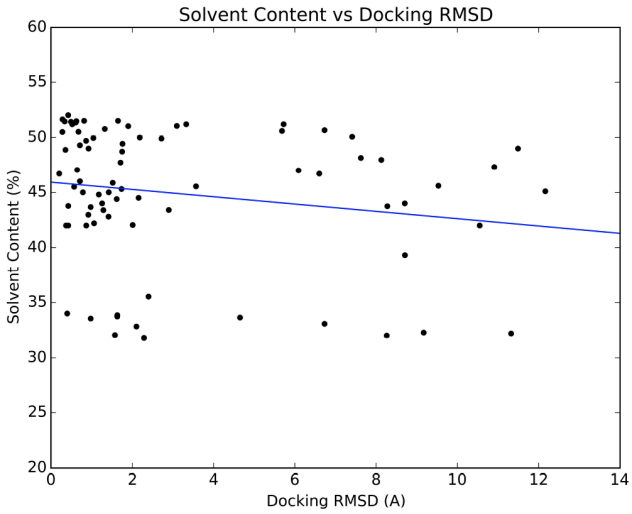

Fig. (2). Scatter plot for solvent content *vs* docking RMSD.

3.3. Ensemble Docking

The structures in this data set have been employed in a previous docking study [38], which allows us to carry out comparative analysis. For this data set, our results indicated docking accuracies of 64.024 % and 64.939 %, for DA1(2,3) and DA2(2,3,4), respectively. These results are higher than docking accuracy (DA1) reported for the same data set [38]. Table **3** shows the analysis of the correlation between scoring functions and docking RMSD for all 82 crystallographic structures in the data set. The highest correlation coefficients (ρ= 0.736 and p-value1 < 0.001; R^2 =0.349 and p-value2 < 0.001) were observed for MolDock Score (Fig. (**3**)). The docking RMSD for the lowest value of MolDock Score is 0.678 Å, better than the mean docking RMSD (1.728 Å) described for a previous study [38] that used the following programs: Glide (Schrödinger, http://www.schrodinger.com/), Gold [39]; Autodock [40], and Audodock Vina [41].

Table 3. Ensemble docking results.

Score	ρ	p-value1	R^2	p-value2
Plants	-0.192	$8.437.10^{-02}$	0.125	$1.140.10^{-03}$
MolDock	0.736	$3.454.10^{-15}$	0.349	$5.272.10^{-09}$
Re-rank	0.450	$2.203.10^{-05}$	0.107	$2.671.10^{-03}$
Interaction	0.680	$2.201.10^{-12}$	0.312	$4.905.10^{-08}$
Co-factor	-0.003	$9.776.10^{-01}$	0.001	$8.104.10^{-01}$
Protein	0.676	$3.048.10^{-12}$	0.315	$4.215.10^{-08}$
Internal	0.021	$8.494.10^{-01}$	0.001	$7.583.10^{-01}$
Electro	-0.008	$9.432.10^{-01}$	0.003	$6.270.10^{-01}$
Electro Long	-0.052	$6.398.10^{-01}$	0.025	$1.520.10^{-01}$
H-Bond	0.157	$1.602.10^{-01}$	0.007	$4.453.10^{-01}$
LE1	0.619	$5.669.10^{-10}$	0.063	$2.308.10^{-02}$
LE3	0.637	$1.277.10^{-10}$	0.127	$1.028.10^{-03}$
Docking	-0.192	$8.437.10^{-02}$	0.125	$1.140.10^{-03}$

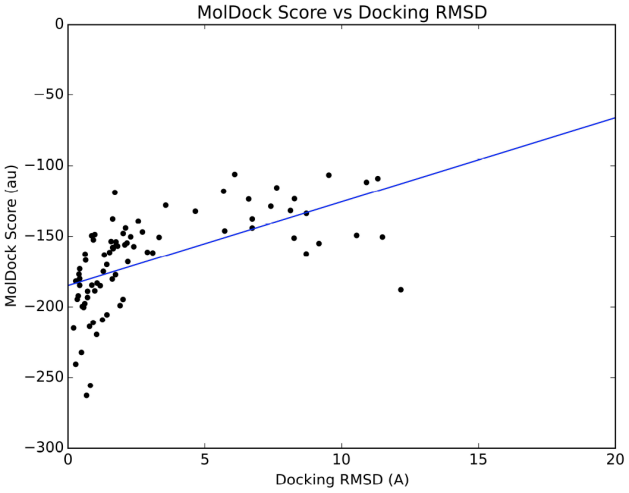

Fig. (3). Scatter plot for MolDock score *vs* docking RMSD.

3.4. Scoring Functions

Analysis of correlation coefficients between scoring functions and log(K_i) for data set is shown in Table **4**.

The highest correlation was observed for MolDock Score (ρ= 0.736 and p-value1 < 0.001, R^2 = 0.349 and p-value2 < 0.001). This correlation is higher than a recently published study [38] (ρ ranging from -0.1382 to +0.0954 and R^2 ranging from 0.005863 to 0.015903), involving the same docking programs previously mentioned [39-41].

In addition, we also applied the polynomial scoring function methodology to data set. Table **5** summarizes the results. The best result was obtained for polynomial equation 110 with ρ= 0.56 (p-value < 0.001) for the training set (57 structures) and ρ = 0.435 (p-value = 0.02975) for a test set (25 structures). Fig. (**4**) shows the scatter plot for polynomial equation 110 vs log(K_i), with training set data.

Table 4. Correlation between scores and log(K_i).

Score	ρ	p-value1	R^2	p-value2
MolDock	0.736	$3.454.10^{-15}$	0.349	$5.272.10^{-09}$
Re-rank	0.450	$2.203.10^{-05}$	0.107	$2.671.10^{-03}$
Interaction	0.680	$2.201.10^{-12}$	0.312	$4.905.10^{-08}$
Co-factor	-0.003	$9.776.10^{-01}$	0.001	$8.104.10^{-01}$
Protein	0.676	$3.048.10^{-12}$	0.315	$4.215.10^{-08}$
Internal	0.021	$8.494.10^{-01}$	0.001	$7.583.10^{-01}$
Electro	-0.008	$9.432.10^{-01}$	0.003	$6.270.10^{-01}$
Electro Long	-0.052	$6.398.10^{-01}$	0.025	$1.520.10^{-01}$
H-bond	0.157	$1.602.10^{-01}$	0.007	$4.453.10^{-01}$
LE1	0.619	$5.669.10^{-10}$	0.063	$2.308.10^{-02}$
LE3	0.637	$1.277.10^{-10}$	0.127	$1.028.10^{-03}$

Below we have polynomial equation 110, with coefficients determined by regression analysis,

$$\text{score}_{110} = 1.603905 - 0.006304.(\text{Electro Score}) - 0.005256.(\text{MolDock Score}).(\text{Interaction Score}) - 0.000028.(\text{Interaction Score}).(\text{Electro Score}) + 0.002801.(\text{MolDock Score2})^2 + 0.002439.(\text{Interaction Score})^2$$

The polynomial equation 110 uses Electro Score, (MolDock Score).(Interaction Score), (Interaction Score).(Electro Score), (MolDock Score)^2 and (Interaction Score)^2 functions as independent variables. The highest coefficient in the regression model was obtained for Electro Score, which considers short-range interactions (interatomic distance < 4.5 Å)[20]. In polynomial equation 110 the Electro Score is the second term in equation (8), MolDock Score is defined in the equation (7). The interaction score is the total interaction energy between the ligand and the protein, as defined in the equation (8).

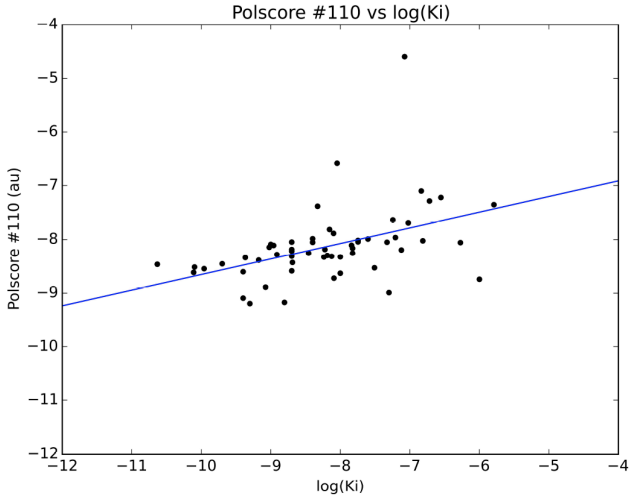

Fig. (4). Scatter plot for predicted and experimental binding affinity.

Table 5. Results for training and test sets.

Score	ρ (training set)	p-value (training set)	ρ (test set)	p-value (test set)
MolDock	0.160	$2.335.10^{-01}$	0.396	$4.995.10^{-02}$
Re-rank	0.181	$1.766.10^{-01}$	0.333	$1.038.10^{-01}$
Interaction	0.190	$1.574.10^{-01}$	0.174	$4.055.10^{-01}$
Co-factor	0.140	$2.999.10^{-01}$	0.024	$9.087.10^{-01}$
Protein	0.140	$2.976.10^{-01}$	0.141	$5.005.10^{-01}$
Internal	-0.005	$9.696.10^{-01}$	0.079	$7.077.10^{-01}$
Electro	-0.083	$5.370.10^{-01}$	-0.325	$1.132.10^{-01}$
Electro Long	0.075	$5.794.10^{-01}$	-0.231	$2.666.10^{-01}$
H-bond	0.129	$3.390.10^{-01}$	-0.127	$5.463.10^{-01}$
LE1	0.004	$9.759.10^{-01}$	0.466	$1.882.10^{-02}$
LE3	0.151	$2.612.10^{-01}$	0.296	$1.507.10^{-01}$
Polscore #38	0.525	$2.809.10^{-05}$	0.193	$3.546.10^{-01}$
Polscore #506	0.600	$8.166.10^{-07}$	0.044	$8.364.10^{-01}$
Polscore #442	0.500	$7.479.10^{-05}$	0.032	$8.795.10^{-01}$
Polscore #510	0.539	$1.539.10^{-05}$	0.077	$7.159.10^{-01}$
Polscore #110	0.560	$5.920.10^{-06}$	0.435	$2.975.10^{-02}$
Polscore #126	0.561	$5.624.10^{-06}$	0.317	$1.228.10^{-01}$
Polscore #422	0.499	$7.870.10^{-05}$	0.306	$1.362.10^{-01}$
Polscore #166	0.519	$3.576.10^{-05}$	0.226	$2.782.10^{-01}$
Polscore #294	0.520	$3.391.10^{-05}$	0.226	$2.782.10^{-01}$
Polscore #478	0.574	$3.109.10^{-06}$	0.049	$8.150.10^{-01}$

The prevalence of electrostatic intermolecular interactions may be due to the presence of charged residues in the binding pocket (Arg 143, Gln 192, and Asp 189), as shown in Fig. (**5**). It has been shown the importance of electrostatic interactions such as dipolar interactions for ligand binding [33]. In addition, analysis of several potent factor Xa inhibitors [33, 42, 43] indicate basic amine residues to fill this binding pocket which, in protonated form, are implicated in efficient cation-π interactions involving residues Tyr 99, Phe 174, and Trp 216 as shown in Fig. (**5**).

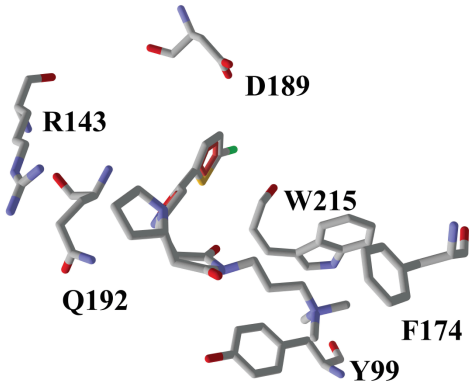

Fig. (5). Residues involved in intermolecular electrostatic interactions for factor Xa.

We built a data set with the ligands identified in the 82 complex structures of the Xa data set as actives and added 738 decoy ligands randomly selected from DUD-E [13] database to generate a data set with actives and decoys. This data set was used to run a small VS using MVD (docking protocol 23). Table **6** shows the statistical analysis of VS results. If we consider the previous results, we could say that polynomial equation 110 shows a better overall performance when compared with original scoring functions shown in Table **4**. Fig. (**6**) shows ROC curve for polynomial equation 110. Furthermore, equation 110 shows a better performance compared to the previously published docking results (EF1 = 14.6 and EF20 = 3.8) [44].

4. CONCLUSION

The big data generated during protein-ligand docking simulations lacked an integrated computational tool for statistical analysis docking results, influence of structural parameters on docking results, and scoring function performance [44]. Here we described SAnDReS, an integrated computational environment for statistical analysis of docking results and development of scoring functions. SAnDReS was written in Python 3 using scientific computing libraries. Application of SAnDReS to analyze docking results for structures in a docking benchmark was able to generate high-quality plots for docking results, which

Table 6. Statistical analysis of virtual screening results.

Scores	AUC (%)	EF1	EF2	EF5	EF10	EF20
MolDock	86.475	0.000	0.000	24.167	10.500	4.909
Re-rank	80.622	0.000	0.000	35.556	11.026	5.185
Interaction	88.539	0.000	70.000	58.333	20.370	5.769
Co-factor	37.646	0.000	0.000	3.667	8.222	3.226
Protein	88.739	0.000	0.000	58.333	20.370	5.769
Internal	57.170	1.429	0.667	0.789	1.081	1.156
Electro	55.076	16.667	6.000	12.778	18.276	5.327
Electro Long	58.650	6.000	6.000	3.667	4.386	5.619
H-bond	36.409	0.000	0.667	0.250	0.649	0.649
LE1	74.012	1.429	4.545	4.138	3.443	3.016
LE3	68.126	0.000	0.000	2.424	3.443	2.519
Polscore #38	67.762	0.000	150.0	14.118	6.735	2.813
Polscore #506	88.294	0.000	70.0	58.333	20.370	5.769
Polscore #442	42.792	0.000	0.000	0.000	0.000	0.123
Polscore #510	13.937	0.000	0.000	0.000	0.000	0.000
Polscore #110	85.599	16.667	43.333	21.538	9.070	4.909
Polscore #126	85.683	16.667	43.333	21.538	9.070	4.909
Polscore #422	13.905	0.000	0.000	0.000	0.000	0.000
Polscore #166	85.662	0.000	0.000	21.538	10.000	4.775
Polscore #294	68.592	0.000	150.0	15.625	6.735	2.913
Polscore #478	13.950	0.000	0.000	0.000	0.000	0.000

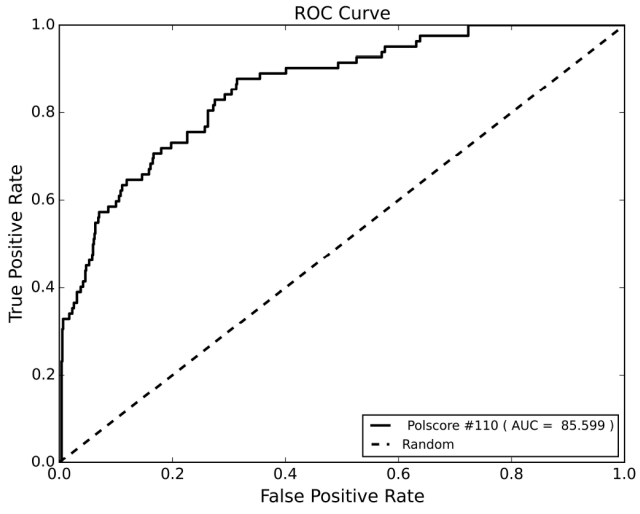

Fig. (6). Receiver operating characteristic curve for Xa data set.

facilitates analysis of docking results. In addition, SAnDReS was able to identify the correlation between structural parameters and docking results. In this study, we identified the correlation of unexpected structural parameters such crystal solvent content with docking RMSD. In addition, SAnDReS is able to build new polynomial scoring functions to predict binding affinity with better performance than well-established scoring functions such the ones available in Autodock, Autodock Vina, Gold, Glide, and Molegro Virtual Docker. SAnDReS is a free software that can be used to analyze docking results, not only for the coagulation factor Xa described here, but also for any ensemble of protein structures. Due to the importance of docking results for the initial stages of drug discovery, we believe that SAnDReS is a computational tool with potential of improve accuracy in docking projects. Being used in the analysis of the docking results and/or employed in generation of new scoring functions to predict binding affinity tailored to the biological system under study.

ABBREVIATIONS

AUC = Area under curve

AU = Arbitrary unit

CSAR = Community-Structure Active Resource

CSV = Comma separated values

DUD-E = Directory of useful decoys, enhanced

EC = Enzyme classification

EC_{50} = Half maximal effective concentration

EF = Enrichment factor

FPR = False positive rate

K_d	=	Dissociation constant
K_i	=	Inhibition constant
IC_{50}	=	Half-maximal inhibitory concentration
LE	=	Ligand Efficiency
MVD	=	Molegro virtual docker
PDB	=	Protein Data Bank
p-value1	=	P-value for Spearman's rank correlation coefficient (ρ)
p-value2	=	P-value for Pearson's correlation coefficient (R)
RMSD	=	Root-mean-square deviation
ROC	=	Receiver operating characteristics
SAnDReS	=	Statistical analysis of Docking results and Scoring functions
SDF	=	Structure data format
TPR	=	True positive rate
VS	=	Virtual screening

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank Prof. Peter Rose, Ph.D. (Site Head, RCSB Protein Data Bank West (http://www.rcsb.org), San Diego Supercomputer Center (http://bioinformatics.sdsc.edu) University of California, San Diego) for valuable discussion about Protein Data Bank functioning. This work was supported by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil) (Process Numbers: 472590/2012-0 and 308883/2014-4). NMBL and MBA acknowledge support from CAPES. MMX, NLC, and GSH acknowledge support from CNPq. VOP acknowledges support from PUCRS/BPA fellowship. WFA is senior researcher for CNPq (Brazil).

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publishers Web site along with the published article.

REFERENCES

[1] Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.,* **2000**, *28*(1), 235-242.

[2] Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J.D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **2002**, *58*(Pt 6 No 1), 899-907.

[3] Westbrook, J.; Feng, Z.; Chen, L.; Yang, H.; Berman, H.M. The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **2003**, *31*(1), 489-491.

[4] Kitano, H. Systems biology: a brief overview. Science, **2002**, *295*(5560) 1662-1664.

[5] Forli, S.; Huey, R.; Pique, M.E.; Sanner, M.F.; Goodsell, D.S.; Olson, A.J. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc*., **2016**, *11*(5), 905-919.

[6] Azevedo, L.S.; Moraes, F.P.; Xavier, M.M.; Pantoja, E.O.; Villavicencio, B.; Finck, J.A.; Proenca, A.M.; Rocha, K.B.; de Azevedo, W.F. Recent Progress of Molecular Docking Simulations Applied to Development of Drugs. *Curr. Bioinform*., **2012**, *7*(4), 352-365.

[7] Heberlé, G.; de Azevedo, W.F Jr. Bio-inspired algorithms applied to molecular docking simulations. *Curr. Med. Chem*., **2011**, *18*(9), 1339-1352.

[8] Sousa, S.F.; Ribeiro, A.J.; Coimbra, J.T.; Neves, R.P.; Martins, S.A.; Moorthy, N.S; Fernandes, P.A.; Ramos, M.J. Protein-ligand docking in the new millennium--a retrospective of 10 years in the field. *Curr. Med. Chem.*, **2013**, *20*(18), 2296-2314.

[9] Grinter, S.Z.; Zou, X. Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design. *Molecules*, **2014**, *19*(7), 10150-10176.

[10] Böhm, 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*(3): 243-256.

[11] Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; Shaw, D.E.; Francis, P.; Shenkin, P.S.; Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.*, **2004**, *47*(7), 1739-1749.

[12] Zar, J.H. Significance Testing of the Spearman Rank Correlation Coefficient. *J. Am. Stat. Assoc*., **1972**, *67*(339), 578-580.

[13] Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem*., **2012**, *55*(14) 6582-6594.

[14] Diederichs, K.; Karplus, P.A. Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nat. Struct. Biol.,* **1997**, *4*(4), 269-275.

[15] de Ávila, M.B.; de Azevedo, W.F. Data Mining of Docking Results. Application to 3-Dehydroquinate Dehydratase. *Curr. Bioinform*., **2014**, *9*(4), 361-379.

[16] Hu, L.; Benson, M.L.; Smith, R.D.; Lerner, M.G.; Carlson, H.A. Binding MOAD (Mother Of All Databases). *Proteins: Struct. Funct. Genet*., **2005**, *60*(3):333-340.

[17] Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R.N.; Gilson, M.K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res*., **2007**, *35*(Database issue), D198-201.

[18] Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem*., **2004**, *47*(12), 2977-2980.

[19] Ballante, F.; Marshall, G.R. An Automated Strategy for Binding-Pose Selection and Docking Assessment in Structure-Based Drug Design. *J. Chem. Inf. Model*., **2016**, *56*(1), 54-72.

[20] Thomsen, R.; Christensen, M.H. MolDock: a new technique for high-accuracy molecular docking. *J. Med. Chem*., **2006**, *49*(11), 3315-3321.

[21] Korb, O.; Stutzle, T.; Exner, T.E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.*, **2009**, *49*(1), 84-96.

[22] De Azevedo, W.F. Jr. MolDock applied to structure-based virtual screening. *Curr. Drug Targets*, **2010**, *11*(3), 327-334.

[23] Filgueira de Azevedo, W. Jr.; dos Santos, G.C.; dos Santos, D.M.; Olivieri, J.R.; Canduri, F.; Silva, R.G.; Basso, L.A.; Renard, G.; da Fonseca, I.O.; Mendes, M.A.; Palma, M.S.; Santos, D.S. Docking and small angle X-ray scattering studies of purine nucleoside phosphorylase. *Biochem. Biophys. Res. Commun*., **2003**, *309*(3), 923-928.

[24] Vieth, M.; Hirst, J.D.; Kolinski, A.; Brooks III, C.L. Assessing Energy Functions for Flexible Docking. *J. Comp. Chem*., **1998**, *19*(14), 1612-1622.

[25] Sheldrick, G.M.; Schneider, T.R. SHELXL: high-resolution refinement. *Methods Enzymol*., **1997**, *277*, 319-343.

[26] Blanc, E.; Roversi, P.; Vonrhein, C.; Flensburg, C.; Lea, S.M.; Bricogne, G. Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr. D Biol. Crystallogr*., **2004**, *60*(Pt 12 Pt 1), 2210-2221.

[27] Adams, P.D.; Afonine, P.V.; Bunkóczi, G.; Chen, V.B.; Davis, I.W.; Echols, N.; Headd, J.J.; Hung, L.W.; Kapral, G.J.; Grosse-Kunstleve, R.W; McCoy, A.J.; Moriarty, N.W.; Oeffner, R.; Read, R.J.; Richardson, D.C.; Richardson, J.S.; Terwilliger, T.C.; Zwart, P.H. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr*., **2010**, *66*(Pt2), 213-221.

[28] Matthews, B.W. Solvent content of protein crystals. *J. Mol. Biol*., **1968**, *33*(2), 491-497.

[29] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E.. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res*., **2011**, *12*, 2825-2830.

[30] De Azevedo, W.F. Jr.; Dias, R. Evaluation of ligand-binding affinity using polynomial empirical scoring functions. *Bioorg. Med. Chem*., **2008**, *16*(20), 9378-9382.

[31] Cichero. E.; Cesarini, S.; Mosti, L.; Fossa, P. CoMFA and CoMSIA analyses on 1,2,3,4-tetrahydropyrrolo[3,4-b]indole and benzimidazole derivatives as selective CB2 receptor agonists. *J. Mol. Model*., **2010**, *16*(9) 1481-1498.

[32] Brooijmans, N.; Kuntz, I.D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomolec. Struct*., **2003**, *32*, 335-373.

[33] Salonen, L,M.; Bucher, C.; Banner, D.W.; Haap, W.; Mary, J.L.; Benz, J.; Kuster, O.; Seiler, P.; Schweizer, W.B.; Diederich, F. Cation-pi interactions at the active site of factor Xa: dramatic enhancement upon stepwise N-alkylation of ammonium ions. *Angew. Chem.-Int. Edit*., **2009**, *48*(4) 811-814.

[34] Karplus, P.A.; Diederichs, K. Linking crystallographic model and data quality. *Science*, **2012**, *336*(6084), 1030-1033.

[35] Diederichs, K.; Karplus, P.A. Better models by discarding data? *Acta Crystallogr. D Biol. Crystallogr*., **2013**, *69*(Pt 7), 1215-1222.

[36] Karplus, P.A.; Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol*., **2015**, *34*, 60-68.

[37] De Azevedo, W.F. Jr.; Canduri, F.; Basso, L.A.; Palma, M.S.; Santos, D.S. Determining the structural basis for specificity of ligands using crystallographic screening. *Cell Biochem. Biophys*., **2006**, *44*(3), 405-411.

[38] Martiny, V.Y.; Martz, F.; Selwa, E.; Iorga, B.I. Blind Pose Prediction, Scoring, and Affinity Ranking of the CSAR 2014 Data set. *J. Chem. Inf. Model*., **2016**, *56*(6), 996-1003.

[39] Verdonk, M.L.; Cole, J.C.; Hartshorn, M.J.; Murray, C.W.; Taylor, R.D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct. Funct. Genet*., **2003**, *52*(4), 609-623.

[40] Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated Docking With Selective Receptor Flexibility. *J. Comput. Chem*., **2009**, *30*(16), 2785-2791.

[41] Trott, O,. Olson, A.J. AutoDock Vina: Improving the Speed and Accuracy of Docking With a New Scoring Function, Efficient Optimization, and Multithreading. *J. Comput. Chem*., **2010**, *31*(2), 455-461.

[42] Ma, J.C.; Dougherty, D.A. The cation-pi interaction. *Chem. Rev*., **1997**, *97*(5), 1303-1324.

[43] Zacharias, N.; Dougherty, D.A. Cation-pi interactions in ligand recognition and catalysis. *Trends Pharmacol. Sci*., **2002**, *23*(6), 281-287.

[44] Huang, N.; Shoichet, B.K.; Irwin, J.J. Benchmarking sets for molecular docking. *J. Med. Chem*., **2006**, *49*(23), 6789-6801.