Biophysical Chemistry 240 (2018) 63–69

Contents lists available at ScienceDirect

Biophysical Chemistry

journal homepage: www.elsevier.com/locate/biophyschem

Research Article

Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes

Gabriela Bitencourt-Ferreira[a], Walter Filgueira de Azevedo[a,b,*]

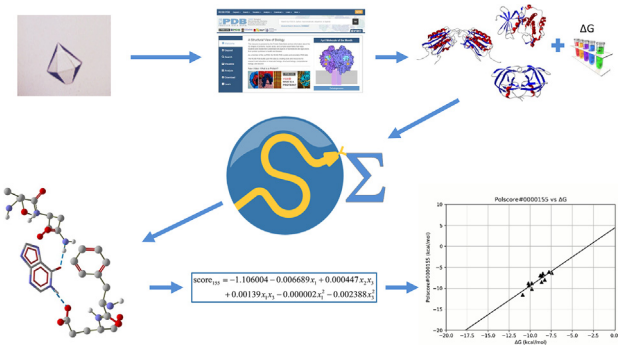
[a] Laboratory of Computational Systems Biology, School of Sciences, Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre, RS 90619-900, Brazil

[b] Graduate Program in Cellular and Molecular Biology, The Pontifical Catholic University of Rio Grande do Sul (PUCRS), Av. Ipiranga, 6681, Porto Alegre, RS 90619-900, Brazil

HIGHLIGHTS

- Development of a machine-learning model to predict free energy of binding for protein-ligand complexes;
- The use of a dataset composed of 48 high-resolution crystallographic structures to be used to build a new scoring function;
- Improved predictive power of the machine learning model to predict ΔG, when compared with classical scoring functions.

GRAPHICAL ABSTRACT



ABSTRACT

The possibility of using the atomic coordinates of protein-ligand complexes to assess binding affinity has a beneficial impact in the early stages of drug development and design. From the computational view, the creation of reliable scoring functions is still an open problem in the simulation of biological systems, and the development of a new generation machine-learning model is an active research field. In this work, we propose a novel scoring function to predict Gibbs free energy of binding (ΔG) based on the crystallographic structure of complexes involving a protein and an active ligand. We made use of the energy terms available the AutoDock Vina scoring function and trained a novel function using the machine learning methods available in the program SAnDReS. We used a training set composed exclusively of high-resolution crystallographic structures for which the ΔG data was available. We describe here the methodology to develop a machine-learning model to predict binding affinity using the program SAnDReS. Statistical analysis of our machine-learning model indicated a superior performance when compared to the MolDock, Plants, AutoDock 4, and AutoDock Vina scoring functions. We expect that this new machine-learning model could improve drug design and development through the application of a reliable scoring function in the analysis virtual screening simulations.

* Corresponding author at: School of Sciences, The Pontifical Catholic University of Rio Grande do Sul, PUCRS, Av. Ipiranga, 6681, Porto Alegre, RS 90619-900, Brazil.
E-mail address: walter@azevedolab.net (W.F. de Azevedo).

https://doi.org/10.1016/j.bpc.2018.05.010
Received 22 April 2018; Received in revised form 28 May 2018; Accepted 30 May 2018
Available online 07 June 2018
0301-4622/

1. Introduction

The application of computational methods to predict ligand-binding affinity based on the atomic coordinates of a complex involving a protein and a small-molecule binder is an open problem in computational chemistry and structural bioinformatics [1, 2]. Through quantum mechanical methods, it is possible to descript the protein-ligand system, which although computational feasible require the use of high-performance computing if we focus on datasets with thousands of complexes involving protein and ligands [3, 4]. On the other hand, molecular dynamics simulations, even being a classical approach, can generate reliable methodologies to determine the binding affinity [5]. Nevertheless, the computational cost of such calculations is still computational demanding and thus time-consuming. One alternative is combinational of standard scoring functions with supervised machine learning techniques [6, 7].

Application of supervised machine learning techniques available in the scientific libraries such as scikit-learn [8] opens the possibility to explore the scoring function space [9]. This mathematical space is a scoring function set that connects the protein sequence space [10] with the chemical space [11]. Through the application of machine learning approaches, we can find an adequate scoring function (element of the scoring function space) that predicts the binding affinity for the biological system of interest. This biological system could be a specific protein with an abundance of structural and binding affinity data or a dataset comprised of structures of several protein families for which experimental binding affinity data is available [12–17]. Here we adopt the second approach. We applied machine-learning methods to calibrate an AutoDock Vina-based scoring function [18] to predict the Gibbs free energy of binding (ΔG). To train our machine-learning model, we used high-resolution crystallographic structures for which experimental binding affinity data were known. Using this approach, we expect to have a reliable dataset of structures and binding affinity data, instead of relying on docked structures for the protein-ligand complexes. All information used to generate machine-learning models was experimental determined: complex structures determined using X-ray diffraction crystallography and experimental binding affinity data obtained through isothermal titration calorimetry [19]. Our machine-learning model was compared with traditional scoring functions such as MolDock Score, Plants Score [20], AutoDock 4 scoring function [21], and AutoDock Vina scoring function [18]. The predictive power of the machine-learning model was superior to the standard scoring functions. The potential of this approach to virtual screening and drug design is described here.

2. Methods

2.1. ΔG dataset

The SAnDReS program [22] was used to download the structures and related binding information from the Protein Data Bank (PDB) [23] to construct a dataset. PDB gathers experimental binding affinity data from three other databases: MOAD (Mother Of All Databases) [24], BindingDB [25], and PDBbind [26]. SAnDReS source code is available from GitHub (https://github.com/azevedolab/sandres). Also, we provide SAnDReS installers for Linux and Windows in the following link: https://drive.google.com/drive/folders/1GXDOTByRUyo6EszY5UJ2aXLtA1uysnTz.

We used a dataset composed of crystallographic structures refined at high-resolution (< 1.5 Å) for which ΔG information is available for the active ligands. SAnDReS was also employed to carry out the filtering of the data to eliminate repeated ligands from the initial dataset. After filtering, we ended up with 48 structures (search carried out on June 22th, 2017). This dataset will be referred as ΔG dataset from now on. Table 1 shows the PDB access codes for all structures in the ΔG dataset. Details about ligand identification and experimental ΔG for all structures in the ΔG dataset are available in the supplementary material 1.

Table 1
PDB access codes for ΔG dataset.

PDB access codes
1A9T, 1AJ6, 1GFW, 1HXD, 1KZK, 1SG0, 1 T64, 1US0, 1YHS, 1ZND, 1ZNG, 1ZNH, 1ZNK, 2AVS, 2BIK, 2BYA, 2C3I, 2DM5, 2FZD, 2G1O, 2G1R, 2G1S, 2I4Q, 2IKH, 2IKO, 2IKU, 2IL2, 2NMZ, 2O3P, 2O9A, 2PDK, 2PZN, 2Q6B, 2QX8, 2UXI, 2UXP, 2UXU, 3AKM, 3CCT, 3CCW, 3CCZ, 3K8Q, 3M4H, 4LCE, 4QA0, 4QOG, 4RXD, 5A14

2.2. Evaluation of binding affinity

In this work, we used the scoring functions available from the programs AutoDock Vina [18], AutoDock4 [21] and Molegro Virtual Docker (MVD) [20] to evaluate binding affinity. This evaluation was performed using the crystallographic position of the active ligand of each structure in the ΔG dataset. As active ligand, we mean the ligand for which ΔG data is available.

2.2.1. AutoDock Vina scoring function

The program AutoDock Vina uses a scoring function based on a combination of methods, which employs knowledge-based potentials and empirical scoring function to estimate the ligand-receptor affinity. Here we briefly describe the scoring function of AutoDock Vina, for a more detailed description, please see [18]. Part of the function of AutoDock Vina works with the following equation,

$$c = c_{\text{inter}} + c_{\text{intra}} = \sum_{i<j} h_{t_i t_j}(d_{ij}) \tag{1}$$

where i is each atom of the biological system that is assigned a type represented by the term t_i. The second part of the equation is the sum of the intermolecular (c_{inter}) and intramolecular (c_{intra}) contributions. The component $h_{t_i t_j}$is a weighted sum of steric interactions (Eqs. (3) to (5)) similar for all the atoms, and the term $d_{ij}$is the distance surface which we define by the following equation.

$$d_{ij} = r_{ij} - R_{t_i} - R_{t_j} \tag{2}$$

In the above equation,r_{ij} represents the interatomic distance and R_t is the van der Waals radius for an atom of type t.

Each atom pair interacts through a steric interaction. The AutoDock Vina program makes use of three terms to access this steric interaction. These three terms are defined as follows.

$$\text{gauss}_1(d) = e^{-(d/0.5\ \text{Å})^2} \tag{3}$$

$$\text{gauss}_2(d) = e^{-((d-3\ \text{Å})2\text{Å})^2} \tag{4}$$

$$\text{repulsion}(d) = \begin{cases} d^2, \text{ if } d < 0 \\ 0, \text{ if } d \geq 0 \end{cases} \tag{5}$$

The hydrophobic and hydrogen bond (Hbond) terms are piecewise linear functions, and they can increase the attraction when modifying the steric interaction. The Eq. (6) is included in the calculation of the steric interaction when both of the atoms in the pair are hydrophobic. We apply the Eq. (7) when the atom pair is composed of a hydrogen bond donor atom and a hydrogen bond acceptor atom [27]. We calculate these two terms as follows.

$$\text{hydrophobic}\,(d) = \begin{cases} 1, \text{ if } d < 0.5\,\text{Å} \\ \text{linearly interpolated if } 0.5\,\text{Å} < d > 1.5\,\text{Å} \\ 0, \text{ if } d > 1.5\,\text{Å} \end{cases} \tag{6}$$