Applied Soft Computing 20 (2014) 119–126

Contents lists available at ScienceDirect

Applied Soft Computing

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

Improving drug discovery using hybrid softcomputing methods

Horacio Pérez-Sánchez [a,*], Gaspar Cano [b], José García-Rodríguez [b]

[a] Computer Science Department, Catholic University of Murcia (UCAM), E30107 Murcia, Spain
[b] Computing Technology Department, University of Alicante, Ap. 99, E03080 Alicante, Spain

ARTICLE INFO

Article history:
Received 29 March 2013
Received in revised form 29 October 2013
Accepted 30 October 2013
Available online 28 November 2013

Keywords:
Neural networks
Support vector machines
Clinical research
Drug discovery
Virtual screening
Parallel computing

ABSTRACT

Virtual screening (VS) methods can considerably aid clinical research, predicting how ligands interact with drug targets. Most VS methods suppose a unique binding site for the target, but it has been demonstrated that diverse ligands interact with unrelated parts of the target and many VS methods do not take into account this relevant fact. This problem is circumvented by a novel VS methodology named BINDSURF that scans the whole protein surface in order to find new hotspots, where ligands might potentially interact with, and which is implemented in last generation massively parallel GPU hardware, allowing fast processing of large ligand databases. BINDSURF can thus be used in drug discovery, drug design, drug repurposing and therefore helps considerably in clinical research. However, the accuracy of most VS methods and concretely BINDSURF is constrained by limitations in the scoring function that describes biomolecular interactions, and even nowadays these uncertainties are not completely understood. In order to improve accuracy of the scoring functions used in BINDSURF we propose a hybrid novel approach where neural networks (NNET) and support vector machines (SVM) methods are trained with databases of known active (drugs) and inactive compounds, being this information exploited afterwards to improve BINDSURF VS predictions.

1. Introduction

In clinical research, it is crucial to determine the safety and effectiveness of current drugs and to accelerate findings in basic research (discovery of new leads and active compounds) into meaningful health outcomes. Both objectives need to process the large data set of protein structures available in biological databases such as PDB [1] and also derived from genomic data using techniques as homology modeling [2]. Screenings in lab and compound optimization are expensive and slow methods [3], but bioinformatics can vastly help clinical research for the mentioned purposes by providing prediction of the toxicity of drugs and activity in non-tested targets, and by evolving discovered active compounds into drugs for the clinical trials.

This can be achieved thanks to the availability of bioinformatics tools and Virtual Screening (VS) methods that allow testing all required hypothesis before clinical trials. Nevertheless current Virtual Screening (VS) methods, such as docking, fail to make good toxicity and activity predictions since they are constrained by the access to computational resources; even the nowadays fastest VS methods cannot process large biological databases in a reasonable time-frame. Therefore, these constraints impose serious limitations in many areas of translational research.

The use of last generation massively parallel hardware architectures such as Graphics Processing Units (GPUs) can tremendously overcome this problem. The GPU has become increasingly popular in the high performance computing arena, by combining impressive computational power with the demanding requirements of real-time graphics and the lucrative mass-market of the gaming industry [4]. Scientists have exploited this power in arguably every computational domain, and the GPU has emerged as a key resource in applications where parallelism is the common denominator [5]. To maintain this momentum, new hardware features have been progressively added by NVIDIA to their range of GPUs, with the Fermi architecture [6] being the most recent milestone in this path. Therefore, GPUs are well suited to overcome the lack of computational resources in VS methods, accelerating the required calculations and allowing the introduction of improvements in the biophysical models not a affordable in the past [7]. We have previously worked in this direction, showing how VS methods can benefit from the use of GPUs [8,9,10]. Moreover, another important lack of VS methods is that they usually take the assumption that the binding site derived from a single crystal structure will be the same for different ligands, while it has been shown that this does not always happen [11], and thus it is crucial to avoid this very basic supposition. In this work, we present a novel VS methodology called BINDSURF [12] which takes advantage of massively parallel

* Corresponding author. Tel.: +34 653067844.
E-mail addresses: hperez@ucam.edu (H. Pérez-Sánchez), gcano@dtic.ua.es (G. Cano), jgarcia@dtic.ua.es (J. García-Rodríguez).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2013.10.033

and high arithmetic intensity of GPUs to speed-up the required calculations in low cost and consumption desktop machines, providing new and useful information about targets and thus improving key toxicity and activity predictions. In BINDSURF a large ligand database is screened against the target protein over its whole surface simultaneously. Afterwards, information obtained about novel potential protein hotspots is used to perform more detailed calculations using particular VS method, but just for a reduced and selected set of ligands.

Other authors have also performed VS studies over whole protein surfaces [13] using different approaches and screening small ligand databases, but as far as we know, none of them have been implemented on GPUs, while BINDSURF has been designed from scratch taken into account the GPU architecture.

However, the accuracy of most VS methods is constrained by limitations in the scoring function that describes biomolecular interactions, and even nowadays these uncertainties are not completely understood. In order to solve this problem we propose a novel hybrid approach where softcomputing methods that includes neural networks (NNET) and support vector machines (SVM) are trained with known active (drugs) and inactive compounds and are later used to improve VS predictions.

The rest of the paper is organized as follows. Section 2 describes the methodology including VS using BINDSURF, NNET and SVM techniques, and molecular properties used in this study. Section 3 presents the experiments carried out to refine the BINDSURF method with the previously mentioned techniques while Section 4 discusses the results obtained. In Section 5 we present our main conclusions and further work.

2. Metholodogy

In this section we describe the methodologies we used for improving the prediction of protein–ligand affinity; (a) the Virtual Screening method BINDSURF, and (b) two different softcomputing techniques are studied; neural networks (NN) and support vector machines (SVM) trained with different molecular properties calculated for known active and inactive compounds selected from standard VS benchmarks. In Fig. 1 a flowchart of the methodology is shown; once a protein target (component A) and a compound database (component B) have been chosen, compounds for which no information about affinity against protein target is available (component C) are docked using BINDSURF (component D) and estimated affinities (component E) and 3D poses (component F) are obtained. Using the methods described in this section, we start selecting compounds from the database for which affinity data is available (component G), so that we can calculate relevant descriptors (component H) and train adequately neural networks and support vector machines (component I) so that affinities obtained in component E are post-processed and we finally obtain improved values for the affinities (component J).

2.1. Virtual Screening with BINDSURF

The main idea underlying our VS method BINDSURF is the protein surface screening method, implemented in parallel on GPUs. Essentially, VS methods screen a large database of molecules in order to find which one fit some established criteria [14]. In the case of the discovery of new leads, compound optimization, toxicity evaluation and additional stages of the drug discovery process, we screen a large compound database to find a small molecule which interacts in a desired way with one or many different receptors. Among the many available VS methods for this purpose we decided to use protein–ligand docking [15,16]. These methods try to obtain rapid and accurate predictions of the 3D conformation a ligand adopts when it interacts with a given protein target, and also the strength of this union, in terms of its scoring function value. Docking simulations are typically carried out using a very concrete part of the protein surface in methods like Autodock [17], Glide [18] and DOCK [19], to name a few. This region is commonly derived from the position of a particular ligand in the crystal structure, or from the crystal structure of the protein without any ligand. The former can be performed when the protein is co-crystallized with the ligand, but it might happen that no crystal structure of this ligand–protein pair is at disposal. Nevertheless, the main problem is to take the assumption, once the binding site is specified, that many different ligands will interact with the protein in the same region, discarding completely the other areas of the protein.

Given this problem we propose to overcome it by dividing the whole protein surface into defined regions. Next, docking simulations for each ligand are performed simultaneously in all the specified protein spots. Following this approach, new hotspots might be found after the examination of the distribution of scoring function values over the entire protein surface. This information could lead to the discovery of novel binding sites. If we compare this approach with a typical docking simulation performed only in a region of the surface, the main drawback of this approach lies on its increased computational cost. We decided to pursue in this direction and show how this limitation can be overcome thanks to GPU hardware and new algorithmic designs.

In essence, in a docking simulation we calculate the ligand–protein interaction energy for a given starting configuration of the system, which is represented by a scoring function [20]. In BINDSURF the scoring function calculates electrostatic (ES), Van der Waals (VDW) and hydrogen bond (HBOND) terms.

Furthermore, in docking methods it is normally assumed [14] that the minima of the scoring function, among all ligand–protein conformations, will accurately represent the conformation the system adopts when the ligand binds to the protein. Thus, when the simulation starts, we try to minimize the value of the scoring function by continuously performing random or predefined perturbations of the system, calculating for each step the new value of the scoring function, and accepting it or not following different approaches like the Monte Carlo minimization method [21] or others. Simulations were always carried out with a total of 500 Monte

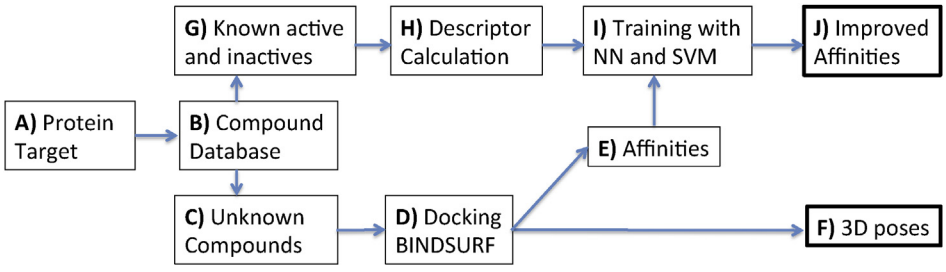

Fig. 1. Flowchart of the methodology used for improving the predictive capability of BINDSURF.