Improving Drug Discovery Using a Neural Networks Based Parallel Scoring Function

Horacio Pérez-Sánchez*, Ginés D. Guerrero, José M. García, Jorge Peña, José M. Cecilia, Gaspar Cano, Sergio Orts-Escolano, and José García-Rodríguez

***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 to find new hotspots, where ligands might potentially interact with, and which is implemented in massively parallel Graphics Processing Units, 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 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 approach where neural networks are trained with databases of known active (drugs) and inactive compounds, and later used to improve VS predictions.**

I. Introduction

In clinical research, it is crucial to determine the safety and effectiveness of current drugs and to accelerate 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, 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.

Manuscript received February 20, 2013.

Horacio Pérez-Sánchez, Jorge Peña and José M. Cecilia are with the Computer Science Department, Catholic University of Murcia (UCAM) 30107 Murcia, Spain e-mail: {horacio,jpena,jmcecilia}@ucam.edu.

Ginés D. Guerrero, José M. García, are with Computer Engineering Department, School of Computer Science, University of Murcia, Spain; e-mail: {horacio, gines.guerrero, jmgarcia}@ditec.um.es).

Gaspar Cano, Sergio Orts-Escolano, and José García-Rodríguez are with the Department of Computers Techonology. University of Alicante, Spain. email: {gcano, sorts, jgarcia}@dtic.ua.es

*Corresponding author

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 massively parallel and throughput-oriented hardware architectures such as Graphics Processing Units (GPUs) can tremendously overcome this problem. The GPU has become increasingly popular in the High Performance Computing (HPC) arena, by combining impressive computational power with the demanding requirements of real-time graphics and the lucrative mass-market of the gaming industry [3]. 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 [4]. To maintain this momentum, new hardware features have been progressively added by NVIDIA to their range of GPUs, with the Kepler architecture [5] 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 affordable in the past [6]. We have previously worked in this direction, showing how VS methods can benefit from the use of GPUs [7-9]. 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 [10], and thus it is crucial to avoid this very basic supposition. In this work, we present a novel VS methodology called BINDSURF, which takes advantage of massively parallel 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 any particular VS method, but just for a reduced and selected set of ligands.

Other authors have also performed VS studies over whole protein surfaces [11] using different approaches and screening small ligand databases, but as far as we know, none of them have been implemented on GPUs and used in the same fashion as BINDSURF.

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 this paper we tackle with this problem, proposing a novel approach where neural networks are trained with databases of known active (drugs) and inactive compounds and later used to improve VS predictions through BINDSURF.

The rest of the paper is organized as follows. Section II briefly introduces the preliminary knowledge to better understand the rest of the article. Section III we introduce our proposal of training neural networks to improve VS predictions before concluding the paper and showing possible directions for future work.

II. Methodology

In this section we describe the methodologies we used for the prediction of protein-ligand affinity; a) the Virtual Screening method BINDSURF, and b) a neural network trained with chemical similarity data of known active and inactive compounds.

A. 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 [12]. 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 [13, 14]. 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 in a very concrete part of the protein surface in methods like Autodock [15], Glide [16] and DOCK [17], 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 [18]. 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 [12] 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 [19] or others. Simulations were always carried out with a total of 500 Monte Carlo steps. For a detailed discussion it is advisable to have a look at our previous BINSURF publication [20].

B. Neural Networks

One of the most dominant application areas of neural networks is non-linear function approximation. There are several types of feed-forward neural networks; the most widely used being multi-layer networks with sigmoidal activation functions (multi-layer perceptrons) and single layer networks with local activation functions (radial basis function networks). The good approximation capability of neural networks has been widely demonstrated by both practical applications and theoretical research. We decided to use a single-hidden-layer neural network with skip-layer connections (see Figure 1) in this study since it has been clearly demonstrated its impact on chemical applications concerning similarity calculations [21].

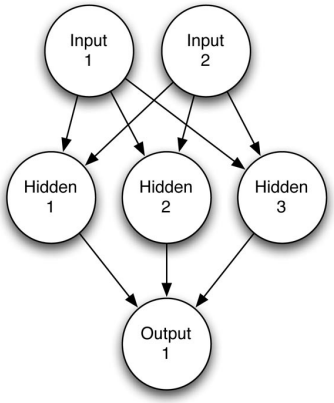

Fig. 1. Single hidden layer neural network structure.

For such purpose we used the nnet function of the R package [22]. The following default parameters described in Table 1 were used.

TABLE 1. MOST RELEVANT PARAMETERS USED FOR THE NEURAL NETWORK USED IN THE NNET FUNCTION OF THE R PACKAGE.

Parameter	*value*
WeighDecayFactor	0.001
CrossValidationFolds	5
MaxNumberIterations	2000
MaxnumberWeights	2000
TraceOptiimization	no

C. Molecular similarity

Extended-connectivity fingerprints (ECFPs), which are implemented in jCompoundMapper [23] were used as structural descriptors for training the Neural Networks. The ECFPs are a class of fingerprints for molecular characterization. Its features correspond to the presence of an exact structure (not a substructure) with limited specified attachment points.

In generating the fingerprints, the program assigns an initial code to each atom. The initial atom code is derived from the number of connections to the atom, the element type, atomic charge, and atomic mass. This corresponds to an ECFP with a neighborhood size of zero. These atom codes are then updated in an iterative manner to reflect the codes of each atoms neighbors. In the next iteration, a hashing scheme is employed to incorporate information from each atom immediate neighbors. Each atoms new code, now describes a molecular structure with a neighborhood size of one. This process is carried out for all atoms in the molecule. When the desired neighborhood size is reached, the process is complete and the set of all features is returned as the fingerprint. For the ECFPs employed in this paper, neighborhood sizes of two, four and six (ECFP 2, ECFP 4, ECFP 6) were used to generate the fingerprints. The resulting ECFPs can represent a much larger set of features than other fingerprints and contain a significant number of different structural units crucial for the molecular comparison, among the compounds.

III. RESULTS AND DISCUSSION

A. Virtual Screening with BINDSURF

We performed VS calculations with BINDSURF using standard benchmark tests, such as the Directory of Useful Decoys (DUD) [24], where VS methods check how efficient they are in differentiating ligands that are known to bind to a given target, from non-binders or decoys. Input data for each molecule of each set contains its molecular structure and whether it is active or not. After BINDSURF calculations, results for three different DUD datasets are shown in the ROC curves of Figure 2. Given the results obtained for the DUD datasets TK, MR and GPB, and characterized by the value of the area under the curve (AUC) for each ROC curve, it could be said that, on average, BINDSURF performs similarly than other docking methods reported for these datasets [25].

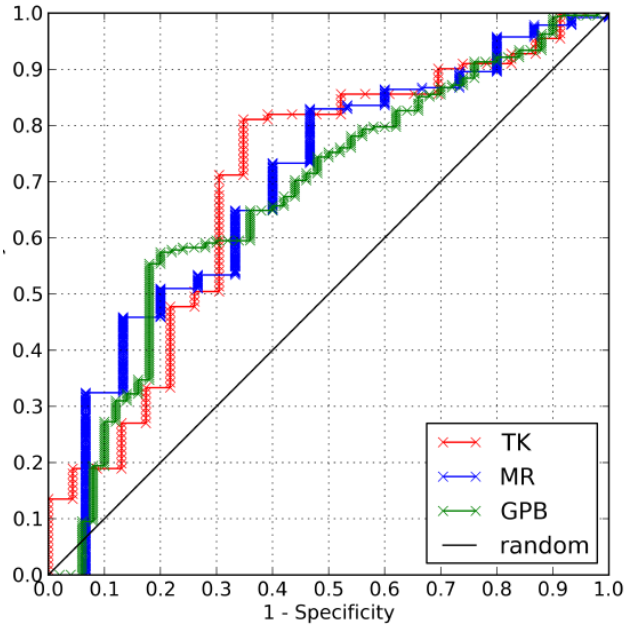

Fig. 2. ROC plots obtained using BINDSURF for the targets of the DUD data set TK (red), MR (blue) and GPB (green). Diagonal line indicates random performance. Obtained values for AUC are 0.700, 0.695 and 0.675, respectively.

Nevertheless, it is clear that there is still room for improvement in the scoring function that BINDSURF uses, and in its energy optimization method (Monte Carlo), since both affect directly to the effectiveness of the direct prediction of binding poses.

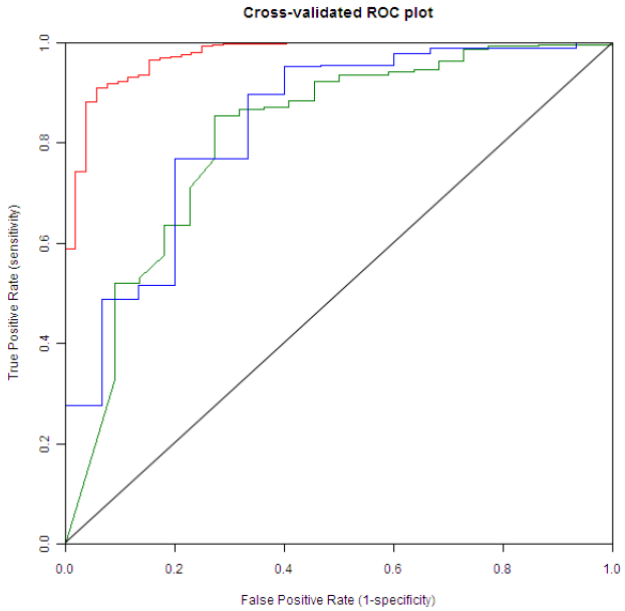

Fig. 3. ROC plots obtained using a neural network for the targets of the DUD data set TK (green), MR (blue) and GPB (red). Diagonal line indicates random performance. Obtained values for AUC are 0.801, 0.812 and 0.963, respectively.

B. Neural Networks based activity prediction based on chemical similarity

Neural networks were trained with the previously mentioned DUD datasets TK, MR and GPB and also using ECFP 2, ECFP 4, ECFP 6 fingerprints; which were calculated for each molecule as described in the methods section. Different combinations of these parameters were tested (only ECFP 2, ECFP 2 plus ECFP 4, etc) and we observed that using simultaneously the three descriptors yielded the best results in terms of AUC for the ROC curves, as can be seen in Figure 3.

If we compare these results with the previous ones obtained by BINDSURF in Figure 2 it is clear that predictive capability increases.

Consequently, and taking into account information obtained by the neural network we can post-process docking results obtained by the scoring function of BINDSURF and neglect resulting compounds that are predicted as inactive. Then we can sort them by the final affinity value predicted by the scoring function for such cases and study visually the top ones.

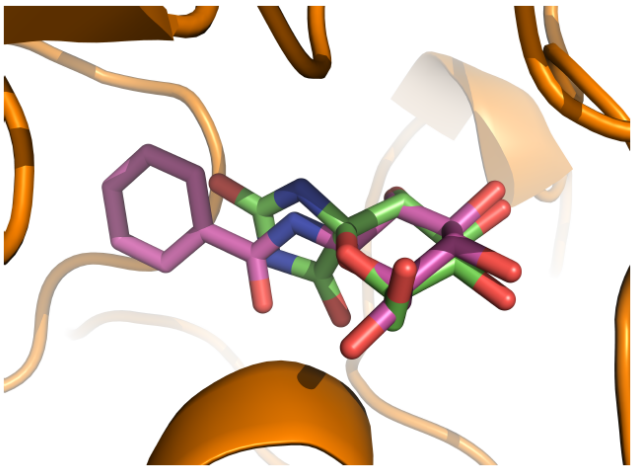

Fig. 4. Depiction of the binding mode found by BINDSURF for ligand number 17 of the DUD data set of GPB (PDB ID: 1A8I) with pink skeleton and crystallographic pose for Beta-D-Glucopyranose Spirohydantoin, with green skeleton.

As an example, we can observe in Figure 4 the good agreement in the comparison between the top predicted compound from the DUD database for GPB and the crystallographic pose. In this case main stabilizing interactions are due to a hydrogen-bond network where the intervening nitrogen and oxygen atoms of the predicted compound fall very close to the same atoms of the crystallographic pose.

IV. Conclusions

In this work we have shown how the predictive capability of the BINDSURF program can be increased using a neural network trained with ligand activity data. It must be mentioned that the neural network approach can only be used when there is data available for active and non-active compounds for a given protein.

This methodology can be used to improve drug discovery, drug design, repurposing and therefore aid considerably in clinical research. In the next steps, we want to substitute the Monte Carlo minimization algorithm for more efficient optimization alternatives, such as the Ant Colony optimization method, which we have already efficiently implemented on GPU [26] and implement also full ligand and receptor flexibility. Lastly, we are also working on improved scoring functions to include efficiently metals, aromatic interactions, and implicit solvation models.

Acknowledgements

This work has been jointly supported by the Fundación Séneca (Agencia Regional de Ciencia y Tecnología de la Región de Murcia) under grant 15290/PI/2010, by the Spanish MINECO and the European Commission FEDER funds under grants TIN2009-14475-C04 and TIN2012-31345, and by the Catholic University of Murcia (UCAM) under grant PMAFI/26/12. We also thank Nvidia Corporation for hardware donation. This work was partially supported by the computing facilities of Extremadura Research Centre for Advanced Technologies (CETA-CIEMAT), funded by the European Regional Development Fund (ERDF). CETA-CIEMAT belongs to CIEMAT and the Government of Spain. The authors also thankfully acknowledge the computer resources and the technical support provided by the Plataforma Andaluza de Bioinformática of the University of Málaga

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 28 (2000) 235-242

[2] Sanchez, R., Sali, A.: Large-Scale Protein Structure Modeling of the Saccharomyces cerevisiae Genome. Proc Natl Acad Sci USA 95(23) (November 1998) 13597-13602

[3] Garland, M., Kirk, D.B.: Understanding throughput-oriented architectures. Commun ACM 53 (November 2010) 58-66

[4] Garland, M., Le Grand, S., Nickolls, J., Anderson, J., Hardwick, J., Morton, S., Phillips, E., Zhang, Y., Volkov, V.: Parallel computing experiences with cuda. IEEE Micro 28 (July 2008) 13-27

[5] NVIDIA: Whitepaper NVIDIA's Next Generation CUDA Compute Architecture: Kepler GK110. (2012)

[6] Pérez-Sánchez, H., Wenzel, W.: Optimization methods for virtual screening on novel computational architectures. Curr Comput Aided Drug Des 7(1) (2011,44-52

[7] Guerrero, G., Pérez-Sánchez, H., Wenzel, W., Cecilia, J.M., García, J.M.: Effective parallelization of non-bonded interactions kernel for virtual screening on gpus. In: 5th International Conference on Practical Applications of Computational Biology; Bioinformatics (PACBB 2011). Volume 93. Springer Berlin / Heidelberg (2011) 63-69

[8] Sánchez-Linares, I., Pérez-Sánchez, H., Guerrero, G.D., Cecilia, J.M., García, J.M.: Accelerating multiple target drug screening on gpus. In: Proceedings of the 9th International Conference on Computational Methods in Systems Biology (CMSB' 11), New York, NY, USA, ACM (2011) 95-102

[9] Sánchez-Linares, I., Pérez-Sánchez, H., García, J.M.: Accelerating grid kernels for virtual screening on graphics processing units. In D'Hollander, E., Padua, D., eds.: Parallel Computing: Proceedings of the

International Conference ParCo 2011. Volume 22., IOS (April 2012) 413-420
[10] Brannigan, G., LeBard, D.N., H_enin, J., Eckenho_, R.G., Klein, M.L.: Multiple binding sites for the general anesthetic isourane identified in the nicotinic acetylcholine receptor transmembrane domain. Proc Natl Acad Sci USA 107(32) (August 2010) 14122-14127
[11] Hetényi, C., van der Spoel, D.: Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 11(7) (July 2002) 1729-1737
[12] Jorgensen, W.: The many roles of computation in drug discovery. Science 303(5665) (2004) 1813-1818
[13] Yuriev, E., Agostino, M., Ramsland, P.A.: Challenges and advances in computational docking: 2009 in review. J Mol Recogn 24(2) (2011) 149-164
[14] Huang, S.Y., Zou, X.: Advances and challenges in protein-ligand docking. Int J Mol Sci 11(8) (2010) 3016-3034
[15] Morris, G., Goodsell, D., Halliday, R., Huey, R., Hart, W., Belew, R., Olson, A.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19(14) (1998) 1639-1662
[16] 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., et al.: Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. J Med Chem 47(7) (2004) 1739-1749
[17] Ewing, T.J.A., Makino, S., Skillman, A.G., Kuntz, I.D.: Dock 4.0: Search strategies for automated molecular docking of exible molecule databases. J Comput-Aided Mol Des 15(5) (2001) 411-428
[18] Wang, R., Lu, Y., Fang, X., Wang, S.: An extensive test of 14 scoring functions using the PDBbind re_ned set of 800 protein-ligand complexes. J Chem Inform Comput Sci 44(6) (2004) 2114-2125
[19] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J Chem Phys 21 (1953) 1087-1092
[20] Sánchez-Linares, I., Pérez-Sánchez, H., Cecilia, J., García, J.: High-throughput parallel blind virtual screening using bindsurf. BMC Bioinformatics 13(Suppl 14) (2012) S13
[21] G., S., P., W.: Artificial neural networks for computer-based molecular design. Progress in Biophysics and Molecular Biology 70(3) (1998) 175-222
[22] Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Fourth edn. Springer, New York (2002) ISBN 0-387-95457-0.
[23] Hinselmann, G., Rosenbaum, L., Jahn, A., Fechner, N., Zell, A.: jcompoundmap-per: An open source java library and command-line tool for chemical fingerprints. Journal of Cheminformatics 3(1) (2011) 3
[24] Huang, N., Shoichet, B.K., Irwin, J.J.: Benchmarking sets for molecular docking. J Med Chem 49(23) (2006) 6789-6801
[25] Cross, J.B., Thompson, D.C., Rai, B.K., Baber, J.C., Fan, K.Y., Hu, Y., Humblet, C.: Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. J Chem Inf Model 49(6) (2009) 1455-1474
[26] Cecilia, J.M., García, J.M., Ujaldon, M., Nisbet, A., Amos, M.: Parallelization strategies for ant colony optimisation on gpus. In: In 14th Int. Workshop on Nature Inspired Distributed Computing -NIDISC11- (in conjunction with IPDPS 2011), IEEE (2011) 339-346.