A Machine Learning Approach to Enhance Scoring Performance in Docking-Based Virtual Screening Experiments: COX-1 as a Case Study

Cândida G. Silva[1,2⋆], Pedro Carreiras[1,3], Elsa S. Henriques[1], Carlos J. V. Simões[1,2], and Rui M. M. Brito[1,2]

[1] Centre for Neuroscience and Cell Biology, University of Coimbra, 3004-517 Coimbra, Portugal
[2] Chemistry Department, University of Coimbra, 3004-535 Coimbra, Portugal
[3] Department of Life Sciences, University of Coimbra, Apartado 3046, 3001-401 Coimbra, Portugal candidasilva@qui.uc.pt

Abstract. Molecular docking can be reasonably successful at reproducing X-ray poses of a ligand in the binding site of a protein. However, scoring functions are typically unsuccessful at correctly ranking ligands according to their binding affinity. Using cyclooxygenase-1 (COX-1), a particularly challenging workhorse in virtual screening (VS) we show how the use of support vector machines (SVMs), trained with the individual energy terms retrieved from docking-based VS experiments, can improve the discrimination between active and inactive compounds. Actives and inactives for COX-1 were obtained from the Directory of Useful Decoys (DUD) and docked into COX-1 with AutoDock Vina (Vina). The energy parameters of Vina's scoring function were used to train classification models with SVM-light. The results show that Vina offers acceptable pose prediction accuracy, but its scoring function performs poorly compared to our SVM classification models. The superior performance of the trained classification models highlights the potential of using non-linear machine learning methods to identify bioactive compounds through docking-based screening.

Keywords: Virtual screening, Machine learning, Docking, AutoDock Vina, SVM, COX-1

1 Introduction

Virtual screening (VS) of compound libraries has become a commonly-employed methodology in modern lead discovery [1, 2]. The goal of VS is to prioritise compounds for biological evaluation by using computational tools and information on the protein target (receptor-based VS) and/or known active ligands (ligand-based VS). Amongst the receptor-based methods used for VS, molecular docking is the most prominent [3]. In molecular docking, small compounds are docked

⋆ Corresponding author

into a particular target and ranked according to their predicted binding affinity or complementarity to the binding site. At a basic level, the process can be dissected into two main stages: conformational search and scoring. First, a search algorithm samples the degrees of freedom of the compound:protein system to include the true binding mode(s). A scoring function attempts to score the allowed interactions in each conformation and thus distinguish the correct binding mode(s) from all configurations explored [3]. The scores may assume different forms according to the chosen scoring function. For example, a Gibbs free energy difference is typically output by force field-based scoring functions. The strongest scores for each compound are then used to rank the entire library and thence select top compounds for experimental validation.

Although docking can be reasonably successful at reproducing ligand poses determined by X-ray crystallography, scoring functions are normally unsuccessful at correctly ranking compounds according to their binding affinity. Ideally, compounds that bind tightly and receive stronger docking scores should be more likely to be truly active, which is rarely the case [4, 5]. Strategies to overcome this problem have included the use of consensus scoring methods, combining different scoring functions into a single prediction, thus taking advantage of the different capabilities of the various types of scoring functions [4]. However, evidence of the success of these methods is still missing. In most scoring functions, a set of weights is assigned to the individual terms that build up the overall score. Force-field based scoring functions, in particular, traditionally assume that individual interactions contribute toward the total binding affinity in an additive fashion, thus deriving their predictions from a linear combination of individual energy terms. This method fails to consider the cooperative effects of non-covalent interactions, which only recently have been acknowledged [6, 7]. Therefore, if reasonable outcomes are to be expected, the scoring function must be trained non-linearly to derive a specific set of weights in a target-dependent manner. In recent years, machine learning methodologies have been increasingly applied to the computational identification of active compounds and have become a viable alternative to conventional approaches [8–11]. In fact, there are examples of their use to derive knowledge-based scoring functions [8, 12]. Inspired by such works, we propose the integration of support vector machines (SVM) in the context of docking-based compound screening to improve the discrimination between active and inactive compounds.

Using cyclooxygenase-1 (COX-1) as case-study, we demonstrate that the use of SVM trained with the individual energy terms retrieved from docking with AutoDock Vina can enhance the ranking of active inhibitors of COX1 over a subset of carefully selected decoys. Vina is a very fast docking program that employs a simple force field-based scoring function, rendering it a suitable choice for this work. Cyclooxygenase (COX) is a membrane-bound enzyme which exists in two isoforms: COX-1 (Fig. 1, left) is constitutively expressed in most cells and tissues; COX-2 is inductively expressed to mediate inflammation [13]. Most non-steroidal anti-inflammatory drugs (NSAIDs, Fig. 1, right) are non-selective because they inhibit both COX-1 and COX-2. Inhibition of COX-1 leads to side

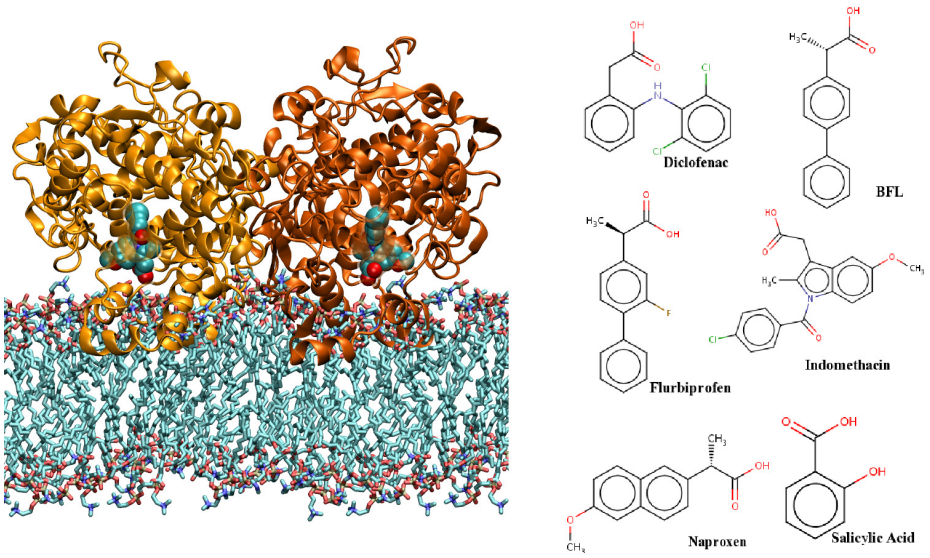

Fig. 1. Left. A ribbon representation of the ovine COX-1 dimer structure (PDB ID 1Q4G [16]) showing the binding site for 2-(1,1'-biphenyl-4-yl)propanoic acid (BFL) (blue), and the putative position of the lumenal leaflet of the ER bilayer. **Right.** Example of NSAIDs known to inhibit COX-1 activity.

effects commonly associated to NSAIDs, such as stomach ulcer. The development of selective NSAIDs, is therefore a long-sought goal and the development of models and methods to predict (and thus avoid) inhibition of COX-1 is a worthy effort.

2 Materials and Methods

Data Set. COX-1 actives and inactives were extracted from the Directory of Useful Decoys (DUD) [14]. DUD comprises data sets of active and inactive compounds for 40 diverse protein targets. For each active, 36 inactives were chosen based on the similarity of their physical properties compared to the actives while still being topologically distinct, making this a challenging test set. A set of 25 actives and 911 inactives is available for COX-1. Of the 911 inactives, 62 were excluded as these were redundant entries (the same inactive was related to multiple active compounds). Thus, a total of 25 actives and 849 inactives were considered in the analysed data set.

Molecular Docking (Pose Prediction). Each compound in the data set was docked into the structure of COX-1 using AutoDock Vina (Vina) [15]. Polar hydrogen atoms were added to the protein using Autodock Tools. In general, the docking parameters were kept to their default values. The targeted site in

the docking calculations was defined by the position of the 2-(1,1-biphenyl-4-yl)propanoic acid (BFL) molecule observed in complex with COX-1 (PDB ID: 1Q4G [16]). A docking box of 18Å×18Å×20Å and centered in point (x,y,z) = (26.6, 33.8, 201.5) was defined. The docking runs were carried out to generate a maximum of 15 poses per compound with a maximum energy difference of 10 kcal/mol.

Pose Prediction Accuracy. The performance at pose prediction was assessed through redocking (a.k.a. self-docking) experiments, by calculating the root mean spare deviation (RMSD, Eq. 1) between the predicted poses and the native pose of the co-crystallised ligand in COX-1 complexes. Six complexes were available in the PDB [17] (http://www.rcsb.org) with the following ligands: BFL (1Q4G); diclofenac (3N8Y); flurbiprofen (3RR3); indomethacin (4COX); naproxen (3NT1), and salicylic acid (1PTH; 3N8Y).

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N_{atoms}} (X_{1,i} - X_{2,i})^2 - (Y_{1,i} - Y_{2,i})^2 - (Z_{1,i} - Z_{2,i})^2}{N_{atoms}}} \quad (1)$$

SVM Classification Models. Vina output provides five different energy terms that contribute to the overall energy score: gauss1, gauss2, repulsion, hydrophobic, hydrogen bonding. SVM-light [18] was used to train a classification model where each molecule was labeled as either active (+1) or inactive (-1), based on the classification provided by DUD, and assigned with a z-normalised feature vector with Vina energy terms.

Given that training data originated from DUD data sets presents a strong bias toward negative examples (1 active: 36 inactive), which may be adequate for testing but not training purposes, a multiple-planar classification model has been proposed [11] to overcome this problem, and is summarised here. After randomly dividing the active and inactive compounds into three subsets for the 3-fold cross validation, each of the three subsets of inactives was randomly partitioned into n subsets, with n ranging from 5 to 36. Thus, for each iteration of the 3-fold cross validation, n models were trained using a different subset of inactives.

For each compound in the test set, the n models predicted a score value. A consensus vote was assigned to each compound based on the sum of the n scores. If the sum of the n scores was greater than zero, the compound was predicted to be active, otherwise the compound was predicted to be inactive. The F score (Eq. 2) was then calculated for each of the three iterations as

$$F = \frac{2 \times precision \times sensitivity}{precision + sensitivity} \quad (2)$$

where precision is the number of true positives divided by the total number of predicted positive results, and sensitivity is defined by equation 4.

The mean F score was assigned to that particular value of n. The optimal number of partitions (n) was decided based on the best mean F score value.

Performance Evaluation. The performance of Vina scoring function and the SVM classification models in discriminating between active ligands and decoys of COX-1 was assessed using two different metrics: enrichment factors (EF) and the area under the Receiver Operator Characteristic (ROC) curves [19]. While enrichment factors measure how many more active compounds are found within a defined "early recognition" fraction of the ranked list in comparison to a random selection, the area under the ROC curve (AUC ROC) provides a measure of the overall performance of the methods applied [19–21].

The enrichment factor (EF) is defined as

$$EF^{x\%} = \frac{n_a^{x\%}/N^{x\%}}{n_a/N} \tag{3}$$

where $n_a^{x\%}$ is the number of actives found at top x% of the database screened, $N^{x\%}$ is the total number of compounds screened at top x% of the database, n_a is the total number of actives in the data set, and N is the total number of compounds in the data set. The $EF^{x\%}$ metric relies on cutoffs made at various points through the ranking and so can be sensitive to small changes in ranking.

The ROC curve is a graphical plot of the sensitivity (true positive rate, Eq. 4) versus 1-specificity (false positive rate, Eq. 5) where sensitivity and specificity are defined as

$$sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$specificity = \frac{TN}{TN + FP} \tag{5}$$

where TP represents the number of correctly identified actives (true positives), TN, the number of correctly identified inactive (true negatives), FP, the number of inactive incorrectly predicted as active (false positives), and FN the number of active incorrectly predicted to be inactive (false negatives).

3 Results and Discussion

A library of 25 active compounds and 849 distinct inactives for COX-1 was retrieved from DUD and docked into COX-1 using Vina, with the aim of assessing Vina's ability to predict and score correct ligand poses. The use of support vector machines (SVMs), trained with the individual energy terms retrieved from Vina, to improve the discrimination between active and inactive compounds was then investigated.

3.1 Pose Prediction Evaluation

Vina pose prediction accuracy was analysed using the RMSD value between the predicted poses and the available experimental conformations for a collection of

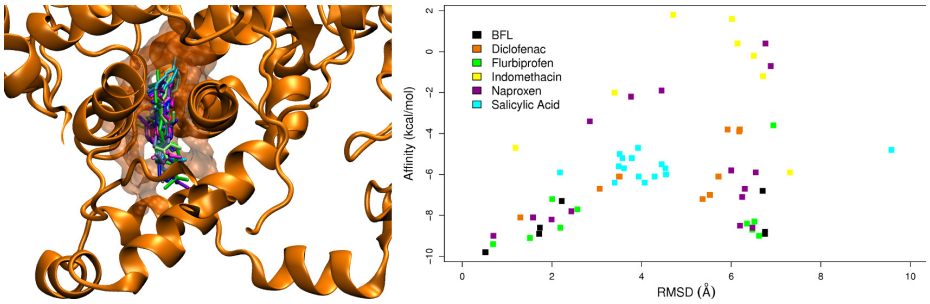

Fig. 2. Left. Best binding pose of compounds BFL, diclofenac, flurbiprofen, indomethacin, naproxen and salicylic acid in COX-1 active site as predicted by Vina. **Right.** Plot of energy *vs.* RMSD for the best ranking poses of each compound.

six different compounds in complexes with COX-1: BFL, diclofenac, flurbiprofen, indomethacin, naproxen and salicylic acid. Fig. 2 (left panel) presents the docked top ranking poses of each of these six reference compounds according to Vina's scoring function. Visual inspection of the poses suggests that Vina's native docking and scoring protocols can capture the key interactions responsible for ligand binding to COX-1. A plot of predicted energy values for each pose as a function of the RMSD values between each pose and the appropriate crystal structure is also shown (Fig. 2, right panel). For this collection of six compounds, the plot confirms "good" pose predictions as judged by the low RMSD value (RMSD $\leq$ 2Å) of the best scored pose of each compound. These positive results at pose prediction offer the basis on top of which improvements to Vina's ranking performance, through training of its scoring function, may be attempted.

3.2 SVM Classification Models

To select the optimal proportion of actives and inactives (decoys) to include in the training sets, the compounds were first divided in three random sets for 3-fold cross validation, and then the sets of inactives were further divided in n subsets, with n ranging from 5 to 36. Given that each of the n models predicts a different classification for each compound in the test sets, a consensus classification was obtained based on the n classifications of the n models. Then, the F score was calculated for each of the three consensus classification obtained from the cross validation, and the average value of the three F scores attributed to each value of n. The best average F score was obtained for $n = 33$. Thus, the classification models evaluated were produced with the number of partitions in the decoy set equal to 33.

3.3 Performance Evaluation

The performance of Vina's scoring function and our SVM classification models in discriminating active from inactive compounds was evaluated using the values

of the area under the ROC curve (AUC) and enrichment factors (EF) (Table 1, Fig. 3). The values of AUC examine the overall performance of the scoring function and classifiers, while EF values account for their performance on the "early recognition" of the active compounds.

Overall, the results show that SVM classification models (ROC AUC $\simeq$ 75%) perform considerably better than the scoring function of Vina (AUC $\simeq$ 36%) (Fig. 3). In fact, AUC values show that for COX-1, Vina scoring performs worse than random. On the other hand, the SVM classification models trained using the energy parameters of Vina's scoring function unquestionably outperform Vina's scoring function for COX-1.

	AUC (%)	**EF 1%**	**EF 5%**	**EF 10%**
Vina Scoring	36.65	0	0.8	1.6
Classifier 1	74.95	10.96	4.8	4
Classifier 2	74.46	12	4.8	3.6
Classifier 3	74.47	8	4.8	4

Table 1. Comparison of the area under the ROC curve (AUC) and enrichment factors (EF) at top 1%, 5% and 10% of the ranked list of compounds screened obtained for the Vina scoring function and the SVM classification models with $n = 33$.

The performance of Vina scoring function and our classification models were further analysed for early recognition, using the enrichment factor (EF). EF values at 1, 5 and 10% of the data set were determined and the results are summarised in Table 1. The results show that Vina completely fails on the early recognition of active compounds. On the other hand, the SVM classification models perform well across the entire test set with average values of 10.32 ± 2.1, 4.8 ± 0 and 3.87 ± 0.23 at levels EF 1, 5 and 10%, respectively. Overall, the SVM classification models trained using the energy parameters of Vina scoring function show a significant improvement in discriminative power.

4 Conclusions

In the present work we make use of a non-linear machine learning method to train a docking scoring function and thus improve its discriminative power towards COX-1 inhibitors. First, we validated AutoDock Vina's ability to predict correct ligand poses in the target active site. The results show that Vina performs well at pose prediction for COX-1 ligands, with RMSD values below 2Å. Conversely, its scoring function fails to discriminate active compounds from inactive. The results also show that our SVM classification models, trained over the energy parameters of AutoDock Vina's scoring function significantly improve its discriminative power for COX-1 ligands, as reflected by the superior

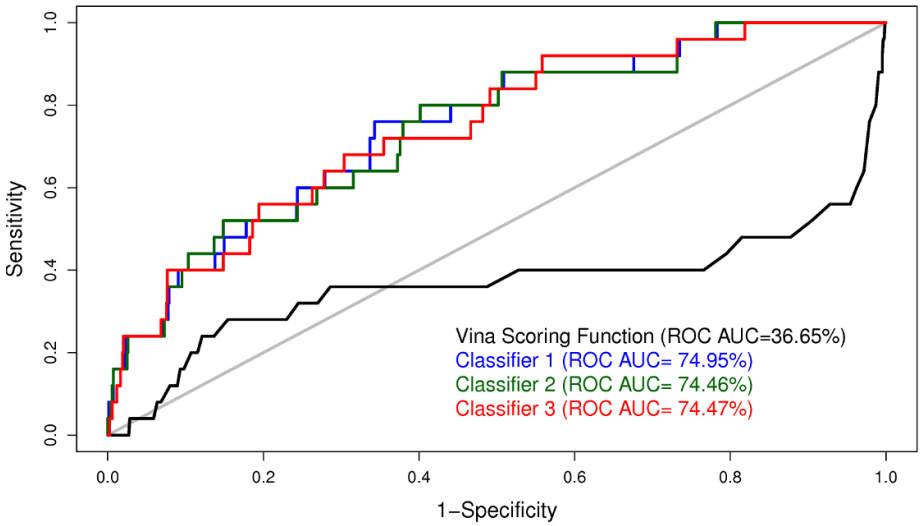

Fig. 3. Area under the ROC curves (AUC) reflecting the performance of Vina scoring function and the SVM classifiers in discriminating actives and inactives for COX-1.

AUC and enrichment factor profiles. The derived models can be thus applied either to the screening of new COX-1 inhibitors or to the design of new drugs devoid of COX-1 activity. The presented approach may be followed for other targets of pharmacological interest in order to increase the likelihood of identifying promising compounds by docking-based virtual screening.

Acknowledgments. This work is funded by ERDF – European Regional Development Fund through the COMPETE Programme (Operational Programme for Competitiveness) and by National Funds through FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) and projects PTDC/QUI-QUI/122900/2010 and Pest-C/SAU/LA0001/2013-2014.

References

1. Green D.V.: Virtual screening of chemical libraries for drug discovery. Expert Opin. Drug Discov. 3(9): 1011–1026 (2008)
2. Kar, S., Roy, K.: How far can virtual screening take us in drug discovery?. Expert Opin. Drug Discov. 8(3), 245–261 (2013)
3. Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J.: Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov. 3(11), 935-949 (2004)
4. Huang, S.Y., Grinter, S.Z., Zou, X.: Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Phys. Chem. Chem. Phys. 12(40), 12899–12908 (2010)

5. Wang, J.C., Lin, J.H.: Scoring functions for prediction of protein-ligand interactions. Curr. Pharm. Des. 19(12), 2174–2182 (2013)
6. Baum, B., Muley, L., Smolinski, M., Heine, A., Hangauer, D., Klebe, G.: Non-additivity of functional group contributions in protein-ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. J. Mol. Biol. 397(4), 1042–1054 (2010)
7. Muley, L., Baum, B., Smolinski, M., Freindorf, M., Heine, A., Klebe, G., Hangauer, D. G.: Enhancement of hydrophobic interactions and hydrogen bond strength by cooperativity: synthesis, modeling, and molecular dynamics simulations of a congeneric series of thrombin inhibitors. J. Med. Chem. 53(5), 2126-2135 (2010)
8. Melville, J.L., Burke, E.K., Hirst, J.D.: Machine learning in virtual screening. Comb. Chem. High Throughput Screen. 12(4), 332–343 (2009)
9. Cannon, E.O., Amini, A., Bender, A., Sternberg, M.J., Muggleton, S.H., Glen, R.C., Mitchell, J.B.: Support vector inductive logic programming outperforms the naive Bayes classifier and inductive logic programming for the classification of bioactive chemical compounds. J. Comput. Aided Mol. Des. 21(5), 269–280 (2007)
10. Geppert, H., Vogt, M., Bajorath, J.: Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J. Chem. Inf. Model. 50(2), 205–216 (2010)
11. Kinnings, S.L., Liu, N., Tonge, P.J., Jackson, R.M., Xie, L., Bourne, P.E.: A machine learning-based method to improve docking scoring functions and its application to drug repurposing. J. Chem. Inf. Model. 51(2), 408–419 (2011)
12. Mitchell, J.B.: Informatics, machine learning and computational medicinal chemistry. Future Med. Chem. 3(4), 451–467 (2011)
13. Garavito, R.M., DeWitt, D.L.: The cyclooxygenase isoforms: structural insights into the conversion of arachidonic acid to prostaglandins. Biochim. Biophys. Acta. 1441(2-3), 278-287 (1999)
14. Huang, N. , Shoichet, B.K., Irwin, J.J.: Benchmarking Sets for Molecular Docking. J. Med. Chem. 49 (23), 6789–6801 (2006)
15. 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. 31, 455–461 (2010)
16. Gupta, K., Selinsky, B.S., Kaub, C.J., Katz, A.K., Loll, P.J.: The 2.0 A resolution crystal structure of prostaglandin H2 synthase-1: structural insights into an unusual peroxidase. J. Mol. Biol. 335, 503–518 (2004)
17. 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 Research, 28, 235–242 (2000)
18. Joachims, T.: Making Large-Scale SVM Learning Practical; MIT Press, Cambridge (1999)
19. Truchon, J.F., Bayly, C.I.: Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. J. Chem. Inf. Model. 47, 488–508 (2007)
20. Zhao, W., Hevener, K.E., White, S.W., Lee, R.E., Boyett, J.M.: A statistical framework to evaluate virtual screening. BMC Bioinformatics 10, 225 (2009)
21. Hamza, A., Wei, N.N., Zhan, C.G.: Ligand-Based Virtual Screening Approach Using a New Scoring Function. J. Chem. Inf. Model. 52, 963–974 (2012)