Consensus virtual screening approaches to predict protein ligands

Andreas Kukol

School of Life Sciences, University of Hertfordshire, College Lane, Hatfield AL10 9AB, United Kingdom
Fax: (+44)(0)1707 284870, E-mail: a.kukol@herts.ac.uk

Abstract – In order to exploit the advantages of receptor-based virtual screening, namely time/cost saving and specificity, it is important to rely on algorithms that predict a high number of active ligands at the top ranks of a small molecule database. Towards that goal consensus methods combining the results of several docking algorithms were developed and compared against the individual algorithms. Furthermore, a recently proposed rescoring method based on drug efficiency indices was evaluated. Among AutoDock Vina 1.0, AutoDock 4.2 and GemDock, AutoDock Vina was the best performing single method in predicting high affinity ligands from a database of known ligands and decoys. The rescoring of predicted binding energies with the water/butanol partition coeffcient did not lead to an improvement averaged over all receptor targets. Various consensus algorithms were investigated and a simple combination of AutoDock and AutoDock Vina results gave the most consistent performance that showed early enrichment of known ligands for all receptor targets investigated. In case a number ligands is known for a specific target, every method proposed in this study should be evaluated.

Keywords: molecular docking, in silico screening, consensus ranking, benchmark, comparison

1. Introduction

Receptor-based virtual screening docks each molecule of a library into a receptor binding site of known or predicted 3D structure. It has been successfully used to predict high affinity protein ligands (Lee et al, 2010; Park et al, 2009). The molecules of the library are ranked according to their predicted binding affinity for the receptor. Apart from saving time and costs in the discovery of ligands for a protein target, an additional benefit is the increased specificity of the predicted ligands, because receptor-based virtual screening is directed against a known binding site or even against a particular receptor conformation (Bruning et al, 2010). This enables the targeting of specific binding sites that are evolutionary conserved in pathogens such as the influenza virus (Darapaneni et al, 2009) or conversely, for endogenous diseases targeting of binding sites that are not conserved among homologous proteins in order to avoid side effects. Critical for the success of a virtual screening experiment is the prediction of binding affinities for the ligand to the receptor. Previous studies have attempted to increase the correlation between predicted and experimental binding affinities. With the consensus docking method VoteDock that combines seven docking algorithms a Pearson correlation coefficient of 0.5 to 0.6 was observed (Plewczynski et al, 2011). Another approach directed at the re-scoring of predicted binding energies with various ligand-derived chemical parameters, such as the water/butanol partition coeffcient achieved correlation coeffcients of better than 0.9 for individual docking methods (Garcia-Sosa et al, 2010). However, these studies did not consider decoy ligands, but only the ligands included in the receptor-ligand complex contained in the PDBbind database (Wang et al, 2005). High correlation coefficients between predicted and experimental binding affinities can be meaningless, if they do not lead to a separation between ligand and decoy molecules as shown in this study.

The present study reveals the first evaluation of the virtual screening performance of the new software AutoDock Vina (Trott & Olson, 2010), the new version of AutoDock 4.2 (Huey et al, 2007) and Gemdock (Yang & Chen, 2004) against a selection of targets from the Database of Useful Decoys (Huang et al, 2006) (DUD). In addition, various strategies of combining the results of two or more docking algorithms were developed and and the utility of recently published ligand efficiency indices (Garcia-Sosa et al, 2010) was evaluated. DUD contains protein targets with known ligands

and decoys that have similar physico-chemical properties to known ligands; thus it provides a challenging test case for virtual screening methods. The performance of virtual screening methods depends on the protein target, therefore, for this work ten targets were selected based on previously reported enrichment factors (Huang et al, 2006) providing a combination of challenging and easy targets.

2. Results and Discussion

The virtual screening performance was analysed with Receiver Operator Characteristic (ROC) curves, where the prediction rate of true ligands (true positives) was plotted against the rate of false positives. Examples of ROC curves for various methods are shown in figure 1 for the target RXRa.

Important for the realisation of the cost- and time-saving features of virtual screening experiments is the prediction of active ligands at the highest ranks of database. In order to capture this 'early' performance the ROC Enrichment (ROCE) factors at 2% of the ligand/decoy database were calculated for all methods and targets. It was obtained as the prediction rate of true ligands (true positives)

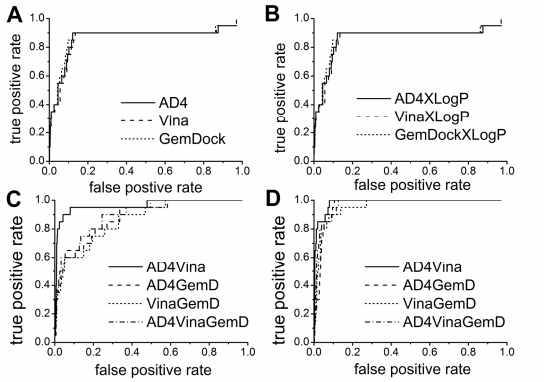

Figure 1: Receiver Operator Characteristics (ROC) curves for virtual screening against the target RXRa. A) ROC curves obtained with standard methods AutoDock, AutoDock Vina and GemDock, B) ROC curves obtained after recalculating the binding affinities with the XlogP3 calculated butanol/water partition, C) ROC curves obtained from a combination of standard methods using the CommonTop algorithm, D) ROC curves obtained from a combination of standard methods with the joining algorithm

Table 1: Receiver Operator Curve enrichment (ROCE) at 2% of ligands for various targets and methods.

Method	Target											
	AchE	CDK2	COMT	FGFr1	HIVRT	InhA	PPARg	PR	RXRa	VEGFr2	**Mean**	StdErr[d]
AD4	1.5	17.1	0.0	2.2	2.6	16.0	0.6	13.1	31.2	7.2	**9.2**	2.1
Vina	1.5	13.3	10.4	0.0	8.8	16.0	35.8	2.0	31.2	6.2	**12.5**	3.7
GemD	1.5	7.1	17.9	4.0	7.4	2.5	9.3	6.9	31.2	7.2	**9.5**	2.7
AD4-P[a]	1.5	5.6	0.0	8.0	3.9	2.5	8.5	0.0	31.2	4.5	**6.6**	2.7
Vina-P	1.5	5.6	0.0	1.3	5.6	2.5	5.9	0.0	31.2	4.5	**5.8**	2.8
GemD-P	2.5	5.6	0.0	2.7	3.9	2.5	8.5	2.0	31.2	4.5	**6.3**	2.7
CT-AV[b]	2.5	13.3	0.0	0.9	2.6	20.9	2.6	2.0	54.9	5.4	**10.5**	5.1
CT-AG[b]	0.0	10.0	0.0	3.1	5.6	20.9	2.6	4.3	31.8	4.5	**8.3**	3.1
CT-VG[b]	0.5	11.3	17.9	1.3	5.6	19.9	15.9	2.0	24.2	6.2	**10.5**	2.6
CT-AVG[b]	0.0	8.2	0.0	1.8	2.6	19.9	3.4	0.0	31.8	6.2	**7.4**	3.1
J-AV[c]	1.5	14.7	10.4	1.7	7.4	19.9	17.1	6.9	146.3	7.2	**23.3**	13.1
J-AG	2.5	10.0	4.4	4.7	7.4	21.7	8.3	9.8	18.2	7.2	**9.4**	1.8
J-VG	2.0	11.3	10.4	2.7	10.9	16.8	24.7	4.3	24.2	7.2	**11.5**	2.5
J-AVG	2.5	8.2	10.4	4.0	8.8	19.9	15.9	6.9	31.8	9.2	**11.8**	2.6

[a] Binding energy scores of the ranked list were recalculated with the computed water/octanol partition coefficient, as new_score = $\log_{10}$ (-score/P).
[b] The ligands, which were common to AutoDock and AutoDock Vina in the top n positions were chosen (with n = 1, 2, 3, ...), AG: AutoDock/GemDock, VG: Vina/GemDock, AVG: AutoDock/Vina/GemDock
[c] Joined rank lists
[d] The standard error was calculated as the sample standard deviation divided by the square root of the number of targets, i.e. ten.

divided by the rate of false positives, where an ROCE factor of one denotes no improvement above random picking of molecules, while an ROCE factor above one denotes better than random performance.

The resulting ROCE factors shown in table 1 are highly variable between methods and targets ranging from $ROCE_{2\%}$ = 0 to 146. There is substantial target variation from easy targets that achieve reasonable performance with any method to difficult targets that achieve a low performance with every method. Among the individual docking methods, AutoDock Vina shows the best performance with an average $ROCE_{2\%}$ = 12.5±3.7, which is also the second best overall performance, while the best consensus approach is the simple joining method of AutoDock and AutoDock Vina rank lists with an average $ROCE_{2\%}$ = 23±13.

It should be noted the standard error reported represent between-target variation, while a repeat docking with the same target yielded almost identical results. The rescoring of the binding energies ΔG with the water/butanol partition coefficient P according to log(-ΔG/P) as suggested by Garcia-Sosa et al. (2010), did not lead to a significant improvement. The authors of this study reported an improvement of correlation coefficient between calculated and experimental binding energies from 0.347 for the AutoDock calculated binding energies to 0.996 after rescoring. A reason for the lack of improvement in the current study may be the challenging DUD decoy set, which was chosen according to physico-chemical similarity with the known ligands (Huang et al, 2006).

Most notably, the simple method of joining of AutoDock and AutoDock Vina rank lists illustrated in figure 2 achieved above random performance for all targets. This is important in situations, where no existing ligand for a target is known. In the situation, when a number of ligands is known for a protein target all methods should be tested as table 1 shows that even methods with a low average performance can perform well on a particular target, e.g. for the target FGFr1 the rescoring with the water/octanol partition coefficient yielded the best performance of $ROCE_{2\%}$ = 8.0.

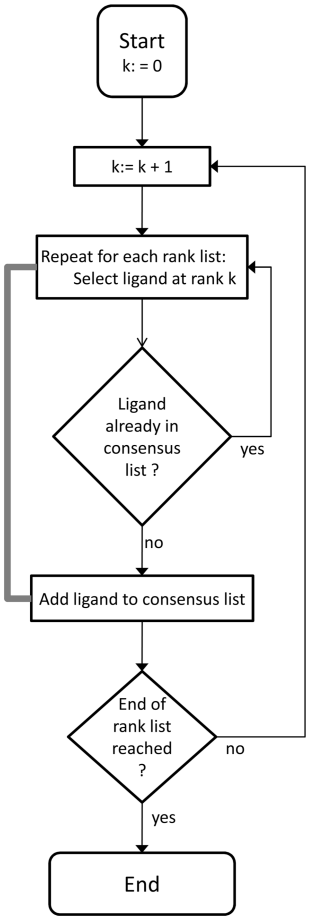

Figure 2: Flowchart of the consensus method based on a simple joining of rank lists from individual methods.

It is interesting, to note the relationship between the ranking of ligands from various docking methods as shown in figure 3 for the example of the target RXRa. If two docking methods were perfect predictors of binding affinity, all rankings between the two methods should fall on a straight line in figure 3. In reality there is almost no correlation between the ranks obtained by two different docking algorithms, even if they were developed in the same research group such as AutoDock and AutoDock Vina (figure 3a). There is a weak correlation for the highest ranked predictions up to rank 120 as highlighted in figure 3. While this lack of correlation shows that docking methods are far from perfect, it also provides the opportunity to combine predictions from different methods as it was exploited in the current study.

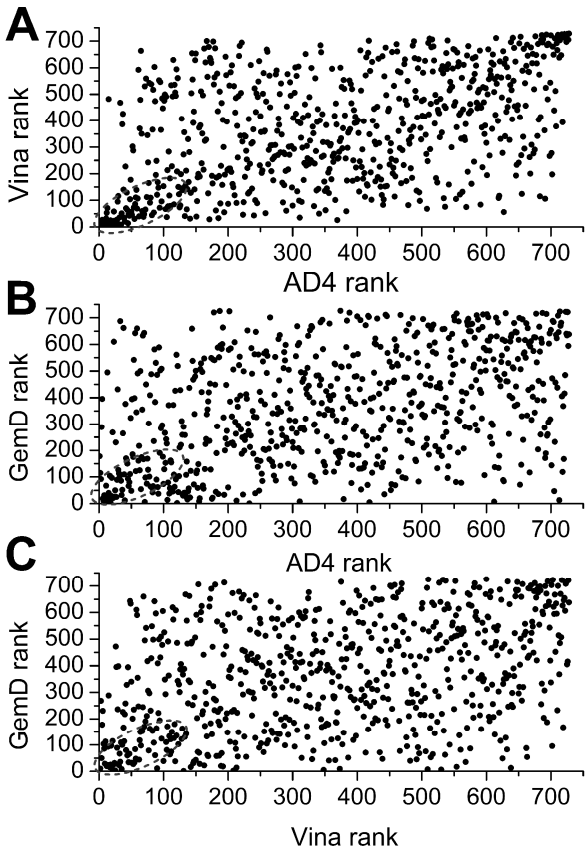

Figure 3: Correlation between the ranking of molecules in the DUD dataset for the target RXRa. The rank obtained with one docking algorithm was plotted agains the rank obtained with another algorithm. The area surrounded by a dashed line highlights some correlation observed at the highest ranking ligands. A) AutoDock Vina ranks plotted against AutoDock ranks, B) GemDock ranks plotted against AutoDock ranks and C) GemDock ranks plotted against AutoDock Vina ranks.

For the two most difficult targets, AchE and FGFr1, further approaches were investigated such as increasing the calculation time of individual methods by a factor of three or combining two or more rank lists by calculating a consensus rank (method 4). The results in table 2 showed that the increase of docking calculation time by a factor of three lead to a minimal or no improvement apart from GemDock. However, the GemDock performance did not improve to any level higher than shown in table 1. The consensus scoring methods did not show any improvements compared to the methods discussed previously.

The following recommendations can be derived from the results reported in this work. If active ligands for a target are known, the performance of predicting new ligands can be improved by choosing between different docking methods or a combination thereof.

Table 2: Receiver Operator Curve enrichment (ROCE) at 2% for the two most difficult targets employing additional methods.

Method	Target	
	AchE	FGFr1
AD4	1.5	2.2
Vina	1.5	0.0
GemD	1.5	4.0
AD4 x3[a]	1.5	2.7
Vina x3[b]	0.95	0.4
GemD x3[a]	2.0	6.5
Cons-AV[c]	2.5	1.0
Cons-AG	1.0	2.5
Cons-VG	0.5	0
Cons-AVG	0.0	1.0

[a] The number of calculations was increased by a factor of 3.

[b] In AutoDock Vina the exhaustiveness parameter was increased from 8 to 24.

[c] Consensus method that calculates an average rank from two or more ranked lists and applies a weighting factor from 1.0 to 0.0 based on the position of a molecule in the individual ranked list, i.e. molecules at the top are contributing more to the average rank.

If no active ligands are known, the performance of ligand prediction can be improved through a consensus method based on a simple combination of AutoDock and AutoDock Vina rank lists. The software for employing the consensus methods is available from the author upon request.

3. Conclusion

In summary, this communication reports about the development of new consensus methods for virtual screening, which are evaluated against three individual molecular docking algorithms. Among the individual docking algorithms AutoDock Vina achieved the best performance in the early enrichment of known ligands followed by GemDock and AutoDock 4.2. A simple joining of AutoDock 4.2 and AutoDock Vina rankings gave the best early enrichment performance that lead to enrichment above random in every case, while individual methods failed for some targets investigated. These results are important for the early stages of drug discovery as well as academic research, where costly and time-

consuming laboratory experiments can be replaced with *in silico* methods.

4. Computational methods

AutoDock 4.2 (Huey et al, 2007; Morris et al, 1998) was used with the Lamarckian Genetic algorithm. Default parameters were used except that the frequency of performing a local search was set to 0.15. AutoDock Vina (Trott & Olson, 2010) version 1.02 was employed with default parameters and GemDock (Yang & Chen, 2004) parameters were adjusted to a population size of 300, 80 generations and 5 solutions resulting in a similar docking time/CPU than AutoDock. Various strategies of improving the recall of known ligands in the top 2% of the ranked ligand databases were employed such as 1) the rescoring of the ranked list with the computed water/octanol partition coefficient (Cheng et al, 2007) P according to the equation new_score = $\log_{10}$ (-score/P) as suggested by Garcia-Sosa et al. (Garcia-Sosa et al, 2010); 2) the combination of two or three ranked ligand lists by choosing the ligands that are common among the top n ligands of each list, whereby n is counted in steps of five to the maximum number of ligands (CommonTop); 3) a simple joining of ranked ligand list by choosing the top n (n = 1, 2, 3, ...) ligand from each list , if it was not chosen previously; 4) a consensus scoring method, where for each ligand a weighted average rank was computed from two or more rank lists. The weight was decreased linearly from 1.0 to 0.0 based on the position of the ligand in the rank list. The processing of rank lists was performed with PERL scripts developed in-house, that are available from the author on request. The virtual screening performance was evaluated as suggest by Nicholls (Nicholls, 2008) with Receiver Operator Characteristics Enrichment (ROCE) by dividing the fraction of true positives by the fraction of false positives at 2% of the ligand/decoy molecules. An $ROCE_{2\%}$ value of 1.0 is expected for random picking of ligands. According to Nicholls (Nicholls, 2008), the ROCE provides a more robust measure of performance than the commonly reported enrichment factor. Enrichment factors for comparison with other studies are shown in table 3.

Table 3: Enrichment factors at 1% of for various targets and methods

Method	Target											
	AchE	CDK2	COMT	FGFr1	HIVRT	InhA	PPARg	PR	RXRa	VEGFr2	**Mean**	StdErr[4]
AD4	1.9	16.2	0.0	3.4	4.9	22.0	1.2	11.2	20.8	6.8	**8.8**	2.5
Vina	1.0	10.1	18.2	0.0	7.4	15.4	23.5	0.0	26.0	8.2	**11.0**	2.9
GemD	2.9	8.1	9.1	5.1	12.3	16.6	11.1	7.5	0.0	9.6	**8.2**	1.4
DOCK[1]	1.9	13.9	0.0	0.0	5.0	0.0	0.0	0.0	24.8	1.3	**4.7**	2.5
AD4-P	1.9	10.1	0.0	10.2	4.9	1.2	7.4	0.0	20.8	6.8	**6.3**	1.9
Vina-P	1.9	10.1	0.0	1.7	4.9	1.2	2.5	0.0	20.8	8.2	**5.1**	1.9
GemD-P	1.0	10.1	0.0	9.3	4.9	1.2	4.9	0.0	15.6	8.2	**5.5**	1.6
CT-AV	2.9	16.2	0.0	3.4	2.5	25.0	1.2	3.7	20.8	8.2	**8.4**	2.7
CT-AG	0.0	12.1	0.0	3.4	7.4	22.6	0.0	7.5	10.4	6.8	**7.0**	2.1
CT-VG	0.0	12.1	27.3	1.7	7.4	22.6	16.1	3.7	10.4	10.9	**11.2**	2.6
CT-AVG	0.0	10.1	0.0	2.5	4.9	22.6	3.7	0.0	15.6	5.5	**6.5**	2.3
J-AV[3]	1.0	10.1	18.2	2.5	7.4	22.6	11.1	3.7	26.0	10.9	**11.4**	2.6
J-AG	2.9	12.1	9.1	5.9	9.8	24.9	11.1	11.2	15.6	9.6	**11.2**	1.8
J-VG	1.9	12.1	18.2	2.5	12.3	19.0	18.5	7.5	15.6	10.9	**11.9**	1.9
J-AVG	2.9	10.1	18.2	4.2	12.3	23.8	12.4	7.5	15.6	13.7	**12.1**	1.9

[1] Data taken from Huang et al. (2006).

Acknowledgements

This work made use of the University of Hertfordshire Science and Technology Research Institute high-performance computing facility. The work was supported by the School of Life Sciences at the Health and Human Sciences Research Institute, University of Hertfordshire, United Kingdom

References

Bruning JB, Parent AA, Gil G, Zhao M, Nowak J, Pace MC, Smith CL, Afonine PV, Adams PD, Katzenellenbogen JA, Nettles KW (2010) Coupling of receptor conformation and ligand orientation determine graded activity. *Nature Chemical Biology* **6:** 837-843

Cheng TJ, Zhao Y, Li X, Lin F, Xu Y, Zhang XL, Li Y, Wang RX, Lai LH (2007) Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *Journal of Chemical Information and Modeling* **47:** 2140-2148

Darapaneni V, Prabhaker VK, Kukol A (2009) Large-scale analysis of influenza A virus sequences reveals potential drug target sites of non-structural proteins. *Journal of General Virology* **90:** 2124-2133

Garcia-Sosa AT, Hetenyi C, Maran U (2010) Drug Efficiency Indices for Improvement of Molecular Docking Scoring Functions. *Journal of Computational Chemistry* **31:** 174-184

Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry* **49:** 6789-6801

Huey R, Morris GM, Olson AJ, Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry* **28:** 1145-1152

Lee K, Jeong KW, Lee Y, Song JY, Kim MS, Lee GS, Kim Y (2010) Pharmacophore modeling and virtual screening studies for new VEGFR-2 kinase inhibitors. *Eur J Med Chem* **45:** 5420-5427

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **19:** 1639-1662

Nicholls A (2008) What do we know and when do we know it? *Journal of Computer-Aided Molecular Design* **22:** 239-255

Park H, Bhattarai BR, Ham SW, Cho H (2009) Structure-based virtual screening approach to identify novel classes of PTP1B inhibitors. *Eur J Med Chem* **44:** 3280-3284

Plewczynski D, Lazniewski M, von Grotthuss M, Rychlewski L, Ginalski K (2011) VoteDock: consensus docking method for prediction of protein-ligand interactions. *J Comput Chem* **32:** 568-581

Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *Journal of Computational Chemistry* **31:** 455-461

Wang RX, Fang XL, Lu YP, Yang CY, Wang SM (2005) The PDBbind database: Methodologies and updates. *Journal of Medicinal Chemistry* **48:** 4111-4119

Yang JM, Chen CC (2004) GEMDOCK: A generic evolutionary method for molecular docking. *Proteins-Structure Function and Bioinformatics* **55:** 288-304