Frontiers in Drug, Chemistry and Clinical Research

Mini Review

ISSN: 2631-5424

Tailoring specialized scoring functions for more efficient virtual screening

Vieira TF, Magalhaes RP and Sousa SF*

UCIBIO@REQUIMTE, BioSIM - Departamento de Biomedicina, Faculdade de Medicina da Universidade do Porto, Alameda Professor Hernâni Monteiro, 4200-319 Porto - Porto, Portugal

Abstract

Protein-ligand docking is a computational method that is commonly applied to predict and rank the structure of the complex formed between a specific protein-target and a small-molecule ligand. Despite its limitations, it is currently an integral part of the drug design and development process and is often used with virtual screening to evaluate large computational databases of molecular compounds, as a first attempt to guide the selection of limited sets of compounds for experimental testing.

More than 30 scoring functions are currently available and are routinely used to evaluate ligand binding in docking. However, their performance in not uniform, and depends on a variety of features such as the specific structural and chemical characteristics of the binding pocket of the target protein, and those of the ligand molecules.

This work reviews some of the limitations of scoring functions in protein-ligand docking, while demonstrating strategies to select the most effective alternatives for particular problems. Special emphasis is given to the design of new consensus scoring functions for more efficient drug discovery campaigns.

Introduction

Protein-ligand docking is a popular computational method used to predict and rank the structure(s) resulting from the association between a small-molecule (the ligand) and a protein target of known 3-D structure, typically an enzyme or a receptor. It is a relevant part of the current drug discovery process [1-7]. Docking involves the search through different ligand conformations (i.e. the pose) within a given target protein, and the measure of the binding affinity of the different alternatives (the scoring). The first is accomplished by the application of the search algorithm, while the latter is handled by the scoring function.

A critical aspect of each docking protocol is the ability to evaluate and rank the ligand conformations predicted by the search algorithm. In fact, being able to generate the right conformation is not enough. It is also necessary to be able to recognize it from among the set of generated conformations. In addition, the scoring function should also be able to distinguish between active and random compounds i.e. molecules that bind to a specific target from non-binding molecules. Over the years, significant improvements in the computational power and software algorithms have enabled substantial progress in the ability of docking software in covering adequately the conformational space of a ligand inside the target's binding pocket. However, presently, the lack of efficient scoring functions, in terms of speed and accuracy is the major bottleneck in docking [6,8]. A large number of scoring functions is currently available [9-13]. However, their performance can vary significantly from problem to problem [12,14-16].

Virtual screening (VS) involves the use of computational techniques capable of using large chemical databases to guide the selection of likely drug candidates for specific pharmacological targets [17-19]. The main objective of virtual screening is to obtain hits of novel chemical structures that bind strongly to a specific target. Structure-based VS methods apply protein-ligand docking to evaluate the interaction between a small-molecule ligand and a protein target, trying to discriminate molecules that bind strongly from molecules that do not. These techniques can be applied to databases containing millions of compounds such as ZINC [20,21]. Because the cost of performing these virtual screens is significantly less than currently existing experimental alternatives, VS has been playing an important role in drug design and discovery, limiting the number of compounds that are experimentally evaluated to a subset of molecules that are more likely to bind effectively, while ensuring a greater exploration of the chemical space. For these reasons, VS and docking have become critical components in the drug design and development process [1,7,19,22].

Despite its huge potential and general use, VS has several limitations, with impact on the accuracy and direct applicability of its conclusions [18,19,23]. One of the main problems of VS comes from the large number of false positives i.e. molecules that are erroneously suggested from docking to bind strongly to the target. An even worse problem in VS comes from the large number of false negatives, i.e. molecules that docking fails to identify as strong ligands, despite their high affinity. In fact, while the first molecules can be easily discarded in the preliminary experimental studies with a relatively small cost, the latter never reach that stage remaining incognito among millions of compounds, despite their sometimes high potential pharmacological, social and economic

Correspondence to: Sergio F Sousa, UCIBIO@REQUIMTE, BioSIM - Departamento de Biomedicina, Faculdade de Medicina da Universidade do Porto, Alameda Professor Hernâni Monteiro, 4200-319 Porto - Porto, Portugal, E-mail: sergiofsousa@med.up.pt

Key words: *docking, virtual screening, computer-aided drug discovery, scoring, consensus*

Received: December16, 2018; **Accepted:** January 04, 2019; **Published:** January 14, 2019

value. Both these problems arise from the imperfections in the currently available scoring functions. While it is not exactly difficult to improve the mathematical description in the scoring functions as to provide more accurate results, it is important to take into account that scoring functions were designed to ensure a fast estimation of the binding free energy between a ligand and a protein for a large numbers of molecules. In fact, more accurate methods such as thermodynamic integration or even quantum mechanics can be applied in principle to evaluate the interaction of a protein and a ligand, but with a computational cost of 10,000 to 1,000 000 times higher than that of docking, rendering the analysis of several molecules unfeasible. Such methods are presently not viable for virtual screening, forcing the adoption of a series of simplifications to reduce the complexity of the scoring function.

So, the inaccuracies in the scoring function continue to be the most important limitation to the success of docking and, therefore, to a more successful use of VS in drug design. The development of new improved scoring functions, more accurate but fast, is therefore a problem of the maximum importance for drug design and of particular interest for the pharmaceutical industry.

Specificities in scoring

Scoring functions are mathematical expressions use to evaluate the interaction of the protein and the ligand. Different scoring functions have different strengths and weaknesses. In general, they are based on different physical principles, which in some cases are related to the statistical representation in selected sets of protein-ligand complexes for which there is extensive and detailed experimental information. The range of alternatives is vast and diverse, comprising a total of more than 30 scoring functions [6,10,24]. These can be divided into three major classes: force-field-based, empirical, and knowledge-based scoring functions.

Force-field based scoring functions evaluate the protein-ligand interaction as the sum of two energies: the interaction energy of the receptor/ligand pair, plus the internal energy of the ligand. This is accomplished through a combination of a van der Waals (lennard-jones) and an electrostatic energy term (columbic potential). Typical limitations of force-field scoring functions include the lack of entropic and solvation terms, and the simplicity in the description of the long-range effects associated to docking.

Empirical scoring functions are based on a rather different principle. These scoring functions are designed to reproduce experimental data. They are fitted to use a sum of several individual uncorrelated terms to reproduce experimental results, such as binding free energies, by means of a regression analysis. The individual terms are normally fast to calculate, but the success of these scoring functions depend on the experimental data used in the parameterization process. This process is not controlled by the user and is easily not transferable.

Knowledge-based scoring functions apply statistical principles obtained from the analysis of collections of 3D structures. These scoring functions are typically based on the frequency of occurrence of different atom-atom pair contacts and other typical interactions if large data-sets of protein ligand complexes of know 3D structure. This simplicity of these scoring functions makes them efficient in screening large compound databases. However, their accuracy is dependent also on the type of structures that were included in the 3D databases that were used in their creation process.

Some of these scoring functions are better in handling protein targets with certain structural and chemical characteristics, while other are more accurate in targets with other specific properties. Examples of the features that can often make a difference include the size and exposure of the binding pocket, the presence of cofactors and metal atoms, the presence of very charged groups around the binding pocket, etc.

Furthermore, the performance of different scoring functions also varies very significantly with the characteristics of molecules that are tested in docking. Features like the protonation state, partial charge, and number of rotatable bonds are just some examples of properties that can affect the performance of a scoring function.

It is generally difficult to anticipate the best scoring function for a particular target. Choice often tends to rely on the availability of a particular software with a specific scoring function to the researcher/user. Some docking programs are freely available, through open source licenses, while other are freely available but only to academic users. However, many alternatives are paid, sometimes through very expensive licences [8,25].

Other common rationale for the choice of the scoring function is the familiarity of the user with that specific software. Researchers tend to use the protein-ligand docking software that they already know and are confident in using from a technical point of view. Also, while some alternatives have extensive background information and a wide availability of manuals, user guides and tutorials, other are difficult to start with, due to lack of information. All these issues have very little to do with the scoring accuracy of a specific scoring function in discriminating between ligands and non-ligand molecules for a specific protein target. However, they continue to be responsible for the final choice for most users.

For accurate results however, it is important to rely on clear and objective guidelines on the scoring functions that should be used for different types of proteins and different types of ligands. However, that information is seldom available, leaving to the user the heavy burden of selecting and validating the particular scoring function to use with a specific protein-target and ligand type.

Relying in scoring

To rely on a specific scoring function for a particular protein target, it is important to understand the limits of validation of the different alternatives. For that, their performance has to be evaluated. That is seldom an easy task.

Most scoring functions and docking programs upon publication report extensive validation tests. In general, such tests demonstrate their superior performance. However, questions often emerge on the existence of bias on the test sets used. When comparing different docking programs and scoring function, it is important that the comparison be done on unbiased validation sets.

One of the most commonly used reference validation sets is the DUD-E (directory of useful decoys – extended). DUD-E is a collection of useful decoys for benchmarking virtual screening containing 22,886 ligands and their affinities against 102 target, set by Huang, et al. [26,27]. For each of the ligands, this database contains a set of 50 "decoys", i.e. molecules with similar 1-D physico-chemical properties to remove bias (e.g. molecular weight, calculated LogP) but dissimilar 2-D topology to be likely non-binders, making it a challenging dataset to test scoring functions and protein-ligand docking algorithms.

Using this dataset, the performance of a scoring function in virtual screening can be expressed through a graphical representation of the

true positive rate versus the false positive rate in terms of receiver operating characteristic (ROC) plots. In ROC plots the True Positive Rate (TPR = TP/P) is plotted versus the False Positive Rate (FPR = FP/N), where TP is the number of True Positives, P is the total number of Positives (actives), FP is the number of False Positives, and N is the total number of Negatives (decoys). An useful measure is the area under the curve (AUC). The higher the AUC value in a ROC curve, the better the discrimination between the true positive and the false positive poses. As a successful scoring function for virtual screening should rank active compounds early on a large score list, the fraction of actives recovered at 0.1%, 1% and 2% decoys recovered (abbreviated to ROC(0.1%), ROC(1%) and ROC(2%)) - called early recognition metrics - can be used to discriminate between the best and the worst scoring functions for each protein target.

This strategy allows a clear and objective definition of the limits of validity of the different scoring functions, in terms of the different characteristics of the type of protein target, and/or of the characteristics of target binding pockets and molecules.

The best scoring function for a specific target

While the strategy outlined above ensures a selection of a reasonable scoring function in protein ligand docking for a general type of protein-target or ligand (e.g. among GPCRs or among metalloenzymes, etc), providing a reasonable starting point for most studies, it is important to stress that even within a specific type of target or ligand, the performance of different scoring functions can vary drastically.

The simplistic nature of most scoring function is seldom able to capture all of the intrinsic characteristics of the protein binding pocket and ligand interaction. For example, among metalloenzymes a particular scoring function can give very good results for one specific protein-target, but bad results for other, depending on the specific metal, its location in the binding pocket, type of dominant interactions, exposure to solvent, etc. Such variations in performance often happen event within sets of related proteins, including among GPCRs.

To ensure that a good result can be obtained for a specific protein target, the best strategy relies in designing and optimizing your own active/decoys test set for your specific target. This strategy is subject to the availability of experimental information on active ligands. Presently, several well-known databases contain detailed information on specific active ligands with experimental information on their binding ability to particular protein targets. Examples include the bindingDB (http://www.bindingdb.org/) [28-32] and the CHEMBL (https://www.ebi.ac.uk/chembl/) [33,34] databases.

Users can search these databases for known active molecules against their specific protein targets. Recorded values include IC50, EC50, Km and percentage of inhibition, etc. Values can be exported in .csv format including the experimental data together with extensive information on the active molecules contained in the database, including molecular weight and other physic-chemical properties, as well as the formula of the different molecules in smiles or sdf format.

From the structure of the active molecules, decoys can be generated using the Generate DUD-E decoys server (http://dude.docking.org/generate) [27]. This online server can be used to generate 50 decoys for each active ligand, generating a final customized active/decoys test set with several hundred molecules. This specific test can then be used to assess the performance of the different scoring function, for the user's own specific protein target, guaranteeing that the best scoring function available to the user is in fact chosen for the virtual screening campaign.

Consensus scoring

In many cases, even the best individual scoring function gives results that are far from ideal. To go beyond the level of success that can be given in these cases, consensus scoring has to be applied.

Consensus Scoring [8,35] combines the information obtained from different scores to compensate for errors from individual scoring functions, therefore improving the probability of finding the correct solution. Several studies have demonstrated the success of consensus scoring methods in relation to the use individual functions schemes [36-39]. Improved consensus scoring can be achieved through a combination of better and more scoring functions, and by a careful parameterization of the weight (coefficients) to attribute to each individual scoring function.

The best possible strategy today consists in creating an active/decoys database, as described above to test the performance of different individual scoring functions, and then combining the best scoring functions, by attributing different coefficients to the individual scores. These coefficients are then optimized, adjusting the weight of the different scoring functions in the final consensus function, as to maximize active/decoys discrimination in the test set prepared, effectively tailoring the scoring function to that specific protein target.

The final optimized consensus scoring is then used in the virtual screening campaign to sample databases containing millions of compounds.

Conclusions

Despite its limitations virtual screening is a computational technique nowadays routinely used in drug discovery efforts to improve the probability of identifying novel chemical entities of potential pharmacological interest in databases containing millions of molecules. While the principles of virtual screening are easy to define, a careful selection of the specific scoring function is paramount for that useful results can be obtained, with the same scoring function displaying quite different performance for different types of targets.

While general guidelines can be anticipated for different types of protein targets, we emphasize the importance of a customized choice of the scoring function for a specific protein target through the creation of active/decoy test sets by the user. This strategy ensures a rational choice of scoring function for a particular target from among the scoring functions available to a specific user.

However, as highlighted, superior performance can be obtained by combining different scoring functions, by attributing different weights to individual scores to maximize active/decoys discriminating in the selected test sets, effectively tailoring highly specialized consensus scoring functions for a particular protein-target.

Acknowledgments

This work was supported by national funds from Fundação para a Ciência e a Tecnologia (SFRH/BD/137844/2018 and IF/00052/2014) and co-financed by the ERDF under the PT2020 Partnership Agreement (POCI-01-0145- FEDER-007728).

References

1. de Ruyck J, Brysbaert G, Blossey R, Lensink MF (2016) Molecular docking as a popular tool in drug design, an in silico travel. *Adv Appl Bioinform Chem* 9: 1-11. [Crossref]

2. Ferreira LG, Dos Santos RN, Oliva G, Andricopulo AD (2015) Molecular docking and structure-based drug design strategies. *Molecules* 20: 13384-13421. [Crossref]

3. Gupta M, Sharma R, Kumar A (2018) Docking techniques in pharmacology: How much promising? *Comput Biol Chem* 76: 210-217. [Crossref]

4. Huang SY, Zou X (2010) Advances and challenges in protein-ligand docking. *Int J Mol Sci* 11: 3016-3034. [Crossref]

5. Lohning AE, Levonis SM, Williams-Noonan B, Schweiker SS (2017) A Practical guide to molecular docking and homology modelling for medicinal chemists. *Curr Top Med Chem* 17: 2023-2040. [Crossref]

6. Sousa SF, Fernandes PA, Ramos MJ (2006) Protein-ligand docking: current status and future challenges. *Proteins* 65: 15-26. [Crossref]

7. Wang G, Zhu W (2016) Molecular docking for drug discovery and development: a widely used approach but far from perfect. *Future Med Chem* 8: 1707-1710. [Crossref]

8. Sousa SF, Ribeiro AJ, Coimbra JT, Neves RP, Martins SA, et al. (2013) Protein-ligand docking in the new millennium--a retrospective of 10 years in the field. *Curr Med Chem* 20: 2296-2314. [Crossref]

9. Adeniyi AA, Soliman MES (2017) Implementing QM in docking calculations: is it a waste of computational time? *Drug Discov Today* 22: 1216-1223. [Crossref]

10. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* 295, 337-356. [Crossref]

11. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins* 47: 409-443. [Crossref]

12. Rarey M, Degen J, Reulecke I (2008) Docking and scoring for structure-based drug design.

13. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, et al. (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* 49: 5912-5931. [Crossref]

14. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) Comparing protein-ligand docking programs is difficult. *Proteins* 60: 325-332. [Crossref]

15. Li C, Liu Y, Lu L, Ma M, Tan J, et al. (2017) Research progresses on scoring function design and rna-binding site identification in protein-RNA docking. *Journal Beijing Univ Technol* 43: 1779-1786.

16. Sotriffer C (2018) Docking of covalent ligands: challenges and approaches. *Mol Inform* 37: e1800062. [Crossref]

17. Cerqueira NMFSA, Gesto D, Oliveira EF, Santos-Martins D, Brás NF, et al. (2015) Receptor-based virtual screening protocol for drug discovery. *Arch Biochem Biophys* 582: 56-67. [Crossref]

18. Fradera X, Babaoglu K (2017) Overview of methods and strategies for conducting virtual small molecule screening. *Curr Protoc Chem Biol* 9: 196-212. [Crossref]

19. Sousa SF, Cerqueira NM, Fernandes PA, Ramos MJ (2010) Virtual screening in drug design and development. *Comb chem high throughput screen* 13, 442-453. [Crossref]

20. Irwin JJ, Shoichet BK (2005) ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45: 177-182. [Crossref]

21. Nicola G, Liu T, Gilson MK (2012) Public domain databases for medicinal chemistry. *J Med Chem* 55: 6987-7002. [Crossref]

22. Crespo A, Rodriguez-Granillo A, Lim VT (2017) Quantum-Mechanics Methodologies in Drug Discovery: Applications of Docking and Scoring in Lead Optimization. *Curr Top Med Chem* 17: 2663-2680. [Crossref]

23. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 3: 935-949. [Crossref]

24. Kroemer RT (2007) Structure-based drug design: docking and scoring. *Curr Protein Pept Sci* 8: 312-328. [Crossref]

25. Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. *Biophys Rev* 9: 91-102. [Crossref]

26. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. *J Med Chem* 49: 6789-6801. [Crossref]

27. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* 55: 6582-6594. [Crossref]

28. Chen X, Liu M, Gilson MK (2001b) Binding DB: A web-accessible molecular recognition database. *Comb Chem High Throughput Screen* 4: 719-725.

29. Chen X, Lin Y, Gilson MK (2001) The binding database: overview and user's guide. *Biopolymers* 61: 127-141. [Crossref]

30. Chen X, Lin Y, Liu M, Gilson MK (2002) The binding database: data management and interface design. *Bioinformatics* 18: 130-139. [Crossref]

31. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, et al. (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44: D1045-1053. [Crossref]

32. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res* 35: D198-201. [Crossref]

33. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, et al. (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res* 42: D1083-1090. [Crossref]

34. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40: D1100-1107. [Crossref]

35. Wang R, Wang S (2001) How does consensus scoring work for virtual library screening? an idealized computer experiment. *J Chem Inf Comput Sci* 41: 1422-1426. [Crossref]

36. Charifson PS, Corkery JJ, Murcko MA, Walters WP (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* 42: 5100-5109. [Crossref]

37. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB (2002) Consensus scoring for ligand/protein interactions. *J Mol Graph Model* 20, 281-295. [Crossref]

38. Houston DR, Walkinshaw MD (2013) Consensus docking: Improving the reliability of docking in a virtual screening context. *J Chem Inf Model* 53, 384-390. [Crossref]

39. Teramoto R, Fukunishi H (2007) Supervised consensus scoring for docking and virtual screening. *J Chem Inf Model* 47: 526-534. [Crossref]