

Research Article

Shape Matters: Improving Docking Results by Prior Analysis of Geometric Attributes of Binding Sites

Renee M Keyes¹, Ervin Pejo², Kazuo Katagiri², Ken Huynh², Aleksandra Rudnitskaya², Boguslaw Stec³ and Kimberly A Stieglitz^{1*}

¹Department of Biotechnology, Roxbury Community College, USA

²Department of Chemistry, University of Massachusetts, USA

³Department of Chemistry, Burnham Institute for Medical Research, USA

***Corresponding author**

Kimberly A Stieglitz, Department of Biotechnology, Roxbury Community College, 1234 Columbus Ave, Boston, MA, USA, Email: kastieglitz@rcc.mass.edu

Submitted: 21 March 2016

Accepted: 07 April 2016

Published: 16 April 2016

ISSN: 2333-6633

Copyright

© 2016 Stieglitz et al.

OPEN ACCESS

Abstract

An important improvement for selection of docking programs has been found. Correlating the attributes of ligand binding pocket shape with an appropriate program in the early stages of automated docking has been proven to increase the success of the procedure. This potentially constitutes an important improvement in structure-based drug design process. A two-stage approach: (1) computing attributes of the binding site and (2) running the appropriate docking algorithm, has been used to screen ~one hundred structures from the Protein Data Bank (PDB). The attributes of the binding pockets used in this study were: the ratio of the volume of the solvent accessible surface to the volume of the molecular surface (V_{sa}/ms) and the area of the solvent accessible surface to the area of the molecular surface (A_{sa}/ms). This study doesn't look at charges and H-bonding or hydrophobic interactions. However it is still very useful to aid in choosing the best docking program possible. The results of initial screening within the bounds of optimally selected parameters indicated that, it is possible to use an algorithm that performs better than others. The study shows that for high numerical values of both ratios all the docking computer programs produced poor results, for medium and medium high values of those ratios, Auto dock and DOCK were the best choice. However, with small values of the ratios all four programs GOLD, Surflex, DOCK and Auto dock produced agreement within 10% difference comparing RMSD of docked versus crystallographic ligand.

INTRODUCTION

Structure-based virtual screening has been thoroughly tested, with substantial success. The automated docking procedure is a vital part of virtual screening and is now a widely used technique in the early stages of drug discovery in most academic and commercial (Big Pharmacy) environments. Improvements in computer processing speeds and multiprocessing methods as well as distributed computational methods using multiple workstations permit en masse screening with storage of large quantities of data. This enables investigators to screen large numbers of compounds available in online libraries [1-3] by docking them into the binding pocket of the target enzyme. Concurrently, the experimental emergence of the 'high-throughput' automated X-ray crystallographic screening techniques has dramatically increased the rate at which

researchers can progress from the over-expression of a target protein to an inhibitor protein complex. The combination of high-throughput X-ray crystallography and molecular docking techniques allows investigators to work efficiently to design potent inhibitors to medically relevant enzyme targets. In automated docking procedures the success critically depends on the accuracy and precision of the process of in silico molecular docking. This, in turn, depends on the choice of software, the protein, how the binding site and search space is defined, and the ligands. The software choice may become the weak link in the process. Many other research groups [4-21] worked on comparisons, and improvement to correlate predicted docking results of ligands with experimental results (x-ray data). The problem of disagreement between virtual versus actual screening of small molecules for tight binding to proteins, remains an active area of investigation. Reliance on software

that is less than optimal in predictions continues to present an obstacle during the research process. This study has investigated whether geometric attributes of binding site pockets that can be defined quickly with on line tools; can be used as one of the several indicators to determine which program is optimal for screening small molecules for large scale docking. Here, the shape of binding pockets is evaluated as an independent indicator of docking performances of several molecular docking programs.

METHODS

One hundred different, unrelated structures were used, to achieve an accurate representation of a variety of proteins from the protein data bank (PDB) as previously described in the Astex study [22]. The three-dimensional protein structures were selected according by the following criteria: (1) The structure had a resolution higher than 2.5 Å; (2) The R factor was less than 0.20; (3) The difference between the R factor and the R free was less than 3% (low model bias); (4) more than 95% of the residues were in allowed regions of the Ramachandran plot, indicating good stereochemistry of the model [23] and the ligand had full occupancy and reasonable average B factor.

Defining Protein Binding Pockets

In order to identify pockets and calculate pocket volume and mouth area from the protein structure, the computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues (CASTp) [24] was used. The CASTp server uses the weighted Delaunay triangulation and the alpha complex method for quantifying the shape measurements of surface accessible pockets and interior cavities. The pocket can be defined by Voroni diagrams and characterized by Delaunay triangulation and alpha shape descriptors using Liang's method [24]. The Delaunay triangulation method sums together adjacent triangles and it is similar to cubic or rectangular definition of the binding pocket cavity. In contrast, alpha shape measurement utilizes concentric circles. Alpha shape approximates a spherical description of the binding pocket in three dimensions. Voroni and Delaunay shape descriptors utilize a method of calculation known as discrete flow, to identify and measure the pockets. In CASTp the molecular surface (ms) is defined as the total surface area and volume of the molecule which is void (cavity) and the pocket surface area and volume of the molecule which is considered solvent accessible (sa) [25]. Thus the ratio of the area or volume of the pocket that is solvent accessible to the total molecular surface area or volume ($A_{sa/ms}$ or $V_{sa/ms}$) yields the fraction or percent solvent accessible area or volume. These two attributes Asa/Ams and Vsa/Vms accurately predict the shape of the protein binding pocket. When the solvent accessible area or volume is low relative to total molecular surface, the protein binding site shape is highly encapsulated (Figure 1A). The converse is also true (Figure 1B) when the calculated surface accessible area or volume is high relative to total molecular surface, the protein binding pocket is more open with a wide mouth to the binding pocket.

The list of interacting pocket residues was matched against pocket data generated from the CASTp server to assign protein-ligand interactions to a pocket ID. The calculated characteristics including partitioning of volume and area of the pockets are

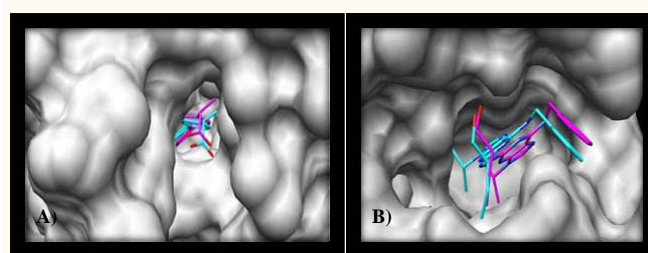


Figure 1 (A) Example of Low Solvent Accessible Surface Area/Volume Prostaglandin H₂ Synthase, (PDB code: 1Q4G) (Ligand: BFL: 2- (1, 1'-BIPHENYL-4-YL) PROPANOICACID). RMSD with Crystallographic Ligand is 1.10 Å with Dock6. (B) Example of High Solvent Accessible Surface Area/Volume Cyclic dependent kinase, (PDB code 1UNL) (Ligand: RRC R-Roscovotine). RMSD with Crystallographic Ligand is 1.10 Å with Auto dock 4.2.6.

shown in Supplemental Materials. Pockets that were identified to contain greater than 70% of the total interactions of the protein or two or more pockets in tandem with greater than 35% of the total interactions of the protein and a crystallographic ligand were used for this study. PDBs that had fewer than 70% of the protein-ligand interactions within the pocket described by CASTp were reevaluated with a similar volume calculation method using the CCP4 (Collaborative Computational Project, 1994) [25] program before they were removed from the study.

Preparation of Ligands

Manual evaluations of crystallographic ligands using the Ligand Explorer software were also performed to create text files of ligand-protein interactions to be compared to pockets created in CASTp. Upon further analysis, the failure in CASTp is due to three factors: (1) The binding pocket for the ligand is composed of two or more subunits of the protein and an incomplete or fragmented pocket is defined; (2) The binding pocket for the ligand is enclosed only on three sides and has a large solvent-accessible entry to the binding site (no mouth); and (3) The critical residues that are in the active site, are on mobile elements such as loops, have large conformational changes upon substrate/natural ligand binding, and are not properly positioned with substrate analogues to be included in the binding pocket. In summary, 4% of the proteins evaluated did not have assigned pockets for ligand binding sites and another 11% of the proteins evaluated had pockets but no detectable mouths to the pockets. Therefore 85 out of 100 PDB's originally selected are included in this study.

Once protein and ligand structures were prepared for computational docking, they were processed in four readily available docking programs: AutoDock4.2 [26], DOCK6.1 [27] and GOLD [28], and Surflex [29]. These four different programs were chosen because they use different approximations for treating the ligands. The Auto Dock 4 performs the docking of the ligand to a set of grids describing the target protein. The Dock algorithm (utilized by Dock6.1) uses a geometric matching system to place a ligand in the "negative image" of the binding pocket. GOLD calculates docking modes of small molecules in protein binding sites and offers full ligand flexibility. In addition, GOLD has improved handling and control of bound water molecules and metal coordination geometries. The Surflex program works by

fragmenting the ligand into stable functional groups to be fit into the binding pocket of the protein (protomol).

Ligand Evaluation

The results and docked ligands were compared with the actual crystallographic ligand. Root mean square deviation (RMSD) of the two molecules was the main criterion for comparison between the docked and actual ligand conformations. The RMSD was calculated in the Quanta software and confirmed in Qmol [30]. Since the binding strength between the protein and the ligand is determined by electrostatic, hydrophilic and hydrophobic interactions, X-Score [31,32], a primary function was used as a common “yardstick” to compare binding energies of top scoring ligands. X-Score was utilized as a common unit of measurement for all four docking programs and is reported in the Supplemental Materials.

RESULTS AND DISCUSSION

Analysis shows that the binding sites are usually located at large indentations (pockets) on the surface of the proteins. Once the binding pocket is identified through the method described in previous section, a detailed characterization of the pocket shape such as volumes and areas of molecular surface and solvent accessible areas were obtained. The CASTp parameters associated with such an arrangement of atoms on the surface of the protein, and found V_{sa}/ms , a ratio of the volume of solvent accessible to volume of total molecular surface, a useful descriptor for binding sites associated with the effectiveness of docking. We also identified the similar parameter Asa/ms , a ratio of area solvent accessible surface to area of total molecular surface. The V_{sa}/ms and Asa/ms ratios of the pocket and mouth, respectively, were analyzed in clusters of 0.5 Å as ligand docking results within this distance are considered equivalent. A. In the supporting material, the V_{sa}/ms and Asa/ms values are shown with RMSDs for each of the four programs. A frequency count for RMSD in 0.5 Å increments is shown in Table 1. This organization was chosen because Auto dock output files group ligand conformations with 0.5 Å as equivalent. Both AutoDock4 and DOCK6.1 showed the highest frequency of success in the RMSD range of 0.5-1.5 Å. More specifically, out of 85 PDB's tested; Auto dock produced 33 structures which had a RMSD 1.5 Å or less from the crystallographic position of the ligand. Dock6 followed closely with 31 structures, and then Surflex with 18 and GOLD with only 13 structures with an RMSD of 1.5 Å or less from the crystallographic position of the ligand (Table 1).

The data presented in Supplemental Materials shows that many ligand-bound protein pockets tend to have a relatively large molecular surface volume relative to the total solvent-accessible volume. This is consistent with Liang's computational method of differentiating binding pockets from indentations on the protein surface. Neither GOLD nor Surflex showed a clear correlation of docking performance with shape descriptor ratios, and overall both programs had lower frequency of success than either Autodock4 or DOCK6.1. Pockets that have a high V_{sa}/ms ratio are more similar to invaginations or indentations and thus perform poorly for predictive automated docking for all four programs. This may be due to the fact that unless specific water molecules are modeled, the low total number of interactions results in

a lower predicted binding in the absence of the mathematical consideration of induced fit. In contrast, when the V_{sa}/ms ratio is low, the pocket is more like a channel for small ions than a pocket for ligand binding. As shown in (Figure 1A) the performance of DOCK 6.1 was acceptable with highly encapsulating docking site (low V_{sa}/ms), whereas in Figure 1B Auto dock 4.2 performed well in a more open docking site (high V_{sa}/ms). (Figure 2A and 2B) show the negative and positive correlation Auto dock and Dock6 docking performance has with percent solvent accessible volume. Neither Surflex nor GOLD showed positive or negative correlation (See Supplemental Material). Overall, Dock6 showed the highest correlation of docking with shape descriptors as the R^2 of a linear regression of (Figure 2B) was ~ 0.72 .

CONCLUSION

Drug development studies focus on inhibitor binding to protein targets. Many Drug companies use molecular docking to find lead compounds to initiate drug discovery process. As much

Table 1: Frequency of Docking Hits by RMSD Cluster.

RMSD Cluster (Å ²)	Auto dock	Dock6	GOLD	Surflex
0-0.5	1	3	0	1
0.5-1.0	12	11	4	6
1.0-1.5	20	17	9	11
1.5-2.0	11	11	10	9
2.0-2.5	7	10	8	5
2.5-3.0	9	8	11	8
3.0-3.5	14	13	12	18
3.5-4.0	10	8	15	10
4.0-4.5	1	3	5	6
4.5-5.0	0	1	4	2
RMSD > 5.0	0	0	7	9

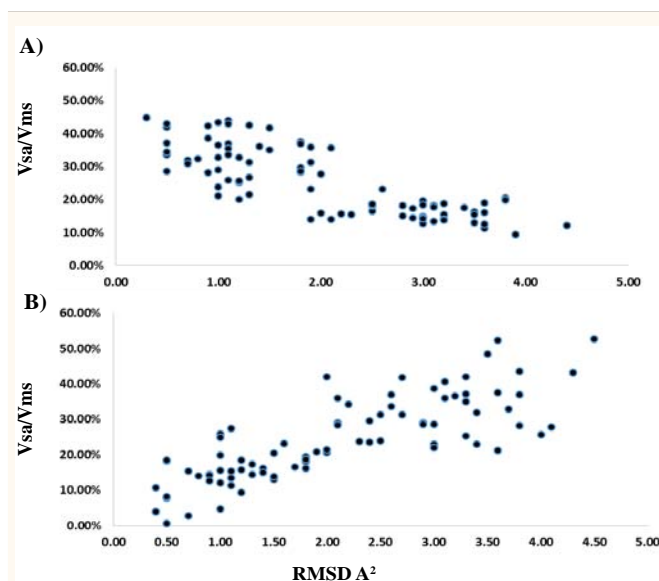


Figure 2 Auto dock (A), Dock6 (B) Shape Descriptors: The Correlation between V_{sa}/V_{ms} (%) of 85 PDB's and RMSD from Crystallographic Ligand.

as possible, an accurate starting point for either synthesized or purchased compounds is necessary. Existing databases of small molecule compounds have thousands/millions of commercially available compound to test their potency against every disease imaginable. These databases are mirrored by the in silico databases (Zinc database is one which has Sigma Aldrich catalogue ~ 3 million compounds) that have 3D coordinates of ligands/inhibitors. The docking programs have limited capacity for careful testing by docking of available ligands/inhibitors. We presented here a procedure to limit number of docking experiments and increase their success by investigating the dependence of this success on the shape of the protein target. When large numbers of small organic molecules are going to be screened to make sure the best docking program is selected, protocols such as the ones described here, need to be developed to choose the best strategy to perform optimally on a particular binding site. This study provides a recipe, a protocol to choose the best performing docking program based on the shape of the binding site.

In summary of the results, the detection of the shape of the docking site is a complex problem for which there is no comprehensive solution providing a simple geometric shape descriptor. Both descriptors (area and volume) proved to be useful in discriminating different programs in their success rate in the entire descriptor space. This result holds even when descriptors are sometimes significantly correlated. The use of a single descriptor offers significant advantages in proper selection of the program for maximum effectiveness especially when using a two-staged approach; first by computing the geometric descriptor of the docking site and then by selecting the most efficient docking algorithm for this particular range of the descriptors.

ACKNOWLEDGMENT

The authors thank the developers of all the programs used during the preparation of this paper, especially for the trial version of Surflex-Dock™ 2.0.

REFERENCES

- Irwin JJ, Shoichet BK. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model.* 2005; 45: 177-182.
- Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 1998; 11: 739-747.
- Kellenberger E, Rodrigo J, Muller P, Rognan D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins.* 2004; 57: 225-242.
- Kontoyianni M, McClellan LM, Sokol GS. Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem.* 2004; 47: 558-565.
- Kontoyianni M, Sokol GS, McClellan LM. Evaluation of library ranking efficacy in virtual screening. *J Comput Chem.* 2005; 26: 11-22.
- Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins.* 2004; 56: 235-249.
- Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, et al. A critical assessment of docking programs and scoring functions. *J Med Chem.* 2006; 49: 5912-5931.
- Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J Med Chem.* 2004; 47: 45-55.
- Zavodszky MI, Rohatgi A, Van Voorst JR, Yan H, Kuhn LA. Scoring ligand similarity in structure-based virtual screening. *J Mol Recognit.* 2009; 22: 280-292.
- Bursulaya BD, Totrov M, Abagyan R, Brooks CL 3rd. Comparative study of several algorithms for flexible ligand docking. *J Comput Aided Mol Des.* 2003; 17: 755-763.
- Nissink JW, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein-ligand interaction. *Proteins.* 2002; 49: 457-471.
- Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem.* 2000; 43: 4759-4767.
- Pérez C, Ortiz AR. Evaluation of docking functions for protein-ligand docking. *J Med Chem.* 2001; 44: 3768-3785.
- Michal Vieth, Jonathan D Hirst, Brian N Dominy, Heidi Daigler, Charles L Brooks III. Assessing search strategies for flexible docking. *J Comput Chem.* 1998; 19: 1623-1631.
- Ha S, Andreani R, Robbins A, Muegge I. Evaluation of docking/scoring approaches: a comparative study based on MMP3 inhibitors. *J Comput Aided Mol Des.* 2000; 14: 435-448.
- Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J Mol Model.* 2003; 9: 47-57.
- Merlitz H, Wenzel W. Comparison of stochastic optimization methods for receptor-ligand docking. *Chem Phys Letters.* 2002; 362: 271-277.
- Mc Conkey B, Sobolev V, Edelman M. The performance of current methods in ligand-protein docking. *Current Sci.* 2002; 83: 845-856.
- Cummings MD, DesJarlais RL, Gibbs AC, Mohan V, Jaeger EP. Comparison of automated docking programs as virtual screening tools. *J Med Chem.* 2005; 48: 962-976.
- Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mater Des.* 2002; 16: 11-26
- Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WT, Mortenson PN, et al. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem.* 2007; 50: 726-741.
- Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. *J Mol Biol.* 1963; 7: 95-99.
- Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* 2006; 34: 116-118.
- Binkowski TA, Naghibzadeh S, Liang J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.* 2003; 31: 3352-3355.
- CCP4. Free Software Foundation. 1999.
- Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, et al. Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. *Com Chem.* 1998; 19: 1639-1662.
- Shoichet BK, Kuntz ID. Protein docking and complementarity. *J Mol*

- Biol. 1991; 221: 327-346.
28. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. *Proteins*. 2003; 52: 609-623.
29. Surflex-Dock™ 2.0 Tripos International. 1699.
30. Gans JD, Shalloway D. Qmol: a program for molecular visualization on Windows-based PCs. *J Mol Graph Model*. 2001; 19: 557-609.
31. Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem*. 2003; 46: 2287-2303.
32. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des*. 2002; 16: 11-26.

Cite this article

Keyes RM, Pejo E, Katagiri K, Huynh K, Rudnitskaya A, et al. (2016) Shape Matters: Improving Docking Results by Prior Analysis of Geometric Attributes of Binding Sites. *JSM Chem* 4(1): 1020.