

Binding Site Druggability Assessment in Fragment-Based Drug Design

Yu Zhou and Niu Huang

Abstract

Target druggability refers to the propensity that a particular target is amenable to bind high-affinity drug-like molecules. A robust yet accurate computational assessment of target druggability would greatly benefit the fields of chemical genomics and drug discovery. Here, we illustrate a structure-based computational protocol to quantitatively assess the target binding-site druggability via *in silico* screening a fragment-like compound library. In particular, we provide guidelines, suggestions, and critical thoughts on different aspects of this computational protocol, including: construction of fragment library, preparation of target structure, *in silico* fragment screening, and analysis of druggability.

Key words Druggability assessment, Fragment screening, Molecular docking, MM-GB/SA rescoring, Hit rate

1 Introduction

Successful drug development requires a disease target of both biological relevance and chemical tractability. With the completion of the human genome, we now have unprecedented access to large numbers of potential therapeutic targets. The question that arises is which specific protein targets can be modulated by a drug-like molecule. Druggability (i.e., propensity that a particular target is amenable to bind high-affinity drug-like molecules) assessment in the process of target selection would reduce drug discovery attrition and put effort on those targets most likely to lead to therapeutic intervention [1].

The first step in evaluating the druggability of a target is to identify the presence of binding pockets with suitable size, shape, and composition to accommodate drug-like molecules. Many approaches for this purpose have been developed that are generally classified as geometry-based [2–5], information-based [6, 7], and energy-based algorithms [8, 9]. Benchmarking studies using

training set data extracted from the Protein Data Bank (PDB), most approaches have demonstrated to correctly detect the true ligand-binding sites. However, the presence of a suitable protein pocket is necessary but not sufficient to guarantee potent binding of drug-like small molecules.

The more difficult step is to quantitatively predict the druggability index of a given binding site. Early studies have predicted target druggability on the basis of sequence and structure homology to known drug targets [10, 11]. However, not all members of the same protein family are equally druggable [12]. More importantly, such methods cannot be used to assess druggability of novel target families. Recently several structure-based target druggability methods have been developed and validated against a set of reference targets where the degree of tractability is known. These methods provide quantitative assessments of druggability using physicochemical descriptors derived from the ligand binding pockets and apply techniques as varied as biophysical modeling [13], linear regression [14, 15], and support vector machines [16].

Hajduk et al. made a seminal contribution by demonstrating that experimental hit rates from the heteronuclear-NMR-based fragment screening could serve as an effective druggability index within a set of 23 protein targets containing 28 different binding sites [17]. Furthermore, they derived a linear regression model to fit the experimentally measured hit rates to physicochemical descriptors of these 28 binding pockets. Applying an appropriate cutoff, this model was assessed using an additionally assembled binding-site dataset, and 33 out of 35 known drug-like ligand-binding sites were correctly identified. Being essentially analog to the NMR-based fragment screening, an *in silico* fragment screening protocol was also developed to assess target binding-site druggability [18]. It makes use of a molecular mechanics-based scoring method for the protein–ligand interaction and the obtained virtual hit rates were demonstrated to correlate with the hits rate measured experimentally from the NMR-based screening method. This protocol can be employed to distinguish known druggable and non-druggable targets, and it is generally applicable without relying on any assembled training data set that potentially extends its capacity toward unexplored target space.

In this chapter, we illustrate the computational details of this *in silico* fragment screening protocol for target druggability assessment (*see* Fig. 1 for a schematic overview). We outline the criteria for the construction of fragment library, discuss the method for the preparation of target structure, and describe the procedure for carrying out the *in silico* fragment screening. Finally, we discuss the druggability analysis from the virtual screening results.

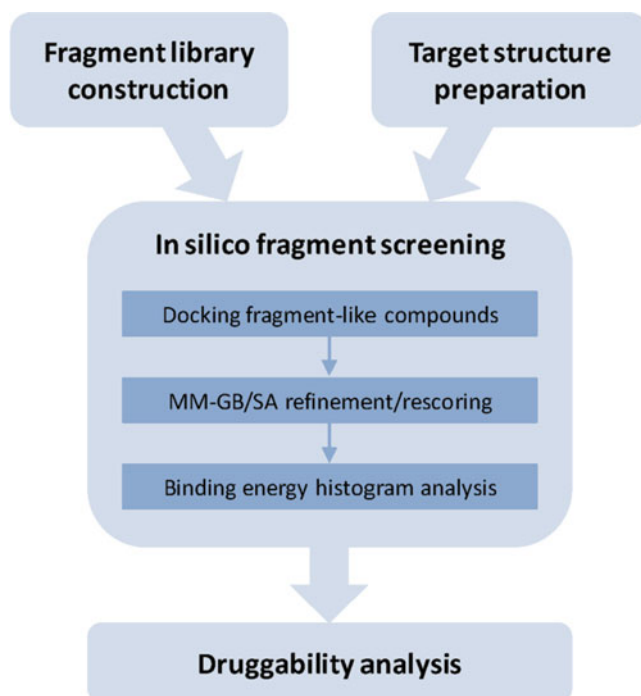


Fig. 1 Schematic illustration of druggability prediction via fragment-based docking and scoring approach

2 Materials

The druggability assessment protocol entails building a fragment-like compound library and performing *in silico* fragment screening experiments, which could be carried out by means of a variety of Web servers and software. The programs listed here are merely the ones used as examples for illustrating this procedure. The diverse set of fragments is selected from the fragment-like subset of the ZINC database [19, 20]. The DOCK 3.5.54 program [21, 22] is used to dock the fragment database into the protein binding site. The Protein Local Optimization Program (PLOP) [23–25] is used to perform MM-GB/SA refinement and rescoring.

3 Methods

3.1 Fragment Library Construction

1. Extract compounds from the fragment-like subset of the ZINC database (*see Note 1*).
2. Eliminate fragments with more than 15 heavy atoms (*see Note 2*).

3. Calculate feature key fingerprints using CACTVS [26], and perform the fingerprint-based similarity analysis with a modified version of the program SUBSET [27] to reduce redundancy of the fragment library (*see Note 3*).

3.2 Target Structure Preparation

1. Select one or more representative structures for the protein target (*see Note 4*).
2. Determine the ligand binding pocket (*see Note 5*). Identify cofactors, metal ions, and structural waters in the target protein and treat them as part of the protein if they are involved in ligand binding.
3. Add hydrogen atoms to the protein. Assign proper protonation states for binding-site residues and optimize the orientations for polar hydrogen atoms using PLOP (*see Note 6*).

3.3 In Silico Fragment Screening

The in silico screening protocol employs a physics-based hierarchical scoring method which consists of two steps: predicting the binding poses of ligands using a docking program, and then refining and rescoring those protein–ligand complexes using a more computationally intensive molecular-mechanics based energy function [28, 29]. This protocol uses a high-throughput docking program to initially orient and score the ZINC fragment-like compounds in the binding site, and subjects the best single docking pose for each docked compound to a rescoring stage in which the ligand is fully minimized inside the binding site and the binding energy is estimated with an all-atom molecular mechanics force field combined with an implicit solvent model. Finally the results of all compounds are analyzed based on the binding energy distribution.

3.3.1 Docking Fragment-Like Compounds Library

1. Identify binding site residues within a certain range (e.g., 12 Å) away from any heavy atom of the crystallographic ligand or the residues used to define the site, using the program FILT (part of the UCSF DOCK suite).
2. Calculate the solvent-accessible molecular surface [30] of the protein binding site with the program DMS [31] using a probe radius of 1.4 Å.
3. Generate receptor-derived spheres with the program SPHGEN (part of the UCSF DOCK suite) [32], in combination with the ligand-derived spheres if necessary (*see Note 7*).
4. Set the grid box dimensions with edges 15 Å beyond the matching spheres initially. Then refine the box dimensions to maximize the coverage of the protein without exceeding 2 million grid points at a predefined grid resolution (three points per angstrom by default). Finally, four scoring grids are generated: an excluded volume grid using DISTMAP [33], a united

atom AMBER-based van der Waals potential grid using CHEMGRID [33], an electrostatic potential grid using DelPhi [34], and a solvent occlusion map using the program SOLVMAP [35].

5. Perform docking with DOCK 3.5.54, a flexible-ligand method that uses a force-field-based scoring function. Ligand conformations are scored on the basis of the total docking energy ($E_{\text{tot}} = E_{\text{ele}} + E_{\text{vdw}} - \Delta G_{\text{lig-solv}}$), which is the sum of electrostatic (E_{ele}) and van der Waals interaction energies (E_{vdw}), corrected by the partial ligand desolvation energy ($\Delta G_{\text{lig-solv}}$).
6. Save a single docking pose with the best total energy score for each docked molecule for the next stage of scoring (*see Note 8*).

3.3.2 MM-GB/SA Refinement and Rescoring

1. Generate OPLS force field parameter for each molecular compound and cofactor (if present), using IMPACT (part of the Schrödinger suite).
2. Submit the free ligand, free protein and docked protein–ligand complex to multi-scale Truncated Newton (MSTN) energy minimization [25] in all-atom OPLS force field [36, 37] and Generalized Born (GB) solvent [38, 39] using PLOP (*see Note 9*).
3. Calculate the binding energy ($E_{\text{bind}} = E^{\text{RL}} - E^{\text{L}} - E^{\text{R}}$) by subtracting the energies of the optimized free ligand in solution (E^{L}) and the free protein in solution (E^{R}) from the optimized protein–ligand complex’s energy in solution (E^{RL}) (*see Note 10*).

3.3.3 Histogram Analysis of Energy Score

1. Report the energy scores distribution for the protein target.

3.4 Druggability Analysis

1. Compute the “hit rate” for the in silico screening based on a chosen energy cutoff value (-40 kcal/mol) (*see Note 11*).
2. Calculate the druggability score which is defined as $\log(\text{hit rate})$.
3. Compare the druggability score with the cutoff value of 0.36 to classify the assessed target as druggable or non-druggable (*see Note 12*).

4 Notes

1. Fragments are molecules of low complexity, which sample chemical space exponentially more effectively than drug-sized molecules. Different estimates exist of the size of chemical space. Here, the fragment-like subset of the ZINC database (version 6, December 2005) contains 49,134 compounds with relatively low molecular weight ($\text{MW} \leq 250$), few rotatable

bonds ($RB < 3$), low hydrophobicity ($-2 < \log P < 3$), and weak hydrogen bonding potentials ($HB_{\text{donor}} < 3$ and $HB_{\text{acceptor}} < 6$).

2. Kuntz et al. observed that the maximal binding free energy increases more slowly for ligands containing more than 15 heavy atoms [40]. Therefore, fragments with more than 15 heavy atoms were eliminated. This filter reduced the library size to 32,717 molecules.
3. Representative structures were selected for each structural cluster with Tanimoto coefficient (T_c) less than 0.9 to other clusters. This further reduced the library to 11,129 diverse molecules. To assess any potential bias resulting from the diversity-based filtering, redo the screening using 32,717 ZINC fragment-like compounds for the training dataset, leads to very similar energy distributions.
4. Targets may have multiple crystal complex structures available and some display significant side-chain movement upon binding to different ligands [41]. In most cases, we found that the changes of the histograms of energy scores and the druggability scores calculated from them are remarkably small when using different crystal structures. Nevertheless, multiple conformations are recommended for the binding sites with large structural variation, especially for the protein–protein interaction (PPI) interfaces. Applying our protocol, specific druggable conformations could also be identified.
5. The identification of the protein binding pocket is straightforward for ligand-bound complex structures. However, the binding site is not known from a 3-D structure or from other experimental data, a “suitable” pocket is required to be detected firstly by pocket detection programs or virtual inspection.
6. Ideally, the target protein should be prepared as if the crystal ligand was absent, as adjusting the protein to favor crystal ligands is a source of bias.
7. Spheres are generated to fill the binding site. Matching spheres required for the orientation of the ligand within the binding site are obtained by augmenting the ligand-derived spheres with receptor-derived spheres. By default, spheres furthest away from ligand-derived spheres, furthest from the centroid of the remaining spheres, too close to receptor atoms, or too close to each other are removed iteratively until the total number of sphere is 35 or less. However, for large binding surfaces like protein–protein interfaces, we use a maximum of 120 matching spheres to ensure adequate ligand sampling.
8. One major limitation of the current protocol is that it relies entirely on the docking algorithm to identify the correct binding pose. A simple extension of this protocol is to subject

a small number of dissimilar binding poses to minimization in the MM-GB/SA rescoring step and use the most favorable binding energy for rank-ordering ligands. Therefore, multiple (usually hundreds of) docking poses could be saved in docking stage and subjected for structural descriptor-based filtering and KGS-penalty function-based conformational clustering [42]. Tens of poses might be finally obtained for next MM-GB/SA rescoring.

9. The molecular mechanics forces are divided into short-range (bond, angle, torsion, and local non-bonded) and long-range components, with the long-range forces updated only intermittently. The algorithm is also optimized for minimizations with GB solvent that increases the computational expense by only a factor of ~ 3 relative to the vacuum. Thus, this scoring approach accounts for accurate and efficient calculations of ligand-protein interaction energies, the ligand/receptor desolvation, and to a lesser extent, ligand strain energies. In this work, the protein was kept rigid during protein-ligand minimization to reduce the computational expense.
10. Accurate free energy calculations depend on a proper balance of many different energetic components. The MM-GB/SA rescoring method strikes a balance between computational speed and accuracy, and in particular neglects entropic loss and protein flexibility. Empirically scaling certain energy components as a post-rescoring process, in a manner similar to LIE scheme, may be useful to compensate for some of these limitations [43]. It has been suggested that the MM-GB/SA scoring function underestimates the nonpolar binding contributions to the free energy of binding [28]. In this study, we empirically scaled the van der Waals energy component by a factor of 2.
11. This cutoff value was empirically chosen to maximally differentiate druggable and non-druggable binding site. We visually inspected the energy distributions for the 13 druggable binding sites and 11 non-druggable binding sites in Hajduk et al. training data set and explored the effect of varying the cutoff with respect to differentiating between druggable and non-druggable binding sites. We found the correlation between the docking screening hit rates and the NMR screening results is relatively insensitive to the value of the energy cutoff within a certain range (from -40 to -34 kcal/mol). In this work, an energy cutoff of -40 kcal/mol was used for computing the in silico hit rate.
12. The calculated druggability scores correlate reasonably well with the NMR-based fragment screening results. Hajduk et al. defined binding sites as “highly druggable” if they have a experimental $\log(\text{hit rate}) > -1.0$. The corresponding value of computational $\log(\text{hit rate})$ is 0.36, and we used this value to

classify proteins as druggable or non-druggable in this work. Although Hajduk et al. distinguish between “highly druggable” and “moderately druggable,” we use a simple binary classification for simplicity. Nevertheless, the higher druggability score a target is assigned, the more druggable it might be.

Acknowledgement

The Chinese Ministry of Science and Technology “973” Grant 2011CB812402 (to N.H.) is acknowledged for financial support, Shoichet Lab at UCSF for the DOCK3.5.54 program and Jacobson Lab at UCSF for PLOP.

References

1. Fauman EB, Rai BK, Huang ES (2011) Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol* 15:463–468
2. Brady GP Jr, Stouten PF (2000) Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* 14:383–401
3. Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model* 15(359–363):389
4. Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 13(323–330):307–328
5. Liang J, Edelsbrunner H, Woodward C (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* 7:1884–1897
6. Soga S, Shirai H, Kobori M, Hirayama N (2007) Use of amino acid composition to predict ligand-binding sites. *J Chem Inf Model* 47:400–406
7. Stuart AC, Ilyin VA, Sali A (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. *Bioinformatics* 18:200–201
8. An J, Totrov M, Abagyan R (2004) Comprehensive identification of “druggable” protein ligand binding sites. *Genome Inform* 15:31–41
9. Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. *Proteins* 62:479–488
10. Hopkins AL, Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* 1:727–730
11. Blundell TL, Sibanda BL, Montalvo RW, Brewerton S, Chelliah V, Worth CL, Harmer NJ, Davies O, Burke D (2006) Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. *Philos Trans R Soc Lond B Biol Sci* 361:413–423
12. Fauman EB, Hopkins AL, Groom CR (2003) Structural bioinformatics in drug discovery. *Methods Biochem Anal* 44:477–497
13. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 25:71–75
14. Sheridan RP, Maiorov VN, Holloway MK, Cornell WD, Gao YD (2010) Drug-like density: a method of quantifying the “bindability” of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *J Chem Inf Model* 50:2029–2040
15. Halgren TA (2009) Identifying and characterizing binding sites and assessing druggability. *J Chem Inf Model* 49:377–389
16. Nayal M, Honig B (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins* 63:892–906
17. Hajduk PJ, Huth JR, Fesik SW (2005) Druggability indices for protein targets derived from NMR-based screening data. *J Med Chem* 48:2518–2525
18. Huang N, Jacobson MP (2010) Binding-site assessment by virtual fragment screening. *PLoS One* 5:e10109

19. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45:177–182
20. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* 52:1757–1768
21. Lorber DM, Shoichet BK (2005) Hierarchical docking of databases of multiple ligand conformations. *Curr Top Med Chem* 5:739–749
22. Wei BQ, Baase WA, Weaver LH, Matthews BW, Shoichet BK (2002) A model binding site for testing scoring functions in molecular docking. *J Mol Biol* 322:339–355
23. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS (2002) Force field validation using protein side chain prediction. *J Phys Chem B* 106:11673–11680
24. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA (2004) A hierarchical approach to all-atom protein loop prediction. *Proteins* 55:351–367
25. Zhu K, Shirts MR, Friesner RA, Jacobson MP (2007) Multiscale optimization of a truncated newton minimization algorithm and application to proteins and protein-ligand complexes. *J Chem Theory Comput* 3:640–648
26. Ihlenfeldt WD, Takahashi Y, Abe S, Sasaki S (1994) Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and flexibility. *J Chem Inf Comput Sci* 34:109–116
27. Voigt JH, Bienfait B, Wang S, Nicklaus MC (2001) Comparison of the NCI open database with seven large chemical structural databases. *J Chem Inf Comput Sci* 41:702–712
28. Huang N, Kalyanaraman C, Bernacki K, Jacobson MP (2006) Molecular mechanics methods for predicting protein-ligand binding. *Phys Chem Chem Phys* 8(44):5166–5177
29. Huang N, Kalyanaraman C, Irwin JJ, Jacobson MP (2006) Physics-based scoring of protein-ligand complexes: enrichment of known inhibitors in large-scale virtual screening. *J Chem Inf Model* 46:243–253
30. Connolly ML (1983) Solvent-accessible surfaces of proteins and nucleic acids. *Science* 221:709–713
31. Ferrini TE, Huang CC, Jarvis LE, Roberts L (1988) The MIDAS display system. *J Mol Graph* 6:13–27
32. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 161:269–288
33. Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. *J Comput Chem* 13:505–524
34. Nicholls A, Honig B (1991) A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. *J Comput Chem* 12:435–445
35. Mysinger MM, Shoichet BK (2010) Rapid context-dependent ligand desolvation in molecular docking. *J Chem Inf Model* 50:1561–1573
36. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 118:11225–11236
37. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J Phys Chem B* 105:6474–6487
38. Gallicchio E, Zhang LY, Levy RM (2002) The SGB/NP hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *J Comput Chem* 23:517–529
39. Ghosh A, Rapp CS, Friesner RA (1998) Generalized born model based on a surface integral formulation. *J Phys Chem B* 102:10983–10990
40. Kuntz ID, Chen K, Sharp KA, Kollman PA (1999) The maximal affinity of ligands. *Proc Natl Acad Sci U S A* 96:9997–10002
41. Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* 49:534–553
42. Peng SM, Zhou Y, Huang N (2013) Improving the accuracy of pose prediction in molecular docking via structural filtering and conformational clustering. *Chin Chem Lett* 24:1001–1004
43. Zhou Z, Madura JD (2004) CoMFA 3D-QSAR analysis of HIV-1 RT non nucleoside inhibitors, TIBO derivatives based on docking conformation and alignment. *J Chem Inf Comput Sci* 44:2167–2178



<http://www.springer.com/978-1-4939-2485-1>

Fragment-Based Methods in Drug Discovery

Klon, A.E. (Ed.)

2015, IX, 230 p. 68 illus., 53 illus. in color., Hardcover

ISBN: 978-1-4939-2485-1

A product of Humana Press