

Application of a Bioinformatics Approach to High-throughput Docking for Drug Discovery

Shuxing Zhang, Lei Du-Cuny

Abstract—we have developed a high-throughput docking (HTD)-based virtual screening scheme, termed HiPCDock, for drug discovery and development. To improve the statistical significance of our screening results, a bioinformatics approach, motivated by sequence alignment package BLAST, was implemented so that we can estimate the confidence of our prediction accuracy. The model was validated by docking ten known thymidine kinase (TK) binders into the enzyme and the real inhibitors showed significant statistics of low probability and expectation values compared to those random compounds extracted from ChemBank. It was also demonstrated that HiPCDock was able to efficiently recover the ten known TK inhibitors from a dataset including 990 decoy compounds randomly selected from ChemBank. Our HiPCDock is currently implemented based on only one configuration template file for job submission, which makes it very easy to use for both computational experts and experimental scientists who have little molecular docking experience. With just one command line, users can submit massive parallel docking jobs for their screening of millions of compounds against a specific target of interest. The statistical model will be very useful to guide the decision-making for lead discovery and optimization. Thus HiPCDock is an elegant, professional integrated package for high-throughput molecular docking and drug discovery. We are currently developing a web-based user interface to further simplify the process.

Index Terms—extreme value distribution, high-performance computing-based docking, virtual screening

Manuscript submitted on August 9, 2007. This work was supported in part by M.D. Anderson Cancer Center Faculty Startup Funding.

S. Zhang is with the Department of Experimental Therapeutics, Unit 36, M. D. Anderson Cancer Center, Houston, TX 77030 USA (corresponding author to provide phone: 713-745-2958; fax: 713-794-5577; e-mail: shuzhang@mdanderson.org).

L. Du-Cuny is with the Department of Experimental Therapeutics, Unit 36, M. D. Anderson Cancer Center, Houston, TX 77030 USA (e-mail: lducuny@mdanderson.org).

I. INTRODUCTION

High-throughput docking (HTD) is one of the most important approaches among structure-based virtual screening methods for hit/lead discovery in both pharmaceutical industry and academia [1]. Several success stories have been summarized recently and nanomolar binders were developed after the optimization of the original identified structures [1]. However, accurate and fast prediction of binding free energy of ligands to receptors is still a very challenging task [2]. Most of the existing scoring functions in molecular docking are over simplified on the one hand. On the other hand, we still rely heavily on these schemes to rank compounds or determine how active they are [2]. As is known, different scoring functions usually give different results of ranking or estimates of binding affinities, and this makes the decision-making difficult during the step of hit selection [2].

To tackle this problem, and motivated by BLAST [3], we propose a new bioinformatics approach widely used in sequence alignment where the accurate statistical estimates for sequence similarity scores makes the homologue identification universal and statistically meaningful [3]. Such a statistical model has been used to evaluate the scores obtained during docking of very large chemical databases. It enabled us to rank or select potential drug candidates based on the statistical significance of the prediction instead of using the absolute docking score values. In addition, the statistical estimate can be used to compare results from the screening of different databases or with different scoring functions.

II. METHODS AND MATERIALS

1. Preparation of the Receptor and Ligands

The target of our interest is thymidine kinase, which is a key enzyme of the salvage pathway and important for antiviral and/or cytostatic treatment

[4]. The enzyme structure (1KIM [4]) was obtained from Protein Data Bank (PDB) [5]. The original ligand and waters were removed from the protein. All hydrogens were loaded to the enzyme and the lone pairs were subjected to removal.

The chemical compounds were downloaded from ChemBank [6] (as of 05/22/2007). In total there were 345,351 2D structures in sdf format. After removing salts, the rest were subject to the conversion to 3D structures and energy minimization after the addition of hydrogens. This process was performed with MOE (Chemical Computing Group, Montreal, Quebec, Canada). The molecules were then converted to PDBQ files using a standalone python program of AutoDockTools (ADT) [7]. The final pdbq files were made ready for our HiPCDock-based virtual screening. Currently this sdf-to-pdbq conversion was automated in our package, as demonstrated in Figure 1 which will be described in the next section. Therefore this enables user to directly use mol2 or sdf files as their input database.

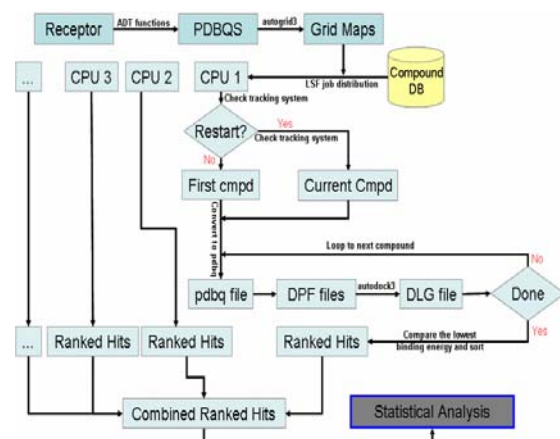


Figure 1. The working flowchart of HiPCDock, a package used for the massive parallel docking-based virtual screening of very large chemical databases as well as for lead optimization in drug development.

2. Automated HiPCDock Approach

Molecular docking is a multi-step process. In addition to the above protein target preparation and ligand database curation, it is also involved in conformational sampling and scoring. The final step is to collect the docked results for the hit identification analysis and visualization. The overall workflow for this HiPCDock is shown in Figure 1. Autodock3 [8] and several other packages such as OpenBabel (<http://openbabel.sourceforge.net>) have

been wrapped into the program. The final hit compounds are collected in a single hit list file.

In each docking session all input parameters are specified via a configuration file and could be easily modified by users with any text editors (e.g., vi). All calculations are fully integrated on our Linux cluster Load Sharing Facility (LSF) queuing system (<http://www.platform.com>) for parallel massive molecular docking job submission and management. This removes many of the complexities associated with large scale high-throughput virtual screening and provides a convenient way for users to take advantage of high-performance computing (HPC) clusters. In addition to LSF, the package currently has also been made compatible with both Portable Batch System (PSB, <http://www.openpbs.org>) and Sun Grid Engine (<http://gridengine.sunsource.net>) queuing system. The job array function is used for the job distribution and scheduling. Once a CPU is available, HiPCDock distributes a job on that CPU and starts docking. Otherwise, it is pending in the queue.

Currently ten protein atom types were used for proteins to generate the grid parameter file (.gpf), including carbon (C), nitrogen (N), oxygen (O), sulfur (S), hydrogen (H), metal (M), phosphorus (P), Zinc (Z), Calcium (L) and X for unknown type. One of the characteristic features of the original AutoDock3 [8] is that the atom types absent in the input ligand are excluded in the calculation of potential grids. This is because the original AutoDock3 [8] was developed for the purpose of precise binding mode analysis of a single ligand instead of database screening. Therefore, it becomes a very inefficient step in virtual screening with the time-consuming 3D grid pre-calculation of interaction energy for each ligand with the original AutoDock3 [8]. In order to overcome this shortcoming, we calculated the 3D grids of interaction energy for all possible atom types at one time. These uniquely defined ligand atom types include non-aromatic carbon (C), aromatic carbon (A), nitrogen (N), oxygen (O), sulfur (S), phosphorus (P), hydrogen (H), metal (M), fluorine (F), chlorine (c), bromine (b), iodine (I), zinc (Z), calcium (L), iron (f) and unknown type (X). They basically cover most of the possible ligand atom types included in databases. As for the center of the common grids, it can be either the center of mass coordinates of the ligand that had been removed

from the binding site of the target protein under consideration, or the geometrical center of a series of key residues provided by users. A script was coded for the generation of the GPF file with our customized atom types and the parameter values provided by users. Based on the GPF file, AutoDock3 [8] will create 16 atom type maps, plus an additional electron density map.

The docking parameter file (.dpf) is automatically generated by HiPCDock based on the users' input. After docking each molecule, the result in its docking log file (.dlg) is analyzed and the lowest estimated binding free energy is recorded. This is used for the comparison with other molecules to determine whether this molecule is a stronger binder or not. If yes, its docking log file is kept, otherwise its related files are deleted in order to save disk space. Since this is the most time-consuming part, a restart function is implemented to improve the robustness of the program. Basically each successfully processed molecule is recorded in a tracking file. Every time HiPCDock runs, it first checks the tracking file and starts from the molecule where the last run was stopped.

Once the jobs on all CPUs are done, HiPCDock post-process starts to analyze the results. It collects all of the individual hit lists together and generates an overall hit list. The list is sorted according to their binding free energy and the top ranked compounds, as requested by users (e.g., 10% of all database compounds), are selected as the final hits. These hits can be further refined by chemists' knowledge combined with molecular visualization. In addition, OpenBabel is utilized to calculate some molecular properties, such as molecular weight and number of rings in the molecules.

3. *Virtual Screening of ChemBank Collection and Statistical Model Generation*

As molecular docking is an optimization process to find the best fit of a ligand to its receptor, theoretically the predicted binding for each compound should follow an extreme value distribution (minimum EVD) for random data. Based on this hypothesis, we devised a statistical model of scores that we would expect at random by docking a large database of ligands to a receptor. The receptor structure was also randomized by shuffling the sequence. In order to get robust scores, the randomization process was arbitrarily repeated

five times and virtual screening was performed for each random structure. The averaged docked energy for each compound was collected and the histogram was plotted. The data was then fitted to EVD using MATLAB R2007a (<http://www.mathworks.com>) after removing those compounds with free energy of binding below -3kcal/mol. The fitted model was then used to compute the probability and expectation values of ten known binders [9] (extracted from 1E2K, 1E2M, 1E2N, 1E2P, 1KI2, 1KI3, 1KI6, 1KI7, 1KIM and 2KI5) to evaluate the validity of the statistical model. The equation to calculate the probability is as follows [10]:

$$(1). P(Z > z) = 1 - \exp(-\exp(z\pi / \sqrt{6}) - \Gamma'(1))$$

Where P is the probability, z is called Z score transformed by

$$(2). z = (S - \mu) / \sigma$$

S is the raw docking score while μ is the mean value and σ is the standard deviation. $\Gamma'(1)$ is the Euler-Mascheroni constant (≈ 0.5772). The expectation value can be calculated by:

$$(3). E(z) = P(z)N_{db}$$

where N_{db} is the number of chemical compounds in the database.

In order to assess the ability of HiPCDock to identify the known binders from a pool of decoys, we examined the recovery of the same set of ten TK inhibitors from a pool of 1000 compounds (ten inhibitors plus 990 randomly selected ChemBank [6] compounds). The receiver operating characteristic (ROC) plot was generated to aid the analysis. This procedure has been popularly employed in virtual screening [9]. ROC is a plot of the sensitivity (fraction of true positives - TP) vs. 1-specificity (the fraction of false positives - FP) as the discrimination threshold is varied. In our virtual screening context, sensitivity would be the percentage of truly binders being selected from the virtual screening workflow, varying between 0 (all binders missed) and 1 (when all binders are selected), given by below equation:

$$(4). \text{Sensitivity} = \frac{TP}{TP + FN}$$

Specificity, on the other hand, is the percentage of truly non-binders being correctly identified by the

test and therefore being discarded, that is:

$$(5). \text{ Specificity} = \frac{TN}{TN + FP}$$

Specificity can give insight on non-binders that are wrongly classified: the false positives.

III. RESULTS

After the curation of the downloaded ChemBank compounds, 340,845 molecules were made ready for virtual screening. The HiPCDock jobs were distributed onto 300 CPUs and it took about 51 hours to complete each job.

The histogram of the binding free energy was plotted and it turned out to be conformed to an extreme value distribution after excluding 541 compounds with estimated free energy of binding below -3kcal/mol , corresponding to milli-molar concentration weak binding. The hypothesis was further confirmed by the well fitting of the data to the extreme value distribution (red curve in Figure 2). This observation allows us, as described in Method section, to calculate the probability and expectation value for any raw data – the binding free energy of a ligand, and thus evaluate the statistical significance of the prediction.

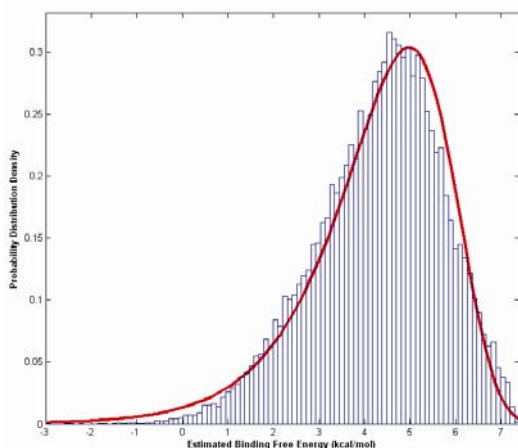


Figure 2. Histogram analysis (blue) and curve fitting (red) showed that the distribution of the binding free energy conformed to an extreme value distribution.

The same screening procedure was applied to the ten known thymidine kinase inhibitors extracted from PDB, and the probability and expectation values were calculated for their predicted binding free energy. We found most of the inhibitors have very low probability and expectation values for those predicted energy values. In particular, the original ligand from 1KIM [4] was ranked top first

with estimated binding free energy of -18.15kcal/mol , and the corresponding root mean square distance (RMSD) between docked conformation and crystal structure is 0.78\AA . This means that, based on our model, the predicted strong binding was statistically significant, with probability of 4.35×10^{-11} and correspondent expectation value of 1.33×10^{-5} . The detailed data for the ten binders was listed in Table I.

Table I. Docking Results for Ten Known TK Binders

Ligands	Free Energy of Binding (kcal/mol)	Probability	Expectation Value
1kim_lig	-18.15	4.35E-11	1.33E-05
1ki3_lig	-10.93	4.56E-07	1.39E-01
1ki2_lig	-10.43	8.66E-07	2.64E-01
2ki5_lig	-9.97	1.57E-06	4.77E-01
1e2k_lig	-9.77	2.02E-06	6.16E-01
1e2n_lig	-9.34	3.52E-06	1.07E+00
1ki6_lig	-8.81	6.94E-06	2.12E+00
1e2p_lig	-7.77	2.65E-05	8.06E+00
1e2m_lig	-7.17	5.72E-05	1.74E+01
1ki7_lig	-5.03	8.84E-04	2.69E+02

Our recovery experiment was used to measure how quickly the known binders were identified from the 1000 compound pool compared to random chance.

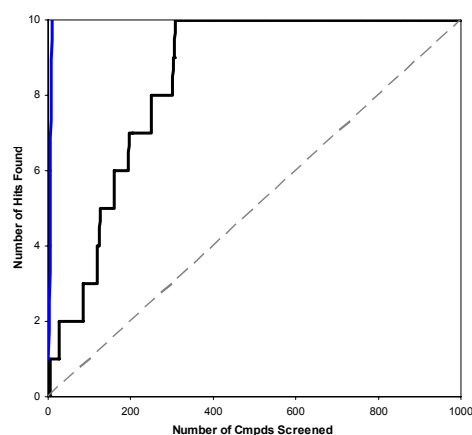


Figure 3. The recovery (black) of ten known TK inhibitors from a 1000-compound pool (990 from ChemBank) with HiPCDock. The grey dashed line represents random picking and the blue is for ideal screening.

As indicated in Figure 3, all of the ten inhibitors were correctly identified in the first 307 compounds,

which were less than 30% of the dataset. A random classification of the compounds is represented by a diagonal rising from the origin to the upper right corner (in dashed grey). For ideal distributions, where known binders are completely separated from the decoys, the curve almost skyrockets vertically and then joins the upper-right corner horizontally (in solid blue). The actual recovery plot is in the middle of the random and the theoretical maximum. With the definition of enrichment used by Warren et al [11], it gives an enrichment of 5.0 if 50% of the known binders are identified within the top 10% of the rank-ordered list. So the enrichment obtained here based on this definition is 4.0 (50%/12.5%), representing that HiPCDock is 4.0 times faster to identify a known inhibitor than random screenings.

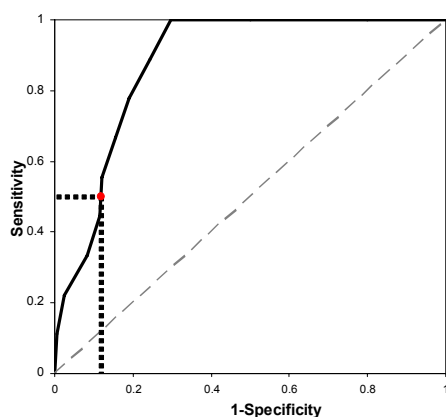


Figure 4. ROC curve obtained with HiPCDock in identifying ten known TK binders from a 1000-compound pool (990 from ChemBank). It is much efficient (solid black) than the random (dashed grey) selection.

In addition to recovery plots, the widely used ROC curve in statistics was also employed in this study to assess the virtual screening ability of HiPCDock. In Figure 4, the ROC curve indicates that the identification of the real binders with HiPCDock (black solid curve) is much better than random screening, which is consistent to the recovery plot. Figure 4 also demonstrated that with the correct identification of 50% of the known binders (ranked on the top 125), 12.5% of the compounds in the dataset are classified as binders (red dot). Similar to the recovery plot, a random classification of the compounds is represented by a diagonal line rising from the origin to the upper right corner (grey dashed line).

IV. DISCUSSIONS

This paper described an implementation of an HPC-based virtual screening scheme with AutoDock3 [8] as the docking engine, enabling us to dock hundreds of thousands of compounds within days. Also for the first time, to the best of our knowledge, we presented an application of a statistical-based bioinformatics model to structure-based high-throughput docking. As aforementioned, there is no a single scoring function that could accurately predict the binding free energy of ligand as of today. Although researchers have realized that consensus scoring [12] might be an option, different scoring schemes unfortunately give quite different results (ranking, binding free energy, etc.) in most of the cases. This makes the hit/lead selection difficult, and thus people have to go through each compound on the top rank using their intuition combined with visualization. It is a very time consuming process, and usually different persons can select quite different compounds. This challenging task provided an impetus for us to develop a statistical-based bioinformatics approach to obtain the confidence of the prediction. Based on the probability and expectation values, the decision-making for hit/lead selection is more objective instead of depending on human subjectivity.

The theoretical rationale can be justified by the fact that docking is a minimum searching process in which the best fitted binding mode of each compound is sought out; therefore, each final docked conformation is the minimum of the process, which is consistent with the definition of extreme value distribution (minimum). The result here is also consistent to some previous report in which the plot implicitly indicated the screening result was an extreme value distribution [8]. The histogram analysis and function fitting validated our hypothesis, thus the resulted model would allow us to calculate the probability and expectation values for any raw docking scores of ligands in that the background data was based on a process of a large number of chemical compounds, of which the probability to be strong binders is small, and random receptor structures were generated through multiple time sequence shuffling. The low probability and expectation values of the known TK binders demonstrated that the docking scores we obtained were statistically significant, thus useful in guiding our hit selection.

The recovery and ROC plotting demonstrated that our HiPCDock is about 4 times more efficient than random screenings. In Figure 4, the recovery curve bends towards up left corner, compared to the random selection (diagonal line). It is worth to note that Figure 4 shows that with the correct identification of 50% of the known binders, 12.5% of the compounds in the dataset were classified as binders. However, this needs to be further validated because this might not be true in reality as it is highly possible that some or even none of the compounds can actually bind to the enzyme. In a ROC curve, the more it bends towards the upper left corner of the diagram, the more distinct the signal appears.

V. CONCLUSIONS

We have developed an elegant and efficient high-throughput docking package, termed HiPCDock, as part of our attempt to integrate cheminformatics and bioinformatics disciplines for drug discovery and development. HiPCDock enables us to screen very large databases (millions of compounds) within a reasonable time period. Also a bioinformatics-based statistical model, motivated by sequence alignment program BLAST, was devised to evaluate the statistical significance of predicted docking scores (binding free energy specifically for AutoDock3). This gave us insight on the confidence of the prediction and will be helpful to guide our decision-making during the hit/lead selection process. We are currently also undergoing further development of its web-based user interface which would make it easier to use for both computational experts and experimental scientists.

ACKNOWLEDGMENT

We would like to thank MDACC Research Information Service for their support of using the institutional HPC clusters. We specially thank those developers of AutoDock, ADT and OpenBabel for allowing us to freely use the software.

REFERENCES

- [1] J. C. Alvarez, "High-throughput docking as a source of novel drug leads," *Curr. Opin. Chem. Biol.*, vol. 8, no. 4, pp. 365-370, Aug. 2004.
- [2] S. Zhang, A. Golbraikh, and A. Tropsha, "Development of quantitative structure-binding affinity relationship models based on novel geometrical chemical descriptors of the protein-ligand interfaces," *J. Med. Chem.*, vol. 49, no. 9, pp. 2713-2724, May 2006.
- [3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389-3402, Sept. 1997.
- [4] J. N. Champness, M. S. Bennett, F. Wien, R. Visse, W. C. Summers, P. Herdewijn, C. E. de, T. Ostrowski, R. L. Jarvest, and M. R. Sanderson, "Exploring the active site of herpes simplex virus type-1 thymidine kinase by X-ray crystallography of complexes with aciclovir and other ligands," *Proteins*, vol. 32, no. 3, pp. 350-361, Aug. 1998.
- [5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235-242, Jan. 2000.
- [6] ChemBank, <http://chembank.broad.harvard.edu>.
- [7] AutoDockTools, Scripps Research Institute, Lo Jolla, CA., 2007.
- [8] AutoDock, Scripps Research Institute, Lo Jolla, CA., 1999.
- [9] C. Bissantz, G. Folkers, and D. Rognan, "Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations," *J. Med. Chem.*, vol. 43, no. 25, pp. 4759-4767, Dec. 2000.
- [10] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nat. Biotechnol.*, vol. 25, no. 2, pp. 197-206, Feb. 2007.
- [11] G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head, "A critical assessment of docking programs and scoring functions," *J. Med. Chem.*, vol. 49, no. 20, pp. 5912-5931, Oct. 2006.
- [12] H. Gohlke and G. Klebe, "Statistical potentials and scoring functions applied to protein-ligand binding," *Curr. Opin. Struct. Biol.*, vol. 11, no. 2, pp. 231-235, Apr. 2001.