Dataset Paper

DockScreen: A Database of *In Silico* Biomolecular Interactions to Support Computational Toxicology

Michael-Rock Goldsmith,¹ Christopher M. Grulke,² Daniel T. Chang,³ Thomas R. Transue,² Stephen B. Little,³ James R. Rabinowitz,³ and Rogelio Tornero-Velez³

¹ Chemical Computing Group, 1010 Sherbrooke Street W., Suite 910, Montreal, QC, Canada H3A 2R7

² Lockheed Martin Information Technology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA

³ Office of Research & Development, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA

Correspondence should be addressed to Rogelio Tornero-Velez; tornero-velez.rogelio@epa.gov

Received 6 June 2014; Accepted 16 September 2014; Published 11 November 2014

Academic Editor: Josefina Pons

Copyright © 2014 Michael-Rock Goldsmith et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We have developed DockScreen, a database of *in silico* biomolecular interactions designed to enable rational molecular toxicological insight within a computational toxicology framework. This database is composed of chemical/target (receptor and enzyme) binding scores calculated by molecular docking of more than 1000 chemicals into 150 protein targets and contains nearly 135 thousand unique ligand/target binding scores. Obtaining this dataset was achieved using eHiTS (Simbiosys Inc.), a fragment-based molecular docking approach with an exhaustive search algorithm, on a heterogeneous distributed high-performance computing framework. The chemical landscape covered in DockScreen comprises selected environmental and therapeutic chemicals. The target landscape covered in DockScreen was selected based on the availability of high-quality crystal structures that covered the assay space of phase I ToxCast *in vitro* assays. This *in silico* data provides continuous information that establishes a means for quantitatively comparing, on a structural biophysical basis, a chemical's profile of biomolecular interactions. The combined minimum-score chemical/target matrix is provided.

1. Introduction

A major challenge with chemicals in consumer products, including but not limited to both pharmaceutical and environmental chemicals, is the ability to fully discover, characterize, and anticipate adverse effects that may result as a consequence of exposure to these chemicals. Classical safety assessment and animal studies are not only cost-prohibitive and lengthy [1, 2]; they often do not include the data required for extrapolations that are inherent in human risk assessment [3]. Developing and evaluating predictive strategies to elucidate the mode of biological activity of environmental chemicals are major undertakings of the US Environmental Protection Agency's Computational Toxicology program (<http://www.epa.gov/comptox/>). Aligning these strategies with the Agency's ongoing chemical-specific risk assessment needs

provision of additional incentive to develop new means of elucidating key determinants of toxicity in the chemical source-to-outcome continuum at a molecular level of accountability. This has provided the motivation for the development of tools such as the Aggregated Computational Toxicology Resource (<http://www.epa.gov/actor>) [1], the DSSTox Toxicoinformatics initiative (<http://www.epa.gov/ncct/dsstox/>) [4], and the ToxRefDB (<http://www.epa.gov/ncct/toxrefdb/>) [5] *in vivo* animal effects database.

In an attempt to fill the inherently large data gaps required for modern risk assessment [6] and to develop both time- and cost-effective approaches for prioritizing the toxicity testing of large numbers of chemicals, the ToxCast program was initiated (<http://www.epa.gov/ncct/toxcast/>) [7]. In Phase 1 of ToxCast, the profiling of >300 well-characterized chemicals (primarily pesticides) in over 400 HTS endpoints was

performed. However, even such large scale *in vitro* screening may not be enough to understand the complex systemic effects seen *in vivo*. Virtual screening has been shown to greatly enhance the success rates of screening experiments [8, 9] and virtual molecular profiling has been shown to be an effective tool for understanding the potential polypharmacology of a chemical leading to probabilistic, data-driven drug discovery [10–12]. It is therefore quite appropriate to apply similar techniques when attempting to understand the polypharmacology that may lead to chemical hazard.

The unparalleled amount of data, low cost, high speed, and rich information content (i.e., high content data) afforded by *in silico* structure-based inquiry (e.g., molecular docking) in addition to the large number of public resources for both target crystal structures and chemical libraries has urged us to consider the development of a structure-based *in silico* database, DockScreen, to complement both the ToxCast program’s screening/prioritization effort and computational toxicology in general. This database contains the biophysical evaluation of molecules within relevant structural constraints of the target proteins (receptors and enzymes) through multiple-chemical, multiple-target molecular docking experiments. We have created a web interface for accessing the data including multiple binding poses and scores for each protein/ligand pairing, but this report is limited to only the most generally useful data: a table containing the highest score obtained for the docking of each ligand to each crystal structure.

2. Methodology

Chemicals were collected and prepared as follows. The chemical landscape covered in DockScreen comprises a selected set of environmental chemicals from ToxCast Phase I (<http://www.epa.gov/ncct/dsstox/sdf.toxcst.html>) [7] and therapeutic chemicals from the FDA MDD database (<http://www.epa.gov/ncct/dsstox/sdf.fdamdd.html>) [13] as drawn from DSSTox [4]. Multiple stereoisomeric forms of ToxCast Phase I chemicals were generated using FLIPPER [14] since many chiral anthropogenic environmental chemicals are unresolved racemic mixtures. Chirality is an important factor in nearly all biomolecular interactions [15] and docking must therefore be carried out using only single isomers. Pregnancy categories for many of the therapeutics were manually extracted from Briggs et al. [16]. Parent-SMILES fields for all chemicals were imported into MOE [17], structures were cleaned, hydrogens were added, and geometries were optimized in a molecular mechanics framework using the MMFFx force-field parameters [18]. (Note: for docking all 3D ligand chemical structure files were submitted as .SDF (MDL) format; however, ligand ID and smiles codes are provided for brevity in supporting information under the SMILES field in Dataset Item 1.)

Targets were selected and prepared as follows. The target landscape covered in DockScreen was selected based on the availability of high-quality crystal structures that covered the assay space of ToxCast Phase I *in vitro* assays (http://www.epa.gov/ncct/toxcast/files/ToxCast_Assays.01aug2007.pdf). A breakdown of the targets selected for study by class is

FIGURE 1: Biomolecular target class breakdown used in computational molecular docking.

available in Figure 1 and more detailed information is contained in the dataset.

The 3D structure files of target proteins were obtained from the Protein Data Bank (PDB), visually inspected in COOT [19], and cleaned up by removing HETATM and solvent waters. HETATM structures, for example, primarily bound ligands, were used as the starting point for clip-file geometries. In some instances (e.g., 2BXK and 1LFO, human serum albumin, and fatty acid binding protein, resp.) multiple binding sites within the same crystal structure were evaluated, in which case the PDBID was augmented by a letter code (a–g) designating a different binding region. There were several redundant target sequences; however, each pocket’s 3D conformation is different providing a unique computational docking experiment.

We chose eHiTS [20] as our initial docking platform since it has a flexible ligand docking method that is exhaustive on the conformations and poses that avoid severe steric clashes between receptor and ligand [21, 22], a potential benefit to computational toxicology where minimizing false negatives is one of many goals. It has also been shown to compete well with other docking software programs in accuracy of docking and enrichment of chemical libraries [23].

In eHiTS, the binding pocket is determined by building a steric grid for the specified receptor binding site and a cavity description is built that consists of thousands of geometric shapes. The ligand is divided into rigid fragments and connecting flexible chains, where the rigid fragments are docked to all possible places in the defined cavity independently of each other. Then, exhaustive matching of compatible rigid fragment pose sets is performed by a rapid hyper-graph clique detection algorithm that enables the elucidation of acceptable

combinations of poses and respective scores. The flexible chains are then fitted to the specific rigid fragment poses that comprise a matching pose set, driven by a scoring function built upon a local energy minimization in the active site of the receptor.

The default clip-file parameters were used for docking in eHiTS, using a square docking box around the desired ligand. An intermediate pose-reconstruction with “Accuracy level” 3 was used to evaluate all poses balancing accuracy with speed. The minimum energy score for each ligand/receptor complex is included in this dataset; however, all poses and scores are retained in the internal DockScreen database. All scores are stored and reported in units of $\log(K_i[M])$.

The total project is comprised of ~1100 ligands on 150 unique protein (receptor/enzyme) binding sites that covered a total of ~100 unique targets. At a 32-pose-per-ligand (maximum) storage capacity, this puts the upper range of calculations at 1.6×10^5 unique ligand/target complex combinations. At a run time of >5 minutes on average, we anticipated a total run-time of >1.5 years; clearly, a job suited for distributed computing architecture. Calculations were run at the US Environmental Protection Agency’s National Computing Center and deployed over a heterogeneous distributed high-performance computing cluster using primarily idle time and resulted in an average performance of ~20 nodes running at any given instant over the period of two (2) months.

Biomolecular interaction profiles require the ligand to reach its target. For many chemicals, absorption, distribution, metabolism, and excretion (ADME) impose a limit on the interactions available. QikProp [24] was applied to the cleaned, stereospecific set of chemicals to provide an initial set of predictions regarding these chemicals’ ADME characteristics. All chemicals were energy minimized using OPLS-AA [25] in MOE [17] to match with the energy minimization technique used in QikProp model generation.

3. Dataset Description

The dataset associated with this Dataset Paper consists of 4 items which are described as follows.

Dataset Item 1 (Table). Listing of 1094 ligands for which docking results are available. Documented for each ligand are its name and stereospecific SMILES along with whether it came from the ToxCast or FDA dataset. DSSTox_CID is included for linking to DSSTox (<http://www.epa.gov/ncct/dsstox/>). FDA therapeutic categories and pregnancy class are included for therapeutic chemicals.

Column 1: ID
 Column 2: Ligand ID
 Column 3: DSSTox CID
 Column 4: Gsid
 Column 5: SMILES
 Column 6: Name
 Column 7: Categories
 Column 8: CASRN

Column 9: ToxCast ID
 Column 10: FDA Therapeutic Class
 Column 11: FDA Pregnancy Class

Dataset Item 2 (Table). Listing of 140 pdb entries from which target protein structure was extracted for docking studies. Contained are the PDB ID for linking to the Protein Data Bank (<http://www.pdb.org/>), a description of the protein, and details on the deposition, origin, and quality of the structure. Included are a target class annotation made manually by the authors and comments on whether multiple chains were used in docking studies.

Column 1: PDB ID
 Column 2: Description
 Column 3: Experimental Technique
 Column 4: Deposition Date
 Column 5: Release Date
 Column 6: Authors
 Column 7: Keywords
 Column 8: Target Class
 Column 9: Resolution
 Column 10: Comment

Dataset Item 3 (Table). Minimum scores resulting from docking the 1094 ligands defined in Dataset Item 1 to the 150 targets (from 140 PDB entries) denoted in Dataset Item 2. Each row contains data for one ligand linkable by Ligand ID. Column names correspond to PDB ID with an additional chain letter if applicable. Data values are the minimum scores obtained from docking that ligand to that target. If no score is listed, no poses were found to have sufficient binding to warrant scoring.

Column 1: Ligand ID
 Column 2: 1A28
 Column 3: 1BEH
 ⋮
 Column 149: 2VEV
 Column 150: 3CBS
 Column 151: 3ERT

Dataset Item 4 (Table). Prediction made using the QikProp program. Contained are several commonly used descriptors and ADME properties. These data form a basis for attempting to evaluate the effects ADME may play on toxic effects with which a particular ToxCast chemical may or may not be associated.

Column 1: Ligand ID
 Column 2: #stars
 Column 3: #amine
 ⋮

Column 50: #nonHatm

Column 51: RuleOfThree

Column 52: Jm

4. Concluding Remarks

Potential methods by which 3D modeling techniques could inform mechanistic toxicology have previously been documented [26], but the wealth of data contained in DockScreen may provide even more options. As can be seen from comparison studies of different docking method [22], the use of only a single software program has its limitations; however, the creation of a large database of docking results across many targets yields advantages for data mining and analysis which are unavailable elsewhere. One apparent use is combining DockScreen with chemical descriptors to model and understand *in vitro* or *in vivo* assay results, as reported previously [27]. This application as a unique source of knowledge in modeling may improve linking chemical structures with *in vitro* and *in vivo* effects in a fully computational approach, thereby increasing *in silico* predictive power and reducing our reliance on animal models. A second use for DockScreen is the population of information in data fusion tools intent on enabling decision support in chemical screening and prioritization. Efforts to build such tools have increased recently, resulting in the creation of “Dashboards” in the EPA’s Chemical Safety and Sustainability program (<http://actor.epa.gov/actor/faces/CSSDashboardLaunch.jsp>). Third, DockScreen contains molecular-level representations that are readily searchable and can be a valuable resource for scientists within the EPA working on molecular-level insight to some of their *in vitro* data efforts. For instance, a DockScreen user can use the system to search for structural analogues of the novel compounds. Similarly, the nature of the data is amenable to probing molecular similarity based on 3-dimensional biophysical interaction profiles (e.g., multiple target vector scores for a given chemical) [11] which are significantly different from 2D Tanimoto similarity based on chemical fingerprints.

Dataset Availability

The dataset associated with this Dataset Paper is dedicated to the public domain using the CC0 waiver and is available at <http://dx.doi.org/10.1155/2014/421693/dataset>. In addition, more information on the calculated properties is available in the QikProp manual downloadable at <http://www.schrodinger.com/supportdocs/18/>.

Disclosure

Lockheed Martin is a contractor to the U.S. Environmental Protection Agency. Michael-Rock Goldsmith and Daniel T. Chang completed the research and a large portion of the manuscript preparation while working at the U.S. Environmental Protection Agency.

Disclaimer

This paper was funded and managed by the United States Environmental Protection Agency through its Office of Research and Development. It has been subjected to the agency administrative review and approved for submission and peer review. Reference to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was enabled by a Material Transfer Agreement with SimBioSys Inc. (http://www.epa.gov/ncct/download_files/partners/SimBioSys.pdf) for a multinode license for eHiTS 5.8 which was deployed on the National Computing Center’s High-Performance Computing infrastructure (<http://www.epa.gov/nesc/>) from 2007 to 2008. Many thanks are due to Dr. Ann Richard (US EPA NCCT) and Maritja Wolf (Lockheed Martin IT) for quality assurance and assistance on chemical structure curation and quality.

References

- [1] R. Judson, A. Richard, D. Dix et al., “ACToR—aggregated computational toxicology resource,” *Toxicology and Applied Pharmacology*, vol. 233, no. 1, pp. 7–13, 2008.
- [2] R. Judson, A. Richard, D. J. Dix et al., “The toxicity data landscape for environmental chemicals,” *Environmental Health Perspectives*, vol. 117, no. 5, pp. 685–695, 2009.
- [3] J. R. Rabinowitz, M.-R. Goldsmith, S. B. Little, and M. A. Pasquini, “Computational molecular modeling for evaluating the toxicity of environmental chemicals: prioritizing bioassay requirements,” *Environmental Health Perspectives*, vol. 116, no. 5, pp. 573–577, 2008.
- [4] A. M. Richard and C. R. Williams, “Distributed structure-searchable toxicity (DSSTox) public database network: a proposal,” *Mutation Research*, vol. 499, pp. 27–52, 2002.
- [5] M. T. Martin, R. S. Judson, D. M. Reif, R. J. Kavlock, and D. J. Dix, “Profiling chemicals based on chronic toxicity results from the U.S. EPA ToxRef database,” *Environmental Health Perspectives*, vol. 117, no. 3, pp. 392–399, 2009.
- [6] National Research Council, *Toxicity Testing in the 21st Century: A Vision and a Strategy*, The National Academies Press, Washington, DC, USA, 2007.
- [7] D. J. Dix, K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer, and R. J. Kavlock, “The toxcast program for prioritizing toxicity testing of environmental chemicals,” *Toxicological Sciences*, vol. 95, no. 1, pp. 5–12, 2007.
- [8] J. Bajorath, “Integration of virtual and high-throughput screening,” *Nature Reviews Drug Discovery*, vol. 1, no. 11, pp. 882–894, 2002.

- [9] D. C. Fara, T. I. Oprea, E. R. Prossnitz, C. G. Bologa, B. S. Edwards, and L. A. Sklar, "Integration of virtual and physical screening," *Drug Discovery Today: Technologies*, vol. 3, no. 4, pp. 377–385, 2006.
- [10] G. V. Paolini, R. H. Shapland, W. P. van Hoorn, J. S. Mason, and A. L. Hopkins, "Global mapping of pharmacological space," *Nature Biotechnology*, vol. 24, no. 7, pp. 805–815, 2006.
- [11] Á. Peragovics, Z. Simon, L. Tombor et al., "Virtual affinity fingerprints for target fishing: a new application of drug profile matching," *Journal of Chemical Information and Modeling*, vol. 53, no. 1, pp. 103–113, 2013.
- [12] D. Hristozov, T. I. Oprea, and J. Gasteiger, "Ligand-based virtual screening by novelty detection with self-organizing maps," *Journal of Chemical Information and Modeling*, vol. 47, no. 6, pp. 2044–2062, 2007.
- [13] E. J. Matthews, N. L. Kruhlak, R. D. Benz, and J. F. Contrera, "Assessment of the health effects of chemicals in humans: I. QSAR estimation of the maximum recommended therapeutic dose (MRTD) and no effect level (NOEL) of organic chemicals based on clinical trial data," *Current Drug Discovery Technologies*, vol. 1, no. 1, pp. 61–76, 2004.
- [14] FLIPPER, "Openeye Scientific Software," Santa Fe, NM, USA, <http://www.eyesopen.com/>.
- [15] A. J. S. Knox, M. J. Meegan, G. Carta, and D. G. J. Lloyd, "Considerations in compound database preparation "hidden" impact on virtual screening results," *Journal of Chemical Information and Modeling*, vol. 45, no. 6, pp. 1908–1919, 2005.
- [16] G. G. Briggs, R. K. Freeman, and S. J. Yaffe, *Drugs in Pregnancy and Lactation: a Reference Guide to Fetal and Neonatal Risk*, Lippincott Williams & Wilkins, 2012.
- [17] Chemical Computing Group, *Molecular Operating Environment (MOE) v2006.08*, Chemical Computing Group, Quebec, Canada, 2006.
- [18] T. A. Halgren, "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94," *Journal of Computational Chemistry*, vol. 17, no. 5-6, pp. 490–519, 1996.
- [19] P. Emsley and K. Cowtan, "Coot: model-building tools for molecular graphics," *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, pp. 2126–2132, 2004.
- [20] eHITS v. 6.1, Simbiosys Inc. Toronto Canada.
- [21] Z. Zsoldos, D. Reid, A. Simon, B. S. Sadjad, and A. P. Johnson, "eHiTS: an innovative approach to the docking and scoring function problems," *Current Protein and Peptide Science*, vol. 7, no. 5, pp. 421–435, 2006.
- [22] Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad, and A. P. Johnson, "eHiTS: a new fast, exhaustive flexible ligand docking system," *Journal of Molecular Graphics and Modelling*, vol. 26, no. 1, pp. 198–212, 2007.
- [23] D. Plewczynski, M. Łażniewski, R. Augustyniak, and K. Ginalski, "Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database," *Journal of Computational Chemistry*, vol. 32, no. 4, pp. 742–755, 2011.
- [24] *Small-Molecule Drug Discovery Suite 2014-1: QikProp*, version 3.9, Schrödinger, LLC, New York, NY, USA, 2014.
- [25] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, "Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids," *Journal of the American Chemical Society*, vol. 118, no. 45, pp. 11225–11236, 1996.
- [26] M. R. Goldsmith, S. D. Peterson, D. T. Chang et al., "Informing mechanistic toxicology with computational molecular models," in *Computational Toxicology*, pp. 139–165, Humana Press, 2012.
- [27] M. R. Goldsmith, D. T. Chang, J. R. Rabinowitz, S. B. Little, and R. R. Tice, *To Hit, or Not to Hit?: In Silico Models of In Vitro Nuclear Receptor Transactivation*, National Toxicology Board of Scientific Counselors, 2010.