Title: DockScreen: A database of *in silico* biomolecular interactions to support computational toxicology

Authors: Michael-Rock Goldsmith[3], Christopher M. Grulke[2], Daniel T. Chang[1], Thomas R. Transue[2], Stephen B. Little[1], James R. Rabinowitz[1], Rogelio-Tornero-Velez[1,*]

Affiliations

1. Office of Research & Development, U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park NC 27711
2. Lockheed Martin, A Contractor to the U.S. Environmental Protection Agency, 109 T.W. Alexander Drive, Research Triangle Park NC 27711
3. Chemical Computing Group, 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec, Canada H3A 2R7

* Corresponding author information:
U.S. Environmental Protection Agency
109 T.W. Alexander Drive (E205-01)
Research Triangle Park, NC 27711
phone: 919-541-9447
tornero-velez.rogelio@epa.gov

Abstract

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 >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.

Keywords: *in silico*, virtual high-throughput screening, comparative molecular interaction profile analysis, affinity matrix, binding affinity, computational toxicology, ToxCast, molecular docking

1. Introduction

A major challenge within the social chemical industries, 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 is a major undertaking 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 provides 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 Toxico-Chemoinformatics 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 both 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 [10,11,12]. 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 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. Methods

Chemical Collection and Preparation

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 GG, *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 within the Ligand tab.)

Target Selection and Preparation

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

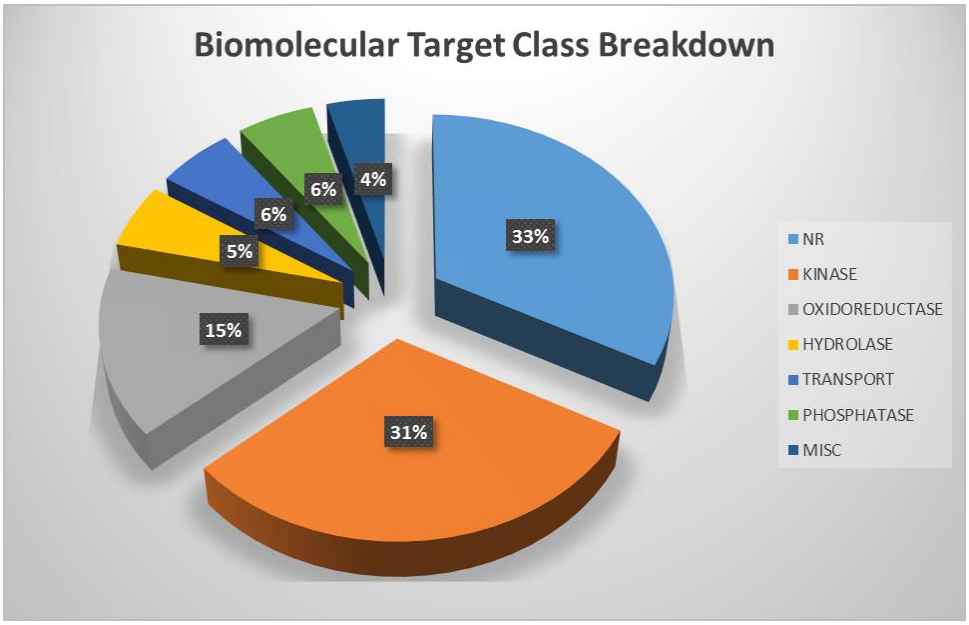

Figure 1: Biomolecular target class breakdown used in computational molecular docking.

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, e.g., 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 respectively) 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.

Docking

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 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 an 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 are 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 $\log(K_i[\mathrm{M}])$.

Computational Complexity

The total project comprised of ~1100 ligands on 150 unique protein (receptor/enzyme) binding sites that covered a total of ~ 100 unique targets. At 32 pose maximum storage 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.

ADME Properties

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 included dataset consists of four data items:

Dataset Item 1 (Table). This item consists of a listing of 1094 ligands for which docking results are available. Documented for each ligand is 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: LigandId
Column 2: DSSTox_CID
Column 3: SMILES
Column 4: Name

Column 5: Categories
Column 6: CASRN
Column 7: ToxCast_ID
Column 8: FDA_Therapeutic_Class
Column 9: FDA_Pregnancy_Class

Dataset Item 2 (Table). This item consists of a listing of 140 pdb entries from which target protein structure were extracted for docking studies. Contained is the PDBId for linking to the Protein Data Bank (www.pdb.org), a description of the protein and details on the deposition, origin, and quality of the structure. Included is a target class annotation made manually by the authors and comments on whether multiple chains were used in docking studies.

Column 1: PDBId
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). This item consists of 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 LigandId. Column names correspond to PDBId 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 where found to have sufficient binding to warrant scoring.

Column 1: LigandId
Column 2-Column152: 1A28 – 3ERT (PDBId)

Dataset Item 4 (Table). This item consists of the 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. More information on the calculated properties is available in the Qikprop manual downloadable at http://www.schrodinger.com/supportdocs/18/.

Column 1: LigandId
Column 2-Column 53: #star - 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 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.

5. Acknowledgements

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

6 . References

[1] Judson, R., Richard, A., Dix, D., Houck, K., Elloumi, F., Martin, M., Cathey, T., Transue, T. R., Spencer, R., and Wolf, M. ACToR — Aggregated Computational Toxicology Resource. Toxicol. Appl. Pharmacol. 233(1) (2008) 7-13, ISSN 0041-008X, http://dx.doi.org/10.1016/j.taap.2007.12.037

[2] Judson, R., Richard, A., Dix, D. J., Houck, K., Martin, M., Kavlock, R. and Smith, E. The toxicity data landscape for environmental chemicals. Environ. Health Perspect. 117(5), (2009) 685

[3] Rabinowitz, J. R., Goldsmith, M.-R., Little, S. B. and Pasquinelli M. A. Computational Molecular Modeling for Evaluating the Toxicity of Environmental Chemicals: Prioritizing Bioassay Requirements. Environ. Health Perspect., 116, (2008) 573-577

[4] Richard, A. M. and Williams, C. R. Mutation Research 499 (2002) 27-52 (b)
http://www.epa.gov/ncct/dsstox/

[5] Martin, M. T., Judson, R. S., Reif, D. M., Kavlock, R. J., and Dix, D. J. Profiling chemicals based on chronic toxicity results from the US EPA ToxRef Database. Environ. Health Perspect. 117(3), (2009) 392

[6] National Research Council. Toxicity Testing in the 21st Century: A Vision and a Strategy. Washington, DC: The National Academies Press (2007)

[7] David, D. J., Houck, K. A., Martin, M. T., Richard, A. M., Setzer, W. R., and Kavlock, R. J. The ToxCast Program for Prioritizing Testing of Environmental Chemicals. Toxicol. Sci. 95(1), (2007) 5-12

[8] Bajorath, J. Integration of virtual and high-throughput screening. Nature Reviews Drug Discovery 1.11 (2002): 882-894

[9] Fara, D. C., Oprea, T. I, Prossnitz, E. R., Bologa, C. G., Edwards, B. S. and Sklar, L. A. Integration of virtual and physical screening Drug Discovery Today: Technologies 2, (2006) 377-385

[10] Paolini, G. V., Shapland, R. H., van Hoorn, W. P., Mason, J. S., and Hopkins, A. L. Global mapping of pharmacological space. Nature Biotechnology, 24(7), (2006) 805-815

[11] Peragovics, A., Simon, Z., Tombor, L., Jelinek, B., Hári, P., Czobor, P. and Málnási-Csizmadia, A. Virtual affinity fingerprints for target fishing: a new application of Drug Profile Matching. J. Chem. Inf. Model. 53(1), (2012) 103-113

[12] Hristozov D., Oprea T. I. and Gasteiger J. Ligand-Based Virtual Screening by Novelty Detection with Self-Organizing Maps. J. Chem. Inf. Model. 47, (2007) 2044-2062

[13] Matthews, E.J., Kruhlak, N.L., Benz, R.D. and Contrera, J.F. 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, 1(1), (2004) 61-76

[14] FLIPPER; Openeye Scientific Software. http://www.eyesopen.com, Santa Fe NM

[15] Knox, A. J. S., Meegan, M. J., Carta, G. and Lloyd, D. G. J. Chem. Inf. Model. 45 (6), (2005) 1908–1919

[16] Briggs, G. G., Freeman, R. K., and Yaffe, S. J. Drugs in pregnancy and lactation: a reference guide to fetal and neonatal risk. Lippincott Williams & Wilkins. (2012)

[17] Molecular Operating Environment (MOE), Chemical Computing Group Inc., Quebec, Canada.

[18] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94, J. Comp. Chem. 17 (5-6), (1996) 490-519

[19] Emsley, P. and Cowtan, K. Acta Cryst. D60, (2004) 2126-2132
http://www.ysbl.york.ac.uk/~emsley/coot/

[20] eHITS v. 6.1, Simbiosys Inc. Toronto Canada

[21] Zsoldos, Z., Reid, D., Simon, A., Sadjad, B. S. and Johnson, A. P. eHiTS: an innovative approach to the docking and scoring function problems. Current Protein and Peptide Science 7(5), (2006) 421-435

[22] Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B., and Johnson, A. P. eHiTS: a new fast, exhaustive flexible ligand docking system. J. Mol. Graph. Model. 26(1), (2007) 198-212

[23] Plewczynski, D., Łaźniewski, M., Augustyniak, R., and Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J. Comp. Chem. 32(4), (2011) 742-755

[24] Small-Molecule Drug Discovery Suite 2014-1: QikProp, version 3.9, Schrödinger, LLC, New York, NY, 2014

[25] Jorgensen W.L., Maxwell, D.S., and Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids; J. Am. Chem. Soc. 117 (1996) 11225–11236.

[26] Goldsmith, M. R., Peterson, S. D., Chang, D. T., Transue, T. R., Tornero-Velez, R., Tan, Y. M., and Dary, C. C. Informing Mechanistic Toxicology with Computational Molecular Models. In *Computational Toxicology*, Humana Press (2012) 139-165

[27] Goldsmith, M. R., Chang D. T., Rabinowitz, J. R., Little, S. B., and Tice, R. R. To hit, or not to hit?: *In silico* models of *in vitro* nuclear receptor transactivation. National Toxicology Board of Scientific Counselors, November 30, 2010