SAMPL5 cyclohexane-water distribution coefficient challenge

version 2, updated Oct. 27, 2015

In this challenge, participants will predict cyclohexane-water distribution coefficients (log D) for small molecules at pH 7.4 for a newly-generated experimental dataset of approximately 100 small molecules.

$\log D_{pH\ 7.4} = \log_{10} ([solute]_{cyclohexane} / [solute]_{aqueous})$

where [solute] denotes the sum of all ionized and unionized species in that phase, and the aqueous phase is buffered to pH 7.4.

These measurements were carried out during the summer of 2015 by Bas Rustenburg (PBSB graduate student at the Weill Cornell Medical College) during a summer research internship at Genentech under the direction of Dan Ortwine, originally organized by JW Feng (now at Denali Therapeutics). We are incredibly grateful to Genentech for supporting this blind community challenge.

Experimental details

Measurements of the log distribution coefficient (log D) between water and cyclohexane of approximately 100 small molecules were performed at Genentech. The protocol was adapted from the UPLC-MS/MS protocol described in *Lin & Pease 2013*[1]. For the aqueous phase, pH 7.4 was attained using phosphate buffered saline (PBS) buffer, which consisted of 136 mM NaCl, 2.6 mM KCl, 7.96 Na_2HPO_4, and 1.46 mM KH_2PO_4. Cyclohexane (ACS grade >=99%, Sigma-Aldrich 179191-2L, batch #00555ME) was used for the cyclohexane phase. Unlike the original protocol, neither phase was presaturated prior to pipetting. Glass vials were used to perform partitioning, as cyclohexane was found to dissolve polystyrene and polypropylene 96-well plates used in the original experimental protocol. For each experiment, 10 uL of 10 mM compound in DMSO and 5 uL of 200 uM propanolol in acetonitrile (an internal standard) were added to 500 uL cyclohexane, followed by the addition of 500 uL of PBS. As in the original protocol, samples were shaken for 50 minutes, centrifuged to separate phases, and then aliquots were taken from each layer for measurement. Before concentration quantification, cyclohexane aliquots were diluted with 90% octanol to prevent accumulation of cyclohexane on the C18 column. Total concentration of all ionization states of the compound were quantified in each phase using the original UPLC-MS/MS method.

The dataset provided contains 64 small molecule drug-like fragments from public catalogs, predicted to span a large dynamic range of log $D_{7.4}$ values. 31 additional compounds used by Genentech as internal controls for octanol-water measurements

[1] Lin & Pease, Combinatorial Chemistry & High Throughput Screening, 2013, 16, 817-825

were also added to the set of cyclohexane-water measurements[1]. Cyclohexane-water distribution coefficient data was provided initially for 95 compounds, but subsequent curation for data quality took this down to the 53 compounds presented here.

Possible factors to consider in your approach

Partition or distribution coefficients are determined by the difference in solvation free energies of the relevant species in the different phases. In fact, they can be estimated from gas-to-solvent transfer free energies into the different solvents. **However, it is important to note that the experimental reality may be more complicated for several reasons.**

First, as noted, the measurements are for distribution coefficients rather than partition coefficients. Thus, for molecules which may be present in multiple protonation states in each environment, or which may undergo a change of protonation states upon transfer between environments, your calculations may need to consider how this may impact the distribution coefficients. We plan to provide predicted pKa values (from Epik) for these compounds, though these may not be available with the initial data package.

Secondly, the experiments are done on phase-separated water and cyclohexane and the solute can impact the distribution of water and cyclohexane themselves. For example, carboxylic acids and some other solutes can carry a bound water with them into the nonaqueous phase, at least in some cases[2]. Additionally, after mixing, the cyclohexane and water phases may no longer be pure---water and/or salts may be found in the cyclohexane phase, and cyclohexane may be found in the aqueous phase. The addition of the internal standard (propanolol) or acetonitrile and DMSO content may also be consideration.

Finally, dimerization or oligomerization of solute molecules in one or more of the phases may also impact results; for example, a polar molecule might dimerize in a non-polar phase, resulting in stabilization in that phase relative to what would be expected based on the monomer's transfer free energy[2].

We are not aware which, if any, of these potential complications are relevant for this competition. We mention them only to ensure all participants are aware of them.

Reference calculations

We plan to perform a set of reference calculations using gas-to-solvent transfer free energies to cyclohexane and water to estimate log D, for the form of each molecule provided in the structure (.sdf and .mol2) files and simulation input files in this package. The purpose of these reference calculations is *not* necessarily to yield accurate partitioning coefficients, but to provide a well-converged set of computed partitioning coefficients calculated via a straightforward protocol, with parameters which are

---

[2] Leo, A., Hansch, C., & Elkins, D. (1971). Partition coefficients and their uses. *Chemical Reviews*, *71*(6), 525–616.

available to everyone. This will provide a point of comparison which other participants can use to separate results which differ due to sampling or simulation method from results which differ because of simulation parameters, selected protonation state or tautomer, etc.

In past iterations of SAMPL, when participants' results disagree for a particular system, it has often been unclear as to whether these differences are due to the sampling or free energy method used, the details of the simulation setup (including things like protonation state or tautomer selected) or the choice of force field. The reference calculations provide an attempt to begin separating these factors, allowing participants who so desire to run a set of calculations with the exact same force field parameters and system composition, separating sampling issues from force field and simulation setup considerations. Hopefully this will allow us to more directly compare performance of different force fields given the same sampling method, and different sampling methods given the same force field.

Molecule and simulation setups

SMILES strings for all compounds were provided by Genentech. We used these, in combination with the OpenEye toolkits, to generate 3D structures of all compounds, calculate AM1-BCC partial charges (which may or may not be useful to participants) and then write out final isomeric SMILES strings and .mol2 and .sdf format files. These are provided. We then used our internal pipeline (based on OpenMolTools and SolvationToolkit, available on GitHub) to build solvated boxes of each system, with GAFF cyclohexane and TIP3P water (solvated in 500 molecules of cyclohexane and 1000 molecules of water, respectively). These were then prepared in GROMACS format, and converted to other formats (AMBER, Desmond, LAMMPS) with custom scripts utilizing ParmEd and InterMol. No equilibration or energy minimization was done on these systems yet; however, they will be used as inputs for our standard calculations.

To validate our conversion of the provided input files into formats for different simulation packages, single-point energies for the provided conformations/systems are be provided in SAMPL5Energies.xls. Results are summarized in the instructions file.

Please feel free to contact us if you notice any errors in the information provided or have questions about SAMPL5: samplchallenge@gmail.com

Acknowledgments

Ariën S. ("Bas") Rustenburg, PBSB Graduate Program, Weill Cornell Graduate School of Medical Sciences // Chodera lab, MSKCC
Dan Ortwine, Genentech
Baiwei Lin, Genentech
Joseph Pease, Genentech
Justin Dancer, Genentech
David L. Mobley, University of California, Irvine

Caitlin Bannan, University of California, Irvine // Mobley Lab
Kalli Burley, University of California, Irvine // Mobley Lab
Michael Shirts, University of Colorado Boulder
JW Feng, Denali Therapeutics
John D. Chodera, Sloan Kettering Institute, MSKCC