Semi-explicit solvent model: theoretical assumptions

A detailed description and validation of model assumptions is presented in our previous publications. Here, we give a brief overview of the most important aspects.

- The space available to solvent is discretised into body centred cubic (BCC) lattice, with 1 Å distance between the nearest lattice points.
- Each lattice point, \( r \), can be either occupied by solvent with density \( \rho(r) = 1 \) or empty, with \( \rho(r) = 0 \).
- Excess water chemical potential is evaluated, by considering an effective Hamiltonian of a rotatable solvent probe:

  \[
  H_{\text{eff}}(r, \{n\}) = -k_B T \ln \left( \sum_\theta e^{-\beta H_{UV}(r,\theta)-\beta H_{VV}(r,\theta,\{n\})} \right),
  \]

  (1)

  where \( \{n\} \) denotes an instantaneous distribution of occupied and empty lattice points, \( H_{UV}(r, \theta) \) is a solute-solvent interaction energy combining electrostatic and Lennard-Jones (LJ) contributions with SPC/E model\(^4\), representing the solvent probe, \( H_{VV}(r, \theta, \{n\}) \) is a mean-field solvent-solvent term, \( k_B T \), and \( \beta \) are the Boltzmann constant times temperature (\( T = 300 \) K), and its inverse respectively, and the summation extends over a set of 12 discrete probe orientations, \( \theta \).

  Note, that due to specific geometry of the SPC/E water model (OH distance of 1 Å, and HOH angle of 109°), the probe can be placed on the BCC lattice such that its oxygen atom occupies point \( r \), and two hydrogen atoms are located at two of its 8 nearest neighbours, with 12 possible spatial orientations.

  - Given a unique solute placement on the grid, solute-solvent interaction potentials (electrostatic and Lennard-Jones) are calculated for all lattice points, giving rise to the \( H_{UV}(r, \theta) \) term.

  - The mean-field solvent solvent term, \( H_{VV}(r, \theta, \{n\}) = k \epsilon_{HB} \), is evaluated based on the assumed energy of water hydrogen bond, \( \epsilon_{HB} \) (one of six adjustable model parameters), and the number \( k \in \{0..4\} \) of occupied lattice points, two of which are located 3 BCC grid spacings away from the central point \( r \) occupied by the oxygen atom in the direction of hydrogen atoms (the resulting 3 Å distance is well within the length limit for water-water hydrogen bond), and two other that are at the same distance at the direction of oxygen lone pairs, i.e. reproducing the tetrahedral arrangement of water hydrogen-bonding.

  - The calculations start with a uniform water distribution, \( \rho = 1 \) at each lattice point. \( H_{\text{eff}} \) is then evaluated for each lattice point, and points with \( H_{\text{eff}} > \eta \) are vacated, with \( \eta \) being a "drying threshold" – an adjustable model parameter. This procedure is repeated until stationary solvent distribution \( \{n\} \) is reached.

  - The resulting map of excess chemical potential is mapped on a finer, 0.5 Å Cartesian grid. The calculations are then repeated for next random BCC grid orientation with respect to the solute (by default \( N = 500 \) independent runs are performed).
An excess chemical potential map accumulated on the Cartesian grid during $N$ independent runs is then partitioned into individual hydration sites. Water binding free energy at each site $i$, $\Delta F_i$, is obtained as a difference between the estimate of water excess chemical potential, $\mu(r_i)$, and the reference, bulk water excess chemical potential, $\mu_b$:

$$\Delta F_i = \mu(r_i) - \mu_b = -k_B T \ln \sum_{r \in V_i} \sum_{N} e^{-\beta H_{eff}^N(r)} + V_{WCA}(r_i) - \mu_b,$$

where the summation over $r$ includes all points belonging to hydration site volume $V_i$, $H_{eff}^N(r)$ is an effective Hamiltonian obtained at run $N$, and $V_{WCA}(r_i)$ corresponds to a repulsive, Weeks-Chandler-Andersen potential accumulated at $r_i$ during solvent partitioning. The latter was parametrised to obtain spacing between surface water molecules in agreement with explicit solvent simulations.

Structures submitted for calculations are parametrised using with non-bonded force field parameters: Amber99 for proteins or GAFF for ligands. In this latter case AM1-BCC atomic partial charges are assigned with the use of antechamber program.

The model has six adjustable parameters: 1) the energy of water-water hydrogen bond, 2) bulk water excess chemical potential compatible with grid representation, 3) the drying threshold, 4) an energy scaling factor, 5) an effective BCC grid solvent dielectric constant (used only in hydration free energy calculations), 6) an effective number density of water on the cartesian grid. Parameters 1-5 were adjusted to reproduce experimental hydration free energies of drug-like solutes (700 neutral and charged compounds, with hydration free energies ranging from +3.4 to 536 kcal/mol.) and the parameter 6 was adjusted to provide agreement of predicted water binding free energies to protein cavities in agreement with reference, explicit solvent free energy calculations.

PQRS file format

The GridSolvate server uses a slightly modified PQR file format to store all necessary information about solvated system. A sample listing is shown in Fig. In addition to data found in a standard PDB file up to columns containing X, Y, Z atomic positions, a PRQS file includes 3 additional records:

- atomic partial charge in $e$ units (float),
- atomic radius, $R$, in Å (float),
- well depth of atomic Lennard Jones potential, $\varepsilon$ in kcal/mol (float).

| ATOM | N | GLU | 16 | 5.968 | 28.822 | 38.723 | 0.0017 | 1.8240 | 0.1700 |
|------|---|-----|----|-------|--------|--------|--------|--------|--------|
| ATOM | CA| GLU | 16 | 7.361 | 28.828 | 38.162 | 0.0588 | 1.9080 | 0.0860 |
| ATOM | C | GLU | 16 | 7.575 | 27.801 | 37.020 | 0.5621 | 1.9080 | 0.0860 |
| ATOM | O | GLU | 16 | 7.067 | 26.660 | 37.055 | -0.5889 | 1.6612 | 0.2100 |
| ATOM | CB| GLU | 16 | 8.428 | 28.677 | 39.267 | 0.0909 | 1.9080 | 0.0860 |
| ATOM | CG| GLU | 16 | 8.639 | 27.254 | 39.786 | -0.0236 | 1.9080 | 0.0860 |
| ATOM | CD| GLU | 16 | 10.078 | 26.983 | 40.238 | 0.8087 | 1.9080 | 0.0860 |
| ATOM | OE1| GLU | 16 | 10.757 | 27.948 | 40.697 | -0.8189 | 1.6612 | 0.2100 |
| ATOM | OE2| GLU | 16 | 10.476 | 29.605 | 42.443 | -0.2188 | 1.6612 | 0.2100 |

Figure 1: A sample listing of PQRS file used by GridSolvate.
Server output files

The following files are generated upon successful job completion:

- **water_added.pdb** containing all added water molecules. In addition to Cartesian coordinates, two extra columns in this file contain an estimate of local excess water chemical potential in kcal/mol (more negative values mean stronger affinity to a given site), and a rough estimate of volume in Å³ accessible to the centre of water molecule at a given site.

- **water_buried.pdb**, **water_5.pdb**, and **water_10.pdb** containing subsets of all added water molecules that are either completely buried within macromolecular structure, or have 5, or 10% of surface area exposed to bulk solvent, respectively. All those files contain the same additional columns as the **water_added.pdb**.

- **grid_s.pdb** containing all sites with measurable excess chemical potential, including those with positive (i.e. unfavourable) values in addition to hydrated sites contained in **water_added.pdb** file.

- **grid_c.pdb** containing a dense Cartesian grid representing a spatial map of excess chemical potential. This file can be used to visualise and analyse hydration propensity of a protein surface.

- **system_hydrated.pdb** containing the submitted macromolecular structure/complex together with all added water molecules.

- **system.pqrs** containing the complete macromolecular structure/complex converted to PQRS format.

- **$job_id.log** a log file with detailed information concerning job execution.

- **config.gsl** a configuration file generated by the server.
Test set structures

A list of protein-ligand complexes used for server tests is provided in the accompanying text file. Each line contains information about a single complex in the following format: PDB id, ligand residue name, PDB chain id, ligand residue number, followed by a list of interface water molecules. If a given water molecule was identified (within 1.4 Å spatial accuracy), an estimated binding free energy (kcal/mol) is provided, otherwise "X" is given. For complexes that were automatically not processed, SQM denotes a failure ligand partial charges calculations, and PQR denotes a failure in protein parametrisation.

Interface water: solvent accessible surface distribution

Solvent accessible surface area (SASA) of interface water molecules was calculated with the VMD program\cite{humphrey1996} using its default SASA method with solvent probe of 1.4 Å radius. The distribution of SASA fraction (i.e. the actual SASA divided by SASA of a free water molecule) among crystallographic and predicted water molecules is presented in Fig. 2.

![Figure 2: SASA fraction among crystallographic and predicted water molecules.](image)

References

[1] Setny, P.; Zacharias, M. J. Phys. Chem. B 2010, 114, 8667–75.
[2] Setny, P. J. Phys. Chem. B 2015, 119, 5970–5978.
[3] Setny, P. J. Chem. Theory Comput. 2015, 11, 5961–5972.
[4] Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. J. Phys. Chem. 1987, 91, 6269–6271.
[5] Wang, J.; Cieplak, P.; Kollman, P. A. J. Comput. Chem. 2000, 21, 1049–1074.
[6] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. J. Comput. Chem. 2004, 25, 1157–1174.
[7] Jakalian, A.; Bush, B. J. Comput. Chem. 2000, 21, 132–146.
[8] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. J. Mol. Graph. Model. 2006, 25, 247–260.
[9] Jurrus, E. et al. Protein Sci. 2018, 27, 112–128.
[10] Humphrey, W.; Dalke, A.; Schulten, K. J. Mol. Graph. 1996, 14, 33–38.