Probing the role of interfacial waters in protein-DNA recognition using a hybrid implicit/explicit solvation model

Shen Li and Philip Bradley

Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle WA, 98109, USA

Supplementary Material

I. Supplementary Methods

Simulation details: Rotamer prediction simulations consisted of 10*nres_flexible Monte Carlo sidechain moves (nres_flexible = number of flexible sidechain positions). In the DNA sequence recovery simulations, 25*nbp_mutable base pair mutation moves were interspersed among these sidechain moves (nbp_mutable = number of sequence-sampled DNA base pairs). Water positions and occupancies were re-optimized after each move in the neighborhood of the perturbation and prior to evaluation of the acceptance criterion. Gradient-based minimization of nearby sidechains, bases, and waters was performed before move evaluation. Global sidechain/base/water optimizations and gradient-based minimizations were performed 10 times, at evenly spaced points within each simulation.

Modification to the hydrogen bonding potential for SP3 acceptors in Rosetta: In preliminary simulations, we found that the hydrogen bond networks among explicit waters often contained interactions with poor geometry at one of the two acceptor hydrogens. By default, Rosetta's hydrogen bonding potential assesses three geometric features: the distance between the hydrogen atom and the acceptor, the angle at the hydrogen, and the angle at the acceptor measured to a single 'acceptor base' atom chosen based on the chemical type of the acceptor. We found that calculating the acceptor-angle potential with respect to both attached hydrogen atoms of the acceptor oxygen and taking the maximum of these two energies improved geometries and gave a more realistic balance between intra-solvent and solute-solvent hydrogen bonding.

Entropy parameters combinations tested: We manually selected a range of entropy parameter combinations aimed at sampling across native-like hydration-site occupancies for several settings of the orientation-dependent parameter *H2O_hbscale*.

3W model, H2O_hbscale= 1.0, H2O_refwt= 0.9,1.1,1.3,1.5
3W model, H2O_hbscale= 0.8, H2O_refwt= 0.7,0.9,1.1,1.3
3W model, H2O_hbscale= 0.6, H2O_refwt= 0.3,0.4,0.5,0.6,0.7,0.9
3W model, H2O_hbscale= 0.4, H2O_refwt= 0.12,0.17,0.2,0.3,0.5

1W model, H2O_hbscale= 1.0, H2O_refwt= 0.7,0.9,1.1,1.3
1W model, H2O_hbscale= 0.6, H2O_refwt= 0.4,0.5,0.6

II. Benchmark Set Details

(a) The 116 PDB structures used for water and sidechain prediction assessment: 1a73, 1bc8, 1cl8, 1d02, 1dc1, 1dfm, 1dp7, 1dsz, 1e3o, 1egw, 1emh, 1fiu, 1gu4, 1h6f, 1jx4, 1k3x, 1kx5, 1l3l, 1llm, 1lmb, 1mnn, 1mus, 1nkp, 1omh, 1orn, 1owf, 1p71, 1puf, 1qna, 1sa3, 1sx5, 1sxq, 1t9i, 1tez, 1tro, 1w0u, 1wte, 1xyi, 1y8z, 1zs4, 2a07, 2bcq, 2bop, 2c7p, 2dp6, 2fmp, 2g1p, 2gb7, 2gig, 2h7g, 2han, 2hdd, 2heo, 2i13, 2ih2, 2itl, 2nll, 2nq9, 2o4a, 2oaa, 2odi, 2ofi, 2py5, 2r1j, 2ve9, 2vjv, 2vla, 2voa, 2w42, 2w7n, 2wbs, 2xhi, 2xqc, 2xzf, 3aaf, 3bam, 3bm3, 3bs1, 3cmy, 3dvo, 3e6c, 3ey1, 3eyi, 3fde, 3fdq, 3fsi, 3g00, 3g9m, 3gox, 3gpu, 3h8r, 3hts, 3i0w, 3igk, 3jxy, 3k59, 3kde, 3kxt, 3l2c, 3m4a, 3m7k, 3mfi, 3mr3, 3nci, 3ndh, 3o1t, 3oqg, 3osg, 3osn, 3pv8, 3pvi, 3py8, 3q23, 3qmd, 3sm4, 3spd

(b) The subset of 62 structures used for DNA sequence recovery calculations: 1a73, 1bc8, 1d02, 1dc1, 1dfm, 1dp7, 1dsz, 1e3o, 1egw, 1fiu, 1gu4, 1h6f, 1l3l, 1lmb, 1mnn, 1nkp, 1orn, 1owf, 1puf, 1sa3, 1sx5, 1t9i, 1tro, 1wte, 1zs4, 2a07, 2bop, 2gb7, 2gig, 2han, 2hdd, 2nll, 2nq9, 2o4a, 2oaa, 2odi, 2r1j, 2vla, 2w7n, 2wbs, 2xzf, 3bam, 3bm3, 3bs1, 3cmy, 3dvo, 3e6c, 3ey1, 3g00, 3g9m, 3gox, 3hts, 3i0w, 3igk, 3jxy, 3l2c, 3m7k, 3ndh, 3oqg, 3osg, 3pvi, 3sm4

(c) The four PDB structures and DNA sequence positions analyzed in detail:

trp **repressor**

PDB ID= 1TRO.
DNA target site positions:
Recognition positions 1-4 (1st half-site): chain I, A4-A7
Recognition positions 5-8 (2nd half site): chain I, T14-T17

Restriction endonuclease *Bam*HI

PDB ID= 2BAM
DNA target site positions: Recognition positions 1-6: chain C, G4-C9

Restriction endonuclease *Eco*RI

PDB ID= 1CKQ
DNA target site positions: Recognition positions 1-6: chain B, G5-C10

Hin Recombinase

PDB ID= 1JJ6
DNA target site positions: Recognition positions 1-3: chain A, G9-T11

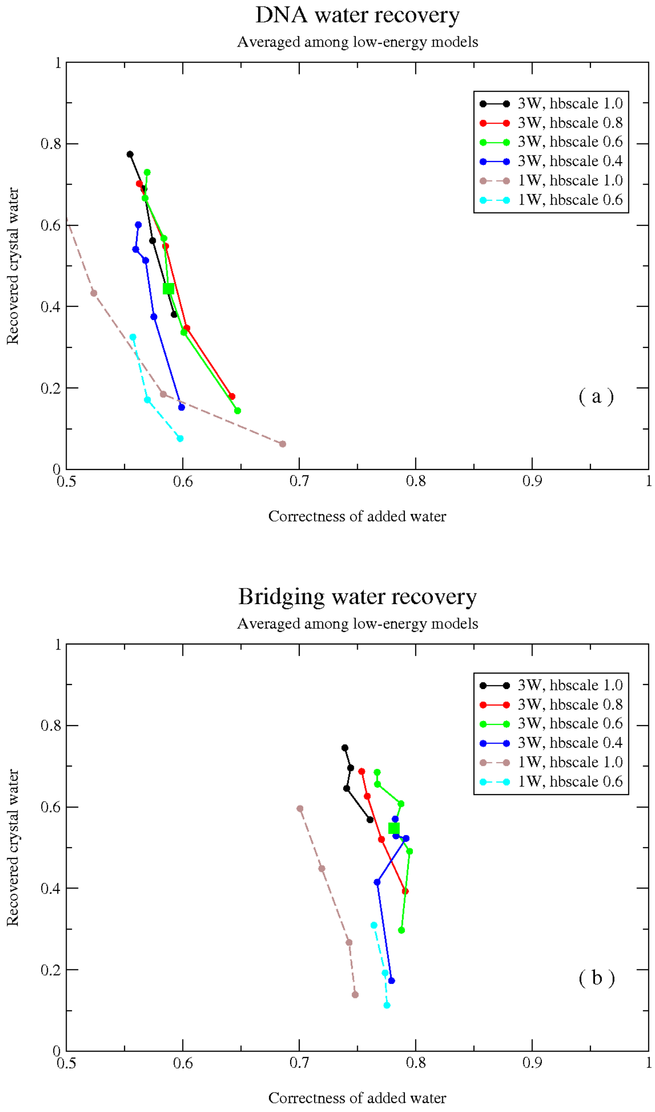

Figure S1. Prediction accuracy for DNA-bound water (a, water that directly contacts the DNA bases) and bridging water (b, water that interacts with both DNA and protein) from low-energy simulations with 3-site ('3W') and single-site ('1W') waters with a range of entropy parameters. The fraction of modeled waters that overlap with crystal waters (x-axis, 1.4Å distance threshold) is plotted against the fraction of crystal waters that overlap with modeled water (y-axis, same threshold). Recovery is averaged across the 20% lowest-energy models for each member of the full benchmark set. Comparing with main text Fig. 2, we see that enriching by energy does not appear to dramatically improve performance, in contrast to the protein sidechain and DNA sequence recovery results.

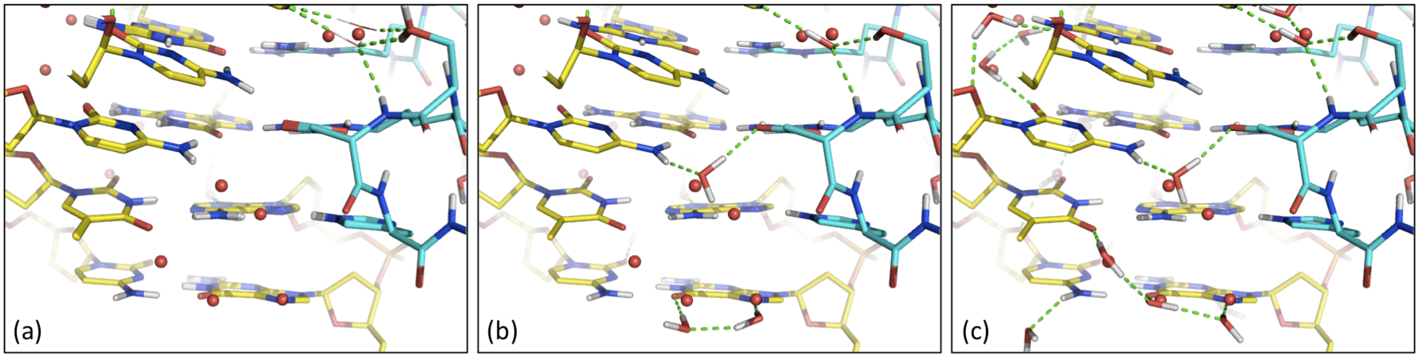

Figure S2. Hydration site occupancy increases as the fixed cost of introducing a water (the *H2O_refwt* parameter) decreases from relatively high (a) to relatively low (c) values. Native waters are shown as red spheres; modeled waters and the protein and DNA are shown in stick representation.

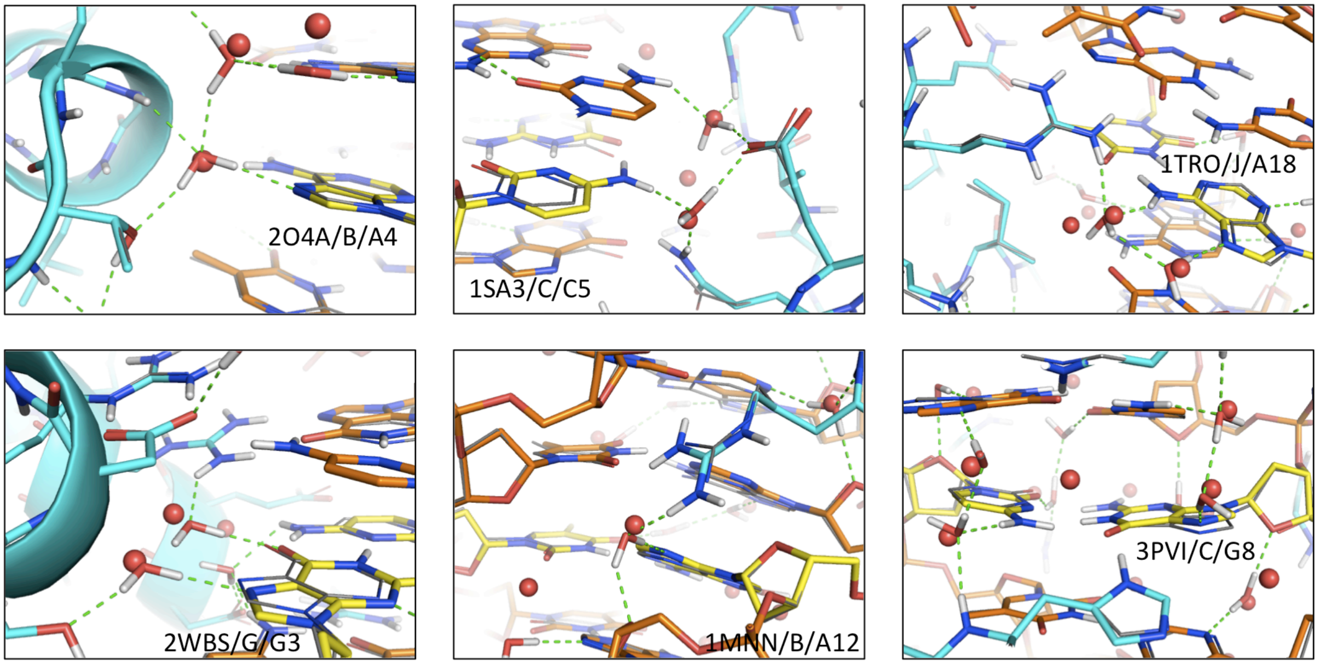

Figure S3. The six DNA sequence positions with the largest improvement in recovery upon addition of explicit water ('3W' constrained simulations). Crystal structure conformations are shown in thin lines; modeled conformations are shown in stick representation. In each case, the improvement is associated with one or more bridging waters observed in the crystal structure (red spheres), which are successfully recapitulated in the models (stick representation) and likely contribute to the improvement in recovery. For each case, the base pair of interest is shown in yellow, and labeled "<PDB ID>/<chain>/<base><resnum>". Hydrogen bonds to modeled waters are shown as green dashed lines.