Relative solvation free energies of amino acid side chains using Bennett's method

Stephen Lillington

December 12, 2019

Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara, CA 93106

**Abstract:** Free energy calculations are a powerful molecular simulation technique with many applications to studying biomolecular systems. Such techniques are useful for predicting protein-ligand binding properties as well as understanding the mechanisms by which amino acid mutations alter a protein's behavior. Critical to correctly measuring the effects of amino acid mutations on a folded peptide's stability are accurate estimations of amino acid solvation energies. Here, I attempt to match experimental and previous results for the relative solvation free energies of alanine, isoleucine, and serine as members of a capped GXG tripeptide, representing the unfolded protein state. Using the Bennett Acceptance Ratio method, I obtain estimates which are in poor agreement for isoleucine, but good agreement for serine. Potential undersampling effects resulting in poor agreement are further discussed.

**Introduction:** The ability to predict via computational methods the effects of amino acid mutations on the properties of a protein has widespread applications for drug discovery and protein engineering. While high throughput library creation and screening (directed evolution) methods have been very successful in experimentally identifying beneficial amino acid substitutions for a number of applications, designing protein-protein interactions is much more challenging due to the difficulty in designing an adequate screen. Free energy calculations coupled with *in silico* amino acid mutation present an attractive method to address this gap, enabling the high throughput computational screening of mutants for desirable protein binding properties such as free energies of binding. Computing relative binding properties of whole proteins, however, is beyond the scope of the project because the simulation time required to ensure adequate equilibration and sampling of each alchemical state is very high. Instead, I compute the relative solvation free energies of different amino acids by modeling solvated and *in vacuo* GXG peptides, where X is mutated from alanine to isoleucine. Such quantities are still very valuable to compute given their contribution to the effects of residue mutations on protein stability[1].

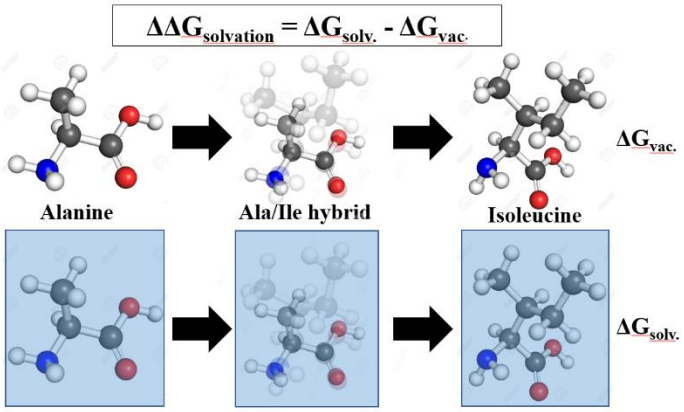


**Figure 1.** The overall goal of this work is to implement an alchemical simulation workflow to estimate the relative solvation free energies of Alanine to other amino acids, e.g. isoleucine. These quantities are calculated by running two sets of simulations - one in solvent and one in vacuum.

The most widely used simulations-based approach to computing relative free energy changes in protein systems is the alchemical method[2], which couples interaction energies to a state parameter, λ, which varies from 0 to 1 to completely turn on or off certain interactions between atoms in a system. Generally, average system energies (or $\langle \frac{dU(or\ H)}{d\lambda} \rangle$) are collected for equilibrium states spanning λ = [0,1] and are used to estimate a free energy difference between the two states by a method such as histogram reweighting[3], thermodynamic integration[4], or Bennett's Acceptance Ratio (BAR) method[5].

**Methods:** The first step in setting up these simulations was the construction of topology files for the start and end state. For these simulations, a dual topology approach was used, in which all state topologies have a capped G(X/Y)G peptide with a middle hybrid residue that contains all atoms from both the start and end sidechain. Dummy atoms are designated to make the residue equivalent to residue X in state A and residue Y in state B. As a result, all inter- and intramolecular interactions involving the changing sidechain atoms are either turned completely 'ON' or completely 'OFF' from state A to state B.

Practically, doing this requires the addition of hybrid residues with Lennard-Jones parameters, partial charges, and masses to an existing force field. It further requires the construction of topology files capturing the dual topology system setup. Fortunately, topologies and force field parameters for all possible GXG mutations in several force fields (AMBER99SB, AMBER99SB*ILDN, Charmm36, Charmm22, and OPLS AA/L) were freely available from the de Groot lab at the Max Planck Institute[6].

**System setup:** The constructed topology file for the alanine-isoleucine mutation as well as the Amber99SB force field from their work were used in this project. All energy minimization and dynamics were done in GROMACS. For the Ala → Ile mutation, the atoms unique to Ile are designated dummy atoms. See Ref. 6 for details on how the topologies were constructed. A cubic simulation box (3.559 nm per side) with periodic boundary conditions centered around the peptide was solvated with 1418 simple point charge (TIP3P) water molecules.

**Simulation parameters**: The GROMACS simulation package was used to perform energy minimization and molecular dynamics. When free energy differences were estimated using slow growth thermodynamic integration, only one system was minimized and equilibrated prior to production MD. To set up free energy calculations using BAR, multiple systems with differing values of λ were constructed and minimized and equilibrated prior to production MD of each system independently.

Non-bonded interactions: Lennard-Jones interactions were implemented with a potential switching function such that $U_{LJ}$ was shifted smoothly to 0 between 1.0 nm and 1.2 nm (where $U_{LJ} = 0$). A Verlet cutoff scheme for generating interacting pair lists was used with a cutoff distance of 1.0 nm. Dispersion corrections to the energy and pressure were applied to compensate for cutoff-induced errors.

Electrostatics were modeled using Particle Mesh Ewald (PME) with 0.1 nm Fourier spacing and an interpolation order of 6. A cutoff of 1.2 nm was used for these interactions.

Temperature and Pressure coupling: NVT, NPT, and production simulations were performed using Langevin dynamics with a friction coefficient of 0.5 $ps^{-1}$. When required, constant pressure was maintained using the Parrinello-Rahman barostat with compressibility set at 4.5e-5 $bar^{-1}$ and a time constant of 1 ps for pressure coupling.

Constraints: For all equilibration and production MD, all bonds were constrained using the LINCS algorithm. This is necessary because bond stretching contributions need to be explicitly accounted for in free energy calculations.

**System minimization, equilibration, and production simulation:** At each value of λ, the system was energy minimized first by steepest descent for 100 steps followed by 100 steps of conjugate gradient minimization. Though max force tolerances of 1.0e-5 were set, these were never achieved for the solvated or vacuum system – the maximum force on any particle was < 1000 N in all simulation starting configurations. Following energy minimization, a 100 ps NVT equilibration using Langevin dynamics with a friction coefficient of 0.5 $ps^{-1}$ was performed. An additional 500 ps NPT equilibration was performed prior to a 3 ns production period for data collection. During each production simulation, changes in the Hamiltonian $\Delta H(\lambda_1, \lambda_2)$ for adjacent λ values were collected every 1000 time steps for use in the free energy calculation.

**Free energy estimation:** Both non-bonded and bonded interactions involving the alchemical atoms were parametrized by λ, yielding potentials of the form:

Bond, angle, improper torsion potential:

$$U_b = \frac{1}{2}[(1-\lambda)k_b^A + \lambda k_b^B][b - (1-\lambda)b_0^A - \lambda b_0^B]^2 \qquad (1)$$

$k_b^i$ are force constants for the bond/angle/improper torsion and $b_0^i$ are the equilibrium bond lengths/angles.

Proper torsion potential:

$$U_d = [(1-\lambda)k_d^A + \lambda k_d^B]\left(1 + \cos\left[n_\phi \phi - (1-\lambda)\phi_s^A - \lambda \phi_s^B\right]\right) \qquad (2)$$

$k_d^i$ are force constants for the dihedral, $\phi_s^i$ is the equilibrium angle in state I, and $n_\phi$ is an integer.

Softcore potentials:

$$U_{sc}(r) = (1-\lambda)V^A(r_A) + \lambda V^B(r_B) \qquad (3)$$

$$r_A = (\alpha\sigma_A^6\lambda^p + r^6)^{\frac{1}{6}} \qquad (4)$$

$$r_B = (\alpha\sigma_B^6(1-\lambda)^p + r^6)^{\frac{1}{6}} \qquad (5)$$

Here, α was set to 0.5 and p set to 1. $\sigma_i$ is the van der Waals radius of an atom in state i. In simulations, it is set as a parameter to 0.3 to be the radius for hydrogen atoms, which cause issues with no LJ interaction. $V^i(r_i)$ is the standard form of the Lennard-Jones or Coulomb potential.

Since atoms are disappearing and appearing at the same time in the alanine to isoleucine transformation, Lennard Jones and electrostatic interactions had to be transformed at the same time, necessitating a softcore form for the electrostatic potential in contrast to the conventionally used linear scaling. All interactions were decoupled at the same rate. Initially, λ values in increments of 0.1 between 0.0 and 1.0 were used, but these values were too few for the desired precision of ± 0.2 kJ/mol. The final set of values used for all decoupling was λ = [0.0, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.0]. The estimated free energy change between adjacent states was computed using the BAR algorithm:

$$\beta\Delta G = ln\frac{\langle\frac{1}{1+e^{-\beta\Delta H(\boldsymbol{p}^N,\boldsymbol{q}^N)+\beta\Delta G}}\rangle_0}{\langle\frac{1}{1+e^{\beta\Delta H(\boldsymbol{p}^N,\boldsymbol{q}^N)-\beta\Delta G}}\rangle_1};\ \Delta H = H(\lambda_1-\lambda_0) \qquad (6)$$

This equation was solved iteratively using the bisect method. I'll note that while I did write my own BAR implementation for processing trajectories from OpenMM, I did not have time to recode this for processing Gromacs output and used the Gromacs BAR implementation to get my results. My implementation along with a script comparing it to standard implementations for computing the chemical potential of liquid argon is included in my submission.

An error estimate for each adjacent ΔG was computed by splitting each 3 ns production simulation into five blocks, yielding five ΔG values from which the average and variance were computed. Standard deviations were successively propagated as $\sigma_{sum} = \sqrt{\sigma_1^2 + \sigma_2^2}$.

We then get the relative solvation free energy from Eqn. 7.

$$\Delta\Delta G_{solvation,A\rightarrow I} = \Delta G_{solvated,A\rightarrow I} - \Delta G_{vacuum,A\rightarrow I} \qquad (7)$$

It wasn't possible to barostat the vacuum system composed of only once molecule without stability issues, so production simulations were done in the NVT ensemble. In the vacuum/ideal gas state, the PV term is invariantly equal to RT, cancelling out in the free energy difference expression, so $\Delta A_{vacuum,A\rightarrow I} = \Delta G_{vacuum,A\rightarrow I}$ .

**Table 1.** Comparing my results to experiment and prior literature shows an unfortunate lack of agreement for the relative free energy of solvation of isoleucine compared to alanine, but surprisingly good agreement for the less dramatic Ala to Ser transformation. All energies are in units of kT. Errors are one standard deviation.

| Mutation | Experiment | Literature | This work (BAR) | This work (My BAR) | This work (Slow growth) |
|---|---|---|---|---|---|
| **GAG to GIG** | 0.21 ± 0.02 – 0.05[7] | 0.19 ± 0.02 – 0.05[7] | -0.59 ± 0.16 | -0.88 ± 0.16 | 1.84 ± 1.81 |
| **GAG to GSG** | -7.00 ± 0.02 – 0.05[7] | -6.75 ± 0.02 – 0.05[7] | -7.19 ± 0.21 | -6.58 ± 0.51 | N/A |

**Results**: The key result of this work is the predicted relative solvation free energy of two amino acid residues as they appear as part of a polypeptide chain. Table 1 contains this result, with comparisons of my measured value using both BAR and slow growth to both experimental and prior simulation values. Note that the previous study estimating the relative solvation free energies

of amino acid side chains did so using the chemical analog of the isolated sidechains (e.g. n-butane for isoleucine), not a capped GXG peptide.

In the process of generating these results, I experienced firsthand the tricky nature of achieving high precision in free energy simulations. For a system this large (>5000 atoms in the solvated case), undersampling is a potential pitfall, as is the inadequate phase space overlap of simulations at different λ values used to estimate ΔG. Both errors were apparent in my first attempt at computing $\Delta\Delta G_{solvation,A\rightarrow I}$, and are especially evident in the slow-growth estimate. Unfortunately, it is likely that I needed to increase the number of intermediate states samples even further for GAG to GIG, particularly towards λ = 1 (Figure 2). In contrast, free energy changes between intermediate states were much smaller for the alanine to serine mutation (Figure 2).

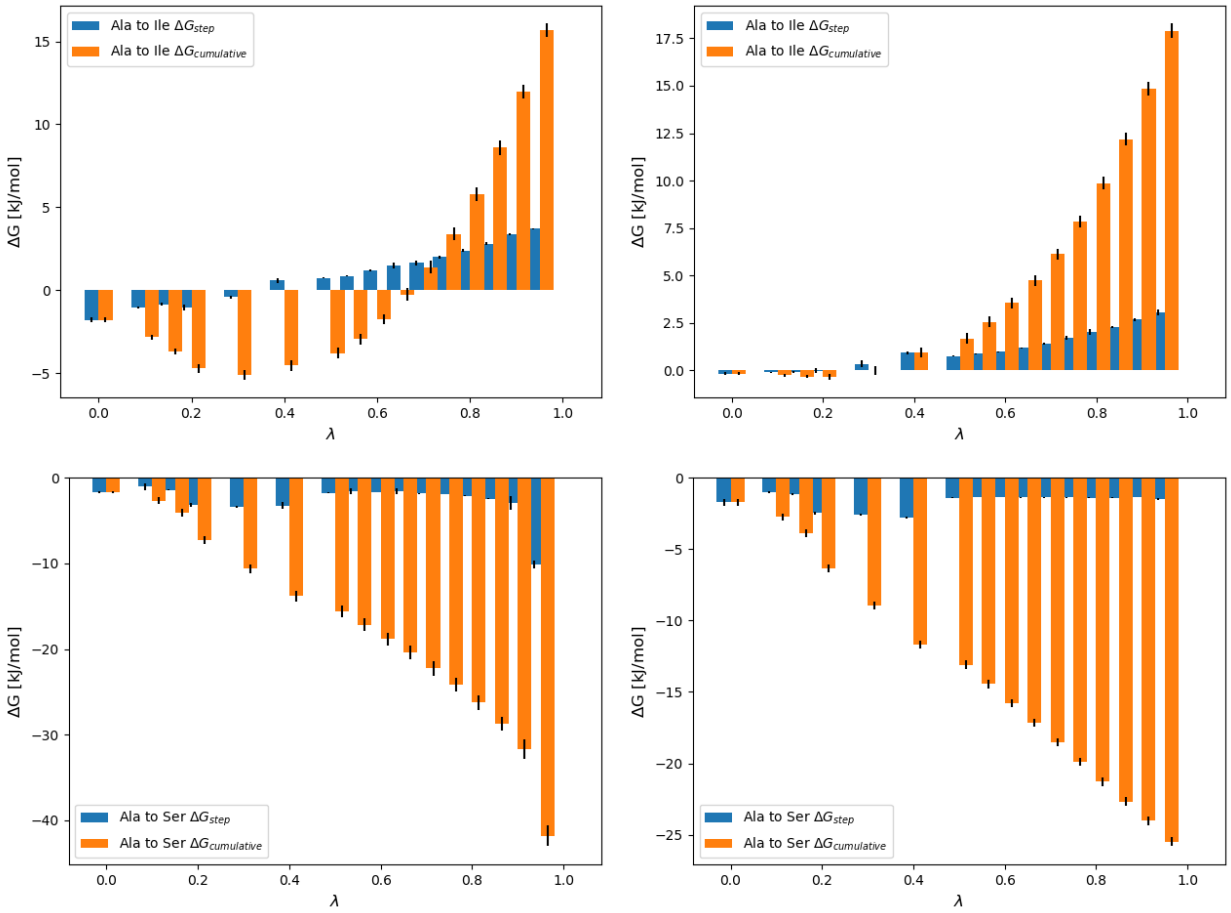


**Figure 2.** (Left) Stepwise and cumulative delta G for GAG to GIG (Top) and GAG to GSG (Bottom) in solvated system. (Right) Same metrics plotted for the vacuum system. Error bars are ± one standard deviation.

**Discussion:** In addition to simulation/sampling quality, there are many possible reasons why free energy estimates may not agree with prior work or experiment. For example, a different force field with different parameters, different implementations of the non-bonded interactions, and of course the fact that my values are calculated for a capped GXG peptide in contrast to the isolated sidechain molecule. Ultimately, my goal in this project was to obtain statistically high quality free energy estimates by using good simulation practices. Obtaining precise estimates should undermine the

possibility that any observed differences between my results and others are due to undersampling or an unequilibrated system. However, the relative failure and success of the Ala to Ile and Ala to Ser mutations while obtaining similar levels of precision is puzzling.

Assuming the fidelity of the force field parameters and the way interactions were calculated, the most apparent source of error is the relative "difficulty" of the transformations taking place. Alanine to isoleucine involves the appearance of three C atoms with accompanying H's, in contrast to alanine to serine which requires only the appearance of an OH group. More difficult transitions will require more intermediate states, as evidenced by the higher stepwise free energy changes in Figure 2 compared to Figure 3. I expect that adding intermediate states to keep stepwise $\Delta G$'s $\leq$ 0.5 kT would improve the accuracy of my $\Delta\Delta G_{solv., A \rightarrow I}$. Future work diving deeper into what drives these relative free energy changes (e.g. enthalpic or entropic effects) will make a nice addition to these efforts as well.

**Conclusions:** In this work, the relative solvation free energies of two peptides, GIG and GSG are estimated using explicit solvent molecular dynamics simulations and the Bennett Acceptance Ratio method of free energy estimation. Significant inaccuracy of the GIG estimate compared to the GSG estimate suggests that inadequate phase space overlap between intermediate states in the GIG, likely exacerbated by the large number of alchemical atoms simultaneously changing, is the primary reason for this inaccuracy, though I cannot rule out issues in the force field parametrization. The fact that the free energy estimates were obtained with similar levels of precision, however, undermines this argument and warrants further investigation.

**References**


[1] D. Seeliger and B.L. de Groot, Biophys. J. **98**, 2309 (2010).

[2] T. Simonson, G. Archontis, and M. Karplus, Acc. Chem. Res. **35**, 430 (2002).

[3] S. Kumar, J.M. Rosenberg, D. Bouzida, R.H. Swendsen, and P.A. Kollman, J. Comput. Chem. **13**, 1011 (1992).

[4] J. Hermans, J. Phys. Chem. **95**, 9029 (1991).

[5] C.H. Bennett, J. Comput. Phys. **22**, 245 (1976).

[6] V. Gapsys, S. Michielssens, D. Seeliger, and B.L. De Groot, J. Comput. Chem. **36**, 348 (2015).

[7] M.R. Shirts and V.S. Pande, J. Chem. Phys. **122**, 1 (2005).