Lab on Host-Guest Binding Free Energies: Data Analysis and Discussion

Emilio Gallicchio, emilio@biomaps.rutgers.edu
Mauro Lapelosa, maurolap@eden.rutgers.edu

April 26, 2011

Theory

Binding free energies are computed using the Binding Energy Distribution Analysis Method (BEDAM) [Gallicchio, Lapelosa, Levy, *J. Chem. Theor. Comput.* 6, 2961-2977 (2010)]. Briefly, the BEDAM method computes the binding free energy ΔG_b° by means of the equation

$$\Delta G_b^\circ = -kT \ln \left[C^\circ V_{\text{site}} \int du\, p_0(u) e^{-\beta u} \right], \tag{1}$$

where $\beta = 1/k_B T$, C° is the standard concentration of ligand molecules, V_{site} is the volume of the binding site, and $p_0(u)$ is the probability distribution of binding energies collected in the special ensemble of conformations of the protein-ligand complex in which the ligand is present in the binding site and the receptor and the ligand are not interacting. The binding energy u is defined for each conformation of the complex as the difference between the effective energies of the bound and separated conformations of the complex without conformational rearrangements:

$$u(x_R, x_L, \zeta_L) = U_{RL}(x_R, x_L, \zeta_L) - [U_R(x_R) + U_L(x_L)], \tag{2}$$

where U_{RL} is the effective potential energy of the complex, x_R and x_L are the internal coordinates of the receptor and ligand, respectively, and ζ_L represents the position and orientation of the ligand relative to the receptor. $U_R(x_R) + U_L(x_L)$ is the sum of the effective potential energies of the receptor and ligand when these are not interacting.

The accurate calculation of the important low binding energy tail of $p_0(u)$ can not be accomplished by brute-force collection of binding energy values from a simulation of the non-interacting complex because this is rarely sampled when the ligand is not guided by the interactions with the receptor. Instead we use parallel Hamiltonian Replica Exchange (HREM) sampling in which swarms of coupled replicas of the system, differing in the value of an interaction parameter $0 \le \lambda \le 1$ controlling the strength of ligand-receptor interactions, are simulated. Each replica uses a $\lambda-$dependent potential of the form:

$$U_\lambda(x_R, x_L, \zeta_L) = U_R(x_R) + U_L(x_L) + \lambda u(x_R, x_L, \zeta_L) \tag{3}$$

The replicas collectively sample a wide range of unfavorable, intermediate and favorable binding energies which are unbiased and combined together by means of the Weighted Histogram Analysis Method (WHAM).

BEDAM employs the OPLS-AA force field and the AGBNP2 implicit solvent model [Gallicchio et al., *J. Chem. Theor. Comput.* **5**, 2544-2564 (2009)], which is based on the Generalized Born electrostatic model and a non-polar hydration free energy estimator.

β-Cyclodextrin Host-Guest Systems

β-cyclodextrin (Fig. 1) is a naturally-occurring cyclic oligosaccharide made of seven maltose monomers. The 3D shape of the molecule can be described as a truncated cone with a cavity in the middle. The narrow

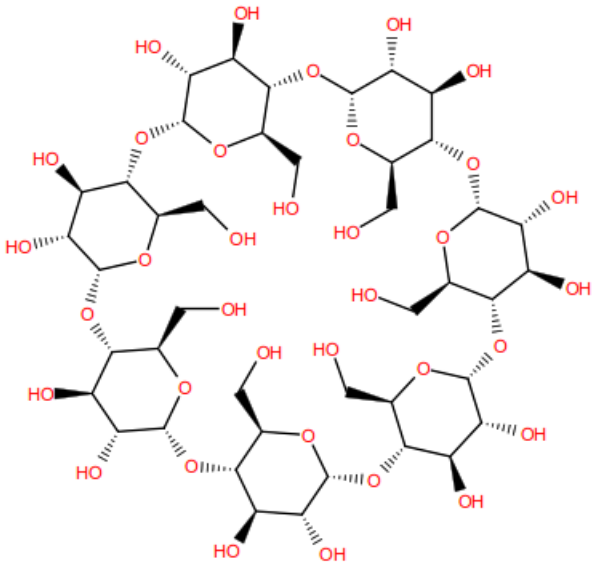

Figure 1: Chemical structure of the β-cyclodextrin host molecule.

entrance to the cavity is surrounded by seven primary hydroxy (OH) groups, whereas the wider rim is is surrounded by fourteen secondary OH groups. Cyclodextrins bind a range of hydrophobic and polar guest molecules in water [see: Rekharsky & Inoue, "complexation thermodynamics of cyclodextrins", *Chem. Rev.*, 98, 1875-1917, (1998)]. For this reason cyclodextrins find applications in drug and food formulations and other areas.

In this lab we will be computing the binding constants (or equivalently the standard binding free energy) of β-cyclodextrin with a series of small molecules using the BEDAM method. We will compare the computational results with experimental data from the compilation of Rekharsky & Inoue. Despite the large quantity of affinity data, little is known about the structures of cyclodextrin host-guest complexes. Rekharsky & Inoue make a number of structural hypotheses based on the pattern of affinity data. We will have access to detailed structural data from the simulations. We will attempt to confirm their hypotheses concerning guest properties such as aliphatic chain length, primary vs. secondary alcohols, aromaticity, cyclization and rigidity. Most of these guest-host systems have not been modelled before. So in a way we will be doing actual novel research.

Simulation Post-Processing

Assuming your BEDAM Replica Exchange calculation has completed on the hugin cluster, analyze the results by going to the working directory on hugin corresponding to the ligand simulation in question, and setting the `SCHRODINGER` environment variable:

```
ssh hugin
cd /net/huginp/u1/<myusername>/<workdir>
export SCHRODINGER=~emilio/schrod/r2012/b1
```

where "`<myusername>`" is your user name and "`<workdir>`" the directory. To extract the binding energies and at the same time compute the binding free energy do:

```
   $SCHRODINGER/run ~emilio/utils/bedam_scripts/bedam_analyze.py \
<jobname>.inp > results.log
```

where "`<jobname>.inp`" is the control file you used to prepare and run the BEDAM calculation (the "\" symbol at the end of the first line is a command line continuation character; just type the full command

on one line). Results are saved in "`results.log`" (or any other log file name you choose). The computed standard binding free energy is reported at the end of this file in kcal/mol plus its estimated uncertainty (one σ). For example:

```
DG(binding) = -2.115927 +- 0.058787 kcal/mol
```

Towards the the end of this file you can also find the values of the WHAM dimensionless free energies f_k as a function of λ:

```
   f_k = [ 0.  0.57360101 0.77648056 0.89487705 0.9243753 0.94830035 0.98696658 1.0181753
1.12446893 1.2505684 1.37202725 1.45939328 1.67777569 1.39997478 -0.31478005 -4.1579129
]
```

The binding free energy at λ is given by $\Delta G_b(\lambda_k) = k_B T f_k$. The first number corresponds to the binding free energy at $\lambda = 0$, which is zero by definition. The last corresponds to the physical binding free energy. Note that these values do not include the binding site volume correction term $-k_B T \ln C^\circ V_{\text{site}}$. This quantity is reported in kcal/mol in the log file, for example:

```
grep dgvsite results.log
dgvsite = 0.361619
```

Note that$-k_B T \ln C^\circ V_{\text{site}}$ is included in the final value of the "`DG(binding)`" binding free energy value above. The initial part of the log file contains the results of the WHAM iterative process to determine the f-values.

The analysis script also generates a directory (called `a/`) with histograms and time trajectories of binding energy data. The directory `a/be_hist/` contains histograms of binding energies at each value of λ. For example the file `h_1.000000.dat` contains the binding energy histogram at $\lambda = 1$. The first column lists the bin edges (in kcal/mol) and the second column is the number of binding energy samples collected in that bin. Note that the binding energy bins vary in width. Also, binding energy values are capped at $1,000$ kcal/mol.

The directory `a/be_hist/` also contains a file called `p0.dat` which contains the distribution of binding energies $p_0(u)$ at $\lambda = 0$ as computed from WHAM. This is the function that plugged into Eq. (1) yields the binding free energy. The first column in this file lists again the locations of the binding energy bins. The second column is the probability p_i of that bin and the third column is the estimate of the probability density $p_0(u_i) = p_i/\Delta u_i$, where Δu_i is the width of the bin. A plot of the second column of `p0.dat` should resemble the histogram in `h_0.000000.dat`, however the data in `p0.dat` is much more accurate at small u.

The directory `a/lbe_trj/` contains the time trajectories of binding energies and λ values for each BEDAM simulation. These are mostly used for debugging. We hope not to have to inspect them.

Extract simulation trajectories by typing:

```
$SCHRODINGER/impact -i <jobname>_readtraj.inp -HOST hugin -LOCAL
```

This will create Maestro structure files named `<jobname>_remd_trj_l<lambda>.maegz` where `<lambda>` is the value of λ for the corresponding replica. These files can be visualized in Maestro. To view trajectories and complete the analysis, copy the simulation results to your workstation. From the workstation terminal type:

```
rsync -avz hugin:~/<workdir> .
```

where `<workdir>` is the remote job directory on hugin.

Discussion

You are likely the first person to have performed these calculations. We are not sure exactly what to expect from the model. Comparison with the experiments is not an absolute indication of whether the calculation is "right" or "wrong". Experimental conditions may differ by those implied by the model. Experiments are not perfect either, artifacts such as aggregation could affect the results. The report will be judged on the level

of effort, and statistical thermodynamics and chemical insight; not on the "accuracy" of the results. Focus the discussion on things that you are able to learn from the calculations and that were not available to the experimentalists. Please address the following points when preparing your report:

1. Compare the computed standard binding free energies for benzene and your assigned guest with the experimental measurements in the review paper by Rekharsky & Inoue [*Chem. Rev.*, 98, 1875-1917, (1998)]. Comment on the accuracy of the calculation. How does it compare with the range of experimental affinities? Is the computed accuracy sufficient to discriminate benzene from your assigned guest? Obtain computed binding free energies values from your fellow students and compare them to the experimental measurements. Comment on the overall accuracy and the ability of the method to discriminate classes of compounds. Compute the correlation coefficient and Spearman rank order correlation coefficient between calculated and experimental binding free energies.

2. Analyze the trajectory of the assigned guest at $\lambda = 1$ (fully coupled state). Show a picture of the most representative conformation of the complex, that is the one that tends to occur most often. Describe the position and orientation of the guest relative to the host. Describe the interactions between the host and the guest in terms of, for example, hydrogen bonds and hydrophobic contacts. Does this structure confirm the hypotheses formulated by Rekharsky & Inoue? Is the complex best described by a single conformation or by multiple conformations? Examine the structures of similar complexes obtained by your fellow students. Do they overall confirm the hypothesis of Rekharsky & Inoue? If not, can you formulate a new structural explanation for the observed affinities?

3. Show the histogram of binding energies at $\lambda = 1$ for benzene and your assigned guest. Comment on the relative position, width, and shape of the histogram. Which guest presents more favorable binding energies? By how much? Is this reflected in their relative binding free energies? Compute the average binding energies $\langle u \rangle_1$ at $\lambda = 1$ and their corresponding reorganization free energies $\Delta G^\circ_{\text{reorg}} = \Delta G^\circ_b - \langle u \rangle_1$. Which of the two guests loses more reorganization free energy in order to bind the host? Rationalize this finding in terms of the structure of the guest and the structure of host-guest complex. Does this result fit with the hypothesis of Rekharsky & Inoue?

Finally, feel free to include in the discussion any aspect of the theory and calculations that you find interesting or intriguing. As described above binding free energies, binding energy data, $p_0(u)$ values, and trajectories at all λ values are available for you to study and discuss.