3. Binding affinity estimation using the Linear Interaction Energy method

Martin Almlöf, Martin Andér, Sinisa Bjelic, Jens Carlsson,
Hugo Gutiérrez de Terán, Martin Nervall, Stefan Trobro and Fredrik Österberg

18th August 2005

Contents

1 Introduction **2**
1.1 Linear Interaction Energy method 2

2 Methods: Prepare and run the MD simulations **3**
2.1 Protein simulation 3
2.2 Water simulation 4

3 Results: Evaluating the simulations **5**
3.1 Analysis of P450cam binding of CAM and CMA 5

1 Introduction

In this practical the free energy of binding will be calculated for two ligands, camphor (CAM) and camphane (CMA), to the P450cam receptor using molecular dynamics (MD) together with the Linear Interaction Energy (LIE) method. The example is taken from the paper Almlöf *et al*, J. Comput. Chem., 25, 1242-1254,(2004), but with some simplifications and automation of the steps.

1.1 Linear Interaction Energy method

The LIE method for the calculation of the *absolute* free energy of binding ($\Delta G_{\rm bind}$) was first proposed by Åqvist *et al* in 1994. It is a semi-empirical method, which is computationally less expensive than the rigorous free energy perturbation (FEP) method used in the previous practical. The LIE approach is similar to FEP, but the method only requires simulations of the the free and bound ligand, *i.e.* it does not involve any transformation process as in FEP. LIE also overcomes some of the difficulties associated with a FEP calculation, such as convergence. The estimated free energy of binding is calculated as a linear combination of the differences in the average ligand-surrounding (l-s) interactions in water and protein. Interaction energies are split into electrostatic (*el*) and van der Waals (*vdw*) terms, and weighted by different factors

$$\Delta G_{\rm bind} = \alpha(\langle V_{l-s}^{vdw}\rangle_{\rm p} - \langle V_{l-s}^{vdw}\rangle_{\rm w}) + \beta(\langle V_{l-s}^{el}\rangle_{\rm p} - \langle V_{l-s}^{el}\rangle_{\rm w}) + \gamma \qquad (1)$$

where the brackets denote thermal averages sampled during the MD simulations of the protein (p) and water (w) environments. The main idea of the method is to consider polar and non-polar contributions to the free energy separately.

- **The polar contribution**, $\beta\Delta\langle V_{l-s}^{el}\rangle$: The scaling factor β=0.5 is theoretically derived from electrostatic linear response theory and yields very good agreement with experimental solvation energies for ionic solutes. For uncharged compounds, FEP calculations have shown that lower β values are necessary and can be assigned by a simple scheme depending on the ligand's chemical nature. For the ligands considered here β=0.43 (group: neutral ligand with no hydroxyl groups).

- **The non polar contribution**, $\alpha\Delta\langle V_{l-s}^{vdw}\rangle$: The non-polar contributions to the free energy of binding are assumed to have a linear relationship with the surrounding van der Waals energies. The scaling factor, α, is empirical, but a value of 0.18 has worked well for a large number of systems, including the system considered here.

  An additional constant, γ, may be required to reproduce absolute binding free energies and is dependent on the hydrophobicity of the binding site, *i.e.* specific for a given protein but conserved on any series of ligands studied on that protein. For P450cam the optimal value of this constant term is γ=-4.5 kcal/mol.

2 Methods: Prepare and run the MD simulations

Two separate MD simulations of the ligand ("water" and "protein" simulations) have to be carried out to get an estimate of the free energy of binding using LIE.

2.1 Protein simulation

Go to the directory named "LIE/protein", where one directory for each of the ligands ("CAM" and "CMA") has been created. In each of these you will find

- The coordinate file (`complex.pdb`) prepared as described in the first practical. For CAM and CMA the coordinates were taken from the files with PDB codes 6CPP and 2CCP.
- The topology `lig.top` generated by `Qprep5` as shown in the first practical, the input files `name.inp`, and a FEP file `lig.fep`.
  - Go to the directory named "CAM". Type `pymol complex.pse` in your terminal window to see the structure of the protein-ligand complex. Repeat this procedure for "CMA".
  - Compare the FEP file for LIE to the one used in the free energy perturbation practical. What are the differences between the files? The ligand-surrounding energies will be evaluated for the group of atoms specified in the FEP file. These are also called "Q atoms" by the program.

The MD simulation is divided into two phases:

- **Equilibration phase**: The equilibration phase consists of stepwise heating from 1 to 300 K, with heavy solute atoms restrained to their crystallographic positions. The equilibration scheme is divided into five steps `eq1.inp` to `eq5.inp` that are similar to that used in the first practical.
  - Compare `eq5.inp` for the protein simulations of CAM and CMA. There is one difference between these files. What? (HINT: For the CMA ligand there is a conserved water bound to the HEME group, but not for CAM)
- **Production phase**: The MD simulation is divided into 15 consecutive blocks named `md1.inp` to `md15.inp`.
  - How long is the total simulation (in ps)?

Running the MD simulations would take about 10 hours. In order to save time the output files have been prepared in the directory "results" for each ligand.

2.2 Water simulation

Now move to the directory "LIE/water", where one directory for each ligand has been created. In these you will find

- The coordinates of the ligand `ligand.pdb` extracted from the original PBD files, but with all the protein atoms removed.
- The topology generated by `Qprep5` with the ligand positioned in the center of a sphere of radius 18Å filled with water molecules, the input files, and a `FEP` file. All files are named according to the same convention as for the protein simulation.
  - Move to the "CAM" or "CMA" directory. Type `pymol ligand.pse` in your terminal window. The ligand is positioned in the center of the sphere and surrounded by water.

As previously the MD simulation is divided into two phases.

- **Equilibration phase**: The equilibration phase (`eq1.inp` to `eq5.inp`) is similar to the protein simulation, but one important change has been made. Since there is no protein present in the simulation, only a restraint that keeps the center of mass of the ligand in the sphere center has been added:

```
[sequence_restraints]
1 27 10 1 2
```

- **Production phase**: The production phase is basically the same as with the complex. The restraint that keeps the ligand in the center of the sphere is maintained to ensure a correct solvation of the ligand.

Again, running the MD simulations would take about 10 hours. In order to save time the output files have been prepared in the directory "results" for each ligand.

3 Results: Evaluating the simulations

Now we will evaluate energies from the MD simulations, the conformations of the system and specific ligand-protein interactions.

3.1 Analysis of P450cam binding of CAM and CMA

- **Convergence:** First we will look at the convergence and results of the ligand-surrounding energies for the ligands in water and protein. By using the `perl` script `energies.pl`, the *electrostatic* and *van der Waals* energies are extracted from the `name.log` files.

  - Move to "protein/CAM/results" directory.
  - Open a `name.log` file and try to find the ligand-surrounding energies for the water and protein simulations.
  - Now run the script `perl energies.pl md 1 15`, which extracts, summarizes and plots all ligand-surrounding energies for the entire production phase (`md1.log` to `md15.log`).
    - Are there any large changes in the energies throughout the simulations?
    - The script has also calculated the difference between the average energies for the first and second halves of the simulation. This can be considered as a measure of the convergence error. Have the simulations converged?
  - Repeat this procedure for all simulations and summarize the ligand-surrounding energies in Table 1.

Ligand	$\langle V_{l-s}^{vdw} \rangle_{\mathrm{p}}$	$\langle V_{l-s}^{vdw} \rangle_{\mathrm{w}}$	$\langle V_{l-s}^{el} \rangle_{\mathrm{p}}$	$\langle V_{l-s}^{el} \rangle_{\mathrm{w}}$
CAM				
CMA				

Table 1: Ligand-surrounding energies for CAM and CMA

- **Binding free energies:** The calculated ligands-surrounding energies can be used to estimate the free energy of binding.

  - Calculate the LIE binding free energies using the parameterization; α=0.18, β=0.43 and γ = -4.5 and summarize these into Table 2.
  - Are the calculated binding free energies in agreement with experiment? Are the ligands correctly ranked?
  - Is the binding due to electrostatic or hydrophobic interactions?

Ligand	$\Delta\langle V_{l-s}^{vdw}\rangle$	$\Delta\langle V_{l-s}^{el}\rangle$	$\Delta G_{bind,calc}$	$\Delta G_{bind,exp}$
CAM				-7.90
CMA				-5.91

Table 2: Differences in ligand-surrounding energies and calculated free energies using LIE

- **Structures:** Viewing the structures after the different parts of the simulation is a very important part of the evaluation of the simulations.

  - Move to a "results" directory.
  - To look at how the ligand moves as a function of time, we will use a `pymol` script. Type `pymol simulation.pse` in your terminal window and the program will load about 150 snapshots from the simulation. Start the simulation by pressing play in the lower-right corner.
  - Calculate the average structures from the MD simulation using `Qprep5` for each protein simulation. Type `Qprep5 < qprep_av1.inp > qprep_av1.out` and then `Qprep5 < qprep_av2.inp > qprep_av2.out` in your command window to get the average structures of the first and second halves of the simulations. Use `pymol` to compare the these two pdb files (`average_md1to7.pdb` and `average_md8to15.pdb`). Are the simulations structurally converged? Also compare the structures to the crystal structure (`masked.pdb`). Do the simulations agree with the crystal structure?

    Here is a list of useful `pymol` tips

    * Load pdb file: type `load file.pdb` for all the files you want to load.
    * Mouse control: left button rotates, middle button translates, right button zooms.
    * Selections: for selecting the ligand, type `sel lig, res 407` and the residue 407 will be the selection "lig". You can change colors or representation on it. See the list of selections on the right side of the interface, and explore all the possibilities

- **Key interactions:** To elucidate which residues contribute most to binding, the energetic interactions of the ligand with its surrounding groups can be calculated. This has been carried out using the program `Qcalc5`.

  - Go to the "protein/CAM/results" directory.
  - The residue-ligand interactions for residue 1 to 406 to the ligand (residue 407) for all production files has been saved to the file `res_lig.txt`. Type `gnuplot` in your terminal window. Now type `load 'residue.plt'` to plot the residue-ligand interactions. Can you identify the residues that play an active role in ligand binding?
  - Repeat the same procedure for CMA.