Tutorial for Pseudo-bond QM/MM Model with Q-Chem/Tinker Interface

Shenglong Wang
Department of Chemistry, New York University

The enzymes responsible for methylation of lysine residues of histones process are histone lysine methyltransferases (HKMTs), which catalyze the transfer of methyl group(s) from cofactor S-adenosyl-methionine (AdoMet) to some specific lysine residues in the N-terminal histone tails. Among the HKMTs, SET7/9 is one of the best characterized experimentally. This is a monomethyltransferase, which can only catalyze the transfer of one methyl group to the unmodified histone lysine residue H3-K4. The cofactor AdoMet and the substrate peptide bind to opposite faces of the SET7/9 and are connected by a narrow channel that has a hydrophobic inner wall. The target lysine residue is inserted into this narrow channel to access the methyl moiety of AdoMet. The reaction is a typical in-line SN2 nucleophilic substitution reaction with a dissociative transition state.

The preparation of the enzyme-substrate complex was based on the crystal structure 1O9S, the system has been equilibriated with Amber and the restart file for molecular dynamics simulation with Amber after 5 ns is *restrt-5000*. This snapshot will be used to setup ab initio QM/MM simulation with Q-Chem/Tinker. For this system, there are 32510 atoms in total, which includes 4032 protein atoms, 9488 water molecules and 14 counter ions Na+.

QM/MM System Setup

1. Convert Amber restart file to PDB file

The executable file ***ptraj*** in Amber can beconvert Amber restart file to the corresponding PDB file. The input file *ptraj.input* is

```
trajin restrt-5000
trajout SET79 pdb
image center
go
```

and the command is $AMBERHOME/exe/ptraj SET79.prmtop < ptraj.input. ***ptraj*** in Amber 10 generates the PDB file with continuous atom serial number that starts from 1, which will make it easy for later QM/MM atom group assignment. If you use other version of Amber, please be careful of the

atom serial number in the generated PDB file. The generated PDB file name is *SET79.1* and is renamed as *SET79-Amber.pdb*. The structure of the simulation system with periodic boundary conditions is shown in Figure 1.

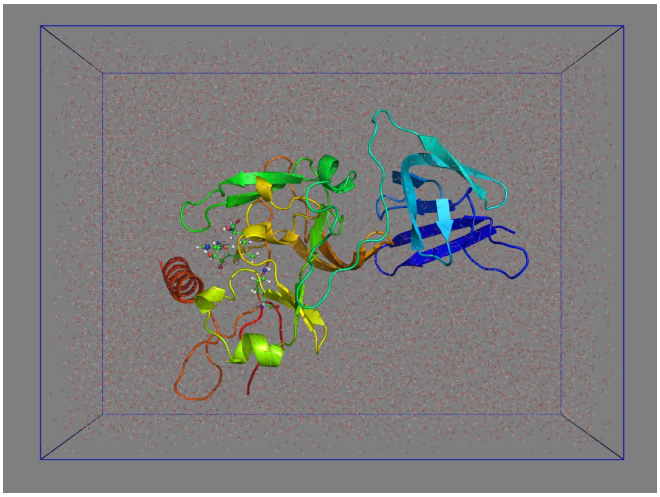

Figure 1. Molecular system from molecular dynamics simulation with periodic boundary conditions

2. Remove solvent molecules for spherical boundary conditions

Spherical boundary conditions will be employed for QM/MM simulations, the molecular system from molecular dynamics simulation with periodic boundary conditions can be cut into a sphere by removing the solvent water molecules far away from the reactive center. In this way, we can reduce some computational cost for MM calculations during QM/MM simulation processes. In another side, the maximum atom number has been defined as 20000 in current Q-Chem/Tinker. For this system, CE atom in residue SAM with coordinates (29.311, 29.267, 31.941) is selected as sphere center , all the solvent water molecules with distance from CE atoms larger than 30 Å will be removed.

The Perl script for this function is ***cut-sphere.pl*** and the corresponding input file *cut-sphere.input* is

```
PDB-Input SET79-Amber.pdb
PDB-Output sphere-30.pdb
Center-Coordinate 29.311  29.267  31.941
Sphere-Radius 30.0
Box-Dimension 85.2212166   62.9545826   60.1530603
```

The command is perl cut-sphere.pl cut-sphere.input, and the generated PDB file is *sphere-30.pdb*. The box dimensions are used here in order to avoid the case that protein molecules are not in the center of the simulation box, in this case, the solvent water molecules in the nearby image boxes should be kept. There are total 12961 atoms in the system with spherical boundary conditions, which includes all the protein atoms, 2975 water molecules and 4 Na+ ions. The structure is shown in Figure 2.

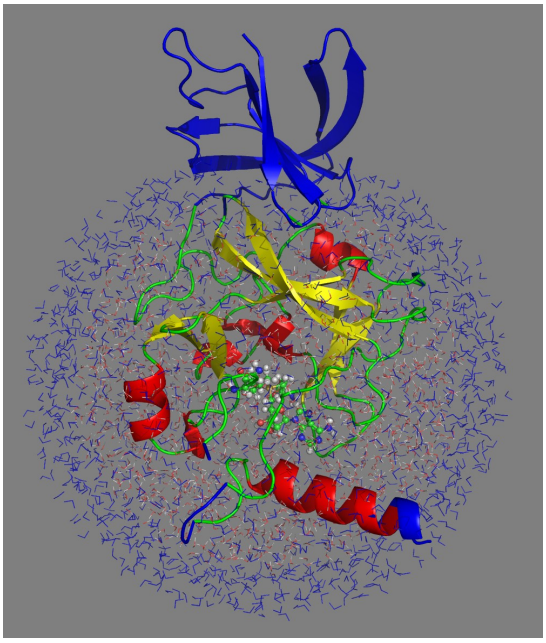

Figure 2. Molecular system with spherical boundary conditions for QM/MM simulations

Sphere radius 25 Å will be used in the later simulations to construct active atom list for spherical boundary conditions, all the atoms with distance from CE atoms less than 25 Å are free to move and all the other atoms are fixed in space. In Figure 2, all the blue atoms (outside the sphere) are fixed in space and all the inside atoms are active atoms.

3. Convert PDB file to ZPDB file

The difference between PDB file and ZPDB file is that Tinker atom type information is added for every atom in the ZPDB file which will be used to assign molecular mechanics force field parameters for Tinker. A Perl hash table (file name *amber2tinker_hmt.table*) has been implemented to match atom types in PDB file to Tinker atom types defined in the Tinker parameter file AMBER-ff99. If there are some new ligands in the systems, you can append the atom type information into the conversion table file and Tinker parameter file. A Perl script *pdb2zpdb.pl* has been implemented for the atom type conversion and the input file *pdb2zpdb.input* is

```
PDB-Input sphere-30.pdb
ZPDB-Output SET79-30.zpdb
N-Terminal-Resid 1 251
C-Terminal-Resid 250 260
Amber-Tinker-Table amber2tinker_hmt.table
```

and the command is perl pdb2zpdb.pl pdb2zpdb.input.
All the N and C terminal residues should be declared explicitly with residue Ids since the force field parameters for the terminal residues are different with the non-terminal residues. Now, we have ZPDB file *SET-30.zpdb*.

4. Convert ZPDB file to QM/MM Tinker XYZ File

Now we can convert ZPDB file to QM/MM Tinker XYZ file. First, we need to divide all the atoms in ZPDB file into 4 different QM/MM atoms groups: QM atoms, pseudo atoms, zero-charge atoms and MM atoms. Atom serial numbers for QM, pseudo and zero charge atoms must be declared with the corresponding atom serial numbers in the ZPDB file. It is easy to obtain these atom serial numbers with the visualization program Visual Molecular Dynamics (VMD) by double click the corresponding atoms to get atom index information. We assume atom index from VMD is just 1 less than the atom serial number in ZPDB file since atom index in VMD begins from 0 while atom serial number is PDB file begins from 1. This assumption is true for ***ptraj*** in Amber 10, please check this assumption if other version of ***ptraj*** is employed.

Perl script ***zpd2xyz.pl*** has been implemented to convert ZPDB file to QM/MM Tinker XYZ file. The script first moves all the QM atoms to the front according to the atom serial number, then pseudo atoms and zero charge atoms, finally MM atoms. The script will check the atom bond information by calculating atom distances between all atom pairs. C++ source code *bond-search.C* has been coded in order to speedup the atom pair distance search, which can be compiled into executable file ***bond-search.exe***. Finally, the script will print out the reordered ZPDB file and the corresponding QM/MM Tinker XYZ file, which includes QM/MM atom group information, atomic coordinates, atom type and connection information for every atom. There are 5 integer numbers in the first line of QM/MM Tinker XYZ files, which are total atom number, QM atom number, pseudo atom number, zero-charge atom number and MM atom number. In order to make QM/MM Tinker XYZ file works properly, total charge and spin multiplicity of QM atoms should be appended to the first line manually, otherwise, the default value 0 for total charge and 1 for spin multiplicity will be used.

Here we'll prepare two QM/MM systems with 34 and 66 atoms in the QM subsystems respectively as shown in Figure 3.

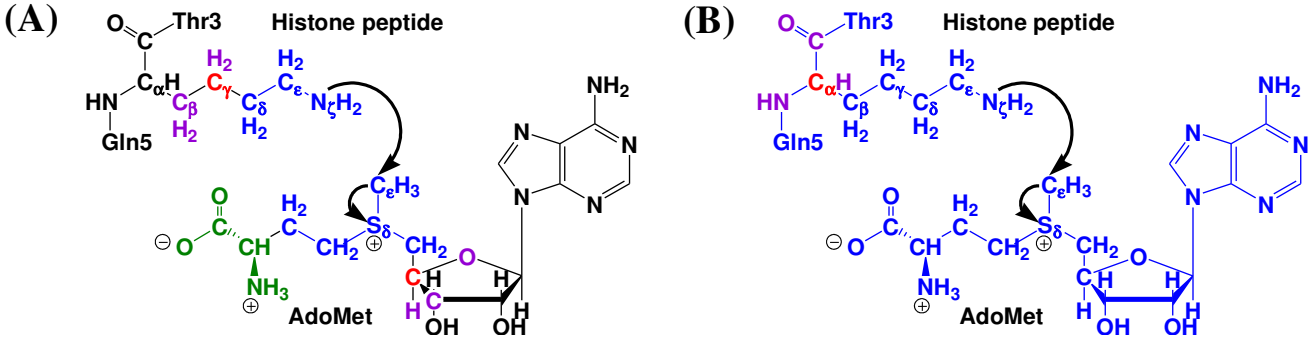

Figure 3. Illustration of the division of QM/MM system for simulating the methyl transfer from AdoMet to histone lysine residue H3-K4. (A) There are 32 QM atoms and 2 pseudo-atoms. The atoms in the reaction center (colored in blue) are assigned 6-31G(d) basis sets, the nearby atoms (colored in green) are assigned 3-21G basis sets. The two pseudo-atoms (colored in red) are treated with the improved pseudo-bond parameters. The atoms connected with pseudo-atoms (colored in violet) are treated as zero-charge atoms. (B) There are 65 QM atoms and 1 pseudo-atom. All atoms colored in blue are

assigned 6-31G(d) basis sets and the pseudo-atom (colored in red) is treated with the improved pseudo-bond parameters. The atoms connected with pseudo-atom (colored in violet) are treated as zero-charge atoms. All the left atoms in the two systems are MM atoms that are treated with Amber force field. All the zero-charge atoms have no electrostatic interaction with QM atoms.

The input file for 34-QM atom system is *zpdb2xyz-34.input*

```
Tinker-Parameter amber99_m.prm
ZPDB-Input SET79-30.zpdb
ZPDB-Output SET79-qmmm-34.zpdb
Tinker-XYZ SET79-qmmm-34.xyz
Bond-Search-Exe ./bond-search.exe
QM-Atoms 3863 3864 3865 3866 3867 3868 3869 3870 3871
QM-Atoms 3983 3984 3985 3986 3987 3988 3989 3990
QM-Atoms 3991 3992 3993 3994 3995 3996 3997 3998 3999
QM-Atoms 4000 4001 4002 4003 4004 4005
Pseudo-Atoms 3860 4006
Zero-Charge-Atoms 3857 3858 3859 3861 3862 4007 4008 4009
```

and the input file for 66-QM atom system is *zpdb2xyz-66.input*

```
Tinker-Parameter amber99_m.prm
ZPDB-Input SET79-30.zpdb
ZPDB-Output SET79-qmmm-66.zpdb
Tinker-XYZ SET79-qmmm-66.xyz
Bond-Search-Exe ./bond-search.exe
QM-Atoms 3983 3984 3985 3986 3987 3988 3989
QM-Atoms 3990 3991 3992 3993 3994 3995 3996 3997 3998 3999
QM-Atoms 4000 4001 4002 4003 4004 4005 4006 4007 4008 4009
QM-Atoms 4010 4011 4012 4013 4014 4015 4016 4017 4018 4019
QM-Atoms 4020 4021 4022 4023 4024 4025 4026 4027 4028 4029
QM-Atoms 4030 4031 4032
QM-Atoms 3857 3858 3859 3860 3861 3862 3863 3864 3865 3866
QM-Atoms 3867 3868 3869 3870 3871
Pseudo-Atoms 3855
Zero-Charge-Atoms 3853 3854 3856 3872 3873
```

The command is perl zpdb2xyz.pl zpdb2xyz-??.input

Now we have two QM/MM Tinker XYZ files *SET79-qmmm-34.xyz* and *SET79-qmmm-66.xyz*, which can be used for QM/MM simulations with Q-Chem/Tinker interface.

Molecular Mechanics Minimization

Now we rename the QM/MM Tinker XYZ file *SET79-qmmm-34.xyz* as *SET79.xyz*, together with *SET79.key* file, we can minimize the structure first with Amber force field method. The minimization executable file ***minimize*** in Tinker has been paralleled with OpenMPI for spherical boundary conditions. A new executable file ***QMHessian*** has been coded to calculate the Hessian matrix for QM atoms from MM force field, which will be used as the initial guessed Hessian matrix for QM atoms in the QM/MM micro-iterative optimization process. ***QMHessian*** will write the the Hessian matrix to file *QM-Hessian*, Q-Chem/Tinker interface can read this file for micro-iterative QM/MM minimization process.

The first line of *QM-Hessian* file is an integer number, which is the dimension of the Hessian matrix. For example, for *SET79-qmmm-34.xyz* system, there are 32 QM atoms and 2 pseudo-atoms, the QM Hessian matrix dimension is 34 x 3 =102. The following 102 x 102 = 10404 lines are the Hessian matrix elements.

QM/MM Micro-Iterative Minimization for Reactant

Several Linux shell scripts are necessary in order to make Q-Chem/Tinker interface works automatically, especially for the jobs running in computer cluster with PBS. There are several tcsh scripts have been provided as template which can be modified for other applications.

HF method has been employed and the atoms near the reaction center have been assigned with 6-31G(d) basis sets and some other atoms have been assigned 3-21G(d) basis sets. The Q-Chem input file *SET79.in* is

```
$Molecule
 Tinker SET79.xyz
$End

$Rem
  JobType Tinker
  Exchange HF
  Correlation None
  Basis Mixed
  ECP Gen
  Symmetry Off
  Print_Input False
  Sym_Ignore True
  MEM_Total 1000
  MEM_Static 500

  SCF_Guess GWH
  !Basis2 STO-3G

  SCF_Convergence 8
  Max_SCF_Cycles 200

  Geom_Opt_Max_Cycles 400
  Geom_Opt_Coords 0
  Geom_Opt_IProj 0
  Geom_Opt_IDB 1
  Geom_Opt_Print 3

  Geom_Opt_DMax 300
  !Geom_Opt_STol 300

$End

$Basis
 C 1
6-31G(d)
****
 H 2
6-31G(d)
****
 H 3
6-31G(d)
****
 C 4
6-31G(d)
****
 H 5
6-31G(d)
****
 H 6
6-31G(d)
****
 N 7
6-31G(d)
****
 H 8
6-31G(d)
****
 H 9
6-31G(d)
```

```
****
 C 19
6-31G(d)
****
 H 20
6-31G(d)
****
 H 21
6-31G(d)
****
 C 22
6-31G(d)
****
 H 23
6-31G(d)
****
 H 24
6-31G(d)
****
 S 25
6-31G(d)
****
 C 26
6-31G(d)
****
 H 27
6-31G(d)
****
 H 28
6-31G(d)
****
 H 29
6-31G(d)
****
 C 30
6-31G(d)
****
 H 31
6-31G(d)
****
 H 32
6-31G(d)
****
 N 10
3-21G(d)
****
 H 11
3-21G(d)
****
 H 12
3-21G(d)
****
 H 13
3-21G(d)
****
 C 14
3-21G(d)
****
 H 15
3-21G(d)
****
 C 16
3-21G(d)
****
 O 17
3-21G(d)
****
 O 18
3-21G(d)
****
 Cc 33
 SP  2  1.00
 1.150000  1.000000  1.000000
 0.200000  0.224750  0.827470
 ****
 Cc 34
 SP  2  1.00
 1.150000  1.000000  1.000000
 0.200000  0.224750  0.827470
 ****
$End

$ECP
 Cc 0
 try1 0 92
 S component
 1
 1   5.80000       5.00000
 ****
$End
```

And the Tinker key file *SET79.key* for QM/MM micro-iterative optimization is

```
Parameters amber99_m.prm

Neutral-Groups
Chg-Cutoff 18.0
Vdw-Cutoff 12.0

Sphere 29.311  29.267  31.941  25.0

SCF-Initial-Guess Yes
QM-Total-Charge 1
QM-Multiplicity 1
Mulliken-Analysis None

NC-Restart-File SET79.nc
Print-Key-Words Yes

Opt-MM-RMS-Gradient 0.02
Opt-QM-Maximum-Cycle 200
Opt-QM-Coordinate Cartesian
Opt-ReFit-ESP-Charges No
Opt-QM-Trajectory-File SET79.arc
Opt-QMMM-xyz-File SET79-opt.xyz
Opt-QM-Print 3
Opt-QM-Hessian-File QM-Hessian
Opt-Read-QM-Hessian Yes
Opt-Read-QM-Hessian-From-NC Yes
Opt-Dump-RC-Energy True

Job-Type QMMM-Iterative-Optimization
```

The sphere center for spherical boundary conditions is same as that has been used to remove water molecules and the radius is 25 Å. Cutoff for electrostatic interaction between MM atoms is 18 Å and cutoff for all VDW interactions is 12 Å, there is no cutoff for electrostatic interaction between QM

atoms and MM atoms and all MM atoms (active atoms and fixed atoms) have interaction with QM atoms. Hessian matrix from MM force field is used as the initial guess for QM minimization. Identity matrix will be used as the initial guess for QM minimization if no Hessian matrix is provided. The SCF convergence is 10^(-8) and the MM minimization tolerance is that root of mean square of gradients is less than 0.02 kcal/mol/ Å. Cartesian coordinates are used both in QM minimization and MM minimization process. NetCDF restart file *SET79.nc* will be used to save all the necessary restart information for Q-Chem/Tinker.

The total potential energy changes during micro-iterative minimization process is shown in Figure 4. The total potential always decrease and the minimization convergences very quickly even with Cartesian coordinates for QM minimization process. It is strongly suggested to use MM Hessian matrix as the initial guess, which will make the QM minimization converges much faster. L-BFGS method have been used for both QM and MM minimizations. The updated approximated Hessian for QM atoms and other necessary restart information are saved to the NetCDF restart file for every QM/MM iterative minimization step, including atomic coordinates, molecular orbital coefficients, electrostatic potential (ESP) charges for QM atoms and some others. The content of the NetCDF restart file can be converted to text file with executable file ***ncdump***. If you want to restart the minimization from NetCDF file, please set the Tinker keyword Opt-Read-QM-Hessian No, Q-Chem/Tinker will not read Hessian matrix from QM-Hessian file anymore, and will just use the Hessian matrix saved in the NetCDF restart file as the initial guess if there is Hessian matrix saved in NetCDF restart file. All the information saved in NetCDF file will be used for the following reaction path scan. The final optimized structure has been saved in the file *SET79-opt.xyz*. If this file exists, it means that the micro-iterative QM/MM minimization is successful.

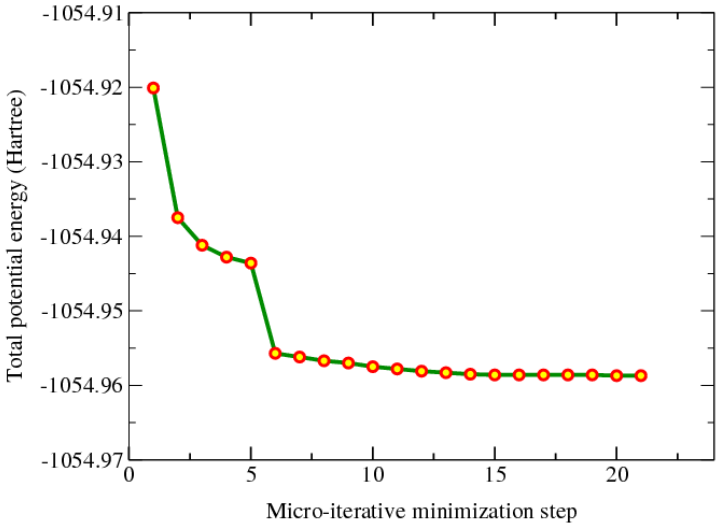

Figure 4. Total potential energy convergency during micro-iterative QM/MM minimization process

The reaction center (QM atoms) structures for the 34-QM atoms and 66-QM atoms are shown in Figure 5, both the two QM/MM systems give the similar structures. The S_δ-C_ε bond length is 1.8 Å and the C_ε-N_ζ is 3.2 Å.

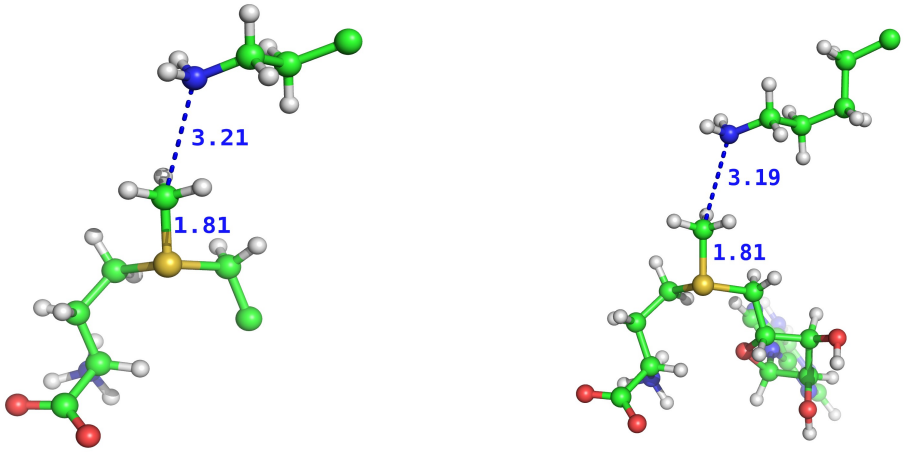

Figure 5. Reactant structures for the methyl transfer from AdoMet to histone lysine residue H3-K4. Left panel is for 34-QM system and right panel is for 66-QM system.

Reaction Path Scan with Reaction Coordinate Driving

Reaction coordinate driving (RCD) method has been approved to be very successful to map the minimum energy path (MEP) for some chemical reactions if a good reaction coordinate is available. RCD method heavily depends on the existence of a good reaction coordinate. A harmonic potential with very big force constant will be applied to constrain the reaction coordinate to be a predefined value along the reaction path. This is the only difference between reactant minimization and minimization for points in the MEP curve.

First define the reaction coordinate as the bond lengths differences between S_δ-C_ε and C_ε-N_ζ,

$$Rc = R(S_\delta\text{-}C_\varepsilon) - R(C_\varepsilon\text{-}N_\zeta)$$

which has been proved can describe methyl transfer reactions very well from reactant, transition state and product.

These several lines about reaction coordinate constrain have been added to Tinker key file *SET79.key*,

(A) 34-QM system

```
Opt-Read-QM-Hessian No

RC-Length RCLENGTH
RC-Force-Constant  2000.0
RC-Bond  25  26   1.0
RC-Bond  26   7   -1.0
RC-Dump-Energy Yes
RC-Term Yes
```

(B) 66-QM system

```
Opt-Read-QM-Hessian No

RC-Length RCLENGTH
RC-Force-Constant  2000.0
RC-Bond  31  32   1.0
RC-Bond  32  13  -1.0
RC-Dump-Energy Yes
RC-Term Yes
```

Here, atom indexes for S_δ, C_ε and N_ζ are 25, 26 and 7 respectively for 34-QM atoms. The macro RCLENGTH will be changed from -1.40 to 2.80 with increment 0.1 gradually. For every point with fixed reaction coordinate, the structure will be minimized with QM/MM micro-iterative minimization method. The information saved in NetCDF restart file from the previous point will be used as initial values for the next point, only the reaction coordinate value is changed. In order to get a smooth enough MEP curve, several times of forward and backward reaction path scan are necessary. It is suggested to start from Rc = -1.40 since this is the reaction coordinate value of the minimized reactant, first scan the reaction path forward to 2.80, then scan the reaction path backward to -1.80, we want to locate the MEP along the reactant as well. The reaction path scan from -1.80 to 2.80 forward and backward can be repeated several times until the smooth MEP curve can be obtained and the curves from forward and backward scans agree with each other very well. The highest energy point along the MEP is supposed to be the transition state point. The minimum points in the reactant and product regions are supposed to be reactant and product respectively.

A set of Shell and Perl scripts have been implemented to drive the reaction path scan and collect the necessary data from output files automatically.

The minimum energy paths for the 34-QM and 66-QM systems are shown in Figure 6. The reaction barrier from reactant to the transition state is 22.4 kcal/mol for the 34-QM systems and 20.0 kcal/mol for the 66-QM system.

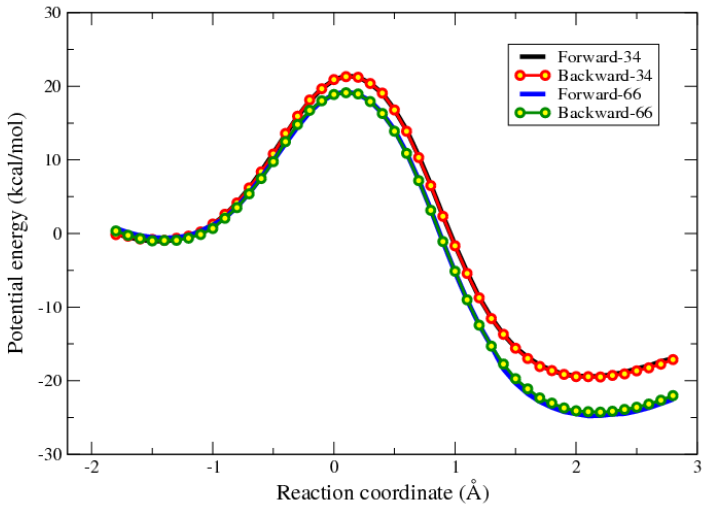

Figure 6. Reaction path for the methyl transfer from AdoMet to histone lysine residue H3-K4 with reaction coordinate driving method. Forward and backward reaction path scan for both 34-QM and 66-QM systems are shown.

The reaction coordinates corresponding to the transition states are around 0.1 Å and the two bond lengths are shown in Figure 7. The structures from the two systems are almost same, which means that the 34-QM system and HF method with mixed 6-31G(d) and 3-21G(d) basis sets can describe the methyl transfer reaction very well.

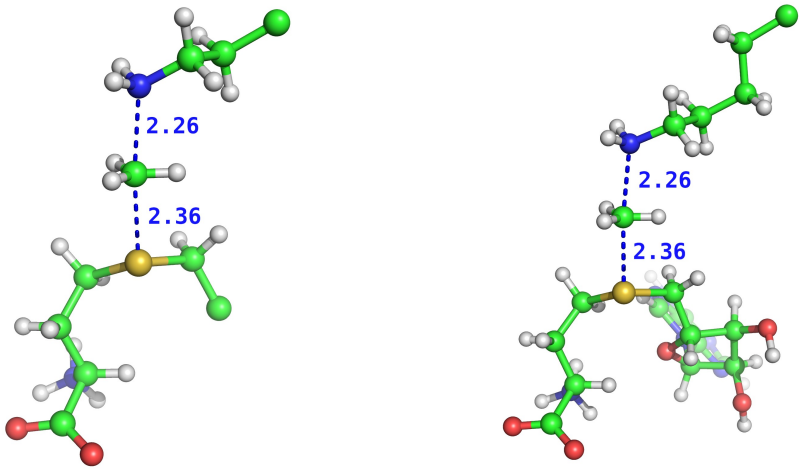

Figure 7. Transition state structures for the methyl transfer from AdoMet to histone lysine residue H3-K4. Left panel is for 34-QM system and right panel is for 66-QM system.

QM/MM Free Energy Perturbation

After we get the smooth reaction path, we can run QM/MM free energy perturbation (QM/MM-FE) to get the approximated free energy profile along the reaction coordinate. Before that we need to calculate the energy between QM atoms and to save the final coordinates and ESP charges for QM atoms, which will be used for QM/MM-FE calculations. These key words can be appended to the Tinker key file and run reaction path scan again or just read the NetCDF restart file to restart the QM/MM iterative optimization for each point along the reaction path.

Opt-QMMM-FE-QM-Energy Yes
Opt-Write-QM-XYZ-Charges Yes
Opt-QM-XYZ-Charges-File SET79-QM.xyzq

The QM energy will be printed to the stdout like

QM energy for QMMM-FE: -1013.8978547871 Hartree

and the QM coordinates and ESP charges will be saved to file *SET79-QM.xyzq*. The format of xyzq file is almost same as XYZ file except that the 5th column is the ESP charges. This file will be used for QM/MM-FE MD simulations.

The following Tinker key words are need to control the molecular dynamics and Twin cutoff are used for electrostatic and VDW interactions. If the MD simulations are not stable, you can decrease the update frequency till to disable the Twin cutoff.

Tau-Temperature 0.5
RandomSeed 164623
Integrate Beeman
MD-Temperature 300.0
MD-Total-Steps 2000000
MD-Steps 500000
MD-Time-Step 1.0
MD-Dump-Time 1.0

```
MD-Save-ESP-Charges No
MD-NetCDF-Trajectory-File SET79-MD.tnc
MD-NetCDF-Trajectory Yes
MD-Reverse-Velocity No


Twin-Vdw Yes
Twin-Vdw-CutOff 10.0
Twin-Vdw-Frequency 5

Twin-Charge Yes
Twin-Charge-CutOff 10.0
Twin-Charge-Frequency 5

QMMM-FE-A-State SET79-QM.xyzq.F.0.00
QMMM-FE-B-State SET79-QM.xyzq.F.-0.20
QMMM-FE-B-State SET79-QM.xyzq.F.0.20

QMMM-FE-Energy-File QMMM-energy

Job-Type QMMM-FE
```

The perturbation energies are saved in file *QMMM-energy*, which can be used to calculate the perturbation free energies.

ab initio QM/MM Molecular Dynamics

The NetCDF restart file from QM/MM-FE can be used to run ab initio QM/MM molecular dynamics. Q-Chem/Tinker will read coordinates for all atoms and velocities for MM atoms. Q-Chem/Tinker will assign velocities for QM atoms first according to Boltzmann distributions. Ab initio QM/MM MD simulations with umbrella sampling can be used to get the potential of mean force (PMF) along the reaction coordinate. These are the part of the Tinker key file about umbrella sampling.

```
RC-Length 0.00
RC-Force-Constant  50.0
RC-Bond  25  26   1.0
RC-Bond  26   7   -1.0
RC-Coordinate-File SET79-RC.dist
RC-Dump-Coordinate Yes
RC-Dump-Energy Yes
RC-Term Yes
```

The Tinker keywords to control ab initio QM/MM MD simulations are same as in QM/MM-FE simulations. The reaction coordinates of every MD step are saved in file *SET79-RC.dist*, which can be used with the weighted histogram analysis method (WHAM). Please create an empty file *SET79-RC.dist* before run the ab initio QM/MM MD jobs.

Tutorial Files

The file *QChemTinker-Tutorial.tar.gz* contains the following directories

MM-Min/ Path-1/ QMMM-FE/ QMMM-MD/ Reactant/ Setup/

Setup: Amber restart file and scripts to setup Tinker QM/MM xyz file
MM-Min: For MM minimization with Tinker ***minimize*** and calculate QM Hessian matrix
Path-1: to scan reaction path with reaction coordinate driving method
QMMM-FE: QM/MM free energy perturbation
QMMM-MD: ab initio QM/MM MD simulations

These files are just templates, which should be modified according to the Q-Chem settings in the cluster. Some scripts will help a lot to setup the simulations in QM/MM-FE and QM/MM-MD simulations.