Umbrella Sampling & WHAM Tutorial

Tanja Kovacevic

Chemistry Department, University of Colorado Denver
Hai Lin Theoretical Chemistry Lab

May 2019

Contents

1 Introduction

This tutorial helps you to learn how to perform umbrella sampling and to use the weighted histogram analysis method (WHAM) to construct potential of mean force (PMF) for the internal rotation (characterized by the periodic changes in a specific dihedral) of amino acid side chain. The tutorial assumes that you have a working knowledge of navigating bash, the NAMD (nanoscale molecular dynamics) program, and the VMD (visual molecular dynamics) program.

1.1 Background on Umbrella Sampling

Umbrella sampling is a biased sampling method to improve sampling for a system for which ergodicity is hindered by the potential energy landscape [1] [2]. Here is a brief, though not very rigorous, description.

A free energy profile $F(x)$, or potential of mean force (PMF), along a reaction coordinate x can in principle be obtained through Boltzmann inversion using the normalized population (i.e. probability) distribution $p(x)$ along x,

$$F(x) = -RTlnp(x) \tag{1}$$

where R is the gas constant and T the temperature in Kelvin. The population distribution can be collected conveniently by dividing the entire range of x into multiple (e.g. 20) bins of often equal width and counting how many times the simulated system possesses the x values that belong to a given bin.

However, in practice, the accuracy of $F(x)$ can be significantly limited if some parts of the area of interest are rarely sampled, i.e. $p(x)$ is very small for certain bins and thus is associated with too large statistical uncertainties. For example, if you report a barrier height as (2$\pm$100) $kcal/mol$, the value of 2 $kcal/mol$ is almost meaningless.

The poorly sample areas are usually associated with high potential energies. Therefore, in umbrella sampling, biasing potentials (usually harmonic potentials) are added to the true potential to change its shape so that the system can now frequently visit these desired parts of area:

$$V_m(x) = V(x) + k(x - x_0)^2 \tag{2}$$

Here, $V_m(x)$ is the modified potential that is easy for sampling, $V(x)$ is the original, or "true" potential, k is the force constant of the biasing potential, and x_0 is the center of the biasing potential, usually near the targeted location to which we want the system to visit. Because we know the biasing potentials, we can remove their effects afterwards (via re-weighting):

$$V(x) = V_m(x) + k(x - x_0)^2 \tag{3}$$

In this tutorial, we will study the PMF for the internal rotation of the side chain of a specific residue (F58) in the HP36 protein. The reaction coordinate is the dihedral angle χ_2 in the range of –180 to +180 deg of the side chain.

1.2 Background on the WHAM Method

Weighted Histogram Analysis Method (WHAM) [3] is a popular approach to reconstruct the PMF from the population distribution resulted in umbrella sampling. Briefly, when we do the Boltzmann inversion, the results actually contain an unknown arbitrary constant:

$$F(x) = -RTlnp(x) + C \tag{4}$$

That is because energy and free energy are relative quantities, i.e. only the changes in energy or free energies are meaningful (and measurable):

$$\begin{aligned} \Delta F &= F(x_1) - F(x_2) \\ &= (-RTlnp(x_1) + C) - (-RTlnp(x_2) + C) \\ &= -RTln\frac{p(x_1)}{p(x_2)} \end{aligned} \tag{5}$$

In umbrella sampling, one often performs a series of sampling runs with (different) biasing potentials at different locations along the reaction coordinate x. (The number of runs need not be the same as the number of bins.) Each Boltzmann inversion will lead to an unknown arbitrary constant C. The WHAM method is to iteratively adjust these constants until self-consistency is achieved, or in other terms, until the reconstructed PMF is as smooth as possible.

2 Required Programs

The following are required to perform umbrella sampling and WHAM analysis:

1. **NAMD**: Nanoscale Molecular Dynamics [MD] software package used for MD simulation. [1] NAMD available for download here:
2. **VMD**: available to download at: https://www.ks.uiuc.edu/Development/Download/download.cgi?PackageName=VMD. [2]
3. **Python**: Make sure to have the latest version downloaded. Available at: https://www.python.org/downloads/ [3]
4. **Text Editor**: VIM will be used as the text editor

3 Umbrella Sampling Tutorial

3.1 Preparing Files

In good practice, one should be working on an equilibrated system with enough MD simulation. As mentioned in the background (Section 1.1), this is needed to have a sufficient population distribution $P(x)$, therefore leading to a more trustworthy PMF.

For this tutorial there will be a directory labeled `equilibrated_system`. An unconstrained, equilibrated system has been provided. This system contains one HP36 protein in water. and has been equilibrated for 2 nanoseconds [ps].

The water is described with TIP3 force fields

3.1.1 Selecting & Saving Coordinates

In this section we will be using the pre-equilibrated system to choose a dihedral angle from one phenylalanine (PHE6) residue of one protein. By the end of this section, we will have a pdb file with coordinates of PHE6 from the protein HP36 at a particular dihedral angle ($n \times 10^\circ = y; (-180^\circ > y > 180^\circ)$).

First load the necessary files into VMD to open the trajectory located within the folder (file.psf, file.pdb, & file.dcd
Within VMD locate the VMD Main window,

- Open the pull down menu *Graphics* > *Representations*
- Now we will isolate the desired protein and find the dihedral angle of interest
  - Under '*selected atoms*' type in "*segname VP1 and resid* 6"

A good exercise is to open the PDB file in VIM and locate residue 6

[1] Any MD software package can be used but some instructions might vary depending on the MD package used
[2] VMD is a free visualization software
[3] Python 3.8 was used for these purposes

- Press 4 on your keyboard and select four consecutively bonded atoms.
  - In this case select, (1) CA, (2) CB, (3) CG, (4) CD1 or (4) CD2 of residue 6

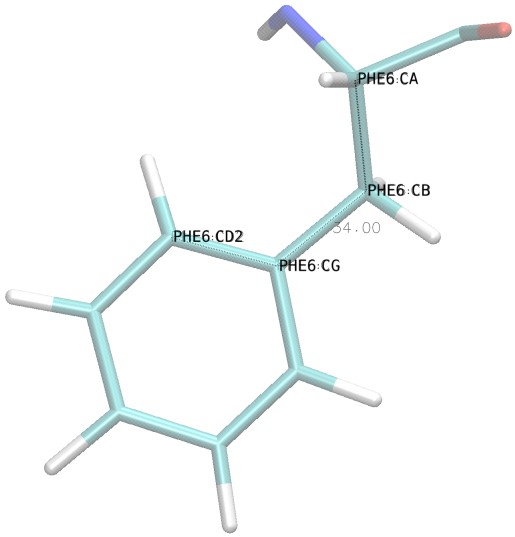

Figure 1: Visualization of the selected atoms making up the dihedral angle

This visualization might look a little different because I've changed visualization settings for tutorial purposes

- Again, in the VMD Main window click on *graphics* > *labels*
- A *Labels* window should have popped up and there should be a pull down window that is currently under *Atoms...* click on it and select *Dihedrals*
  - you should see PHE6:CA (and 3 other atoms in the window)
- Highlight the atoms, click on the *Graph* tab and select the *Graph* button

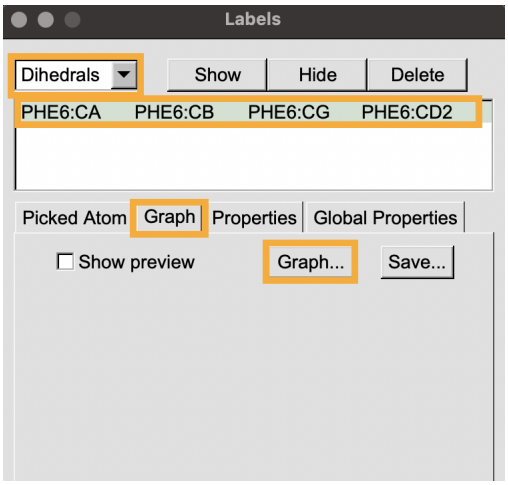

Figure 2: *Labels* window we will use to graph the chosen dihedral angle

- A new window should pop labeled Multiplot, scroll over the blue dots on the black line and locate at the terminal which is open with VMD
- There should be two columns in your terminal (frame + dihedral angle). jot down the frame associated with the dihedral angle closest to the 10° multiple (e.g. 120°, -60°, 80°, etc)

You can choose a dihedral angle within 0.3° (e.g. for 10° → 9.7° or 10.3° is sufficient)

- Now that we have identified the frame with the dihedral angle of interest, we will go back to the *VMD Main* window, right click on the uploaded trajectory, and hit *Save Coordinates*

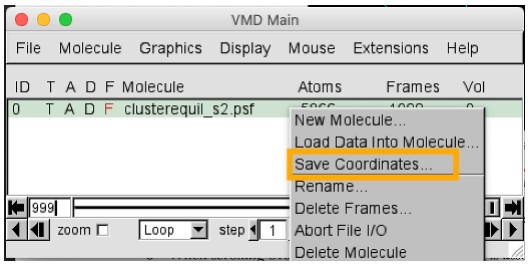

Figure 3: *VMD Main* window where we will save the coordinates

- Another window will pop-up, *Save Trajectory*, and we will do the following:
  - For *Selected atoms* type *all*
  - On the *File Type* drop down menu select *namdbin*
  - In the outlined box labeled *Frames* reference back to your noted frame and enter it here

– Hit save and choose the output name of your choosing

Ex) p10_F6.pdb for an angle of 10° of residue 6

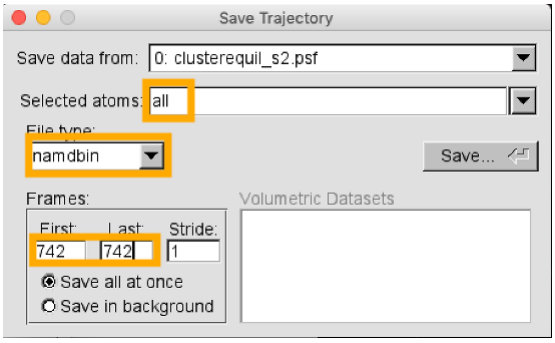

Figure 4: The *Save Trajectory* window where we will save the coordinates for the chosen frame

Yay! You now have input coordinates (.coor file) to begin umbrella sampling!

3.1.2 Preparing input files for Umbrella Sampling

We will now look at the file.conf (configuration file) & file.xsc (extended system) files located in the folder. By the end of this section we will be ready to run MD to obtain the $p(x)$ data for the dihedral angle of interest.

Use vim to open the `file.conf` file:

- Two new lines need to be added to the configuration file under the `#Custom Parameters#` section:

```
Colvars            on
ColvarsConfig      ushp36.in
```

- Then open the `ushp36.in` in vim, and it should look like this:

This is the file we use to constrain the dihedral angle in our MD simulation

```
colvarsTrajFrequency    10
colvarsRestartFrequency 10

colvar {
    name dih

    dihedral {
      group1 { AtomNumbers 245 }
      group2 { AtomNumbers 247 }
      group3 { AtomNumbers 250 }
      group4 { AtomNumbers 251 }
    }
}

  harmonic {
    name tanja
    colvars dih
    centers -60.0
    forceConstant 0.03
  }
```

Figure 5: Example of the `ushp36.in` file

- Let's walk through the ushp36.in file above
  - `Frequency` lines: data will be written every 10 lines
  - `group1-group4`: the `AtomNumbers` corresponding to the chosen atoms for the dihedral angle
  - `harmonic`: this imposes the harmonic bias potential
  - `name`: you can put your name here!
  - `colvars`: what we are constraining `dih` = dihedral in this case
  - `centers`: what dihedral angle you are trying to constrain
  - `forceConstant`:the force k of the harmonic bias potential you are imposing in $[kcal/mol]$

Atom numbers can be found in the pdb file

- Back to the configuration (.conf) file
  - Remember that frame you jotted down earlier? We need it again!
  - Using this frame number we need to find this frame number in file.xst
  - When you find the line needed in your file.xst we will copy and paste this into file.xsc

The first column in file.xst is the frame number

Now `file.xsc` is ready to go!

- **Let's gather our thoughts, so far you should have:**
  - `file.coor`
  - `file.xsc`
  - `file.conf`(updated but not yet complete)

What's next? Let's get comfortable with finish the configuration file!

This URL will give you detailed information about the configuration file:
https://www.ks.uiuc.edu/Training/Tutorials/namd/namd-tutorial-unix-html/node26.html

- Open file.conf in vim and locate the `#adjustable parameters#` section
- The first couple of lines state the files that will be used in the dynamics calculation
  - `structure` = file.psf
  - `coordinates` = file.coor
  - `paraTypeCharmm` = on
  - the three parameters (force fields) we will need for these calculations
    - `parameters` = par_all36_lipid.prm

    - * `parameters` = par_all36_protein.prm
    - * `parameters` = toppar_water_ions.str
  - `outputname` = put your desired output name
  - `seed` = this is the random seed, so pick a unique number that stays the same throughout all of your umbrella sampling
  - `timestep` = 2.0 (2.0 $fs/step$)
  - `output energies` = 10
  - `set temp` = 300 (K)
  - `temperature` = $temp
  - `frequency` = make this a value so that there are 1000 frames saved in the files: `file.dcd`, `file.restart.dcd`, `file.xst`
  - `rigidBonds` = all
  - `rigidTolerance` = 1.0e-8
  - `rigidIterations` = 1000

using your student ID can be any easy to remember!

energy data is written every 10 steps

ambient temperature rounded up

For example if you run a 20ps (20000fs) MD, you should enter `frequency = 2000`

The configuration file in your folder should have most of the values in `#adjustable parameters#` set, but it is still good practice to know what each line means so you can customize your configuration file when needed

- Next, let's locate the `#Restart Parameters#` section
  - `bincoordinates` = file.coor (created earlier)
  - `binvelocities` = should be commented out
  - `extendedSystem` = file.xsc (created earlier)

The `#Custom Parameters#` are set for this tutorial. Reference the URL above to gain a working knowledge of each line in this section. These lines will be important for other simulations you will perform in the future

- Lastly, we will consider the `#Execution Commands#` section
  - `firsttimestep` = this should be commented out when initializing a run
  - `minimization` = 1000 [for the first dihedral calculation, you will need to minimize the system for 1000 steps, when calculating other dihedral angles this can be commented out]
  - `run` = length in time of your MD in fs

If you are continuing a run, you'd set this to the next time step in your MD

Now `file.conf` is ready to go!

3.1.3 Taking account of all files before MD simulation

The files you will need to make sure you have in your simulation folder are:

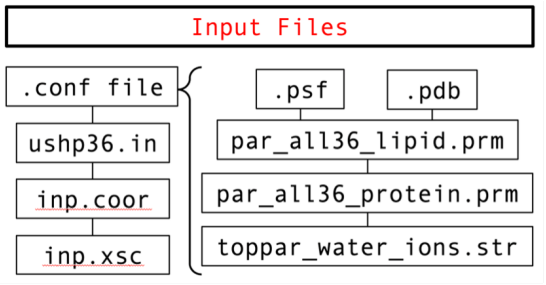

Figure 6: Files needed to perform umbrella sampling

3.2 Running MD for Umbrella Sampling

This is a short section describing how to run NAMD calculations in order to gather data for umbrella sampling. By the end of this section and lots of MD simulations, you will have enough data to begin WHAM Analysis.

- For the initial dihedral angle you have coordinates for you will need to minimize the system for 1000 steps
- The bash command to run NAMD:

```
/path/to/NAMD file.conf > output.log &
```

the ampersand (&) will have your MD simulation run in the background

- After a successful minimization, you can begin running an umbrella sampling production run
- The bash command to run NAMD is the same as above

Now we need to run this for every dihedral angle!

- To expedite the MD performed this is how I've completed it before in the past:
  - Run an initial equilibration for $20ps$
  - Obtain the next dihedral angle (going through all previous steps (Sections 3.1 & 3.2) again and running for $20ps$
  - Repeating until I have all of the dihedral angle 'mini'-equilibrations I need and sending my files to a super computer to run for however long necessary to achieve equilibration (in my case I did $2ns$).

To reiterate after you've completed sections 3.1 & 3.2 you will need to iterate over these sections for each dihedral angle.

4 Weighted Histogram Analysis Method [WHAM] Tutorial

You have finished sampling all possible states of rotation for the dihedral angle of interest (χ_2 of PHE6, in this case). We imposed a force constant by applying the harmonic bias at each window (window is synonymous with dihedral angle). You need to remove the imposed bias throughout this 'rotation' and obtain the unbiased free energy landscape. After completion of this section we should be able to plot a Potential Mean Force Graph (PMF)!! *So much work for one graph, huh? Such is life as a researcher!*

WHAM[4] can be found & downloaded at:
http://membrane.urmc.rochester.edu/?page_id=126

There is a wonderfully written tutorial on the website which can be used to configure & run the package. This tutorial can be found at: http://membrane.urmc.rochester.edu/sites/default/files/wham/doc.pdf

I will give a brief 101 to configuring and running WHAM in this tutorial.

4.1 Configuring WHAM

- Building WHAM 1-D code:
  - After downloading WHAM onto your workstation, you should move WHAM into your preferred directory
  - In the WHAM folder you will need to configure the package by typing into your terminal:

```
cd wham
make clean
make
```

4.2 Preparing Files

When I do WHAM Analysis I also perform convergence tests and histogram analysis. The convergence tests prove, to me, that I have run molecular dynamics on my system long enough to consider it equilibrated (to check that $2ns$ is enough MD time). It is crucially important to have enough simulation time lengths so that the results can be reproduced experimentally. The histogram analysis is another method performed to confirm I have enough data, $p(x)$, to plot an accurate PMF.

We need to collect all of the trajectory files in order to ascertain the data needed for the WHAM package to run its magic. I usually create a

`WHAM` working directory where I collect all of the files necessary to run my analysis. You can do the same if you are inclined to do so!

From this point on, I will assume you have made this directory.

`WHAM` direcotry and I will discuss as if we are working in it.

- Create `timeseries` & `histogram` directories
- Enter the `timeseries` directory and transfer all `file.colvars.traj` files into the `timeseries` directory
- From here we will be cleaning up the trajectory files in order to run executables for the WHAM analysis [4] :
  - We will use a bash command to remove all `#` in the trajectories files

    ```
    sed -i '/#/'d *.traj
    ```

  - Then, execute the `histogram\_script.sh` on all `file.colvar.traj` files and move them into the `histogram` directory

    ```
    ./histogram_script.sh
    mv *.hist ../histogram
    ```

  - In the scripts directory there will be a directory labeled `hist_graph` and you can move this folder into your `histogram` directory.
  - In the `histogram` directory you can run the executable

    ```
    ./make_hist_graphs.sh
    ./graphHistogram.py totalcount.tot
    ```

  - The output you should get is 'pic.png' and this is the histogram of your umbrella sampling

You should have files that consist of data for the dihedral angles in the range $-180^{\circ} \rightarrow 180^{\circ}$

[4] If your rotation is symmetric (phenylalanine's rotation is symmetric) you can use the `wrap_colvars.py` to wrap your files after removing the #'s from your files (e.g. $0^{\circ} \rightarrow 180^{\circ}$)

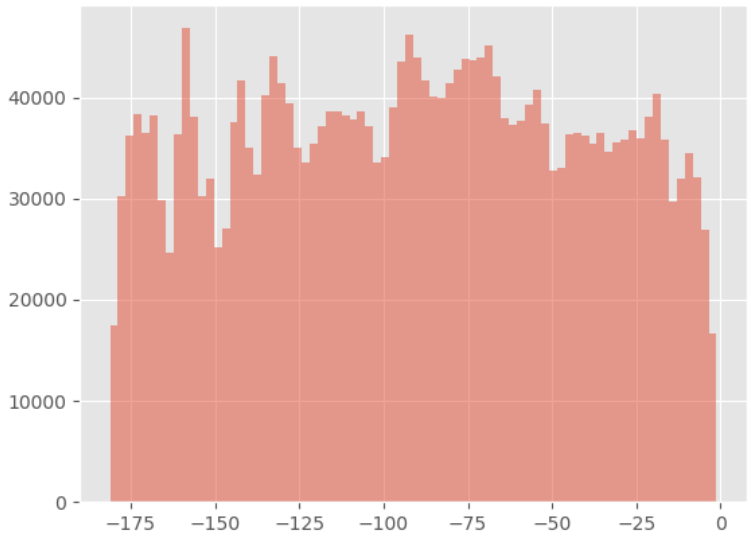

Figure 7: Example of how your histogram figure will appear

- Working our way back to the `timeseries` directory we will execute `create_convergence.sh` script on all `file.colvars.traj` files

```
./create_convergence.sh
```

open the file and make sure the stride and max_lines are correct

  - You will now have several files labeled file.colvars.traj but they will now end with numbers (e.g. 50000, 100000). These correspond to the time lengths (in fs) of each file grouping.

I generally create new folders where I gather all the files based on their time length. (e.g. gathering all 50000 files together into one folder).

- To refocus, we have just completed preparing our files for all of our analysis
  - First, we have the `file.colvars.traj` files for WHAM analysis
  - Then, we have the `file.hist` files to create our histograms
  - Lastly, we have the `file.convergence+(stride)` files to confirm we have done adequate MD sampling
- Our next step will be to create a `metadatafile`

```
timeseries/0.5/n20_f17.colvars.traj.50000   -20.0   0.05
timeseries/0.5/n25_f17.colvars.traj.50000   -25.0   0.04
timeseries/0.5/n30_f17.colvars.traj.50000   -30.0   0.035
timeseries/0.5/n35_f17.colvars.traj.50000   -35.0   0.035
timeseries/0.5/n40_f17.colvars.traj.50000   -40.0   0.035
timeseries/0.5/n50_f17.colvars.traj.50000   -50.0   0.035
timeseries/0.5/n60_f17.colvars.traj.50000   -60.0   0.03
```

Figure 8: Example of the `metadatafile` file

- Column 1 = path to the `colvars.traj` file
- Column 2 = the dihedral angle for each window
- Column 3 = force constant used (this can be found in your ushp36.in file)

The path shown goes to a file ending in 50000 , this is the result of the convergence testing I performed where I am analyzing only $0.5ns$ out of the full $2.0ns$

When you have completed your `metadatafile` you will need to open `runwham.sh` in vim using your terminal

Why are some of my dihedral angles in increments of 5°? It's because I didn't generate enough $p(x)$ data using 10° increments and needed more windows

- Make sure that all command line arguments are correct
- Refer to the WHAM tutorial URL above to find details on command line arguments
- You will be running WHAM on all of the convergence file 'groupings' you have

4.3 Let's run WHAM analysis!

The final steps in generating your PMF! You're almost at the finish line!

- execute the wham shell script in the directory containing your metadatafile

```
./runwham.sh
```

- After the calculation is complete there should be an output file named `freefile`
- Use whichever preferred method of graphing and graph the first two columns of the `freefile`
  - Column 1 = X-values [*degrees*]
  - Column 2 = Y-values [*free energy in kcal/mol*]

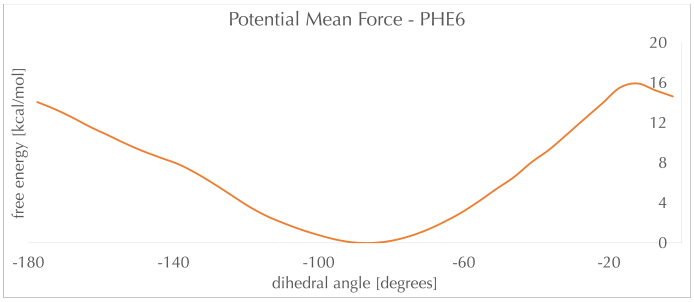

Figure 9: Example PMF graph

Congratulations! You now have a PMF graph

Thank you for using my tutorial on Umbrella Sampling & WHAM Analysis! I wrote this during my time as an undergraduate in the Chemistry Department at CU Denver. I have tried my best at catching any mistakes I had, but there are most likely errors I did not catch. If you have any suggestions/corrections please email me at **tanja_kovacevic@berkeley.edu** and I will be timely in my corrections. Please be sure to site any programs used appropriately. This includes NAMD, VMD, and WHAM. You can also site this tutorial: Kovacevic, Tanja "Umbrella Sampling WHAM Tutorial" 2019

References

[1] G. Torrie and J. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling," *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, 1977.

[2] B. Roux, "The calculation of the potential of mean force using computer simulations," *Computer Physics Communications*, vol. 91, pp. 275–282, Sept. 1995.

[3] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," *Journal of Computational Chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.

[4] A. Grossfield, "Wham: the weighted histogram analysis method," vol. 2.0.10.