Fitting to torsional profiles

Ivan Welsh

February 17, 2016

1 Molecules

A set of twenty, simple organic molecules were chosen for this investigation. The aim of the investigation was to determine the bonded distance required between a dihedral of interest and a functional group, for the functional group to have minimalism effect on the dihedral rotation profile. The molecules used are shown in figure 1.
In order to investigate the dihedral rotation profiles, QM calculations where the dihedral of interest (ψ in figure 1) has been fixed at a given value, for a range of values. At each fixed value, the structure was optimised. Energy differences between the fixed values give the dihedral rotational profile. All calculations were performed with the B3LYP density functional.

2 Basis Sets

A quick look at the effect of basis set on the dihedral rotation profile was under taken. The molecules shown in row 0 of figure 1 were exclusively used for this investigation. ψ was fixed at 15 degree increments in the range of 0 to 360 degrees, to give 24 regularly spaced conformations for each molecule. All calculations were performed at five different basis set levels (6-31G(d), 6-31+G(d), 6-311+G(2d,p), 6-311++G(2d,2p) and 6-311++G(3df,3p)) to determine the basis set error. RMSD results are given in table 1.

Results indicate that it should acceptable for further calculations to be performed at the B3LYP/6-31G(d) level of theory. The two RMSD values reported, indicate different things. The first (RMSD w.r.t QM) is the RMSD between the QM calculated energies and energies calculated by performing a two-pass, linear least squares fitting to the QM energies, using the equation $U = \sum k_i \cdot \cos(m_i\theta + \delta)$. The least squares solves for k_i and (indirectly) δ. A negative k_i gives a δ of 180, positive a δ of 0. The two-pass nature performs a first pass, where values of $m_i = 0 \rightarrow 6$ are used to generate linear equations, which are solved for k_i and sorted based on the magnitude of k_i. The most important N values of m_i are then set as the only possible values, and passed through the same fitting process again. $m_i = 0$ is always in this fitting to account for any DC shift within the data. The RMSD obtained between the QM energies and the fitted energies gives an indication of how good the fit is. As shown in table 1, the RMSD between the QM calculated energies and the energies calculated from fits sits slightly higher than k_bT

`Images/molecules.png`

Figure 1: Structures of molecules torsional profiles attained for

Need to add the -1 row

Table 1: RMSD values obtained from QM calculations of the dihedral rotational profile a the 0 row of molecules. Two RMSD values are given. The first is the RMSD between the QM calculated energy values and the energy values determined from a two pass fitting of three dihedral terms to the energy profile. The second is the RMSD between the energies determined from fits at the given basis set against the smallest basis set. All energies are in kJ mol^{-1}.

Molecule	Basis Set	RMSD w.r.t QM	RMSD w.r.t 6-31G(d)
AMINO0	6-31G(d)	2.07	0.00
	6-31+G(d)	2.59	0.43
	6-311+G(2d,p)	2.55	0.48
	6-311++G(2d,2p)	2.52	0.48
	6-311++G(3df,3p)	2.51	0.37
CHLORO0	6-31G(d)	2.91	0.00
	6-31+G(d)	3.09	0.14
	6-311+G(2d,p)	3.13	0.63
	6-311++G(2d,2p)	3.13	0.70
	6-311++G(3df,3p)	3.14	0.51
HYDRO0	6-31G(d)	1.69	0.00
	6-31+G(d)	2.59	1.04
	6-311+G(2d,p)	2.48	1.10
	6-311++G(2d,2p)	2.45	1.13
	6-311++G(3df,3p)	2.45	1.02
METH0	6-31G(d)	3.26	0.00
	6-31+G(d)	3.44	0.21
	6-311+G(2d,p)	3.37	0.15
	6-311++G(2d,2p)	3.36	0.15
	6-311++G(3df,3p)	3.39	0.15
THIO0	6-31G(d)	3.36	0.00
	6-31+G(d)	3.56	0.18
	6-311+G(2d,p)	3.53	0.38
	6-311++G(2d,2p)	3.52	0.42
	6-311++G(3df,3p)	3.56	0.37

BasisSets/Figures/METH0/Combined.png

Figure 2: Dihedral rotational energy profile of METH0 ψ as calculated with 5 different basis sets

(this RMSD drops to less than 0.5 kJ mol^{-1} when phase is also fitted), which shows a reasonable fit is obtained.

The second RMSD value reported in 1 are the RMSD values between the energies as calculated from the fits of the basis set of interest, and the smallest basis set used here in (6-31G(d)). In all cases, these values remain below $\frac{1}{2}k_bT$, indicating that, in these molecules, there is not much gained from using a much larger basis set. This is further reinforced through inspection of the overlaid energy profiles for the least (figure 2) and most (figure 3) deviated examples. Even in the largest deviation present in the HYDRO0 molecule energy profiles, the 6-31G(d) basis set still performs remarkably well when compared with the larger basis sets.

As such, all further QM calculations will be performed at the B3LYP/6-31G(d) level of theory.

BasisSets/Figures/HYDRO0/Combined.png

Figure 3: Dihedral rotational energy profile of HYDRO0 ψ as calculated with 5 different basis sets

3 Fitting

Torsional energy profiles can be described with a Fourier series, a potentially infinite set of cosines. For molecular dynamics simulations, it is desirable to have a finite series of cosine terms describing the torsional energy profile. Going from the raw data of the energy profile, to the explicit series of cosines requires a fitting process to be under taken. Given a data set spanning a full revolution of a torsional angle, there are a number of possible methods to fit a finite series of cosines to this data set. In each case, there is a need to perform a linear least squares fitting procedure.

3.1 Using simultaneous linear equations to fit to a Fourier series

A Fourier series describing the torsional potential of conformation j is given by

$$V_j = \sum_{i=1}^{N} k_i(1 + \cos(m_i\theta_j - \delta_i)) \tag{1}$$

where V_j is the potential of the system at conformation j, k_i is the amplitude of the i-th term, m_i is the multiplicity associated with the term, θ_j is the dihedral angle of conformation j, and δ_i is the phase shift associated with the i-th term. In general usage, δ_i is limited to be either 0 or π. With this limit imposed, the Fourier series becomes:

$$V_j = C + \sum_{i=1}^{N} k_i \cos(m_i\theta_j)\cos\delta_i \tag{2}$$

where $C = \sum_{i=1}^{N} k_i$. As δ_i is limited to being either 0 or π, $\cos\delta_i = \pm 1$, meaning it can ignored (a negative k_i value can be converted to a positive value with a δ_i of π). By providing a list of n, m_i values (multiplicities are integer values, and so do not need to be fit to), this becomes a series of N simultaneous equations in n variables, the k_i values. This can then be written in the matrix form of $b = Ax$:

$$\begin{bmatrix} V_1 - c^V \\ V_2 - c^V \\ \vdots \\ V_N - c^V \end{bmatrix} = \begin{bmatrix} \cos(m_1\theta_1) - c_1^A & \cos(m_2\theta_1) - c_2^A & \dots & \cos(m_n\theta_1) - c_n^A \\ \cos(m_1\theta_2) - c_1^A & \cos(m_2\theta_2) - c_2^A & \dots & \cos(m_n\theta_2) - c_n^A \\ \vdots & \vdots & \ddots & \vdots \\ \cos(m_1\theta_N) - c_1^A & \cos(m_2\theta_N) - c_2^A & \dots & \cos(m_n\theta_N) - c_n^A \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ \vdots \\ k_n \end{bmatrix} \tag{3}$$

where $c^V = \frac{1}{N}\sum_{j=1}^{N} V_j$ and $c_i^A = \frac{1}{N}\sum_{j=1}^{N}\cos(m_i\theta_j)$. A least squares solving of this matrix equation can then be undertaken, and the values of k_i obtained. It is trivial to extended this method to fit to multiple torsional angles simultaneously.

Expanding to variable δ

As defined thus far, the matrix means of solving the simultaneous equations is unable to solve the equations if δ_i is allowed to vary from $0/\pi$. This is because the equations will become nonlinear, and linear least squares is unable to solve this. However, there is a way to convert the nonlinear, variable phase (ie having to fit k_i and δ_i) equations into linear equations than can be fit.

If the individual torsional energy terms are treated as phasors, we can use the phasor addition rule to show that:

$$k_i(1 + \cos(m_i\theta - \delta_i)) = k_{i,a}(1 + \cos(m_i\theta - \delta_a)) + k_{i,b}(1 + \cos(m_i\theta - \delta_b)) \tag{4}$$

Then, if δ_a and δ_b are set to predetermined, non equal values, we are left with linear simultaneous equations which can be solved as per above. The solutions will give the values $k_{i,a}$ and $k_{i,b}$, but the desired results are k_i and δ_i. It is trivial to obtain these values from the fit determined values:

$$\begin{aligned} k_i^2 = &\quad (k_{i,a}\cos\delta_a + k_{i,b}\cos\delta_b)^2 + (k_{i,a}\sin\delta_a + k_{i,b}\sin\delta_b)^2 & (5) \\ \delta_i = &\quad \arctan\left(\frac{k_{i,a}\sin\delta_a + k_{i,b}\sin\delta_b}{k_{i,a}\cos\delta_a + k_{i,b}\cos\delta_b}\right) & (6) \end{aligned}$$

3.2 Fitting methods

a) **Single pass:** a single pass fit involves a single generation and solve of the linear equations. The data, a list of the multiplicity values to be included in the fit, and an optional limit, N to the number of output multiplicities are provided. The matrices are generated and solved using all the provided multiplicity values, giving the k_i and δ_i obtained for each multiplicity m_i. The multiplicities are sorted in descending order based on the k_i value determined. If a limit has been provided, the first N multiplicity, amplitude and phase tuples are returned as the fitted results.

b) **Twin pass:** a twin pass fit performs two single pass fits. The first single pass is as described above, with a limit of N values to be returned. The second pass takes the returned N multiplicities and performs a fit with only those values. In this way, the refitted parameters will hopefully be a better fit to the data than the 'truncated' parameters, as they compensate slightly for the fitting provided by the discarded parameters.

c) **Multi pass:** instead of relying on a sorted list of fit parameters to determine the N parameters to use, a multi pass method performs fits with all possible combinations of multiplicity values of length N. Thus, if the possible multiplicities are (1, 2, 3, 6), and $N = 3$, the fits performed will have multiplicities (1, 2, 3), (1, 2, 6), (1, 3, 6), and (2, 3, 6). Each fit performed has an error, ϵ, associated with it, between the fit and the raw data. The fit with the lowest such error would thus be determined to be the best fit, and the one to use. A simple error estimation function to use would be the RMSD:

$$\epsilon = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\left(V_{\text{raw}}(\theta_j) - V_{\text{fit}}(\theta_j)\right)^2} \tag{7}$$

Of course, any other error estimation function is also usable.

3.3 Pruning outliers(???)

3.4 Difference data

Determining MD torsional parameters is the aim of this fitting procedure. Data obtained from QM calculations and MD minimisations (where the torsional terms fitting for have been removed/zeroed out of the topology) provide the reference data. The difference between the relative energies of QM and MD data is the data to be fit by the fitting process. In order to obtain this difference data, there are a number of possible methods to use.

1. Difference of raw data. In this approach, each raw QM datum has the MD datum at the same, fixed torsional value subtracted from it, in order to give the reference

value for that point. Though this approach is the most obvious means of determining the difference data, it has a downside. Both the QM data (which is expensive to calculate) and the MD data (which has negligible computational cost), must be determined at the same points, which may not always be desirable or possible.

2. Subtract one set of raw data from calculated data determined from a fit to the other set of raw data. In this approach, two fitting processes are performed. The first is to set to one set of the raw data, for example a high density sampling MD data set, and the second to the difference between the raw data of the other, and the calculated of the first. This method does allow for differing sampling densities, which may be useful.

3. Fit both sets of raw data, and then calculate and fit the difference between them. This is probably the most versatile fitting method. For example, one perform a one pass fit, with no limit, to the raw QM and MD data, utilising the ability to fit phase, and then perform a multi pass fit to the calculated difference, without fitting phase. In this way, very good fits to the QM and MD data could be obtained, while allowing the difference data to be fit in a manner consistent with the force field parameterisation philosophy.

4 Results

4.1 Fitting methods

Each of the three fitting methods described were tested on 50,000 sets of randomly generated periodic data, and 113 sets of real data, from QM, all-atom MD and united atom MD. Random data was generated by from a Fourier series of all multiplicities up to 12. Each term has a randomly generated amplitude, and phase as well as a small amount of noise applied. Real data came from a set of QM, AA-MD and UA-MD calculations performed for later testing purposes. Fitting was performed to generate a 3 term fit.

Fitting to the random data gave identical results within all the phased and non-phased fits. Non-phased fits naturally had a larger RMSD when compared with the phased fit, but all three non-phased fits results were the same, as were all three phased fit results. Real data fitting gave similar results. Two pass and multi pass fitting methods gave identical results, as would be expected. One pass fitting had RMSD values slightly larger than the two/multi pass methods; on average a 0.001% variation (from the RMSD of the two/multi pass methods) for the non-phased fit, and a 0.05% variation for the phased fits. The difference between the random and real data fitting (ie that real had slight differences between the one pass and other fitting methods when random did not) probably arises from the way noise was provided to the random generation method. In the real data, there is small amounts of noise (defining noise as differing from the data falling on a perfect periodic function) in each of the data points. However, a few data points have much larger deviation from the curve. This larger deviation was not modelled by the random generation, as noise was applied to the generating cosine functions, not the generated data points. Regardless, these results indicate that it should be perfectly reasonable to perform fitting with the cheapest one-pass method, without having any noticeable difference in the quality of the fits.

4.2 Difference data

The three methods to determine the difference data are tested. The aim of these tests is to determine which methods provide the best means of calculating the required torsional terms. Tests are run on the -1 series of molecules. The tests are:

- $\text{QM}_{\text{raw}} - \text{MD}_{\text{raw}} = \text{DIFF}$. Fit phased and non-phased to DIFF. MD using both AA and UA. (4 total tests)
- Fit to MD_{raw} using phased and non-phased. $\text{QM}_{\text{raw}} - \text{MD}_{\text{fit}} = \text{DIFF}$. Fit phased and non-phased to DIFF. MD using both AA and UA. (8 total tests)
- Fit to QM_{raw} using phased and non-phased. $\text{QM}_{\text{fit}} - \text{MD}_{\text{raw}} = \text{DIFF}$. Fit phased and non-phased to DIFF. MD using both AA and UA. (8 total tests)
- Fit to both MD_{raw} and QM_{raw}, using phased and non-phased (but *not* mixing the fits). $\text{QM}_{\text{fit}} - \text{MD}_{\text{fit}} = \text{DIFF}$. Fit phased and non-phased to DIFF (non-phased can be fit to phased raw data, but phased cannot be fit to non-phased raw data). MD using both AA and UA. (6 total tests)

These give a total of 26 different tests to be performed, on the 5 molecules in the -1 series. Fits to raw data will be performed with an unlimited one pass fitting. Fits to the diff data will be performed with a limit of 3 terms in a one pass fitting process.