Paramfit: Automated Optimization of Force Field Parameters for Molecular Dynamics Simulations

Robin M. Betz[a] and Ross C. Walker*[a,b]

The generation of bond, angle, and torsion parameters for classical molecular dynamics force fields typically requires fitting parameters such that classical properties such as energies and gradients match precalculated quantum data for structures that scan the value of interest. We present a program, Paramfit, distributed as part of the AmberTools software package that automates and extends this fitting process, allowing for simplified parameter generation for applications ranging from single molecules to entire force fields. Paramfit implements a novel combination of a genetic and simplex algorithm to find the optimal set of parameters that replicate either quantum energy or force data. The program allows for the derivation of multiple parameters simultaneously using significantly fewer quantum calculations than previous methods, and can also fit parameters across multiple molecules with applications to force field development. Paramfit has been applied successfully to systems with a sparse number of structures, and has already proven crucial in the development of the Assisted Model Building with Energy Refinement Lipid14 force field. © 2014 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23775

Introduction

Classical molecular dynamics (MD) simulations integrate Newton's equations of motion over a molecule for a set time step. This method has been used to study condensed phase biomolecular systems including proteins, nucleic acids, carbohydrates, and lipids on biological (μs) timescales. Critical to the success of classical MD simulations is the accuracy of the underlying parameters, collectively termed a force field.

Assisted Model Building with Energy Refinement (AMBER) is a MD software suite widely used by researchers to simulate proteins and biomolecules.[1,2] The potential energy is described in terms of the following equation: [3]

$$V_{\mathrm{AMBER}} = \sum_{i}^{n_{\mathrm{bonds}}} b_i (r_i - r_{i,\mathrm{eq}})^2 + \sum_{i}^{n_{\mathrm{angles}}} a_i (\theta_i - \theta_{i,\mathrm{eq}})^2 + \sum_{i}^{n_{\mathrm{dihedral}}} \sum_{n}^{n_{i,\max}} (V_{i,n}/2)[1 + \cos(n\phi_i - \gamma_{i,n})] + \sum_{i<j}^{n_{\mathrm{atoms}}{}'} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) + \sum_{i<j}^{n_{\mathrm{atoms}}{}'} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \quad (1)$$

This equation calculates energy as the sum of a harmonic potential for bonds and angles, a truncated Fourier series for dihedrals, and Lennard-Jones and pairwise electrostatic potential function for nonbonded forces, with the prime on the nonbonded term sum indicating that the calculation is only performed for atoms in different molecules or separated by at least three bonds. Partial derivatives of this equation with respect to atom position in the *x*, *y*, and *z* directions provide the forces from which to propagate Newton's equations of motion.

The force field parameters in the AMBER Hamiltonian are typically refined to fit quantum level equations at an appropriate level of theory and basis set, typically the highest possible level of theory that can be completed in a reasonable timescale without known biases. Force fields are validated through comparison of simulation derived values (e.g., heat of vaporization, density, or area per molecule) with experimental ones.

The parameters that describe the harmonic potential for bonds and angles may be obtained simply from scanning a set of structures containing a sampling of bond or angle values and plotting the energy of the resulting structures.[4,5] The equilibrium value parameter corresponds to the bond or angle value resulting in a minimal energy structure, and the force constant is described by fitting a quadratic function around this minimum.

Dihedrals are represented by a more complex potential function, but are parameterized in a similar way—a scan of the energy of structures with many possible torsion angles for that dihedral is conducted, and the resulting plot is fit to a truncated Fourier series with typically up to six terms.

The equilibrium value for bonds and angles may also be obtained from experimental data such as infrared, microwave, or neutron diffraction studies. Nonbonded forces are defined by the partial charges and Van der Waals potentials on each

[a] *R. M. Betz, R. C. Walker*
San Diego Supercomputer Center, La Jolla, California
[b] *R. C. Walker*
Department of Chemistry and Biochemistry, UC San Diego, La Jolla, California
E-mail: ross@rosswalker.co.uk

Contract grant sponsor: National Science Foundation Scientific Software Innovations Institutes Program—NSF SI2-SSE (R.C.W.); Contract grant number: NSF1047875 and NSF1148276; Contract grant sponsor: University of California (R.C.W.); Contract grant number: UC Lab 09-LR-06-117792; Contract grant sponsor: University of California Institute for Mexico and the United States (UC MEXUS) and the Consejo Nacional de Ciencia y Tecnología de México (CONACYT); Contract grant number: CN-13-554; Contract grant sponsor: CUDA fellowship (R.C.W.) (NVIDIA Inc.)

atom, and well defined methods such as RESP exist for their derivation.[6]

Obtaining the truncated Fourier series describing dihedrals presents a significant obstacle to force field development. Usually, dihedral terms are derived by fitting to a quantum-level rotational scan about a dihedral of interest or to several stationary points on the dihedral potential[7] either manually[8] or with algorithms such as Monte Carlo simulated annealing.[9] However, this method requires a large number of expensive quantum calculations requiring significant computational investment. Furthermore, it can prohibit fitting multiple dihedral types simultaneously, which is problematic given the coupled nature of most molecular dihedrals.

Existing methods for parameterization of custom small molecules, such as Antechamber[10] or Paramchem,[11,12] usually retrieve parameters by analogy to similar molecules that have already been parameterized as part of an existing forcefield. However, the assumption that parameters from seemingly analogous molecules are identical may not be valid, especially if the analogy is determined by software. Methods used to derive parameters as part of force field development rely on a small number of conformational samples,[7] involve expensive calculations of vibrational spectra and geometry optimization at a quantum level of theory,[13] or require hand-tuning of resulting parameters.[14] These methods additionally have little to no support for the simultaneous fitting of multiple parameters, resulting in more quantum calculations and potential neglect of coupling effects as each parameter must be fitted individually to a set of quantum data that sample that parameter rather than using a common set of calculations to derive all parameters.

There is a dearth of software that can assist the average computational chemist in obtaining parameters for a small molecule—if assigning parameters by analogy to other molecules fails to accurately describe the system of interest, the researcher must become familiar with force field development to obtain parameters from first principles. Several projects aimed at addressing this need are still under development or defunct—the ParamChem gateway[11] does not yet provide functionality to generate dihedral parameters, and the visual molecular dynamics (VMD) plugin ParaTool, which aimed to derive Chemistry at HARvard Macromolecular Mechanics (CHARMM) parameters from quantum mechanical (QM) calculations, halted in the development stage years ago. Similarly, the ParmScan[15] program, which could obtain the Fourier series for a dihedral using a genetic algorithm, is not publicly available. Existing user-friendly tools such as ffTK[16] and general automated atomic model parameterization (GAAMP)[17] can fit parameters to QM data, and can derive a single dihedral force constant at a time through a rotational scan.

We present here a program named Paramfit that is designed to address derivation of bonded terms in the AMBER equation in a systematic way with an emphasis on minimizing the amount of necessary quantum calculations. Paramfit's interface guides users through the creation of *ab initio* calculation input files to the generation of parameter files for AMBER's preparatory programs for simulation.

Paramfit is capable of refining any parameter in the bonded terms, including force constants K_r and K_θ, equilibrium bond length r_{eq} or angle θ_{eq}, dihedral barrier height V_n, dihedral phase γ or periodicity n. Any combination of these parameters can be fit simultaneously, given a single set of input structures. In fact, Paramfit can fit multiple parameters at once given any reasonable conformational set.

The program is designed to address the needs of all users who wish to generate force field parameters for use in AMBER or other programs that use the AMBER force field. Paramfit played an integral role in the refinement of the AMBER lipid force fields GAFFLipid[18] and Lipid14,[19] where it was used to fit multiple coupled torsional terms in glycerophospholipid molecules. Paramfit provides a powerful tool for new molecular systems for where there are incomplete or insufficient parameter sets available, offering an efficient method for manual parameterization.

Methodology

Force fields are parameterized against *ab initio* quantum data or experimental measurements so that energies or forces calculated with force field parameters match the given quantum or experimental data. Traditionally, terms used in the AMBER force fields have been parameterized by conducting a scan across a variety of structures sampling the torsion angles of interest. Paramfit is designed for fitting to *ab initio* quantum data, and can fit to either single-point quantum energies or atomic forces. When fitting to energies, the program optimizes parameters with the goal of minimizing the following least squares fitness function:

$$f(N, E_{QM}, K) = \sum_{i=1}^{N} \left[(E_{MM}(i) - E_{QM}(i) + K)^2 \right] \quad (2)$$

where N is the number of molecular conformations to consider, E_{QM} is the single-point quantum energy of each conformation, and E_{MM} is the energy calculated using the AMBER equation [eq. (1)] with a potential parameter set. K is a constant offset that accounts for the different origins in the quantum and AMBER energies, and allows minimization to zero to be conducted.

Fitting to first derivatives operates under a similar approach, using the vector norm to quantify differences in forces:

$$f(N, N_{atoms}, F_{QM}) = \sum_{i=1}^{N} \sum_{atom=1}^{N_{atoms}} |\mathbf{F}(i, atom)_{MM} - \mathbf{F}(i, atom)_{QM}|^2 \quad (3)$$

Forces are summed either for all atoms in the molecule, or only for those involved in a parameter to be optimized. This option can reduce noise and ambiguity that is often present in the energy landscape, especially for structures that are not minimized before parameter fitting.

This optimization poses an extremely challenging problem, especially when fitting more than one parameter. The fitness landscape is very complicated, with a number of minima, and often features attractive local minima with parameters that are

physically unreasonable. The dimensionality of the problem is often very high in a typical use case, and given the possible parameter space to search, it is difficult to sample a representative landscape. To be efficient, any minimization algorithm implementation must quantify the number of samples required for convergence and reduce the number of function evaluations to find the minimum, and provide reproducible results while exploring a maximal amount of the search space.[20]

Paramfit implements a hybrid genetic algorithm to conduct the minimization, with refinement using a simplex algorithm to accelerate convergence. A genetic algorithm minimizes in a method analogous to biological evolution—an initial population of size η is created at random and selection, recombination, and mutation operations are carried out on this population in successive generations until an optimum is reached.

The genetic algorithm starts with an initial population of randomly generated parameter sets within a physically reasonable range. The sets are then ranked according to their performance on the fitness function. A certain percentage as defined by the algorithm parameter ρ of the sets is allowed to proceed to the recombination step.

For recombination, two parents are chosen uniquely from the selection pool and are combined in one of two ways, selected at random. Several recombination methods common to genetic algorithms were tested and extensive trials found the combination of these two methods that produced the most effective convergence.[21]

The intermediate recombination method sets each child parameter randomly within the range between that of the two parents.[22] This method is most beneficial to the population when a parameter is close to the optimum. The linear crossover recombination method chooses a split point at random; all parameters in the set that occur before the split point come from one parent, and the remainder from the other.[23] This method improves fitness by allowing parameters to be inherited independently of each other but retain the favorable value found in the parent.

Following recombination, the mutation operation takes place on a randomly chosen amount of parameters in the population as defined by the algorithm parameter δ. The fitness of each member of the population is recalculated, the population is sorted, and the next generation begins.

Convergence is reached when the best fitness within the population remains unchanged for a threshold number of generations τ. In Paramfit's hybrid genetic algorithm, when the best fitness remains unchanged from one generation to the next, a simplex algorithm is run with weak convergence criteria starting with the parameters specified by each of a random 5% of the population.

This novel combination of genetic and simplex algorithms results in increased convergence speed compared to a genetic algorithm alone, as shown in Figure 1. Genetic algorithms excel at producing an exponential decrease in function value within the first few generations, but following finding the neighborhood of the minimum begin to stagnate and rely on

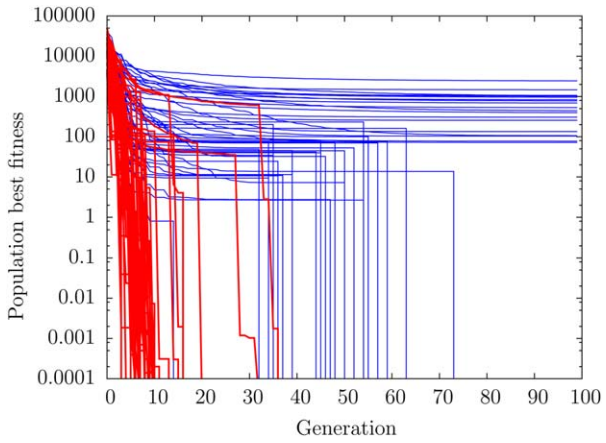

Figure 1. Algorithm convergence with simplex iterations for 50 runs on blocked alanine tetrapeptide. In red are the runs with simplex refinement, and in blue those without. All runs with simplex found the correct global minimum within 35 generations, but many without continued to run for 100 generations or more without convergence.

rare mutation events to improve fitness. Conversely, simplex algorithms are excellent at finding a minimum given its neighborhood, and improve the genetic algorithm population in a much more directed manner than random mutation.

The algorithm is quite sensitive to the choice of its own internal parameters, and as such they must be chosen carefully to ensure optimal performance. For example, if the mutation rate δ is too high, good solutions will be frequently eliminated as the mutation operation is usually destructive, but if it is too low the algorithm may become trapped in local minima. This δ was optimized by running Paramfit repeatedly to scan this value, obtaining a function-independent optimum at δ=0.05. This ρ=0.35 was obtained via the same method, and is also function-independent as shown in Figure 2.

Perhaps the most important algorithm parameter is the number of generations to converge, τ. If τ is too small, values that do not represent the global minimum will be returned, and if it is too large, computing power will be wasted running needless generations. A pure genetic algorithm requires a large value for τ, as infrequent mutations will often improve fitness in later generations; however, Paramfit's combined genetic-simplex approach reduces τ considerably, as the simplex iterations will always result in population improvement unless the minimum has been found. A value of $\tau = 5$ is the default for the program, and results in a notably decreased number of function evaluations.

All of the values for these parameters may be adjusted in Paramfit's input to establish stricter convergence criteria.

The population size η is function-dependent. Large values of η provide greater sampling of the solution space, which may result in faster initial progress, but require many more function evaluations for convergence. A smaller value of η can result in slower convergence and also more function evaluations, depending on the initial generation. However, following convergence the algorithm will retrieve the same parameters regardless of η. The default population size of 50 results in a near-minimal number of function evaluations for the majority

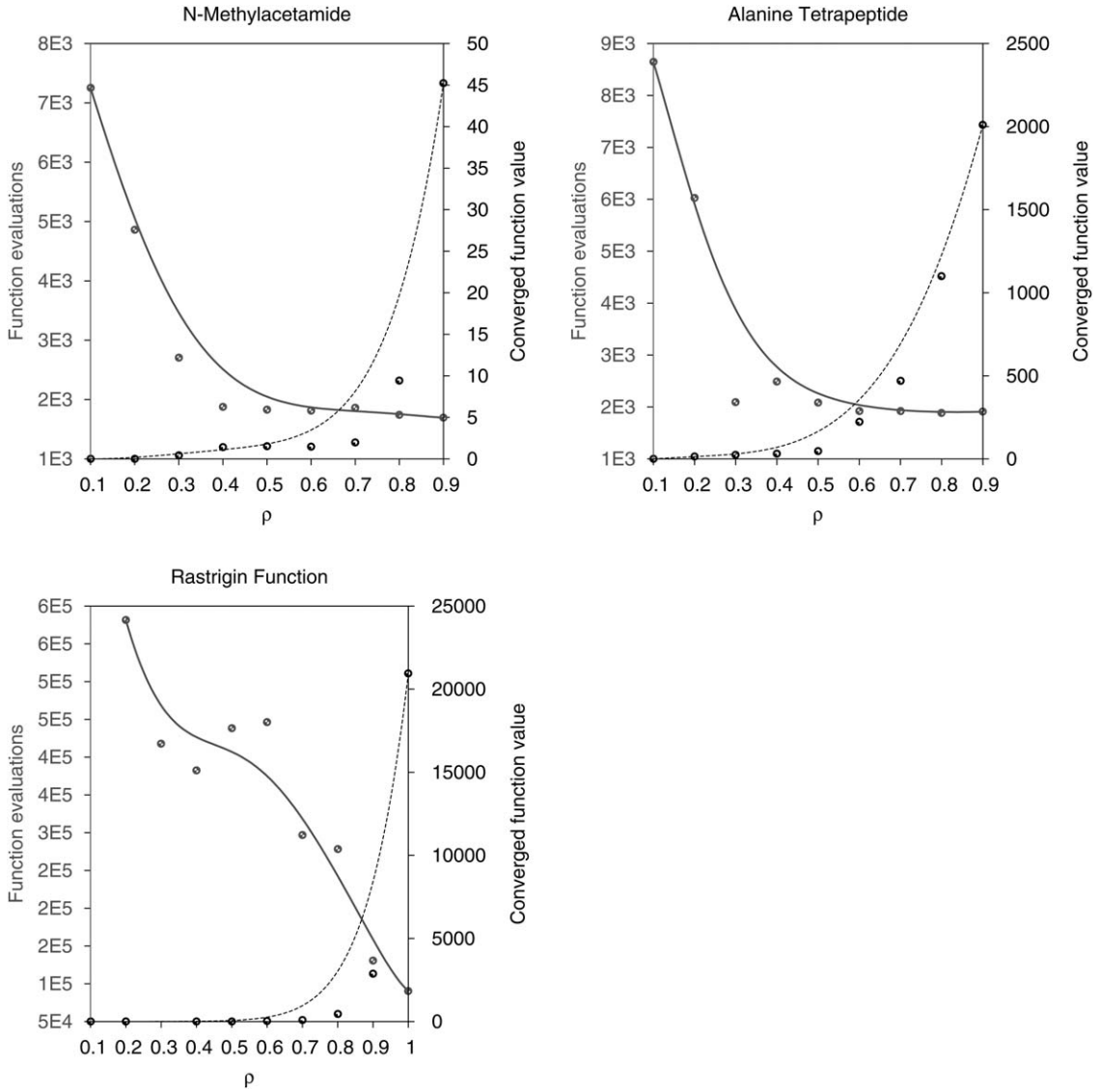

Figure 2. A plot of the mean converged function value (grey, triangles) and the mean number of function evaluations (black, circles) over several values of the parent percentage parameter ρ. The optimal value of ρ preserves accuracy with the fewest number of function evaluations, and is system-independent, with a value of approximately 0.35. Multiple amide backbone dihedrals were fit over 100 and 500 conformations of alanine tetrapeptide and N-methylacetamide, respectively, for 10 runs at each ρ value. A scan of ρ was also performed for 30 trials at each value for minimizing a 10-dimensional Rastrigin function (see Results).

of molecular fits, however when the algorithm is applied to other minimization problems, η should be rederived as shown in Figure 6.

Additional fitting features

To facilitate force field development, fitting may be performed to one or more parameters over several different molecules in independent input structures. The fitness function used by the algorithm then becomes:

$$f(N, E_{QM}, \text{molecules}, K) = \sum_{\text{molecules}} \sum_{i=1}^{N} [(E_{MM}(i) - E_{QM}(i))^2 + K] \quad (4)$$

This fitness function is minimized by the same algorithm as single-molecule fits, and the resulting parameters will be applicable over all of the molecules given. This enables the development of general parameters that describe classes of molecules rather than individual structures, and has been applied to phospholipid force field development in Lipid14 (see Results section).

Additionally, input structures may be given weights within the fitting algorithm. This method may be used, for example, in Lipid14 to fit alkane chain torsions. High energy structures with less favorable conformations can be given lower weights. These results in the following fitness function being used for fitting to energies:

$$f(N, \mathbf{w}, E_{QM}, K) = \sum_{i=1}^{N} w_i [(E_{MM}(i) - E_{QM}(i))^2 + K] \quad (5)$$

A higher w_i for some structure i will increase the relative energy agreement of that structure.

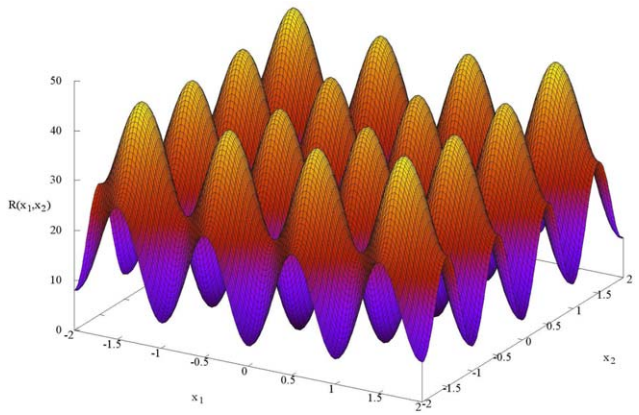

Figure 3. Three-dimensional view of the Rastrigin function in two dimensions from $x\epsilon[-2, 2]$. The landscape is complicated, with multiple local minima that are very close to the global minimum at 0.

Usability

Paramfit contains a number of features designed to make it accessible to computational researchers. Automated functionality is included to create input files for several quantum programs given an input set of structures in AMBER coordinate or restart format.

Paramfit parses the resulting outputs from the quantum package (currently automatic parsing of Gaussian, Amsterdam Density Functional (ADF), and general atomic and molecular electronic structure system (GAMESS) outputs is supported) to extract the relevant energy or forces for each structure, eliminating the need for scripting.

Prompts allow the user to define the specific parameters to be fit efficiently, and the list of these parameters is saved for use in subsequent runs.

There are several structure validation tools included with Paramfit to evaluate the quality of the input conformations, which are crucial for parameter refinement. Paramfit includes functionality to generate plots demonstrating sampling of relevant bond, angle, and dihedral length, angle, or torsion values, and will give warnings in program output when structures inadequately sample certain conformational space, potentially leading to poorly refined parameters. For example, dihedral parameterization requires a structure sampling at least every 10° for the dihedral of interest. Bond and angle parameters require the final converged equilibrium angle value be within $\pi/20$ of a sampled conformational value. These thresholds may be changed by the user to adjust the strictness of the bounds checking functionality.

Run time options, including the format of input and output as well as algorithm parameters, are specified in an input job control file. Given the numerous features of Paramfit, a wizard is included that assists the user in the creation of the input files.

Results

Three case studies of Paramfit's usage in a variety of scenarios were examined to assess the program's performance and robustness.

The Rastrigin function

The efficacy of Paramfit's core algorithm was verified by minimizing to the Rastrigin function, a function commonly used in minimization algorithm testing that features numerous local minima and one global minimum.[24] The function in n dimensions is as follows:

$$R(x)=10n+\sum_{i=1}^{n}(x_i^2-10\cos 2\pi x_i) \quad (6)$$

The function is defined for all x_i, and the global minimum is at $\forall x_i=0$, $R(x)=0$ (Figure 3).

Paramfit's genetic algorithm was used to minimize the Rastrigin function in a variety of dimensions, and was able to successfully identify the global minimum in all trials with an efficient number of function evaluations.

The algorithm started with an initial population of 50–500, with initial values randomly selected in the range $x_i\epsilon[-1000, 1000]$.

The algorithm was able to successfully find the global minimum to at least 4 decimal places in all cases, and was tested on systems up to 15 dimensions. Although the number of function evaluations required predictably increases with dimensionality, the algorithm scales well, as shown in Figure 4. The algorithm is less efficient than other genetic algorithms on this problem,[25] but this can be attributed to its stricter convergence criteria and use of simplex iterations to ensure the bottom of the well is reached.

Interestingly, when the algorithm was given a uniform initial population with -1000 for each value, convergence to the correct global minimum was achieved, albeit with a very high number of function evaluations. This demonstrates that the algorithm's success does not depend on the values contained within the randomly selected initial population (Fig. 5).

To verify algorithm parameter tuning, the algorithm was run 100 times with a variety of values for the initial population size η on an eight-dimensional Rastrigin function and the number of function evaluations required for convergence averaged. The resulting curve, shown in Figure 6, illustrates how the optimal population size of approximately 400 is evident.

Alanine tetrapeptide

To verify the algorithm's ability to fit a realistic molecular system, Paramfit was used to generate the ff99SB modified version[26] of the ff99 force field using the same system used in the original derivation. The original AMBER ff99 [27] parameters misrepresented torsion terms on the amide backbone, resulting in over-stabilization of α-helices in protein simulations, and ff99SB corrected this bias by adjusting dihedral torsion terms.

The correction was derived on blocked alanine tetrapeptide (Fig. 7) by fitting each dihedral individually to an *ab initio* level quantum scan. Paramfit was used to do a similar derivation while fitting all backbone dihedrals simultaneously. To confirm that the algorithm finds the global minimum energy difference for realistic systems, the fit is conducted to classical energies calculated with the ff99SB parameters.

Initial input structures were generated from a scan of ϕ and ψ from 0 to 180° every 5°. The energy of the structures was calculated using the AMBER equation and ff99SB parameter

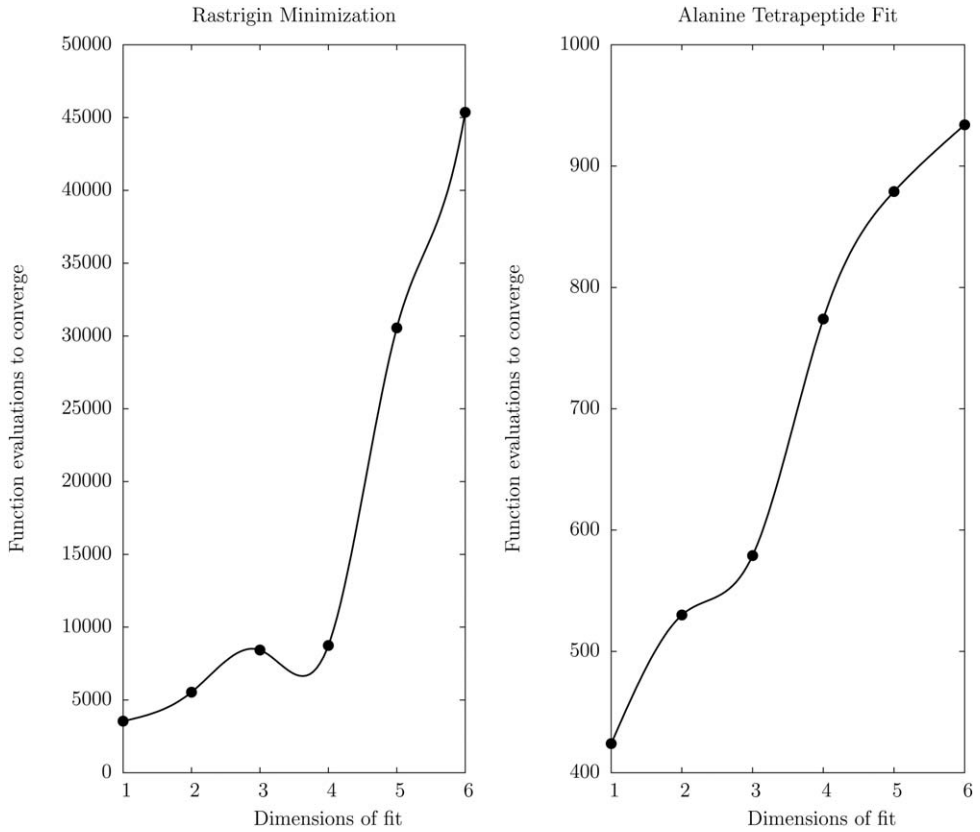

Figure 4. Scaling of algorithm function evaluations with minimizing increasing dimensions of fit for both the Rastrigin test case and a fit to 50 conformations of blocked alanine tetrapeptide (Fig. 7). To account for variance between individual fits, the algorithm was run 50 times on each dimension.

set. Any structures with an energy higher than 2000 kcal/mol (representing severe clash or even overlap between atoms) were discarded. This left a number of strained structures in the remaining 1301 valid structures, but prevented the fit from being biased by attempting to describe extremely rare and high energy conformations.

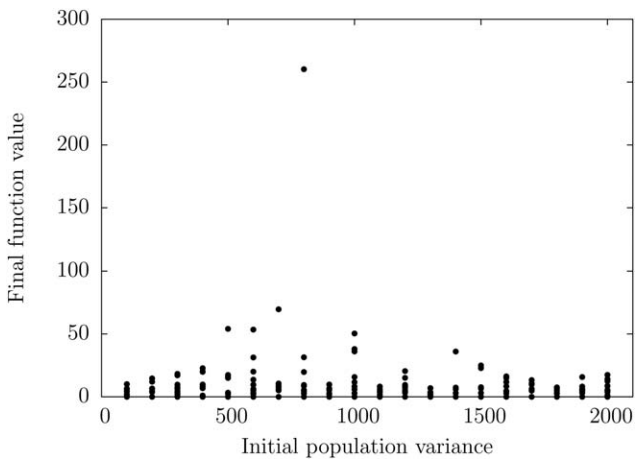

Figure 5. Paramfit's algorithm functions is capable of finding the global minimum independently of initial population variance, demonstrating random sampling of solution space is not required for successful convergence. The algorithm was modified to sample initial populations with $x_i \epsilon (-1000, -1000+\text{variance})$, with variance = 2000 representing a normal use case's completely random sampling. The algorithm run 50 times on the six-dimensional Rastrigin function, and the final function value plotted.

Paramfit was used to fit the six dihedral torsion values for ϕ and ψ. The initial parameters given were set to a value of 1.00 for each term, although this value is not important as for this experiment the program did not base its search on the initial value. (If desired, Paramfit can be used to refine existing parameters). A random selection of structures were chosen to fit to, ranging in number from 2 to all 1301 and the fit was performed.

In all cases, Paramfit fit the objective function to 0.000. With structures numbering greater than 10, the program always

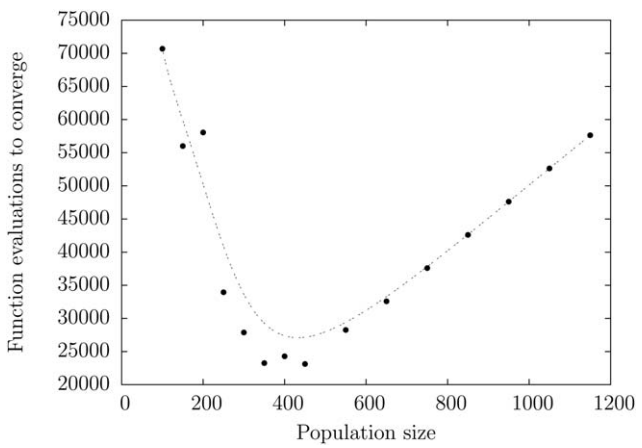

Figure 6. Number of function evaluations vs. η for the Rastrigin function in eight dimensions with a line of best fit. The number of function evaluations refers to the mean evaluations required for convergence over 100 different runs of the algorithm.

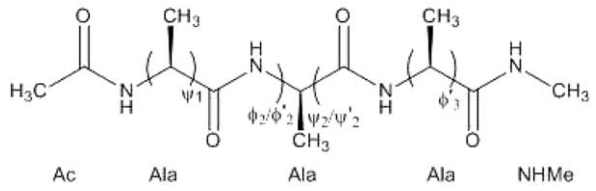

Figure 7. Blocked alanine tetrapeptide. The carboxy carbons are assigned atom type C, while the alpha carbons have atom type CT. Dihedral ϕ consists of atoms C-N-CT-C, and ψ is N-CT-C-N, and each is represented three times in the molecule.

correctly recovered the ff99SB parameters. With a smaller amount of structures, the algorithm still minimized the sum squares energy difference to zero, but recovered a similar but nonidentical dihedral torsion profile for the parameters (Table 1).

The structures used in the fit were chosen at random from a set of 1301 each time the algorithm was run. The program's success at recovering the ff99SB parameters each time using as few as 10 structures from the set demonstrates how a complete scan of each dihedral is not necessary for parameter generation. However, some degree of sampling of torsion space is required for the parameters to be applicable in simulations. If structures that represent only a small range of torsion angles are used in fitting, the resulting parameters will result in accurate energy calculations when that dihedral is within that torsion range, but are not guaranteed to provide accurate results for other values (Fig. 8).

Lipid 14

Paramfit was used in the development of the GAFFLipid[18] and Lipid14[19] force fields to generate dihedral parameters for lipid tails, resulting in the current Lipid14 parameter set that allows tensionless simulation of lipid bilayers with AMBER. Paramfit was used to fit torsion parameters for tail and ester linkage regions of several lipids during the development of these force fields.

The CH_2—CH_2—CH_2—CH_2 alkane torsion potential was fit with Paramfit to the energy of structures from torsion scans performed on hexane and octane molecules evaluated using the hybrid method for interaction energies (HM-IE).[28] Initial structures were generated from a 15° torsion scan of hexane and octane and then optimized at the MP2/cc-pVDZ level before performing the single-point energy calculation (Fig. 9). To emphasize physically reasonable conformations, the *tgt* hexane local minima and *tgttt* octane local minima were given a weighting of 10, all other local minima were given a weight of 4. The remainder of the structures was weighted at 1, while the high-energy *cis* conformers were given a weight of 0.1. The resulting torsion parameters reproduce the quantum energy profile with considerably more success than standard general amber force field (GAFF) parameters.

Paramfit was also used to generate torsion parameters involving atoms of the ester linkage region between the lipid head group and tail group, bringing their energy into agreement with quantum data.

Lipid14 used the new force field refinement features of Paramfit including the hybrid genetic algorithm, structure weighting, and fitting to multiple molecules. Validation of the new parameter set was conducted by comparison to multiple experimental lipid bilayer properties including density and X-Ray scattering profiles. Paramfit was able to successfully parameterize critical lipid torsions from quantum mechanics energies that reproduce experimental lipid bilayer properties.

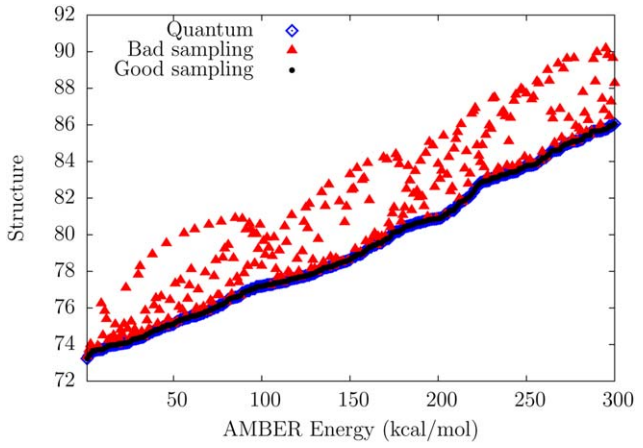

Figure 8. The conformational space sampling of the input structures to Paramfit determines the quality of the resulting parameters. The energy of 500 structures was evaluated with parameters obtained from a 10 structures fit to ff99SB energies randomly sampling dihedral space for ϕ and ψ (blue circles), and with those obtained from a 10 structures fit sampling a 15° range of both dihedrals (red stars). Each fit successfully minimized the algorithm function to 0.00. The parameters resulting from the adequately sampled run reproduce the true ff99SB parameters (black) while the other set can be off by over 6 kcal/mol. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Discussion

Paramfit greatly simplifies the generation of bonded parameters for use with AMBER MD simulations. Previously, the

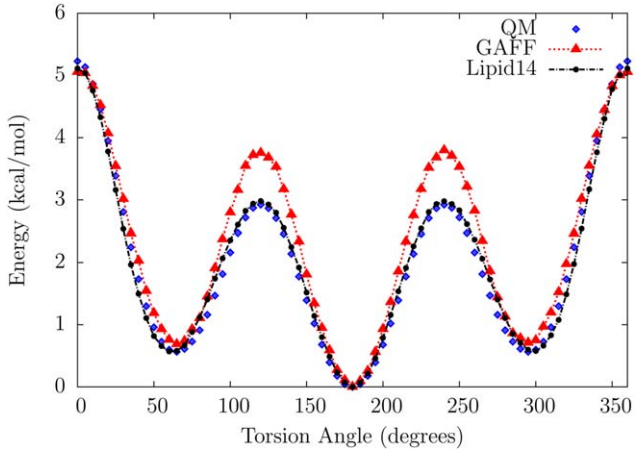

Figure 9. An energy scan of the CH_2—CH_2—CH_2—CH_2 torsion angle of octane at the QM level with the HM-IE relation (blue diamonds), the GAFF parameters (red triangles) and Lipid14 parameters (black circles) demonstrates improved energy calculations following the use of Paramfit. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 1. Paramfit was able to recover ff99SB torsion terms correctly with as few as 10 structures.

# Structures	ψ_1	ψ_2	ψ_2	ϕ_1	ϕ_2	ϕ_3
1000	0.4500	1.5800	0.5500	−0.0000	0.2700	0.4200
750	0.4500	1.5800	0.5500	0.0000	0.2700	0.4200
500	0.4500	1.5800	0.5500	−0.0000	0.2700	0.4200
250	0.4500	1.5800	0.5500	0.0000	0.2700	0.4200
100	0.4500	1.5800	0.5500	0.0000	0.2700	0.4200
50	0.4500	1.5800	0.5500	0.0000	0.2700	0.4200
20	0.4500	1.5800	0.5500	−0.0000	0.2700	0.4200
10	0.4500	1.5800	0.5500	0.0000	0.2700	0.4200
7	0.4500	1.5799	0.5501	−0.0000	0.2699	0.4200
5	0.4716	1.4492	0.4848	0.4871	0.0874	0.9257
ff99SB	0.45	1.58	0.55	0.00	0.27	0.42

Structures were selected randomly from a scan set and the algorithm was run. The runs were repeated 10 times with a new random structure assortment each time, and representative parameters are shown in the table. All runs retrieved the correct parameters to four decimal places for $n \geq 10$.

problem of generating bonded parameters required a large number of quantum calculations and a serial process of fitting one parameter at a time.[12,17] Paramfit simplifies this process by automating the fit using a novel combination of a genetic and simplex algorithm.

This algorithm is able to consistently find the global minimum within a poorly sampled, multidimensional landscape with many local minima, as verified by its performance on the Rastrigin function as well as its ability to recover known parameters.

Obtaining parameters with Paramfit requires far fewer quantum calculations than other methods due to its novel ability to fit to an arbitrary set of input structures rather than requiring a scan over parameters of interest. Additionally, Paramfit can fit multiple parameters simultaneous, allowing for parameter coupling to be accounted for in the resultant force field as well as further reducing the number of quantum calculations required to obtain a comprehensive set of parameters for a system.

Paramfit also greatly simplifies the development of force fields that describe entire classes of molecules, by allowing the derivation of one parameter across multiple molecules. For example, data from every amino acid could be weighted and used to generate terms that describe the behavior of dihedrals in protein backbones in general.

Aimed at both average users and force field developers, the program's options streamline the parameter derivation workflow at each step, from writing the quantum input files to fitting parameters to generating an output that can be easily read into preparatory programs for simulation. A text-based wizard can walk users who wish to generate parameters for a specific system of interest through the entire process, while powerful features such as weighting individual structures and fitting multiple molecules allow the force field developer to produce general parameters with ease.

Conclusions

Paramfit is open-source and distributed with AmberTools, the free component of the AMBER suite of programs. Its algorithm minimizes the energy or force fitness function in a naïve manner that allows it to be applied to any minimization problem, and has been demonstrated to successfully find the global minimum of complicated landscapes such as the Rastrigin function. This core algorithm may be applied to other minimization problems outside of MD, and within MD Paramfit's range of potential applications is large. This program is a powerful and efficient solution for refining force fields.

We aim to apply Paramfit to further force field development problems, including the extension of the Lipid14 force field to other lipids or membrane components such as cholesterol. Its release as part of AmberTools and emphasis on useability also make it attractive to users seeking to refine general parameters such as those from GAFF to better describe small molecules. Finally, Paramfit's powerful minimization algorithm may be applied to other MD force fields through the addition of support for other force field equations.

Acknowledgment

The authors thank Prof. David Case of Rutgers University for helpful comments and assistance during preliminary development of Paramfit.

Keywords: classical molecular dynamics · force fields · parameterization · minimization algorithms

How to cite this article: R. M. Betz, R. C. Walker. *J. Comput. Chem.* **2014**, DOI: 10.1002/jcc.23775

[1] P. K. Weiner, P. A. Kollman, *J. Comput. Chem.* **1981**, *2*, 287.
[2] D. A. Case, T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Götz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko, and P. A. Kollman, AMBER 12, University of California, San Francisco **2012**.
[3] R. Salomon-Ferrer, A. W. Göetz, D. Poole, S. Le Grand, R. C. Walker, *J. Chem. Theory Comput.* **2013**, *9*, 3878.
[4] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, P. Weiner, *J. Am. Chem. Soc.* **1984**, *106*, 765.
[5] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, P. A. Kollman, *J. Am. Chem. Soc.* **1995**, *117*, 5179.
[6] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *J. Phys. Chem.* **1993**, *97*, 10269.
[7] D. Q. McDonald, W. C. Still, *Tetrahedron Lett.* **1992**, *33*, 7743.
[8] J. P. Jämbeck, A. P. Lyubartsev, *J. Phys. Chem. B* **2012**, *116*, 3164.
[9] O. Guvench, A. D. MacKerell, Jr., *J. Mol. Model.* **2008**, *14*, 667.
[10] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157.
[11] J. Ghosh, N. Singh, Y. Fan, S. Marru, K. Vanomesslaeghe, S. Pamidighantam, Molecular Parameter Optimization Gateway (ParamChem), In Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery (ACM, 2011), Salt Lake City, UT, USA.
[12] K. Vanommeslaeghe, E. Prabhu Raman, J. A. D. MacKerel, *J. Chem. Inf. Model.* **2012**, *52*, 3155.
[13] S. E. Feller, D. Yin, R. W. Pastor, A. D. MacKerell, Jr., *Biophys. J.* **1997**, *73*, 2269.

[14] O. Krämer-Fuhrmann, J. Neisius, N. Gehlen, D. Reith, K. N. Kirschner, *J. Chem. Inf. Model.* **2013**, *53*, 802.
[15] J. Wang, P. Kollman, *J. Comput. Chem.* **2001**, *22*, 1219.
[16] C. Mayne, J. Saam, S. K., E. Tajkhorshid, J. Gumbart, *J. Comput. Chem.* **2013**, *34*, 2757.
[17] L. Huang, B. Roux, *J. Chem. Theory Comput.* **2013**, *9*, 3543.
[18] C. J. Dickson, L. Rosso, R. M. Betz, R. C. Walker, I. R. Gould, *Soft Matter* **2012**, *8*, 9617.
[19] C. J. Dickson, B. D. Madej, Å. A. Skjevik, R. M. Betz, K. Teigen, I. R. Gould, R. C. Walker, *J. Chem. Theory Comput.* **2014**, *10*, 865.
[20] E. P. Ron Wehrens, L. M. Buydens, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 151.
[21] D. Whitley, *Stat. Comput.* **1994**, *4*, 65.
[22] H. Mühlenbein, D. Schlierkamp-Voosen, *Evol. Comput.* **1993**, *1*, 25.
[23] A. H. Wright, In Foundations of Genetic Algorithms; G. J. E. Rawlins, Ed.; Morgan Kaufmann: Burlington, MA **1990**, pp. 205–218.
[24] L. A. Rastrigin, Adaptive Control Systems. Nauka: Moscow **1974**.
[25] H. Mühlenbein, M. Schomisch, J. Born, *Parallel Comput.* **1991**, *17*, 619.
[26] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins: Struct. Funct. Bioinf.* **2006**, *65*, 712.
[27] T. E. Cheatham, III, P. Cieplak, P. A. Kollman, *J. Biomol. Struct. Dyn.* **1999**, *16*, 845.
[28] J. B. Klauda, S. L. Garrison, J. Jiang, G. Arora, S. I. Sandler, *J. Phys. Chem. A* **2004**, *108*, 107.

Received: 13 May 2014
Revised: 2 October 2014
Accepted: 12 October 2014
Published online on 00 Month 2014